Comparative genomics and epidemiology of the amphibian-killing fungus *Batrachochytrium dendrobatidis*

Rhys Alexander Farrer

2013

Submitted for PhD. examination

Department of Infectious Disease Epidemiology
Imperial College London
CID: 00622103

Supervisors:
Dr. Matthew C. Fisher, Imperial College
Dr. Trenton W. J. Garner, Institute of Zoology
Prof. François Balloux, University College London

Statement of originality

All ideas presented within this thesis were formulated by the candidate under the guidance of Matthew Fisher, Trent Garner, and François Balloux.

I declare that the work presented in this thesis is that of my own and that I have acknowledged the work of others where relevant.

The results presented in many sections of this thesis have been published in peer-reviewed journals and presented as posters or talks at meetings and conferences as described in the Appendices.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

The primary aim of this thesis was to study the population structure and epidemiology of the fungal pathogen of amphibians, *Batrachochytrium dendrobatidis* (*Bd*). I have addressed these questions by collecting and isolating *Bd* from multiple infected host species. I have then extracted and sequenced whole nuclear and mitochondrial genomes for 50 isolates of *Bd* using the ABI SOLiD 3 and Illumina HiSeq 2000 platforms.

The first aspect of the analysis was to tailor a new method for identifying variant sites amongst the isolates, as well as verifying the accuracy of the alignment and SNP-calling methods. Next, using a number of phylogenetic methods, I identified a population split into at least three deeply divergent lineages. Two of these lineages were found in multiple continents and are associated with known introductions by anthropogenic means. Isolates belonging to one clade, which we named the Global Panzootic Lineage (*Bd*GPL), have emerged across at least five continents and are associated with the onset of epizootics in all five continents we tested. Dating the divergence between *Bd*GPL isolates suggested a recent common ancestor in the 20^{th} Century, and that the widespread trade of amphibians is an important mechanism of transmission. In contrast, *Bd*GPL diverged from the other two lineages approximately 1000 years ago, clearly refuting a single emergence hypothesis. The two newly identified divergent lineages were the Cape lineage (*Bd*CAPE) that appeared to have originated from the Cape Province in South Africa and a Swiss lineage (*Bd*CH) comprised of a single isolate from a pond in Gamlikon, Switzerland.

The secondary aim of this thesis was to identify and compare virulence determinants and other genomic features responsible for known differences in phenotypes. Using a variety of statistical and computational methods, I identified compelling evidence for genetic recombination targeting virulence factors, selection of those and other virulence factors, and rapid changes in ploidy and aneuploidy amongst the isolates of all three lineages. These genomic features shed light on the emergence, patterns of global spread, and modes of evolution in the pathogen(s) responsible for contemporary disease-driven losses in amphibian biodiversity.

Finally, I discuss how these findings update our understanding of *Bd* and the importance for tracking and understanding the dynamics of other current emerging pathogens in an increasingly globalised habitat.

Contents

List of Tables 4
List of Figures 10
Acknowledgements 22
Acronyms 22

1 Introduction **24**
1.1 General introduction to the fungus *Batrachochytrium dendrobatidis* (*Bd*) . . 24
1.2 Host-pathogen interactions and pathogenicity of *Bd* 30
1.2.1 Pathogen specific features that impact outcome of infection 30
1.2.2 Host-specific features that impact the outcome of infection 34
1.2.3 Environmental factors that impact the outcome of infection 38
1.3 Mapping the contemporary global spread of *Bd* 39
1.4 Vectors and origins of the *Bd* panzootic 42
1.5 High-throughput/next-generation sequencing 45
1.5.1 Introduction to using NGS to study *Bd* 45
1.5.2 Quality control of NGS fungal datasets 47
1.5.3 Calling polymorphic sites and genome assembly *de novo* 48
1.6 Reconstruction of the evolutionary relationships between isolates 49
1.7 Population genetics approaches to understanding the evolutionary processes of *Bd* 52
1.8 Collecting, isolating and sequencing *Bd* 56
1.9 Aims and objectives 59
1.10 Funding 62

2 Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects **63**
2.1 Abstract 64

2.2 Introduction . . . 64
2.3 Methods . . . 67
2.4 Results . . . 69
2.5 Discussion . . . 75
2.6 Acknowledgements . . . 77
2.7 Author Contributions . . . 77
2.8 Competing Financial Interests . . . 77

3 Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent recombinant lineage 78
3.1 Abstract . . . 80
3.2 Introduction . . . 80
3.3 Materials and Methods . . . 81
3.3.1 Library preparation and sequencing . . . 81
3.3.2 Optimisation of alignments and SNP calling parameters . . . 82
3.3.3 Phylogenetic analysis of nuclear and mitochondrial genomes . . . 87
3.3.4 Patterns of mutation . . . 88
3.3.5 Recombination and break points in phylogeny . . . 89
3.3.6 Experimental assessment of host response to challenge by *Bd* isolates 91
3.3.7 Phenotypic and drug susceptibility testing . . . 92
3.3.8 Patterns and timing of *Bd* spread . . . 93
3.4 Results . . . 94
3.4.1 SNP discovery and phylogenetics . . . 94
3.4.2 Mutation patterns and copy number variation (CNV) . . . 100
3.4.3 Recombination and break points in phylogeny . . . 106
3.4.4 Differences between *Bd* lineages: host response, phenotypic and drug susceptibility . . . 115
3.4.5 Dating the emergence . . . 120
3.5 Discussion . . . 126
3.6 Acknowledgments . . . 127
3.7 Publicity, interviews and aftermath . . . 128
3.7.1 Tamera Jones, Natural Environment Research Council - Frog trade link to killer fungus revealed . . . 128
3.7.2 Michael Marshall, New Scientist - Frog-killer disease was born in trade 130
3.7.3 Richard Black, BBC - Killer frog fungus 'spread by trade' . . . 132

3.7.4 Rhys Farrer & Matthew Fisher, Frog log - Multiple emergences of genetically diverse amphibian infecting chytrids include a globalized hypervirulent recombinant lineage . . . 134
3.7.5 Faculty of 1000 (F1000) . . . 137

4 Chromosomal copy number variation and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity 139
4.1 Abstract . . . 140
4.2 Introduction . . . 140
4.3 Materials and Methods . . . 142
4.3.1 Library preparation and sequencing . . . 142
4.3.2 *in vitro* divergence of independent replicate lines of *Bd*CH . . . 144
4.3.3 Optimisation of alignments and SNP calling parameters . . . 144
4.3.4 Phylogenetic analysis of the nuclear genomes . . . 147
4.3.5 Identifying chromosome copy number variation (CCNV) . . . 147
4.3.6 Identifying gene groups and names . . . 148
4.3.7 Detecting recombination and hybridization amongst *Bd* isolates . . . 149
4.3.8 Detecting patterns of mutation and selection in *Bd* . . . 154
4.4 Results . . . 155
4.4.1 SNP discovery and phylogenetics . . . 155
4.4.2 Chromosomal Copy Number Variation (CCNV) . . . 160
4.4.3 Recombination among and/or within *Bd* . . . 167
4.4.4 Patterns of mutation and selection . . . 184
4.5 Discussion . . . 192
4.6 Acknowledgments . . . 197

5 Conclusions and Discussion 198
5.1 Introduction . . . 198
5.2 Epidemiology of *Bd* . . . 199
5.2.1 Spatial-temporal distribution of each of the *Bd* lineages . . . 199
5.2.2 The origin(s) of *Bd* . . . 201
5.3 *Bd*'s pathogenicity determinants and mechanisms of infection . . . 204
5.3.1 Evolution of virulence . . . 204
5.3.2 Antimicrobial peptide (AMP) and drug resistance . . . 208
5.3.3 Aneuploidy and its association with virulence in *Bd* . . . 210

5.4 The life cycle of *Bd* 213

Bibliography **216**

Appendix A: Culturing *Bd* 237

Appendix B: Sequencing and cryo-preserving *Bd* 238

Appendix C: Published articles from chapters presented in this thesis 241

List of Tables

1.1 Fifty isolates used for whole-genome sequencing during this study. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger L *et al.*, 1998]. Amphibian hosts are *Amietia angolensis* (Angola River Frog), *Afrixalus enseticola* (Ethiopian Banana Frog), *Afrana fuscigula* (Cape River Frog), *Alytes muletensis* (Mallorcan Midwife Toad), *Alytes obstetricans* (Common Midwife Toad), *Amietia vertebralis* (Ice Frog), *Bufo boreas* (Western Toad), *Hadromophryne natalensis* (Natal Ghost Frog), *Litoria caerulea* (Green Tree Frog), *Lissotriton vulgaris* (Smooth Newt), *Leptodactylus fallax* (Mountain Chicken Frog), *Lithobates catesbeianus* (American Bullfrog), *Mesotriton alpestris* (Alpine newt), *Phyllomedusa lemur* (Lemur Leaf Frog), *Rana perezi* (Iberian Green Frog). AC (Andrew Cunningham), CM = Claude Miaud, CW = Che Weldon, JEL = Joyce Longcore, LB = Lee Berger, MF = Matthew Fisher, RF = Rhys Farrer, TG = Trent Garner. 59

3.1 Samples used for SOLiD sequencing and details of alignments. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger L, 1998]. The number of reads (millions) aligning to the *Bd* JEL423 genome assembly and the corresponding depth of coverage. Amphibian hosts include *Afrana fuscigula* (Cape River Frog), *Alytes muletensis* (Mallorcan midwife toad), *Alytes obstetricans* (Common Midwife Toad), *Bufo boreas* (Western Toad), *Hadromophryne natalensis* (Natal Ghost Frog), *Litoria caerulea* (Green Tree Frog), *Lissotriton vulgaris* (Smooth Newt), *Leptodactylus fallax* (Mountain Chicken Frog), *Lithobates catesbeianus* (American Bullfrog), *Phyllomedusa lemur* (Lemur Leaf Frog). Collectors are AC (Andrew Cunningham), JEL (Joyce Longcore), LB (Lee Berger), MF (Matthew Fisher), TG (Trent Garner), CW (Che Weldon). CdsB = Cocó de sa Bova, TdF = Torrent des Ferrerets 83

3.2 Percent of shared entirely covered in all (ECA) homozygous SNPs (top) and percent of shared ECA heterozygous positions (below). 97

3.3 Alignments to *Bd* JEL423 mitochondrial sequence had an average depth of 51X deep from reads that were >99.2% uniquely aligned to the mitochondrial genome. We identified ECA homozygous SNPs in 9/20 isolates using BiSCaP v0.1. In total, 306 non-redundant polymorphic sites were identified in the mitochondrial genome with our panel of 20 isolates. 99

3.4 Twenty-six recombination breakpoints across the genome were identified using GARD analysis. The *Bd* JEL423 supercontig (SC) number and position of each breakpoints is given in the first column. The left hand side (LHS) and right hand side (RHS) raw probabilities are shown, and adjusted probabilities after Bonferroni correction in parethesis. The overlying feature type and the genes top Blastp hit using the Blast2Go server [Conesa A *et al.* 2005] is shown in the final column. *Significant breakpoint. 117

3.5 Itraconazole treatment of twelve isolates. MC_{50} (μg/ml) values from day 15 for 12 individual isolates spanning the three lineages, *Bd*GPL, *Bd*CAPE and *Bd*CH 122

3.6 Time to most recent common ancestor (tMRCA) for the 16 *Bd*GPL isolates for the fifteen significant breakpoint segments. Median dates for the emergence of the global hypervirulent lineage range from 35-257 years before present (ybp). Lower and upper 95% confidence intervals (CI) for each gene tree are also shown. 125

4.1 Samples used and details of alignments. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger *et al.*, 1998]. TdF = Torrent des Ferrerets, AP = Affluente Pisharoni. Passage numbers (PN) are best approximations from records prior to DNA extractions in January and May 2011. The sequenced depth and aligned depth were calculated from the number of nucleotides in all or aligned reads respectively and divided by 24Mb (the length of the *Bd* JEL423 genome assembly). Amphibian hosts include *Afrixalus enseticola* (Ethiopian Banana frog), *Alytes muletensis* (Mallorcan Midwife Toad), *Alytes obstetricans* (Common Midwife Toad), *Amietia angolensis* (Angola River Frog), *Amietia fuscigula* (Cape River Frog), *Amietia vertebralis* (Ice Frog), *Discoglossus sardus* (Tyrrhenian Painted Frog), *Epidalea calamita* (Natterjack Toad), *Leptopelis sp.* (Big eyed Tree Frog), *Lithobates catesbeianus* (American Bullfrog), *Phyllomedusa lemur* (Lemur Leaf Frog). CM = Claude Miaud, DG = David Gower, JEL = Joyce Longcore, MF = Matthew Fisher, PH = Phineas Hamilton, PM = Peter Minting, RF = Rhys Farrer, TG = Trent Garner. 143

4.2 Polymorphisms and reference bases were identified in 22 *Bd* nuclear genomes relative to *Bd* JEL423 using BiSCaP v0.11 with default settings. Shown are tallies of each category of loci found in each separate isolate. 156

4.3 The percent of ECVA polymorphic sites shared between each of the 22 isolates. Greater overlap (≥30%) highlighted in red. (**A**) The overlap of homozygous SNPs varied between 3% and 97% (**B**) The overlap of heterozygous positions varied between 3% and 75%. 159

4.4 The percent of uniquely mapped reads over each type of category of loci. Biallelic heterozygous positions had a reduced percent of uniquely mapped reads in the 2 divergent lineages of *Bd*, which may result from structural variants. 160

4.5 The two most common allele frequencies over each base of each supercontig were determined by percent of read agreement with the reference base. Using 1000 Bootstrap replicates of these values, we recorded how often 47-53% reads agreeing with an allele predominated over 30-36% or 63-69% reads agreeing with an allele. Supercontigs with >95% of replicates with predominantly bi-alleles have support for an even-ploidy. Equally, supercontigs with <5% of replicates with predominantly bi-alleles have support for an odd-ploidy. Supercontigs with 5% <x <95% bootstrap support do not have support either an even or odd ploidy (and are therefore ambiguous). 161
4.6 Pairwise comparisons for shared phased heterozygous positions. (top) Total numbers of matching phased heterozygous positions in same phase (kilobase) (below) Percent of matching phase positions from the total number of shared phased positions. 172
4.7 Phased heterozygous positions demonstrating crossovers were identified between every isolate. (top) Total numbers of crossovers identified. (below) Percent of crossovers from the total number of shared phased positions. . . . 173
4.8 Haplotypes from isolates belonging to each of the separate lineages were tested for disequilibrium using the index of association (I_A), $\overline{r}_d$ and the 4-gamete test. To check differences between lineages were not resulting from different numbers of isolates, 2 subsets were made from *Bd*GPL. Subset 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. Over 30% of the *Bd*GPL haplotypes from any of the subsets were in significant disequilibrium, whilst only 11% of the haplotypes in *Bd*CH and 16% of the haplotypes in *Bd*CAPE were in disequilibrium, suggesting these populations are recombining more than the clonal *Bd*GPL. 175
4.9 Genes were tested for enrichment in non-redundant (NR; at unique loci) crossovers (XO) and NR XO / NR phased positions (NRPP), comparing each to the values obtained for all genes using Hypergeometric tests (HgT) and *t*-tests respectively. For *t*-tests, all genes with <2 NRPP (the minimum required for a crossover) were excluded. Although both CRN-like and uncharacterised (secreted) were significantly (SIG.) enriched for XOs at unique loci (non-redundant), only CRN-like (between lineages) were SIG. enriched for XO/NRPP. *p*-values for HgT and *t*-tests are shown as P≥0.01 (-), P<0.01 (*), P<0.001 (**) and P<0.0001 (***). NA=not applicable 180

4.10 Haplotypes over coding sequence that failed the four-gamete test were predominantly from coding-regions. Haplotypes overlapping a number of genes were included in the counts for each gene (385 extra counts to total number of haplotypes). After accounting for these extra counts, an additional 1,162 haplotypes were still found to come from coding regions compared with those from intergenic or intron regions. However, no gene group had a clear enrichment for haplotypes that failed the four-gamete test. *Bd*GPL subset (s.s.) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. 181
4.11 The numbers of variable sites per locus amongst haplotypes, demonstrating all lineages to be as likely to have arisen from out-crossing. *Bd*GPL subset (ss) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. 184
4.12 Only five presence absence (PA) polymorphisms relative to *Bd*GPL JEL423 were identified amongst *Bd*CAPE and *Bd*CH isolates (read depth=0), whilst none were identified amongst *Bd*GPL isolates. 184
4.13 Homozygous polymorphisms were found in the coding and non-coding regions of the *Bd* nuclear genomes. SNPs in the coding regions are also tallied according to whether they result in a synonymous (Syn.) or non-synonymous (Non-syn.) amino acid change. The total numbers of each variant-type are followed by their numbers per kilobase of genomic region in parentheses. . . 185
4.14 Heterozygous polymorphisms were found in the coding and non-coding regions of the *Bd* nuclear genomes. For heterozygous positions, the affect on the transcript (synonymous, Syn. or non-synonymous, Non-syn.) was determined using the alternative allele. Where two alternative alleles to the reference sequence were found (infrequently), the first present within the VCF was chosen. The total numbers of each variant-type are followed by their numbers per kilobase of genomic region in parentheses. 186

4.15 Gene categories were tested for enrichment in non-redundant (NR; at unique loci) homozygous SNPs, NR synonymous (Syn.) and non-synonymous (Non-syn.) amino acid changes, NR heterozygous (het.) positions and NR phased positions (PP). For each; the total number (sum), average ($\overline{x}$), and standard deviation (S.D.) were calculated for all isolates, and lineage specific isolates. Hypergeometric tests (HgT) were used to identify significant (SIG.) enrichment for variants, where P≥0.01 (-), P<0.01 (*), P<0.001 (**) and P<0.0001 (***). NA=not applicable 188
4.16 Number and category of genes with $\omega \geq 1$ in each of the three lineages. NS = non-secreted, S = secreted. Hypergeometric tests were used to identify significant enrichment for gene categories. Significant enrichment for Crinkler-like genes and uncharacterised (secreted) genes with $\omega \geq 1$ were found in both *Bd*CAPE and *Bd*CH. NA=not applicable 192
4.17 Number and category of genes with $\omega \geq 1$ and 2D'≥11.4076 (1% significance after Bonferroni correction) and 11.4076>2D'>8.1887 (5% significance after Bonferroni correction) in each of the three lineages. NS = non-secreted, S = secreted. Hypergeometric tests (HgT) were used to identify significant enrichment from genes with $\omega \geq 1$. NA=not applicable 192

List of Figures

1.1 Life cycle of *Bd* from aquatic motile zoospore to embedded zoosporangia. Taken from [Rosenblum EB *et al.*, 2010]. 28
1.2 A snapshot of the global *Bd* database taken from *Bd*-maps (7.11.12) indicating *Bd*-positive samples per country. 40
1.3 Expected genotypic frequencies under Hardy-Weinberg Equilibrium (HWE) assumptions for an allele at frequency q. Here, the simplest scenario of a diploid with two alleles at both loci is shown. However, HWE can be expanded to greater ploidy and/or a larger number of alleles. 54
1.4 Collecting and isolating *Bd*. (**A**) An isolation kit, a 1 litre bottle of mT-GhL, and *Bd* growing in liquid and solid media. (**B**) Ingredients and scales for preparing media. (**C**) The hood where isolates can be handled without contamination. (**D**) Collecting *A. obstetricians* in Switzerland, 2010 (**E**) All equipment and boots were treated with itraconazole to prevent *Bd* spread between sites. (**F**) The *Bd* isolates (as of 29/07/2012). (**G**) The Waltensberg, CH site where *Bd* BEW2 was collected. 60
1.5 Locations that *Bd* isolates were taken from for whole-genome sequencing. Black squares signify isolates chosen for the first round of SOLiD sequencing, and red squares signify isolates chosen for Illumina sequencing. Some locations had multiple isolates taken, and some (such as JEL423 from Panama) were sequenced by both platforms. 61

2.1 A flow diagram showing the steps to verify false discovery rate and call polymorphic sites using cFDR and BiSCaP. Specifically, the reference (genome) sequence has random nucleotides changed. Next, reads are aligned to the modified genome and SNPs called. Finally, the cFDR script can be used to distinguish true positive and false positive SNPs. Given poor results, alignment and SNP-calling parameters as well as read pre- or post-processing methods can then be adjusted. 70

2.2 False discovery rates for variants were ascertained using cFDR for three fungal NGS datasets. (**A**) Dataset-specific error rates were identified for both homozygous SNPs and heterozygous positions after alignment with BWA and calling SNPs using SAM/BCFtools with default settings, which persisted after trimming to 30mers, aligning random 10X deep subsets of aligned reads, and only considering SNPs that fell over CDS regions thereby reducing genome size as a factor. (**B**) Combinations of alignments and SNP-calling methods resulted in different accuracies from the *Bd* JEL423 30mer NGS dataset. Experimenting with a variety of methods can therefore reveal the most suitable method for a given dataset based on these metrics of accuracy. Sam/Bcftools only takes the strongest non-reference allele so was not included in the assessment of heterozygous accuracy. Parameters include the percent cut-off for inclusion as a SNP or heterozygous base and minimum depth (MD). 72

2.3 Erroneous base calls (homozygous SNPs) from BiSCaP and percent cutoff methods were compared for proximity and depth of read coverage using full dataset 30mer *Bd*, *Sc* and *Pt* reads. (**A**) False negatives were predominantly more closely associated for all 3 datasets (SNPs called with BiSCaP). False negatives were almost entirely caused by lack of coverage in each of these datasets demonstrating the most divergent part of each of the genomes is the most poorly resolved. (**B**) Homozygous errors were more frequently called over lower depth regions using strict cut-off methods than with BiSCaP. . . . 74

3.1 Single Nucleotide Polymorphism (SNP) calling was optimized using False Discovery Rate analysis. (**A**) Over 100 combinations of alignment and SNP-calling tools and parameters including minimum depth cutoffs (MD) and read trim lengths (RL) had their accuracy at calling simulated SNPs assessed. BWA and a binomial SNP-calling method (BiSCaP) using a minimum read depth (MRD) of four was chosen for this study (shown here as a red triangle) (**B**) Simulated reads were used to assess accuracy at calling simulated heterozygous positions (**C**) The number of true positives and false positives found using BiSCaP over each depth of coverage, the MRD used in this study is indicated by the dotted line. 86

3.2 Six random subsets of *Bd* isolate JEL423 30mers (ranging from 1X to 11X deep) were re-aligned back to the JEL423 genome sequence. (**Left**) Breadth of coverage increased over whole genome including the coding regions (CDS) as subset size increased. At 5X deep coverage (all our isolates had 5X or greater) >80% of the genome and >90% of the genes. (**Right**) Depth of coverage over CDS and the genome as a whole increased as subsets of data increased. Using a 5X deep subset, CDS were covered by an average of 5.23 reads per base. . 87

3.3 The total numbers of polymorphic nucleotide positions found in 20 nuclear genomes of *Bd*. (**A**) Sixteen isolates had between 0.351Kb and 2.883Kb homozygous SNPs and four isolates had >10Kb homozygous SNPs. A greater number of heterozygous positions (>5.87Kb) were found for all 20 genomes. (**B**) Polymorphic sites that were covered by ≥ 4 reads in all 20 nuclear genomes (ECA) closely resembled the total number of polymorphic sites, which were used for phylogenetic analysis. 96

3.4 Tree of the nuclear genomes made using the UPGMA algorithm in PAUP. The three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH) identified from the mitochondrial genome analysis also occur in the nuclear genome comparisons. * Signifies the JEL423 genome isolate 98

3.5 ECA polymorphic sites caused a similar excess of transitions in all 20 isolates of *Bd*. (**A**) Average frequency of transition (Ts) and transversion (Tv) rates in the 20 genomes relative to the reference sequence *Bd* JEL423 was calculated using ECA SNPs. For all 20 isolates there was an excess of transitions within a very narrow standard deviations (shown by error bars). (**B**) The Ts/Tv ratio was between 7-10, with little or no differences found between lineages. . 102

3.6 Homozygous SNPs identified within the CDS regions were categorised by their effect on the transcripts. (**A**) Approximately twice as many synonymous changes were identified in total as non-synonymous changes. (**B**) Heterozygous positions were also categorised based on their affect to the transcript, again showing similar excess of heterozygous positions. 103
3.7 Genes undergoing diversifying selection were predominantly metabolic genes. Genes undergoing diversifying selection were identified by (**A**) DoS and (**B and C**) dN/dS >1 for isolates *Bd*CAPE CCB1 and *Bd*CH 0739 respectively. A small number of genes were identified (4 DoS and 110 dN/dS >1) that were categorised by GO terms. Many of the GO terms for divergent genes in *Bd*CAPE CCB1 and *Bd*CH 0739 were involved in metabolism (>50% and >40% respectively). 104
3.8 dN, dS and ω values were calculated for all *Bd*CAPE genes from alignments to *Bd*GPL JEL423. Predicted proteases (blue), secreted genes (red) and secreted proteases (purple) and all other genes (grey) show no obvious pattern in terms of selection (where ω>1 = positive or relaxed selection and ω<1 = purifying selection. 105
3.9 The genome assembly of *Bd* JEL423 and ECA positions had a similar AT content bias. (**A**) The composition of bases in the JEL423 genome by genomic features. (**B**) The composition of bases covered by ≥4 reads in all 20 samples (ECA). While composition of ECA bases showed a similar AT content bias, they covered a greater percent of CDS than intergenic or introns. 107
3.10 The Average Read Depth (ARD) over genes. Boxplots of ARD over all 8795 genes for each of the 20 isolates with outliers omitted. The number of reads sequenced, and therefore aligned, largely affected the distributions. Lineage specific bias was not identified. 108
3.11 Most of the aligned reads mapped to unique locations in both the nuclear and mitochondrial genomes. (**A**) The nuclear genome had a lower proportion of uniquely mapped reads compared to the mitochondrial genome. Reads that aligned to the nuclear genome and specified homozygous SNPs or heterozygous sites for all global lineage isolates were more comparatively (>80%) aligned to unique locations. (**B**) The lowest proportion of uniquely mapped reads was found over heterozygous positions in the nuclear genomes for the isolates comprising the two divergent lineages. 109

3.12 Total number of heterozygous positions in each genome belonging to *Bd*GPL was not affected by different dates since isolation. (**A**) A small decrease in the number of heterozygous positions was found since the date of isolation from the host. (**B**) However, most of the trend can be explained by the recently isolated divergent isolates. 110
3.13 Phylogenetic analysis of the 20 resequenced *Bd* mitochondrial genomes demonstrates three divergent lineages. The locations of the isolates belonging to the different lineages are shown using the same colors as in the phylogeny. Each genome is represented next to the right of the phylogeny. A non-overlapping sliding window of SNPs minus heterozygous positions across the genome illustrates regions where heterozygosity predominates (blue) and where homozygosity predominates (red), illustrating the hallmark of loss-of-heterozygosity in the pan-global *Bd*GPL lineage. The block below the 20 genomes denotes the supercontigs with black lines and the GARD recombination breakpoints are shown in red dotted lines. The star signifies the reference genome JEL423, crosses represent isolates that have been recovered from epizootics. 112
3.14 A range of sliding non-overlapping window lengths revealed the highly uneven distribution of homozygous and heterozygosity. Shown here are window lengths 714nt (left) and 2814nt (right) corresponding to half and double that used in Fig. 3.13) 113
3.15 Isolates from the *Bd*CAPE lineage were aligned against the consensus sequence for the *Bd*CAPE lineage. (**A**) A non-overlapping sliding window of SNPs minus heterozygous positions does not reveal hallmarks of loss-of-heterozygosity. (**B**) Details of the alignments show that there are similar levels of SNPs within the *Bd*CAPE linage as there are within the *Bd*GPL. These alignments also achieve a greater depth from which to call those mutations. . 114

3.16 Two possible mechanisms of achieving the uneven distribution of heterozygous and homozygous SNPs throughout the *Bd* genome. Each black and white bar represents a haplotype identified in a parental isolate, and the charts in the middle represent the plots illustrating regions where heterozygosity predominates (blue) and where homozygosity predominates (red for homozygous SNPs and empty for homozygous identical to reference). On the left of the diagram, meiosis generates recombinant haploid genomes that then are united via syngamy into new diploid offspring with patchy heterozygosity. Meiosis involving fusion of diploid gametes could result in similar patterns if chromosomal segregation remains independent. On the right of the diagram, mitotic gene conversion generates patches of homozygous sites via homologous DNA repair in diploid progeny. 116
3.17 Maximum clade credibility trees for the fifteen significant recombination breakpoint segments of the nuclear genome. The four isolates from Valencia, Spain are highlighted in red in each of the phylogenies showing that they form a monophyletic group in ten of the fifteen trees. A representation of the genome is given below the trees showing the location of supercontigs (black lines) and the corresponding numbers for recombination breakpoints (red dotted lines) detected by the GARD method. 118
3.18 Proportion of animals infected at time of death or by the end of the experiment. (**A**) within the control treatment, *Bd*GPL and *Bd*CAPE, and (**B**) within the six *Bd*GPL isolates. 121
3.19 Kaplan-Meier survival curves illustrating post-metamorphic survival of animals exposed to *Bd*GPL isolates (red) and *Bd*CAPE lineage isolates (blue). *Bd*GPL isolates were significantly different from both *Bd*CAPE lineage isolates and the negative controls. 122
3.20 Diameter of sporangia and hyphae (μm) 20 days post initial culture (dpi) and 10dpi respectively. (Top) Measurements were taken for the 12 isolates individually. (Below) Measurements of isolates grouped according to their three lineages. Sporangia sizes between the three lineages were significantly different. 123
3.21 MC_{50} (μg/ml) for Itraconazole from isolates grouped into three lineages. Readings were taken from day 15. No significant difference was observed between the three lineages. 124

3.22 The Panamanian golden frog (*Atelopus zeteki*) is a critically endangered toad which is endemic to Panama. 128
3.23 Moment of truth for an Atelopus frog in Ecuador (Image: Joel Sartore/ National Geographic/Getty) 131
3.24 The scarlet frog of Venezuela may have been a victim of chytridiomycosis . . 132
3.25 A different strain of the chytrid fungus was recently found on Japanese giant salamanders. 133
3.26 Midwive mass mortalities (*Alytes obstetricans*) in the French Pyrenees, 2010. Photo: Matt Fisher. 135
3.27 Moribund and dead Midwife metamorphs suffering from terminal chytridiomycosis. Pyrenees 2010. Photo: Matt Fisher. 136

4.1 The previous SOLiD reads [Farrer RA *et al.*, 2011] and the new Illumina paired end reads of *Bd* isolate JEL423 were aligned to a modified JEL423 reference sequence. Additionally, simulated reads from a heterozygous reference sequence were made to the depths of the Illumina and SOLiD datasets. Single Nucleotide Polymorphisms (SNPs) and heterozygous positions were then called and the False Discovery Rates (FDR) ascertained. The SNP-caller BiSCaP v0.11 was tested using default settings, and SAM/BCFTools with VCFUtils was tested for its ability to call SNPs using its default settings. SNPs were also filtered for those found without first modifying the reference sequence (f=filtered). (**A**) 1nt/Kb simulated SNPs or heterozygous positions (12,458 in total) within the coding region (CDS) (**B**) 1nt/100nt simulated SNPs or heterozygous positions (124,588 in total) within the CDS region. The new Illumina data was able to recover >95% of true positive SNPs and >80% true positive heterozygous positions using BiSCaP v0.11, outperforming the previous lower-depth SOLiD sequences. 146
4.2 Illustrations of how phased haplotypes were extracted from the alignment. Heterozygous positions that did not pass the minimum depth or percent phased cut-offs, along with examples of pairwise crossovers and outcomes for a four-gamete test between three isolates. 150

4.3 Phylogenetic trees were made using the UPGMA algorithm in PAUP from ECVA polymorphic positions identified in the nuclear genomes demonstrating three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH shown in red blue and green respectively). (**A**) A tree from 275Kb ECVA polymorphic positions identified from the Illumina sequencing (**B**) A tree from 36Kb ECVA polymorphic positions from Illumina and SOLiD sequencing. 157
4.4 Phylogenetic trees were made using the UPGMA algorithm in PAUP from ECVA homozygous positions (either reference or SNP) identified in the nuclear genomes demonstrating three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH shown in red blue and green respectively). (**A**) A tree from 218Kb ECVA polymorphic positions identified from the Illumina sequencing (**B**) A tree from 8Kb ECVA polymorphic positions from Illumina and SOLiD sequencing. 158
4.5 Read depth across 22 genomes was normalised by total alignment depth and plotted against location in the genome using a 10Kb long non-overlapping sliding window (red lines signify borders of supercontigs). Base ploidy levels (shown at the beginning of each isolates genome) were determined using allele frequencies for supercontig 1 and shown at the start of each plot. Intra-chromosome read depth is largely consistent amongst the isolates, except over supercontig 14 due to a long stretch of rDNA. Shifts in read-depth between chromosomes demonstrate variation in chromosome copy number. 162
4.6 *t-tests* for the mean depth of read coverage across each chromosome against chromosome 1 revealed significant p-values demonstrating uneven chromosome copy number. Stringent cut-offs for ploidy differences relative to the largest chromosome (chr. 1) of each isolate were chosen: $p<5^{-10}$. Chromosomes with p-values below this cut-off, with a mean depth that is greater than chromosome 1 are highlighted in blue, while those with a mean depth lower than chr. 1 are shown in green. All 308 chromosomal p-values (excluding chr. 1) are shown in the bottom plot ordered from smallest to greatest. 163

4.7 The percent of reads specifying the two most frequent alleles per chromosome using 2 representative isolates from each lineage of *Bd*. The most common allele is shown in black and the second most common allele is shown in blue. Bins were used to summarise the expected peaks for odd, even and odd numbers of chromosomes and shown in red (lines show bin value cut-offs and dots show values). Individual chromosomes with a predominantly bi-allelic value are shown with a blue border, and those with a predominant tri-allelic value are shown with a black border. 164
4.8 CCNV in the *Bd* nuclear genomes was identified using allele-frequencies and mean read depths across each chromosome normalised to the alignment depth for each isolate. Many *Bd*GPL isolates can be seen to include more copies of chromosome 2 and 3, while the 3 *Bd*CH and 3 of the 5 *Bd*CAPE isolates have fewer copies of chromosome 9 and 11. Fewer copies of chromosome 9, 11 and 16 appear to be found in many of the isolates. 166
4.9 Chromosome copy number variation was identified across the three *Bd*CH isolates (ACON and its progenitors CON2A and APEP) following 40 generations in culture with or without the addition of anti-microbial peptides (AMP), respectively. Read depth is normalised to total alignment depth. A tally of all loci (per kilobase) with between 25-75% reads agreeing with the reference nucleotide are shown below, and summarised by the most common allele (black line), the second most common allele (blue line), and bins between 32-34, 49-51 and 65-67% (red circles). ACON is putatively triploid across the largest six supercontigs, whereas CON2A has lost a copy of supercontig IV and gained a copy of supercontigs V. APEP has gained a copy of supercontigs V. 167
4.10 Sliding non-overlapping windows of 10Kb across the 22 *Bd* nuclear genomes showing homozygous SNPs minus heterozygous positions. Predominance of homozygous SNPs is shown in red and predominance of heterozygous positions in shown in blue. Windows across *Bd*GPL isolates demonstrate highly uneven distribution of heterozygosity attributed to recombination whereas polymorphisms are more evenly spread across the genomes of *Bd*CAPE and *Bd*CH isolates. 168

4.11 Heterozygous positions had their phase determined using overlapping reads. Reads from each isolate are shown as a separate black line on the graphs. Only bi-allelic polymorphisms were compared for phasing. Predominantly, overlapping reads agreed with a single bi-allelic phase. (**A**) All reads over all phased positions. A 90% cut-off was used to filter ambiguous phased positions or those with an excess of mismatches as shown by the red line. (**B**) Positions that agreed 90-100% for a single phase are shown as a percent of all reads. (**C**) the total number of phased positions (including phased SNPs) and the percent of uniquely mapped reads over those positions. 170
4.12 Lengths of haplotypes (in nucleotides) that included at least two alleles per loci in every isolate of a given group, and were therefore suitable for population genetic analysis. *Bd*GPL subset (s.s.) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. 175
4.13 Intra-lineage heterozygote's, the percent of heterozygote's that were phased, the number of phased positions that demonstrated a crossover (XO) and the $\overline{r}_d$ were plotted using non-overlapping windows across the genome (length 10Kb). Both phased positions and crossovers were found across each of the chromosomes in each of the lineages of *Bd*, suggesting recombination is not confined to small or large chromosomes, or the ends of any given chromosome. The same is seen with $\overline{r}_d$ values. 176
4.14 The total numbers of crossovers found within genes demonstrated variation between gene families. All crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size. 178
4.15 Crossovers at unique locations (non-redundant, NR) occurred differentially across gene families. NR crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size. 179
4.16 Boxplots for eight non-overlapping gene categories comprising every gene were compared for numbers of crossovers per phased positions (PP) within each gene ($\geq$2PP) for all isolates. 182

4.17 The fixation index (F_{ST}) was calculated for each pairwise lineage across window lengths of 1.4Kb (**A**) and 10Kb (**B**). All three lineages are differentiated from one another across each chromosome, with some intra-chromosomal variation. Notably, the stretch of rDNA located at the start of chromosome 14 appears to have a reduced genetic distance between each of the three lineages of *Bd*. *Bd*GPL subset (ss) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. 183
4.18 The ratio of non-synonymous mutation per non-synonymous site (dN) vs synonymous mutation per synonymous site (dS) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the *Bd*GPL. The line designates the ω value (dN/dS), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation. 190
4.19 The ratio of non-synonymous mutation per non-synonymous site (dN) vs synonymous mutation per synonymous site (dS) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the three lineages. The lines designate the ω value (dN/dS), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation. Summaries of ω values for all genes in each of the three lineages are shown in the final three plots. 191
4.20 A Venn diagram showing the number of genes undergoing positive selection according to the Branch site model (BSM), where genes had 2D'>8.1887. The nine genes identified in all three lineages were four uncharacterised (secreted) with transcript ID's 05565, 02533, 00379, 06783 and five uncharacterised (non-secreted) with transcript ID's 03962, 07794, 05877, 02935, 08088. 193

Acknowledgements

I would like to firstly thank my primary supervisor Dr. Matthew Fisher, who not only devised the premise for this project, but also gave a huge amount of help in terms of hypothesising how the results could be obtained, interpreted and which directions to take the project in. I would like to thank my 2^{nd} supervisor Dr. Trenton Garner for much of the fieldwork in South Africa and Europe, as well as the *in vivo Bd* virulence experiments, both of which were essential for the work carried out in this thesis. My 3^{rd} supervisor Prof. François Balloux was invaluable for discussions on the genetic analysis and methods used. Without these isolates, phenotypes, and ideas this study would not have been as comprehensive, and indeed, many of the discoveries may not have been made. In addition, I would like to thank Dr. Ursina Tobler for help with every aspect of the fieldwork in Switzerland and Dr. Lucy Weinert for work on dating the emergence of *Bd*. On a day-to-day basis, Dr. Daniel Henk gave a huge amount of advice, support and insight, particularly for the phylogenetic aspects of this project. Others who I had the opportunity to work and discuss much of this work with were Julia Halder, Frances Clare, Olivia Edwards, Rocio Rodriguez, Peter Minting, Jon Bielby, Freya Smith, and many others. Finally, I would like to also thank friends, family and girlfriend Daisy for their support during this project.

Acronyms

AMP: Anti-microbial peptides
ARD: Average read depth
Bd: *Batrachochytrium dendrobatidis*
BEAST: Bayesian Evolutionary Analysis Sampling Trees
BiSCaP: Binomial SNP Caller from Pileup
BLAST: Basic local alignment search tool
BWA: Burrows-Wheeler Aligner
CBM: Carbohydrate binding module
CDS: Coding Regions
CNV: Copy Number Variation
CRN: Crinkler
ECA: Entirely Covered in All
ECVA: Entirely Covered and Verified in All
EID: Emerging Infectious Disease
EPH: Endemic Pathogen Hypothesis

FDR: False Discovery Rate
GATK: Genome Analysis Toolkit
GO: Gene Ontologies
GPL: Global Panzootic Lineage
HGT: Horizontal Gene Transfer
HWE: Hardy-Weinberg equilibrium
I_A: Index of Association
Ig: Immunoglobulin
Indel: Insertion/deletion
ITS: Internal transcribed spacer
LD: Linkage Disequilibrium
LE: Linkage Equilibrium
MCMC: Markov chain Monte Carlo
MHC: Major Histocompatibility Complex
ML: Maximum Likelihood
MLST: Multi-locus sequence types
MP: Maximum Parsimony
MRCA: Most Recent Common Ancestor
mTGhL: Media Tryptone Gelatin hydrolysate Lactose
NGS: Next-generation sequencing
NJ: Neighbour joining
NPH: Novel Pathogen Hypothesis
PCA: Principle Component Analysis
PCR: Polymerase Chain Reaction
RACE: Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity
SHRiMP: Short-Read Mapping Package
SNP: Single Nucleotide Polymorphism
Ti/Tv: transition transversion
UPGMA: Unweighted pair group method with Arithmetic Mean
VCF: Variant Call Format

Chapter 1

Introduction

1.1 General introduction to the fungus *Batrachochytrium dendrobatidis* (*Bd*)

Emerging and re-emerging infectious diseases (EIDs) are infections that have newly appeared in a population, or have previously existed but are rapidly increasing in incidence or geographic range [Morse SS *et al.*, 1995] and present a growing threat to natural populations of animal and plant species [Fisher MC *et al.*, 2012]. EIDs of free living animals (wildlife EIDs) have been classified into three major groups: 1) associated with "spill-over" from domestic animals to wildlife populations, 2) related directly to human intervention and 3) those with no overt human or domestic animal involvement [Daszak P *et al.*, 2000]. Each of these categories poses a substantial risk to the conservation of global biodiversity [Daszak P *et al.*, 2000].

The kingdom Fungi contains an increasing proportion of the species that cause EIDs [Fisher MC *et al.*, 2012], which overwhelmingly belong to the subkingdom Dikarya (consisting of the phyla Ascomycota and Basidiomycota)[1]. For example, the previously undescribed Ascomycota *Geomyces destructans* is overwhelming little brown bat populations across the US [Gargas A *et al.*, 2009; Frick WF *et al.*, 2010]. The Basidiomycota *Puccinia striiformis* (the yellow rust fungus) pathotypes PstS1 and 2 are evolving to overcome cultivar resistance

[1]However, the overwhelming majority of described fungal species are members of the subkingdom Dikarya [Hibbett DS *et al.*, 2007].

resulting in escalating losses for agribusiness [Hovmøller MS *et al.*, 2010], and the Basidiomycota *Cryptococcus gattii* is expanding its range into non-endemic environments with a consequential increase in incidence in humans [Fraser JA *et al.*, 2005; Byrnes EJ 3^{rd} *et al.*, 2010; Simwami SP *et al.*, 2011]. However, it is a newly described species belonging to the basal phylum Chytridiomycota (chytrid) that is having the greatest impact of any wildlife EID on its hosts.

Chytrids are globally ubiquitous and comprise approximately one thousand described species of fungi [James TY *et al.*, 2006] and appear to be the main class of fungi that dominate at high altitudes [Freeman KR *et al.*, 2009]. Chytrids are broadly split into two classes: the Chytridiomycetes (which reproduce by isogamy - gametes of similar morphology) and the Monoblepharidomycetes (which reproduce by oogamy - gametes of different morphology) [Hibbett DS *et al.*, 2007]. Both classes include species that reproduce asexually or sexually, resulting in the production of single flagellated motile spores (zoospores). Chytrids live predominantly in aquatic environments and acquire their energy from non-living organic matter (saprotrophic nutrition) by breaking down macromolecules such as chitin and cellulose. Other species of chytrid parasitize phyto- and zooplankton, fungi, invertebrates, and plants [Gleason FH *et al.*, 2008], for instance, the Chytridiomycete *Synchytrium endobioticum* (order Chytridiales), which causes potato wart disease, and *Rhizophydium graminis* (order Rhizophydiales[2]), which is a parasite of wheat roots [Barr DJS, 1973].

The Chytridiomycete *Batrachochytrium dendrobatidis* (*Bd*) - order Rhizophydiales - was discovered in 1997 and named in 1999 as an infection causing the rapidly progressing and often-fatal cutaneous[3] disease, chytridiomycosis, in anuran (frog-like) and caudate (salamander-like) amphibians [Berger L *et al.*, 1998; Longcore JE *et al.*, 1999]. *Bd* is the only known chytrid pathogen of vertebrates [Berger L *et al.*, 1998, Pessier AP *et al.*, 1999] and was named after the host genus of frogs it was first identified on [Longcore JE *et al.*, 1999][4].

Prior to the discovery of *Bd*, scientists had determined that amphibians were facing an extinction crisis that threatened approximately one-third of all species [Stuart SN *et al.*, 2004]. While habitat loss was known to be the main driver of amphibian species loss, it was

[2]Formally designated as the Rhizophidium clade in the order Chytridiales [James TY *et al.*, 2006]

[3]Cutaneous means "of the skin" from the Latin *cutis*, skin

[4]*Batrachochytrium dendrobatidis* (*Bd*) was first identified and named after a genus of poison dart frogs native to South America called *Dendrobates*, which in turn is derived from the Greek words *dendron* (a tree) and *bateyo* (I mount).

also recognized that many declines were found to occur in pristine, protected environments where known threats (such as habitat loss or species overharvesting) did not occur; these mysterious losses were recorded as "enigmatic declines" by the International Union for Conservation of Nature (IUCN) Red List of Threatened Species. *Bd* was subsequently discovered as the proximate driver of multiple-species enigmatic declines [Daszak P *et al.*, 1999; Skerratt LF *et al.*, 2007; Lötters S *et al.*, 2009] following the observation of simultaneous waves of population declines in Central America and Australia [Berger L *et al.*, 1998; Lips KR *et al.*, 2006], and the observation of local introduction and spread of the fungus [Vredenburg VT *et al.*, 2010; Walker SF *et al.*, 2010].

Bd is now recognised as infecting over 442 species of amphibians in 49 countries on all continents except for the amphibian-free Antarctic [Fisher MC *et al.*, 2009, *Bd*-maps]. The widespread occurrence, and spread, is in part attributed to the worldwide trade[5] in amphibians [Fisher MC *et al.*, 2007]. In many host-species, *Bd* is able to survive as an asymptomatic infection until metamorphosis [Berger L *et al.*, 1999] when hosts are most likely to succumb to their infection [Garner TWJ *et al.*, 2009]. This is due in part to amphibians becoming fully keratinized[6], which is a substrate for *Bd*. Other host species, such as the North American Bullfrog *Lithobates catesbeianus* (formerly *Rana catesbiana*) rarely succumb to the disease (despite also undergoing keratinisation during metamorphosis) but exhibit asymptomatic infections - they are therefore tolerant of infection. This has the consequence that they can act as natural reservoirs in habitats with multiple host species.

The ability for a pathogen to exist in multiple hosts and in natural reservoirs is a feature for many other EIDs such as avian influenza (H5N5) in Waterfowl [Gu M *et al.*, 2011] and severe acute respiratory syndrome coronavirus (SARS-CoV) in Bats [Li W *et al.*, 2005], which both complicate and impede disease control measures. Recently it was discovered that *Bd* is also able to infect syntopic crayfish (*Procambarus spp.* and *Orconectes virilis*), maintain infection for up to at least 12 weeks, and transmit the infection to amphibians [McMahon TA *et al.*, 2013]. In addition, *Bd* infectious stages are suspected to persist in environments such as in the soil or water, and have been detected for up to 12 weeks in experimental microcosms [Walker SF *et al.*, 2007]. Both of these studies demonstrate how the disease is able to persist, even at low host densities. However, no resting spore or dormant life

[5] In this thesis, the term trade will be used to refer to the business of buying and selling commodities; commerce.

[6] *Bd* colonises keratinised host cells, which is broken down and provides a nutritive substrate. Keratin is a family of fibrous structural proteins found in skin, hair and nails that supercoil to form a very stable, left-handed superhelical filament. The word Keratin derives from a Greek word $\kappa\acute{\epsilon}\rho\alpha\varsigma$ meaning "horn".

stage for *Bd* has yet been discovered. In a separate study, it was shown that *Bd* is able to adhere and even proliferate on the toes of aquatic birds suggesting waterfowl are another potential environmental reservoir and mechanism by which *Bd* propagates [Garmyn A *et al.*, 2012]. The ability of *Bd* to live in multiple hosts and remain in the environment allows the pathogen to largely forgo normal host-pathogen fluctuations, thereby driving extinctions in susceptible species of amphibian.

Bd appears to have a wholly asexual life cycle that takes between 4-5 days at 22°C in culture (Fig. 1.1). Development typically begins when a zoospore encounters live or dead keratinized amphibian tissue such as keratinized layers (stratum corneum) of amphibian skin or larval mouthparts [Berger L *et al.*, 1998]. Zoospores exhibit chemotaxis towards a number of substrates including sugars, proteins, and amino acids [Moss AS *et al.*, 2008] and encyst into keratin tissue once encountered [Berger L *et al.*, 2005]. After 1-day post infection, zoospore cysts germinate short tubular structures called germ tubes able to penetrate the stratum corneum [Van Rooij P *et al.*, 2012]. The ability of germ tubes to proceed later on to deeper skin layers appears to be host-dependent (present in *A. muletensis* and *L. caerulea* which are highly susceptible to infection, but not in *X. laevis*, which is relatively tolerant) [Van Rooij P *et al.*, 2012]. Zoospore cysts later give rise to a structure that produces and contains zoospores called a zoosporangium. These in turn release new spores either to the environment or to surrounding keratinized cells [Berger L *et al.*, 2005] as shown in Fig 1.1.

Intact skin function is essential for amphibians owing to their need to exchange gas through the skin, which enables adult amphibians to respire without rising to the surface of water and to hibernate at the bottom of ponds. In diseased individuals, a pronounced imbalance in electrolyte levels occurs as a consequence of epidermal sodium and chloride channels becoming inhibited, leading to hypokalemia (low plasma potassium) and hyponatremia (low plasma sodium). Host death appears to result from electrolyte imbalance and cardiac arrest [Voyles *et al.*, 2009; Voyles *et al.*, 2012] between 10-50 days post-infection with a mortality rate of between 50-100% depending on a number of factors including host genus, environment, and the infecting strain of *Bd*. Whether the release of a fungal toxin or direct damage to infected host cells results in the disruption of osmoregulatory function is not yet known.

Whilst some species of the chytridiomyctes are known to undergo sex, such as the order Chytridiales (including the species *Chytriomyces hyalinus* [Moore ED *et al.*, 1973] and

Siphonaria petersenii [Blackwell WH *et al.*, 2012]) and even reported in the order Rhizophydiales for which *Bd* belongs [Sparrow FK, 1960], although unconfirmed, others such as *Bd* are thought to reproduce solely clonally and therefore not undergo meiosis. This is based firstly on the observation of fixed heterozygous multilocus sequence markers found amongst a number of globally collected samples [Morehouse *et al.*, 2003]. Numerous studies have identified *Bd* as having a diploid genome, with no haploid life stage [Morehouse *et al.*, 2003; James TY *et al.*, 2009; Morgan JAT *et al.*, 2007]. Furthermore, no sexual structure or mating-type locus (MAT) encoding homeobox domain and pheromone genes have been identified.

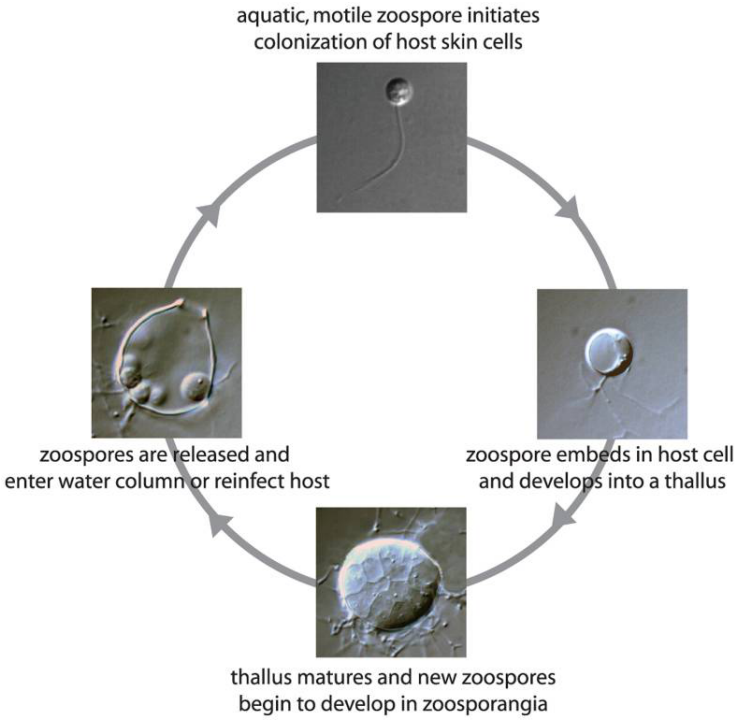


Figure 1.1: Life cycle of *Bd* from aquatic motile zoospore to embedded zoosporangia. Taken from [Rosenblum EB *et al.*, 2010].

Whether a pathogen is sexual or asexual can have an important influence on the amount and type of genetic variation that is generated, and its consequential response to natural selection in and out of a host. Two main advantages of sexual reproduction are that it allows for 1) rapid evolution (generating new combinations of genes), and 2) the ability to purge deleterious mutations (mutational determinism [Kondrashov AS, 1988]) through such mechanisms as DNA repair and chromosomal crossover during meiosis. These two advantages give rise to two alternative hypotheses for the maintenance of sex in the population: 1) the

environmental deterministic hypotheses (EDH) whereby novel genotypes are important for coping with changing environments and 2) the mutational hypotheses (MH) where purging deleterious mutations is important in a stable environment [Taylor JW *et al.*, 1999]. One manifestation of the EDH is the Red Queen Hypothesis, which states that continuing evolution is needed in order to maintain fitness relative to the systems it is co-evolving with [Hamilton W, 1980; Jaenike J, 1978]. The theory further speculates that the origin of sex may have been driven by the need to respond to and overcome pathogens or parasites that have a faster generation time (and therefore faster evolution).

Substantial costs are also associated with sexual reproduction including the requirement to find the opposite sex to reproduce, a further two-fold cost to population size [Smith JM, 1978], and the resulting recombination may break up favourable gene combinations more frequently than it creates them (under a model with a deterministic random mating population with constant fitness) [Eshel I *et al.*, 1970].

Testament to the trade-off between asexual and sexual propagation, many fungal species appear to be neither sexual nor asexual, and instead maintain the ability for both forms of reproduction. Switching can then occur during optimum conditions, such as in response to nutritional signals in the dairy yeast *K. lactis* [Booth LN *et al.*, 2010]. Other fungi retain cryptic, modified sexual cycles that enable both inbreeding and outcrossing [Heitman J, 2010], for example, through homothallism (self-fertilizing), which is a feature for the majority of species in the genus *Aspergillus* (phyla Ascomycota) [Geiser DM, 2009]. Genetic recombination in a population can also be generated in a number of non-meiotic ways including mitotic recombination/gene conversion, whereby homologous chromosomes crossover during mitosis.

Highlighting the importance of sex determination in fungi is the pathogenic yeast *Candida albicans*, which in 2000 was shown to have a MAT that specified either an α or a-type sex [Miller MG *et al.*, 2002; Johnson A, 2003]. The epigenetic phenomenon of white-opaque switching was found to be a critical step in mating [Lockhart SR *et al.*, 2002], and the outcome allowed white cells to fare better by infecting the blood stream, while opaque cells are better optimised for colonising skin [O'Day *et al.*, 2007]. Therefore, the mating-type of *C. albicans* is also an indicator of it whether it is adjusted to live as a commensal or pathogenic yeast.

Determining if *Bd* is able to sexually reproduce has important implications regarding its behaviour in a population, the principal mode and rate by which it evolves, and as

illustrated by *C. albicans*, potentially how it affects the host. Although *Bd* has thus far only been shown to exhibit the hallmarks of a clonal organism, future research may yet identify a sexual structure through microscopy, a mating-type locus from genome analysis, or evidence of linkage equilibrium (concordant with meiosis) through population genetics analysis.

In the following sections I introduce and examine the mechanisms of *Bd* infection. For example, which specific genes or genetic features enable attachment and entry into the host? How have these mechanisms related to pathogenicity evolved? And to what extent has natural selection, horizontal gene transfer (HGT), hybridization or geographical and environmental aspects played in its development? How long has *Bd* co-existed with its hosts, and at what stage in history did it develop features characteristic of a pathogen? Was this development gradual or transitional? And if so, what were the causes? How do hosts respond and to what extent are host-pathogen interactions variable? And finally, what are the epidemiological parameters and origin(s) of the disease? The following sections attempt to summarise the multitude of studies from around the world that are attempting to answer these or similar questions, thereby helping to unravel the many enigmas surrounding this devastating wildlife disease and in a broader sense, the on-going Holocene extinction.

1.2 Host-pathogen interactions and pathogenicity of *Bd*

1.2.1 Pathogen specific features that impact outcome of infection

Numerous studies have searched for unique genotypes, phenotypes, and morphotypes of *Bd* from geographically diverse habitats in order to understand *Bd* on a molecular level. Comparing sequences from the mitochondrial genome, nuclear ribosomal DNA internal transcribed spacers (ITS), numerous microsatellites [Morgan JAT *et al.*, 2007] and other multi-locus sequence types (MLST) typically found very little or no variation between individuals, leading to the conclusion that *Bd* exhibits low genetic variation [Morehouse EA *et al.*, 2003]. This has led to the development of a hypothesis accounting for the pathogens global distribution is a result of a recent and single clonal expansion. It has also been reported that no pathogen-host specificity exists [Morgan JAT *et al.*, 2007].

However, the view that the *Bd* population contained very low genetic variation was challenged by reports that certain strains differ significantly in their morphology, and also their virulence. For example, strains from Mallorca produce smaller sporangia, which give rise to less zoospores than other isolates [Fisher MC *et al.*, 2009], whilst hosts infected with isolates with different passage histories vary in time to death [Berger L *et al.*, 2005] or the extent of observed mortality [Retallick RWR *et al.*, 2007]. The causes and extent of such phenotypic and genotypic variation was therefore poorly understood, but was speculated to have in part occurred by loss of heterozygosity through mitotic recombination [James TY *et al.*, 2009].

The recent completion of the two *Bd* genome sequences, isolate JEL423 [MIT JEL423 sequence] from a Lemur leaf frog (*Phyllomedusa lemur*) in Panama and isolate JAM81 [JGI JAM81 sequence] from a mountain yellow-legged frog (*Rana muscosa*) frog in the California Sierra Nevada, provided the first opportunity to understand the evolution and the genes that contribute to pathogenicity. A detailed comparison of the two genomes using long sequence alignment software such as Mercator and Mavid [Bray N *et al.*, 2004] or Mummer [Delcher AL *et al.*, 2002] has not yet been carried out. Preliminary work on these genomes included automated annotation of the JEL423 sequence assembly, revealing 8,819 genes, 25 of which were splice variants. However, >50% carried non-specific annotation such as "(conserved) hypothetical/predicted protein".

Amongst the *Bd* genes that have a preliminary putative function relating to pathogenesis are the carbohydrate-binding module (CBM) proteins, which are thought to facilitate attachment of zoospores to host keratinized tissue. *Bd* JEL423 has undergone a large expansion of CBM genes, including 67 copies of one protein family (Pfam) domain [Punta M *et al.*, 2012] called CBM18 compared with the closely related, but non-pathogenic chytrid *Homolaphlyctis polyrhiza*, that had only ten CBM18 copies [Abramyan J *et al.*, 2012]. CBM18 are also found in greater copy numbers amongst pathogenic species in the phylum Ascomycota compared with non-pathogenic species [Soanes DM *et al.*, 2008], where chitin binding has been demonstrated in a number of species (*Trichoderma viride* and *Cladosporium*). However, it is thought CBM18 are responsible for multiple functions, of which cell-adhesion is just one possibility amongst the chytrids.

Analysis of the *Bd* genome has also shown that fungalysin metallopeptidase (also known as peptidase M36) and serine proteases[7] (such as those containing S41 domains) have under-

[7]The term protease is synonymous with peptidase.

gone extensive expansions in the *Bd* genome [Rosenblum EB *et al.*, 2010]. M36 proteinases have so far only been found to be encoded by fungi [Monod M *et al.*, 2002] where they hydrolyse structural proteins[8] using a conserved active site consisting of a HEXXH motif bound to a zinc ion and a water molecule [Rawlings ND *et al.*, 2010], and are thought to degrade host proteins when secreted by pathogens. S41 peptidases have been identified in all kingdoms of life except viruses and have been categorised into two evolutionary related subfamilies (A and B) with unprecedented differences in their active site within a single family [Rawlings ND *et al.*, 2010]. However, both subfamilies of S41 peptidase consist of three domains: a helix bundle (several α-helices), a PDZ-like domain (important for substrate specificity) and a catalytic domain that uses serine as the nucleophilic amino acid. There are over 4,000 deposited sequences for S41 peptidases in the protease database Merops, which include a wide range of substrates [Rawlings ND *et al.*, 2010]. Both M36 and S41 endopeptidases are secreted by many other species of pathogenic fungi including *Aspergillus fumigates* and *Trichophyton rubrum* [Monod M *et al.*, 2002].

In *Bd*, these proteinases have been detected in analyses of the *Bd* proteome for a number of different isolates, and are thought to play a role in degrading host-cellular components. Furthermore, they are translated at levels sufficiently high to be detected by two-dimensional gel approaches [Fisher MC *et al.*, 2009]. The fungalysin metallopeptidase gene family is differentially expressed between two different life-history stages of *Bd*, the zoospore and the sporangia, which lends support to their putative key role in the infection process [Rosenblum EB *et al.*, 2008]. In particular, of the 25 genes with a peptidase M36 domain that included microarray probes, 18 exhibited higher levels of expression in the sporangia. Only one gene (identifier 046850) was identified with a zoospore-specific upregulation of expression, suggesting this one gene may have a unique function - such as in the initial stages of infection.

Metallopeptidases have similarly undergone expansion in the human-infecting dermatophytes *Trichophyton spp.* and *Microsporum spp.* where they are highly upregulated and account for up to 36% of total secreted protein extracts [Jousson O *et al.*, 2004; Burmester A *et al.*, 2011]. Dermatophyte fungi, like *Bd*, are keratinophilic and comparative genomics of *Arthrodema benhamiae* and *Trichophyton verrucosum* show >235 predicted protease-encoding genes in each species [Burmester A *et al.*, 2011], where they break down structural proteins to gain entry and to reduce proteins to further digestion. Such secreted proteins

[8]Hydrolysis means to break a chemical bond by the addition of a water molecule. Peptidase M36 can hydrolyse laminins, elastin, collagen and keratin.

are therefore prime candidates as virulence factors in *Bd* as well as in other skin-infecting fungi.

Because *Bd* grows intracellularly and in close contact with its host, it seems plausible that it would secrete proteins that affect the host and host-response. Genes that encode for proteins able to modulate host immunity and enable parasitic infection are called effectors, and play an important role in many plant and animal pathogens. Such effector genes may encode proteins that target host defence mechanisms to enable the microbe to gain access to the host cell or avoid detection (either innate or acquired immunity for example). One example is the gene AVR3a (belonging to a group that have an RXLR or RXLQ motif, and collectively known as 'RXLRs'), which is found in the *Phytophthora* genus. AVR3a (specifically AVR3aKI that contains amino acids C19, K80 and I103) causes suppression of a hypersensitive response (apoptosis) in potatoes that lack the necessary resistance gene R3a [Bos JI *et al.*, 2006], thereby facilitating in its initial biotrophic stage of growth. Separately, several genes from a class called Crinklers (Crn) can either trigger cell death (such as PsCRN63) or inhibit cell death (such as PsCRN115) when expressed inside plant cells [Liu T *et al.*, 2011].

In *Bd*, a group of 110 genes has been identified that share sequence similarity with the C-terminal of the Crn family of effector proteins found in the *Phytophthora* genus [Joneson S *et al.*, 2011; Sun G *et al.*, 2011]. These genes have not yet been found in any other fungi (previously only found in oomycetes), and may play a role in the host-pathogen interface as Crn's are able to translocate into plant cells and target the nuclei [Schornack *et al.*, 2010] amongst *Phytophthora* species. Furthermore, many of these genes are expressed in *Bd* [Rosenblum EB *et al.*, 2012]. However, the absence of a recognisable secretion signal suggests that Crn's, should they indeed form viable proteins, remain intracellular in *Bd*. Because Crn's have not yet been found in other fungi, it has been suggested that the Crn-like genes were acquired recently, perhaps through Horizontal Gene Transfer (HGT) [Sun G *et al.*, 2011].

Experimental confirmation is needed to establish a link between Crn-like genes and the virulence of *Bd* as their functionality has yet to be proven. Indeed, no *Bd* gene or gene family has yet been experimentally proven to be responsible for any of the attributes of infection. Future experiments utilising gene knockouts or transformations may yield clues as to the way pathogenesis is maintaining and regulated. Population genome studies may also shed light on the evolution of these and other putative pathogenic elements of the pathogen.

Should an effector gene family be discovered in *Bd*, which interacts and functions solely in the host, would suggest those genes arose from a long-standing relationship with an amphibian host(s), either as part of the *Bd* genome or a separate species prior to HGT. Conceptually, if a *Bd* effector were found to solely function in a host species, then it could be considered that the pathogen creates both fungal and host amphibian proteins (such as in Richard Dawkins' Extended Phenotype concept), a finding never before documented.

1.2.2 Host-specific features that impact the outcome of infection

The burden of infection caused by *Bd* varies according to the availability of food for the host as well as its life stage, with an increase in mortality rate when food is limited, or if the host is undergoing metamorphosis [Garner TWJ *et al.*, 2009]. Increased doses of *Bd* decrease survival rates in a number of hosts, such as the highly susceptible Western Toad (*Bufo boreas*) [Carey C *et al.*, 2006]. In another study using the Common Toad (*Bufo bufo*), tadpoles were treated with repeat high doses (3000 - 15,000 zoospores) or low doses (300 - 1,500 zoospores). Of those that died before completing metamorphosis, 100% of the tadpoles treated with the high-dose had a detectable infection, compared with only 26% of tadpoles treated with a low-dose [Garner TWJ *et al.*, 2009], suggesting that *Bufo bufo* are able to clear a low-level infection, although they may still later succumb to the stresses of metamorphosis or subclinical infection. Additionally, post-metamorphic animals were treated with high-doses and low-doses. All *Bufo bufo* with low body mass died within 20 days, whilst >40% of those with high body mass were still alive by 20 days, demonstrating that body size is a better predictor of time till death from *Bd* than degree of exposure [Garner TWJ *et al.*, 2009].

In addition to infection burden, life-stage of the amphibian, and host species, different populations of a given species of amphibian (even under similar environmental conditions and disease burden) may have different infection outcomes due to differing levels of genetic variability amongst the population. In one study, three different populations of European Tree Frog (*Hyla arborea*), two of which were suffering from inbreeding, were compared. Although all three populations overcame their infection, the effect of *Bd* was more severe in the population lacking genetic diversity in terms of decreased mass and postmetamorphic life span [Luquet E *et al.*, 2012]. This might help explain why some populations appear to co-exist with the disease for some length of time before noticeable population declines are observed.

Molecular and cellular host responses can be broadly split into innate immune responses and adaptive immune responses, both of which are likely to widely differ between the hugely diverse species that *Bd* is able to infect. This is exemplified by amphibian species such as the North American Bullfrog (*Lithobates catesbeianus*) that show tolerance to *Bd*, whilst others do not, for instance the Australian Green Tree Frog (*Litoria caerulea*) [Murphy PJ *et al.*, 2009]. In one experiment [Woodhams D *et al.*, 2007], four frog species were thoroughly exposed to *Bd* (in a horizontal jar), quantities of *Bd* standardized (c. 5000 zoospores) and possible strain differences controlled for (isolate GibboRiver-Llesueuri-00-LB-1). The outcome in survival rates after 60 days ranged from 1/20 for *L. caerulea*, 14/20 for the Red-eyed Tree Frog (*Litoria chloris*) and no mortalities (20/20) for the Spotted Marsh Frog (*Limnodynastes tasmaniensis*).

The first line of defence against *Bd* is the skin of amphibians, responsible for a number of innate immune functions such as a barrier to entry, inflammation, and the presence of a complex range of antimicrobial peptides (AMP) secreted through the skin [Ramsey JP *et al.*, 2010]. AMP's have been shown to interact with the growth of *Bd*, and survival rates are shown to increase with the efficacy of particular antimicrobial peptides secreted by some species of frog [Woodhams *et al.*, 2007], some of which work optimally in synergy [Rollins-Smith *et al.*, 2002] to overcome *Bd*.

Functional approaches to investigating the amphibian response against *Bd* have focused largely on monitoring host transcriptional responses across different time points between infected and uninfected animals. By studying the Western Clawed Frog, *Silurana* (*Xenopus*) *tropicalis*, numerous AMPs have been identified that may provide resistance to some pathogens. In one study, seven separate peptides (named XT-1 $\rightarrow$ XT-7) were isolated from skin secretions, one of which (XT-7) was shown to exhibit antimicrobial activity against 22 microbial species including Gram-positive bacteria such as MRSA and the fungus *Candida albicans* [Ali MF *et al.*, 2001], whilst not testing for *Bd* specifically. In another study, the AMP preprocaerulein type III precursor (PPCP), which is the most potent AMP identified in *S. tropicalis*, was found to express a 59-fold change above *Bd*-negative controls in the spleen [Ribas L *et al.*, 2009]. PPCP is also a principal constituent of the skin secretion for a variety of amphibian species [Ribas L *et al.*, 2009]. These findings correlate with previous observations, showing that amphibian species that tolerate infection (such as *L. tasmaniensis*) produce antimicrobial peptides with greater efficacy against *Bd* [Woodhams DC *et al.*, 2007].

Bd acts on the keratinized skin-cells of the host, which contain large numbers of fine blood capillaries supporting rapid water absorption, which in turn contain a number of different white blood cells (leukocytes) of the innate and adaptive immune system. Leukocytes are all derived from hematopoietic stem cells present in bone marrow and are thought to respond to *Bd* infection. Because increased capillary blood flow is activated by intrinsic stimuli relating to the animals' hydration status [Hillyard SD *et al.*, 2011] and pathogen load is associated with disruptions in fluid and electrolyte balance [Voyles *et al.*, 2012], it may follow that leukocytes play an increasingly active role in later or more advanced stages of infection.

The innate leukocytes include cells known as granulocytes (which include Neutrophils, Eosinophils and Basophils) due to the presence of granules that contain toxic proteins and free oxygen radicals that can damage or kill pathogens such as *Bd*. Neutrophils along with macrophages and dendritic cells can engulf (in a process called phagocytosis) and either break down or inhibit pathogens using respiratory bursts. Macrophages are able to move outside the vascular system, which *Bd* is not thought to actively colonise, thereby making it a potentially important feature in combating infection. Furthermore, extracellular *Bd*, like other fungi, is not prone to phagocytosis [Lippincott's Illustrated Reviews, 2007]. Other cells called Mast cells are present in connective tissue and mucous membranes where they secrete a variety of proteins in response to an identified pathogen such as histamine, which causes inflammation and recruits Neutrophils and Macrophages. At cold temperatures, *S. tropicalis* manifests a more pronounced inflammatory response to *Bd*, whilst having a reduced ability to mount a PPCP-based innate response and increased disease burden [Ribas L *et al.*, 2009], suggesting these innate immune cells of amphibians are temperature-dependent, and that greater production of AMP's is the optimal response.

Other cells of the innate immune system are the Natural killer cells (NK-cells), which are able to detect infected cells through a cell surface marker called Major Histocompatibility Complex (MHC) and initiate cell death (apoptosis). *X. laevis* tadpoles with different MHC haplotypes have been shown to have differing survival rates to the bacterial pathogen *Aeromonas hydrophila* [Barribeau SM *et al.*, 2008]. However, whether MHC genotypes might confer differential tolerance to *Bd* pathogen load is currently unknown.

In addition to the innate immune system, amphibians - like all vertebrates - have an adaptive immune system, which is activated by the evolutionarily more ancient innate immune system. The adaptive immune system comprises a type of leukocytes called lympho-

cytes (major types include B-cells, T-cells and NK-cells). B-cells originate from the bone marrow, and are responsible for creating and secreting antibodies (immunoglobulin's, Ig) that can recognize and specifically respond (initially through binding to complementary cell surface proteins) to a particular pathogen. T-cells originate from the thymus and contain a T-cell receptor, which recognises pathogen antigens bound to MHC class I or II depending on the type of T-cell. T helper cells promote maturation of B-cells and activate Macrophages and cytotoxic T-cells, which release cytotoxins into target cells and are able to induce apoptosis via cell surface interactions.

Two key studies focusing on the Western Clawed Frog, *S. tropicalis*, as a model species showed little evidence of an adaptive immune response in the skin, liver, and spleen of infected frogs [Ribas L *et al.*, 2009; Rosenblum EB *et al.*, 2009]. In these studies, the only significant protective response observed was the induction of components of host innate immunity [Ribas L *et al.*, 2009]. However, a separate study identified IgM, IgY, and IgX antibodies present in *X. laevis* mucus secretions that bind to *Bd* [Ramsey JP *et al.*, 2010]. Furthermore, all three of these antibody types were elevated significantly in frogs previously exposed to *Bd* [Ramsey JP *et al.*, 2010]. In a separate study, *Bd* Infection was shown to cause a decrease in expression of immune genes such as the immune regulating and signalling molecules ITAM and interleukin-1, which may be a result from specific *Bd*-host interactions [Rosenblum *et al.* 2008]. Together with the identification of increased levels of *Bd*-specific antibodies, these results suggest amphibians are able to mount adaptive immune responses, raising the possibility of a vaccine for susceptible amphibians that harbour similar immune systems to *X. laevis*.

The host microbiome has also been shown to play a role in mediating infection dynamics. For example, the bacterial species *Janthinobacterium lividum* is found on several species of amphibians, such as *Rana muscosa* [Harris RN *et al.*, 2009]. *J. lividum* produces a number of natural AMPs including violacein, which has been shown to prevent morbidity and mortality to *R. muscosa* caused by the pathogen [Harris RN *et al.*, 2009]. The addition of another bacterial species, *Pseudomonas reactans*, to Red-backed salamanders (*Plethodon cinereus*) also reduced the severity of a disease symptom (significant loss in body mass by 46 days post infection) [Harris RN *et al.*, 2009 (2)]. The differences identified in host responses to *Bd* are therefore likely to involve both the innate and adaptive immune systems, as well as the host microbiome. Future comparative and functional genomics on amphibians and microbiome should provide new insights into host-pathogen interactions and the mechanisms underlying different survival rates and responses amongst amphibians [Calboli FCF *et al.*, 2011].

1.2.3 Environmental factors that impact the outcome of infection

Environment parameters play a large role in determining not only the environmental envelope allowing the survival of both actors, but also the emergent dynamics of the host and pathogen interaction. They also provide the setting to which both host and pathogen may either be adapted or disadvantaged by, and therefore contribute to the infection outcome. However, distinguishing the multivariate environmental factors from their effects on both host and pathogen can be experimentally challenging. *Bd* appears to maintain itself in the environment for up to 12 weeks even without the host [Walker SF *et al.*, 2007]. Amphibians are also declining in some regions in the absence of *Bd* or other diseases, due to pollution, cattle damage, introduction of fish and habitat destruction by logging [Berger *et al.*, 2009, Hayes MP *et al.*, 1986; Tyler MJ, 1997].

The survival, growth, and reproduction of *Bd* are all highly temperature and pH dependent, ranging between 4-25°C and pH 4-8 [Piotrowski JS *et al.*, 2004], and grows optimally at 17-25°C and pH 6-7. *Bd* does not grow well above 25°C and 50% mortality is incurred by cultures subject to 30°C for 8 days [Piotrowski JS *et al.*, 2004]. However, *Bd* has been documented in the Peruvian Andes where soil surface temperatures fall to -13.5°C and may reach as high as 30°C [Seimon *et al.*, 2007], thereby showing *Bd* can persist at greater extremes. From a widespread survey across the Iberian Peninsula, no evidence of any relationship between the presence of *Bd* and environmental variables was found [Walker SF *et al.*, 2010]. However, a negative relationship has been found between the conditional prevalence (the probability of infection at a site where *Bd* is present) and temperature/ultraviolet light [Walker *et al.*, 2010]. In this study, altitude was strongly correlated with levels of mortality, suggesting that hosts are more susceptible at higher altitudes. It has been further suggested that changes in surface and air temperatures linked to global warming have shifted towards the temperature optimum for *Bd* growth (17-25°C), thereby facilitating infections [Pounds *et al.*, 2006], although this paper has been strongly criticised by at least two studies that found only a weak causal relationship between climate change and extinction in the Harlequin Frogs (genus *Atelopus*) and in contrast, greater support for those extinctions resulting from the introduction and spread of a novel pathogen (NPH) [Rohr JR *et al.*, 2008; Lips KR *et al.*, 2008].

Temperature has also been shown to have variable effects on different host species [Kil-

patrick AM *et al.*, 2009]. For example, no effect on survival times has been found for *B. boreas* toads held at 12 and 23°C with *Bd* [Carey C *et al.*, 2006], but greater survival rates were identified in the mountain yellow-legged frog (*Rana muscosa*) [Andre SE *et al.*, 2008] and *M. fasciolatus* [Berger L *et al.*, 2004] kept at higher temperatures (27°C compared to 17-23°C). The incidence of chytridiomycosis was also higher amongst a number of species in eastern Australia throughout winter [Berger L *et al.*, 2004]. An increased disease burden was also identified in *S. tropicalis* kept at 18°C over those at 26°C, and that the animals kept in the low temperature had a reduction in their ability to mount a PPCP-based innate response [Ribas L *et al.*, 2009]. However, it is likely that the temperature-dependent nature of *Bd* infection reflects a combination of different host and pathogen responses.

Unpublished results from Garmyn A *et al.* (described at a Risk Assessment of Chytridiomycosis to European amphibian biodiversity (RACE) meeting in Madrid, ES 2012) identified that the survival rates of some frogs correlated with the presence of certain fresh-water living microbes that predate on *Bd*. The mineral content and other solutes in the water may also have an impact on the health and therefore the infection outcome for amphibians - this observation may also contribute to explaining the increased burden of infection at higher altitudes, as microbial diversity is much lower at high altitudes.

Separate infectious amphibian pathogens existing in the environment or in hosts such as *Ranavirus*, *Basidiobolus ranarum* and *Aeromonas hydrophila* have also been linked to amphibian population declines, which in combination with *Bd* further complicate prognosis. The multitude of interlinked environmental, host and pathogen interactions create a unique challenge for determining the specific cause of amphibian declines in each habitat, and making meaningful comparisons between different habitats.

1.3 Mapping the contemporary global spread of *Bd*

The emergence of *Bd* was first detected by the observation that waves of amphibian die-offs were occurring simultaneously in Central America and the Australian Wet Tropics [Berger L *et al.*, 1998]. Subsequently, numerous spatiotemporal emergences and spread were detected in the Mesoamerican peninsula, the northern tip of South America, the Sierra Nevada (US), eastern Australia, and the European Pyrenees [Fisher MC *et al.*, 2009]. The *Bd* epidemic in El Copé, Panama arrived in 2004 and followed a reduction in amphibian taxonomic, lineage and phylogenetic diversity surveyed in the following 2-4 years [Crawford AJ *et al.*,

2010]. In this study site, as many as 30 species were lost including five undescribed species, representing 41% of the total documented amphibian lineage diversity present.

The development of a highly specific and sensitive molecular, diagnostic, quantitative TaqMan® real-time PCR assay (Life Technologies, Inc., Carlsbad, CA, USA) that utilises a minor groove-binding probe to the *Bd* ribosomal DNA array (the internal transcribed spacer; ITS, and 5.8S regions) [Boyle DG *et al.*, 2004] to a large extent standardised nation-wide surveys for the presence/absence of *Bd*. Subsequently, a multiphase project focused on rapidly acquiring and compiling global *Bd* data was initiated - the global *Bd*-mapping project (*Bd*-Maps). Using a web-based system for researchers to deposit records including spatial coordinates of confirmed *Bd* infection was developed at http://www.bd-maps.net [*Bd* -Maps]. This development has enabled the extent of infection, the species infected, and their locations to be tracked and visualised. Currently, the database holds prevalence data for >34 000 samples from 3 600 locations, 79 countries, and 1 095 species of amphibian (Fig 1.2). Future detailed meta-analysis of these data should provide greater insight into global and local spatial patterns of *Bd* infections.

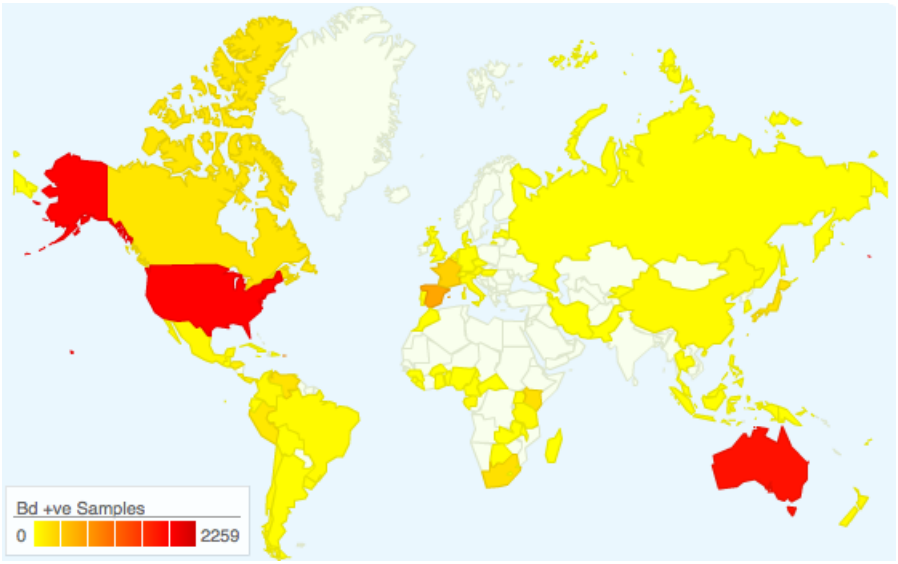


Figure 1.2: A snapshot of the global *Bd* database taken from *Bd*-maps (7.11.12) indicating *Bd*-positive samples per country.

Given the rapidity of the emergence of *Bd*, and the large numbers of species affected, it is critical that hitherto uninfected regions remain biosecure and protected from introduction of the pathogen. The mapped distribution of *Bd* shows that, while infection is widespread, it remains patchy and several areas exist that contain high amphibian biodiversity but so

far appear uninfected. The most notable of these regions is the island of Madagascar, which contains >460 species of amphibian [Vieites DR *et al.*, 2009]. The potential for *Bd* to extirpate this unique and mega-diverse community of amphibians has led to calls for a high degree of biosecurity to be implemented [Andreone F *et al.*, 2008, Lötters SD *et al.*, 2011].

The use of publicly available global mapping tools such as *Bd*-maps [*Bd*-maps] by scientists, clinicians, and policy-makers, is necessary for allocating resources and informing policies to curb the impact of the disease. A promising aspect of this is the use of mobile phone-based data-acquisition using customized applications allowing two-way communication between online databases and workers in the field. EpiCollect (www.epicollect.net) is one such application [Aanensen DM *et al.*, 2009], allowing fieldworkers to collect project-specific information focused on monitoring the spread of *Bd*, and to synchronise these data with *Bd*-Maps online and via third-generation (3G) mobile telephone networks.

Each year, tens of millions of live amphibians are globally traded for pets and food [Schloegel LM *et al.*, 2010]. The data collected from mapping *Bd*'s occurrence should be used to regulate this trade, such as informing the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). CITES aims to determine species that should have their trade regulated or prevented, and has so far been signed by over 170 countries[9]. The three groups/appendices that species can be placed are those that are either 1) threatened with extinction and affected by trade, 2) affected by trade which may later threaten them with extinction, or 3) threatened with extirpation or are of special concern to a member country[10]. However, of the 7044 documented species of amphibian [Frost DR, 2013], only 116 species of amphibian are listed in CITES (16 species in appendix 1 and 100 species in appendix 2) - in total less than 2%. Countries that have a high incidence of *Bd*, or those with high amphibian biodiversity could make better use of this treaty following greater dissemination of *Bd* surveillance data.

Another relevant regulatory framework comes from the World Organization for Animal Health (OIE), which has so far been signed by 178 countries[11]. This treaty aims to collect, analyse and disseminate relevant scientific information on animal diseases and welfare, such as determining the distribution of *Bd* in each country and sharing such info amongst

[9]List of countries that have signed the CITES treaty can be accessed at http://www.cites.org/

[10]Further details of appendices can be accessed at http://www.cites.org/eng/app/appendices.php

[11]List of countries that have signed the OIE treaty can be accessed at http://www.oie.int/about-us/our-members/member-countries/

signatories[12]. Finally, release of non-native amphibians in the UK (traded or otherwise) is regulated by the Wildlife and Countryside Act 1981[13], which states that any person who does not have permission to release non-native plants or animals is prosecutable.

1.4 Vectors and origins of the *Bd* panzootic

EIDs of free-living animals are often associated either with spill over from domestic animals or anthropogenic means [Daszak P *et al.*, 2000]. The contemporary global spread of *Bd* appears to be largely driven by the global trade in infected amphibians [Fisher MC *et al.*, 2009], although examples of spill over from captive animals have also been documented [Walker SF *et al.*, 2008]. A meta-analysis showed that at least 28 species of amphibian are known to be carriers of *Bd*, and have invaded novel ecosystems when introduced as alien species [Fisher MC *et al.*, 2007]. Several of these species are known to support asymptomatic infections of *Bd*, and have been introduced on an enormous scale; chief culprits are the African Clawed Frog *X. laevis*, the North American Bullfrog *Lithobates catesbeianus*, and the South American Cane Toad *Rhinella marina*. These species have established feral populations in the Americas, Europe, Australia, Asia, as well as many oceanic and coastal islands, and are associated with high prevalence of infection both across their native ranges and in regions where they have been introduced [Fisher MC *et al.*, 2007; Garner TWJ *et al.*, 2006]. Historically (1930s-50s), female frogs and toads (such as *X. laevis* and *Bufo bufo*) were transported in large numbers between Europe and America for use as a pregnancy test [Weldon C *et al.*, 2004], where a positive result was indicated by the production of eggs in the presence of human chorionic gonadotropin (hCG).

The globally farmed North American Bullfrog is thought to have acted as a "super-spreader" of infection for a number of reasons: the species is widely infected by *Bd* across its native range in the US [Longcore JR *et al.*, 2007] and tolerates high burdens of infection, the species is farmed in huge numbers and so infection is amplified [Mazzoni R, 2003], live and infected animals are exported globally [Garner TWJ *et al.*, 2006], and the species is widely invasive outside of its native range following uncontrolled introduction [Garner TWJ *et al.*, 2009]. Analyses of polymorphisms in the ribosomal ITS have shown that specific haplotypes

[12]Full list of OIE aims can be accessed at http://www.oie.int/support-to-oie-members/

[13]The Wildlife and Countryside Act 1981 can be accessed at http://www.legislation.gov.uk/ukpga/1981/69/enacted

of *Bd* are associated with bullfrogs, and these have spread into native populations of amphibians following introduction [Fisher MC, 2009; Goka K *et al.*, 2009], demonstrating the status of this species as a vector of the pathogen. Indeed, the first time *Bd* was found in wild amphibians in Britain was on two introduced North American Bullfrogs from a series of ponds on the East Sussex/Kent border [Cunningham AA *et al.*, 2005].

The African Clawed Frog has also been suspected to act as a superspreader, which, like Bullfrogs, are widely infected and farmed, globally transported for research purposes, and can tolerate high burdens of infection. Direct "spillover" of *Bd* from wild-caught African *Xenopus spp.* into a naïve *Bd*-susceptible Mallorcan Midwife Toad *Alytes muletensis* has been demonstrated following accidental co-housing of the two species during a captive breeding programme and subsequent release [Walker SF *et al.*, 2008]. Subsequently, in two of the *A. muletensis* populations on the Mediterranean island of Mallorca with established infections (Cocó de sa Bova and Torrent des Ferrerets), a prevalence of (or almost) 100% was recorded [Walker SF *et al.*, 2008]. Together, these studies illustrate the following important features about the transcontinental movement of infectious fungi:

- Anthropogenic activity is causing the rapid, continent-wide translocation of infectious fungi vectored by multiple host species or as saprophytes in contaminated biological material.
- There are likely as-yet-unidentified pathogens of future consequence that could be spread by trade.
- Modern, intensive-farming processes can act to amplify infection *in situ*.
- Pathogen spill over can occur from tolerant (vector/reservoir) to susceptible host species, and this risk can be high for pathogenic fungi of animals that manifest broad host-ranges and can characteristically infect multiple host species.

Anthropogenic movement of *Bd* via the global trade of amphibians appears to explain much of the rapid global emergence of chytridiomycosis, although other modes of propagation (such as on the feet of waterfowl [Garmyn A *et al.*, 2012]) have also been suggested, but not proven. However, the origins of the infection remain a mystery. Historical surveys have shown that some of the earliest records of *Bd* appear to stem from southern Africa (such as South Africa, Lesotho, and Swaziland[14]) in the early half of the twentieth century, with

[14]Archived specimens were housed in five South African institutions, Bayworld (Port Elizabeth), Natal Museum (Pietermaritzburg), National Museum (Bloemfontein), South African Museum (Cape Town), and

preserved specimens of *Xenopus* showing signs of cutaneous infection by sporangia from as early as 1938 [Weldon C *et al.*, 2004]. However, the earliest record of *Bd* comes from a specimen of Fraser's Clawed Frog (*Xenopus fraseri*) collected from the Batouri District, Cameroon in 1933, followed by specimens of *Xenopus laevis* from Lake Bunyoni, Uganda in 1934 [Soto-Azat C *et al.*, 2010].

Each of these specimens pre-dates those from other regions of the world, where the earliest record of infection in the Americas was from a specimen in the state of Quebec, Canada, in 1961 [Ouellet M *et al.*, 2005]. *Xenopus spp.* are widely infected in Africa but can tolerate infection. As widespread global amphibian declines were initially observed in the 1960s [Houlahan JE *et al.*, 2000], a temporal 'Out of Africa' pattern of *Bd* emergence appears plausible. However, wider spatial surveys from museum collections and the use of molecular analysis of globally distributed isolates of *Bd* are necessary to dissect the exact pattern of the global spread of *Bd*, and to test the validity of the Out of Africa hypothesis for the emergence of *Bd*.

Indeed, such molecular analyses have already revealed aspects of the population genetic structure of *Bd* that are in conflict with the Out of Africa model. For instance, Goka *et al.* have shown that there is a divergent lineage of *Bd* that infects *Andrias japonicas* (Japanese giant salamanders) and that this appears to pre-date the arrival of strains of the pathogen that occurred via introduction of the North American Bullfrog [Goka K *et al.*, 2009]. This lineage appears to be asymptomatic, suggesting a history of co-evolution with salamanders. Co-evolution or long standing relationships following changing host-pathogen interactions are considered part of an Endemic Pathogen Hypothesis (EPH), whilst the single worldwide clonal expansion with a possible origin in Southern Africa is described as the Novel Pathogen Hypothesis (NPH). While numerous examples of *Bd* conforming to the NPH have been found (such as in El Copé, Panama), whether *Bd* is or has been endemic to a spatial region remains unclear.

As more genetic data accrue on isolates of amphibian-associated chytrids worldwide, it is entirely plausible that multiple lineages of *Bd* will be found and that more than one of these lineages may contribute to the spread of infection. Whether these lineages all have the same capacity to cause serious infection in new host species and environments, or whether different lineages support either the NPH and EPH, remains to be seen, and is a key challenge for future research. The development of high-throughput sequencing methods presents the

Transvaal Museum (Pretoria) [Weldon C *et al.*, 2004].

opportunity to study these questions, as well as the mechanisms of pathogenicity and the evolutionary history of *Bd*.

1.5 High-throughput/next-generation sequencing

1.5.1 Introduction to using NGS to study *Bd*

While detecting the presence or fungal load of *Bd* in an amphibian host remains in the remit of diagnostic PCR methods, identifying the temporal sequence underlying the emergence of *Bd* requires the application of molecular epidemiological approaches. James TY *et al.* generated 17 sequence-based markers from 59 *Bd* isolates from five continents and 30 host amphibian species. This comprehensive study showed that the entire global diversity of *Bd* could be explained by the dispersal of a single diploid individual, and levels of genetic diversity were among the lowest recorded for a eukaryotic pathogen [James TY *et al.*, 2009]. While these findings are consistent with a recent globalization of *Bd*, they do not support the Out of Africa hypothesis of Che Weldon *et al.* [Weldon C *et al.*, 2004], as levels of genetic diversity in isolates from North America were found to be as high as they were in Africa. The extreme paucity of polymorphisms across these loci coupled with a sampling strategy targeting New World populations of amphibians, suggests that a greater depth of both sampling and genotyping would be necessary to more effectively address the question of the origin of *Bd* [James TY *et al.*, 2009].

Subsequent studies are now focused on investigating the molecular epidemiology of *Bd* using next-generation sequencing (NGS) to enable whole genome sequence typing of the pathogen. Such approaches have already found use in molecular epidemiological applications for other fungal pathogens, such as determining linkage between cases of coccidioidomycosis disseminated by organ transplants [Engelthaler DM et al., 2011]. Platforms such as Illumina HiSeq™(Illumina, Inc., San Diego, CA, USA), Applied Biosystems Sequencing by Oligonucleotide Ligation and Detection (ABI SOLiD™), or 454 Sequencing™(454 Life Sciences, Inc., Bradford, CT, USA) can provide the entire genomic sequences for numerous isolates of *Bd*, which can then be assembled (reviewed here: [Farrer RA *et al.*, 2009]) or aligned (reviewed here: [Trapnell C *et al.*, 2009]) to either of the two publically available genomes

of *Bd* (JEL423[15] or JAM81[16]), revealing polymorphic sites among those samples across the whole genome. This approach vastly increases the analytical power as tens of thousands of single nucleotide polymorphisms (SNPs; i.e. sites that differ between two isolates) are scored, as opposed to the dozens that typify earlier approaches such as that used by James TY *et al.* [James TY *et al.*, 2009].

The application of these methods can be illustrated by aligning the Sanger sequencing reads of *Bd* JAM81 to the genome of *Bd* JEL423 using Burrows-Wheeler Aligner (BWA) with default settings [Li H *et al.*, 2009], and scoring mutations using Sequencing Alignment/Map tools (SAMtools) [Li H *et al.*, 2009]. From this method, the location of approximately 10,000 SNPs (0.4 per Kb), 31 000 heterozygous sites (1.3 per Kb), and 9000 insertion/deletions can be identified between these genomes. These polymorphisms can be identified in just a few hours using only a desktop computer and a freely available alignment tool such as BWA, which is able to utilise any of the aforementioned (Illumina, 454, SOLiD) platforms. While the availability of two *Bd* genomes and high levels of heterozygosity make alignments a more attractive alternative to ascertain these differences than assemblies *de novo*, the suitability of an alignment strategy will vary depending on the type and size of the input dataset, as well as the extent of error present - all of which may affect downstream analysis. Therefore, it is important to make an initial assessment for a number of methods and parameters in terms of accuracy and utilization of the NGS dataset.

By comparing variant sites across the genome for multiple isolates, a phylogenetic tree can be constructed, revealing the evolutionary history of those organisms. By ascertaining patterns of relatedness, past demographic events and transmission events between continents/countries or sites of introduction, can be reconstructed. Numerous population genetics and comparative genomics approaches can also be applied, revealing population structure, the evolution of gene families including expansions and contractions, as well as the evidence of selection across functionally related gene families. Although many of the NGS platforms are reasonably well established in terms of ability to sequence multiple genomes, there remain a number of bioinformatics challenges that need to be considered - such as choosing the correct tools and parameters, addressing quality control, and setting up suitable and potentially reusable pipelines [Nowrousian M *et al.*, 2010].

[15] *Bd* JEL423 sequence and annotation is available for download at http://www.broadinstitute.org/annotation/genome/batrachochytrium_dendrobatidis/MultiHome.html

[16] *Bd* JAM81 sequence and annotation is available for download at http://genome.jgi-psf.org/Batde5/Batde5.home.html

1.5.2 Quality control of NGS fungal datasets

There are a large number of alignment tools designed specifically for dealing with NGS datasets such as BWA or Bowtie [Nowrousian M *et al.*, 2010; Li H *et al.*, 2009; Langmead B *et al.*,2009], some of which are designed for specific platforms. For instance, the Short-Read Mapping Package (SHRiMP) [Rumble SM *et al.*, 2009] is tailored to the di-base (colour-space) sequencing format of ABI SOLiD reads, which requires a different set of algorithms to correctly resolve consecutive SNPs or miscalled bases compared with Illumina systems [Sasson A *et al.*, 2010; Rumble SM *et al.*, 2009]. In addition to choosing an alignment tool, there are often a number of adjustable parameters that may affect the alignment such as the maximum number of mismatches allowed between a read and the reference. Common alignment optimisation options include the ability to remove reads containing (or averaging below) a Phred quality score[17], or by trimming low quality 3' ends that would otherwise prevent usable data from aligning.

The most successful alignment strategy can be determined when a closely related strain or species has already had its full genome sequenced. By including this 'reference strain', for example the *Bd* isolate JEL423, within the panel being sequenced, a control alignment can be generated. Using this alignment, the alignment parameters and SNP-calling algorithms can be fine-tuned in terms of breadth of read-coverage (ideally higher) and the number of SNPs called (ideally lower). Next, to optimise true positive/ false-negative SNP calling, the same reads can be aligned to an artificially 'mutated' version of the reference genome. This will reveal the power of the tools, within the context of the dataset, to correctly determine the number of correct (introduced) SNPs, or numbers of correctly covered genes.

A false discovery rate (FDR) approach (calculating false positives / true positives) is a good way to measure how completely the genomes from the panel of isolates are covered, to determine the error rate associated with the dataset, and to help identify the most suitable alignment strategy. For example, a dataset of whole genome sequences can be identified as incomplete or of low quality when the 'control alignment' has genes uncovered, or a high number of discrepancies are found, even after the optimisation of parameters. While these results may also arise from a poor quality reference genome, this should be less likely when the genome has been produced at a specialist sequencing centre using Sanger sequencing, which is still considered the 'gold standard' for sequencing.

[17]Software for assigning a Phred quality value DNA sequencing trace files available from http://www.phrap.org

1.5.3 Calling polymorphic sites and genome assembly *de novo*

After aligning the NGS dataset to a reference sequence, polymorphic sites can be identified using either a pre-made SNP-calling tool such as SAMtools [Li H *et al.*, 2009], diBayes [Tang S *et al.*, 2008] or The Genome Analysis Toolkit (GATK) [McKenna A *et al.*, 2010]. For example, GATK uses sophisticated statistical frameworks based around known features of genetic data, such as repetitive elements, PCR clones in the dataset, transition transversion (Ti/Tv) rates in host species, alignment and sequencing quality scores, and the ability to modify calls based on databases of polymorphisms or with training datasets. However, one drawback of pre-made tools is a lack of flexibility or transparency (black-box) for parameter optimisation. An alternative method for calling SNPs is writing a custom program or script that is specifically fine-tuned to a dataset and/or alignment, and therefore may out-perform other generic packages.

If this strategy is adopted, an initial starting place would be to first define a cut-off for the number, or percentage, of reads disagreeing with a reference nucleotide (e.g. >70% or >90%). Next, a minimum read depth can be applied to reduce low-coverage sequencing errors being called, or a minimum base or alignment quality score for inclusion applied, which would, again, reduce the number of sequencing errors mistaken for real SNPs. Depending on the length of the reads, and to what extent the quality drops at the 3' end, it might also be worth including read position filters or the pre-processing of reads. Because it is common to call greater numbers of false positive heterozygotes as well as SNPs miscalled as heterozygotes, more stringent filters could be applied over these base calls.

The numbers of putative SNPs will likely change based on the strategy and dataset used. Therefore, a number of methods should be tested against the reference isolate in order to identify the most suitable method before applying it to the rest of the panel of isolates. Maximising the ratio of true positives to false positives is important for successfully tracking the spread of *Bd* and other pathogenic fungi, especially on very fine geographical scales where isolates may differ by only a handful of SNPs.

Although NGS is primarily used for alignments, if a closely related reference sequence is unavailable (as might be expected for a novel emerging infection), high-depth NGS datasets can be used for assemblies *de novo*. Recent improvements made to read length and assembly tools designed specifically for NGS such as Velvet [Zerbino D *et al.*, 2008], are facilitating the assembly of fungal genomes in combination with Sanger sequencing [DiGuistini S *et*

al., 2009] or from entirely NGS datasets [Nowrousian M *et al.*, 2010]. Assemblies *de novo* are substantially more memory intensive and problematic in deciphering true discrepancies between assembled genomes. Despite these drawbacks, assemblies *de novo* are arguably better suited to the identification of genomic rearrangements, novel sequences not present in other reference sequences, and to give greater resolution to long indels such as gene presence/absence polymorphisms.

1.6 Reconstruction of the evolutionary relationships between isolates

Newly emerging pathogenic fungi are characterised as having low levels of genetic variation relative to their point of origin, owing to the multiple population bottlenecks that are associated with rapid spread. In order to determine how species or individuals are related, marker based sequencing is usually undertaken and qualitatively compared with geographical or spatial information. Given sufficient genetic variation either from large sample sizes, large stretches of sequences, or a mixture of the two, phylogenetic analysis is usually then performed to reconstruct evolutionary relationships.

All species harbour genomic regions that are changing more or less rapidly, and therefore different sequences will be suitable for different phylogenetic analysis or comparisons. For example, to understand the relationship of distantly (or more anciently) related species, a set of conserved genes such as alpha and beta tubulins, ribosomal RNA (rRNA) and translation elongation factors [Keeling PJ *et al.*, 2000] might be most suitable. Non-concordant or conflicting multiple gene genealogies can be explained and reveal such processes as HGT, hybridization, introgression or convergent evolution (homoplasy). Likewise, fully concordant gene trees are one of the best methods for defining a fungal species compared with a subpopulation [Taylor JW *et al.*, 2000]. However, as with all phylogenetic methods of defining species (as opposed to morphological methods), describing the limits of genetic differentiation, as well as other aforementioned processes, mean even the concept of a species remains controversial.

A separate complication is experienced in recombining populations, where different polymorphisms do not share the same demographic history, which can produce misleading and incorrect phylogenetic trees/dendrograms that do not reflect the true evolutionary relation-

ships between individuals. In this instance, a valid phylogenetic tree can still be constructed using polymorphisms within the mitochondrial genome, where recombination is thought to occur only rarely amongst fungi that inherit biparental mtDNA [Saville BJ *et al.*, 1998].

Two common methods to verify if sequences are representative of an individual are to bootstrap - or less commonly jackknife - the data. Both methods rely on constructing trees iteratively using random subsets of the data. The methods differ depending on whether sampling occurs with replacement or non-replacement respectively. 100% bootstrap or jack-knife support for a given tree or part of a tree signifies a stronger confidence in those relationships.

In addition to species or individuals in a population any other form of genetic entity such as gene family or motif within an individual could be used for phylogenetic analysis. The resulting trees show each individual at the tip of a branch (shown as a line), which are joined at nodes. Trees may either be rooted, whereby the most recent common ancestor (MRCA) is predicted and illustrated as the base node, or un-rooted, which does not specify the MRCA and instead just the relatedness of each individual. A common method for rooting trees is to include a more distantly related species to any that are being compared, which is known as the outgroup. Most trees show divergence by both the length of the branches and number of branches between any two individuals (Phylogram). Cladograms on the other hand, have fixed branch lengths. Depending on the algorithm used to construct the tree, it may contain exactly two branches at each node (bifurcating) or many branches (multifurcating). Informing branch lengths is either the use, or not, of a fixed mutation rate/molecular clock that assumes sequences are evolving at a constant rate over time and among different organisms. Although this may be suitable in some comparisons, it can result in inaccuracies. For example, older divergence times have had time to purge deleterious mutations that will be present in the tips. Other factors such as selection pressures may also result in a non-linear molecular clock.

The most common phylogenetic algorithms can be categorised into distance-based (which includes neighbour joining (NJ) and unweighted pair group method with Arithmetic Mean (UPGMA)), maximum parsimony (MP), maximum likelihood (ML), and Bayesian approaches. The most simple to implement and computationally inexpensive of these are the distance-based algorithms. First, pairwise differences in sequence are counted, with indels often omitted, producing a distance-matrix. Next, from smallest to largest distance, individuals are joined up, first by branches (of a length dependent on the number of differences), or

secondly nodes, if branches have already been set. Nodes can be joined up based on the mean, median, minimum or maximum number of differences between individuals in a given cluster (group of individuals). A key difference between the distance-based methods is that UPGMA assumes a fixed clock, whilst NJ does not.

MP methods attempt to identify the fewest number of sequence changes, usually through intermediate sequences, to link each of the individuals in a tree. This method can involve keeping or breaking the order of the sequence, as well as weighting certain changes such as Ti/Tv or 3^{rd} position codon changes that are most likely to encode a synonymous change and therefore less likely to be selected against. ML is similar to MP, but considers changes across different sites over different lineages to be statistically independent from one another. Although ML and MP are much more computationally expensive, they are thought to produce generally more accurate trees. For example, studies have compared distance-based, MP and ML, using simulated sequences and relationships with variable parameters such as sequence length, substitution rates and Ti/Tv rates across branches (although keeping them constant across each branch; clock assumption) and the outgroup (e.g. [Holland BR *et al.*, 2003; Kuhner *et al.*, 1994]). In these studies, each method was less accurate when internal branches were short [Holland BR *et al.*, 2003], or when rates of evolution varied among different sites [Kuhner *et al.*, 1994]. In both of these studies, ML with the clock assumption performed best overall compared with the other methods tested.

Bayesian approaches such as that implemented by the software Bayesian Evolutionary Analysis Sampling Trees (BEAST) have been used more recently, and apply prior probability distributions and models of evolution (using parameters such as isolate dates; tip dating, or models of mutations) that can reveal additional information about isolates, such as predicted dates of divergence [Drummond AJ *et al.*, 2007]. Briefly, a random seed relationship is made between each of the isolates. Next, the software iteratively changes tree topology and samples their probability, often using a Markov chain Monte Carlo (MCMC) method. Any new given tree probability therefore may go up or down, and based on some parameter will search around identified optimal trees to look for trees with an even greater probability. In this way, a "probability landscape" can be traversed, usually arriving at a stable location after millions of iterations. Bayesian approaches can be enormously computationally expensive, depending on the frequency of the sampling and number of iterations, along with sequence specific variables such as length and number of sequences. The end result is a set of probabilities that can be visualised as a tree. Usually the initial steps where tree topologies are changing and have low probabilities are removed from the analysis in a process called "burnin". The

ability to gather predicted dates of divergence at tips (if this information is given as a prior) makes this method particularly useful for ascertaining how and when pathogens such as *Bd* have spread.

The anticipated low levels of genetic differences amongst *Bd* isolates that in turn would result in short internal branches, along with uncertain recombination rates and no suitable outgroup (most related sequences are those from the chytrid *H. Polyrhiza* that is estimated to have diverged hundreds of millions of years ago [Abramyan J *et al.*, 2012]) will all contribute to the challenge of applying phylogenetics to *Bd*. However, by using suitable methods on whole-genome polymorphisms, and testing a number of methods for comparable results, we should be able to infer robust information about the evolutionary relationships and divergence times of our *Bd* isolates, which in turn may inform transmission events.

1.7 Population genetics approaches to understanding the evolutionary processes of *Bd*

Unlike with comparative genomics where genomes from different species are compared for fixed differences, population genetics (and genomics) aims to understand allele (variants at a given loci) frequencies amongst a population of a single species. Using population genetics analysis on whole genome datasets is a relatively new field, and offers greater power in terms of the number of loci and the different parts of the genome they represent, but often tends to have reduced power in terms of number of individuals sampled from the population. Some of the most common statistical tests that can be categorised into the field of population genetics are testing against the Hardy-Weinberg equilibrium model (HWE)[18], the index of association (I_A) [Smith JM *et al.*, 1993], $\overline{r}_d$ [Agapow PM *et al.*, 2001], the four-gamete test [Hudson R *et al.*, 1985], D [Lewontin RC *et al.*, 1964] and fixation indices (θ, F_{ST}, F_{CT}) [Wright S, 1931]. The main questions these types of analyses can help answer are the detection of adaptation, population size and structure, inter and intra- species genetic differentiation, and how variants are linked across the population (linkage analysis), which offers insight into sexual cycles. Recently, linkage has been studied in human populations (HapMap Project [International HapMap Consortium, 2003]), where the aim is to find genes associated with disease and response to pharmaceuticals.

[18]The Hardy-Weinberg principal is named after Godfrey Harold Hardy and Wilhelm Weinberg, who independently developed the principle of genetic equilibrium.

In linkage analysis, alleles that are always or frequently found together are said to be in linkage disequilibrium (LD), while alleles that are found in different combinations in the population are said to be in linkage equilibrium (LE). LD is a hallmark of either clonality or that the alleles are closely associated physically on a chromosome, which can in turn can allow genetic hitchhiking whereby an unfavourable allele remains in a population due to its association with a separate favourable allele. LD can be identified by alleles failing the four-gamete test, an $\overline{r}_d$ close to the value of one, or an I_A or D that differs significantly from zero.

A fundamental theoretical aspect of population genetics, and which many population genetics analyses use, is a neutral expectation of allele frequencies in the population such as defined by the HWE. The model states that in an ideal state (which population genetic data can be tested against), any two or more (linked or unlinked) alleles within a population do not undergo changes in frequency between generations. A theoretical HWE can only be maintained without deviation if these eight criteria are fulfilled:

1) Random mating
2) No selective pressures
3) No mutations
4) No genetic drift (random changes in allele frequencies)
5) No gene-flow between other populations
6) No genes that undergo or respond abnormally to meiosis such as sex-ratio distorting genes (meiotic drive)
7) The population size is infinite
8) Alleles are sufficiently separated on a chromosome to not have any affect on LE (i.e. alleles next to one another are more unlikely to have a meiotic breakpoint between them).

Although no wild population could probably completely achieve any one of these criteria, many allele frequencies in populations are found in the proportions expected under the HWE. Indeed, the discovery of a population with alleles that do not fall within the equilibrium is evidence that one or more of these assumptions are being broken, which may have important conservation or evolutionary implications (e.g. a small population or clonality). Testing for significant deviation between observed data (either genotypic or phenotypic or a combination of the two) to the expected HWE can be done using a statistical test such as Pearson's chi-

squared test (χ^2) or the Fisher's exact test.

The simplest scenario to illustrate the HWE is in a diploid sexual population with two alleles: A and a, with frequencies p and q respectively. The three possible combinations in the population are AA (p^2), Aa ($2pq$) and aa (q^2) and their combined probabilities add up to one as illustrated in Figure 1.3. Within a haploid individual there would be no Aa ($2pq$) present, and each individual would either be A or a, and therefore cannot be used for HWE analysis.

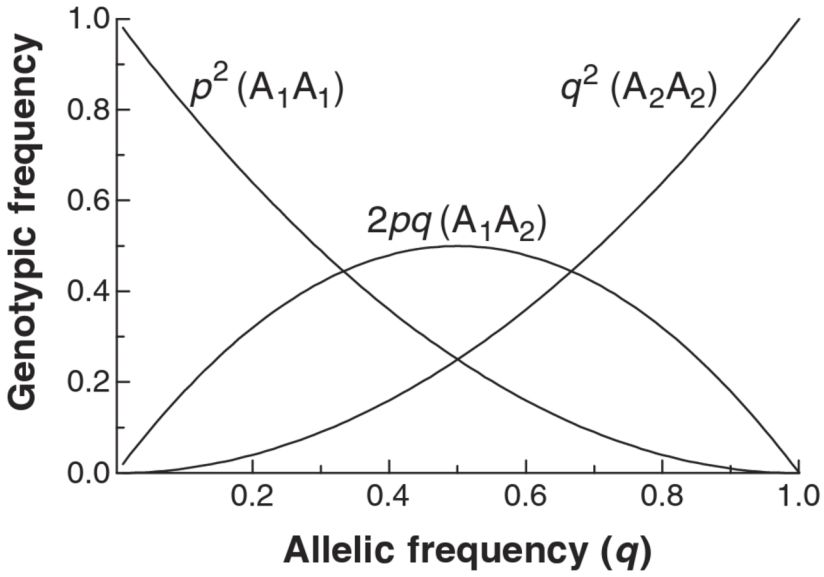


Figure 1.3: Expected genotypic frequencies under Hardy-Weinberg Equilibrium (HWE) assumptions for an allele at frequency q. Here, the simplest scenario of a diploid with two alleles at both loci is shown. However, HWE can be expanded to greater ploidy and/or a larger number of alleles.

The genotype frequencies of an individual with any type of ploidy including haploidy can be predicted with separate (non-HWE) models such as the branching model or the Wright-Fisher idealized models. Both of these models use many of the same assumptions as the HWE (with a notable exception of point 7), with the former more suitable for very small populations that will compute extinction possibilities. The example of two alleles in a diploid can be written as a binomial: $(p + q)^2 = p^2 + 2pq + q^2$. If the organism has more than two alleles then we can account for them using extra letters, e.g. $(p_1 + \ + p_n)^2$, and if

the organism is triploid (or polyploid) we can raise it to that power e.g. $(p + q)^c$ where c is the ploidy. Therefore, the expected frequency (under the HWE) of any number of alleles at two loci in an organism with any ploidy greater or equal to diploid can be expressed as:

$$(p_1 + ... + p_n)^c = \sum_{k_1,...,k_n \in \mathbb{N}: k_1+...+k_n=c} \binom{c}{k_1, ..., k_n} p_1^{k_1} ... p_n^{k_n} \tag{1.1}$$

One straightforward test for LD is the four-gamete test, where given two polymorphic loci, all four combinations are found, e.g. aa, AA, aA, Aa. Separately, the index of association (I_A) tests whether individuals with a shared locus are more likely to be the same at another locus with a null hypothesis of being the same (i.e. LD). An I_A value of zero is obtained if there is no LD, and a value of $0<x\leq1$ if there is some evidence of LD. One problem noted in the software manual of Multilocus [Agapow PM *et al.*, 2001] is that for I_A values greater than zero, the I_A will increase with the number of loci, making it less reliable and difficult to make comparisons between studies. An alternative statistic to I_A that is said to "largely remove this dependency" is $\overline{r}_d$ [Agapow P *et al.*, 2001], which uses the same equations for var_j and V_D (see equations 4.2 and 4.3 in chapter 4). Both I_A and $\overline{r}_d$ are based on unordered alleles, and both give values of between zero (no LD) and one (LD).

The different methods of detected LD are likely to differ in sensitivity and specificity given different datasets, and no single standard method has yet been decided upon. However, a number of different programs have been developed to calculate these values rather than needing to implement them *de novo*, such as Multilocus [Agapow PM *et al.*, 2001], or PLINK [Purcell S *et al.*, 2007], which is tailored to use on NGS. However, program-specific or non-standard file formats, not performing the required tests, as well as potentially overlooking new or updated equations, make implementing them in a script or program another option.

Another group of population genetic tests are the F-statistics that include F_{IS} (inbreeding coefficient), F_{ST} (fixation index) and F_{IT} (overall fixation index), which measure the correlation of genotypes within subpopulations relative to the entire population. F_{ST} values range from zero (no differentiation) to one (complete differentiation). Low F_{ST} values can therefore indicate interbreeding/hybridisation, introgression/backcrossing, or HGT.

In one key study on *Bd* population genetics [Morgan JAT *et al.*, 2007], 104 isolates from eight infected amphibian species of Sierra Nevada of California were assessed using 76 microsatellites (simple sequence repeats) - themselves derived from mining 9,100 sequences corresponding to 20% of the *Bd* genome. The authors discovered an excess of heterozygosity and no two genotypes at a given site, which was attributed to clonality supporting previous work [Morehouse EA *et al.*, 2003]. A non-significant result was found from using the Fisher-exact test to identify if alleles were in HWE. Furthermore, the I_A and $\overline{r}_d$ was used to test clonal vs. sexual reproduction, both of which identified significant LD in each of the thirteen collection sites except for two populations, Little Indian Valley and Hitchcock Lakes, where sexual reproduction could not be ruled out, and may demonstrate rare sexual recombination events. Finally, the F_{ST} revealed >50% of the observed genetic variance occurs within sites, irrespective of how the populations are grouped, supporting a recent single clonal expansion into California.

It remains to be seen whether population genomics on a global panel of isolates would shed further light on the population structure, and in particular the LD that has so far been predominantly found. If sexual reproduction or genetic exchange is identified within the population, then some parts of an individual's genome may be differentially affected i.e. recombination 'hotspots' or across a mating type locus where recombination may not occur. These and other methods of measuring selection have yet to be carried out and may shed light on the evolution and selective sweeps that have occurred since its introduction into new populations around the world.

1.8 Collecting, isolating and sequencing *Bd*

Many of the isolates used in this thesis were collected and isolated by members of the Fisher lab or other members of RACE before or over the duration of the study (see Table 1.1 for a full list of isolates and their respective collectors). The isolates chosen for sequencing were thought to best represent the global genetic diversity of *Bd*, and to be suitable for identifying new introductions or evidence of multiple-introductions of the disease into countries and continents. We therefore sampled across large geographic regions, obtaining isolates from five separate continents (Europe, North America, South America, Africa and Australia) as well as across smaller geographic distances including four isolates from a single pond near Valencia, Spain, but on different hosts (*R. perezi* and *A. obstetricans*). Many isolates were

sampled from amphibians experiencing die-offs (such as JEL423, an isolate that is also used as a control for resequencing by comparison to the available Sanger-sequenced JEL423 genome [MIT JEL423 sequence]). Other isolates chosen for sequencing came from habitats not experiencing die-offs and had already been identified as having markedly different microscopic characteristics such as smaller sporangia that in turn give rise to fewer zoospores compared with other isolates [Fisher MC *et al.*, 2009]. Although our list of isolates was not exhaustive, it was at the time of publishing in 2011, the largest intra-specific panel of eukaryotic pathogen genomes to date. Since our first publication we have more than doubled the number of *Bd* genomes sequenced.

These isolates and others collected over the duration of this study were cultured according to an unpublished document called "Recognizing, isolating and culturing *Bd* from amphibians" by Joyce E. Longcore (see [Appendix A: Culturing *Bd*]). Briefly, toe-clips or keratinised mouthparts of the tadpoles were placed in mTGhL with antibiotics (Ab) as part of an 'isolation kit' in the field, and on agar plates (w/ Ab) in the lab. Presence for *Bd* is checked in the following week using microscopy, and if confirmed, transferred to mTGhL media without Ab. The key ingredients and some aspects of fieldwork are shown in Figure 1.4.

In addition to the isolates collected by the Fisher lab or other members of RACE during the course of this study, I also visited and collected infected *A. obstetricans* from six sites across Switzerland, and successfully isolated *Bd* from two of those sites (situated within the municipalities Itingen and Waltensberg). A number of isolation kits were sent back through the post, or suspected *Bd*+ amphibians brought over by collaborators, which I then isolated from. These isolates included a number from Spain, Sardinia, France, South Africa, Ethiopia and England [Cunningham AA *et al.*, 2005]. In total, 50 isolates from 41 sites and 15 host-species were sequenced during the course of this study, which are summarised in Table 1.1. The locations of all the isolates sequenced over the course of this study are shown in Figure 1.5.

Collection site	Amphibian host	Year	Collector	Culture Reference
America, Colorado	*B. boreas*	1999	JEL	JEL274 (US)
America, PtReyes	*L. catesbeianus*	1999	JEL	JEL270 (US)
Australia, Rockhampton	*L. caerulea*	1999	LB	Lcaerulae98 (AU)
Canada, Quebec	*L. catesbeianus*	1999	JEL	JEL261 (CA)
Canada, Vancouver Island	*L. catesbeianus*	2009	PH	VC1 (CA)
England, Cumbria	*E. calamita*	2010	PM	SFBC014 (GB)

England, Kent	*L. vulgaris*	2007	MF	UKTvB (GB)
England, Kent	*M. alpestris*	2011	MF	MAL05 (GB)
Ethiopia, Hotcho	*A. enseticola*	2011	DG	ETH2 (ET)
Ethiopia, Telilia Stream	*Leptopelis sp.*	2011	DG	ETH4 (ET)
France, Ansabere	*A. obstetricans*	2007	MF	0711 (FR)
France, Lac Arlet	*A. obstetricans*	2008	MF	PNP08489 (FR)
France, Lac d'Aule	*A. obstetricans*	2010	MF	AUL (FR)
France, Loire et Cher	*L. catesbeianus*	2010	CM	RC5.1 (FR)
France, Madamette	*A. obstetricans*	2010	MF	MAD (FR)
France, Pyrenees	*A. obstetricans*	2011	MF	LHURS8 (CH)
Mallorca, Cocó de sa Bova	*A. muletensis*	2007	MF	CCB1 (FR)
Mallorca, Torrent des Ferrerets	*A. muletensis*	2007	MF	TF5a1 (ES)
Montserrat	*L. fallax*	2009	AC	L2203 (MS)
Panama, Guabal	*P. lemur*	2004	JEL	JEL423 (PA)
Sardinia, Affluente Pisharoni	*D. sardus*	2010	TG	AP15 (IT)
Sardinia, Monte Olia	*D. sardus*	2010	TG	MODS27 (IT)
Sardinia, Monte Olia	*D. sardus*	2010	TG	MODS28 (IT)
Sardinia, Scuponi	*D. sardus*	2010	TG	SP10 (IT)
South Africa, Cape province	*A. fuscigula*	2009	CW	MCT8 (ZA)
South Africa, Cape province	*H. natalensis*	2009	CW	MC55 (ZA)
South Africa, Mount, KZN	*A. vertebralis*	2010	TG	MG1 (ZA)
South Africa, Pinetown, KZN	*A. angolensis*	2011	TG	SA1d (ZA)
South Africa, Pinetown, KZN	*A. angolensis*	2010	TG	SA4c (ZA)
South Africa, SilverMine, KZN	*A. fuscigula*	2010	TG	MG4 (ZA)
Spain, Ibon Acherito	*A. obstetricans*	2004	TG	IA042 (ES)
Spain, Mainland	*A. obstetricans*	2011	MF	TORO1 (ES)
Spain, Penalara	*A. obstetricans*	2002	MF	C2A (ES)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo2 (ES)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo4 (ES)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo5 (ES)
Spain, Valencia	*R. perezi*	2008	MF	VRp1 (ES)
Switzerland, Brandsitengraben	*A. obstetricans*	2011	MF	BR1 (CH)
Switzerland, Chalchofen	*A. obstetricans*	2011	MF	CHAL3 (CH)
Switzerland, Chnubel	*A. obstetricans*	2011	MF	CH4 (CH)
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	ACON (CH)
Switzerland, Gamlikon	*A. obstetricans*	2008	TG	APEP (CH)
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	CON2A (CH)
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	0739 (CH)
Switzerland, Hergiswald	*A. obstetricans*	2011	MF	HER3 (CH)
Switzerland, Itingen	*A. obstetricans*	2010	RF	BLI1 (CH)
Switzerland, Raschenhaus	*A. obstetricans*	2011	MF	RAS3 (CH)
Switzerland, Schauensee	*A. obstetricans*	2011	MF	SCH (CH)

Switzerland, Strickrain	*A. obstetricans*	2011	MF	STR3 (CH)
Switzerland, Vorder Birnbaum	*A. obstetricans*	2011	MF	VOR3 (CH)
Switzerland, Waltensberg	*A. obstetricans*	2010	RF	BEW2 (CH)

Table 1.1: Fifty isolates used for whole-genome sequencing during this study. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger L *et al.*, 1998]. Amphibian hosts are *Amietia angolensis* (Angola River Frog), *Afrixalus enseticola* (Ethiopian Banana Frog), *Afrana fuscigula* (Cape River Frog), *Alytes muletensis* (Mallorcan Midwife Toad), *Alytes obstetricans* (Common Midwife Toad), *Amietia vertebralis* (Ice Frog), *Bufo boreas* (Western Toad), *Hadromophryne natalensis* (Natal Ghost Frog), *Litoria caerulea* (Green Tree Frog), *Lissotriton vulgaris* (Smooth Newt), *Leptodactylus fallax* (Mountain Chicken Frog), *Lithobates catesbeianus* (American Bullfrog), *Mesotriton alpestris* (Alpine newt), *Phyllomedusa lemur* (Lemur Leaf Frog), *Rana perezi* (Iberian Green Frog). AC (Andrew Cunningham), CM = Claude Miaud, CW = Che Weldon, JEL = Joyce Longcore, LB = Lee Berger, MF = Matthew Fisher, RF = Rhys Farrer, TG = Trent Garner.

For genome sequencing, DNA was first extracted and purified using the Qiagen Blood and Cell Culture DNA mini kit (cat. No. 13323) according to the protocol provided. Finally, Single-end SOLiD libraries were constructed according to the protocols provided by Life Technologies (Fragment library kit) and sequenced on two flowcells on an Applied Biosystems SOLiD 3 machine. For our second round of sequencing, paired-end Illumina libraries were constructed according to the protocols provided by Illumina sequencing (Truseq kit) and sequenced using the HiSeq 2000 platform. Finally, each isolate was cryo-preserved for future reference (see [Appendix B: sequencing and cryo-preserving *Bd*] for full details).

1.9 Aims and objectives

During this thesis, I aim to collect and isolate *Bd* from around the world by visiting infected sites (such as Switzerland, 2010 and Spain, 2012) and by distributing 'isolation kits' to members of RACE for use during their fieldwork (see [Appendix A: Culturing *Bd*]). From these isolates, I then aim to sequence and analyse their genome sequences. A key analytical step for the whole-genome datasets will be to extensively check the quality of the sequence



Figure 1.4: Collecting and isolating *Bd*. (**A**) An isolation kit, a 1 litre bottle of mTGhL, and *Bd* growing in liquid and solid media. (**B**) Ingredients and scales for preparing media. (**C**) The hood where isolates can be handled without contamination. (**D**) Collecting *A. obstetricians* in Switzerland, 2010 (**E**) All equipment and boots were treated with itraconazole to prevent *Bd* spread between sites. (**F**) The *Bd* isolates (as of 29/07/2012). (**G**) The Waltensberg, CH site where *Bd* BEW2 was collected.

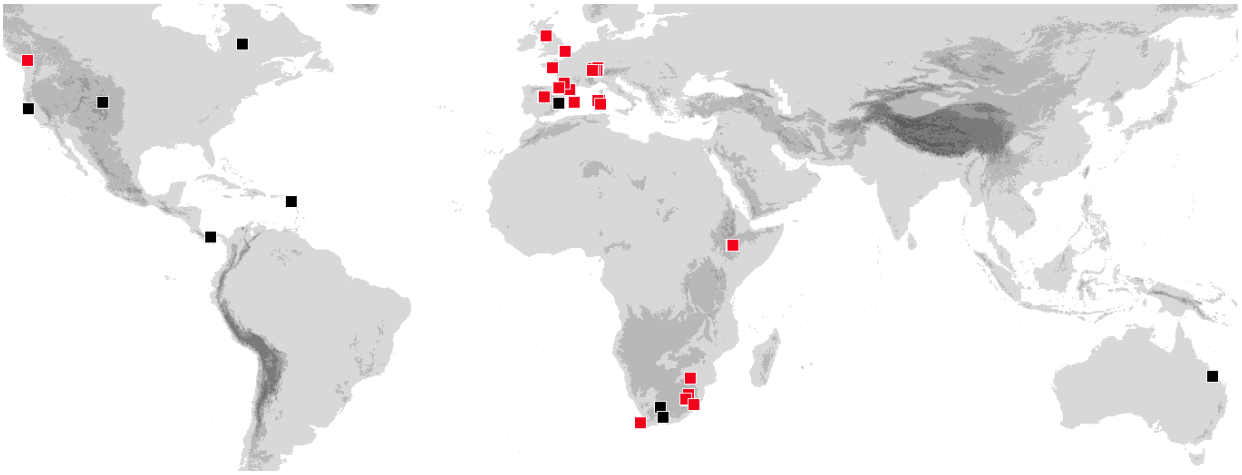


Figure 1.5: Locations that *Bd* isolates were taken from for whole-genome sequencing. Black squares signify isolates chosen for the first round of SOLiD sequencing, and red squares signify isolates chosen for Illumina sequencing. Some locations had multiple isolates taken, and some (such as JEL423 from Panama) were sequenced by both platforms.

data and benchmark the alignment tools using a custom approach. Next, using a combination of comparative genomics and population genomics, I can answer fundamental questions about the biology and evolution of *Bd*.

Initially, from the alignments I will obtain information on 1) read-depth, which will inform Gene presence/absence polymorphisms and Copy Number Variation (CNV) for each isolate, and 2), read-discrepancies, which will be used to identify polymorphic loci such as SNPs or indels between the query (sequenced isolates) and the reference (JEL423). Next, subsets of the polymorphisms will be used for phylogenetic analyses, including building trees and identifying epidemiological signals such as disease transmission between sites or host-jumps. From this analysis, I can then help answer the extent that either the NPH or EPH explain the current pandemic. Additionally, I will test for selection and statistical enrichment of polymorphisms across families of virulence genes. Finally, using a number of population genetic approaches, I aim to study the role that recombination plays in the apportionment of genetic diversity and its impact on the population structure.

Novel genomic approaches such as the use of NGS for whole-genome sequencing provide new insights into the fundamental evolutionary modes in many organisms, and so should also be applied to *Bd*. The extensive literature and research on *Bd* are raising it to a model status in terms of understanding the drivers that govern the emergence of virulent pathogenic fungi. Given the recent emergence of this and other fungi as serious pathogens, there is an

urgent need to increase our research into the drivers that underlie emerging fungal pathogens in order to effectively take action and to mitigate the effect of such infections.

1.10 Funding

The UK Natural Environment Research Council (NERC) and the European Research Area Network project "Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity" (RACE) have provided funding for this research.

Chapter 2

Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects

Rhys A. Farrer[1], Daniel A. Henk[1], Dan MacLean[2], David J. Studholme[3], Matthew C. Fisher[1]

[1]Department of Infectious Disease Epidemiology, St Mary's Hospital, Imperial College London, London, UK [2]The Sainsbury Laboratory, Norwich Research Park, Norwich, UK [3]College of Life and Environmental Sciences, University of Exeter, Exeter, UK

The contents of this chapter were published in: Farrer, R.A., Henk, D.A., MacLean, D., Studholme, D.J. & Fisher, M.C. Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects. *Sci. Rep.* **3**, 1512; DOI:10.1038/srep01512 (2013).

Author contributions: R.A.F., D.A.H. and M.C.F. wrote the main manuscript text. R.A.F. prepared the figures. R.A.F., D.M. and D.J.S. conceived of the study.

2.1 Abstract

Sequence alignments form the basis for many comparative and population genomic studies. Alignment tools provide a range of accuracies dependent on the divergence between the sequences and the alignment methods. Despite widespread use, there is no standard method for assessing the accuracy of a dataset and alignment strategy after resequencing. We present a framework and tool for determining the overall accuracies of an input read dataset, alignment and SNP-calling method providing an isolate in that dataset has a corresponding, or closely related reference sequence available. In addition to this tool for comparing False Discovery Rates (FDR), we include a method for determining homozygous and heterozygous positions from an alignment using binomial probabilities for an expected error rate. We benchmark this method against other SNP callers using our FDR method with three fungal genomes, finding that it was able achieve a high level of accuracy. These tools are available at http://cfdr.sourceforge.net/

2.2 Introduction

Sources of error within next-generation sequencing (NGS) projects can result in poorly resolved genotypes [DePristo MA *et al.*, 2011; The 1000 Genomes Project Consortium; Kiezun A *et al.*, 2012]. Without an assessment of the false discovery rate (FDR[1]) for genome or transcriptome projects, misaligned reads and inaccurate base calls can go unnoticed thereby propagating into SNP discovery. From the onset, datasets comprising sequenced fragments (reads) can harbour a range of potential sequencing errors such as PCR amplification bias in Illumina data [Aird D *et al.*, 2011], polyclonal errors in SOLiD data [Sasson A *et al.*, 2010] and homopolymers in 454 data [Huse SM *et al.*, 2007]. Despite a general trend towards higher success rate at correctly deciphering bases in reads, individual runs can still harbor unexpected levels of error from low quality DNA extractions or library preparations.

In addition to read-quality, an arguably more substantial source of error arises during read-alignment [Landan G *et al.*, 2009], the process whereby sequenced reads are mapped to a closely related reference genome. This is the first and most fundamental analysis undertaken once the DNA sequence has been produced [Flicek P *et al.*, 2009]. Alignments

[1]The False Discovery Rate is a method to assess accuracy, and calculated by taking the number of false positives divided by the number of true positives

are often preferable to *de novo* assemblies due to increased speed and reduced memory requirements. Given a low read-depth or highly heterogeneous sequence, alignments may also recover more genetic data than assembling without a reference due to current limitations in assembly algorithms [Alkan C *et al.*, 2011]. As is true for *de novo* assemblies, the accuracy of alignments varies considerably depending on the software and its parameters used [Frith MC *et al.*, 2010], the type and size of the dataset and the amount of erroneous base calls. Further, alignments are also affected by the genetic distance between reference and newly sequenced genomes.

Most alignment tools score the placement of a read based on the uniqueness of its match. For example, BWA [Li H *et al.*, 2009] and Bowtie [Langmead B *et al.*, 2009] are frequently used alignment tools that are based on indexing a reference sequence using a Burrows-Wheeler transformation (BWT) [Burrows M *et al.*, 1994]. These tools can align in base (Illumina, 454, Ion Torrent, PacBio) or color space (SOLiD) using a combination of base and alignment mapping quality scores to determine the correct positions of reads and assigning a genotype, whilst ignoring reads containing low quality base calls or low mapping scores. The alignment tool SHRiMP [Rumble SM *et al.*, 2009] is tailored to the platform specific biases in the color-space format of ABI SOLiD reads, using a different set of algorithms to resolve consecutive single nucleotide polymorphisms (SNPs) or miscalled bases, which base-space' aligners would unsuccessfully determine. Many other examples of short-read alignment tools such as Maq [Li H *et al.*, 2008], SOAP [Li R *et al.*, 2009] and Zoom [Lin H *et al.*, 2008] are available and may be preferential given specific experimental requirements such as speed, accuracy or additional features [Li H *et al.*, 2010].

Pre-processing reads prior to alignment by removing low quality reads or 3' ends is a common initial step to improve alignment accuracy. Sequencing errors can also be detected and removed during the pre-processing of reads using software such as Quake [Kelley DR *et al.*, 2010] or EDAR [Zhao X *et al.*, 2010], which search the datasets for subsequences that occur in low frequency, and therefore likely to be due to sequencing error. Some experiments may also only consider unique matches to the reference sequence. The software Seal [Ruffalo M *et al.*, 2011] was recently developed to evaluate alignment tools using simulated reads by testing them for correctly aligned reads, error rates and run-times. However, simulated data will not always correctly recreate the types of systematic biases that may be present within an actual dataset.

More recently, data post-alignment processing is being performed to filter erroneous

sites or increase resolution of insertions and deletions (indels). For example, the Genome Analysis Toolkit (GATK) [McKenna A *et al.*, 2010], which was used by the 1000 Genome Project Consortium found that a local realignment around indels could reduce misaligned reads and nearby false positive SNPs. In addition, GATK uses base and variant quality score recalibrations, whereby the quality scores of reads are re-adjusted according to several covariates over known variant sites as well as use expected transition / transversion ratios (Ti/Tv) [McKenna A *et al.*, 2010]. The option of removing duplicate reads arising from a common progenitor DNA molecule have also been applied as both pre- [Li H *et al.*, 2009] and post-processing steps [McKenna A *et al.*, 2010], ensuring greater accuracy of subsequent SNP-calling.

Some alignment tools also have inbuilt genotyping such as Maq [Li H *et al.*, 2008], which is able to call SNPs based on a Bayesian statistical model using an expected rate of heterozygotes within the genome and a dependency coefficient. However, most SNP-callers are used post-alignment, such as the UnifiedGenotyper of GATK [McKenna A *et al.*, 2010], or the Sam/Bcftools' SNP caller [Li H *et al.*, 2009] that considers a number of factors when determining the presence of a polymorphism such as minimum and maximum number of mismatches or filtering by Phred quality scores. Most SNP-callers consider homozygous and heterozygous (bi-allelic) sequences, whilst others such as Bcftools do not properly handle multi-allelic variants and only take the strongest non-reference allele. Many other SNP-callers have been developed, which may be tailored to data-types or expected levels of variation. The number of possible tools and their rate of development make benchmarking an issue that needs to be frequently readdressed.

Increasingly, panels of resequenced isolates include a reference strain for which an assembled genome is already available, in order to determine suitable depth of read coverage for the other isolates, and to refine the alignment and SNP-calling parameters to reveal acceptable levels of discrepancies. However, this approach still does not address divergence between the comparison strain and the rates of true positives that will later be called from the alignments. Furthermore, heterozygosity in diploid organisms is still difficult to verify, with few methods developed that do not rely on verification with previously identified sites. Although there have been numerous comparisons of DNA sequence assembly tools for length and accuracy of the contigs, running time, ability to resolve repetitive genomic elements [Farrer RA *et al.*, 2009; Miller MJ *et al.*, 1994; Zhang W *et al.*, 2011] and numerous methods have been proposed for reducing sequencing errors [Salmela L, 2011; Sasson A *et al.*, 2010; Huse SM *et al.*, 2007] there is currently no standard method for assessing the accuracy of

correctly identifying mutations from an alignment and SNP-calling strategy.

Here, we introduce a series of Perl scripts that can reveal, and compare the false discovery rates (FDR) for a given NGS dataset used for alignment and SNP calling, requiring only an available reference genome sequence that is closely related to the sequenced strain. Comparing the FDR of methods of alignment and SNP-calling simultaneously can reveal the best combination of tools. In addition to providing a method for determining FDR, we have developed our own tool for calling polymorphisms post-alignment based on cumulative binomial expectations for the number of reads agreeing with a polymorphism, given a depth and an expected error rate. These expectations are stored in lookup tables and used along with the Samtools mpileup format [Li H *et al.*, 2009] as an input. The method is able to identify homozygous and heterozygous mutations with appropriate accuracy. We have benchmarked this method (Binomial SNP Caller from Pileup; BiSCaP), and others, using the scripts for assessing FDR.

2.3 Methods

The genome sequence and feature files for *Saccharomyces cerevisiae* S288C were downloaded from the SGD on 31.3.11 (http://www.yeastgenome.org/). Genomes of *Puccinia triticina* race 1 isolate 1-1 and *Batrachochytrium dendrobatidis* (*Bd*) JEL423 were downloaded from the Broad Institute (http://www.broadinstitute.org/). Illumina reads were obtained from the Short Read Archive under accessions SRR003681 and SRA009871 for *S. cerevisiae* S288C and *P. triticina* respectively. We previously resequenced the genome of *Bd* JEL423 [Farrer RA *et al.*, 2011; Chapter 3] using SOLiD, which is available for download in the Short Read Archive under accession SRA030504. Genome sequences were modified by randomly choosing and modifying 1nt/Kb within the coding sequence (CDS) using a script that is part of the toolset (Introduce Random Mutation into Sequence; IRMS.pl).

SRA files were converted to FASTQ and aligned to their modified reference genome sequences using BWA v0.5.9 [Li H *et al.*, 2009] with default parameters and SHRiMP v2 [Rumble SM *et al.*, 2009] with an 80% identity threshold for read alignment. Pileups were made using SAMTools v0.1.18 [Li H *et al.*, 2009] and polymorphisms called using the mpileup command piped to Bcftools v0.1.17-dev and filtered using vcfutils.pl with default parameters. In order to assess heterozygous variants, we randomly chose and modified 1nt/Kb within the CDS using the "HET" setting of IRMS.pl, which first generates a duplicate (homologous)

genome. We then simulated single-end reads from these modified sequences to the same depth as the 'real' data using simLibrary and simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/) using the default runfile (s_3_4x), which describes how noise and cluster intensities are distributed in a real run of an Illumina machine, and aligned those reads to the non-modified reference genomes.

The Genome Analysis Toolkit (GATK) v2.1-9 [McKenna A *et al.*, 2010] was assessed according to the "Best Practice Variant Detection with the GATK v4, for release 2.0" detailed on the Broad Institute website. Briefly, Picard Tools v1.68 (http://picard.sourceforge.net) was first used for marking duplicates. Indel-realignment was performed using the GATK2 RealignerTargetCreator and IndelRealigner tools. Next, the UnifiedGenotyper was used to output raw variants that were used for base quality score recalibration (BaseRecalibrator and PrintReads). The UnifiedGenotyper was then assessed using default parameters on the new BAM file. Without a training dataset, variant recalibration is still considered experimental, so this step was left out for each of the three fungal genomes. Percent cut-offs with variable minimum read-depths and using the Binomial SNP caller from Pileup (BiSCaP) v0.11 (presented here) were also used to call polymorphisms. Each alignment/SNP-calling combination was assessed for accuracy using the Comparison of False Discovery Rate script (cFDR).

BiSCaP is based on the binomial expectations for the number of reads agreeing with a reference base over a given locus. These expectations allow for polymorphisms to be called with different levels of leniency for sequencing errors dependent on the depth of read coverage and for heterozygous positions to be called without a bias for the read depth. Briefly, an expected alignment and base calling error rate (e.g. 0.1 or 0.01) is used to generate a list of binomial probabilities p for sequencing and aligning k number of correct bases, given a read depth of n (number of trials) or $f(k;\ n,\ p)$. The probability a base is homozygous (h^1) can be considered as P(1 - error rate). In a diploid, the probability of a heterozygous base (h^2) can be considered as P((h^1/ 2) + ((error rate * 0.5)), where half the sequencing or alignment errors' are now specifying the two correct bases. Equally, a heterozygous allele in a triploid sequence (h^3) can be considered as either P((h1/ 3) + ((error rate * 0.25)) or P((h^1 / 3) * 2) + ((error rate * 0. 5)). From these, a cumulative h^1 probability for the lower tail can be calculated (ch^1) in addition to the minimum values found from the cumulative h^2 or h^3 upper tail and the cumulative h^2 or h^3 lower tail (ch^2 and ch^3 respectively). Binomial values are generated by the script GBiD.pl (Generate Binomial Distributions) and stored in a lookup table. Pre-calculated tables for error rates of 0.1 and 0.01 up to a read depth of five hundred

are provided in the current version. BiSCaP then uses one such look up table to infer the most probable consensus nucleotides from the alignment.

Briefly, the algorithm for determining the consensus sequence is to tally each of the four possible aligned bases, each of which needs to be $\geq$ the minimum read depth to be considered a consensus allele. The most common base is considered homozygous where ch^1 $\geq$ error rate and ch^1 $>ch^2$. The most common and 2^{nd} most common base are considered heterozygous where $ch^2 \geq$ error rate and $ch^2 > ch^1$ for both bases. A triploid heterozygous site is considered when the ch^3 for each of the three most common bases $\geq$ error rate. Indels are treated separately but using the same criteria and sub routine.

BiSCaP by default provides details of polymorphic sites in Variant Call Format (VCF) [Danecek P *et al.*, 2011] and can also output pileup lines into separate files based on the identified mutation-type. Other optional parameters include different lookup table based on error rate, minimum read depth, ploidy, stringency for heterozygous SNP calling, and a Phred quality score filter. If the read depth is greater than the lookup table depth (default 500 for error rates 0.1 and 0.01), reads up to the maximum lookup table depth can be used to determine the genotype (default), or printed to a separate file named outside-distribution'. The cFDR script considers Percent True Positive (TP) homozygous SNPs as ((N^o TP hom. SNPs / N^o Introduced mutations) x 100) and the Percent False Positive (FP) homozygous SNPs as ((N^o FP hom. SNPs / (N^o TP and FP hom. SNPs)) x 100).

2.4 Results

Using fungal NGS datasets and genomes from three separate phyla: *Saccharomyces cerevisiae* S288C (Ascomycota), *Batrachochytrium dendrobatidis* (*Bd*) JEL423 (Chytridiomycota) and *Puccinia triticina* race 1 isolate 1-1 (Basidiomycota), we assessed the accuracy of the alignment program BWA v0.5.9 [Flicek P *et al.*, 2009] with Samtools piped to Bcftools [Li H *et al.*, 2009] by comparing their false discovery rate (cFDR) on 1nt/Kb test SNPs within the coding sequence (CDS) of their corresponding reference genome (workflow shown in Figure 2.1). We found that these tools resulted in a highly variable accuracy rate for both homozygous SNPs and heterozygous positions dependent on the input datasets, even after normalising the datasets for read length, depth of reads in the alignment and by only considering mutations that were found within the CDS regions (Fig. 2.2A).

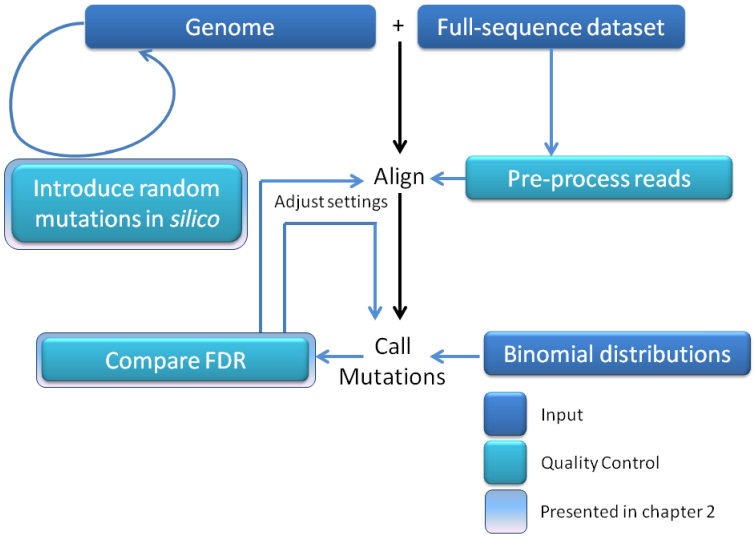


Figure 2.1: A flow diagram showing the steps to verify false discovery rate and call polymorphic sites using cFDR and BiSCaP. Specifically, the reference (genome) sequence has random nucleotides changed. Next, reads are aligned to the modified genome and SNPs called. Finally, the cFDR script can be used to distinguish true positive and false positive SNPs. Given poor results, alignment and SNP-calling parameters as well as read pre- or post-processing methods can then be adjusted.

Pre-processing input datasets was briefly assessed for ability to improve the accuracy of downstream SNP calls. Specifically, trimming 3' ends from all *Bd* JEL423 reads increased the number of reads that were aligned using BWA nearly 5 fold from 2.2 million to 10.4 million, in turn, enabling >4 times the number of true positives SNPs and heterozygous positions to be called by Sam/Bcftools. Furthermore, read-trimming decreased the percent of false positives SNPs from 15.6% to 5.9% and false positive heterozygous positions from 14.4% to 3.4%, demonstrating how much error can, and was present within this NGS dataset, often at greater frequencies towards the 3' ends of reads. Trimming the reads only achieved an improvement in the number of true positive and false positives with the *Bd* JEL423 reads, and conversely reduced both the number of false and true positives in the *P. triticina* and *S. cerevisiae* datasets. The variation between each of these datasets demonstrates the importance of assessing quality control as a preliminary step for resequencing projects.

To demonstrate how alignment and SNP-calling varied, we compared combinations of methods and parameters on the *Bd* JEL423 SOLiD 30-mer dataset (Fig. 2.2B). Owing primarily to the high levels of sequencing errors, even after removing low quality 3' ends, none of the tested methods called >73% true positives SNPs or <5% false positive SNPs. The alignment program SHRiMP, which is specifically designed for the color-space' reads of SOLiD sequencing aligned 60% of the *Bd* reads and 70% after trimming to 30mers, compared with BWA that aligned just 9% full length and 41% 30mers. Despite the differences in the output alignment depth between the two programs, BWA (with the *Bd* 30mers) resulted in a greater number of true positives and approximately equal number of false positives than SHRiMP called using either read length and any of the SNP-calling methods we tested. This comparison shows that BWA is a more appropriate tool for this particular dataset. The comparably low false discovery rate and number of reads aligned by BWA on the SOLiD *Bd* dataset was not found to the same extent in the Illumina *Sc* and *Pt* datasets, where >59% full-length reads and >68% trimmed reads were aligned to their reference sequence. For heterozygous base-calls, GATK2 had the greatest accuracy of any of the methods tested with 86.78% true positives and remarkably not a single false positive.

A comparison of the FDR for SNPs achieved by the alignment program BWA and the SNP calling method presented here (BiSCaP) using default settings on the *Bd* 30mer dataset revealed 6582/12458 (52.83%) true positives and 494 (6.98%) false positive SNPs that were covered $\geq$ minimum required depth (MD). This result is a considerably preferable outcome compared to using SAM/Bcftools for SNP-calling (as shown in Fig. 2.2B), whilst the UnifiedGenotyper of GATK performed competatively in terms of specificity at a small

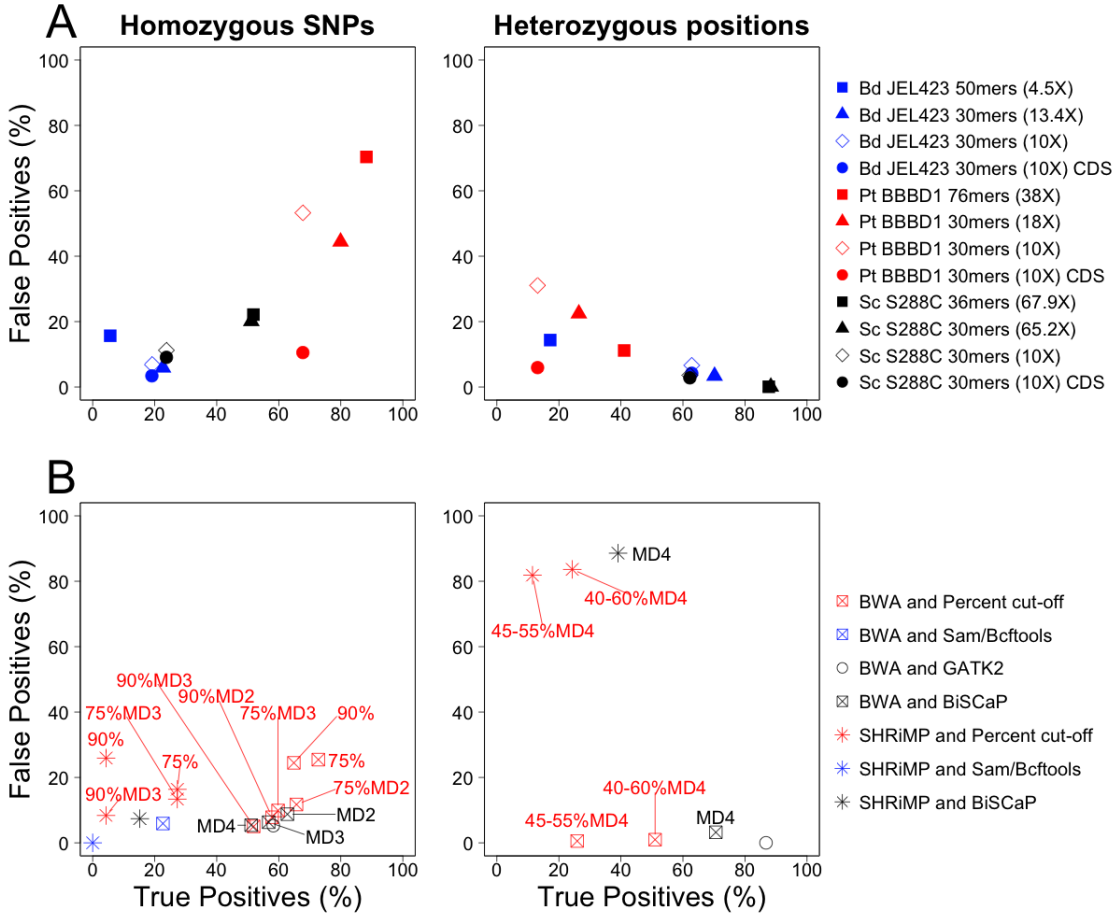


Figure 2.2: False discovery rates for variants were ascertained using cFDR for three fungal NGS datasets. (**A**) Dataset-specific error rates were identified for both homozygous SNPs and heterozygous positions after alignment with BWA and calling SNPs using SAM/BCFtools with default settings, which persisted after trimming to 30mers, aligning random 10X deep subsets of aligned reads, and only considering SNPs that fell over CDS regions thereby reducing genome size as a factor. (**B**) Combinations of alignments and SNP-calling methods resulted in different accuracies from the *Bd* JEL423 30mer NGS dataset. Experimenting with a variety of methods can therefore reveal the most suitable method for a given dataset based on these metrics of accuracy. Sam/Bcftools only takes the strongest non-reference allele so was not included in the assessment of heterozygous accuracy. Parameters include the percent cut-off for inclusion as a SNP or heterozygous base and minimum depth (MD).

expense in sensitivity compared with one of the tested settings of BiSCaP (MD4). This variation of FDR demonstrates the importance of comparing candidate tools for a given study. Of the false positives called by BiSCaP, 328 were identified when aligning to the non-modified reference genome (351 total), and 321/328 were covered by 100% uniquely aligned reads, suggesting that they may be real genetic changes occurring between the separate batches of isolate *Bd* JEL423, or genuine mistakes in the reference genome.

The remaining 45.83% of modified positions in *Bd* that were not identified, consisted of 98.5% that were uncovered by the minimum depth, 1.47% called as heterozygous, and a single incorrectly called homozygous SNP. Of the 84 bases that were incorrectly identified as heterozygous, 83 sites correctly called the modified base but also inferred the presence of the reference allele and one consisted of the reference base and a deletion. None of these false positives were identified without first modifying the reference for cFDR, which may suggest they arise from misaligned reads. These results demonstrate that by far the greatest impact on alignment/SNP-calling error on the *Bd* 13.4X deep dataset arises from lack of coverage. Surprisingly however, the *S. cerevisiae* 67.9X deep dataset (Fig. 2.2A) also revealed a similar situation where 1920/1929 of the false negatives were due to below required depth compared to the remaining nine that were incorrectly called as heterozygous positions.

To explain this pattern of enrichment for uncovered polymorphic sites (*Bd* and *S. cerevisiae* had 98.4% and >99.9% CDS covered respectively, compared to 45.8% and 15.7% of the introduced mutations within the CDS), we compared the distances between the introduced SNPs (Fig. 2.3A). Firstly, we found that *Sc* and *Pt* tended to have more closely associated introduced mutations (both had 19% occurring within 20nt of each other) than *Bd* (only 2% within 20nt). In terms of number of contigs and genes, number and average length of exons, and length of coding and introns sequence, *Bd* is situated between *Pt* and *Sc*. However, a clear difference is found in the number of genes selected by IRMS for introducing mutations in: 953, 1976 and 5701 for *Sc*, *Pt* and *Bd* respectively, which could be caused by one or more computational or biological differences such as differing numbers of genes specified with overlapping exons or splice variants specified in the feature file, which IRMS excludes. In any case, the difference in the number of modified genes likely explains the difference in distance between mutations. Despite this, false negatives were predominately more closely associated in *Bd* and *Pt*, whilst false negatives in *Sc* appeared to not be so clearly correlated with distance from other mutations.

We also looked at minimum depth cut-off points for different numbers of reads agreeing

or disagreeing for a particular base, a parameter often taken into account by mutation callers, but compounded by variable depths over each of base in the genome. To examine this issue, we looked at the depth of read coverage over false positives and false negatives called using the percent cut-off methods and BiSCaP, which uses the depth-dependent cut-offs (Fig. 2.3B). As expected, homozygous erroneous base calls using either BiSCaP or percent cut-off methods were more frequently called over lower read depth positions in the alignment. BiSCaP achieved fewer false positive/negatives over the lowest read depths for *Bd* and *Sc*, and also over bases covered by more than 5 reads. The dataset *Pt* had fewest false base calls with the 90% cut-off method, which again highlights that different algorithms are suitable for different datasets, and only after testing them for FDR could you be certain of the success rate, or determine the most appropriate method.

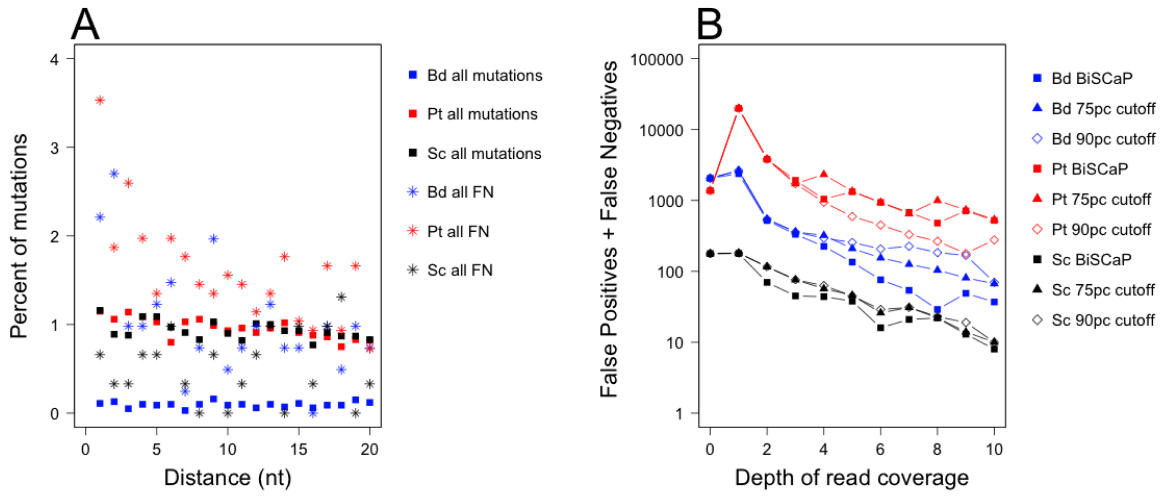


Figure 2.3: Erroneous base calls (homozygous SNPs) from BiSCaP and percent cutoff methods were compared for proximity and depth of read coverage using full dataset 30mer *Bd*, *Sc* and *Pt* reads. (**A**) False negatives were predominantly more closely associated for all 3 datasets (SNPs called with BiSCaP). False negatives were almost entirely caused by lack of coverage in each of these datasets demonstrating the most divergent part of each of the genomes is the most poorly resolved. (**B**) Homozygous errors were more frequently called over lower depth regions using strict cut-off methods than with BiSCaP.

2.5 Discussion

The rapid rate of increase in large-scale population studies using genome resequencing for SNP detection necessitates the development of improved tools to assess the quality of resequencing projects. Here we describe the development and efficacy of two such tools. We have tested the comparison of false discovery rate (cFDR) and Binomial SNP Caller from pileup (BiSCaP) scripts on sequence data from fungal genomes from three separate phyla: *Saccharomyces cerevisiae* S288C (Ascomycota), *Batrachochytrium dendrobatidis* (*Bd*) JEL423 (Chytridiomycota) and *Puccinia triticina* race 1 isolate 1-1 (Basidiomycota). These fungal genomes were chosen to represent a range of genome sizes and structures in terms of introns numbers, repeat richness and sequence heterogeneity. These methods can be applied to any resequencing study regardless of taxon. A large number of similar alignment, SNP-calling tools or pre-processing methods could have also been tested using these FDR scripts in addition to those we have tested here (BWA [Li H *et al.*, 2009], SHRiMP [Rumble SM *et al.*, 2009], GATK [McKenna A *et al.*, 2010], SAMTools [Li H *et al.*, 2009]).

To identify the factors involved in the FDR variation between the three datasets composed of equal alignment depth, length, and modified genetic distance to its reference sequence, we extracted the positions corresponding to true and false positives. A large amount of variation between datasets, alignments, and SNP calling accuracy was identified using these tools, and was used to identify the most suitable combination of methods to accurately detect variants. Similar approaches to the generalised tools we present here have already been used by a number of NGS projects [Raffaele S *et al.*, 2010; Farrer RA *et al.*, 2011], and facilitated by the release of these packages, a wider range of projects could also make use of either, or both cFDR and BiSCaP. Each combination of methods performed differently on the *Bd* JEL423 genome with one simulated divergence rate, and from this test we could decide a single set of methods and parameters that performed optimally. However, FDR validation and SNP-calling should be readdressed for every new dataset. For example, GATK would be most suitably benchmarked against a dataset for which a training set of known variation is available. However, even without variant quality score recalibration, GATK performed well on the fungal JEL423 genome, in particuar over bi-allelic heterozygotes.

Alternative methods for assessing homozygous variants that rely largely on simulated reads [Ruffalo M *et al.*, 2011] or cross checking databases of polymorphisms [Musumeci L *et al.*, 2010] to determine FDR would have less power than our method in terms of realistic

read and alignment error, or relying on resources that harbour their own sources of error. The alignment tool Maq [Li H *et al.*, 2008] includes a command for introducing mutations into a reference sequence, with the intent of assessing alignment accuracy from simulating reads. Our method is able to make use of and assess real read data, which can harbor any number of platform or non-platform specific errors that may influence ability to call SNPs, in contrast to simulated reads or use of databases. This feature makes our method the only currently available technique to simultaneously assess the quality of data generated and the quality of methods used to analyze those data.

BiSCaP has been designed for variant-calling across haploid, diploid or triploid sequences, with corresponding binomial probabilities provided. In its current form, BiSCaP takes longer than the other assessed SNP-callers to complete: roughly one hour on a desktop computer on a modestly sized genome and dataset (the 13.4X *Bd* dataset to the 23Mb genome), which is likely to persist until scripts are converted into a lower level programming language. The FDR method is also able to verify heterozygous alleles called by either GATK or BiSCaP using simulated reads, and the cFDR scripts finish running within a few minutes using the ouput from either of these (and SAMTools) SNP-calling tools. BiSCaP is able to call polymorphisms from standard input and output formats, making it a versatile tool for projects utalising these formats. We have not assessed how quality scores could be used to improve accuracy of those SNP-calls although both GATK and BiSCaP are able to filter potential SNPs based on these scores, so could be incorporated into the analysis.

Each of the methods presented here rely on a reference genome strain that is both high quality in terms of accurately assembled and with correct base calls, and has few discrepancies to the consensus (resequenced) isolate. For example, more distantly related isolates (reference and consensus) will result in a greater number of 'false positives' called by cFDR. Furthermore, if the reference sequence is poorly resolved (missing sequence or low quality or repetitive areas), the method may identify genuine polymorphisms that will also be considered false positives. This limitation can be partially resolved using a separate quality control measure for those false positives using either quality scores or called without first modifying the reference.

We found the ideas for, and implementation for SNP-calling based on cumulative binomial probabilities a suitable method for determining polymorphisms from an alignment. We tested both of these methods on three unique fungal pathogens, each of which are thought to be predominantly diploid, and therefore had both homozygous and heterozygous polymor-

phic positions called, which we found homozygous polymorphisms using the cFDR scripts to have a high level of accuracy. Either or both of these tools could be used with any other sequenced panel of diploid or haploid isolates to gauge the accuracy of alignment and SNP-calling.

2.6 Acknowledgements

Funding: R.A.F. was funded by the Natural Environment Research Council (NERC). D.A.H. and M.C.F were supported by the Wellcome Trust. No additional external funding received for this study

2.7 Author Contributions

RAF, DAH and MCF wrote the main manuscript text. RAF prepared the figures. RAF, DM and DJS conceived of the study.

2.8 Competing Financial Interests

The author declares no competing financial interests.

Chapter 3

Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent recombinant lineage

Rhys A. Farrer[1,2], Lucy A. Weinert[1], Jon Bielby[2], Trenton W. J. Garner[2], François Balloux[1], Frances Clare[1,2], Jaime Bosch[3], Andrew A. Cunningham[2], Che Weldon[4], Louis H. du Preez[4], Lucy Anderson[2], Sergei L. Kosakovsky Pond[5], Revital Shahar-Golan[1], Daniel A. Henk[1] and Matthew C. Fisher[1]

[1]The Dpt. Infectious Disease Epidemiology, Imperial College, W2 1PG, UK; [2]Institute of Zoology, Zoological Society of London, NW1 4RY, UK; [3]Museo Nacional de Ciencias Naturales (CSIC), Madrid, 28006, Spain; [4]Unit for Environmental Sciences and Management, North-West University, Potchefstroom 2520, South Africa; [5]Dpt. Medicine, University of California San Diego, San Diego CA, 9500, USA

The contents of this chapter were published in: Farrer RA, Weinert LA, Bielby J, Garner TW, Balloux F, *et al.* Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalized hypervirulent recombinant lineage. *Proc Natl Acad Sci U S A.* 2011 Nov 15;**108**(46):18732-6. doi: 10.1073/pnas.1111915108.

Author contributions: D.A.H. and M.C.F. designed research; R.A.F., L.A.W., J. Bielby, T.W.J.G., F.C., J. Bosch, A.A.C., C.W., L.H.d.P., L.A., R.S.-G., D.A.H., and M.C.F. performed research; J. Bosch, A.A.C., C.W., L.H.d.P., and S.L.K.P. contributed new reagents/analytic tools; R.A.F., L.A.W., J. Bielby, T.W.J.G., F.B., F.C., D.A.H., and M.C.F. analyzed data; and R.A.F., L.A.W., J. Bielby, T.W.J.G., F.B., D.A.H., and M.C.F. wrote the paper.

3.1 Abstract

Batrachochytrium dendrobatidis (*Bd*) is a globally ubiquitous fungal infection that has emerged to become a primary driver of amphibian biodiversity loss. Despite widespread effort to understand the emergence of this panzootic, the origins of the infection, its patterns of global spread and principle mode of evolution remain largely unknown. Using comparative population genomics, we discovered three deeply diverged lineages of *Bd* associated with amphibians. Two of these lineages were found in multiple continents and are associated with known introductions by the amphibian trade. We found that isolates belonging to one clade (the Global Panzootic Lineage; *Bd*GPL) have emerged across at least five continents during the 20^{th} Century and are associated with the onset of epizootics in North America, Central America, the Caribbean, Australia and Europe. The two newly identified divergent lineages (*Bd*CAPE, *Bd*CH) were found to differ in morphological traits when compared against one another and *Bd*GPL, and we show that *Bd*GPL is hypervirulent. *Bd*GPL uniquely bears the hallmarks of genomic recombination, manifested as extensive inter-genomic phylogenetic conflict and patchily distributed heterozygosity. We postulate that contact between previously genetically isolated allopatric populations of *Bd* may have allowed recombination to occur, resulting in the generation, spread and invasion of the hypervirulent *Bd*GPL leading to contemporary disease-driven losses in amphibian biodiversity.

3.2 Introduction

Emerging fungal diseases present a growing threat to the biodiversity of free-ranging animal species [Fisher MC *et al.*, 2009, Frick WF *et al.*, 2010]. In recent years, a single species within a basal clade of fungi little recognized for their pathogenicity, the Chytridiomycota, has gained substantial notoriety owing to its impact on global amphibian biodiversity [Fisher MC *et al.*, 2009]. *Batrachochytrium dendrobatidis* (*Bd*) is known to have driven the local extinction (extirpation) of up to 40% of species in affected communities [Crawford AJ *et al.*, 2010], and to have spread rapidly through diverse environments [Fisher MC *et al.*, 2009]. Despite widespread research efforts, the geographic origin of this emerging infection and its subsequent patterns of global spread remain largely unknown [Frick WF *et al.*, 2010; James TY *et al.*, 2009; Walker SF *et al.*, 2008]. For example, the hypothesis that *Bd* originated in Africa and spread via the global trade in *Xenopus spp.* during the first half of the 20^{th}

century [Weldon C *et al.*, 2004] is disputed by the detection of lower genetic diversity in African isolates of *Bd* compared to isolates from North America [James TY *et al.*, 2009]. This observation, however, was based on only five African isolates collected from two sites in the South African Cape, compared to 29 USA isolates collected from multiple sites across the United States. The genetically depauperate nature of African *Bd* has been further challenged by the discovery of isolates of African descent exhibiting pronounced differences in morphology and virulence [Walker SF *et al.*, 2008; Fisher MC *et al.*, 2009].

A separate marker based study conducted by Morehouse *et al.* [Morehouse EA *et al.*, 2003] using ten loci from 35 isolates found very low levels of polymorphism (five variable positions) and fixed heterozygous sites suggesting a primarily clonal mode of reproduction, although with some evidence for spatially localised genetic recombination [Morgan JAT *et al.*, 2007]. These results, and other marker-based studies [James TY *et al.*, 2009; Weldon C *et al.*, 2004], support the Novel Pathogen Hypothesis, by suggesting that *Bd* is a recently emerged pathogen [Skerratt LF *et al.*, 2007]. Although these results describe the population structure of *Bd* at a coarse scale, patchily sampled genomes combined with a chronic lack of genetic diversity at the sequenced loci have prevented a reliable inference of *Bd*'s evolutionary history. Recently, new whole-genome typing methods have greatly increased our ability to decipher genealogies by enabling unbiased sampling of the entire genome, thus increasing our power to date the coalescence of lineages and to identify recombination events. In this chapter, I present the analyses from the first whole-genome sequencing dataset of *Bd* isolates.

3.3 Materials and Methods

3.3.1 Library preparation and sequencing

To examine the recent evolutionary history of *Bd* and its patterns of global genome diversity, a panel of 20 isolates previously obtained and kept in culture were chosen (full details in Table 3.1). These isolates were chosen to represent both the global and a local distribution of *Bd*. For the global aspect of the panel, we included isolates from five separate continents (Europe, North America, South America, Africa and Australia). To study the localised populations, we included four isolates from a single pond but from different hosts (*R. perezi* and *A. obstetricans*) near Valencia, Spain. Most of the isolates were sampled from amphibians that

had experienced die-offs (e.g. JEL423 from Panama) or were experiencing die-offs (e.g. L2203 from Spain). Isolate JEL423 is also suitable for using as a control for the sequencing and alignment methods, as it is this isolate that we are using as the reference Sanger-sequenced genome [MIT JEL423 sequence]. Other isolates chosen for sequencing came from habitats not experiencing die-offs (E.g. MC55 from South Africa) and have been identified as having markedly different microscopic characteristics such as smaller sporangia that in turn give rise to less zoospores compared with other isolates [Fisher MC *et al.*, 2009].

Genome libraries were constructed according to the protocols provided by Life Technologies (Fragment library kit). Fragment library sequencing was performed on 2 flowcells on an Applied Biosystems SOLiD 3 machine with an anticipated sequencing output of 7.5 Gigabase per flowcell. Briefly, genomic DNA was sheared into 105bp fragments using a Covaris S2 sonicator. After end-repair and column purification, double strand barcoded adapters were ligated to the fragments. After a size selection step the fragments were nick-translated and PCR amplified using 6-10 cycles, followed by column purification and another size selection. The final libraries were quality-controlled on a bioanalyzer and quantified with qPCR. Two pools of libraries were required: Pool 1 contained libraries 1-8 with barcodes 1-8, and pool 2 contained libraries 9-20 with barcodes 1-12. Pooled barcoded libraries were united, and ePCR was performed according to Life Technologies‘ (“full scale”) specification (templated bead preparation kits). After a WFA run, a total amount of 260 - 290 million beads was loaded onto the flowcells. The output read length was 50bp.

3.3.2 Optimisation of alignments and SNP calling parameters

By resequencing the genome for isolate JEL423, we had a control for sequencing, alignment and SNP calling. Firstly, the genome sequence and feature file for *Bd* strain JEL423 was downloaded from http://www.broadinstitute.org/ (GenBank project accession number AATT00000000). In order to avoid misleading counts for gene-absences or polymorphisms, the feature file for JEL423 also had all but the longest splice variants removed for each gene leaving 8794/8819 genes. The full length reads from JEL423 were then aligned to the nuclear and mitochondrial genome sequences using Burrows-Wheeler Aligner (BWA) v0.5.8 [Li H *et al.*, 2009] with default parameters.

Due to the poor quality of the reads, 6.27Mb (27%) of the genome was completely uncovered (no reads aligned). By trimming the low-quality bases from the 3’-end of the

Collection site	Amphibian host	Year	Collector	Culture reference	Reads & Depth
America, Colorado	*B. boreas*	1999	JEL	JEL274	4.7 (6X)
America, PtReyes	*L. catesbeianus*	1999	JEL	JEL270	9.4 (12X)
Australia, Rockhampton	*L. caerulea*	1999	LB	Lcaerulae	10.1 (12.9X)
Canada, Quebec	*L. catesbeianus*	1999	JEL	JEL261	7 (9X)
England, Kent	*L. vulgaris*	2007	MF	UKTvB	8.3 (10.6X)
France, Ansabere	*A. obstetricans*	2007	MF	0711/1	6.3 (8.1X)
France, Lac Arlet	*A. obstetricans*	2008	MF	PNP08489	6 (7.7X)
Mallorca, CdsB	*A. muletensis*	2007	MF	CCB1	10.5 (13.4X)
Mallorca, TdF	*A. muletensis*	2007	MF	TF5a1	4.3 (5.4X)
Montserrat	*L. fallax*	2009	AC	L2203	5.8 (7.5X)
Panama, Guabal	*P. lemur*	2004	JEL	JEL423	10.5 (13.5X)
South Africa	*A. fuscigula*	2009	CW	MCT8	5.5 (7X)
South Africa, Cape	*H. natalensis*	2009	CW	MC55	9.3 (11.9X)
Spain, Ibon Acherito	*A. obstetricans*	2004	TG	IA042	6.9 (8.8X)
Spain, Penalara	*A. obstetricans*	2002	MF	C2A	11.6 (14.8X)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo2	6.3 (8X)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo4	8 (10.3X)
Spain, Valencia	*A. obstetricans*	2008	MF	VAo5	7.4 (9.5X)
Spain, Valencia	*R. perezi*	2008	MF	VRp1	6.8 (8.7X)
Switzerland	*A. obstetricans*	2007	TG	0739	6.2 (8X)

Table 3.1: Samples used for SOLiD sequencing and details of alignments. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger L, 1998]. The number of reads (millions) aligning to the *Bd* JEL423 genome assembly and the corresponding depth of coverage. Amphibian hosts include *Afrana fuscigula* (Cape River Frog), *Alytes muletensis* (Mallorcan midwife toad), *Alytes obstetricans* (Common Midwife Toad), *Bufo boreas* (Western Toad), *Hadromophryne natalensis* (Natal Ghost Frog), *Litoria caerulea* (Green Tree Frog), *Lissotriton vulgaris* (Smooth Newt), *Leptodactylus fallax* (Mountain Chicken Frog), *Lithobates catesbeianus* (American Bullfrog), *Phyllomedusa lemur* (Lemur Leaf Frog). Collectors are AC (Andrew Cunningham), JEL (Joyce Longcore), LB (Lee Berger), MF (Matthew Fisher), TG (Trent Garner), CW (Che Weldon). CdsB = Cocó de sa Bova, TdF = Torrent des Ferrerets

reads, we obtained an increase in number of reads aligning and depth of coverage over the genome. To identify the optimum read-length, I realigned 42mers, 36mers, 30mers, 24mers, 18mers and 12mers to the genome. 18mers obtained the greatest genome depth of any of the dataset. Uncovered bases in the genome continued to decrease by trimming reads to 12mers. However, indiscriminately trimming also has the potential for loosing good sequence data, and as reads are reduced in length, there is also a greater chance for them being incorrectly aligned due to redundant sequences in the genome. Therefore, in addition to the depth of coverage, the most suitable pre-processing methods were chosen on the ability to correctly call the discrepancies between reference and query sequences.

To identify the most accurate method (including trim-length) for calling bases in the genome, a False Discovery Rate (FDR) approach was used (see Chapter 2). Firstly, 12,458nt within the coding regions (CDS) were randomly selected (amounting to 0.1% of the total gene sequence of *Bd* JEL423) and randomly changed to a different nucleotide. Next, the reads (following different trim lengths or pre-processing for low quality bases) from *Bd* JEL423 were aligned to this modified sequence, and SNPs called (Fig. 3.1A). This approach was used to assess the alignment tool Short-Read Mapping Package (SHRiMP) v2 [Rumble SM et al., 2009] tailored specifically for SOLiD data. For testing SHRiMP, I used the 80% identity threshold for read alignment. Additionally, I tested the alignment tool Burrows-Wheeler Aligner (BWA) v0.5.8 [Li H *et al.*, 2009] using both default settings and "-k 3 -N" which increases the maximum differences allowed in the seed alignment (parameter k) from two to three and searches up to the maximum number of differences (parameter N).

For SNP-calling from each of these alignments, I tested three separate percent cut-offs (>90, >75 and >50%) of reads agreeing with a given base, and minimum read depths (MRD) of 1, 2, 3 and 4. SNPs were also called by Samtools.pl [Li H *et al.*, 2009] with and without the varFilter (default parameters), and the output of Samtools piped to Bcftools v0.1.17-dev and filtered using vcfutils.pl with default parameters (as recommended). We also developed and tested a method based on three depth-dependent binomial distributions (see Chapter 2) called Binomial SNP-Caller from Pileup format (BiSCaP).

From each of these methods, a variety of accuracies for calling polymorphisms were achieved (true positives; TP >false positives; FP) (Fig. 3.1A). Firstly, 30mers were found to provide the best balance between read-coverage and accuracy rates of any of the trim lengths tested. The alignment tool BWA resulted in a greater number of true positives and approximately equal numbers of false positives than SHRiMP called using any read length

and any of the SNP-calling methods we tested. For SNP-calling Samtools achieved up to 82.3% TP (TP / 12,458) at the expense of 71.74% FP (FP / TP + FP) or as little as 12.11% FP with 69.05% TP. Therefore, none of the 14 methods involving Samtools resulted in <12% FP. Using either 75% or 90% cut-offs with >3MRD as well as BiSCaP with default settings performed fairly well.

We also assessed the ability for calling heterozygous variants, which has been recognised as an important feature of *Bd*'s genome [Morehouse *et al.*, 2003; James TY *et al.*, 2009; Morgan JAT *et al.*, 2007]. First, we concatenated the unmodified JEL423 genome to the modified JEL423 genome and then simulated 30nt single-end reads from these modified sequences to the same depth as the 'real' data (13.4X) using simLibrary and simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/) using the default runfile (s_3_4x), which describes how noise and cluster intensities are distributed in a real run of an Illumina machine, and aligned those reads to the non-modified reference genomes. While an Illumina runfile does not represent the true intensities and sequencing errors present within the SOLiD reads, a SOLiD runfile could not be found when this analysis was performed, and many shared features such as reduced quality at the 3' ends of reads make it at least comparable. Finally, heterozygous positions were called by each of the methods (Fig. 3.1B). Based on this set of benchmarking, we used trimmed the SOLiD reads to 30mers, aligned them with BWA using default settings, and called SNPs using BiSCaP. Additionally we chose a MRD of 4, thereby ignoring low-covered bases that made up a substantial number of false positives (Fig. 3.1C).

Finally, to test for adequate coverage of 30mer mapped reads in the alignments, we took 6 random subsets of aligned *Bd* JEL423 reads (1, 3, 5, 7, 9 and 11X deep) and aligned them separately to the reference *Bd* JEL423 genome (Fig. 3.2). From these alignments, we found that breadth and depth of read coverage over the genome and exon sequences increased as we increased subset size. By 5X deep (mean depth of 20 alignments was 9.5X, lowest 5.4X) nearly 80% of the genome and >90% CDS were covered. Indeed, even with the smallest subset size, no gene was completely uncovered by reads, suggesting that the depth of sequencing is adequate for whole-genome analysis. Furthermore, genes that are found to be completely uncovered for other isolates in this dataset would therefore be reasonable evidence for gene presence/absence polymorphism relative to this isolate.

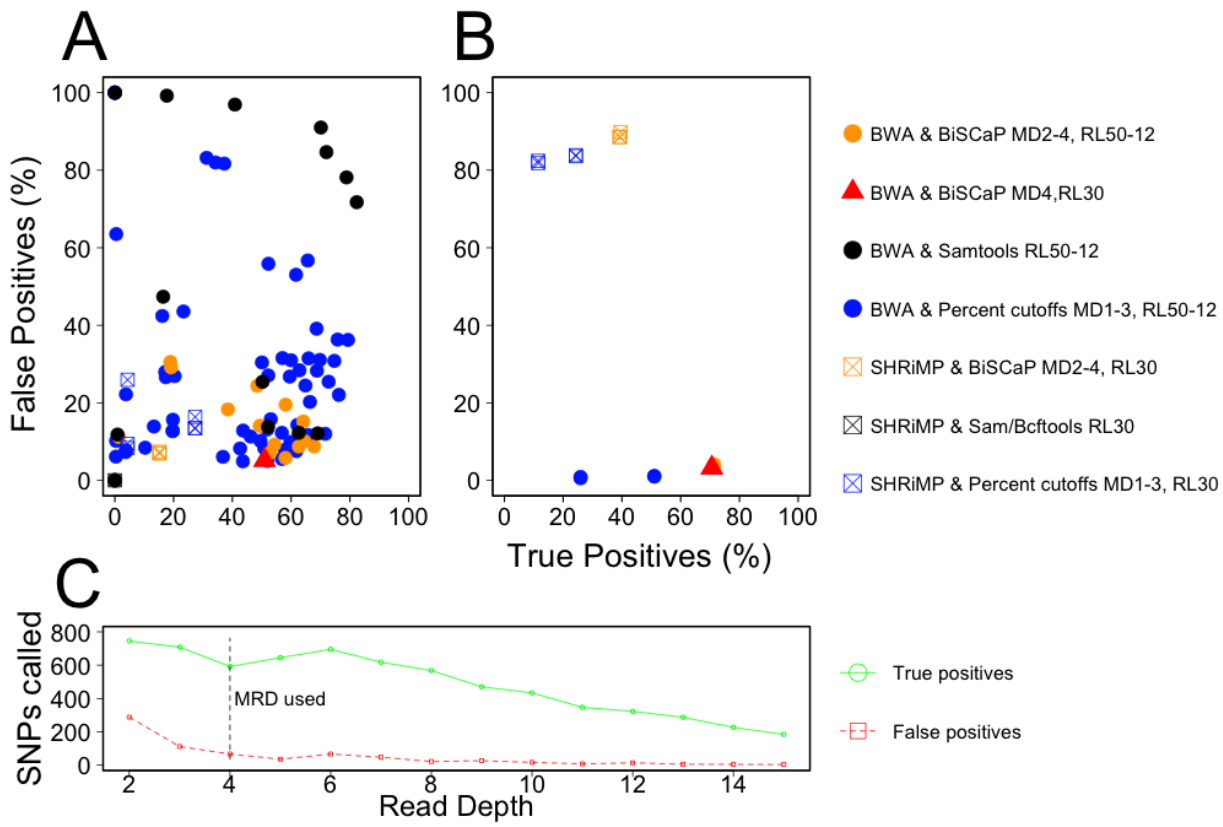


Figure 3.1: Single Nucleotide Polymorphism (SNP) calling was optimized using False Discovery Rate analysis. (**A**) Over 100 combinations of alignment and SNP-calling tools and parameters including minimum depth cutoffs (MD) and read trim lengths (RL) had their accuracy at calling simulated SNPs assessed. BWA and a binomial SNP-calling method (BiSCaP) using a minimum read depth (MRD) of four was chosen for this study (shown here as a red triangle) (**B**) Simulated reads were used to assess accuracy at calling simulated heterozygous positions (**C**) The number of true positives and false positives found using BiSCaP over each depth of coverage, the MRD used in this study is indicated by the dotted line.

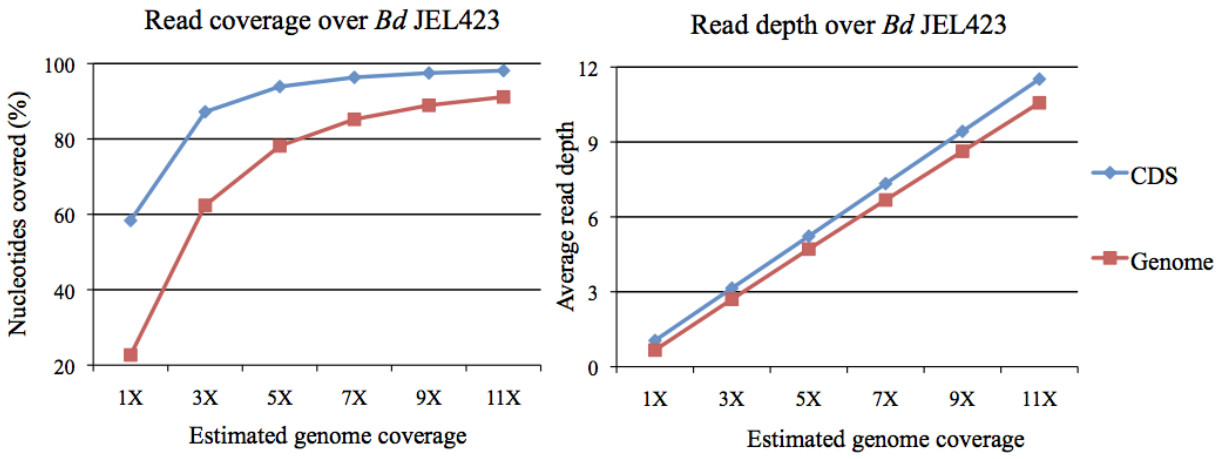


Figure 3.2: Six random subsets of *Bd* isolate JEL423 30mers (ranging from 1X to 11X deep) were re-aligned back to the JEL423 genome sequence. (**Left**) Breadth of coverage increased over whole genome including the coding regions (CDS) as subset size increased. At 5X deep coverage (all our isolates had 5X or greater) >80% of the genome and >90% of the genes. (**Right**) Depth of coverage over CDS and the genome as a whole increased as subsets of data increased. Using a 5X deep subset, CDS were covered by an average of 5.23 reads per base.

3.3.3 Phylogenetic analysis of nuclear and mitochondrial genomes

Having ascertained a suitable method for the alignment and SNP-calling for this dataset using the control isolate JEL423, we then applied these methods to the rest of the isolates. For phylogenetic analysis, we used a subset of these sites (loci), which we named "entirely covered in all" or ECA. ECA sites were those that passed the following criteria:

1) Homozygous SNP or heterozygous position in at least 1 of the isolates.
2) Covered by $\geq$4 reads (minimum read depth) for all isolates
3) Loci absent of any insertions or deletions in any isolate

In total, 22,181 ECA sites were obtained for the nuclear genome and 306 for the mitochondrial genome. Both sets were individually concatenated into FASTA files, which were converted to nexus file using a Perl script and then to XML using Beauti [Drummond AJ *et al.*, 2007]. Beauti is used to calculate some of the prior values, such as suitable root height and population size of 0.097.

Because recombination of DNA sequences will result in inaccurate or misleading trees

(e.g. relationships between isolates will incorrectly appear closes following recombination), phylogenetic trees were made using the mitochondrial ECA positions (which are thought not to undergo recombination in *Bd*) using Bayesian Monte Carlo Markov Chain (MCMC) analysis implemented by BEAST v1.6.1 [Drummond AJ *et al.*, 2007], with the HKY substitution model [Hasegawa *et al.*, 1985] with a gamma distribution of rates, under a strict clock. We ran three separate chains of over 50 million generations (after discarding the first 10% as burn in) to estimate the posterior probability distributions of all model parameters and topologies, sampling every 1,000th generation. We next combined the estimated topologies from the three different runs and used the program TreeAnnotator v1.6.1 from the BEAST package [Drummond AJ *et al.*, 2007] to construct the maximum clade consensus tree for the mitochondrial genomes.

For the nuclear genomes, which are more likely to have undergone recombination events, we used the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm in PAUP to construct a tree. UPGMA trees are based on the overall genetic distance inferred from the ECA polymorphic sites, and not the evolutionary or historical relationships between isolates, therefore making it suitable for recombining populations.

3.3.4 Patterns of mutation

To look for mutation bias, all ECA polymorphic sites identified in the CDS regions were categorised by the composition of bases amount genome features and into non-strand specific transition (*Ts*, A<>G or C<>T) and transversion (*Tv*, A<>C or G<>T) rates. We further categorised each of the mutations into synonymous and non-synonymous mutations. To measure selection across the genome we calculated the rates of synonymous substitution per synonymous site (denoted dS or Ks), non-synonymous substitution per non-synonymous site (denoted dN or Ka) and omega (ω, dN/dS) for every gene in every isolate using the yn00 program of PAML [Yang Z *et al.*, 2007] implementing the Yang and Nielsen method [Yang Z *et al.*, 2000], which takes into account Ts/Tv rate bias and base/codon frequency bias in the sequence.

Because these values are all for fixed differences (same in the population), which is not even the case for our intra-population isolates, we also used the direction of selection (DoS) measure proposed by Stoletski and Eyre-Walker for genes with few substitutions [Stoletzki N *et al.*, 2011]. DoS is based on the McDonald-Kreitman (MK) test, where the count of

fixed synonymous (Ds) and fixed non-synonymous (Dn) is used in conjunction with the numbers of polymorphisms (in this test defined as sites with any variation within species) and denoted Pn for non-silent and Ps for silent polymorphisms. Next, using an MK 2x2 contingency table [Mcdonald JH *et al.*, 1991], deviation from the neutrality index (NI = $DsPn/DnPs$ or $(Pn/Ps)/(Dn/Ds)$) can be detected, and indicate positive selection where $Dn/Pn > Ds/Ps$. However, being a ratio of two ratios, NI is undefined when either Dn or Ps is 0 and tends to be biased and to have a large variance, especially when numbers of observations are small [Stoletzki N *et al.*, 2011]. The DoS measure does not have these issues, and so is suitable when the data is sparse. DoS is calculated as $Dn/(Dn + Ds)$ - $Pn/(Pn + Ps)$, and significance of deviation from neutrality can still be calculated using a MK 2x2 contingency table.

To classify genes undergoing diversifying selection, Gene ontology (GO) mapping was performed on each of the predicted *Bd* JEL423 proteins using the Blast2Go server [Conesa A *et al.*, 2005]. In addition, proteins were searched for homology to peptidases and secretion signals using Merops [Rawlings ND *et al.*, 2010] and SignalP 3.0 [Bendtsen JD *et al.*, 2004] respectively.

The Average Read Depth (ARD) over genes was calculated to look for evidence of copy number variation (CNV). To compare the distributions of ARD, we performed pairwise, two-sided Kolmogorov-Smirnov (KS) tests, which is a nonparametric test with a null hypothesis that the samples are being drawn from the same distribution. To account for differences in the number of reads generated for each sample, we divided the ARD per gene by the average depth in the alignment.

3.3.5 Recombination and break points in phylogeny

To summarise the locations of genomic differences in the alignments, we tallied polymorphisms using a non-overlapping sliding window. This information was collected and printed into a dataframe using a Perl script, which then generated the figures using R. Window lengths of 1.4Kb and 10Kb were used, and counts for homozygous SNPs and heterozygous positions were taken individually across the genomes (consisting of two plots per genome or contig). These counts were also summarised into single plots by taking the number of homozygous SNPs minus heterozygous positions.

Separately, Dr. Weinert performed a genetic algorithm for recombination detection

(GARD) analysis [Pond SLK *et al.*, 2006] on an alignment of the ECA positions (22,181 sires) for the 20 genomes, with ambiguous character codes used to represent heterozygous SNPs. The analysis treats these ambiguous (heterozygous) positions as partially missing data during phylogenetic analyses. Briefly, maximum likelihood (ML) trees were made for each alignment fragment of the full sequence. ML is based on an explicit probabilistic model of the process of base substitution, in contrast to maximum parsimony that weighs each substitution equally. Next, a genetic algorithm (GA) is used to search for breakpoint locations that result in different branches in the resulting tree. The GA terminates if the best score remains unchanged over 100 consecutive generations.

The General time reversible (GTR) [Tavaré S, 1986] model of nucleotide substitution with a 4-bin general discrete distribution to account for site-to-site rate variation was employed to evaluate the goodness-of-fit for competing topologies. The optimum number of topologies (or breakpoints) was selected by the minimum value obtained from small-sample corrected Akaike Information Criterion (AICc) [Sugiura N, 1978; Pond SLK *et al.*, 2006]:

$$AICc = 2k - 2ln(L) + \frac{2k(k+1)}{n-k-1} \tag{3.1}$$

Where k is the number of parameters in the statistical model and L is the maximized value of the likelihood function for the estimated model. AICc therefore does not test the model in terms of a null hypothesis, and instead gives a measurement based on penalties for overfitting (using many parameters) and rewards for goodness-of-fit (the model to the data).

Topological incongruence between all neighbouring fragments was examined by sequential Kishino-Hasegawa (KH) (10,000 replicates) testing [Hasegawa M *et al.*, 1989] with Bonferroni correction for multiple testing. KH calculated the variance between generated trees (as an alternative to the more computationally expensive bootstrapping process), and Bonferroni correction controls for the probability of false positives. After assessing the most likely breakpoints and resulting tree topologies using these methods, we reconstructed separate topologies for all breakpoint segments using Bayesian phylogenetic inference techniques implemented in BEAST [Drummond AJ *et al.*, 2007].

3.3.6 Experimental assessment of host response to challenge by *Bd* isolates

Complementing the genome-analysis, a number of the isolates that had been sequenced were also used for *in vivo* experiments by Dr. Bielby and Dr. Garner. These experimental procedures were subject to ethical review at the Institute of Zoology and Imperial College and licensed by the UK Home Office. Briefly, common toad (*Bufo bufo*) larvae (Gosner stage 24, Gosner 1960) hatched from 12 egg strings were randomly allocated to one of ten experimental treatment groups (9 *Bd* isolates and 1 negative control). Isolates used were IA042, IA043, 0711 (Pyrenees), VAo2, VAo4, VAo5 (Valencia), CCB1, TF5a1 and TF1.1 (Mallorca). While isolates IA043 and TF1.1 were not included in the phylogenetic component of this study, previous work has shown that IA043 clusters with other Pyrenean isolates [Walker SF *et al.*, 2010] and that TF1.1 clusters with isolates derived from Mallorcan amphibians [Walker SF *et al.*, 2008].

The experiment was undertaken in a climate-controlled (approx. 16°C constant temp) room with a 12:12 hour day/night light schedule. Tadpoles in *Bd+* treatments were exposed to high doses (3000 - 17,000 active zoospores per exposure in liquid media) of *Bd* zoospores every four days a total of eight times. Zoospore counts and volume of media were standardized among isolates for each exposure. Tadpoles in the negative control treatment were exposed on the same schedule to an equivalent volume of sterile media as was used for *Bd+* treatments, but lacking any *Bd*.

The experiment lasted 122 days after administering the first dose. We recorded time to metamorphosis, mass at metamorphosis and either time to death or survival to the end of the experiment for each animal. Metamorphosis was classified as both forelimbs emerged, tail reduced to a stub (Gosner stage 45) and the first day where we observed that the individual had moved persistently out of standing water. All animals surviving to the end of the experiment were humanely euthanised and all experimental animals were preserved in 70% ethanol for quantitative PCR (qPCR) analysis of infection status and fungal load. One hind limb foot of each experimental animal was used for qPCR amplification following the protocol of Garner *et al.* [Garner TWJ *et al.*, 2009] and Boyle *et al.* [Boyle DG *et al.*, 2004]. All samples were tested in duplicate and against a set of known standards. In the event that only one replicate from any sample did amplify, this sample was run a third time. If this third amplification attempt did not result in an amplification profile, the sample was

scored as negative for infection.

3.3.7 Phenotypic and drug susceptibility testing

In order to investigate whether the isolates *Bd* used in this study showed variation in phenotype or drug susceptibility, Frances Clare conducted three experiments to determine whether sporangia diameter, the length of pseudohyphae, and the susceptibility to the antifungal drug Itraconazole[1] varied amongst the isolates. For each experiment we used twelve isolates (IA042, IA043, 0711.1 VAo2, VAo4, VAo5, MCT8 and JEL423, TF1.1, TF5a1 and CCB1 and 0739). All isolates were grown and maintained in TGhL broth at 18°C [Longcore JE *et al.*, 1999] within 75ml tissue culture flasks (NUNC).

Initial photographs were taken of the floor of the tissue culture flask where sporangia had settled three days post culture using a canon EOS 350D (3456 x 2304 pixel field). We sub cultured the isolates into 12 well tissue culture plates (NUNC), with three replicas per isolate at a concentration of c. 8,000 zoospores per well (calculated using an improved Neubauer haemocytometer). The bottoms of the wells were photographed at 10, 15 and 20 days post-initial culture. Two to three images were obtained from each well for each isolate. Using ImageJ software [Abramoff MD *et al.*, 2004] we measured the diameter of the largest sporangia contained in the field of view ($n = 20$). Measurements taken on day 20 were used to compare sporangia diameter, as at this time point all sporangia appeared mature. Individual isolate results were grouped and analysed using a series of Wilcoxon Rank Sum Tests, to account for uneven distribution, using R [R Development Core Team, 2009].

Photographs from 10 days post initial culture were used to measure hyphal growth of sporangia. For hyphae, 10 days post initial culture was chosen over the 20 days for sporangia as hyphae grew more quickly, and by 20 days could no longer be individually measured. The longest length of hyphae protruding from each sporangia was measured for each isolate (n = 7-25) using ImageJ software. Results from each isolate were analysed using a series of Wilcoxon Rank Sum Tests using the statistical package R.

The susceptibility of isolates to the antifungal drug Itraconazole was also determined by calculating the Midpoint Cytotoxicity (MC_{50}), which is the concentration of a given drug at which the growth of an organism is reduced by 50%, relative to control. We determined

[1]Itraconazole is a triazole antifungal agent commonly used to treat captive amphibians [Tamukai K *et al.*, 2011; Brannelly LA *et al.*, 2012].

the MC_{50} to Itraconazole for all 12 isolates. A 10μg/ml solution was formulated and concentrations of Itraconazole ranging from 0.07μg/ml to 2.1875ng/ml were used [Berger L *et al.*, 2009]. The culture of IA043 had the lowest zoospore concentration ($5.75x10^4ml^{-1}$), which was used as the standard concentration to which the concentration of the subsequent 11 isolates was generated. Using 96 well plates, decreasing dilutions of Itraconazole within 100μl culture were placed in each well (0.07μg/ml, 0.035μg/ml, 0.175μg/ml, 0.00875μg/ml, 0.004375μg/ml and 0.0021875μg/ml), followed by an estimated 5750 *Bd* zoospores in 100μl of culture. The plate also included a negative control (100μl of given *Bd* and 100μl of media) and a blank (200μl media alone). Each isolate was replicated three times.

We incubated plates at 18°C and read optical densities using a BioTek Absorbance Microplate Reader (ELx808), using an absorbance of 450nm. Readings were taken on days 3, 5, 8, 10 and 12. We analysed results generated from day 10 as between days 8-10 results had plateaued and by day 12 there appeared to be some clumping of *Bd* within some wells (Visualised under the microscope), which distorted the results. The MC_{50} was calculated using the curve created with the mean of the three replicas per Itraconazole.

3.3.8 Patterns and timing of *Bd* spread

Because our isolates were collected over a ten-year time period, this allowed us to calculate a mutation rate based on the observed mutations, which have accumulated over this time on the branches of the phylogeny. As the genealogies were reconstructed under a coalescent process, this information could be used to infer the timing of divergence between the isolates. Firstly, due to our inability to resolve the individual sequences (the phase of the sequence) of the two different haploid chromosomes (assuming diploidy), Dr. Weinert tailored a model of evolution. First, each possible diploid genotype at a site were individually coded, yielding 10 possible states in all (AA, CC, GG, TT, AC, AG, AT, CG, CT and GT). Their evolution was then modelled by a time-reversible transition matrix, just as would be the case for a standard 4-state DNA substitution model although in this case, some of the changes might represent both substitutions and gene conversion processes. However, we did not allow double changes to take place instantaneously (For example, AG could not evolve to CT in a single step). Accordingly, 21 of the 45 matrix entries were fixed at zero, leaving 24 parameters to estimate in total.

The analyses were run with a prior gamma distribution of rates [Lemey P *et al.*, 2009]

assuming a strict clock (a relaxed clock with an uncorrelated lognormal distribution of rates did not give a significantly better fit to the data). All other default parameters were used although the clock rate was uniformly bounded between 0 and 1. Two separate MCMCs were run for each breakpoint segment alignment to check for convergence to the same region of parameter space and a proportion of each chain was discarded as burn-in, all visualised using the program Tracer v1.5 [Drummond AJ *et al.*, 2007].

3.4 Results

3.4.1 SNP discovery and phylogenetics

Using ABI SOLiD we achieved a mean 9.5X deep coverage of the c. 24Mb nuclear genomes for 20 globally distributed (Europe, North and Central America, South Africa, Australia) isolates of *Bd* from eleven amphibian host species. Eight isolates were from regions where epizootics have been documented (Table 3.1). We aligned the trimmed (30mer) reads to the genome sequence of isolate JEL423 [MIT JEL423 sequence] and searched for discrepancies using the depth-dependent binomial method (See Chapter 2 and Section 3.3.2) finding in total 51,915 non-redundant (unique loci) of homozygous SNPs and 87,121 non-redundant heterozygous positions in the nuclear genome (Fig. 3.3A). Of these, 21% of the homozygous SNP positions and 19% of the heterozygous positions (22,181 polymorphic positions in total) were covered $\geq$4 reads in all 20 samples (Entirely Covered in All; ECA), shown in Fig. 3.3B, which were therefore suitable for using in phylogenetic analysis.

Extensive overlap of these ECA positions was found by calculating for each of the 20 samples ((sample1 $\cap$ sample2) / ((sample1 + sample2) / 2)) x 100. In general, there were greater levels of variation between SNPs than between heterozygous positions (Table 3.2). For example, as much as 91% of SNPs were independently found in isolate CCB1 (ES) and isolate MC55 (ZA), whilst fewer (55%) heterozygous positions were at shared locations. The maximum overlap of heterozygous positions was 63% between PNP08489 (FR) and VA04 (ES). Importantly, sixteen of the twenty samples, including the reference strain JEL423, were >99.9% genetically identical. ECA sites were next used to generate UPGMA trees, demonstrating these isolates to belong to a single highly supported clade (Fig. 3.4). This lineage includes all previously genotyped isolates of *Bd* and all the isolates in our panel that are associated with regional epizootics, recovered from five continents [James TY *et al.*

2009]. We therefore named this lineage the 'Global Panzootic Lineage' (*Bd*GPL).

In addition to the genome alignments, we aligned our 30mer reads to the 175,284nt *Bd* JEL423 mitochondrial genome (Table 3.3). We achieved 23-87X deep alignments (mean 51X) from reads that were >99.2% un-aligned to the nuclear genome, indicating a high specificity to the mitochondrial genome. We again used the SNP calling method based on 3 binomial distributions with a minimum required read depth of 4. In total, we identified 231 ECA non-redundant unique polymorphic SNPs in the mitochondrial sequences, only 2 of which were from the *Bd*GPL. The first was found at position 108,615 (reference T → consensus C), which was independently found in 4 isolates taken from a single pond in Valencia (VAo2, VAo4, VAo5 and VRp1) and the second, at position 152,201 (reference A → consensus G) was found in the isolate from Kent, England (UKTvB). We identified between 114 and 141 SNPs in the isolates belonging to the two divergent lineages, supporting their divergence identified in the nuclear genome. Therefore, both nuclear and mitochondrial alignments supported three distinct lineages of *Bd*.

The remaining four newly sampled isolates formed two, novel and deeply divergent, highly supported lineages. One of these lineages comprised two isolates from the island of Mallorca, and one from the Cape Province, South Africa. Dr. Susan Walker *et al.* had previously hypothesised that spillover of *Bd* from captive South African Cape Clawed Frogs (*Xenopus gilli*) into Mallorcan Midwife Toads (*Alytes muletensis*) led to the introduction of *Bd* onto the island through a captive breeding and reintroduction program at the Durrell Wildlife Conservation Trust (DWCT), Jersey in 1991 [Walker SF *et al.*, 2008]. The relatedness of South African and Mallorcan isolates support Walker's hypothesis, and we therefore named this lineage *Bd*CAPE after its possible origin. Furthermore, the discovery of *Bd*CAPE as a separate lineage to previously genotyped isolates demonstrates that multiple emergences of amphibian chytridiomycosis have occurred, as well as confirming that anthropogenic spread of amphibians is directly responsible for at least one of these emergences.

A third novel lineage ('Swiss Lineage'; *Bd*CH) is composed of a single isolate derived from a Common Midwife Toad (*Alytes obstetricans*) from a pond near the village of Gamlikon, Switzerland. Further sampling is necessary to establish whether *Bd*CH is a European-endemic isolate, or more broadly distributed.

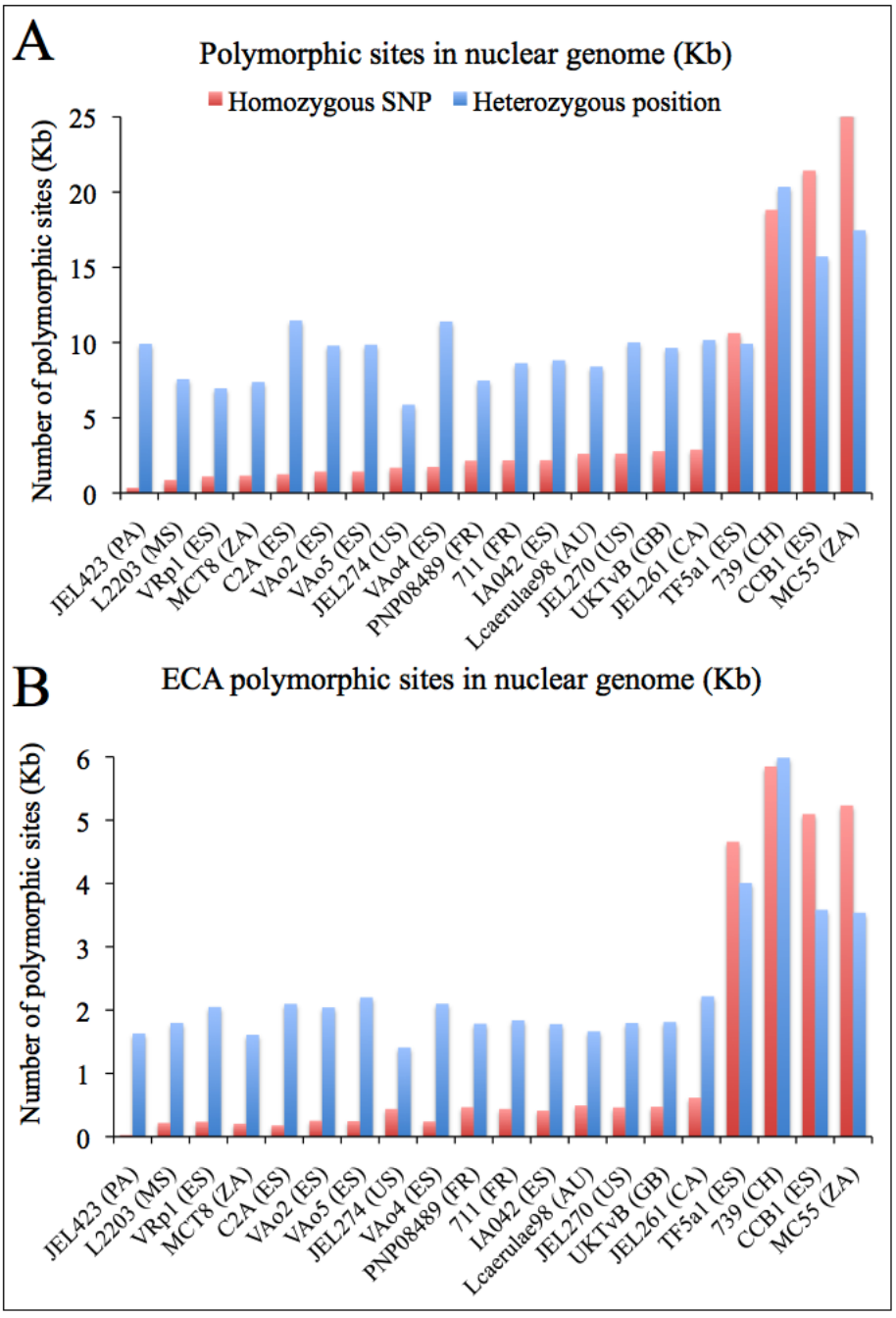


Figure 3.3: The total numbers of polymorphic nucleotide positions found in 20 nuclear genomes of *Bd*. (**A**) Sixteen isolates had between 0.351Kb and 2.883Kb homozygous SNPs and four isolates had >10Kb homozygous SNPs. A greater number of heterozygous positions (>5.87Kb) were found for all 20 genomes. (**B**) Polymorphic sites that were covered by ≥ 4 reads in all 20 nuclear genomes (ECA) closely resembled the total number of polymorphic sites, which were used for phylogenetic analysis.

Homozygous SNPs:

Isolate	L2203 (MS)	VRp1 (ES)	MCT8 (ZA)	C2A (ES)	VAo2 (ES)	VAo5 (ES)	JEL274 (US)	VAo4 (ES)	PNP08489 (FR)	711 (FR)	IA042 (ES)	Lcaerulae98 (AU)	JEL270 (US)	UKTvB (GB)	JEL261 (CA)	TF5a1 (ES)	CCB1 (ES)	MC55 (ZA)	739 (CH)
JEL423 (PA)	13	11	13	15	11	10	6	12	8	8	8	7	7	7	4	1	1	1	1
L2203 (MS)		25	25	9	23	27	19	26	16	8	9	28	8	12	15	4	4	3	3
VRp1 (ES)			50	9	79	79	29	88	18	10	10	19	12	13	15	4	4	5	4
MCT8 (ZA)				9	42	41	28	51	19	10	10	22	17	12	13	4	4	4	3
C2A (ES)					10	9	6	10	8	7	8	6	18	20	17	3	3	3	3
VAo2 (ES)						80	30	79	15	8	8	15	7	17	13	5	5	5	4
VAo5 (ES)							30	81	18	10	9	15	6	17	16	5	4	5	4
JEL274 (US)								31	16	7	8	35	13	21	24	7	6	6	5
VAo4 (ES)									17	9	9	20	11	12	13	4	4	4	4
PNP08489 (FR)										78	79	13	17	41	33	6	7	6	5
711 (FR)											87	6	23	38	34	6	6	6	5
IA042 (ES)												6	23	39	34	6	6	6	5
Lcaerulae98 (AU)													13	21	33	7	7	7	6
JEL270 (US)														9	32	7	7	6	6
UKTvB (GB)															41	7	6	6	5
JEL261 (CA)																9	8	8	7
TF5a1 (ES)																	87	85	29
CCB1 (ES)																		91	31
MC55 (ZA)																			31

Heterozygous positions:

JEL423 (PA)	52	49	55	56	49	49	45	52	41	44	43	60	49	51	50	7	6	6	6
L2203 (MS)		46	48	50	40	48	42	44	41	41	38	52	48	46	43	7	6	6	6
VRp1 (ES)			47	52	60	61	38	63	43	45	44	49	44	46	53	7	7	7	7
MCT8 (ZA)				48	47	46	42	48	40	39	39	53	45	46	47	7	7	6	6
C2A (ES)					50	56	40	53	47	52	50	52	56	50	57	8	8	8	7
VAo2 (ES)						59	36	63	43	45	45	45	43	45	53	8	7	7	7
VAo5 (ES)							40	63	47	48	48	49	50	49	55	8	7	7	7
JEL274 (US)								40	33	35	35	46	41	40	39	6	6	5	5
VAo4 (ES)									45	47	46	50	44	49	57	8	7	7	7
PNP08489 (FR)										57	55	37	44	44	50	7	8	8	6
711 (FR)											61	38	49	47	54	8	8	8	6
IA042 (ES)												39	48	47	53	8	8	8	6
Lcaerulae98 (AU)													50	48	48	6	6	6	7
JEL270 (US)														51	46	7	7	7	7
UKTvB (GB)															53	8	7	7	6
JEL261 (CA)																9	8	8	7
TF5a1 (ES)																	49	44	8
CCB1 (ES)																		53	8
MC55 (ZA)																			8

Table 3.2: Percent of shared entirely covered in all (ECA) homozygous SNPs (top) and percent of shared ECA heterozygous positions (below).

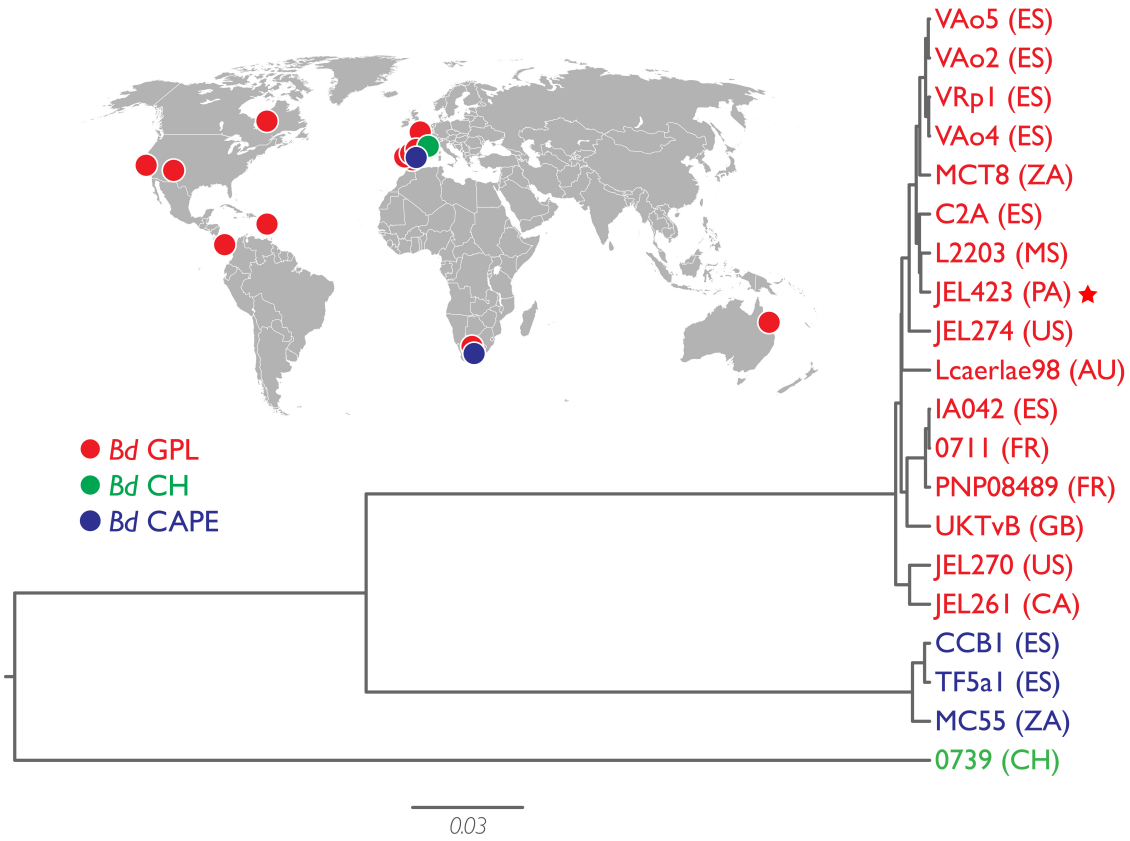


Figure 3.4: Tree of the nuclear genomes made using the UPGMA algorithm in PAUP. The three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH) identified from the mitochondrial genome analysis also occur in the nuclear genome comparisons. * Signifies the JEL423 genome isolate

Culture reference	Reads aligned	Depth (X)	Uniquely aligned	SNPs	heterozygous
0711 (FR)	140,341	23	99.83	0	4
C2A (ES)	497,492	82	99.42	0	1
IA042 (ES)	179,953	30	99.68	0	1
JEL261 (CA)	237,849	39	99.57	0	2
JEL270 (US)	435,666	72	99.51	0	0
JEL274 (US)	152,453	25	99.63	0	5
JEL423 (PA)	297,956	49	99.53	0	2
L2203 (MS)	313,058	52	99.71	0	1
Lcaerulae98 (AU)	420,971	70	99.23	0	0
MCT8 (ZA)	204,824	34	99.65	0	3
PNP08489 (FR)	331,753	55	99.74	0	2
UKTvB (GB)	360,674	60	99.65	1	2
VAo2 (ES)	379,503	63	99.62	1	5
VAo4 (ES)	290,377	48	99.74	1	1
VAo5 (ES)	322,774	53	99.57	1	0
VRp1 (ES)	333,700	55	99.62	1	5
739 (CH)	250,181	41	99.94	114	39
TF5a1 (ES)	212,978	35	99.80	137	50
MC55 (ZA)	252,421	42	99.84	139	47
CCB1 (ES)	526,968	87	99.81	141	40

Table 3.3: Alignments to *Bd* JEL423 mitochondrial sequence had an average depth of 51X deep from reads that were >99.2% uniquely aligned to the mitochondrial genome. We identified ECA homozygous SNPs in 9/20 isolates using BiSCaP v0.1. In total, 306 non-redundant polymorphic sites were identified in the mitochondrial genome with our panel of 20 isolates.

3.4.2 Mutation patterns and copy number variation (CNV)

To ascertain whether genomic features were associated with these lineages, we searched for evolutionary differences by looking for mutation biases, copy number variation (CNV) and genes undergoing diversifying selection in the *Bd*CAPE and *Bd*CH lineages when compared to *Bd*GPL. First, ECA polymorphic sites identified in the CDS regions were categorised into non-strand specific transition (Ts) and transversion (Tv) rates (shown in Fig. 3.5A). Most (89%) of these polymorphic sites caused a transition (A:T↔G:C) with a mean Ts/Tv of 8.03. This finding is consistent with Ts/Tv rates for other eukaryotes [Belle EMS *et al.*, 2005], caused by known molecular mechanisms such as spontaneous deamination of methylated cytosines [Ebersberger I *et al.*, 2002]. The standard deviation (shown as error bars) between the 20 isolates was within a narrow range for transitions (1.88-2.1) and even more so for transversion (0.15-0.27). Therefore, the mechanism causing transition bias is acting on each of the isolates independently of divergence from the reference strain JEL423 as shown in Fig. 3.5B (i.e. no lineage specific biases).

We further categorised each of the mutations into synonymous and non-synonymous mutations (Fig. 3.6). SNPs in gene encoding regions at homozygous positions (Fig. 3.6A) accounted for 21,030 non-redundant synonymous changes and 9,520 non-redundant non-synonymous changes (Ratio 2.21). Similarly, SNPs at heterozygous positions (Fig. 3.6B) accounted for a greater number of non-redundant synonymous changes (25,758) than non-redundant non-synonymous changes (20,039, ratio 1.29). With the exception of the reference strain *Bd* JEL423, the ratios of synonymous to non-synonymous change were between 1.66-2.4 and 1.04-1.92 for homozygous and heterozygous positions respectively. The greater rates of synonymous change are concordant with the fact that most transitional mutations at twofold degenerate sites are synonymous.

To measure selection across the genome we used the DoS measure of selection [Stoletzki N *et al.*, 2011] with the McDonald-Kreitman 2x2 contingency table [Mcdonald JH *et al.*, 1991]. Polymorphisms was calculated from ECA heterozygous and homozygous positions in the global lineage while fixed differences were from the homozygous SNPs in 0739 (*Bd*CH lineage) or CCB1 (*Bd*CAPE lineage). Overall per gene polymorphism was low with an average of 1.6 polymorphic sites per gene calculated and over 20% of those polymorphisms were still segregating between the two groups. We calculated DoS for a total of 753 genes in the *Bd*CH to *Bd*GPL comparison and 633 genes in the *Bd*CAPE comparison. In the *Bd*CH comparison, 23.0% of genes had DoS >0 and 37.2% had DoS <0. In the *Bd*CAPE compari-

son, 18.9% of genes had DoS >0 and 35.8% had DoS <0. The numbers of polymorphic sites and fixed differences were nearly identical in the *Bd*CAPE comparison, 1024 polymorphic sites with 1025 fixed differences, and for the *Bd*CH comparison, 1230 polymorphic sites with 1212 fixed differences. Each lineage shared a similar percentage of polymorphic sites with the *Bd*GPL lineage, 9.4% of total sites in the Swiss lineage and 9.9% of sites in the Cape lineage. Only five genes were significantly different from neutral expectations at a 0.05% probability, and each had very few substitutions (Fig. 3.7A).

We also measured the rates of synonymous substitution per synonymous site (dS), non-synonymous substitution per non-synonymous site (dN) and omega ($\omega = dN/dS$) for every gene in isolates CCB1 (*Bd*CAPE) and 0739 (*Bd*CH) using the yn00 program of PAML [Yang Z *et al.*, 2007] implementing the Yang and Nielsen method [Yang Z *et al.*, 2000]. We identified 73 and 41 genes with ω>1 for *Bd*CAPE and *Bd*CH respectively, with only 4 genes identified in both (Transcript ID's 00801, 03487, 04077 and 07589). GO mapping was performed on each of the predicted proteins with ω>1 using the Blast2Go server [Conesa A *et al.*, 2005]. These GO terms were summarised for CCB1 and 0739 in Figs. 3.7B and 3.7C respectively. To further investigate the three *Bd*CAPE isolates (CCB1, MC55 and TF5a1), we increased our read coverage by aligning all the reads from the *Bd*CAPE isolates to JEL423, providing a 28.78X deep alignment. Using BiSCaP v0.1, we could now identify 42Kb homozygous SNPs and 25Kb heterozygous positions. dN/dS values were identified for each of the genes (Fig. 3.8). Those that were predicted proteases by Merops and secreted genes by SignalP were identified, but were not largely distinguishable compared with the rest of the transcriptome.

The composition of bases among genome features and polymorphic sites were compared to identify mutational or experimental bias (Fig. 3.9). The *Bd* JEL423 genome is comprised of 60% adensine and thymine, which differ depending on the overlying features (57.43%, 63.8% and 65.85% for genes, intergenic and introns respectively). ECA bases showed a similar composition for each of the features they covered, but interestingly, were not covering each of the genomic features equally: 32.61% of genes, 12.15% intergenic regions and 11.13% introns. Differences in coverage can be explained most parsimoniously by greater difficulties in alignment of reads over non-coding regions due to repetitive elements. Alternatively, chromosomal length polymorphisms identified between several strains of *Bd* [Morehouse EA *et al.* 2003] and accompanying indels may reduce alignment coverage.

The Average Read Depth (ARD) over genes was calculated to look for evidence of

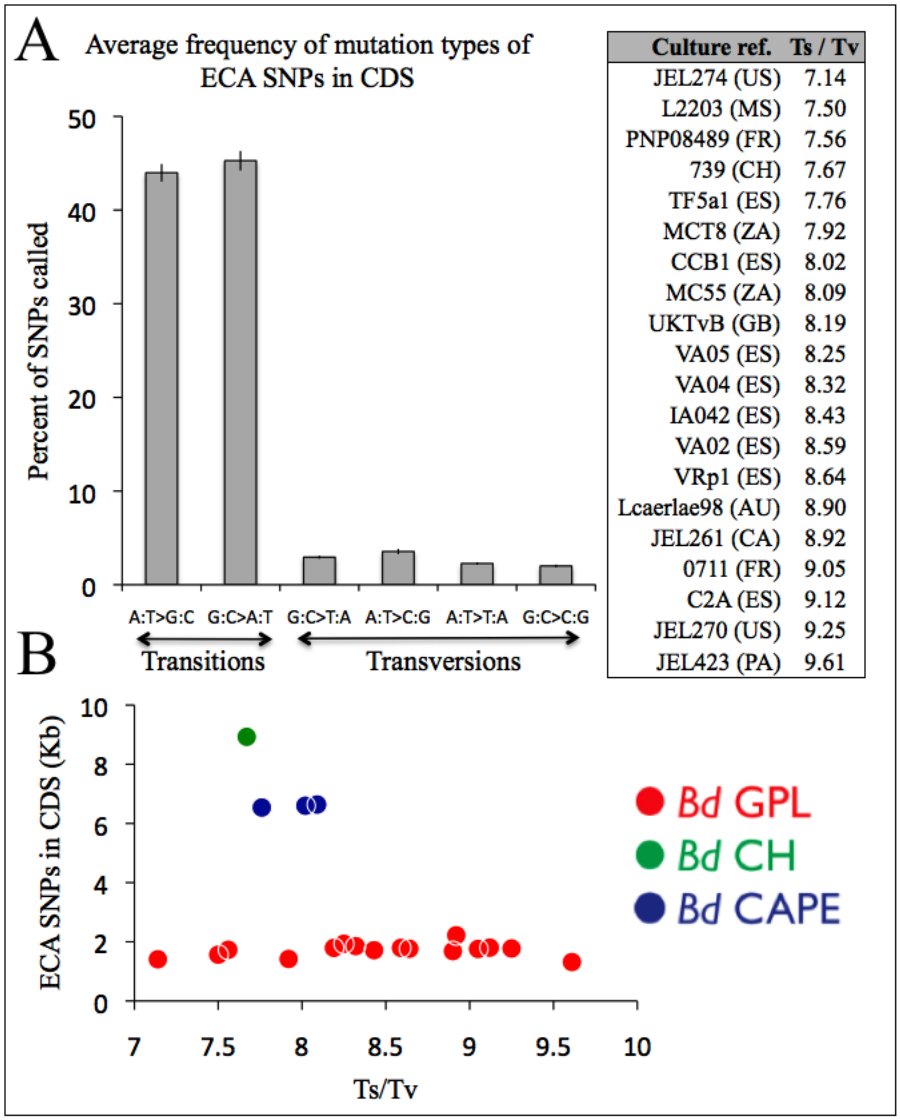


Figure 3.5: ECA polymorphic sites caused a similar excess of transitions in all 20 isolates of *Bd*. (**A**) Average frequency of transition (Ts) and transversion (Tv) rates in the 20 genomes relative to the reference sequence *Bd* JEL423 was calculated using ECA SNPs. For all 20 isolates there was an excess of transitions within a very narrow standard deviations (shown by error bars). (**B**) The Ts/Tv ratio was between 7-10, with little or no differences found between lineages.

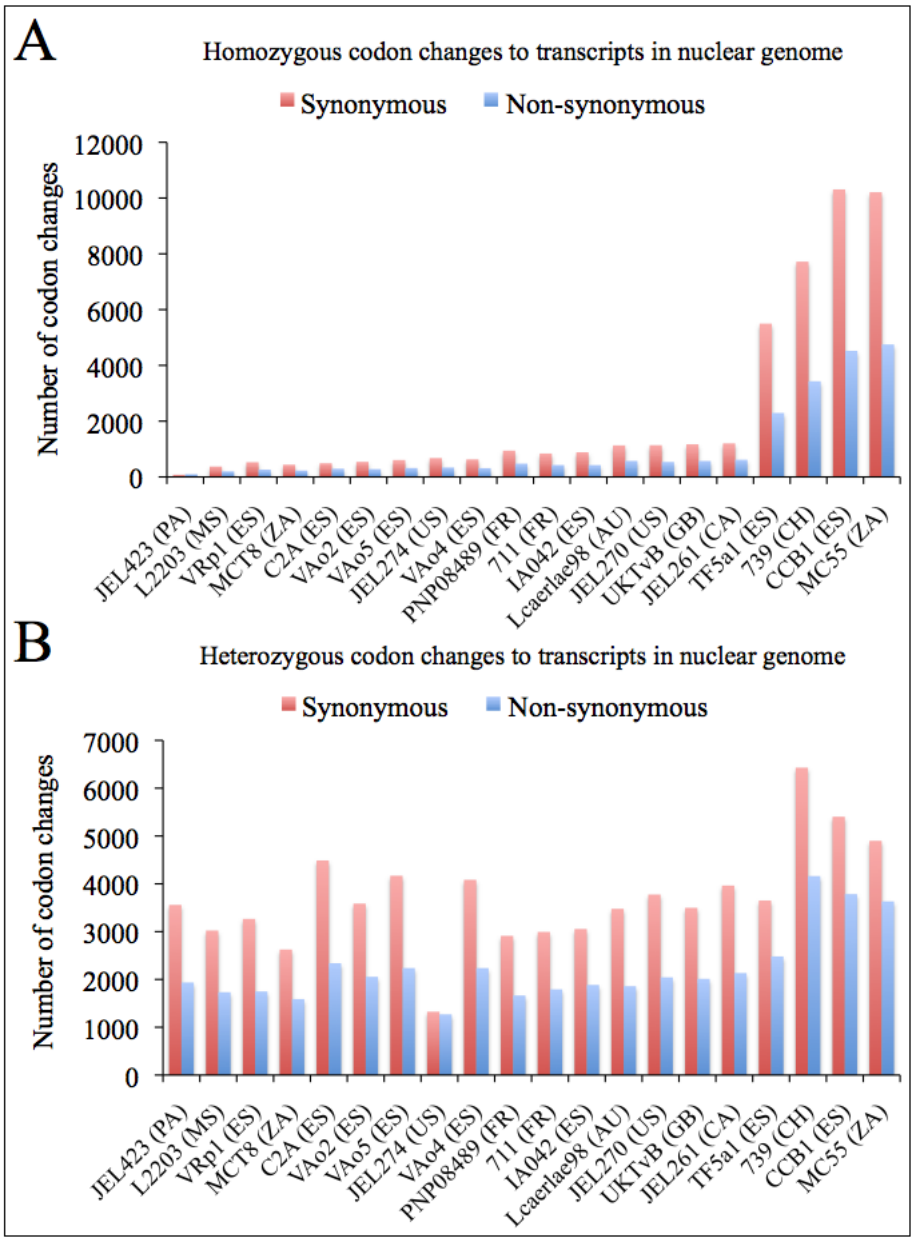
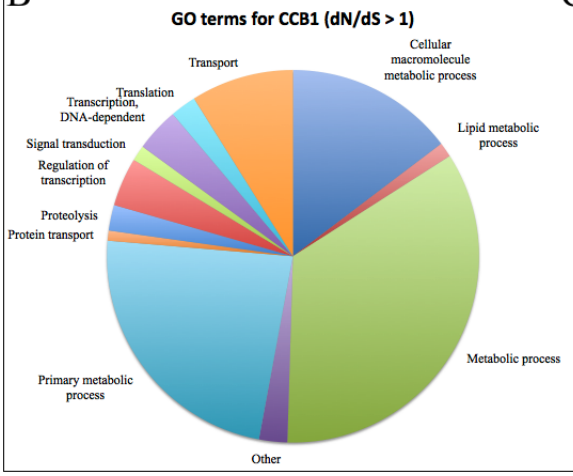


Figure 3.6: Homozygous SNPs identified within the CDS regions were categorised by their effect on the transcripts. (**A**) Approximately twice as many synonymous changes were identified in total as non-synonymous changes. (**B**) Heterozygous positions were also categorised based on their affect to the transcript, again showing similar excess of heterozygous positions.

A

Gene identifier	DoS	P-values	Lineage	PolySyn	PolyNon	FixedSyn	FixNon	Yang dN/dS (ω)	Seq. description
Bdet_04823	1	0.0253	Cape	3	0	0	2	0.2730	nucleoporin 214kda
Bdet_01770	1	0.0455	Cape	2	0	0	2	0.3553	adenylate cyclase-associated protein
Bdet_03893	1	0.0455	Swiss	1	0	0	3	1.168	N/A
Bdet_01538	1	0.0455	Swiss	2	0	0	2	N/A	truncated cell wall surface protein
Bdet_07338	1	0.0455	Swiss	2	0	0	2	N/A	N/A

B



C

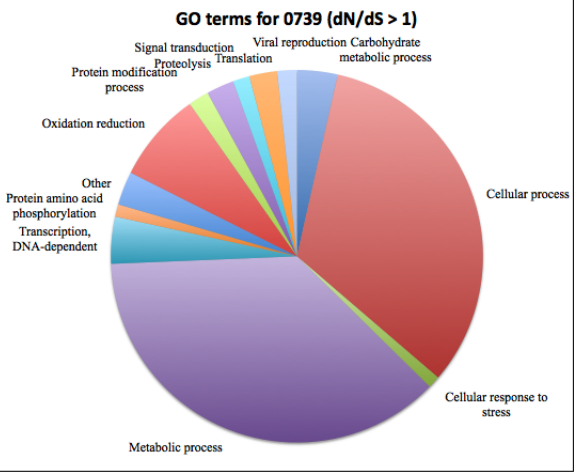


Figure 3.7: Genes undergoing diversifying selection were predominantly metabolic genes. Genes undergoing diversifying selection were identified by (**A**) DoS and (**B and C**) dN/dS >1 for isolates *Bd*CAPE CCB1 and *Bd*CH 0739 respectively. A small number of genes were identified (4 DoS and 110 dN/dS >1) that were categorised by GO terms. Many of the GO terms for divergent genes in *Bd*CAPE CCB1 and *Bd*CH 0739 were involved in metabolism (>50% and >40% respectively).

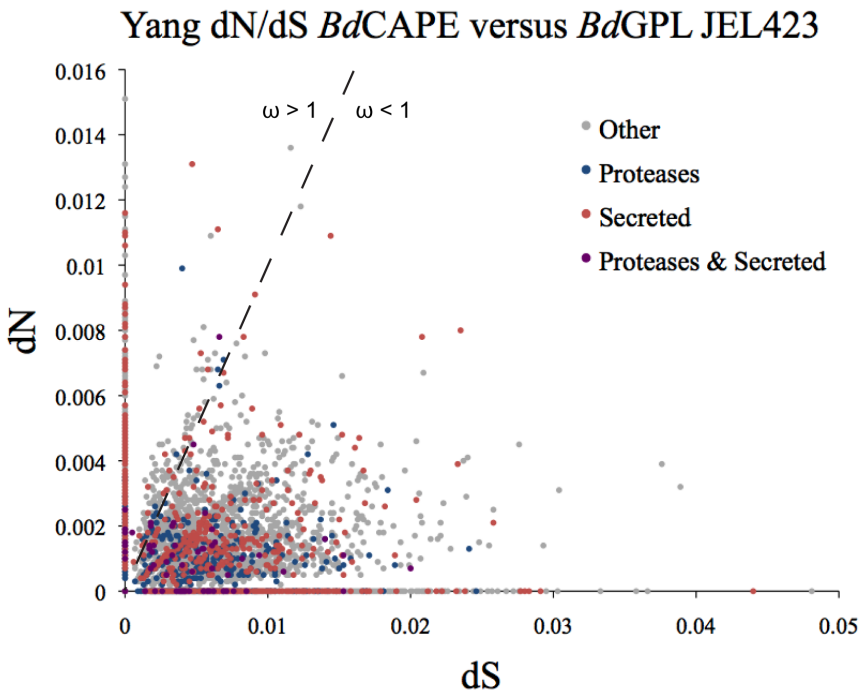


Figure 3.8: *dN*, *dS* and ω values were calculated for all *Bd*CAPE genes from alignments to *Bd*GPL JEL423. Predicted proteases (blue), secreted genes (red) and secreted proteases (purple) and all other genes (grey) show no obvious pattern in terms of selection (where $\omega>1$ = positive or relaxed selection and $\omega<1$ = purifying selection.

copy number variation (CNV; Fig. 3.10). Although there were many outliers (not shown for clarity), no observable 'second peaks' indicative of frequent CNV were identified. To compare the distributions of ARD, we performed pair-wise, two-sided Kolmogorov-Smirnov (KS) tests. The maximum vertical deviation between the cumulative fraction (D) was lowest between isolates L2203 - PNP08489 (0.04) and 0739 - VAo2 (0.05), whilst other KS tests had 0.07 <D<0.9 indicating different ARD distributions for each isolate. To account for differences in the number of reads generated for each sample, we divided the ARD per gene by the average depth in the alignment. After normalization, nine pairwise KS tests had D <0.05 including six that were between lineages (four *Bd*GPL - *Bd*CAPE, one*Bd*GPL - *Bd*CH and one *Bd*CAPE - *Bd*CH), while all KS tests had a D<0.394. Therefore, while there are different distributions of ARD, there are no differences suggestive of lineage specific CNV.

We examined the percent of reads that aligned uniquely to only one position in the JEL423 nuclear and mitochondrial genome (Fig. 3.11). Whilst almost all reads aligned uniquely to the mitochondrial genome (>96%), only between 73.5 and 82% of reads were aligned uniquely in the nuclear genome. Using subsets of the reads that specified SNPs and heterozygous positions, we found a greater percent of uniquely mapped reads over homozygous SNPs, but a lower percent of uniquely mapped reads over heterozygous positions. The lower percent of uniquely aligned reads over heterozygous positions was specific to just the *Bd*CAPE and *Bd*CH isolates (Fig. 3.11B), which may be a result of the greater abundance of heterozygous positions in these lineages compared with *Bd*GPL. To test that the observed decrease in unique reads was not caused by the abundance of homozygous SNPs, we aligned all samples to the consensus sequence for isolate MC55 (ZA), and found a similar (a slight drop in uniquely mapped reads) for all isolates, and the expected fall in homozygous SNPs for *Bd*CAPE isolates. A negative correlation between the date since isolation and the number of heterozygous positions was identified (Fig. 3.12). However, this is largely explained by the recent isolation of the *Bd*CAPE and *Bd*CH isolates.

3.4.3 Recombination and break points in phylogeny

Although no obvious inter-lineage variation was uncovered by these analyses, clear differences emerged when we examined the locations of the polymorphic sites within the genome. To summarise the locations of genomic differences in the alignments we used non-overlapping sliding windows. Using a window length of 1414nt which corresponds to the mean length of

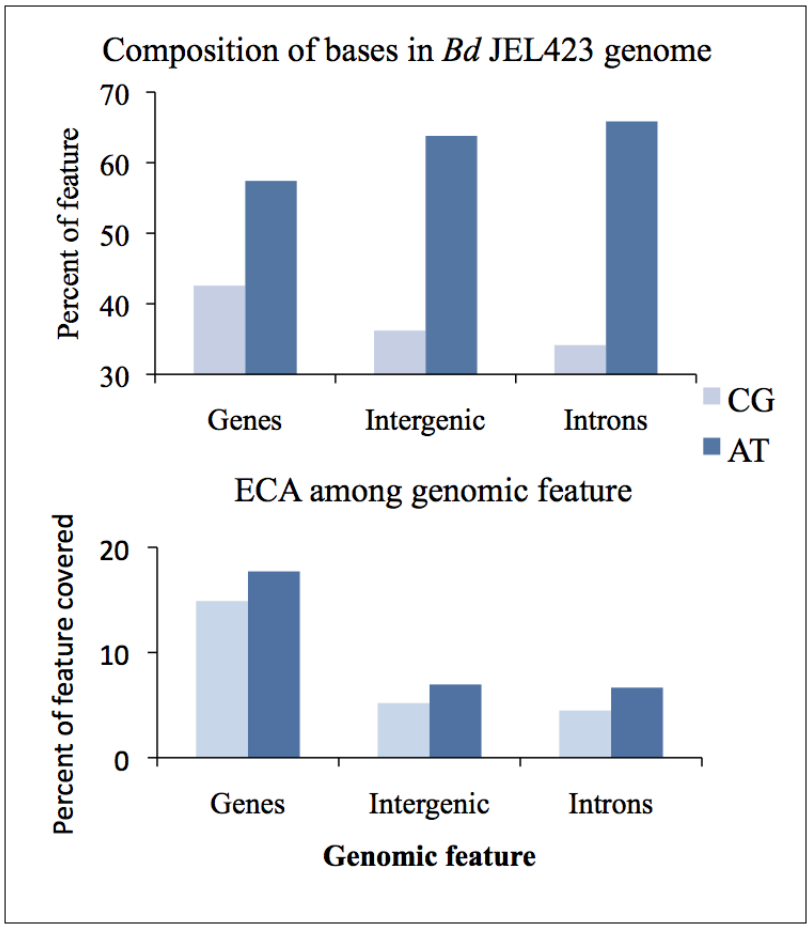


Figure 3.9: The genome assembly of *Bd* JEL423 and ECA positions had a similar AT content bias. (**A**) The composition of bases in the JEL423 genome by genomic features. (**B**) The composition of bases covered by $\geq$4 reads in all 20 samples (ECA). While composition of ECA bases showed a similar AT content bias, they covered a greater percent of CDS than intergenic or introns.

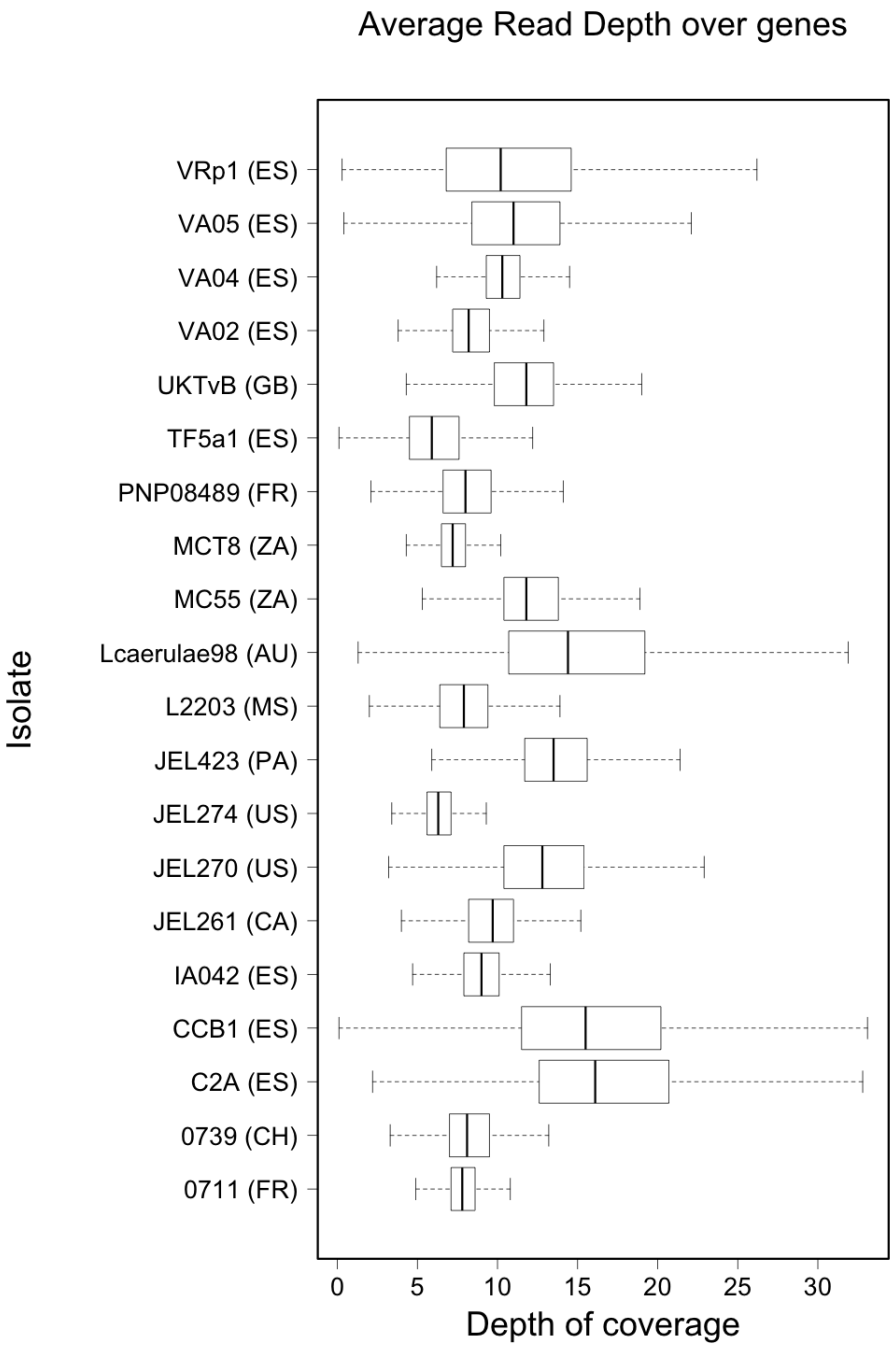


Figure 3.10: The Average Read Depth (ARD) over genes. Boxplots of ARD over all 8795 genes for each of the 20 isolates with outliers omitted. The number of reads sequenced, and therefore aligned, largely affected the distributions. Lineage specific bias was not identified.

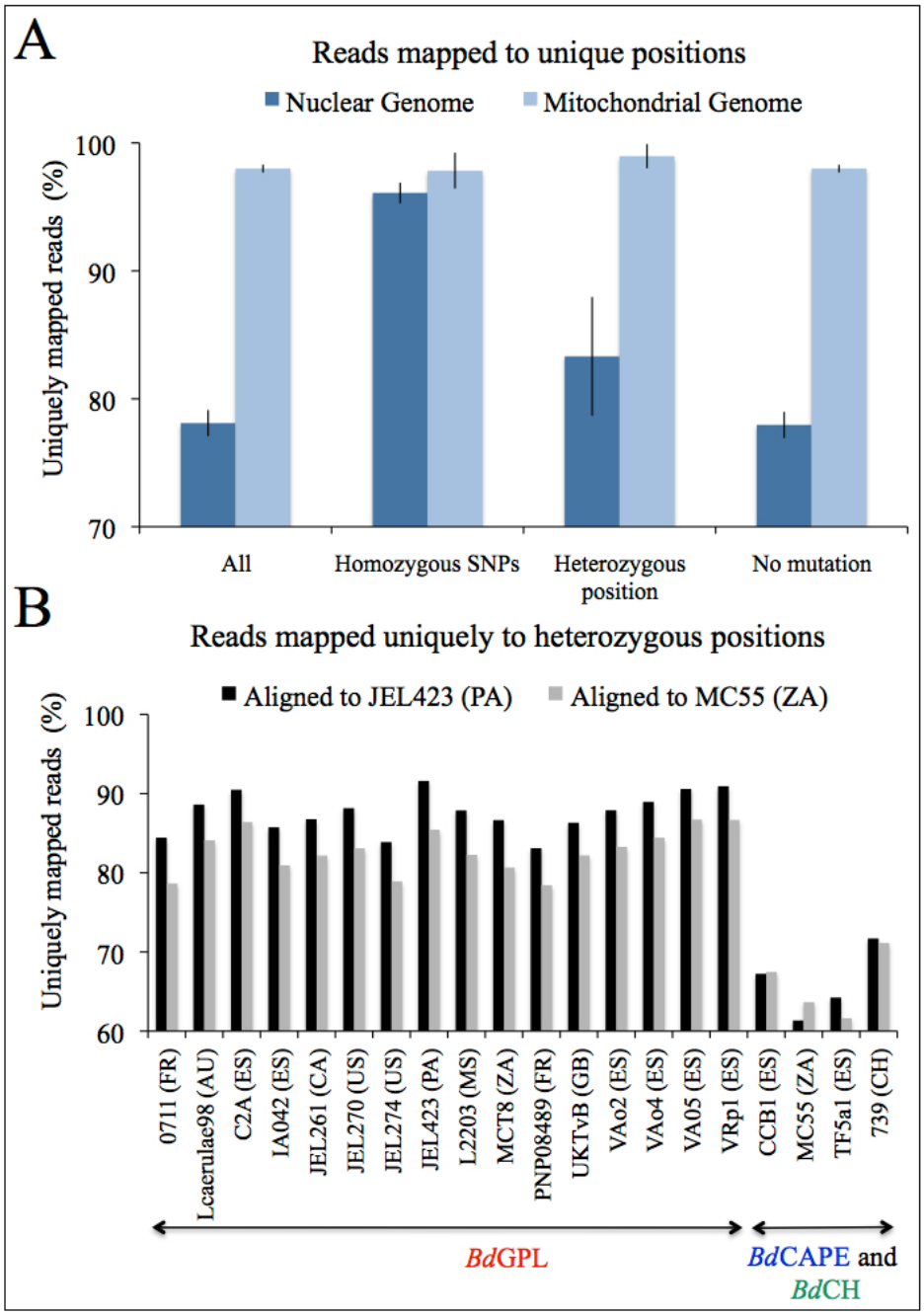


Figure 3.11: Most of the aligned reads mapped to unique locations in both the nuclear and mitochondrial genomes. (**A**) The nuclear genome had a lower proportion of uniquely mapped reads compared to the mitochondrial genome. Reads that aligned to the nuclear genome and specified homozygous SNPs or heterozygous sites for all global lineage isolates were more comparatively (>80%) aligned to unique locations. (**B**) The lowest proportion of uniquely mapped reads was found over heterozygous positions in the nuclear genomes for the isolates comprising the two divergent lineages.

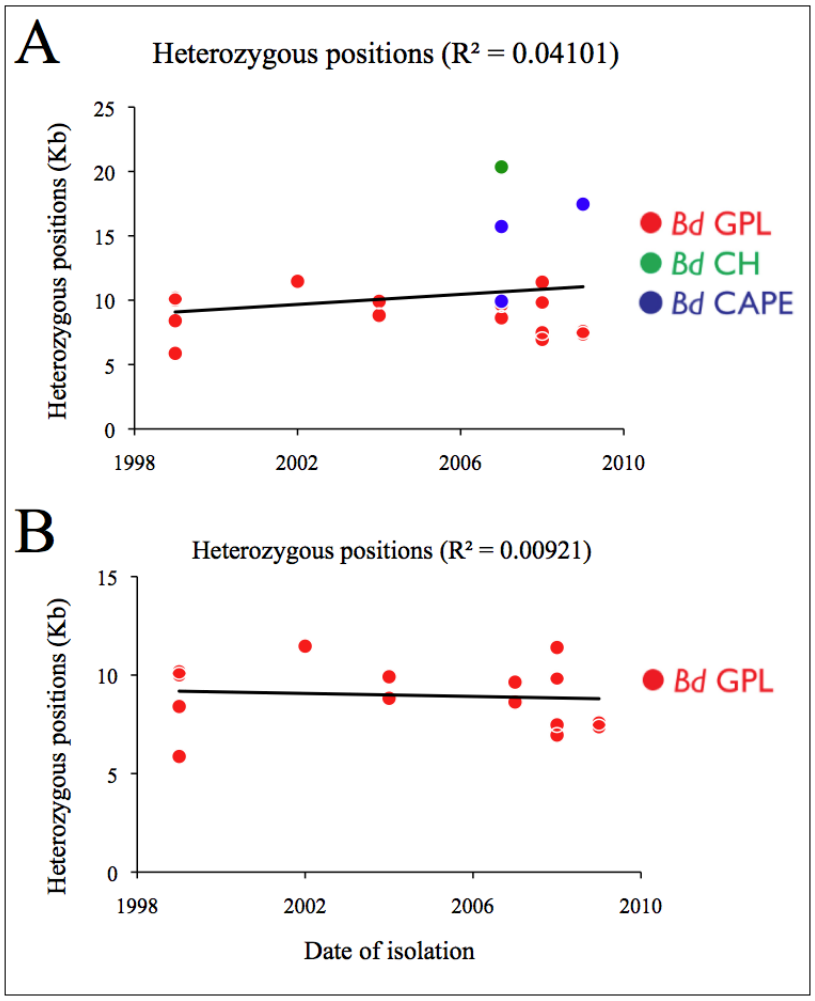


Figure 3.12: Total number of heterozygous positions in each genome belonging to *Bd*GPL was not affected by different dates since isolation. (**A**) A small decrease in the number of heterozygous positions was found since the date of isolation from the host. (**B**) However, most of the trend can be explained by the recently isolated divergent isolates.

Bd ORFs, we identified a highly uneven distribution of homozygous SNPs and heterozygous positions across the 16 *Bd*GPL isolates (Fig. 3.13). To check if this pattern was a function of window length, we re-ran the sliding windows using length 714nt (half length of average CDS) and 2814nt (Fig. 3.14). The same uneven distribution was identified in both of these window lengths. To test *Bd*CAPE for the hallmarks of loss-of-heterozygosity found in the *Bd*GPL, we constructed a consensus sequence from the *Bd*CAPE alignment. By aligning each of the *Bd*CAPE isolates and again using a non-overlapping sliding window, we observed a more even distribution of SNPs (Fig. 3.15), which therefore suggests that recombination has only been occurring within the *Bd*GPL.

Within the *Bd*GPL, the patterns of high or low levels of heterozygosity appeared relatively similar between isolates. For example, we observed a region of low heterozygosity spanning supercontig 2 that was common to all isolates within *Bd*GPL. To investigate if this clustered distribution was due to recombination events, we tested whether the observed pattern was explained by multiple different evolutionary histories using the GARD method [Pond SLK *et al.*, 2006] on an alignment of 22,181 positions comprising all variable sites in the 20 genomes, with ambiguous character codes used to represent heterozygous SNPs.

We identified 26 recombination breakpoints across the nuclear genome, 14 of which remained significant after KH testing (Fig. 3.17 and Table 3.4). Seven of these breakpoints were found within chromosomes and seven were found occurring between chromosomes. This suggests that the independent assortment of individual chromosomes expected under a model of frequent meiosis has not eroded the congruent phylogenetic signal between chromosomes, suggesting that meiosis is rare. Taken together, these data support the hypothesis of a single hybrid origin of *Bd*GPL via an ancestral meiosis as proposed by James TY *et al.* [James TY *et al.*, 2009].

To further characterise the differences between the 15 significantly different recombination segments, we reconstructed separate topologies for each segment, but sharing the same model of evolution, using Bayesian phylogenetic inference techniques implemented in BEAST v.1.6.1 [Drummond AJ *et al.*, 2007] (Fig. 3.17). The three distinct lineages were recovered in all 15 segment trees thereby ruling out recombination between them. In contrast, the topologies amongst recombination segments within *Bd*GPL were markedly different from one another. This pattern can be explained by two alternative biological explanations: 1) different segments have different genealogies reflecting meiotic events or 2) a highly heterozygous ancestor has since undergone recurrent mitotic gene conversion events at different places

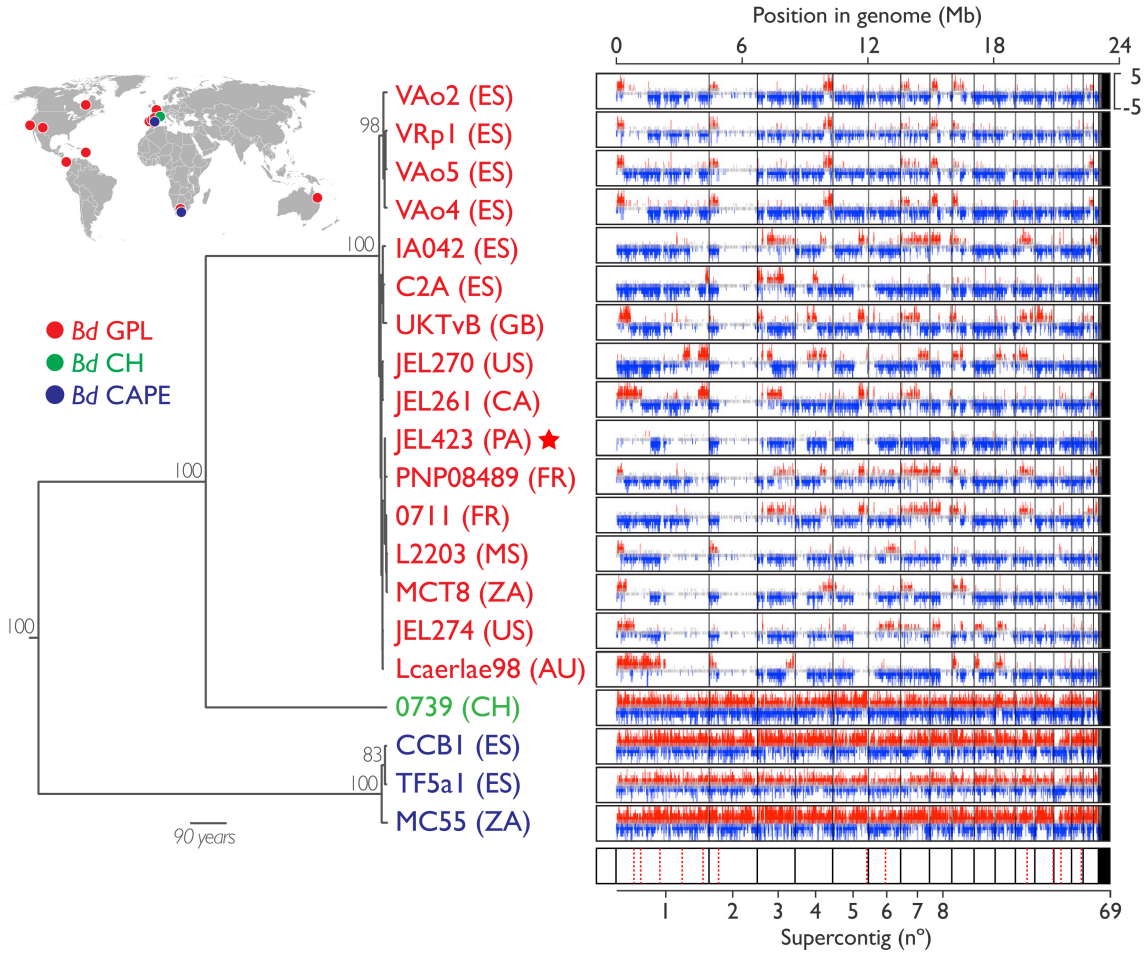


Figure 3.13: Phylogenetic analysis of the 20 resequenced *Bd* mitochondrial genomes demonstrates three divergent lineages. The locations of the isolates belonging to the different lineages are shown using the same colors as in the phylogeny. Each genome is represented next to the right of the phylogeny. A non-overlapping sliding window of SNPs minus heterozygous positions across the genome illustrates regions where heterozygosity predominates (blue) and where homozygosity predominates (red), illustrating the hallmark of loss-of-heterozygosity in the pan-global *Bd*GPL lineage. The block below the 20 genomes denotes the supercontigs with black lines and the GARD recombination breakpoints are shown in red dotted lines. The star signifies the reference genome JEL423, crosses represent isolates that have been recovered from epizootics.

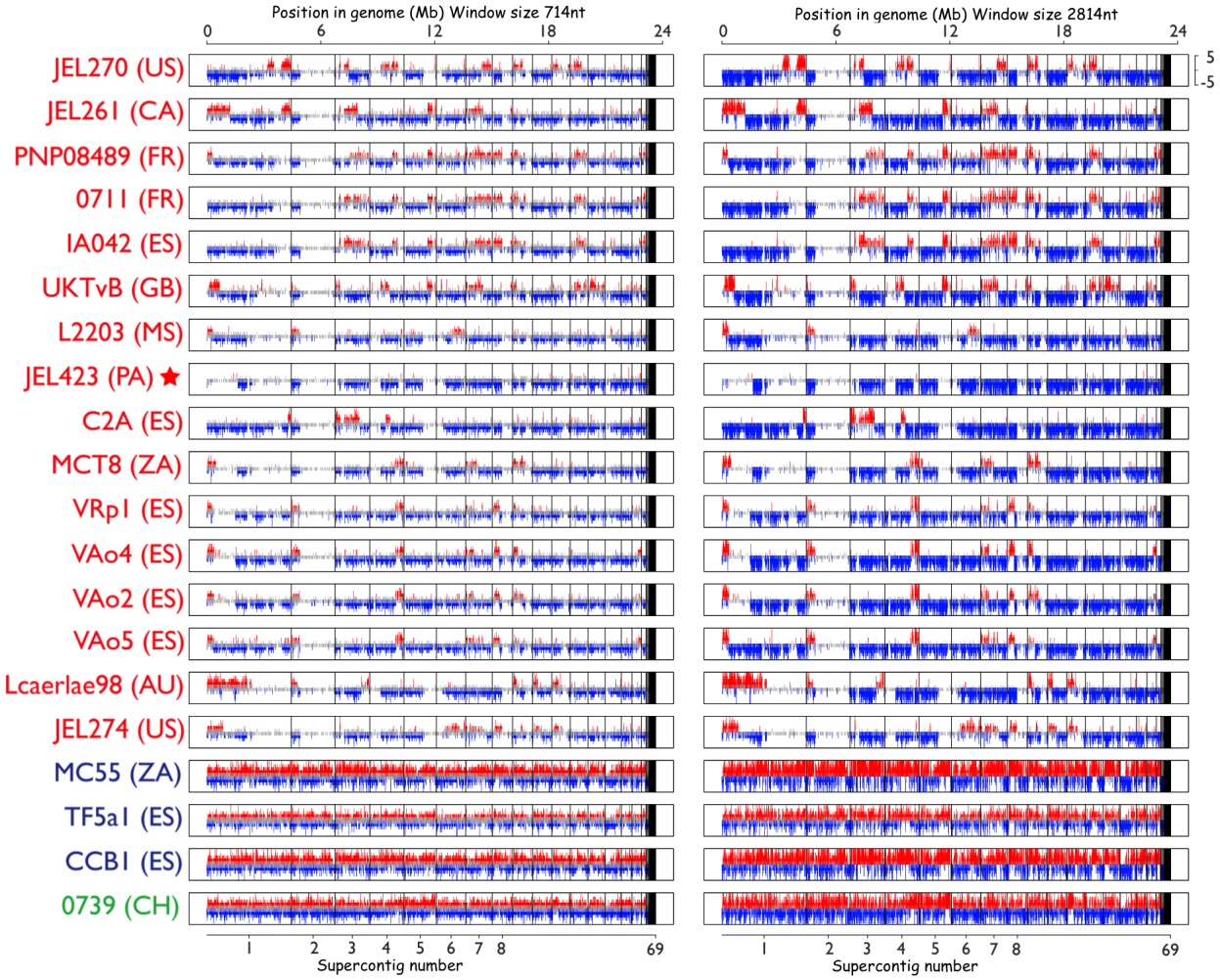
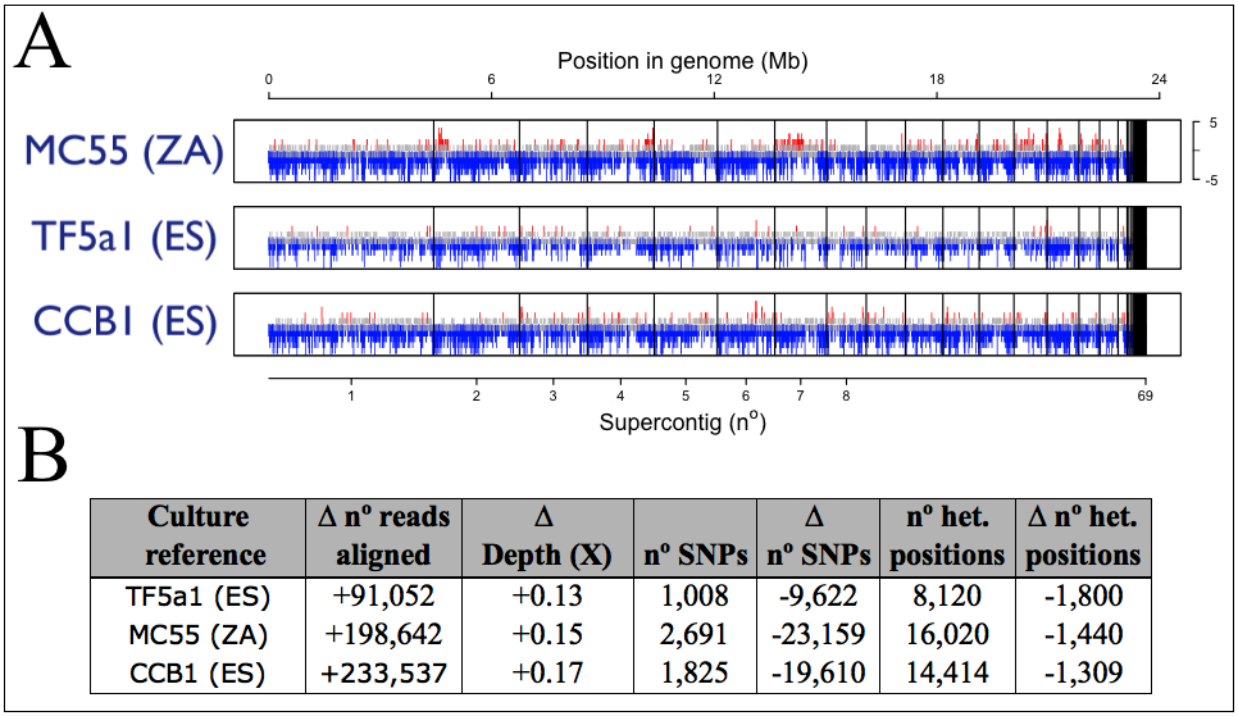


Figure 3.14: A range of sliding non-overlapping window lengths revealed the highly uneven distribution of homozygous and heterozygosity. Shown here are window lengths 714nt (left) and 2814nt (right) corresponding to half and double that used in Fig. 3.13)



B

Culture reference	Δ n° reads aligned	Δ Depth (X)	n° SNPs	Δ n° SNPs	n° het. positions	Δ n° het. positions
TF5a1 (ES)	+91,052	+0.13	1,008	-9,622	8,120	-1,800
MC55 (ZA)	+198,642	+0.15	2,691	-23,159	16,020	-1,440
CCB1 (ES)	+233,537	+0.17	1,825	-19,610	14,414	-1,309

Figure 3.15: Isolates from the *Bd*CAPE lineage were aligned against the consensus sequence for the *Bd*CAPE lineage. (**A**) A non-overlapping sliding window of SNPs minus heterozygous positions does not reveal hallmarks of loss-of-heterozygosity. (**B**) Details of the alignments show that there are similar levels of SNPs within the *Bd*CAPE linage as there are within the *Bd*GPL. These alignments also achieve a greater depth from which to call those mutations.

in the genome (Fig. 3.16). While we cannot formally reject either hypothesis, rare meiotic recombination is the more parsimonious explanation for the observed pattern. In particular, four of our sequenced isolates (VRp1, VAo4, VAo2 and VAo5) originate from the same pond (near Valencia, Spain) at the same time point. The mitochondrial tree (Fig. 3.13), which is arguably expected to reflect the species tree best given its non-recombining nature, gives high posterior support for this group being monophyletic. It therefore seems likely that the Valencia isolates originated from a single ancestor introduced to the pond. However, only 10 of the 15 gene trees (shown in Fig. 3.17) give posterior support for this group being monophyletic, with consistently different groupings observed for the other cases. In order for this pattern to be compatible with a model of gene conversion, as described by hypothesis 2, we would need to assume that the population that colonised the pond near Valencia had levels of diversity comparable to *Bd*GPL as a whole, and that this diversity was lost through rapid gene conversion only after the local colonisation event. Distinguishing between these two hypotheses awaits the experimental determination of relative rates of meiotic versus mitotic recombination.

3.4.4 Differences between *Bd* lineages: host response, phenotypic and drug susceptibility

To ascertain if *Bd* virulence recapitulated phylogeny, we experimentally exposed common toad (*Bufo bufo*) tadpoles to repeated, high concentration doses of *Bd* using multiple isolates from the two lineages with replication (*Bd*GPL[2] and *Bd*CAPE[3]) and from two clusters within *Bd*GPL (Valencia and the Pyrenees). All but four animals completed metamorphosis, and those four were excluded from analysis.

First, we used analysis of variance (ANOVA) to determine if time to metamorphosis or mass at metamorphosis varied at either level. While we saw no effect at the lineage level (mass at metamorphosis degrees of freedom, d.f.=2, F=0.7635, $p>0.05$; time until metamorphosis d.f.=2, F= 1.1856, $p>0.05$), mass at metamorphosis did vary among isolates in the global lineage (d.f.= 6, F = 2.2188, $p<0.05$) while time to metamorphosis did not (d.f.= 6, F = 0.264, $p>0.05$). *Post hoc* pairwise comparison (ANOVA) of global lineage isolates against the control treatment revealed that animals exposed to isolate VAo5 were significantly larger

[2] *Bd*GPL isolates included IA042, IA043, 0711.1 VAo2, VAo4, VAo5, MCT8 and JEL423
[3] *Bd*CAPE isolates included TF1.1, TF5a1 and CCB1

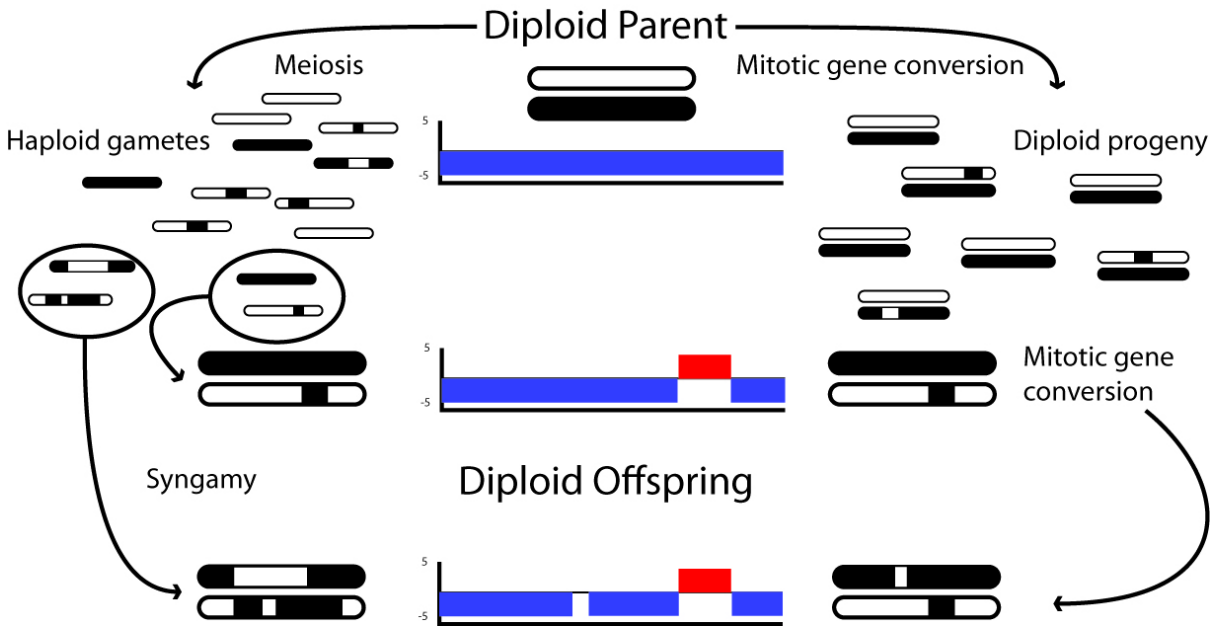


Figure 3.16: Two possible mechanisms of achieving the uneven distribution of heterozygous and homozygous SNPs throughout the *Bd* genome. Each black and white bar represents a haplotype identified in a parental isolate, and the charts in the middle represent the plots illustrating regions where heterozygosity predominates (blue) and where homozygosity predominates (red for homozygous SNPs and empty for homozygous identical to reference). On the left of the diagram, meiosis generates recombinant haploid genomes that then are united via syngamy into new diploid offspring with patchy heterozygosity. Meiosis involving fusion of diploid gametes could result in similar patterns if chromosomal segregation remains independent. On the right of the diagram, mitotic gene conversion generates patches of homozygous sites via homologous DNA repair in diploid progeny.

Breakpoint		LHS	RHS	
SC	position	p (adjusted p)	p (adjusted p)	Feature
1	421590	0.0001 (0.0052)	0.0001 (0.0052)	exon: mepce protein
1	855497*	0.0005 (0.0260)	0.0001 (0.0052)	exon: mgc81149 protein
1	1162181*	0.0032 (0.1664)	0.0056 (0.2912)	exon: bifunctional endo-β-xylanase
1	2103486*	0.0034 (0.1768)	0.0001 (0.0052)	intron: nuc156 family protein
1	3175083*	0.0001 (0.0052)	0.0073 (0.3796)	intergenic
1	4160547*	0.1910 (1.0000)	0.0272 (1.0000)	exon: circumsporozoite protein
1	4384574	0.0001 (0.0052)	0.0001 (0.0052)	exon: swi-snf complex subunit
2	479122*	0.6608 (1.0000)	0.0001 (0.0052)	exon: f-box wd-repeat
2	2146771	0.0001 (0.0052)	0.0001 (0.0052)	intergenic
3	806984	0.0001 (0.0052)	0.0001 (0.0052)	exon: c-4 methyl sterol oxidase
3	1762853	0.0001 (0.0052)	0.0001 (0.0052)	exon: protein
4	1100968	0.0001 (0.0052)	0.0001 (0.0052)	exon: protein
4	1781757	0.0001 (0.0052)	0.0001 (0.0052)	intron: No hit
5	996490	0.0001 (0.0052)	0.0001 (0.0052)	exon: protein
5	1655427*	0.0241 (1.0000)	0.0001 (0.0052)	exon: No hit
6	837975*	0.0123 (0.6396)	0.0026 (0.1352)	exon: unnamed protein (*P. anserina*)
6	1505912	0.0001 (0.0052)	0.0001 (0.0052)	exon: cytosolic regulator pianissimo
7	793086	0.0001 (0.0052)	0.0001 (0.0052)	exon: conserved hypothetical protein
8	492953	0.0001 (0.0052)	0.0001 (0.0052)	exon: No hit
8	1059351	0.0001 (0.0052)	0.0001 (0.0052)	intergenic
9	649616	0.0001 (0.0052)	0.0001 (0.0052)	exon: carboxypeptidase D
11	888027	0.0001 (0.0052)	0.0001 (0.0052)	exon: nadph oxidase isoform 2
12	568725*	0.0032 (0.1664)	0.0001 (0.0052)	exon: d-lactate dehydrogenase
13	885486*	0.0001 (0.0052)	0.5179 (1.0000)	intergenic
14	373467*	0.0155 (0.8060)	0.0001 (0.0052)	exon: No hit
15	472449*	0.0001 (0.0052)	0.0025 (0.1300)	exon: No hit

Table 3.4: Twenty-six recombination breakpoints across the genome were identified using GARD analysis. The *Bd* JEL423 supercontig (SC) number and position of each breakpoints is given in the first column. The left hand side (LHS) and right hand side (RHS) raw probabilities are shown, and adjusted probabilities after Bonferroni correction in parethesis. The overlying feature type and the genes top Blastp hit using the Blast2Go server [Conesa A *et al.* 2005] is shown in the final column. *Significant breakpoint.

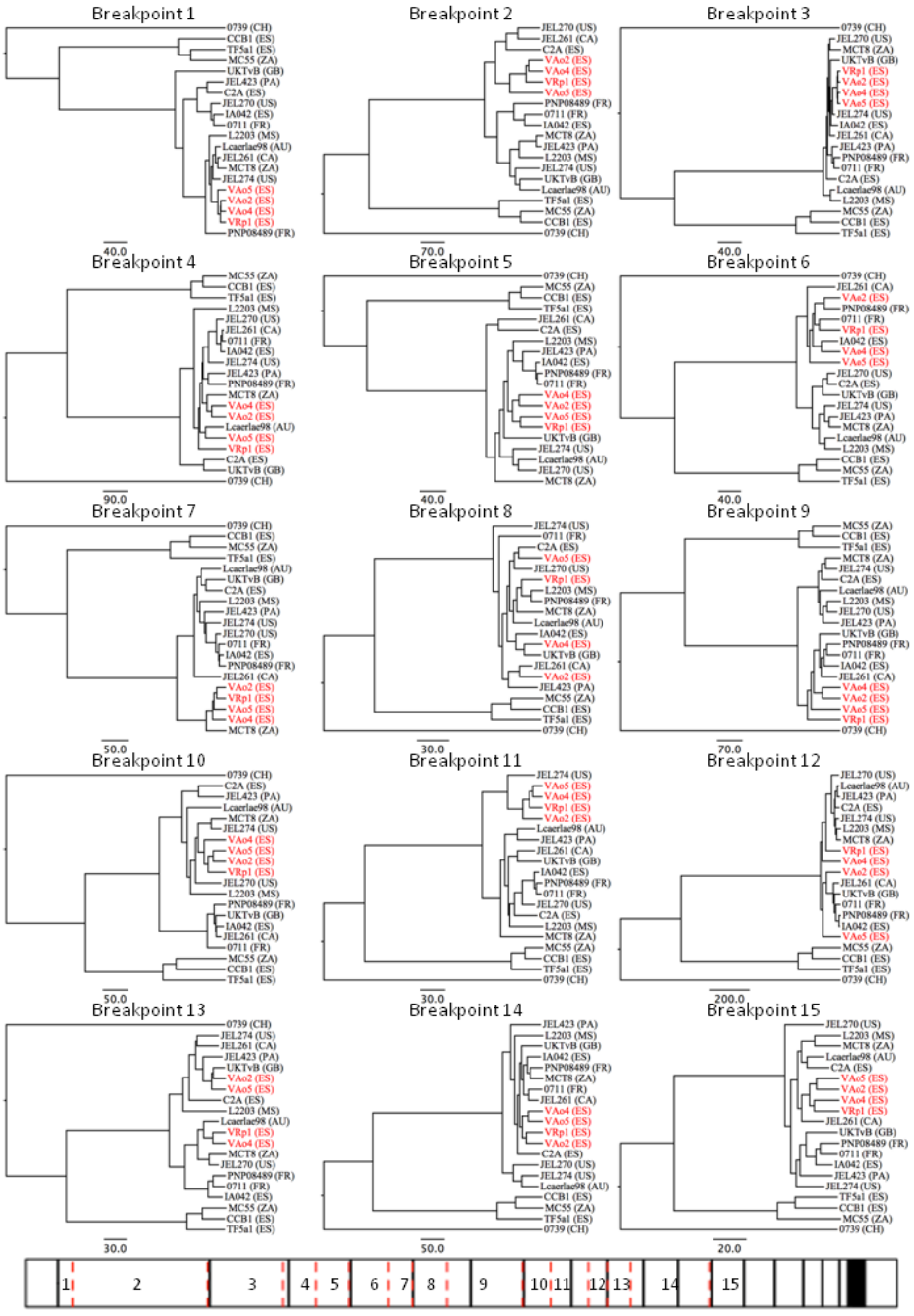


Figure 3.17: Maximum clade credibility trees for the fifteen significant recombination breakpoint segments of the nuclear genome. The four isolates from Valencia, Spain are highlighted in red in each of the phylogenies showing that they form a monophyletic group in ten of the fifteen trees. A representation of the genome is given below the trees showing the location of supercontigs (black lines) and the corresponding numbers for recombination breakpoints (red dotted lines) detected by the GARD method.

at metamorphosis than controls (t = -2.0228, df = 63.385, $p = 0.04731$).

However, by comparing infection status of toads exposed to *Bd*GPL and *Bd*CAPE, while low for all, we found exposure to *Bd*CAPE resulted in significantly reduced infection and mortality compared to those exposed to *Bd*GPL (Figs. 3.18A and 3.19 respectively) (Chi-square test, chi = 088, df = 2, $p = 4.013e^{-13}$). Significant variation in virulence was also detected among isolates within the *Bd*GPL (Chi-square test chi = 54.4, df = 5, $p<0.001$, Fig. 3.18B).

Kaplan-Meier survival curves illustrate how post-metamorphic survival of animals exposed to global lineage isolates was significantly different from both *Bd*CAPE lineage isolates and the negative controls (Fig. 3.19, Cox proportional hazards model, CPH; d.f.=4, z=3.124, $p<0.01$). When treatment group was accounted for, infection (CPH; d.f.=4, z=10.216, $p<0.001$) and increased fungal load (higher qPCR genomic equivalent score, CPH; d.f.=4 , z=4.459, $p<0.001$) increased the risk of mortality. This last result is in agreement with the findings of both Briggs *et al.* [Briggs CJ *et al.*, 2010] and Stockwell *et al.* [Stockwell MP *et al.*, 2010] where the likelihood of mortality increases with increasing *Bd* intensity of infection.

Significant variation in survival was observed among global lineage isolates (Fig. 3.18B). Specifically, tadpoles exposed to isolates IA043 (CPH; d.f.=12, z=-2.592, $p = 0.00953$) and VAo5 (CPH; d.f.=12, z=-2.172, $p = 0.02984$) experienced significantly reduced risk of death post-metamorphosis than those exposed to other global lineage isolates. Lastly, infection (CPH; d.f.=12, z=11.606 $p<2e^{-16}$) and smaller mass at metamorphosis (CPH; d.f.=12, z=-3.140 $p = 0.00169$) both also significantly increased the risk of death in animals exposed to global pandemic lineage isolates. Relationships between post-metamorphic body mass and infection with *Bd* have previously been reported for toads [Garner TWJ *et al.*, 2011].

Environmental conditions were standardised and animals were kept in isolation from each other precluding any effect of environmental forcing. However, the possibility does exist that the observed variation among *Bd*GPL treatments could be due to variation in host susceptibility, as mass at metamorphosis and infection dynamics of individual toadlets both play a role in determining mortality. Nevertheless, any effect of host, or the effects of variation among isolates from *Bd*GPL on virulence, is minor when compared to the observed effect of lineage on post-metamorphic mortality.

The discovery of *Bd*CH came too late for this lineage to be included in our *in vivo*

experimental framework. However, extensive population surveillance of *Bd* infected and uninfected populations across Switzerland has demonstrated a lack of association between infection status and population decline (http://www.bd-maps.net/maps); If *Bd*CH is representative of the lineages infecting other areas of Switzerland, then this is circumstantial evidence that *Bd*CH, like *Bd*CAPE, is hypovirulent.

Previous studies of morphological variation amongst *Bd* isolates have shown that phenotypic profiles are linked to the virulence of isolates [Fisher MC *et al.*, 2009] and that isolates belonging to *Bd*CAPE exhibit smaller sporangial sizes compared to isolates belonging to *Bd*GPL. We repeated this study using a greater number of *Bd*GPL isolates from other continents and including *Bd*CH.

Differences in sporangia size were observed amongst individual isolates (Fig. 3.20). Once isolates had been grouped into the relevant lineage, significant differences were seen between all three lineages, with *Bd*CAPE being significantly smaller in the following comparisons: *Bd*CAPE vs. *Bd*GPL ($p<0.001$); *Bd*CAPE vs. *Bd*CH ($p<0.001$); *Bd*GPL vs. *Bd*CH ($p = 0.00614$). These results corroborate the differences in average size of sporangia between lineages first described by Fisher *et al.* [Fisher MC *et al.*, 2009]. Differences in hyphal length were also observed amongst individual isolates (Fig. 3.20). Amongst lineages, *Bd*CAPE showed increased hyphal length relative to *Bd*GPL ($p = 0.01298$). *Bd*CH showed no significant difference in hyphal length to either *Bd*GPL or *Bd*CAPE and is therefore not diagnostic for the lineage. Together, these data show that genetic differentiation between lineages of *Bd* has resulted in significant morphological variation.

To test if different isolates or lineages varied in their susceptibility to Itraconazole treatment, we recorded their MC_{50} values from day 15 (Table 3.5). We found no significant difference between the MC_{50} values between *Bd*GPL and *Bd*CAPE. Results from *Bd*CH could not be included in analysis as there was only a single value for this singleton (isolate 0739); however it is within the range observed for the other two isolates suggesting that lineage-specific differences in susceptibility to itraconazole do not occur.

3.4.5 Dating the emergence

Surveillance data shows that *Bd* is increasing its global range, and rapid emergence across geographic regions has been documented in the United States [Vredenburg VT *et al.*, 2010], Central America [Lips KR *et al.*, 2006] and Australia [Skerratt LF *et al.*, 2007]. Eight of our

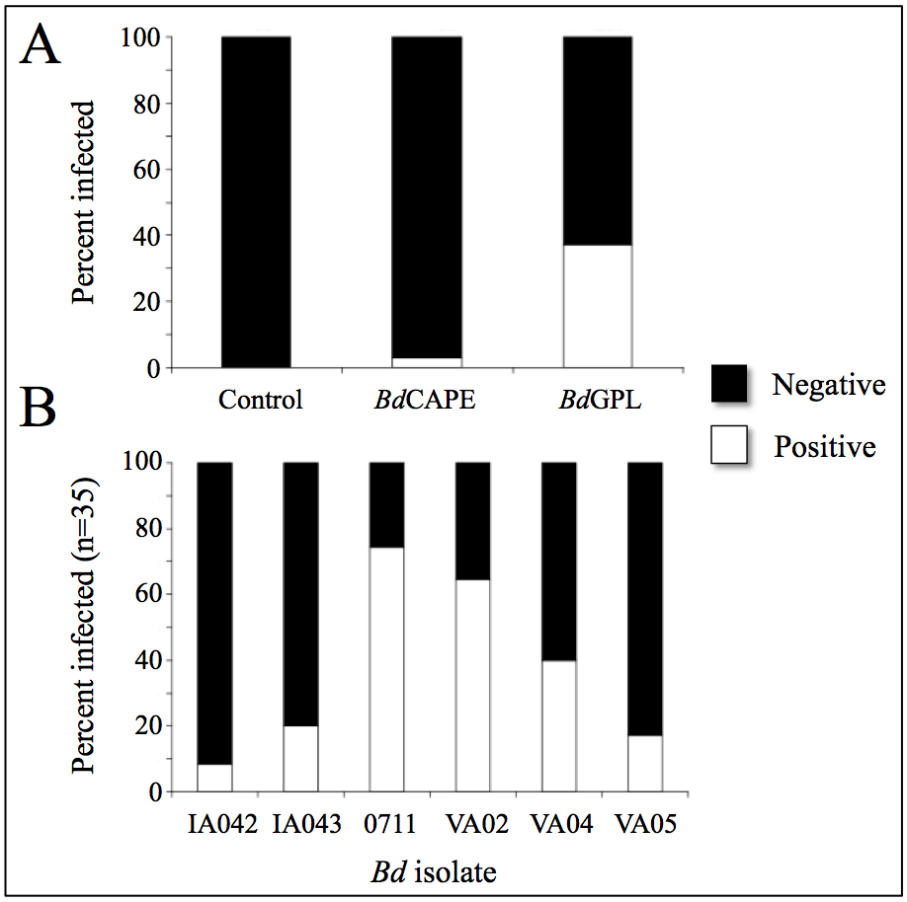


Figure 3.18: Proportion of animals infected at time of death or by the end of the experiment. (**A**) within the control treatment, *Bd*GPL and *Bd*CAPE, and (**B**) within the six *Bd*GPL isolates.

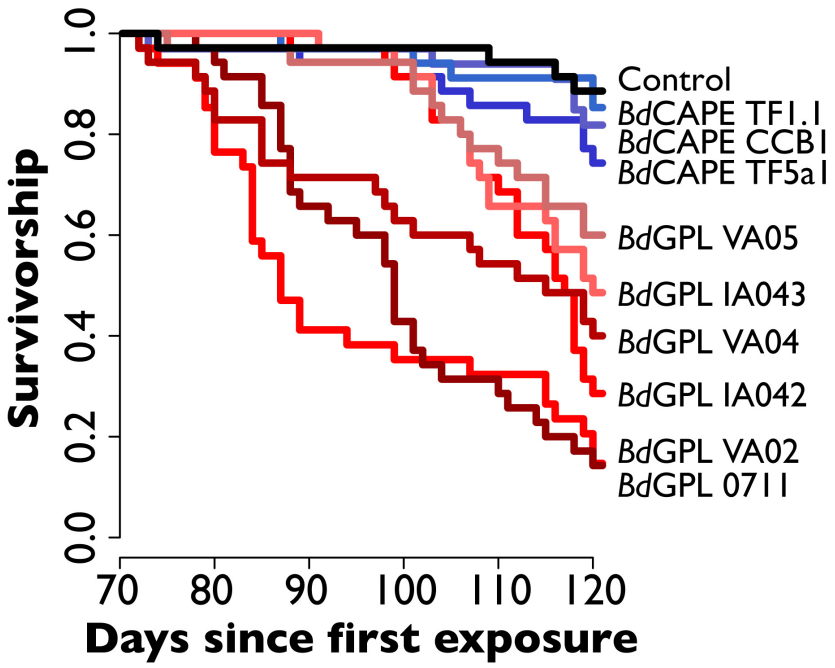


Figure 3.19: Kaplan-Meier survival curves illustrating post-metamorphic survival of animals exposed to *Bd*GPL isolates (red) and *Bd*CAPE lineage isolates (blue). *Bd*GPL isolates were significantly different from both *Bd*CAPE lineage isolates and the negative controls.

Lineage	Isolate	Itraconazole MC_{50} (μg/ml)
*Bd*GPL	0711 (FR)	0.01652
*Bd*GPL	IA042 (ES)	0.00600
*Bd*GPL	IA043 (ES)	0.01257
*Bd*GPL	JEL423 (PA)	0.01200
*Bd*GPL	MCT8 (ZA)	0.01100
*Bd*GPL	VAo2 (ES)	0.00600
*Bd*GPL	VAo4 (ES)	0.00875
*Bd*GPL	VAo5 (ES)	0.01609
*Bd*CAPE	CCB1 (ES)	0.01964
*Bd*CAPE	TF1.1 (ES)	0.01281
*Bd*CAPE	TF5a1 (ES)	0.01132
*Bd*CH	0739 (CH)	0.01260

Table 3.5: Itraconazole treatment of twelve isolates. MC_{50} (μg/ml) values from day 15 for 12 individual isolates spanning the three lineages, *Bd*GPL, *Bd*CAPE and *Bd*CH

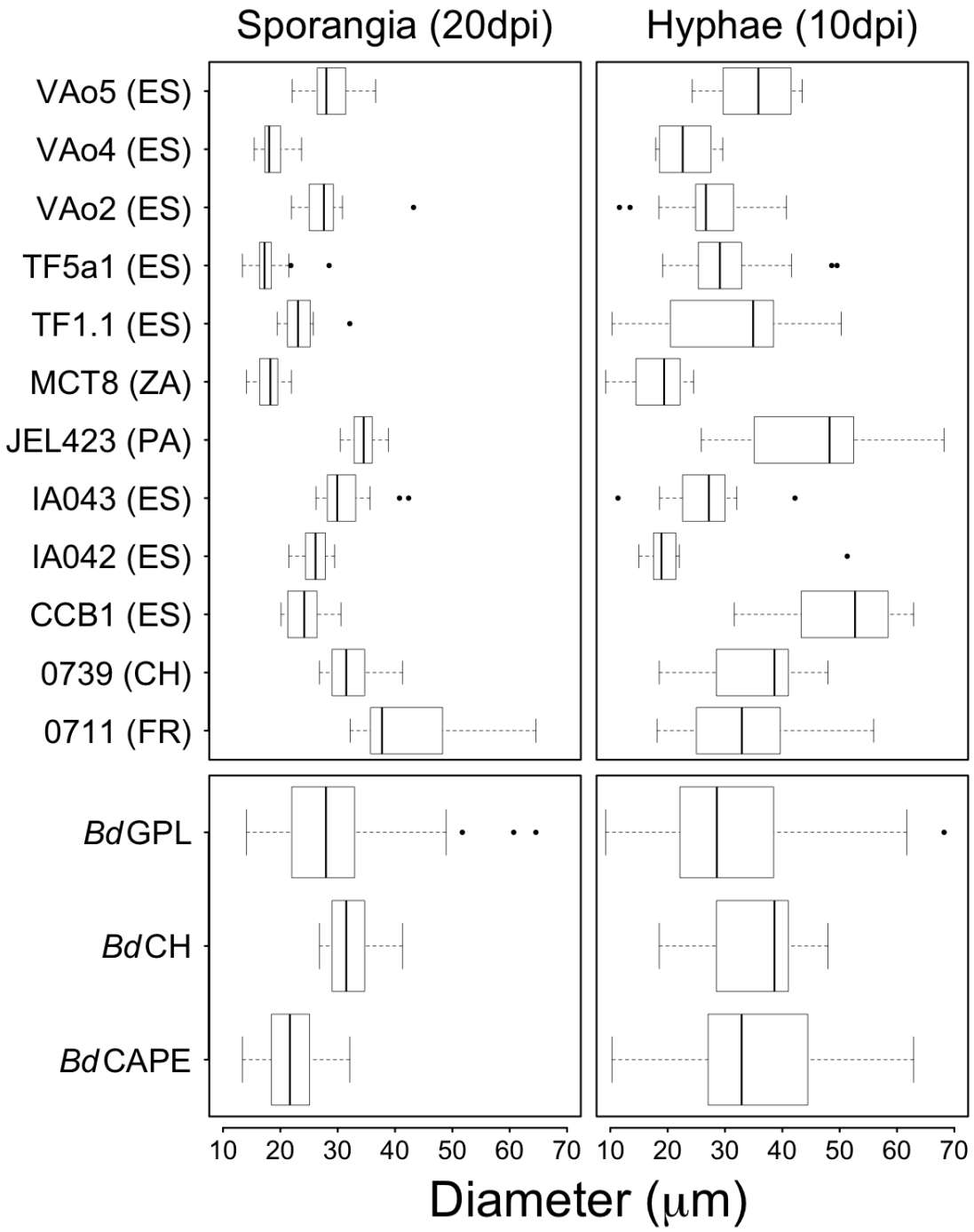


Figure 3.20: Diameter of sporangia and hyphae (μm) 20 days post initial culture (dpi) and 10dpi respectively. (Top) Measurements were taken for the 12 isolates individually. (Below) Measurements of isolates grouped according to their three lineages. Sporangia sizes between the three lineages were significantly different.

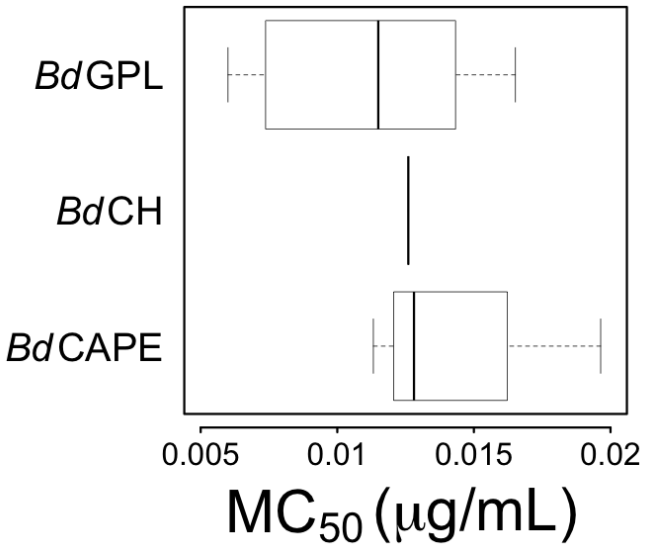


Figure 3.21: MC_{50} (μg/ml) for Itraconazole from isolates grouped into three lineages. Readings were taken from day 15. No significant difference was observed between the three lineages.

sequenced isolates are associated with rapid population/community extirpations in diverse biomes across the New World, Europe and Australia [Crawford AJ *et al.*, 2010], exemplified by the rapid and virtual extirpation of the critically endangered Mountain Chicken Frog *Leptodactylus fallax* in Montserrat by *Bd*GPL isolate L2203. The widespread occurrence of *Bd*GPL across five continents suggests that the emergence of this lineage is responsible for these *Bd*-driven collapses in diversity. To date the emergence of this lineage, we took advantage of the fact that the samples were collected over a period of ten years and used this to inform a strict clock across the phylogenies of the 15-breakpoint segments (Fig. 3.17). The median dates for the emergence of the global hypervirulent lineage range from 35-257 years before present (ybp) (Table 3.6). Amphibian declines were first noted in the 1970's simultaneously in the USA (Sierra Nevadas), Central America and Australia [Berger L *et al.*, 1998; Houlahan JE *et al.*, 2000], therefore our dating suggests that the globalization of *Bd*GPL is compatible with its spread within the amphibian trade.

Gene tree number	tMRCA for *Bd*GPL			tMRCA for 3 *Bd* lineages		
		95% CI			95% CI	
	median	lower	upper	median	lower	upper
1	144	74	231	591	303	942
2	237	180	303	930	701	1194
3	35	27	45	490	361	660
4	145	96	206	611	399	863
5	111	68	165	389	231	584
6	94	59	137	447	279	663
7	130	79	195	543	328	815
8	52	35	75	232	152	346
9	101	67	148	527	347	786
10	167	105	245	528	330	769
11	117	72	174	443	268	665
12	92	55	139	921	550	1384
13	81	46	130	326	179	535
14	60	41	87	325	213	481
15	36	31	42	140	117	170

Table 3.6: Time to most recent common ancestor (tMRCA) for the 16 *Bd*GPL isolates for the fifteen significant breakpoint segments. Median dates for the emergence of the global hypervirulent lineage range from 35-257 years before present (ybp). Lower and upper 95% confidence intervals (CI) for each gene tree are also shown.

3.5 Discussion

One of the more puzzling aspects of the emergence of amphibian chytridiomycosis has been that, while epizootics have been widely observed, many susceptible amphibian communities apparently coexist alongside *Bd* with no evidence of disease. Such coexistence has been attributed to the context-dependent nature of susceptibility to disease, and while this is undoubtedly an important factor [Walker SF *et al.* 2010], our data show that *Bd* genotype is also an important epidemiological determinant. Here, we found that there is a much greater diversity of *Bd* than was previously recognised, and that multiple lineages are being vectored between continents by the trade of amphibians (*Bd*GPL and *Bd*CAPE). We have characterised hypervirulence in *Bd*GPL suggesting that the emergence and spread of chytridiomycosis largely owes to the globalisation of this recently emerged recombinant lineage [Fisher MC *et al.*, 2007].

Research has recently suggested the existence of a *Bd* lineage that is associated with the Japanese giant salamander, *Andrias japonicus* [Goka K *et al.*, 2009]. This lineage, defined by sequencing a short fragment of the ribosomal DNA, is dissimilar to the rDNA sequences of *Bd*GPL, which non-native North American bullfrogs have introduced to Japan. Therefore, it appears that we can now provisionally recognise at least four lineages of *Bd*, two of which are possibly endemic (*Bd*CH and Japan), one of which may have been previously endemic to South Africa but was then vectored to Mallorca (*Bd*CAPE) and one of which has a pan-global distribution (*Bd*GPL). This diversity was uncovered from sampling only 20 genomes from a cohort biased towards sampling amphibian populations experiencing chytridiomycosis (and hence infected with the *Bd*GPL), therefore our data suggest a more extensive diversity of amphibian-associated chytrid lineages is pending discovery. Whether these are *Batrachochytrium dendrobatidis sensu stricto*, or represent cryptic species, remains to be determined. However, observations that populations infected with non-*Bd*GPL lineages do not undergo epizootics suggest that the diversity of *Bd* represents a patchwork of genetically and phenotypically diverse lineages. In this case, determining the geographical origin(s) of the parental genotypes of *Bd*GPL will likely remain an elusive challenge until a much wider sample of amphibian-associated chytrid lineages has been sampled in order to identify the full phylogeographic range of diversity held within the order Rhizophydiales.

The origin of novel virulence in fungal species via recombination/hybridization is a well-recognised pathway underpinning disease emergence for increasing numbers of plant

and animal pathogens [Fisher MC *et al.* 2009; Stukenbrock EH *et al.* 2008]. Genomic rearrangement between allopatric fungal lineages that have not evolved reproductive barriers is promoted when anthropogenically-mediated dispersal increases the rate of lineage mixing. The resulting novel inter-lineage recombinants exhibit a diversity of virulence profiles, some of which can initiate epidemics; contemporary examples include the evolution of hypervirulence in the Vancouver Island outbreak of *Cryptococcus gattii* [Fraser JA *et al.*, 2005], and many novel aggressive plant pathogens that increasingly threaten global food security [Strange RN *et al.*, 2005] as well as natural populations [Goss EM *et al.*, 2009].

We postulate that the anthropogenic mixing of allopatric lineages of *Bd* has led to the generation of the hypervirulent *Bd*GPL via an ancestral meiosis, and that, as previously suggested [Morgan JAT *et al.*, 2007], this lineage is undergoing further diversification by either mitotic or sexual recombination. We show that the global trade in amphibians is resulting in contact and cross-transmission of *Bd* amongst previously naive host species, resulting in intercontinental pathogen spread. As the rate of inter-lineage recombination between fungi will be proportional to their contact-rates, we predict that such globalization will increase the frequency that recombinant genotypes are generated. Theory and experimentation has shown that in genetically diverse infections, virulent lineages can have a competitive advantage resulting in increased transmission [Nowak MA *et al.*, 1994; de Roode JC *et al.*, 2005]. As a consequence, we predict the evolution of further hypervirulent fungal lineages across a diverse range of host species and biomes in the absence of tighter biosecurity [Fisher MC *et al.*, 2009].

3.6 Acknowledgments

We thank Joyce Longcore for supplying isolates of *Bd* that we sequenced in this project, Beni Schmidt and Ursina Tobler for facilitating of the collection of animals from which *Bd*CH was isolated and Andrew Rambaut and Philippe Lemey for advice on the phylogenetic analysis. The Government of Montserrat issued permits for the isolation and export of *Bd* from *L. fallax*. This project was funded by the UK Natural Environmental Research Council (NERC) grant NE/E006701/1, the UK Department for Environment, Food and Rural Affairs (DEFRA) grant FC1195, the Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/H008802/1, the European Research Council (ERC) grant 260801-BIG_IDEA) and the Biodiversa project RACE: Risk Assessment of Chytridiomycosis to European Amphib-

ian Biodiversity (http://www.bd-maps.eu). Computational analysis was supported by the UCSD Center for AIDS Research BEAST Core (NIH AI 036214). SOLiD sequencing performed by IoSSB and CISBIC, Imperial College. Sequences are deposited in the NCBI Short Read Archive under the submission accession number SRA030504.

3.7 Publicity, interviews and aftermath

3.7.1 Tamera Jones, Natural Environment Research Council - Frog trade link to killer fungus revealed

The global trade in frogs, toads and other amphibians may have accidentally helped create and spread the deadly fungal disease, chytridiomycosis, which has devastated amphibian populations worldwide.

The Panamanian golden frog (*Atelopus zeteki*) is a critically endangered toad which is endemic to Panama. What's more, researchers say that unless the trade is regulated, even deadlier strains of the disease may soon emerge.



Figure 3.22: The Panamanian golden frog (*Atelopus zeteki*) is a critically endangered toad which is endemic to Panama.

An international team of scientists, led by Dr Matthew Fisher from Imperial College London found that the trade may have let non-lethal strains of the chytrid fungus from different parts of the world come into contact with each other.

This means they've exchanged genes in a process called recombination, creating a new and lethal strain which has decimated frog populations around the world in recent years.

'It's likely that the amphibian trade has allowed different populations of the fungus to come into contact with each other, allowing recombination to occur,' says Rhys Farrer from Imperial College London and ZSL's Institute of Zoology, lead author of the study, published in *Proceedings of the National Academy of Sciences*.

'This has created a hypervirulent strain leading to losses in amphibian biodiversity.'

The chytrid fungus, or *Batrachochytrium dendrobatidis* (*Bd*) as it is sometimes called, infects the skins of amphibians like frogs, toads, salamanders and newts.

Declines in many amphibian populations around the world are due to the disease and over 200 species are suspected to have become extinct as a result. In Central America alone, chytridiomycosis has led to the loss of up to 40 per cent of wild amphibians including the Panamanian Golden Frog.

Despite much research on the disease, scientists have struggled to figure out where it came from or explain how it spread. The problem is even more puzzling because some amphibians coexist alongside *Bd* with no sign of disease.

'This strongly suggested there may be more than one type of strain of chytrid fungus,' says Farrer. So, he and his colleagues decided to sequence and compare *Bd* genomes from 20 disease samples isolated from 11 amphibian species worldwide to find out more about the fungus' ancestry.

They found three different strains. One of these, the Global Panzootic Lineage (GPL), has made its way to at least five continents and has caused infections in North America, Central America, the Caribbean, Australia and Europe.

The researchers found evidence of gene exchange in this lineage, which turned out to be the deadliest of the three strains.

In one example, a captive breeding and reintroduction program to boost numbers of the

endangered Mallorcan midwife toad may have helped *Bd* spread from captive African Cape clawed frogs to the toads.

The fact that they found three strains in just 20 samples also suggests *Bd* is much more diverse than previously thought.

'What's interesting is that they're not all causing disease,' says Fisher. 'Only one lineage is a killer, and it has evolved very recently'. Scientists had until now thought there was just one strain of *Bd*.

Farrer, Fisher and their colleagues also found that the start of the decline of amphibians around the 1970s coincides with the emergence of the amphibian trade.

'The age of the lethal *Bd*GPL lineage coincides with the start of the amphibian trade in the 20th century, when we started moving many frogs and toads around the world,' says Fisher.

'The horse has well and truly bolted, but to halt the further spread of this disease, we really need to increase global biosecurity,' says Fisher.

3.7.2 Michael Marshall, New Scientist - Frog-killer disease was born in trade

The global amphibian trade spread the lethal chytrid fungus, which is decimating frogs around the planet, and it now looks like it may have created the disease in the first place.

The team behind this finding are calling for an amphibian quarantine to help slow the disease's spread.

Rhys Farrer of Imperial College London and colleagues sequenced the genomes of 20 samples of the offending fungus, *Batrachochytrium dendrobatidis* (*Bd*), collected in Europe, Africa, North and South America and Australia.

They found that 16 of the 20 samples were genetically identical, belonging to a single strain called *Bd*GPL that had spread to all five continents. Tests on tadpoles also revealed that the strain was extremely virulent.

*Bd*GPL's genome showed that it had formed when two strains mated, some time in the past 100 years. The best and simplest explanation is that 20th-century trade, which

shipped amphibians all over the world, enabled the mating, says Farrer's supervisor Matthew Fisher.

"We've got to restrict trade, or at least make sure that amphibians are not contaminated," says Fisher. One approach would be for countries to quarantine all imported amphibians and only allow them to stay if they are uninfected.



Figure 3.23: Moment of truth for an Atelopus frog in Ecuador (Image: Joel Sartore/National Geographic/Getty)

When it emerged that trade was spreading chytrid, the World Organisation for Animal Health made the disease notifiable, meaning that countries must report whether they have it or not. But that doesn't stop it spreading.

The two places in most urgent need of protection are Madagascar and south-east Asia, says Fisher: "They're the last redoubts of uninfected amphibian species." Both are hotspots of amphibian diversity, and are clear of *Bd*GPL. Madagascar remains uninfected despite rampant *Bd*GPL in Africa, and a recent survey shows that Asian chytrid strains are not very virulent (*PLoS One*, DOI: 10.1371/journal.pone.0023179).

If *Bd*GPL reaches these places, it could quickly devastate their frogs. Within months of it reaching Montserrat, in the West Indies, in early 2009, conservationists had to fly giant ditch frogs - also known as mountain chickens - out of the country to save them from extinction.

Countries that already have *Bd*GPL should also institute quarantine, says Peter Daszak, president of EcoHealth Alliance in New York. "This research shows that recombination can occur and give rise to new virulent strains," he says. "Blocking introduction of new strains will cut down on this."

Daszak adds: "It will be hard to stop the spread of new lineages of *Bd*, but if we look at the devastation that this pathogen has already caused, we desperately need to try."

3.7.3 Richard Black, BBC - Killer frog fungus 'spread by trade'

The fungus killing frogs around the world comes in several forms, and has almost certainly been distributed by trade in amphibians, research shows.

Scientists led from Imperial College London found three distinct lineages of the chytrid fungus in various nations.

The most widespread and lethal form was probably created by a crossing of two prior forms, they report in *Proceedings of the National Academy of Sciences*.



Figure 3.24: The scarlet frog of Venezuela may have been a victim of chytridiomycosis

Chytrid is now found on every continent and has wiped out a number of species.

"It's safe to say that it arose in the 20^{th} Century, and that's in the realm of time for the trade in amphibians" Rhys Farrer, Imperial College

Identified just over a decade ago, it kills amphibians by blocking the transfer of vital substances through their skins, eventually causing cardiac arrest.

Its origins are believed to lie in southern Africa.

"Before this study, no-one knew there were any different lineages," said Rhys Farrer, the project leader from Imperial.

"This work comes from using the new whole-genome sequencing technique, combining data from all over the world.

"And it's obviously important, as chytrid is one of the most devastating wildlife diseases with the largest host range of any, and responsible for dozens of species extinctions and many more extirpations of local populations."

The team took samples from amphibians in 20 sites spanning Europe, North America, Central America, the Caribbean and South Africa.

The majority carried *Batrachochytrium dendrobatidis* (*Bd*) fungus of the type that has a truly global spread, which they dubbed *Bd*GPL.

But their Swiss sample showed a different form, or clade, named *Bd*CH, while a third clade (named *Bd*CAPE) turned up in the Cape Province of South Africa and the Mediterranean island of Mallorca.

The Mallorcan chytrid was almost certainly carried from South Africa, probably via the trade in amphibians for zoos or private collections.



Figure 3.25: A different strain of the chytrid fungus was recently found on Japanese giant salamanders.

The Swiss form probably came via a similar route, researchers believe.

Laboratory tests showed that *Bd*CAPE was substantially less damaging to amphibians than *Bd*GPL. (The Swiss form was identified too late in the project to be tested in this way.)

The genetic differences that make *Bd*GPL more lethal have not been identified. But the team believes it became so deadly through a chance encounter between two or more prior strains.

"We think we are seeing unique evidence of recombination within *Bd*GPL - we can't say for sure if it's a hybridisation event but it's the most likely explanation,' said Mr Farrer.

"From the dating work we've done it's safe to say that it arose in the 20^{th} Century, and that's in the realm of time for the trade in amphibians."

Although the transport of exotic amphibians for pets is a prime suspect, another theory holds that the lethal *Bd*GPL chytrid spread through the importation of frogs from Africa to North America and Europe for use in pregnancy testing.

However, yet another form of the fungus was recently discovered in Japan, its relationship to African-derived lineages uncertain.

The latest research marks a new staging post on a fast and fascinating voyage of scientific discovery. Whether it can help combat the disease is another matter.

Analysing the genomes of the various strains may show scientists what makes some virulent and others relatively benign.

The Imperial team believes it is also worth investigating whether the less virulent forms can be used to give amphibians a degree of resistance, in the same way that some vaccines do through using attenuated forms of disease-causing microbes.

3.7.4 Rhys Farrer & Matthew Fisher, Frog log - Multiple emergences of genetically diverse amphibian infecting chytrids include a globalized hypervirulent recombinant lineage

Since the discovery of the amphibian infecting pathogenic fungus *Batrachochytrium dendrobatidis* (*Bd*) during the 1990s, it has been found on every continent on Earth (except Antarctica!) and responsible for dozens of Amphibian extinctions and local extinctions (extirpations). The current global distribution of *Bd* and results of surveillance can be found at the *Bd* Global Mapping project www.bd-maps.net. One of the puzzling aspects of the disease is the extremely high genetic similarity between isolates found from diverse habitats worldwide. This finding supported a hypothesis whereby the fungus has been recently spread

'far and wide' by anthropogenic means such as the pet trade, human consumption, zoos and for medical purposes.

Because genetic variation between isolates can provide clues regarding the origin, evolution and spread of diseases, we examined the entire genomes from a global panel of *Bd* isolates that had been isolated from amphibian habitats experiencing die-offs (such as the recent mountain chicken *Leptodactylus fallax* epizootic on Montserrat), and others that appeared to be only harboring the disease. In particular, we focused on Europe, and took samples spanning Iberia, the mountains of the French Pyrenees, northern Mallorca, Switzerland and the UK, and by doing so recovered isolates of *Bd* that appeared to be markedly different from one another in their microscopic characteristics.



Figure 3.26: Midwive mass mortalities (*Alytes obstetricans*) in the French Pyrenees, 2010. Photo: Matt Fisher.

By examining how the *Bd* genomes differed from one and other, we identified three separate and divergent lineages that we have named *Bd*GPL, *Bd*CAPE and *Bd*CH. The first two are found in more than one continent, and the third only in Switzerland. Of these lineages, *Bd*GPL also has features within the genome that may have resulted from hybridization between two parental strains of the disease. Our data shows that this is the most common type of *Bd* (found on all 5 continents we tested), and also the type associated

with mass-mortalities and extinction, by invading regions such as the US Sierra Nevadas, Central America, Australia and Montserrat. We therefore called this lineage the Global Panzootic Lineage (*Bd*GPL). The divergence between our *Bd*GPL isolates pointed to an emergence during the 20^{th} Century, which is within the realm of time for the widespread trade of amphibians. *Bd*GPL diverged from the other two lineages at least 1000 years ago.



Figure 3.27: Moribund and dead Midwife metamorphs suffering from terminal chytridiomycosis. Pyrenees 2010. Photo: Matt Fisher.

Isolates belonging to *Bd*CAPE were found on both the island of Mallorca and in the Western Cape of South Africa. This finding supported a hypothesis by Dr. Susan Walker that the island of Mallorca was accidently infected by the release of infected mallorcan midwive toads Alytes muletensis that had aquired their infection from a South African *Xenopus gilli* at a Mallorcan breeding facility in 1991. We called this lineage *Bd*CAPE after its likely origin in Cape province, South Africa.

We also performed assessment of host response of the common toad (*Bufo bufo*) to 9 *Bd* isolates (three *Bd*CAPE and six *Bd*GPL) and 1 negative control. We found that postmetamorphic survival of animals exposed to *Bd*GPL isolates was significantly reduced when compared to isolates of *Bd*CAPE and the negative controls. The discovery of the third lineage, which we identified only in Switzerland, came too late to include in the host response

experiments.

Morphological variation between the two lineages was also compared as previous studies of morphological variation among *Bd* isolates have shown that phenotypic profiles are linked to the virulence of isolates. We found that *Bd*CAPE exhibits significantly smaller sporangial sizes and increased hyphal length compared with isolates belonging to *Bd*GPL. Together, these findings show that genetic differentiation between lineages of *Bd* has resulted in significant morphological variation. However, we did not detect differences in sensitivity to the fungicide itraconazole among lineages.

The discovery of multiple emergences of *Bd* sheds light on a number of extant questions. Firstly, *Bd* is not a single clone as was previously thought, and different lineages exhibit different morphology, genome structure and virulence. Secondly, that the infection comprises at least two Novel Pathogens, which have both been moved via the trade in amphibians: Once by accidental cohousing in a zoo (*Bd*CAPE), and *Bd*GPL, which has been transported many times by international trade in, principally, north American bullfrogs.

If hybridization occurred between two parental strains to generate *Bd*GPL, it is likely that they arose from separate geographical regions that were brought together by some form of trade. If this hypothesis is true, extinctions from *Bd* are not solely caused by introduction of the pathogen into naive populations, but that the largely unregulated trade in amphibians may have inadvertently created this fungal superbug. However, until further sampling has been done to verify either parental strains, or lab-based experiments have identified a sexual phase of *Bd*, questions regarding the origin and life cycle of this disease remain to be found.

Our study highlights the urgent action required to prevent future panzootics, which may rest in readdressing the measures used to prevent transmission of infectious diseases (biosecurity). Like many other threatened species (little brown bats in North America, European crayfish, Zambezi Tilapia), tightening the biosecurity associated with trade is key to preventing accelerated evolution and spread of hyper-virulent diseases in the future.

3.7.5 Faculty of 1000 (F1000)

Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalized hypervirulent recombinant lineage. (*Proc. Natl. Acad. Sci. U. S. A.* 2011 Nov 15), has been

selected and evaluated by Adam Jones and Richard Frankham, Members of the Faculty of 1000 (F1000), which places your work in our library of the top 2% of published articles in biology and medicine.

Comments (Adam Jones, Texas A&M University, TX, USA):

This paper is interesting because it sheds light on the emergence and evolution of the fungus (*Batrachochytrium dendrobatidis* or *Bd*) that has been implicated in the global decline of amphibian populations. The results of the reported whole-genome, population genomic study show that one hypervirulent strain emerged recently and spread around the globe, probably through the aquarium trade. This dataset provides interesting insights into the evolution of this important pathogen and will be instrumental in understanding its role in global amphibian declines. This study is also noteworthy in that it uses population genomic data to shed new light on the evolution of *Bd*. The authors have used sequencing by oligonucleotide ligation and detection (SOLiD) to generate genome sequence data from 20 isolates of *Bd* from around the globe. In addition to the globally distributed, hypervirulent strain, the data reveal two other divergent lineages of *Bd*, which are shown to be less virulent. In addition to providing an interesting case study, this work thus demonstrates the utility of population genomics in the study of pathogen evolution and conservation biology.

Comments (Richard Frankham, Macquarie University, NSW):

This study shows that the recently emerged hypervirulent chytrid fungus that is devastating amphibian populations across the planet, *Batrachochytrium dendrobatidis*, arose by recombination between different strains. The authors postulate that it arose following human-mediated contact between previously isolated populations. This represents a further example of the emergence of virulent plant and animal fungal pathogens following recombination/hybridization. The study involves whole-genome sequencing of 20 global isolates of the fungal species from 11 amphibian host species. The results support the hypothesis of a single hybrid origin of the hypervirulent strain, but the geographic origins of the parental strains have yet to be identified. The study shows that the global trade in amphibians is resulting in contact and cross transmission of the chytrid fungus among previously nave host species. It represents another example of the deleterious effects of trade on biodiversity.

Chapter 4

Chromosomal copy number variation and uneven rates of recombination reveal cryptic genome diversity linked to pathogenicity

Rhys A. Farrer[1,2], Daniel A. Henk[1], Trenton W. J. Garner[2], François Balloux[3], Douglas C. Woodhams[4] and Matthew C. Fisher[1]

[1]The Dpt. Infectious Disease Epidemiology, Imperial College London, UK; [2]Institute of Zoology, Zoological Society of London, UK; [3]Dpt. of Genetics, Evolution and Environment, University College London, UK [4]Dept. of Ecology & Evolutionary Biology, University of Colorado, USA

The contents of this chapter were published in: Farrer RA, Henk DA, Garner TWJ, Balloux F, Woodhams DC, *et al.* (2013) Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. *PLoS Genet* **9**(8): e1003703. doi:10.1371/journal.pgen.1003703

Author contributions: Conceived and designed the experiments: RAF DAH FB DCW MCF. Performed the experiments: RAF DAH DCW. Analyzed the data: RAF DAH DCW. Contributed reagents/materials/analysis tools: RAF TWJG. Wrote the paper: RAF DAH TWJG FB DCW MCF.

4.1 Abstract

Pathogenic fungi constitute a growing threat to both plant and animal species across a global scale. Despite a clonal mode of reproduction dominating the population genetic structure of many fungi, putatively asexual species are known to adapt rapidly when confronted by efforts to control their growth and transmission. However, the mechanisms by which adaptive diversity is generated across a clonal background are often poorly understood. We sequenced a global panel of the emergent amphibian pathogen, *Batrachochytrium dendrobatidis* (*Bd*), to high depth and characterized rapidly changing features of its genome that we believe hold the key to the worldwide success of this organism. Our analyses show threefold processes that contribute to the generation of *de novo* diversity. Firstly, we show that the majority of wild isolates manifest chromosomal copy number variation that changes over short timescales. Secondly, we show that cryptic recombination occurs within all lineages of *Bd* leading to large regions of the genome being in linkage equilibrium and is preferentially associated with classes of genes of known importance for virulence in other pathosystems. Finally, we show that these classes of genes are under directional selection and some are hotspots for recombination. These analyses show that *Bd* manifests an unusually dynamic genome that is potentially being shaped by its association with the amphibian host. The rates of variation that we document likely explain the phenotypic variability that has been reported for *Bd*, and indicate that the highly dynamic genome of this pathogen has contributed to its success across multiple biomes and host-species.

4.2 Introduction

A diverse cadre of fungi and fungal-like oomycetes have recently taken centre stage as emerging infectious diseases (EIDs) owing to their increasing impact on animals, plants and wider ecosystem health [Fisher MC *et al.*, 2012]. The widespread emergence of this class of pathogens shows that they are able to successfully adapt to infect diverse hosts and ecological niches, suggesting that their genomes are able to respond rapidly to natural selection [Fisher MC *et al.*, 2012; Raffaele R *et al.*, 2012]. This idea finds widespread support; for

example, horizontal transfer of whole chromosomes [Ma L *et al.*, 2010] and accelerated evolution across functional domains in effector genes [Win J *et al.*, 2007] are associated with rapid host-adaptation and changes in virulence across lineages and species. Maintaining the pool of genetic diversity necessary to respond to selection is facilitated by the ability of fungi to utilise multiple reproductive modes, including cryptic recombination that enables inbreeding, outcrossing, hybridization, and the generation of diversity via parasexual mechanisms [Heitman J, 2010]. These features are suspected to have contributed to the rise of contemporary fungal EIDs, which play a major role in host population declines across a broad swathe of plant and animal species [Fisher MC *et al.*, 2012; Wittenberg AHJ *et al.*, 2009; Cooke DEL *et al.*, 2012].

In recent years, whole genome sequencing has led to the characterization of novel mechanisms driving dynamic genome structure in microbial eukaryotes. In particular, it is increasingly apparent that pathogenic fungi manifest highly plastic genome architecture in the form of variable numbers of individual chromosomes, known as chromosomal copy-number variation (CCNV) or aneuploidy. This feature has been identified across the fungal phylum Ascomycota, ranging from *Botrytis cinerea* [Büttner P *et al.*, 1994], *Histoplasma capsulatum* [Carr J *et al.*, 1998], *Saccharomyces cerevisiae* [Sheltzer JM *et al.*, 2011], *Candida albicans* [Abbey D *et al.*, 2011] and the Basidiomycota *Cryptococcus neoformans* [Lengeler KB *et al.*, 2001; Hu G *et al.*, 2011; Sionov E *et al.*, 2010]. The mechanism(s) generating chromosomal CCNV in fungi are not yet well understood, but are thought to occur as a consequence of nondisjunction following meiotic or mitotic segregation [Reedy JL*et al.*, 2009], followed by selection operating to stabilise the chromosomal aneuploidies [Hu G *et al.*, 2011] . Although stress occurring as a consequence of either host response or exposure to antifungal drugs has been linked to a rapid rate of CCNV in *Candida* [Forche A*et al.*, 2009], it is currently unclear to what extent this contributes to broader rates of CCNV in fungi. However, dynamic numbers of chromosomes could offer routes to potentially advantageous phenotypic changes *via* several mechanisms such as over expression of virulence-factors [Hu G *et al.*, 2011] or drug efflux pumps [Kwon-Chung KJ *et al.*, 2012], the maintenance of diversity through homologous recombination [Forche A *et al.*, 2008], increased rates of mutation and larger effective population sizes [Arnold B *et al.*, 2012], or by purging deleterious mutations through non-disjunction during chromosomal segregation [Schoustra SE *et al.*, 2007]. Thus, CCNV likely represents an important, yet uncharacterized, source of *de novo* variation and adaptive potential in many fungi and other non-model microbial pathogens.

A contemporary EID that is gaining substantial notoriety is the aquatic chytrid fungus

Batrachochytrium dendrobatidis (*Bd*), which has so far been identified in over 50 countries worldwide and infecting over 500 species of amphibians [Fisher MC *et al.*, 2009; *Bd*-maps]. One of the most enigmatic aspects of *Bd*'s population genetic structure has been the low levels of genetic variation identified between globally distributed isolates. However, recent studies have shown the existence of up to five separate lineages [Farrer RA *et al.*, 2011; Goka K *et al.*, 2009; Schloegel LM *et al.*, 2012], one of which is shown to have undergone a worldwide range expansion in the 20^{th} Century. We recently compared the genomic diversity of this 'Global Panzootic Lineage' (*Bd*GPL) against that of a separate, distantly related ($\sim$ 1,000 ybp) lineage that appears to have originated in South Africa (named *Bd*CAPE), using SOLiD sequencing. *Bd*GPL was found to harbour evidence of historical recombination, manifested as patchily distributed heterozygosity, and phylogenetic incongruency across small spatial scales that we hypothesised has resulted from ongoing recombination [Farrer RA *et al.*, 2011]. Therefore, despite the lack of any known sexual meiotic mechanisms in its life cycle, *Bd* clearly has a more dynamic genome than a purely clonal, mitotic mode of reproduction would suggest. Here, we further our previous study with a new global panel of isolates that were subjected to high-depth Illumina sequencing in order to better understand cryptic genomic features that are associated with the rapid ascendancy of this pathogen.

4.3 Materials and Methods

4.3.1 Library preparation and sequencing

Twenty-two isolates that had been collected from nine countries and four continents were chosen for sequencing (Table 4.1). Paired-end Libraries were constructed according to the protocols provided by Illumina sequencing (Truseq kit). DNA was sheared into 150-500bp fragments using a Covaris S2 sonicator. After end-repair and size selection using magnetic beads, barcoded adapters (Illumina 12plex tags) were ligated to the fragments and PCR amplified using 10 cycles. The final libraries were quality-controlled on a bioanalyzer and quantified with qPCR. Two pools of Libraries were used, and sequenced on 2 lanes of an Illumina HiSeq 2000 machine. The output read length was 100bp. Sequence data was processed using RTA version 1.12.4.2, with default filter and quality settings. Sequence files were generated with CASAVA 1.7. Fastq Conversions were performed with a custom script.

The genome sequence and feature file for the chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*) strain JEL423 was downloaded from http://www.broadinstitute.org/ (GenBank project accession number AATT00000000). The feature file for JEL423 had all but the longest splice variants removed for each gene leaving 8794/8819 genes. We aligned all of our reads to the genome sequence using Burrows-Wheeler Aligner (BWA) v0.5.9 [Li H *et al.*, 2009] with default parameters and converted to Samtools mpileup format using SAMtools v.0.1.18 [Li H *et al.*, 2009].

Collection site	Amphibian host	Year	Collector	PN	Culture reference	Sequenced depth (X)	Aligned depth (X)
Canada, Vancouver Island	*L. catesbeianus*	2009	PH	>3	VC1 (CA)	52.75	49.17
England, Cumbria	*E. calamita*	2010	PM	2	SFBC014 (GB)	115.32	106.96
Ethiopia, Hotcho	*A. enseticola*	2011	DG	2	ETH2 (ET)	68.90	58.66
Ethiopia, Telilia Stream	*Leptopelis sp.*	2011	DG	2	ETH4 (ET)	166.03	152.87
France, Lac d'Aule	*A. obstetricans*	2010	MF	2	AUL (FR)	195.89	175.48
France, Loire et Cher	*L. catesbeianus*	2010	CM	3	RC5.1 (FR)	85.28	67.65
France, Madamette	*A. obstetricans*	2010	MF	2	MAD (FR)	127.41	110.29
Mallorca, TdF	*A. muletensis*	2007	MF	>3	TF5a1 (ES)	150.33	133.33
Panama, Guabal	*P. lemur*	2004	JEL	>3	JEL423 (PA)	53.32	48.48
Sardinia, AP	*D. sardus*	2010	TG	2	AP15 (IT)	179.93	164.91
Sardinia, Monte Olia	*D. sardus*	2010	TG	2	MODS27 (IT)	52.84	49.17
Sardinia, Monte Olia	*D. sardus*	2010	TG	2	MODS28 (IT)	160.15	148.47
Sardinia, Scuponi	*D. sardus*	2010	TG	2	SP10 (IT)	129.60	115.46
South Africa, Mount	*A. vertebralis*	2010	TG	2	MG1 (ZA)	81.71	61.64
South Africa, Pinetown	*A. angolensis*	2011	TG	3	SA1d (ZA)	148.72	136.69
South Africa, Pinetown	*A. angolensis*	2010	TG	2	SA4c (ZA)	180.52	161.82
South Africa, SilverMine	*A. fuscigula*	2010	TG	2	MG4 (ZA)	131.16	122.30
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	>3	ACON (CH)	167.19	144.62
Switzerland, Gamlikon	*A. obstetricans*	2008	TG	>43	APEP (CH)	110.43	100.99
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	>43	CON2A (CH)	115.29	102.17
Switzerland, Itingen	*A. obstetricans*	2010	RF	2	BLI1 (CH)	52.76	49.32
Switzerland, Waltensberg	*A. obstetricans*	2010	RF	2	BEW2 (CH)	144.54	132.07

Table 4.1: Samples used and details of alignments. *Bd* isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [Berger *et al.*, 1998]. TdF = Torrent des Ferrerets, AP = Affluente Pisharoni. Passage numbers (PN) are best approximations from records prior to DNA extractions in January and May 2011. The sequenced depth and aligned depth were calculated from the number of nucleotides in all or aligned reads respectively and divided by 24Mb (the length of the *Bd* JEL423 genome assembly). Amphibian hosts include *Afrixalus enseticola* (Ethiopian Banana frog), *Alytes muletensis* (Mallorcan Midwife Toad), *Alytes obstetricans* (Common Midwife Toad), *Amietia angolensis* (Angola River Frog), *Amietia fuscigula* (Cape River Frog), *Amietia vertebralis* (Ice Frog), *Discoglossus sardus* (Tyrrhenian Painted Frog), *Epidalea calamita* (Natterjack Toad), *Leptopelis sp.* (Big eyed Tree Frog), *Lithobates catesbeianus* (American Bullfrog), *Phyllomedusa lemur* (Lemur Leaf Frog). CM = Claude Miaud, DG = David Gower, JEL = Joyce Longcore, MF = Matthew Fisher, PH = Phineas Hamilton, PM = Peter Minting, RF = Rhys Farrer, TG = Trent Garner.

4.3.2 *in vitro* divergence of independent replicate lines of *Bd*CH

An isolate of *Bd* from a Swiss *Alytes obstetricans* (isolate 0739) was subcultured into control (ACON) and peptide-treated (APEP) culture flasks containing 10ml 1% tryptone media supplemented with 1% penicillin-streptomycin (Sigma) to reduce the risk of bacterial contamination. Cultures were incubated at 18°C and passaged every 4-5 days by scraping the side of the flask and transferring 1ml into 9ml fresh media. Peptide-treatment included addition to the media of 80μg ml^{-1} skin defense peptides collected from Pelophylax esculentus (n=15 combined) according to Daum JM *et al.* [Daum JM *et al.*, 2012]. This was equivalent to the IC_{50}, or the concentration at which growth of *Bd* was inhibited by 50%. These three isolates were included in the panel of 22 for whole-genome sequencing.

Phenotypic changes following selection were assessed on the above sequenced isolates and additional replicates from treatment and control conditions. Growth of the isolates in five replicate wells in media containing 100μg ml^{-1} skin defense peptides in water was compared to growth of the isolate with water added (positive control) or with water added and heat killed (negative control) according to standard protocols for determining growth inhibition [Daum JM *et al.*, 2012].

After 40 passages with or without amphibian skin defence peptides added to media, *Bd* growth was measured to determine the effect of selection on resistance to peptides. Mean growth inhibition (%±SE) of *Bd* given 100μg ml^{-1} peptide was: 14.23±11.48 for *Bd*CH APEP, 59.44 ± 14.50 for *Bd*CH ACON, and 55.74±5.98 for *Bd*GPL. When replicate cultures were compared, the average growth inhibition of *Bd*CH after selection for peptide resistance was 11.43±3.02. The value was significantly reduced compared to controls at 48.81±6.72 (independent t-test, t=-3.980, df=3, p=0.028).

4.3.3 Optimisation of alignments and SNP calling parameters

We resequenced the genome for isolate JEL423 to act as a control for sequencing, alignment and SNP calling. To test for ability to accurately call polymorphisms from the alignments, we aligned the reads from JEL423 to two modified reference *Bd* JEL423 genomes (1nt/1Kb and 1nt/100nt within the coding regions; CDS) reflecting the different levels of sequence divergence found within and between different lineages of *Bd* (Fig. 4.1) using the comparison of false discovery rate (FDR) method [Farrer RA *et al.*, 2013; Chapter 2]. We next tested the

SNP-callers: Binomial SNP-Caller from Pileup (BiSCaP) v0.1 and v0.11 using a minimum depth of 4 reads [Farrer RA *et al.*, 2013; Chapter 2] and the SAMtools mpileup command piped to Bcftools v0.1.17-dev and filtered using vcfutils.pl with default parameters [Li H *et al.*, 2009]. We tested each method with variable minimum required read-depths and compared results after filtering SNPs called from alignment to a non-modified reference.

In order to assess heterozygous variants, we randomly chose and modified 1nt/Kb and 1nt/100nt within the CDS using the "HET" setting of IRMS.pl, which first generates a duplicate (homologous) genome. We then simulated single-end reads from these modified sequences to the same depth as the 'real' data using simLibrary and simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/) using the default runfile (s_3_4x), which describes how "noise and cluster intensities are distributed in a real run of an Illumina machine", and aligned those reads to the non-modified reference genomes. Finally, the same SNP-callers used for the FDR of homozygous SNPs were used to call heterozygous bases. To compare the heterozygous-base calls with those achievable from our previous SOLiD datasets, we applied the same method using simulated 10.17X deep 30mers.

We achieved a variety of success rates (true positives, TP >false positives, FP) at calling the polymorphisms depending on the SNP-calling method and its settings. SNP-calling from our previous *Bd* JEL423 SOLiD resequencing sequences [Farrer RA *et al.*, 2011; Chapter 3] achieved 40-50% TP for both simulated divergences and both versions of BiSCaP, whilst the Illumina data presented here achieved >95% TP using either BiSCaP versions 0.1 or v0.11 or SAM/BCFTools for both simulated divergences. Simulated reads to the depth and read-length of the previous SOLiD sequences were used to assess its ability at called heterozygous positions. BiSCaP v0.11 achieved 40-60% TP and 1-6% FP for both simulated divergences. Aligning simulated reads to the depth and read-length of the new Illumina sequences achieved 81.38% TP and 0.29% FP for the greater divergence and 92.17% TP and 0.23% FP for the lesser divergence. Filtering SNPs called from a non-modified reference revealed that many of the FP SNPs identified by either BiSCaP or SAM/BCFTools were independently found without the simulated divergence, and are therefore likely to reflect either real errors in the reference genome or discrepancies between the two *Bd* JEL423 isolates. Using BiSCaP v0.11 with a minimum of four reads (after taking into account polymorphic positions identified using a non-modified reference) called the optimum numbers of both TP and FP and was therefore used to call homozygous SNPs and bi-allelic heterozygous positions across each of the 22 genomes.

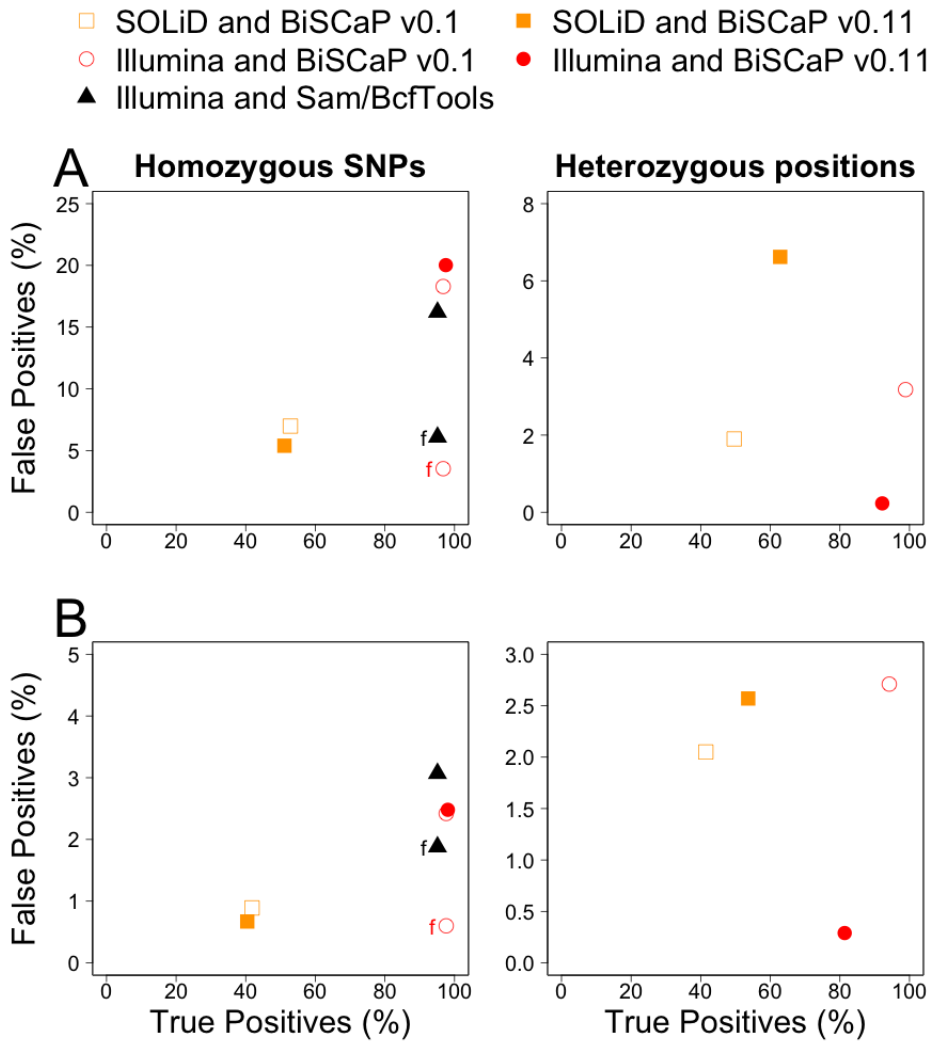


Figure 4.1: The previous SOLiD reads [Farrer RA *et al.*, 2011] and the new Illumina paired end reads of *Bd* isolate JEL423 were aligned to a modified JEL423 reference sequence. Additionally, simulated reads from a heterozygous reference sequence were made to the depths of the Illumina and SOLiD datasets. Single Nucleotide Polymorphisms (SNPs) and heterozygous positions were then called and the False Discovery Rates (FDR) ascertained. The SNP-caller BiSCaP v0.11 was tested using default settings, and SAM/BCFTools with VCFUtils was tested for its ability to call SNPs using its default settings. SNPs were also filtered for those found without first modifying the reference sequence (f=filtered). (**A**) 1nt/Kb simulated SNPs or heterozygous positions (12,458 in total) within the coding region (CDS) (**B**) 1nt/100nt simulated SNPs or heterozygous positions (124,588 in total) within the CDS region. The new Illumina data was able to recover >95% of true positive SNPs and >80% true positive heterozygous positions using BiSCaP v0.11, outperforming the previous lower-depth SOLiD sequences.

To check our variants were not a function of difficulties in mapping reads accurately, we checked the number of best hits in the alignment (SAM file) for every read over every variant. For this analysis, we ignored entries missing the optional field specifying the number of best hits (XO tag). Reads with exactly 1 best hit were considered uniquely mapped, whilst those with zero or >1 were considered as non-unique.

4.3.4 Phylogenetic analysis of the nuclear genomes

Entirely covered and verified in all (ECVA) polymorphisms comprising 275,009 positions were extracted from each of the isolates and concatenated into FASTA files for phylogenetic analysis. Bi-allelic loci were represented in the FASTA file by the corresponding ambiguity codes. FASTA files were converted into Nexus files and a tree constructed using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm in PAUP and visualised using Figtree [Drummond AJ *et al.*, 2007]. In addition, we extracted 36,309 positions fulfilling the same criteria for the newly sequenced isolates and the previous 20 isolates sequenced using the SOLiD platform [Farrer RA *et al.*, 2011; Chapter 3] comprising 42 isolates including *Bd*GPL JEL423 (PA) and *Bd*CAPE TF5a1 (ES) sequenced by both platforms. Finally, we performed the same analysis using just the homozygous positions (reference or SNP in all) comprising 218,269 positions across the Illumina sequenced isolates and 8,457 positions across the Illumina and SOLiD sequenced isolates.

4.3.5 Identifying chromosome copy number variation (CCNV)

We predicted ploidy levels across the genomes using depth of coverage and percent of reads specifying two most frequent alleles at a loci. Depth of read coverage across the genome was visualised using non-overlapping sliding windows of length 1414nt (half the mean length of the *Bd* transcripts) and 10Kb. To quantify changes, we performed *t*-tests on the mean depths across the largest supercontig (supercontig 1) against each subsequent supercontig for each isolate using a cut-off of $p<5^{-10}$. Chromosome 14 had a large peak of depth in all isolates, which occurred over a long stretch of rDNA sequence identified using the non-redundant BLAST database. We therefore excluded Chr14 from the following analysis.

To determine the ploidy of the largest chromosomes (1-13,15-16) of the 23 isolates (97% of the total genome length), we calculated the percent of reads specifying the two most

frequent alleles for each chromosome in each isolate separately with a minimum depth cut-off of 4 reads for both alleles. We then binned the percent of reads aligned over each base in the genome agreeing with a nucleotide between 47-53% (expected even ploidy/bi-allelic) and 30-36% and 63-69% (expected odd ploidy/tri-allelic). To check that these values could be found given different subsets of each chromosome in each isolate, we calculated 1000 bootstraps for either predominance of bi-allelic or tri-allelic peaks using a 5% cut-off (5%<x <95%).

We then combined both the read-depth and allele-frequencies to predict chromosome copy number and ploidy for each isolate. Specifically, when supercontig 1 for an isolate had predominantly tri-allelic peaks, then all chromosomes with no significant change in mean depth and also predominantly tri-allelic peaks were considered triploid. Conversely, should supercontig 1 have predominantly bi-allelic peaks, then all supercontigs with no significant change in mean depth and also predominantly bi-allelic peaks were considered diploid. When a supercontig had different allele peaks (bi- instead of tri- or visa-versa) to that of supercontig 1, and a significant change in depth, it was considered a ploidy below or above that of supercontig 1. When mean depth and allelic-peaks did not match, depth alone could be used to predict the ploidy.

4.3.6 Identifying gene groups and names

All 8819 *Bd* JEL423 transcripts were searched for secretion signals, protease domains and carbohydrate binding protein domains using SignalP3.0 [Bendtsen JD *et al.*, 2004], Merops [Rawlings ND *et al.*, 2010] and Procarb604 v1 [Malik A *et al.*, 2010] respectively. We were also interested in a number of genes that showed homology to the C-terminal of the Crinkler (CRN) family of oomycete effector proteins found in the *Phytophthora* genus [Joneson S *et al.*, 2011; Sun G *et al.*, 2011]. First, gene annotation (downloaded) and blastx searching was performed against the non-redundant BLAST database with Blast2Go [Götz S *et al.*, 2008] using a 1e-05 e-value cut-off, which found hits to >64% of the transcripts. BLAST annotations were also screened for the key terms 'crinkler', 'peptidase', 'protease', 'metalloproteinase', 'chitin', 'abc transporter' and 'abc multidrug transporter' to further identify candidate gene groups. In total we identified 599 non-secreted proteases, 93 secreted proteases, 41 secreted chitin-associated genes, 48 non-secreted chitin-associated genes, 110 Crinkler (CRN-like), 985 non-categorised secreted genes, 29 ABC transporters, and the remaining 6897 (78%) uncharacterized genes. Many genes were identified in the same functional group using the

different methods, and no gene was identified in more than one group.

In addition to grouping genes using naming and searching protease, carbohydrate and secretion peptide databases, we grouped genes entirely based on sequence. First, all protein sequences were Blastp searched against themselves, again with a $1e^{-5}$ e-value cut-off. Gene tribes were identified using the MCL software [Dongen SV, 2000] with settings stream-mirror, stream-neg-log10 and I=2 (recommended settings). Where genes had multiple transcripts, only the longest was considered. From these 2 methods we identified 621 clusters of genes with 2 or more transcripts that matched a gene from the NCBI Sequence database, encompassing 3132 genes (36% of total transcripts).

The largest gene tribe (Tribe 1) included 564 genes encompassing genes from all 8 categories of genes, including a large number of CRN-like, Proteases and ABC Transporter genes (from largest to smallest): 75/110 (68%) CRN-like genes, 14/29 (48%) ABC-transporters, 39/93 (42%) proteases (secreted), 15/48 (31%) chitin-associated (non-secreted), 7/41 (17%) chitin-associated (secreted), 69/599 (12%) proteases (non-secreted), 33/985 (3%) uncharacterised (secreted) genes and the rest (305) uncharacterized (non-secreted). Although this prevented using this grouping method along side the method based on blast searches, it demonstrates that some shared homology exist between many of these genes that are putatively involved in virulence, despite having potentially very different roles and undergoing different selection pressures.

4.3.7 Detecting recombination and hybridization amongst *Bd* isolates

To investigate recombination in *Bd*, we phased the bi-allelic heterozygous positions (in Variant Call Format) identified by BiSCaP v0.11 according to overlapping reads. Firstly, the corresponding SAMTools sorted BAM file to each isolates VCF was searched for reads that mapped to the genome without any indels, and also covered two or more heterozygous positions. The phase for each of the overlapping heterozygous positions was then calculated. To ensure well supported haplotypes corresponding to only two alleles per loci, any heterozygotes with less than four overlapping reads over each of the bases, and those with <90% agreeing with a single phase (such as heterozygous positions in different phases or different bases, either of which may be due to polyploidy or error) were filtered from all analysis. Phased heterozygotes were given unique numbered identifiers and placed in an ordered VCF

for each isolate, which could be used to distinguish which positions were in phase with others. Homozygous SNPs within phased groups were also combined into that group. A diagram illustrating the method for phasing is shown in Fig. 4.2.

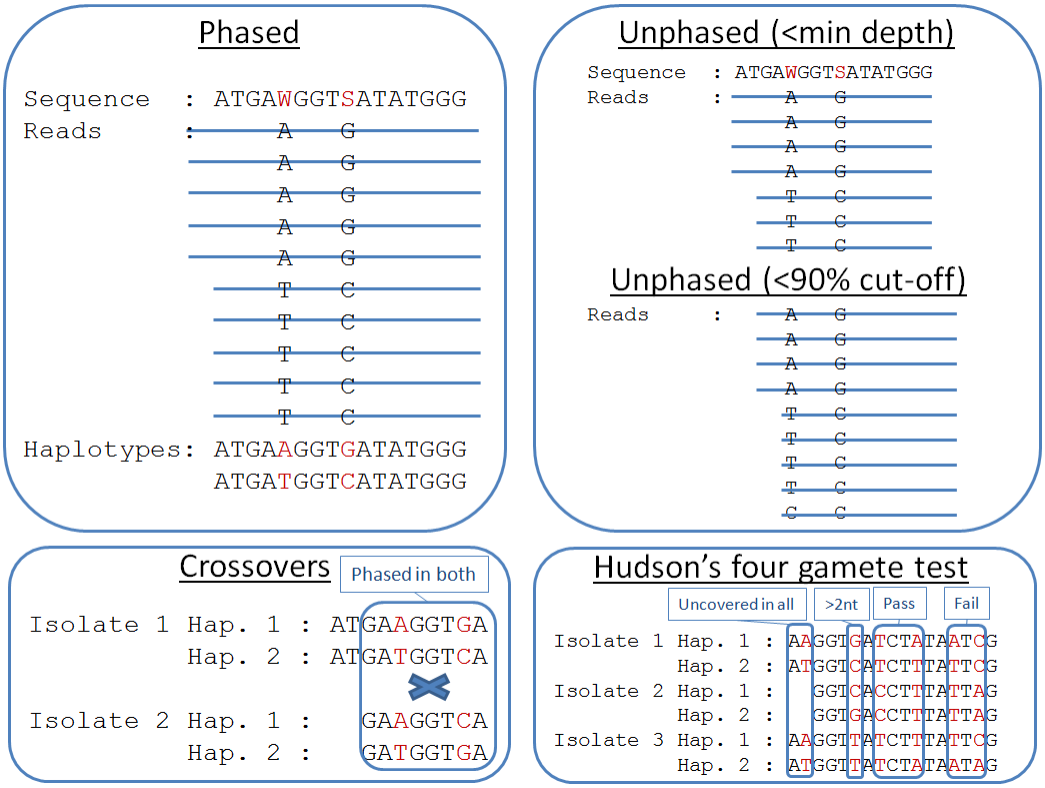


Figure 4.2: Illustrations of how phased haplotypes were extracted from the alignment. Heterozygous positions that did not pass the minimum depth or percent phased cut-offs, along with examples of pairwise crossovers and outcomes for a four-gamete test between three isolates.

In order to extend haplotypes, we included all homozygous SNPs within a given phased group and created new phase groups consisting of consecutive homozygous SNPs. Finally, we identified consecutive "SNP - Bi-allelic HET - SNP", which can therefore also be considered either a new separate phase group or joined to the previous phase group if no other un-phased variant occurs between them.

Next, a 'phased in all' file (PIA) was constructed from overlapping phase groups identified across all VCFs within a given lineage. Phased groups that were <10nt long were filtered out to avoid short repeats or closely associated heterozygotes resulting from poorly resolved indels as reported by Mckenna A *et al.* [McKenna A *et al.*, 2010]. Next, overlying variants

(homozygous SNPs and phased heterozygous positions) were extracted from each VCF. Isolates lacking any variant at those positions had their corresponding SAMTools pileup lines extracted, and reference-base calling was performed using the same default settings of BiSCaP v0.11 [Farrer RA *et al.*, 2013; Chapter 2) as used for the variants. All positions that were not called as a reference-base were excluded from further analysis. Positions lacking any variant, but were called as a reference base were included. Next, all fixed loci in a given lineage were filtered out. Haplotypes with less than two remaining loci or any that have been reduced to <10nt were also filtered. The remaining haplotypes consisting of those polymorphic loci were constructed (2 haplotypes per isolate as phased groups were extracted from bi-allelic regions).

From these haplotypes, we were then able to perform 4-gamete tests between every combination of loci in a haplotype. We also calculated the Index of Association (I_A) [Smith JM *et al.*, 1993] to examine LD in *Bd* using the calculations given in the software Multilocus [Agapow PM *et al.*, 2001]. Firstly, the mean distance (loci that are different) h_j is calculated using the number of individuals in the test (haplotypes or haploid isolates) n, which carry the $_i$th allele at the $_j$th locus, using the equation:

$$h_j = 1 - \frac{\sum n_i(n_i - 1)}{n(n-1)} = \frac{n}{n-1}\left(1 - \sum p_i^2\right) \tag{4.1}$$

Next, the number of possible pairs (n_p) can be calculated as $n_p = n(n\text{-}1)/2$. Then for the distance (zero or one) d at locus $_j$ (a single loci of h_j shown in 4.1), the variance of the mean difference (var_j) can be calculated as:

$$var_j = \frac{\sum d^2 - \frac{(\sum d)^2}{n_p}}{n_p} \tag{4.2}$$

The total distance between two isolates or haplotypes can then be calculated and expressed as D. The variance of distance (V_D) is calculated from D and n_p with:

$$V_D = \frac{\sum D^2 - \frac{(\sum D)^2}{n_p}}{n_p} \tag{4.3}$$

There are a number of different methods for ascertaining significance for V_D, which can then be used to provide confidence in measures of LD or LE. One such method is to calculate

a one-tailed 95% critical value that V_D must be greater than to demonstrate a confidence in LD [Haubold B *et al.*, 1998; Brown AHD *et al.*, 1980], which is written as L:

$$L = E(V_D) + 1.645\sqrt{Var(V_D)} \tag{4.4}$$

Where the expected variance of distance or $E(V_D)$ is obtained by summing the h_j value times one minus h_j for each loci, written as:

$$E(V_D) = \sum_{j=1}^{r} h_j(1 - h_j) \tag{4.5}$$

A close approximation to the variance of V_D suggested by Haubold B *et al.* [Haubold B *et al.*, 1998] is that given by Anthony Brown [Brown AHD *et al.*, 1980]:

$$Var(V_D)_{old} = \frac{(n-1)^2}{n^3} m4 - \frac{(n-1)(n-3)}{n^3} E(V_D)^2 \approx \frac{m4 - E(V_D)^2}{n} \tag{4.6}$$

where:

$$m4 = \sum h_j - 7\sum h_j^2 + 12h_j^3 - 6h_j^4 + 3(\sum h_j - \sum h_j^2)^2 \tag{4.7}$$

Separately, the I_A can be calculated using var_j and V_D with:

$$I_A = \frac{V_D}{\sum var_j} - 1 \tag{4.8}$$

As an alternative statistic, we calculated the $\overline{r}_d$ value, which uses the same equations for var_j and V_D and largely accounts for the I_A dependency on the number of loci (which biases values greater than zero). Both I_A and $\overline{r}_d$ are based on unordered alleles, and both give values of between zero (no LD) and one (LD). The equation for calculating $\overline{r}_d$ is given by:

$$\begin{aligned}\overline{r}_d &= \frac{\sum\sum cov_{j,k}}{\sum\sum\sqrt{var_j \cdot var_k}} \\ &= \frac{V_D - \sum var_j}{2\sum\sum\sqrt{var_j \cdot var_k}}\end{aligned} \tag{4.9}$$

To study whether recombination had occurred between the separate *Bd* lineages since their divergence, we calculated Weir's [Weir BS, 1996] formulation of Wright's fixation index (θ) using the equations set out in the software Multilocus [Agapow PM *et al.*, 2001]:

$$\theta = \frac{\sum Q_2 - \sum Q_3}{\sum(1 - Q_3)} \tag{4.10}$$

Where Q2 and Q3 are the probabilities two intra- or inter-population alleles are identical by descent, respectively, and given by the equations:

$$\begin{aligned}Q2 &= \frac{X - r}{r(\overline{n} - 1)} \\ Q3 &= \frac{1}{(r(r-1)\overline{n}n_c}\left(\gamma - \frac{\overline{n}(n_c - 1)}{\overline{n} - 1}X\right) + \frac{\overline{n} - n_c}{n_c(\overline{n} - 1)}\left(1 - \frac{1}{r-1}X\right)\end{aligned} \tag{4.11}$$

Which uses:

$$\begin{aligned}X &= \sum_u \sum_{i=1}^{r} n_i \tilde{P}_{ui}^2 \\ \gamma &= \sum_u \left(\sum_{i=1}^{r} n_i \tilde{P}_{ui}\right)^2 \\ n_c &= \frac{1}{r-1}\left(\sum_{i=1}^{r} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}\right)\end{aligned} \tag{4.12}$$

Where r = the number of populations, n_i = the number of individuals sampled from population i, $\overline{n}$ = mean number of individuals sampled per population, and $\tilde{P}_{ui}$ = the fre-

quency of the uth allele in the ith population.

Finally, to study if certain gene categories were recombining; we ran Hypergeometric tests for non-redundant crossovers (NR XO) in each of the identified lineages separately across each of the gene categories.

4.3.8 Detecting patterns of mutation and selection in *Bd*

To identify genes that are present in the reference sequence and absent in our panel of isolates (presence/absence polymorphism), we identified any genes with zero depth of coverage (no reads aligned to them). To study the patterns of mutation across the nuclear genome, we categorised each of the variant-types by their location in the genome in terms of coding regions (CDS), introns and intergenic regions. Next, we categorised each of the mutations within CDS regions as synonymous and non-synonymous mutations. Each gene category was tested for significant enrichment for SNPs, synonymous mutations, non-synonymous mutations, heterozygous positions and phased positions using Hypergeometric tests.

To further measure selection across the genome we measured the rates of synonymous substitution (dS), non-synonymous substitution (dN) and omega ($dN/dS = \omega$) for every gene in every isolate using the yn00 program of PAML [Yang Z, 2007] implementing the Yang and Nielsen method [Yang Z *et al.*, 2000]. To further look for selection along the three lineages of *Bd*, we analysed each of those genes with the Branch site model (BSM) A (model=2, NSsites=2, fix_omega=0) compared with the null model (model=2, NSsites=2, fix_omega=1, omega=1) for each lineage (using the UPGMA tree from all verified homozygous loci in all isolates) implemented in the program codeml in PAML [Yang Z, 2007]. Next, we calculated 2 * the log likelihood difference between the two compared models (2D') with two degrees of freedom, and identified any with values greater than 8.1887 and 11.4076 (5% and 1% significance after Bonferroni correction respectively).

4.4 Results

4.4.1 SNP discovery and phylogenetics

We sequenced 22 isolates of *Bd* with a geographical distribution spanning five continents to a high depth (52-195X; Table 4.1) using the Illumina HiSeq 2000 platform. Sequences are deposited in the NCBI Short Read Archive under the submission accession number SRA058657. These reads were then aligned to a reference sequence assembly for isolate JEL423 [MIT JEL423 sequence] using BWA [Li H *et al.*, 2009] and polymorphisms were identified using BiSCaP v0.11 [Farrer RA *et al.*, 2013; Chapter 2]. In total, we identified 904,000 SNPs, 761,000 bi-allelic heterozygous positions and 95,000 multi-allelic heterozygous positions (Table S1), which were distributed across 425,000 loci. Those loci in turn were made up of 218Kb non-redundant (at unique loci) SNPs (average 9.3SNPs/Kb) and 279Kb non-redundant bi-allelic heterozygous positions (average 11.9hets/Kb) (Table 4.2).

Across the *Bd* JEL423 genome, 96% was covered by at least four reads in every isolate. Additionally, 65% of the total identified variant sites (275,000; 11.8/Kb) were called as either reference or polymorphic in all 22 isolates (entirely covered and verified in all; ECVA) amounting to >10X the number of 'covered in all' polymorphic loci previously found using the ABI SOLiD 3 platform [Farrer RA *et al.*, 2011], owing in part to the higher depth of sequencing coverage (Table 4.2). UPGMA trees were made from both the 275Kb EVCA polymorphic sites and the 218Kb EVCA homozygous sites (Figs. 4.3A and 4.4A respectively). In addition, we identified 36Kb EVCA polymorphic loci and 8,457 EVCA homozygous loci across all of the 42 sequenced isolates (Illumina dataset presented here, SOLiD dataset presented previously [Farrer RA *et al.*, 2011; Chapter 3] (Figs. 4.3B and 4.4B respectively). All four trees demonstrated three divergent lineages previously identified [Farrer RA *et al.*, 2011; Chapter 3].

Amongst *Bd*GPL, JEL423 for both platforms came together in both trees. Among the more distant lineages, the Illumina sequenced *Bd*CAPE isolates clustered separately from the SOLiD sequenced *Bd*CAPE isolates, including in the replicate isolate TF5a1, likely demonstrating sequencing errors and lack of depth in the previous SOLiD dataset. In addition to showing that our new panel of isolates belonged to three (*Bd*GPL, *Bd*CAPE, *Bd*CH) of the five suspected lineages of *Bd* [Farrer RA *et al.*, 2011; Goka K *et al.*, 2009; Schloegel LM *et al.*, 2012], our new panel extended both *Bd*GPL and *Bd*CAPEs known geographic range

(*Bd*GPL into Switzerland and Ethiopia and *Bd*CAPE into France).

Lineage	Culture reference	Homozygous Reference	Homozygous SNPs	Homozygous INS	Homozygous DEL	Bi-allelic Heterozygous Positions	Bi-allelic Heterozygous INS	Bi-allelic Heterozygous DEL	Bi-allelic positions with tri-allelic probabilities Positions	Bi-allelic positions with tri-allelic probabilities INS	Bi-allelic positions with tri-allelic probabilities DEL	Tri-allelic positions & indels
GPL	AP15 (IT)	23,301,357	7,799	651	1,347	26,815	403	401	2,264	576	645	8
GPL	AUL (FR)	23,225,020	17,062	699	1,354	23,380	322	353	1,944	613	673	13
GPL	BEW2 (CH)	23,305,979	5,050	644	1,323	25,274	430	404	4,206	773	857	5
GPL	BLI1 (CH)	23,266,824	3,001	505	978	29,532	443	456	6,654	884	976	8
GPL	ETH2 (ET)	23,251,245	7,885	677	1,276	33,596	449	475	5,939	958	1,070	13
GPL	ETH4 (ET)	23,289,744	8,390	606	1,186	28,782	391	435	2,615	602	655	6
GPL	JEL423 (PA)	23,210,747	2,845	599	1,182	24,147	353	338	6,207	776	885	8
GPL	MAD (FR)	23,286,937	13,445	687	1,355	22,503	338	348	3,873	963	1,071	12
GPL	MG4 (ZA)	23,296,876	5,231	600	1,262	27,398	456	420	4,327	691	798	10
GPL	MODS27 (IT)	23,270,109	4,524	526	991	27,541	439	423	6,547	847	911	10
GPL	MODS28 (IT)	23,315,018	5,094	611	1,272	26,905	383	426	3,248	678	733	8
GPL	SFBC014 (GB)	23,260,363	10,928	645	1,144	28,495	412	431	4,291	806	849	12
GPL	SP10 (IT)	23,299,834	7,515	659	1,364	26,134	382	438	3,782	792	826	2
GPL	VC1 (CA)	23,210,508	16,497	689	1,121	20,555	313	314	4,952	826	993	8
CH	ACON (CH)	22,870,756	100,582	1,518	2,116	47,560	637	799	5,629	1,673	2,234	89
CH	APEP (CH)	22,858,911	115,279	2,719	3,771	60,804	1,064	1,205	5,081	2,503	3,193	145
CH	CON2A (CH)	22,849,053	100,453	1,669	2,331	53,982	662	904	5,760	1,683	2,272	72
CAPE	MG1 (ZA)	22,829,514	87,846	1,895	2,293	47,717	486	659	4,492	1,652	2,025	78
CAPE	RC5.1 (FR)	22,796,731	88,969	1,908	2,349	45,529	529	644	4,365	1,578	1,922	71
CAPE	SA1d (ZA)	23,027,304	101,244	1,983	2,651	46,235	676	838	2,270	1,502	1,820	116
CAPE	SA4c (ZA)	22,990,652	97,061	1,741	2,292	44,106	557	727	2,380	1,212	1,694	72
CAPE	TF5a1 (ES)	22,999,308	97,252	1,865	2,480	44,114	592	764	3,478	1,520	1,955	76

Table 4.2: Polymorphisms and reference bases were identified in 22 *Bd* nuclear genomes relative to *Bd* JEL423 using BiSCaP v0.11 with default settings. Shown are tallies of each category of loci found in each separate isolate.

Extensive overlap of ECVA polymorphisms was found by calculating for every two pairwise isolate ((sample1 ∩ sample2) / ((sample1 + sample2) / 2)) x 100 as shown in Table 4.3. We found the greatest levels of variation amongst SNPs within *Bd*GPL (between 17% and 95%) compared with *Bd*CAPE (93-96%) or *Bd*CH (92-97%) (Fig. S2A). *Bd*GPL also had the greatest variation amongst heterozygous positions (between 32% and 75%) compared with *Bd*CAPE (44-55%) or *Bd*CH (34-55%) (Fig. S2B).

To check our variants were not a function of difficulties in mapping reads accurately, we checked the percent of uniquely mapped reads over each variant type in each isolate (Table 4.4). We found that on average >85% of our reads were uniquely mapped across the genomes, with homozygous SNPs having >90% uniquely mapped reads and bi-allelic heterozygous positions having >82% uniquely mapped reads. However, isolates belonging to the two divergent lineages had less uniquely mapped reads over bi-allelic heterozygous (averaging 68%, compared with 91% in *Bd*GPL). This reduction in uniquely mapped reads may result from structural variations such as gene presence/absence polymorphisms. Therefore, false positive rates over heterozygous positions may be greater than that found with our FDR

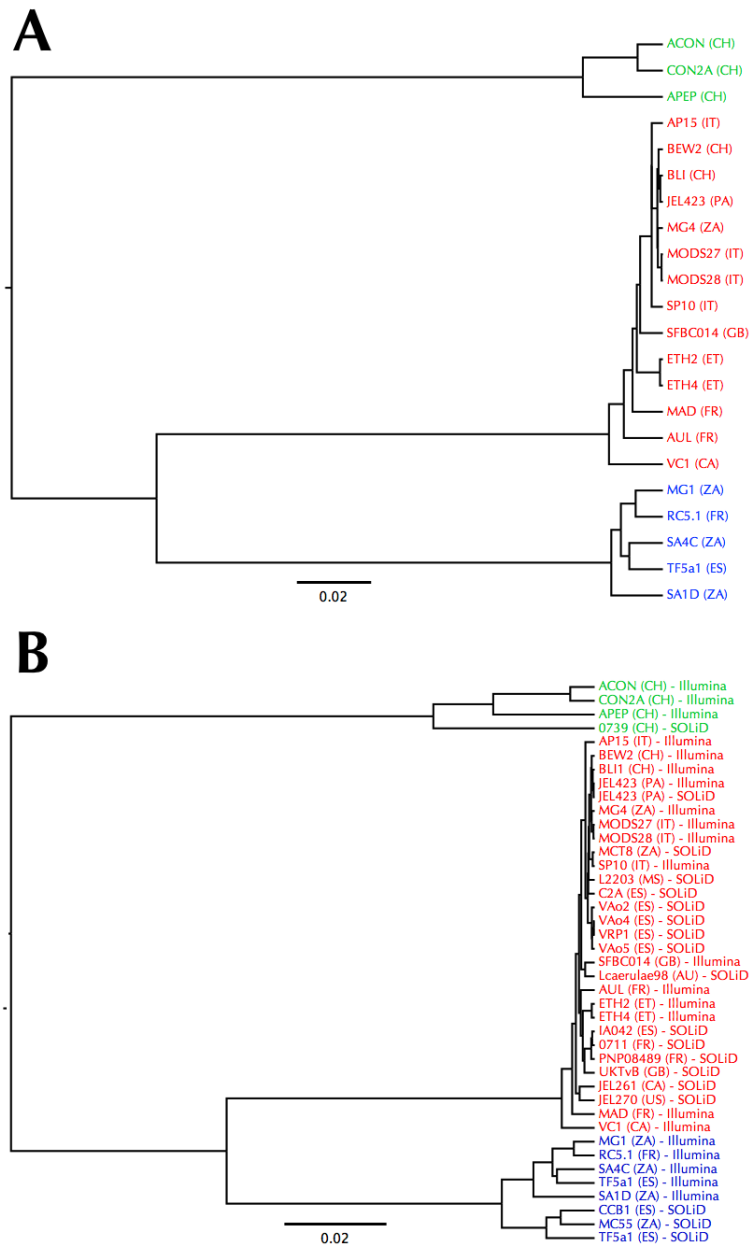


Figure 4.3: Phylogenetic trees were made using the UPGMA algorithm in PAUP from ECVA polymorphic positions identified in the nuclear genomes demonstrating three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH shown in red blue and green respectively). (**A**) A tree from 275Kb ECVA polymorphic positions identified from the Illumina sequencing (**B**) A tree from 36Kb ECVA polymorphic positions from Illumina and SOLiD sequencing.

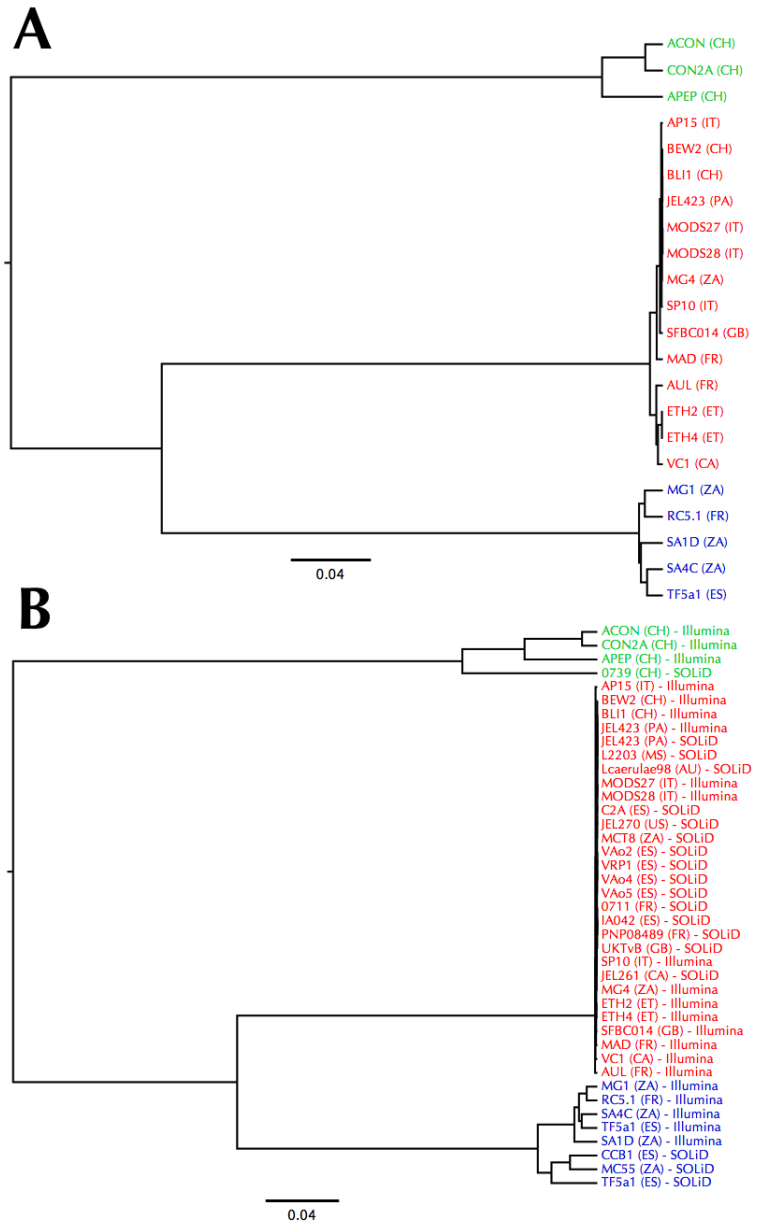


Figure 4.4: Phylogenetic trees were made using the UPGMA algorithm in PAUP from ECVA homozygous positions (either reference or SNP) identified in the nuclear genomes demonstrating three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH shown in red blue and green respectively). (**A**) A tree from 218Kb ECVA polymorphic positions identified from the Illumina sequencing (**B**) A tree from 8Kb ECVA polymorphic positions from Illumina and SOLiD sequencing.

Homozygous SNPs:

Isolate	BLI1 (CH)	MODS27 (IT)	BEW2 (CH)	MODS28 (IT)	MG4 (ZA)	SP10 (IT)	AP15 (IT)	ETH2 (ET)	ETH4 (ET)	SFBC014 (GB)	MAD (FR)	VC1 (CA)	AUL (FR)	MG1 (ZA)	RC5.1 (FR)	SA4c (ZA)	TF5a1 (ES)	SA1d (ZA)	CON2A (CH)	APEP (CH)	ACON (CH)
JEL423 (PA)	74	59	59	60	57	41	42	35	36	29	24	19	20	3	3	3	3	3	3	3	3
BLI1 (CH)		58	65	58	57	40	42	36	36	29	26	19	20	4	4	4	4	3	3	3	3
MODS27 (IT)			48	95	61	37	60	31	32	29	25	26	23	5	5	5	5	5	4	4	4
BEW2 (CH)				48	48	35	37	32	33	26	30	24	19	5	5	5	5	4	4	4	4
MODS28 (IT)					62	38	61	32	32	29	25	26	23	5	5	5	5	4	4	4	4
MG4 (ZA)						37	53	32	32	27	24	23	19	5	5	5	5	4	4	4	4
SP10 (IT)							53	26	26	42	28	18	38	7	7	7	7	7	6	6	6
AP15 (IT)								27	27	42	26	24	32	6	6	6	6	6	5	5	5
ETH2 (ET)									93	21	33	40	43	7	7	7	7	7	6	6	6
ETH4 (ET)										21	33	41	44	7	7	7	7	7	6	6	6
SFBC014 (GB)											25	17	53	9	9	9	9	9	8	7	8
MAD (FR)												41	34	12	12	12	12	11	10	9	10
VC1 (CA)													35	14	14	14	14	14	12	11	12
AUL (FR)														13	13	12	13	12	11	10	11
MG1 (ZA)															96	94	93	93	38	38	39
RC5.1 (FR)																94	93	93	38	38	38
SA4c (ZA)																	95	93	38	38	38
TF5a1 (ES)																		93	38	38	39
SA1d (ZA)																			38	38	39
CON2A (CH)																				92	97
APEP (CH)																					92

Heterozygous positions:

Isolate	BLI1 (CH)	MODS27 (IT)	BEW2 (CH)	MODS28 (IT)	MG4 (ZA)	SP10 (IT)	AP15 (IT)	ETH2 (ET)	ETH4 (ET)	SFBC014 (GB)	MAD (FR)	VC1 (CA)	AUL (FR)	MG1 (ZA)	RC5.1 (FR)	SA4c (ZA)	TF5a1 (ES)	SA1d (ZA)	CON2A (CH)	APEP (CH)	ACON (CH)
JEL423 (PA)	61	60	63	59	58	63	57	56	54	58	36	41	45	4	4	4	4	5	5	4	4
BLI1 (CH)		69	65	70	71	61	64	64	62	60	42	40	45	5	5	5	5	5	5	4	4
MODS27 (IT)			59	74	68	59	67	61	59	58	37	43	42	5	5	5	4	5	5	4	4
BEW2 (CH)				60	61	62	60	57	56	59	37	42	45	4	4	4	4	4	4	4	4
MODS28 (IT)					71	59	69	61	60	58	39	41	43	4	4	5	4	4	4	4	4
MG4 (ZA)						59	67	61	60	57	41	38	42	4	4	5	4	4	5	4	4
SP10 (IT)							58	60	59	58	40	43	48	4	4	4	4	4	4	4	4
AP15 (IT)								59	59	55	36	40	41	4	4	4	4	4	4	4	4
ETH2 (ET)									75	55	39	43	51	5	5	6	5	5	6	5	5
ETH4 (ET)										54	38	42	51	4	4	5	4	4	5	4	4
SFBC014 (GB)											35	38	52	4	4	5	4	4	5	5	4
MAD (FR)												32	33	5	5	6	5	5	4	4	4
VC1 (CA)													32	5	4	4	5	4	5	5	5
AUL (FR)														4	4	5	4	4	4	3	4
MG1 (ZA)															55	45	46	47	8	8	8
RC5.1 (FR)																44	44	46	8	8	8
SA4c (ZA)																	49	50	7	7	7
TF5a1 (ES)																		50	7	7	7
SA1d (ZA)																			7	7	7
CON2A (CH)																				39	55
APEP (CH)																					34

Table 4.3: The percent of ECVA polymorphic sites shared between each of the 22 isolates. Greater overlap (≥30%) highlighted in red. (**A**) The overlap of homozygous SNPs varied between 3% and 97% (**B**) The overlap of heterozygous positions varied between 3% and 75%.

experiment.

Lineage	Culture reference	All Positions	Homozygous SNPs	Homozygous INS	Homozygous DEL	Bi-allelic Heterozygous Positions	Bi-allelic Heterozygous INS	Bi-allelic Heterozygous DEL	Bi-allelic positions with tri-allelic probabilities Positions	Bi-allelic positions with tri-allelic probabilities INS	Bi-allelic positions with tri-allelic probabilities DEL	Tri-allelic positions & indels
GPL	AP15 (IT)	87.49	90.67	98.63	99.42	93.38	93.03	94.98	73.24	91.16	86.91	93.88
GPL	AUL (FR)	85.27	94.97	98.05	99.21	91.36	93.36	94.06	71.32	87.69	81.37	95.06
GPL	BEW2 (CH)	86.94	87.18	98.95	99.19	91.41	94.5	90.23	85.05	85.14	90.4	78.13
GPL	BLI1 (CH)	85.2	84.4	99.16	98.78	92.58	94.87	96.34	89.16	93.25	89.33	83.55
GPL	ETH2 (ET)	87.37	93.51	98.47	99.1	90.95	95.38	91.64	83.78	88.58	88.37	84.23
GPL	ETH4 (ET)	84.77	87.72	98.96	98.91	92.27	93.72	93.39	73.04	91.49	86.52	81.01
GPL	JEL423 (PA)	85.9	89.07	99.27	99.29	92.52	94.64	95.31	88.48	91.68	91.54	84.38
GPL	MAD (FR)	85.58	93.82	98.55	99.06	82.78	92.65	87.35	74.11	84.95	84.98	80.68
GPL	MG4 (ZA)	86.34	88.11	99.21	99.29	93.97	96.12	95.77	86.86	91.22	91	87.42
GPL	MODS27 (IT)	85.11	89.88	99.45	99.29	91.61	93.45	95.31	87.84	92.83	91.06	99.35
GPL	MODS28 (IT)	84.38	89.71	99.27	99.32	94.28	96.12	96.71	82.56	92.15	88.89	87.02
GPL	SFBC014 (GB)	85.38	93.55	98.97	99.24	92.65	94.42	93.73	84.99	92.1	90.04	69.57
GPL	SP10 (IT)	86.55	93.73	98.18	99.28	93.1	93.81	95.67	81.19	89	87.21	95.92
GPL	VC1 (CA)	82.84	89.95	98.51	98.87	80.03	85.66	93.19	77.69	86.31	84.2	95.24
CH	ACON (CH)	86.06	91.14	94.64	94.61	70.57	83.59	78.23	83.48	78.08	76.57	79.99
CH	APEP (CH)	87.44	90.58	93.72	94.35	70.54	76.20	73.44	84.61	72.82	74.56	73.38
CH	CON2A (CH)	85.82	90.92	95.06	95.41	74.61	84.97	77.71	79.4	77.6	75.28	88.49
CAPE	MG1 (ZA)	79.86	88.55	81.39	87.94	60.97	78.9	73.67	70.37	78.13	74.08	60.23
CAPE	RC5.1 (FR)	79.44	88.04	81.31	87.29	59.52	73.11	77.19	65.29	71.14	75.36	56.29
CAPE	SA1d (ZA)	86.24	91.75	95.12	93.6	70.64	78.34	81.54	55.64	70.66	67.97	62.6
CAPE	SA4c (ZA)	85.45	91.42	92.9	93.86	73.79	78.31	83.49	59.13	75.13	73.65	67.99
CAPE	TF5a1 (ES)	83.41	90.08	95.05	91.27	66.49	62.93	80.44	71.09	69.05	71.74	58.65

Table 4.4: The percent of uniquely mapped reads over each type of category of loci. Bi-allelic heterozygous positions had a reduced percent of uniquely mapped reads in the 2 divergent lineages of *Bd*, which may result from structural variants.

4.4.2 Chromosomal Copy Number Variation (CCNV)

Comparing the depth of read coverage over each chromosome using non-overlapping sliding windows revealed CCNV present in isolates belonging to all three lineages of *Bd* and affecting nine of the largest fifteen supercontigs (Fig. 4.5). *t*-tests on the mean depths across windows compared with those in the largest supercontig confirmed a significant increase in read-depth across 36 supercontigs, and a significant decrease in depth across 25 supercontigs in 18 of the 22 sequenced isolates (Fig. 4.6). To further verify relative ploidy within an isolate and the order of ploidy-changes, we inferred whether individual bases were 'evenly' - or 'oddly'-distributed across Illumina reads within a single genome by binning their frequencies into histograms for each chromosome. The expectation here is that a chromosome with an even ploidy will tend towards a 50:50 distribution across each single SNP, while chromosomes with an odd ploidy will tend towards a 33:66 or 33:33:33 ratio across SNP-calls (Table 4.7). This method identified even- or odd-ploidies for 92% of the chromosomes tested with >95%

bootstrap support (Table 4.5).

Culture reference	Supercontig number														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
AP15 (IT)	100	82	100	100	100	100	100	100	100	100	100	100	100	100	100
AUL (FR)	100	100	0	100	100	100	83	100	100	100	100	99	100	100	100
BEW2 (CH)	100	0	0	100	100	100	100	100	100	100	100	0	100	100	100
BLI1 (CH)	100	98	0	100	100	100	100	22	100	100	100	2	100	100	100
ETH2 (ET)	100	100	100	100	100	100	100	88	100	100	100	100	100	100	100
ETH4 (ET)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
JEL423 (PA)	100	0	0	100	100	100	100	100	0	100	0	0	100	100	100
MAD (FR)	100	0	8	50	100	100	100	100	0	0	100	100	100	100	100
MG4 (ZA)	100	0	0	100	100	100	100	100	100	0	100	100	100	100	100
MODS27.3 (IT)	100	8	100	100	100	100	100	100	100	100	2	100	100	100	100
MODS28 (IT)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
SFBC014 (GB)	100	0	0	100	15	100	100	0	100	100	100	100	100	100	100
SP10 (IT)	100	100	100	100	100	100	100	100	0	100	100	100	100	100	100
VC1 (CA)	1	0	100	51	0	10	0	100	100	100	0	50	0	100	37
ACON (CH)	0	0	0	0	0	0	0	0	100	0	100	0	96	0	0
APEP (CH)	0	0	0	0	0	0	0	0	100	100	100	0	3	100	0
CON2A (CH)	0	0	0	100	100	0	0	0	100	100	100	0	88	100	0
MG1 (ZA)	0	0	44	12	0	100	100	12	100	0	99	45	100	38	0
RC5.1 (FR)	0	0	98	2	0	100	100	8	100	0	54	9	97	7	2
SA1d (ZA)	100	100	100	100	100	100	100	77	100	100	0	100	100	100	100
SA4c (ZA)	100	0	100	100	100	100	100	100	100	100	100	100	100	100	100
TF5a1 (ES)	0	100	0	0	0	100	22	0	100	0	100	0	20	100	0

Table 4.5: The two most common allele frequencies over each base of each supercontig were determined by percent of read agreement with the reference base. Using 1000 Bootstrap replicates of these values, we recorded how often 47-53% reads agreeing with an allele predominated over 30-36% or 63-69% reads agreeing with an allele. Supercontigs with >95% of replicates with predominantly bi-alleles have support for an even-ploidy. Equally, supercontigs with <5% of replicates with predominantly bi-alleles have support for an odd-ploidy. Supercontigs with 5% <x <95% bootstrap support do not have support either an even or odd ploidy (and are therefore ambiguous).

Thirteen *Bd*GPL and two *Bd*CAPE isolates had greater numbers of bi-alleles than tri-alleles (corresponding to an even ploidy that most parsimoniously corresponds to diploidy) (Tables 4.6 and 4.5), and six isolates belonging to all three separate lineages that had greater numbers of tri-alleles than bi-alleles (corresponding to an odd ploidy that most parsimoniously corresponds to triploidy). The remaining four isolates (*Bd*GPL JEL423 & MODS27, *Bd*CAPE SA1d & SA4c) had significant p-values showing between 1-3 chromosomes in lower ploidy levels relative to the remaining bi-allelic genome. Over these genomes we observed greater numbers of tri-alleles than bi-alleles and no decrease in heterozygous base-calls (both of which should occur if these chromosomes were haploid). We therefore conclude that these

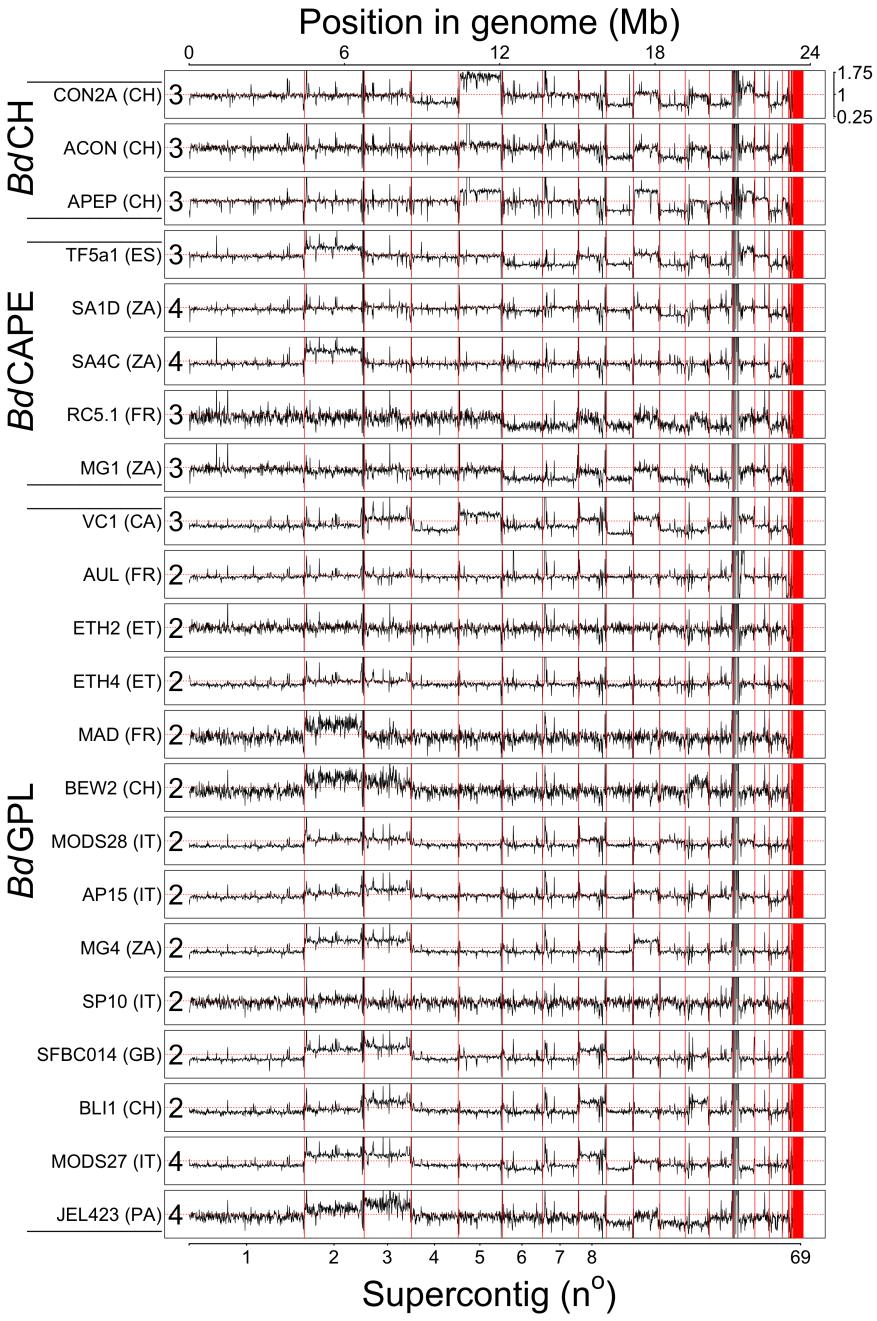


Figure 4.5: Read depth across 22 genomes was normalised by total alignment depth and plotted against location in the genome using a 10Kb long non-overlapping sliding window (red lines signify borders of supercontigs). Base ploidy levels (shown at the beginning of each isolates genome) were determined using allele frequencies for supercontig 1 and shown at the start of each plot. Intra-chromosome read depth is largely consistent amongst the isolates, except over supercontig 14 due to a long stretch of rDNA. Shifts in read-depth between chromosomes demonstrate variation in chromosome copy number.

Lineage	Culture reference	Mean Depth Chromosome 1	Normalised Mean Depth Chromosome 1	Standard Deviation Normalised Depth Chromosome 1	T test p-values for average read depths over chromosomes compared with chromosome 1 — Chromosome number														
					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GPL	AP15 (IT)	143.537	0.870	0.065	1.00E+00	4.13E-11	< 2.2e-16	5.79E-08	5.57E-05	2.29E-02	2.57E-03	1.25E-02	1.67E-02	< 2.2e-16	1.86E-01	5.62E-04	5.65E-02	4.31E-04	2.25E-01
GPL	AUL (FR)	154.023	0.878	0.068	1.00E+00	2.78E-03	4.34E-04	2.09E-01	9.47E-01	1.48E-01	3.20E-02	2.08E-01	3.44E-01	4.57E-01	9.43E-01	1.05E-01	7.60E-02	1.66E-05	2.04E-01
GPL	BEW2 (CH)	109.512	0.829	0.132	1.00E+00	< 2.2e-16	< 2.2e-16	4.80E-01	1.14E-01	8.21E-01	8.78E-01	6.21E-01	6.80E-01	7.99E-01	2.69E-01	3.90E-10	3.19E-01	1.93E-04	4.22E-01
GPL	BLI1 (CH)	40.054	0.812	0.070	1.00E+00	9.47E-06	< 2.2e-16	< 2.2e-16	2.45E-01	9.78E-02	4.89E-02	< 2.2e-16	6.29E-01	5.26E-02	9.33E-02	< 2.2e-16	7.32E-02	7.15E-05	7.00E-05
GPL	ETH2 (ET)	51.858	0.884	0.110	1.00E+00	3.19E-02	8.88E-01	1.16E-03	6.37E-03	1.84E-01	6.04E-01	2.71E-01	1.05E-01	4.68E-02	8.27E-04	6.67E-07	3.90E-01	9.67E-03	3.63E-01
GPL	ETH4 (ET)	129.191	0.845	0.069	1.00E+00	2.79E-08	4.85E-07	5.79E-01	2.35E-01	7.74E-02	3.51E-03	5.42E-02	4.26E-02	1.18E-01	1.78E-01	5.81E-01	1.25E-02	7.92E-05	3.92E-02
GPL	JEL423 (PA)	40.804	0.842	0.109	1.00E+00	4.90E-16	< 2.2e-16	2.43E-01	9.92E-01	7.95E-01	4.00E-01	9.02E-01	< 2.2e-16	3.03E-01	< 2.2e-16	< 2.2e-16	5.18E-01	2.11E-04	8.69E-01
GPL	MAD (FR)	90.746	0.823	0.137	1.00E+00	< 2.2e-16	2.52E-01	9.42E-01	6.15E-01	4.29E-01	8.59E-01	5.86E-01	4.09E-01	7.03E-01	5.61E-01	3.09E-04	3.56E-01	1.73E-04	6.71E-01
GPL	MG4 (ZA)	99.872	0.817	0.058	1.00E+00	< 2.2e-16	< 2.2e-16	3.36E-05	3.72E-01	8.04E-02	1.72E-02	9.88E-02	9.11E-01	< 2.2e-16	2.09E-01	8.11E-01	7.45E-02	1.80E-04	8.26E-02
GPL	MODS27.3 (IT)	40.103	0.816	0.062	1.00E+00	< 2.2e-16	< 2.2e-16	3.03E-02	5.88E-01	1.14E-02	8.81E-02	6.51E-16	< 2.2e-16	4.03E-15	1.59E-01	9.85E-01	7.49E-02	5.65E-04	2.73E-01
GPL	MODS28 (IT)	118.114	0.796	0.057	1.00E+00	< 2.2e-16	< 2.2e-16	1.15E-14	1.34E-04	1.78E-02	2.45E-03	7.00E-08	1.43E-01	1.51E-02	< 2.2e-16	1.00E-01	2.17E-02	7.87E-05	1.45E-02
GPL	SFBC014 (GB)	84.316	0.788	0.067	1.00E+00	< 2.2e-16	< 2.2e-16	2.37E-03	2.92E-11	2.97E-02	1.70E-02	3.81E-16	7.54E-01	2.28E-02	7.49E-01	8.55E-05	3.98E-02	7.25E-05	1.50E-01
GPL	SP10 (IT)	101.836	0.882	0.112	1.00E+00	5.88E-04	6.61E-01	6.91E-01	9.55E-01	9.70E-01	2.68E-01	9.88E-01	5.65E-01	6.66E-01	2.38E-01	2.63E-04	3.72E-01	4.17E-04	6.51E-01
GPL	VC1 (CA)	39.025	0.794	0.074	1.00E+00	3.46E-03	7.93E-12	< 2.2e-16	< 2.2e-16	1.17E-01	9.66E-03	1.59E-05	< 2.2e-16	4.01E-10	< 2.2e-16	1.41E-09	9.32E-02	6.10E-06	8.28E-02
CH	ACON (CH)	132.681	0.917	0.103	1.00E+00	2.29E-02	7.49E-02	6.73E-01	1.94E-07	3.66E-01	2.22E-05	6.53E-01	< 2.2e-16	4.10E-01	< 2.2e-16	1.03E-05	4.63E-04	4.43E-04	2.25E-01
CH	APEP (CH)	90.325	0.894	0.102	1.00E+00	9.75E-02	8.04E-01	9.40E-01	< 2.2e-16	4.68E-01	3.80E-01	8.87E-01	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.67E-07	3.06E-01	8.51E-04	1.38E-03
CH	CON2A (CH)	91.315	0.894	0.089	1.00E+00	2.36E-02	1.14E-02	< 2.2e-16	< 2.2e-16	7.51E-02	1.25E-02	2.97E-01	< 2.2e-16	7.34E-04	< 2.2e-16	1.80E-04	2.70E-03	1.15E-06	9.25E-02
CAPE	MG1 (ZA)	50.405	0.818	0.113	1.00E+00	3.78E-01	8.99E-01	< 2.2e-16	6.78E-01	2.13E-06	6.15E-07	8.32E-01	< 2.2e-16	2.57E-01	< 2.2e-16	6.41E-11	9.77E-01	9.80E-05	5.37E-01
CAPE	RC5.1 (FR)	55.440	0.820	0.161	1.00E+00	1.27E-01	9.17E-01	7.87E-03	5.91E-01	1.32E-05	6.14E-04	5.61E-01	< 2.2e-16	2.86E-01	< 2.2e-16	7.18E-14	9.07E-01	6.92E-05	6.97E-01
CAPE	SA1d (ZA)	122.677	0.897	0.064	1.00E+00	5.06E-04	3.62E-07	3.75E-02	1.45E-05	1.96E-01	1.69E-03	9.89E-02	8.30E-06	1.60E-04	< 2.2e-16	2.50E-02	3.57E-02	1.21E-04	4.20E-04
CAPE	SA4c (ZA)	140.103	0.866	0.093	1.00E+00	< 2.2e-16	1.88E-01	5.54E-01	3.59E-01	2.90E-01	3.58E-01	4.20E-01	3.83E-02	6.71E-02	8.26E-01	8.98E-04	1.06E-01	1.57E-04	1.38E-01
CAPE	TF5a1 (ES)	118.095	0.886	0.076	1.00E+00	4.67E-15	8.82E-02	2.37E-05	3.32E-01	5.77E-07	6.30E-12	3.31E-01	< 2.2e-16	2.31E-03	< 2.2e-16	1.08E-08	9.62E-01	4.59E-06	1.26E-01

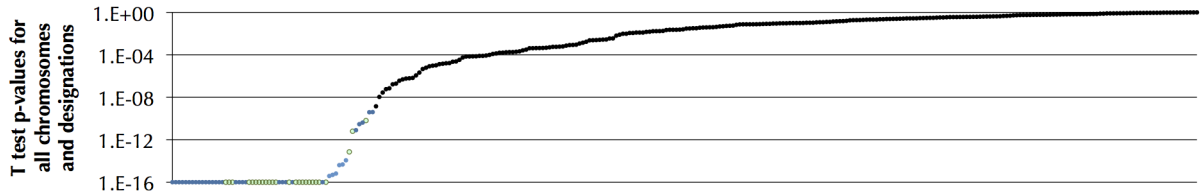


Figure 4.6: *t-tests* for the mean depth of read coverage across each chromosome against chromosome 1 revealed significant p-values demonstrating uneven chromosome copy number. Stringent cut-offs for ploidy differences relative to the largest chromosome (chr. 1) of each isolate were chosen: $p<5^{-10}$. Chromosomes with p-values below this cut-off, with a mean depth that is greater than chromosome 1 are highlighted in blue, while those with a mean depth lower than chr. 1 are shown in green. All 308 chromosomal p-values (excluding chr. 1) are shown in the bottom plot ordered from smallest to greatest.

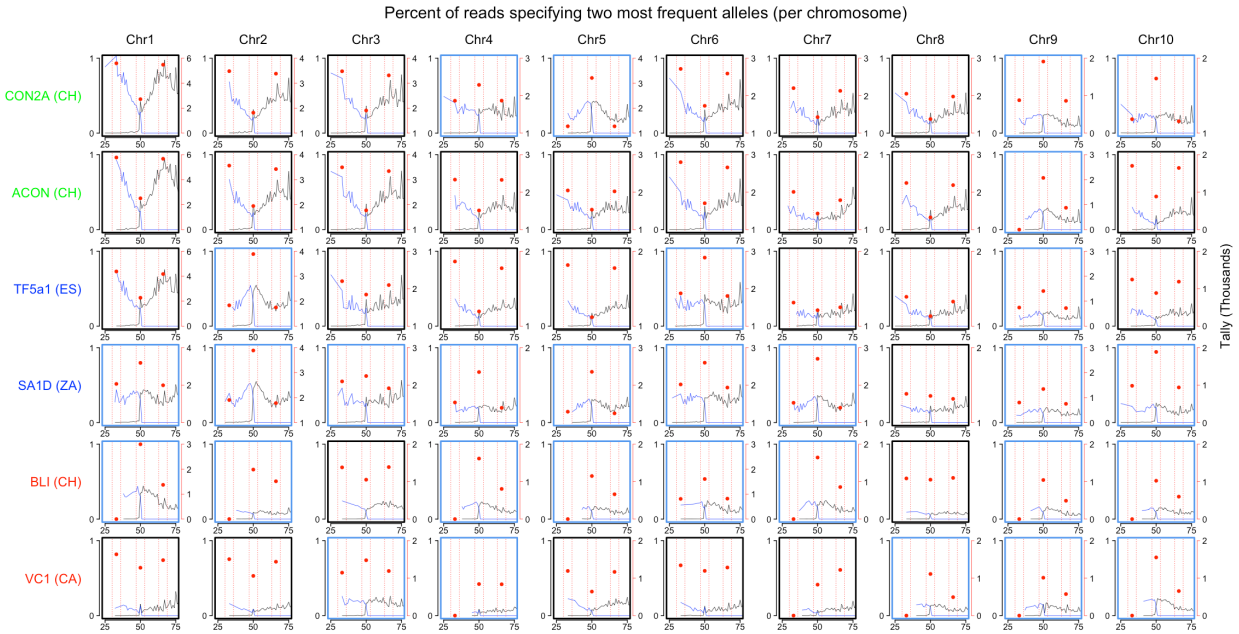


Figure 4.7: The percent of reads specifying the two most frequent alleles per chromosome using 2 representative isolates from each lineage of *Bd*. The most common allele is shown in black and the second most common allele is shown in blue. Bins were used to summarise the expected peaks for odd, even and odd numbers of chromosomes and shown in red (lines show bin value cut-offs and dots show values). Individual chromosomes with a predominantly bi-allelic value are shown with a blue border, and those with a predominant tri-allelic value are shown with a black border.

isolates have tetraploid genomes with the identified losses in read-depth corresponding to chromosomes that have lost a single copy and are now trisomic. Our predictions of CCNV and ploidy based on each of these analyses is summarised in Fig. 4.8.

We were able to take advantage of replicate lines of *Bd*CH, which were passaged for 40 generations with and without exposure to skin antimicrobial peptides collected from the water frog *Pelophylax esculentus*. In these culture lines, the ancestral putatively triploid isolate (*Bd*CH ACON) differentially lost and gained copies of supercontig IV and V respectively when passaged without selection (*Bd*CH CON2A), and gained a copy of supercontig V following treatment with antimicrobial peptide (*Bd*CH APEP), which resulted in a significant reduction in mean growth inhibition (Fig. 4.9). Due to the fact that most of our isolates exhibiting CCNV were sequenced shortly following isolation from nature without sequential passage, we know that CCNV is occurring frequently in both wild and cultured isolates. The rapidity that these mutations are accumulating across our isolates shows that aneuploidies in *Bd* are occurring at rates that will generate genome diversity within the timescale of a single host infection. We can only speculate about the impact that hypervariable CCNV has on the evolution of *Bd*, however it is likely that these alterations in dosage of regulatory and coding elements of the genome will lead to widespread changes in the downstream phenotype, and hence fitness, of the pathogen.

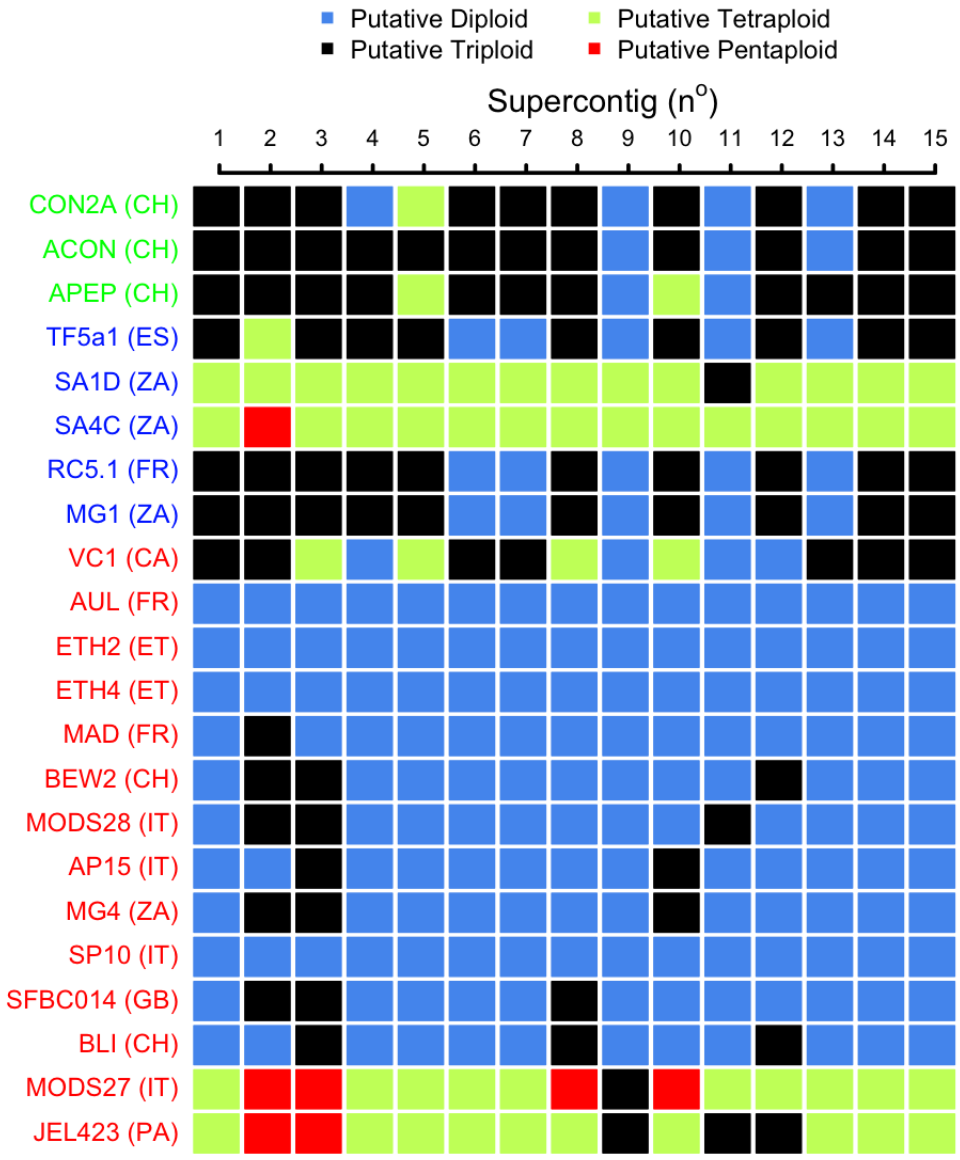


Figure 4.8: CCNV in the *Bd* nuclear genomes was identified using allele-frequencies and mean read depths across each chromosome normalised to the alignment depth for each isolate. Many *Bd*GPL isolates can be seen to include more copies of chromosome 2 and 3, while the 3 *Bd*CH and 3 of the 5 *Bd*CAPE isolates have fewer copies of chromosome 9 and 11. Fewer copies of chromosome 9, 11 and 16 appear to be found in many of the isolates.

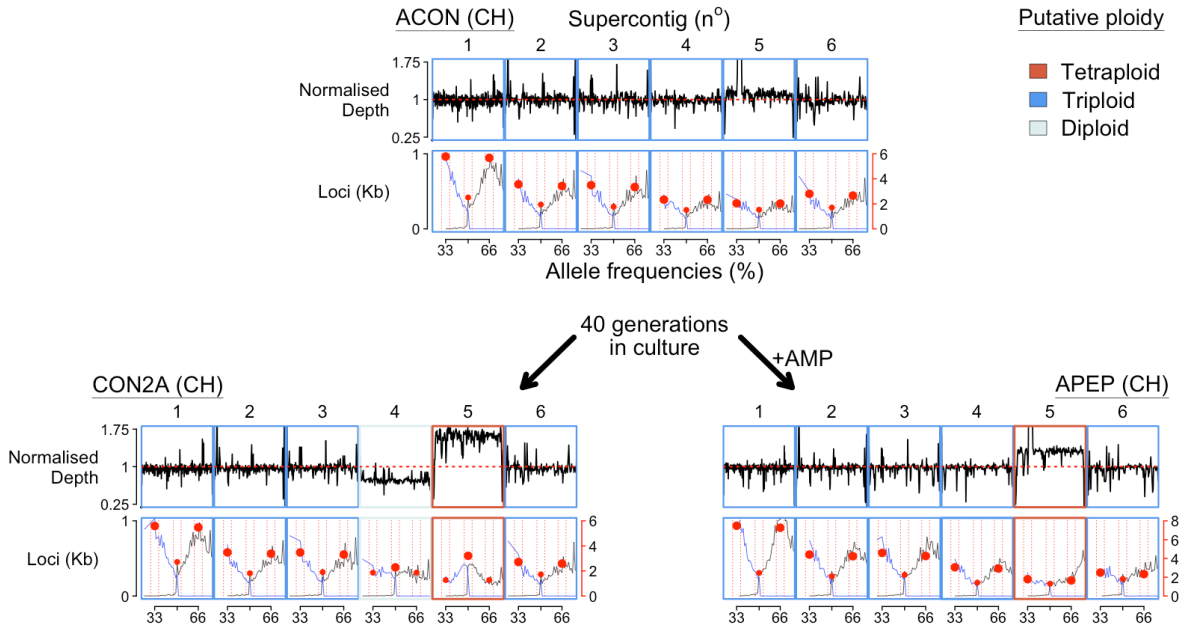


Figure 4.9: Chromosome copy number variation was identified across the three *Bd*CH isolates (ACON and its progenitors CON2A and APEP) following 40 generations in culture with or without the addition of anti-microbial peptides (AMP), respectively. Read depth is normalised to total alignment depth. A tally of all loci (per kilobase) with between 25-75% reads agreeing with the reference nucleotide are shown below, and summarised by the most common allele (black line), the second most common allele (blue line), and bins between 32-34, 49-51 and 65-67% (red circles). ACON is putatively triploid across the largest six supercontigs, whereas CON2A has lost a copy of supercontig IV and gained a copy of supercontigs V. APEP has gained a copy of supercontigs V.

4.4.3 Recombination among and/or within *Bd*

We identified >10X the number of 'covered in all' polymorphic loci from our previous round of SOLiD sequencing [Farrer RA *et al.*, 2011] due at least in part to increased sequence depth and read-length. To check if the same pattern of patchy heterozygosity could be found, we made new plots of 'SNPs minus heterozygous positions' across the genome using a 10Kb non-overlapping window. We identified the same pattern across the genome including an absence of heterozygosity across the majority of chromosome 2 in the *Bd*GPL (Fig. 4.10). The mechanism generating this uneven distribution remains unknown however may reflect a shared ancient meiotic segregation that unifies all isolates of *Bd*GPL, and is unique to *Bd*GPL.

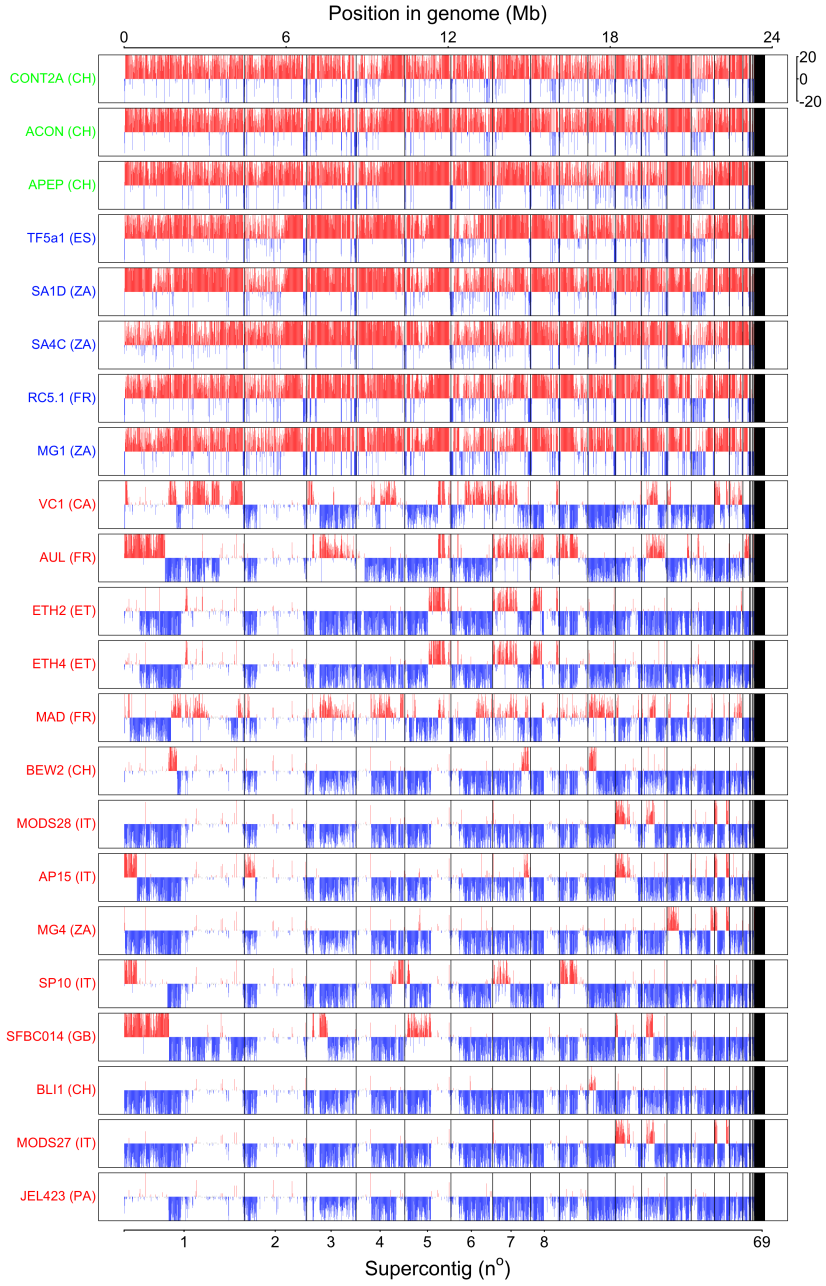
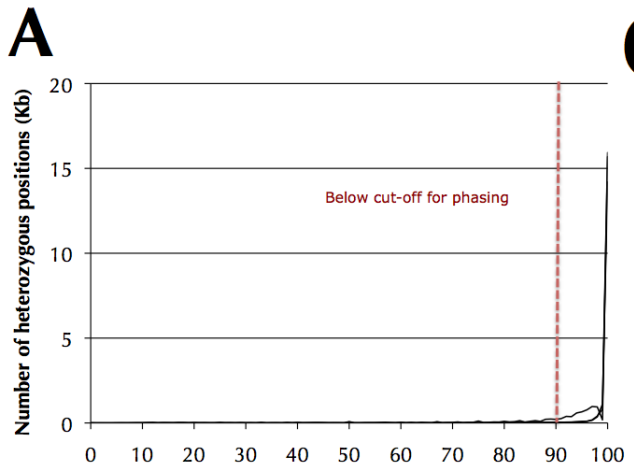


Figure 4.10: Sliding non-overlapping windows of 10Kb across the 22 *Bd* nuclear genomes showing homozygous SNPs minus heterozygous positions. Predominance of homozygous SNPs is shown in red and predominance of heterozygous positions in shown in blue. Windows across *Bd*GPL isolates demonstrate highly uneven distribution of heterozygosity attributed to recombination whereas polymorphisms are more evenly spread across the genomes of *Bd*CAPE and *Bd*CH isolates.

In order to further investigate the presence and frequency of recombination events we determined the phase of bi-allelic heterozygous polymorphisms (Table 4.2). Across the 22 isolates, 71% of the identified heterozygous positions were eligible for pre-filtered phasing whereby each had at least four reads overlapping that and at least one other heterozygous locus and >90% of the reads specified the two nucleotides of the heterozygote (thereby further filtering any potential tri-allelic heterozygous position wrongly called as bi-allelic). Of these 604,895 positions, 487,605 (80.61%) of the overlapping reads inferred a single phase (using the >90% cut-off; Fig. 4.11). In total, 412,883 (84.67%) of the eligible heterozygous positions inferred a single bi-allelic phase without any exceptions (100% reads agreeing with phase). Finally, we extended and increased the number of haplotypes by including all homozygous SNPs within a given phased group, creating new phase groups consisting of consecutive homozygous SNPs, and consecutive "SNP, Bi-allelic HET, SNP". As a proportion of the total homozygous SNPs and bi-allelic heterozygous positions called across each of the 22 isolates, 72.48% were phased using these methods.

By lineage, *Bd*CAPE and *Bd*CH isolates had a lower proportion of resolved bi-allelic phased positions (70 and 74% respectively) compared with 93% in *Bd*GPL isolates, which may have resulted from greater numbers of tri-allelic loci in trisomic chromosomes. A separate possibility is due to the greater divergence of *Bd*CAPE and *Bd*CH isolates to the reference genome, and the subsequent issues with alignments. Although the corresponding tested divergence had no increased FDR for SNP-calls, other issues such as gene-duplications followed by mutation may increase the false positive rate. Although bi-allelic heterozygous positions had a reduced number of uniquely mapped reads, phased positions (comprising both homozygous SNPs and heterozygous positions) had between 86-94% uniquely mapped reads in any given isolate, while the average across the genomes were between 79-87% uniquely mapped reads (Fig. 4.11C), suggesting this issue did not propagate into poorly resolved haplotypes.

To compare linkage between isolates, we performed pairwise comparisons of shared phased positions between each of our isolates. Between *Bd*GPL isolates, >99.8% of the phased in both heterozygous positions were found in the same phase. Within *Bd*CAPE and *Bd*CH isolates, >98.9% and >99% were within the same phase respectively. Even pairwise comparisons for isolates belonging to separate lineages had >92% positions in the same phase, thereby supporting the hypothesis that the population of *Bd* is predominantly clonal and non-recombining (Table 4.6). However, we also identified 4,974 haplotypes demonstrating crossovers (Table 4.7) where all four pairwise combinations of bases were observed. Of



C

Lineage	Culture Reference	Total number of phased positions	Percent uniquely mapped reads
GPL	AP15 (IT)	21,681	93.24
GPL	AUL (FR)	28,066	94.14
GPL	BEW2 (CH)	19,176	91.53
GPL	BLI1 (CH)	21,136	92.25
GPL	ETH2 (ET)	23,411	92.77
GPL	ETH4 (ET)	23,245	91.95
GPL	JEL423 (PA)	15,920	92.38
GPL	MAD (FR)	23,479	90.9
GPL	MG4 (ZA)	20,869	93.65
GPL	MODS27 (IT)	21,468	92.00
GPL	MODS28 (IT)	19,458	93.62
GPL	SFBC014 (GB)	27,150	93.58
GPL	SP10 (IT)	21,363	93.79
GPL	VC1 (CA)	27,683	87.66
CH	ACON (CH)	114,697	89.7
CH	APEP (CH)	134,187	88.89
CH	CON2A (CH)	118,525	89.33
CAPE	MG1 (ZA)	95,682	86.19
CAPE	RC5.1 (FR)	97,329	86.08
CAPE	SA1d (ZA)	113,415	89.84
CAPE	SA4c (ZA)	109,872	89.68
CAPE	TF5a1 (ES)	109,048	88.06

Figure 4.11: Heterozygous positions had their phase determined using overlapping reads. Reads from each isolate are shown as a separate black line on the graphs. Only bi-allelic polymorphisms were compared for phasing. Predominantly, overlapping reads agreed with a single bi-allelic phase. (**A**) All reads over all phased positions. A 90% cut-off was used to filter ambiguous phased positions or those with an excess of mismatches as shown by the red line. (**B**) Positions that agreed 90-100% for a single phase are shown as a percent of all reads. (**C**) the total number of phased positions (including phased SNPs) and the percent of uniquely mapped reads over those positions.

these, 2,007 occurred at unique positions/loci in the genome. Every pair of isolates that we compared (except between *Bd*GPL isolates MAD (FR) and AUL (FR)) showed at least one haplotype that included an inferred crossover (Table 4.7), even when comparing isolates that likely share a very recent common ancestor. For instance, we found that two isolates (MODS27 and MODS28) which were recovered from *Discoglossus sardus* at a single site in Sardinia on a single collection trip and are closely related (Figs. 4.4 and 4.3) had accumulated three crossovers. This shows that recombinant genotypes can accumulate even within highly-related free-living populations of *Bd*, a feature of this chytrid's population genetics that was first remarked upon by Morgan *et al.* in populations of Sierra Nevadan *Rana muscosa* [Morgan JAT *et al.*, 2007].

The greatest proportion of phased positions demonstrating crossovers were found between lineages, likely demonstrating an accumulation (multiple rounds of recombination) following longer divergence times than within lineages (Table 4.7). For example, as many as 7.3% of the phased positions revealed crossovers between isolates *Bd*GPL AP15 (IT) and *Bd*CH ACON (CH). By lineage, crossovers were most prevalent amongst *Bd*CAPE and *Bd*CH isolates (between 0.6 and 1.1% of phased positions) compared with 0-0.2% in *Bd*GPL. This was a surprising finding given that the three *Bd*CH isolates were separated by only 40 passages in the lab, derived from a single isolate that had been recently isolated in 2007, and given our previous results demonstrating unique hallmarks of recombination only in the *Bd*GPL. This suggests two hypotheses: Either *in vitro* passage under selective conditions promotes rapid recombination, or our isolate of *Bd*CH is descended from a population of *Bd* that is more recombinogenic the *Bd*GPL.

Because a greater abundance of crossovers were also found between *Bd*CAPE isolates, then perhaps *Bd*CAPE and *Bd*CH are able to recombine within their lineage (either by selfing or out-crossing) and *Bd*GPL (possibly following an ancestral meiotic recombination event) has had a principally clonal expansion with reduced opportunities for meiosis. Alternatively, *Bd*GPL has been undergoing non-meiotic recombination such as gene-conversion but not selfing or out-crossing (thereby showing patchy heterozygosity, and a reduced level of crossovers). However, it is uncertain how the rate of recombination within a given lineage may have changed since its divergence with other lineages. A small proportion of the identified crossovers could have arisen from two separate mutations *de novo* at that locus. However, the consistency found between isolates in a given lineage suggests this is not having a large effect, if any. Finally, as a separate speculative comment, it is worth noting the association between higher levels of recombination within *Bd*CAPE and *Bd*CH and their

Phased positions Kb: Isolate	BLI1 (CH)	MODS27 (IT)	BEW2 (CH)	MODS28 (IT)	MG4 (ZA)	SP10 (IT)	AP15 (IT)	ETH2 (ET)	ETH4 (ET)	SFBC014 (GB)	MAD (FR)	VC1 (CA)	AUL (FR)	MG1 (ZA)	RC5.1 (FR)	SA4c (ZA)	TF5a1 (ES)	SA1d (ZA)	CON2A (CH)	APEP (CH)	ACON (CH)
JEL423 (PA)	7.7	7.2	6.6	6.2	6.3	6.6	6.0	5.9	5.8	6.8	3.4	4.2	4.4	0.4	0.4	0.4	0.4	0.5	0.6	0.6	0.5
BLI1 (CH)		10.0	8.6	8.9	9.6	8.0	8.3	8.3	8.4	8.8	4.7	5.1	5.7	0.5	0.5	0.6	0.6	0.5	0.9	0.8	0.6
MODS27 (IT)			7.4	9.2	8.7	7.4	8.4	7.5	8.1	8.0	4.1	5.3	5.0	0.5	0.5	0.6	0.5	0.6	0.8	0.8	0.6
BEW2 (CH)				7.0	7.5	7.1	6.8	6.5	6.9	7.6	3.8	4.4	4.9	0.3	0.3	0.5	0.4	0.4	0.7	0.6	0.4
MODS28 (IT)					8.6	6.7	8.3	7.1	7.5	7.8	3.8	4.2	4.9	0.4	0.4	0.5	0.4	0.5	0.7	0.6	0.5
MG4 (ZA)						7.1	8.2	7.3	7.8	7.9	4.4	4.3	4.7	0.5	0.4	0.6	0.5	0.5	0.8	0.7	0.6
SP10 (IT)							6.7	6.9	7.2	7.4	4.2	4.5	5.1	0.4	0.4	0.5	0.5	0.5	0.7	0.6	0.6
AP15 (IT)								6.8	7.4	7.5	3.5	4.4	4.6	0.3	0.4	0.5	0.4	0.5	0.7	0.7	0.6
ETH2 (ET)									9.2	7.0	3.9	4.6	5.5	0.4	0.5	0.5	0.5	0.5	0.8	0.8	0.7
ETH4 (ET)										7.7	3.9	4.8	6.0	0.4	0.4	0.5	0.5	0.5	0.8	0.7	0.6
SFBC014 (GB)											3.9	4.4	6.5	0.6	0.5	0.6	0.6	0.5	0.8	0.8	0.6
MAD (FR)												2.9	3.1	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.6
VC1 (CA)													2.9	0.7	0.5	0.4	0.5	0.5	0.7	0.8	0.9
AUL (FR)														0.3	0.5	0.5	0.6	0.5	0.7	0.5	0.5
MG1 (ZA)															11.2	10.1	10.3	11.1	1.8	2.2	2.5
RC5.1 (FR)																9.5	10.0	9.9	1.8	2.2	2.2
SA4c (ZA)																	12.3	14.5	2.1	2.2	1.9
TF5a1 (ES)																		12.3	2.2	2.6	2.4
SA1d (ZA)																			2.0	2.2	2.1
CON2A (CH)																				18.7	20.1
APEP (CH)																					16.2
Same phase %:																					
JEL423 (PA)	100	100	100	100	100	100	100	100	100	100	100	100	100	97	98	97	96	97	96	95	94
BLI1 (CH)		100	100	100	100	100	100	100	100	100	100	100	100	97	97	95	94	96	95	95	94
MODS27 (IT)			100	100	100	100	100	100	100	100	100	100	100	97	98	96	95	97	95	96	96
BEW2 (CH)				100	100	100	100	100	100	100	100	100	100	96	98	95	94	96	95	95	95
MODS28 (IT)					100	100	100	100	100	100	100	100	100	97	97	96	94	96	95	94	95
MG4 (ZA)						100	100	100	100	100	100	100	100	97	96	96	95	96	94	94	94
SP10 (IT)							100	100	100	100	100	100	100	97	97	95	96	96	96	96	96
AP15 (IT)								100	100	100	100	100	100	97	97	95	96	96	95	94	93
ETH2 (ET)									100	100	100	100	100	95	96	95	95	97	95	96	96
ETH4 (ET)										100	100	100	100	95	96	95	94	95	95	96	95
SFBC014 (GB)											100	100	100	97	98	96	96	96	95	95	95
MAD (FR)												100	100	99	96	97	97	97	96	97	98
VC1 (CA)													100	99	97	96	97	97	97	97	98
AUL (FR)														97	97	95	97	97	95	96	97
MG1 (ZA)															99	99	99	99	98	98	99
RC5.1 (FR)																99	99	99	99	98	99
SA4c (ZA)																	99	99	98	98	98
TF5a1 (ES)																		99	99	98	99
SA1d (ZA)																			98	98	99
CON2A (CH)																				99	99
APEP (CH)																					99

Table 4.6: Pairwise comparisons for shared phased heterozygous positions. (top) Total numbers of matching phased heterozygous positions in same phase (kilobase) (below) Percent of matching phase positions from the total number of shared phased positions.

Crossovers:

Isolate	BLI1 (CH)	MODS27 (IT)	BEW2 (CH)	MODS28 (IT)	MG4 (ZA)	SP10 (IT)	AP15 (IT)	ETH2 (ET)	ETH4 (ET)	SFBC014 (GB)	MAD (FR)	VC1 (CA)	AUL (FR)	MG1 (ZA)	RC5.1 (FR)	SA4c (ZA)	TF5a1 (ES)	SA1d (ZA)	CON2A (CH)	APEP (CH)	ACON (CH)
JEL423 (PA)	6	7	9	5	7	5	9	8	6	8	5	6	4	14	10	12	18	12	26	28	28
BLI1 (CH)		3	3	2	7	14	8	5	7	4	3	4	3	16	13	30	34	22	42	38	36
MODS27 (IT)			4	3	8	7	3	6	2	3	5	9	1	14	12	27	24	19	43	35	26
BEW2 (CH)				6	5	5	6	6	2	6	1	1	3	15	8	21	28	16	33	30	21
MODS28 (IT)					11	5	9	8	3	8	2	4	2	10	13	19	23	17	38	40	23
MG4 (ZA)						6	5	7	7	8	2	3	4	15	16	23	26	23	45	40	31
SP10 (IT)							6	10	6	10	6	6	7	11	13	24	22	17	27	28	26
AP15 (IT)								6	5	5	5	3	3	11	12	23	16	17	41	43	42
ETH2 (ET)									12	6	6	2	4	22	21	26	25	16	42	31	28
ETH4 (ET)										6	4	4	6	19	17	29	29	25	36	28	29
SFBC014 (GB)											3	7	5	16	9	28	24	20	39	41	32
MAD (FR)												1	0	7	19	16	16	13	21	15	12
VC1 (CA)													1	8	14	17	16	14	24	27	19
AUL (FR)														11	16	26	19	18	31	22	15
MG1 (ZA)															95	88	99	63	29	34	27
RC5.1 (FR)																90	114	94	20	34	14
SA4c (ZA)																	125	118	33	47	32
TF5a1 (ES)																		128	32	43	33
SA1d (ZA)																			36	39	30
CON2A (CH)																				114	202
APEP (CH)																					108

%:

JEL423 (PA)	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	3.4	2.4	2.9	4.2	2.6	4.0	4.7	5.7
BLI1 (CH)		0.0	0.0	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	3.0	2.8	4.6	5.6	4.0	4.7	5.0	5.6
MODS27 (IT)			0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.2	0.0	2.7	2.3	4.5	4.9	3.3	5.2	4.4	4.5
BEW2 (CH)				0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.1	4.5	2.4	4.6	6.5	3.7	5.0	5.0	4.7
MODS28 (IT)					0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.0	2.7	3.4	3.8	6.2	3.6	5.3	6.3	4.6
MG4 (ZA)						0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	3.2	4.2	3.9	4.9	4.3	5.8	5.7	5.6
SP10 (IT)							0.1	0.1	0.1	0.1	0.1	0.1	0.1	3.0	3.2	5.3	4.2	3.7	3.9	4.4	4.5
AP15 (IT)								0.1	0.1	0.1	0.1	0.1	0.1	3.3	3.4	4.7	4.1	3.6	5.5	6.1	7.3
ETH2 (ET)									0.1	0.1	0.2	0.0	0.1	5.5	4.5	4.9	5.2	3.0	5.0	4.0	4.0
ETH4 (ET)										0.1	0.1	0.1	0.1	4.9	4.0	5.3	5.8	5.0	4.7	4.0	4.9
SFBC014 (GB)											0.1	0.2	0.1	2.8	2.0	4.4	4.3	3.7	4.6	5.1	4.9
MAD (FR)												0.0	0.0	1.5	3.6	3.2	3.1	2.9	4.0	2.8	1.9
VC1 (CA)													0.0	1.2	3.0	4.1	3.4	2.8	3.3	3.4	2.2
AUL (FR)														3.2	3.2	5.4	3.4	3.5	4.7	4.1	2.9
MG1 (ZA)															0.9	0.9	1.0	0.6	1.6	1.6	1.1
RC5.1 (FR)																1.0	1.1	1.0	1.1	1.6	0.7
SA4c (ZA)																	1.0	0.8	1.6	2.2	1.7
TF5a1 (ES)																		1.0	1.4	1.7	1.4
SA1d (ZA)																			1.8	1.8	1.4
CON2A (CH)																				0.6	1.0
APEP (CH)																					0.7

Table 4.7: Phased heterozygous positions demonstrating crossovers were identified between every isolate. (top) Total numbers of crossovers identified. (below) Percent of crossovers from the total number of shared phased positions.

hypovirulence compared with *Bd*GPL.

To further study the amount of recombination within lineages and between isolates, we extracted haplotypes that were phased across all of the isolates within a given lineage and contained at least two alleles per loci (ranging in length from 11nt to 33.3Kb: Fig. 4.12). In total we obtained 2,275 haplotypes for *Bd*CAPE and 5,215 haplotypes for *Bd*CH were extracted ranging in length from 10nt up to 21.744Kb, with similar lengths found in both lineages (Fig. 4.12). Because only 35 haplotypes were retained for the entire panel of *Bd*GPL isolates, we also extracted haplotypes from two *Bd*GPL subsets consisting of 3 and 5 isolates, thus allowing higher numbers of crossovers to be retained. Subset 1 consisted of VC1, AP15 and JEL423 (919 haplotypes), and the 2^{nd} subset including subset 1 as well as ETH4 and MODS27 (438 haplotypes), which were chosen based on their representation of the total diversity in the UPGMA tree (Figs. 4.4 and 4.3). From each of these sets of haplotypes, we calculated the index of association (I_A), the multilocus measure of linkage disequilibrium (the standardised index of association $\overline{r}_d$, [Agapow P *et al.*, 2001]) and applied Hudson's four-gamete test [Hudson R *et al.*, 1985] in order to quantify the amount of recombination amongst isolates within each lineage (Table 4.8).

Across the *Bd*GPL groups, >30% of phased positions were in significant disequilibrium compared with 16% and 11% for *Bd*CH and *Bd*CAPE respectively. $\overline{r}_d$ appeared to be robust against sample size differences, and gave values from *Bd*GPL values of 0.79-0.82 compared against 0.58 and 0.61 for *Bd*CH and *Bd*CAPE. Finally, a smaller proportion of *Bd*GPL subset haplotypes failed the four-gamete test compared with *Bd*CAPE or *Bd*CH isolates. Each of these findings shows that recombination is causing diversity within each of the lineages. However, the emergent *Bd*GPL is far more clonal than either of the other two lineages.

To check for genome-positional enrichment, the locations of each intra-lineage heterozygote, the percent of which were phased, and those that demonstrated a crossover were plotted across the genome using a non-overlapping 10Kb windows (Fig. 4.13). The location of each haplotype, its corresponding $\overline{r}_d$ value, those that failed the 4 gamete test, as well as the mean $\overline{r}_d$ were also plotted. Both phased positions and crossovers were identified in every major chromosome in all three lineages of *Bd*, suggesting recombination is not confined to small or large chromosomes, or the ends of any given chromosome. The same is seen with $\overline{r}_d$ values. Finally, crossovers were predominantly identified in intergenic regions (143 Mb^{-1} compared with 57 Mb^{-1} for coding regions and 65 Mb^{-1} for introns) (Fig. 4.13).

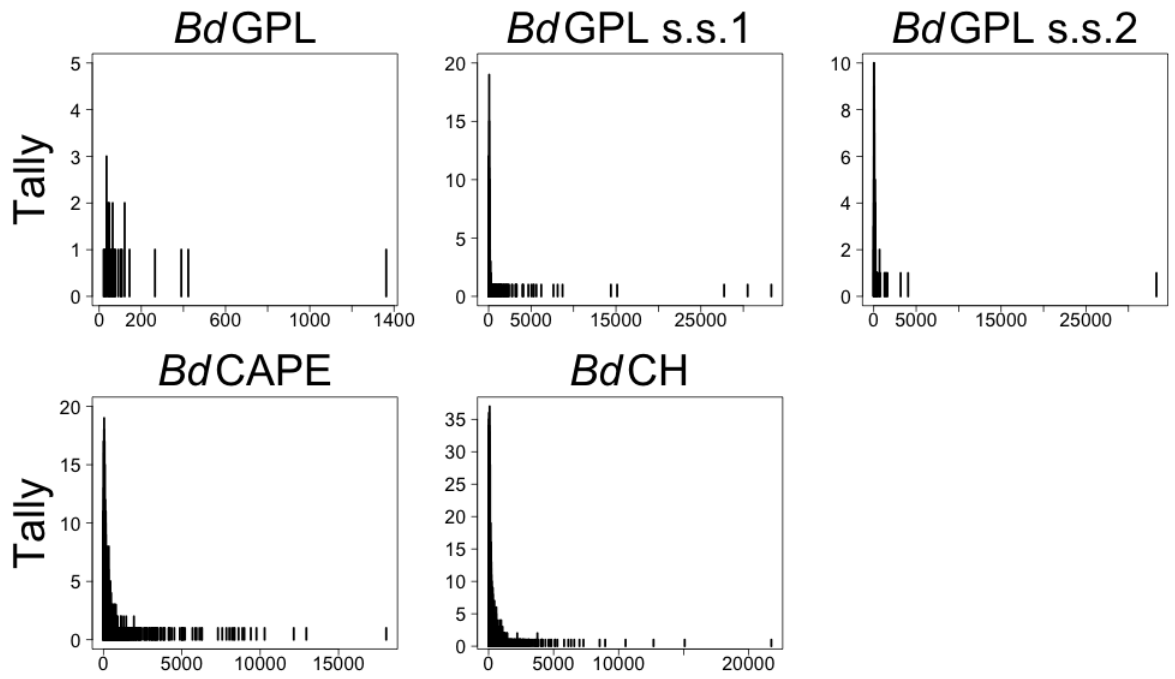


Figure 4.12: Lengths of haplotypes (in nucleotides) that included at least two alleles per loci in every isolate of a given group, and were therefore suitable for population genetic analysis. *Bd*GPL subset (s.s.) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27.

Lineage	Isolates	Haplotypes	Length (nt)	Loci	Significant Disequilibrium (%)	Mean $\overline{r}_d$	Fail 4-gamete test (%)
*Bd*GPL	14	35	4,409	95	68.57	0.82	7 (20%)
*Bd*GPL s.s 1	3	919	341,325	2,822	31.12	0.79	61 (6.64%)
*Bd*GPL s.s 2	5	438	83,414	1,232	41.32	0.82	36 (8.22%)
*Bd*CAPE	5	2,275	952,307	7,212	11.47	0.61	197 (8.66%)
*Bd*CH	3	5,215	1,537,742	16,612	16.36	0.58	655 (12.56%)

Table 4.8: Haplotypes from isolates belonging to each of the separate lineages were tested for disequilibrium using the index of association (I_A), $\overline{r}_d$ and the 4-gamete test. To check differences between lineages were not resulting from different numbers of isolates, 2 subsets were made from *Bd*GPL. Subset 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. Over 30% of the *Bd*GPL haplotypes from any of the subsets were in significant disequilibrium, whilst only 11% of the haplotypes in *Bd*CH and 16% of the haplotypes in *Bd*CAPE were in disequilibrium, suggesting these populations are recombining more than the clonal *Bd*GPL.

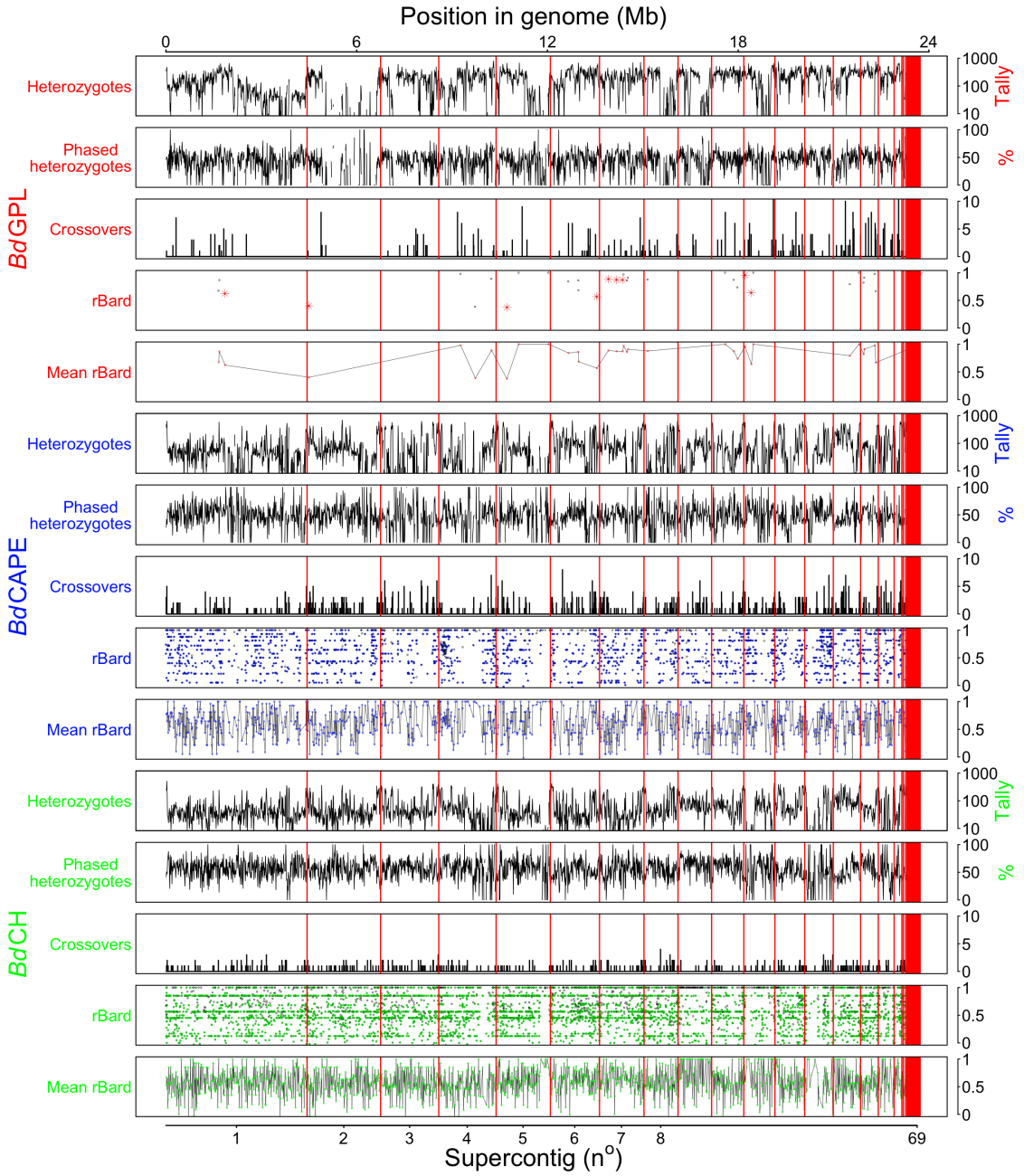


Figure 4.13: Intra-lineage heterozygote's, the percent of heterozygote's that were phased, the number of phased positions that demonstrated a crossover (XO) and the $\overline{r}_d$ were plotted using non-overlapping windows across the genome (length 10Kb). Both phased positions and crossovers were found across each of the chromosomes in each of the lineages of *Bd*, suggesting recombination is not confined to small or large chromosomes, or the ends of any given chromosome. The same is seen with $\overline{r}_d$ values.

We next determined whether certain categories of genes were associated with higher-than-average rates of recombination using t-tests on the total numbers of crossovers and those at unique loci (non-redundant; NR) as shown in Table 4.9 and Figs. 4.14 and 4.15. Surprisingly, all three lineages had significant crossover-enrichment for genes showing homology to the C-terminal of the Crinkler (CRN) family of oomycete effector proteins. In addition, *Bd* (interlineage) and *Bd*CAPE had enrichment for uncharacterised (secreted) genes, both of which might be implicated in *Bd* pathogenicity (Table 4.9, Figs. 4.14 and 4.15).

To check if this was simply a result of greater heterozygosity providing an increased proportion that was phased and therefore our ability to detect recombination, we normalised to the number of phased positions (PP) from which crossovers could be called (NRXO/NRPP). Again we tested for enrichment using t-tests (Table 4.9), now finding that only CRN-like genes remained enriched amongst all *Bd* isolates (inter-lineage; Fig. 4.16). It is unclear and remains to be seen whether these genes are recombining at a greater rate due to an unrelated process, or if they are actively involved in recombination (e.g. a recombination motif) between homologous sequences. Haplotypes that failed the four-gamete test were predominantly from coding-regions, but had no clear pattern of enrichment for any gene category (Table 4.9).

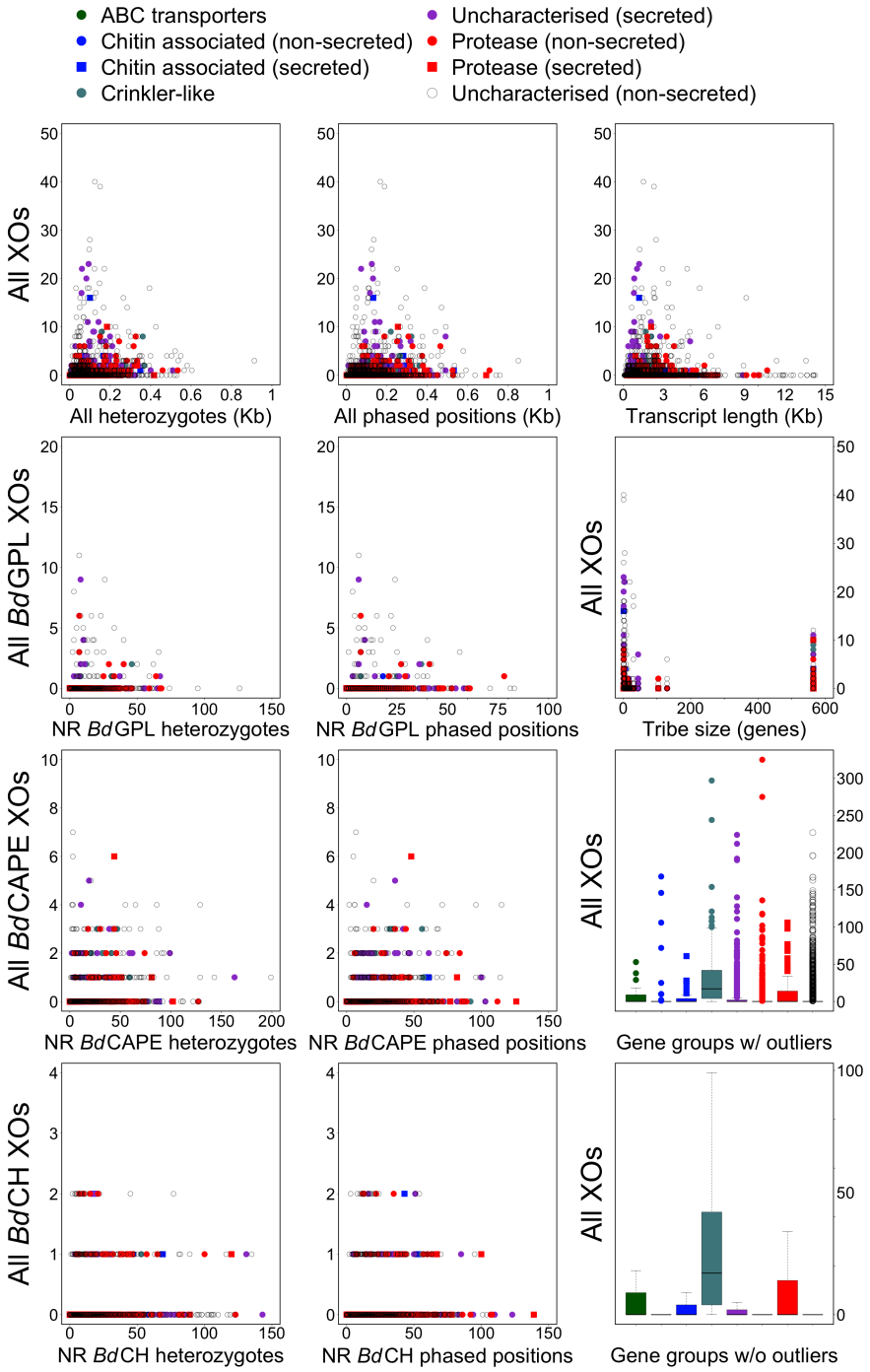


Figure 4.14: The total numbers of crossovers found within genes demonstrated variation between gene families. All crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size.

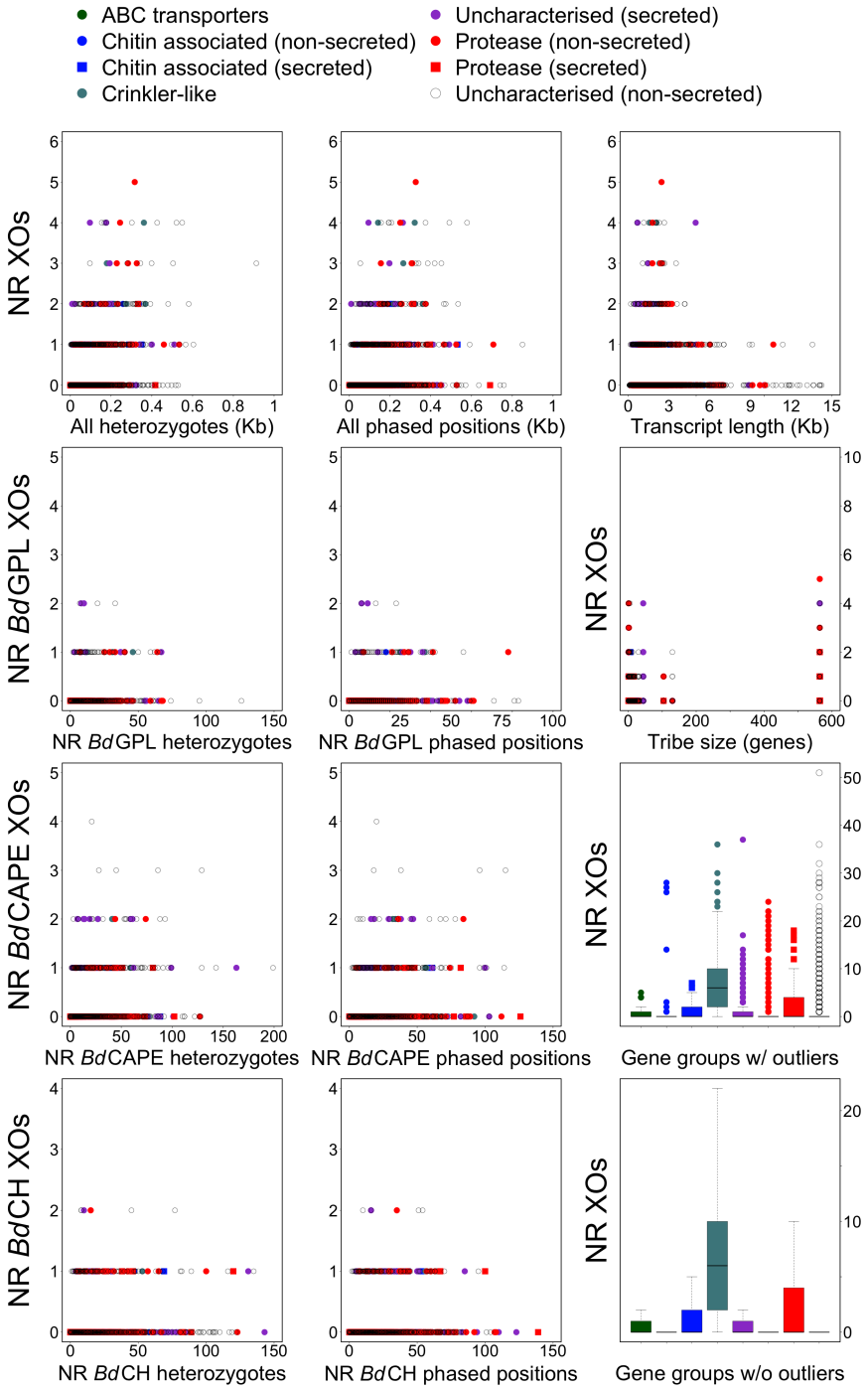


Figure 4.15: Crossovers at unique locations (non-redundant, NR) occurred differentially across gene families. NR crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size.

Gene names	Isolates	NR XO	(NR XO / NRPP)	HgT for NR XO P(X>x)	SIG.	*t*-test for (NR XO / / NRPP) P(X>x)	SIG.
ABC transporters	All	0	NA	NA	-	NA	-
	*Bd*GPL	0	NA	NA	-	NA	-
	*Bd*CAPE	0	NA	NA	-	NA	-
	*Bd*CH	0	NA	NA	-	NA	-
Chitin associated (non-secreted)	All	6	0.01	3.593E-01	-	8.864E-01	-
	*Bd*GPL	1	<0.01	1.133E-01	-	6.803E-01	-
	*Bd*CAPE	0	<0.01	8.878E-01	-	1.000E+00	-
	*Bd*CH	2	<0.01	1.270E-01	-	NA	-
Chitin associated (secreted)	All	10	0.01	3.687E-02	-	7.045E-01	-
	*Bd*GPL	1	<0.01	3.069E-01	-	NA	-
	*Bd*CAPE	1	<0.01	4.137E-01	-	NA	-
	*Bd*CH	2	<0.01	3.239E-02	-	9.974E-01	-
CRN-like	All	49	0.01	3.418E-22	***	1.850E-04	**
	*Bd*GPL	4	<0.01	1.913E-03	*	3.633E-01	-
	*Bd*CAPE	22	0.01	7.801E-13	***	1.097E-02	-
	*Bd*CH	9	<0.01	2.293E-05	***	7.002E-01	-
Uncharacterised (secreted)	All	115	0.01	3.833E-11	***	3.772E-02	-
	*Bd*GPL	11	<0.01	1.773E-02	-	7.469E-01	-
	*Bd*CAPE	48	<0.01	4.865E-07	***	6.130E-01	-
	*Bd*CH	19	<0.01	3.849E-02	-	1.000E+00	-
Protease (non-secreted)	All	75	0.01	2.999E-01	-	7.822E-01	-
	*Bd*GPL	7	<0.01	4.167E-01	-	9.997E-01	-
	*Bd*CAPE	21	<0.01	8.645E-01	-	9.995E-01	-
	*Bd*CH	22	<0.01	2.954E-02	-	9.997E-01	-
Protease (secreted)	All	11	<0.01	2.128E-01	-	9.725E-01	-
	*Bd*GPL	0	NA	NA	-	NA	-
	*Bd*CAPE	5	<0.01	1.349E-01	-	9.928E-01	-
	*Bd*CH	3	<0.01	1.335E-01	-	1.000E+00	-
Uncharacterised (non-secreted)	All	440	0.01	1.000E+00	-	7.690E-01	-
	*Bd*GPL	47	<0.01	9.580E-01	-	1.000E+00	-
	*Bd*CAPE	166	<0.01	1.000E+00	-	1.000E+00	-
	*Bd*CH	93	<0.01	1.000E+00	-	1.000E+00	-

Table 4.9: Genes were tested for enrichment in non-redundant (NR; at unique loci) crossovers (XO) and NR XO / NR phased positions (NRPP), comparing each to the values obtained for all genes using Hypergeometric tests (HgT) and *t*-tests respectively. For *t*-tests, all genes with <2 NRPP (the minimum required for a crossover) were excluded. Although both CRN-like and uncharacterised (secreted) were significantly (SIG.) enriched for XOs at unique loci (non-redundant), only CRN-like (between lineages) were SIG. enriched for XO/NRPP. *p*-values for HgT and *t*-tests are shown as P≥0.01 (-), P<0.01 (*), P<0.001 (**) and P<0.0001 (***). NA=not applicable

Lineage	*Bd*GPL			*Bd*GPL s.s.1			*Bd*GPL s.s.2			*Bd*CAPE			*Bd*CH		
4-gamete test	fail	don't fail	ratio	fail	don't fail	ratio	fail	don't fail	ratio	fail	don't fail	ratio	fail	don't fail	ratio
ABC Transporters	0	0	N/A	0	2	N/A	0	1	N/A	1	3	0.33	1	9	0.11
Chitin associated (non-secreted)	0	0	N/A	1	2	0.50	0	0	N/A	3	10	0.30	7	14	0.50
Chitin associated (secreted)	0	0	N/A	1	3	0.33	0	2	N/A	2	8	0.25	2	16	0.13
Crinkler-like	0	0	N/A	1	5	0.20	0	2	N/A	2	15	0.13	9	39	0.23
Protease (non-secreted)	0	3	N/A	3	50	0.06	3	23	0.13	18	133	0.14	38	237	0.16
Protease (secreted)	0	0	N/A	1	4	0.25	0	1	N/A	1	23	0.04	10	38	0.26
Uncharacterised (non-secreted)	6	16	0.38	47	458	0.10	31	213	0.15	82	1020	0.08	282	1852	0.15
Uncharacterised (secreted)	1	0	N/A	8	64	0.13	6	28	0.21	20	178	0.11	53	295	0.18
Non gene	1	9	0.11	18	320	0.06	9	135	0.07	70	819	0.09	282	2197	0.13

Table 4.10: Haplotypes over coding sequence that failed the four-gamete test were predominantly from coding-regions. Haplotypes overlapping a number of genes were included in the counts for each gene (385 extra counts to total number of haplotypes). After accounting for these extra counts, an additional 1,162 haplotypes were still found to come from coding regions compared with those from intergenic or intron regions. However, no gene group had a clear enrichment for haplotypes that failed the four-gamete test. *Bd*GPL subset (s.s.) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27.

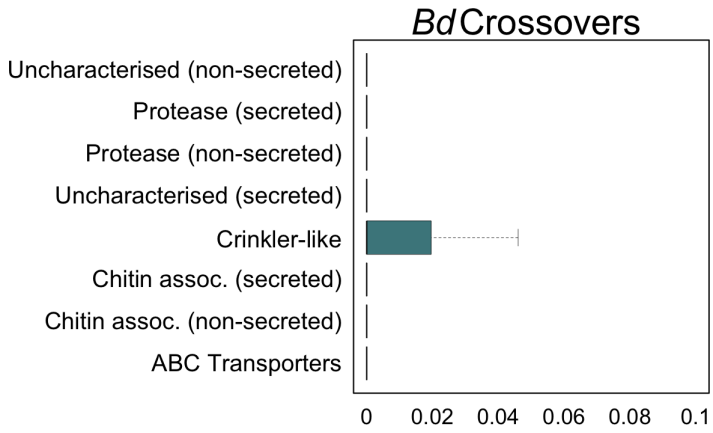


Figure 4.16: Boxplots for eight non-overlapping gene categories comprising every gene were compared for numbers of crossovers per phased positions (PP) within each gene (≥2PP) for all isolates.

To look for evidence of hybridization, we counted the number of variable sites per locus across the haplotypes with an expectation that a hybrid lineage should have a greater number of variable sites at a given locus. Between 70-75% of *Bd*GPL subsets 1 and 2 loci consisted of 2 different bases compared with 75% in *Bd*CAPE and 72% in *Bd*CH (Table 4.11). Additionally, *Bd*GPL subsets had between 20-24% of loci consisting of 3 different bases compared to 24% in *Bd*CAPE and 25% in *Bd*CH. These values suggest that none of the three lineages stand out as an expected hybrid given the expectation for a greater percent of variable sites per locus.

We next calculated θ, Weir's (31) formulation of Wrights fixation index (F_{ST}) for each pairwise lineage across window lengths of 1.4Kb and 10Kb (Fig. 4.17) using all homozygous (SNPS or reference) and bi-allelic heterozygous positions (and excluding all loci with neither in all). We found that all three lineages were highly differentiated from one another across each chromosome, with only minor intra-chromosomal variation, such as the stretch of rDNA located at the start of chromosome 14. Both the F_{ST} and the number of different bases at haplotypes loci suggest that *Bd*GPL, *Bd*CAPE and *Bd*CH are not recombining with one another.

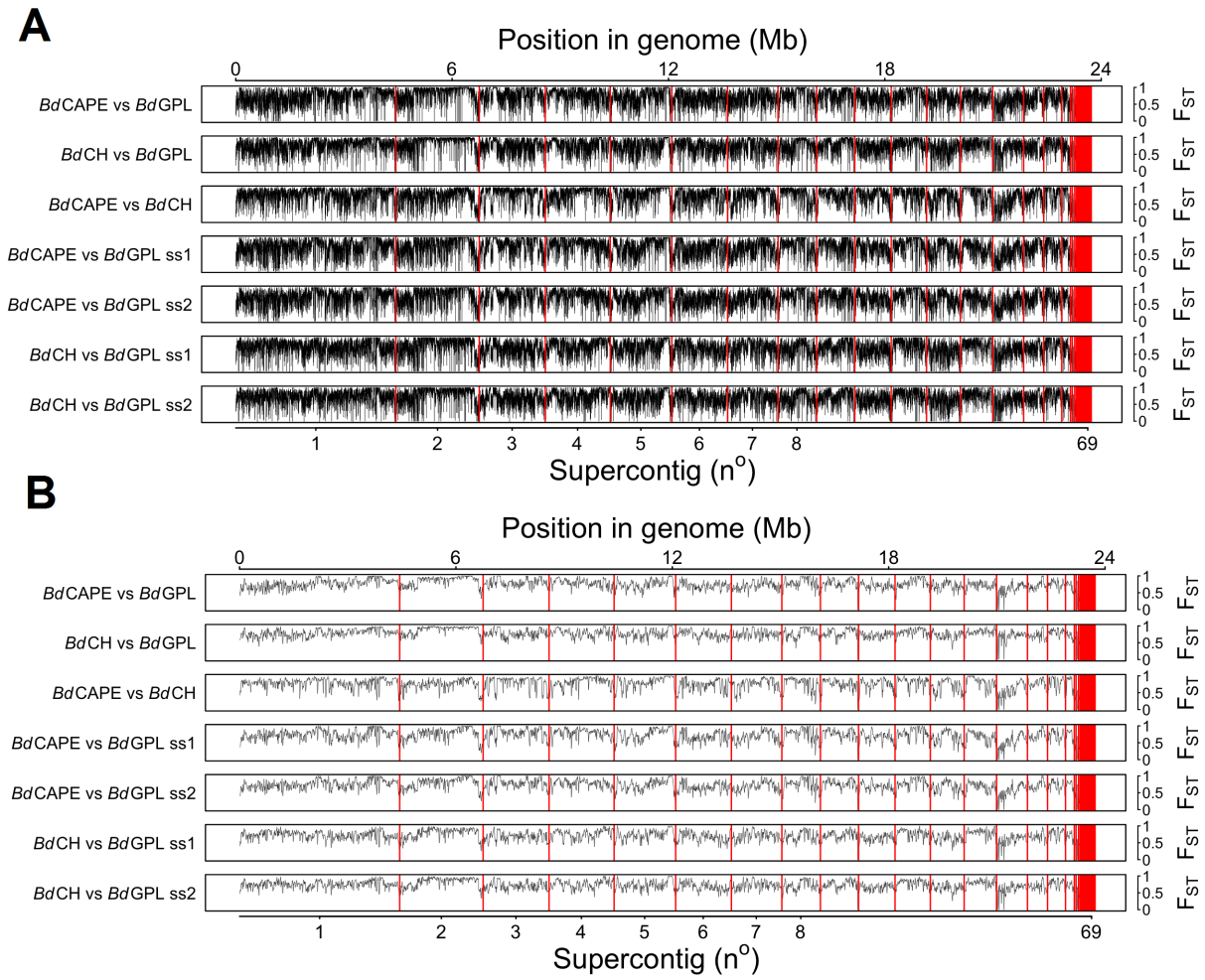


Figure 4.17: The fixation index (F_{ST}) was calculated for each pairwise lineage across window lengths of 1.4Kb (**A**) and 10Kb (**B**). All three lineages are differentiated from one another across each chromosome, with some intra-chromosomal variation. Notably, the stretch of rDNA located at the start of chromosome 14 appears to have a reduced genetic distance between each of the three lineages of *Bd*. *Bd*GPL subset (ss) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27.

Lineage	2nt per locus	3nt per locus	4nt per locus
*Bd*GPL	54 (56.84%)	28 (29.47%)	13 (13.68%)
*Bd*GPL s.s 1	2,118 (75.05%)	575 (20.38%)	129 (4.57%)
*Bd*GPL s.s 2	861 (69.89%)	301 (24.43%)	70 (5.68%)
*Bd*CAPE	5,377 (74.55%)	1,709 (23.70%)	126 (1.75%)
*Bd*CH	11,920 (71.76%)	4,184 (25.19%)	508 (3.06%)

Table 4.11: The numbers of variable sites per locus amongst haplotypes, demonstrating all lineages to be as likely to have arisen from out-crossing. *Bd*GPL subset (ss) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27.

4.4.4 Patterns of mutation and selection

To identify genes that are present in the reference sequence and absent in our panel of isolates (presence/absence polymorphism), we examined the read-depth across each of the genes. From this analysis we identified only 5 genes that were absent from our panel (Table 4.12). None of these genes were absent amongst *Bd*GPL isolates. Three of these genes were uncharacterised (non-secreted) genes and were *Bd*CAPE specific. The remaining two were an uncharacterised (non-secreted) gene and a protease (non-secreted) gene that were *Bd*CH specific. Therefore, whilst high-levels of aneuploidy are occurring, it does not appear to be resulting in frequent gene loss.

Supercontig	Positions	Gene ID	Gene category	Isolates with PA polymorphism
1	1,576,787 - 1,577,857	BDET_00586	Protease (non-secreted)	ACON (CH), CON2A (CH), APEP (CH)
4	1,094,582 - 1,095,022	BDET_03604	Uncharacterised (non-secreted)	MG1 (ZA)
9	903,715 - 904,536	BDET_06336	Uncharacterised (non-secreted)	SA1d (ZA), SA4c (ZA)
12	270,695 - 271,108	BDET_07282	Uncharacterised (non-secreted)	MG1 (ZA), RC5.1 (FR), SA4c (ZA), TF5a1 (ES)
16	237,147 - 237,968	BDET_08449	Uncharacterised (non-secreted)	APEP (CH)

Table 4.12: Only five presence absence (PA) polymorphisms relative to *Bd*GPL JEL423 were identified amongst *Bd*CAPE and *Bd*CH isolates (read depth=0), whilst none were identified amongst *Bd*GPL isolates.

To study the patterns of mutation across the nuclear genome, we categorised each of the mutations by their location in the genome in terms of coding regions (CDS), introns and intergenic regions (Tables 4.13 and 4.14 for homozygous and heterozygous polymor-

phisms respectively). In every isolate we sequenced, every variant type was found in greater abundance per kilobase in the non-coding regions (with the exception of 0.01/Kb fewer heterozygous positions in the introns compared with the CDS for isolate MG1). Specifically, introns had 1.43-1.90X SNPs, 1-1.79X heterozygous positions, 3.01-8.72X homozygous indels and 3.14-10.62X heterozygous indels compared with the CDS for each of the isolates (per Kb). Intergenic regions had 1.25-2.52X SNPs, 1.33-2.47X heterozygous positions, 2.44-7.19X homozygous indels and 5.75-9.48X heterozygous indels than the CDS for each isolate (per Kb). We further grouped heterozygous and homozygous indels in each of the three regions, identifying the introns to have 4.37-6.73X more than the CDS, and the intergenic regions to have 4.12-6.79X more than the CDS (again, both per kb). This overall pattern can be explained through selection purging deleterious mutations from the CDS. In addition, we found homozygous polymorphisms to be highly supported in all lineages in terms of uniquely mapped reads, whilst un-phased bi-allelic heterozygous positions had a smaller total proportion in the divergent lineages compared with *Bd*GPL, suggesting some heterozygous positions may be miscalled due to paralogs (Table 4.4).

Culture	Coding regions				Introns		Intergenic	
reference	SNPs	Syn.	Non-syn.	Indels	SNPs	Indels	SNPs	Indels
AP15 (IT)	2,854 (0.23)	984	1,542	703 (0.06)	1,118 (0.33)	608 (0.18)	3,526 (0.46)	1,369 (0.18)
AUL (FR)	7,218 (0.58)	2,981	3,647	693 (0.06)	2,999 (0.89)	780 (0.23)	6,571 (0.87)	1,436 (0.19)
BEW2 (CH)	1,846 (0.15)	591	999	684 (0.05)	768 (0.23)	605 (0.18)	2,189 (0.29)	1,236 (0.16)
BLI1 (CH)	1,040 (0.08)	289	577	543 (0.04)	473 (0.14)	447 (0.13)	1,466 (0.19)	805 (0.11)
ETH2 (ET)	3,321 (0.27)	1,298	1,680	656 (0.05)	1,431 (0.42)	606 (0.18)	3,127 (0.41)	1,123 (0.15)
ETH4 (ET)	3,353 (0.27)	1,281	1,707	648 (0.05)	1,383 (0.41)	592 (0.18)	3,379 (0.45)	1,184 (0.16)
JEL423 (PA)	969 (0.08)	225	552	616 (0.05)	494 (0.15)	507 (0.15)	1,516 (0.20)	977 (0.13)
MAD (FR)	6,235 (0.50)	2,679	3,122	706 (0.06)	2,417 (0.72)	704 (0.21)	4,782 (0.63)	1,394 (0.18)
MG4 (ZA)	1,862 (0.15)	554	1,034	678 (0.05)	763 (0.23)	563 (0.17)	2,387 (0.31)	1,184 (0.16)
MODS27 (IT)	1,739 (0.14)	547	964	534 (0.04)	673 (0.20)	449 (0.13)	2,078 (0.27)	835 (0.11)
MODS28 (IT)	1,795 (0.14)	552	994	681 (0.05)	703 (0.21)	572 (0.17)	2,332 (0.31)	1,218 (0.16)
SFBC014 (GB)	4,579 (0.37)	1,851	2,338	602 (0.05)	1,906 (0.56)	649 (0.19)	4,246 (0.56)	1,116 (0.15)
SP10 (IT)	2,811 (0.23)	1,099	1,417	697 (0.06)	1,338 (0.40)	624 (0.18)	3,197 (0.42)	1,293 (0.17)
VC1 (CA)	7,463 (0.60)	3,264	3,695	560 (0.05)	3,135 (0.93)	653 (0.19)	6,055 (0.80)	1,107 (0.15)
ACON (CH)	44,044 (3.54)	19,495	21,910	730 (0.06)	17,691 (5.24)	1,734 (0.51)	38,982 (5.14)	2,976 (0.39)
APEP (CH)	48,695 (3.91)	20,689	24,575	1,248 (0.10)	18,649 (5.53)	2,092 (0.62)	47,654 (6.28)	4,707 (0.62)
CON2A (CH)	43,945 (3.53)	19,488	21,825	752 (0.06)	17,679 (5.24)	1,777 (0.53)	39,224 (5.17)	3,175 (0.42)
MG1 (ZA)	36,622 (3.18)	17,692	19,799	788 (0.06)	16,339 (4.84)	1,723 (0.51)	33,310 (4.39)	2,940 (0.39)
RC5.1 (FR)	40,450 (3.25)	17,896	20,328	815 (0.07)	16,421 (4.87)	1,733 (0.51)	33,458 (4.41)	2,851 (0.38)
SA1d (ZA)	44,213 (3.55)	18,882	22,539	1,018 (0.08)	17,369 (5.15)	1,980 (0.59)	39,179 (5.16)	3,643 (0.48)
SA4c (ZA)	42,527 (3.42)	18,206	21,600	917 (0.07)	16,749 (4.96)	1,831 (0.54)	37,247 (4.91)	3,213 (0.42)
TF5a1 (ES)	42,704 (3.43)	18,379	21,660	949 (0.08)	16,862 (5.00)	1,919 (0.57)	37,385 (4.93)	3,364 (0.44)

Table 4.13: Homozygous polymorphisms were found in the coding and non-coding regions of the *Bd* nuclear genomes. SNPs in the coding regions are also tallied according to whether they result in a synonymous (Syn.) or non-synonymous (Non-syn.) amino acid change. The total numbers of each variant-type are followed by their numbers per kilobase of genomic region in parentheses.

Culture reference	Coding regions				Introns		Intergenic	
	SNPs	Syn.	Non-syn.	Indels	SNPs	Indels	SNPs	Indels
AP15 (IT)	27,159 (2.18)	10,492	13,803	1,242 (0.10)	09,837 (2.91)	1,217 (0.36)	31,299 (4.13)	04,648 (0.61)
AUL (FR)	24,671 (1.98)	09,072	12,684	1,206 (0.10)	08,340 (2.47)	1,010 (0.30)	30,077 (3.96)	04,232 (0.56)
BEW2 (CH)	26,051 (2.09)	09,909	13,389	1,135 (0.09)	08,935 (2.65)	0,985 (0.29)	28,475 (3.75)	03,502 (0.46)
BLI1 (CH)	25,921 (2.08)	10,657	13,030	0,490 (0.04)	09,475 (2.81)	0,969 (0.29)	26,080 (3.44)	02,330 (0.31)
ETH2 (ET)	27,351 (2.20)	10,516	13,872	0,930 (0.07)	09,544 (2.83)	1,079 (0.32)	30,744 (4.05)	03,077 (0.41)
ETH4 (ET)	28,696 (2.31)	10,883	14,732	1,099 (0.09)	10,094 (2.99)	1,176 (0.35)	33,147 (4.37)	04,302 (0.57)
JEL423 (PA)	22,616 (1.82)	09,208	11,351	0,536 (0.04)	07,822 (2.32)	0,800 (0.24)	20,273 (2.67)	01,972 (0.26)
MAD (FR)	24,663 (1.98)	08,745	12,919	1,301 (0.10)	08,127 (2.41)	1,067 (0.32)	29,484 (3.89)	03,742 (0.49)
MG4 (ZA)	27,415 (2.20)	10,864	13,883	0,969 (0.08)	09,892 (2.93)	1,159 (0.34)	29,958 (3.95)	03,704 (0.49)
MODS27.3 (IT)	24,467 (1.97)	10,025	12,326	0,484 (0.04)	09,061 (2.68)	0,908 (0.27)	25,077 (3.31)	02,255 (0.30)
MODS28 (IT)	27,504 (2.21)	10,783	13,950	1,126 (0.09)	09,965 (2.95)	1,180 (0.35)	30,157 (3.97)	04,116 (0.54)
SFBC014 (GB)	25,858 (2.08)	10,260	12,918	0,826 (0.07)	09,392 (2.78)	1,020 (0.30)	28,160 (3.71)	03,152 (0.42)
SP10 (IT)	25,606 (2.06)	09,644	13,166	1,054 (0.08)	08,724 (2.58)	1,041 (0.31)	28,012 (3.69)	03,495 (0.46)
VC1 (CA)	20,172 (1.62)	07,461	10,555	0,519 (0.04)	06,569 (1.95)	0,679 (0.20)	25,229 (3.33)	02,236 (0.29)
ACON (CH)	53,383 (4.29)	15,741	26,508	2,452 (0.20)	15,961 (4.73)	1,909 (0.57)	72,276 (9.53)	07,459 (0.98)
APEP (CH)	61,131 (4.91)	19,600	34,434	3,843 (0.31)	16,655 (4.93)	2,098 (0.62)	82,880 (10.9)	11,217 (1.48)
CON2A (CH)	51,126 (4.11)	15,564	25,315	1,874 (0.15)	15,737 (4.66)	1,812 (0.54)	72,869 (9.60)	07,312 (0.96)
MG1 (ZA)	35,722 (2.87)	09,781	18,249	1,588 (0.13)	09,053 (2.68)	1,141 (0.34)	54,606 (7.20)	04,881 (0.64)
RC5.1 (FR)	36,426 (2.93)	09,993	18,608	1,521 (0.12)	08,694 (2.58)	1,043 (0.31)	52,298 (6.89)	04,468 (0.59)
SA1d (ZA)	42,917 (3.45)	10,372	22,081	2,688 (0.22)	10,478 (3.10)	1,416 (0.42)	72,941 (9.61)	09,557 (1.26)
SA4c (ZA)	42,643 (3.43)	10,893	21,698	2,516 (0.20)	11,134 (3.30)	1,492 (0.44)	69,005 (9.09)	08,480 (1.12)
TF5a1 (ES)	42,481 (3.41)	21,683	21,683	2,431 (0.20)	10,678 (3.16)	1,374 (0.41)	69,193 (9.12)	08,296 (1.09)

Table 4.14: Heterozygous polymorphisms were found in the coding and non-coding regions of the *Bd* nuclear genomes. For heterozygous positions, the affect on the transcript (synonymous, Syn. or non-synonymous, Non-syn.) was determined using the alternative allele. Where two alternative alleles to the reference sequence were found (infrequently), the first present within the VCF was chosen. The total numbers of each variant-type are followed by their numbers per kilobase of genomic region in parentheses.

We categorised each of the mutations within the CDS into synonymous and non-synonymous mutations (Tables 4.13 and 4.14). A total of 391Kb of homozygous SNPs were detected amongst the 22 isolates within the CDS, accounting for 169,000 synonymous changes and 197,000 non-synonymous changes. Similarly, heterozygous positions accounted for a greater number of non-synonymous changes (179Kb) than synonymous changes (138Kb). With the exception of the reference strain *Bd* JEL423, the ratios of synonymous to non-synonymous change were between 0.49-0.91 and 0.56-0.97 for homozygous and heterozygous positions respectively. These results are in contrast to our previous SOLiD sequencing [Farrer RA *et al.*, 2011; Chapter 3], whereby greater rates of synonymous change were identified. Although shared covered positions revealed similar SNPs and therefore comparable phylogenetic trees (Figs. 4.3B and 4.4B), the greater depth of sequencing, longer 100nt paired-end reads (compared to ∼10X deep single 30nt SOLiD reads), and improved SNP-calling (Fig. 4.1) each enable more robust variants to be identified, especially over hyper-variable regions such as mutations that were are found in close proximity.

We then grouped genes by searching for secretion signals, protease domains and carbohydrate binding domains, and tested each of these for enrichment of homozygous SNPs and heterozygous positions (Table 4.15) using hypergeometric tests. We found that gene groups that carried a secretion signal (proteases, chitin-binding and uncharacterized secreted) as well as CRN-like genes, were significantly enriched for both homozygous and heterozygous polymorphisms relative to the whole set of genes. Predicted chitin-binding proteins that lacked a secretion peptide were not enriched for either homozygous SNPs or heterozygous positions, and non-secreted proteases were only enriched for synonymous amino acid changes. Conversely, CRN-like genes are only enriched for non-synonymous homozygous SNPs and not synonymous SNPs. This is further demonstrated by the ratio of 2.86-3.17:1 non-synonymous to synonymous SNPs at unique loci across all three lineages of *Bd*, compared with <1.76:1 for any other gene families tested. These findings are the first demonstration that gene families are evolving differentially in *Bd*, and that genes putatively interacting with the amphibian host as well as CRN-like genes (whose function are unknown) are undergoing increased diversifying selection compared with those that remain within the pathogen.

		All *Bd* isolates					*Bd*GPL					*Bd*CAPE					*Bd*CH				
Gene names	Enrichment in	sum	$\overline{x}$	S.D.	HgT P(X>x)	SIG.	sum	$\overline{x}$	S.D.	HgT P(X>x)	SIG.	sum	$\overline{x}$	S.D.	HgT P(X>x)	SIG.	sum	$\overline{x}$	S.D.	HgT P(X>x)	SIG.
ABC transporters	NR hom. SNP	286	10	6	1.00E+00	-	38	1	2	1.00E+00	-	182	6	4	1.00E+00	-	181	6	5	1.00E+00	-
	NR Syn.	162	6	5	8.28E-01	-	23	1	1	9.97E-01	-	101	4	3	3.62E-01	-	107	4	4	2.70E-01	-
	NR Non-Syn.	118	4	4	1.00E+00	-	14	0	1	1.00E+00	-	78	3	3	1.00E+00	-	71	3	2	1.00E+00	-
	NR het.	333	12	9	1.00E+00	-	107	4	5	1.00E+00	-	129	5	4	1.00E+00	-	137	5	4	1.00E+00	-
	NR PP	413	15	9	1.00E+00	-	79	3	4	1.00E+00	-	243	9	5	1.00E+00	-	246	9	7	1.00E+00	-
Chitin associated	NR hom. SNP	742	16	13	6.09E-01	-	134	3	6	9.75E-01	-	424	9	8	4.30E-01	-	414	9	8	8.46E-01	-
(non-secreted)	NR Syn.	347	7	6	1.72E-02	-	65	1	2	6.81E-01	-	204	4	4	1.12E-02	-	202	4	4	4.57E-02	-
	NR Non-Syn.	334	7	8	9.96E-01	-	61	1	4	9.77E-01	-	189	4	5	9.78E-01	-	182	4	5	9.98E-01	-
	NR het.	860	18	30	2.73E-01	-	256	5	9	9.03E-01	-	359	8	17	1.04E-01	-	454	10	17	3.61E-06	***
	NR PP	1,055	22	22	7.55E-02	-	206	4	8	9.35E-01	-	563	12	11	3.29E-01	-	644	14	13	1.87E-02	-
Chitin associated	NR hom. SNP	765	19	18	6.53E-48	***	144	4	5	1.52E-07	***	410	10	10	4.80E-23	***	464	12	12	8.23E-34	***
(secreted)	NR Syn.	289	7	5	1.28E-14	***	64	2	2	1.85E-04	**	161	4	3	8.98E-09	***	168	4	4	2.36E-09	***
	NR Non-Syn.	389	10	11	1.95E-24	***	65	2	3	2.34E-03	*	208	5	6	8.50E-12	***	236	6	7	9.77E-18	***
	NR het.	1,007	25	32	4.48E-96	***	250	6	9	2.12E-11	***	492	12	19	1.14E-72	***	470	12	16	7.38E-54	***
	NR PP	1,097	27	28	8.62E-82	***	207	5	7	6.52E-10	***	605	15	17	9.25E-47	***	642	16	17	4.96E-48	***
CRN-like	NR hom. SNP	2386	22	17	1.50E-230	***	344	3	5	2.22E-10	***	1353	12	11	1.45E-134	***	1345	12	10	2.06E-121	***
	NR Syn.	528	5	5	5.67E-03	*	72	1	1	1.00E+00	-	284	3	3	1.19E-01	-	299	3	3	6.39E-02	-
	NR Non-Syn.	1541	14	10	1.63E-238	***	218	2	3	1.92E-16	***	900	8	7	5.68E-148	***	881	8	6	5.93E-134	***
	NR het.	5,381	49	26	0.00E+00	***	1,222	11	10	5.03E-223	***	2,777	25	15	0.00E+00	***	2,490	23	14	0.00E+00	***
	NR PP	4,479	41	25	0.00E+00	***	850	8	8	9.94E-116	***	2,576	24	16	0.00E+00	***	2,291	21	14	0.00E+00	***
Uncharacterised	NR hom. SNP	14241	14	18	0.00E+00	***	2428	2	5	2.78E-71	***	7745	8	10	0.00E+00	***	8395	9	12	0.00E+00	***
(secreted)	NR Syn.	4573	5	6	1.21E-108	***	886	1	2	1.95E-08	***	2507	3	4	6.16E-53	***	2680	3	4	1.55E-68	***
	NR Non-Syn.	7936	8	11	0.00E+00	***	1277	1	3	8.54E-51	***	4342	4	7	0.00E+00	***	4715	5	7	0.00E+00	***
	NR het.	16,057	16	24	0.00E+00	***	4,353	4	7	1.69E-141	***	6,946	7	14	0.00E+00	***	7,316	7	13	0.00E+00	***
	NR PP	18,681	19	23	0.00E+00	***	3,712	4	6	1.43E-141	***	10,040	10	13	0.00E+00	***	10,976	11	14	0.00E+00	***
Protease	NR hom. SNP	9856	16	19	4.83E-11	***	2065	3	6	5.46E-03	*	5529	9	10	6.87E-07	***	5658	9	12	2.88E-05	***
(non-secreted)	NR Syn.	4235	7	8	2.42E-12	***	933	2	3	3.56E-03	*	2423	4	5	2.42E-10	***	2408	4	5	1.15E-05	***
	NR Non-Syn.	4891	8	11	2.65E-02	-	993	2	4	2.08E-01	-	2737	5	6	2.01E-01	-	2801	5	7	2.04E-01	-
	NR het.	9,872	17	25	1.00E+00	-	3,458	6	9	3.03E-01	-	3,814	6	13	1.00E+00	-	4,105	7	13	1.00E+00	-
	NR PP	12,975	22	26	1.16E-06	***	3,008	5	8	1.73E-04	**	7,034	12	14	4.58E-03	*	7,593	13	15	5.59E-04	**
Protease	NR hom. SNP	2506	27	31	8.60E-249	***	405	4	6	1.16E-19	***	1356	15	17	8.38E-124	***	1475	16	19	1.26E-151	***
(secreted)	NR Syn.	877	10	9	8.69E-57	***	154	2	2	2.91E-05	***	481	5	5	5.50E-30	***	509	6	5	6.10E-34	***
	NR Non-Syn.	1305	14	19	6.92E-134	***	209	2	4	1.29E-12	***	715	8	11	7.01E-69	***	768	8	12	1.65E-82	***
	NR het.	3,113	34	39	0.00E+00	***	731	8	8	1.13E-37	***	1,423	15	21	1.85E-228	***	1,453	16	22	2.91E-208	***
	NR PP	3,321	36	40	1.82E-317	***	609	7	8	6.59E-33	***	1,784	19	22	7.84E-163	***	1,949	21	24	7.93E-186	***
Uncharacterised	NR hom. SNP	61154	9	11	1.00E+00	-	13776	2	4	1.00E+00	-	34436	5	7	1.00E+00	-	35294	5	7	1.00E+00	-
(non-secreted)	NR Syn.	26805	4	5	1.00E+00	-	6268	1	2	1.00E+00	-	15026	2	3	1.00E+00	-	15526	2	3	1.00E+00	-
	NR Non-Syn.	30627	4	7	1.00E+00	-	6730	1	2	1.00E+00	-	17469	3	4	1.00E+00	-	17617	3	4	1.00E+00	-
	NR het.	66,797	10	19	1.00E+00	-	23,524	3	6	1.00E+00	-	25,109	4	10	1.00E+00	-	28,267	4	9	1.00E+00	-
	NR PP	82,086	12	16	1.00E+00	-	19,258	3	5	1.00E+00	-	44,830	7	9	1.00E+00	-	48,298	7	9	1.00E+00	-

Table 4.15: Gene categories were tested for enrichment in non-redundant (NR; at unique loci) homozygous SNPs, NR synonymous (Syn.) and non-synonymous (Non-syn.) amino acid changes, NR heterozygous (het.) positions and NR phased positions (PP). For each; the total number (sum), average ($\overline{x}$), and standard deviation (S.D.) were calculated for all isolates, and lineage specific isolates. Hypergeometric tests (HgT) were used to identify significant (SIG.) enrichment for variants, where P$\geq$0.01 (-), P<0.01 (*), P<0.001 (**) and P<0.0001 (***). NA=not applicable

We next measured the rates of synonymous substitution (dS), non-synonymous substitution (dN) and omega ($dN/dS = \omega$) for every gene in every isolate and compared values by grouping isolates into their lineages (Figs. 4.18 and 4.19). In total, we identified 1,450 genes with $\omega \geq 1$ in at least one of our isolates, suggesting positive or diversifying selection. Of those genes, 283 were identified among *Bd*GPL isolates, 816 were identified among *Bd*CAPE isolates and 746 were identified among *Bd*CH isolates. Although no clear pattern could be distinguished within *Bd*GPL (Fig. 4.18) owing to the high degree of relatedness amongst isolates, CRN-like genes in both *Bd*CAPE and *Bd*CH had the greatest median, upper quartile and upper tail values of omega (Fig. 4.19). In addition, average ω values for secreted chitin-associated genes and proteases were marginally higher than their non-secreted counter parts. Uncharacterised secreted genes also had a greater ω than either of those non-secreted gene groups. Finally, a significant enrichment of CRN-like genes and uncharacterised (secreted) genes with $\omega \geq 1$ were identified in both *Bd*CAPE and *Bd*CH (Table 4.16).

By analysing each of these 1,450 genes with $\omega \geq 1$ using Branch site models (BSM) along each of the three lineages of *Bd* (Fig. 4.4A), we identified a subset of 482 genes that show evidence for positive selection in at least one of the lineages (2D'>8.1887, Table 4.17). While no clear pattern between gene categories and lineage overlap could be seen, nine genes were identified in all three lineages: four uncharacterised (secreted) with transcript ID's 05565, 02533, 00379, 06783 and five uncharacterised (non-secreted) with transcript ID's 03962, 07794, 05877, 02935, 08088 (Fig. 4.20). A greater percent of each of the secreted gene categories were found to have accumulated an excess of non-synonymous mutations compared with their non-secreted counterpart gene categories in both *Bd*CAPE and *Bd*CH (Table 4.17). Specifically, 21% (28% for *Bd*CAPE and 14% for *Bd*CH) of the chitin-associated (secreted) with $\omega \geq 1$ also had 2D'>8.1887, while only 14% (18% for *Bd*CAPE and 9% for *Bd*CH) of the chitin-associated (non-secreted) also had 2D' >8.1887.13% of the protease (secreted) genes with $\omega \geq 1$ also had 2D'>8.1887, while only 6.9% of the protease (non-secreted) genes also had 2D'>8.1887. Finally, 13% uncharacterised (secreted) with $\omega \geq 1$ also had 2D'>8.1887, while only 5.7% of uncharacterised (non-secreted) genes also had 2D'>8.1887. For *Bd*CAPE and *Bd*CH, a moderate enrichment for uncharacterized (secreted) were identified with 2D'>8.1887 from those genes with $\omega \geq 1$, and a small enrichment for protease (secreted) in *Bd*CH (Table 4.17).

Although no clear enrichment for genes with $\omega \geq 1$ and 2D'>8.1887 was identified for *Bd*GPL, the most striking finding of this analyses was found among *Bd*GPL isolates, 349/482 (72%) of the genes with a signature of positive selection were identified on this lineage

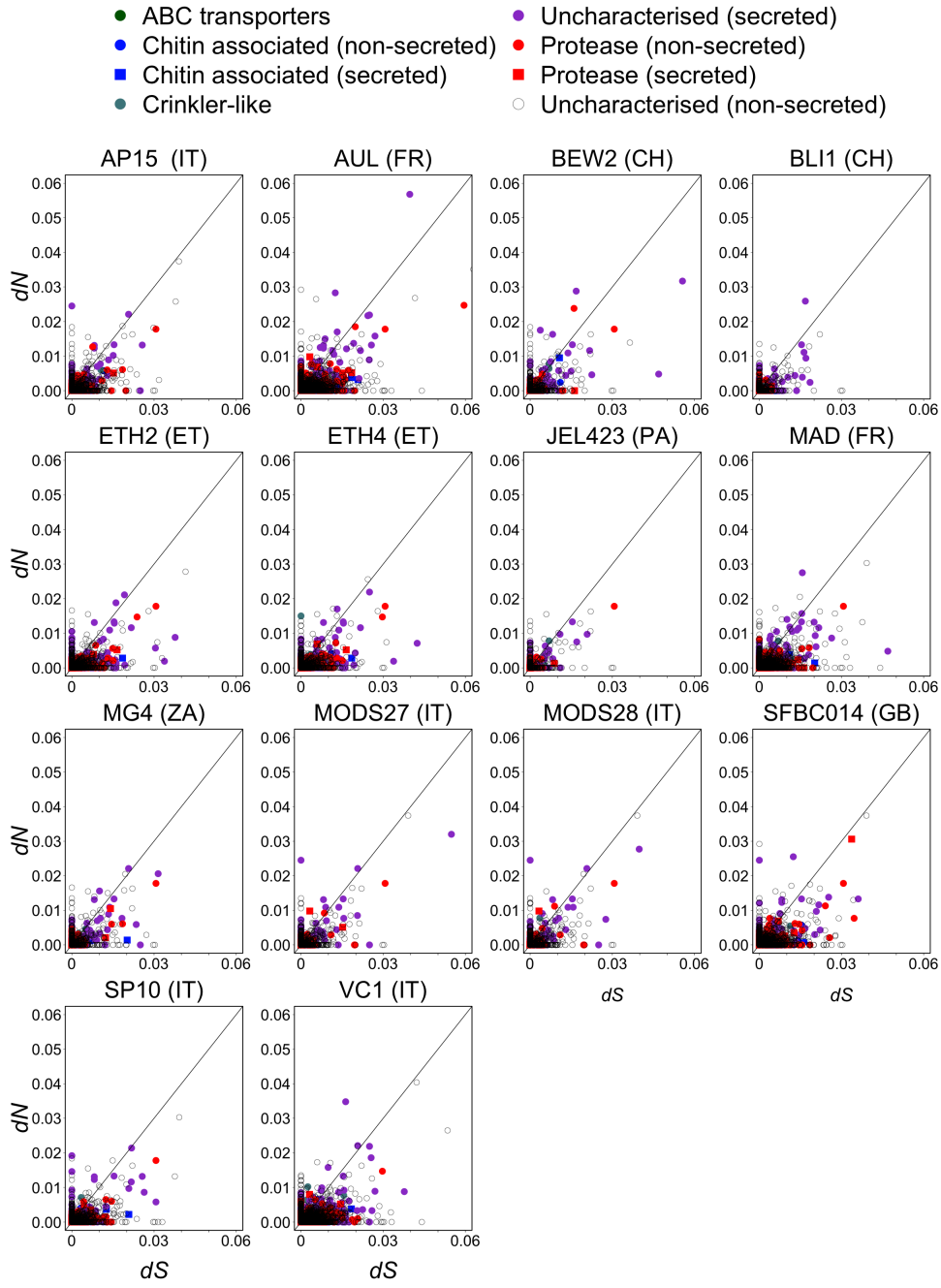


Figure 4.18: The ratio of non-synonymous mutation per non-synonymous site (dN) vs synonymous mutation per synonymous site (dS) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the *Bd*GPL. The line designates the ω value (dN/dS), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation.

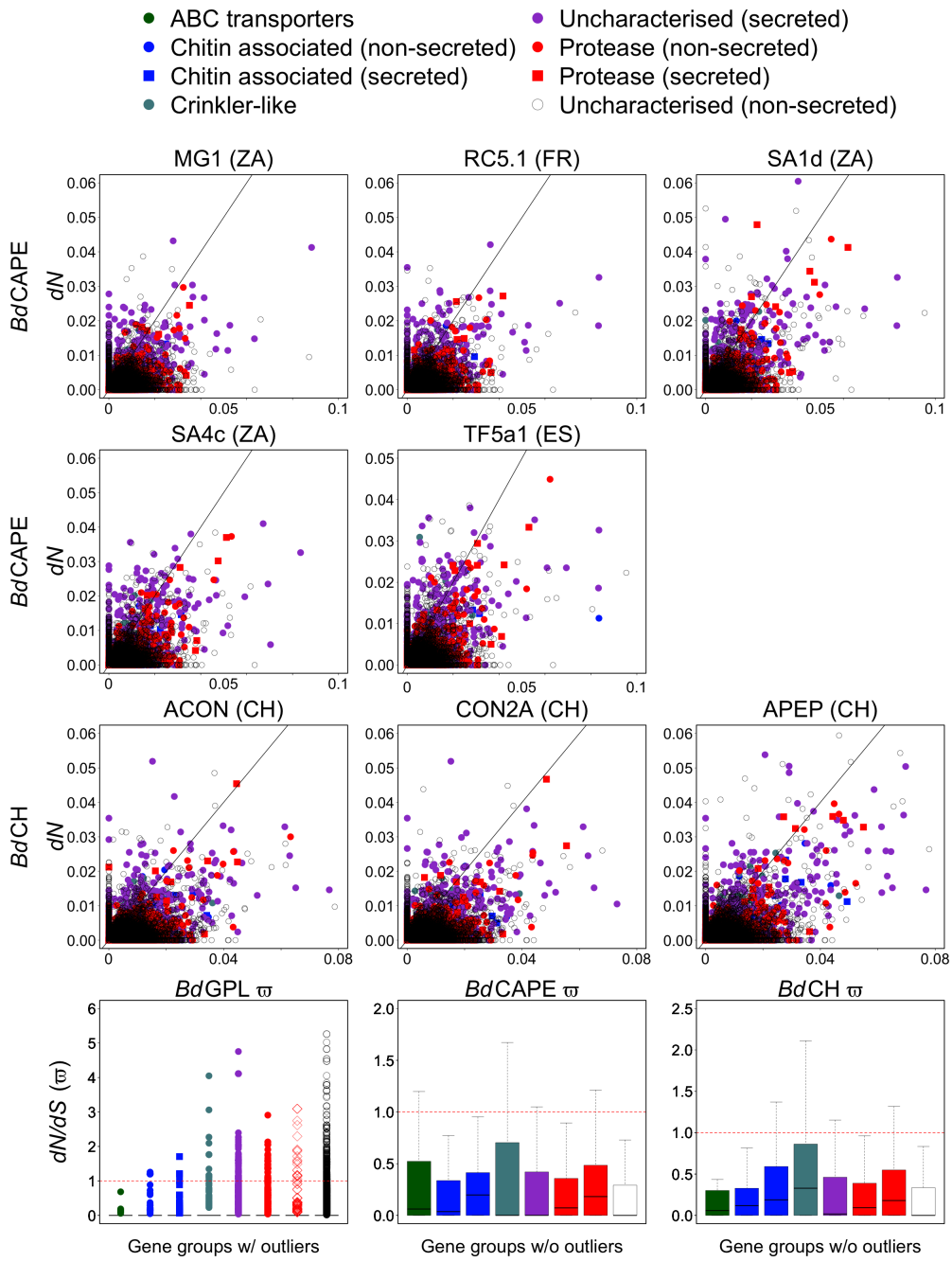


Figure 4.19: The ratio of non-synonymous mutation per non-synonymous site (dN) vs synonymous mutation per synonymous site (dS) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the three lineages. The lines designate the ω value (dN/dS), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation. Summaries of ω values for all genes in each of the three lineages are shown in the final three plots.

compared with only 23% for each of the other two lineages. This finding suggests that *Bd*GPL has been undergoing greater levels of positive selection than either *Bd*CAPE or *Bd*CH, and that much of the variation and patterns of selection identified in *Bd*CAPE and *Bd*CH may stem from the ancestral *Bd*GPL. Finally, our discovery of chromosome copy number variation among most of the isolates of *Bd* likely to both introduce and increase levels of genomic variation, is not great enough to erode this signal of selection.

	Total number	Genes w/	*Bd*GPL		*Bd*CAPE		*Bd*CH	
Category	of genes	$\omega\geq1$ *Bd*	$\omega\geq1$	P(X>x)	$\omega\geq1$	P(X>x)	$\omega\geq1$	P(X>x)
ABC transporter	29	7	0	NA	6	1.44E-02	3	2.27E-01
Chitin-associated (NS)	48	11	2	1.99E-01	7	7.14E-02	4	3.83E-01
Chitin-associated (S)	41	7	3	4.15E-02	3	5.33E-01	4	2.64E-01
Crinkler-like	110	61	6	6.32E-02	41	1.42E-16	39	2.46E-16
Protease (NS)	599	101	28	1.69E-02	48	8.44E-01	55	2.29E-01
Protease (S)	93	26	6	2.96E-02	15	1.07E-02	11	9.19E-02
Uncharacterised (NS)	6,897	975	195	1.00E+00	539	1.00E+00	493	1.00E+00
Uncharacterised (S)	985	262	43	1.40E-02	157	3.26E-13	137	3.88E-10

Table 4.16: Number and category of genes with $\omega\geq1$ in each of the three lineages. NS = non-secreted, S = secreted. Hypergeometric tests were used to identify significant enrichment for gene categories. Significant enrichment for Crinkler-like genes and uncharacterised (secreted) genes with $\omega\geq1$ were found in both *Bd*CAPE and *Bd*CH. NA=not applicable

	*Bd*GPL				*Bd*CAPE				*Bd*CH			
	2D' SIG.			HgT	2D' SIG.			HgT	2D' SIG.			HgT
Category	1%	5%	1-5%	P(X>x)	1%	5%	1-5%	P(X>x)	1%	5%	1-5%	P(X>x)
ABC transporter	2	0	2 (28.57%)	2.24E-01	0	0	0 (0%)	NA	0	0	0 (0%)	NA
Chitin-associated (NS)	2	1	3 (27.27%)	2.61E-01	2	0	2 (18.18%)	4.37E-02	0	1	1 (9.09%)	1.92E-01
Chitin-associated (S)	3	1	4 (57.14%)	1.07E-02	1	1	2 (28.57%)	1.16E-02	1	0	1 (14.29%)	8.88E-02
Crinkler-like	8	2	10 (16.39%)	9.03E-01	3	2	5 (8.20%)	3.07E-01	3	2	5 (8.20%)	2.92E-01
Protease (NS)	21	4	25 (24.75%)	3.81E-01	4	1	5 (4.95%)	7.89E-01	5	4	9 (8.91%)	2.04E-01
Protease (S)	5	1	6 (23.08%)	4.40E-01	2	0	2 (7.69%)	3.10E-01	3	2	5 (19.23%)	9.56E-03
Uncharacterised (NS)	194	46	240 (24.62%)	2.23E-01	36	24	60 (6.15%)	9.96E-01	30	22	52 (5.33%)	1.00E+00
Uncharacterised (S)	49	10	59 (22.52%)	7.13E-01	21	12	33 (12.60%)	3.80E-04	22	12	34 (12.98%)	1.05E-04

Table 4.17: Number and category of genes with $\omega\geq1$ and 2D'$\geq$11.4076 (1% significance after Bonferroni correction) and 11.4076>2D'>8.1887 (5% significance after Bonferroni correction) in each of the three lineages. NS = non-secreted, S = secreted. Hypergeometric tests (HgT) were used to identify significant enrichment from genes with $\omega\geq1$. NA=not applicable

4.5 Discussion

Recent studies have attributed aspects of *Bd*'s pathogenesis to the presence of a number of putative virulence factors that include proteases and chitin-binding proteins [Joneson S *et*

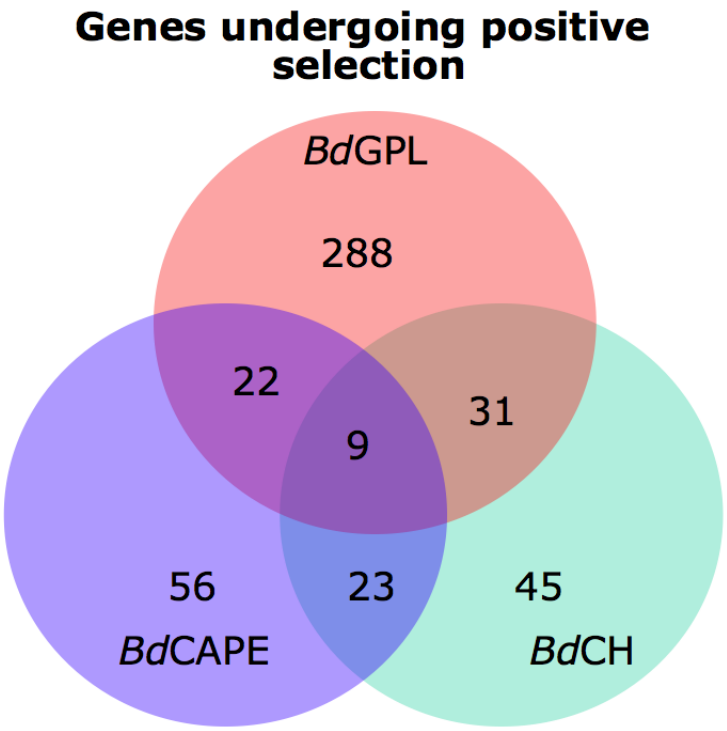


Figure 4.20: A Venn diagram showing the number of genes undergoing positive selection according to the Branch site model (BSM), where genes had 2D'>8.1887. The nine genes identified in all three lineages were four uncharacterised (secreted) with transcript ID's 05565, 02533, 00379, 06783 and five uncharacterised (non-secreted) with transcript ID's 03962, 07794, 05877, 02935, 08088.

al., 2011; Sun G *et al.*, 2011;, Abramyan J *et al.*, 2012]. The former category contain M36 or S41 domains that are thought to degrade host-cellular components, and these protease families are known to have undergone extensive expansions in *Bd* since its divergence from free-living saprobes such as *Homolaphlyctis polyrhiza* [Joneson S *et al.*, 2011]. Chitin binding proteins are thought to be involved in pathogenesis by allowing *Bd* to bind to keratinized host cells and to subsequently enter the host cells [Abramyan J *et al.*, 2012]. To date, the functional nature of the crinkler-like family in *Bd* has only been inferred owing to their sequence similarity to host-translocated proteins of known virulence in oomycetes [Sun G *et al.*, 2011],. Our data show that, across this global panel of 22 isolates and three lineages, the secretome and crinkler-like family of genes manifest higher diversity of homozygous and heterozygous SNPs, enrichment for non-synonymous mutations and greater dN/dS (ω) ratios when compared against classes of genes that do not contain a signal peptide. This shows that these gene families are evolving most rapidly in *Bd*, and that gene-products that interact with the amphibian host are undergoing diversifying (or reduced purifying) selection when compared with those gene-products that remain intracellular. These findings suggest that *Bd* has had an evolutionary association with amphibians that predates the radiation of the lineages that we have characterised here, and is further evidence that this chytrid has an obligate rather than an opportunistic association with its amphibian hosts.

By mapping read-depth and SNPs across these genomes, we discovered that widespread genomic variation occurs within and amongst *Bd* isolates from the level of SNPs up to heterogeneity in ploidy amongst genomes and amongst chromosomes within a single isolate. Individuals from all three lineages harboured chromosome copy number variation along with predominantly or even entirely diploid, triploid and tetraploid genomes, and is the first example of CCNV ocurring in a basal fungal lineage. This variation may itself, reflect only part of the full diversity in *Bd*, as +2/+3 shifts in ploidy, whole genomes in tetraploid, or chromosomes in pentaploid or greater, may occur and await discovery. Chromosomal genotype was shown to be highly plastic as significant changes in CCNV occurred in as few as 40 generations in culture. It is not known whether other chytrid species also undergo CCNV, or if this is a unique feature of *Bd* and hence may be intrinsic to its parasitic mode of life. Currently, CCNV is known to occur in a variety of protist microbial pathogens, including fungi, however it is currently not known whether this genomic-feature is specific to a parasitic life-style, or is a more general feature of eukaryote microbes; identifying the ubiquity of CCNV or otherwise across nonpathogenic species will therefore be of great interest. Further, the manner in which the plasticity of CCNV in *Bd* affects patterns of global transcription

and hence the phenotype of each isolate also remains to be studied. However, it is clear from research on yeast, *Candida* and *Cryptococcus* that CCNV significantly contributes to generating altered transcriptomic profiles, phenotypic diversity and rates of adaptive evolution even in the face of quantifiable costs; understanding the relationship between CCNV and *Bd*-phenotype will therefore likely be key to understanding its patterns of evolution at both micro- and macro-scales.

Whilst differing numbers of individual chromosomes presents a potential barrier to the standard model of meiosis, homologous recombination may still be occurring via mitotic processes within compatible genomes. In order to study recombination amongst our isolates, we determined the phase of our reads and constructed haplotypes that were suitable for traditional population genetic tests. This showed that, whilst the majority of the genomes from all three lineages manifest widespread linkage disequilibrium, recombination could still be detected across each chromosome and in all genomes. Crossovers (measured both as the proportion of SNPs that change phase and the numbers of haplotypes failing Hudson's four-gamete test) were found to occur much more frequently within the *Bd*CAPE and *Bd*CH lineages compared to *Bd*GPL, and these two lineages accordingly manifest lower average linkage disequilibrium. All of the *Bd*CH genomes that we sequenced stem from a single initial passaged isolate originating in 2007. This suggests that either the high rates seen here have accrued since the isolate was taken into culture (suggesting a very rapid rate of recombination *in vitro*) or that we are characterising recombination events that occurred prior to the isolation of *Bd*CH and have been preserved as a consequence of the multiple-ploidy nature of *Bd*. In support of the latter hypothesis, comparisons between population-level data for *Bd*GPL and *Bd*CAPE show that *Bd*GPL is far less recombinogenic and has been undergoing a largely clonal expansion since its emergence, consistent with previous observations made by James TY *et al.* [James TY *et al.*, 2009]. These data suggests that the global *Bd*GPL population is derived from a less recombinogenic ancestor than either *Bd*CH or *Bd*CAPE, that contemporary recombination is not occurring at a rapid rate and, where it occurs, is the result of a selfing rather than outcrossing events.

The discovery of a lower proportion of variable sites across haplotypes in addition to the lower proportion of heterozygous positions in *Bd*GPL compared against *Bd*CAPE or *Bd*CH does not support the notion that *Bd*GPL is an outbred hybrid lineage as previously proposed [Farrer RA *et al.*, 2011]. The discovery of a new *Bd* lineage found in Brazil (*Bd*Brazil) along with an isolate that is a likely *Bd*GPL/*Bd*Brazil recombinant [Schloegel LM *et al.*, 2012] strongly implies that *Bd* retains the ability to outcross, despite having a

primarily clonal genome and life cycle. However, values of F_{ST} across our dataset show no introgression between the three lineages; this demonstrates that they have remained separate since their divergence and suggests that outcrossing between lineages of *Bd* is rare or spatially restricted. Further broad-scale collections of isolates and extension of our comparative-population genomic analyses will allow the assignment of more accurate rates of introgression across evolutionary timescales.

We show that rates of recombination are uneven across the genome, with CRN-like genes enriched for crossovers suggesting either that CRN-like genes might have features that favour recombination or that recombinants of these genes have a fitness advantage and are thus more likely to reach fixation than recombinants at other locations in the genome. CRNs were also enriched for non-synonymous polymorphisms and are amongst the most polymorphic genes in *Bd*'s genome. Within the oomycete genus *Phytophthora*, CRNs manifest diverse carboxy-terminal domains and high rates of homologous recombination targeted to the conserved HVLVXXP motif, suggesting that the mosaic domains of CRNs are being shuffled by recombination [Raffaele R *et al.*, 2012]. Recently, a number of *Bd* CRNs have been shown to be highly expressed on host tissue *in vitro* [Rosenblum EB *et al.*, 2012]. Therefore, whilst these genes in *Bd* lack a secretion signal, their expression, accumulation of genetic variation in terms of recombination and ω values, and similarities with oomycete CRNs strongly suggest that a number of these CRN-like genes are functional in *Bd*. However, whether they contribute directly or indirectly to the virulence of *Bd* remains to be shown.

Our demonstration of multiple hierarchies of cryptic genomic variation in *Bd* in terms of CCNV, ubiquitous and potentially targeted recombination, and selection, points to an ability to generate diversity without the necessity of an obligate sexual stage. Our study has uncovered high levels of genotypic plasticity that are likely to cause widespread phenotypic plasticity even without the need to invoke outcrossing. These large and small-scale changes are therefore likely to contribute to rapid evolutionary rates in the face of an effective host response. Such 'genomic instability' may explain the diverse phenotypic responses observed in *Bd* [Woodhams DC *et al.*, 2008], and may also explain the enormous diversity of hosts and biomes that this generalist pathogen has managed to infect.

4.6 Acknowledgments

We thank Benedict Schmidt, David Gower, Claude Miaud, Ché Weldon, Jaime Bosch, Giulia Tessa and Phineas Hamilton for collection of isolates, and Hanna Larner for assistance with Illumina sequencing. This project was funded by the UK Natural Environmental Research Council (NERC) grant NE/E006701/1, the European Research Council (ERC) grant 260801-BIG_IDEA, the Swiss National Science Foundation grant 31-125099 and the Biodiversa project RACE: Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity (http://www.bd-maps.eu).

Chapter 5

Conclusions and Discussion

5.1 Introduction

Since *Bd*'s discovery in the 1990s, it has been shown as a proximate driver of extinctions and extirpations of amphibians. However, *Bd* is no exception to the fungal kingdom, as other widespread emergences of disparate fungal diseases continue unabated. Perhaps the most recent example to gain widespread attention (as of November 2012) is the sexual form of the Ascomycota *Chalara fraxinea* (*Hymenoscyphus albidus*) [Kirisits T *et al.*, 2012], which has recently been identified in the UK thereby threatening the 80 million (3^{rd} most common[1]) native ash trees. Other recent examples include *Geomyces destructans* threatening little brown bat (*Myotis lucifugus*) populations across the US [Gargas A *et al.* 2009, Frick WF *et al.* 2010] and *Cryptococcus* species that cause more than 1 million new infections per year, killing more than 650,000 people per annum, most commonly in AIDS patients in Africa [Park BJ *et al.*, 2009].

Understanding the biology of emerging infectious diseases is crucial for informing the most suitable policies to curb or control their impact. During this project, I have discovered two previously unknown lineages of *Bd*, thereby updating previous evidence that erroneously concluded there was a paucity of genetic variation occurring amongst *Bd* isolates. Furthermore, using comparative and population genetic approaches, I have gained new insights into the evolution of *Bd* including evidence for the generation of rapid cryptic genotypic diversity, which again refutes genetic homogeneity and presents new candidate mechanisms that are

[1] http://www.british-trees.com

likely to contribute to *Bd*'s success across multiple biomes and host-species.

In this chapter I aim to summarise and discuss our findings presented in the previous chapters, with a focus on 1) the epidemiological features of *Bd* (section 5.2), the pathogenicity determinants of *Bd* (section 5.3), and the life-cycle of *Bd* (section 5.4). Where necessary, I will introduce key studies and hypotheses that can be extrapolated to the study of *Bd*. Finally, I will discuss and integrate these findings with the wider disease and EID literature and suggest where future work could be directed.

5.2 Epidemiology of *Bd*

5.2.1 Spatial-temporal distribution of each of the *Bd* lineages

Over the course of this project we have genotyped *Bd* from five separate continents, thereby identifying two novel and deeply divergent lineages from those belonging to the 'Global Panzootic Lineage' (*Bd*GPL). However, the large majority of our isolates did fall within the *Bd*GPL, including previously genotyped isolates of *Bd* [James TY *et al.*, 2009], both isolates chosen for whole genome Sanger sequencing (isolates JEL423 [MIT JEL423 genome sequence] from a Lemur leaf frog (*Phyllomedusa lemur*) in Panama and isolate JAM81 [JGI JAM81 genome sequence] from a mountain yellow-legged frog (*Rana muscosa*) in the California Sierra Nevada), and all the isolates in our panel that are associated with regional epizootics, recovered from five separate regions where *Bd* has been identified (North America, Central America, the Caribbean, Australia, and Europe).

The newly identified 'Cape Lineage' (*Bd*CAPE) includes two isolates from the island of Mallorca, and one from the Cape Province, South Africa. This clade supports the hypothesis of Walker SF *et al.* [Walker SF *et al.*, 2008], which reported accidental spill-over of *Bd* from captive Cape clawed frogs (*Xenopus gilli*) to Mallorcan midwife toads (*Alytes muletensis*) through a captive breeding and reintroduction program for this endangered species. We subsequently identified *Bd*CAPE in the KwaZulu Natal province of South Africa and the Loire et Cher department of France, further demonstrating that multiple emergences of amphibian chytridiomycosis have occurred (Chapter 4), as well as confirming that anthropogenic spread of amphibians is directly responsible for at least one of these emergences.

The third novel lineage we identified was composed of a single isolate (named 0739),

derived from a common midwife toad (*Alytes obstetricans*) inhabiting a pond near the village of Gamlikon, Switzerland (Chapter 3), which we therefore named the 'Swiss Lineage' (*Bd*CH). Whilst we have subsequently sequenced *Bd* from nine separate municipalities in Switzerland, all were later identified as *Bd*GPL isolates (Chapter 4). Therefore, the origin and distribution of this third lineage remains the most enigmatic of those we identified, even after a specific attempt to find it. Further sampling, including with a focus around other sites near Gamlikon is necessary to establish whether *Bd*CH is a European-endemic isolate, or an introduction from elsewhere.

In addition to the lineages we discovered, the existence of a potentially fourth *Bd* lineage associated with the Japanese giant salamander, *Andrias japonicus* [Goka K *et al.*, 2009] has been identified, and currently defined by a short fragment of the ribosomal DNA dissimilar to the rDNA sequences of *Bd*GPL. However, it remains to be seen whether this lineage is endemic to Japan (in parallel to our *Bd*CH), or exists elsewhere in the world. Indeed, as of writing this, we're still uncertain whether this *Bd*JAPAN is genetically distinct from *Bd*CH or *Bd*CAPE. The discovery of a fifth *Bd* lineage found in Brazil (*Bd*Brazil) confirms that the diversity of *Bd* represents a patchwork of genetically and at least in the case of *Bd*GPL and *Bd*CAPE, phenotypically diverse lineages.

Recently, an isolate was detected to have a hybrid genotype between *Bd*GPL and *Bd*Brazil [Schloegel LM *et al.*, 2012], which was identified using the software New Hybrids [Anderson EC *et al.* 2002], which implements Bayesian model-based clustering approach to identify hybrids, along with principle component analysis (PCA) of MLST data and a neighbour-joining tree supporting its relationship to the two lineages. This finding shows that *Bd*GPL has maintained its ability to outcross (Schloegel LM *et al.*, 2012), and suggests that a single hybrid origin of *Bd*GPL via ancestral meiosis is possible (as we suggested in chapter 3). Although we later discovered that *Bd*GPL has no differences in number of variable sites in its haplotypes compared with *Bd*CAPE or *Bd*CH (Chapter 4), the ancestral meiosis could have occurred between parents with spatially or genetically limited distances [Henk DA *et al.*, 2012]. Finally, our F_{ST} values between each of the three lineages (Chapter 4) revealed a lack of evidence for introgression between those lineages, suggesting these lineages have remained separate since their divergence. However, as shown in the Schloegel LM *et al.* study [Schloegel LM *et al.*, 2012], additional sampling may still reveal hybrids from these lineages.

Future work is needed to incorporate the *Bd*Brazil and *Bd*JAPAN lineages, with those

identified during this thesis. Most pressing of all is to verify their status as distinct lineages. In either case, this will give a clearer picture of the distribution and genetic diversity that exists amongst *Bd* isolates. Furthermore, if any have been independently found and wrongly named as distinct lineages, this could suggest new candidate hosts or trade-routes between Asia and Europe (e.g. if *Bd*CH is *Bd*JAPAN, then *Bd*CH may not be an endemic lineage to Switzerland, and the transmission may have been acquired from Asia).

Despite uncertainties regarding the full extent of the population genetic structure and the phylogenetic relationships between some of these lineages, the (up until recent) conclusion that *Bd* consisted of a single clonal pathogen, has now been quite clearly replaced with the view that *Bd* consists of a patchwork of genetically distinct lineages, and with potentially new hybrid lineages emerging where more than one lineage exists. Each of these populations is then being geographically dispersed by a number of routes including the trade of amphibians. Given this trend, future sampling seems likely to unravel additional lineages, as well as help distinguish endemic from novel lineages in any given location.

Undoubtedly, anthropogenic movement of *Bd via* the global trade and conservation of amphibians appears to explain much of the recent global emergence of chytridiomycosis and we confirmed this for *Bd*GPL and *Bd*CAPE. However, the spread of *Bd* is both historical and contemporary, and measures at controlling it remain lacking. Fortunately, several areas that contain high amphibian biodiversity appear to be free of infection, so far. The most notable of these regions is the island of Madagascar, which contains >460 species of amphibian [Vieites DR *et al.*, 2009]. The potential for *Bd* to extirpate multiple species in this unique and highly diverse community of amphibians has led to calls for a high degree of biosecurity to be implemented [Andreone F *et al.*, 2008; Lötters SD *et al.*, 2011]. Knowing which countries and host-species are most at risk of infection should aid policy-makers in targeting the most likely sources of spread. For example, the importation of known vector species (such as the *Bd*-tolerant Bullfrogs and African Clawed Frogs) into Madagascar, especially those from *Bd*+ regions, carry a high risk to susceptible species such as those from the amphibian family Mantellidae, which are found only in Madagascar and the northern island Mayotte.

5.2.2 The origin(s) of *Bd*

To date, the oldest *Bd*+ samples were identified from a specimen of Fraser's Clawed Frog (*Xenopus fraseri*) collected from the Batouri District, Cameroon in 1933, followed by spec-

imens of *Xenopus laevis* from Lake Bunyoni, Uganda in 1934 [Soto-Azat C *et al.*, 2010], concordant with the "Out of Africa" hypothesis. However, as increasingly old specimens are sequenced for residual *Bd* DNA, new historical locations of importance are likely to be identified. In our studies, *Bd*GPL and *Bd*CAPE were both identified across multiple South African provinces, with directionality of *Bd*CAPE going from South Africa to Mallorca (Farrer RA *et al.*, 2011; Chapter 3). More recently, *Bd*GPL was also identified from Ethiopia (Chapter 4), which again is concordant with an African origin, although where in Africa remains uncertain.

In chapter 3, we presented a separate hypothesis for the origin of *Bd*GPL, whereby anthropogenic mixing of allopatric lineages of *Bd* (which may or may not have originated in Africa) could have led to the generation of the hypervirulent lineage *via* hybridisation. However, our subsequent discovery of a lower proportion of differences in haplotypes per loci in *Bd*GPL compared against *Bd*CAPE or *Bd*CH (Chapter 4) does not support this hypothesis - although neither does it rule it out, as outcrossing could have still occurred between two closely related lineages, as suggested by Henk DA *et al.* [Henk DA *et al.*, 2012].

In order to determine the origin of *Bd*, assessments needs to be made of:

1. Recent local introductions and expansions such as the simultaneous waves of population declines documented in Central America and Australia [Berger L *et al.*, 1998; Lips KR *et al.*, 2006]

2. Recent intercontinental introductions such as that caused by the transport of South African Clawed frogs (*Xenopus*) infected with *Bd*CAPE to the Durrell Wildlife Conservation Trust (DWCT), Jersey in 1991, where they were co-housed with Mallorcan Midwife Toads (*Alytes muletensis*) prior to their re-introduction to Cocó de sa Bova in Mallorca [Walker SF *et al.*, 2008; Farrer RA *et al.*, 2011]

3. Evidence of historically infected amphibians such as those museum specimens from Cameroon and Uganda in the 1930s [Soto-Azat C *et al.*, 2010]. Each of these assessments have so far been achieved using PCR assays for *Bd*s nuclear ribosomal ITS regions or MLST for a number of polymorphic loci, however our studies (Chapters 3 and 4) have been the first to use whole-genome sequencing typing (WGST) to track *Bd*'s spatial-temporal distribution.

An increasing number of studies are using WGST to track the spread of pathogens,

date their emergence, and identify their origin, including amongst an outbreak of the bacteria *Mycobacterium tuberculosis* (*Mtb*) across the northern states of Germany [Roetzer A *et al.*, 2013], *invasive non-typhoidal Salmonella* (*iNTS*) across sub-Saharan Africa [Okoro CK *et al.*, 2012] and the spread of the fungus *Apophysomyces* in the US state of Missouri [Etienne KA *et al.*, 2012]. Comparisons between WGS has been shown to be superior to conventional genotyping methods in terms of accuracy [Roetzer A *et al.*, 2013], and the ability to more reliably use molecular dating (as was also performed for *Mtb*, dating the outbreak to between 1993-1997 [Roetzer A *et al.*, 2013]). Despite these advantages, marker based genotyping studies are still more common than WGST for outbreak strains of both bacteria and fungi due to limitations in cost and time. However, with the increasing number of genomic centers, development of smaller and cheaper platforms such as Illumina's MiSeq and Life Technologies' Ion Torrent, as well as reducing costs for both the chemistry and platform, this seems likely to change in the near future.

During this study we used our first round of whole-genome sequences of *Bd* to date its emergence for the first time, whereby variation between isolates separated by a known quantity of time can be used to estimate the observed variation between isolates separated by an unknown time (known as tip dating). Whilst tip dating can be prone to underestimating age due to greater numbers of deleterious mutations that selection and population bottlenecks have not had time to purge, confidence can be assessed using bootstrapping and sampling from different chromosomes that could be expected to have similar predicted ages. Our study was the first to date the emergence of *Bd*GPL, as well as its divergence with *Bd*CAPE and *Bd*CH. The median dates for the emergence of *Bd*GPL ranged from 35-257 years before present (ybp), which is compatible with its spread within the amphibian trade. The median dates for the divergence between all of our isolates sequenced in our first panel of *Bd* isolates sequenced with SOLiD ranged from 140-930ybp. Therefore, whilst *Bd*GPL arose more recently, the ancestral population that each of the lineages arose from was far older and predates the global trade of amphibians.

To ascertain *Bd*s ancestral evolution from other species of chytridiomycetes, comparative genomics can be used to reveal fixed large-scale changes in genes and genome structure. Recently, this approach has been successfully applied to *Bd* using comparisons to the chytrid *Homolaphlyctis polyrhiza*, revealing gene-family expansions of carbohydrate-binding module (CBM) proteins [Abramyan J *et al.*, 2012], fungalysin metallopeptidases and serine proteases [Joneson S *et al.*, 2011], each of which are thought to be important in the transition of pathogenicity since its divergence millions of years ago. Unfortunately, a more recently

diverged extant chytridiomycetes has not yet been identified, which would enable a more detailed description of *Bd*'s evolutionary history. Should a geographic origin of *Bd* be identified, this region of the world would also likely be a good candidate for identifying more ancient relatives of *Bd*. Alternatively, future work identifying new species of chytrids is also likely to identify a more common recent ancestor of *Bd*, and potentially give clues to *Bd*'s origin.

Future work needs to clarify how well the newest sequence datasets (either ours, or other high-quality next-generation sequencing datasets yet to be published) consisting of greater depth, coverage and base calling accuracy would support these dates. In addition, our recent successes with extracting haplotypes from short-read data should also allow a more accurate model of evolution to be used over these sites. For example, our tip-dating without phase data coded each nucleotide through 10 possible states (AA, CC, GG, TT, AC, AG, AT, CG, CT and GT). With the phase data, we could use a standard 4-state DNA substitution model such as the GTR [Tavaré S, 1986] or HKY85 model [Hasegawa M *et al.*, 1985] for individual haplotypes, thereby eliminating incorrect initial states of sequences, and taking into account such features as Ti/Tv or unequal base frequencies.

5.3 *Bd*'s pathogenicity determinants and mechanisms of infection

5.3.1 Evolution of virulence

A number of theories predict why pathogens should evolve to do harm to their hosts, which they ultimately rely on to reproduce. Based on this premise, the avirulence hypothesis suggests that parasites (and later, by extension, pathogens) should evolve towards avirulence because the death or disability of the host is harmful to the coloniser [Smith T, 1904] (recently reviewed: [Alizon S *et al.*, 2009]). However, this does not explain why many pathogens (including those that are endemic) do harm their hosts [Ball GH, 1943]. One explanation is that the some pathogens may have evolved for a separate host or habitat, which by coincidence causes virulence in a new host. It has also been argued that while mutualism can evolve out of parasitism, the reverse can also occur [Allen ES, 1934]. This assertion forms part of the short-sighted evolution hypothesis, whereby any improved ability to reproduce

will quickly replace the population regardless of its eventual effect on the host [Levin BR *et al.* 1994]. Finally, the trade-off hypothesis [Anderson RM *et al.*, 1982] balances the assertions of the avirulence hypothesis with the predicted benefits of virulence. These benefits include greater access to host resources and the potential to transmit to a new host before harm is done to the pathogen. Thus, the hypothesis suggests that there is a trade-off between virulence and transmission.

These different hypotheses are each suitable for explaining the evolution of virulence in a range different pathogens, depending on their individual selection pressures [Alizon S *et al.*, 2009]. For example:

- Coincidental virulence has been reported in the B2 phylogenetic group of *Escherichia coli*, where virulence genes involved in iron uptake may have been originally selected and further maintained for resistance to digestion by free-living protozoa [Adiba S *et al.*, 2010].
- The short-sighted evolution hypothesis is supported by the Ebola hemorrhagic fever virus, which is characterised by sporadic outbreaks, high fatality rates and relatively short incubation-times [Casillas AM *et al.*, 2003].
- The trade-off hypothesis is supported by asymptomatic (set-point) viral load in patients with HIV-1, which is clustered around those that correlate with both decreased time to AIDS and increased infectiousness, thereby suggesting it could have evolved to optimise its transmissibility [Fraser C *et al.*, 2007].

The ability of *Bd* to exist in multiple host species [Fisher MC *et al.*, 2009] and remain in the environment [Walker SF *et al.*, 2007] suggests that it should largely avoid the normal host-pathogen fluctuations and the trade-offs associated with hypervirulence. *Bd* infection is therefore concordant with both the short-sighted and trade-off hypothesis. In chapter 3, we presented the first evidence for multiple *Bd* lineages with distinct virulence profiles, whereby common toad (Bufo bufo) tadpoles exposed to repeat high doses of *Bd*GPL had significantly increased infection and mortality status after 122 days post first exposure compared with toads exposed to *Bd*CAPE isolates. Because both *Bd*GPL and *Bd*CAPE are now recognised as globalised lineages, it is unclear why both hyper- and hypo- virulence should have evolved, as no evidence suggests that *Bd*CAPE carries any advantage (for example in its transmission) over *Bd*GPL. Curiously, significant variation in virulence was also identified between *Bd*GPL isolates, suggesting intra-lineage genetic variation exists - perhaps as it becomes endemic in

these locations (such as with the avirulence hypothesis) or as it becomes increasingly virulent (as in the short-sighted hypothesis). Alternatively, this variation arises from different host response within the species, as the experimental design we used involved tadpoles from multiple matrilines (and probably patrilines). Future studies that are able to determine the virulence-profiles in conjunction with the dates of introduction for multiple *Bd* isolates may be able to predict *Bd*s evolutionary trajectory.

To identify the genetic underpinnings of virulence in *Bd*, a number of studies have begun to examine gene-family expansions and gene homologies to other known virulence factors. For example, a large expansion of carbohydrate-binding module (CBM) proteins have been identified in *Bd* relative to the closely related, but non-pathogenic chytrid *Homolaphlyctis polyrhiza* [Abramyan J *et al.* 2012]. CBMs are thought to facilitate attachment of zoospores to keratin, and are known virulence factors in other dermatophytes [Jousson O *et al.*, 2004; Burmester A *et al.*, 2011], which therefore makes them good candidates in the evolution from presumed free-living saprobe to pathogen (coincidental virulence hypothesis) [Joneson S *et al.* 2011]. Other gene-family expansions include the fungalysin metallopeptidase and serine proteases, which are both candidates for *Bd*s ability to degrade of host cellular components. Many of these proteases have also been shown to have differential expression between zoospores and sporangia life-stages [Rosenblum EB *et al.*, 2008], again suggesting they harbour a functional role.

Supporting the evidence for proteases and CBMs as likely candidate virulence factors, we identified evidence for selection across *Bd*s secretome in terms of significant enrichments for heterozygous and non-synonymous polymorphisms and raised omega values since *Bd*GPL divergence with the other divergent lineages (Chapter 4). It is these secreted proteins that are most likely to interact with the host. In contrast, proteases and CBMs lacking a secretion signal were also found to be lacking enrichment for these polymorphisms.

A separate large family of *Bd* genes that share homology with the C-terminal of the Crinkler (Crn) family of effector proteins found in the *Phytophthora genus* [Joneson S *et al.*, 2011, Sun G *et al.*, 2011] have also recently gained attention. It has even been suggested that *Bd* acquired Crn's via horizontal gene transfer (HGT), which was responsible for the onset of pathogenicity [Sun G *et al.*, 2011]. Addressing the first of these hypotheses (HGT), the authors failed to look for Crn's in other fungal species, such as in the chytrid *Spizellomyces punctatus*, which also has at least one Crn (identifiable homology using a blastx search of the nr database) - suggesting that *Bd*'s Crn genes may have an older evolutionary history

amongst the chytridiomycetes. Furthermore, the suggestion that Crn's are functional effector genes in *Bd* is contentious for at least two reasons. Firstly, in *Bd* they lack the secretion signal essential for host-translocation in the *Phytophthora* genus. Secondly, while their function has been shown as an effector gene in plants, there is no evidence, nor does it seem intuitive, that they should have a shared common target between plant and amphibian, especially given the short time frame in which their acquisition is purported. Therefore, given the current evidence, both HGT and their role as virulence factors are both lacking evidence and require substantial future work.

In chapter 4, we found a significant enrichment of recombination occurring across the Crn group of genes in *Bd*GPL and between lineages. Given that Crn's manifest diverse carboxy-terminal domains in *Phytophthora sp.*, it has been suggested that the mosaic domains of Crn's are being shuffled by recombination [Raffaele S *et al.*, 2012]. Perhaps, Crn's promote homologous recombination, and this is a shared function of these genes. In addition, we also identified the Crn's to be significantly enriched for both homozygous and heterozygous polymorphisms relative to the whole set of genes. They were also shown to have the greatest median, upper quartile and upper tail values of omega of any of our gene groups in both *Bd*CAPE and *Bd*CH. Recently, a number of CRN's have been shown to be highly expressed on host tissue in vitro [Rosenblum EB *et al.*, 2012]. Therefore, whilst these genes lack a secretion signal in *Bd*, their large gene-family size, patterns of genetic variation and expression, and similarities with oomycete CRN's, all strongly suggest that they are functional in *Bd*. However, as with each the other putative virulence factors, future work involving further experimental and computational characterisation is required to resolve what that function is.

Indeed, the validity of each of our *in silico* analyses would hugely benefit from further wet-lab experiments (such as the experimental assessment of host response to *Bd* isolates as described in section 3.3.6). Other complementary wet-lab experiments that were able to demonstrate gain of function in an otherwise non-pathogenic organism expressing certain *Bd* genes would support the role of those genes as virulence factors. Should gene silencing or knockout techniques be used in *Bd*, and show a measurable alteration to virulence, especially if this change could then be reversed, would provide convincing support for the function of those genes.

Another interesting experiment would be to use fluorescent microscopy (such as with GFP-tagged proteins) to show how virulence factors are translocated into the host, their

timing during infection, and whether they target the nucleus (such as the virulence factor lipophosphoglycan produced by the protozoan genus *Leishmania*, which impair the nuclear translocation of NF-κB in monocytes [Argueta-Donohué *et al.*, 1998]), the cytoplasm (such as the small proteins PagK1, PagK2 and PagJ produced by *Salmonella enterica sv. Typhimurium*, which manipulate host cellular activities and promoting bacterial proliferation [Yoon H *et al.*, 2011]), or other cell compartments, and thereby help distinguish their function.

5.3.2 Antimicrobial peptide (AMP) and drug resistance

Drug resistance amongst pathogens presents an increasingly important challenge for controlling and mitigating their effects. Drug resistance is found across a large number of bacterial pathogens including Methicillin-resistant *Staphylococcus aureus* (MRSA) and Multidrug Resistant *Mycobacterium tuberculosis*. Fungi under the selection imposed by polypeptides that exert antimicrobial activity (AMPs) or antifungal drugs[2] also frequently evolve resistance [Anderson JB *et al.*, 2005]. While AMPs and antifungal drugs (such as azoles[3]) can be very different in their chemical properties, many have the same targets, such as inhibiting the enzyme lanosterol 14 α-demethylase[4] ([Grimaldi M *et al.*, 2010] and [Sanglard D *et al.*, 1999] respectively). Furthermore, some AMPs carry the imidazole ring characteristic of azole compounds [Grimaldi M *et al.*, 2010]. Other antifungal agents interact with fungal membrane sterols post- synthesis, and a 3^{rd} category inhibits macromolecular synthesis [Ghannoum MA *et al.*, 1999].

A common mechanism for drug resistance in fungi, such as *Candida* isolates to azoles, is the over-expression of plasma membrane efflux pumps belonging to the ATP-binding cassette (ABC) class of transporters [Sanglard D *et al.*, 1999], which tend to have broad specificity to substrates, and the major facilitator superfamily (MFS) class of transporters, which tend to have a limited range of substrates [Cannon RD *et al.*, 2009]. Aneuploidy has been shown to result in the over-expression of genes on affected chromosomes - including efflux pumps,

[2]For the purpose of this subsection, I will consider antimicrobial drugs as a distinct class of agents consisting of non-protein based synthesised agents, whilst AMPs will be considered as a class of proteins with antimicrobial properties.

[3]Azoles are a group of cyclic compounds with 5 atoms in its ring and 2 double bonds. The ring contains carbon, 1 or 2 nitrogen atoms, and some also contain a sulphur or oxygen atom. Azoles are therefore a class of 5-membered heterocyclic compound.

[4]lanosterol 14 α-demethylase used to convert lanosterol to ergosterol (a component of the fungal cell membranes where it serves to maintain the plasma membrane structure and function)

and thereby conferring resistance to antifungal drugs (including azole stress) and fitness advantages in a number of pathogenic fungi such as *Candida sp.* and *Cryptococcus neoformans* [Kwon-Chung KJ *et al.*, 2012]. Other mechanisms that may contribute to the development of resistance to antifungal agents include alteration to the drug target by mutation, or down-regulation of that target enzyme [Ghannoum MA *et al.*, 1999].

Over the course of this thesis, we have tested *Bd*'s response to both antifungal drugs (Chapter 3) and AMPs (Chapter 4). For example, *Bd* infections of captive amphibians are commonly treated with 5-10 minute baths containing c. 0.0025% - 0.01% Itraconazole[5] each day for a week [Tamukai K *et al.*, 2011; Brannelly LA *et al.*, 2012]. To test if susceptibility to Itraconazole varied amongst the isolates and lineages, we tested 6 concentrations on 12 isolates including all three lineages (Chapter 3). However, we found no significant difference between the MC_{50} values between *Bd*GPL and *Bd*CAPE.

In chapter 4, we report on the passaging of *Bd*CH isolate 0739 for 40 generations with or without selection by 100μg ml^{-1} skin AMPs collected from the water frog *Pelophylax esculentus*, which gave rise to the isolates APEP and CON2A, respectively. APEP was found to show significantly greater tolerance for those AMPs, whilst ACON showed similar levels of susceptibility - suggesting that APEP had developed this phenotype through this *in vitro* divergence experiment. Specifically, APEP had 14.23 $\pm$ 11.48 Mean growth inhibition (%$\pm$ SE) compared with 59.44 $\pm$ 14.50 for *Bd*CH CON2A and 55.74 $\pm$ 5.98 for *Bd*GPL.

To study how the differences between *Bd*CH isolates APEP and CON2A arose, we included these isolates for sequencing in Chapter 4, and compared their chromosome content and mutations across the genome. From this analysis, we identified changes in chromosome copy number over supercontig 4 and 5 in CON2A (decrease and increase respectively) and an increase in the chromosome copy number of supercontigs 5 for APEP. It is interesting that the same supercontigs should undergo CCNV in the two isolates, perhaps suggesting a common mechanism driving instability over this supercontigs amongst *Bd*CH isolates. Indeed, general patterns of CCNV were seen amongst the three lineages, with a common increase throughout chromosomes 2 and 3 in *Bd*GPL for example. However, whilst an increase was identified across both isolates, the normalised mean depth over this supercontig in APEP was 1.22, while for CON2A it was 1.47, a difference of 0.25. Although I opted for a cautious estimate in this thesis, whereby each significant shift in depth was considered a single +/- ploidy

[5]Itraconazole is a triazole containing a five-membered ring of two carbon atoms and three nitrogen atoms. Like the azoles, it inhibits the production ergosterols used in the plasma membrane of fungi.

level, these putative ploidies may prove to be an understatement. Determining the exact functions aneuploidy plays in drug resistance, in addition to other aspects of *Bd*'s biology and virulence, will therefore require additional future experiments to characterise and verify changes using such methods as qPCR.

Aneuploidy, as a general feature, was identified in the negative controls for AMP resistance, and across much of the rest of the panel of sequenced isolates in chapter 4, suggesting it does not occur (at least solely) as an adaptive response to AMPs. Instead, the changes from ACON to APEP and CON2A offer an insight into the rapidity of change amongst *Bd* isolates, regardless of apparent selection pressures.

A clearer difference between APEP and either ACON and CON2A was identified from the SNP-calls, whereby APEP had a much higher number of homozygous SNPs and double the number of new heterozygous positions (gain of heterozygosity) compared with those lost (loss of heterozygosity; LOH). In contrast, ACON $\rightarrow$ CON2A had fewer of every type of polymorphism (except for a single additional change from heterozygous to a different heterozygous base - double mutation). Acquisition of newly acquired heterozygous positions was only marginally (2%) higher than LOH.

Discerning which changes might be responsible for the observed differences between growth inhibition from AMP treatment remains a task for future research. Two possible ways to identify the genetic elements responsible (as used in chapters 3 and 4) are the Yang and Nielsen method for measuring dN/dS ratios [Yang Z *et al.*, 2000] as well as the direction of selection (DoS) measure [Stoletzki N *et al.*, 2011], in conjunction with Gene ontology (GO) mapping. For the purpose of identifying the mechanism of AMP resistance, it would also be useful to focus on candidate genes and their associated TFBS, such as ABC and MFS transporters, both of which have been previously implicated in antifungal resistance amongst other pathogenic fungi [Sanglard D *et al.*, 1999; Cannon RD *et al.*, 2009].

5.3.3 Aneuploidy and its association with virulence in *Bd*

In recent years, it has become increasingly apparent that pathogenic fungi manifest highly plastic genome architecture in the form of variable numbers of individual chromosomes, known as chromosomal copy-number variation (CCNV) or aneuploidy. This feature has been identified across the Ascomycota phylum, ranging from *Botrytis cinerea* [Büttner P *et al.*, 1994], *Histoplasma capsulatum* [Carr J *et al.*, 1998], *Saccharomyces cerevisiae* [Sheltzer JM

et al., 2011], *Candida albicans* [Abbey D *et al.*, 2011] and also in *Cryptococcus neoformans* of the Basidiomycota phylum [Lengeler KB *et al.*, 2001; Hu G *et al.*, 2011; Sionov E *et al.*, 2010]. The mechanism(s) generating aneuploidy in fungi is thought to occur as a consequence of chromosome non-disjunction followed by either meiotic or mitotic segregation [Reedy JL *et al.*, 2009], followed by selection operating to stabilise chromosome numbers [Hu G *et al.*, 2011]. Host response or exposure to antifungal agents (section 5.3.2) has also been linked to aneuploidy in *C. albicans* [Forche A *et al.*, 2009]. The generation of dynamic numbers of chromosomes offers potentially advantageous phenotypic changes *via* over expression of virulence-factors [Hu G *et al.*, 2011] or efflux pumps [Kwon-Chung KJ *et al.*, 2012], the maintenance of diversity through homologous recombination [Forche A *et al.*, 2008], increased rates of mutation and larger effective population sizes [Arnold B *et al.*, 2012], or by purging deleterious mutations through non-disjunction during chromosomal segregation [Schoustra SE *et al.*, 2007].

As briefly described in the previous subsection (section 5.3.2), a striking feature of the *Bd* genome is the high rates of aneuploidy that we found across all three lineages (Chapter 4). Using statistical tests with both the non-overlapping sliding windows of the mean depth and allele-frequencies, we found evidence for 18/22 sequenced isolates exhibiting aneuploidy, 11/22 isolates that were predominantly diploid, 7/22 isolates that were predominantly triploid and 4/22 isolates that were predominantly tetraploid. From this analysis, only 4/22 of isolates appeared to be entirely diploid across their genome, as was previously thought the case for *Bd* [James TY *et al.*, 2009]. Furthermore, our estimate of aneuploidy is still arguably cautious, as each significant shift in depth was considered a single +/- ploidy level. However, normalised mean depth (NMD) increased at different intervals between isolates. For example, supercontigs 5 in isolate APEP had a NMD of 1.22, while isolate CON2A had a NMD 1.47, a difference of 0.25 - however, these were both interpreted as tetraploid supercontigs against triploid baseline. Separately, these isolates were separated by only 40 passages in culture, demonstrating the rapidity of change in the genomic content of isolates.

Given the presence of aneuploidy across all three lineages of *Bd* would suggest it is a general feature of *Bd*, rather than just a certain group or lineage of *Bd*. However, whether it is occurring in free-living isolates in addition to those we sequenced - which have all been passaged at least twice before sequencing, and likely to have been in contact with antibiotics when first isolated in order to avoid contamination. Therefore, these isolates may have undergone both bottlenecking and stresses (or lack of host-stresses) that may affect the tempo or patterns of aneuploidy (as a deleterious anomaly rather than a functional

feature of the genome). However, given that we have found no correlation between ploidy or CCNV with time in culture, would suggest that the phenomonen is not a result of being in culture. A separate hypothesis is that cultured isolates have undergone greater or even reduced levels of aneuploidy than would occur in the host - thereby hindering our attempts to identify signatures of selection and adaptation from those patterns. To date, whole-genome sequencing of *Bd in vivo* has not yet been achieved due to difficulties separating fungal and frog DNA. However, such an experiment may shed light on whether aneuploidy is occurring within its naturally-occurring parasitic form.

We have so far only described the general patterns of aneuploidy in *Bd*. In other fungi, the resulting over-expression has been described, and the effect it has on individual genes attributed to phenotype. For example, In *C. albicans*, the genes ERG11 and TAC1 encode a transcription regulator of drug efflux pumps. Amplification of these two genes was determined to be the major mechanism responsible for increased azole resistance in i(5L) aneuploids [Kwon-Chung KJ *et al.*, 2012][6]. A separate study performed a meta-analysis on gene expression from aneuploid cells in diverse organisms, including yeast, plants, mice, and humans found highly related gene expression patterns that are conserved between species [Sheltzer JM *et al.*, 2012]. Interestingly, this pattern consisted of both up-regulation of genes involved in the response to stress, and down-regulation of genes associated with the cell cycle and proliferation [Sheltzer JM *et al.*, 2012], suggesting similar or conserved cellular pathways for dealing with aneuploids.

In *Bd*, future work is first required to verify both the haploid number of chromosomes and their corresponding sequence in *Bd* (rather than the current 69 supercontigs of JEL423), and the chromosomes exhibiting aneuploidy to allow greater characterisation of how it affects individual genes. Another follow-up experiment will be to sequence the transcriptome (or part of) for *Bd*, which will help clarify which genes are functional, likely reveal new genes and splice-variants, and explain how aneuploidy affects the transcription profile of the fungus which itself may reveal why it undergoes such a process. Finally, as chapter 4 is the first whole-genome population study of any chytrid, it would be interesting to see whether aneuploidy is present in either the other lineages of *Bd* (i.e. *Bd*JAPAN [Goka K *et al.*, 2009] and *Bd*Brazil [Schloegel LM *et al.*, 2012]), other pathogenic chytrids (e.g. *Synchytrium*

[6] i(5L) corresponds to the primarily trisomic chromosome 5 in *C. albicans*, which also exhibited a high frequency of segmental aneuploidy comprising the two left arms of Chr5 flanked by a single centromere. Sometimes, i(5L) was attached to a homolog of Chr5. The left arm of Chr5 houses the ERG11 and TAC1 genes, therefore each of these additional homologs offer further increases in drug resistance.

endobioticum causing potato wart disease), as well as other non-pathogenic chytrids (e.g. the free-living saprobe *Homolaphlyctis polyrhiza*). While I expect the other lineages of *Bd* to also exhibit aneuploidy, it is less certain whether it will turn out to be a general feature of chytrids, chytrid pathogens, or whether it is a unique feature of *Bd* amongst the Chytridiomycota.

5.4 The life cycle of *Bd*

An important aspect of the biology and often the pathogenicity in fungi are their ability to utilise multiple reproductive modes, including cryptic recombination that enables inbreeding, outcrossing, hybridization, and the generation of diversity via parasexual or sexual mechanisms [Heitman J, 2010]. It is known that some species belonging to the class chytridiomyctes reproduce by sex, such as the some species belonging to the order Chytridiales (including the species *Chytriomyces hyalinus* [Moore ED *et al.*, 1973] and *Siphonaria petersenii* [Blackwell WH *et al.*, 2012]) and has been reported in the order Rhizophydiales to which *Bd* belongs [Sparrow FK, 1960], although this one study from 1960 has not since been followed up.

The first population-wide studies of *Bd* [Morgan JAT *et al.*, 2007; Morehouse *et al.*, 2003] identified *Bd* populations consisting of highly clonal genotypes (including shared heterozygous positions [Morehouse *et al.*, 2003]). Later studies confirmed numerous shared heterozygous loci [James TY *et al.*, 2009], again suggesting that *Bd* was undergoing a clonal expansion. However, in addition to shared heterozygous loci, Morgan JAT *et al.* [Morgan JAT *et al.*, 2009] also detected diversity in two *Bd*GPL populations in the Sierra Nevada of California suggesting segregation, not clonality - the first such evidence for recombination in *Bd*. More recently, a number of fungi have been found to have infrequent or parasexual cycles [Heitman J, 2010], thereby offering a solution to these contrasted conclusions. Furthermore, shared heterozygous loci have also been attributed to sexually reproducing fungal populations with spatially or genetically close distances [Henk DA *et al.*, 2012].

Over the course of my project, we found support for recombination amongst *Bd* isolates by the discovery of phylogenetic breakpoints (Chapter 3), observations of patchily-distributed heterozygosity, and later, directly using shared phased heterozygous positions (Chapter 4). Although each of these findings could be partly explained by non-recombination mechanisms, such as mutational hotspots, or through double mutations at those loci, it is unlikely this should account for so many sites. However, these discoveries could result from either mi-

totic recombination of homologous chromosomes within a single individual (such as through gene conversion), by meiosis within a single individual, or between two parental isolates (as depicted in Fig. 3.16), or a combination of these.

Supporting the hypothesis that *Bd* isolates are able to undergo meioses was the discovery of a hybrid genotype between *Bd*GPL and *Bd*Brazil [Schloegel LM *et al.*, 2012], which was found using the software New Hybrids [Anderson EC *et al.* 2002] and PCA analysis. These findings suggest either a cryptic or rare sexual cycle and/or a mechanism of driving mitotic recombination resulting in loss of heterozygosity and gene conversion. Using population genetic analysis, we found the emergent *Bd*GPL is far more clonal than the other naturally occurring lineage, *Bd*CAPE (Chapter 4), suggesting that the life cycle of *Bd* may differ depending on its lineage. Whether this difference is caused by differences in genotype, environmental constraints or phenotypic plasticity, and the extent recombination contributes to *Bd*s phenotypic profile each remains to be tested.

If *Bd* is able to undergo meiosis as suggested by hybrid genotypes, then determining the processes that enables exchange of genetic material (such as pseudo-hyphal fusion) and environmental stimuli promoting genetic exchange (such as within particular host species) would also provide additional information on the life-cycle and biology of *Bd* and possibly new features that could be targeted. Until sexual structures are discovered by microscopy, or by determining segregation from crosses (e.g. as used for *S. Cerevisiae* [Ross LO *et al.*, 1996] and *C. albicans* [Forche A *et al.*, 2008]), the current genetic studies are arguably insufficient to prove that *Bd* undergoes a sexual cycle.

Recent work with microscopy in *Bd* has identified zoospore cysts germinating short tubular structures called germ tubes able to penetrate the stratum corneum [Van Rooij P *et al.*, 2012], confirming *Bd* as an intracellular pathogen. One possible explanation for the lack of a known sexual cycle in *Bd* is that the structures and signals promoting meiosis (e.g. *Schizosaccharomyces pombe* (fission yeast) reacting to nitrogen starvation [Shiozaki K *et al.* 1996; Sukegawa Y *et al.*, 2011]) are produced within the host (and as of yet, not been experimentally replicated in cell culture). Furthermore, a large proportion of genes in the *Bd* genome show increased expression when grown in host tissue compared with standard mTGhL [Rosenblum EB *et al.*, 2012], suggesting aspects of its life-cycle could also be different when experiencing conditions closer to *in vivo*.

In addition to heterothallic mating (opposite sexes), homothallic (same-sex) mating has been reported to occur in other fungi such as *C. neoformans* and *C. albicans* [Alby K *et*

al., 2009; Lin X *et al.*, 2005; Fraser JA *et al.*, 2005]. Indeed, in *C. neoformans* it appears to be the predominant mode of reproduction as the majority of natural isolates are of the α-mating type [Heitman, 2006]. In either case, there is a requirement for at least a single mating-type locus that has so far not been identified in *Bd*.

In *C. albicans*, mating results in a tetraploid a/α cell that must undergo a reductional division to return to the diploid state [Forche A *et al.*, 2008]. Mating between divergent *C. neoformans* strains belonging to serotype A and D led to diploid or aneuploid progeny, which would normally undergo meiosis given less genetic divergence [Lin X *et al.*, 2007]. Therefore, aneuploidy has been attributed to both a successful round of meiosis and a barrier to meiosis in these two model organisms. The way in which aneuploidy (as presented and discussed in chapter 4 and section 5.3.3) affects the life-cycle of *Bd* and the mechanism(s) driving aneuploidy are currently unknown, but future work detecting this feature in both parental and hybrid genotypes should reveal clues to its purpose and impact.

Bibliography

[1000 Genomes Project Consortium *et al.*, 2010] 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population- scale sequencing. *Nature* **467**:1061-73 (2010).

[Aanensen DM *et al.*, 2009] Aanensen DM, Huntley DM, Feil EJ *et al.* EpiCollect: linking smart phones to Web applications for epidemiology, ecology and community data collection. *PLoS One* **4**:e6968 (2009).

[Abbey D *et al.*, 2011] Abbey D, Hickman M, Gresham D, Berman J. High-resolution SNP/CGH microarrays reveal the accumulation of loss of heterozygosity in commonly used *Candida albicans* strains. *G3: Genes Genomes Genetics* **1**:523-30 (2011).

[Abramoff MD *et al.*, 2004] Abramoff MD, Magelhães PJ, Ram SJ. Image Processing with ImageJ. *Biophotonics International* **11**:36-42 (2004).

[Abramyan J *et al.*, 2012] Abramyan J, Stajich JE. Species-specific chitin-binding module 18 expansion in the amphibian pathogen *Batrachochytrium dendrobatidis*. *mBio* **3**:e00150-12 (2012).

[Adiba S *et al.*, 2010] Adiba S, Nizak C, van Baalen M, Denamur E, Depaulis F. From grazing resistance to pathogenesis: the coincidental evolution of virulence factors. *PLoS One* **5**:e11882 (2010).

[Agapow PM *et al.*, 2001] Agapow PM, Burt A. Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**:101-102 (2001).

[Aird D *et al.*, 2011] Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**:R18 (2011).

[Alby K *et al.*, 2009] Alby K, Schaefer D, Bennett RJ. Homothallic and heterothallic mating in the opportunistic pathogen *Candida albicans*. *Nature* **460**:890-93 (2009).

[Ali MF *et al.*, 2001] Ali MF, Soto A, Knoop FC, Conlon JM. Antimicrobial peptides isolated from skin secretions of the diploid frog, *Xenopus tropicalis* (Pipidae). *Biochim. Biophys. Acta.* **1550**:81-89 (2001).

[Alizon S *et al.*, 2009] Alizon S, Hurford A, Mideo N, Van Baalen M. Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future. *J. Evol. Biol.* **22**:245-59 (2009).

[Alkan C *et al.*, 2011] Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**:61-65 (2011).

[Allen ES, 1934] Allen ES. Review: V. A. Kostitzin, Symbiose, Parasitisme et volution. *Bull. Amer. Math. Soc.* **40**:788 (1934).

[Altschul SF *et al.*, 1990] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* **215**:403-10 (1990).

[Anderson EC *et al.*, 2002] Anderson EC, Thompson EA. A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**:1217-1229 (2002).

[Anderson JB *et al.*, 2005] Anderson JB. Evolution of antifungal-drug resistance mechanisms and pathogen fitness. *Nat. Rev. Microbiol.* **3**:547-56 (2005).

[Anderson RM *et al.*, 1982] Anderson RM, May RM. Coevolution of hosts and parasites. *Parasitology* **85**:411-426 (1982).

[Andre SE *et al.*, 2008] Andre SE, Parker J, Briggs CJ. Effect of temperature on host response to *Batrachochytrium dendrobatidis* infection in the mountain yellow-legged frog (*Rana muscosa*). *J. Wildl. Dis.* **44**:716-720 (2008).

[Andreone F *et al.*, 2008] Andreone F, Carpenter AI, Cox N, du Preez L, Freeman K, Furrer S, Garcia G, Glaw F, Glos J, Knox D, Köhler J, Mendelson JR 3rd, Mercurio V, Mittermeier RA, Moore RD, Rabibisoa NH, Randriamahazo H, Randrianasolo H, Raminosoa NR, Ramilijaona OR, Raxworthy CJ, Vallan D, Vences M, Vieites DR, Weldon C. The challenge of conserving amphibian megadiversity in Madagascar. *PLoS. Biol.* **6**:e118 (2008).

[Argueta-Donohué J *et al.*, 1998] Argueta-Donohué J, Carrillo N, Valdés-Reyes L, Zentella A, Aguirre-García M, Becker I, Gutiérrez-Kobeh L: *Leishmania mexicana*: participation of NF-kappaB in the differential production of IL-12 in dendritic cells and monocytes induced by lipophosphoglycan (LPG). *Exp. Parasitol.* **120**:1-9 (1998).

[Arnold B *et al.*, 2012] Arnold B, Bomblies K, Wakeley J. Extending coalescent theory to autotetraploids. *Genetics* **192**:195-204 (2012).

[Ball GH, 1943] Ball GH. Parasitism and evolution. *Am. Nat.* **77**:345-364 (1943).

[Barr DJS, 1973] Barr DJS. *Rhizophydium graminis* (Chytridiales): morphology, host range, and temperature effect. *Can. Plant Dis. Survey* **53**:191-193 (1973).

[Barribeau SM *et al.*, 2008] Barribeau SM, Villinger J, Waldman B. Major histocompatibility complex based resistance to a common bacterial pathogen of amphibians. *PLoS ONE* **3**:e2692 (2008).

[*Bd*-maps] *Bd*-maps and the *Bd* global mapping project. *Bd*-maps and the *Bd* global mapping project is a collaboration between Aanensen D, Fisher MC, Powell C, and Spratt B (Imperial College London, London, UK), and Olson D (US forest service). Available at http://www.bd-maps.net.

[Belle EMS *et al.*, 2005] Belle EMS, Piganeau G, Gardner M, Eyre-Walker A. An investigation of the variation in the transition bias among various animal mitochondrial DNA *Gene* **355**:58-66 (2005).

[Bendtsen JD *et al.*, 2004] Bendtsen JD, Nielsen H, Heijne GV, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J. *Mol. Biol.* **340**:783-795 (2004).

[Berger L *et al.*, 1998] Berger L, Speare R, Daszak P, Green DE, Cunningham AA, Goggin CL, Slocombe R, Ragan MA, Hyatt AD, McDonald KR, Hines HB, Lips KR, Marantelli G, Parkes H. Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and central America. *Proc. Natl. Acad. Sci. U.S.A.* **95**: 9031-6 (1998).

[Berger L *et al.*, 1999] Berger L, Speare R, and Hyatt A. Chytrid fungi and amphibian declines: overview, implications and future directions. Pages 23-33 in A. Campbell, editor. Declines and Disappearances of Australian Frogs. Environment Australia, Canberra (1999).

[Berger L *et al.*, 2004] Berger L, Speare R, Hines HB, Marantelli G, Hyatt AD, McDonald KR, Skerratt LF, Olsen V, Clarke JM, Gillespie G, Mahony M, Sheppard N, Williams C, Tyler MJ. Effect of season and temperature on mortality in amphibians due to chytridiomycosis. *Aust. Vet. J.* **82**:434-9 (2004).

[Berger L *et al.*, 2005] Berger L, Marantelli G, Skerratt LF, Speare R. Virulence of the amphibian chytrid fungus *Batrachochytium dendrobatidis* varies with the strain. *Dis. Aquat. Org.* **68**:47-50 (2005).

[Berger L *et al.*, 2005] Berger L, Hyatt AD, Speare R, Longcore JE. Life cycle stages of the amphibian chytrid *Batrachochytrium dendrobatidis*. *Dis. Aquat. Org.* **68**:51-63 (2005).

[Berger L *et al.*, 2009] Berger L, Speare R, Marantelli G, Skerratt LF. A zoospore inhibition technique to evaluate the activity of antifungal compounds against *Batrachochytrium dendrobatidis* and unsuccessful treatment of experimentally infected green tree frogs (*Litoria caerulea*) by fluconazole and benzalkonium chloride. *Res. Vet. Sci.* **87**:106-10 (2009).

[Blackwell WH *et al.*, 2012] Blackwell WH, Letcher PM, Powell MJ. Synopsis of Obelidium (Chytridiomycota). *Phytologia* **94**:103-117 (2012).

[Booth LN *et al.*, 2010] Booth LN, Tuch BB, Johnson AD. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* **468**:959-63 (2010).

[Bos JI *et al.*, 2006] Bos JI, Kanneganti TD, Young C, Cakir C, Huitema E, Win J, Armstrong MR, Birch PR, Kamoun S. The C-terminal half of *Phytophthora infestans* RXLR effector AVR3a is sufficient to trigger R3a-mediated hypersensitivity and suppress INF1-induced cell death in *Nicotiana benthamiana*. *Plant J.* **48**:165-76 (2006).

[Boyle DG *et al.*, 2003] Boyle DG, Hyatt AD, Daszak P, Berger L, Longcore JE, Porter D, Hengstberger SG, Olsen V. Cryo-archiving of *Batrachochytrium dendrobatidis* and other chytridiomycetes. *Dis. Aquat. Org.* **56**:59-64 (2003).

[Boyle DG *et al.*, 2004] Boyle DG, Boyle DB, Olsen V, Morgan JA, Hyatt AD. Rapid quantitative detection of Chytridiomycosis (*Batrachochytrium dendrobatidis*) in amphibian samples using real-time Taqman PCR assay. *Dis. Aquat. Org.* **60**:141-8 (2004).

[Brannelly LA *et al.*, 2012] Brannelly LA, Richards-Zawacki CL, Pessier AP. Clinical trials with itraconazole as a treatment for chytrid fungal infections in amphibians. *Dis. Aquat. Organ.*

10:95-104. (2012).

[Bray N *et al.*, 2004] Bray N, Pachter L. Mavid: constrained ancestral alignment of multiple sequences. *Genome Res.* **14**:693-9 (2004).

[Briggs CJ *et al.*, 2010] Briggs CJ, Knapp RA, Vredenburg VT. Enzootic and epizootic of the chytrid fungal pathogen of amphibians. *Proc. Natl. Acad. Sci. U.S.A.* **107**:9695-9700 (2010).

[Brown AHD *et al.*, 1980] Brown AHD, Feldman MW, Nevo E. Multilocus structure of natural populations of Hordeum spontaneum. *Genetics* **96**:523-536 (1980).

[Burmester A *et al.*, 2011] Burmester A, Shelest E, Glockner G, Heddergott C, Schindler S, Staib P, Heidel A, Felder M, Petzold A, Szafranski K, Feuermann M, Pedruzzi I, Priebe S, Groth M, Winkler R, Li W, Kniemeyer O, Schroeckh V, Hertweck C, Hube B, White TC, Platzer M, Guthke R, Heitman J, Wöstemeyer J, Zipfel PF, Monod M, Brakhage AA. Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. *Genome Biol.* **12**:R7 (2011).

[Burrows M *et al.*, 1994] Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical report 124. Palo Alto, CA: Digital Equipment Corporation (1994).

[Büttner P *et al.*, 1994] Büttner P, Koch F, Voigt K, Quidde T, Risch S, Blaich R, Brückner B, Tudzynski P. Variations in ploidy among isolates of *Botrytis cinerea*: implications for genetic and molecular analyses. *Curr. Genet.* **25**:445-50 (1994).

[Byrnes EJ 3rd *et al.*, 2010] Byrnes EJ 3rd, Li W, Lewit Y, Ma H, Voelz K, Ren P, Carter DA, Chaturvedi V, Bildfell RJ, May RC, Heitman J. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. *PLoS Pathog.* **6**:e1000850 (2010).

[Calboli FCF *et al.*, 2011] Calboli FCF, Fisher MC, Garner TWJ, Jehle R. The need for jumpstarting amphibian genome projects, *Trends Ecol. Evol.* **26**:378-9 (2011).

[Cannon RD *et al.*, 2009] Cannon RD, Lamping E, Holmes AR, Niimi K, Baret PV, Keniya MV, Tanabe K, Niimi M, Goffeau A, Monk BC. Efflux-mediated antifungal drug resistance. *Clin. Microbiol. Rev.* **22**:291-321 (2009).

[Carey C *et al.*, 2006] Carey C, Bruzgul JE, Livo LJ, Walling ML, Kuehl KA, Dixon BF, Pessier AP, Alford RA, Rogers KB. Experimental exposures of boreal toads (*Bufo boreas*) to a pathogenic chytrid fungus (*Batrachochytrium dendrobatidis*). *EcoHealth* **3**:5-21 (2006).

[Carr J *et al.*, 1998] Carr J, Shearer Jr G. Genome size, complexity, and ploidy of the pathogenic fungus *Histoplasma capsulatum*. *J. Bacteriol.* **180**:6697-6703 (1998).

[Casillas AM *et al.*, 2003] Casillas AM, Nyamathi AM, Sosa A, Wilder CL, Sands H. A current review of Ebola virus: pathogenesis, clinical presentation, and diagnostic assessment. *Biol. Res. Nurs.* **4**:268-75 (2003).

[Conesa A *et al.*, 2005] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.

Bioinformatics **21**: 3674-6 (2005).

[Cooke DEL *et al.*, 2012] Cooke DEL, Cano LM, Raffaele S, Bain RA, Cooke LR, Etherington GJ, Deahl KL, Farrer RA, Gilroy EM, Goss EM, Grünwald NJ, Hein I, MacLean D, McNicol JW, Randall E, Oliva RF, Pel MA, Shaw DS, Squires JN, Taylor MC, Vleeshouwers VG, Birch PR, Lees AK, Kamoun S. Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS Pathog.* **8**:e1002940 (2012).

[Crawford AJ *et al.*, 2010] Crawford AJ, Lips KR, Bermingham E. Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proc. Natl. Acad. Sci. U. S. A.* **107**:13777-82 (2010).

[Cunningham AA *et al.*, 2005] Cunningham AA, Garner TWJ, Aguilar-Sanchez V, Banks B, Foster J, Sainsbury AW, Perkins M, Walker SF, Hyatt AD, Fisher MC. Emergence of amphibian chytridiomycosis in Britain. *Vet. Rec.* **157**:386-7 (2005).

[Danecek P *et al.*, 2011] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics* **27**:2156-8 (2011).

[Daszak P *et al.*, 1999] Daszak P, Berger L, Cunningham AA, Hyatt AD, Green DE, Speare R. Emerging infectious diseases and amphibian population declines. *Emerging Infect. Dis.* **5**:735-48 (1999).

[Daszak P *et al.*, 2000] Daszak P, Cunningham AA, Hyatt AD. Emerging infectious diseases of wildlife-threats to biodiversity and human health. *Science* **287**:443-9 (2000).

[Daum JM *et al.*, 2012] Daum JM, Davis LR, Bigler L, Woodhams DC. Hybrid advantage in skin peptide immune defenses of water frogs (*Pelophylax esculentus*) at risk from emerging pathogens. *Infect. Genet. Evol.* **12**:18541864 (2012).

[de Roode JC *et al.*, 2005] de Roode JC, Pansini R, Cheesman SJ, Helinski ME, Huijben S, Wargo AR, Bell AS, Chan BH, Walliker D, Read AF. Virulence and competitive ability in genetically diverse malaria infections. *Proc. Natl. Acad. Sci. U.S.A.* **102**:7624 (2005).

[Delcher AL *et al.*, 2002] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**: 2478-83 (2002).

[DePristo MA *et al.*, 2011] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**: 491-498 (2011).

[DiGuistini S *et al.*, 2009] DiGuistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, Mardis E, Marra MA, Hamelin RC, Bohlmann J, Breuil C, Jones SJ. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol.* **10**:R94 (2009).

[Dongen SV, 2000] Dongen SV. A cluster algorithm for graphs, Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, (2000).

[Drummond AJ *et al.*, 2007] Drummond AJ, Andrew Rambaut. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.***7**:214 (2007).

[Ebersberger I *et al.*, 2002] Ebersberger I, Metzler D, Schwarz C, Pääbo S. Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.* **70**: 1490-7 (2002).

[Engelthaler DM *et al.*, 2011] Engelthaler DM, Chiller T, Schupp JA, Colvin J, Beckstrom-Sternberg SM, Driebe EM, Moses T, Tembe W, Sinari S, Beckstrom-Sternberg JS, Christoforides A, Pearson JV, Carpten J, Keim P, Peterson A, Terashita D, Balajee SA. Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. *Emerg. Infect. Dis.* **17**:227-32 (2011).

[Eshel I *et al.*, 1970] Eshel I, Feldman MW. On the evolutionary effect of recombination. *Theor. Popul. Biol.* **1**:88-100 (1970).

[Etienne KA *et al.*, 2012] Etienne KA, Gillece J, Hilsabeck R, Schupp JM, Colman R, Lockhart SR, Gade L, Thompson EH, Sutton DA, Neblett-Fanfair R, Park BJ, Turabelidze G, Keim P, Brandt ME, Deak E, Engelthaler DM. Whole Genome Sequence Typing to Investigate the *Apophysomyces* outbreak following a tornado in Joplin, Missouri, 2011. *PLoS ONE* **7**: e49989 (2012).

[Farrer RA *et al.*, 2009] Farrer RA, Kemen E, Jones JD, Studholme DJ. *De novo* assembly of the *Pseudomonas syringae pv. syringae b728a* genome using illumina/solexa short sequence reads. *FEMS Microbiol. Lett.* **291**: 103-111 (2009).

[Farrer RA *et al.*, 2011] Farrer RA, Weinert LA, Bielby J, Garner TWJ, Balloux F, Clare F, Bosch J, Cunningham AA, Weldon C, du Preez LH, Anderson L, Pond SL, Shahar-Golan R, Henk DA, Fisher MC. Multiple emergence of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent lineage. *Proc. Natl. Acad. Sci. U.S.A.* **108**: 18732-6 (2011).

[Farrer RA *et al.*, 2013] Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC. Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Sci. Rep.* **3**:1512 (2013).

[Fisher MC *et al.*, 2007] Fisher MC, Garner TWJ. The relationship between the emergence of *Batrachochytrium dendrobatidis*, the international trade in amphibians and introduced amphibian species. *Fungal Biol. Rev.* **21**:2-9 (2007).

[Fisher MC, 2009] Fisher MC. Endemic and introduced haplotypes of *Batrachochytrium dendroba tidis* in Japanese amphibians: sink or source? *Mol. Ecol.* **18**:4731-3 (2009).

[Fisher MC *et al.*, 2009] Fisher MC, Bosch J, Yin Z, Stead DA, Walker J, Selway L, Brown AJP, Walker LA, Gow NAR, Stajich JE, Garner TWJ. Proteomic and phenotypic profiling of the

amphibian pathogen *Batrachochytrium dendrobatidis* shows that genotype is linked to virulence. *Mol. Ecol.* **18**: 415-29 (2009).

[Fisher MC *et al.*, 2009] Fisher MC, Garner TWJ, Walker SF. Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. *Annu. Rev. Microbiol.* **63**: 291-310 (2009).

[Fisher MC *et al.*, 2012] Fisher MC, Henk DA, Briggs CJ, Brownstein JSMadoff LC, McCraw SL, Gurr SJ. Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**:186-194 (2012).

[Fisher MC *et al.*, 2012] Fisher MC, Stajich J, Farrer RA. Emergence of the chytrid fungus *Batrachochytrium dendrobatidis* and global amphibian declines. Evolution of Virulence in Eukaryotic Microbes, eds Heitman J, Sibley D, Howlett B (2012).

[Flicek P *et al.*, 2009] Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods* **6**:S6-12 (2009).

[Forche A *et al.*, 2008] Forche A, Alby K, Schaefer D, Johnson AD, Berman J, Bennett RJ. The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biol.* **6**:e110 (2008).

[Forche A *et al.*, 2009] Forche A, Magee PT, Selmecki A, Berman J, May G. Evolution in *Candida albicans* populations during a single passage through a mouse host. *Genetics* **182**:799-811 (2009).

[Fraser C *et al.*, 2007] Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **104**:17441-6 (2007).

[Fraser JA *et al.*, 2005] Fraser JA, Giles SS, Wenink EC, Geunes-Boyer SG, Wright JR, Diezmann S, Allen A, Stajich JE, Dietrich FS, Perfect JR, Heitman J. Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak. *Nature* **437**:1360-4 (2005).

[Freeman KR *et al.*, 2009] Freeman KR, Martin AP, Karki D, Lynch RC, Mitter MS, Meyer AF, Longcore JE, Simmons DR, Schmidt SK. Evidence that chytrids dominate fungal communities in high-elevation soils. *Proc. Natl. Acad. Sci. U. S. A.* **106**:18315-20 (2009).

[Frick WF *et al.*, 2010] Frick WF, Pollock JF, Hicks AC, Langwig KE, Reynolds DS, Turner GG, Butchkoski CM, Kunz TH. An emerging disease causes regional population collapse of a common North American bat species. *Science* **329**:679-82 (2010).

[Frith MC *et al.*, 2010] Frith MC, Hamada M, Horton P. Parameters for accurate genome alignment. *BMC Bioinformatics* **11**:80 (2010).

[Frost DR, 2013] Frost DR. Amphibian Species of the World: an Online Reference. Version 5.6 (9 January 2013). Electronic Database accessible at http://research.amnh.org/herpetology/amphibia/index.html. American Museum of Natural History, New York, USA. (2013).

[Gargas A *et al.*, 2009] Gargas A, Trest MT, Christensen M, Volk TJ, Blehert DS. *Geomyces*

destructans sp. nov. associated with bat white-nose syndrome. *Mycotaxon* **108**:147-54 (2009).

[Garmyn A *et al.*, 2012] Garmyn A, van Rooij P, Pasmans F, Hellebuyck T, Van Den Broeck W, Haesebrouck F, Martel A. Waterfowl: potential environmental reservoirs of the chytrid fungus *Batrachochytrium dendrobatidis*. *PLoS ONE* **7**:e35038 (2012).

[Garner TWJ *et al.*, 2006] Garner TWJ, Perkins MW, Govindarajulu P, Seglie D, Walker S, Cunningham AA, Fisher MC. The emerging amphibian pathogen *Batrachochytrium dendrobatidis* globally infects introduced populations of the North American bullfrog, *Rana catesbeiana*. *Biol. Lett.* **2**:455-9 (2006).

[Garner TWJ *et al.*, 2009] Garner TWJ, Stephen I, Wombwell E, Fisher MC. The amphibian trade: bans or best practice? *Ecohealth* **6**:148-51 (2009).

[Garner TWJ *et al.*, 2009] Garner TWJ, Walker S, Bosch J, Leech S, Rowcliffe JM, Cunningham AA, Fisher MC. Life history tradeoffs influence mortality associated with the amphibian pathogen *Batrachochytrium dendrobatidis*. *Oikos* **118**:783-91 (2009).

[Garner TWJ *et al.*, 2011] Garner TWJ, Rowcliffe JM, Fisher MC. Climate change, chytridiomycosis or condition: an experimental test of amphibian survival. *Glob. Chang. Biol.* **17**:667-675 (2011).

[Geiser DM, 2009] Geiser DM. Sexual structures in *Aspergillus*: morphology, importance and genomics. *Med. Mycol.* **47**:S21-S26 (2009).

[Ghannoum MA *et al.*, 1999] Ghannoum MA, Rice LB. Antifungal Agents: Mode of Action, Mechanisms of Resistance, and Correlation of These Mechanisms with Bacterial Resistance. *Clin. Microbiol. Rev.* **12**:501-517 (1999).

[Gleason FH *et al.*, 2008] Gleason FH, Kagami M, Lefevre E, Sime-Ngando T. The ecology of chytrids in aquatic ecosystems: roles in food web dynamics. *Fungal Biol. Rev.* **22**:17-25 (2008).

[Goka K *et al.*, 2009] Goka K, Yokoyama J, Une Y, Kuroki T, Suzuki K, Nakahara M, Kobayashi A, Inaba S, Mizutani T, Hyatt AD. Amphibian chytridiomycosis in Japan: distribution, haplotypes and possible route of entry into Japan. *Mol. Ecol.* **18**:4757-74 (2009).

[Goss EM *et al.*, 2009] Goss EM, Carbone I, Grünwald NJ. Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen *Phytophthora ramorum*. *Mol. Ecol.* **18**:1161-74 (2009).

[Götz S *et al.*, 2008] Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**:3420-3435 (2008).

[Grimaldi M *et al.*, 2010] Grimaldi M, De Rosa M, Di Marino S, Scrima M, Posteraro B, Sanguinetti M, Fadda G, Soriente A, D'Ursi AM. Synthesis of new antifungal peptides selective against *Cryptococcus neoformans*. *Bioorg. Med. Chem.* **18**:7985-90 (2010).

[Gu M *et al.*, 2011] Gu M, Liu W, Cao Y, Peng D, Wang X, Wan H, Zhao G, Xu Q, Zhang W,

Song Q, Li Y, Liu X. Novel reassortant highly pathogenic avian influenza (H5N5) viruses in domestic ducks, China. *Emerg. Infect. Dis.* **17**:1060-3 (2011).

[Hamilton W, 1980] Hamilton W. Sex versus non-sex versus parasite. *Oikos* **35**:282-290 (1980).

[Harris RN *et al.*, 2009] Harris RN, Brucker RM, Walke JB, Becker MH, Schwantes CR, Flaherty DC, Lam BA, Woodhams DC, Briggs CJ, Vredenburg VT, Minbiole KP. Skin microbes on frogs prevent morbidity and mortality caused by a lethal skin fungus. *ISME J.* 3:818-24 (2009).

[Harris RN *et al.*, 2009 (2)] Harris RN, Lauer A, Simon MA, Banning JL, Alford RA. Addition of antifungal skin bacteria to salamanders ameliorates the effects of chytridiomycosis. *Dis. Aquat. Organ.* 83:11-6 (2009).

[Hasegawa M *et al.*, 1985] Hasegawa M, Kishino H, Yano T. Dating of human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160-174 (1985).

[Hasegawa M *et al.*, 1989] Hasegawa M, Kishino H. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* **43**:672-677 (1989).

[Haubold B *et al.*, 1998] Haubold B, Travisano M, Rainey PB, Hudson RR. Detecting linkage disequilibrium in bacterial populations. *Genetics* **150**:1341-1348 (1998).

[Hayes MP *et al.*, 1986] Hayes MP, Jennings MR. Decline of ranid frogs species in western North America: Are bullfrogs (*Rana catesbiana*) responsible? *J. Herpetol.* **20**:490-509 (1986).

[Heitman J, 2006] Heitman J. Sexual reproduction and the evolution of microbial pathogens. *Curr. Biol.* **16**:R711-R725 (2006).

[Heitman J, 2010] Heitman J. Evolution of eukaryotic microbial pathogens via covert sexual reproduction. *Cell Host & Microbe* **8**:86-99 (2010).

[Henk DA *et al.*, 2012] Henk DA, Shahar-Golan R, Devi KR, Boyce KJ, Zhan N, Fedorova ND, Nierman WC, Hsueh PR, Yuen KY, Sieu TP, Kinh NV, Wertheim H, Baker SG, Day JN, Vanittanakom N, Bignell EM, Andrianopoulos A, Fisher MC. Clonality despite sex: the evolution of host-associated sexual neighborhoods in the pathogenic fungus *Penicillium marnef fei*. *PLoS Pathog.* **8**:e1002851 (2012).

[Hibbett DS *et al.*, 2007] Hibbett DS, Binder M, Bischoff JF, Blackwell M, Cannon PF, Eriksson OE, Huhndorf S, James T, Kirk PM, Lücking R, Thorsten Lumbsch H, Lutzoni F, Matheny PB, McLaughlin DJ, Powell MJ, Redhead S, Schoch CL, Spatafora JW, Stalpers JA, Vilgalys R, Aime MC, Aptroot A, Bauer R, Begerow D, Benny GL, Castlebury LA, Crous PW, Dai YC, Gams W, Geiser DM, Griffith GW, Gueidan C, Hawksworth DL, Hestmark G, Hosaka K, Humber RA, Hyde KD, Ironside JE, Kõljalg U, Kurtzman CP, Larsson KH, Lichtwardt R, Longcore J, Miadlikowska J, Miller A, Moncalvo JM, Mozley-Standridge S, Oberwinkler F, Parmasto E, Reeb V, Rogers JD, Roux C, Ryvarden L, Sampaio JP, Schüssler A, Sugiyama J, Thorn RG, Tibell L, Untereiner WA, Walker C, Wang Z, Weir A, Weiss M, White MM, Winka K, Yao YJ, Zhang N. A higher-level phylogenetic classification of the Fungi. *Mycol.*

Res. **111**:509-47 (2007).

[Hillyard SD *et al.*, 2011] Hillyard SD, Willumsen NJ. Chemosensory function of amphibian skin: integrating epithelial transport, capillary blood flow and behaviour. *Acta. Physiol.* **202**:533-48 (2011).

[Holland BR *et al.*, 2003] Holland BR, Penny D, Hendy MD. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - a simulation study. *Syst. Biol.* **52**:229-238 (2003).

[Houlahan JE *et al.*, 2000] Houlahan JE, Findlay CS, Schmidt BR, Meyer AH, Kuzmin SL. Quantitative evidence for global amphibian population declines. *Nature* **404**:752-5 (2000).

[Hovmøller MS *et al.*, 2010] Hovmøller MS, Walter S, Justesen AF. Escalating threat of wheat rusts. *Science* **329**:369 (2010).

[Hu G *et al.*, 2011] Hu G, Wang J, Choi J, Jung WH, Liu I, Litvintseva AP, Bicanic T, Aurora R, Mitchell TG, Perfect JR, Kronstad JW. Variation in chromosome copy number influences the virulence of *Cryptococcus neoformans* and occurs in isolates from AIDS patients. *BMC Genomics* **12**:526 (2011).

[Hudson R *et al.*, 1985] Hudson R, Kaplan N. Statistical properties of the number of recombination events in the history of a sample of sequences. *Genetics* **111**:147-164 (1985).

[Huse SM *et al.*, 2007] Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**:R143 (2007).

[Inderbitzin P *et al.*, 2011] Inderbitzin P, Davis RM, Bostock RM, Subbarao KV. The ascomycete *Verticillium longisporum* is a hybrid and a plant pathogen with an expanded host range. *PLoS ONE* **6**:e18260 (2011).

[International HapMap Consortium, 2003] International HapMap Consortium. The International HapMap Project. *Nature* **426**:789-96 (2003).

[Jaenike J, 1978] Jaenike J. An hypothesis to account for the maintenance of sex within populations. *Evol. Theor.* **3**:191-4 (1978).

[James TY *et al.*, 2006] James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüssler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R. Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature* **443**: 818-22 (2006).

[James TY *et al.*, 2006] James TY, Letcher PM, Longcore JE, Mozley-Standridge SE, Porter D, Powell MJ, Griffith GW, Vilgalys R. A molecular phylogeny of the agellated fungi (chytridiomycota) and description of a new phylum (blastocladiomycota). *Mycologia* **98**:860-71 (2006).

[James TY *et al.*, 2009] James TY, Litvintseva AP, Vilgalys R, Morgan JAT, Taylor JW, Fisher MC, Berger L, Weldon C, du Preez L, Longcore JE. Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. *PLoS Pathog.* **5**:1-12, (2009).

[JGI JAM81 sequence] These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community.

[Johnson A, 2003] Johnson A. The biology of mating in *Candida albicans*. *Nat. Rev. Microbiol.* **1**:106-116 (2003).

[Joneson S *et al.*, 2011] Joneson S, Stajich JE, Shiu S-H, Rosenblum EB. Genomic transition to pathogenicity in chytrid fungi. *PLoS Pathog.* **7**:e1002338 (2011).

[Jousson O *et al.*, 2004] Jousson O, Léchenne B, Bontems O, Capoccia S, Mignon B, Barblan J, Quadroni M, Monod M. Multiplication of an ancestral gene encoding secreted fungalysin preceded species differentiation in the dermatophytes *Trichophyton* and *Microsporum*. *Micro biol.* **150**:301-10 (2004).

[Keeling PJ *et al.*, 2000] Keeling PJ, Luker MA, Palmer JD. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol. Biol. Evol.* **17**:23-31 (2000).

[Kelley DR *et al.*, 2010] Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**:R116. (2010).

[Kiezun A *et al.*, 2012] Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, Sunyaev SR. Exome sequencing and the genetic basis of complex traits. *Nat. Genet.***44**:623-630 (2012).

[Kilpatrick AM *et al.*, 2009] Kilpatrick AM, Briggs CJ, Daszak P. The ecology and impact of chytridiomycosis: an emerging disease of amphibians. *TREE* **25**:109-18 (2009).

[Kirisits T *et al.*, 2012] Kirisits T, Kritsch P, Kräutler K, Matlakova M, Halmschlager E. Ash dieback associated with *Hymenoscyphus pseudoalbidus* in forest nurseries in Austria. *J. Agric. Ext. Rural Dev.* **4**:230-235 (2012).

[Kondrashov AS, 1988] Kondrashov AS. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336**:435-40 (1988).

[Kosakovsky Pond SL *et al.*, 2006] Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: A genetic algorithm for recombination detection. *Bioinformatics* **22**: 3096-8 (2006).

[Kuhner MK *et al.*, 1994] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates, *Mol. Biol. Evol.* **11**:459-468 (1994).

[Kwon-Chung KJ *et al.*, 2012] Kwon-Chung KJ, Chang YC. Aneuploidy and drug resistance in pathogenic fungi. *PLoS Path.* **8**:e1003022 (2012).

[Landan G *et al.*, 2009] Landan G, Graur D. Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**:141-7 (2009).

[Langmead B *et al.*, 2009] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**:R25 (2009).

[Lemey P *et al.*, 2009] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5 (2009).

[Lengeler KB *et al.*, 2001] Lengeler KB, Cox GM, Heitman J. Serotype AD strains of *Cryptococcus neoformans* are diploid or aneuploid and are heterozygous at the mating-type locus. *Infect. Immun.* **69**:115-22 (2001).

[Levin BR *et al.*, 1994] Levin BR, Bull JJ. Short-sighted evolution and the virulence of pathogenic microorganisms. *Trends Microbiol.* **2**:76-81 (1994).

[Lewontin RC *et al.*, 1964] Lewontin RC, The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**:49-67 (1964).

[Li H *et al.*, 2008] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**: 1851-8 (2008).

[Li H *et al.*, 2009] Li H, Durbin R, Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**:1754-60 (2009).

[Li H *et al.*, 2009] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078-9 (2009).

[Li H *et al.*, 2010] Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics* **11**:473-483 (2010).

[Li R *et al.*, 2009] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**:1966-7 (2009).

[Li W *et al.*, 2005] Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton BT, Zhang S, Wang LF. Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**:676-9 (2005).

[Lin H *et al.*, 2008] Lin H, Zhang Z, Zhang MQ, Ma B, Li M. ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**:2431-7 (2008).

[Lin X *et al.*, 2005] Lin X, Hull CM, Heitman J. Sexual reproduction between partners of the same mating type in *Cryptococcus neoformans*. *Nature* **434**:1017-21 (2005).

[Lin X *et al.*, 2007] Lin X, Litvintseva AP, Nielsen K, Patel S, Floyd A, Mitchell TG, Heitman J. Alpha AD alpha hybrids of *Cryptococcus neoformans*: evidence of same-sex mating in nature and hybrid fitness. *PLoS Genet.* **3**: e186 (2007).

[Lippincott's Illustrated Reviews, 2007] Lippincott's Illustrated Reviews: Immunology. Publisher: Lippincott Williams & Wilkins; ISBN 0-7817-9543-5. Page 172 (2007).

[Lips KR *et al.*, 2006] Lips KR, Brem F, Brenes R, Reeve JD, Alford RA, Voyles J, Carey C, Livo L, Pessier AP, Collins JP. Emerging infectious disease and the loss of biodiversity in a neotropical amphibian community. *Proc. Natl. Acad. Sci. U. S. A.* **103**:3165-70 (2006).

[Lips KR *et al.*, 2008] Lips KR, Diffendorfer J, Mendelson III JR, Sears MW Riding the wave: Reconciling the roles of disease and climate change in amphibian declines. *PLoS Biol.* **6**:e72 (2008).

[Liu T *et al.*, 2011] Liu T, Ye W, Ru Y, Yang X, Gu B, Tao K, Lu S, Dong S, Zheng X, Shan W, Wang Y, Dou D. Two host cytoplasmic effectors are required for pathogenesis of *Phytophthora sojae* by suppression of host defenses. *Plant Physiol.* **155**:490-501 (2011).

[Lockhart SR *et al.*, 2002] Lockhart SR, Pujol C, Daniels KJ, Miller MG, Johnson AD, Pfaller MA, Soll DR. In *Candida albicans*, white-opaque switchers are homozygous for mating type. *Genetics* **162**:737-745 (2002).

[Longcore JE *et al.*, 1999] Longcore JE, Pessier AP, Nichols DK. *Batrachochytrium dendrobatidis gen. et sp.* nov., a chytrids pathogenic to amphibians. *Mycologia* **91**:219-227, (1999).

[Longcore JR *et al.*, 2007] Longcore JR, Longcore JE, Pessier AP, Halteman WA. Chytridiomycosis widespread in Anurans of northeastern United States. *J. Wildl. Manage.* **71**:435-44 (2007).

[Lötters S *et al.*, 2009] Lötters S, Kielgast J, Bielby J, Schmidtlein S, Bosch J, Veith M, Walker SF, Fisher MC, Rödder D. The link between rapid enigmatic amphibian decline and the globally emerging chytrid fungus. *Ecohealth* **6**:358-72 (2009).

[Lötters S *et al.*, 2011] Lötters S, Kielgast J, Glaw F. Hotspots, conservation and diseases: Madagascar's megadiverse amphibians and the potential impact of chytridiomycosis. In: Habel JC, Zachos F, editors. Biodiversity Hotspots - Evolution and Conservation, 1st edition. Berlin and Heidelberg: Springer Verlag (2011).

[Luquet E *et al.*, 2012] Luquet E, Garner TWJ, Léna J, Bruel C, Joly P, Lengagne T, Grolet O, Plénet S. Genetic erosion in wild populations makes resistance to a pathogen more costly. *Evolution* **66**:1942-52 (2012).

[Ma L *et al.*, 2010] Ma L, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M, et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature* **464**:367-73 (2010).

[Malik A *et al.*, 2010] Malik A, Firoz A, Jha V, Ahmad S. PROCARB: A Database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools. *Adv. Bioinformatics* 436036 (2010).

[Mazzoni R, 2003] Mazzoni R. Emerging pathogen of wild amphibians in frogs (*Rana catesbeiana*) farmed for international trade. *Emerg. Infect. Dis.* **9**:995-8 (2003).

[Mcdonald JH *et al.*, 1991] Mcdonald JH, Kreitman M. Adaptive protein evolution at the Adh locus

in *Drosophila*. *Nature* **351**:652 (1991).

[McMahon TA *et al.*, 2013] McMahon TA, Brannelly LA, Chatfield MW, Johnson PT, Joseph MB, McKenzie VJ, Richards-Zawacki CL, Venesky MD, Rohr JR. Chytrid fungus *Batrachochytrium dendrobatidis* has nonamphibian hosts and releases chemicals that cause pathology in the absence of infection. *Proc. Natl. Acad. Sci. U.S.A.* **110**:210-5 (2013).

[McKenna A *et al.*, 2010] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next- generation DNA sequencing data. *Genome Res.* **20**:1297-303 (2010).

[Miller MG *et al.*, 2002] Miller MG, Johnson AD. White-opaque switching in *Candida albicans* is controlled by mating-type locus homeodomain proteins and allows efficient mating. *Cell* **110**:293-302 (2002).

[Miller MJ *et al.*, 1994] Miller MJ, Powell JI. A quantitative comparison of DNA sequence assembly programs. *J. Comput. Biol.* **1**:257-69 (1994).

[MIT JEL423 sequence] *Batrachochytrium dendrobatidis* sequencing project. Broad institute of Harvard and MIT (http://www.broad.mit.edu).

[Monod M *et al.*, 2002] Monod M, Capoccia S, Léchenne B, Zaugg C, Holdom M, Jousson O. Secreted proteases from pathogenic fungi. *Int. J. Med. Microbiol.* **292**:405-419 (2002).

[Moore ED *et al.*, 1973] Moore ED, Miller CE. Resting body formation by rhizoidal fusion in *Chytriomyces hyalinus*. *Mycologia* **65**:145-154 (1973).

[Morehouse EA *et al.*, 2003] Morehouse EA, James TY, Ganley ARD, Vilgalys R, Berger L, Murphy PJ, Longcore JE. Multilocus sequence typing suggests the chytrid pathogen of amphibians is a recently emerged clone. *Mol. Ecol.* **12**:395-403 (2003).

[Morgan JAT *et al.*, 2007] Morgan JAT, Vredenburg VT, Rachowicz LR, Knapp RA, Stice MJ, Tunstall T, Bingham RE, Parker JM, Longcore JE, Moritz C, Briggs CJ, Taylor JW. Population genetics of the frog-killing fungus *Batrachochytrium dendrobatidis. Proc. Natl. Acad. Sci. U.S.A.* **104**:13845-50 (2007).

[Moss AS *et al.*, 2008] Moss AS, Reddy NS, Dortaj IM, San Francisco MJ. Chemotaxis of the amphibian pathogen *Batrachochytrium dendrobatidis* and its response to a variety of attractants. *Mycologia* **100**: 1-5 (2008).

[Murphy PJ *et al.*, 2009] Murphy PJ, St-Hilaire S, Bruer S, Corn PS, Peterson CR. Distribution and pathogenicity of *Batrachochytrium dendrobatidis* in boreal toads from the Grand Teton Area of western Wyoming. *EcoHealth* **6**:109-20 (2009).

[Musumeci L *et al.*, 2010] Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Human mutat.* **31**:67-73 (2010).

[Ni M *et al.*, 2011] Ni M, Feretzaki M, Sun S, Wang X, Heitman J. Sex in Fungi. **45**: 405-430

(2011).

[Nowak MA *et al.*, 1994] Nowak MA, May RM. Superinfection and the Evolution of Parasite Virulence. *Proc. R. Soc. London Ser. B.* **255**:81-9 (1994).

[Nowrousian M *et al.*, 2010] Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC, Osiewacz HD, Pöggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kück U, Freitag M. *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet.* **6**:e1000891 (2010).

[Nowrousian M, 2010] Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot. Cell* **9**:1300-10 (2010).

[O'Day K, 2007] O'Day K. Completing the *Candida* loop. *PLoS Biol.* **5**:e270 (2007).

[Okoro CK *et al.*, 2012] Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, Kariuki S, Msefula CL, Gordon MA, de Pinna E, Wain J, Heyderman RS, Obaro S, Alonso PL, Mandomando I, MacLennan CA, Tapia MD, Levine MM, Tennant SM, Parkhill J, Dougan G. Intracontinental spread of human invasive *Salmonella Typhimurium* pathovariants in sub-Saharan Africa. *Nat. Genet.* **44**:1215-21 (2012).

[Ouellet M *et al.*, 2005] Ouellet M, Mikaelian I, Pauli BD, Rodrigue J, Green DM. Historical evidence of widespread chytrid infection in North American amphibian populations. *Conserv. Biol.* **19**:1431-40 (2005).

[Park BJ *et al.*, 2009] Park BJ, Wannemuehler KA, Marston BJ, Govender N, Pappas PG, Chiller TM. Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. *AIDS* **23**:525-530 (2009).

[Pessier AP *et al.*, 1999] Pessier AP, Nichols DK, Longcore JE, Fuller MS. Cutaneous chytridiomycosis in poison dart frogs (*Dendrobates spp.*) and white's tree frogs (*Litoria caerulea*). *J. Vet. Diagn. Invest.* **11**:194-9 (1999).

[Piotrowski JS *et al.*, 2004] Piotrowski JS, Annis SL, Longcore JE. Physiology of *Batrachochytrium dendrobatidis*, a chytrid pathogen of amphibians. *Mycologia* **96**:9-15 (2004).

[Pond SLK *et al.*, 2006] Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SD. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* **23**:1891-1901 (2006).

[Pounds JA *et al.*, 2006] Pounds JA, Bustamante MR, Coloma LA, Consuegra JA, Fogden MP, Foster PN, La Marca E, Masters KL, Merino-Viteri A, Puschendorf R, Ron SR, Sánchez-Azofeifa GA, Still CJ, Young BE. Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature* **439**:161-7 (2006).

[Punta M *et al.*, 2012] Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman

A, Finn RD. The Pfam protein families database, *Nucleic Acids Res.* **40**:D290-D301 (2012).

[Purcell S *et al.*, 2007] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**:559-575 (2007).

[R Development Core Team, 2009] R Development Core Team. R: A Language and Environment for Statistical Computing. R foundation for Statistical Computing: Vienna, Austria (2009).

[Raffaele S *et al.*, 2010] Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, Jiang RH, Zody MC, Kunjeti SG, Donofrio NM, Meyers BC, Nusbaum C, Kamoun S. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science* **330**:1540-3 (2010).

[Raffaele R *et al.*, 2012] Raffaele R, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* **10**:417-430 (2012).

[Ramsey JP *et al.*, 2010] Ramsey JP, Reinert LK, Harper LK, Woodhams DC, Rollins-Smith LA. Immune defences against *Batrachochytrium dendrobatidis*, a fungus linked to global amphibian declines, in the South African clawed frog, *Xenopus laevis*. *Infect. Immun.* **78**:3981-92 (2010).

[Rawlings ND *et al.*, 2010] Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38**:D227-D233 (2010).

[Reedy JL *et al.*, 2009] Reedy JL, Floyd AM, Heitman J. Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. *Curr. Biol.* **19**:891-9 (2009).

[Retallick RWR *et al.*, 2007] Retallick RWR, Miera V. Strain differences in the amphibian chytrid *Batrachochytrium dendrobatidis* and non-permanent, sub-lethal effects of infection. *Dis. Aquat. Org.* **75**:201-7 (2007).

[Ribas L *et al.*, 2009] Ribas L, Li MS, Doddington BJ, Robert J, Seidel JA, Kroll JS, Zimmerman LB, Grassly NC, Garner TW, Fisher MC. Expression profiling the temperature-dependent amphibian response to infection by *Batrachochytrium dendrobatidis*. *PLoS One* **4**:e8408 (2009).

[Roetzer A *et al.*, 2013] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S, Supply P, Kalinowski J, Niemann S. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. *PLoS Med.* **10**:e1001387 (2013).

[Rohr JR *et al.*, 2008] Rohr JR, Raffel TR, Romansic JM, McCallum H, Hudson PJ. Evaluating the links between climate, disease spread, and amphibian declines. *Proc. Natl. Acad. Sci. U. S. A.* **105**:17436-17441 (2008).

[Rollins-Smith LA *et al.*, 2002] Rollins-Smith LA, Doersam JK, Longcore JE, Taylor SK, Shamblin JC, Carey C, Zaslo MA. Antimicrobial peptide defences against pathogens associated with global amphibian declines. *Dev. Comp. Immunol.* **26**:63-72 (2002).

[Rosenblum EB *et al.*, 2010] Rosenblum EB, Fisher MC, James TY, Stajich JE, Longcore JE,

Gentry LR, Poorten TJ. A molecular perspective: biology of the emerging pathogen *Batrachochytrium dendrobatidis*. *Dis. Aquat. Organs.* **92**:131-47 (2010).

[Rosenblum EB *et al.*, 2009] Rosenblum EB, Poorten TJ, Settles M, Murdoch GK, Robert J, Maddox N, Eisen MB. Genome-wide transcriptional response of *Silurana* (*Xenopus*) *tropicalis* to infection with the deadly chytrid fungus. *PLoS One* **4**:e6494 (2009).

[Rosenblum EB *et al.*, 2008] Rosenblum EB, Stajich JE, Maddox N, Eisen MB. Global gene expression profiles for life stages of the deadly amphibian pathogen *Batrachochytrium dendrobatidis*. *Proc. Natl. Acad. Sci. U. S. A.* **105**:17034-9 (2008).

[Rosenblum EB *et al.*, 2012] Rosenblum EB, Poorten TJ, Joneson S, Settles M. Substrate-specific gene expression in *Batrachochytrium dendrobatidis*, the chytrid pathogen of amphibians. *PLoS One* **7**:e49924 (2012).

[Ross LO *et al.*, 1996] Ross LO, Maxfield R, Dawson D. Exchanges are not equally able to enhance meiotic chromosome segregation in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **93**:4979-83 (1996).

[Ruffalo M *et al.*, 2011] Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**:2790-6 (2011).

[Rumble SM *et al.*, 2009] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. Shrimp: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**:e1000386 (2009).

[Salmela L, 2011] Salmela L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* **26**:1284-90 (2011).

[Sanglard D *et al.*, 1999] Sanglard D, Ischer F, Calabrese D, Majcherczyk PA, Bille J. The ATP binding cassette transporter gene CgCDR1 from *Candida glabrata* is involved in the resistance of clinical isolates to azole antifungal agents. *Antimicrob. Agents Chemother.* **43**:2753-65 (1999).

[Sasson A *et al.*, 2010] Sasson A, Michael TP. Filtering error from SOLiD Output. *Bioinformatics* **26**:849-850 (2010).

[Saville BJ *et al.*, 1998] Saville BJ, Kohli Y, Anderson JB. mtDNA recombination in a natural population. *Proc. Natl. Acad. Sci. U. S. A.* **95**:1331-1335 (1998).

[Schloegel LM *et al.*, 2010] Schloegel LM, Daszak P, Cunningham AA, Speare R, Hill B. Two amphibian diseases, chytridiomycosis and ranaviral disease, are now globally notifiable to the World Organization for Animal Health (OIE): an assessment. *Dis. Aquat. Organ.* **92**:101-8 (2010).

[Schloegel LM *et al.*, 2012] Schloegel LM , Toledo LF, Longcore JE, Greenspan SE, Vieira CA, Lee M, Zhao S, Wangen C, Ferreira CM, Hipolito M, Davies AJ, Cuomo CA, Daszak P, James TY. Novel, panzootic and hybrid genotypes of amphibian chytridiomycosis associated with the bullfrog trade. *Mol. Ecol.* **21**:5162-5177 (2012).

[Schoustra SE *et al.*, 2007] Schoustra SE, Debets AJM, Slakhorst M, Hoekstra RF. Mitotic recom-

bination accelerates adaptation in the fungus *Aspergillus nidulans*. *PLoS Genet.* **3**:e68 (2007).

[Seimon TA *et al.*, 2007] Seimon TA, Seimon A, Daszak P, Halloy SRP, Schloegel LM, Aguilar CA, Sowell P, Hyatt AD, Konecky B, Simmons JE. Upward range extension of Andean anurans and chytridiomycosis to extreme elevations in response to tropical deglaciation. *Glob. Chang. Biol.* **13**:288-299 (2007).

[Sheltzer JM *et al.*, 2011] Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, Humpton TJ, Brito IL, Hiraoka Y, Niwa O, Amon A. Aneuploidy drives genomic instability in yeast. *Science* **333**:1026-30 (2011).

[Sheltzer JM *et al.*, 2012] Sheltzer JM, Torres EM, Dunham MJ, Amon A. Transcriptional consequences of aneuploidy. *Proc. Natl. Acad. Sci. U. S. A.* **109**:12644-9 (2012).

[Shiozaki K *et al.*, 1996] Shiozaki K, Russell P. Conjugation, meiosis, and the osmotic stress response are regulated by Spc1 kinase through Atf1 transcription factor in fission yeast. *Genes Dev.* **10**:2276-88 (1996).

[Simwami SP *et al.*, 2011] Simwami SP, Khayhan K, Henk DA, Aanensen DM, Boekhout T, Hagen F, Brouwer AE, Harrison TS, Donnelly CA, Fisher MC. Low diversity *Cryptococcus neoformans variety grubii* multilocus sequence types from Thailand are consistent with an ancestral African origin. *PLoS Pathog.* **7**:e1001343 (2011).

[Sionov E *et al.*, 2010] Sionov E, Lee H, Chang YC, Kwon-Chung KJ. *Cryptococcus neoformans* overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. *PLoS Pathog.* **6**:e1000848 (2010).

[Skerratt LF *et al.*, 2007] Skerratt LF, Berger L, Speare R, Cashins S, McDonald KR, Phillott AD, Hines HB, Kenyon N. Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. *Ecohealth* **4**:125-34 (2007).

[Smith JM, 1978] Smith JM. The evolution of sex. (1978).

[Smith JM *et al.*, 1993] Smith JM, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? *Proc. Natl. Acad. Sci. U. S. A.* **90**:4384-8 (1993).

[Smith T, 1904] Smith T. Some problems in the life-history of pathogenic microorganism. *Science* **20**:817-832 (1904).

[Soanes DM *et al.*, 2008] Soanes DM, Alam I, Cornell M, Wong HM, Hedeler C, Paton NW, Rattray M, Hubbard SJ, Oliver SG, Talbot NJ. Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS One* **3**:e2300 (2008).

[Soto-Azat C *et al.*, 2010] Soto-Azat C, Clarke BT, Poynton JC, Cunningham AA. Widespread historical presence of *Batrachochytrium dendrobatidis* in African pipid frogs. *Divers. Distrib.* **16**:126-31 (2010).

[Sparrow FK, 1960] Sparrow FK. Aquatic Phycomycetes. 2^{nd} revised edition. Ann Arbor, MI. University of Michigan Press (1960).

[Stockwell MP *et al.*, 2010] Stockwell MP, Clulow J, Mahoney MJ. Host species determines whether infection load increases beyond disease-causing thresholds following exposure to the amphibian chytrid fungus. *Anim. Conserv.* **13**, Suppl. 1, 62-71 (2010).

[Stoletzki N *et al.*, 2011] Stoletzki N, Eyre-Walker A. Estimation of the Neutrality Index. *Mol. Biol. Evol.* **28**:63 (2011).

[Strange RN *et al.*, 2005] Strange RN, Scott PR. Plant disease: a threat to global food security. *Annu. Rev. Phytopathol.* **43**:83-116 (2005).

[Stuart SN *et al.*, 2004] Stuart SN, Chanson JS, Cox NA, Young BE, Rodrigues AS, Fischman DL, Waller RW. Status and trends of amphibian declines and extinctions worldwide. *Science* **306**:1783-6 (2004).

[Stukenbrock EH *et al.*, 2008] Stukenbrock EH, McDonald BA. The origins of plant pathogens in agro-ecosystems. *Ann. Rev. Phytopathol.* **46**:75-100 (2008).

[Sugiura N, 1978] Sugiura N. Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Stat. Theory And Meth.* **7**:13-26 (1978).

[Sukegawa Y *et al.*, 2011] Sukegawa Y, Yamashita A, Yamamoto M. The fission yeast stress-responsive MAPK pathway promotes meiosis via the phosphorylation of Pol II CTD in response to environmental and feedback cues. *PLoS Genet.* **7**:e1002387 (2011).

[Sun G *et al.*, 2011] Sun G, Yang Z, Kosch T, Summers K, Huang J. Evidence for acquisition of virulence effectors in pathogenic chytrids. *BMC Evol. Biol.* **11**:195 (2011).

[Tamukai K *et al.*, 2011] Tamukai K, Une Y, Tominaga A, Suzuki K, Goka K. Treatment of spontaneous chytridiomycosis in captive amphibians using itraconazole. *J. Vet. Med. Sci.* **73**:155-9 (2011).

[Tang S *et al.*, 2008] Tang S, Hyland FCL, Wessel TC, Sorenson J, Peckham H, Vega FMDL. DiBayes: A SNP detection algorithm for next-generation Dibase sequencing. Available from http://solidsoftwaretools.com/gf/project/dibayes/ (2008).

[Tavaré S, 1986] Tavaré S. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures Math. Life Sci.* **17**:57-86 (1986).

[Taylor JW *et al.*, 1999] Taylor JW, Jacobson DJ, Fisher MC. The evolution of asexual fungi: reproduction, speciation and classification. *Annu. Rev. Phytopathol.* **37**:197-246 (1999).

[Taylor JW *et al.*, 2000] Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM. Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* **31**:21-32 (2000).

[Trapnell C *et al.*, 2009] Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **27**:455-7 (2009).

[Tyler MJ, 1997] Tyler MJ. The action plan for Australian frogs. Wildlife Australia, Canberra (1997).

[Van Rooij P *et al.*, 2012] Van Rooij P, Martel A, D'Herde K, Brutyn M, Croubels S, Ducatelle R, Haesebrouck F, Pasmans F. Germ tube mediated invasion of *Batrachochytrium dendroba-*

tidis in amphibian skin is host dependent. *PLoS ONE* **7**:e41481. (2012).

[Vieites DR *et al.*, 2009] Vieites DR, Wollenberg KC, Andreone F, Köhler J, Glaw F, Vences M. Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. *Proc. Natl. Acad. Sci. U. S. A.* **106**:8267-72 (2009).

[Voyles J *et al.*, 2009] Voyles J, Young S, Berger L, Campbell C, Voyles WF, Dinudom A, Cook D, Webb R, Alford RA, Skerratt LF, Speare R. Pathogenesis of chytridiomycosis, a cause of catastrophic amphibian declines. *Science* **326**:582-5 (2009).

[Voyles J *et al.*, 2012] Voyles J, Vredenburg VT, Tunstall TS, Parker JM, Briggs CJ, Rosenblum EB. Pathophysiology in mountain yellow-legged frogs (*Rana muscosa*) during a chytridiomycosis outbreak. *PLoS ONE* **7**:e35374 (2012).

[Vredenburg VT *et al.*, 2010] Vredenburg VT, Knapp RA, Tunstall TS, Briggs CJ. Dynamics of an emerging disease drive largescale amphibian population extinctions. *Proc. Natl. Acad. Sci. U. S. A.* **107**:9689-94 (2010).

[Walker SF *et al.*, 2007] Walker SF, Salas MB, Jenkins D, Garner TW, Cunningham AA, Hyatt AD, Bosch J, Fisher MC. Environmental detection of *Batrachochytrium dendrobatidis* in a temperate climate. *Dis. Aquat. Organ.* **77**:105-12 (2007).

[Walker SF *et al.*, 2008] Walker SF, Bosch J, James TY, Litvintseva AP, Oliver Valls JA, Piña S, García G, Rosa GA, Cunningham AA, Hole S, Griffiths R, Fisher MC. Invasive pathogens threaten species recovery programs. *Curr. Biol.* **18**:R853-4 (2008).

[Walker SF *et al.*, 2010] Walker SF, Bosch J, Gomez V, Garner TW, Cunningham AA, Schmeller DS, Ninyerola M, Henk DA, Ginestet C, Arthur CP, Fisher MC. Factors driving pathogenicity versus prevalence of amphibian panzootic chytridiomycosis in Iberia. *Ecol. Lett.* **13**: 372-82 (2010).

[Weir BS, 1996] Weir BS. Genetic Data Analysis II. Sinauer, Sunderland (1996).

[Weldon C *et al.*, 2004] Weldon C, du Preez LH, Hyatt AD, Muller R, Spears R. Origin of the amphibian chytrid fungus. *Emerg. Infect. Dis.* **10**:2100-5 (2004).

[Weldon C *et al.*, 2007] Weldon C, De Villiers AL, du Preez LH. Quantification of the trade in *Xenopus laevis* from South Africa, with implications for biodiversity conservation. *Afr. J. Herpetol.* **56**: 77-83 (2007).

[Win J *et al.*, 2007] Win J, Morgan W, Bos J, Krasileva KV, Cano LM, Chaparro-Garcia A, Ammar R, Staskawicz BJ, Kamoun S. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic Oomycetes. *Plant Cell* **19**:2349-2369 (2007).

[Wittenberg AH *et al.*, 2009] Wittenberg AH, van der Lee TA, Ben M'barek S, Ware SB, Goodwin SB, Kilian A, Visser RG, Kema GH, Schouten HJ. Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. *PLoS ONE* **4**:e5863 (2009).

[Woodhams DC *et al.*, 2007] Woodhams DC, Ardipradja K, Alford R. Resistance to chytridio-

mycosis varies among amphibian species and is correlated with skin peptide defences. *Anim. Conserv.* **10**:409-17 (2007).

[Woodhams DC *et al.*, 2008] Woodhams DC, Alford RA, Briggs CJ, Johnson M, Rollins-Smith LA. Life-history trade-offs influence disease in changing climates: strategies of an amphibian pathogen. *Ecology* **89**:1627-39 (2008).

[Wright S, 1931] Wright S. Evolution in Mendelian populations. *Genetics* **16**:97-159 (1931).

[Yang Z *et al.*, 2000] Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**:32-43. (2000).

[Yang Z, 2007] Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**:1586-91 (2007).

[Yoon H *et al.*, 2011] Yoon H, Ansong C, Adkins JN, Heffron F. Discovery of *Salmonella* virulence factors translocated via outer membrane vesicles to murine macrophages. *Infect. Immun.* **79**:2182-92 (2011).

[Zerbino D *et al.*, 2008] Zerbino D, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**:821-9 (2008).

[Zhang W *et al.*, 2011] Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS One* **6**:e17915 (2011).

[Zhao X *et al.*, 2010] Zhao X, Palmer LE, Bolanos R, Mircean C, Fasulo D, Wittenberg GM. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* **17**: 1549-60 (2010).

Appendix A: Culturing *Bd*

The isolates collected over the duration of this study were cultured according to an unpublished document called "Recognizing, isolating and culturing *Bd* from amphibians" by Joyce E. Longcore. Firstly, media called mTGhL (acronym of the ingredients) is made by adding 8g tryptone, 2g gelatin hydrolysate, 4g lactose and 1 litre of distilled water in a clean jar (or proportional amounts in smaller or larger jars) and autoclaved. The addition of 10g agar will make the final media solidify at c. 40°C, and therefore needs to be kept in a waterbath at 50°C and then poured onto plates. *Bd* can then be either kept alive on solid media for longer periods of time, or grown up in larger quantities (such as for sequencing) in liquid media. Preparing mTGhL plates, as well as passaging from either flasks or plates, was done in a Biosafety cabinet to prevent bacterial or fungal contamination. The addition of antibiotics (Ab) to the media (200mg/L penicillin-G and 200-500mg/L streptomycin sulfate) after autoclaving and cooling can further reduce bacterial contaminations, but has the undesired effect of slowing or inhibiting *Bd* growth through an as yet unknown mechanism. Media was kept in a 10°C fridge, and it has a shelf life of approximately a month before it starts to lose its ability to sustain *Bd*. These ingredients were used both in the lab, and in the field (either by me or other collectors) with 'isolation kits' that I put together, consisting of mTGhL in screw cap tubes and plates (both with antibiotics). The key ingredients and some aspects of fieldwork are shown in Figure 1.4.

Although swabbing is suitable for detecting *Bd*, the only methods that have successfully isolated the fungus involve taking toe-clips from adult frogs (non-lethal) kept in mTGhL w/Ab or with whole (euthanised) tadpoles, whereby tadpoles are bathed in 10g MS 222/litre water, which is absorbed through amphibian skin and prevents the generation and conduction of nerve impulses. Mouthparts can then be separated from the rest of the tadpole using a sterile scalpel blade. To remove other fungi or bacteria from mouthparts of toe-clips, it can be dragged through solid media with antibiotics. Next, small pieces of tissue can be placed in individual wells of a 12 well plate. Because contaminations can occur very easily, sterile tools (ethanol, fire or new from factories) are used in the hood, and experimental design includes many replicates, which decrease the likelihood of bacterial or fungal contamination and therefore successfully isolating *Bd*. Identifying *Bd*+ growth is done using a light microscope at 100X magnification in the following 2days-1 week.

Appendix B: Sequencing and cryo-preserving *Bd*

Once a panel of isolates were chosen for sequencing, the first step was to grow it up in large enough quantities: usually ranging between 4-6 100ml flasks and 2-3 plates. DNA was then extracted and purified using the Qiagen Blood and Cell Culture DNA mini kit (cat. No. 13323) according to the protocol provided for each of the isolates in Table 1.1. Specifically, after c. 10 days growing at 22°C, all flasks and plates were scraped and transferred into 50ml falcon flasks and centrifuged for 5mins at 1700 rotations per minute (rpm) to remove supernatant (media). The final pellet of cells was then digested with 2ml Buffer G2, transferred into autoclaved screw cap tubes with glass beads and 4μl RNAse A, and put in a BeadBeater at high speed (homogenise setting) for 1.5 minutes. Tubes were centrifuged at maximum speed, and the supernatant moved to a 15ml falcon tubes leaving behind cell walls and large proteins. Next, 100μl of proteinase K was added to digest remaining proteins, vortexed (shaken) and incubated at 50°C overnight. The following day, falcon tubes were spun at 3400rpm for 10 minutes. Qiagen filter tips were placed in new 15ml falcon tubes and 2ml Buffer QBT put through, followed by the supernatant (leaving behind a pellet of remaining proteins and other cell contents), and then washed with 3ml of Buffer QC. Genomic DNA in the filter was then eluted into new 15ml falcon tubes with 2ml Buffer QF. To purify the DNA, 1.4ml isopropanol[7] was added, centrifuged for 30min at 3400rpm, and supernatant removed. A second step was to add 1ml cold 70% ethanol, centrifuged for 10 minutes at 3400rpm, supernatant removed, and placed in the hood to air-dry for between 10 minutes and 1 hour. Finally, DNA was re-suspended with 100μl-deionised water, left in a water bath for 1-2 hours at 55°C, vortexed, and transferred into a small tube placed in the -20°C freezer until library preparation or quantification.

DNA quantification was initially carried out on a NanoDrop spectrophotometer, where 1μl of solution was analysed using different light wavelengths providing a concentration in ng/μL. Later, all DNA was more accurately measured using a Qubit fluorometer, whereby a fluorescent dye and high and low standards are used (100ng/μL and no DNA respectively). Although some extractions failed to recover any DNA, measurements were typically between 30-60μg/mL, with 216μg/mL (therefore, 21.6μg total) the most I recovered from a single isolate. Should a more concentrated sample of DNA be needed, the small-tubes could be used in the speed-vac, which removed excess water.

Genomic DNA is prepared differently following extraction and purification, according to the platform used to sequence it and the type of dataset being made. The cheapest and most straightforward NGS dataset consists of 'single-end' reads from a single sample. Owing to the large number

[7]Synonyms include 2-propanol, isopropanol and isopropyl alcohol.

of sequences and depth produced by most sequencing platforms, isolates are often 'pooled' together, and separated after sequencing, which requires additional indexed primers. Finally, two reads can be derived from DNA molecules longer than the two reads (called paired-end), with a specific space called 'insert-sizes' (e.g. 500nt) between them, which can later be used to help resolve larger structural re-arrangements such as indels and inversions, assemble genomes *de novo*, as well as give greater resolution across repetitive regions.

In this study we used single-end SOLiD reads and paired-end Illumina reads, which were done in part by an out-of-department sequencing facility (Hammersmith Hospital and South Kensington Imperial College). However, Dr. Hannah Larner was responsible for performing most of the library preparation for the Illumina sequencing, whom I learnt the procedure from. Briefly, DNA was placed in Covaris tubes and fragmented by a Sonicator machine (replacing the previous Nebulisation methods), transferred into a 96 well PCR plate, followed by end-repair using proprietary end-repair mixes in a thermal engine at 30°C for 30 minutes in the proportions 1:1. Tubes were spun at 1000rpm for 20 seconds, and a magnetic bead and ethanol clean up was performed[8]. The plate was then spun down for 20 seconds at 1000rpm, and put back on the magnetic stand for 30 seconds, and the 15μl fragmented DNA solution from each well transferred to a clean plate and left in the -20°C freezer.

The following day, a 1:1 mix was added to each well (e.g. 15μl consisting of 12.5μl A-tailing mix and 12.5μl EB Buffer) and incubated at 27°C for 30 minutes in the thermocycler, which added a string of adenines to a single strand at each end of the DNA fragments. Next, specific coded adapters were joined to the A-tails using a mix of 2.5μl adapters and 2.5μl EB Buffer for each well, plus a specific 2.5μl indexed primer set to each sample, which could later be used to distinguish pooled samples. The plate was incubated for 10 minutes at 30C to allow the reaction to take place, and then 5μl STOP solution was added per well to stop the reaction. The bead and ethanol clean up was performed a second time, except with 55μl EB Buffer, 50μl of which was moved to a clean plate, and 5μl could be kept to analyse later to identify the problem, should there be one.

Finally, size-selection was performed to remove primer dimers as well as poorly fragmented DNA that would not sequence properly. First, 40μl beads were added to the 50μl solution and left for 10 minutes. Next, the plate was put on the magnetic stand, and 85μl was moved to a new plate, leaving behind beads binding long strands of DNA that could join more than one cluster on the sequencing flow cell and give erroneous readings. Next, 100μl new beads (1:2 ratio) were added,

[8]A magnetic bead and ethanol clean up was performed by adding 1:1 Agencourt AMPure XP magnetic beads to the samples (E.g. the first time 100μl beads: 100μl DNA-solution) and leaving for 10 minutes, which gives it time to bind to the DNA. The 96 well plate was then put on a magnetic plate and left till the solution became clear, and bound beads moved to the bottom of the wells. Beads were washed with 200μl 80% ethanol, and 20μl EB buffer added to elute the DNA.

left for 10 minutes, put on the magnetic stand for a few minutes, and all supernatant removed. The beads were given two ethanol clean-ups (the final part of the footnote), and DNA is eluted with 25μl EB Buffer. In order to increase the DNA content of the solution, a PCR step can be performed using 25μl MasterMix and 5μl primers in the thermocycler for 98°C for 30 seconds, and then 10 cycles of 98°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds. The PCR ends with 72°C for 5 minutes and left at 4°C. A further bead and ethanol clean up was performed and left in 35μl EB Buffer, 30μl of which was transferred to a new plate for sequencing and 5μl for the Bioanalyser, which is able to size and quantify the DNA before loading onto an Illumina flow-cell. In the Illumina machine, DNA strands bound to flow cells are replicated using polymerases and fluorescently labelled terminators that are imaged as each dNTP is added and then cleaved to allow incorporation of the next base. The latest Illumina HiSeq 2500 model (as of September 2012) produces terabytes of images, which are processed down into c. 8Gbytes fastq files holding up to 6 billion 100nt paired-end reads. The software, chemistry and pipelines are continually being developed, resulting in a rapidly increasing quantity and quality of data.

A final aspect of Lab work was the cryogenic preservation of *Bd* isolates, which provided insurance should the live cultures die-out or change following passaging and replication, and we needed to go back to either a group of isolates of interest detected by sequencing. Fortunately, we did not need to recover any cryogenically frozen isolates over the duration of this study, but they may still be useful for future studies. The procedure is as instructed by Boyle DG *et al.* [Boyle DG *et al.*, 2003] and involved growing a fresh batch of *Bd* for 10 days, scraping the flask and then transferring into a 15ml falcon tube. Cells were spun down at 1700rpm for 10 minutes and the pellet was resuspended in 1ml of a solution containing 10% Dimethyl sulfoxide (DMSO) that protects against ice formation and 10% Fetal Calf Serum (FCS) that acts as nutrition when reactivated. The solution is then put into a 2ml cryotube and placed in a circular cryo-container with isopropanol for slower cooling, and left in the -80°C freezer. If these samples are wanted at a later date, cryo-tubes should be placed directly into water at 43°C and placed on a TGhL plates with the addition of 1ml DS salt solution as suggested by Boyle *et al.* [Boyle DG *et al.*, 2003].

Appendix C: Published articles from chapters presented in this thesis

Outbreaks and the Emergence of Novel Fungal Infections: Lessons from the Panzootic of Amphibian Chytridiomycosis

Matthew C Fisher, BSc, PhD, and Rhys A Farrer, BSc, MSc

Department of Infectious Disease Epidemiology, St Mary's Hospital, Imperial College London, London, UK



Share this article with a colleague, visit
www.invasivefungalinfections.com

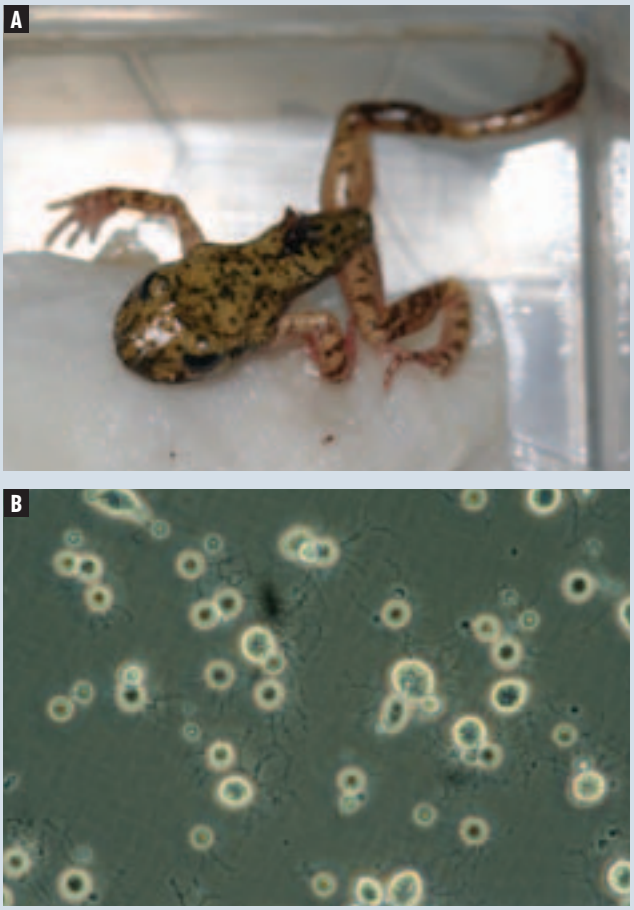
Chytridiomycosis is a cutaneous infection of amphibians caused by the chytridiomycete fungal pathogen *Batrachochytrium dendrobatidis* (*Bd*). Despite being in a phylum not known for pathogenicity in vertebrates, *Bd* is now recognized as a primary driver of amphibian declines. Data show that this novel pathogen emerged in the 20th century to colonize amphibians worldwide. Such rapid emergence of a previously unrecognized pathogen illustrates many aspects of emerging fungal infections that threaten human health, namely long-distance human-mediated dispersal, multihost reservoirs, and altered virulence. In order to combat *Bd*, new tools have been developed to track its global spread and to analyze in parallel whole-genome diversity. This article details how such tools have applications to tracking and managing human fungal infections. *J Invasive Fungal Infect* 2011;**5**(3):73–81.

Emerging pathogenic fungi are an increasing threat to natural populations and species of animals, plants, and humans [1]. For example, new strains of fungi are evolving to overcome cultivar resistance resulting in escalating losses for agribusiness [2], little brown bats across the US are being overwhelmed by the previously undescribed ascomycete fungus *Geomyces destructans* [3,4], and *Cryptococcus gattii* is expanding its range into non-endemic environments with a consequential increase in incidence seen in humans [5,6]. However, the record for the greatest impact of any pathogen on its host is held by a member of a previously little-known basal phylum of fungi, the Phylum Chytridiomycota. The chytrid *Batrachochytrium dendrobatidis* (*Bd*) was discovered in 1997 and named in 1999 as an infection causing the rapidly progressing and often fatal cutaneous disease, chytridiomycosis, in anuran (frog-like) and caudate (salamander-like) amphibians (**Figure 1**) [7,8]. Prior to the discovery of *Bd*, it was recognized that amphibians were facing an extinction crisis that threatened approximately one-third of all species [9]. While habitat loss was known to be the main driver of amphibian species loss, it was also recognized that many declines were found to occur in pristine, protected environments where known threats (such as habitat loss or species overharvesting) did not occur; these mysterious losses were recorded as "enigmatic declines" by the International Union for Conservation of Nature (IUCN) Red List of Threatened Species. *Bd* was subsequently discovered as the proximate driver of multiple-species enigmatic declines [10,11] following the observation of simultaneous waves of population declines in Central America and Australia [7,12], and the observation of local introduction and spread of the fungus [13,14]. The fungus has now been found to infect >442 species of amphibians in 49 countries on all continents except for the amphibian-free Antarctic [15,16].

We ask the question: "how does a previously little-recognized clade of fungi rise to infamy as containing one of the most destructive pathogens ever witnessed?". We show that many of the characteristics that have enabled *Bd* to cause a panzootic are broadly shared across the fungal kingdom. We argue that the emergence of *Bd* and chytridiomycosis is a direct consequence of anthropogenic activity, and that new tools are required to rapidly detect and manage this, and other new, emerging fungal pathogens. Finally, we show that lessons learned from the emergence of chytridomycosis, and the current toolkit developed to combat the emergence, are

Address for correspondence: Matthew C Fisher, Department of Infectious Disease Epidemiology, St Mary's Hospital, Imperial College London, London, W2 1PG, UK.
Email: matthew.fisher@imperial.ac.uk

Figure 1. A: The Mallorcan midwife toad (*Alytes muletensis*). **B:** *Batrachochytrium dendrobatidis* growing in mTGHL liquid media showing zoospores, zoosporangia, and pseudohyphae.



mTGhL: 8 g tryptone, 2 g gelatin hydrolysate, 4 g lactose, 1 L distilled water. Available in colour at www.invasivefungalinfections.com.

relevant to preparing for future fungal pathogens that could impact human health.

Host–pathogen interactions and pathogenicity in *Bd*

Chytrids are widely distributed across soil and aquatic ecosystems, and appear to be the main class of fungi that dominate at high altitudes [17]. Many species of the Phylum Chytridiomycota are aquatic saprobes, which break down macromolecules such as chitin and cellulose. While only *Bd* is pathogenic to vertebrates, many species of chytrid parasitize phyto- and zooplankton, fungi, invertebrates, and plants [18]. During its reproductive, parasitic phase, sporangia release flagellate zoospores into water, where they exhibit chemotaxis towards a number of substrates including sugars, proteins, and amino acids [19]. These zoospores can be identified by specific environmental quantitative polymerase chain reaction (PCR) tests [20]. Following attachment to host cells, likely facilitated by carbohydrate-binding proteins, cell entry and formation of new sporangia occurs, completing the infectious life-cycle.

While *Bd* is not a dermatophyte *sensu stricto*, it has close similarities to other cutaneous fungi of vertebrates in its utilization of keratin as a primary substrate. This feature limits the distribution of infection to the stratum corneum and stratum granulosum of adult amphibian skin, where sporangia develop within keratinized tissues. In larval amphibians (tadpoles), the infection is localized to the keratinized mouthparts, and consequently larvae do not suffer the dramatic pathologies that are associated with adult infection [21]. While the mechanism of infection has not been fully determined, a comparative analysis of the *Bd* genome compared with that of other non-pathogenic fungi has shown that fungalysin metallopeptidase (also known as peptidase M36) and serine protease gene families have undergone extensive expansions in the *Bd* genome [22]. Proteinases have been detected in analyses of the *Bd* proteome for a number of different isolates, showing that these open-reading frames are translated at levels sufficiently high to be detected by two-dimensional gel approaches [23]. The fungalysin metallopeptidase gene family was shown to be differentially expressed between two different life-history stages of *Bd*, the zoospore and the sporangia, which lends support to their putative key role in the infection process [24]. Metallopeptidases have similarly undergone expansion in the human-infecting dermatophytes *Trichophyton* spp. and *Microsporum* spp. where they are highly upregulated and account for up to 36% of total secreted protein extracts [25,26]. Dermatophyte fungi, like *Bd*, are keratinophilic and comparative genomics of *Arthrodema benhamiae* and *Trichophyton verrucosum* show >235 predicted protease-encoding genes in eacxh species, many of which are shared with other closely related species in the order Onygenales such as *Coccidioides immitis* [26]. RNA-seq experiments in *A benhamiae* after co-inoculation with and without keratinocytes demonstrated the differential expression of >40 genes encoding putatively secreted proteins, showing the capacity of skin-infecting fungi to modify gene expression according to their metabolic substrate [26]. Such secreted proteins are therefore prime candidates as virulence factors in *Bd* as well as in other skin-infecting fungi.

In common with vertebrate dermatophyte fungi, many of the mechanisms underlying the interactions between *Bd* and its hosts are currently unclear. For example, there appears to be a minimal host reaction to infection other than hyperplasia and hyperkeratosis of the stratum corneum, and noticeable lesions are usually not observed. Death of the host appears to result from pathophysiological changes related to electrolyte imbalances. Intact skin function is essential for amphibians owing to their need to actively maintain a hyperosmotic internal environment as a consequence of having highly permeable skin. In diseased individuals, a pronounced imbalance in electrolyte levels occurs as a consequence of epidermal sodium and chloride

channels becoming inhibited, leading to hypokalemia (low plasma potassium) and hyponatremia (low plasma sodium). The ultimate cause of death is asystolic cardiac arrest, resulting from the electrolytic imbalances [27]. However, whether the release of a fungal toxin or direct damage to infected host cells results in the disruption of osmoregulatory function is not yet known. Furthermore, no genes in *Bd* have yet been proven to interact directly or indirectly with any host response to infection.

"Innate immunity in vertebrates is a key defense against a wide variety of cutaneous fungi"

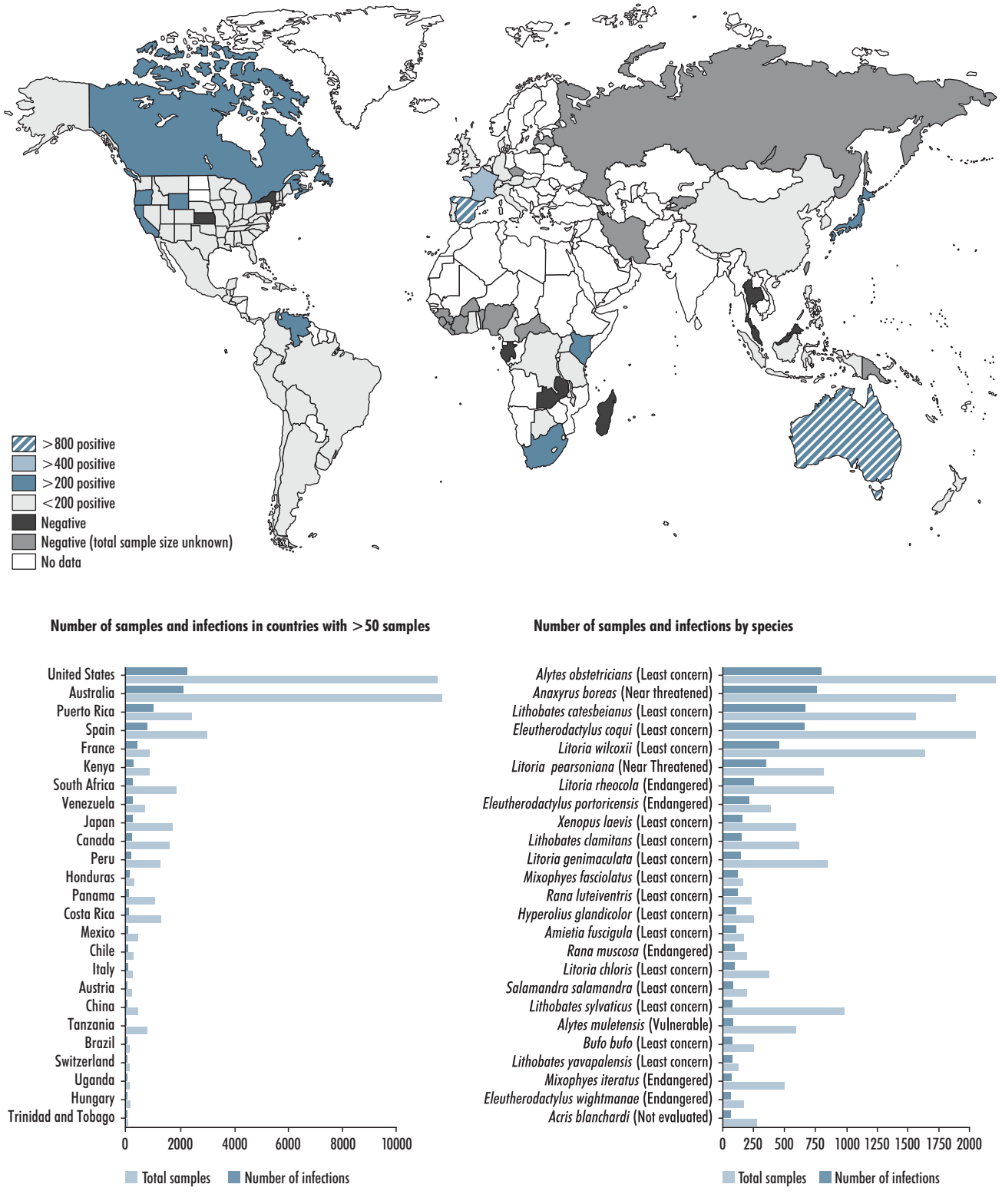
Functional approaches to investigating the amphibian response against *Bd* have focused on measuring differences in time courses of gene-expression between infected and uninfected animals. Two key studies focusing on the Western clawed frog, *Silurana* (*Xenopus*) *tropicalis*, as a model species showed little evidence for an adaptive immune response in the skin, liver, and spleen of infected frogs [28,29]. In these studies, the only significant protective responses observed were the induction of components of host innate immunity including the expression of genes associated with the production of the antimicrobial skin peptide preprocareulein, and proinflammatory responses [28]. These findings correlate with previous observations, showing that amphibian species that tolerate infection produce antimicrobial peptides with greater efficacy against *Bd* [30]. A parallel study on a related species of clawed frog, *Xenopus laevis,* confirmed that skin peptide secretion was necessary to combat infection, and could be ablated by using norepinephrine to discharge skin peptides prior to infection [31]. However, the study also demonstrated the production of anti-*Bd* immunoglobulins following immunization with crude heat-killed *Bd* preparations, suggesting that an adaptive immune response was also being mounted. Together, these studies clearly show that an intact antimicrobial innate response is necessary to defend against infection; however, the role of adaptive immunity still needs to be defined. More broadly, it appears that intact innate immunity in vertebrates is a key defense against a wide variety of cutaneous fungi, and that a systems-based functional genomic analysis will provide novel insights into the host–pathogen interaction that is broadly applicable across the fungal kingdom.

Mapping the contemporary global spread of *Bd*

The emergence of *Bd* was first detected by the observation that waves of amphibian die-offs were occurring simultaneously in Central America and the Australian Wet Tropics [7]. Subsequently, five main systems have been identified in which spatiotemporal emergence and spread have been occurring: the Mesoamerican peninsula, the northern tip of South America, the Sierra Nevada (US), eastern Australia, and the European Pyrenees [15]. In several systems, species extinction has occurred; >40% (n=30) of host species were documented as extirpated in the El Copé study site (Panama, Central America) [32]. Initial responses to the crisis were *ad hoc* and fragmented; however, the development of a highly specific and sensitive molecular, diagnostic, quantitative TaqMan® real-time PCR assay (Life Technologies, Inc., Carlsbad, CA, USA) that utilizes a minor groove-binding probe to the *Bd* ribosomal DNA array (the internal transcribed spacer [ITS] and 5.8S regions) [33], standardized nationwide surveys for the presence/absence of *Bd* to a large extent. Subsequently, a multiphase project focused on rapidly acquiring and compiling global *Bd* data was initiated – the global *Bd*-mapping project (*Bd*-Maps). A novel, web-based system for depositing records of confirmed *Bd* infection was developed at www.bd-maps.net [16], where data on predefined data fields can be uploaded alongside spatial coordinates. Such mapping allows real-time regional aggregation of spatial and temporal data across a variety of scales, allowing the extent of infection, and the species infected, to be tracked rapidly through online maps and data summaries (**Figure 2**). Currently, the database holds prevalence data for >34 000 samples from 3600 locations, 79 countries, and 1095 species of amphibian; mapping on this scale has given precise insights into where infection is present and, importantly, where it is absent. Given the rapidity of the emergence of *Bd*, and the large numbers of species affected, it is critical that hitherto uninfected regions remain biosecure and protected from introduction of the pathogen. The mapped distribution of *Bd* shows that, while infection is widespread, it remains patchy, and several areas exist that contain high amphibian biodiversity but are so far negative for infection (**Figure 2**). The most notable of these regions is the island of Madagascar, which contains >460 species of amphibian [34]. The potential for *Bd* to extirpate this unique and megadiverse community of amphibians has led to calls for a high degree of biosecurity to be implemented [35,36].

The use of publically available global mapping tools (e.g. www.bd-maps.net) that allow simultaneous use by scientists, clinicians, and policy-makers, is currently experiencing intense interest. Mobile phone-based data-acquisition using customized applications allows users to become independent of desktop computing while allowing two-way communication between online databases and workers in the field. EpiCollect (www.epicollect.net) is one such application [37], allowing fieldworkers to collect project-specific information focused on monitoring the spread of *Bd*, and to synchronize these data with *Bd*-Maps online and via third-generation (3G) mobile telephone networks. Such technological applications are being rapidly developed and are widely applicable for the surveillance of

Figure 2. Top panel: The current global distribution of *Bd* samples as of May 5, 2011, visualized using the custom mapping tool. **Bottom panel:** Summary of *Bd*-Maps metrics detailing the worldwide spatial prevalence of *Bd* [16]. *Bd*-Maps was developed by David Aanensen, Imperial College London, in collaboration with Dede Olson and Matthew Fisher.



Bd: Batrachochytrium dendrobatidis.

emerging human pathogens. Examples include ProMed (www.promedmail.org) and HealthMaps (http://healthmap.org/en) where disease trends are identified in real-time, leading to the immediate dispersal of alerts when anomalous increases in the incidence of pathogens are reported.

Vectors and origins of the *Bd* panzootic

The contemporary global spread of *Bd* appears to be driven by the global trade in infected amphibians. A meta-analysis by Fisher and Garner showed that ≥28 species of amphibian are known to be carriers of *Bd*, and have invaded novel ecosystems when introduced as alien species [38]. Several of these species are known to support asymptomatic infections of *Bd*, and have been introduced on an enormous scale; chief culprits are the African clawed frog *X laevis*, the North American bullfrog *Lithobates catesbeianus* (formerly *Rana catesbeiana*), and the South American cane toad *Rhinella marina* (formerly *Bufo marinus*). These species have established feral populations in the Americas, Europe, Australia, Asia, as well as many oceanic and coastal islands, and are associated with a high prevalence of infection both across their native ranges and in regions where they have been introduced [38,39]. The globally farmed North American bullfrog is thought to act as a "superspreader" of infection for a number of reasons: the species is widely infected by *Bd* across its native range in the US [40] and tolerates high burdens of infection, the species is farmed in huge numbers and so infection is amplified [41], live and infected animals are exported globally [39], and the species is widely invasive outside of its native range following uncontrolled introduction [42]. Analyses of polymorphisms in the ribosomal ITS have shown that specific haplotypes of *Bd* are associated with bullfrogs, and these have spread into native populations of amphibians following introduction [43,44], which demonstrates the status of this species as a vector of the pathogen. A similar "superspreader" status has been awarded to *X laevis*, which, like bullfrogs, are widely infected and farmed, globally transported for research purposes, and can tolerate high burdens of infection. Direct "spillover" of *Bd* from wild-caught African *Xenopus* spp. into a naïve *Bd*-susceptible Mallorcan midwife toad *Alytes muletensis* (**Figure 1**) has been demonstrated within the environs of zoos, followed by the introduction of *Bd*-infected *A muletensis* onto the Mediterranean island of Mallorca where infection has now become established [45].

Together, these studies illustrate the following important features about the transcontinental movement of infectious fungi:

- Anthropogenic activity is causing the rapid, continent-wide translocation of infectious fungi vectored by multiple host species or as saphrophytes in contaminated biological material.
- There are likely as-yet-unidentified pathogens of future consequence that could be spread by trade.
- Modern, intensive-farming processes can act to amplify infection *in situ*.
- Pathogen spillover can occur from tolerant (vector/reservoir) to susceptible host species, and this risk can be high for pathogenic fungi of animals that characteristically infect multiple host species.

The final point is of importance to clinicians as many fungi present extraordinarily broad host ranges, and humans are likely to be exposed to novel fungal pathogens owing to high rates of international trade in goods that might contain infectious fungal spores, whether these be plant or animal products. While *Bd* is not infectious to humans or livestock, there is no guarantee that no other fungal pathogens exist that are better adapted to infecting human hosts. There is a precedent for this phenomenon as such "pathogen spillover" from domesticated or wild animals to humans likely explains the origin and expansion of several important human fungal pathogens [46]. Dermatophyte fungi in the order Onygenales, such as *A benhamiae*, *T verrucosum*, and other *Trichophyton* spp. are widely zoophilic, with natural reservoirs in species such as rodents and equids; however, these fungi also have the potential to cross-infect and cause highly inflammatory infections in humans. *Coccidioides immitis* (also in the order Onygenales) appears to have colonized South America at the end of the Pleistocene era (*circa* 10 000 years ago) through being vectored by early human migrations following spillover of infection into humans from small mammal reservoirs in the Southwestern US [47]. A similar pattern of expansion via the anthropogenic dispersal of pigeons as vectors appears to have caused the expansion of the Basidiomycete fungus *C neoformans* var. *grubii* out of Africa to colonize Southeast Asia [48]. Therefore, there is ample precedence showing that the ability of infectious fungi to infect a wide range of hosts, twinned with global movements of infected plant and animal material, has the potential to establish new infections in naïve, previously uninfected regions.

While anthropogenic movement of *Bd* via the global trade of amphibians appears to explain the rapid global emergence of chytridiomycosis, the origin of the infection remains shrouded in mystery. Historical surveys have shown that the earliest records of *Bd* appear to stem from Africa in the early half of the 20th century, with preserved specimens of *Xenopus* showing signs of cutaneous infection by sporangia from the 1920s and 1930s [49,50]. These specimens pre-date those from other regions of the world, where the earliest record of infection in the Americas was from a specimen in the state of Quebec, Canada, in 1961 [51]. *Xenopus* spp. are widely infected in Africa but can tolerate infection; they were traded globally as a test for human pregnancies in the first half of the 20th century [49]. As widespread amphibian declines were initially observed in the 1960s [52], a temporal pattern for

Bd emergence "Out of Africa" appears plausible. However, wider spatial surveys from museum collections and the use of molecular analysis of globally distributed isolates of *Bd* are necessary to dissect the exact pattern of the global spread of *Bd*, and to test the validity of the "Out of Africa" hypothesis for the emergence of *Bd*. Indeed, such molecular analyses have already revealed aspects of the genetic structure of the *Bd* population that are in conflict with the "Out of Africa" model. For instance, Goka et al. have shown that there is a divergent lineage of *Bd* that infects Japanese giant salamanders (*Andrias japonicus*) and that this appears to pre-date the arrival of strains of the pathogen that occurred via bullfrog introduction [44]. This lineage appears to be asymptomatic, suggesting a history of coevolution with the salamanders. As more genetic data accrue on isolates of amphibian-associated chytrids worldwide, it is entirely plausible that multiple lineages of *Bd* will be found and that more than one of these lineages may contribute to the spread of infection. Whether these lineages all have the same capacity to cause serious infection in new host species/environments remains to be seen, and this is a key challenge for future research.

Population genetics approaches to tracking the emergence of *Bd*

While detecting the presence or fungal load of *Bd* in an amphibian host remains in the remit of diagnostic PCR methods, identifying the temporal sequence underlying the emergence of *Bd* requires the application of molecular epidemiological approaches. James et al. generated 17 sequence-based markers from 59 *Bd* isolates from five continents and 30 host amphibian species. This comprehensive study showed that the entire global diversity of *Bd* could be explained by the dispersal of a single diploid individual, and levels of genetic diversity were among the lowest recorded for a eukaryotic pathogen [53]. While these findings are consistent with a recent globalization of *Bd*, they do not support Weldon et al.'s "Out of Africa" hypothesis [49], as levels of genetic diversity in isolates from North America were found to be as high as they were in Africa. The extreme paucity of polymorphisms across these loci coupled with a sampling strategy that was targeted at New World populations of amphibians, suggested that a greater depth of both sampling and genotyping would be necessary to more effectively address the question of the origin of *Bd* [53].

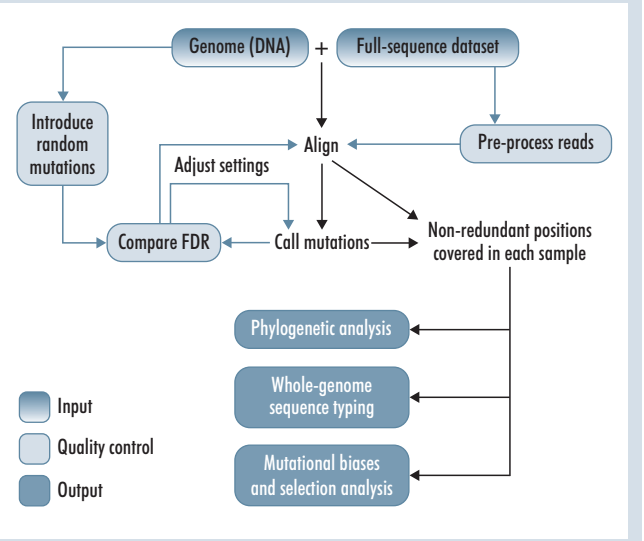
Subsequent studies are now focused on investigating the molecular epidemiology of *Bd* using next-generation sequencing (NGS) to enable whole genome sequence typing of the pathogen. Such approaches have already found use in molecular epidemiological applications for other fungal pathogens, such as determining linkage between cases of coccidioidomycosis disseminated by organ transplants [54]. Platforms such as HiSeq™ (Illumina, Inc., San Diego, CA, USA), Applied Biosystems Sequencing by Oligonucleotide Ligation and Detection (ABI SOLiD™), or 454 Sequencing™ (454 Life Sciences, Inc., Bradford, CT, USA) can provide the entire genomic sequences for numerous isolates of *Bd*, which can then be assembled or aligned to either of the two publically available genomes of *Bd* (JEL423 [www.broadinstitute.org] or JAM81 [www.jgi.doe.gov]), revealing polymorphic sites among those samples across the whole genome. This approach vastly increases the analytical power as tens of thousands of single-nucleotide polymorphisms (SNPs; i.e. sites that differ between two isolates) are scored, as opposed to the dozens that typify earlier approaches such as that used by James et al. [53]. Shared polymorphic sites at a number of loci can then be used to construct a phylogenetic tree, showing past transmission events between continents/countries or sites of introduction. Although many of the NGS platforms are reasonably well established in terms of ability to sequence multiple genomes, there remain a number of bioinformatic challenges that need to be considered such as choosing the correct tools and parameters, addressing quality control, and setting up suitable and potentially reusable pipelines [55]. These methods are largely applicable and transferable to understanding fungal pathogens of humans.

Determining genetic variation with NGS alignment tools

Figure 3 shows a conceptual bioinformatic pipeline that we currently use to manage NGS data. While we illustrate how such a pipeline is used to handle data from *Bd* sequencing projects, these methods are directly applicable for any human pathogenic fungus for which a genome sequence already exists.

The nuclear genome of *Bd* has a haploid genome size of approximately 24 Mb. This is a reasonably standard size for a fungal genome. Recent NGS data for a global panel of isolates has shown that the majority of isolates sequenced from around the world fall within a single, highly-related lineage exhibiting high levels of heterozygosity [56], and contain representative isolates that were previously genotyped by James et al. [53]. For example, alignment of the Sanger sequencing reads of JAM81 to the genome of JEL423 using Burrows–Wheeler Aligner (BWA) with default settings [57], and scoring mutations using Sequencing Alignment/Map tools (SAMtools) [58], identifies approximately 10 000 SNPs (0.4 per Kb), 31 000 heterozygous sites (1.3 per Kb), and 9000 insertion/deletions between these two reference genomes. These polymorphisms can be identified in just a few hours using only a desktop computer and a freely available alignment tool such as BWA, which is able to utilize any of the aforementioned platforms. While the availability of two *Bd* genomes and high levels of heterozygosity make alignments a more attractive alternative to ascertain these differences than assemblies *de novo*, the suitability of an alignment strategy

Figure 3. Schematic illustration of the informatic pipeline used to manipulate and align next-generation sequencing data from *Batrachochytrium dendrobatidis*.



FDR: false discovery rate

will vary depending on the type and size of the input dataset, as well as the extent of error present – all of which may affect downstream analysis. Therefore, it is important to assess a number of methods and parameters in terms of accuracy and utilization of the NGS dataset.

Quality control of NGS fungal datasets

There are a large number of alignment tools designed specifically for dealing with NGS datasets such as BWA or Bowtie [55,57,59], some of which are designed for specific platforms. The Short-Read Mapping Package (SHRiMP), for instance, is tailored to the di-base (color-space) sequencing format of ABI SOLiD reads, which require a different set of algorithms to correctly resolve consecutive SNPs or miscalled bases compared with Illumina systems [60]. In addition to choosing an alignment tool, there are often a number of adjustable parameters that may affect the alignment such as the maximum number of mismatches allowed between a read and the reference. Common alignment optimization options include the ability to remove reads containing, or averaging below, a Phred quality score (www.phrap.org), or by trimming low quality 3′ ends that would otherwise prevent usable data from aligning.

The most successful alignment strategy can be determined when a closely related strain or species has already had its full genome sequenced. By including this "reference strain", for example the *Bd* isolate JEL423, within the panel being sequenced, a control alignment can be generated. Using this alignment, the alignment parameters and SNP-calling algorithms can be fine-tuned in terms of breadth of read-coverage (ideally higher) and the number of SNPs called (ideally lower). Next, to optimize true-positive/false-negative SNP calling, the same reads can be aligned to an artificially "mutated" version of the reference genome. This will reveal the power of the tools, within the context of the dataset, to correctly determine the number of correct (introduced) SNPs, or numbers of correctly covered genes. Using a false discovery rate (FDR) approach is a good way to measure how completely the genomes from the panel of isolates are covered, to determine the error rate associated with the dataset, and to help identify the most suitable alignment strategy. For example, a dataset of whole genome sequences can be identified using FDR as incomplete or of low quality when the "control alignment" has genes uncovered, or as manifesting a high number of discrepancies, even after the optimization of parameters.

Calling polymorphic sites and genome assembly *de novo*

After aligning the NGS dataset to a reference sequence, polymorphic sites can be identified using either a pre-made SNP-calling tool such as SAMtools [58], the Genome Analysis Toolkit [61], diBayes (http://solidsoftwaretools.com/gf/project/dibayes), or with a custom program that can be specifically fine-tuned to a dataset and/or alignment, and therefore may out-perform other generic packages. If this strategy is adopted, an initial starting place is to first define a cut-off for the number, or percentage, of reads disagreeing with a reference nucleotide (e.g. >70% or >90%). Next, a minimum read depth could be applied to reduce low-coverage sequencing errors being called, or a minimum base or alignment quality score for inclusion applied, which would, again, reduce the number of sequencing errors mistaken for real SNPs. Depending on the length of the reads, and to what extent the quality drops at the 3′ end, it might also be worth including read position filters. As the number of putative SNPs will likely change based on the strategy and dataset used, a number of methods should be tested against the reference isolate in order to identify the most suitable method before applying it to the rest of the panel of isolates. Maximizing the ratio of true positives to false positives is important for successfully tracking the spread of *Bd* and other pathogenic fungi, especially on very fine geographical scales where isolates may differ by only a handful of SNPs.

While NGS is primarily used in alignments, if a closely related reference sequence is unavailable (as might be expected for a novel emerging infection), high-depth NGS datasets can be used for assemblies *de novo*. Improvements made in read length and assembly tools designed specifically for NGS, such as Velvet [62], are facilitating the assembly of fungal genomes from entirely NGS datasets [63], or in combination with Sanger sequences [64]. While substantially more memory intensive, as well as identifying SNPs, these methods have the potential to

additionally reveal genomic rearrangements, reveal genes not present in the reference strain, and to give greater resolution to long indels such as gene presence/absence polymorphisms.

Reconstruction of the evolutionary relationships between isolates

Newly emerging pathogenic fungi are characterized by having low levels of genetic variation relative to their point of origin, owing to the multiple population bottlenecks that are associated with rapid spread. For instance, the genomes of isolates from the emergent lineage of *Bd* have low levels of genetic variation, differing by approximately 1000–10 000 SNPs. However, genome sequences have also been assembled from isolates of *Bd* recovered from Africa and Europe that differ from the newly emerged lineage by approximately five-times the number of SNPs, showing that the global genetic diversity of *Bd* is likely to be far higher than was previously expected. Once a set of polymorphic sites has been identified in a panel of isolates, phylogenetic trees can be constructed from a subset of those sites that can be verified in each of the samples (i.e. entirely covered by NGS reads in all samples). Phylogenetic methods such as neighbor-joining (NJ) or the unweighted pair group method (UPGMA) with arithmetic mean, cluster polymorphic sites based on their genetic distances. Highly associated samples within the trees are an indicator of transmission events between their collection sites, and have recently been used to confirm specific cases of intercontinental transmission of *Bd* via the amphibian trade [45]. For instance, NGS has been successfully used to show that isolates of *Bd*, which have emerged in Mallorca, have their ancestry in Africa. This approach has confirmed the hypothesis previously raised by Walker et al. [45]. *Xenopus* frogs collected in Africa were held in the same zoo as Mallorcan *Alytes muletensis*; cross-transmission of *Bd* subsequently occurred and this was followed by establishment on the island via the release of *A muletensis* in an attempt to bolster native endangered populations [45]. This example shows the utility of using NGS datasets for mapping the routes of pathogenic fungi spread with relevance to human health.

Bayesian approaches such as that implemented by the software Bayesian Evolutionary Analysis Sampling Trees (BEAST) apply models of evolution that can reveal additional information about isolates, such as predicted dates of divergence [65]. These statistical genetic techniques show the origin of the *Bd* panzootic as having occurred within the 20th century, confirming that amphibian declines in the late 1960s/early 1970s were due to the globalization of *Bd* [56]. At present, however, the origin of the panzootic lineages of *Bd* have not been determined and it is likely that further NGS of a wider diversity of isolates from different regions, environments, and hosts will be necessary to determine the origins of the panzootic.

Summary

The emergence of *Bd* stems from the globalization of a single, aggressive lineage during the 20th century. Although other lineages of *Bd* are now known to occur, these have not had the same impact on amphibian biodiversity, probably owing to their lower virulence [23]. While the origins of the panzootic of *Bd* remain unknown, it is evident that the global trade in amphibians, and perhaps other yet-unidentified vectors, has rapidly disseminated this keratinophilic cutaneous chytrid into naïve amphibian populations with some catastrophic consequences. Novel genomic approaches, such as the use of NGS for whole genome sequencing or global gene-expression profiling, are providing insights into the fundamental evolutionary modes and virulence determinants in *Bd*. These technologies show great promise to illuminating the processes underlying other emerging fungal infections. While *Bd* does not infect other vertebrates, the extensive literature and research on this pathogen are raising it to a model status in terms of understanding the drivers that govern the emergence of virulent pathogenic fungi. Given the recent emergence of other fungi as serious pathogens, there is an urgent need to increase our research capability in order to rapidly ascertain the drivers that underlie emerging fungal pathogens, in order to effectively take action and to mitigate the effect of such infections. This article on the emergence of *Bd* in amphibian species should be used to inform researchers working on human pathogens in the commonalities of pathogenesis, and research approach, that exist between the disparate fields of human and wildlife disease.

You can submit comments and questions on this article at:

www.invasivefungalinfections.com

Acknowledgments

The UK Natural Environmental Research Council and the European Research Area Network project "Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity" (RACE) have provided funding for this research.

Disclosures

The authors have no relevant financial interests to disclose.

References

1. The Institute of Medicine. Olson L, Choffnes ER, Relman DA et al., editors. *Fungal Diseases: An Emerging Threat to Human, Animal, and Plant Health – Workshop Summary*. Washington, DC: The National Academies Press, 2011. http://www.iom.edu/Reports/2011/Fungal-Diseases-An-Emerging-Challenge-To-Human-Animal-And-Plant-Health-Workshop-Summary.aspx Accessed on September 27, 2011.
2. Hovmøller MS, Walter S, Justesen AF. Escalating threat of wheat rusts. *Science* 2010;**329**:369.
3. Gargas A, Trest MT, Christensen M et al. *Geomyces destructans* sp. nov. associated with bat white-nose syndrome. *Mycotaxon* 2009;**108**:147–54.

4. Frick WF, Pollock JF, Hicks AC et al. An emerging disease causes regional population collapse of a common North American bat species. *Science* 2010;**329**:679–82.
5. Fraser JA, Giles SS, Wenink EC et al. Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak. *Nature* 2005;**437**:1360–4.
6. Byrnes EJ 3rd, Li W, Lewit Y et al. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. *PLoS Pathog* 2010;**6**:e1000850.
7. Berger L, Speare R, Daszak P et al. Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proc Natl Acad Sci U S A* 1998;**95**:9031–6.
8. Longcore JE, Pessier AP, Nichols DK. *Batrachochytrium dendrobatidis* gen. et sp. nov., a chytrid pathogenic to amphibians. *Mycologia* 1999;**91**:219–27.
9. Stuart SN, Chanson JS, Cox NA et al. Status and trends of amphibian declines and extinctions worldwide. *Science* 2004;**306**:1783–6.
10. Skerratt LF, Berger L, Speare R et al. Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. *Ecohealth* 2007;**4**:125–34.
11. Lötters S, Kielgast J, Bielby J et al. The link between rapid enigmatic amphibian decline and the globally emerging chytrid fungus. *Ecohealth* 2009;**6**:358–72.
12. Lips KR, Brem F, Brenes R et al. Emerging infectious disease and the loss of biodiversity in a neotropical amphibian community. *Proc Natl Acad Sci U S A* 2006;**103**:3165–70.
13. Vredenburg VT, Knapp RA, Tunstall TS et al. Dynamics of an emerging disease drive large-scale amphibian population extinctions. *Proc Natl Acad Sci U S A* 2010;**107**:9689–94.
14. Walker SF, Bosch J, Gomez V et al. Factors driving pathogenicity vs. prevalence of amphibian panzootic chytridiomycosis in Iberia. *Ecol Lett* 2010;**13**:372–82.
15. Fisher MC, Garner TW, Walker SF. Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. *Annu Rev Microbiol* 2009;**63**:291–310.
16. *Bd*-Maps and the *Bd* Global Mapping Project. *Bd*-Maps and the *Bd* Global Mapping Project is a collaboration between Aanensen D, Fisher MC, Powell C, and Spratt B (Imperial College London, London, UK), and Olson D (US Forest Service). Available at http://www.bd-maps.net Accessed on September 27, 2011.
17. Freeman KR, Martin AP, Karki D et al. Evidence that chytrids dominate fungal communities in high-elevation soils. *Proc Natl Acad Sci U S A* 2009;**106**:18315–20.
18. Gleason FH, Kagami M, Lefevre E et al. The ecology of chytrids in aquatic ecosystems: roles in food web dynamics. *Fungal Biology Reviews* 2008;**22**:17–25.
19. Moss AS, Reddy NS, Dortaj IM et al. Chemotaxis of the amphibian pathogen *Batrachochytrium dendrobatidis* and its response to a variety of attractants. *Mycologia* 2008;**100**:1–5.
20. Walker SF, Salas MB, Jenkins D et al. Environmental detection of *Batrachochytrium dendrobatidis* in a temperate climate. *Dis Aquat Organ* 2007;**77**:105–12.
21. Garner TW, Walker S, Bosch J et al. Life history tradeoffs influence mortality associated with the amphibian pathogen *Batrachochytrium dendrobatidis*. *Oikos* 2009;**118**:783–91.
22. Rosenblum EB, Fisher MC, James TY et al. A molecular perspective: biology of the emerging pathogen *Batrachochytrium dendrobatidis*. *Dis Aquat Organ* 2010;**92**:131–47.
23. Fisher MC, Bosch J, Yin Z et al. Proteomic and phenotypic profiling of the amphibian pathogen *Batrachochytrium dendrobatidis* shows that genotype is linked to virulence. *Mol Ecol* 2009;**18**:415–29.
24. Rosenblum EB, Stajich JE, Maddox N et al. Global gene expression profiles for life stages of the deadly amphibian pathogen *Batrachochytrium dendrobatidis*. *Proc Natl Acad Sci U S A* 2008;**105**:17034–9.
25. Jousson O, Léchenne B, Bontems O et al. Multiplication of an ancestral gene encoding secreted fungalysin preceded species differentiation in the dermatophytes *Trichophyton* and *Microsporum*. *Microbiology* 2004;**150**:301–10.
26. Burmester A, Shelest E, Glöckner G et al. Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. *Genome Biol* 2011;**12**:R7. In press.
27. Voyles J, Young S, Berger L et al. Pathogenesis of chytridiomycosis, a cause of catastrophic amphibian declines. *Science* 2009;**326**:582–5.
28. Ribas L, Li MS, Doddington BJ et al. Expression profiling the temperature-dependent amphibian response to infection by *Batrachochytrium dendrobatidis*. *PLoS One* 2009;**4**:e8408.
29. Rosenblum EB, Poorten TJ, Settles M et al. Genome-wide transcriptional response of *Silurana* (*Xenopus*) *tropicalis* to infection with the deadly chytrid fungus. *PLoS One* 2009;**4**:e6494.
30. Woodhams DC, Ardipradja K, Alford RA et al. Resistance to chytridiomycosis varies among amphibian species and is correlated with skin peptide defenses. *Animal Conservation* 2007;**10**:409–17.
31. Ramsey JP, Reinert LK, Harper LK et al. Immune defenses against *Batrachochytrium dendrobatidis*, a fungus linked to global amphibian declines, in the South African clawed frog, *Xenopus laevis*. *Infect Immun* 2010;**78**:3981–92.
32. Crawford AJ, Lips KR, Bermingham E. Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proc Natl Acad Sci U S A* 2010;**107**:13777–82.
33. Boyle DG, Boyle DB, Olsen V et al. Rapid quantitative detection of chytridiomycosis (*Batrachochytrium dendrobatidis*) in amphibian samples using real-time Taqman PCR assay. *Dis Aquat Organ* 2004;**60**:141–8.
34. Vieites DR, Wollenberg KC, Andreone F et al. Vast underestimation of Madagascar's biodiversity evidenced by an integrative amphibian inventory. *Proc Natl Acad Sci U S A* 2009;**106**:8267–72.
35. Andreone F, Carpenter AI, Cox N et al. The challenge of conserving amphibian megadiversity in Madagascar. *PLoS Biol* 2008;**6**:e118.
36. Lötters SD, Kielgast J, Glaw F. Hotspots, conservation and diseases: Madagascar's megadiverse amphibians and the potential impact of chytridiomycosis. In: Habel JC, Zachos F, editors. *Biodiversity Hotspots – Evolution and Conservation*, 1st edition. Berlin and Heidelberg: Springer Verlag, 2011.
37. Aanensen DM, Huntley DM, Feil EJ et al. EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. *PLoS One* 2009;**4**:e6968.
38. Fisher MC, Garner TW. The relationship between the introduction of *Batrachochytrium dendrobatidis*, the international trade in amphibians and introduced amphibian species. *Fungal Biology Reviews* 2007;**21**:2–9.
39. Garner TW, Perkins MW, Govindarajulu P et al. The emerging amphibian pathogen *Batrachochytrium dendrobatidis* globally infects introduced populations of the North American bullfrog, *Rana catesbeiana*. *Biol Lett* 2006;**2**:455–9.
40. Longcore JR, Longcore JE, Pessier AP et al. Chytridiomycosis widespread in anurans of northeastern United States. *Journal of Wildlife Management* 2007;71:435–44.
41. Mazzoni R. Emerging pathogen of wild amphibians in frogs (*Rana catesbeiana*) farmed for international trade (erratum in *Emerg Infect Dis* 2003;**9**:1188). *Emerg Infect Dis* 2003;**9**:995–8.
42. Garner TW, Stephen I, Wombwell E et al. The amphibian trade: bans or best practice? *Ecohealth* 2009;**6**:148–51.
43. Fisher MC. Endemic and introduced haplotypes of *Batrachochytrium dendrobatidis* in Japanese amphibians: sink or source? *Mol Ecol* 2009;**18**:4731–3.
44. Goka K, Yokoyama J, Une Y et al. Amphibian chytridiomycosis in Japan: distribution, haplotypes and possible route of entry into Japan. *Mol Ecol* 2009;**18**:4757–74.
45. Walker SF, Bosch J, James TY et al. Invasive pathogens threaten species recovery programs. *Curr Biol* 2008;**18**:R853–4.
46. Daszak P, Cunningham AA, Hyatt AD. Emerging infectious diseases of wildlife – threats to biodiversity and human health. *Science* 2000;**287**:443–9.
47. Fisher MC, Koenig GL, White TJ et al. Biogeographic range expansion into South America by *Coccidioides immitis* mirrors New World patterns of human migration. *Proc Natl Acad Sci U S A* 2001;**98**:4558–62.
48. Simwami SP, Khayhan K, Henk DA et al. Low diversity *Cryptococcus neoformans* variety *grubii* multilocus sequence types from Thailand are consistent with an ancestral African origin. *PLoS Pathog* 2011;**7**:e1001343.
49. Weldon C, du Preez LH, Hyatt AD et al. Origin of the amphibian chytrid fungus. *Emerg Infect Dis* 2004;**10**:2100–5.
50. Soto-Azat C, Clarke BT, Poynton JC et al. Widespread historical presence of *Batrachochytrium dendrobatidis* in African pipid frogs. *Diversity and Distributions* 2010;**16**:126–31.
51. Ouellet M, Mikaelian I, Pauli BD et al. Historical evidence of widespread chytrid infection in North American amphibian populations. *Conservation Biology* 2005;**19**:1431–40.
52. Houlahan JE, Findlay CS, Schmidt BR et al. Quantitative evidence for global amphibian population declines. *Nature* 2000;**404**:752–5.
53. James TY, Litvintseva AP, Vilgalys R et al. Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. *PLoS Pathog* 2009;**5**:e1000458.
54. Engelthaler DM, Chiller T, Schupp JA et al. Next-generation sequencing of *Coccidioides immitis* isolated during cluster investigation. *Emerg Infect Dis* 2011;**17**:227–32.
55. Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell* 2010;**9**:1300–10.
56. Farrer RA, Weinert LA, Bielby J et al. Multiple emergence of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent lineage. *Proc Natl Acad Sci U S A* 2011. In press.
57. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
58. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
59. Langmead B, Trapnell C, Pop M et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
60. Rumble SM, Lacroute P, Dalca AV et al. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* 2009;**5**:e1000386.
61. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
62. Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.
63. Nowrousian M, Stajich JE, Chu M et al. *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet* 2010;**6**:e1000891.
64. Diguistini S, Liao NY, Platt D et al. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol* 2009;**10**:R94.
65. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007;**7**:214.

SCIENTIFIC REPORTS



OPEN

SUBJECT AREAS:
SOFTWARE
COMPARATIVE GENOMICS
RARE VARIANTS
GENOMICS

Received
9 November 2012

Accepted
25 February 2013

Published
21 March 2013

Correspondence and requests for materials should be addressed to R.A.F. (r.farrer09@imperial.ac.uk)

Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects

Rhys A. Farrer[1], Daniel A. Henk[1], Dan MacLean[2], David J. Studholme[3] & Matthew C. Fisher[1]

[1]Department of Infectious Disease Epidemiology, St Mary's Hospital, Imperial College London, London, UK, [2]The Sainsbury Laboratory, Norwich Research Park, Norwich, UK, [3]College of Life and Environmental Sciences, University of Exeter, Exeter, UK.

Sequence alignments form the basis for many comparative and population genomic studies. Alignment tools provide a range of accuracies dependent on the divergence between the sequences and the alignment methods. Despite widespread use, there is no standard method for assessing the accuracy of a dataset and alignment strategy after resequencing. We present a framework and tool for determining the overall accuracies of an input read dataset, alignment and SNP-calling method providing an isolate in that dataset has a corresponding, or closely related reference sequence available. In addition to this tool for comparing False Discovery Rates (FDR), we include a method for determining homozygous and heterozygous positions from an alignment using binomial probabilities for an expected error rate. We benchmark this method against other SNP callers using our FDR method with three fungal genomes, finding that it was able achieve a high level of accuracy. These tools are available at http://cfdr.sourceforge.net/.

Sources of error within next-generation sequencing (NGS) projects can result in poorly resolved genotypes[1–3]. Without an assessment of the false discovery rate (FDR) for genome or transcriptome projects, misaligned reads and inaccurate base calls can go unnoticed thereby propagating into SNP discovery. From the onset, datasets comprising sequenced fragments (reads) can harbor a range of potential sequencing errors such as PCR amplification bias in Illumina data[4], polyclonal errors in SOLiD data[5] and homopolymers in 454 data[6]. Despite a general trend towards higher success rate at correctly deciphering bases in reads, individual runs can still harbor unexpected levels of error from low quality DNA extractions or library preparations.

In addition to read-quality, an arguably more substantial source of error arises during read-alignment[7], the process whereby sequenced reads are mapped to a closely related reference genome. This is the first and most fundamental analysis undertaken once the DNA sequence has been produced[8]. Alignments are often preferable to *de novo* assemblies due to increased speed and reduced memory requirements. Given a low read-depth or highly heterogeneous sequence, alignments may also recover more genetic data than assembling without a reference due to current limitations in assembly algorithms[9]. As is true for *de novo* assemblies, the accuracy of alignments varies considerably depending on the software and its parameters used[10], the type and size of the dataset and the amount of erroneous base calls. Further, alignments are also affected by the genetic distance between reference and newly sequenced genomes.

Most alignment tools score the placement of a read based on the uniqueness of its match. For example, BWA[11] and Bowtie[12] are frequently used alignment tools that are based on indexing a reference sequence using a Burrows-Wheeler transformation (BWT)[13]. These tools can align in base (Illumina, 454, Ion Torrent, PacBio) or color space (SOLiD) using a combination of base and alignment mapping quality scores to determine the correct positions of reads and assigning a genotype, whilst ignoring reads containing low quality base calls or low mapping scores. The alignment tool SHRiMP[14] is tailored to the platform specific biases in the color-space format of ABI SOLiD reads, using a different set of algorithms to resolve consecutive single nucleotide polymorphisms (SNPs) or miscalled bases, which 'base-space' aligners would unsuccessfully determine. Many other examples of short-read alignment tools such as Maq[15], SOAP[16] and Zoom[17] are available and may be preferential given specific experimental requirements such as speed, accuracy or additional features[18].

Pre-processing reads prior to alignment by removing low quality reads or 3′ ends is a common initial step to improve alignment accuracy. Sequencing errors can also be detected and removed during the pre-processing of reads using software such as Quake[19] or EDAR[20], which search the datasets for subsequences that occur in low frequency, and therefore likely to be due to sequencing error. Some experiments may also only consider unique

matches to the reference sequence. The software Seal[21] was recently developed to evaluate alignment tools using simulated reads by testing them for correctly aligned reads, error rates and run-times. However, simulated data will not always correctly recreate the types of systematic biases that may be present within an actual dataset.

More recently, data post-alignment processing is being performed to filter erroneous sites or increase resolution of insertions and deletions (indels). For example, the Genome Analysis Toolkit (GATK)[22], which was used by the 1000 Genome Project Consortium found that a local realignment around indels could reduce misaligned reads and nearby false positive SNPs. In addition, GATK uses base and variant quality score recalibrations, whereby the quality scores of reads are re-adjusted according to several covariates over known variant sites as well as use expected transition/transversion ratios (Ti/Tv)[22]. The option of removing duplicate reads arising from a common progenitor DNA molecule have also been applied as both pre-[11] and post-processing steps[22], ensuring greater accuracy of subsequent SNP-calling.

Some alignment tools also have inbuilt genotyping such as Maq[15], which is able to call SNPs based on a Bayesian statistical model using an expected rate of heterozygotes within the genome and a dependency coefficient. However, most SNP-callers are used post-alignment, such as the UnifiedGenotyper of GATK[22], or the Sam/Bcftools' SNP caller[23] that considers a number of factors when determining the presence of a polymorphism such as minimum and maximum number of mismatches or filtering by Phred quality scores. Most SNP-callers consider homozygous and heterozygous (bi-allelic) sequences, whilst others such as Bcftools do not properly handle multi-allelic variants and only take the strongest non-reference allele. Many other SNP-callers have been developed, which may be tailored to data-types or expected levels of variation. The number of possible tools and their rate of development make benchmarking an issue that needs to be frequently readdressed.

Increasingly, panels of resequenced isolates include a reference strain for which an assembled genome is already available, in order to determine suitable depth of read coverage for the other isolates, and to refine the alignment and SNP-calling parameters to reveal acceptable levels of discrepancies. However, this approach still does not address divergence between the comparison strain and the rates of true positives that will later be called from the alignments. Furthermore, heterozygosity in diploid organisms is still difficult to verify, with few methods developed that do not rely on verification with previously identified sites. Although there have been numerous comparisons of DNA sequence assembly tools for length and accuracy of the contigs, running time, ability to resolve repetitive genomic elements[24–26] and numerous methods have been proposed for reducing sequencing errors[27,5,6] there is currently no standard method for assessing the accuracy of correctly identifying mutations from an alignment and SNP-calling strategy.

Here, we introduce a series of Perl scripts that can reveal, and compare the false discovery rates (FDR) for a given NGS dataset used for alignment and SNP calling, requiring only an available reference genome sequence that is closely related to the sequenced strain. Comparing the FDR of methods of alignment and SNP-calling simultaneously can reveal the best combination of tools. In addition to providing a method for determining FDR, we have developed our own tool for calling polymorphisms post-alignment based on cumulative binomial expectations for the number of reads agreeing with a polymorphism, given a depth and an expected error rate. These expectations are stored in lookup tables and used along with the Samtools mpileup format[23] as an input. The method is able to identify homozygous and heterozygous mutations with appropriate accuracy. We have benchmarked this method (Binomial SNP Caller from Pileup; BiSCaP), and others, using the scripts for assessing FDR.

Results

Using fungal NGS datasets and genomes from three separate phyla: *Saccharomyces cerevisiae* S288C (Ascomycota), *Batrachochytrium dendrobatidis* (*Bd*) JEL423 (Chytridiomycota) and *Puccinia triticina* race 1 isolate 1-1 (Basidiomycota), we assessed the accuracy of the alignment program BWA v0.5.9[8] with Samtools piped to Bcftools[20] by comparing their false discovery rate (cFDR) on 1nt/Kb test SNPs within the coding sequence (CDS) of their corresponding reference genome (workflow shown in Figure 1). We found that these tools resulted in a highly variable accuracy rate for both homozygous SNPs and heterozygous positions dependent on the input datasets, even after normalising the datasets for read length, depth of reads in the alignment and by only considering mutations that were found within the CDS regions (Fig. 2A).

Pre-processing input datasets was briefly assessed for ability to improve the accuracy of downstream SNP calls. Specifically, trimming 3′ ends from all *Bd* JEL423 reads increased the number of reads that were aligned using BWA nearly 5 fold from 2.2 million to 10.4 million, in turn, enabling $>$ 4 times the number of true positives SNPs and heterozygous positions to be called by Sam/Bcftools. Furthermore, read-trimming decreased the percent of false positives SNPs from 15.6% to 5.9% and false positive heterozygous positions from 14.4% to 3.4%, demonstrating how much error can, and was

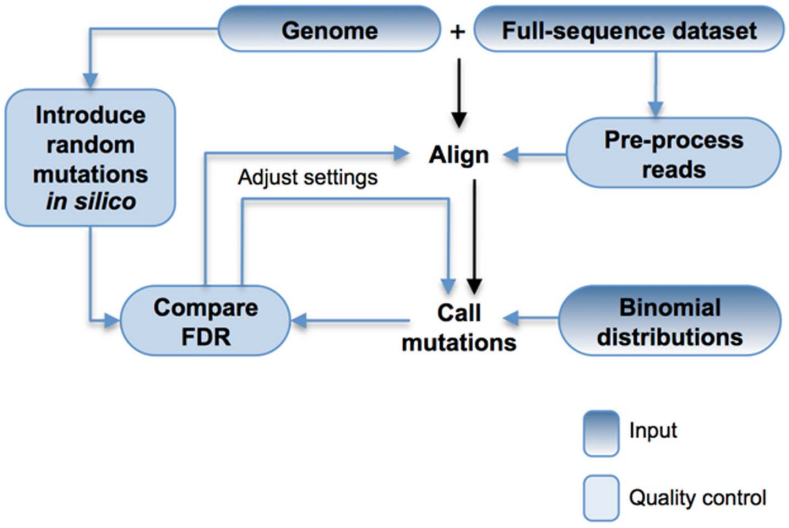


Figure 1 | A flow diagram showing the steps to verify false discovery rate and call polymorphic sites using cFDR and BiSCaP.

present within this NGS dataset, often at greater frequencies towards the 3′ ends of reads. Trimming the reads only achieved an improvement in the number of true positive and false positives with the *Bd* JEL423 reads, and conversely reduced both the number of false and true positives in the *P. triticina* and *S. cerevisiae* datasets. The variation between each of these datasets demonstrates the importance of assessing quality control as a preliminary step for resequencing projects.

To demonstrate how alignment and SNP-calling varied, we compared combinations of methods and parameters on the *Bd* JEL423 SOLiD 30-mer dataset (Fig. 2B). Owing primarily to the high levels of sequencing errors, even after removing low quality 3′ ends, none of the tested methods called > 73% true positives SNPs or < 5% false positive SNPs. The alignment program SHRiMP, which is specifically designed for the 'color-space' reads of SOLiD sequencing aligned 60% of the *Bd* reads and 70% after trimming to 30 mers, compared with BWA that aligned just 9% full length and 41% 30 mers. Despite the differences in the output alignment depth between the two programs, BWA (with the *Bd* 30 mers) resulted in a greater number of true positives and approximately equal number of false positives than SHRiMP called using either read length and any of the SNP-calling methods we tested. This comparison shows that BWA is a more appropriate tool for this particular dataset. The comparably low false discovery rate and number of reads aligned by BWA on the SOLiD *Bd* dataset was not found to the same extent in the Illumina *Sc* and *Pt* datasets, where > 59% full-length reads and > 68% trimmed reads were aligned to their reference sequence. For heterozygous base-calls, GATK2 had the greatest accuracy of any of the methods tested with 86.78% true positives and remarkably not a single false positive.

A comparison of the FDR for SNPs achieved by the alignment program BWA and the SNP calling method presented here (BiSCaP) using default settings on the *Bd* 30-mer dataset revealed 6582/12458 (52.83%) true positives and 494 (6.98%) false positive SNPs that were covered ≥ minimum required depth (MD). This result is a considerably preferable outcome compared to using SAM/Bcftools for SNP-calling (as shown in Fig. 2B), whilst the UnifiedGenotyper of GATK performed competatively in terms of specificity at a small expense in sensitivity compared with one of the tested settings of BiSCaP (MD 4). This variation of FDR demonstrates the importance of comparing candidate tools for a given study. Of the false positives called by BiSCaP, 328 were identified when aligning to the non-modified reference genome (351 total), and 321/328 were covered by 100% uniquely aligned reads, suggesting that they may be real

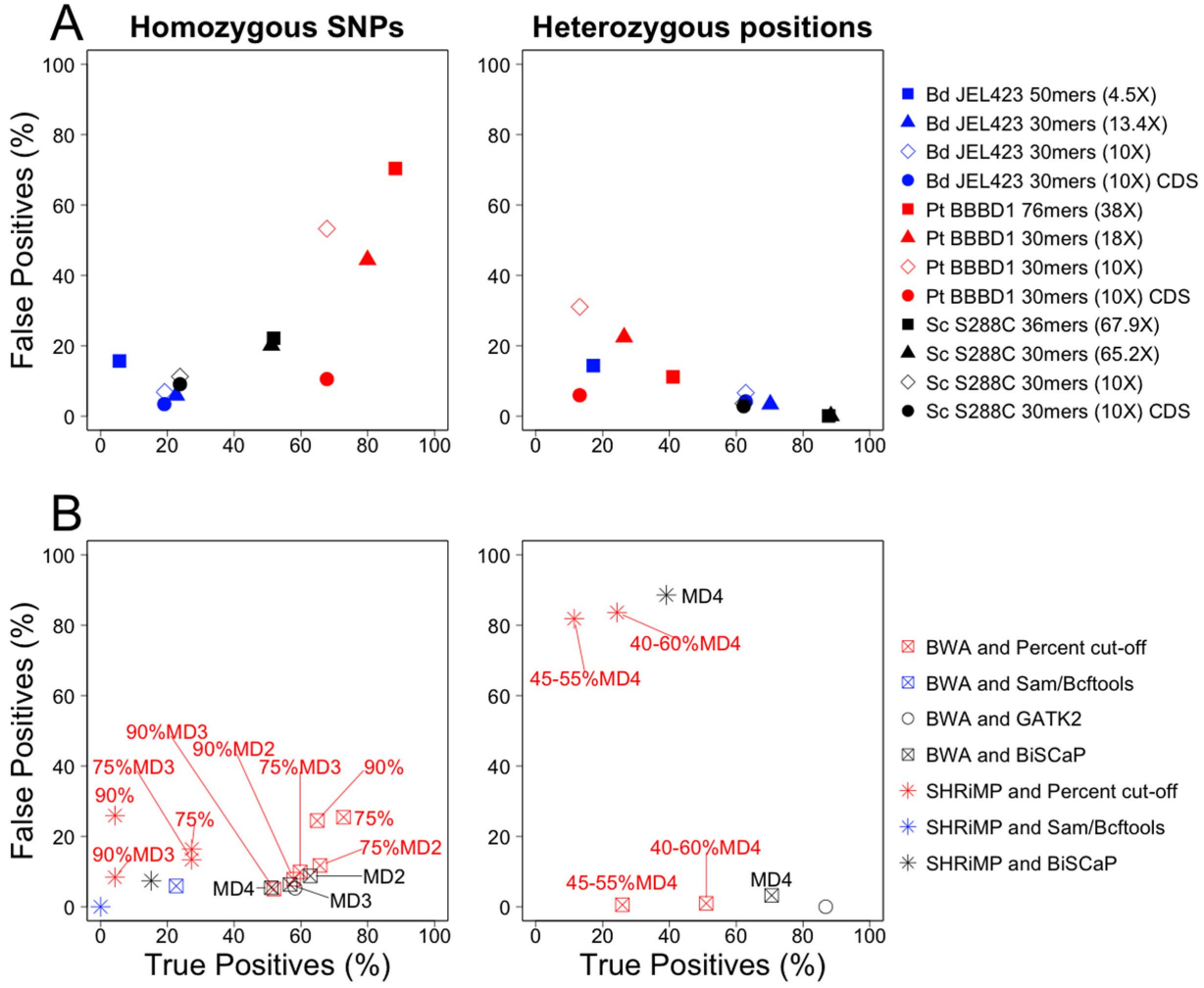


Figure 2 | False discovery rates for variants were ascertained using cFDR for three fungal NGS datasets. (A) Dataset-specific error rates were identified for both homozygous SNPs and heterozygous positions after alignment with BWA and calling SNPs using SAM/BCFtools with default settings, which persisted after trimming to 30 mers, aligning random 10× deep subsets of aligned reads, and only considering SNPs that fell over CDS regions thereby reducing genome size as a factor. (B) Combinations of alignments and SNP-calling methods resulted in different accuracies from the *Bd* JEL423 30 mer NGS dataset. Experimenting with a variety of methods can therefore reveal the most suitable method for a given dataset based on these metrics of accuracy. Sam/Bcftools only takes the strongest non-reference allele so was not included in the assessment of heterozygous accuracy. Parameters include the percent cut-off for inclusion as a SNP or heterozygous base and minimum depth (MD).

genetic changes occurring between the separate batches of isolate *Bd* JEL423, or genuine mistakes in the reference genome.

The remaining 45.83% of modified positions in *Bd* that were not identified, consisted of 98.5% that were uncovered by the minimum depth, 1.47% called as heterozygous or ambiguous if haploid setting is used, and a single incorrectly called homozygous SNP. Of the 84 bases that were incorrectly identified as heterozygous, 83 sites correctly called the modified base but also inferred the presence of the reference allele' and one consisted of the reference base and a deletion. None of these false positives were identified without first modifying the reference for cFDR, which may suggest they arise from misaligned reads. These results demonstrate that by far the greatest impact on alignment/SNP-calling error on the *Bd* 13.4× deep dataset arises from lack of coverage. Surprisingly however, the *S. cerevisiae* 67.9× deep dataset (Fig. 1A) also revealed a similar situation where 1920/1929 of the false negatives were due to below required depth compared to the remaining 9 that were incorrectly called as heterozygous positions.

To explain this pattern of enrichment for uncovered polymorphic sites (*Bd* and *S. cerevisiae* had 98.4% and >99.9% CDS covered respectively, compared to 45.8% and 15.7% of the introduced mutations within the CDS), we compared the distances between the introduced SNPs (Fig. 3A). Firstly, we found that *Sc* and *Pt* tended to have more closely associated introduced mutations (both had 19% occurring within 20 nt of each other) than *Bd* (only 2% within 20 nt). In terms of number of contigs and genes, number and average length of exons, and length of coding and introns sequence, *Bd* is situated between *Pt* and *Sc*. However, a clear difference is found in the number of genes selected by IRMS for introducing mutations in: 953, 1976 and 5701 for *Sc*, *Pt* and *Bd* respectively, which could be caused by one or more computational or biological differences such as differing numbers of genes specified with overlapping exons or splice variants specified in the feature file, which IRMS excludes. In any case, the difference in the number of modified genes likely explains the difference in distance between mutations. Despite this, false negatives were predominately more closely associated in *Bd* and *Pt*, whilst false negatives in *Sc* appeared to not be so clearly correlated with distance from other mutations.

We also looked at minimum depth cut-off points for different numbers of reads agreeing or disagreeing for a particular base, a parameter often taken into account by mutation callers, but compounded by variable depths over each of base in the genome. To examine this issue, we looked at the depth of read coverage over false positives and false negatives called using the percent cut-off methods and BiSCaP, which uses the depth-dependent cut-offs (Fig. 3B). As expected, homozygous erroneous base calls using either BiSCaP or percent cut-off methods were more frequently called over lower read depth positions in the alignment. BiSCaP achieved fewer false positive/negatives over the lowest read depths for *Bd* and *Sc*, and also over bases covered by more than 5 reads. The dataset *Pt* had fewest false base calls with the 90% cut-off method, which again highlights that different algorithms are suitable for different datasets, and only after testing them for FDR could you be certain of the success rate, or determine the most appropriate method.

Discussion

The rapid rate of increase in large-scale population studies using genome resequencing for SNP detection necessitates the development of improved tools to assess the quality of resequencing projects. Here we describe the development and efficacy of two such tools. We have tested the comparison of false discovery rate (cFDR) and Binomial SNP Caller from pileup (BiSCaP) scripts on sequence data from fungal genomes from three separate phyla: *Saccharomyces cerevisiae* S288C (Ascomycota), *Batrachochytrium dendrobatidis* (*Bd*) JEL423 (Chytridiomycota) and *Puccinia triticina* race 1 isolate 1-1 (Basidiomycota). These fungal genomes were chosen to represent a range of genome sizes and structures in terms of introns numbers, repeat richness and sequence heterogeneity. These methods can be applied to any resequencing study regardless of taxon. A large number of similar alignment, SNP-calling tools or pre-processing methods could have also been tested using these FDR scripts in addition to those we have tested here (BWA[11], SHRiMP[14], GATK[22], SAMTools[23]).

To identify the factors involved in the FDR variation between the three datasets composed of equal alignment depth, length, and modified genetic distance to its reference sequence, we extracted the positions corresponding to true and false positives. A large amount of variation between datasets, alignments, and SNP calling accuracy was identified using these tools, and was used to identify the most suitable combination of methods to accurately detect variants. Similar approaches to the generalised tools we present here have already been used by a number of NGS projects[28,29], and facilitated by the release of these packages, a wider range of projects could also make use of either, or both cFDR and BiSCaP. Each combination of

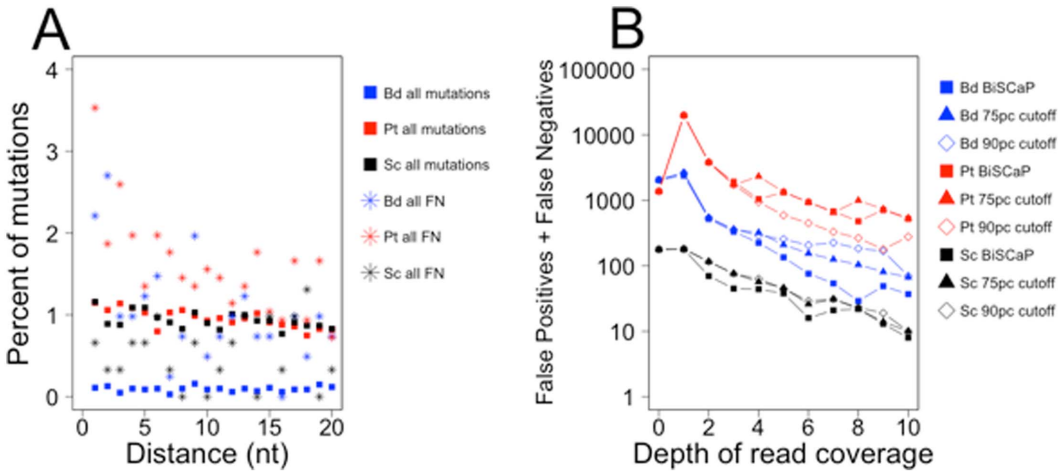


Figure 3 | Erroneous base calls (homozygous SNPs) from BiSCaP and percent cutoff methods were compared for proximity and depth of read coverage using full dataset 30 mer *Bd*, *Sc* and *Pt* reads. (A) False negatives were predominantly more closely associated for all 3 datasets (SNPs called with BiSCaP). False negatives were almost entirely caused by lack of coverage in each of these datasets demonstrating the most divergent part of each of the genomes is the most poorly resolved. (B) Homozygous errors were more frequently called over lower depth regions using strict cut-off methods than with BiSCaP.

methods performed differently on the *Bd* JEL423 genome with one simulated divergence rate, and from this test we could decide a single set of methods and parameters that performed optimumally. However, FDR validation and SNP-calling should be readdressed for every new dataset. For example, GATK would be most suitably benchmarked against a dataset for which a training set of known variation is available. However, even without variant quality score recalibration, GATK performed well on the fungal JEL423 genome, in particuar over bi-allelic heterozygotes.

Alternative methods for assessing homozygous variants that rely largely on simulated reads[21] or cross checking databases of polymorphisms[30] to determine FDR would have less power than our method in terms of realistic read and alignment error, or relying on resources that harbour their own sources of error. The alignment tool Maq[15] includes a command for introducing mutations into a reference sequence, with the intent of assessing alignment accuracy from simulating reads. Our method is able to make use of and assess real read data, which can harbor any number of platform or non-platform specific errors that may influence ability to call SNPs, in contrast to simulated reads or use of databases. This feature makes our method the only currently available technique to simultaneously assess the quality of data generated and the quality of methods used to analyze those data.

BiSCaP has been designed for variant-calling across haploid, diploid or triploid sequences, with corresponding binomial probabilities provided. In its current form, BiSCaP takes longer than the other assessed SNP-callers to complete: roughly one hour on a desktop computer on a modestly sized genome and dataset (the 13.4× *Bd* dataset to the 23 Mb genome), which is likely to persist until scripts are converted into a lower level programming language. The FDR method is also able to verify heterozygous alleles called by either GATK or BiSCaP using simulated reads, and the cFDR scripts finish running within a few minutes using the ouput from either of these (and SAMTools) SNP-calling tools. BiSCaP is able to call polymorphisms from standard input and output formats, making it a versatile tool for projects utalising these formats. We have not assessed how quality scores could be used to improve accuracy of those SNP-calls although both GATK and BiSCaP are able to filter potential SNPs based on these scores, so could be incorporated into the analysis.

Each of the methods presented here rely on a reference genome strain that is both high quality in terms of accurately assembled and with correct base calls, and has few discrepancies to the consensus (resequenced) isolate. For example, more distantly related isolates (reference and consensus) will result in a greater number of 'false positives' called by cFDR. Furthermore, if the reference sequence is poorly resolved (missing sequence or low quality or repetitive areas), the method may identify genuine polymorphisms that will also be considered false positives. This limitation can be partially resolved using a separate quality control measure for those false positives using either quality scores or called without first modifying the reference.

We found the ideas for, and implementation for SNP-calling based on cumulative binomial probabilities a suitable method for determining polymorphisms from an alignment. We tested both of these methods on three unique fungal pathogens, each of which are thought to be predominantly diploid, and therefore had both homozygous and heterozygous polymorphic positions called, which we found homozygous polymorphisms using the cFDR scripts to have a high level of accuracy. Either or both of these tools could be used with any other sequenced panel of diploid or haploid isolates to gauge the accuracy of alignment and SNP-calling.

Methods

The genome sequence and feature files for *Saccharomyces cerevisiae* S288C were downloaded from the SGD on 31.3.11 (http://www.yeastgenome.org/). Genomes of *Puccinia triticina* race 1 isolate 1-1 and *Batrachochytrium dendrobatidis* (*Bd*) JEL423 were downloaded from the Broad Institute (http://www.broadinstitute.org/). Illumina reads were obtained from the Short Read Archive under accessions SRR003681 and SRA009871 for *S. cerevisiae* S288C and *P. triticina* respectively. We previously resequenced the genome of *Bd* JEL423[29] using SOLiD, which is available for download in the Short Read Archive under accession SRA030504. Genome sequences were modified by randomly choosing and modifying 1 nt/Kb within the coding sequence (CDS) using a script that is part of the toolset (Introduce Random Mutation into Sequence; IRMS.pl).

SRA files were converted to FASTQ and aligned to their modified reference genome sequences using BWA v0.5.9[11] with default parameters and SHRiMP v2[11] with an 80% identity threshold for read alignment. Pileups were made using SAMTools v0.1.18[23] and polymorphisms called using the mpileup command piped to Bcftools v0.1.17-dev and filtered using vcfutils.pl with default parameters. In order to assess heterozygous variants, we randomly chose and modified 1 nt/Kb within the CDS using the "HET" setting of IRMS.pl, which first generates a duplicate (homologous) genome. We then simulated single-end reads from these modified sequences to the same depth as the 'real' data using simLibrary and simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/) using the default runfile (s_3_4x), which describes how "noise and cluster intensities are distributed in a real run of an Illumina machine", and aligned those reads to the non-modified reference genomes.

The Genome Analysis Toolkit (GATK) v2.1-9[22] was assessed according to the "Best Practice Variant Detection with the GATK v4, for release 2.0" detailed on the Broad Institute website. Briefly, Picard Tools v1.68 (http://picard.sourceforge.net) was first used for marking duplicates. Indel-realignment was performed using the GATK2 RealignerTargetCreator and IndelRealigner tools. Next, the UnifiedGenotyper was used to output raw variants that were used for base quality score recalibration (BaseRecalibrator and PrintReads). The UnifiedGenotyper was then assessed using default parameters on the new BAM file. Without a training dataset, variant recalibration is still considered experimental, so this step was left out for each of the three fungal genomes. Percent cut-offs with variable minimum read-depths and using the Binomial SNP caller from Pileup (BiSCaP) v0.11 (presented here) were also used to call polymorphisms. Each alignment/SNP-calling combination was assessed for accuracy using the Comparison of False Discovery Rate script (cFDR).

BiSCaP is based on the binomial expectations for the number of reads agreeing with a reference base over a given locus. These expectations allow for polymorphisms to be called with different levels of leniency for sequencing errors dependent on the depth of read coverage and for heterozygous positions to be called without a bias for the read depth. Briefly, an expected alignment and base calling error rate (e.g. 0.1 or 0.01) is used to generate a list of binomial probabilities p for sequencing and aligning k number of correct bases, given a read depth of n (number of trials) or $f(k; n, p)$. The probability a base is homozygous (h^1) can be considered as P (1 – error rate). In a diploid, the probability of a heterozygous base (h^2) can be considered as P (($h^1/2$) + ((error rate * 0.5)), where half the sequencing or alignment 'errors' are now specifying the two correct bases. Equally, a heterozygous allele in a triploid sequence (h^3) can be considered as either P (($h^1/3$) + ((error rate * 0.25)) or P (($h^1/3$) * 2) + ((error rate * 0. 5)). From these, a cumulative h^1 probability for the lower tail can be calculated (ch^1) in addition to the minimum values found from the cumulative h^2 or h^3 upper tail and the cumulative h^2 or h^3 lower tail (ch^2 and ch^3 respectively). Binomial values are generated by the script GBiD.pl (Generate Binomial Distributions) and stored in a lookup table. Pre-calculated tables for error rates of 0.1 and 0.01 up to a read depth of five hundred are provided in the current version. BiSCaP then uses one such look up table to infer the most probable consensus nucleotides from the alignment.

Briefly, the algorithm for determining the consensus sequence is to tally each of the four possible aligned bases, each of which needs to be ≥ the minimum read depth to be considered a consensus allele. The most common base is considered homozygous where $ch^1 \geq$ error rate and $ch^1 > ch^2$. The most common and 2nd most common base are considered heterozygous where $ch^2 \geq$ error rate and $ch^2 > ch^1$ for both bases. A triploid heterozygous site is considered when the ch^3 for each of the three most common bases ≥ error rate. Indels are treated separately but using the same criteria and sub routine.

BiSCaP by default provides details of polymorphic sites in Variant Call Format (VCF)[31] and can also output pileup lines into separate files based on the identified mutation-type. Other optional parameters include different lookup table based on error rate, minimum read depth, ploidy, stringency for heterozygous SNP calling, and a Phred quality score filter. If the read depth is greater than the lookup table depth (default 500 for error rates 0.1 and 0.01), reads up to the maximum lookup table depth can be used to determine the genotype (default), or printed to a separate file named 'outside-distribution'. The cFDR script considers Percent True Positive (TP) homozygous SNPs as ((N° TP hom. SNPs/N° Introduced mutations) × 100) and the Percent False Positive (FP) homozygous SNPs as ((N° FP hom. SNPs/(N° TP and FP hom. SNPs)) × 100).

1. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
3. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, 44, 623–630 (2012).
4. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina Sequencing libraries. *Genome Biol.* **12**, R18 (2011).

5. Sasson, A. & Michael1, T. P. Filtering error from SOLiD Output. *Bioinformatics* **26**, 849–850 (2010).
6. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).
7. Landan, G. & Graur, D. Characterization of pairwise and multiple sequence alignment errors. *Gene* **441**, 141–7 (2009).
8. Flicek, P. & Birney, E. Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**, S6-12 (2009).
9. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
10. Frith, M. C., Hamada, M. & Horton, P. Parameters for accurate genome alignment. *BMC Bioinformatics* 2010, **11**, 80 (2010).
11. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
12. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
13. Burrows, M. & Wheeler, D. J. A block-sorting lossless data compression algorithm. Technical report 124. Palo Alto, CA: Digital Equipment Corporation (1994).
14. Rumble, S. M. *et al.* SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Computational Biology* **5** (2009).
15. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–8 (2008).
16. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–7 (2009).
17. Lin, H., Zhang, Z., Zhang, M. Q., Ma, B. & Li, M. ZOOM! Zillions of oligos mapped. *Bioinformatics* **24**, 2431–7 (2008).
18. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* **11**, 473–483 (2010).
19. Kelley, D. R., Schatz, M. C. & Salzberg, S. L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology* **11**, R116 (2010).
20. Zhao, X. *et al.* EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol.* **17**, 1549–60 (2010).
21. Ruffalo, M., LaFramboise, T. & Koyutürk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790–6 (2011).
22. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010).
23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
24. Farrer, R. A., Kemen, E., Jones, J. D. & Studholme, D. J. *De novo* assembly of the Pseudomonas syringae pv. Syringae B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol. Lett* **1**, 103–111 (2009).
25. Miller, M. J. & Powell, J. I. A quantitative comparison of DNA sequence assembly programs. *J. Comput. Biol.* **1**, 257–69 (1994).
26. Zhang, W. *et al.* A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLoS One* **6** (2011).
27. Salmela, L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* **26**, 1284–90 (2011).
28. Raffaele, S. *et al.* Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science.* **330**, 1540–3 (2010).
29. Farrer, R. A. *et al.* Multiple emergence of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent lineage. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18732–6 (2011).
30. Musumeci, L. *et al.* Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Human mutation* **31**, 67–73 (2010).
31. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–8 (2011).

Acknowledgements

Funding: R.A.F. was funded by the Natural Environment Research Council (NERC). D.A.H. and M.C.F. were supported by the Wellcome Trust. No additional external funding received for this study.

Author contributions

R.A.F., D.A.H. and M.C.F. wrote the main manuscript text. R.A.F. prepared the figures. R.A.F., D.M. and D.J.S. conceived of the study.

Additional information

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

How to cite this article: Farrer, R.A., Henk, D.A., MacLean, D., Studholme, D.J. & Fisher, M.C. Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects. *Sci. Rep.* **3**, 1512; DOI:10.1038/srep01512 (2013).

Multiple emergences of genetically diverse amphibian-infecting chytrids include a globalized hypervirulent recombinant lineage

Rhys A. Farrer[a,b,1], Lucy A. Weinert[a], Jon Bielby[b], Trenton W. J. Garner[b], Francois Balloux[a], Frances Clare[a,b], Jaime Bosch[c], Andrew A. Cunningham[b], Che Weldon[d], Louis H. du Preez[d], Lucy Anderson[b], Sergei L. Kosakovsky Pond[e], Revital Shahar-Golan[a], Daniel A. Henk[a], and Matthew C. Fisher[a]

[a]Department Infectious Disease Epidemiology, Imperial College, London W2 1PG, United Kingdom; [b]Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom; [c]Museo Nacional de Ciencias Naturales, 28006 Madrid, Spain; [d]Unit for Environmental Sciences and Management, North-West University, Potchefstroom 2520, South Africa; and [e]Department of Medicine, University of California at San Diego, La Jolla, CA 92093

Edited* by David B. Wake, University of California, Berkeley, CA, and approved October 10, 2011 (received for review July 22, 2011)

***Batrachochytrium dendrobatidis* (*Bd*) is a globally ubiquitous fungal infection that has emerged to become a primary driver of amphibian biodiversity loss. Despite widespread effort to understand the emergence of this panzootic, the origins of the infection, its patterns of global spread, and principle mode of evolution remain largely unknown. Using comparative population genomics, we discovered three deeply diverged lineages of *Bd* associated with amphibians. Two of these lineages were found in multiple continents and are associated with known introductions by the amphibian trade. We found that isolates belonging to one clade, the global panzootic lineage (*Bd*GPL) have emerged across at least five continents during the 20th century and are associated with the onset of epizootics in North America, Central America, the Caribbean, Australia, and Europe. The two newly identified divergent lineages, Cape lineage (*Bd*CAPE) and Swiss lineage (*Bd*CH), were found to differ in morphological traits when compared against one another and *Bd*GPL, and we show that *Bd*GPL is hypervirulent. *Bd*GPL uniquely bears the hallmarks of genomic recombination, manifested as extensive intergenomic phylogenetic conflict and patchily distributed heterozygosity. We postulate that contact between previously genetically isolated allopatric populations of *Bd* may have allowed recombination to occur, resulting in the generation, spread, and invasion of the hypervirulent *Bd*GPL leading to contemporary disease-driven losses in amphibian biodiversity.**

chytridiomycosis | infectious disease | extinction | epidemiology

Emerging fungal diseases present a growing threat to the biodiversity of free-ranging animal species (1, 2). In recent years, a single species within a basal clade of fungi little recognized for their pathogenicity, the *Chytridiomycota*, has gained substantial notoriety owing to its impact on global amphibian biodiversity (1). *Batrachochytrium dendrobatidis* (*Bd*) is known to have driven the local extinction (extirpation) of up to 40% of species in affected communities (3), and to have spread rapidly through diverse environments (1). Despite widespread research efforts, the geographic origin of this emerging infection and its subsequent patterns of global spread remain largely unknown (2, 4, 5). For example, the hypothesis that *Bd* originated in Africa and spread via the global trade in *Xenopus* spp. during the first half of the 20th century (6) is disputed by the detection of lower genetic diversity in African isolates of *Bd* compared with isolates from North America (4). This observation, however, was based on only five African isolates collected from two sites in the South African Cape, compared with 29 US isolates collected from multiple sites across the United States. The genetically depauperate nature of African *Bd* has been further challenged by the discovery of isolates of African descent exhibiting pronounced differences in morphology and virulence (5, 7).

A separate marker-based study conducted by Morehouse et al. (8) using 10 loci from 35 isolates found very low levels of polymorphism (five variable positions) and fixed heterozygous sites, suggesting a primarily clonal mode of reproduction, although with some evidence for spatially localized genetic recombination (9). These results, and other marker-based studies (4, 6), support the novel pathogen hypothesis, by suggesting that *Bd* is a recently emerged pathogen (10). Although these results describe the population structure of *Bd* at a coarse scale, patchily sampled genomes combined with a chronic lack of genetic diversity at the sequenced loci have prevented a reliable inference of *Bd*'s evolutionary history. Recently, new whole-genome typing methods have greatly increased our ability to decipher genealogies by enabling unbiased sampling of the entire genome, thus increasing our power to date the coalescence of lineages and to identify recombination events. Here, we compare the whole genomes of 20 isolates of *Bd* to examine the recent evolutionary history of *Bd* and its patterns of global genome diversity.

Results

Using ABI's SOLiD system (sequencing by oligonucleotide ligation and detection), we achieved high-density coverage (mean 9.5× deep) of the 24-Mb genome for 20 globally distributed (Europe, North and Central America, South Africa, Australia) isolates of *Bd* from 11 amphibian host species. Eight isolates were from regions where epizootics have been documented (Table 1). We aligned the reads to the genome sequence of isolate JEL423 (http://www.broadinstitute.org/; 11) and searched for discrepancies using a depth-dependent binomial method (*SI Appendix*, Figs. S1 and S2), finding in total 51,915 nonredundant homozygous SNPs and 87,121 nonredundant heterozygous positions (*SI Appendix*, Figs. S3 and S4). Of these, 21% of the homozygous SNP positions and 19% of the heterozygous positions (22 Kb total) were covered ≥4 reads in all 20 samples and were used for phylogenetic analysis. Sixteen of the 20 samples, including the reference strain JEL423, were >99.9% genetically identical and fell within a single highly supported clade (Fig. 1 and *SI Appendix*, Fig. S5 and Table S1). This "global panzootic lineage" (*Bd*GPL) includes all previously genotyped isolates of *Bd* and all of the isolates in our panel that are associated with regional epizootics, recovered from five continents (4).

The remaining four newly sampled isolates form two novel, deeply divergent highly supported lineages. The "Cape lineage"

Author contributions: D.A.H. and M.C.F. designed research; R.A.F., L.A.W., J. Bielby, T.W.J.G., F.C., J. Bosch, A.A.C., C.W., L.H.d.P., L.A., R.S.-G., D.A.H., and M.C.F. performed research; J. Bosch, A.A.C., C.W., L.H.d.P., and S.L.K.P. contributed new reagents/analytic tools; R.A.F., L.A.W., J. Bielby, T.W.J.G., F.B., F.C., D.A.H., and M.C.F. analyzed data; and R.A.F., L.A.W., J. Bielby, T.W.J.G., F.B., D.A.H., and M.C.F. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: Sequences are deposited in the NCBI Short Read Archive (accession no. SRA030504).

[1]To whom correspondence should be addressed. E-mail: r.farrer09@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1111915108/-/DCSupplemental.

Table 1. The samples used and details of alignments

Collection site	Amphibian host	Year	Collector	Culture reference	Reads aligned (millions)	Depth (X)
America, Colorado	*B. boreas*	1999	JEL	JEL274	4.7	6.0
America, PtReyes	*R. catesbeiana*	1999	JEL	JEL270	9.4	12.0
Australia, Rockhampton	*L. caerulea*	1999	LB	Lcaerulae98	10.1	12.9
Canada, Quebec	*R. catesbeiana*	1999	JEL	JEL261	7.0	9.0
England, Kent	*L. vulgaris*	2007	MF	U.K.TvB	8.3	10.6
France, Ansabere	*A. obstetricans*	2007	MF	0711/1	6.3	8.1
France, Lac Arlet	*A. obstetricans*	2008	MF	PNP08489	6.0	7.7
Mallorca, Cocó de sa Bova	*A. muletensis*	2007	MF	CCB1	10.5	13.4
Mallorca, Torrent des Ferrerets	*A. muletensis*	2007	MF	TF5a1	4.3	5.4
Montserrat	*L. fallax*	2009	TG	L2203	5.8	7.5
Panama, Guabal	*P. lemur*	2004	JEL	JEL423	10.5	13.5
South Africa	*A. fuscigula*	2009	CW	MCT8	5.5	7.0
South Africa	*H. natalensis*	2009	CW	MC55	9.3	11.9
Spain, Ibon Acherito	*A. obstetricans*	2004	TG	IA042	6.9	8.8
Spain, Penalara	*A. obstetricans*	2002	MF	C2A	11.6	14.8
Spain, Valencia	*A. obstetricans*	2008	MF	VAo2	6.3	8.0
Spain, Valencia	*A. obstetricans*	2008	MF	VAo4	8.0	10.3
Spain, Valencia	*A. obstetricans*	2008	MF	VAo5	7.4	9.5
Spain, Valencia	*R. perezi*	2008	MF	VRp1	6.8	8.7
Switzerland	*A. obstetricans*	2007	TG	739	6.2	8.0

Bd isolates and locations that were resequenced. The first four columns provide information for the recommended naming scheme outlined by Berger et al. (18). The number of reads (millions) aligning to the *Bd* JEL423 genome assembly and the corresponding depth of coverage. Amphibian hosts include *Afrana fuscigula* (Cape river frog), *Alytes muletensis* (Mallorcan midwife toad), *Alytes obstetricans* (common midwife toad), *Bufo boreas* (Western toad), *Hadromophryne natalensis* (natal ghost frog), *Litoria caerulea* (green tree frog), *Lissotriton vulgaris* (smooth newt), *Litoria fallax* (Eastern dwarf tree frog), *Peltophryne lemur* (Puerto Rican toad), *Rana catesbeiana* (American bullfrog), *Rana perezi* (Iberian green frog). JEL, Joyce Longcore; LB, Lee Berger; MF, Matthew Fisher; TG, Trent Garner; and CW, Che Weldon.

(*Bd*CAPE) includes two isolated from the island of Mallorca, and one from the Cape Province, South Africa. This clade supports the hypothesis of Walker et al. (5) that spillover of *Bd* from captive Cape clawed frogs (*Xenopus gilli*) into Mallorcan midwife toads (*Alytes muletensis*) led to the introduction of *Bd* onto the island through a captive breeding and reintroduction program for this endangered species. The discovery of *Bd*CAPE as a separate lineage to previously genotyped isolates demonstrates that multiple emergences of amphibian chytridiomycosis have occurred, as well as confirming that the trade of amphibians is directly responsible for at least one of these emergences. A third novel lineage, "Swiss lineage," (*Bd*CH) is composed of a single isolate derived from a common midwife toad (*Alytes obstetricans*) from a pond near the village of Gamlikon, Switzerland. Further sampling is necessary to establish whether *Bd*CH is a European-endemic isolate or more broadly distributed.

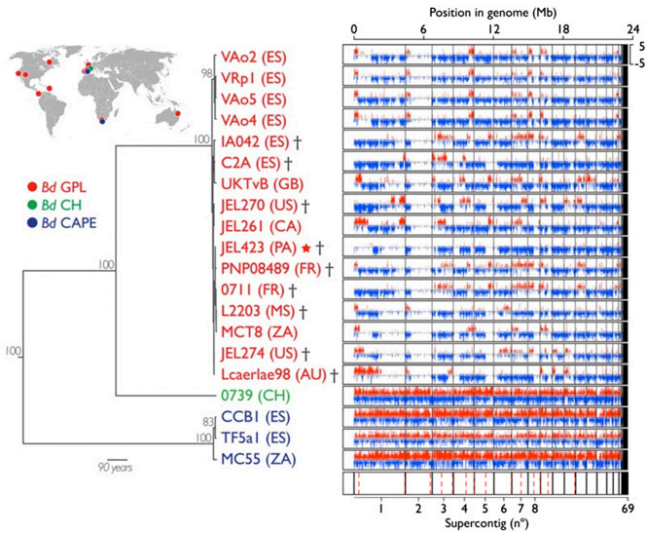


Fig. 1. Phylogenetic analysis of the 20 resequenced *Bd* mitochondrial genomes demonstrates three divergent lineages. The locations of the isolates belonging to the different lineages are shown using the same colors as in the phylogeny. Each genome is represented to the *Right* of the phylogeny. A nonoverlapping sliding window of SNPs minus heterozygous positions across the genome illustrates regions where heterozygosity predominates (blue) and where homozygosity predominates (red), illustrating the hallmark loss-of-heterozygosity events in the pan-global *Bd*GPL lineage. The block below the 20 genomes denotes the supercontigs with black lines and the GARD recombination breakpoints are shown in red dotted lines. The star signifies the reference genome JEL423; crosses represent isolates that have been recovered from epizootics.

To ascertain whether genomic features were associated with these lineages, we searched for evolutionary differences by looking for mutation biases, copy number variation (CNV), and genes undergoing diversifying selection in the *Bd*CAPE and *Bd*CH lineages (*SI Appendix*, Figs. S6–S13) compared with *Bd*GPL. We found that most (89%) of the polymorphisms in complementary determining sequence (CDS) regions caused a transition (A:T↔G:C) with a mean T_s/T_v of 8.03 with no lineage-specific biases. Synonymous mutations accounted for most of the polymorphisms in the CDS (ratio of 2.21 to nonsynonymous) concordant with the fact that most transitional mutations at twofold-degenerate sites are synonymous. Using the DoS measure of selection proposed by Stoletski and Eyre-Walker (12) and d_N/d_S ratios (13), we identified 114 genes in the *Bd*CAPE and *Bd*CH lineages that were significantly different from neutral expectations or had a $d_N/d_S > 1$ (*SI Appendix*, Fig. S9). Most of these genes had no significant blastp matches in the National Center for Biotechnology Information (NCBI) nr database and were otherwise indistinguishable from the rest of the transcriptome. We also detected no evidence for deviations in CNV between lineages using average read depth over each gene as a proxy.

Although no obvious interlineage variation was uncovered by these analyses, clear differences emerged when we examined the locations of the polymorphic sites within the genome. We identified a highly uneven distribution of homozygous SNPs and heterozygous positions across the 16 *Bd*GPL isolates (Fig. 1), which could be observed using a range of window sizes (*SI Appendix*,

Fig. S14). Furthermore, this pattern was found to be unique to the *Bd*GPL (*SI Appendix*, Fig. S15) and was relatively similar between isolates. For example, we observed a region of low heterozygosity spanning supercontig 2 that was common to all isolates within *Bd*GPL. To investigate whether this clustered distribution was due to recombination events, we tested whether the observed pattern was explained by multiple different evolutionary histories using the GARD (genetic algorithm for recombination detection) method (14). We identified 26 recombination breakpoints across the nuclear genome, 14 of which remained significant after Kishino–Hasegawa (KH) testing (*SI Appendix*, Fig. S16 and Table S3). Seven of these breakpoints were found within chromosomes and seven were found occurring between chromosomes. This suggests that the independent assortment of individual chromosomes expected under a model of frequent meiosis has not eroded the congruent phylogenetic signal between chromosomes, suggesting that meiosis is rare. Taken together, these data support the hypothesis of a single hybrid origin of *Bd*GPL via an ancestral meiosis as proposed by James et al. (4).

To further characterize the differences between the 15 significantly different recombination segments, we reconstructed separate topologies for each segment, but sharing the same model of evolution, using Bayesian phylogenetic inference techniques implemented in BEAST v.1.6.1 (15) (*SI Appendix*, Fig. S16). The three distinct lineages were recovered in all 15 segment trees, thereby ruling out recombination between them. In contrast, the topologies among recombination segments within *Bd*GPL were markedly different from one another. This pattern can be explained by two alternative biological explanations: (*i*) different segments have different genealogies reflecting meiotic events or (*ii*) a highly heterozygous ancestor has since undergone recurrent mitotic gene conversion events at different places in the genome (Fig. 2). Although we cannot formally reject either hypothesis, rare meiotic recombination is the more parsimonious explanation for the observed pattern. In particular, four of our sequenced isolates (VRp1, VAo4, VAo2, and VAo5) originate from the same pond (near Valencia, Spain) at the same time point. The mitochondrial tree (Fig. 1), which is arguably expected to reflect the species tree best given its nonrecombining nature, gives high posterior support for this group being monophyletic. It therefore seems likely that the Valencia isolates originated from a single ancestor introduced to the pond. However, only 10 of the 15 gene trees (shown in *SI Appendix*, Fig. S20) give posterior support for this group being monophyletic, with consistently different groupings observed for the other cases. For this pattern to be compatible with a model of gene conversion, as described by hypothesis 2, we would need to assume that the population that colonized the pond near Valencia had levels of diversity comparable to *Bd*GPL as a whole, and that this diversity was lost through rapid gene conversion only after the local colonization event. Distinguishing between these two hypotheses awaits the experimental determination of relative rates of meiotic versus mitotic recombination.

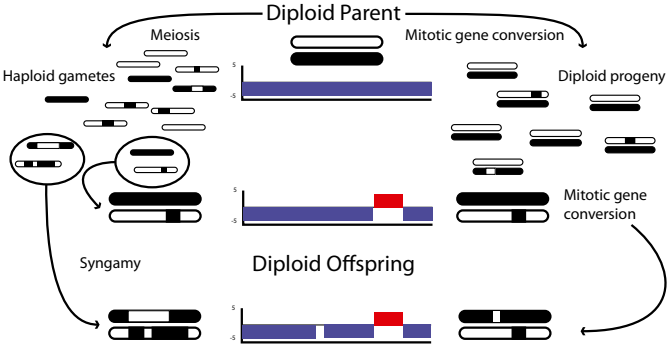


Fig. 2. Two possible mechanisms of achieving the uneven distribution of heterozygous and homozygous SNPs throughout the *Bd* genome. Each black and white bar represents a haplotype identified in a parental isolate, and the charts in the *Middle* represent the plots illustrating regions where heterozygosity predominates (blue) and where homozygosity predominates (red for homozygous SNPs and empty for homozygous identical to reference). On the *Left* of the diagram, meiosis generates recombinant haploid genomes that then are united via syngamy into new diploid offspring with patchy heterozygosity. Meiosis involving fusion of diploid gametes could result in similar patterns if chromosomal segregation remains independent. On the *Right* of the diagram, mitotic gene conversion generates patches of homozygous sites via homologous DNA repair in diploid progeny.

To ascertain whether *Bd* virulence recapitulated phylogeny, we experimentally exposed common toad (*Bufo bufo*) tadpoles to repeated, high-concentration doses of *Bd* using multiple isolates from the two lineages with replication (*Bd*GPL and *Bd*CAPE) and from two clusters within *Bd*GPL (Valencia and the Pyrenees). Toads exposed to *Bd*CAPE experienced significantly reduced infection and mortality compared with those exposed to *Bd*GPL (Fig. 3*A* and *SI Appendix*, Fig. S17). Significant variation in virulence was also detected among isolates within the *Bd*GPL (Fig. 1). Environmental conditions were standardized and animals were kept in isolation from each other precluding any effect of environmental forcing. The possibility does exist that the observed variation among treatments where *Bd*GPL isolates were used could be due to variation in host susceptibility, as mass at metamorphosis and infection dynamics of individual toadlets both played a role in determining mortality. Nevertheless, any effect of host, or the effects of variation among isolates from *Bd*GPL on virulence, is minor compared with the observed effect of lineage on postmetamorphic mortality.

The discovery of *Bd*CH came too late for this lineage to be included in our in vivo experimental framework. However, extensive population surveillance of *Bd* infected and uninfected populations across Switzerland has demonstrated a lack of association between infection status and population decline (http://www.bd-maps.net/maps); if *Bd*CH is representative of the lineages infecting other areas of Switzerland, then this is circumstantial evidence that *Bd*CH, like *Bd*CAPE, is hypovirulent. Previous studies of morphological variation among *Bd* isolates have shown that phenotypic profiles are linked to the virulence of isolates (7) and that isolates belonging to *Bd*CAPE exhibit smaller sporangial sizes compared with isolates belonging to *Bd*GPL. We repeated this study using a greater number of *Bd*GPL isolates from other continents and including *Bd*CH. These data (Fig. 3*B* and *SI Appendix*, Figs. S18 and S19) show that *Bd*CAPE does exhibit significantly smaller sporangia and increased hyphal length relative to *Bd*GPL; however, these characters are not diagnostic for the two lineages. *Bd*CH exhibits larger sporangia than the other two lineages, but no difference in hyphal length. Together, these data show that genetic differentiation between lineages of *Bd* has resulted in significant morphological variation. However, we did not detect differences in sensitivity to itraconazole among lineages (Fig. 3*C* and *SI Appendix*, Fig. S20 and Table S2).

Surveillance data shows that *Bd* is increasing its global range, and rapid emergence across geographic regions has been documented in the United States (16), Central America (17), and Australia (10). Eight of our sequenced isolates are associated with rapid population/community extirpations in diverse biomes across the New World, Europe, and Australia (3), exemplified by the rapid and virtual extirpation of the critically endangered mountain chicken frog *Leptodactylus fallax* in Montserrat by *Bd*GPL isolate L2203. The widespread occurrence of *Bd*GPL across five continents suggests that the emergence of this lineage is responsible for these *Bd*-driven collapses in diversity. To date the emergence of this lineage, we took advantage of the fact that the samples were collected over a period of 10 y and used this to inform a strict clock across the phylogenies of the 15 breakpoint segments (*SI Appendix*, Fig. S20). The median dates for the emergence of the global hypervirulent lineage range from 35 to 257 y before present (ybp) (*SI Appendix*, Table S4). Amphibian declines were first noted in the 1970s simultaneously in the United States (Sierra Nevadas), Central America, and Australia

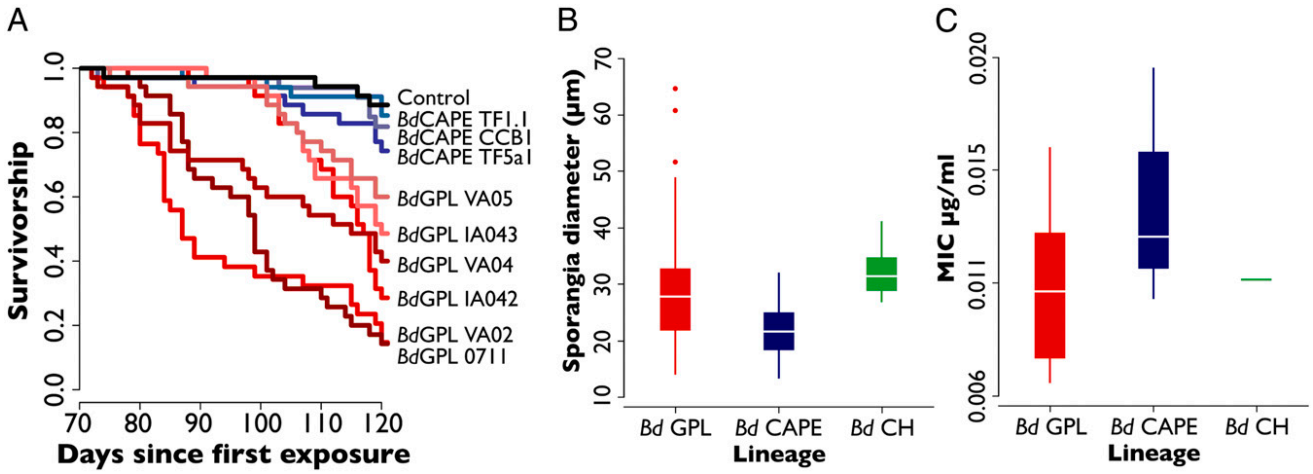


Fig. 3. Kaplan–Meier survival curves illustrating postmetamorphic survival of animals exposed to *Bd*GPL isolates (red) and *Bd*CAPE lineage isolates (blue). (*A*) *Bd*GPL isolates were significantly different from both *Bd*CAPE lineage isolates and the negative controls. (*B*) Significant differences in sporangia size between the three lineages. (*C*) No significant difference was observed between the three lineages in minimum inhibitory concentration values.

(18, 19); therefore, our dating suggests that the globalization of *Bd*GPL is compatible with its spread within the amphibian trade.

Discussion

One of the more puzzling aspects of the emergence of amphibian chytridiomycosis has been that, whereas epizootics have been widely observed, many susceptible amphibian communities apparently coexist alongside *Bd* with no evidence of disease. Such coexistence has been attributed to the context-dependent nature of susceptibility to disease, and whereas this is undoubtedly an important factor (20), our data show that *Bd* genotype is also an important epidemiological determinant. Here, we found that there is a much greater diversity of *Bd* than was previously recognized, and that multiple lineages are being vectored between continents by the trade of amphibians (*Bd*GPL and *Bd*CAPE). We have characterized hypervirulence in *Bd*GPL, suggesting that the emergence and spread of chytridiomycosis largely owes to the globalization of this recently emerged recombinant lineage (21).

Research has recently suggested the existence of a *Bd* lineage that is associated with the Japanese giant salamander, *Andrias japonicus* (22). This lineage, defined by sequencing a short fragment of the ribosomal DNA, is dissimilar to the rDNA sequences of *Bd*GPL, which nonnative North American bullfrogs have introduced to Japan. Therefore, it appears that we can now provisionally recognize at least four lineages of *Bd*, two of which are possibly endemic (*Bd*CH and Japan), one of which may have been previously endemic to South Africa but was then vectored to Mallorca (*Bd*CAPE), and one of which has a pan-global distribution (*Bd*GPL). This diversity was uncovered from sampling only 20 genomes from a cohort biased toward sampling amphibian populations experiencing chytridiomycosis (and hence infected with the *Bd*GPL); therefore, our data suggest that a more extensive diversity of amphibian-associated chytrid lineages is pending discovery. Whether these are *B. dendrobatidis* sensu stricto, or represent cryptic species, remains to be determined. However, observations that populations infected with non-*Bd*GPL lineages do not undergo epizootics suggest that the diversity of *Bd* represents a patchwork of genetically and phenotypically diverse lineages. In this case, determining the geographical origin(s) of the parental genotypes of *Bd*GPL will likely remain an elusive challenge until a much wider sample of amphibian-associated chytrid lineages has been sampled to identify the full phylogeographic range of diversity held within the order Rhizophydiales.

The origin of novel virulence in fungal species via recombination/hybridization is a well-recognized pathway underpinning disease emergence for increasing numbers of plant and animal pathogens (1, 23). Genomic rearrangement between allopatric fungal lineages that have not evolved reproductive barriers is promoted when anthropogenically mediated dispersal increases the rate of lineage mixing. The resulting novel interlineage recombinants exhibit a diversity of virulence profiles, some of which can initiate epidemics; contemporary examples include the evolution of hypervirulence in the Vancouver Island outbreak of *Cryptococcus gattii* (24), and many novel aggressive plant pathogens that increasingly threaten global food security (25) as well as natural populations (26).

We postulate that the anthropogenic mixing of allopatric lineages of *Bd* has led to the generation of the hypervirulent *Bd*GPL via an ancestral meiosis, and that, as previously suggested (9), this lineage is undergoing further diversification by either mitotic or sexual recombination. We show that the global trade in amphibians is resulting in contact and cross-transmission of *Bd* among previously naive host species, resulting in intercontinental pathogen spread. As the rate of interlineage recombination between fungi will be proportional to their contact rates, we predict that such globalization will increase the frequency that recombinant genotypes are generated. Theory and experimentation has shown that in genetically diverse infections, virulent lineages can have a competitive advantage, resulting in increased transmission (27, 28). As a consequence, we predict the evolution of further hypervirulent fungal lineages across a diverse range of host species and biomes in the absence of tighter biosecurity (1).

Materials and Methods

Full details are given in *SI Appendix*. Libraries were constructed according to the protocols provided by Life Technologies (Fragment Library kit). Fragment library sequencing was performed on two flowcells on an Applied Biosystems SOLiD 3 machine. Two pools of libraries were required: pool 1 contained libraries 1–8 with barcodes 1–8, and pool 2 contained libraries 9–20 with barcodes 1–12 (Table 1). Pooled barcoded libraries were united, and ePCR was performed according to Life Technologies' ("full scale") specification (templated bead preparation kits). After a run, a total amount of 260–290 million beads was loaded onto the flowcells. The output read length was 50 bp. The genome sequence and feature file for the chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*) strain JEL423 was downloaded from http://www.broadinstitute.org/ (GenBank project accession no. AATT00000000). The feature file for JEL423 had all but the longest splice variants removed for each gene leaving 8,794/8,819 genes. We trimmed the ABI SOLiD reads to 30 nt to remove low-quality bases from the 3′-end and aligned to the nuclear genome and mitochondrial sequence using Burrows-Wheeler Aligner (BWA) v0.5.8 (29) with default parameters.

The method we used for SNP calling was chosen after assessing the false discovery rate of 97 different settings and three separate methods, and was based on three depth-dependent binomial distributions (*SI Appendix*, Fig. S2*B*). For each base in the genome, we asked, given the number of bases agreeing with the reference base (k), and the depth of coverage (n), and a

set prior probability of finding the right base, 90% (p), what is its probability $f(k; n, p)$? We considered that over any given nucleotide, there are three potential circumstances, each with its own probabilities: (*i*) agree with reference/homozygous agree $P(0.9)$, (*ii*) heterozygous allele $P(0.45)$, and (*iii*) disagree with ref/homozygous SNP $P(0.03)$. Therefore, the nucleotide identified by which possible allele had the highest probability. In addition, we required a minimum depth of at least four reads to be considered different or the same as the reference (*SI Appendix*, Fig. S2*C*).

Trees were made using Bayesian Monte Carlo Markov chain (MCMC) analysis implemented by BEAST v1.6.1 (15), with the Hasegawa–Kishino–Yano substitution model with a γ-distribution of rates, under a strict clock. After discarding the first 10% as burn-in, we ran three separate chains of over 50 million generations, sampling every 1,000th generation. We combined the estimated topologies from the three different runs to construct the maximum clade consensus tree shown in Fig. 1. GARD analysis was run on the alignment of all variable and covered positions in the genome, with ambiguous character codes used to represent heterozygous SNPs. The analysis treated such positions as partially missing data during phylogenetic analyses. General time reversible (30) model of nucleotide substitution with a four-bin general discrete distribution to account for site-to-site rate variation was used to evaluate the goodness-of-fit for competing models. The best model was selected by small-sample corrected Akaike information criterion (31, 32). Because recombination was detected in the nuclear genomes, we used the unweighted pair group method with arithmetic mean (UPGMA) algorithm in phylogenetic analysis using parsimony (PAUP) to construct a dendrogram for the nuclear genome.

Experimental assessment of host response to challenge by *Bd* isolates were done according and subject to ethical review at the Institute of Zoology and Imperial College and licensed by the UK Home Office. Common toad (*B. bufo*) larvae (Gosner stage 24, Gosner 1960) hatched from 12 egg strings were randomly allocated to 1 of 10 experimental treatment groups (9 *Bd* isolates and 1 negative control). The experiment was undertaken in a climate-controlled (approximately 16 °C constant temperature) room with a 12:12 h day/night light schedule. Tadpoles in Bd^+ treatments were exposed to high doses (3,000–17,000 active zoospores per exposure in liquid media) of *Bd* zoospores every 4 d for a total of eight times. Zoospore counts and volume of media were standardized among isolates for each exposure. Tadpoles in the negative control treatment were exposed on the same schedule to an equivalent volume of sterile media as was used for Bd^+ treatments, but lacking any *Bd*. All animals surviving to the end of the experiment were humanely killed.

Photography of *Bd* used in assessing phenotypic differences between lineages was conducted using a Canon EOS 350D (3,456 × 2,304 pixel field). Initial photographs were taken of the floor of the tissue culture flask where sporangia had settled 3 d postculture. We subcultured the isolates into 12-well tissue culture plates (Nunc), with three replicas per isolate at a concentration of ∼8,000 zoospores per well (calculated using an improved Neubauer hemocytometer). The bottoms of the wells were photographed at 10, 15, and 20 d postinitial culture. Two to three images were obtained from each well for each isolate. Using ImageJ software (33) we measured the diameter of the largest sporangia contained in the field of view. We determined the MC_{50} to itraconazole for 12 isolates (estimated 5,750 *Bd* zoospores in 100 μL of culture) using concentrations of itraconazole ranging from 0.07 μg/mL to 2.1875 ng/mL (7). Each isolate was incubated at 18 °C and replicated three times. Optical densities were read on days 3, 5, 8, 10, and 12 using a BioTek Absorbance microplate reader (ELx808), using an absorbance of 450 nm.

ACKNOWLEDGMENTS. We thank Joyce Longcore for supplying isolates of *Bd* that we sequenced in this project, Beni Schmidt and Ursina Tobler for facilitating the collection of animals from which *Bd*CH was isolated, and Andrew Rambaut and Philippe Lemey for advice on the phylogenetic analysis. The government of Montserrat issued permits for the isolation and export of *Bd* from *L. fallax*. Computational analysis was supported by the University of California San Diego's Center for AIDS Research BEAST Core (National Institutes of Health AI 036214). SOLiD sequencing was performed by International Society for Systems Biology and Centre for Integrative Systems Biology, Imperial College. This project was funded by the UK Natural Environmental Research Council Grant NE/E006701/1, the UK Department for Environment, Food and Rural Affairs Grant FC1195, the Biotechnology and Biological Sciences Research Council Grant BB/H008802/1, the European Research Council Grant 260801-BIG_IDEA, and the Biodiversa project *RACE*: Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity (http://www.bd-maps.eu).

1. Fisher MC, Garner TWJ, Walker SF (2009) Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. *Annu Rev Microbiol* 63:291–310.
2. Frick WF, et al. (2010) An emerging disease causes regional population collapse of a common North American bat species. *Science* 329:679–682.
3. Crawford AJ, Lips KR, Bermingham E (2010) Epidemic disease decimates amphibian abundance, species diversity, and evolutionary history in the highlands of central Panama. *Proc Natl Acad Sci USA* 107:13777–13782.
4. James TY, et al. (2009) Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. *PLoS Pathog* 5:e1000458.
5. Walker SF, et al. (2008) Invasive pathogens threaten species recovery programs. *Curr Biol* 18:R853–R854.
6. Weldon C, du Preez LH, Hyatt AD, Muller R, Spears R (2004) Origin of the amphibian chytrid fungus. *Emerg Infect Dis* 10:2100–2105.
7. Fisher MC, et al. (2009) Proteomic and phenotypic profiling of the amphibian pathogen *Batrachochytrium dendrobatidis* shows that genotype is linked to virulence. *Mol Ecol* 18:415–429.
8. Morehouse EA, et al. (2003) Multilocus sequence typing suggests the chytrid pathogen of amphibians is a recently emerged clone. *Mol Ecol* 12:395–403.
9. Morgan JA, et al. (2007) Population genetics of the frog-killing fungus *Batrachochytrium dendrobatidis*. *Proc Natl Acad Sci USA* 104:13845–13850.
10. Skerratt LF, et al. (2007) Spread of chytridiomycosis has caused the rapid global decline and extinction of frogs. *EcoHealth* 4:125–134.
11. Fisher MC, Stajich J, Farrer RA (2012) Emergence of the chytrid fungus *Batrachochytrium dendrobatidis* and global amphibian declines. *Evolution of Virulence in Eukaryotic Microbes*, eds Heitman J, Sibley D, Howlett B (American Society for Microbiology, Washington, DC), in press.
12. Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. *Mol Biol Evol* 28:63–70.
13. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43.
14. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) GARD: A genetic algorithm for recombination detection. *Bioinformatics* 22:3096–3098.
15. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
16. Vredenburg VT, Knapp RA, Tunstall TS, Briggs CJ (2010) Dynamics of an emerging disease drive large-scale amphibian population extinctions. *Proc Natl Acad Sci USA* 107:9689–9694.
17. Lips KR, et al. (2006) Emerging infectious disease and the loss of biodiversity in a Neotropical amphibian community. *Proc Natl Acad Sci USA* 103:3165–3170.
18. Berger L, et al. (1998) Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *Proc Natl Acad Sci USA* 95:9031–9036.
19. Houlahan JE, Findlay CS, Schmidt BR, Meyer AH, Kuzmin SL (2000) Quantiative evidence for global amphibian population declines. *Nature* 412:499–500.
20. Walker SF, et al. (2010) Factors driving pathogenicity vs. prevalence of amphibian panzootic chytridiomycosis in Iberia. *Ecol Lett* 13:372–382.
21. Fisher MC, Garner TWJ (2007) The relationship between the introduction of *Batrachochytrium dendrobatidis*, the international trade in amphibians and introduced amphibian species. *Fungal Biol Rev* 21:2–9.
22. Goka K, et al. (2009) Amphibian chytridiomycosis in Japan: Distribution, haplotypes and possible route of entry into Japan. *Mol Ecol* 18:4757–4774.
23. Stukenbrock EH, McDonald BA (2008) The origins of plant pathogens in agro-ecosystems. *Annu Rev Phytopathol* 46:75–100.
24. Fraser JA, et al. (2005) Same-sex mating and the origin of the Vancouver Island Cryptococcus gattii outbreak. *Nature* 437:1360–1364.
25. Strange RN, Scott PR (2005) Plant disease: A threat to global food security. *Annu Rev Phytopathol* 43:83–116.
26. Goss EM, Carbone I, Grünwald NJ (2009) Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen *Phytophthora ramorum*. *Mol Ecol* 18:1161–1174.
27. Nowak MA, May RM (1994) Superinfection and the evolution of parasite virulence. *Proc Biol Sci* 255:81–89.
28. de Roode JC, et al. (2005) Virulence and competitive ability in genetically diverse malaria infections. *Proc Natl Acad Sci USA* 102:7624–7628.
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
30. Tavaré S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect Math Life Sci* 17:57–86.
31. Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm Statist Theory Methods* 7:13–26.
32. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23:1891–1901.
33. Abramoff MD, Magelhaes PJ, Ram SJ (2004) Image Processing with ImageJ. *Biophotonics International* 11:36–42.

OPEN ACCESS Freely available online

PLOS | GENETICS

Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity

Rhys A. Farrer[1,2]*, Daniel A. Henk[1], Trenton W. J. Garner[2], Francois Balloux[3], Douglas C. Woodhams[4], Matthew C. Fisher[1]*

1 The Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom, **2** Institute of Zoology, Zoological Society of London, London, United Kingdom, **3** Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, **4** Department of Ecology & Evolutionary Biology, University of Colorado, Boulder, Colorado, United States of America

Abstract

Pathogenic fungi constitute a growing threat to both plant and animal species on a global scale. Despite a clonal mode of reproduction dominating the population genetic structure of many fungi, putatively asexual species are known to adapt rapidly when confronted by efforts to control their growth and transmission. However, the mechanisms by which adaptive diversity is generated across a clonal background are often poorly understood. We sequenced a global panel of the emergent amphibian pathogen, *Batrachochytrium dendrobatidis* (*Bd*), to high depth and characterized rapidly changing features of its genome that we believe hold the key to the worldwide success of this organism. Our analyses show three processes that contribute to the generation of *de novo* diversity. Firstly, we show that the majority of wild isolates manifest chromosomal copy number variation that changes over short timescales. Secondly, we show that cryptic recombination occurs within all lineages of *Bd,* leading to large regions of the genome being in linkage equilibrium, and is preferentially associated with classes of genes of known importance for virulence in other pathosystems. Finally, we show that these classes of genes are under directional selection, and that this has predominantly targeted the Global Panzootic Lineage (*Bd*GPL). Our analyses show that *Bd* manifests an unusually dynamic genome that may have been shaped by its association with the amphibian host. The rates of variation that we document likely explain the high levels of phenotypic variability that have been reported for *Bd*, and suggests that the dynamic genome of this pathogen has contributed to its success across multiple biomes and host-species.

Citation: Farrer RA, Henk DA, Garner TWJ, Balloux F, Woodhams DC, et al. (2013) Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. PLoS Genet 9(8): e1003703. doi:10.1371/journal.pgen.1003703

Editor: Joseph Heitman, Duke University Medical Center, United States of America

Received December 20, 2012; **Accepted** June 21, 2013; **Published** August 15, 2013

Copyright: © 2013 Farrer et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This project was funded by the UK Natural Environmental Research Council (NERC) grant NE/E006701/1, the European Research Council (ERC) grant 260801-BIG_IDEA, the Swiss National Science Foundation grant 31-125099 and the Biodiversa project RACE: Risk Assessment of Chytridiomycosis to European Amphibian Biodiversity (http://www.bd-maps.eu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: r.farrer09@imperial.ac.uk (RAF); matthew.fisher@imperial.ac.uk (MCF)

Introduction

A diverse cadre of fungi and fungal-like oomycetes have recently taken centre stage as emerging infectious diseases (EIDs) owing to their increasing impact on animals, plants and wider ecosystem health [1]. The widespread emergence of this class of pathogens shows that they are able to successfully adapt to infect diverse hosts and ecological niches, suggesting that their genomes are able to respond rapidly to natural selection [1,2]. This idea finds widespread support; for example, horizontal transfer of whole chromosomes [3] and accelerated evolution across functional domains in effector genes [4] are associated with rapid host-adaptation and changes in virulence across lineages and species. Maintaining the pool of genetic diversity necessary to respond to selection is facilitated by the ability of fungi to utilise multiple reproductive modes, including cryptic recombination that enables inbreeding, outcrossing, hybridization, and the generation of diversity *via* parasexual mechanisms [5]. These features are suspected to have contributed to the rise of contemporary fungal EIDs, which play a major role in host population declines across a broad swathe of plant and animal species [1,6,7].

In recent years, whole genome sequencing has led to the characterization of novel mechanisms driving dynamic genome structure in microbial eukaryotes. In particular, it is increasingly apparent that pathogenic fungi manifest highly plastic genome architecture in the form of variable numbers of individual chromosomes, known as chromosomal copy-number variation (CCNV) or aneuploidy. This feature has been identified across the fungal phylum Ascomycota, ranging from *Botrytis cinerea* [8], *Histoplasma capsulatum* [9], *Saccharomyces cerevisiae* [10], *Candida albicans* [11] and the Basidiomycota *Cryptococcus neoformans* [12,13,14]. The mechanism(s) generating chromosomal CCNV in fungi are not yet well understood, but are thought to occur as a consequence of nondisjunction following meiotic or mitotic segregation [15], followed by selection operating to stabilise the chromosomal aneuploidies [13]. Although stress occurring as a consequence of either host response or exposure to antifungal

Author Summary

Pathogenic fungi constitute a growing threat to both plant and animal species on a global scale. However, many features of the fungal genome that enable them to successfully adapt to infect diverse hosts and ecological niches remain cryptic, especially for newly evolved emerging lineages. In this paper, we report three novel features of genome diversity linked to pathogenicity in the emerging amphibian pathogen, *Batrachochytrium dendrobatidis* (*Bd*). Firstly, we identified widespread chromosome copy number variation (CCNV) across our lineages, with individual isolates harboring between 2 to 5 copies of each chromosome and rapid rates of CCNV occurring in culture. In addition, by using *in vitro* divergence of replicate lines of *Bd*, we showed that changes in ploidy can occur within as few as 40 generations. Secondly, we identified uneven rates of recombination across the genomes and lineages, revealing hot spots in known classes of virulence factors. Finally we identified significant evidence of diversifying selection across the secretome of *Bd*, and showed that selection also targets putative virulence factors. These findings add to our knowledge of genome-dynamicity and modes of evolution manifested by eukaryote microbial pathogens, and may explain the varied phenotypic responses observed in *Bd*.

drugs has been linked to a rapid rate of CCNV in *Candida* [16], it is currently unclear to what extent this contributes to broader rates of CCNV in fungi. However, dynamic numbers of chromosomes could offer routes to potentially advantageous phenotypic changes *via* several mechanisms such as over expression of virulence-factors [13] or drug efflux pumps [17], the maintenance of diversity through homologous recombination [18], increased rates of mutation and larger effective population sizes [19], or by purging deleterious mutations through non-disjunction during chromosomal segregation [20]. Thus, CCNV likely represents an important, yet uncharacterized, source of *de novo* variation and adaptive potential in many fungi and other non-model eukaryote microbial pathogens.

A contemporary EID that is gaining substantial notoriety is the aquatic chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*), which has so far been identified in over 50 countries worldwide and infecting over 500 species of amphibians [21] (http://www.bd-maps.net). One of the most enigmatic aspects of *Bd*'s population genetic structure has been the low levels of genetic variation identified between globally distributed isolates. However, recent studies have shown the existence of up to five separate lineages [22,23,24], one of which is shown to have undergone a worldwide range expansion in the 20th Century. We recently compared the genomic diversity of this 'Global Panzootic Lineage' (*Bd*GPL) against that of a separate, distantly related (~1,000 ybp) lineage that appears to have originated in South Africa (named *Bd*CAPE), using SOLiD sequencing. *Bd*GPL was found to harbour evidence of historical recombination, manifested as patchily distributed heterozygosity, and phylogenetic incongruency across small spatial scales that we hypothesised has resulted from ongoing recombination [22]. Therefore, despite the lack of any known sexual meiotic mechanisms in its life cycle, *Bd* clearly has a more dynamic genome than a purely clonal, mitotic mode of reproduction would suggest. Here, we describe a new global panel of isolates that were subjected to high-depth Illumina sequencing in order to better understand cryptic genomic features that are associated with the rapid ascendancy of this pathogen.

Results

We sequenced 22 isolates of *Bd* with a geographical distribution spanning five continents to a high depth (52–195X; Table 1) using the Illumina HiSeq 2000 platform. Sequences are deposited in the NCBI Short Read Archive under the submission accession number SRA058657. These reads were then aligned to a reference sequence assembly for isolate JEL423 [25] using BWA [26] and polymorphisms were identified using BiSCaP [27] (Text S1: Optimization of alignments and SNP calling parameters). In total, we identified 904,000 SNPs, 761,000 bi-allelic heterozygous positions and 95,000 multi-allelic heterozygous positions (Table S1), which were distributed across 425,000 loci. Of those sites, 8,457 were homozygous across all of the 42 sequenced isolates (Illumina dataset presented here, SOLiD dataset presented previously [22]), which when combined resulted in concordant phylogenetic trees (Fig. S3). These phylogenies showed that our new panel of isolates belonged to three (*Bd*GPL, *Bd*CAPE, *Bd*CH) of the five suspected lineages of *Bd* [22,23,24] (Figs. S1, S2, S3) and extended both *Bd*GPL and *Bd*CAPE's known geographic range (*Bd*GPL into Switzerland and Ethiopia and *Bd*CAPE into France). Across the *Bd* JEL423 genome, 96% was covered by at least four reads in every isolate. Additionally, 65% of the total identified variant sites (275,000; 11.8 Kb^{1}) were called as either reference or polymorphic in all 22 isolates amounting to >10X the number of 'covered in all' polymorphic loci previously found using the ABI SOLiD 3 platform [22], owing to the higher depth of sequencing coverage (Table 1).

Chromosomal Copy Number Variation (CCNV)

Comparing the depth of read coverage over each chromosome using 10 Kb non-overlapping sliding windows revealed CCNV present in isolates belonging to all three lineages of *Bd* and affecting nine of the largest fifteen supercontigs (Figs. 1 and S4). *t*-tests on the mean depths across windows compared with those in the largest supercontig confirmed a significant increase in read-depth across 36 supercontigs, and a significant decrease in depth across 25 supercontigs in 18 of the 22 sequenced isolates (Fig. S5). To further verify relative ploidy within an isolate and the order of ploidy-changes, we inferred whether individual bases were 'evenly'- or 'oddly'-distributed across Illumina reads within a single genome by binning their frequencies into histograms for each chromosome. The expectation here is that a chromosome with an even ploidy will tend towards a 50:50 distribution across each single SNP, while chromosomes with an odd ploidy will tend towards a 33:66 or 33:33:33 ratio across SNP-calls (Figs. S5, S6). This method identified even- or odd-ploidies for 92% of the chromosomes tested with >95% bootstrap support (Table S2).

Thirteen *Bd*GPL and two *Bd*CAPE isolates had greater numbers of bi-alleles than tri-alleles (corresponding to an even ploidy that most parsimoniously corresponds to diploidy) (Table S2), and six isolates belonging to all three separate lineages that had greater numbers of tri-alleles than bi-alleles (corresponding to an odd ploidy that most parsimoniously corresponds to triploidy). The remaining four isolates (*Bd*GPL JEL423 & MODS27, *Bd*CAPE SA1d & SA4c) had significant *p*-values showing between 1–3 chromosomes in lower ploidy levels relative to the remaining bi-allelic genome. Over these lower-ploidy chromosomes we observed greater numbers of tri-alleles than bi-alleles and no decrease in heterozygous base-calls (both of which should occur if these chromosomes were haploid). We therefore conclude that these four isolates have tetraploid genomes with the identified losses in read-depth corresponding to chromosomes that have lost a single copy and are now trisomic.

Table 1. Samples used and details of alignments.

Collection site	Amphibian host	Year	Collector	Culture reference	Passage number	Sequenced depth (X)	Aligned depth (X)
Canada, Vancouver Island	*L. catesbeianus*	2009	PH	VC1 (CA)	>3	52.75	49.17
England, Cumbria	*E. calamita*	2010	PM	SFBC014 (GB)	2	115.32	106.96
Ethiopia, Hotcho	*A. enseticola*	2011	DG	ETH2 (ET)	2	68.90	58.66
Ethiopia, Telilia Stream near Rira	*Leptopelis sp.*	2011	DG	ETH4 (ET)	2	166.03	152.87
France, Lac d'Aule	*A. obstetricans*	2010	MF	AUL (FR)	2	195.89	175.48
France, Loire et Cher	*L. catesbeianus*	2010	CM	RC5.1 (FR)	3	85.28	67.65
France, Madamette	*A. obstetricans*	2010	MF	MAD (FR)	2	127.41	110.29
Mallorca, Torrent des Ferrerets	*A. muletensis*	2007	MF	TF5a1 (ES)	>3	150.33	133.33
Panama, Guabal	*P. lemur*	2004	JEL	JEL423 (PA)	>3	53.32	48.48
Sardinia, Affluente Pisharoni	*D. sardus*	2010	TG	AP15 (IT)	2	179.93	164.91
Sardinia, Monte Olia	*D. sardus*	2010	TG	MODS27 (IT)	2	52.84	49.17
Sardinia, Monte Olia	*D. sardus*	2010	TG	MODS28 (IT)	2	160.15	148.47
Sardinia, Scuponi	*D. sardus*	2010	TG	SP10 (IT)	2	129.60	115.46
South Africa, Mount, KZN	*A. vertebralis*	2010	TG	MG1 (ZA)	2	81.71	61.64
South Africa, Pinetown Kwazulu	*A. angolensis*	2011	TG	SA1d (ZA)	2/3	148.72	136.69
South Africa, Pinetown Kwazulu	*A. angolensis*	2010	TG	SA4c (ZA)	2	180.52	161.82
South Africa, SilverMine, KZN	*A. fuscigula*	2010	TG	MG4 (ZA)	2	131.16	122.30
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	ACON (CH)	>3	167.19	144.62
Switzerland, Gamlikon	*A. obstetricans*	2008	TG	APEP (CH)	>43	110.43	100.99
Switzerland, Gamlikon	*A. obstetricans*	2007	TG	CON2A (CH)	>43	115.29	102.17
Switzerland, Itingen	*A. obstetricans*	2010	RF	BLI1 (CH)	2	52.76	49.32
Switzerland, Waltisberg	*A. obstetricans*	2010	RF	BEW2 (CH)	2	144.54	132.07

Bd isolates and locations that were resequenced. The first 4 columns provide information for the recommended naming scheme outlined by Berger *et al.* [47]. Passage numbers are best approximations from records prior to DNA extractions in January and May 2011. The sequenced depth and aligned depth were calculated from the number of nucleotides in all or aligned reads respectively and divided by 24 Mb (the length of the *Bd* JEL423 genome assembly). All isolates represent novel sequences, apart from JEL423 and TF5a1 [22]. Amphibian hosts include *Afrixalus enseticola* (Ethiopian Banana frog), *Alytes muletensis* (Mallorcan Midwife Toad), *Alytes obstetricans* (Common Midwife Toad), *Amietia angolensis* (Angola River Frog), *Amietia fuscigula* (Cape River Frog), *Amietia vertebralis* (Ice Frog), *Discoglossus sardus* (Tyrrhenian Painted Frog), *Epidalea calamita* (Natterjack Toad), *Leptopelis sp.* (Big eyed Tree Frog), *Lithobates catesbeianus* (American Bullfrog), *Phyllomedusa lemur* (Lemur Leaf Frog). CM = Claude Miaud, DG = David Gower, JEL = Joyce Longcore, MF = Matthew Fisher, PH = Phineas Hamilton, PM = Peter Minting, RF = Rhys Farrer, TG = Trent Garner.
doi:10.1371/journal.pgen.1003703.t001

We were able to take advantage of replicate lines of *Bd*CH, which were passaged for 40 generations with and without exposure to skin antimicrobial peptides collected from the water frog *Pelophylax esculentus*. In these culture lines, the ancestral putatively triploid isolate (*Bd*CH ACON) differentially lost and gained copies of supercontig IV and V respectively when passaged without selection (*Bd*CH CON2A), and gained a copy of supercontig V following treatment with antimicrobial peptide (*Bd*CH APEP), which resulted in a significant reduction in mean growth inhibition (Text S1: *In vitro* Divergence of Independent Replicate Lines of *Bd*CH; Fig. 2). Due to the fact that most of our isolates exhibiting CCNV were sequenced shortly following isolation from nature without sequential passage, we know that CCNV is occurring frequently in both wild and cultured isolates. The rapidity that these mutations are accumulating across our isolates shows that aneuploidies in *Bd* are occurring at rates that will generate genome diversity within the timescale of a single host infection.

Recombination

In order to detect the presence and frequency of recombination events we determined the phase of bi-allelic heterozygous polymorphisms (Table S1, Figs. S7, S8, S9). We focused our attention on SNPs that were supported by a high percent of uniquely mapped reads (Table S1) and reads agreeing with the phasing (Fig. S9). By performing pairwise comparisons of shared phased positions between each of our isolates, we found >99% of these sites remained in the same phase for intra-lineage comparisons and >92% for inter-lineage comparisons (Fig. S10). However, we also identified 4,974 haplotypes demonstrating crossovers (Fig. S11) where all four pairwise combinations of bases were observed. Of these, 2,007 occurred at unique positions/loci in the genome. Every pair of isolates that we compared (except between *Bd*GPL isolates MAD (FR) and AUL (FR)) showed at least one haplotype that included an inferred crossover (Fig. S11). This was surprising given many of the isolates share a very recent common ancestor. For instance, we found that two isolates (MODS27 and MODS28) which were recovered from *Discoglossus sardus* at a single site in Sardinia on a single collection trip and are closely related (Fig. S2) had accumulated three crossovers. This shows that recombinant genotypes can accumulate even within highly-related free-living populations of *Bd*, a feature of this chytrid's population genetics that was first remarked upon by Morgan *et al.* in populations of Sierra Nevadan *Rana muscosa* [28].

The greatest proportion of phased positions demonstrating crossovers were found to occur between the three lineages, demonstrating an accumulation of recombinant haplotypes that scales with time of divergence (Fig. S11). For example, as many as 7.3% of the phased positions revealed crossovers that have

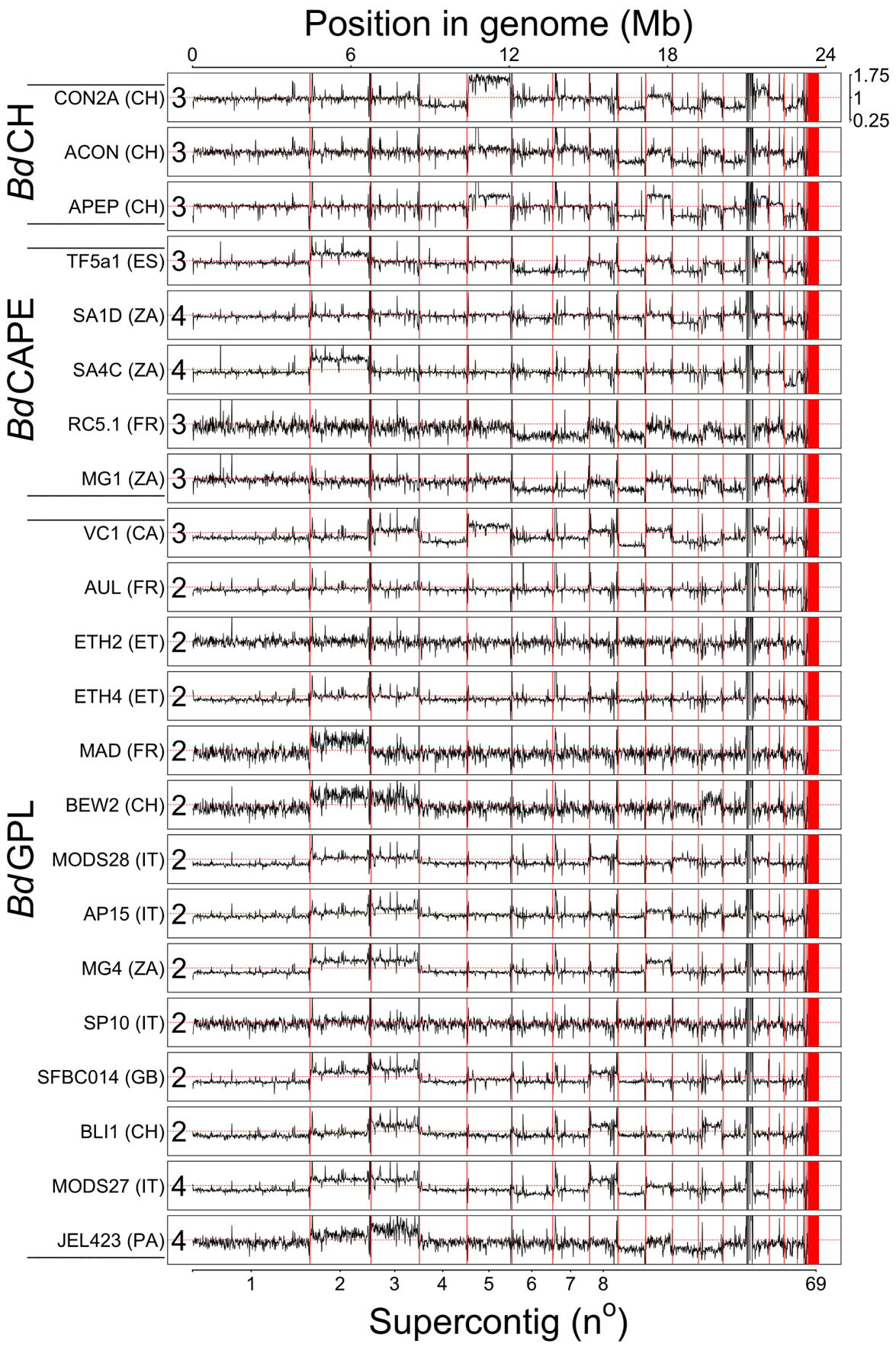

Position in genome (Mb)
0
6
12
18
24
1.75
1
0.25
BdCH
CON2A (CH) 3
ACON (CH) 3
APEP (CH) 3
BdCAPE
TF5a1 (ES) 3
SA1D (ZA) 4
SA4C (ZA) 4
RC5.1 (FR) 3
MG1 (ZA) 3
VC1 (CA) 3
BdGPL
AUL (FR) 2
ETH2 (ET) 2
ETH4 (ET) 2
MAD (FR) 2
BEW2 (CH) 2
MODS28 (IT) 2
AP15 (IT) 2
MG4 (ZA) 2
SP10 (IT) 2
SFBC014 (GB) 2
BLI1 (CH) 2
MODS27 (IT) 4
JEL423 (PA) 4
1
2
3
4
5
6
7
8
69
Supercontig (n°)

Figure 1. Read depth across 22 genomes was normalised by total alignment depth and plotted against location in the genome using a 10 Kb long non-overlapping sliding window. Base ploidy levels were determined using allele frequencies for supercontig 1 and shown at the start of each plot. Intra-chromosome read depth is largely consistent amongst the isolates, except over supercontig 14 due to a long stretch of rDNA. Shifts in read-depth between chromosomes demonstrate variation in chromosome copy number.
doi:10.1371/journal.pgen.1003703.g001

accumulated since isolates *Bd*GPL AP15 (IT) and *Bd*CH ACON (CH) were separated. Crossovers were identified in every major chromosome, and predominantly identified in intergenic regions (143 Mb^{-1} compared with 57 Mb^{-1} for coding regions and 65 Mb^{-1} for introns) (Figs. 3A, S12, S13).

Crossovers were also found to occur with a higher frequency amongst isolates belonging to the lineages *Bd*CAPE and *Bd*CH (between 0.6 and 1.1% of phased positions, respectively) compared with 0–0.2% in *Bd*GPL. This was a surprising finding given that the three *Bd*CH isolates were separated by only 40 passages in the lab and were derived from a single isolate that had been relatively recently isolated in 2007. This suggests two hypotheses: Either *in vitro* passage under selective conditions promotes rapid recombination, or our isolate of *Bd*CH is descended from a population of *Bd* that is more recombinogenic than *Bd*GPL. To further study the amount of recombination within lineages and between isolates, we extracted haplotypes that were phased across all of the isolates within a given lineage and contained at least two alleles per loci (ranging in length from 11 nt to 33.3 Kb: Fig. S12). Because only 35 haplotypes were retained for the entire panel of *Bd*GPL isolates, we also extracted haplotypes from two *Bd*GPL subsets consisting of 3 and 5 isolates respectively, thus allowing higher numbers of crossovers to be retained. From each of these sets of haplotypes, we calculated a multilocus measure of linkage disequilibrium (the standardised index of association rBarD [29]) and applied Hudson's four-gamete test [30] in order to quantify the amount of recombination amongst isolates within each lineage (Table 2). Across the *Bd*GPL groups, >30% of phased positions were in significant disequilibrium compared with 16% and 11% for *Bd*CH and *Bd*CAPE respectively. RbarD appeared to be robust against sample size differences, and gave values from *Bd*GPL values of 0.79–0.82 compared against 0.58 and 0.61 for *Bd*CH and *Bd*CAPE. Finally, a smaller proportion of *Bd*GPL subset haplotypes failed the four-gamete test compared with *Bd*CAPE or *Bd*CH isolates. Each of these findings shows that recombination is causing diversity within each of the lineages. However, the emergent *Bd*GPL is far more clonal than either of the other two lineages.

We next investigated whether recombination had occurred between these three lineages since their divergence, by calculating θ, Weir's [31] formulation of Wright's fixation index (F_{ST}) for pairwise comparisons of each lineage across window lengths of 1.4 Kb and 10 Kb (Figs. S14 and 3B). We found that all three lineages were highly differentiated from one another across each chromosome, with only minor intra-chromosomal regions of high similarity (which mainly comprised a long stretch of rDNA located at the start of chromosome 14). This indicates that recombination amongst these lineages has not occurred since their separation. We then determined whether certain categories of genes were associated with higher-than-average rates of recombination using *t*-tests on numbers of crossovers after accounting for differing levels

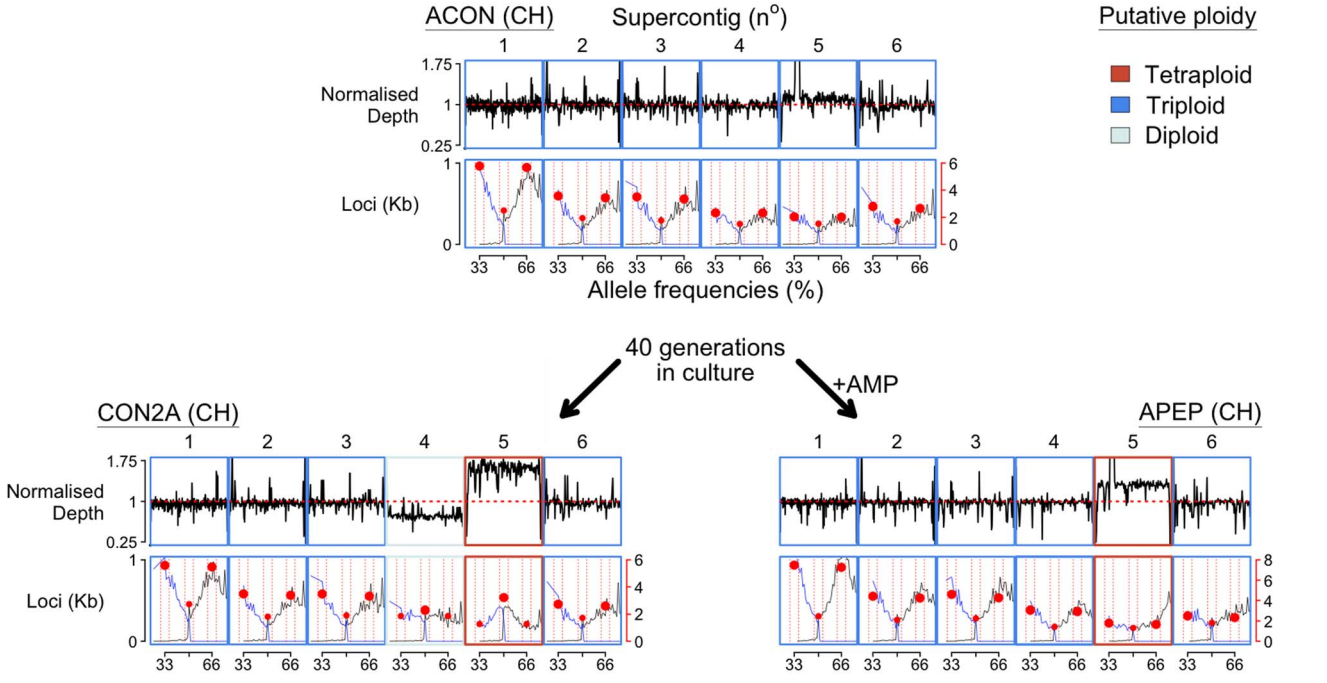


Figure 2. Chromosome copy number variation was identified across the three *Bd*CH isolates (ACON and its progenitors CON2A and APEP) following 40 generations in culture with or without the addition of anti-microbial peptides (AMP), respectively. Read depth is normalised to total alignment depth. A tally of all loci (per kilobase) with between 25–75% reads agreeing with the reference nucleotide are shown below, and summarised by the most common allele (black line), the second most common allele (blue line), and bins between 32–34, 49–51 and 65–67% (red circles). ACON is putatively triploid across the largest six supercontigs, whereas CON2A has lost a copy of supercontig IV and gained a copy of supercontigs V. APEP has gained a copy of supercontigs V.
doi:10.1371/journal.pgen.1003703.g002

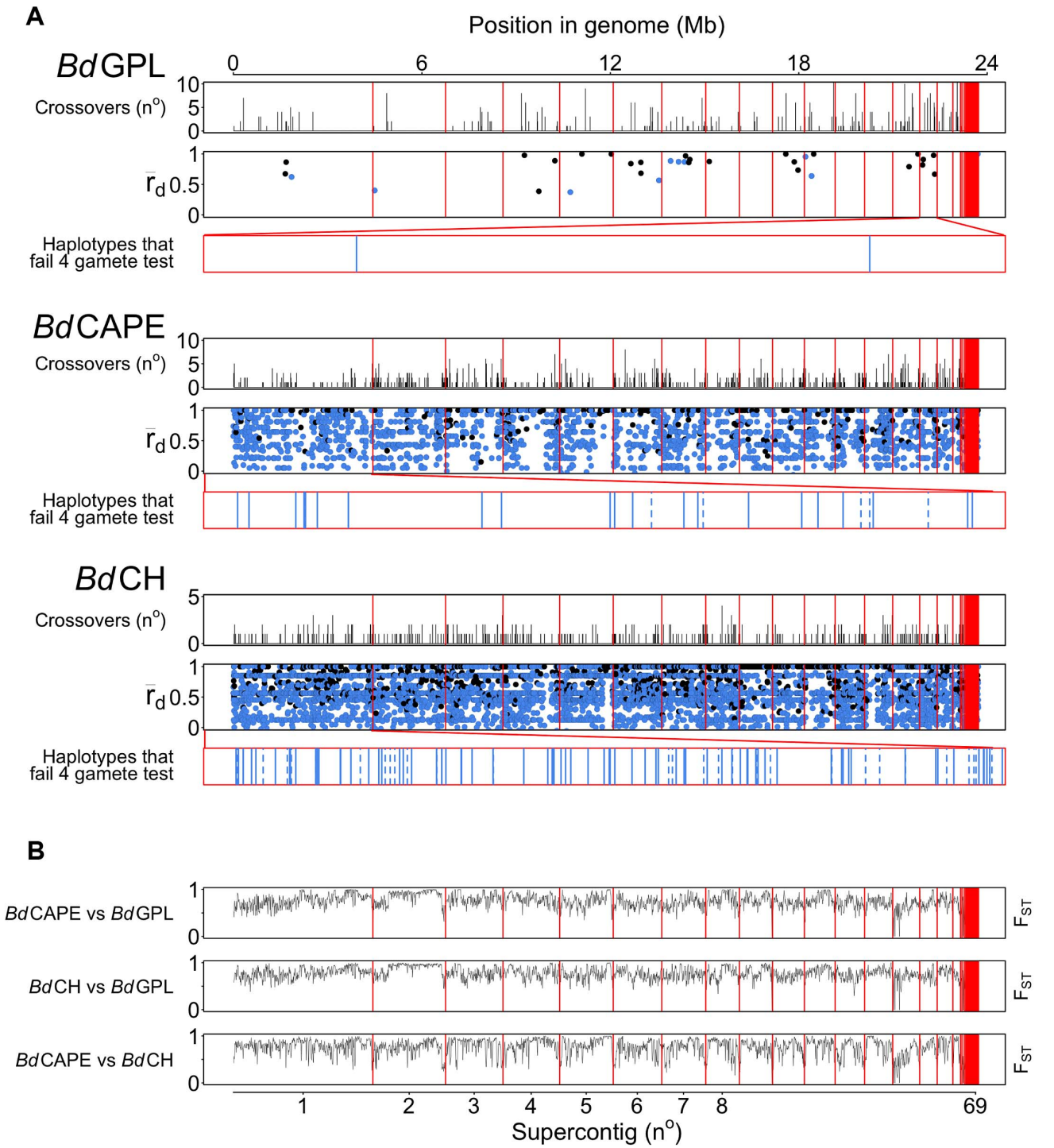


Figure 3. Crossovers were detected with pairwise comparisons for each *Bd* isolate across every supercontig. (A) Crossovers detected between isolates from each of the lineages were tallied and plotted across 10 Kb windows. rBarD was calculated for all haplotypes taken across phased regions of the genome. Haplotypes in linkage equilibrium are shown in blue and those in disequilibrium are shown in black. The supercontigs with the greatest number of haplotypes in linkage equilibrium are shown below rBarD values in red boxes. Haplotypes over genes are shown as a solid black line and haplotypes over intergenic regions are shown with a dotted line. (B) Fixation Indices (F_{ST}) were calculated between each of the lineages using 10 Kb windows revealing no strong evidence for introgression between each of the three lineages sequenced.
doi:10.1371/journal.pgen.1003703.g003

of heterozygosity and density of phased-sequences (Text S1: Identifying gene groups and names; Table S3, Figs. 4 and S15, S16). Surprisingly, we found only one group showing significant enrichment for crossovers: those showing homology to the C-terminal of the Crinkler (CRN) family of oomycete effector proteins found in the *Phytophthora* genus [32,33]. Enrichment was found in both *Bd*GPL and *Bd*CAPE, whilst not in *Bd*CH. Haplotypes that failed the four-gamete test were predominantly

Table 2. Haplotypes from isolates belonging to each of the separate lineages were tested for linkage disequilibrium using the index of association (I_A), rBarD and the 4-gamete test.

Lineage	Isolates	Haplotypes	Length (nucleotides)	Loci	Significant Disequilibrium (%)	Mean rBarD	Fail 4-gamete test (%)	Variable sites per locus
*Bd*GPL	14	35	4,409	95	68.57	0.82	7 (20%)	2 nt = 54 (56.84%) 3 nt = 28 (29.47%) 4 nt = 13 (13.68%)
*Bd*GPL subset 1	3	919	341,325	2,822	31.12	0.79	61 (6.64%)	2 nt = 2,118 (75.05%) 3 nt = 575 (20.38%) 4 nt = 129 (4.57%)
*Bd*GPL subset 2	5	438	83,414	1,232	41.32	0.82	36 (8.22%)	2 nt = 861 (69.89%) 3 nt = 301 (24.43%) 4 nt = 70 (5.68%)
*Bd*CAPE	5	2,275	952,307	7,212	11.47	0.61	197 (8.66%)	2 nt = 5,377 (74.55%) 3 nt = 1,709 (23.70%) 4 nt = 126 (1.75%)
*Bd*CH	3	5,215	1,537,742	16,612	16.36	0.58	655 (12.56%)	2 nt = 11,920 (71.76%) 3 nt = 4,184 (25.19%) 4 nt = 508 (3.06%)

To check differences between lineages were not resulting from different numbers of isolates, 2 subsets were made from *Bd*GPL. Subset 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27. For each isolate subset, the total length (in nucleotides) of all haplotypes and the total number of loci with ≥2 alleles is given. Over 30% of the *Bd*GPL haplotypes from any of the subsets were in significant disequilibrium, whilst only 11% of the haplotypes in *Bd*CH and 16% of the haplotypes in *Bd*CAPE were in disequilibrium, suggesting these populations are recombining more than the clonal *Bd*GPL. The numbers of variable sites per locus are also shown, demonstrating all lineages to be as likely to have arisen from out-crossing.

doi:10.1371/journal.pgen.1003703.t002

from coding-regions, but had no clear pattern of enrichment for any gene category (Table S4).

Patterns of mutation and selection

To identify genes that are present in the reference sequence and absent in our panel of isolates (presence/absence polymorphism), we examined the read-depth across each of the genes. Only five genes were identified from our panel (Table S5), including three amongst *Bd*CAPE isolates and two amongst *Bd*CH isolates. Therefore, whilst high-levels of aneuploidy are occurring, it does not appear to be resulting in frequent gene loss. To study the patterns of mutation across the nuclear genome, we categorized each of the mutations by their location in the genome in terms of coding regions (CDS), introns and intergenic regions (Table S6). In every isolate we sequenced, every variant type was found in greater abundance per kilobase in the non-coding regions (with the exception of 0.01 Kb^{-1} fewer heterozygous positions in the introns compared with the CDS for isolate MG1). This overall pattern can be explained through selection purging deleterious mutations from the CDS. In addition, we found homozygous polymorphisms to be highly supported in all lineages in terms of uniquely mapped reads, whilst un-phased bi-allelic heterozygous positions had a smaller total proportion in the divergent lineages compared with *Bd*GPL, suggesting some heterozygous positions may be miscalled due to paralogs.

We categorised each of the mutations within the CDS into synonymous and non-synonymous mutations (Table S6). SNPs were responsible for 169,000 synonymous changes and 197,000 non-synonymous changes. Genes with putative roles in pathogenicity were grouped by searching for secretion signals, protease domains and carbohydrate binding domains (Text S1: Identifying gene groups and names), and tested each of these for enrichment of homozygous SNPs (Tables S7 and S8) and heterozygous positions (Table S9) using hypergeometric tests. We found that gene groups that carried a secretion signal (proteases, chitin-binding and uncharacterized secreted) as well as CRN-like genes,

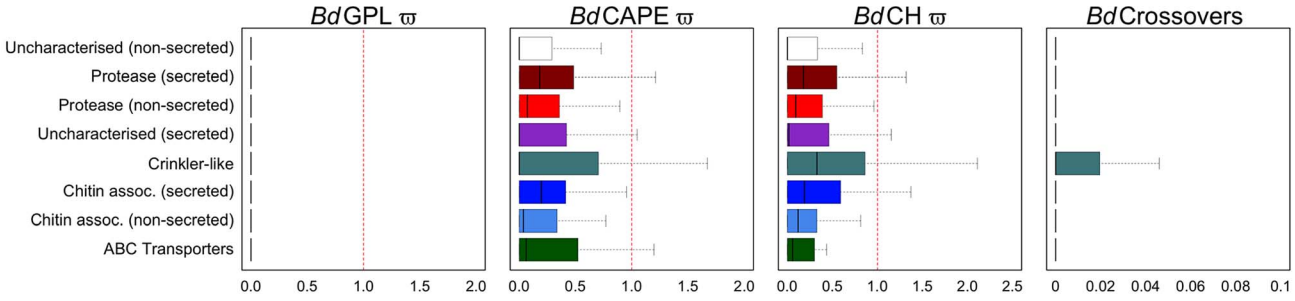


Figure 4. Boxplots for eight non-overlapping gene categories comprising every gene were compared for ratios of non-synonymous to synonymous mutations for each of the three lineages (dN/dS) and numbers of crossovers per phased positions (PP) within each gene (≥2PP) for all isolates (outliers omitted for both). Proteases and chitin-associated genes with predicted signal peptides had greater *dN/dS* ratios than those without for both *Bd*CAPE and *Bd*CH. CRN-like genes had the greatest upper quartile and upper tail showing these to be the most variable genes in the genome.

doi:10.1371/journal.pgen.1003703.g004

were significantly enriched for both homozygous and heterozygous polymorphisms relative to the whole set of genes. Predicted chitin-binding proteins that lacked a secretion peptide were not enriched for either homozygous SNPs or heterozygous positions (Tables S7 and S9), and non-secreted proteases were only enriched for synonymous amino acid changes. Conversely, CRN-like genes are only enriched for non-synonymous homozygous SNPs and not synonymous SNPs.

We next measured the rates of synonymous substitution (*dS*), non-synonymous substitution (*dN*) and omega ($dN/dS = \omega$) for every gene in every isolate and compared values by grouping isolates into their lineages (Fig. 4, Table S10). In total, we identified 1,450 genes with $\omega \geq 1$ in at least one of our isolates (*Bd*CAPE = 816; *Bd*CH = 746; *Bd*GPL = 283), suggesting positive or diversifying selection. Although no clear pattern could be distinguished within *Bd*GPL (Fig. S17) owing to the high degree of relatedness amongst isolates and thus relative paucity of polymorphism, CRN-like genes in both *Bd*CAPE and *Bd*CH had the greatest median, upper quartile and upper tail values of omega (Fig. S18). In addition, average ω values for secreted chitin-associated genes and proteases were marginally higher than their non-secreted counter parts. Uncharacterized secreted genes also had a greater ω than either of those non-secreted gene groups. Finally, a significant enrichment of both CRN-like genes and uncharacterized (secreted) genes with $\omega \geq 1$ were identified in both *Bd*CAPE and *Bd*CH (Table S11).

By analysing each of these 1,450 genes with $\omega \geq 1$using branch site models (BSM) in PAML along each of the three lineages of *Bd*, we identified a subset of 482 genes that show evidence for positive selection in at least one of the lineages. For *Bd*CAPE and *Bd*CH, a greater percent of each of the secreted gene categories were found to have accumulated an excess of non-synonymous mutations compared with their non-secreted counterpart gene categories (Table S11). Nine genes were also identified in all three lineages (Fig. S19), including four uncharacterized secreted and five uncharacterised non-secreted genes. However, the most striking finding of this analysis was found among *Bd*GPL isolates where 349/482 (72%) of the genes showed a signature of positive selection compared with only 23% for each of the other two lineages. This finding suggests that *Bd*GPL has been undergoing greater levels of positive selection than either *Bd*CAPE or *Bd*CH, despite the low numbers of sites under selection owing to the high levels of relatedness within this lineage.

Discussion

Recent studies have attributed aspects of *Bd*'s pathogenesis to the presence of a number of putative virulence factors that include proteases and chitin-binding proteins [32,33,34]. The former category contain M36 or S41 domains that are thought to degrade host-cellular components, and these protease families are known to have undergone extensive expansions in *Bd* since its divergence from free-living saprobes such as *Homolaphlyctis polyrhiza* [32]. Chitin binding proteins are thought to be involved in pathogenesis by allowing *Bd* to bind to keratinized host cells and to subsequently enter the host cells [34]. To date, the functional nature of the crinkler-like family in *Bd* has only been inferred owing to their homology to host-translocated proteins of known virulence in oomycetes [33]. Our data show that, across this global panel of 22 isolates and three lineages, the secretome and crinkler-like family of genes manifest higher diversity of homozygous and heterozygous SNPs, enrichment for non-synonymous mutations and greater dN/dS (ϖ) ratios when compared against classes of genes that do not contain a signal peptide. This shows that these gene families are evolving most rapidly in *Bd*, and that gene-products that interact with the amphibian host are undergoing diversifying (or reduced purifying) selection when compared with those gene-products that remain intracellular. Our findings suggest that *Bd* has had an evolutionary association with amphibians that predates the radiation of the lineages that we have characterised here, and is further evidence that this chytrid has an obligate rather than an opportunistic association with its amphibian hosts.

By mapping read-depth and SNPs across these genomes, we discovered that widespread genomic variation occurs within and amongst *Bd* isolates from the level of SNPs up to heterogeneity in ploidy amongst genomes and amongst chromosomes within a single genome. Individuals from all three lineages harboured CCNV along with predominantly or even entirely diploid, triploid and tetraploid genomes. Recent research by Rosenblum *et al.* [35] has also identified widespread CCNV across diverse lineages of *Bd* recovered largely from infected amphibians in the Americas, including a single haploid chromosome in isolate *Bd*GPL JEL289. This variation may itself, reflect only part of the full diversity in *Bd*,pathogensas +2/+3 shifts in ploidy, whole genomes in tetraploid, or chromosomes in pentaploid or greater, may occur and await discovery. Chromosomal genotype was shown to be highly plastic as significant changes in CCNV occurred in as few as 40 generations in culture. It is not known whether other chytrid species also undergo CCNV, or if this is a unique feature of *Bd* and hence may be intrinsic to its parasitic mode of life. Currently, CCNV is known to occur in a variety of protist microbial pathogens, including fungi, however it is currently not known whether this genomic-feature is specific to a parasitic life-style, or is a more general feature of eukaryote microbes; identifying the ubiquity of CCNV or otherwise across nonpathogenic species will therefore be of great interest. Further, the manner in which the plasticity of CCNV in *Bd* affects patterns of global transcription and hence the phenotype of each isolate also remains to be studied. However, it is clear from research on yeast, *Candida* and *Cryptococcus*, that CCNV significantly contributes to generating altered transcriptomic profiles, phenotypic diversity and rates of adaptive evolution even in the face of quantifiable costs; understanding the relationship between CCNV and *Bd*-phenotype will therefore likely be key to understanding its patterns of evolution at both micro- and macro-scales.

Whilst differing numbers of individual chromosomes presents a potential barrier to the standard model of meiosis, homologous recombination may still be occurring *via* mitotic processes within compatible genomes. In order to study recombination amongst our isolates, we determined the phase of our reads and constructed haplotypes that were suitable for traditional population genetic tests. This showed that, whilst the majority of the genomes from all three lineages manifest widespread linkage disequilibrium, recombination could still be detected across each chromosome and in all genomes. Crossovers (measured both as the proportion of SNPs that change phase and the numbers of haplotypes failing Hudson's four-gamete test) were found to occur much more frequently within the *Bd*CAPE and *Bd*CH lineages compared to *Bd*GPL, and these two lineages accordingly manifest lower average linkage disequilibrium. All of the *Bd*CH genomes that we sequenced stem from a single isolate collected in 2007. This suggests that either the high rates seen here have accrued since the isolate was taken into culture (suggesting a very rapid rate of *in vitro* recombination), or that we are characterising recombination events that occurred *prior* to the isolation of *Bd*CH and are segregating as a consequence of the multiple-ploidy nature of *Bd*. In support of the latter hypothesis, comparisons between population-level data for *Bd*GPL and *Bd*CAPE show that *Bd*GPL is far less recombinogenic and has

been undergoing a largely clonal expansion since its emergence, consistent with previous observations made by James *et al.* [36]. These data suggests that the global *Bd*GPL population is derived from a less recombinogenic ancestor than either *Bd*CH or *Bd*CAPE, that contemporary recombination is not occurring at a rapid rate and, where it occurs, is the result of a selfing rather than outcrossing events.

The discovery of a lower proportion of variable sites across haplotypes in addition to the lower proportion of heterozygous positions in *Bd*GPL compared against *Bd*CAPE or *Bd*CH does not support the notion that *Bd*GPL is an outbred hybrid lineage as previously proposed [22]. The discovery of a new *Bd* lineage found in Brazil (*Bd*Brazil) along with an isolate that is a likely *Bd*GPL/*Bd*Brazil recombinant [24] strongly implies that *Bd* retains the ability to outcross, despite having a primarily clonal genome and life cycle. However, values of F_{ST} across our dataset show no introgression between the three lineages; this demonstrates that they have remained largely separate since their divergence and suggests that outcrossing between lineages of *Bd* is rare or, if it has occurs, remains spatially restricted. Further broad-scale collections of isolates and extension of our comparative-population genomic analyses will allow the assignment of more accurate rates of introgression across evolutionary timescales.

We show that rates of recombination are uneven across the genome, with CRN-like genes enriched for crossovers, suggests that either CRN-like genes might have features that favour recombination or that recombinants of these genes have a fitness advantage and are thus more likely to reach fixation than recombinants at other locations in the genome. CRNs were also enriched for non-synonymous polymorphisms, are characterised by a signal of directional selection, and are amongst the most polymorphic genes in *Bd*'s genome. Within the oomycete genus *Phytophthora*, CRNs manifest diverse carboxy-terminal domains and high rates of homologous recombination targeted to the conserved HVLVXXP motif, suggesting that the mosaic domains of CRNs are being shuffled by recombination [2]. Recently, a number of *Bd* CRNs have been shown to be highly expressed on host tissue *in vitro* [37]. Therefore, whilst these genes in *Bd* lack a secretion signal, their expression, accumulation of genetic variation in terms of recombination and ϖ values, and similarities with oomycete CRNs strongly suggest that a number of these CRN-like genes are functional in *Bd*. However, whether they contribute directly or indirectly to the virulence of *Bd* remains to be determined.

Our demonstration of multiple hierarchies of cryptic genomic variation in *Bd* in terms of CCNV, ubiquitous and potentially targeted recombination, and natural selection, points to an ability to generate diversity without the necessity of an obligate sexual stage. Our study has uncovered high levels of genotypic plasticity that are likely to cause widespread phenotypic plasticity even without the need to invoke outcrossing. These large and small-scale changes are therefore likely to contribute to rapid evolutionary rates in the face of an effective host response. Such 'genomic instability' may explain the diverse phenotypic responses observed in *Bd* [38], and may also explain the enormous diversity of hosts and biomes that this generalist pathogen has managed to infect.

Materials and Methods

Full details are given in Text S1, Supplemental Materials and Methods. Briefly, twenty-two isolates that had been collected from nine countries and four continents were chosen for sequencing (Table 1). Paired-end Libraries were constructed according to the protocols provided by Illumina sequencing (Truseq kit). The genome sequence and feature file for the chytrid fungus *Batrachochytrium dendrobatidis* (*Bd*) strain JEL423 was downloaded from http://www.broadinstitute.org/ (GenBank project accession number AATT00000000). The feature file for JEL423 had all but the longest splice variants removed for each gene leaving 8794/8819 genes. We aligned our reads to the genome sequence using Burrows-Wheeler Aligner (BWA) v0.5.9 [26] with default parameters, converted to Samtools mpileup format using SAMtools v.0.1.18 [39] and polymorphisms called using the Binomial SNP-Caller from Pileup (BiSCaP) v0.11 [27]. For phylogenetic analysis we extracted polymorphisms covered ≥ 4 reads in all 22 isolates. FASTA files were converted into Nexus files and trees constructed using the Un-weighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm in PAUP and visualised using Figtree [40] (Fig. S3). Gene groups were identified using gene-annotations, blastx searches ($1e^{-05}$ e-value cut-off) to the non-redundant BLAST database, SignalP3.0 [41], Merops [42] and Procarb604 v1 [43].

Chromosome copy number variation (CCNV) was identified using changes in both depth of coverage and percent of reads specifying two most frequent alleles at any locus. To quantify these changes, we first performed *t-tests* (with a cut-off of $p<5^{-10}$) on the mean depths across the largest supercontig (supercontig 1) against each subsequent supercontig for each isolate (Fig. S5). Next, we calculated the percent of reads specifying the two most frequent alleles (Fig. S7) for each chromosome in each isolate separately using a minimum depth of 4 reads for both alleles and binned values falling between 47–53% (expected even ploidy/bi-allelic) and 30–36% and 63–69% (expected odd ploidy/tri-allelic). To account for depth and mutation variation within a chromosome, we performed 1000 bootstraps for either predominance of bi-allelic or tri-allelic peaks (Table S2). Using a 5% cut-off ($5\%<x<95\%$) we found 305/330 largest 15 chromosomes gave confident odd or even allelic peaks and was largely concordant with changes predicted by *t*-tests.

To detect recombination, we identified haplotypes using reads that overlapped two or more bi-allelic heterozygous positions. Haplotypes from each isolate were then compared to haplotypes in other isolates. We also calculated the Index of association (I_A), detecting linkage disequilibrium for a given set of haplotypes if $VD>L$ (L_{old}). We also calculated rBarD values and performed 4 gamete tests between every combination of loci in a haplotype (Fig. S9) to quantify the amount of recombination occurring within populations. In addition, we applied Weir's [30] estimator of Wright's Fixation Index (F_{ST}) according to the equations given in Multilocus 1.3 [29].

For selection, we used the yn00 and codeml programs of PAML [44] implementing the Yang and Nielsen method [45] on every gene in every isolate and those with $\omega \geq 1$ respectively. For codeml, we used the Branch site model (BSM) A (model = 2, NSsites = 2, fix_omega = 0) compared with the null model (model = 2, NSsites = 2, fix_omega = 1, omega = 1). Next, we calculated 2 * the log likelihood difference between the two compared models ($2D'$) with two degrees of freedom, and identified any with values greater than 8.1887 and 11.4076 (5% and 1% significance after Bonferroni correction). Enrichment for crossovers and polymorphisms was detected using hypergeometric tests and *t-tests*.

For *in vitro* divergence, an isolate of *B. dendrobatidis* from a Swiss *Alytes obstetricans* (isolate 0739) was subcultured into control (ACON) and peptide-treated (APEP) culture flasks containing 10 ml 1% tryptone media supplemented with 1% penicillin-streptomycin (Sigma) to reduce the risk of bacterial contamination. Cultures were incubated at 18°C and passaged every 4–5 d by scraping the side of the flask and transferring 1 ml into 9 ml fresh

media. Peptide-treatment included addition to the media of 80 µg ml-1 skin defense peptides collected from *Pelophylax esculentus* (n = 15 combined) according to Daum *et al.* (2012) [46]. This was equivalent to the IC_{50}, or the concentration at which growth of *Bd* was inhibited by 50%.

Supporting Information

Figure S1 The previous SOLiD reads (5) and the new Illumina paired end reads of *Bd* isolate JEL423 were aligned to a modified JEL423 reference sequence. Additionally, simulated reads from a heterozygous reference sequence were made to the depths of the Illumina and SOLiD datasets. Single Nucleotide Polymorphisms (SNPs) and heterozygous positions were then called and the False Discovery Rates (FDR) ascertained. The SNP-caller BiSCaP v0.11 was tested using default settings, and SAM/BCFTools with VCFUtils was tested for its ability to call SNPs using its default settings. SNPs were also filtered for those found without first modifying the reference sequence (f = filtered). (**A**) 1 nt/Kb simulated SNPs or heterozygous positions (12,458 in total) within the coding region (CDS) (**B**) 1 nt/100 nt simulated SNPs or heterozygous positions (124,588 in total) within the CDS region. The new Illumina data was able to recover >95% of true positive SNPs and >80% true positive heterozygous positions using BiSCaP v0.11, outperforming the previous lower-depth SOLiD sequences.
(PNG)

Figure S2 The percent of ECVA polymorphic sites shared between each of the 22 isolates. Greater overlap (≥30%) highlighted in red. (**A**) The overlap of homozygous SNPs varied between 3% and 97% (**B**) The overlap of heterozygous positions varied between 3% and 75%.
(PNG)

Figure S3 Phylogenetic trees were made using the UPGMA algorithm in PAUP from ECVA polymorphic positions identified in the nuclear genomes demonstrating three divergent lineages (*Bd*GPL, *Bd*CAPE and *Bd*CH shown in red blue and green respectively). (**A**) A tree from 275 Kb ECVA polymorphic positions identified from Illumina sequencing. (**B**) A tree from 36 Kb ECVA polymorphic positions from Illumina and SOLiD sequencing. (**C**) A tree from 218 Kb EVCA homozygous positions identified from Illumina sequencing. (**D**) A tree from 8 Kb EVCA homozygous positions identified from Illumina and SOLiD sequencing.
(PNG)

Figure S4 CCNV in the *Bd* nuclear genomes was identified using allele-frequencies and mean read depths across each chromosome normalised to the alignment depth for each isolate. Many *Bd*GPL isolates can be seen to include more copies of chromosome 2 and 3, while the 3 *Bd*CH and 3 of the 5 *Bd*CAPE isolates have fewer copies of chromosome 9 and 11. Fewer copies of chromosome 9, 11 and 16 appear to be found in many of the isolates.
(PNG)

Figure S5 *t*-tests for the mean depth of read coverage across each chromosome against chromosome 1 revealed significant p-values demonstrating uneven chromosome copy number. Stringent cut-offs for ploidy differences relative to the largest chromosome (Chr. 1) of each isolate were chosen: $p<5^{-10}$. Chromosomes with p-values below this cut-off, with a mean depth that is greater than chromosome 1 are highlighted in blue, while those with a mean depth lower than chromosome 1 are shown in green. All 308 chromosomal p-values (excluding chr1) are shown in the bottom plot ordered from smallest to greatest.
(PNG)

Figure S6 The percent of reads specifying the two most frequent alleles per chromosome using 2 representative isolates from each lineage of *Bd*. The most common allele is shown in black and the second most common allele is shown in blue. Bins were used to summarise the expected peaks for odd, even and odd numbers of chromosomes and shown in red (lines show bin value cut-offs and dots show values). Individual chromosomes with a predominantly bi-allelic value are shown with a blue border, and those with a predominant tri-allelic value are shown with a black border.
(PNG)

Figure S7 Sliding non-overlapping windows of 10 Kb across the 22 *Bd* nuclear genomes showing homozygous SNPs minus heterozygous positions. Predominance of homozygous SNPs is shown in red and predominance of heterozygous positions in shown in blue. Windows across *Bd*GPL isolates demonstrate highly uneven distribution of heterozygosity attributed to recombination whereas polymorphisms are more evenly spread across the genomes of *Bd*CAPE and *Bd*CH isolates.
(PNG)

Figure S8 Heterozygous positions had their phase determined using overlapping reads. Reads from each isolate are shown as a separate black line on the graphs. Only bi-allelic polymorphisms were compared for phasing. Predominantly, overlapping reads agreed with a single bi-allelic phase. (**A**) All reads over all phased positions. A 90% cut-off was used to filter ambiguous phased positions or those with an excess of mismatches as shown by the red line. (**B**) Positions that agreed 90–100% for a single phase are shown as a percent of all reads.
(PNG)

Figure S9 Illustrations of how phased haplotypes were extracted from the alignment. Heterozygous positions that did not pass the minimum depth or percent phased cut-offs, along with examples of pairwise crossovers and outcomes for a four-gamete test between three isolates.
(PNG)

Figure S10 Pairwise comparisons for shared phased heterozygous positions. (**A**) Total numbers of matching phased heterozygous positions in same phase (Kb) (**B**) Percent of matching phase positions from the total number of shared phased positions.
(PNG)

Figure S11 Phased heterozygous positions demonstrating crossovers were identified between every isolate. (**A**) Total numbers of crossovers identified. (**B**) Percent of crossovers from the total number of shared phased positions.
(PNG)

Figure S12 Lengths of haplotypes (in nucleotides) that included at least two alleles per loci in every isolate of a given group, and were therefore suitable for population genetic analysis. *Bd*GPL subset (s.s.) 1 consists of isolates VC1, AP15 and JEL423. Subset 2 consists of subset 1, ETH4 and MODS27.
(PNG)

Figure S13 Intra-lineage heterozygote's, the percent of heterozygote's that were phased (PP), the percent of PP's that demonstrated a crossover (XO) and the RbarD were plotted using non-overlapping windows across the genome (length 10 Kb). Both phased positions and crossovers were found across each of the chromosomes in each of the lineages of *Bd*, suggesting recombi-

nation is not confined to small or large chromosomes, or the ends of any given chromosome. The same is seen with rBarD values.
(PNG)

Figure S14 The Fixation Index (F_{ST}) was calculated for each pairwise lineage across window lengths of 1.4 Kb (**A**) and 10 Kb (**B**). All three lineages are differentiated from one another across each chromosome, with some intra-chromosomal variation. Notably, the stretch of rDNA located at the start of chromosome 14 appears to have a reduced genetic distance between each of the three lineages of *Bd*. *Bd*GPL subset (ss) 1 consists of isolates VC1, AP15 and JEL423. Subset 2 consists of subset 1, ETH4 and MODS27.
(PNG)

Figure S15 The total numbers of crossovers found within genes demonstrated variation between gene families. All crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size. Proteases and chitin recognition proteins had a greater number of crossovers than would be expected by random over their combined number of phased positions.
(PNG)

Figure S16 Crossovers at unique locations (non-redundant, NR) occurred differentially across gene families. NR crossovers were compared against total number of heterozygous and phased positions, transcript length and tribe size. Proteases and chitin recognition proteins had a greater number of crossovers than would be expected by random over their combined number of phased positions.
(PNG)

Figure S17 The ratio of non-synonymous mutation per non-synonymous site (*dN*) vs synonymous mutation per synonymous site (*dS*) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the *Bd*GPL. The line designates the ω value (*dN*/*dS*), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation.
(PNG)

Figure S18 The ratio of non-synonymous mutation per non-synonymous site (*dN*) vs synonymous mutation per synonymous site (*dS*) from alignments to *Bd* JEL423 for each of the gene families for all isolates belonging to the three lineages. The lines designate the ω value (*dN*/*dS*), whereby everything above the line has ω>1 and represents genes undergoing the greatest levels of variation. Summaries of ω values for all genes in each of the three lineages are shown in the final three plots.
(PNG)

Figure S19 A Venn diagram showing the total number of genes undergoing positive selection according to the Branch site model (BSM), where genes had 2D′>8.1887. The nine genes were identified in all three lineages were four uncharacterised (secreted) with transcript ID's 05565, 02533, 00379, 06783 and five uncharacterised (non-secreted) with transcript ID's 03962, 07794, 05877, 02935, 08088.
(PNG)

Table S1 Polymorphisms and reference bases were identified in 22 *Bd* nuclear genomes relative to *Bd* JEL423 using BiSCaP v0.11 with default settings. (**A**) Tallies of each category of loci found in each separate isolate. (**B**) The percent of uniquely mapped reads over each type of category of loci. Bi-allelic heterozygous positions had a reduced percent of uniquely mapped reads in the 2 divergent lineages of *Bd*, which may result from structural variants. Additionally, 72.48% of the homozygous SNPs and heterozygous positions were phased, which came from reads >86% uniquely mapped to the genome in any given isolate.
(PNG)

Table S2 The two most common allele frequencies over each base of each chromosome were determined by percent of read agreement with the reference base. Using 1000 Bootstrap replicates of these values, we recorded how often 47–53% reads agreeing with an allele predominated over 30–36% or 63–69% reads agreeing with an allele. Shown in white are chromosomes with >95% of replicates showing a predominantly bi-allele signature (even-ploidies). Chromosomes with <5% bootstrap support for an even number of chromosomes therefore had a high support for unbalanced allele frequencies (odd-ploidies), and shown in blue. Chromosomes not fulfilling these criteria are shown in green and considered ambiguous.
(PNG)

Table S3 Genes were tested for enrichment in non-redundant (NR; at unique loci) crossovers (XO) and NR XO/NR phased position (NRPP) compared to the values for all genes using Hypergeometric tests and *t*-tests respectively. For *t*-tests, all genes with <2 NRPP (the minimum required for a crossover) were excluded. Although both CRN-like and uncharacterised (secreted) were enriched for crossovers at unique loci (non-redundant), only CRN-like (between lineages) were enriched for XO/NRPP.
(PNG)

Table S4 Haplotypes over coding sequence that failed the four-gamete test were predominantly from coding-regions. Haplotypes overlapping a number of genes were included in the counts for each gene (385 extra counts to total number of haplotypes). After accounting for these extra counts, an additional 1,162 haplotypes were still found to come from coding regions compared with those from intergenic or intron regions. However, no gene group had a clear enrichment for haplotypes that failed the four-gamete test. *Bd*GPL subset (s.s.) 1 consisted of isolates VC1, AP15 and JEL423. Subset 2 consisted of subset 1, ETH4 and MODS27.
(PNG)

Table S5 Only five presence absence (PA) polymorphisms relative to *Bd*GPL JEL423 were identified amongst *Bd*CAPE and *Bd*CH isolates, whilst none were identified amongst *Bd*GPL isolates.
(PNG)

Table S6 Homozygous (A) and bi-allelic heterozygous (B) polymorphisms were found in the coding and non-coding regions of the *Bd* nuclear genomes. The total numbers of each variant-type are followed by their numbers per kilobase of genomic region in parentheses. For heterozygous positions, the affect on the transcript (synonymous/non-synonymous) was determined using the alternative allele. Where two alternative alleles to the reference sequence were found (infrequently), the first present within the VCF was chosen. With the exception of the reference strain *Bd* JEL423, the ratios of non-synonymous to synonymous changes were between 1.12–2.00 and 1.22–2.13 for homozygous and heterozygous positions respectively.
(PNG)

Table S7 ABC transporters, Chitin associated genes and CRN-like genes were tested for enrichment in homozygous SNPs. The total number, average and standard deviation of non-redundant homozygous SNPs for each gene family were calculated for all isolates, and lineage specific isolates. A Hypergeometric test was used to identify significant enrichment for variants where P<0.01 (*), P<0.001 (**) and P<0.0001 (***).
(PNG)

Table S8 Proteases, uncharacterized secreted genes and uncharacterized genes were tested for enrichment in homozygous SNPs. The total number, average and standard deviation of non-redundant homozygous SNPs for each gene family were calculated for all isolates, and lineage specific isolates. A Hypergeometric test was used to identify significant enrichment for variants where $P<0.01$ (*), $P<0.001$ (**) and $P<0.0001$ (***).
(PNG)

Table S9 Secreted and CRN-like genes are significantly enriched for heterozygous positions at unique loci. The total number, average and standard deviation of non-redundant heterozygous and phased positions for each gene family were calculated for all isolates and lineage specific isolates. A Hypergeometric test was used to identify significant enrichment for heterozygosity where $P<0.01$ (*), $P<0.001$ (**) and $P<0.0001$ (***).
(PNG)

Table S10 The average rates of synonymous substitution (*dS*), non-synonymous substitution (*dN*) and omega ($dN/dS=\omega$) for every gene in every isolate.
(PNG)

Table S11 Number and category of genes with $\omega\geq1$ (1,450 in total) that also were found to have undergone positive selection using the Branch Site Models in codeml. (**A**) The total numbers of genes, the numbers of genes with $\omega\geq1$ among all isolates, and how many of those genes had $2D'\geq11.4076$ (1% significance after Bonferroni correction) and $11.4076>2D'>8.1887$ (5% significance after Bonferroni correction). The final column shows the percent of genes with $2D'>8.1887$ from those with $\omega\geq1$. (**B**) For each lineage, the numbers of genes with $\omega\geq1$ and those that also had $2D'>8.1887$. Following both of these columns are the results from a hypergeometric test for enrichment. For the genes with $\omega\geq1$, the test is for enrichment from the entire set of genes, whilst for the genes with $2D'>8.1887$, the test is for enrichment from just the genes with $\omega\geq1$. (**C**) Overlap of genes with $2D'>8.1887$.
(PNG)

Text S1 Supplemental Materials and Methods. Full details of methods and analysis described in this manuscript.
(DOC)

Acknowledgments

We thank Benedict Schmidt, David Gower, Claude Miaud, Ché Weldon, Jaime Bosch, Giulia Tessa and Phineas Hamilton for collection of isolates, and Hanna Larner for assistance with Illumina sequencing.

Author Contributions

Conceived and designed the experiments: RAF DAH FB DCW MCF. Performed the experiments: RAF DAH DCW. Analyzed the data: RAF DAH DCW. Contributed reagents/materials/analysis tools: RAF TWJG. Wrote the paper: RAF DAH TWJG FB DCW MCF.

References

1. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, et al. (2012) Emerging fungal threats to animal, plant and ecosystem health. Nature 484:186–194.
2. Raffaele R, Kamoun S (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 10:417–430.
3. Ma L, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M, et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464:367–73.
4. Win J, Morgan W, Bos J, Krasileva KV, Cano LM, et al. (2007) Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic Oomycetes. Plant Cell 19:2349–2369.
5. Heitman J (2010) Evolution of eukaryotic microbial pathogens via covert sexual reproduction. Cell Host Microbe 8:86–99.
6. Wittenberg AHJ, van der Lee TAJ, M'Barek SB, Ware SB, Goodwin SB, et al. (2009) Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. PLoS ONE 4:e5863.
7. Cooke DEL, Cano LM, Raffaele S, Bain RA, Cooke LR, et al. (2012) Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. PLoS Pathog 8:e1002940.
8. Büttner P, Koch F, Voigt K, Quidde T, Risch S, et al. (1994) Variations in ploidy among isolates of *Botrytis cinerea*: implications for genetic and molecular analyses. Curr Genet 25:445–50.
9. Carr J, Shearer Jr G (1998) Genome size, complexity, and ploidy of the pathogenic fungus *Histoplasma capsulatum*. J Bacteriol 180:6697–6703.
10. Sheltzer JM, Blank HM, Pfau SJ, Tange Y, George BM, et al. (2011) Aneuploidy drives genomic instability in Yeast. Science 333:1026–30.
11. Abbey D, Hickman M, Gresham D, Berman J (2011) High-resolution SNP/CGH microarrays reveal the accumulation of loss of heterozygosity in commonly used *Candida albicans* strains. G3: Genes Genomes Genetics 1: 523–30.
12. Lengeler KB, Cox GM, Heitman J (2001) Serotype AD strains of *Cryptococcus neoformans* are diploid or aneuploid and are heterozygous at the mating-type locus. Infect Immun 69:115–22.
13. Hu G, Wang J, Choi J, Jung WH, Liu I, et al. (2011) Variation in chromosome copy number influences the virulence of *Cryptococcus neoformans* and occurs in isolates from AIDS patients. BMC Genomics 12:526.
14. Sionov E, Lee H, Chang YC, Kwon-Chung KJ (2010) *Cryptococcus neoformans* overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. PLoS Pathog 6: e1000848.
15. Reedy JL, Floyd AM, Heitman J (2009) Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. Curr Biol 19:891–9.
16. Forche A, Magee PT, Selmecki A, Berman J, May G (2009) Evolution in *Candida albicans* populations during a single passage through a mouse host. Genetics 182:799–811.
17. Kwon-Chung KJ, Chang YC (2012) Aneuploidy and drug resistance in pathogenic fungi. PLoS Path 8:e1003022.
18. Forche A, Alby K, Schaefer D, Johnson AD, Berman J, et al. (2008) The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. PLoS Biol 6:e110.
19. Arnold B, Bomblies K, Wakeley J (2012) Extending coalescent theory to autotetraploids. Genetics 192:195–204.
20. Schoustra SE, Debets AJM, Slakhorst M, Hoekstra RF (2007) Mitotic recombination accelerates adaptation in the fungus *Aspergillus nidulans*. PLoS Genet 3: e68.
21. Fisher MC, Garner TWJ, Walker SF (2009) Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. Annu Rev of Microbiol 63:291–310.
22. Farrer RA, Weinert LA, Bielby J, Garner TWJ, Balloux F, et al. (2011) Multiple emergence of genetically diverse amphibian-infecting chytrids include a globalised hypervirulent lineage. Proc Natl Acad Sci U S A 108:18732–6.
23. Goka K, Yokoyama J, Une Y, Kuroki T, Suzuki K, et al. (2009) Amphibian chytridiomycosis in Japan: Distribution, haplotypes and possible route of entry into Japan. Mol Ecol 18:4757–74.
24. Schloegel LM, Toledo LF, Longcore JE, Greenspan SE, Vieira CA, et al. (2012) Novel, panzootic and hybrid genotypes of amphibian chytridiomycosis associated with the bullfrog trade. Mol Ecol 21:5162–5177.
25. *Batrachochytrium dendrobatidis* Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/).
26. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.
27. Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC (2013) Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. Sci Rep 3:1512.
28. Morgan JAT, Vredenburg VT, Rachowicz LJ, Knapp RA, Stice MJ, et al. (2007) Population genetics of the frog-killing fungus *Batrachochytrium dendrobatidis*. Proc Natl Acad Sci USA 104: 13845–50.
29. Agapow P, Burt A (2001) Indices of multilocus linkage disequilibrium. Mol Ecol Notes 1: 101–2.
30. Hudson R, Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of sequences. Genetics 111:147–164.
31. Weir BS (1996) Genetic Data Analysis II. Sinauer, Sunderland.
32. Joneson S, Stajich JE, Shiu SH, Rosenblum EB (2011) Genomic transition to pathogenicity in chytrid fungi. PLoS Pathog 7:e1002338.
33. Sun G, Yang Z, Kosch T, Summers K, Huang J (2011) Evidence for acquisition of virulence effectors in pathogenic chytrids. BMC Evol Biol 11:195.
34. Abramyan J, Stajich JE (2012) Species-specific chitin binding module 18 (CBM18) expansion in the amphibian pathogen *Batrachochytrium dendrobatidis*. mBio 3: e00150–12.

35. Rosenblum EB, James TY, Zamudio KR, Poorten TJ, Rodriguez D, et al. (2013) Complex history of the amphibian-killing chytrid fungus revealed with genome resequencing data. Proc Natl Acad Sci U S A 110: 9385–9390.
36. James TY, Litvintseva AP, Vilgalys R, Morgan JAT, Taylor JW, et al. (2009) Rapid global expansion of the fungal disease chytridiomycosis into declining and healthy amphibian populations. PLoS Pathog 5:1–12.
37. Rosenblum EB, Poorten TJ, Joneson S, Settles M (2012) Substrate-specific gene expression in *Batrachochytrium dendrobatidis*, the chytrid pathogen of amphibians. PLoS ONE 7: e49924.
38. Woodhams DC, Alford RA, Briggs CJ, Johnson M, Rollins-Smith LA (2008) Life-history trade-offs influence disease in changing climates: strategies of an amphibian pathogen. Ecology 89:1627–39.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.
40. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214.
41. Bendtsen JD, Nielsen H, Heijne GV, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795.
42. Rawlings ND, Barrett AJ, Bateman A (2010) MEROPS: the peptidase database. Nucleic Acids Res 38:D227–D233.
43. Malik A, Firoz A, Jha V, Ahmad S (2010) PROCARB: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools. Adv Bioinformatics 436036.
44. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24:1586–91.
45. Yang Z, Nielsen R (2000) Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol Biol Evol 17:32–43.
46. Daum JM, Davis LR, Bigler L, Woodhams DC (2012) Hybrid advantage in skin peptide immune defenses of water frogs (*Pelophylax esculentus*) at risk from emerging pathogens. Infect Genet Evol 12:1854–1864.
47. Berger L, Speare R, Daszak P, Green DE, Cunningham AA, et al. (1998) Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and central America. Proc Natl Acad Sci U S A 95:9031–6.