The Ecology and Evolution of

Selfish Genes

Matthew Robert Goddard

University of London

Imperial College of Science, Technology and Medicine

Department of Biology

Silwood Park

A thesis submitted for the degree of Doctor of Philosophy in the year 2000
Acknowledgements

I thank Dr. Duncan Greig for the construction of the DH 89, 90, 91 and 95 Saccharomyces cerevisiae strains.

I thank Dr. Vassiliki Koufopanou and Stephanie Schaffer for their help with yeast sampling and for completing the majority of the Saccharomyces paradoxus laboratory isolations.

I thank Dr. Tim Barraclough and Dr. Rob Belshaw for their comments on preliminary drafts Chapters 1 and 6.

I thank Dr. Austin Burt for his ideas, supervision and more importantly for his role as mentor.

I thank Mary for her understanding during the three years it took to complete this thesis, her invaluable help in the writing of this thesis, for keeping me sane, and essentially for everything (not least agreeing to marry me!).
Abstract

Homing endonuclease genes (HEGs) are a family of optional, non-essential genes that have been discovered in all kingdoms of life. Studies concerning HEG molecular biology prove they are actively able to invade genes, meaning they may effect super-Mendelian inheritance. HEGs are termed 'selfish genes' since they will theoretically spread through a population even if detrimental. Very little is known about the ecology and evolution of HEGs. This thesis has taken detailed knowledge of HEG molecular biology and used this to test various hypotheses concerning HEG ecology and evolution. Firstly, theoretical predictions concerning HEG spread were addressed with experimentally outcrossed and inbred populations of yeast. It was found that HEGs do indeed increase in frequency in sexually outcrossed populations despite a selective disadvantage. This contrasts with their decrease in frequency in inbred populations (demonstrating a parasitic load). Natural yeast are predisposed to inbreeding, and inbreeding reduces the effectiveness of HEG spread. Therefore, the degree to which natural populations of yeast outcross was estimated using microsatellite markers. It was estimated that, at most, 4% of the population were outcrossing. Given this rate HEGs are predicted to increase in frequency 100 times slower than is possible in a 100% outcrossed population. None of the 30 natural isolates of yeast contained the two HEGs surveyed for. Next, questions concerning the longer term fate of HEGs were addressed. Using phylogenetic methods, and novel tests, it was found that HEGs are horizontally transferred approximately once every 1.59 million years. This estimation, combined with population genetic reasoning, allowed the construction of a cyclical model of HEG evolution describing how HEGs recurrently invade differing genomes via horizontal transfer events. Once at fixation within a
lineage HEGs degenerate and are ultimately lost. Recurrent invasions allow HEGs to persist. This work concludes that HEGs show a highly dynamic evolutionary history and are justly classified 'selfish genes'.
# Table of contents

ACKNOWLEDGEMENTS.......................................................................................................................2

ABSTRACT............................................................................................................................................3

CHAPTER ONE

INTRODUCTION...................................................................................................................................10

WHAT ARE SELFISH GENES?.............................................................................................................11

SEX AND SELFISH GENES................................................................................................................12

MODES OF NON-MENDELIAN INHERITANCE................................................................................15

Meiotic drive ....................................................................................................................................15

B - chromosomes ............................................................................................................................17

Transposable Genetic Elements ......................................................................................................18

HOMING ENDONUCLEASE GENES..................................................................................................19

The molecular biology of HEGs .......................................................................................................19

Homing: the non-Mendelian inheritance of HEGs .........................................................................20

The location of HEGs .......................................................................................................................22

EXPERIMENTAL ORGANISM OF CHOICE: YEAST.......................................................................26

THE AIMS OF THIS THESIS............................................................................................................28

CHAPTER TWO

HEG POPULATION DYNAMICS .........................................................................................................30

INTRODUCTION ...............................................................................................................................31

MATERIALS AND METHODS ...........................................................................................................34

Strains ...............................................................................................................................................34
Culture procedures ............................................................................................................. 37
Molecular techniques ........................................................................................................ 39

RESULTS AND DISCUSSION ......................................................................................... 43
VDE frequency change ...................................................................................................... 43
The parasitic load of VDE ................................................................................................. 47

CONCLUSIONS ................................................................................................................ 50

CHAPTER THREE

AN ESTIMATION OF NATURAL YEAST OUTCROSSING RATE....................................... 53
INTRODUCTION .................................................................................................................. 54
The Yeast life cycle .............................................................................................................. 55
Saccharomyces paradoxus – a natural yeast species found within the U.K. ...................... 58
MATERIALS AND METHODS ............................................................................................ 59
Isolation and identification of natural Saccharomyces paradoxus ......................................... 59
The scoring of microsatellite loci ....................................................................................... 62
Homing endonuclease survey ............................................................................................ 63

RESULTS AND DISCUSSION ......................................................................................... 65
S. paradoxus isolation ........................................................................................................ 65
HEG survey ....................................................................................................................... 65
Microsatellite polymorphism ............................................................................................. 67
S. paradoxus population structure ..................................................................................... 68
An estimation of outcrossing rate ...................................................................................... 69

CONCLUSIONS ................................................................................................................ 79

CHAPTER FOUR

THE LONG-TERM FATE OF HEGs .................................................................................. 85
INTRODUCTION .................................................................................................................. 86
MATERIALS AND METHODS ............................................................................................ 91
List of Tables

**TABLE 1.1** Homing endonucleases discovered in *S. cerevisiae* ..........................................................27

**TABLE 2.1** Yeast strains used for the VDE meiotic drive experiments .....................................................34

**TABLE 2.2** The VDE\(^+\) allele frequencies observed in the meiotic drive experiment ..........................44

**TABLE 2.3** The significance of change in VDE\(^+\) allele frequency ....................................................46

**TABLE 2.4** The relative fitness of the VDE\(^+\) allele .................................................................48

**TABLE 3.1** Details of the origins of the wild *S. paradoxus* isolates .....................................................66

**TABLE 4.1** Species, strains and Genbank accession numbers for sequences used in this study ............................92

**TABLE 4.2** Sequences of the primers used to study the ecology and evolution of the \(\omega\) element .........................93

**TABLE 4.3** Results of an algorithm to find the minimum number of horizontal transfers of a selfish gene ...............107
Chapter One

INTRODUCTION
What are selfish genes?

It is well understood why genes coding for products which have a functionally beneficial role in cellular or organismal processes are maintained within populations. Heritable phenotypic traits are coded for by genes, and genes which enhance the phenotypic fitness of individuals are more likely to be inherited by successive generations. This theory predicts that natural selection should have streamlined genomes and removed genes which are detrimental or phenotypically neutral (both could be termed ‘non-beneficial’) with respect to the host organism (Dawkins, 1982). However, some non-beneficial genes are found in contemporary organisms, suggesting that selection is not one hundred percent efficient (Crow, 1979; Werren et al., 1988). Certain non-beneficial genes persist not because of an absence of selection, but because they have evolved mechanisms to evade the forces of selection. These selfish genes persist over evolutionary time by enhancing their transmission into the next generation, meaning they may spread within an outcrossed population even if they are of no immediate benefit to the individual which carries them (Hickey, 1982). Selection, then, ultimately acts at the level of the gene (Williams, 1966). Indeed, all genes may be thought of as having a ‘selfish’ nature: in the Darwinian sense ‘ordinary’ genes ensure their persistence not by over-representing themselves in the next generation relative to the rest of their genome, but by aiding the construction and maintenance of organisms which are more likely to survive relative to their competitors (Williams, 1966). The concerted effort of ‘ordinary’ genes is, however, not necessarily as harmonious a process as we are lead to believe. In sexual species there appears to be conflict between genes at all levels, once again
reinforcing the concept that the gene is the fundamental unit of selection (Partridge and Hurst, 1998).

**Sex and selfish genes**

For the most part, selfish genes are only able to over-represent themselves in successive generations because of one phenomenon: sex. The recombination and subsequent assortment of genetic material from two individuals among progeny allows selfish genes to persist (Futcher et al., 1988; Hickey, 1982). In asexual populations clones which possess a phenotypically detrimental gene will be selected against, and the degree of selection against the detrimental gene will dictate the rapidity with which it is removed from the population. The same is true in sexual populations: detrimental genes which are inherited in a Mendelian fashion will eventually be removed from the population due to selection (Mettler and Gregg, 1969) (though the time it takes to remove a gene will depend on the recessive/dominant status of the allele). However, if a gene possesses a mechanism which enables it to over-represent itself among the progeny of a sexual population (i.e. can be inherited in a non-Mendelian manner) then it may increase in frequency despite reducing the fitness of individuals carrying it, see Figure 1.1) (Hickey, 1982). Conceptually, as long as a gene is inherited by more progeny than, on average, fail to reproduce because of the effects of that gene, then it will persist in a population. These predictions are based upon the equation shown below when ‘selfish gene+’ and ‘selfish gene−’ are considered alternate alleles of the same locus, and no migration and complete outcrossing are also assumed. The frequency of a detrimental selfish gene in the next generation is predicted by:
where \( p \) is the frequency of the 'selfish gene+' allele in the starting generation; \( p' \) is the frequency of the 'selfish gene+' allele in the next generation; \( q \) is the frequency of the alternative 'selfish gene~' allele; \( s \) is the selection coefficient of the 'selfish gene+' allele (the detrimental load); and \( d \) defines the degree to which an allele is over-represented in the next generation (\( d = 0.5 \): Mendelian). This model is based upon a two allele system, and it is assumed the 'selfish gene+' heterozygote has the same fitness as a 'selfish gene+' homozygote. \( \bar{w} \) represents the average fitness of the population, which may be calculated since:

\[
\bar{w} = p^2(1-s) + 2pq(1-s) + q^2
\]

[Eq. 1.2]

The speed with which a selfish gene is predicted to increase will depend upon the degree to which it is over-represented in the next generation and its selection coefficient. Figure 1.1 shows the rate of increase of a selfish gene that has a detrimental load of 10\% (\( s = 0.1 \)) with varying degrees of over-representation in the next generation. Genes with less of a detrimental effect will increase in frequency at a greater rate. If a 'selfish gene' is not a single gene but comprises a cluster of tightly linked genes, which may also have
Figure 1.1 The predicted change in frequency of a detrimental ‘selfish gene+’ allele under various rates of super-Mendelian inheritance \((d = 0.6 \rightarrow 0.9)\) compared with a Mendelian pattern of inheritance \((d = 0.5)\), in a panmictic population. The allele frequency changes were calculated using a selection coefficient \((s)\) of 0.1 and were simulated in Excel using Eq. 1.1.

The additional complication of external modifying genes, then models to predict their population dynamics will understandably be more complicated (e.g. Charlesworth and Hartl, 1978).

The paradoxical existence of non-beneficial genes is partly understood given some knowledge of their patterns of non-Mendelian inheritance. However, little is known of the actual population dynamics of selfish genes. Given the above theoretical model, a selfish gene should quickly sweep through a population. Is this true in practise, and if so what happens after fixation? What is the frequency of selfish genes in natural populations? What
life history strategies do selfish genes adopt? How large a detrimental load can a selfish gene convey to its host before the host is reproductively crippled? Do selfish genes remain with host lineages, or are they horizontally transferred among species?

**Modes of non-Mendelian inheritance**

Genes that are able to effect non-Mendelian inheritance may do so in a number of differing ways, and attempts have been made to classify these modes (Werren et al., 1988). This thesis is concerned with the study of a family of selfish genes called Homing Endonuclease Genes (HEGs). HEGs do not easily fall into any of the currently accepted modes of non-Mendelian inheritance. First, I will briefly discuss three major modes of non-Mendelian inheritance, before moving on to discuss HEGs and how they relate to other known selfish genes.

**Meiotic drive**

Meiotic drive is the phenomena whereby segments of DNA are able to over-represent themselves, relative to their homologues, in the gametes resulting from a heterozygous cross. In other words, meiotic drive is when one of two differing heterozygous alleles or chromosomes 'drive' themselves into more than the predicted 50% of progeny. The classic, and probably best characterised, example of a meiotic drive system are the SD elements in *Drosophila* (Crow, 1979). Segregation distorser (SD) in *Drosophila melanogaster* is due to an autosomal cluster of tightly linked loci, found near the centromere on the second
chromosome, which acts only in males. Sperm can either be Sd or Sd*, and 95% or more Sd* sperm are prevented from development when in the presence of Sd sperm (Charlesworth and Hartl, 1978). Hence, males which are heterozygous at the Sd loci just transmit one of the two alleles by only producing sperm that contain Sd. A closely linked responder locus (Rsp) dictates sensitivity to Sd. Chromosomes and can be insensitive to the driving Sd (Rsp^), or can show varying degrees of sensitivity [Rsp^ → Rsp^ss], a chromosome which is Sd Rsp^ is suicidal (!]), which is correlated with the number of AT-rich repeats (the larger the repeat region the more sensitive) found at the Rsp locus (Werren et al., 1988). The precise mode of action of SD is unknown, but the simplest model proposes that one or more products of the Sd bind to the Rsp repeats and inhibit replication (Hurst, 1998; Werren et al., 1988). However, more recent studies have narrowed down the functional SD product to a truncated version of the RanGAP nuclear transport protein (Merrill et al., 1999). Unexpectedly, over 70% of naturally occurring chromosomes are SD sensitive (Temin and Marthas, 1984), and long term selection experiments suggest that sensitive, tandem repeat bearing chromosomes, are fitter in comparison to insensitive chromosomes (Werren et al., 1988).

Sex ratio (SR) is another meiotic drive system found in Drosophila (Beckenbach, 1983). Here males that have SR on their X chromosome do not transmit any Y bearing sperm, and as a result only female offspring are produced. Theoretically SR should quickly drive through the population and cause extinction. Natural population surveys indicate that SR exceeds 20% only in some populations and that in laboratory populations it is often lost. The detrimental effect of SR in SR homozygous females and the half sperm viability of SR carrying males may
explain the low natural frequency of SR. The evolution of suppressors may also explain the low frequency, though none have so far been found.

Meiotic drive is also known in such systems as mosquitoes, mice (t-locus), flour beetles and even a 'social meiotic drive system' in red fire ants, (Beeman et al., 1992; Crow, 1979; Keller and Ross, 1998; Lewontin, 1962). All the meiotic drive systems so far described appear to be polygenic and are effective at the gametic level, or higher in the case of the fire ants; all these systems therefore have significant fitness effects. The molecular basis is not well understood for any meiotic drive system, but hypotheses often include the expression of a toxin and antidote system.

**B - chromosomes**

B - chromosomes are whole extra chromosomes, and were the first genetic elements to be labelled as parasitic when Ostergren (1945) pointed out that B - chromosomes need not be useful to the individual, they only need be useful to themselves. Since then, B - chromosomes have been discovered in hundreds of animal and plant species (Burt and Trivers, 1998; Werren et al., 1988). Most Bs have a negative fitness effect but persist within populations due to their non-Mendelian inheritance which is achieved by gonotaxis (whereby during cell division Bs preferentially move towards the germline and away from the soma). Within British plants it has been highlighted that the breeding system is of crucial importance to the observed distribution of Bs. A population genetic model predicts that parasitic Bs are better able to persist in outcrossed plant species when compared to those that inbreed. These predictions were borne out using data on the incidence of Bs in British plants, where it was
found that outcrossed plants do indeed tend to have more Bs (Burt and Trivers, 1998). This again underlines the fact that parasitic genetic elements are able to persist within populations because of sex.

Transposable Genetic Elements

Transposable genetic elements (TGEs) are segments of DNA that are able to cut or copy themselves from one genomic location and insert themselves into another (Griffiths et al., 1993). They are found in both pro- and eukaryotes and are probably the most abundant forms of selfish gene. There are many different families of TGEs and these elements clearly have their own agenda: they are therefore justly considered selfish genetic elements. TGEs that are virus-like may be considered as separate organisms which are capable of surviving outside of the host cell, even though they are not able to replicate their DNA without the help of the host's cellular machinery. TGEs that are not able to infect contemporary members of a population are only able to increase in frequency within a population via meiosis. Transposition within a TGE heterozygote will ensure that all offspring will contain at least some copies of the TGE, and in this way TGEs may effect super-Mendelian inheritance. It is the autonomous replication and movement of TGEs which allows them to persist and spread in populations. The parasitic nature of these elements may come in the form of an economic load upon the host cell, but it is the detrimental insertion of TGEs into crucial coding areas of the host's genome which is more likely to be the major parasitic effect (Zeyl et al., 1996). However, it has been suggested that such insertions may be rarely beneficial and are possible catalysts for organismic evolution (McDonald, 1995).
The above is not intended as an exhaustive review, but rather to illustrate the current perception of selfish genetic elements. Only some could really be described as constituting a single gene, most are more than that. Also, most of the above categories affect large proportions of the host’s genome, either by killing whole gametes or by making many copies of themselves. HEGs are, in contrast, a family of single genes which are only found at specified loci.

Homing Endonuclease Genes

The molecular biology of HEGs

HEGs are a family of optional, non-essential genes that are roughly one thousand basepairs (bp) in length. HEG protein products (homing endonucleases) possess, as their name describes, endonuclease function: they are able to cut DNA (Mueller et al., 1993). These homing endonucleases do not cause random double stranded breaks within a genome - they are targeted toward certain recognition sites (Belfort and Roberts, 1997). The size and specificity of target site varies among homing endonucleases: certain HEGs code for endonucleases that will only recognise and therefore only cleave a single site within a genome, whilst other endonucleases appear to be more promiscuous. Double stranded breaks in DNA are potentially fatal to the cell (Fairhead and Dujon, 1993), such breaks are therefore repaired via double-strand-break repair pathways (Szostak et al., 1983).
**Homing: the non-Mendelian inheritance of HEGs**

According to the rules of Mendelian genetics alleles inherited from heterozygous parents segregate in a 50:50 ratio. However, parents heterozygous for HEGs pass on HEGs to over 95% of their progeny (Gimble and Thorner, 1992; Jacquier and Dujon, 1983). HEGs manage to over-represent themselves in successive generations via a process known as 'homing' (Mueller et al., 1993; Szostak et al., 1983). Consider an individual who is heterozygous for an HEG at a certain locus (i.e. one chromosome contains an HEG and the other does not). The HEG codes for an enzyme which only cleaves DNA at a specific genomic site, and this site is the homologous allele which does not contain a copy of the HEG – a homologous 'empty' allele (Figure 1.2). The specificity of these enzymes is achieved by their relatively long DNA recognition sequences (15-30 bp) since the enzyme physically binds to the double stranded DNA before cleavage (Dalgaard et al., 1997). The endonuclease’s action results in a double stranded break in the ‘empty’ allele, and the break is made at the exact position where the HEG is present in the other intact chromosome. This break initiates the double-strand-break repair pathway (Szostak et al., 1983). The cleaved DNA is not simply re-ligated (joined together) as presumably certain information may be missing; rather, the cell looks to the other intact allele and uses this as a template for repair. The other intact allele has an HEG present and, since this is used as the template, the HEG is copied into the gap to repair the broken allele. Once the repair process is finished both alleles contain an HEG, i.e. the cell is now homozygous for the HEG at that locus. This process is illustrated in Figure 1.2. Homing explains how HEGs are super-inherited. As explained earlier, this will mean that an HEG can theoretically spread quickly once in an outcrossing population.
Figure 1.2
Model of double-strand break-repair pathway for transfer of endonuclease coding sequences.

1. Donor allele with HEG (red lines) and recipient allele.
2. Site specific cleavage (double stranded break) of recipient allele by homing endonuclease.
3. Exonuclease degradation and alignment of recipient allele with donor allele.
4. Once aligned, a 3' tail of the recipient invades the donor, which acts as a template for DNA repair.
5. Resolution of duplexes.
DNA sequences close to the HEG may be copied over when an HEG homes by a process known as flanking co-conversion. Indeed, this is how the first HEG, \( \omega \) ("omega"), was discovered in yeast: a selective marker, which turned out to be close to \( \omega \), demonstrated non-Mendelian inheritance (Dujon, 1980). The probability with which flanking sequences are carried over decreases with increasing distance from the homing site, although in yeast the process still appears to be around 80% effective at around one kilobase pairs from the homing site (Jacquier and Dujon, 1985). This co-conversion effect is probably due to the fact that a certain amount of the recipient’s chromosome is degraded by exonucleases during the homing process (see Figure 1.2). The observation that sequences close to HEGs undergo co-conversion introduces the possibility that other genes, beneficial or otherwise, may 'hitchhike' along with an HEG as it sweeps through a population. To date, this theoretical possibility has not been investigated.

Additionally, the double-strand-break repair process is associated with crossing over at meiosis. Certain resolutions of the duplexes (see Figure 1.2, step 4) may allow crossing over, and in yeast this occurs at a frequency of around 50% (Szostak et al., 1983). HEG homing therefore increases the recombination rate of the host cell. Sex and recombination are arguably beneficial so, in this respect, HEGs may have hidden long term benefits.

**The location of HEGs**

Of the HEGs that have been demonstrated to show homing ability, 86% are found within genes that are crucial for the host cell's function such as ribosomal and metabolic pathway
genes; the remaining 14% are found in non-coding areas (Belfort and Roberts, 1997). Ordinarily a large insert within a gene would destroy its function. HEGs avoid this potential suicidal effect (the death of the host cell also means the demise of the HEG) by splicing out of the host gene in one of two ways: either at the mRNA level; or, at the protein level (Closton and Davis, 1994; Lambowitz and Belfort, 1993). These splicing events leave a functional host gene product and a functional endonuclease (see Figure 1.3). HEGs that splice at the mRNA level are unable to achieve this independently, and so they need the assistance of a surrounding self-splicing intron (group I, group II or archaeal) within which they find a 'safe haven'. The secondary and tertiary structures formed by these introns allow them to splice out of mRNA during mRNA processing. Such introns are able to tolerate large insertions within their loop structures (see Figure 1.3), and it is here where HEGs may 'hide' (Cech, 1990), since HEGs are removed as part of the intron. The intron and HEG appear to have a symbiotic relationship: the intron allows the HEG to be tolerated within host genes, and in turn the HEG confers mobility to the intron since it is carried over with the HEG when it homes. It is interesting to note that many group I and II introns have been discovered which do not contain HEGs (Dujon et al., 1986; Lambowitz and Belfort, 1993). Also horizontal transfer
Figure 1.3 The two methods of processing which allow homing endonucleases (red) to inhabit host genes (black). Intronic HEGs rely upon associated self splicing introns (white) for excision at the RNA level, here a group I intron is shown. HEGs may be tolerated in the intron loop structures A-E, here the HEG is in loop E by way of an example. Other HEGs do not require an accessory intron, they are able to splice themselves out at the protein level, these are known as inteins.
(cross-species transmission) of group I introns has been inferred (Sellem and Belcour, 1997; Vaughn et al., 1995; Yamada et al., 1994), but prior to this thesis no inference of HEG cross-species transmission had been made. The other method of HEG-splicing occurs at the protein level (Closton and Davis, 1994; Gimble and Thorner, 1992) and this is achieved without the aid of an accessory intron (any accessory intron would be removed during RNA processing).

The host gene and HEG are transcribed and translated as one contiguous molecule, and only at the protein level will the HEG splice to yield a functional host gene product and a functional endonuclease (Figure 1.3). These latter HEGs are known as inteins, to distinguish them from HEGs that splice at the RNA level (referred to as introns) (Dujon et al., 1989). Inteins therefore have a dual function at the protein level: as well as having an ability to cleave DNA at a specific site, they are also able to self-splice. It is thought that the self-splicing and HEG domains have separate evolutionary histories. These two domains appear to form a symbiotic relationship that parallels the self-splicing intron/HEG relationship at the protein level. Phylogenetic analysis of many known and putative HEGs suggest that intein and intron HEGs independently cluster together and that intron HEGs appear to be more ancestral (Dalgaard et al., 1993). Presumably this is because the evolution of self-splicing activity is a rare event among HEGs, thought it is worth noting that some intron HEGs have maturase activity – they are thought to help in intron splicing (Lambowitz and Belfort, 1993).

HEGs may also be categorised into sub-families according to certain amino acid motifs that are found within them (Belfort and Roberts, 1997). The most common family contains a LAGLIDADG protein motif and Genbank searches suggest there are up to 130 such sequences that conform to this motif (Dalgaard et al., 1997). Other smaller families include
the N-H-N, Gİ-Y-IİG and the His-Cys Box motifs; it is not known how the amino acid configurations that define these families affect endonuclease function. Over 35 homing endonucleases with demonstrable function have been discovered from all kingdoms of life (Belfort and Roberts, 1997). Within Eukaryotes HEGs have been found in mitochondrial, chloroplast and nuclear genomes; therefore, more than one HEG has been found in every kingdom and cellular compartment. However, HEGs are predominantly found in microorganisms, and no HEG has been found in an animal of greater complexity than a sea anemone (Beagley et al., 1996).

**Experimental organism of choice: yeast**

The first HEG to be discovered, I-Scel or ω, was in *Saccharomyces cerevisiae* (bakers yeast) (Dujon, 1980), and since then a further ten have been discovered in S. cerevisiae (Table 1.1). Many of these HEGs have also been reported from a variety of other yeast species (Dalgaard et al., 1997). Yeasts are extremely attractive organisms in which to study the ecology and evolution of HEGs since:

1. a vast wealth of yeast molecular data is available (e.g. the whole S. cerevisiae genome has been sequenced);
2. yeast are easily manipulated in the laboratory and have a very quick generation time (roughly 90 minutes);
3. yeast can easily be made to clone, inbreed or outcross, and;
4. many close relatives of *S. cerevisiae* are available from culture collections and the natural environment.

<table>
<thead>
<tr>
<th>Name(^a)</th>
<th>Location</th>
<th>Host Gene</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-Scel</td>
<td>Mito</td>
<td>large RNA</td>
<td>Intron - group I</td>
</tr>
<tr>
<td>I-Scel II</td>
<td>Mito</td>
<td>cox I</td>
<td>Intron - group I</td>
</tr>
<tr>
<td>I-Scel III</td>
<td>Mito</td>
<td>cox I</td>
<td>Intron - group I</td>
</tr>
<tr>
<td>I-Scel IV</td>
<td>Mito</td>
<td>cox I</td>
<td>Intron - group I</td>
</tr>
<tr>
<td>I-Scel V</td>
<td>Mito</td>
<td>cox I</td>
<td>Intron - group II</td>
</tr>
<tr>
<td>I-Scel VI</td>
<td>Mito</td>
<td>cox I</td>
<td>Intron - group II</td>
</tr>
<tr>
<td>I-Scel VII</td>
<td>Mito</td>
<td>cob</td>
<td>Intron - group I</td>
</tr>
<tr>
<td>Pl-Scel</td>
<td>Nuclear</td>
<td>VMA1</td>
<td>Intein</td>
</tr>
<tr>
<td>F-Scel</td>
<td>Mito</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F-Scel II</td>
<td>Nuclear</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1.1 Homing endonucleases discovered in *S. cerevisiae* (Belfort and Roberts, 1997). All have demonstrated function. F-Scel is the endonuclease that allows yeast to mate type switch: the endonuclease allows one of two differing mating type determining genes to be copied into the mating type locus via gene conversion. \(\ast\), intron encoded; PI, protein insert (intein); F, freestanding. \(\S\) endonuclease activity only activated by mutation. Mito = mitochondrion.
The aims of this thesis

Clearly, HEG molecular biology is extremely well characterised; however, very little is known of the ecology and evolution of these selfish genes. The dynamics of HEGs at the population level and greater have not been explored. This thesis will address these questions. Three differing approaches, covering three evolutionary scales, will be taken to help better understand the ecology and evolution of these genes.

I. HEG population dynamics

Firstly, the population dynamics of HEGs on the short term will be investigated since little is known of HEGs population genetics. HEGs are theoretically predicted to rapidly increase in frequency over a few generations in a fully outcrossed population. It will be investigated if this is true in practice by assaying the frequency of an HEG in inbred and outcrossed yeast populations. It is not known whether HEGs convey a parasitic load to the host they inhabit. This question will be investigated using experiments with isogenic yeast.

II. An estimation of natural yeast outcrossing rate

As has hopefully been made clear, sex is of critical importance to the persistence of selfish genes. Yeast are only facultatively sexual, the amount of sex that wild yeast actually have will therefore be of crucial importance to the fate of HEGs; this is currently unknown and will be
estimated using microsatellite loci. The microsatellite data will also provide information on natural yeast population structure. The frequency of HEGs within wild populations of yeast is also unknown. Additionally, yeast isolated from the natural environment will be surveyed for the presence of HEGs.

III The long-term fate of HEGs

HEGs found at the same insertion site have been sporadically reported from differing species of yeast, but the overall phylogenetic distribution of HEGs is not well understood. A collection of closely related yeasts will be surveyed for an HEG, and the survey results will then be mapped onto a host phylogeny. Further, an HEG phylogeny will be reconstructed as knowledge of both phylogenies will then allow tests for HEG horizontal transfer to be conducted. The distribution of HEGs upon a yeast host phylogeny will indicate whether HEGs have been solely vertically inherited, or if they have also been horizontally transferred. Further questions concerning HEG evolution will also be addressed: if HEGs do quickly go to fixation once in a population or species, what then? Will HEGs be fixed in all populations? If they are not why not? What other possible mechanisms could be acting upon HEGs?
Chapter Two

HEG POPULATION DYNAMICS
Chapter Two: HEG population dynamics

Introduction

Major questions concerning the meiotic drive and parasitic load of HEGs have not previously been tackled; the population dynamic properties of HEGs will have a crucial bearing upon their long term fate. Do HEGs increase in frequency in an outcrossed population? Do HEGs confer a parasitic load to the host cell? The efficiency of homing has previously been estimated for only two yeast HEGs: VDE and ω (omega). VDE is an HEG intein that is found interrupting the single copy nuclear VMA1 gene (Hirata et al., 1990), and ω is an HEG intron found interrupting the mitochondrial large sub unit rRNA gene (Dujon, 1980). The efficiency with which these HEGs home has been estimated from single generation crosses: HEG+ and HEG− yeast were mated, and the progeny were assayed for their HEG status. These assays proved that both these HEGs may drive themselves into the next generation. VDE drives with a frequency of 88% (Gimble and Thorner, 1992) (n = 26), and ω drives into over 99% of the progeny (Jacquier and Dujon, 1985) (n = 293). However, whilst such observations allow one to theoretically predict the spread of such driving genes (Hickey, 1982, and see Figure 1.1), they do not necessarily prove that HEGs will spread in an outcrossed population. Only the direct observation of changes in allele frequency will allow such statements to be made with confidence. Two types selfish element, the 2 μm plasmid and the Ty3 retrotransposon, have been observed to increase in frequency in an outcrossed yeast population (Futcher et al., 1988; Zeyl et al., 1996). The 2 μm plasmid is a circular double-stranded DNA molecule of roughly six kilobase in size (Futcher, 1988), and is found in the nuclear envelope of nearly all strains of S. cerevisiae. The plasmid only has four genes, and these provide it with mitotic
partitioning and copy number amplification systems. Competition experiments have proved that the plasmid confers a parasitic load to the host cell of around 1% (Futcher et al., 1988). Additionally, populations containing 2 μm⁺ and 2 μm⁻ yeast were passed through four rounds of meiosis. Yeast undergo sporulation when subjected to nutrient limited conditions (Kurtzman and Fell, 1998). One diploid yeast cell produces four ascospores which are contained within a tough ascus (Latin for sac). These ascospores have the mating type designations of either a or a, and they may only mate with other ascospores of the opposite mating type (Berry, 1982). Yeast may easily be made to either outcross or self fertilise in the laboratory since the ascus may either be left intact or be enzymatically and physically destroyed (Guthrie and Fink, 1991; and see Chapter three of this thesis). If the ascus is left intact the ascospores will mate with their asci (pl. ascus) partners and effect self fertilisation. If the ascus is destroyed the ascospores will be released which subsequently allows the random union of ascospores. In the 2 μm plasmid experiment one set of the yeast populations were allowed to randomly outcross whilst the other set were only allowed to self fertilise. The yeast populations were subjected to four rounds of sporulation over which the frequency of the plasmid was monitored. It was found that the 2 μm plasmid only increased in frequency in the outcrossed populations, despite having a parasitic load, whilst the plasmid frequency remained constant in the inbred populations. The increase in plasmid frequency was attributed to the fact that the plasmid was being transmitted to new hosts with each round of mating and sporulation (Futcher et al., 1988). Similar increases in Ty3 frequency were observed in outcrossed populations of yeast when compared to asexual populations (Zeyl et al., 1996). Theoretical considerations of VDE molecular biology suggest that it should also increase in frequency in outcrossed populations since VDE is able to home into vde⁻ VMA1
genes (Chong et al., 1996; Gimble and Thomer, 1992). To test this hypothesis populations of
\textit{VMA1::VDE} (VDE$^+$) and \textit{VMA1\textDelta VDE} (vde$^-$) isogenic yeast were subjected to similar
inbreeding and outcrossing experiments. The isogenic nature of the yeast will control for the
possibility of confounding genetic background effects. The relative change in VDE$^+$ allele
frequency will indicate the population dynamic nature of HEGs.

HEGs are clearly not essential for host cell function since isolates may be observed which
lack HEGs (Goddard and Burt, 1999; Ragnini et al., 1991; Wilson and Fukuhara, 1991). The
extent to which HEGs convey a parasitic load to the host cell is unknown since there are no
empirical data available concerning the relative fitness of HEG$^+$ alleles when compared to
HEG$^-$ ones. Questions concerning the parasitic load conferred on the host cell by the
presence of VDE may be answered using such experiments.
Materials and methods

Strains

The yeast strains used for the meiotic drive assays are shown in Table 2.1. These were kindly constructed by D. Grieg and derive from the *S. cerevisiae* Y55 wild type strain. These strains are genotypically extremely similar and are described as being 'isogenic' since they are the progeny of two highly inbred, genotypically equivalent strains of Y55. The only difference between these strains is thought to be in the *VMA1* gene which either does or does not contain *VDE* (*VMA1* derived endonuclease). However, since these strains were not constructed by autodiploidization (where a haploid yeast mates with a clone mate to produce a truly homozygous diploid, see Chapter Three for further details) there is a small possibility of hidden background genetic differences.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH89</td>
<td>α, ura3, vde⁻</td>
</tr>
<tr>
<td>DH90</td>
<td>a, ura3, vde⁻</td>
</tr>
<tr>
<td>DH91</td>
<td>α, ura3, VDE⁺</td>
</tr>
<tr>
<td>DH95</td>
<td>a, ura3, VDE⁺</td>
</tr>
</tbody>
</table>

Table 2.1. Yeast strains used for the VDE meiotic drive experiments. DH89 and DH90 derive from a cross between Y55 2366 x 2189; DH91 and DH95 derive from a cross between VDE transformed DH89 x DH90.

To construct the VDE⁺ strains DH89 and DH90 were independently transformed with the YEp VMA1 plasmid (kindly supplied by F. Gimble; Gimble and Thorner, 1993) which contains the
Chapter Two: HEG population dynamics

VMA1::VDE gene. These transformants were then mated with each other and then sporulated allowing VDE to home into the nuclear VMA1 gene from the plasmid. The resulting asci were disrupted to release single spores which were isolated and allowed to grow as haploids. These haploids were cured of the plasmid and screened for successful homing events via PCR. Two haploid VDE+ strains were selected (DH91 and 95) and both of these originated from different tetrads. These two haploids were then mated to produce a DH91/95 VDE+ diploid. It was necessary to obtain vde− strains which had identical breeding histories to DH 91/95 VDE+. DH89 and DH90 haploids were first mated and the resulting diploids sporulated. The resulting asci were disrupted releasing the ascospores, and these spores were then allowed to mate randomly with one another meaning that spores from differing tetrads combined to form zygotes. Six of these DH89/90 vde− diploids were selected for experimentation. The starting strains used to assay the homing ability of VDE were therefore both one random mating away from their identical parent strains (see Figure 2.1). Individual colonies from these cultures were then picked and checked for sporulation efficiency. The six DH89/90 clones were competed independently against DH91/95 to produce a total of six replicate populations. The possibility of VDE dysfunction was checked by mating DH91/95 to the prototrophic Y55 vde− S. cerevisiae strain, and successful homing events were confirmed by PCR.
Figure 2.1 The construction of the DH89/90 vde- and DH91/95 VDE+ diploid yeast that were used for the VDE meiotic drive experiments. α and α refer to the mating type of the haploid. All strains are ura3 auxotrophs.
Culture procedures

Pure cultures of all seven diploid strains (6 x DH89/90 + 1 x DH91/95) were grown in 100mL of YPD in a 500mL conical flask with shaking at 25°C until stationary phase was reached (approx. 24 hrs). Serial dilutions of these cultures were then made and plated on YPD (1% yeast extract, 2% glucose, 2% tryptone) for culture density estimation. The seven culture densities did not significantly differ.

Since the stationary phase cultures had equivalent densities the starting 2 : 8 VDE+: vde− ratio was achieved by mixing the stationary phase cultures in the appropriate proportions. 200μL of the VDE+ culture was mixed with 800μL of the vde− to give a total of 1 mL. Each of the six DH89/90 vde− isolates were mixed separately with DH91/95 VDE+ isolates to give a total of six replicates. The six replicates were then synchronously put through four rounds of mating.

Immediately after mixing, the populations were vigorously vortexed and samples were removed for allele frequency estimation and for storage in glycerol at -80°C. 200 μL of the 2 : 8 mixtures were then plated onto sporulation medium (2% Potassium acetate) and incubated for four days at 30°C. After four days the cultures were microscopically checked for sporulation. The resulting spores were exposed to diethyl-ether vapour since this kills any unsporulated vegetative cells (Guthrie and Fink, 1991), meaning that only the progeny from meiotic divisions can contribute to the next generation. Sporulated yeast were then removed.
from the plates with a sterile loop and placed into 400 µL of sterile water: these were then
vigorously vortexed to homogenise the sample.

Each of the six replicates was then divided in two: one half constituted the inbred population
the other the outcrossed population. The division of replicates yielded a total of 12
populations (6 inbred, 6 outcrossed). This protocol meant that each of the inbred and
outcrossed replicate populations derived from the same starting culture.

For yeast to outcross the tough ascus that encapsulates the four ascospores must be
removed, and if this is not achieved the ascospores will usually mate within the ascus to yield
inbred progeny. In the outcrossed populations the asci were enzymatically digested with
Sulfatase (from Helix pomatia: Sigma no. S9626). 100 µL of 22mg/mL Sulfatase was added
to each of the six outcrossed populations and incubated at 30°C overnight in sterile water.
The inbred populations were incubated in the same conditions without Sulfatase. The
following day the outcrossed populations were added to 3mL of sterile water and sonicated at
20kHz for five three second bursts. The sonication step not only breaks open any resistant
asci but also randomises ascospores: this allows true outcrossing since any one individual
ascospore is no more likely to be physically nearer an asci mate than any other spore in the
population. Ascus disruption was checked microscopically. The sonicated population was
then spun down and resuspended in 200 µL of sterile water. The inbred populations were not
sonicated.
Both the inbred and outcrossed populations were then densely spread onto YPD plates, and incubated at 30°C for three days, which allowed mating to occur. After three days samples of yeast were removed from the plates with a sterile loop and placed in 1mL of sterile water and then vortexed. 200 µL of this homogenised suspension was spread onto sporulation media to complete the cycle. This cycle was repeated for a total of four rounds of mating and this protocol is illustrated in Figure 2.2. At every round of mating samples were removed for storage in glycerol at -80°C. The VDE+: vde− allele frequencies were estimated at the start and end of the experiment.

**Molecular techniques**

Starting cultures were double checked for their VDE status with a novel direct colony PCR protocol. Primers VMA01 (5' - ATT TAT ATC CCA AGA GGT ATT - 3') and VMA02 (5' ATT CCA TCA AGA CTT CTG C - 3') were designed which flank the VDE insertion site and are based on an alignment of *S. cerevisiae* and *Candida tropicalis* VMA1 genes. Colonies were grown overnight on YPD at 30°C, and the following day individual colonies were removed with a sterile tooth pick and placed in a ‘boil buffer’ which comprised: 1x PCR buffer (1.5 mM Mg2+), 50µM dNTPs and 10pmols of
**Figure 2.2.** The culture procedures conducted to enable populations of yeast to inbreed and outcross. All 12 populations (six replicates each containing one inbred and one outcrossed population) were put through the cycle contemporaneously. Sporulation media = 2% Potassium acetate, mating media = YPD (1% yeast extract, 2% glucose, 2% tryptone). Sporulation and mating were both conducted on solid media.
VMA01/02 primers to a final volume of 40μL. This was then vortexed well and incubated at 98°C for five minutes, vortexed well once more and then centrifuged at high speed for 10 minutes. 20 μL of the boil buffer supernatant was used as a template for PCR amplification. Amplification of VDE was carried out under the following conditions: 30 seconds at 95°C; 1 minute at 50°C; 3 minutes at 72°C for fifty cycles and then a 10 minute final extension step at 72°C.

VDE+ and vde− allele frequencies were estimated via colony hybridization; first serial dilutions were made and plated out to a density of around 300 colony forming units per plate. Starting frequencies were estimated by individually removing one hundred colonies from each replicate and streaking these onto a YPD plate overlain with a gridded Hybond N nylon membrane. These plates were then incubated overnight allowing further colony growth. End frequencies were estimated from serial dilution plates overlain with a Hybond N nylon membrane averaging around 300 colonies. The following day the resulting colonies were lysed in situ and their DNA fixed to the nylon membrane in their relative positions according to the protocol outline in Kaiser et al., (1994). Membranes were then incubated with a hybridisation solution (5x SSC, 0.5% SDS, 5x Denhardt’s solution, 130 ng denatured salmon sperm DNA) in a rotating oven at 65°C with an α-32P dCTP incorporated VDE DNA sequence following Sambrook et al. (1989). The VDE probe was amplified from DH 91/95 using internal primers VMAI1 (5′ – GGG TAG GAA TGT TTT AAT GGG GG – 3′) and VDEI2 (5′ – GGA CGA CAA GGT GGT TGG C – 3′), and contains none of the flanking VMA1 sequence. The resulting autoradiograms indicated the number of colonies that were VDE+. Finally the colony
blots were probed with a positive control amplified from the nuclear SAG1 pheromone receptor gene to check the efficiency of the yeast colony blot.
Results and discussion

VDE frequency change

The frequency of the VDE+ and vde− alleles were estimated at the beginning of the experiment and after the fourth round of mating. VDE only homes during meiosis (Gimble and Thorner, 1992), in contrast to other HEGs, such as ω, which home in mitotic cells (Dujon, 1989). Since all yeast were homozygous for their VDE status at the start of the experiment, the opportunity for VDE to home only occurred in three of the four sporulation cycles. The frequency of the VDE+ allele at the start of the experiment was calculated by dividing the number of colonies which showed a positive hybridisation signal, when probed with VDE, by the number showing a positive hybridisation signal with SAG. If it is assumed that no outcrossing occurred in the inbred treated populations (that none of the asci were disrupted), then the same calculation may be used to estimate frequency of the VDE+ allele for the inbred replicates. However, the same calculation may not be used for the outcrossed populations since a positive signal may be produced by both VDEVDEVDE+ homozygote and VDE+vde− heterozygote colonies. If it is assumed that complete outcrossing occurred in the outcrossed populations then Hardy-Weinberg equilibrium will have been achieved. Given this assumption, the frequency of the VDE+ allele may be calculated since the frequency of vde− homozygotes may be estimated from the proportion of colonies that showed no hybridisation signal with the VDE probe. The frequency of the vde− allele (q) may be found by taking the square root of the frequency of the vde− homozygotes. The frequency of the VDE+ allele (p)
is found simply since \( p = 1 - q \). The frequencies of VDE\(^+\) and vde\(^-\) alleles are shown in Table 2.2.

<table>
<thead>
<tr>
<th>Rep.</th>
<th>VDE(^+) Start</th>
<th>Arcsine</th>
<th>VDE(^+) Inbred</th>
<th>Arcsine</th>
<th>VDE(^+) Outcrossed</th>
<th>Arcsine</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.07</td>
<td>15.34</td>
<td>0.07</td>
<td>15.34</td>
<td>0.13</td>
<td>30.79</td>
</tr>
<tr>
<td>B</td>
<td>0.20</td>
<td>26.57</td>
<td>0.10</td>
<td>18.43</td>
<td>0.23</td>
<td>28.52</td>
</tr>
<tr>
<td>C</td>
<td>0.21</td>
<td>27.27</td>
<td>0.13</td>
<td>21.13</td>
<td>0.46</td>
<td>41.84</td>
</tr>
<tr>
<td>D</td>
<td>0.24</td>
<td>29.33</td>
<td>0.11</td>
<td>19.37</td>
<td>0.37</td>
<td>37.58</td>
</tr>
<tr>
<td>E</td>
<td>0.22</td>
<td>27.97</td>
<td>0.14</td>
<td>21.97</td>
<td>0.29</td>
<td>32.39</td>
</tr>
<tr>
<td>F</td>
<td>0.20</td>
<td>26.57</td>
<td>0.15</td>
<td>22.79</td>
<td>0.33</td>
<td>35.12</td>
</tr>
<tr>
<td>Mean</td>
<td>0.19</td>
<td>25.508</td>
<td>0.12</td>
<td>19.838</td>
<td>0.30</td>
<td>32.707</td>
</tr>
<tr>
<td>S. D.</td>
<td>0.06</td>
<td>5.088</td>
<td>0.030</td>
<td>2.734</td>
<td>0.107</td>
<td>7.384</td>
</tr>
</tbody>
</table>

Table 2.2. The VDE\(^+\) allele frequencies observed in the meiotic drive experiment at the start, and after four generations of inbreeding and complete outcrossing, in six replicate populations A – F. The frequency data were arcsine transformed to meet the assumptions requirements of a paired t-test (Sokal and Rohlf, 1981). The mean and standard deviation of the replicates are also shown. Start n = 100; inbred n = 200-600; outcrossed n = 117-330.

Five of the six replicates had the desired starting VDE\(^+\) allele frequency of around 0.2, however, replicate ‘A’ showed a lower starting frequency. Regardless of VDE\(^+\) starting frequency, Table 2.2 shows that the frequency of the VDE\(^+\) allele was seen to increase in all
of outcrossed populations. The frequency of VDE⁺ alleles was seen to decrease in five of the six inbred replicates. Are these changes in VDE⁺ frequency significant? The design of the experiment allows the use of a paired t-test to statistically determine the significance of the change in allele frequency. The power of this test will not be reduced by the anomalous initial frequency observed in replicate ‘A’ since a paired t-test only makes use of the change in allele frequency and not the absolute frequencies. Frequency data defies assumptions of the t-test, since it is not normally distributed and in a binomial distribution the variance is dependant upon the mean. The frequency data may be arcsine transformed to prevent these properties and allow use in a t-test (Sokal and Rohlf, 1981). This transformation finds the arcsine of the square root of the frequency. The transformed frequency data are shown in Table 2.2. A paired t-test is performed by calculating the difference in value for each pair, and then by finding the mean of the differences for all the pairs and then dividing this by the standard error of these differences. The significance of the VDE⁺ allele frequency change (under the null hypothesis that the difference in frequency = 0) is obtained by comparing the resulting t value to the table of critical values of student’s t-distribution (two-tailed). The VDE⁺ frequency results allow three comparisons to be made, and these are shown in Table 2.3.
<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean of differences</th>
<th>S.E. of mean of differences</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start v Inbred</td>
<td>5.670</td>
<td>1.421</td>
<td>3.991</td>
<td>0.0104</td>
</tr>
<tr>
<td>Start v Outcrossed</td>
<td>7.198</td>
<td>1.785</td>
<td>4.032</td>
<td>0.0100</td>
</tr>
<tr>
<td>Inbred v Outcrossed</td>
<td>12.868</td>
<td>2.303</td>
<td>5.589</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

Table 2.3. The significance of change in VDE⁺ allele frequency in inbred and outcrossed populations over four generations of meiosis, and between the end frequencies of inbred and outcrossed populations.

The results of the t-test prove that the VDE⁺ allele significantly increased in frequency within the outcrossed population. This increase in frequency is not due to any selective advantage since the frequency of the VDE⁺ allele significantly decreased in the inbred populations. That VDE was observed to significantly increase in frequency in the six outcrossed populations must be due to the fact that VDE was able to effect non-Mendelian inheritance by homing. This observed increase in frequency therefore agrees with the theoretical population genetic predictions based upon the knowledge of HEG homing ability (Mueller et al., 1993). It is clear that VDE is only able to persist within an outcrossed population; in an inbreeding population the VDE⁺ allele frequency significantly decreases over time. Therefore, the observed increase in frequency of VDE, despite a selective disadvantage, is attributed to its ability to home.
The parasitic load of VDE

These results show that in the absence of outcrossing VDE does not increase in frequency within a population. They also suggest that VDE may confer a parasitic load on the host cell which carries it since VDE frequency is seen to decrease in most of the inbred populations. In these inbred populations VDE is unable to override any possible selective forces through homing. This decrease in frequency within inbred populations may be due to the fact that either VDE confers a mitotic parasitic load to the host cell, or that VDE confers a meiotic parasitic load to the host cell, or both. It is possible that VDE interferes with the sporulation process, however, this was not immediate since both the VDE+ strain and all six vde− strains had similar sporulation efficiencies (between 70 – 75%, data not shown). The relative fitness of the VDE+ allele in an inbred population may be calculated, if it is assumed the change in frequency of the VDE+ allele in the populations is due to selection, by:

\[
\frac{p_t}{q_t} = \frac{p_0}{q_0} \left[ \frac{W_A}{W_B} \right]^t
\]

[Eq. 2.1]

where p and q are the frequencies of alleles at generations 0 and t, and \( W_A \) and \( W_B \) are the relative fitnesses of these alleles. If the mean values for the frequency of the VDE+ allele at the start and end of the experiment are substituted for \( p_0 \) and \( p_t \) respectively, then the VDE+ allele is estimated to confer an a 12% average disadvantage to the host cell when compared to the vde− allele (Table 2.4). In outcrossed populations, where VDE may home, the relative fitness of the HEG+ alleles becomes positive.
### Chapter Two: HEG population dynamics

<table>
<thead>
<tr>
<th>Replicate</th>
<th>VDE$^+$ allele relative fitness when compared to the vde$^-$ allele</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inbred</td>
</tr>
<tr>
<td>A</td>
<td>1.0 (-0%)</td>
</tr>
<tr>
<td>B</td>
<td>0.816 (-18.4%)</td>
</tr>
<tr>
<td>C</td>
<td>0.866 (-13.4%)</td>
</tr>
<tr>
<td>D</td>
<td>0.791 (-20.9%)</td>
</tr>
<tr>
<td>E</td>
<td>0.872 (-12.8%)</td>
</tr>
<tr>
<td>F</td>
<td>0.917 (-8.3%)</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>0.877 (-12.3%)</strong></td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>0.075 (± 7.5%)</td>
</tr>
</tbody>
</table>

Table 2.4. The relative fitness of the VDE$^+$ allele when compared with the vde$^-$ allele in the experimental inbred and outcrossed populations of yeast. The relative fitnesses were calculated with equation 2.1 using the data seen in Table 2.2. The relative selective load of the VDE$^+$ allele is shown in parentheses.

If this selection coefficient, and the previously estimated value for the rate of VDE drive of 88% (Gimble and Thorner, 1992), are substituted for $s$ and $d$ in Equation 1.1, then the theoretical rate of VDE spread in a fully outcrossed population may be estimated. With a starting frequency of 0.19, the VDE$^+$ allele is predicted to reach a frequency of 0.56 after three rounds of complete outcrossing (see Figure 2.3). The observed value of 0.30 is considerably lower than this predicted value suggesting that either the efficiency of VDE homing was less than 88%, or that complete outcrossing was not achieved in the experimental populations.
Figure 2.3 The predicted change in frequency of \textit{VDE} in fully outcrossed populations when \textit{VDE} drives ($d$) with a rate of 88% and with no drive (50%), and with a selection coefficient ($s$) of 12%. This is compared with the observed change in \textit{VDE} frequency in the experimentally outcrossed populations (error bars represent ± 1 standard deviation).
Conclusions

It is apparent that VDE only increases in frequency in outcrossed yeast populations. Given the detailed knowledge of VDE molecular biology (Chong et al., 1996; Gimble and Thorner, 1992) it is clear that homing is of critical importance for spread since inbred populations, which lack homing opportunities for VDE, experience a decrease in VDE frequency. These results agree with the theoretical predictions concerning the spread of HEGs, and selfish genes in general (Hickey, 1982). These results also agree with other observations of selfish element spread made in experimental populations of yeast (Futcher et al., 1988; Zeyl et al., 1996). Moreover, such observations agree with studies concerning selfish element spread and abundance in other organisms (Burt and Trivers, 1998; Good et al., 1989), and together these serve to reinforce the concept that outcrossed sex is of critical importance for the spread of selfish genes. The estimation that VDE confers a parasitic load of 12% to the host cell seems high when it is considered that the 2μm plasmid has been estimated to confer a parasitic load of only 1%. The 2μm plasmid is roughly 6 kbp (kilobase pairs) in size and has an average copy number of 60-100 (Futcher, 1988), this plasmid therefore contributes roughly an additional 360-600 kbp of DNA to a diploid yeast cell (≈4% of the genomic DNA). In contrast VDE will only contribute an additional two kbp to a diploid yeast cell (≈0.02% of the genomic DNA). If the parasitic load of these selfish elements comes purely in the form of an economic load, then it is hard to reconcile the differences in parasitic load between these two elements since the 2μm plasmid contributes over two orders of magnitude more DNA when compared to VDE.
Alternatively, the parasitic load of VDE may originate at the protein level from Pl-Sce I (the endonuclease protein encoded by VDE) (Gimble and Thorner, 1993; Perler et al., 1994). Pl-Sce I may have deleterious effects which impart this serious selection coefficient. If this is the case then such a parasitic load may not be observed in mitotic cells since Pl-Sce I is only seen to mediate homing events during meiosis. This is somewhat perplexing since VMA1 (and therefore VDE) is expressed during mitotic cell growth; however, no homing of VDE is observed during mitotic growth (Gimble and Thorner, 1992). An additional mitotic competition experiment between isogenic VDE+ and vde− yeast would answer such questions concerning differences in VDE selection coefficient.

It is also possible that the parasitic load associated with VDE is a result of the genetic engineering procedures that were carried out during the construction of the VDE+ strains since the VDE+ allele was 'engineered in'. Transformation of the DH89 and 90 haploids with the Yep VMA1 plasmid may have resulted in hidden genetic disruptions which ultimately may slow the growth of these strains. Indeed, three VDE+ strains were originally obtained for use in the meiotic drive experiments, but two of these three strains displayed obvious slow growth and were therefore discarded. This problem may be overcome by 'engineering out' VDE from the VDE+ allele strains, and then competing the 'engineered out' vde− strains with the relatively undisrupted VDE+ strains. If VDE still shows a parasitic load with these competition experiments the possibility of hidden genetic disruptions may be discounted since it would be the vde− allele strains that have been engineered the most recently. Nevertheless, even if the selective load attributed to VDE is artificially high, this goes even further to prove that VDE may increase in frequency in a host population despite being parasitic.
That the predicted VDE frequency of 0.56 (when $d = 0.88$ and $s = 0.12$) was not reached in three generations may be due to the incomplete outcrossing of the yeast population. The degree to which the experimental populations actually outcrossed may be gauged by determining the heterozygotes frequency within the population. In principle, this may be achieved by first probing for VDE$^+$ alleles and then probing for the vde$^{-}$ alleles: any colonies that show a positive signal with both may be identified as heterozygotes. The degree to which the population deviate from Hardy-Weinberg equilibrium will indicate the extent to which the population has outcrossed. The proportion of the population that outcrossed is likely to have a critical bearing on the ability of an HEG to spread through that population. A partially inbreeding yeast population may reflect the more natural yeast state since the yeast mating system leaves them predisposed to selfing (Naumov et al., 1997). The degree to which natural populations of yeast outcross is therefore of paramount importance to the longer term fate of HEGs, and selfish genes, in general.
Chapter Three

AN ESTIMATION OF NATURAL YEAST OUTCROSSING RATE
Chapter Three: An estimation of natural yeast outcrossing rate

Introduction

The work detailed in the preceding chapter demonstrates that HEGs may only increase in frequency within sexually outcrossed populations. In the absence of outcrossing HEGs are unlikely to increase in frequency since they have no selective advantage. Populations which are obligate outcrossers will therefore experience the spread of such driving genes if they are present in the population. Indeed, this appears to be true for plants where it has been found that parasitic B - chromosomes are much more likely to be reported from obligately outcrossed species when compared with inbred ones (Burt and Trivers, 1998). Yeast, however, are not so restricted - they may outcross, but they may also reproduce via the most severe form of inbreeding: self fertilisation (selfing). The proportion of yeast which outcross within a population will therefore critically determine the rate at which HEGs, and other such driving genes, spread in that population. One study has made use of allozyme data from Saccharomyces paradoxus (a close relative of Saccharomyces cerevisiae) to infer reproductive isolation between isolates from Europe and the far East (Naumov et al., 1997). However, there has been no previous attempt to estimate the outcrossing rate in natural populations of yeast.

The extent to which HEGs are found in natural populations of yeast is currently unknown. Are HEGs found in the wild or only in domesticated laboratory strains of yeast? If HEGs are present in natural populations of yeast, at what frequency will they be found? These questions will be addressed since wild isolates of S. paradoxus will be surveyed for both a nuclear and mitochondrial HEG.
The Yeast life cycle

Yeast have a well characterised sexual life cycle in which the haploid and diploid phases may carry out vegetative reproduction (Kurtzman and Fell, 1998). This life cycle is illustrated in Figure 3.1. Yeast are usually found in the diploid state and may mitotically divide giving rise to many diploid clones, and under suitable conditions this may occur indefinitely. Under conditions of nitrogen starvation mitotic division ceases and sporulation begins (Berry, 1982). The meiotic division of diploid yeast cells gives rise to four haploid ascospores contained within a tough ascus which is the original diploid’s modified cell wall. Haploid ascospores possess mating type designations as defined by one of two alleles found at the MAT (mating type) locus: a and α (alpha). Haploid ascospores may complete the life cycle by subsequently mating with another ascospore of the opposite mating type to give rise to diploid cells. It is unlikely that the ascus will be broken down in the natural environment. The total degeneration of the ascus and subsequent release of the ascospores is only achievable in the laboratory by first subjecting asci to Sulfatase, and then sonicating to completely break open the asci (Guthrie and Fink, 1991; Futcher et al., 1988). It can be argued that such chemical and physical conditions will seldom be met in the natural
Figure 3.1. The yeast life cycle. See text for details. Cell sizes are not to scale.
environment. In the absence of asci degeneration ascospores will mate with their asci partners and effect self fertilisation. If the ascus is broken down in the natural environment then the four ascospores are released and may individually germinate and mitotically divide giving rise to haploid clones. However, haploid clones will not persist since wild type haploid yeast are able to 'mate type switch' (Mortimer et al., 1994). Apart from the functional a or α gene found at the MAT locus an additional copy of each mating type gene is found at one of two silent associated loci (HML and HMR). Mating type switching is achieved by swapping the gene present in the active MAT locus for the opposite one located at either the HML or HMR loci - this is accomplished by gene conversion. The HO gene codes for an endonuclease which cleaves in the MAT locus, and this break is repaired using the opposite copy of the mating type gene, found at one of the silent loci, as a template (the same process as seen in Figure 1.2). The ability to mate type switch further increases the probability that yeast will show high levels of inbreeding. Any ascospores which fail to mate within the ascus (either because some spores are not viable or because the ascus has been broken) may divide mitotically for a few generations, by which time some individual cells will have mate type switched allowing them to mate with their clones. With the exception of the MAT locus, such matings will produce individuals who are homozygous at every loci. Indeed, such a process is thought to occur in S. cerevisiae found in wine fermentations (Mortimer et al., 1994).
Chapter Three: An estimation of natural yeast outcrossing rate

*Saccharomyces paradoxus* - a natural yeast species found within the U.K.

*Saccharomyces cerevisiae* is considered a domesticated species and is rarely isolated from the natural environment (Naumov et al., 1997). An outcrossing rate estimated from populations of *S. cerevisiae* is therefore unlikely to reflect the rate at which natural yeast populations outcross since *S. cerevisiae* inhabits man-made, artificial environments. *Saccharomyces paradoxus* is a close relative of *S. cerevisiae* and they are capable of hybridising. Reassociation studies suggest that the total DNA homology between these organisms is around 50% (Vaughan-Martini and Kurtzman, 1985). *Saccharomyces paradoxus* has previously been isolated from the natural environments of both Europe and the Far East, and such isolates represent members of truly wild populations of yeast (Naumov et al., 1997). The use of wild *S. paradoxus* to estimate outcrossing rate will therefore give a more accurate picture of this rate. *S. paradoxus* has also previously been isolated from Oak (*Quercus* sp) within the U.K. and Europe (Naumov et al., 1995; Greig, 1999).

Microsatellite loci will be used to score the degree of heterozygosity found in *S. paradoxus* isolates obtained from Silwood Park and Windsor Great Park. Microsatellite loci are assumed to be relatively stable neutral genetic markers (Field and Wills, 1998). The degree of heterozygosity, and its deviation from predicted Hardy-Weinberg values, may then be used to estimate an outcrossing rate since selfing will tend to decrease heterozygosity (Falconer, 1981). Such microsatellite markers may also give some insight as to the geographic population structure of yeast.
Materials and methods

Isolation and identification of natural *Saccharomyces paradoxus*

Oak (mostly *Quercus robur*) trees from six locations in Silwood park and four locations in Windsor Great Park were chosen to sample for yeast (Figure 3.2*). Six trees (three large and three small) from each of the Silwood locations had two bark scraping samples removed. These samples were from the east and west sides of the trunk approximately 1.7 metres from the ground. Twelve trees from each of the Windsor Great Park locations also had two scraping samples removed, these were from the north and south aspects of the trees' trunks, again, approximately 1.7 meters from the ground. Each sample position was visited twice between December 1997 and December 1998, therefore, a total of 344 (2x172) samples were removed from 86 different trees (see appendix I). A few grams of bark were removed from each sample position with a sterile scalpel and placed immediately into a sterile plastic sealable bag. These samples were placed in acidified liquid malt media (5% malt extract, 0.4% lactic acid w/v) under aseptic conditions in the laboratory, and incubated, with shaking, for two days at 30°C. The low pH deters the growth of bacteria and the shaking deters the growth of...

*Figure 3.2. (over) The geographic location of sites sampled for *S. paradoxus* in Silwood Park (S1 to S6) and Windsor Great Park (W1 to W4).*
Chapter Three: An estimation of natural yeast outcrossing rate

Figure 3.2
filamentous fungi. A vast array of microbes were present in the media after the two day growth period and therefore a selection procedure was designed to isolate *Saccharomyces paradoxus*. Samples from the two day cultures were taken and plated on acidified malt agar and again incubated at 30°C (the lactic acid must be added to the medium after autoclaving) for 24 hrs. The resulting colony forming units were visually inspected and likely yeast colonies were picked, placed on solid YPD (1% yeast extract, 2% glucose, 2% tryptone) and incubated at 30°C overnight. The subsequent colonies' phenotypes were then inspected via microscopy and hyphal forming and bacterial colonies were removed (Barnett et al., 1990). Along with other yeast *S. paradoxus* produces distinct spore forming structures, known as tetrads, when starved of nitrogen sources (Kurtzman and Fell, 1998). Therefore, the candidate colonies were replica plated onto sporulation inducing media (2% potassium acetate) and incubated at 30°C for two days; any resulting tetrads were visualised via microscopy. In order to further eliminate unwanted colonies, the Internal Transcribed Spacer (ITS)1–5.8S–ITS2 region of the tetrad forming colonies was amplified using primers ITS1 and ITS4 (see Table 4.2) (White et al., 1990), and then visualised via electrophoresis through 1% agarose gels. The ITS1-5.8S-ITS2 region in *S. paradoxus* is known to be roughly 765 base pairs in length (Genbank accession no. AJ229059, my deposition), allowing the rejection of colonies that produced significantly different sized amplicons. ITS1-5.8S-ITS2 amplicons of roughly the correct size were sequenced and those with sequences identical to *S. paradoxus* (CBS isolate 432) were identified as such.
The scoring of microsatellite loci

Open reading frames containing variable tandem repeats have previously been identified in *S. cerevisiae* (Field and Wills, 1998). In this earlier study primers were designed and successfully employed to amplify the homologous area in five other closely related yeasts. Thirteen microsatellite loci reported in this paper are found in coding genes, these all therefore have repeats which are multiples of three since these will maintain the reading frame. Primers for these thirteen loci were obtained in an attempt to amplify the homologous areas of *S. paradoxus*. Under a large variety of conditions only three pairs of these primers successfully amplified a distinct product from *S. paradoxus* [a failure of the vast majority of the primers in *S. paradoxus* was confirmed by the original author (D. Field pers. comm.)]; these were the primers designed against the *S. cerevisiae* genes *TFA1*, *SRP40* and *FUN12* (loci designations SCU12825, YSCSRP40X and YSCFUN12A respectively). These three loci were then amplified from the wild *S. paradoxus* isolates.

To check for length polymorphism and within isolate heterozygosity the three amplicons were electrophoresed through Elchrome Spreadex EL400 mini-gels. These gels are agarose based and are reportedly able to resolve differences in as little as one base pair by retarding the migration of DNA molecules larger than 400bp (different exclusion limits are available). However, the results obtained with these gels proved to be imprecise, and in order to obtain better amplicon resolution polyacrylamide gel electrophoresis (PAGE) was utilised. First, the forward primers for each of the three loci were end labelled with $^{32}$P molecules with the enzyme T4 Polynucleotide kinase (PNK). These radiolabelled primers were used in conjunction with the unlabelled reverse primers to amplify the loci from the wild isolates.
radiolabelled PCR products were then electrophoresed through 6% polyacrylamide denaturing gels. The gels were run at 1,700 volts, 35 watts and 100 milliamps for between two and four hours depending upon the size of the amplicon (Sambrook et al., 1989). After electrophoresis the gels, and the DNA within them, were stabilised by immersion in 10% methanol: 10% acetic acid and then dried under heat and vacuum. The dried gels were then allowed to exposed X-ray film at -80°C for three days. The resulting autoradiograms were then scored for length polymorphism and isolate heterozygosity.

Some of the TFA1 amplicons were directly sequenced from PCR products. Forward and reverse sequences were obtained using the dye dideoxy chain termination method (Sambrook et al., 1989) using a Perkin Elmer dye terminator cycle sequencing ready reaction kit (part no. 402079) and the resulting products were electrophoresed with an ABI 373 automated sequencer.

**Homing endonuclease survey**

The wild isolates were surveyed for two HEGs, VDE and ω, using PCR. VDE is found interrupting the nuclear VMA1 gene (Gimble and Thorner, 1992) and ω is found interrupting the mitochondrial LSU rRNA gene (Jacquier and Dujon, 1983). Primers ω01 (5' - GAT AAC GAA TAA AAG TTA CGC TAG GG – 3'), ω02 (5' - CTT CAG CAG ATA GGA ACC ATA CTG – 3'), VMA01 (5' ATT TAT ATC CCA AGA GGT ATT) and VMA02 (5' ATT CCA TCA AGA CTT CTG C) were designed to flank the HEG insertion sites and are based on an alignment
of *S. cerevisiae* and *Candida tropicalis* VMA1 and mt LSU genes. The sizes of the resulting amplicons indicate whether or not a HEG is present at the insertion site.
Results and discussion

S. paradoxus isolation

Just under nine percent of the bark samples yielded S. paradoxus isolates, meaning that 30 isolates were obtained from the 344 bark samples. It appears that S. paradoxus may be relatively easily and consistently isolated from Oak trees in southern England which agrees with other studies (Greig, 1999). The origins of the 30 isolates are detailed in Table 3.1.

HEG survey

None of the 30 wild isolates contain VDE. This agrees with survey results obtained from a variety of other S. paradoxus isolates, both isolated from the wild and from culture collections (results not shown). None of the 30 wild isolates contain the ω HEG, but they do all contain the group I intron which is associated with the ω HEG (see Chapter Four for more detail). This also agrees with previous studies (Ragnini, 1991). This group I intron, as the introduction explained, is not able to home, and simply allows the splicing of the ω element at the RNA level and therefore the tolerance of the ω HEG at the LSU insertion site. It therefore appears that these two HEGs are absent (or at a very low frequency) from the natural population of S. paradoxus in Silwood and Windsor Great Park.

<table>
<thead>
<tr>
<th>Isolate²</th>
<th>Sample area information$</th>
<th>Date Isolated</th>
</tr>
</thead>
<tbody>
<tr>
<td>T4b ⊕</td>
<td>S1:1s:W</td>
<td>May '98</td>
</tr>
</tbody>
</table>
### Table 3.1 Details of the origins of the wild *S. paradoxus* isolates.

<table>
<thead>
<tr>
<th>Code</th>
<th>Location</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>T8.1</td>
<td>S1:2s:W</td>
<td>May '98</td>
</tr>
<tr>
<td>W7</td>
<td>Sliwood Park</td>
<td>Oct '96</td>
</tr>
<tr>
<td>T32.1</td>
<td>S2:3L:W</td>
<td>July '98</td>
</tr>
<tr>
<td>T28.3</td>
<td>S2:2s:E</td>
<td>July '98</td>
</tr>
<tr>
<td>S36.7</td>
<td>S2:3s:W</td>
<td>Dec '97</td>
</tr>
<tr>
<td>T27.3</td>
<td>S2:3s:E</td>
<td>July '98</td>
</tr>
<tr>
<td>T76.6</td>
<td>S3:1s:W</td>
<td>July '98</td>
</tr>
<tr>
<td>T18a</td>
<td>S4:2L:W</td>
<td>May '98</td>
</tr>
<tr>
<td>T18.2</td>
<td>S4:2L:W</td>
<td>May '98</td>
</tr>
<tr>
<td>T21.4</td>
<td>S4:3s:E</td>
<td>May '98</td>
</tr>
<tr>
<td>T22.1</td>
<td>S4:3s:W</td>
<td>May '98</td>
</tr>
<tr>
<td>T62.1</td>
<td>S6:1L:W</td>
<td>July '98</td>
</tr>
<tr>
<td>T68.2</td>
<td>S6:1S:W</td>
<td>July '98</td>
</tr>
<tr>
<td>Q6.1</td>
<td>W1:2L:S</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q14.4</td>
<td>W1:4L:S</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q15.1</td>
<td>W1:5L:N</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q16.1</td>
<td>W1:5L:S</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q4.1</td>
<td>W1:1s:S</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q27.1</td>
<td>W2:1:N</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q31.4</td>
<td>W2:4:N</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q32.3</td>
<td>W2:4:S</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q43.5</td>
<td>W2:10:N</td>
<td>Sept '98</td>
</tr>
<tr>
<td>Q59.1</td>
<td>W3:6:N</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q62.5</td>
<td>W3:7:S</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q69.8</td>
<td>W3:11:N</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q70.8</td>
<td>W3:11:S</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q74.4</td>
<td>W4:1:S</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q89.9</td>
<td>W4:9:N</td>
<td>Nov '98</td>
</tr>
<tr>
<td>Q95.3</td>
<td>W4:12:N</td>
<td>Nov '98</td>
</tr>
</tbody>
</table>

* Those isolates with the same symbol have the same allele at the TFA1 locus, see text for further details. 

* The three phrase code (1:2:3) identifying the sample location: first phrase = sample area (see Figure 3.2); second phrase = tree number; third phrase sample aspect (North, South, East or West). See Appendix I for more detailed sampling information.
Microsatellite polymorphism

The TFA1 and FUN12 loci gave clear, sharp bands when run on polyacrylamide gels but the SRP40 locus gave smeared, diffuse and therefore unscoreable bands. The FUN12 locus proved to be uninformative for outcrossing analysis since 29 of the isolates amplified the same banding patterns; the only isolate to yield a different banding pattern was T18a. This difference is noteworthy since this is one of the only two isolates, along with T18.2, that derived from a single bark sample. The difference in banding patterns between these two yeasts indicates that they are not clones, and it therefore appears that yeast from distinct, different clonal lineages may contemporaneously inhabit the same small area of tree bark. All the isolates amplified two bands roughly five b.p. (base pairs) apart for the FUN12 locus, and these two bands were around 200 b.p. in size. The bands amplified from T18a were the same distance apart, however, these two bands were shorter (roughly 3 b.p.) than the corresponding bands in the other isolates. The sizes could not be measured exactly due to an absence of PAGE size marker. A sequence reaction could have been run with the amplicons to estimate size differences; however, such a size marker is extremely expensive and since the FUN12 amplicons are uninformative their exact sizes are irrelevant with respect to outcrossing analysis. The presence of two bands for every isolate could be due to three reasons: 1, The two bands represent two alleles from the FUN12 locus and that all these isolates are heterozygous. This would indicate extreme selection against any homozygous individuals, results from the TFA1 locus (see later) indicate that this is not the case; 2, that the FUN12 gene has been duplicated, and that one copy has a different number of repeats to the other (this locus is not duplicated in S. cerevisiae), or; 3, that the primers designed against the S. cerevisiae FUN12 gene also happen to amplify another area in S. paradoxus and that this
area is of a similar size to the FUN12 amplicon. The sequencing of these bands would discriminate between these alternatives.

Amplicons for all but the Q21.7 isolate were obtained from the TFA1 locus, and these amplicons gave more information concerning the outcrossing rate of wild S. paradoxus since the isolates display polymorphism at this locus. A total of four alleles were found but none of the isolates was heterozygous. The distribution of the homozygous alleles is shown in Table 3.1. The different band sizes were apparent on the autoradiogram of the acrylamide gel, but as no size marker was available representatives from each allele were directly sequenced to check: a, that they were definitely homozygous; and b, the difference in repeat number between alleles. Sequencing confirmed that all these bands represented homozygotes and further that, in principle, any heterozygotes would be detected since the smallest difference in repeat is only six base pairs. The observed allele frequencies for the four TFA1 alleles are (using the symbols for alleles seen in Table 3.1):

\[
\begin{align*}
\text{p (©, 20 repeats)} &= 15/29 = 0.518 \\
\text{q (∨, 22 repeats)} &= 11/29 = 0.379 \\
\text{r (☆, 26 repeats)} &= 2/29 = 0.067 \\
\text{s (Ξ, 28 repeats)} &= 1/29 = 0.034.
\end{align*}
\]

S. paradoxus population structure

The polymorphism found at the TFA1 locus demonstrates that the wild isolates are not all clones of the same lineage. The alleles found at the TFA1 locus are not obviously geographically or contemporaneously structured since individuals with the same allele may be
isolated from different trees within and between samples sites (indeed the same allele is found in both Silwood and Windsor Great Park). Further, different allele types were isolated from the same area of the same tree on two differing occasions. It therefore appears that allele distribution differs over space and time; this could be due to the physical movement of yeasts. Of course the sample size is small and only one locus is being considered, so it may be that many different alleles exist contemporaneously in many different areas. This study did not explore the extent of yeast diversity on a very small scale, as typically only one colony of yeast was selected from each bark sample. There was one case where two yeast colonies were isolated from the same bark sample (T18a and T18.2), these proved to have different alleles for both the TFA1 and FUN12 loci, suggesting that a small area of bark may not be dominated by one genetically distinct clonal lineage.

An estimation of outcrossing rate

The observed frequency of heterozygous genotypes within a population may give an indication as to the amount of inbreeding that population experiences. Inbreeding is defined as ‘sexual reproduction involving fertilisation between gametes from closely related individuals, or in its most extreme form between gametes from the same individual (self fertilisation = selfing)’ (Thain and Hickman, 1994). If the population is experiencing no inbreeding then the frequency of genotypes will be at the equilibrium predicted by the Hardy-Weinberg equation (assuming no selection, migration and mutation), i.e. for two alleles at the same locus with frequencies $p$ and $q$, the expected heterozygous genotype frequency will be $2pq$. If the entire population then experiences strict self fertilisation, what will happen to the
frequencies of the genotypes in the next generation? Consider a single locus with alleles $A$ and $a$ over one generation. Either homozygote will only produce one kind of gamete and these will combine to give the same genotype as that of its parent. A heterozygote, on the other hand, will produce gametes that carry both $A$ and $a$ alleles. When yeasts self it means that ascospores will fuse with their ascus mates, and these will combine to give the three genotypes, which occur on average in the ratio $\frac{1}{6} AA$, $\frac{2}{3} Aa$, and $\frac{1}{6} aa$ (see Figure 3.3). These frequencies occur because the ascospores can only mate with other ascospores that derive from the same meiosis. This frequency differs from that expected with a diploid organism where selfing is achieved through fusion of gametes that derive from independent meiosis, and here one half of the progeny of heterozygotes are expected to be $Aa$ (Mettler and Gregg, 1969). This means that yeast lose heterozygosity less quickly when subjected to a strict system of selfing when compared to other unconstrained systems of selfing. Only $\frac{2}{3}$ of the progeny of heterozygotes are themselves heterozygotes, resulting in an average
Chapter Three: An estimation of natural yeast outcrossing rate

A. All possible pairs with one gamete each from independent meiosis

<table>
<thead>
<tr>
<th>Gametes produced</th>
<th>Potential gametes to mate with</th>
<th>Progeny</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A A a</td>
<td>A A a</td>
<td>A A</td>
<td>1/4</td>
</tr>
<tr>
<td></td>
<td>A a</td>
<td>A a</td>
<td>1/4</td>
</tr>
<tr>
<td></td>
<td>a A</td>
<td>a A</td>
<td>1/4</td>
</tr>
<tr>
<td></td>
<td>a a</td>
<td>a a</td>
<td>1/4</td>
</tr>
</tbody>
</table>

B. All possible pairs from a single meiosis

<table>
<thead>
<tr>
<th>Gametes produced</th>
<th>Potential gametes to mate with</th>
<th>Progeny</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A a</td>
<td>A A a</td>
<td>A A</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>A a</td>
<td>A a</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>a A</td>
<td>a A</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>a a</td>
<td>a a</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Figure 3.3. Diagram to demonstrate expected genotype frequencies given two breeding systems; modelled by a single locus with two alleles represented by A and a. A: the frequency of progeny genotypes expected when a organism that produces independent gametes self fertilizes and the progeny result from the union of two gametes from two independent meioses. B: The frequency of progeny expected when a yeast self fertilizes by mating within the ascus.

71
decrease in heterozygosity of \( \frac{1}{3} \). There is a corresponding evenly divided increase in the two homozygous classes. Since the number of heterozygotes decreases by \( \frac{1}{3} \) every generation of complete selfing, eventually all individuals will be homozygous if selfing persists. The frequency of heterozygotes will be restored to the Hardy-Weinberg equilibrium if any one generation completely outcrosses.

Considerations concerning the yeast mating system that were outlined in the introduction suggest that wild \( S. \) paradoxus may be predisposed to selfing. However, what if a proportion of the \( S. \) paradoxus population are outbreeding? In this case the number of heterozygotes in the next generation will be determined by:

\[
Het' = \frac{3}{2} Het(s) + 2pq(t) \quad ,
\]

[Eq. 3.1]

where \( Het \) equals the frequency of heterozygotes in one generation, \( Het' \) equals the frequency of heterozygotes in the next generation, \( s \) is the proportion of the population which are derived from selfed matings, and \( t \) is the proportion which are derived from outcrossed matings (\( s + t = 1 \)).

If it assumed that there is no selection, that \( t \) is constant in every generation, and that an individual’s mode of mating is independent of its parent’s mode, then the frequency of observed heterozygotes will reach an equilibrium. This equilibrium can be found by letting \( Het' = Het \) and then solving for \( Het' \):

72
Chapter Three: An estimation of natural yeast outcrossing rate

\[
\begin{align*}
\begin{align*}
H_{\text{et}} &= \frac{2pqf}{1 - s \frac{2}{3}} \\
\text{[Eq. 3.2]}
\end{align*}
\end{align*}
\]

The frequency of observed heterozygotes will be critically determined by \( s \) and \( t \). The greater the proportion of the population that outcross, the closer to \( 2pq \) (the Hardy-Weinberg value) will the observed heterozygote frequency be. If we assume that the \textit{S. paradoxus} population has reached this equilibrium then the degree to which this equilibrium differs from the predicted Hardy-Weinberg genotype frequency value may be calculated. If \( H_{\text{et}} \) equals the observed heterozygote frequency, which has reached equilibrium (then \( H_{\text{et}} = H_{\text{et}}^{*} \)), as it would if the population is subjected to a constant \( s \) and \( t \), then the difference from the predicted Hardy-Weinberg heterozygous genotype frequency (\( H_{\text{et-H-W}} \)) is:

\[
\frac{H_{\text{et}}}{H_{\text{et-H-W}}} = \frac{2pqf}{2pq} = \frac{t}{1 - s \frac{2}{3}} \quad \text{[Eq. 3.3]}
\]

and thus the average inbreeding coefficient (Wright's \( F \)) of a yeast population with a constant \( t \) is:

\[
\bar{F} = 1 - \frac{H_{\text{et}}}{H_{\text{et-H-W}}} = 1 - \frac{t}{1 - s \frac{2}{3}} \quad \text{[Eq. 3.4]}
\]

and substituting \( s \) for \( 1-t \) this simplifies to:
Equation 3.5 is the equivalent to Falconer's Equation 5.15 (Falconer, 1981) where he calculates the average inbreeding coefficient at equilibrium for an organism which self fertilises through gametes from independent meiosis.

It is now possible to find the value for $t$ given the data concerning the TFA1 locus since Equation 3.5 allows one to calculate the degree to which the average inbreeding coefficient of the observed population differs from the predicted Hardy-Weinberg equilibrium. First Equation 3.5 must be put in terms of $t$:

$$t = \frac{1 - \tilde{F}}{2\tilde{F} + 1}$$  

[Eq. 3.6]

The value of $\tilde{F}$ for the TFA1 locus may be calculated using Equation 3.4. The frequency of observed heterozygotes is zero and the expected value given the frequencies of the four alleles is 0.582 ($2pq + 2pr + 2ps + 2qr + 2qs + 2rs$, where $p$, $q$, $r$ and $s$ are the frequencies of the four alleles). This equation may now be solved to give a best estimate for $t$ of zero. Therefore, under the assumption that the S. paradoxus population has constant a rate of outcrossing, and that this rate occurs randomly to individuals each generation, it is estimated, given the data concerning the TFA1 locus, that none of the population are outcrossing.
This estimation is unsurprising given that no heterozygotes were observed in the S. paradoxus wild isolates. However, the above calculations will be of future use when more data concerning the frequency of heterozygotes, both for this locus and other loci, are gathered. Earlier considerations of yeast biology are in keeping with this estimate since the morphology of the ascus and the ability to mate type switch leave yeast predisposed to inbreeding. The lack of data (only one informative locus) and small sample size mean that a relatively large degree of sampling error will surround this estimate.

An upper bound of the outcrossing rate may be calculated using maximum likelihood methods. There will be values of \( t \) above zero that will still mean that 29 homozygous individuals may be randomly selected from a population (only 29/30 wild isolates amplified a product with the TFA1 primers). To find this upper bound the maximum likelihood estimate of \( t \) given the data and model must be found. Then, the upper bound for the outcrossing rate will equal the 2 unit Ln likelihood support value for \( t \) (Edwards, 1972).

A population with a constant \( t \), and where the mode of reproduction (selfing or outcrossing) for an individual is independent of their parents mode, will contain different individuals who will have experienced a differing number of generations since they last outcrossed. Even with a constant \( t \) there will be, by chance, some individuals who have been through many generations since they last outcrossed and some that have been through very few. These 'classes' will form a frequency distribution. The probability that a randomly selected individual has gone through \( n \) generations of selfing since they last outcrossed is given by \( (1-t)^n \). The higher the value of \( t \) the less chance that a random individual will have experienced many
successive generations of inbreeding (i.e. as \( n \) increases the probability becomes rapidly smaller with a larger value of \( t \) when compared to a smaller value of \( t \)). Further, an individual that has experienced \( n \) generations of selfing still has a probability of being heterozygous. For example starting at \( H_{het}W \) \((2pq)\) after one generation of selfing there is \( \frac{2}{3} \cdot 2pq \) chance that that a selfed individual is heterozygous (see Equation 3.1). The probability that an individual with \( n \) generations of selfing will be heterozygous is therefore \((\frac{2}{3})^n 2pq\).

To find an average probability of being heterozygous, the probability that an individual with \( n \) generations of selfing is heterozygous \( [(\frac{2}{3})^n 2pq] \) is multiplied by the frequency of that class in the population \([(1-t)^n f]\). The probability of a random individual being heterozygous (\( p_{Het} \)) is found by summing the product of these probabilities for all values of \( n \), thus finding the weighted average of these probabilities:

\[
p_{Het} = \sum_{n=0}^{\infty} (1-t)^n f \left(\frac{2}{3}\right)^n 2pq
\]

[Eq. 3.7]

The probability of randomly selecting a homozygote (\( p_{Hom} \)) is \( 1-p_{Het} \). The probability of randomly selecting 29 homozygotes is therefore:

\[
p_{Hom}^{29} = \left[ 1 - \sum_{n=0}^{\infty} (1-t)^n f \left(\frac{2}{3}\right)^n 2pq \right]^{29}
\]

[Eq. 3.8]

The four alleles found at the \( TFA1 \) locus have the estimated frequencies of \( p = 0.518 \), \( q = 0.379 \), \( r = 0.067 \) and \( s = 0.034 \), and this gives a predicted Hardy-Weinberg heterozygote
genotype frequency of 0.582. Substituting 0.582 in Equation 3.8 for $2pq$, the maximum likelihood estimate for $t$ given the observed data is found to be 0 (see Figure 3.4). However, the upper 2 unit support limit for $t$ is found to be 0.041 [calculated with Mathematica v 3.0 (Wolfram, 1996), file available on request]. This still means that the best estimate for the natural S. paradoxus population outcrossing rate is 0, but given these data a rate of up to 4% cannot be discounted.
Figure 3.4 Log likelihood surface for values of $t$, calculated with Equation 3.8 in Mathematica v3.0 (Wolfram, 1996). Given the data concerning the TFA1 locus the likelihood is maximised when $t = 0$, and the 2 unit log likelihood support limit is 0.041.
Conclusions

The very low estimated levels of outcrossing in the natural S. paradoxus population of Silwood and Windsor Great Park may explain the absence (or very low frequencies) of the \( \omega \) and VDE HEGs from this population. With such a low outcrossing rate the increase in frequency of any HEG (or any driving gene) would be substantially lower than could be realised in a fully outcrossed population. But how much less? A model to predict the increase in frequency of a driving allele given a certain constant rate of outcrossing has previously been constructed (Burt and Trivers, 1998). The model predicts the change in frequency of B – chromosomes in plant populations whilst taking into account certain levels of inbreeding. Here, this model has been modified to take into account the differing rate of heterozygote loss in yeast due to intra-ascus mating (see Figure 3.3), and the fact that drive occurs equally in all gametes (ascospores) of yeast. The change in frequency of HEGs may be modelled by allowing HEGs to represent two alleles (HEG\(^{-}\) and HEG\(^{+}\)) found at a single locus. Consider a hypothetical population of yeast with non-overlapping generations, and with no selection against a HEG\(^{+}\) allele, in which spores may come from either selfed or randomly mated yeast. If the frequency of the HEG\(^{-}\) and HEG\(^{+}\) alleles are \( p \) and \( q \) respectively these will give three genotype frequencies \( x \), \( y \) and \( z \), where:

\[
x = p^2 + Fpq, \quad y = 2pq(1 - F), \quad z = q^2 + Fpq
\]

[Eq. 3.9]

and \( F \) is the average inbreeding coefficient of the population (defined in Equation 3.4). The spores produced from these diploid parents will have an HEG\(^{+}\) frequency of:
Chapter Three: An estimation of natural yeast outcrossing rate

\[ u = yd + z \quad , \quad \text{[Eq. 3.10]} \]

where \( d \) is the rate of drive displayed by the HEG\(^+\) allele. The frequency of diploids in the next generation will be:

\[
\begin{align*}
x' &= s(x + yJ) + t(1 - u)^2 \\
y' &= sGy + t[2u(1 - u)] \\
z' &= s(yHz) + tu^2
\end{align*}
\quad , \quad \text{[Eq. 3.11]}
\]

and \( J, G \) and \( H \) are the frequencies of HEG\(^-\)/HEG\(^-\), HEG\(^-\)/HEG\(^+\), and HEG\(^+\)/HEG\(^+\) genotypes, that result from heterozygote yeast that mate within the ascus. These are defined by:

\[
\begin{align*}
J &= \frac{1}{6}(1 - d)^2 \\
G &= \frac{3}{4}(1 - d)^2 + (1 - d)d \\
H &= \frac{4}{5}(1 - d)^2 + (1 - d)d + d^2
\end{align*}
\quad \text{[Eq. 3.12]}
\]

The frequency of the HEG\(^-\) and HEG\(^+\) alleles in the next generation will therefore be:

\[
\begin{align*}
p' &= x' + 0.5y' \\
g' &= z' + 0.5y'
\end{align*}
\quad \text{[Eq. 3.13]}
\]
Chapter Three: An estimation of natural yeast outcrossing rate

The average inbreeding coefficient of the next generation may be found by \( F' = 1 - y f(2p'q') \).

These equations were used to calculate the change in frequency of an HEG (with no selection coefficient) over successive generations when \( t = 0.04 \), which is the upper bound previously calculated for the *S. paradoxus* population of Silwood and Windsor Great Park. This HEG change in frequency in a highly inbred yeast population is compared with an HEG change in frequency in a completely outcrossed population in Figure 3.5.

![Graph showing the predicted change in frequency of an HEG in highly inbred and fully outcrossed populations of yeast. The graph demonstrates that an HEG increases in frequency at a much slower rate within highly inbred populations when compared with outcrossed ones.](image)

**Figure 3.5.** The predicted change in frequency of an HEG in highly inbred \((t = 0.04)\) and fully outcrossed \((t = 1)\) populations of yeast. These frequency changes were calculated in Mathematica v 3.0 (Wolfram, 1996) using Equation 3.13 where the HEG+ allele starting frequency \((q)\) was 0.01 and \( d = 0.95 \). The population was started at Hardy-Weinberg equilibrium \((F = 0)\) and then constantly inbred with a rate of \(1-t\). There is no selection against the HEG+ allele.

It can be seen from Figure 3.5 that an HEG increases in frequency at a much slower rate within highly inbred populations when compared with outcrossed ones. This formal model reinforces verbal reasoning concerning the fate of HEGs within inbred populations. In a 96% inbred population it takes roughly 50 generations for an HEG to go from a frequency of 0.01 to 0.1; in a fully outcrossed population this takes just four generations (an order of magnitude
less). Such driving genes are unable to realise the potential for quick spread as they are rarely able to infect inbred HEG− lineages. Such a slow increase in frequency may allow selection to act more effectively on such alleles meaning that small HEG selection coefficients may ultimately mean their removal from highly inbred populations. The plot shown in Figure 3.5 is a general case for HEGs where there is no selection against the HEG+ allele and the rate of HEG drive (d) is 95%. If these changes in frequency are recalculated for the more specific VDE case, where \(d = 0.88\) (see Chapter Two), then the relative fitness of a VDE+ allele may be calculated (using Equation 2.1). When VDE is at a low frequency (\(\leq 0.01\)) within a 4% outcrossing population, then its relative fitness (due to the ability to home in heterozygous yeast) is +1.04%. This value is not enough to counterbalance the —12% relative fitness deficit calculated for VDE in Chapter Two. With such a large (probably artificial) selection coefficient it is unsurprising that VDE is predicted to be lost from largely inbreeding populations. In such populations there are very few opportunities for VDE to override selection via homing. Indeed, selection upon individual yeast cells may be facilitated by such breeding systems. In S. cerevisiae cultures found in fermentants there is evidence to suggest that certain ‘fit’ haploids, that have mate type switched thus becoming completely homozygous diploids, may subsequently go on to out compete, and therefore replace, their heterozygous parents (Mortimer et al., 1994). If such processes occur in natural yeast populations they may serve to remove detrimental alleles from the population by reducing population heterozygosity.

\[\text{Calculated with Equation 3.13 where } q = 0.01 \rightarrow 0.00000001, d = 0.88, t = 0.04, F = 0.958\]
The above model assumes that the outcrossing rate within a population remains constant over time; the violation of this assumption may have radical effects on the change in frequency of selfish genes. Even a small increase in outcrossing rate may mean that driving genes will be able to more effectively permeate a population. For example, if all other parameters in the model are kept the same as those used to simulate the change in HEG frequency seen in Figure 3.4, apart from \( t \) which is raised from 0.04 to 0.1 (approximately doubled), then the HEG\(^+\) allele is predicted to spread at twice the speed within the population. From a starting frequency of 0.01 the HEG\(^+\) allele reaches a frequency of approximately 0.5 in only 50 generations compared to the 100 generations needed when \( t = 0.04 \). If the outcrossing rate in natural populations of yeast fluctuates this may well have radical effects on the frequency of any driving genes within these populations. It is currently unknown if the outcrossing rate fluctuates in wild populations of yeast. Indeed, the average time between generations (from sporulation to sporulation) is not even known for natural yeast.

The lack of HEGs within the sampled \( S. \) paradoxus population would be predicted with the very low (perhaps non existent) levels of outcrossing since HEGs are not predicted to spread in largely inbred yeast populations. But what of the longer term fate of HEGs? It may be that the common ancestor of \( S. \) paradoxus and \( S. \) cerevisiae did not contain either \( VDE \) or \( \omega \), and that \( S. \) cerevisiae has subsequently gained these two HEGs by horizontal transfer events. Considerations concerning the HEG status of more ancestral yeast, and the observation that all of the wild isolates contain the \( \omega \) element group I intron, suggest that this may not be the case. Therefore, alternatively, it may be that \( S. \) paradoxus has lost these HEGs. The levels of outcrossing estimated for the Silwood and Windsor Great Park population may be a fair
reflection of the outcrossing rate for all natural populations of *S. paradoxus*. If this is the case, the breeding system in *S. paradoxus* may have facilitated the selection against and removal of HEGs from the species. If this is so why has the \( \omega \) element group I intron not been removed? The detrimental load estimated for VDE in Chapter Two may be an unfair reflection of the degree to which HEGs hinder their host cell. If the parasitic load that HEGs confer is much lower than 12%, then only relatively a small amount of outcrossing will be needed to keep slightly deleterious allele within the populations. For example, Futcher et al (1987) estimated that the 2\( \mu \)m plasmid, which has a parasitic load of 1%, only requires an outcrossing rate of 0.0002% per generation to keep the plasmid at a 99% frequency in that population. It is therefore possible that some other processes are affecting the longer term population dynamics of HEGs, and these possibilities are addressed in the next chapter.
Chapter Four

THE LONG-TERM FATE OF HEGs

Note: the information contained within this chapter may also be found, in an alternative form, as a journal article: M. R. Goddard and A. Burt (1999) Recurrent invasion and extinction of a selfish gene. Proceedings of the National Academy of Sciences USA. Vol. 96, No.24, pp13880-13885.
Chapter Four: The long-term fate of HEGs

Introduction

In the preceding chapters it was demonstrated that the theoretical prediction of HEG spread, despite having no selective advantage, is indeed observed in experimental populations. These observations appear to explain the existence of HEGs in the short term at least. But what of HEGs, and selfish genes in general, over longer time scales: once at fixation, what then? When an HEG is at fixation there are, by definition, no 'empty' sites to home into: the positive selective pressure upon HEGs to maintain function will therefore be reduced. Mutations will then slowly accumulate to the point where function is destroyed, and beyond this point HEG loss cannot be counter balanced by homing since they cease to exhibit non-Mendelian behaviour. The loss of an HEG will be selectively favoured if the HEG has a detrimental effect. One could therefore predict that within a population HEGs will display a quick boom followed by a slow bust. The idea that HEGs may be lost from within a population is reflected in the results of the wild isolate survey, in which none of those studied contained either VDE (nuclear HEG) or $\omega$ (mitochondrial HEG; see Chapter Three). Although this S. paradoxus population appears not to contain these two HEGs, an ever increasing number of HEGs are being reported. If it is true that HEG degeneration is followed by loss then how does one explain this ever increasing list (Belfort and Roberts, 1997; Mueller et al., 1993)? One hypothesis suggests that the sporadic observation of such genes is due to the fact that they are transient in nature, and that HEGs are able to escape degeneration either by copying themselves into new genomic locations (transposition), or by moving to a different species genome altogether (horizontal transfer). Once in a new location an HEG would once more
quickly go to fixation before slowly degenerating. If such movements are frequent enough they may allow HEGs to evade extinction through the recurrent invasion of new genomic sites or new genomes. Comparative sequence analysis shows that the horizontal transfer of self-splicing group I introns does occur occasionally but, to date, there has been no evidence for HEG movement (Paquin et al., 1994; Turmel et al., 1995; Vaughn et al., 1995). Even if one were to assume that all group I intron mobility is facilitated by HEGs (see Chapter One), then the data thus far are still too diffuse to allow one to estimate the frequency of movement. To address this question directly a particular element called $\omega$ ("omega", also known as r1 and Sc LSU.1) (Dujon, 1980; Jacquier and Dujon, 1983) has been analysed in a series of closely related saccharomycete yeast.

The $\omega$ element comprises an HEG (the $\omega$ HEG product is termed l-Scel) and a group I intron (see Figure 4.1): such associations are common for HEGs (Lambowitz and Belfort, 1993). $\omega$ was first discovered in S. cerevisiae interrupting the mitochondrial large sub-unit (LSU) rRNA gene (Dujon, 1980). The secondary and tertiary structure possessed by group I introns allows them to self-splice out of mRNA transcripts, the $\omega$ HEG being therefore spliced out along with the rest of the intron (Cech, 1990). Without a
Figure 4.1. A: Structure of ω. White boxes represent the mitochondrial LSU rDNA, black boxes represent the ω group I intron, and the stippled box represents the ω HEG. The two boxes with horizontal lines represent the endonuclease recognition site, which is interrupted by the presence of ω. The arrows show the PCR primer binding sites. Not to scale. B: Consensus primary and secondary structure of the group I introns showing base-paired regions P1 to P9.3. If present, HEGs are found in region P8. Solid black lines represent areas of secondary structure that are present in all introns but are not sequentially conserved. Dashed arrows connect nucleotides that have been separated for ease of display. Stem-loop structures in grey are optional or unalignable (P6b is only found in K. debilwattii, K. lactis, Z. hisporus and Z. rouxii; P9.2 is only found in S. castelli and S. mukitii) and were not included in the phylogenetic analysis. Areas A-H, indicated by thick grey lines, are homologous areas that were aligned independently and used in phylogenetic analysis. Exon sequences are in lower case letters.
surrounding group I intron the $\omega$ HEG presumably could not be tolerated within the rRNA LSU gene as this massive insertion would destroy the gene's function. Following crosses between $\omega^+$ and $\omega^-$ individuals 99% of progeny are $\omega^+$ (noting that yeast mitochondria are biparentally inherited and recombine) (Jacquier and Dujon, 1985). This pattern of non-Mendelian inheritance means that $\omega$ is predicted to quickly spread in an outcrossed population; indeed this spread has been observed for another HEG (see Chapter Two). It will take about 15 fully outcrossed generations for $\omega$ to go from a frequency of 0.001 to 0.999 (calculated using Equation 1.1).

Elements very similar to $\omega$ have been sequenced from *Kluyveromyces thermotolerans* and *Pichia canadensis*, and both are inserted at the homologous position in the LSU rRNA (Jacquier and Dujon, 1983). Similar homologous elements have also been found in *S. paradoxus* and *K. lactis*, although curiously these consist of group I introns which do not contain HEGs (Ragnini et al., 1991; Wilson and Fukuhara, 1991). Without an HEG these elements are not expected to home, though this has yet to be experimentally demonstrated. *K. marxianus* also has an HEG$^-$ group I intron similar in sequence to $\omega$ in its mitochondrial ATPase subunit 9 gene (Dujon et al., 1986) (see GenBank accession no. U75348 for a similar element in the same gene of *K. lactis*). Further, various genomes of yeast have been probed with radiolabelled $\omega$ sequences, and the results prove that many taxa do not contain $\omega$, or if they do it has diverged so greatly in sequence so as to be undetectable (Skelly and Maleszka, 1991; Wilson and Fukuhara, 1991). There have been no previous attempts to determine the
phylogenetic distribution of $\omega$. In the absence of phylogenetic information tests for the horizontal transfer of $\omega$ cannot be performed.

To address the question of $\omega$ phylogenetic distribution and horizontal transfer twenty closely related species of yeast will be surveyed for the $\omega$ element using external PCR primers. The size of the resulting amplicons will reveal if an element is present and if so whether it contains an HEG (see Figure 4.1). Phylogenetic trees will then be reconstructed for both the host genome and $\omega$ element. The survey results and subsequent phylogenetic analyses will allow, firstly, tests for $\omega$ horizontal transfer and secondly, an estimation of the rate of $\omega$ movement.
Chapter Four: The long-term fate of HEGs

Materials and Methods

Strains

Eighteen species of Saccharomycetaceae (Ascomycota: Hemiascomycetes: Saccharomycetales) yeast were obtained from Centraalbureau Voor Schimmelcultures (CBS), spanning four genera: Saccharomyces, Torulaspora, Kluyveromyces and Zygosaccharomyces. Table 4.1 shows the species, with their CBS numbers, which were surveyed for $\omega$. In addition two species, S. mikatii and S. cariocus (kindly supplied by E. Louis), were included in the survey. These isolates have been placed tentatively in the Saccharomyces sensu stricto group (Vaughan-Martini and Kurtzman, 1985), based on hybridization and chromosome analysis (Naumov et al., 1995a; Naumov et al., 1995b).
<table>
<thead>
<tr>
<th>Taxon</th>
<th>Strain Number</th>
<th>Genbank Accession Number</th>
<th>Intron‡</th>
<th>ITS1-5.8S-ITS2</th>
<th>18S</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>1171&lt;sup&gt;T&lt;/sup&gt;</td>
<td>V00684</td>
<td>AJ229057</td>
<td>Z75578</td>
<td></td>
</tr>
<tr>
<td><em>S. paradoxus</em></td>
<td>432&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229046</td>
<td>AJ229059</td>
<td>X97806</td>
<td></td>
</tr>
<tr>
<td><em>S. bayanus</em></td>
<td>380&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ229058</td>
<td>X97777</td>
<td></td>
</tr>
<tr>
<td><em>S. pastorianus</em></td>
<td>1538&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ229060</td>
<td>X97805</td>
<td></td>
</tr>
<tr>
<td><em>S. cariocus</em></td>
<td>-</td>
<td>AJ229045</td>
<td>AJ229061</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td><em>S. mikatii</em></td>
<td>-</td>
<td>AJ229048</td>
<td>AJ229064</td>
<td>n/a</td>
<td></td>
</tr>
<tr>
<td><em>S. exiguus</em></td>
<td>379&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229047</td>
<td>AJ229063</td>
<td>X98868</td>
<td></td>
</tr>
<tr>
<td><em>S. unisporus</em></td>
<td>398&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ229065</td>
<td>Z75582</td>
<td></td>
</tr>
<tr>
<td><em>S. castellii</em></td>
<td>4309&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229049</td>
<td>AJ229062</td>
<td>Z75881</td>
<td></td>
</tr>
<tr>
<td><em>S. dairenensis</em></td>
<td>421&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ2290672</td>
<td>Z75579</td>
<td></td>
</tr>
<tr>
<td><em>T. pretoriensis</em></td>
<td>5080</td>
<td>AJ229051</td>
<td>AJ229066</td>
<td>X84638</td>
<td></td>
</tr>
<tr>
<td><em>T. delbrueckii</em></td>
<td>404&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229052</td>
<td>AJ229075</td>
<td>X98120</td>
<td></td>
</tr>
<tr>
<td><em>T. globosa</em></td>
<td>764&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229053</td>
<td>AJ229074</td>
<td>X84639</td>
<td></td>
</tr>
<tr>
<td><em>Z. rouxii</em></td>
<td>688</td>
<td>AJ229050</td>
<td>AJ229071</td>
<td>X58057</td>
<td></td>
</tr>
<tr>
<td><em>Z. bailii</em></td>
<td>685&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ22906770</td>
<td>X91083</td>
<td></td>
</tr>
<tr>
<td><em>Z. bisporus</em></td>
<td>702&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229056</td>
<td>AJ229176</td>
<td>X91084</td>
<td></td>
</tr>
<tr>
<td><em>K. polysporus</em></td>
<td>2163&lt;sup&gt;T&lt;/sup&gt;</td>
<td>-</td>
<td>AJ229076</td>
<td>X83825</td>
<td></td>
</tr>
<tr>
<td><em>K. thermotolerans</em></td>
<td>6924</td>
<td>X00143</td>
<td>AJ229073</td>
<td>X89526</td>
<td></td>
</tr>
<tr>
<td><em>K. dozhanskii</em></td>
<td>2104&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229054</td>
<td>AJ229068</td>
<td>D83430</td>
<td></td>
</tr>
<tr>
<td><em>K. lactis</em></td>
<td>683&lt;sup&gt;T&lt;/sup&gt;</td>
<td>AJ229055</td>
<td>AJ229069</td>
<td>X51830</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1. Species, strains and Genbank accession numbers for sequences used in this study. † denotes a type strain (type strains are reference strains for that species). ‡ Accession numbers in bold are ω elements which consist of group I introns that contain an HEG.
Molecular Methods

DNA was extracted from overnight cultures of yeast following Strathern and Higgins (1991). The PCR was used to assay for the presence or absence of the intron using primers ω01 and ω02 which were designed from an alignment of *S. cerevisiae* and *Pichia canadensis* mt LSU rRNA genes. Amplicons containing introns were sequenced directly using the Amersham Thermo Sequence™ dye terminator cycle sequencing kit, and then analysed with an ABI 373 automated sequencer. Forward and reverse sequences were obtained in every case. The internal primers ω la and ω lb were designed to facilitate the sequencing of the larger PCR products, primer sequences are shown in Table 4.2.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Sequence (5' – 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>ω01</td>
<td>GATAACGAATAAAAGTTACGCTAGGG</td>
</tr>
<tr>
<td>ω02</td>
<td>CTTCAGCAGATAGGAACCATACTG</td>
</tr>
<tr>
<td>ω05</td>
<td>CTACTCTTTGCTATATTACC</td>
</tr>
<tr>
<td>ω06</td>
<td>GGGTAATATAGCGAAGAGTAG</td>
</tr>
<tr>
<td>ω la*</td>
<td>GARAGAGTWAATCATTTA</td>
</tr>
<tr>
<td>ω lb*</td>
<td>TGGTTAAAWGTTGAGCYCC</td>
</tr>
<tr>
<td>ITS1§</td>
<td>TCCGTAGGTGAACCTGCGG</td>
</tr>
<tr>
<td>ITS4§</td>
<td>TCCTCCGCTTATTGATATGC</td>
</tr>
</tbody>
</table>

Table 4.2. Sequences of the primers used to study the ecology and evolution of the ω element in the Saccharomycetaceae yeast. Asterisked primers contain degenerate bases. § primer sequences derive from White et al. (1990).
Additionally, the sequences of \( \omega \) from \( S. \text{ cerevisiae} \) (Sc LSU.1) and \( K. \text{ thermotolerans} \) (Kt LSU.1), described by Jacquier and Dujon, (1983), are included in this study. To address the question of possible \( \omega \) horizontal transfer to or from taxa outside of the Saccharomycetaceae, Genbank was searched for non-Saccharomycetaceae taxa that contain an intron at the \( \omega \) insertion site. Genbank (release 105.0) contains seven non-Saccharomycetaceae mt LSU rRNA sequences with a group I intron at the \( \omega \) insertion site, and four of these seven group I introns contain an open reading frame (ORF) within the group I intron. However, none of these ORFs aligned at all well, either totally or partially, with the HEGs found in the 20 yeast surveyed. Only one other homologous element was found, from \( P. \text{ canadensis} \), a more distantly related Saccharomycetaceae yeast, and this sequence was included for analysis (D31785; a group I intron with an HEG).

To discover whether or not \( \omega \) is present at sites ectopic to the LSU position, that is to address the question of \( \omega \) transposition, a Southern blot of all 20 yeast species’ genomes was made. Yeast genomes were first cut with Hae III (a restriction enzyme that has a four base pair recognition site of GG\( \underline{CC} \)) and then electrophoresed through a 1% agarose gel. The electrophoresed DNA was then capillary transferred to a nylon membrane following Sambrook et al. (1989). This resulting 'zooblot' (a Southern blot of many species) was probed with a pool of \( ^{32} \text{P} \) labelled \( S. \text{ paradoxus} \) and \( K. \text{ lactis} \) \( \omega \) introns. The membrane was hybridised at 55°C in the presence of 5xSSC, 0.5% SDS, 5x Denhardt’s solution with 130ng of denatured salmon sperm DNA; after hybridisation it was washed in 1xSSC:0.1% SDS at 65°C for 15 minutes. In order to minimise the amount of LSU sequence present in the probes.
the S. paradoxus and K. lactis ω products were amplified using primers ω01 and ω05 (see Figure 4.1 A); positive controls proved that the small amount of LSU sequence present in these probes was not sufficient to produce false positives. The predicted number of Hae III cleavage sites within each ω intron were compared to the number of bands visualised on the autoradiogram, and then to a positive control where the zooblot had been probed with 200bp of the 3' flanking LSU amplified using primers ω06 and ω02.

To reconstruct a host genome phylogeny the nuclear Internal Transcribed Spacer region (ITS1 - 5.8S - ITS2) was amplified and directly sequenced for all the 20 yeast species using primers ITS1 and ITS4. Table 4.1 shows the accession numbers for the ITS1-5.8S-ITS2 sequences. Additionally, 18S sequence data for the eighteen CBS yeast species were obtained from Genbank (Table 4.1).

**Analyses**

An initial alignment of the group I introns was ambiguous. In order to assist alignment the secondary structure of all the group I introns were obtained by folding them in sections with the m-fold (v2.3) server (http://www.ibc.wustl.edu/~zuker/rna): folding temperatures were set to 25°C and all other parameters were kept as default. When folding, the overall secondary structures of the group I introns were obtained using the characteristic group I intron secondary structure as a guide (Cech, 1988). These secondary structures defined eight homologous group I intronic areas which were then aligned separately. These areas are shown in Figure 4.1 marked A-H, and exclude unalignable loops and optional stem-loop
regions. All sequence alignments were carried out using ClustalW (Thompson et al., 1994) for the Apple Macintosh, with all the parameters kept as default. Base pairs found on the reciprocal sides of stem regions are unlikely to be independent, since a change in one of the pairing will select for a compensatory change in the other in order to preserve the secondary and tertiary structure. One half of the base pairs in every stem structure were therefore removed for the construction of the final data matrix. The 18S and ITS1-5.8S-ITS2 sequences were utilised since they are assumed to represent the evolutionary history of the yeasts' genomes (James, et al., 1986; James, et al, 1987). The 18S sequences aligned well across all taxa and in the final data set gaps were treated as a fifth base. Of the ITS1-5.8S-ITS2 region only the 5.8S gene and a small area of the ITS2 were alignable across all taxa; these data were combined with the 18S data as bootstrap trees reconstructed from each data set contained no branches over 50% which contradicted each other. This combined data still left the closely related sensu stricto complex unresolved. However, the complete ITS data aligned well for these taxa and were therefore employed to resolve this group. HEG sequences were aligned at both the DNA and amino acid level. Phylogenetic analyses were carried out using PAUP* (test version 4.0d64) (Swofford, 1999). Partition homogeneity and Kishino-Hasegawa tests were carried out within PAUP* and intron assignment permutations were carried out with MacClade (Maddison and Maddison, 1992).

The method described for estimating the minimum number of horizontal transfer events involves taking a tree and generating all possible rearrangements which could result from a single horizontal transfer event. This was simulated with the subtree pruning regrafting (SPR) algorithm found in PAUP*. To ensure PAUP* makes only one round of SPR swapping on any
one starting tree (so that one horizontal transfer event may be analysed at a time in a stepwise procedure) a degenerate data set must be constructed (a data set where all characters are the same). With a degenerate data set PAUP* will construct all one SPR rearrangements of the starting tree: since the data set is degenerate none of these trees will be better and therefore PAUP* will terminate the search. More specific details of this procedure may be found in the Appendix II. Not all the SPR rearrangements will be valid if a rooted tree is to be used: some infer impossible horizontal transfer to an ancestor of the donor, hence rearrangements that are utilised must be manually checked. Finally, the maximum likelihood transition estimates were calculated by Dr. A. Burt using a specifically written Mathematica (v 3.0.1; Wolfram, 1996) file.
Results and discussion

ω survey

Fourteen of the twenty species contained an intron at the ω insertion site, but only five of these fourteen introns contained an HEG (see Figure 4.2). Two of the HEGs, those found in S. exigus and T. delbrueckii, have insertions within them which destroy their reading frames: these HEGs are presumably non-functional. The most common state at the ω insertion site is therefore ‘non-functional ω element’ (non-functional either because there is no HEG, or because the HEG present is itself non-functional), the next most common state is ‘empty site’, and the least common ‘functional ω element’. Firstly the question of ω transposition must be addressed, does an ω element exist elsewhere in the genome? To answer this question the 20 species genomes were probed with the S. paradoxus and K. lactis LSU ω sequences. The 14 taxa predicted to show strong hybridisation signals did so; further they showed one strong band indicating no ectopic genomic locations of ω. However, additional weaker bands were observed in K. lactis and K. dozhanskii. These bands are presumably the ATPase subunit 9 introns which were mentioned in the introduction since they show limited similarity to the ω group I introns. These results suggest that the transposition of ω is an extremely rare event, and the fact that the ATPase intron’s similarity to ω is weak implies a single, ancient transposition event, although, without further data, the direction of movement cannot be determined.
\( \omega \) phylogenetic correlation

Next \( \omega \) states were mapped onto the host phylogeny (Figure 4.2). The distribution of the intron upon the phylogeny suggests that \( \omega \) horizontal transfer has occurred. If no horizontal transfer had occurred then loss at a deep node would have produced an entire intron free clade: this would have the effect of clumping the \( \omega \) intron states upon the host phylogeny. A visual inspection of the intron state distribution suggests no clumping. However, clumping and loss were statistically tested by randomising the intron state assignment 100 times; this essentially simulated an infinite number of \( \omega \) horizontal transfer events. For each of the 100 randomisations the amount of clumping was ascertained by counting the number of intron losses, and these were then combined into a frequency distribution. The inferred number of losses from the observed data (nine HEG losses and four intron losses) were then compared to the randomised frequency distribution; the results of this test show that the \( \omega \) element is not significantly clumped upon the host phylogeny (\( p=0.82 \) and \( p=0.19 \), see Figure 4.2). These results support the alternate hypothesis which invokes an explanation of \( \omega \) evolution that includes horizontal transfer of as well as loss.

A far more robust test for horizontal transfer makes use of the phylogenetic information from both the \( \omega \) and the host sequence data. Farris's partition homogeneity test (PHT) (Farris et al., 1995) may be implemented in PAUP*: this tests two data sets for differing phylogenetic signals. The two data sets should share the same evolutionary history (phylogenetic signal) if there has been no \( \omega \) horizontal transfer; however, if \( \omega \) horizontal transfer has occurred then
the two will have significantly different signals. The PHT initially finds the shortest tree from each data partition (host and intron) and then sums these minimum tree lengths. Next the characters are randomly shuffled between the two data partitions \( n \) times to produce \( n \) replicate data sets that are a chimera of characters from each. The shortest trees are then found and summed for each of the shuffled data partitions and the \( n \) summations are combined to produce a frequency distribution to which the original, observed, tree length sum is compared. If the two data partitions share the same phylogenetic signal then a chimera of the characters will converge upon very similar trees which will have similar lengths, and the mean of the frequency distribution for the shuffled data should therefore be similar to the original observed sum. However, if the two data sets have different phylogenetic signals then the inclusion of characters from one data set into another will reduce the strength of phylogenetic signal by introducing homoplasy. The mean of the tree lengths resulting from the shuffled replicates will be significantly longer when compared to the observed tree length sum. The \( \omega \) element comprises of two distinct areas: the group I intron and HEG. The difference in HEG and group I intron phylogenetic signals were tested first as there have been suggestions that HEGs may move independently between introns (Bechhofer et al., 1994) (although no mechanisms have yet been proposed). The PHT (with 1000 randomisations) indicates that the HEG and group I intron share the same evolutionary
Figure 4.3. Phylogenetic comparisons between intron, HEG, and host data sets. The histogram associated with each tanglegram shows the distribution of summed tree lengths for random partitions of the data, and arrows show the observed length for the actual partition. All phylogenies were reconstructed via a maximum parsimony branch and bound bootstrap analysis (1000 replicates, bootstrap scores shown), and trees reconstructed with maximum likelihood methods gave the same topologies.
history \((p=0.38, \text{Figure 4.3})\). Next, the HEG was compared with the host genome and it was found that these do have significantly different histories \((p=0.004)\). Lastly, the group I intron and host were compared and again it was found that these also have significantly different evolutionary histories \((p<0.001)\). The results of these PHTs strongly indicate that the group I intron and HEG share the same evolutionary history and that this significantly differs from the evolutionary history of the host. Such phylogenetic incongruence is compelling evidence for \(\omega\) horizontal transfer.

The PHT test informs us that horizontal transfer has occurred, but it tells us nothing of the relationship between the donor and recipient species. Has \(\omega\) exchange occurred only between the Saccharomycetaceae, or also with taxa that are more distantly related? To address this question Genbank (release 111.0) was searched using BLASTn (v2.0.8) with the sequences of each of the five HEGs reported (after they had been deposited). In each case six Saccharomycetaceae \(\omega\) HEG sequences were retrieved: the five HEG used in this study plus an HEG found at the homologous site of a more distantly related Saccharomycetaceae, \textit{Pichia canadensis}. All other known HEGs (including many at other locations in the 20 yeasts here) are too divergent at the nucleotide level to be recovered by BLASTn (Dalgaard et al., 1997). Therefore, the HEGs found in the LSU rRNA of Saccharomycetaceae are monophyletic: they are more closely related to one another than to any other known HEG, despite their horizontally transmission. These two results allow the conclusion that \(\omega\) horizontal transfer amongst the LSU rRNA genes of Saccharomycetaceae yeasts occurs at a higher rate than transfers involving either more distantly related species, or other genomic sites within these species. The concept of more frequent horizontal transfer among closer
relatives is continued on a smaller phyletic scale within the Saccharomycetaceae: the HEG sequence of *Pichia canadensis* (which is an outgroup with respect to the yeast studied here) is the most divergent of all the Saccharomycetaceae sequences. This trend is also apparent in the group I intron data. BLASTn searches with the 14 Saccharomycetaceae intron sequences recovered group I introns from a number of genes and taxa, but none of these ever scored higher than the Saccharomycetaceae LSU introns, with the occasional exceptions of the intron in the *K. lactis* ATPase gene. The ATPase intron of *K. lactis* is flanked by sequences that in no way resemble the HEG’s recognition site, and so presumably is not able to home into this site. However, it is possible that this potential transposition event was facilitated by a change in HEG. An alternate hypothesis suggests non-endonuclease mediated movement. Reverse-splicing of a previously excised group I intron into an ATPase mRNA transcript may then be followed by reverse transcription back into DNA (perhaps by rogue viral reverse transcriptases). Finally, homologous recombination of this cDNA with its ‘in situ’ ATPase gene target could place the group I intron at the ATPase locus (Woodson and Cech, 1989). This mechanism in no way needs an endonuclease to facilitate movement. More data concerning the ATPase intron in the *Kluyveromyces* is needed to differentiate between these two hypotheses.

**Calculating the horizontal transfer rate**

The results of the above analyses show that horizontal transfer has occurred; they also suggest a tendency for increased transfer between closer relatives. However, these results say nothing about the rate of transfer: is this a rare process, or is it frequent enough to allow HEGs to recurrently invade genomes and thus escape degeneration? A rate of horizontal
transfer may now be estimated, because the time at which the *Saccharomyces* – *Kluyveromyces* split has been previously calculated. This split occurred roughly 70 MYA (Berbee and Taylor, 1993; calibrated with fossil data), thus allowing one to calculate the total amount of time represented by the host phylogeny. The total amount of time covered was calculated with the aid of the most parsimonious host phylogeny which had a molecular clock enforced (branch lengths calculated with maximum likelihood, see Figure 4.5). The time from the root to any one tip equals 70 Mys, therefore the total amount of time represented by the entire phylogeny is estimated to be 420 Mys.

Two methods were then used to estimate the number of ω transfer events which have occurred over the 420 million years.
Chapter Four: The long-term fate of HEGs

Method one: a lower bound calculation based upon parsimony

First, a parsimony method was employed to calculate the minimum number of horizontal transfer events that have occurred. A horizontal transfer event may be simulated by removing any one taxa or clade from a starting phylogeny and then reconnecting the removed portion, in the correct orientation (i.e. reconnecting a clade with its relative root), to a different area of the same phylogeny (Hein, 1993). The number of these transformations needed to convert one tree into another would indicate the minimum number of horizontal transfers needed to reconcile the two trees. This process may be simulated using the sub-tree pruning regrafting (SPR) branch swapping algorithm in PAUP*. The number of such SPR rearrangements that are needed to convert the 14 taxa host phylogeny (Figure 4.3 C) into the most likely intron phylogeny, the 'goal' tree (a better resolution of the bootstrap parsimony phylogeny seen in Figure 4.3 C), was calculated. A fully resolved 14 taxa phylogeny has 462 different possible one SPR rearrangements and the intron 'goal' tree is not one of these. To simulate a second horizontal transfer event all these 462 trees must be used as starting trees for the next round of SPR branch swapping. As the simulated number of horizontal transfers increases, the number of trees that must be evaluated dramatically expands (to simulate four events $462^4 = 4.6\times10^{10}$ trees must be evaluated). Therefore, at each round the trees were filtered by scoring them with a likelihood Kishino-Hasegawa test to see how well they fit the intron data (Kishino and Hasegawa, 1989). The tree that best fit the intron data and all the trees that did not differ from the best tree at the $p<0.5$ level were used as starting trees in the next round. After five such SPR rounds the most likely intron tree was found, although a tree that was not
significantly worse than the intron ‘goal’ tree (p>0.05 with a Kishino-Hasegawa test) was arrived at after just three (see Table 4.3, and Appendix II).

However, this best estimate of five horizontal transfer events is likely to be an underestimate for two reasons. Firstly, not all movement will be phylogenetically detectable. If, for example, the donor and recipient species are sister taxa then a transfer between these two will not alter the topology of the phylogeny (though such an event may in principle be detected by changes in branch length). Coalescent modelling suggests that for 14 taxa 2/3 of the horizontal transfer events will be missed (calculated by A. Burt, following Hudson and Kaplan, 1985). This now raises the best estimate to 15 horizontal transfer events, since these 15 have occurred over 420 million years this equals a minimum rate of 0.04 horizontal transfer events per million years. This correction is still likely to be an underestimate as the parsimony method assumes that transfer probability is independent of phylogeny. The earlier results suggest that transfer events appear to be more common between closer relatives, and that horizontal transfer is therefore not phyletically independent. It is possible that a large amount of transfer amongst closer relatives is occurring and this method would unable to detect such events.

<table>
<thead>
<tr>
<th>No. horizontal transfers</th>
<th>Ln likelihood of the most likely tree</th>
<th>Significantly worse than intron tree?§</th>
<th>No. trees†</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (Host tree)</td>
<td>-458.8</td>
<td>Yes (p &lt; 0.0001)</td>
<td>n/a</td>
</tr>
<tr>
<td>1</td>
<td>-429.3</td>
<td>Yes (p &lt; 0.0001)</td>
<td>4</td>
</tr>
</tbody>
</table>
Chapter Four: The long-term fate of HEGs

Table 4.3. Results from an algorithm to find the minimum number of rearrangements (horizontal transfers) necessary to change the host phylogeny (Fig. 4.3C) into one of the three most likely intron phylogenies. These are the 3 possible resolutions of the following consensus tree: (((((S.paradoxus, S. cariocus), S. mikatii), ((S.cerevisiae, T.delbrueckii), T.globosa)), K.thermotolerans, (S.exiguus, S.castellii)), T.pretoriensis), Z.bisporus), Z.rouxii), (K.dobzhanskii, K.lactis)). Likelihoods were calculated using the HKY model with the transition:transversion ratio and proportion of invariant sites estimated from the data. § Test for whether the most likely tree is significantly worse than the maximum likelihood intron tree by the Kishino-Hasegawa test, † Number of trees found, including the most likely tree and all others not significantly worse (at p=0.5). These were used as starting trees in the next iteration. n/a, not applicable. See Appendix II for further details of the protocol.

Table 4.3:

<table>
<thead>
<tr>
<th></th>
<th>Likelihood</th>
<th>Result</th>
<th>p Value</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-405.5</td>
<td>Yes</td>
<td>0.04</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>-397.6</td>
<td>No</td>
<td>0.22</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>-391.9</td>
<td>No</td>
<td>0.83</td>
<td>16</td>
</tr>
<tr>
<td>5 (Intron tree)</td>
<td>-390.8</td>
<td>No</td>
<td>1</td>
<td>170</td>
</tr>
</tbody>
</table>

Secondly, as this method calculates the minimum number of transfers that are needed to reconcile two phylogenies, multiple transfers of an element between two lineages will not be detected. This method will therefore further underestimate the true number of transfers in the same way as parsimony underestimates the actual number of character changes on a phylogeny. For example, two phylogenies cannot differ by more than n-2 rearrangements (n-3 for rooted trees), which means that no more than 12 horizontal transfer events may be inferred using this method. To explore the extent of detectable horizontal transfer events with this method, an infinite rate of horizontal transfer was simulated by randomising the terminal taxa on the
intron phylogeny. These randomised phylogenies were then treated as ‘goal’ trees and subjected to the parsimony SPR algorithm. Four such replicates gave estimates of 7(3), 6(4), 6(5), and 7(4) events (lower bound in brackets). This suggests that even with an infinite rate of \( \omega \) horizontal transfer the theoretical, artificial maximum of 12 transfer events is not realised, this being due to ‘multiple hits’, as mentioned above. It appears that the inferred five \( \omega \) transfer events (with a lower bound of three) are close to the point where this method reaches saturation, meaning that an upper bound cannot be put on the rate of \( \omega \) horizontal transfer.

**Method two: estimating the most likely frequency of state change in a cyclical model**

An alternative approach to estimating the frequency of \( \omega \) horizontal transfer is to make use of the phylogenetic information, the terminal taxa states, and population genetic reasoning outlined in the introduction to construct a cyclical evolutionary model. This model consists of the three states previously mentioned: ‘non-functional \( \omega \) element’ (non-functional either because there is no HEG, or because the HEG present is itself non-functional); ‘empty site’; and ‘functional \( \omega \) element’. Only three of the six possible character transitions are possible (Figure 4.4). An empty site may only be infected by a functional \( \omega \) element. It is possible that other rare transposition events may insert non-functional copies of \( \omega \); however, with no homing ability these degenerated elements are unable to drive through the population. Rare non-functional elements are therefore highly likely to be lost. Once the functional \( \omega \) element is fixed in the population no empty sites are encountered, and the positive selective pressure to
maintain HEG function is reduced. However, the selective pressure remains upon the group I intron since any loss of splicing ability will be fatal to the host cell since the rRNA LSU function is destroyed. Mutations will accumulate in the HEG, introducing premature stop codons; HEGs may be lost entirely from the group I intron, possibly increasing the splicing ability. The 'non-functional ω element' state is thus arrived at. The non-functional elements are unable to home and therefore unable to effect non-Mendelian inheritance, meaning that any loss of ω cannot be counterbalanced. Only precise intron losses which restore the ω recognition site are possible, since partial deletions will leave fragmented group I introns. The removal of non-Mendelian inheritance effects means that ω may not override negative selective pressures. Any individuals that have lost ω may be selectively favoured which will facilitate the transition to 'empty site'. Once a

![Diagram](image_url)

**Empty site**

Precise loss (0.34)

Horizonal transmission and fixation (0.63)

Intron with no, or non-functional, HEG

[11]

Degeneration (1.14)

Intron with functional HEG

[3]
Figure 4.4. Cyclical model of $\omega$ gain and loss. Numbers in square brackets indicate the number of taxa, out of the 20 surveyed, with that intron state. The numbers in parentheses are the maximum likelihood transition rates in events per million years, calculated by mapping the character states onto the most parsimonious tree with branch lengths estimated by maximum likelihood, and forcing a molecular clock (Figure 4.5).

population is predominated by 'empty site' individuals the potential for $\omega$ reinfection may be realised by the horizontal transfer of a functional $\omega$ element from another species. This new $\omega$ element will then quickly go to fixation and start the cycle turning once more. Taking into account the host phylogeny and using maximum likelihood methods, Dr. A Burt calculated the speed at which this cycle must turn in order to arrive at the terminal taxa states we observed in the laboratory (see Figure 4.5). The maximum likelihood transition rate calculations prove that a model which includes a horizontal transfer step is significantly more likely that one which does not ($\Delta \text{ln lik} = 13.2; \Delta \text{df}=1$). A model where each of the three transition rates are allowed to vary is no more likely than a model where the three rates are constrained to be the same. Under the constrained three step model a complete turn of the cycle takes 5.7 million years, with a lower bound of 0 and an upper bound of 42 million years. This converts to an $\omega$ horizontal transfer rate of 0.53 events per million years, with 2-unit support limits of 0.07 and infinity.
Figure 4.5. The most parsimonious host phylogeny, with branch length calculated by maximum likelihood (HKY + pI) where a molecular clock was enforced, which was used in the transition rate calculations. The three states of the cyclical model are shown mapped on to the terminal taxa (obtained from the PCR results). Conceptually, it was asked how often the three step model must turn, over the entire phylogeny, to arrive at the observed terminal states.
Conclusions

The survey results and subsequent phylogenetic analyses prove that $\omega$ has a highly dynamic history. These results support the population genetic speculations outlined in the introduction: within any one lineage $\omega$ appears to be going to fixation, and once there the lack of positive selection allows the HEG to degenerate and be lost. Only the horizontal transfer of $\omega$ to a $\omega^-$ species will restore positive selective pressures (due to spread), which will allow $\omega$ to maintain function through the recurrent invasion of lineages, possibly in a frequency dependent way. Within the donor species $\omega$ will eventually degenerate and then be lost and the horizontal transfer of a functional $\omega$ from another species will then start the cycle turning once more. This 'boom and then move before bust' cycle means that a minimum rate of movement is needed for the long term persistence of such elements (Hurst and McVean, 1996). If horizontal transfer is unlikely to occur before the degeneration of an element then loss is inevitable. If a minimum rate of transfer is needed this may go some way to explain why HEGs are almost exclusively found in microbes. With the exception of a sea anemone (Beagley et al., 1996), no HEGs have been found in animals, where a horizontal transfer event would have to place an HEG in a germ line cell. Further, the predominance of HEGs within microbial organelles may be explained by the need for a base line horizontal transfer rate. It can be suggested that the transfer of organelles among sexual species is more probable than the transfer of nuclear genomes when processes such as introgression (a horizontal transfer caused by a brief period of hybridisation with another species) are considered. That horizontal transfer appears more frequent among close relatives also prompts the suggestion that hybridisation may be a key mechanism for such movements, as
arguably isolated hybridisation occurs more frequently among close relatives. At least in yeast, the introduction of genes into a species via introgression may also take place because of unsuccessful hybridisation attempts. Yeast are known to go through a process known as ‘kissing’ where they begin to mate by fusing cytoplasmic constituents, but the process is aborted before any nuclear exchange, producing mixis of only the cytoplasm.

The methods employed here to estimate horizontal transfer rate suggest that $\omega$ is moving between species once every $10^6$-$10^7$ years, and perhaps much more frequently than this. As far as I am aware, this is the first attempt to quantify gene horizontal transfer rate and these methods should be widely applicable to other studies in gene horizontal transmission. These results show that the $\omega$ element is much more likely to move to the LSU rRNA gene of another species than to an ectopic location. HEG transposition events are presumably rarer than horizontal transfer since they require considerable mutation to achieve a parallel change in HEG recognition site. The relatively high horizontal transfer rate of this selfish gene is unexpected since HEGs have no extra-chromosomal step in their life cycle. It does, however, suggest that DNA can be moved between species and effectively maintain its integrity. The probability of observing a horizontal transfer event will, theoretically, depend upon: (i) the probability with which a segment of DNA is physically transferred between cells (species); and, (ii) the probability with which the transferred DNA will increase in frequency once within the recipient population. The fact that other genes do not appear to exhibit a high frequency of horizontal transfer (though this has by no means been tested rigorously) is presumably not because of lack of opportunity, but because of the lack of positive selective advantage (or a non-Mendelian inheritance) once in the recipient population.
Chapter Five

HEG CONCLUSIONS
Short term HEG dynamics

This thesis has taken detailed knowledge of HEG molecular biology and used it to construct and test various hypotheses concerning HEG ecology and evolution. The fact that HEGs are seen to effect non-Mendelian inheritance in single generation studies, by 'homing', allows one to predict that such genes will increase in frequency within a population, despite conferring no selective advantage to the individuals who carry these genes (Gimble and Thorner, 1992; Hickey, 1982; Jacquier and Dujon, 1985). One HEG, VDE, was tested for the ability to increase in frequency in populations of yeast. An increase in VDE frequency could theoretically be due to two reasons, either: (1) VDE confers a selective advantage to the host cell and increases in frequency due to natural selection at the cellular level; or, (2) VDE confers no selective advantage to the host cell and the increase in frequency is purely a consequence of VDE homing events. The change in VDE frequency was assayed in both inbred and outcrossed populations to ascertain which of these two hypotheses is the more likely. Within the inbred populations VDE decreased significantly in frequency, and the relative starting and finishing proportions allowed a VDE selection coefficient of 0.12 to be estimated. The VDE frequency increased significantly within the outcrossed populations. These observations allow one to conclude that VDE is only able to increase in frequency by constantly homing into vde⁻ alleles. The infection of vde⁻ alleles is only possible via homing events, and these homing events are only made possible when sex unites VDE+ and vde⁻ alleles in heterozygous genotypes. The 12% parasitic load attributed to VDE may not reflect the true selection coefficient of VDE due to possible genetic engineering artefacts (Futcher et al., 1988). However, this artificially high selection coefficient only serves to further underline
that fact that HEGs may increase in frequency despite being of considerable detriment to the
individuals which carry them. Therefore, HEGs do, as theory predicts, increase in frequency
within an outcrossing population, despite being of no selective advantage to individuals.
HEGs are therefore justly considered selfish genes.

Selection level?

Such observations are applicable to questions concerning the level at which natural selection
operates. G. C. Williams outlines a doctrine for his work on Adaptation and Natural Selection
(1966, p 4-5):

“When recognised adaptation [due to natural selection] should be attributed to no higher a level of organization than is demanded by the evidence”.

HEGs demonstrably increase in frequency within populations because of an advantage at the
genic level. An HEG is in direct competition with the HEG~ allele within a host’s genome.
HEGs do not contribute to the phenotypic fitness of the host organism and so will not increase
in frequency due to the host organism’s preferential survival. HEGs increase in frequency purely as a consequence of their ability to home. This ability allows HEG+ alleles to out-compete HEG~ alleles since HEG+ alleles are able to over-represent themselves in the next
generation. Selection is clearly acting at the level of the gene where HEGs are concerned.
Conceptual HEG population dynamics

The ability of HEGs to infect populations leads to the formulation of hypotheses that propose an increasingly dynamic life history when longer time scales are considered. Knowledge of HEG spread, along with the evidence presented for HEG horizontal transfer, and laboratory observations, permit the construction of the cyclical model shown in Figure 4.4. This cycle proposes that within a host species HEGs of differing 'strains' will show a fluctuating frequency as they invade, go to fixation, degenerate, and then are removed from the species. Gene horizontal transfer allows individual 'strains' of HEG to recurrently invade differing species and therefore maintain selective pressure for function. If the transfer of a particular HEG 'strain' is not achieved before non-functional mutations are accrued, then that 'strain' of HEG has a high probability of consequent extinction (only back mutations will restore function). Gene horizontal transfer allows HEGs to persist in the long term. Without horizontal transfer HEGs would be unable to escape from their hosts, where they will degenerate and ultimately be lost. The persistence of HEGs appears to depend upon the frequency with which an HEG is likely to be transferred to another species. HEGs are unable to effect their own horizontal transfer (unlike viruses) and so presumably must rely upon random events to supply them with the opportunity for HEG allele invasion. A particular HEG may be horizontally transferred to an uninfected species where it will begin to increase in frequency. The speed with which the HEG will increase in frequency will depend upon the selection coefficient associated with it and, more importantly, the degree to which the recipient host population is outcrossing. Low levels of outcrossing will mean long lag times between HEG invasion and fixation, when compared to largely outcrossed populations (assuming that the
parasitic load attributed to the HEG is minimal). These ideas concerning HEG fluctuation within a host species are illustrated in Figure 5.1. Maximum likelihood transition rates were calculated using the $\omega$ data [under a model where the three transitions were allowed to vary (Figure 4.4), note: this model was not significantly more likely than a model where all the transitions were constrained to be the same]. The average time taken to go from any one of the one steps to another may be calculated using the inverse of this rate. This is defined as the 'average waiting time'. These times are illustrated in Figure 5.1:A where the average time spent waiting is proportional to the distance between the states. These waiting times are combined with the predicted frequency of the $\omega$ HEG within a population, at differing points within the cycle, to produce an $\omega$ HEG frequency fluctuation over time plot (Figure 5.1:B). Starting with a HEG$^{+}$ population it takes on average 1.59 million years for an HEG to be
Figure 5.1 Conceptual frequency dynamics of HEGs within a host population. A: the cyclical model maximum likelihood waiting times to move from one state to the next (inverse of the values seen in Figure 4.4); the distance between states is proportional to the amount of time. B: the predicted fluctuation of an HEG within a host population. At any given point in time the frequency of a particular HEG in the population is proportional to the area represented under the line. Division of the 1.59 MY into time spent waiting for a horizontal transfer event and the time taken for the HEG to go to fixation is unknown. Theoretically this division will differ: it will take much longer for HEGs to increase in frequency in inbred populations compared to more outcrossed ones. It is unlikely that average time waiting for an HEG horizontal transfer events is longer than 1.59 MY since this would be reflected in the transition rate calculation; it could, however, be shorter. C: The ‘rain’ of HEG horizontal transfer events into the population that are predicted to occur on average every 1.59 million years, resulting in the possibility of observing two strains in one population. F = failed HEG invasion since all sites are infected by another strain, see text for details.
for HEG to go to fixation' are unknown. With only 4% of a population outcrossing (upper estimate for natural S. paradoxus outcrossing rate), it is predicted to take roughly $2 \times 10^6$ generations for an HEG to go from a frequency of $0.1 \times 10^{-6}$ to 0.98 (using Equation 3.13 with no selection against the HEG and where $d = 0.95$). The frequency with which yeast sporulate (go through meiosis) in the natural environment is unknown and these sporulation events will determine the effective number of generations natural yeast go through over a given period of time. If, for arguments sake natural yeast go through an average of one sporulation event (generation) every year (probably a conservative estimate) then it would take approximately 2 million years for an HEG with a starting frequency of one in a million to go to fixation in a 96% inbreeding population. If that population were fully outcrossing the waiting time is only 25 years (five orders of magnitude less time). In a predominately outcrossing population rare HEGs will quickly go to fixation showing a steep climb on the frequency fluctuation over time plot. In a predominantly inbreeding population HEGs will take considerably longer to go to fixation and show a more shallow climb to fixation.

Once at fixation it takes on average 0.88 million years before the ‘non-functional co element’ state is reached. This is an estimate at the population level, not the individual level. Once an HEG is at fixation a plateau is reached (see Figure 5.1), and at the beginning of this plateau most copies of the HEG will be functional. Over time random non-functional mutations will occur in increasingly larger numbers of HEG copies within the population. Initially the relatively rare number of individuals with precise HEG deletion events, due to reverse transcription and then reverse splicing (Woodson and Cech, 1989), will be re-infected by functional copies from other members of the population. However, as time increases more HEGs mutate and eventually the population reaches a point where HEG loss cannot be counterbalanced by homing since the HEG population is dominated by non-functional copies. At this point the frequency of the HEG will start to decline (see Figure 5.1:B).
According to the data available for the $\omega$ element it takes on average 2.94 million years until the population is again dominated by 'empty site' individuals. The decrease in frequency of the $\omega$ element is presumably due to a build up of individuals with precise $\omega$ losses coupled with positive selection for those individuals which have lost the $\omega$ element.

The horizontal transfer of $\omega$ is calculated to occur on average once every 1.59 million years. If horizontal transfer events were occurring significantly less frequently than this, a longer 'time waiting for HEG horizontal transfer and fixation' would be estimated from the data. Horizontal transfer may be occurring more frequently, but, as mentioned earlier, it is unknown how the 'time waiting for horizontal transfer' and 'time waiting for HEG fixation' are divided between 1.59 million years; it is not known which of these two is the limiting factor.

According to the data $\omega$ will be horizontally transferred into a species about once every 1.59 million years. If this is true then, starting at the beginning of the cycle, another two HEG transfer events into the species are predicted to occur before the 5.41 million year period needed to complete the cycle (Figure 5.1:C). The first of these events is predicted to occur when the first $\omega$ 'strain' (strain1) is at fixation. The transfer of a second $\omega$ 'strain' (strain2) will never come to fruition since there are no empty sites for 'strain2' to home into. Another 1.59 million years later another $\omega$ horizontal transfer event is predicted to occur. $\omega$ 'strain3' finds a recipient population which contains the $\omega$ 'strain1' in the decline phase of its life cycle. This means that a proportion of the host population have lost $\omega$ and are in the 'empty site' state.
The newly invaded ω 'strain3' is able to infect these empty sites and sweep through the 'empty site' sub-population. This model allows the prediction that one host species may contain multiple 'strains' of ω element due to recurrent invasion, and that these 'strains' will all be at differing stages in their life cycles (Figure 5.1:C). The ω survey conducted in Chapter Four helped to construct the ω cyclical evolutionary model and here only one individual from each species was surveyed. The validity of the predictions concerning the presence of 'multiple strains' of HEG within a species may be addressed by surveying many individuals from a species. The survey results will detail the frequency and relatedness of any HEGs the species may contain. Indeed, laboratory observations show that differing isolates of S. cerevisiae and S. cariocus harbour differing ω states (results not shown). All these processes may occur at the population level if a species is significantly fragmented into sub-populations so as to effect a reproductive barrier.

The highly dynamic life style of HEGs is unlikely to be reflected by other genes. Only genes which are able to sweep through populations may show such population dynamics. Only two types of gene are predicted to sweep through a population when rare – beneficial genes and genes that are able to effect non-Mendelian inheritance.

The mode of action and population dynamic nature of HEGs may lead one to consider them distinct organisms. Viruses are little more than encapsulated DNA or RNA and are considered separate organisms; the simplest viruses are only a little more complicated than HEGs, and in many ways the life histories of the two are similar. HEGs are certainly seen to 'parasitise' a gene within a genome. The life history strategy of HEGs allows them to move between different hosts and then spread. HEGs predominantly target essential genes (e.g. ribosomal
and metabolic pathway genes) and these genes are relatively conserved across taxa allowing
HEGs to potentially invade a wide range of differing genomes.

The life history of HEGs may not be reflected by other ‘selfish genes’. The meiotic drive
systems touched on in the introduction appear to be capable of spreading in populations, and
the loci that constitute the driving system will increase, despite selective forces against them.
The molecular biology of meiotic drive systems is less clear. Whether or not meiotic drive
genes are able to make the most of horizontal transfer events remains to be seen – meiotic
drive systems that work in one species may not work in another. B - chromosomes are
predicted to increase in frequency in outcrossed populations, but it is less likely that such
large segments of DNA are able to be horizontally transferred between species (apart from
via introgression events).

Such selfish genes are powerful tools for molecular evolutionary investigations. Genes such
as HEGs underscore the fact that genes may be viewed as separate units; and the rate at
which these units change, move around and interact with other units can vary dramatically. A
genome appears to be a dynamic system of such units. The vast majority of these units
contribute toward the construction of a larger entity. Genomes that constitute individuals are,
in sexual species at least, transient collections of genes which were individually successful in
past generations. Their success in the future generations depends upon the results of the
complex interactions with all of the other genes in the current genome as well as phenotypic
selection pressures. Most of these genes will contribute toward the construction and
Chapter Five: Conclusions

maintenance of the organism, however, as HEGs patently demonstrate, a certain degree of intra-genomic gene warfare may occur.

This work has addressed questions concerning the ecology and evolution of HEGs. The population dynamics of HEGs on the short, medium and long term have been investigated. The results presented, and the conclusions drawn, constitute a thesis on the evolutionary processes determining the ecology and evolution of selfish genes.
Chapter Six

LARGER SCALES OF YEAST EVOLUTION - ESTIMATION OF HEMIASCOMYCETE PHYLOGENY
Chapter Six: Yeast evolution

Introduction

Past and present studies of yeasts has been extensive, and has produced a vast amount of molecular and genetic data (Dujon, 1996). The majority of workers have studied the baker's yeast *Saccharomyces cerevisiae*, which has resulted in a detailed knowledge of its biochemical pathways and genetic architecture. However, an increasing number of gene sequences have been reported from other species of yeast (24% of Hemiascomycete sequences are non-*S. cerevisiae*, Genbank release 113.0). The potential use of yeast as a powerful tool for ecological and evolutionary based studies is slowly being realised (e.g. Birdsell and Wills, 1996; Naumov et al., 1997). The specific and detailed knowledge of *S. cerevisiae* permits comparisons with close relatives, which then allow one to better understand the evolution of certain traits. Yeast taxonomic classification has traditionally been based upon morphological and biochemical characters (e.g. the appearance of spore forming structures and the ability to utilise certain chemical substrates, Kurtzman and Fell, 1998). However, such classifications can be spurious and say little of the relationships between groups, since the genetic basis of these phenotypic characters can be unclear (James et al., 1997). More recently, molecular data have been used to ascertain relationships among yeasts (e.g. Kurtzman and Robnett, 1991; Vaughan-Martini and Kurtzman, 1985; Wilmotte et al., 1993). DNA sequence data is presumed to be more reliable for use in phylogenetic reconstruction since DNA sequences offer more objective characters for analysis. Even so, it must be highlighted that the evolutionary history of certain genes found within an organism need not necessarily agree with the evolutionary history of the host organism (Bechhofer et al., 1994; Bhattacharya et al., 1996; Hibbett, 1996).
Chapter Six: Yeast evolution

The construction of accurate yeast molecular phylogenies is of paramount importance if they are to be used in evolutionary and ecological studies. This accuracy must be reinforced with a robustness if the extent to which transposable and selfish genetic element evolutionary histories deviate from that of the host is to be measured. Inferences of selfish gene horizontal transfer are only valid if they are made using well supported phylogenies. Yeast molecular phylogenies may also be employed to look at questions concerning gene trees, species trees and gene horizontal transfer, since, as mentioned earlier, the evolutionary history of one particular gene may not reflect that of the species. Additionally, yeast classification may be better defined using the information from such molecular phylogenies, since traditional morphological classifications can be vague (Kurtzman and Fell, 1998).

There have been previous attempts to reconstruct yeast molecular phylogenies, the majority of which have used ribosomal (rRNA) coding genes, or the spacers found between these rRNA genes (James et al., 1994; James et al., 1996). These areas are assumed to accurately reflect the species phylogeny. Previous studies have mostly utilised partial areas of genes, and only a very few studies have analysed whole gene sequences (Ando et al., 1996; Oda et al., 1997). Even rarer is the use of two molecules to reconstruct a yeast phylogeny (Cai et al., 1996). No previous analysis has focused exclusively on Hemiascomycetes (Class). Analyses with fewer taxa predominate, though there are examples of larger yeast phylogenies which include many Hemiascomycetes as well as the more distantly related Euascomycetes (Class) and Basidiomycetes (Phylum) (Cai et al., 1996; Wilmotte et al., 1993). However, these larger analyses exclude many Hemiascomycete taxa of interest. Further, previous studies have
predominately used distance methods (mostly neighbour joining) to reconstruct phylogenies. These methods can produce less reliable estimates than either maximum parsimony (MP) or maximum likelihood (ML) methods. Distance methods summarise a set of sequences in the form of a pairwise distance matrix, and this matrix loses information since the phylogenetic signal contained within individual characters is collapsed to an overall sequence difference. Additionally, the branch lengths estimated by some distance methods may not be evolutionary interpretable (Page and Holmes, 1998).

Therefore to reconstruct a more robust yeast phylogeny two ribosomal genes have been utilised (complete 18S, partial 26S), and both MP and ML methods are employed to analyse these data.
Materials and methods

Sequence data

Two molecules were used to reconstruct a Hemiascomycete phylogeny: the nuclear small ribosomal subunit (18S) and the nuclear large ribosomal subunit (26S). The 18S sequences were obtained from Genbank and the 26S sequences were kindly supplied by C. Kurtzman, and are described in (Kurtzman and Robnett, 1998). 18S and 26S sequence data are available for 59 taxa - all these taxa were included for analysis. The 18S and 26S genes are found adjacent to one another in ribosomal gene families that are repeated many times in the genome. The 18S sequence data originally derive from two source papers (Cai et al., 1996; James et al., 1997) with the exception of two sequences (Candida dubliniensis and Issatchenkia orientalis; these were present in Genbank, accession numbers X99399 and M55528, but not found in these papers). The original authors obtained the 18S data by first amplifying three fragments of the 18S molecule and then by directly sequencing these three overlapping fragments. A further two non-Hemiascomycete taxa, Schizosaccharomyces pombe and Schizosaccharomyces japonicus (these are yeast that grow mitotically by fission as opposed to the budding method employed by the majority of Hemiascomycetes), were included in the analysis for use as an outgroup, for a total of 61 taxa.
Sequence alignments

An alignment of the 18S data based upon the ribosomal secondary structure is available from http://rrna.uia.ac.be/ssu/ (Van de Peer et al., 1998). Unfortunately only the secondary structure alignments of ten taxa are available for the 26S data (De Rijk et al., 1998). As well as utilising the secondary structure alignment an alignment was made using the ClustalW software for the Apple Macintosh, with all the parameters kept at default settings (Thompson et al., 1994). No manual adjustments of the alignments were performed.

Phylogenetic analyses

Analyses of the 18S and 26S data were performed using PAUP* (alpha test versions 4d63-d65 and the beta test version 4.0b2a) (Swofford, 1999) and MacClade (v. 3.07) (Maddison and Maddison, 1992) with an Apple Macintosh Performa containing a 180 MHz processor. Genetic distances between taxa were calculated with PAUP* and visualised with Microsoft Excel 98.
Results and discussion

Initial phylogenetic reconstruction

The ClustalW alignment results were first visually inspected. The taxa aligned extremely well across the vast majority of the gene producing a pattern of large, well aligned blocks with no postulated insertion-deletion events (indels). However, the areas between these blocks showed ambiguous alignment which were characterised by a relatively large number of inferred indels. The more 'jumbled' appearance of these areas result in uncertain nucleotide homology. Characters with uncertain homology defy the assumptions of tree reconstruction methods, and therefore must be removed from the data matrix.

The resulting 18S and 26S data were then analysed using a maximum parsimony optimality criterion. Ideally all possible trees for the 61 taxa would be assessed to find those which postulate the least amount of change (shortest trees). However, for 61 taxa there are $2.417 \times 10^{178}$ binary rooted trees (Quicke, 1993). To assess such a vast amount of trees on a relatively underpowered computer would be impracticable, and so a heuristic search algorithm was used. If the lengths of all possible trees are conceptualised as a landscape, with the shortest trees forming the peaks of mountains, then heuristic algorithms are able to climb these mountains by rearranging the branches of a starting tree. As the branches are 'swapped' rearrangements may be found which are shorter. The branches of these shorter trees are then swapped in an attempt to find even shorter trees, and this process is iterated until no decrease in tree length is seen – the top of the mountain has been reached. However,
there may be many mountains in the landscape and the top of a sub-optimal peak may have been found. To minimise this possibility many different starting trees, scattered randomly over the tree landscape, are used; these starting trees are obtained by process known as random addition (Swofford et al., 1996). Such strategies attempt to find the shortest tree. Shortest trees postulate the least amount of evolutionary change but give no indication as to the degree of confidence (support) of individual branches. Some branches may not truly represent the phylogenetic history of the Hemiascomycetes and may simply be sampling error artefacts. The support for individual branches in the tree may be gauged using a statistical process known as bootstrapping. The bootstrap process re-samples the original data set to produce a pseudoreplicate, and then the shortest tree is found for this pseudoreplicate data set. If many pseudoreplicates of the data are made (typically 100) the degree of support for a particular branch equals the percentage of pseudoreplicate trees in which that branch is found (Page and Holmes, 1998). Bootstrap analysis indicates the degree of robustness in a tree. Such a heuristic bootstrap analysis was therefore performed on the 18S+26S data set to produce the tree seen in Figure 6.1. This resulting tree shows strong support for some deep splits within these taxa and there are a considerable number of branches with bootstrap scores above 95%. The bifurcating relationships between certain taxa are very well represented, for
Chapter Six: Yeast evolution

Figure 6.1. A phylogeny of Hemiascomycetes reconstructed from 18S and 26S data with an unweighted maximum parsimony bootstrap analysis. 100 bootstrap replicates were performed with a heuristic search algorithm: NNI branch swapping, 200 random additions per bootstrap, and no more than 100 trees retained per addition. Ca. = Candida, L. = Longisporus, Deb. = Debaryomyces, Dek. = Dekkera, I. = Issatchenkiia, Pic. = Pichia, K. = Kluyveromyces, Z. = Zygosaccharomyces, S. = Saccharomyces, T. = Torulaspora, Me. =
example the *Candida*, *Debaryomyces* and *Dekkera* clade seen toward the top of Figure 6.1. Also, some of the relationships toward the tips of the tree show strong bootstrap support; for example, some of the *Kluyveromyces* and *Zygosaccharomyces* are independently grouped together with bootstrap scores of over 90%. However, the majority of taxa have poorly resolved relationships in the mid-tree area. This is exemplified by the clade of yeast grouped by a very well supported branch (100% bootstrap score) but within which a 15-way polytomy is seen. However, it is interesting to note that resolution does increase toward the tips of this clade. This strongly supported but poorly resolved clade represents no previously recognised taxonomic group: all the yeast found in this clade belong to the *Saccharomycetaceae*, but *Saccharomycetaceae* yeast are also found outside of this group (e.g. *Dekkera* and *Debaryomyces*). This is somewhat perplexing since there is clearly enough phylogenetic signal in the data to resolve deep, tip and some mid-range branches: why is there not enough phylogenetic signal to resolve the mid-range branches of this clade?

**A polytomy problem**

There are a number of hypotheses that may be responsible for the poor resolution of this clade. The more likely hypotheses are:

1. recombination between certain taxa's 18S or 26S molecules;
2. inefficient alignment methods;
3. poor phylogenetic signal due to conflict between the 18S and 26S data;
4. poor phylogenetic signal *per se*;
5. that this clade represents a 'hard polytomy' – a true multiple speciation event; or,

6. that the rate of ribosomal evolution varies across the phylogeny.

One or more of these hypotheses may explain the poor resolution seen in the Hemiascomycete phylogeny. A variety of tests were performed in order to distinguish which of these factors are the more important.

**Tests for recombination**

Recombination between molecules can lead to problems with reliable tree reconstruction (Hein, 1993). The evolutionary history contained within a recombined data set is best reflected by a network. A tree reconstructed from recombined data will not truly represent the conflicts between differing areas of the recombined sequences. This will result in poor phylogenetic signal. The 18S data were obtained via three separate fragments (see Figure 6.2). To check the three fragments of the 18S data for recombination and/or data deposition errors, the uncorrected 'p' pairwise distances were calculated for all possible pairs of taxa for all three fragments. If there has been no recombination then the three fragments of a particular taxon's 18S gene should be equally related (have proportional distance) with respect to
Figure 6.2. A: the relative positions of the three separately sequenced fragments of the 18S gene; these were amplified with the primers P108, M1190, P2130, M2130, M3490, and M3989 (Cai et al., 1996). B-C: the pairwise distance comparisons of the three 18S fragments for all taxa. Regression analyses performed, and results plotted, with Microsoft Excel.
all other taxa. A recombination event (or a data deposition error) would produce unequal relatedness across all three fragments; for example fragments A and B of taxon y may be very closely related to fragments A' and B' of taxon x, but the third fragment, fragment C, may be distantly related if it has recombined with a third taxa. Such fragment relatedness deviations were checked by plotting all the pairwise distances for each fragment against the distances for the other fragments. The vast majority of taxa will not have recombined and the distances between fragments will be roughly equal producing a tight correlation. A recombination event will show up as a cluster of points falling away from the average regression line since the distances of the two fragments will differ. The plots of the three fragment comparisons (Frags. A v B, A v C and B v C) are shown in Figure 6.2. Visual inspection suggests that there has been no recombination (or data deposition errors) since there are no significant outliers from the regression line. Unfortunately, no such fragment data is available for the 26S sequence preventing a similar analysis. These plots suggest that the poor resolution seen in the Hemiascomycete phylogeny is unlikely to be due to conflicts in signal that derive from recombination between 18S molecules.

**Inefficient alignment methods?**

The method and parameters used to align sequences can radically affect the phylogenetic estimates which are recovered. There are suggestions that the ability to recover the correct tree is more influenced by alignment method than tree reconstruction method (Morrison and Ellis, 1997). There is evidence to suggest that alignment procedures which incorporate information on the secondary structure of molecules may produce more accurate results
(Goddard and Burt, ; Knudsen and Hein, 1999). This is especially applicable to ribosomal genes since the secondary and tertiary structures they form on translation into RNA critically determines their function (Griffiths et al., 1993). To test the hypothesis that an alignment based upon secondary structure would be more informative, the 18S alignments available from http://rrna.uia.ac.be/ssu/ were utilised. A visual inspection of this alignment showed it to contain many indels (postulated insertion deletion events). This is presumably because the alignment was made with a great number of more distantly related taxa, and with a model designed for more general Eukaryote ribosomal secondary structure (Van de Peer et al., 1998). Figure 6.3 shows a tree reconstructed from the data set which has been aligned taking into account secondary structure, and using a heuristic search algorithm under maximum parsimony optimality criterion. There are a number of branches with reasonable bootstrap scores, but the overall resolution of the tree is no better than the tree reconstructed using the ClustalW alignment. Further, the resolution of the large polytomy is worse when using the secondary structure alignment: a nineteen way polytomy is recovered as opposed to a fifteen way one. There is also a database of 26S sequences aligned according to the secondary structure they form, but unfortunately this database only contains ten Hemiascomycete sequences. It therefore appears that the poor resolution seen in the Hemiascomycete
Figure 6.3. A phylogeny of Hemiascomycetes estimated from 18S ribosomal DNA data using an alignment based upon the secondary structure of the ribosome (Van de Peer 1998). A maximum parsimony bootstrap analyses (100 random additions per bootstrap replicate, 100 bootstrap replicates, with NNI branch swapping) produced the tree shown. The numbers on branches are bootstrap scores. Some taxon name are abbreviated for ease of display, the full name may be found in Figure 6.1.
Yeast evolution phylogeny cannot be improved using an 18S alignment based upon secondary structure.

18S – 26S conflict

It may be that the two data partitions (18S and 26S) contain conflicting phylogenetic signals. If the characters within one data set are contradicted by characters from the other, then a tree reconstructed from a chimera of these data sets will reflect this conflict in the form of homoplasy (similarity between taxa that is independently derived and not inherited), poor resolution, and low bootstrap support. To address this question bootstrap trees were independently reconstructed from each data set using identical parsimony search conditions (Figure 6.4). Only two branches from both the trees above 61% contradict one another, and this results from a difference in the placement of only two taxa. Overall these two trees have very similar topologies. Further, the clade that contains the large polytomy is not resolved by either the 18S or the 26S data alone. The separate analyses of the two data partitions indicate that the poor resolution seen in certain areas of the Hemiascomycete phylogeny is unlikely to be due to conflicting signals between the 18S and 26S data.
Figure 6.4. Phylogenies independently reconstructed from the 18S and 26S data with maximum parsimony optimality criteria. The numbers on the branches signify bootstrap scores. The bootstrap analyses were performed with a heuristic search algorithm: 100 bootstrap replicates, NNI branch swapping, 30 random additions per bootstrap with no more than 100 trees saved per addition. Some taxon name have been abbreviated for ease of display, the entire name may be found in Figure 6.1.
Tests for inherently poor phylogenetic signal

Another possible reason for poor resolution may be due to a large amount of homoplasy within these data sets. Homoplastic characters will reduce the strength of phylogenetic signal since they indicate that certain taxa should be grouped together even though they are not related by descent. The uncoordinated signal produced by homoplastic characters may swamp the signal produced by more reliable characters and thus create uncertainties in the relationships of the taxa: this translates into polytomies. In order to test whether or not this is the case for the Hemiascomycete data, a method must be employed which can identify and down-weight homoplastic characters.

The problem with Hemiascomycete resolution is only restricted to the one clade and it was decided to separately analyse this clade since independent analysis may recover more signal. The taxa that constitute the poorly resolved clade first had their 18S and 26S data realigned, with all other sequences left out, except for the Candida albicans, Pichia angusta and Pichia anomala sequences which were included for use as an outgroup (see Figure 6.1). Again, the alignment was visually inspected and areas of uncertain nucleotide homology were removed before analysis. This realignment increased the number of parsimony informative sites by 54 (from 148 to 202) since an alignment containing more closely related taxa had fewer areas of questionable nucleotide homology. In order to identify and down-weight homoplastic characters in the data a modified method of successive approximation described by Farris (1969) was employed. Farris points out that all characters are not equally well correlated with the true phylogenetic history because of convergence, parallelism and reversals: all sources of homoplasy. He suggests that the efficiency of tree building methods may be increased by
weighting characters according to their degree of correlation with each other. Farris’s method essentially involves building an initial tree and then using this to re-weight the data: characters which disagree with the topology will be down weighted as presumably these are sources of homoplasy and are therefore interfering with the phylogenetic signal. The weighted data is then used to reconstruct a second tree. Farris suggests this process be iterated until no change in topology is observed since presumably this is the point at which homoplastic characters are maximally down weighted. When the successive approximation method was first described Farris employed the consistency indices (CI – a measure of how well the data fit the tree, Farris, 1969) to weight the characters; however, he later developed the rescaled consistency (RC) index and suggested that this is of far more use when successively weighting (Farris, 1989). With the RC characters that have as much homoplasy as possible can achieve a weight of zero: this is not true for the CI. The major drawback of the successive weighting method derives from the fact that it is heavily reliant upon a previously estimated tree. Farris originally recommended that the shortest tree be used to re-weight the data. One is always uncertain as to the confidence of branches within the shortest tree, and the tree may therefore contain ‘incorrect’ branches. Re-weighting using a tree containing incorrect branches will lead to the incorrect weighting of characters and therefore the subsequent recovery of increasingly incorrect phylogenetic estimates. This method has been modified to decrease the probability of re-weighting with partially incorrect trees by only utilising 95% bootstrap consensus trees. I am confident that the branches present in a 95% bootstrap tree are a fair reflection of the phylogenetic signal within the data and am therefore equally confident that re-weighting using these trees will only amplify the signal of ‘good’ characters (that is characters that reflect the evolutionary history of these two ribosomal genes and show
little homoplasy). Therefore a 95% bootstrap consensus tree of the 18S + 26S data was produced, and this phylogeny was subsequently used to re-weight the data.

The programme with which the data are to be re-weighted must be chosen carefully since PAUP* and MacClade can treat polytomies differently. PAUP* treats all polytomies as multiple speciation events (hard polytomies), and therefore all characters are assumed to have been independently gained along each branch of a polytomy (autapomorphies). Within a polytomy, even if many taxa have the same state for a given character each will be treated as an independent gain – that character will appear more homoplastic than it really may be. PAUP* is unable to distinguish between the resolutions of a polytomy and therefore re-weighting using it will never allow a polytomy resolution. MacClade, on the other hand, will allow for uncertainties in the resolution of a polytomy (soft polytomies). With this interpretation a character’s evolution is traced and re-weighted assuming the polytomy is resolved in the most favourable way for that character (Maddison and Maddison, 1992). Many taxa in a polytomy may have the same character – MacClade will assume that all these taxa share a common ancestor and infer only one change of that character thus minimising homoplasy. Therefore, re-weighting data using trees that contain polytomies in MacClade (with the soft polytomy option) will more accurately reflect the characters changes over the topology (given the assumptions of parsimony), and this may therefore allow subsequent polytomy resolution.

The 95% bootstrap tree obtained in PAUP* was imported into MacClade and the characters re-weighted according to the RC index. These re-weighted data were then used to reconstruct another 95% bootstrap tree and the procedure was repeated until no change in 95%
bootstrap tree topology was seen, i.e. homoplastic characters had been maximally down-weighted. After only one round of weighting no further change in topology was observed. There is a high degree of confidence that all branches found in the successively approximated 95% bootstrap tree reflect the phylogenetic signal in the 18S+26S data and that the homoplastic characters had all been maximally down-weighted; however, the resolution was still poor.

A bootstrap tree reconstructed from the weighted data is only slightly better resolved than a tree reconstructed from unweighted data (two extra internal branches above 95% were recovered). It therefore appears that the down-weighting of homoplastic characters has not significantly helped with the confident resolution of this clade.

**Possible multiple speciation event**

It is possible that this poorly resolved clade represents a true multiple speciation event (hard polytomy) where the taxa are not related by a bifurcating tree. A multiple speciation event implies the independent evolution of many taxa from one common ancestor. These descendent taxa would be no more closely related to one another than they would to the common ancestor. Such an event may mean there would be many most parsimonious fully resolved trees since no one topology would be any better than another, and a strict consensus of these many most parsimonious trees will collapse into a polytomy. However, if there is a bifurcating relationship described in the data then only a very few most parsimonious trees may be recovered. Further, these few trees should all converge upon similar topologies and a strict consensus of these should show reasonable resolution.
A heuristic search shows that the poorly resolved clade has only four most parsimonious trees (a very small number considering 37 taxa have $4.5 \times 10^{92}$ possible fully resolved rooted trees), and a strict consensus of these is extremely well resolved (Figure 6.5). The very small number of most parsimonious trees suggests that the phylogenetic signal contained within these data agrees upon very similar topologies. That this signal is not reflected in a bootstrap analysis indicates that this signal is positive but weak, and that the polytomy is not 'hard' but 'soft'.
Figure 6.5. A strict consensus of the four most parsimonious trees of the independently analysed, poorly resolved clade seen in Figure 6.1. The trees were found under heuristic search conditions: SPR branch swapping, 200 random additions with no more than 200 trees retained at each addition. The tree is rooted as inferred form the larger Hemisacomycete analysis. The bootstrap scores seen in Figure 6.1 have been placed upon the corresponding branches in this diagram to indicate the degree of support for various branches. Some taxon names have been abbreviated for ease of display, the full names may be found in Figure 6.1.
Clock test – variation in the rate of ribosomal evolution?

The above results suggest that these taxa are related by a bifurcating tree. It is possible that the rate of ribosomal evolution varies across the phylogeny. If one were using a molecule that mutates relatively slowly to reconstruct a phylogeny, then areas of the phylogeny for which there are many taxa will appear to have accrued fewer sequence changes. The phylogenetic relationship between these taxa will therefore be weakly represented in the data: it will have a weak phylogenetic signal. Very few informative sites will not be adequately represented in an unweighted bootstrap analysis but may be detected in the most parsimonious trees.

To get an impression of the rates of evolution over the entire phylogeny the branch lengths of the tree, where the best estimate for the independently analysed clade has been grafted onto the main phylogeny in the appropriate place, were calculated using maximum likelihood (ML). If all clades have evolved at the same rate, i.e. if the poorly resolved independently analysed clade’s 18S+26S sequences evolved at the same rate as the rest of the tree then the branch lengths should not significantly differ and the tree will appear to be ultrametric (clock-like - where all the tips will be roughly equidistant from the root). The phylogeny was therefore tested to see if is significantly different from an ultrametric phylogeny. The likelihood of the phylogeny, given the F81 model with the proportion of invariant sites estimated with ML, was calculated with and without an enforced molecular clock. Other more general models of sequence evolution could have been employed to compute the likelihood of the phylogeny at a computational cost, but for purposes of speed the F81 model (where base composition is allowed to vary and all nucleotide substitution rates are constrained to be the same) was employed. However, given the nature of the data set the proportion of invariant sites were
taken into account in the model and were estimated from the data (a model which includes a parameter that allows the proportion of invariant sites to be estimated is more likely that one than does not). A phylogeny is not significantly different from clock-like if the likelihood under a model which requires 2n-3 branching times (all branches are free to vary in length = non-clock like) is not significantly more likely than a model which requires n-1 branching times (sister taxa have the same branch length = clock like). The difference in Ln likelihoods proved to be -276.832, and since twice the difference in Ln likelihood has a $\chi^2$ distribution with 2n-3-(n-1) = n-2 degrees of freedom, it may be compared to the appropriate critical value. The degrees of freedom are the difference in number of branches that are allowed to vary under the two models. A likelihood ratio test therefore proves that this tree deviates significantly from a clock since twice the difference in log likelihoods even exceeds the $\chi^2$ 0.001 critical value of 98.32 ($2\Delta -\text{Ln} = 553.664$, n-2 = 59 d.f.) (Huelsenbeck and Rannala, 1997). The resulting phylogram is shown in Figure 6.6 (here the branch lengths were calculated using the same ML model as was used to test for the clock). It can be seen that the branch lengths of the clade of interest are much shorter than the rest of the phylogeny – indicating that relatively less change has occurred along these branches.
Figure 6.6. A phylogram of the best estimate of the Hemiascomycete phylogeny. The branch lengths were calculated with maximum likelihood under the F81 model with the proportional of invariant sites estimates with maximum likelihood. The phylogeny significantly deviates from being clock like as proved by the likelihood ratio test - see text for more details.
Conclusions

Maximum parsimony bootstrap analysis of 18S and 26S data for a group of 59 hemiascomycete yeasts recovered a tree with an sporadic pattern of resolution. Certain areas of the phylogeny were very well resolved with good bootstrap support. However, one clade, containing 37 yeasts, showed extensive areas of poor support which was expressed as a polytomy within the bootstrap consensus tree. The pattern of resolution within this clade moved from poor resolution at its base to reasonably well resolved toward the tips.

Several tests were performed in an attempt to discover the source of this poor support. It is unlikely that recombination, poor alignment or 18S - 26S data conflict are responsible for the poor support of the Hemiascomycete phylogeny. However, intrinsically poor phylogenetic signal cannot be discounted as the source of this poor resolution. The possibility of a multiple speciation event proved to be unlikely since the data describe only four most parsimonious trees which combine to produce a highly resolved strict consensus. Such analysis suggest that there is positive but weak signal in the data indicating that these yeast have not undergone a multiple speciation event.

A likely explanation of the intrinsically poor phylogenetic signal suggests that this clade may have experienced a slowing in the number of changes of the 18S and 26S ribosomal genes. It appears that selection constrained the rate of evolution of the ribosomal molecules within this clade. It is possible that the secondary and tertiary structures formed by these sequences were stable and/or efficient – selection may have discouraged deviations from these well adapted sequences thus slowing the rate of evolution. The relatively slow evolution of this
clade combined with the fact that there are many taxa results in little phylogenetic information and thus creates polytomies. Using 18S sequence divergence calibrated with fossil evidence *Saccharomyces cerevisiae* – *Kluyveromyces lactis* are estimated to have diverged roughly 70 MYA (Berbee and Taylor, 1993). The phylogram, seen in Figure 6.6, suggests that this clade experienced a slowing in ribosomal evolution shortly after this split: this is evident from the poor bootstrap support and short branch lengths seen. The evidence suggests that this is a ‘soft’ polytomy since there are only four most parsimonious reconstructions of these poorly supported areas. A hard polytomy (multiple speciation event) would leave no discernible phylogenetic footprint and more than four most parsimonious trees would be predicted from such a data set. The poor bootstrap support, but few number of short trees, suggest that the few informative sites for these areas agree, thus giving a weak but positive reflection of yeast evolutionary history. The rate of ribosomal evolution in this clade appears to increase as one nears contemporary taxa as reflected by increasing branch lengths and bootstrap support nearer the tree tips.

It is possible that the slowing in evolution in this clade is due to a change in ecological niche, that the secondary and tertiary ribosomal structures were ideally adapted to a new niche which was then rapidly exploited. However, no obvious shift in ecological niche is apparent in this clade (Kurtzman and Fell, 1998). The reconstruction of a Hemiascomycete yeast phylogeny based upon biochemical resource utilisation may help to better answer this question. Finally, the acquisition of further sequence data for Hemiascomycetes will allow an increasing number of genes to be utilised in phylogenetic estimation. Different genes clearly evolve at differing rates: for example the internal transcribed spacer (ITS) sequences between
ribosomal genes are only useful for resolving very closely related organisms (Goddard and Burt, 1999; Oda et al., 1997). As long as the appropriate genes are used to resolve the appropriate phylogenetic level the efficiency of phylogenetic estimation should increase. Providing that assumptions of recombination and gene horizontal transfer are checked, an increase in sequence data will further serve to elucidate yeast evolutionary history.
Bibliography


translocating adenosine triphosphate from vacuolar membranes of Saccharomyces cerevisiae. J. Biol. Chem. 265, 6726-6733.


Appendices
## Appendix I

Details concerning the origin of the natural *Saccharomyces paradoxus* isolates.

<table>
<thead>
<tr>
<th>Thesis name</th>
<th>Sample name</th>
<th>Tree information</th>
<th>Isolate details</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>A-Silwood</td>
<td>1L West+East</td>
<td>West May '98 = T4b</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1S West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2S West+East</td>
<td>West May '98 = T8.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3S West+East</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>B-Silwood</td>
<td>1L West+East</td>
<td>?Oct '96 = W7?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3L West+East</td>
<td>West July '98 = T32.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1S West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2S West+East</td>
<td>East July '98 = T26.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3S West+East</td>
<td>West Dec '97 = S36.7; East July '98 = T27.3</td>
</tr>
<tr>
<td>S3</td>
<td>C-Silwood</td>
<td>1L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1S West+East</td>
<td>West July '98 = T76.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2S West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3S West+East</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>D-Silwood</td>
<td>1L West+East</td>
<td>West May '98 = T18a/T18.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1S West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2S West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3S West+East</td>
<td>East May '98 = T21.4; West May '98 = T22.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thesis name</th>
<th>Sample name</th>
<th>Tree information</th>
<th>Isolate details</th>
</tr>
</thead>
<tbody>
<tr>
<td>S5</td>
<td>E-Silwood</td>
<td>1L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3L West+East</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1S West+East</td>
<td></td>
</tr>
</tbody>
</table>
Appendices

<table>
<thead>
<tr>
<th>Thesis name</th>
<th>Sample name</th>
<th>Tree information</th>
<th>Isolate details</th>
</tr>
</thead>
<tbody>
<tr>
<td>S6 F-Silwood</td>
<td>2S West+East</td>
<td>West July '98 = T62.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3S West+East</td>
<td>West July '98 = T68.2</td>
<td></td>
</tr>
<tr>
<td>W1 A-Windsor Great Pk.</td>
<td>1L North+South</td>
<td>South Sept '98 = Q6.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2L North+South</td>
<td>South Sept '98 = Q14.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3L North+South</td>
<td>North Sept '98 = Q15.1; South Sept '98 = Q16.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9L North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1S North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2S North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3S North+South</td>
<td>South Sept '98 = Q4.1</td>
<td></td>
</tr>
<tr>
<td>W2 B-Windsor Great Pk.</td>
<td>1 North and South</td>
<td>North Sept '98 = Q27.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9 North and South</td>
<td>North Sept '98 = Q31.4; South Sept '98 = Q32.3</td>
<td></td>
</tr>
<tr>
<td>W3 C-Windsor Great Pk.</td>
<td>1 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 North and South</td>
<td>North Nov '98 = Q59.1</td>
<td></td>
</tr>
</tbody>
</table>
7 North and South
8 North and South
9 North and South
10 North and South
11 North and South
12 North and South

South Nov '98 = Q62.5

W4 D-Windsor Great Pk.
1 North and South
2 North and South
3 North and South
4 North and South
5 North and South
6 North and South
7 North and South
8 North and South
9 North and South
10 North and South
11 North and South
12 North and South

North Nov '98 = Q89.8

Total trees = 86
Total samples = 344
Appendix II

The protocol for finding the minimum number of horizontal transfers given two incongruent phylogenies. This test should be used only after a more robust test, such as the Partition Homogeneity test, has indicated that the two data sets have significantly different evolutionary histories. "Text" means supply your own file name; text signifies a PAUP* menu or dialogue box title; 'host' = host/organism/genome data; 'parasite' = parasite/selfish gene/gene suspected of horizontal transfer data; 'goal tree' = the best/most likely parasite tree.

1. Construct 'Host' replicate data set in MacClade, making sure all the taxa names are the same as the data file that contains the host data. Have only one character for each taxa and make all these are the same character. 2. Go to the tree window and construct the 'host' tree topology. Save the file ("redundant host.nex").

3. Open "redundant host.nex" in PAUP*.

4. Set to parsimony optimality criteria; in parsimony options uncheck collapse zero branches box.

5. Select Heuristic search... from Analysis menu.

6. In General search options check the keep N Best trees box (N = number of 1 SPR rearrangements = $4n^2 - 26n + 42$, for $n$ taxa. This should be checked by constructing a perfect data set for a binary topology with $n$ taxa and the subjecting that to a SPR search)

7. In Starting tree options check use tree(s) in memory.
8. In **Branch-Swapping options** check **SPR** and uncheck **save all optimal trees**.

9. Save all the trees found from this heuristic search. This will be the $N$ **SPR** rearrangements + the original starting tree(s).

10. Open the "parasite" data file in PAUP* (and clear any trees).

11. Import the trees from step 3.

12. Check for 'goal tree' by filtering all trees with 'goal tree'. If not found proceed.

13. Under the **File** menu **Log output to disk**. Under **Tree: Tree score: Likelihood** menu select the appropriate Likelihood model. Then check the **Kishino-Hasegawa** box, select all trees and perform the K-H test.

14. Manually scroll through the saved log and identify the best tree and all trees not worse than the best at a chosen value (0.5 for omega data) with the P* column. Make a note of all the tree numbers.

15. Save the best trees to a file. These are the trees which will be used as starting trees for the next round of **SPR** swapping.

16. Open the "redundant host.nex" in PAUP* and clear the tree from memory. Go to **Trees:Get trees from file** menu and select the trees saved in step 9. Continue from Step 5. Iterate until the 'goal tree' is found.

17. At each round keep a note of the best trees likelihood score and the number of best trees. At each round score the best tree to see if it significantly differs from the 'goal tree' with a likelihood K-H test.