Developing genetic and biological systems for the study of Epstein Barr virus strain diversity

Mohammed M. Ba abdullah

A thesis submitted for the degree of Doctor of Philosophy

Section of Virology

Department of Medicine

Imperial College London

March 2018

ABSTRACT

Epstein-Barr virus is a ubiquitous gamma herpesvirus that can cause several types of lymphoma and carcinoma. Different viruses have different properties such as lytic reactivation, cell tropism, and oncogenesis, but the genetic determinants of these varied phenotypes are not fully known. There is a need for a better understanding of what constitutes wild-type EBV, as a background against which to identify mutations and/or polymorphisms associated with diseases. Most recent EBV sequencing projects through next-generation sequencing have ignored analysing large repeat regions as their assembly is technically challenging with short-read sequencing technologies. In this thesis, a strategy was developed to assemble and analyse in-depth repeat regions using short-read sequencing data. I have assembled and analysed internal repeat 1 (IR1) for 76 strains and the terminal repeat (TR) for 31 strains. The IR1 sequences showed that the latency promoter Wp, the EBNA-LP W1 and W2 exons, and the short intron between them are highly conserved, as are two regions of unknown function within BWRF1 open reading frame. The length of BWRF1 is not the same in all strains. I show that TR is highly variable across strains but includes conserved regions of known functions (e.g. packaging signals sites) and unknown function.

Analysis showed that IR1 and TR are generally heterogeneous and this heterogeneity seems to arise from spontaneous mutation as well as from interstrain recombination. IR1 heterogeneity in tumour-derived strains suggested a larger degree of inter-strain recombination, raising the possibility that this may contribute to EBV pathology. We also identified a novel stop codon in one copy of the W1 exon of the commonly used laboratory strain of EBV, B95.8 and showed that this mutation reduced EBNA-LP protein production and the quality of transformation. Our strategy has highlighted the importance of studying the heterogeneity of repeat regions and shown some approaches that can improve the quality of the final sequence.

To study the biological properties of EBV, I tried to develop a method to produce circulating EBV viruses by inducing lytic cycle in spontaneous LCLs (sLCLs) and also by cloning EBVs as BACs. sLCLs were very resistant to lytic induction stimuli. I succeeded in cloning the EBV strain P3HR1 as BAC, but was unable to repeat this consistently, highlighting the difficulty of generating EBV-BACs in latent B cells. The constructed P3HR1-BAC would potentially be useful for studying EBV type 2 biology.

ACKNOWLEDGMENTS

I would like to take this opportunity to express my heartfelt gratitude towards my supervisor Dr Rob White for continual patience, unfailing support, and great ideas throughout my PhD I am deeply thankful for every lesson and advice you gave me both scientifically and personally. Also, I would like to thank my second supervisor Prof Paul Farrell for valuable suggestions during the lab meetings.

To former members of White group; Dr Agnieszka Szymula and Dr Richard Palermo, I miss our conversations and jokes in the lab, and I wish you had been here while I was writing this thesis. To Amr Bayoumy, thank you for endless discussions about cloning in St. Mary's hospital cafe, and I hope you are more fortunate than me with cloning stuff. Many thanks to all members of the Allday group, Farrell group, and Maertens group.

To my beautiful wife, Eman, words cannot express my feeling. I would not have been able to finish this without you. Thank you so much for your sacrifices and your patience with my bad mood when things did not go well.

Dear son Jassim, as a father, the first advice I want to give you is "if you are willing to get a PhD, do not EVER think of having a baby before you are finished.", but I have to thank you for your refreshing and lovely smiles every morning.

Last but not least, many thanks to my family and friends for support and encouragement.

PUBLICATIONS

- **Ba abdullah, Mohammed**, Richard Palermo, Anne Palser, Nicholas E Grayson, Paul Kellam, Samantha Correia, Agnieszka Szymula, and Rob White. 2017. "Heterogeneity of the Epstein-Barr Virus Major Internal Repeat Reveals Evolutionary Mechanisms of EBV and a Functional Defect in the Prototype EBV Strain B95-8." Journal of Virology 91 (23): JVI.00920-17. doi:10.1128/JVI.00920-17.

- Palser, Anne L, Nicholas E Grayson, Robert E White, Craig Corton, Samantha Correia, **Mohammed Ba abdullah**, Simon J Watson, et al. 2015. "Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection." Journal of Virology 89 (10): 5222–37. doi:10.1128/JVI.03614-14.

- Szymula, Agnieszka, Richard D Palermo, Amr Bayoumy, Ian J Groves, **Mohammed Ba Abdullah**, Beth Holder, and Robert E White. 2018. "Epstein-Barr Virus Nuclear Antigen EBNA-LP Is Essential for Transforming Naive B Cells, and Facilitates Recruitment of Transcription Factors to the Viral Genome." PLoS Pathogens 14 (2). United States: e1006890. doi:10.1371/journal.ppat.1006890.

DECLARATION OF ORIGINALITY

I hereby declare that all work presented in this thesis is the result of my own work (unless otherwise stated). This work was carried out at the Section of Virology, Faculty of Medicine, Imperial College London under the supervision of Dr. Rob White and co-supervision of Prof. Paul Farrell.

Mohammed M. Ba Abdullah

March 2018

COPYRIGHT DECLARATION

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

ABBREVIATIONS

aa	amino acid
AID	Activation-Induced cytidine Deaminase
BAC	Bacterial Artificial Chromosome
BART	BamHI A Rightward Transcripts
BCR	B Cell Receptor
BL	Burkitt Lymphoma
BWA	Burrows-Wheeler Aligner
ChIP	Chromatin Immunoprecipitation
Cm	Chloramphenicol
CMV	Cytomegalovirus
CNV	copy number of variants
Cp	C Promoter
DLBCL	Diffuse Large B Cell Lymphoma
DMSO	Dimethyl Sulfoxide
DS	Dyad Symmetry
EBER	Epstein–Barr virus-Encoded small RNA
EBNA	Epstein-Barr Nuclear Antigens
EBV	Epstein-Barr Virus
EDTA	Ethylenediamine Tetra-Acetic Acid
EtBr	ethidium bromide
FBS	Foetal Bovine Serum
FSC	Forward Scatter
Fwd	forward
GaC	Gastric Carcinoma
GC	Germinal Centre
GFP	Green Fluorescent Protein
HDAC	Histone Deacetylase
HEK	Human Embryonic Kidney
HL	Hodgkin Lymphoma
HRP	Horseradish Peroxidase
HRS	Hodgkin Reed Sternberg

HSV	Herpes Simplex Virus
HT	4-Hydroxytamoxifen
IFN	Interferon
Ig	Immunoglobulin
IL	Interleukin
IM	Infectious Mononucleosis
IR	Internal Repeat
JNK	C-Jun N-Terminal Kinase
KSHV	Kaposi's Sarcoma-associated Herpesvirus
LB	Luria Broth
LCL	Lymphoblastoid Cell Lines
LCV	Lymphocryptovirus
LMP	Latent Membrane Protein
LP	Leader Protein
miRNA	Micro RNA
MS	Multiple Sclerosis
MSA	multiple sequence alignment
MV	minor variant
NaB	Sodium butyrate
NEB	New England Biolabs
NF-κB	Nuclear Factor Kappa-light-chain-enhancer of activated B cells
NPC	Nasopharyngeal Carcinoma
ORF	Open Reading Frames
OriLyt	Origin of Lytic replication
OriP	Origin of Plasmid replication
PAGE	Poly-Acrylamide Gel Electrophoresis
PBL	Peripheral-Blood Lymphocyte
PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
PI	Propidium Iodide
PTLD	Post-Transplant Lymphoproliferative Disease
Qp	Q Promoter
RBPJ	Recombination Binding Protein-J

RDA	Read depth analysis
REV	Revertant
RIPA	Radioimmunoprecipitation Assay
rpm	round per minuts
SB	Super Broth
SDS	Sodium Dodecyl Sulfate
sisRNA	Stable Intronic Sequences RNA
sLCL	Spontaneous LCL
SNP	Single Nucleotide Polymorphisms
SS	Sjögren's Syndrome
SSC	Side Scatter
STET	Sucrose Triton EDTA Tris
SV40	Simian Virus 40
TBE	Tris-Borate-EDTA
TE	Tris-EDTA
TEMED	NNN'N'-Tetramethylethylenediamine
TGF-β	Transforming Growth Factor-Beta
TPA	12-O-Tetradecanoylphorbol-13-Acetate
TR	Terminal Repeat
WB	Western Blotting
Wp	W Promoter
WT	Wild Type

TABLE OF CONTENTS

ABSTRACT **2**

ACKNOWLEDGMENTS **3**

PUBLICATIONS **4**

DECLARATION OF ORIGINALITY **5**

COPYRIGHT DECLARATION **5**

ABBREVIATIONS **6**

TABLE OF CONTENTS **9**

TABLE OF FIGURES **16**

TABLE OF TABLES **19**

1 Introduction **20**

1.1 Herpesviruses 20

1.1.1 Alphaherpesvirinae 21

1.1.2 Betaherpesvirinae 21

1.1.3 Gammaherpesvirinae 21

1.2 Herpesvirus genomes 23

1.3 Epstein Barr Virus 24

1.3.1 Discovery of EBV 24

1.3.2 Taxonomy 26

1.3.3 EBV genome structure 26

1.3.4 EBV latent infection in vivo 28

1.3.5 EBV latent infection in vitro. 31

1.3.6 EBV tropism to B cells and epithelial cells. 32

1.3.7 EBV latency promoters 32

1.3.7.1 Wp promoter 33

1.3.7.2 Cp promoter 34

1.3.7.3 Qp promoter 35

1.3.7.4 LMP1p promoter 35

1.3.7.5 LMP2Ap promoter 36

1.3.8 EBV latency proteins 36
1.3.8.1 EBNA1 37
1.3.8.2 EBNA2 38
1.3.8.3 EBNA-LP 39
1.3.8.4 EBNA3 family 41
1.3.8.5 LMP1 42
1.3.8.6 LMP2A and LMP2B 42
1.3.8.7 EBERs 43
1.3.8.8 BamHI-A Rightward Transcripts (BARTs) 44
1.3.8.9 EBV microRNAs 44
1.3.9 The diseases of EBV 45
1.3.9.1 Infectious mononucleosis 46
1.3.9.2 Hodgkin lymphoma 47
1.3.9.3 Burkitt lymphoma 47
1.3.9.4 Post-transplant lymphoproliferative disease (PTLD) 48
1.3.9.5 Nasopharyngeal carcinoma 49
1.3.9.6 Gastric cancer (GaC) 49
1.3.9.7 Autoimmune diseases. 50
1.4 What is a wild-type EBV? 50
1.5 EBV diversities that change EBV properties. 51
1.6 Genetic analysis of EBV function 54
1.6.1 Lytic cycle study in vitro 54
1.6.2 Production of recombinant viruses in eukaryotic cells 56
1.6.3 Manipulating EBV genomes in bacteria 57
1.6.4 Using CRISPR-cas9 to support homologous recombination 60
1.6.5 Methods of DNA genome sequencing. 61
1.6.5.1 First generation (Sanger) sequencing 62
1.6.5.2 Next generation sequencing 63
1.6.5.3 Third Generation (long read) Sequencing technologies 66
1.7 Repeat regions in EBV 68
1.7.1 The family of repeats (FR) 69
1.7.2 Internal repeat 1 (IR1) 70
1.7.3 Terminal repeats (TR) 74
1.8 Aims of the thesis 77

2 Material and methods 78
2.1 Sequence data analysis 78
2.1.1 Sequence data source 78
2.1.2 Constructing IR1 theoretical templates 79
2.1.3 Constructing TR unit template 80
2.1.4 Analysis of IR1 and TR sequences 80
2.1.4.1 Single nucleotide polymorphism (SNP) density 80
2.1.4.2 Phylogenetic tree 81
2.1.4.3 BWRF1 ORF prediction 81
2.1.4.4 Variant calling 81
2.1.4.5 Statistics 82
2.1.4.6 Detecting potential recombination events 82
2.2 Solutions and buffers 83
2.3 Antibiotics 85
2.4 Chemical inducers of EBV lytic cycle 85
2.5 Cloning and DNA analysis 85
2.5.1 Restriction enzyme digestion and ligation 85
2.5.2 Transformation 87
2.5.3 Plasmid miniprep 87
2.5.4 Plasmid maxiprep 88
2.5.5 CsCl maxiprep 88
2.5.6 Agarose gel electrophoresis 89
2.5.7 Gel extraction 90
2.5.8 Pulsed-field gel electrophoresis 90
2.6 Cell culture 90
2.7 Lentivirus production and sLCLs infection 91
2.7.1 Lentivirus production 91
2.7.2 Infecting sLCLs with lentivirus 91
2.8 DNA extraction 92
2.9 Polymerase chain reaction PCR 92
2.10 Real-Time PCR 94
2.11 Protein analysis 95
2.11.1 Protein extraction 95
2.11.2 SDS-polyacrylamide gel electrophoresis (SDS-PAGE) 95

2.11.3 Transferring proteins from SDS-PAGE gel to a membrane 96
2.11.4 Western blotting 96
2.12 Cell transfection 97
2.12.1 Transfection of 293T cells 97
2.12.2 Electroporation of LCLs 97
2.12.3 GFP assay 98
2.12.4 Low molecular weight DNA extraction 99
2.12.5 Episome rescue 99
2.13 Generating LCLs 100
2.13.1 Virus titration 100
2.13.2 Infection of peripheral blood leukocytes (PBLs) 100
2.14 sLCL proliferation assay 100
2.14.1 CellTrace Violet staining 100
2.14.2 CellTrace intensity-dilution-based proliferation assay 101
2.14.3 Flow cytometry procedure 101

3 Internal repeat 1 (IR1) analysis 102

3.1 Introduction 102
3.2 Constructing a theoretical template to simplify repeat sequence analysis 104
3.3 Read depth analysis (RDA) verifies the integrity of constructed templates 106
3.4 Read depth analysis identifies the junctions between unique sequences and IR1 110
3.5 IR1 has conserved regions including Wp, W2 and regions of ORF BWRF1 116
3.6 The length of the BWRF1 ORF is not the same in all EBV strains 117
3.7 The hairpin is retained in within IR1 122
3.8 EBNA-LP exons are more variant than were previously reported and have restricted combinations 123
3.9 Developing a methodology to distinguish heterogeneity within repeats from technical errors in sequencing 127
3.9.1 The quality of minor variants can be defined by their frequency and read bias 129
3.9.2 Minor variants at error-prone sites can be detected by the coefficient of variation…. 130
3.9.3 SNPs towards the edges of IR1 can be identified by read bias and frequency 131
3.9.4 Defining a strategy for calling MVs. 132
3.9.5 Origin-based classification of minor variants 133

3.10 Variations in terminal IR1 units are protected from gene conversion 135
3.11 IR1 units are typically heterogeneous, and diversity is more common in tumour and tumour-derived cell lines 137
3.12 Strains with high number of MVs show evidence of interstrain recombination ... 138
3.13 There is no correlation between number of MVs and IR1 copy number..... 141
3.14 Diversity in IR1 is not caused by a mixture of strains in the sample..... 142
3.15 Heterogeneity in IR1 arises regionally and may undergo a process of purification to homogeneity..... 143
3.16 BL36 is probably a lab-associated intertypic recombinant strains 148
3.17 The prototype strain B95-8 contains clustered of low frequency minor variants that include a stop codon mutation in EBNA-LP 148
3.18 The stop codon in a single W1 exon in B95-8 interrupts EBNA-LP isoforms expression and slightly attenuates virus transformation ability..... 154
3.19 Summary..... 157
4 Terminal repeat analysis 158
4.1 Introduction..... 158
4.2 Diversity of the TR hindered making a theoretical template to simplify the analysis…..... 158
4.3 The TR region in EBV genomes is highly diverse 160
4.4 Regions of functional importance in TR are highly conserved 166
4.5 The read bias score threshold was deactivated in defining minor variant within TRs…….. 167
4.6 The B95-8-cell line isolate from Hong Kong does not have the short TR unit 169
4.7 Spontaneous mutations are common in the TR region, but recombination-derived diversity is more common in tumour and tumour-derived cell lines 171
4.8 Summary..... 174
5 Producing circulating EBV viruses 175
5.1 Attempts to obtain viruses from cell lines by inducing the lytic cycle. 175
5.1.1 BZLF1 induction..... 176
5.1.1.1 Infecting cells with a CMV promoter-based lentiviral vector expressing conditional BZLF1 176
5.1.1.1.1 BZLF1 induction did not enhance viral proliferation but induced endogenous BZLF1 expression..... 178

5.1.1.2 Infecting cells with a CAG promoter-based lentiviral vector expressing conditional BZLF1 179
5.1.1.2.1 Constructing CAG promoter-based lentivectors 180
5.1.1.2.2 CAG promoter-derived BZLF1 induction has no effect on lytic cycle activation 182
5.1.2 Inducing the lytic cycle by treating cells with chemical inducers 183
5.2 Cloning EBV genomes as BACs 185
5.2.1 Identifying potential insertion sites in the TR to be targeted by Cas9 187
5.2.2 Targeting the EBV TR region using a donor BAC plasmid with short homology sequences and conditional Cas9 188
5.2.2.1 Constructing CRISPR components 188
5.2.2.2 Constructing donor BACs 190
5.2.2.3 Targeting EBV using a donor BAC plasmid with short homology sequences of the TR region and conditional CRISPR/Cas9 components 193
5.2.3 Targeting a unique region to check the feasibility of cloning selected EBV strains 195
5.2.4 Targeting the EBV TR region using a donor BAC plasmid with long homology sequences and an unconditional Cas9/sgRNA plasmid 200
5.2.4.1 Constructing donor BACs 200
5.2.4.2 Construction and validation of CRISPR components 203
5.2.4.3 EBV genome targeting and screening of bacterial colonies 206
5.2.5 Modifications performed to improve cloning strategy 210
5.2.6 Troubleshooting to understand reasons for obtained results showed in previous attempts to clone EBV as BACs 213
5.2.6.1 High copy number plasmid does not affect cloning process 213
5.2.6.2 The BAC cassette may get into the EBV genome and then be excised 215
5.3 Summary 218

6 Discussion 219

6.1 Assembling and analysing repeat regions is challenging 219
6.2 IR1 and TR have conserved regions of known and unknown functions across EBV strains 221
6.2.1 EBNA-LP exons 222
6.2.2 Stable non-coding RNAs and the BWRF1 ORF 224

6.2.3 Conserved regions in the terminal repeats 226
6.3 Heterogeneity in IR1 and TR regions 229
6.4 Polymorphisms in the prototype strain B95-8 232
6.4.1 IR1 diversity 232
6.4.2 Effect of the stop codon mutation on EBNA-LP production and viral transformation ability 233
6.4.3 The short TR is missing in at least one clone of the B95-8 cell line 234
6.5 Interstrain recombination in IR1 235
6.6 EBV lytic cycle induction 236
6.7 Cloning EBV as BAC 239
6.8 Conclusion 244
7 References 246
Appendix 1. EBNA-LP and BWRF1 subtypes of diverse viruses 267
Appendix 2. DNA sequence with annotations of the B95-8 IR1 theoretical template 273
Appendix 3. BWRF1 models 285
Appendix 4. Secondary structure of the long hairpin within the IR1 of B95-8 288
Appendix 5. SNPs in the flanks of IR1 289
Appendix 6. Distribution of MVs in viruses with the most interstrain recombination 293
Appendix 7. Common SNPs in B95-8 and BL36 295
Appendix 8. DNA sequence with annotations of the B95-8 TR theoretical template 296

TABLE OF FIGURES

Figure 1.1 Schematic of herpesvirus genome structures. 24
Figure 1.2 Morphology of herpesvirus virions. 25
Figure 1.3 Schematic of EBV genome. 28
Figure 1.4 Proposed model of how EBV establishes and maintains latency infection....30
Figure 1.5 Latency promoters activity during the EBV latency programmes. 33
Figure 1.6 Positions of latency genes and promoters. 37
Figure 1.7 Exon organization of EBNA-LP gene transcripts. 40
Figure 1.8 Schematic of the alternative splicing of EBNA3 mRNAs. 41
Figure 1.9 Schematic illustration of the IR1 position and components. 71
Figure 1.10 Alternative splicing that organizes EBNA-LP and EBNA-2 gene transcripts. 72
Figure 1.11 Terminal repeat structure. 75
Figure 3.1 mapping multi-reads. 103
Figure 3.2 Structure of the theoretical template. 105
Figure 3.3 Testing the ability of RDA in predicting deletions in sequences. 108
Figure 3.4 Read depth analysis predicted low read depth within BamHI Y in Nmalwa. 110
Figure 3.5 Contigs assembly of P3HR1. 111
Figure 3.6 Read depth analysis to identify defective junctions. 112
Figure 3.7 Visualisation of a deleted junction within IR1 by scanning mapped reads. 113
Figure 3.8 Schematic representation of deleted regions in strains with defective junction upstream and downstream IR1 region. 115
Figure 3.9 . Polymorphism density in IR1 region across 85 strains. 117
Figure 3.10 BWRF1 models. 119
Figure 3.11 Phylogenetic tree classified by IR1 diversity. 120
Figure 3.12 Diversity within BWRF1 at nucleotides and amino acids level. 122
Figure 3.13 Disruption in the hairpin within the IR1. 123
Figure 3.14 Variants in the amino acid sequence of the EBNA-LP protein. 124
Figure 3.15 mapping reads to the theoretical template. 129
Figure 3.16 Identifying flank-SNPs. 132
Figure 3.17 Visualizing MVs and Defining a threshold for calling genuine minor variants. 134
Figure 3.18 Overall distribution of MVs. 138
Figure 3.19 IR1 diversity in spontaneous LCLs and tumour-derived samples. 139

Figure 3.20 Minor variants accumulation in spontaneous LCLs and tumour-derived samples. 140
Figure 3.21 There is no correlation between IR1 size and number of minor variants. 142
Figure 3.22 T-RECs analysis of genetic exchange in IR1 regions of sLCL-IS2.01 and TRL1. 145
Figure 3.23 Bootstrap analysis of genetic exchange in IR1 regions of sLCL-IS2.01 and TRL1. 147
Figure 3.24 Minor variants in IR1 of the prototype lab EBV strain B95-8. 150
Figure 3.25 Gel electrophoresis of PCR-amplified EBNA-LP transcript (W1W2 exon) obtained from HB9. 151
Figure 3.26 PCR experiment to identify IR1 copy with stop mutation. 153
Figure 3.27 Western blot analysis of EBNA-LP extracted from LCLs 30 days post infection. 155
Figure 3.28 Effect of the stop codon mutation on the ability of EBV in transforming B cells. 156
Figure 4.1 variability of TR length. 161
Figure 4. 2 TRs regions are highly variable. 162
Figure 4.3 Visualizing MVs in TRs region. 168
Figure 4.4 B95-8 terminal repeat alignment. 169
Figure 4.5 B95-8 cell line is missing the known short TR unit. 170
Figure 4.6 Minor variants accumulation in TR region in spontaneous LCLs and tumour-derived strains. 172
Figure 4.7 TR region diversity in spontaneous LCLs and tumour-derived samples. 173
Figure 5.1 Structure of lentivirus vectors. 177
Figure 5.2 Evaluating the lytic cycle induction by activating pCMV-derived BZLF1. 178
Figure 5.3 Western blot analysis of BZLF1 expression. 179
Figure 5.4 CAG promoter-based lentivectors construction strategy. 181
Figure 5.5 Evaluating the lytic cycle induction by inducing pCAG-derived BZLF1. 183
Figure 5.6 Evaluating the lytic cycle induction using chemical inducers. 184
Figure 5.7 Positions of Cas9 targets in TR. 188
Figure 5. 8 Pulsed-field analysis of the functionality of CRISPR components in targeting terminal repeats in B95-8 BAC. 190
Figure 5.9 Donor BAC (pBACR6K-AG876-sg#) construction strategy. 192
Figure 5.10 Electroporation optimization. 193

Figure 5.11 Pulsed-field analysis of miniprep DNA obtained from transfected AG876 cells. 195
Figure 5.12 Structure of donor plasmid pBS_SNU719_BVRF1 and Cas9-sgRNA plasmid pX330.....196
Figure 5.13 Pulsed-field analysis of miniprep DNA obtained from cloning P3HR1 as BAC. 197
Figure 5.14 Pulsed filed analysis of P3HR1-BAC.....199
Figure 5.15 Donor plasmid BAC-AG876-TR# construction strategy. 202
Figure 5.16 Schematic of constructing pX330 and GFP reporter plasmid pCAG-EG-TR-FP. 204
Figure 5.17 Verifying the functionality of Cas9 plasmid. 205
Figure 5.18 Pulsed-field analysis of miniprep DNA obtained from transfected Jijoye and AG876.....206
Figure 5.19 Structure of pBAC-4908-TR1 and 3. 207
Figure 5.20 Pulsed-field analysis of recovered BACs that lacked EBV genome. 208
Figure 5. 21 Pulsed-field analysis of miniprep DNA obtained from transfected P3HR1 cells. 209
Figure 5.22 Structure of donor BAC with I-PpoI site.....211
Figure 5.23 The expected structure of EBV-BAC after homologous recombination.....212
Figure 5.24 Structure of donor BAC with I-PpoI site.....214
Figure 5.25 PCR analysis of homologous recombination between EBV genome and BACs in two time points.....217

TABLE OF TABLES

Table 1.1 Known human herpesviruses. 22
Table 1.2 EBV latent transcription programs. 31
Table 2.1 Antibiotics 85
Table 2.2 Chemical inducers of EBV lytic induction 85
Table 2.3 Oligonucleotides 86
Table 2.4 PCR primers 93
Table 2.5 Primers for Real-Time polymerase chain reaction (qPCR). 94
Table 2.6 Electroporation parameters 98
Table 3. 1 Strains with deletions in unique sequences flanking IR1. 111
Table 3.2 Deletions overlapping the junctions of IR1. 115
Table 3. 3 Subtypes of EBNA-LP 125
Table 3.4 Example of reported MVs at error-prone sites within IR1 unit. 131
Table 3.5 Strains with homogeneous IR1 137
Table 4. 1 List of used strains for terminal repeats analysis. 160
Table 4.2 Potential 9- and 11- bp element sequences 164
Table 4.3 Potential 9- and 11-bp element styles in studied strains 165
Table 5. 1 Comparison between Kanda lab's strategy and my strategy. 200

1 Introduction

1.1 Herpesviruses

The name Herpesvirus is derived from the Greek word “herpein” meaning “to creep or crawl”. This name describes one of the most prominent features of herpesviruses which is that once the virus has infected the host, it persists over the lifetime. Herpesviruses are a group of very ancient viruses that infect human and almost all animal species (Mettenleiter et al. 2009).

All Herpesviruses have common structural features; the virus core holds a densely-packed double-stranded DNA genome of 120-250-kb within an icosahedral protein capsid that is around 125 nm in diameter and contains 162 surface capsomeres. Between the capsid and the viral envelope, there is the tegument that contains virus-coded proteins. The structure of the tegument is generally asymmetrical. This structure is wrapped by the outermost structure, the lipid envelope, which is derived from the host nuclear membrane and contains embedded viral glycoproteins. The overall size of herpesviruses ranges from 120 to 300 nm in diameter owing to the size of the tegument (Jenkins & Hoffman 2000; McGeoch et al. 2006).

Nine herpesviruses are known to infect humans and listed in table 1.1. The Herpesviridae family is divided into three subfamilies named Alpha-, Beta- and Gammaherpesvirinae based on genetic features (e.g. sequence homology) and biologic properties such as host range, duration of the replicative cycle in the host, growth pattern in cell culture and site of latency establishment (Jenkins & Hoffman 2000; Pellett & Roizman 2013):

1.1.1 *Alphaherpesvirinae*

Alphaherpesviruses have a variable host range, a short reproductive cycle and a rapid spread in cell culture. They establish the latency cycle in the sensory ganglia and, as a result, they are referred to as neurotropic herpesviruses in general. The human herpesviruses that belong to this family are HSV-1, HSV-2 and VZV.

1.1.2 *Betaherpesvirinae*

Unlike Alphaherpesviruses, betaherpesviruses have restricted host range (mammalian hosts only), a long reproductive cycle in host cells and slow infection progress in cultured cells compared to Alphaherpesvirinae. Betaherpesviruses maintain latency in several types of cells including lymphocytes, secretory glands, and cells of the kidney. This subfamily contains the human herpesviruses cytomegalovirus (CMV), human herpesvirus 6 (HHV-6 A and B), and human herpesvirus 7 (HHV-7).

1.1.3 *Gammaherpesvirinae*

The replication and latency of gammaherpesviruses occur in lymphoid tissues. Viruses in this group are generally restricted to either to B or T cells. The Gammaherpesvirinae is further divided into four genera: Lymphocryptovirus, Rhadinovirus, Macavirus and Percavirus. Among the human herpesviruses, Epstein–Barr virus (EBV) is a member of the Lymphocryptovirus genus while KSHV is a member of the Rhadinovirus genus.

Table 1.1 Known human herpesviruses.
Classification, latency sites and symptoms of human herpesviruses. (Modified from Riddell et al. 2018; Thorley-Lawson & Allday 2008)

Common name	Abbreviation	Nomenclature	Subfamily	Common symptoms during primary infection	Latent sites	Example of pathogenesis
Herpes simplex virus type 1	HSV-1	HHV-1	α	Mucocutaneous lesions	Sensory ganglia	Recurrent mucocutaneous lesions
Herpes simplex virus type 2	HSV-2	HHV-2	α	Mucocutaneous lesions	Sensory ganglia	Recurrent mucocutaneous lesions
Varicella-zoster virus	VZV	HHV-3	α	Chickenpox	Sensory ganglia	Shingles (zoster)
Human Cytomegalovirus	HCMV	HHV-5	β	Asymptomatic, infectious mononucleosis-like syndrome, congenital infection of fetus	Mononuclear cells, epithelial cells	Non-specific CMV syndrome, retinitis, oesophagitis, colitis, hepatitis
Human herpesvirus 6A	HHV-6A	HHV-6A	β	Roseola infantum	T cells	Encephalitis
Human herpesvirus 6B	HHV-6B	HHV-6B				
Human herpesvirus 7	HHV-7	HHV-7	β	Similar to HHV-6	T cells	Unknown
Epstein Barr virus	EBV	HHV-4	ɣ	Asymptomatic, infectious mononucleosis	B cells	Lymphomas (e.g. Burkitt's, and Hodgkin's) and nasopharyngeal carcinoma
Kaposi sarcoma-associated herpesvirus	KSHV	HHV-8	ɣ	Asymptomatic	Lymphocytes	Kaposi's sarcoma, primary effusion lymphoma, multicentric Castleman disease

1.2 Herpesvirus genomes

Herpesvirus genomes are not a mere linear DNA of a unique sequence, but they include either direct or inverted repeats. Based on the structure of repeats, herpesvirus genomes can be divided into six major classes, named A to F (figure 1.1) (Davison 2007).

Class A genomes consist of a unique sequence flanked by direct terminal repeats. Betaherpesviruses HHV-6 and HHV-7 genomes represent the class A. Class B genomes also have direct terminal repeats like class A but with variable copy numbers of a tandemly repeated sequence. This arrangement is represented among most Gammaherpesvirinae in the Rhadinovirus genus, such as herpesvirus saimiri (HVS) and KSHV. Class C is sub-class of class B and is exemplified by EBV, where the genome has an internal set of internal repeats but is unrelated to the terminal repeats.

Class D genomes, exemplified by VZV, have two unique regions and flanked by inverted repeat pairs. Class E is similar to class D except that the terminal repeat regions are larger and they are repeated in an inverted direction internally dividing the genome into two unique regions flanked by inverted repeats. The class E is characteristic of HSV-1 and HCMV. Class F differs from other by lacking long repetitive DNA sequences. There is no human herpesvirus genome characterised as class F which is represented by tupaia herpesvirus, which has 2,000 direct and 1,611 inverted repeats but in short length (no longer than 25-bp) (Bahr & Darai 2001).

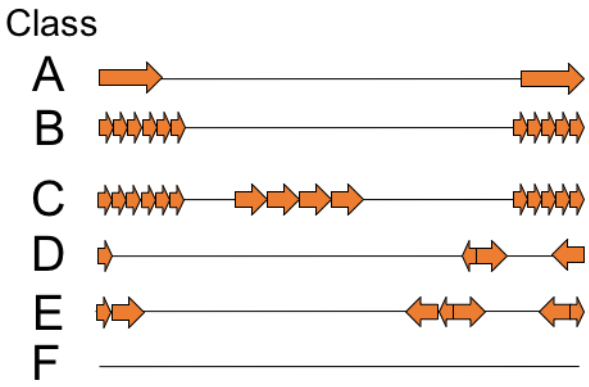


Figure 1.1 Schematic of herpesvirus genome structures.
Names of classes are shown to the left. Unique regions are presented as black lines while repeat regions and their orientations are shown in orange arrows (Figure obtained from Davison 2007 with modifications).

1.3 Epstein Barr Virus

1.3.1 Discovery of EBV

The journey towards the discovery of EBV started on March 22nd 1961 in the Courtauld lecture theatre at Middlesex Hospital London when Dr Antony Epstein attended a talk entitled "The commonest children's cancer in tropical Africa: A hitherto unrecognised syndrome" given by Denis Burkitt (an Irish surgeon working in Uganda). In that very lecture, Dr Burkitt described a mysterious malignant tumour of children in Africa, which is now known as Burkitt lymphoma (Burkitt 1958), and how its incidence depended on temperature and rainfall. Antony Epstein deduced that this bizarre tumour affected by climate should be caused by a biological agent. Since then, Epstein had been receiving samples from Burkitt in an attempt to rescue virus from the tumour cells. However, these attempts were in vain until December 1963 when a sample was delayed by fog, and arrived in cloudy media. This sample could have been discarded as the turbidity suggested that the lengthy journey had allowed bacteria to grow in the sample, but instead, Epstein tested the sample under a light microscope

and found a large number of free-floating tumour cells. This sample was the first culture of Burkitt's lymphoma-derived cells line designated EB1 (E for Epstein and B for Mrs Barr, a research assistant). In February 1963, Epstein tested EB1 sample under an electronic microscope and identified viruses similar to HSV but smaller in size than HSV suggesting a different type of herpesviruses (figure 1.2). The biological inertness of the virus and its immunological singularity was reported announcing of a new member of the herpes family called EBV after EB cell line (Epstein et al. 1965; Henle & Henle 1966). EBV is also known as human herpesvirus 4 (HHV4). The oncogenic potential of EBV was demonstrated by reporting that mixed cultivation of irradiated EBV producer BL cell line (Jijoye) (causing cell lysis and release of EBV DNA) and normal peripheral leukocytes lead to the proliferation of hematopoietic cells (Henle 1968). Subsequently, EBV DNA was found within cells from Burkitt's lymphoma (BL) and nasopharyngeal carcinoma (NPC) (zur Hausen et al. 1970; Henle 1968). Because of these observations and others by the late 1970s, EBV was recognised as the first human tumour virus and the idea that virus can cause cancer in humans has been proved.

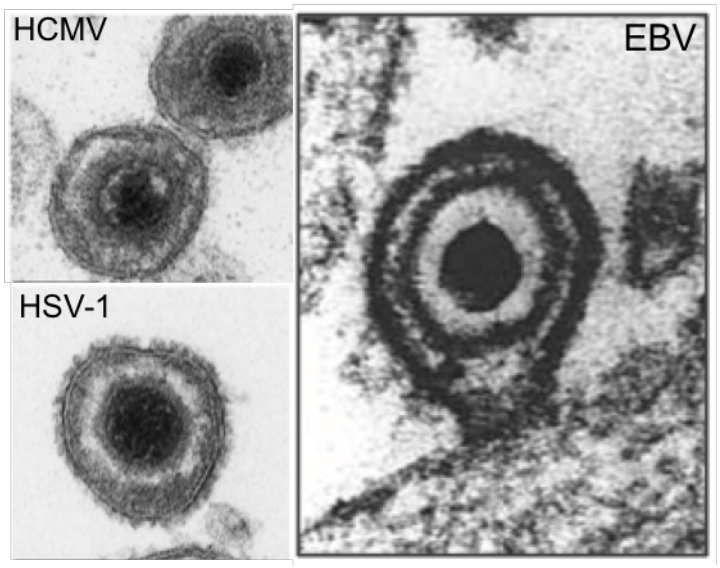


Figure 1.2 Morphology of herpesvirus virions.
Electron micrographs of Epstein Barr virus (EBV), human cytomegalovirus (HCMV) and Herpes simplex virus 1 (HSV-1). (Modified from Mettenleiter et al. 2009; Young et al. 2009).

1.3.2 Taxonomy

EBV (designated as human herpesvirus-HHV4), is the prototype of the lymphocryptovirus (LCV) genus within the Gammaherpesvirus subfamily of the Herpesviridae. EBV is the only lymphocryptovirus that infects humans, but it shares sequences and genetic organisation with the rhesus monkey LCV, and they both maintain infection in B cells and the oropharynx (Moghaddam 1997).

1.3.3 EBV genome structure

The EBV genome is composed of a linear, double-stranded, approximately 172-kb DNA molecule (based on the prototype EBV genome NC_007605 - de Jesus et al. 2003), which characterised by distinctive iterations sequences and encodes more than 85 genes (Baer et al. 1984). The linear EBV genome contains multiple numbers of direct repeat units (2-5 copies in B95-8) at both ends that mediate the circularisation of EBV genome, creating EBV episomes in infected cells (figure 1.3A) (Zimmermann & Hammerschmidt 1995). It also includes four classes of internal direct repeats that divide the genome into five unique regions (figure 1.3B) (Young et al. 2009) (for more details about repeat regions, see section 1.7). The mean of GC-content across EBV genome is 59.5%. Compared to Alpha- and Betaherpesvirinae, most of Gammaherpesvirinae genomes have a lower percentage of GC-content (Honess et al. 1989). The depletion of GC-content is attributed to the dividing activity of host cells where 5-methylcytosine residues can be spontaneously deaminated to thymidine residues, a change that is fixed during DNA replication. Since latent viruses are obligated to replicate as host cells divide, exposure to this process could make the GC-content in EBV, which establish latency in dividing B cells, is depleted compared to those in neurotropic herpesviruses HSV-1 which

persists in non-dividing neurons (Honess et al. 1989). Recently, it has been reported that EBNA3C induces the expression of Activation-induced cytidine deaminase (AID) in EBV-infected B cells (Kalchschmidt et al. 2016). AID is essential in B cell differentiation by deamination of cytosine and turns it into uridine. Thus, AID might be another factor involved in GC-content reduction in EBV genome.

The nomenclature of genes in EBV genome was based on regions defined by the BamHI-restriction fragments named from A to Z and then a to e, where A is the largest fragment and e the smallest. The open reading frames (ORFs) were designated according to position and whether they extend rightward or leftward on the EBV genome. For example, BARF1 ORF is found in the BamHI A fragment, extends rightward, and is the first ORF in that fragment (Skare & Strominger 1980).

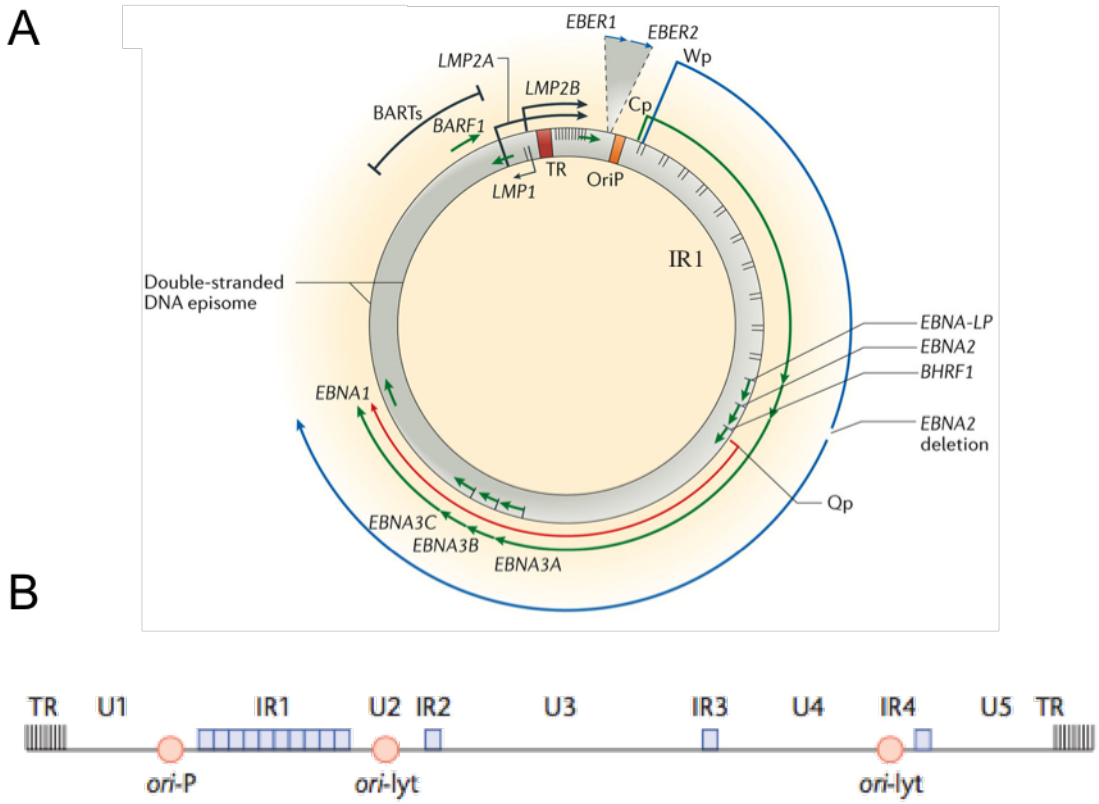


Figure 1.3 Schematic of EBV genome.
(A). Structure of EBV episome and transcription of EBV latency genes. The inner short green arrows show locations and transcription direction of latency genes. The long green arrow in the middle shows EBV transcription during latency III from either the Wp or Cp promoter. The outer long blue arrow show the transcript in Wp-restricted latency in which EBV genome is lacking EBNA2. The short blue arrows show the non-polyadenylated EBV- encoded RNAs (EBER1 and EBER2). The inner red arrow the direction of EBNA1 transcription from Qp promoter during latency I and II. **(B)** A schematic of linear EBV genome showing the regions of terminal repeats (TR), internal repeats (IR1 to IR4) and unique sequences (U1 to U5). *ori*-P, origin of replication; *ori*-lyt, lytic origin of DNA replication. (Modified from Young et al. 2016; Young et al. 2009).

1.3.4 EBV latent infection *in vivo*

EBV spreads between individuals through saliva (Thorley-Lawson 2001)**,** and the primary infection is usually asymptomatic , after which the virus persists in resting memory B cells over host's life asymptomatically in most cases (Young et al. 2009). In infected B cells, EBV adopts different programmes of gene expression depending on the differentiation state of

infected B cells. The current model of persistence was proposed by David Thorley-Lawson (Thorley-Lawson & Babcock 1999) (figure 1.4). Initially, EBV enters the epithelium where it initiates a lytic cycle for virus amplification. Then, the virus infects a naïve B cell, probably in Waldeyer's ring in the oropharynx, and establishes a pattern of transcription called (latency state III) where it expresses all latent proteins: the six EBV nuclear antigens (EBNA 1, 2, 3A, 3B, 3C and LP); three latent membrane proteins (LMP1, LMP2A, LMP2B); two small RNA molecules RNA named EBV-encoded RNAs (EBER1 and EBER2); and microRNAs (miRNAs) – 22 derived from introns of rightward transcripts from the BamHI A fragment (BARTs) and three from the BHRF1 transcript (Kang & Kieff 2015) (table 1.2). The naïve infected B cells then enter the germinal centre (GC), where they proliferate and clonally expand, thus increasing the EBV-infected pool of cells. These infected cells in the GC establish a restricted transcription programme (latency II) in which the EBNAs (except EBNA1) and BHRF1 miRNAs have been shut off. This programme is proposed to let infected cells survive in the GC regardless of antibody specificity, and differentiate to exit as resting memory B cells, which seem to represent the long-term reservoir for EBV. In memory B cells, the true latency programme (latency 0) begins, in which no virus proteins are translated. When these B cells occasionally divide, they transiently express latency programme I, whereby only one viral protein - EBNA-1 - is synthesised, as it is required for the replication and segregation of the EBV episomes (Young & Murray 2003). When infected memory B cells undergo plasma-cell differentiation, viral replication is triggered. The resulting virus shed into saliva, and it could infect other B cells within the same host or infect another host (Thorley-Lawson 2001; Thorley-Lawson & Allday 2008). The mechanism orchestrates EBV latency-type switching is largely unknown. However, both cellular and viral epigenetic regulation of the EBV latency promoters has a role (Takacs et al. 2010). Defects in this process can transform the infected cell into a malignant cell.

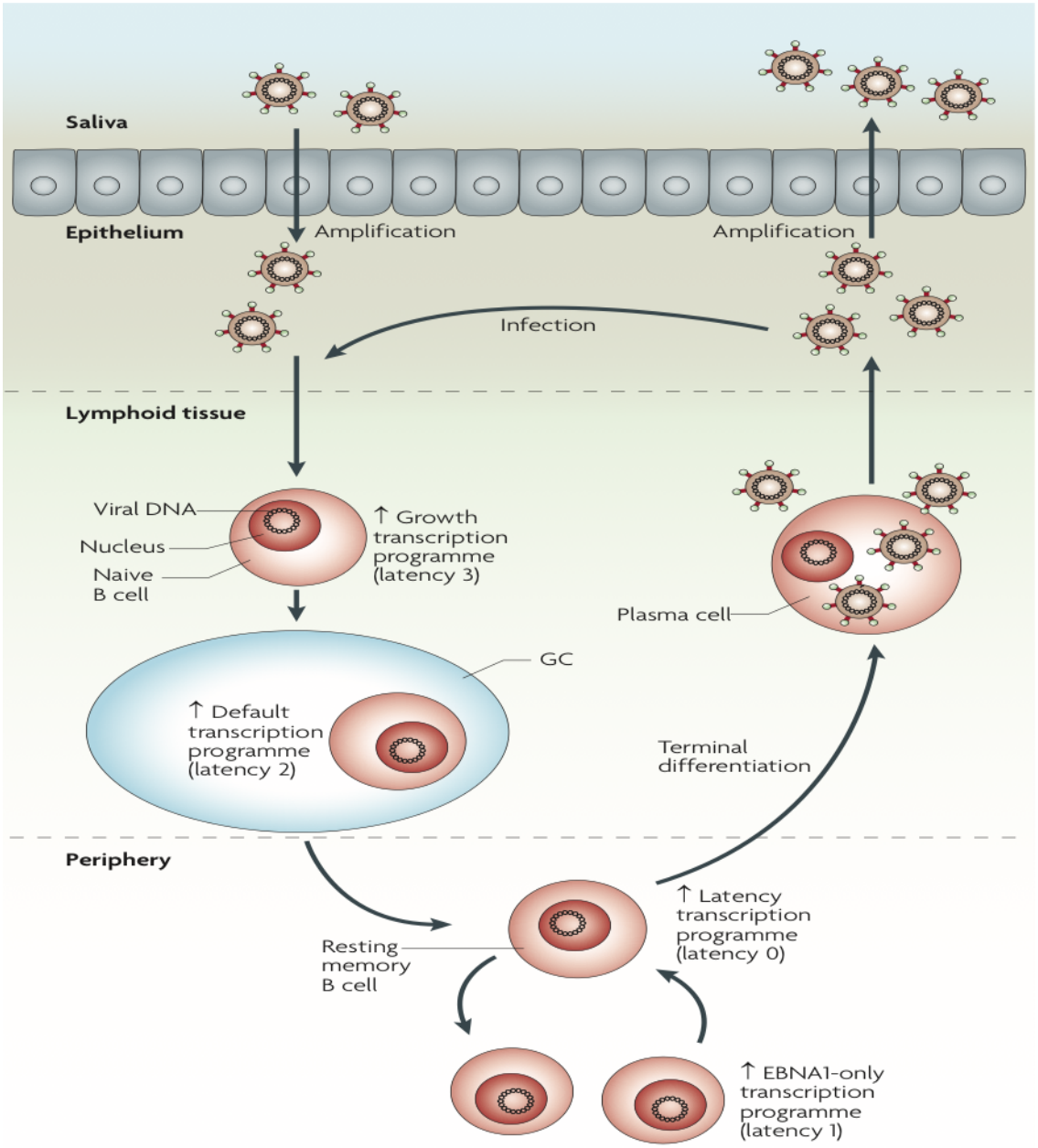


Figure 1.4 Proposed model of how EBV establishes and maintains latency infection. Refers to text in section 1.3.4 for details and references. (Modified from Thorley-Lawson & Allday 2008).

Table 1.2 EBV latent transcription programs.

Latency programmes	Expressed latency genes	EBNA promoters	Cell type where occurs	Associated diseases
0	EBER-1 & -2 BARTs	Unknown	Resting memory B cells	
I	EBNA-1 EBER-1 & -2 BARTs	Qp	Dividing memory cells	Burkitt's lymphoma
II	EBNA-1 LMP-1 LMP-2A & B EBER-1 & -2 BARTs	Qp	Germinal center B cells	Hodgkin's lymphoma
III	EBNA-1 EBNA-2 EBNA-LP EBNA-3A EBNA-3B EBNA-3C LMP-1 LMP-2A & B EBER-1 & -2 BARTs	Wp -> Cp	Newly infected B cells and lymphoblastoid cell lines (LCLs)	Post-transplant lymphoproliferative disease (PTLD)

1.3.5 EBV latent infection *in vitro*.

In vitro, EBV readily infects resting B lymphocytes and transforms them from a resting state to become continually proliferating lymphoblastoid cell lines (LCLs). EBV binding to B lymphocytes is mediated by a direct interaction of the viral glycoprotein gp340/220 with complement receptor 2 (CR2, also known as CD21) on B lymphocytes (Sinclair & Farrell 1995). EBV entry to B lymphocytes differs from entry to epithelial cells. In B cells EBV enters through endocytosis and fuses with the endosome membrane at low pH, while infection of epithelial cells occurs at the cell surface in the absence of endocytosis (Kieff & Rickinson 2007). LCLs express all known latency genes and RNAs in latency III (table 1.2) (for latency programmes see section 1.3.4.). EBV is divided into two subtypes: type 1 and type 2 defined

by differences in the sequence of EBNA-2, 3A, 3B, 3C genes and based on their ability to transform human B cells into lymphoblastoid cell lines (Sample et al. 1990). Type 1 EBV transforms B cells into LCLs more efficiently than type 2 and it is the most prevalent worldwide while type 2 is most common in equatorial Africa (Tzellos et al. 2014).

1.3.6 EBV tropism to B cells and epithelial cells.

The current model suggests that viruses are shed into saliva from epithelial cells in the salivary gland because of switching in the tropism of EBV. EBV has three-part and two-part complexes of envelope glycoproteins, which drive the tropism of EBV to B cells or epithelial cells (Wang et al. 1998). These complexes consist of either gH, gL and gp42 or gH and gL only. B-cell entry requires binding of gp42 to HLA class II while epithelial cell infection, where there is no HLA II, EBV entry requires two-part complex gH-gL (without gp42). It has been shown that HLA class II in B cells degrades three-part gHgLgp42 complexes and increase the proportion of the gHgL protein complex. In contrast, viruses produced in epithelial cells (which are HLA II-negative) gp42 would not be affected and thus the relative amount of gp42 would increase. Consequently, EBV produced from B cells better infects epithelial cells whereas EBV from epithelial cells is better at infecting B cells (Borza & Hutt-Fletcher 2002).

1.3.7 EBV latency promoters

During the latency III programme, production of EBNA occurs from a single large transcript that spans around 100 kb, initiated at either Wp or Cp (figure 1.5). The initial transcript undergoes a combination of alternative splicing that leads to the synthesis of the

various mature EBNAs transcripts. During the restricted latency programmes (I and II), EBNA1 expression arises from Qp while Wp and Cp are silent (Nonkwelo et al. 1996). The elevated level of EBNA1 down regulates Qp activity and EBNA1 bind to oriP (origin of episomal replication) to enhance Cp and Wp. Thus, the major feature that distinguishes latency programmes is the activity of EBNA genes promoters (Cp/Wp/Qp) (figure 1.5) (Paulson & Speck 1999).

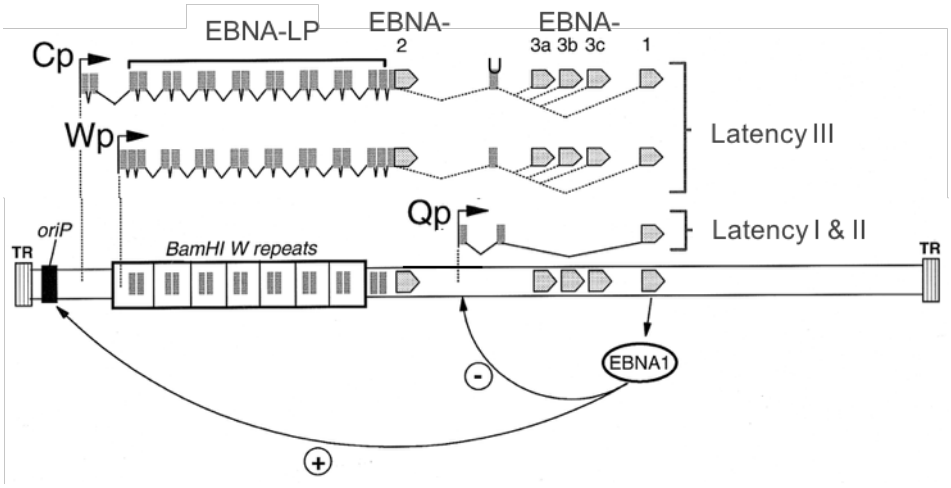


Figure 1.5 Latency promoter activity during the EBV latency programmes.
A schematic illustration of the activity of latency promoters during latency programme I, II, and III. The linear EBV genome with the location of EBNAs are shown at the bottom. During latency III, all six EBNAs are generated from a complex alternative splicing of a long primary transcript that initiates at either Cp or Wp promoters. In latency I and II, only EBNA-1 is expressed and the transcript is derived from Qp. Ori-P; the origin of episomal replication, TR; terminal repeats. (Modified from Paulson & Speck 1999).

1.3.7.1 Wp promoter

The BamHI W promoter, Wp, is present in every unit of the tandem IR1 repeats, which vary in number between viral isolates, usually 5.6 to 8.6 units (Tierney et al. 2011; Palser et al. 2015). Wp is the first promoter to be used upon infection, and it is activated within 12-16 h post infection (Allday et al. 1989; Alfieri et al. 1991).Wp is important for B cell transformation.

It has been shown that Wp was 11- to 190- fold more active in B cells than in non-B cells by quantifying luciferase expression at 16-24 h post transfection (Bell et al. 1998). The cell lineage-dependent activity of Wp is attributed to three regions upstream Wp, namely UAS1, UAS2, and UAS3 at positions 14244-14297, 14032-14120 and 13216-13944 (according to the first IR1 unit of the prototype EBV sequence NC_007605 (hybrid sequence of B95-8 strain with the known deletion filled with the appropriate sequence of the Raji genome) (de Jesus et al. 2003). UAS3 and UAS2 are cell lineage-independent while UAS1 is B cell specific and the key for Wp activity in B cells (Bell et al. 1998).

Mechanisms underlying the activation of Wp are not fully clear, but Wp is mainly regulated by cellular factors. Initially, NF-kB (B cell regulator) was identified as the first cellular factor that binds and activates Wp (Sugano et al. 1997). However, this observation could not be repeated by another lab (Bell et al. 1998). The ubiquitous transcription factor Yin-Yang 1 (YY1) has a binding site in UAS2 that is required for Wp activation in non-B cells (but mutation of the YY1 site has less effect in B cells) (Bell et al. 1998). Analysing UAS1 revealed four binding sites; two sites for the B cell-specific transcriptional activator Pax5, one for the cAMP-dependent CREB/ATF factors and one site for the ubiquitous regulatory factor X (RFX) (Tierney et al. 2000; Kirby et al. 2000; Bell et al. 1998). All these transcription factors bind to UAS1 simultaneously and are required for Wp activity *in vitro* (Tierney et al. 2000).

1.3.7.2 Cp promoter

The BamHI C promoter (Cp) is active only in B cells and it is the main promoter driving EBNA expression in LCLs (Contreras-Brodin et al. 1996). Upon B cell infection with EBV, Wp is responsible for EBNA transcripts but after 3 to 6 days a switch to use Cp occurs (Woisetschlaeger et al. 1990). However, Cp is dispensable for B cell transformation *in vivo*.

Viruses with mutant or deleted Cp were able to transform B cells as efficiently as wild-type Cp EBVs, through Wp usage to produce the EBNA proteins (Swaminathan 1996). Primary infection using a virus lacking EBNA2 gene (P3HR1) utilized Wp and failed to switch to Cp, indicating that EBNA2 is required to activate Cp. In reporter assays, EBNA1 enhances Cp and Wp via its binding to oriP (Paulson & Speck 1999) (figure 1.5). Several cellular transcription factors regulate Cp. For instance, RBP-JK and CBF2 are required for EBNA2-dependent Cp activation (Fuentes-Pananá & Ling 1998). Also, NF-Y and Sp1 are essential for Cp activation in the presence of EBNA2, which suggests that these factors may have a role in the Wp/Cp switch (Borestrom et al. 2003).

1.3.7.3 Qp promoter

The BamHI Q promoter, Qp, is an alternative promoter for EBNA1 that drives EBNA1 expression in the absence of the other EBNAs (Nonkwelo et al. 1996). Qp is a TATA-less promoter, has many features of housekeeping gene promoters (Schaefer et al. 1995). Unlike Cp and Wp, DNA methylation does not control Qp (Takacs et al. 2010). Interferon regulatory factors (IRFs) regulate Qp activity. IRF1 and IRF2 upregulate Qp (Schaefer et al. 1997). Also, it has been shown that STATs positively regulates Qp (Chen et al. 1999). In contrast, Qp is down-regulated by EBNA1 binding to two recognition sites downstream of Qp, which is likely a mechanism by which EBNA1 regulates its own expression (Sample et al. 1992).

1.3.7.4 LMP1p promoter

The LMP1 promoter is regulated by EBNA2, which does not bind directly to the DNA but it is rather tethered to the DNA via binding to the transcription factor RBP-Jk (Fåhraeus et

al. 1990; Szymula et al. 2018). The LMP1 transcription regulatory sequence (LRS), upstream of the LMP1p, contains two binding sites for RBP-Jk, but only the site closest to the promoter has a role in transcriptional activation (Johannsen et al. 1995). LMP1 activation via either EBNA2-dependent or independent transactivation also requires an intact cAMP-responsive element (CRE) in the LRS. Mutations in CRE result in a reduction in the promoter activity (Fåhraeus et al. 1990). PU.1-dependent transcription from LMP1p can be activated by a cooperation between EBNA3C and EBNA2 (Zhao & Sample 2000).

In epithelial cells, LMP1p is inactive, but LMP1 is generated from an alternate TATA-less promoter (called ED-L1) within the first terminal repeat unit (Sadler & Raab-Traub 1995). Therefore transcripts are larger than those expressed from LMP1p in B cells (Sadler & Raab-Traub 1995). ED-L1 is regulated by the cellular Sp1 family proteins (Sp1 and Sp3) interaction with the GC box (transcriptional regulatory element that contains the sequence GGGCGG) (Tsai et al. 1999).

1.3.7.5 LMP2Ap promoter

The LMP2A promoter is located around 3.2 kb downstream of the LMP1 transcription start site, on the opposite DNA strand. LMP2A has two binding sites for RBP-Jk (Salamon et al. 2003; Hofelmayr et al. 1999). LMP2B expression is derived from LMP2Bp in the same bidirectional promoter region as LMP1. However, the controlling factor of LMP2B is unknown.

1.3.8 EBV latency proteins

EBV undergoes sequential latency programmes to reach persistence in a resting

memory B cell by progressively reducing EBV protein expression from latency state III to latency state 0. During latency III, EBV expresses all latent proteins while there is no viral protein during latency state 0 (table 1.2 and figure 1.6).

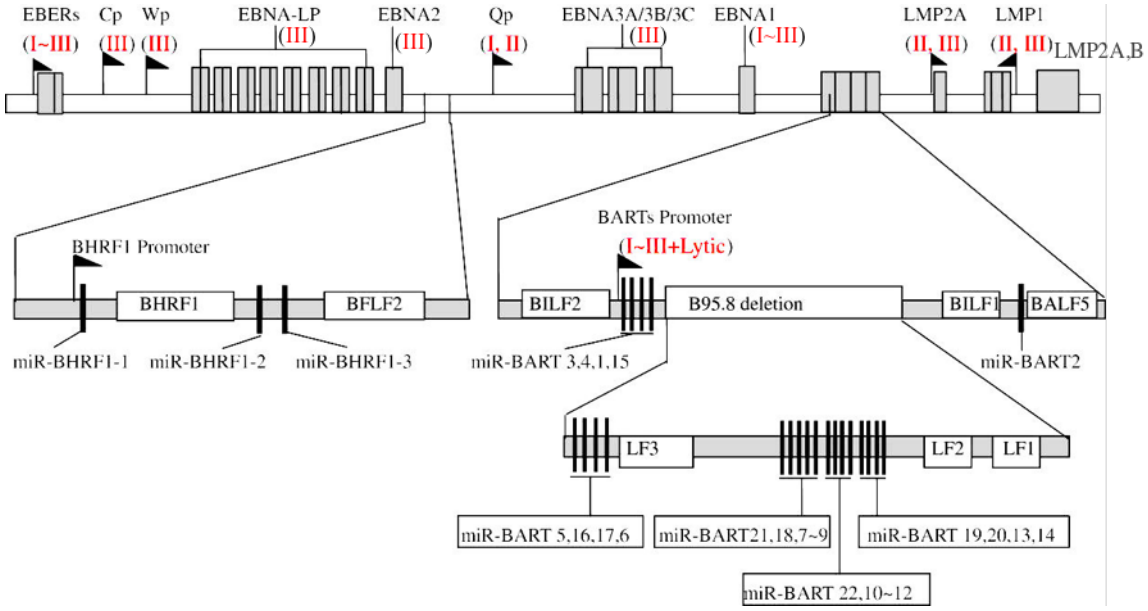


Figure 1.6 Positions of latency genes and promoters.
The top panel illustrates the positions of latency genes (grey boxes), latency state, and driving promoters (pennants). The middle panel shows the positions of the EBV miRNAs; BARTs and BHRF1. The bottom panel illustrates the BARTs that are located in the deletion region in B95-8. (Modified from Yang et al. 2013).

1.3.8.1 EBNA1

EBNA1 is expressed in all EBV-positive tumours (and all proliferating EBV-positive cells) as it is required for the replication and maintenance of the episomal EBV genome in infected cells. This function is performed via the binding of EBNA1 to the dyad symmetry (DS) element of the plasmid origin of viral replication, oriP (Kieff & Rickinson 2007).

EBNA1 down-regulates its expression by interacting with two sites immediately downstream Qp promoter, which drives EBNA1 in restricted latency programmes (Nonkwelo

et al. 1996). Since EBNA1 has more affinity for the DS and the family of repeat elements than for its two recognition sites downstream of Qp, EBNA1 would bind to the lower affinity sites in Qp only when its level is high enough to saturate FR and DS, providing a feedback mechanism to suppress EBNA1 expression (R F Ambinder et al. 1990). Based on luciferase assays, EBNA1 upregulates the Cp promoter and the LMP1 promoter, suggesting that EBNA1 has a role in transcriptional activation of latent genes (Gahn & Sugden 1995).

The EBNA1 protein sequence contains a glycine-glycine-alanine (Gly-Ala) repeat sequence which separates the protein sequence into amino- and carboxy-terminal domains and varies in size in different EBV isolate (~230 amino acids long in the B95-8 strain). This repeat domain acts as a *cis*-acting inhibitor of major histocompatibility complex (MHC) class I-restricted presentation of peptides by inhibiting antigen processing. As a result, EBNA1 is not presented on MHC class 1, so its presence in cells cannot be detected by T-cells, making it invisible to the immune system (Levitskaya et al. 1995).

1.3.8.2 EBNA2

EBNA2 is essential for transforming B cells into LCLs. The role of EBNA2 in cell transformation was demonstrated by showing that the EBV strain P3HR1 (carrying a deletion of the EBNA2 gene, and truncated EBNA-LP) was unable to transform B cells *in vitro* and that restoration of the EBNA2 from B95-8 into P3HR1 restored this ability, unequivocally confirming the importance of EBNA2 in B-cell transformation (Hammerschmidt & Sugden 1989). One difference in a single amino acid of EBNA2 (Ser in type 1 EBV, Asp in type 2 EBV at position 442) confers the superior ability of type 1 EBV in transforming B cells into proliferating LCLs (Tzellos et al. 2014) (see section 1.5).

EBNA2 and EBNA-LP are the first proteins to be expressed during the establishment of latent infection in B lymphocytes (Allday et al. 1989). EBNA2 is an acidic phosphoprotein and upregulates certain B-cell antigens, such as CD21 and CD23 and the key viral genes LMP1 and LMP2A (Young & Rickinson 2004). EBNA2 also trans-activates the viral Cp promoter, inducing the switch from the immediate early Wp to Cp, which remains the dominant promoter in latency III (table 1.2). The proto-oncogene c-MYC (involved in cell cycle progression, apoptosis, survival and transformation) is induced by EBNA2, and this effect seems to be important for proliferation of EBV-infected B cells (Kaiser et al. 1999).

1.3.8.3 EBNA-LP

The EBV nuclear antigen leader protein (EBNA-LP) is encoded across the major internal repeat (IR1) of EBV. As a result, it has a repeat domain composed typically of two to seven 66 amino acid segments encoded by the W1 and W2 exons and joined to a unique 45 amino acid segment at the C-terminal end that is encoded by the Y1 and Y2 exons (encoding 11 and 34 aa respectively). EBNA-LP can be transcribed from either Wp (the promoter present in each IR1 repeat unit) or Cp (located upstream of IR1) (Ling et al. 2009). EBNA-LP is expressed in multiple isoforms with variable sizes according to the number of IR1 repeats contained by EBV isolates (figure 1.7). There is no definite explanation for how multiple isoforms of EBNA-LP arise. From Wp it is easy to envisage transcription initiation from Wp in each of the IR1 repeat units. For Cp-derived transcripts, this phenomenon can be attributed to alternative splicing across IR1 region creating several transcripts encoding EBNA-LP with variable sizes (Kempkes & Ling 2015).

EBNA-LP is not required for B-cell transformation in vitro (Hammerschmidt & Sugden 1989), but during the course of this PhD, our lab showed that EBNA-LP is essential for

transforming umbilical cord B cells (and probably naive B cells in general) but not adult memory B cells (Szymula et al. 2018). The mechanism of EBNA-LP with EBNA2 in transforming cells is not well known, but it has been deduced that EBNA-LP does not interact directly with EBNA2 but instead facilitates the recruitment of transcription factors to EBV genome, to enable transcription of viral but not host genes (Szymula et al. 2018).

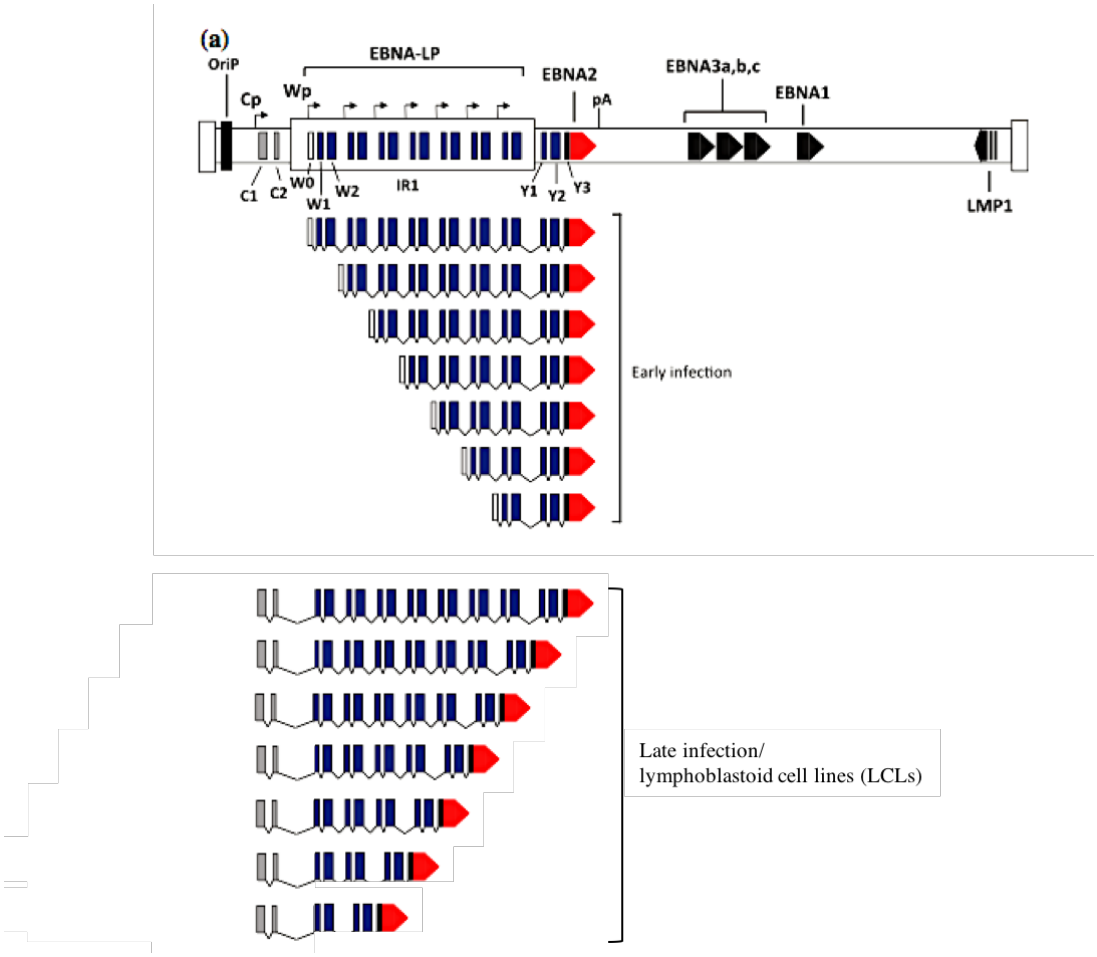


Figure 1.7 Exon organization of EBNA-LP gene transcripts.
EBNA-LP is transcribed from either the Wp promoter (during early infection) or the Cp promoter (during late infection). The latent origin of replication ori-P, location of EBNAs and LMP and the polyadenylation site (pA) for EBNA-LP/EBNA-2 transcripts are shown. The non-coding exons (C1, C2, and W0) are in grey. The coding exons for EBNA-LP (W1, W2, Y1 and Y2) are in blue and the EBNA2 exon Y3 in red. The schematic shows the proposed mechanism that produces multiple isoforms of EBNA-LP. (Modified from Kempkes & Ling 2015).

1.3.8.4 EBNA3 family

The three members of EBNA-3 family EBNA-3A (or EBNA-3), EBNA-3B (or EBNA-4), and EBNA-3C (or EBNA-6) are expressed from a polycistronic transcription unit (figure 1.8). EBNA3 proteins share partial homology near the *N* terminus and have similar gene structure (Allday et al. 2015). This conserved domain in the three members mediates the binding to RBP-Jk. EBNA-3A and EBNA-3C binding to RBP-Jk has been proposed to downregulate activity of EBNA-2-mediated transactivation of the RBP-Jk-dependent Cp, LMP-1 and LMP-2A promoters (Waltzer et al. 1996). EBNA-3A and EBNA-3C have been shown to be necessary for efficient in vitro EBV-infected B cell proliferation as they interface with tumour suppressor pathways (Allday et al. 2015). In contrast, EBNA-3B is dispensable for transformation. Indeed, it appears to have a tumour suppression activity, as deletion of EBNA3B promotes lymphomagenesis in a humanized mouse model, and EBVs with mutated EBNA3B genes have been identified in various B cell malignancies (White et al. 2012).

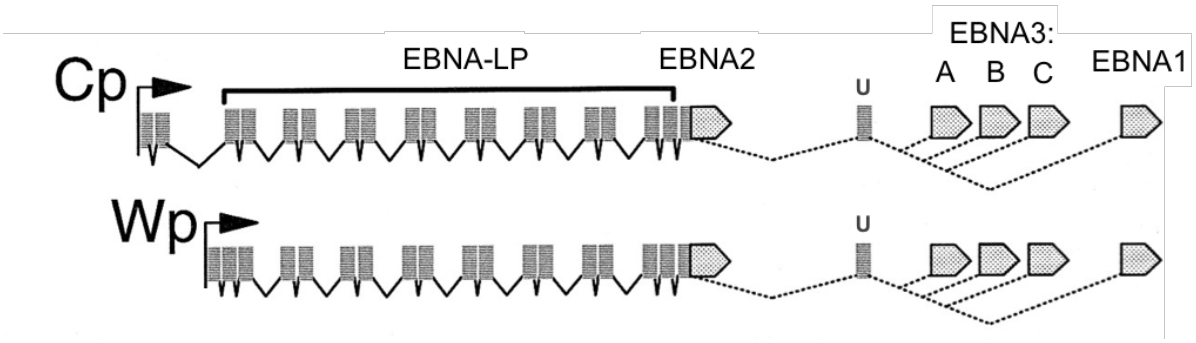


Figure 1.8 Schematic of the alternative splicing of EBNA3 mRNAs.
During latency III, alternative splicing (depicted by dotted lines) of long polysistronic transcripts derived from either Wp or Cp produces individual EBNA1 and EBNA3 mRNAs. (Modified from Paulson & Speck 1999).

1.3.8.5 LMP1

Latent membrane protein-1 gene is located entirely within the LMP 2A gene on the reverse DNA strand. Thus, LMP1 transcription occurs in the reverse orientation. It is essential for EBV-induced transformation, and it is regarded as a major EBV-encoded oncogene: it mimics the signals that are induced by CD40, a member of the tumour necrosis factor receptor (TNF-R) family (Waltzer et al. 1996). LMP1, also, induces several of NF-kB-dependent genes (Gahn & Sugden 1995). This induction contributes to many of LMP1 effects including the protection of apoptosis (Young & Murray 2003).

1.3.8.6 LMP2A and LMP2B

The LMP2 gene spans both ends of the linear genome (ie the terminal repeats lie between the first LMP2 exons, and exons 2-9) so it can only be expressed from a circularised EBV episome. LMP2 transcription can arise from two separate promoters. The LMP2A open reading frame begins in its first exon, whereas LMP2B transcripts start their translation in the common exon 2, meaning that LMP2A has an additional N-terminal region relative to LMP2B. Neither of the LMP2 proteins is essential for B-cell transformation *in vitro* (Longnecker et al. 1993). LMP2A supresses the activation of lytic cycle by blocking the activation of the promoter for BZLF1 (the key protein in the switch from latency state to productive replicative cycle) in LCLs (Jenkins & Binne 2000). In BLs such as the Akata strain, the expression of LMP2 is very low or abolished. However, co-expression of LMP2A and LMP2B abolishes the effect of LMP2A on BZLF1 and results in an increased lytic cycle, suggesting that LMP2B may exert a dominant negative effect on LMP2A (Rechsteiner et al. 2008).

LMP2A is one of the viral genes that helps EBV to accomplish its latency strategy. Just as LMP1 mimics the activity of an activated CD40, LMP2A constitutively assembles a signal transduction complex similar to that of an activated membrane-bound antibody (the B cell

receptor). Together these two signals can allow survival of B cells in the germinal centre, which suggests the likely importance of their expression (via latency II) in germinal centre cells, and for the persistence of GC cancers such as Hodgkin lymphoma (Fruehling et al. 1998).

1.3.8.7 EBERs

Epstein-**B**arr virus-**e**ncoded **R**NAs (EBER1 and EBER2) are small non-coding, non-polyadenylated RNAs. They are transcribed by RNA polymerase III, but have promoter elements more commonly found in RNA polymerase II-transcribed genes (Howe & Shu 1989). EBER1 is 166 nt long while EBER2 is 172 nt long and they are separated by 166 nt. The EBERs are expressed at very high abundance in all latently infected cells, which has made them an ideal target molecule for the in situ hybridization (ISH) detection of EBV in tissues. However, the EBERs are not expressed in some diseases where EBV is replicating in epithelial cells, including oral hairy leukoplakia, Sjögren's syndrome and oral papilloma (Iwakiri 2014).

Even though the EBERs share low primary sequence similarity (54%), they have a very similar secondary structure which resembles that of adenovirus-associated RNAs (VAs) (Iwakiri 2014). Functions of EBERs are still poorly understood and controversial. For instance, Akata-derived viruses in which the EBERs were inactivated showed a ~100-fold decrease in the transformation efficiency compared to wild-type virus (Yajima et al. 2005). Using the same strain (Akata), the role of EBER2 in cell transformation is superior to that of EBER1. Transformation ability of recombinant EBVs expressing EBER1 only was impaired and was similar to that of EBV lacking both EBER1 and EBER2. In contrast, transformation ability of recombinant EBVs expressing EBER2 only was as high as intact EBVs (Wu et al. 2007). The efficiency of EBER2 is attributed to its efficient induction of IL6 which in turn enhances LCL growth (Wu et al. 2007). However, deletion of the EBERs from B95-8-derived viruses had no

measurable difference in transforming B cells compared to the wild-type virus (Gregorovic et al. 2011). These inconsistencies are possibly due to using different strains suggesting that EBERs are not essential for cell transformation.

1.3.8.8 BamHI-A Rightward Transcripts (BARTs)

BARTs are another type of viral RNA present abundantly in all infected cell types. They are a series of transcripts produced via a complicated series of alternative splicing events and they share a common 3' terminus (Marquitz & Raab-Traub 2012). Twenty-two EBV pre-miRNAs (from which miRNA is cleaved) are encoded in the introns of the BART transcripts (Skalsky & Cullen 2010). The function of BARTs is still to be identified. The detection of BARTs in cells from a healthy donor and EBV-associated tumours suggest that BARTs have a role in virus persistence (Young & Murray 2003). Strikingly, the prototype lap strain of EBV, B95-8, has a deletion of around 11 kb that crosses the entire BART region, and thus lacks 17 miRNAs. Even though, B95-8 is fully able to transform B cells in vitro indicating that BARTs are dispensable for B cell immortalisation.

1.3.8.9 EBV microRNAs

MicroRNAs are small (~21–24 nucleotides in length) regulatory noncoding RNAs expressed by all eukaryotes and also by many, if not all, herpesviruses (Skalsky & Cullen 2015). miRNAs post-transcriptionally regulate gene expression by guiding the RNA-induced silencing complex (RISC) complex, which binds partially or entirely to the complementary sequences on target mRNAs. Depending on the degree of complementarity between the target mRNA and the miRNA, mRNA translation is inhibited, or (where there is very high

complementarity) the mRNA is degraded. Twenty-five EBV pre-miRNAs (from which miRNA is cleaved) has been reported (Skalsky & Cullen 2010). Three pre-miRNA are encoded around the BHRF1 ORF and expressed during latency III only. The BHRF1 transcripts have been shown to inhibit apoptosis and enhance the proliferation during the early phase of infected B cells (Seto et al. 2010). The remaining 22 are encoded in introns of the BART transcripts, in three clusters (figure 1.6)

1.3.9 The diseases of EBV

More that 90% of adults are infected benignly with EBV (Thorley-Lawson 2001). It is generally spread through salivary contact, and the primary infection in children is asymptomatic (or indistinguishable from other childhood diseases), after which the virus persists in resting memory B-cells over the host's life in most cases causing no disease. However, when EBV-infected cells suffer defects in the processes leading to latency, EBV may cause various diseases, most of them cancerous.

Although EBV is a ubiquitous virus, the prevalence of EBV-associated malignancies varies dramatically with geography (Chang et al. 2009). The increased incidence of EBV-positive Burkitt lymphoma in Africa and Indonesia is due to the high level of malaria in these areas. But the factors that affect the incidence of other malignancies remain unknown. Nasopharyngeal carcinoma, for instance, is much more prevalent in Southern China and southeast Asia than elsewhere in the world. The incidence of EBV-associated Hodgkin lymphoma is relatively high in South America and Africa, but intermediate in North America, Europe and China (Chang et al. 2009). However, there is no a clear relation between EBV infection and variable incidence of EBV-associated malignancy in different geographic

regions. Several factors may play a role in the geographically variable incidence of EBV cancers, either solely or with other factors. These might include the environmental differences, culture, diet and/or the genetics of the host. In addition, the genetic variation of EBV may also play a role, and may explain why some EBV-infected individuals develop cancers while most remain cancer free.

1.3.9.1 Infectious mononucleosis

Although primary infection by EBV is asymptomatic in most case, EBV causes acute infectious mononucleosis (IM; also called glandular fever) in some adolescents who were not infected in childhood. The reason why the infection in adolescence causes IM (around 25%) is unknown (Crawford et al. 2006). Since EBV is usually transferred via saliva, IM is more likely to occur after the onset of sexual activity, and the incidence of IM may be attributed to an intense immune response rather than the virus itself (Thorley-lawson 2015), as the initial virus dose could be greater via the oral transmission through kissing than other modes such as shared utensils (Rickinson & Kieff 2007).

IM is characterised by a large number of the atypical lymphocytes (predominantly CD8+ T cells) as a vigorous immune response to the virus (Thorley-lawson 2015). The general symptoms of IM are transient fever, sore throat and general fatigue. During the early stage of IM, EBV-infected memory B cells circulate in blood without proliferating and, therefore, infected-cells remain non-pathogenic (Thorley-Lawson & Gross 2004). However, there is an association between a history of IM and development of Hodgkin lymphoma (Glaser et al. 1997).

1.3.9.2 Hodgkin lymphoma

Hodgkin lymphoma is described as a tumour of germinal centre cells (express latency II programme, table 1.2), which are characterised by Hodgkin Reed Sternberg (HRS) cells. Around 40% of diagnosed HL are EBV-positive (Glaser et al. 1997; Roschewski 2012). People who were diagnosed with IM have an increased risk 3 fold of developing HL (Glaser et al. 1997; Hjalgrim et al. 2007). Thorley-Lawson proposed a scenario for the association EBV with Hodgkin lymphoma and IM. During infectious mononucleosis, a germinal centre B cell may acquire a mutation that blocks cell differentiation. If that cell harbours EBV, it can constitutively expressing LMP1 and LMP2 which provide survival and growth signals that promote tumour growth if the cell were to acquire additional mutations (Thorley-Lawson & Gross 2004).

1.3.9.3 Burkitt lymphoma

Burkitt lymphoma (BL) was first described by Denis Burkitt in children in equatorial Africa (Burkitt 1958), and it was where EBV was found for the first time (Epstein et al. 1965). BL is a highly aggressive lymphoma and expresses the latency I viral transcription programme (table 1.2). There are three distinct classes of BL: endemic Burkitt lymphoma; sporadic Burkitt lymphoma; and AIDS-associated Burkitt lymphoma. The three forms of BL differ in their geographic distribution of incidence and EBV association (Thorley-Lawson & Allday 2008). Endemic BL usually presents as a tumour in the jaw or neck (but can arise anywhere) and occurs in areas of malarial infection such as equatorial Africa and Papua New Guinea. The EBV genome is detected in almost all endemic BL tumours (Thorley-Lawson & Allday 2008). In contrast, only 20-30% of sporadic BL (distributed worldwide) are EBV-positive and typically present as an abdominal tumour. In AIDS-associated BL, around 30 to 40% are EBV-

positive, with a similar site of presentation to sporadic (Hsu & Glaser 2000). While the majority of endemic BLs express the latency I programme (where only EBNA1 is expressed from the Qp promoter), 15% of endemic tumours have a deletion of the EBNA2 gene region and express EBNA1 along with the EBNA3 family and the viral Bcl2 homologue BHRF1 from the Wp promoter, leading to their description as 'Wp-restricted BL' (Kelly et al. 2013). Cell lines established from these type of tumours showed more resistance to inducers of apoptosis in vitro (Kelly et al. 2009). The scenario that can lead to BLs has been proposed by Thorley-Lawson; when a translocation occurs in the c-myc gene (a regulator gene) in an EBV-infected centroblast on its way to becoming a memory cell, Myc activity will keep this cell in a proliferative state and therefore constitutively express only EBNA-1 (Thorley-Lawson & Gross 2004).

1.3.9.4 Post-transplant lymphoproliferative disease (PTLD)

Patients who undergo immunosuppressive therapy after allogeneic transplantation of either solid organs or hematopoietic stem cells are at risk for developing a disease such as post-transplant lymphoproliferative disease (PTLD). this is a heterogeneous collection of B-cell tumours that usually carry EBV expressing the latency III programme (table 1.2). The immune suppression leads to a failure of the T cell compartment to supress the uninhibited growth of EBV-infected cells. However, it is not that simple; otherwise, all patients would develop PTLD. To develop such lymphoma, the infected B cell expressing the latency III programme probably also needs to lose the ability to differentiate to become a resting memory cell. Thus, these cells will keep proliferating the reduced level of T-cell immunity. This mechanism may explain the heterogeneity of PTLD and why it is rare as only one or two out of millions of infected cells grow out of control and develop a tumour, while the vast majority of infected

cells grow under tight regulation even in a weakened immune system (Thorley-Lawson & Gross 2004).

1.3.9.5 Nasopharyngeal carcinoma

Nasopharyngeal carcinoma (NPC) is classified into three histologic types based on the degree of tumour-cell differentiation; squamous (type I), non-keratinizing (type II), and undifferentiated (type III). The most form of NPC associated with EBV is the undifferentiated type UNPC (Young & Murray 2003). The association of EBV with the other two forms (type I and II) is controversial. UNPC incidence is common in China and South-East Asia, and EBV expresses latency II programme (table 1.2). The role of EBV in developing NPC is unknown. The incidence rates are high particularly in Cantonese males, suggesting a genetic predisposition to develop NPC in addition to environmental cofactors that may contribute to the aetiology of NPC, such as dietary components (e.g. nitrosamines in salted fish) (Young & Murray 2003).

1.3.9.6 Gastric cancer (GaC)

Tumours resembling UNPC can arise in the stomach, displaying a similar latency pattern to that of NPC. Around 10% of GaC tumours are EBV positive. As in NPC, the role of EBV in GaC is not clear. However, the incidence of GaC varies geographically (e.g. Japan and Colombia have high rates of GaC), which might be attributed to genetic and environmental factors (Young et al. 2016).

1.3.9.7 Autoimmune diseases.

EBV is a plausible infectious agent for playing a role in the pathophysiology of several autoimmune disorders such as multiple sclerosis (MS), systemic lupus erythematosus (SLE), Sjögren's syndrome and rheumatoid arthritis. The association of EBV with these diseases has been attributed to elevated titers of anti-EBV antibody, impaired T cells response against EBV antigen and high EBV viral load in peripheral blood. The relation between autoimmune diseases and EBV is not clear. However, it is likely that EBV proteins involved in immune evasion and suppression of apoptosis of transformed infected lymphocytes results in loss of immune tolerance and development of autoimmunity (Toussirot & Roudier 2008; Draborg et al. 2013).

1.4 What is a wild-type EBV?

Although EBV is a ubiquitous virus, there is not yet a wild-type EBV sequence that can be used as a background against which to identify mutations and/or polymorphisms associated with disease. The incidence of many EBV-associated malignancies varies with geography (Chang et al. 2009) (see section 1.3.9). The relationship between EBV infection and the development of EBV-associated cancers is not clear. Human genetic variation, environmental factors and co-infection with other infectious agents (such as KSHV or malaria) might contribute to cancer incidence, as could the genetic variation of EBV strains. It is not reasonable to link patterns of EBV genetic variation with certain diseases unless the prevalence of such variants in the normal population has been studied. Also, to verify the relation between viral genotype and phenotype, several experiments that must be conducted in the lab (e.g. phenotypic studies) are hampered by EBV's properties, such as, the difficulty of producing virus, the large size of its genome, and the lack of availability of an animal model system.

There is a general belief that double-stranded DNA viruses (such as EBV) are inherently stable due to the high fidelity of eukaryotic DNA polymerase and error correction steps during DNA replication. However, several reports have demonstrated that EBVs have different levels of diversity, raising the need to understand the nature of EBV diversity and what the wild-type EBV is (Farrell 2015; Palser et al. 2015; Chiara et al. 2016). The vast majority of EBV genome sequencing projects have been conducted on strains isolated from either diseased tissue (cancer biopsies or cancer-derived cell lines) or transformed infected B cells (LCLs).

Both of these two conditions (selection of cancers and producing LCLs) may induce unintentional selective pressure on viral genomes, selecting for a subset of strains with specific properties, or inter-strain recombinant viruses, allowing a new genotype to arise (Renner & Szpara 2017). Therefore, there is not yet a good representative of naturally circulating EBV. Palser et al 2015 sequenced one EBV genome isolated from the saliva of a healthy donor. Thanks to advances in genome sequencing and analysis, it is now possible to sequence more EBV genomes (from diseased samples and healthy donors) to get a better understanding of what constitutes wild-type EBV and what genetic variations are associated with malignancies.

1.5 EBV diversities that change EBV properties.

EBV type 1 and type 2 are defined by differences in the sequences of the EBNA-2, -3A, -3B and - 3C genes (Sample et al. 1990). The difference between type 1 and type 2 in the EBNA2 sequence is around 30% at the nucleotide level and 46% in the protein sequence. The degree of difference in the EBNA3s between type 1 and type 2 is less than in EBNA2. The B95-8 (type 1) and AG876 EBV (type 2) showed nucleotide and amino acid difference levels of 10 and 16%, 12 and 20% and 19 and 28% for the EBNA-3A, -3B and -3C respecively

(Sample et al. 1990). However, there is no a clear association between the diversity of the main viral oncogenes (EBNA-2, -3s and LMP1) and diseases. The observed EBV genomic variations are more ascribed to the geographic origin (Chang et al. 2009; Palser et al. 2015; Tzellos & Farrell 2012).

In vitro, both EBV type 1 and type 2 transform B cells into LCLs, but type 1 EBVs are superior in their ability to transform B cells. Recently, it has been identified one difference in a single amino acid of EBNA2 (Ser in type 1 EBV, Asp in type 2 EBV at position 442) that confers the superior ability of type 1 EBV in transforming B cells into proliferating LCLs (Tzellos et al. 2014). Replacing Asp in type 2 with Ser in the transactivation domain (position 442) changed the transformation efficiency of EBV type 2 to be similar to that observed with type 1 and enhanced the induction of LMP1(Tzellos et al. 2014). The superior ability of type 1 EBV in transforming B cells is probably due to the ability of type 1 EBNA2 to induce transcription of LMP1, which is required for growth transformation by EBV (Lucchesi et al. 2008). In contrast, it has been shown that EBV type 2 has a unique tropism to $CD8^+$ T cells and readily infects T cells in vitro compared to EBV type 1 (Coleman et al. 2015). However, it is not clear whether T cells contribute to the persistence of EBV type 2 in vivo.

Focusing only on type 1 EBV strains, the Delecluse lab reported that the M81 strain (isolated from NPC) has an increased tropism for epithelial cells compared to B95-8 (Tsai et al. 2013). The authors conclude that changes in the sequences and expression levels of the virion glycoproteins (gp350, gp220 and gp110) mediate epitheliotropism of the EBV strain M81. Polymorphisms in the BZLF1 promoter (lytic transactivator protein) in M81 elevated (partially) lytic replication (Tsai et al. 2013). They then conducted a comparative characterisation of more strains: SNU719, GP202 and YCCEL1(isolated from GaC), M81, Akata (BL) and the prototype EBV strain B95-8 (Tsai et al. 2017). B95-8 and Akata infected

B cells with more efficient than M81 and YCCEL1. These differences could not be attributed to binding efficiency as Akata was less efficient in binding to B cells than M81 (equal to YCCEL1) but transformed B cells more efficiently. Reciprocally, M81 and YCCEL1 infected primary epithelial cells more efficiently than Akata or B95-8 (Tsai et al. 2017).

Also M81, YCCEL1 and SNU719 expressed a higher level of LMP1 than other tested strains one-month post-infection and showed, at the same time point, the lowest cell growth rate (Tsai et al. 2017). This would fit with the observation that elevated level of LMP1 expression is toxic to the infected cells (Hammerschmidt et al. 1989). In addition, all strains isolated from GaC, spontaneously initiated lytic activity 18-20 dpi in B cells, expressing BZLF1 (immediate early lytic transactivator) and gp350 (late viral glycoprotein), whereas B95-8 and Akata did not show any sign of lytic replication. This observation supports the idea that inducing lytic cycle is restricted to few strains (see section 1.6.1 and 5.1). These comparative characterisation studies conducted by Tsai et al. 2017 showed different behaviour of different EBV strains, but have not yet attributed these different phenotypes to specific genes with different genotypes.

In addition to natural sequence variation, examples of EBV mutations have been reported in EBV-induced cancers. Probable mutations in EBNA3B have been identified in several EBV-positive lymphoma types (White et al. 2012). Mutation of EBNA3B promotes B cell transformation and a propensity to develop diffuse large B cell lymphoma (DLBL) in a humanised mouse model (White et al. 2012). Defective EBNA3B was also observed in tumour-derived cells from a patient with aggressive post-transplant lymphoproliferative disease (PTLD), where an oligoclonal lymphoma (TRL1-pre) developed into a monoclonal tumour (TRL1-post) (Gottschalk et al. 2001).

As mentioned earlier (Section 1.3.9.3), around 10-20% of BLs contain a deletion of EBNA2 and the C terminal domain of EBNA-LP (Kelly et al. 2002). While this mutation abolishes the ability of EBV to transform resting B cells (Rabson et al. 1982; Woisetschlaeger et al. 1991), these Wp-restricted BLs are more resistant to apoptosis than those carrying an intact EBV genomes (Kelly et al. 2005). This improved survival is likely through the expression of EBNA3A, -3B, and -3C and BHRF1 which are not expressed in latency I BLs (Kelly et al. 2005; Kelly et al. 2009; Kelly et al. 2013).

1.6 Genetic analysis of EBV function

1.6.1 Lytic cycle study *in vitro*

In cells, EBV predominantly establishes latent infection programmes (see section 1.3.4). However, switching to the lytic (or productive) cycle is required for EBV to spread to new cells and individuals (Gorres et al. 2014). It has been suggested that the frequency of EBV lytic cycle reactivation may partially explain why some individuals develop lymphoproliferative disease while others remain cancer free (Bhaduri-McIntosh & Miller 2006; Ma et al. 2011). Switching from latency to lytic cycle involves transcription of *BZLF1 and BRLF1,* which encode the viral transcription factors, BZLF1 (also known as Z, Zta and ZEBRA) and BRLF1 (also known as Rta), respectively. These two proteins initiate an EBV lytic cascade by up-regulating their own promoters and, in concert, the promoters of EBV early lytic genes (Kenney & Mertz 2014). During latency programmes (table 1.2), neither BZLF1 nor BRLF1 is expressed as they are silenced by multiple cellular transcription repressors (Kenney & Mertz 2014). The endogenous stimulation of the lytic cycle is poorly characterised. Furthermore, only a small percentage of latently infected cells activate lytic cycle and produce

progeny viruses. One of the challenging steps in studying EBV properties is persuading EBV to enter the lytic cycle.

In vitro, different cell lines vary dramatically in their abilities to respond to different lytic cycle-inducing agents (Gradoville et al. 2002). In some (both B cells and epithelial cells), the lytic cycle can be triggered by exogenous expression of BZLF1, while BRLF1 expression induces the lytic cycle efficiently in epithelial cells (Zalani et al. 1996). This can be achieved either by transiently transfecting latently infected cells with a plasmid expressing BZLF1 (McDonald et al. 2010), or transiently infecting cells with adenovirus vectors expressing BZLF1 (Westphal et al. 1999). Alternatively, producer cell lines can be produced by stably incorporating an inducible or conditional system for making BZLF1, such as was done by fusing BZLF1 to a truncated estrogen receptor (ER), to make BZLF1 activity dependent on the presence of 4-hydroxytamoxifen (4HT) (Johannsen et al. 2004). However, the efficiency of lytic induction relies on how permissive EBV-infected cells are for lytic induction.

Several chemical agents can also induce lytic reactivation. These agents include the phorbol esters such as 12-0-tetradecanoyl phorbol-13-acetate (TPA), which is a protein kinase C agonist, and sodium butyrate (NaB) which is histone deacetylase (HDAC) inhibitor. However, cell lines differ in their responses to different agents (Gradoville et al. 2002). For instance, B95-8 responded to TPA but not NaB, while HH514-16 (sub-clone of P3HR1, BL) was activated by NaB but not TPA (Gradoville et al. 2002). TPA induces the lytic cycle in some cell lines (probably by activating protein kinase C (PKC) pathway), whereas NaB reactivates EBV in some cell lines by increasing histone acetylation, which in turn results in chromatin remodeling allowing binding of transcription factors to the BZLF1 promoter.

Another physiological process that can stimulate of the lytic cycle is cross-linking the B-cell receptor (BCR) on the surface of infected cells with anti-immunoglobulin antibodies. This process mimics IgG-mediated activation of BCR and leads to activation of Zp (BZLF1 promoter) and the EBV lytic cycle (Takada & Ono 1989). LMP2A, which mimics BCR, prevents lytic cycle reactivation by interfering with BCR signaling (Bryant & Farrell 2002). Akata BL cells are responsive to cross-linking BCR with up to 20% of cells reactivate the lytic cycle, and they do not express LMP1 nor LMP2A (Bryant & Farrell 2002).

1.6.2 Production of recombinant viruses in eukaryotic cells

Viral genes can be studied by introducing mutations into the genome and then analysing the phenotype. Because of the large size of herpesvirus genomes, conventional restriction cloning techniques are difficult to apply. Therefore, homologous recombination has been used to precisely alter herpesvirus genomes genetically. Homologous recombination involves introducing a targeting vector carrying the desired mutation flanked by regions that allow the precise replacement of the wild-type sequence with the mutant sequence. The efficiency of homologous recombination in eukaryotic cells relies on viral replication. Unlike α- and β-herpesviruses, ɣ-herpesviruses are not highly replicative in cells. Besides, recombinant viruses are usually produced first as a mixtures with the wild-type ones, and purification of the EBV mutant virus stocks is much more difficult because EBV-infected B cells (that can be used for rescuing based on transformation competence) do not produce progeny viruses, and EBV cannot be plaque-purified (Delecluse & Hammerschmidt 2000).

Therefore, the immortalising property of EBV has been utilised to isolate recombinant viruses. The type 1 EBNA2 gene (including the adjacent EBNA-LP exons Y1 and Y2) was

reintroduced into the Wp-restricted P3HR1 genome. Recombining EBNA2 into P3HR1 genomes restored their immortalising ability. Hence, a plasmid carrying the IR1/EBNA2 genome region is delivered – along with a mutant allele of a viral gene of interest – into cells harbouring P3HR1. Only successfully recombined viruses will be able to immortalise resting primary B cells (Delecluse & Hammerschmidt 2000). Virus supernatant will contain both recombinant viruses as well as the parental P3HR1. Limiting dilution enables the isolation of EBV-immortalised cell clones that harbour only the recombinant virus. However, the virus cannot be subsequently induced to enter the lytic cycle in these cells, and so cannot be used to produce recombinant virus.

This technique is also restricted to genes that are not essential for the transformation potential of the virus (Tomkinson & Kieff 1992)(Tomkinson et al. 1993). An alternative approach relies on introducing a selectable marker such as an antibiotic-resistance gene via homologous recombination along with the desired mutation, and then growing infected cells in media with the corresponding antibiotic. All strategies described above take place in eukaryotic cells where the genetic manipulation is much harder than prokaryotic cells, although the advent of CRISPR technologies is beginning to change this (section 1.6.4). Additionally, introducing loss of function mutations into elements required for the lytic cycle, or for the growth/survival of the host cell line, is not possible in eukaryotic cells, since a complete lytic cycle is a prerequisite for isolating the recombinant genome from wild-type ones.

1.6.3 Manipulating EBV genomes in bacteria

As a result of the limitations of working in eukaryotic cells, virologists working on various diverse viruses have sought ways to clone viral genomes in bacteria in ways that allow

them to be used to produce virus in eukaryotic cells. Bacterial artificial chromosome (BAC) technology was utilised to construct a complete herpesvirus genomes in *E. coli*. A prerequisite of this strategy is the introduction of the F-plasmid replicon into the complete viral genome via homologous recombination and colonies the resultant as a whole in *E. coli*. B95-8 was the first EBV strain to be cloned as BAC clone. The F-plasmid was inserted within the deleted region in B95-8 (Delecluse et al. 1998). EBV-BACs can be easily modified, and foreign sequences (e.g. antibiotic cassette) can be introduced. There are two methods to introduce a mutation into EBV-BAC. One strategy uses a linearized targeting vector carrying an antibiotic resistance gene flanked by Flp-recombinase target (FRT) sites next to the mutant gene. In this case, the gene of interest (in the targeting vector) will be inserted in the position of the wild-type gene. The antibiotic cassette will be excised in transfected cells by transient transfection of the Flp recombinase (reviewed in Feederle et al. 2010).

The second approach called 'chromosomal building' uses a circular targeting vector carrying the mutant gene of interest flanked by homologous sequences, both positive and negative selection markers, which typically includes a temperature-sensitive origin of replication, and recombinase gene (RecA). Upon production of RecA, the recombination process initiates via one region of homology resulting in a co-integrant carrying both the wild-type and the mutated sequence, each with an identical set of homology flanks. Growing the bacterial cells on a media with chloramphenicol (from the BAC) and ampicillin or kanamycin (whichever is present in the target vector) allows the selection of co-integrants. To remove non-recombined targeting vectors, cells are cultured at the temperature that is non-permissive for the temperature-sensitive origin of replication (reviewed in Feederle et al. 2010). After genetic modification in *E. coli*, resultant EBV-BAC can be isolated and used to transfect EBV negative cells (i.e. HEK 293) that support the lytic cycle to establish EBV-producing cells.

Successful homologous recombination in eukaryotic cells relies on active lytic replication which is not available in EBV. This may explain the fact that at the time of writing this section, there are only 7 EBV BACs from 5 strains that have been published. All these EBV-BACs are from mutated lab strains (B95-8), or patients with EBV-related diseases, and there is no EBV-BAC representing circulating wild-type EBVs, and no example of a type 2 EBV-BAC. The two B95-8 BACs differ in the position of the F-factor, either in the deleted region in B95-8 (Delecluse et al. 1998) or the major internal repeat region (Chen et al. 2005). In Akata-BAC, the F-factor was inserted in the BXLF1 ORF (encoding thymidine kinase) which was shown to be unnecessary in vitro (Kanda et al. 2004). However, the selectable marker on the BAC was poorly expressed from this locus, necessitating the introduction of a second selectable marker near the site of the B95-8 deletion, which is more transcriptionally active (Kanda et al. 2004). Tsai et al. managed to clone the more epitheliotrophic NPC-derived M81 strain and GaC-derived YCCEL1 as BACs by inserting the F-factor within the terminal repeat region (Tsai et al. 2013; Tsai et al. 2017). Inserting the F-factor within the TR region will not, theoretically, interrupt coding genes and upon the lytic reactivation, the F-factor will be excised as a consequence of recombination between the two flanking terminal repeats (Feederle et al. 2010). Recently, Kanda et al. have reported the cloning of two GC-derived strains YCCEL1 and SNU719 using CRISPR to facilitate the homologous recombination between the EBV genome and the donor BAC, and have inserted the F-factor between the BVRF1 and BVRF2 ORFs, very close to the location of the B95-8 deletion (Kanda, Furuse, Oshitani, et al. 2016).

Using EBV-BACs allows for reverse genetic analysis, by which functions of viral genes can be delineated. Mutations can be introduced to test the functions of genes, and revertant viruses are used as controls to verify that any observed phenotype is due to the specific mutation or deleted gene. Knockouts of all of the EBV latency genes have been studied by

generating knockout gene EBV-BAC, including EBNA1 (Humme et al. 2003), EBNA2 (Kelly et al. 2005), EBNA-LP (Szymula et al. 2018), EBNA-3A (Hertle et al. 2009), EBNA-3B and -3C (Anderton et al. 2008), LMP1 (Dirmeier et al. 2003), LMP2 (Mancao & Hammerschmidt 2011), and EBERs (Gregorovic et al. 2011)(Wu et al. 2007). More sophisticated mutations have also been made, including conditional proteins (by estrogen receptor fusion) (Maruo et al. 2006) or mutants of specific binding sites, such as the mutants of the binding site for CtBP in EBNA-3A and -3C (Skalska et al. 2010) and the single amino acid change restoring transformation of a EBNA2 type 2 mutant in a type 1 B95-8 BAC (Tzellos et al. 2014).

1.6.4 Using CRISPR-cas9 to support homologous recombination

Clustered regularly interspaced short palendromic repeat (CRISPR) is a naturally occurring, adaptive microbial immune system that uses RNA-guided nucleases to cleave foreign genetic elements. Three types of CRISPR-Cas system have been functionally identified in different microbial species (types I, II and III). CRISPR types I and III mediate the recognition and degradation of target DNA sequence, while CRISPR type II recognises but only cleaves the target sequence. CRISPR type II systems consists of endonuclease protein called Cas9 and guide RNAs whose sequence includes 20 nucleotides complementary to the target region, and a constant sequence that binds to the nuclease. The target sequence of 20-bp must end in a protospacer-adjacent motif (PAM). For the CRISPR/cas9 system that is most widely used in biotechnology applications, the PAM is “NGG”, a common motif. The Cas9 is guided by the sgRNA to the target sequence where the sgRNA pairs to the target region via the complementary sequence allowing Cas9 to cleave the DNA, leaving a double-stranded break (DSB). After cleavage, the double-stranded break can be repaired either by the error-prone non-homologous end joining (NHEJ) – commonly used to make gene knockouts, based on

frame-shifting the target gene – or, in the presence of a homologous template, the more high-fidelity homology-directed repair (Ran et al. 2013).

In addition to its application to editing cellular genomes, CRISPR-Cas9 has been used to edit virus genomes for reverse genetic analysis. For instance, CRISPR has been used to generate several human epithelial cell lines does not express BART by targeting sites flanking the BART promoter (Yuen et al. 2015). More therapeutically, in an attempt to eliminate EBV genome from latently infected cells, CRISPR system was used in EBV Raji and Akata strains to target several regions including EBNA1 and ori-P (Wang & Quake 2014; van Diemen et al. 2016). Both studies show a substantial reduction in viral load.

In cloning, using CRISPR to introduce the DSB was designed to enhance the homologous recombination of the BAC element into the EBV genome. Using CRISPR in the same way that I set out to do (see section 5.2), Kanda et al. (2016) utilized CRISPR to facilitate the cloning of two strains YCCEL1 and SNU719 from gastric cancer cell lines.

1.6.5 Methods of DNA genome sequencing.

The first published complete EBV genome sequence was of B95-8 in 1984 (Baer et al., 1984). At the time, it was the largest contiguous DNA sequence known and represented 10% of the total EMBL nucleotide sequence library (Tzellos & Farrell, 2012). Remarkable advances in next-generation sequencing make genome sequencing far quicker and less expensive than ever before. It will allow sequencing hundreds of EBV genomes, which in turn furthers our understanding of the nature of EBV genetic variation and its association with malignancies.

One of the challenges for virus genome sequencing has been getting adequate amounts of virus DNA from viruses that (like EBV) cannot be readily purified away from their host cells. However, enriching the viral sequences using oligonucleotide enrichment methods has enabled capturing of EBV from saliva and blood with no need for culturing (Palser et al. 2015; Weiss et al. 2017). Oligonucleotide enrichment uses customised probes based on known genome sequences that can hybridise specifically with small amounts of target genomes in any mixed sample. These hybridized fragments are then isolated and prepared for sequencing.

1.6.5.1 First generation (Sanger) sequencing

The original concept underlying the first DNA sequencing technology, which is known as Sanger sequencing, after Dr Fred Sanger who developed it in 1977, relies on sequencing by synthesis, using a "chain-termination" or dideoxy technique (Sanger et al. 1977). In this approach, the sample will mix template DNA with a primer, DNA polymerase, and a mixture of the four 2'-deoxynucleotide triphosphates (dNTPs) and lower concentration of either a single radioactively labelled 2',3'-dideoxynucleotide triphosphate (ddNTP), or all four ddNTPs, each labelled a different fluorescent dye. ddNTPs lack the 3' hydroxyl group that is required for covalent bonding to the next dNTP's 5' phosphate. During the sequencing cycle, a standard PCR reaction takes place, and the elongation process carries out. The extension of the new DNA strand terminates every time the synthesised DNA incorporates the corresponding ddNTP. Strands with different lengths are then separated on a gel (within capillaries on current machines), and the terminal fluorescent-ddNTP is identified by its fluorescence (Heather & Chain 2016). The reliable read length of modern Sanger sequencing is typically around 1kb in length, but was nearer 500 bases when this technology used by the Human Genome Project and Craig Venter's company "Celera" to generate the first complete

draft of the human genome. While these public and private project used the same seqeuncing technology, they differed in their approach to assembling the human genome sequence. The Human Genome Project used a "clone by clone" method in which large DNA fragments (typically BACs and the smaller P1 artificial chromosomes – PACs) of known position in the genome were shredded into small fragments, sequenced and assembled independently. Then, whole genome sequence data were assembled based on sequence overlaps between the BACs and PACs. Celera used a "whole genome shotgun assembly" method in which the whole genome is shredded into small fragments and sequenced. These much shorter sequences were then computationally put back together based on sequence overlaps.

Sanger sequencing is still the choice for verifying specific sequences (e.g. sequencing PCR products and plasmids). First-generation sequencing provided references for many viruses including EBV strains B95-8 (Baer et al. 1984) and AG876 (Dolan et al. 2006). However, large sequencing projects have shifted to next generation sequencing, which uses the shotgun-based approach for genome assembly pioneered by Craig Venter, but with new technologies for producing sequence.

1.6.5.2 Next generation sequencing

The core chemistry of next generation sequencing (NGS) platforms (also known as second generation sequencing) varies between competing platforms. It is built on DNA extension by either DNA polymerase or DNA ligase. There are four sequencing platforms representing second-generation sequencing; Roche 454 uses pyrosequencing technique, Illumina and Helicos both use a polymerase-based reversible dye-terminator technique, and ABI/SoLiD uses sequencing by ligation of single fluorescent nucleotides. Unlike Sanger sequencing, NGS platforms do not require a cloning step for sample preparation. Rather they

independently determine the sequence of many DNA molecules in a mixture of many different sequences. The DNA sample is used to make a library of DNA fragments with adapters (synthetic DNA) attached covalently to each fragment end. The adapters are to amplify the DNA fragments during specific steps of the sequencing process, and can also be used to add a barcode to allow multiple independent samples to be combined in a single sequencing run. NGS platforms generate a massive amount of sequencing data compared to Sanger sequencing, but all NGS platforms produce shorter read lengths than does Sanger sequencing. Illumina platforms are the most widely used in sequencing EBV genomes (Palser et al. 2015; Kwok et al. 2014; Weiss et al. 2017) due to the high output and the relative low cost compared to other platforms (costs of all available market sequencing platforms are reviewed in Goodwin et al. 2016).

- **Illumina sequencing**

The DNA molecule is fragmented (usually by sonication). DNA fragments are then repaired (by polymerase and exonuclease) to be blunt and phosphorylated, adding an adenosine overhang. The fragments are ligated to sequencing adapters, and typically amplified by PCR. Template molecules are attached to a solid surface and then amplified to generate clonal clusters on the surface, ready for sequencing. The surface is a silica slide consisting of eight lanes. This slide enables running of eight different samples at the same time. To attach and amplify the template, the DNA library is denatured and introduced into the silica slide where forward and reverse strands bind covalently to primers attached the slide surface. Strands bound the primers to serve as templates and 3'-extended. Free ends of the newly synthesised templates bind to adjacent primers to form U-shaped bridges. The U-shaped template form the newly bound primer (3'-extension). The resultant dsDNA is denatured, and the free end

hybridises to primers on the slide surface for the new extension. This process is repeated many times (~ 35 times) to create a cluster of around 2000 molecules. Again, the clusters are denatured, and the free 3' ends are blocked to prevent non-specific primer binding. Sequencing primers bind to the free ends of the DNA strands. Sequencing starts by adding the four nucleotides each labelled with different fluorophores and carrying a 3'-reversible terminator. Unlike pyrosequencing methods, the DNA template can only be extended by one nucleotide at a time. After recording the fluorescence emitted from the incorporated nucleotide, the fluorophore and the 3'-reversible terminators are chemically removed, allowing DNA extension in the next cycle. There are several versions of HiSeq platform (Shendure & Ji 2008). The maximum read length of Illumina HiSeq 2000 (the source of sequence data used in this thesis) is 2X100bp, and the output data is up to 600 Gb/run with an accuracy of 98% (Liu et al. 2012). The accuracy of Illumina decreases with increasing nucleotide addition steps. Also, incomplete fluorescent label cleavage creates noise signals in the following cycle from residual fluorescence resulting in errors.

- **Paired-end reads in Illumina sequencing.**

Paired-end reads is a technique that generates two reads by sequencing the same fragment form the opposite ends. It provides informative data to enhance the confidence in mapping reads. The increased length of the region covered by paired-end reads and the large amount of data output facilitate *de novo* assembly, detection of inversions, indels and directional transcriptomics.

After the forward strand has been sequenced (as described above), the newly synthesised strands are denatured, and the free 3' ends are unblocked. Bridge amplification (a process in which each single-stranded molecule binds at both ends to the oligo primers on the

silica slide) takes place creating dsDNA cluster. The forward strands are, then, removed by making a cleavage at the base of the forward primer (incorporated to the forward strand) leaving only the newly constructed reverse strands as ssDNA. The reverse strands are sequenced as mentioned above to generate paired-end data.

1.6.5.3 Third Generation (long read) Sequencing technologies

The main features of third generation sequencing technology are using new and different chemistry and require less operation time consumption (Hui 2014). There are three leading third-generation sequencing platforms; PacBio (adopts real-time single molecule sequencing methods), Ion Torrent (uses changes in pH for monitoring and detecting sequence), and Nanopore (detects sequence by monitoring the transition of nucleotides through a small channel).

- **Pacific Biosciences (PacBio)**

PacBio combines the molecular biology with nanotechnology and uses a very sensitive fluorescence detector. It is the first technology developed for long-read single molecule sequencing. DNA is sequenced several times, which eliminates the associated sequence bias and errors. The instrument uses a chamber made of silicon composed of thousands of individual picolitre wells with a transparent surface. Each well has one engineered polymerase molecule fixed to the bottom. The polymerase has a decreased rate of polymerisation (so the incorporation of dNTPs is not too fast to measure) and the ability to incorporate phospholinked nucleotides labelled with fluorophores, which are added sequentially. By having a fixed location of incorporation owing to the fixed polymerase, the detection system only detects the

single incorporated molecule, without seeing the fluorescent nucleotides that are floating in solution.

To produce the sequencing library, the genomic DNA molecule is fragmented to the desired size (which can be multi-kb size) followed by end repair and adapter ligation. The main feature of PacBio adapters is that they have a hairpin structure (SMRT loop adaptors), thus topologically, the double stranded linear DNA fragment will become a circular single-stranded DNA molecule after ligation and binds to the polymerase to form an active polymerisation complex. Once the fluorescent nucleotides are introduced into the chamber, the instrument optics continuously survey the incorporating nucleotides in real time. The average read length is more than 14 Kb (very useful for de novo assemblies), and the output data are up to 1Gb. However, PacBio has a relatively high error rate of 10-15% per nucleotide, most of which are insertion/deletion errors. However, the errors occur randomly, and their frequency increases towards the end of reads (like Illumina) (Reuter et al. 2015), but by sequencing both strands of the circle several times, this error rate can be substantially reduced.

- **Ion Torrent**

Ion Torrent has been commercialised as the personal genomic machine. Ion Torrent does not use optical sensing; it uses semiconductor chips that predict the incorporated nucleotide(s) by measuring pH changes induced by the release of H^+ as each dNTP is incorporated. Since Ion Torrent does not use labelled dNTPs, there is no noise generated from fluorescence, allowing longer read lengths. However, Ion Torrent struggles in sequencing homopolymers (the same issue that Roche 454 struggles with). The read length is around 400

bp, and data output ranges from 50Mb to 15 Gb (depending on the type of chip that is used) (Goodwin et al. 2016).

- **Nanopore**

Nanopore-based sequencing identifies the sequence of a native ssDNA molecule. The sequence relies on the transition of DNA through a pore. Unlike the techniques described above, nanopore sequencing does not monitor the incorporation of hybridised nucleotides guided by a DNA fragment, but directly measures the primary DNA molecule. The first commercially produced nanopore-based sequencing technology is minION made by Oxford Nanopore. It uses a flow cell containing hundreds of micro-wells, each containing a lipid bilayer perforated by nanopores. As the DNA fragment translocate through the pore (guided by ligated adapters and motor proteins) a characteristic change in voltage through the pore is measured. The read length of Nanopore (MinION) can approach 200Kb with output data up to 1.g Gb (Goodwin et al. 2016). However, it has a high error rate of 12%, and it has been reported that Nanopore has high run failure rate (Reuter et al. 2015).

1.7 Repeat regions in EBV

The EBV genome structure is characterised by the presence of terminal repeats and additional unrelated internal repeated sequences (figure 1.1) (Davison 2007). The EBV genome has several internal regions that contain tandem repeat sequences. Among the largest of these are four internal repeats (IR1 through IR4) and the family of repeats (FR) region of

the latent replication origin oriP, in addition to the terminal repeats (TR) on both ends of the linear genome. IR1 and TR have large repeat units (~3071 bp and ~538 bp respectively) while other repeat regions have much smaller units. For instance, IR2 and IR4 units are 123 bp and 103 bp (Cheung & Kieff 1982), while IR3 units are 6-9 bp composed of different nucleotides triplets (Heller et al. 1982). The copy number of IR1 and TR varies between isolates, which accounts for the genome size variation (Lacoste et al. 2010).

1.7.1 The family of repeats (FR)

In latently EBV-infected cells, EBV genomes are maintained as episomes mediated by EBNA1 binding to oriP (Frappier 2015). In the B95-8 strain, the oriP region contains two functional elements separated by around 1 kb; the dyad symmetry (DS) element and FR. The DS includes four copies of the EBNA1 DNA binding site, while FR consists of 20-30 imperfect tandem copies of a 30 bp repeat unit, each of which contains a 16-bp palindromic sequence that works as EBNA-1 binding sites (Richard F Ambinder et al. 1990; Fruscalzo et al. 2001). The DS is essential for plasmid replication in the presence of EBNA-1. The FR maintains indirectly the oriP containing plasmid by tethering it to cellular chromosomes via EBNA1, so minimising plasmid loss from nuclei during cell dividing (Frappier 2015). EBNA1 binding to the FR enhances the activity of Cp, which in turn drive the expression of viral proteins (e.g. EBNAs) in LCLs. Number of units in FRs varies between strains (from 23 to 30), and they are stably maintained during long-term passaging (Fruscalzo et al. 2001). Seven to nine copy of the FR are sufficient for full ori-P activity (Wysokenski & Yates 1989). However, it has been reported that increasing the number of FR repeat units does not necessarily enhance activity of ori-P (Ali et al. 2009).

1.7.2 Internal repeat 1 (IR1)

IR1 (also known as the major internal repeat and the W repeats after the BamHI W fragment it contains) is the largest repeat region in the EBV genome. In B95-8, IR1 repeat units are 3072 bp (66.9% GC-content), and their number varies between isolates, commonly from 5.6 to 8.6 units (Tierney et al. 2011; Palser et al. 2015). The last IR1 unit is ideally a partial repeat, around 1852-bp in length (Cheung & Kieff 1982)(Baer et al. 1984). In the prototype EBV sequence (NC_007605), which has been given 7.6 repeat units, IR1 starts at position 12001 and ends at 35355 (de Jesus et al. 2003).

IR1 contains features that are biologically important to EBV. First, each repeat unit contains a copy of Wp, the first promoter to be used following the initial infection of resting B cells that drives the expression of EBNA-LP and EBNA-2, which are important for cell transformation (Woisetschlaeger et al. 1990) (see section 1.3.7.1). Second, each copy also contains exons W1 and W2, which together encode the repeat domain of EBNA-LP (figure 1.9). Viruses with 1.6 or fewer copies of IR1 units have lost their ability to transform B cells and 4.6 repeats are the minimum number for optimal transformation efficiency (Tierney et al. 2011). This efficiency in transformation was attributed to the transcription level of EBNA-2 and EBNA-LP that is affected by the number of Wp promoters.

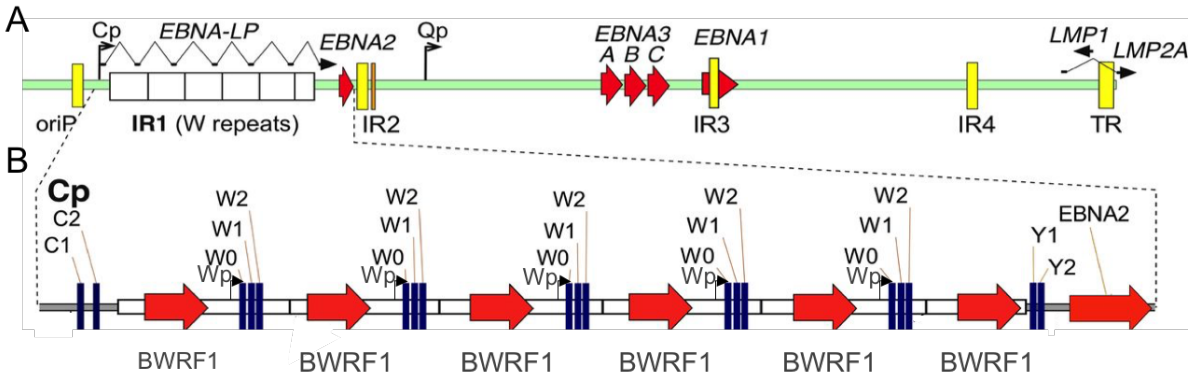


Figure 1.9 Schematic illustration of the IR1 position and components.
(A) Schematic representing the structure of the linear EBV genome and position of internal repeat 1, IR1, (white boxes). Yellow boxes represent the smaller internal repeats (IR2 to IR4), family repeats in the origin of replication (ori-P) and the terminal repeats (yellow boxes). EBNA genes are shown as red arrows. Latency promoter Cp, Qp, LMP1 and LMP2A are shown as bent arrowheads. (**B**) schematic of IR1 units (presented here as 5.6 copies) showing the components of each repeat unit. W exons are in navy blue lines. the ORF of unknown function BWRF1 is presented as red arrow. The Wp promoters are presented as bent arrowheads. (Figure produced by Dr Rob White).

As well as encoding EBNA-LP, all EBNA transcripts (derived from either Wp or Cp) contain repeats of the W1 and W2 exons (Rogers et al. 1990). The transcripts are generated by a complex and not fully characterised pattern of splicing. Exon W1 has two 5' splice acceptors: W1 and W1', the latter of which makes an exon that is 5 nucleotides shorter (figure 1.10). Transcripts derived from the first Wp exon (W0) or the first Cp exons (C1 then C2) can join to either splice acceptor W1 or W1'. Splicing exon W0 or C2 to W1' generates the start codon (AUG) that allows EBNA-LP translation. Subsequent W2 exons splice to W1, making additional arrays of exons W1/W2 from each downstream IR1 unit followed by exons Y1 and Y2 (located downstream of IR1). On the other hand, splicing to exon W1 does not create this AUG codon, enabling EBNA-2 translation from the next start codon (Rogers et al. 1990). To generate downstream EBNAs, exon Y2 splices to either Y3 (a splice donor within EBNA-2 exon YH) or directly to the exon U to generate EBNA3s and EBNA1 (Allday et al. 2015).

Since the U exon contains an internal ribosome entry site (IRES) that facilitates EBNA3/1 translation, these transcripts may be able to be translated into both EBNA-LP and the downstream EBNA.

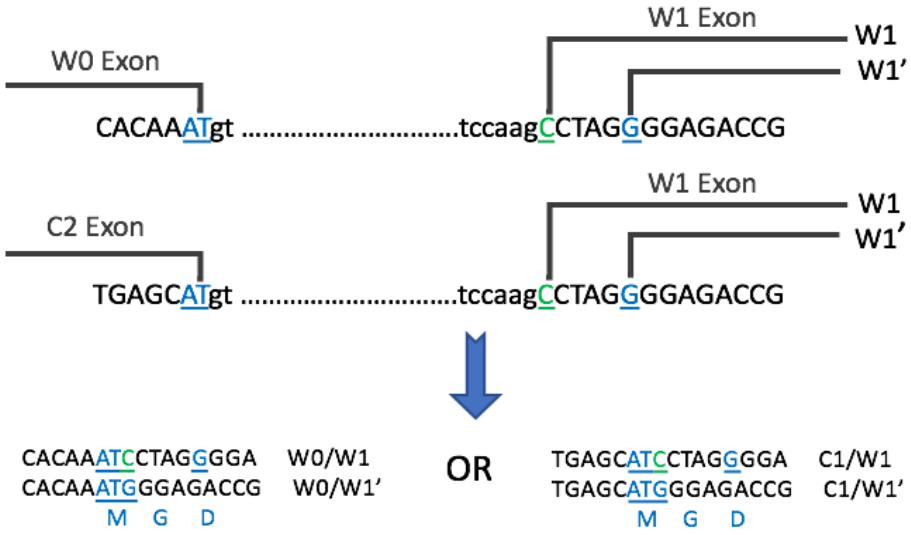


Figure 1.10 Alternative splicing that organizes the EBNA-LP and EBNA-2 gene transcripts.
Splicing from W0 exon (derived from Wp) or C2 exon (derived from Cp) to the shorter W1' exon generates an AUG initiation codon (in blue) for EBNA-LP, while splicing to the longer W1 exon does not create an AUG enabling EBNA-2 translation from the next start codon. (Modified from Kempkes & Ling 2015).

As well as the coding potential in IR1, there is some evidence that IR1 gives rise to important non-coding RNAs. Exons W1 and W2 are separated by short (81 nt) and long (2791 nt) introns. The short and long introns have been identified as conserved and structurally stable namely; stable intronic sequence RNA-1 (sisRNA-1) and stable intronic sequence RNA-2 (sisRNA-2) respectively (Moss & Steitz 2013; Moss et al. 2014).

sisRNA-1 and 2 were detected by RNA-seq, and qPCR suggests that sisRNA-1 (which could also be detected by northern blotting) is the most abundant viral non-coding RNA after

the EBERs (Moss & Steitz 2013). The sisRNA-1 is deduced to be conserved between type 1 and type 2 EBVs, and considerable homology extends to include other lymphocryptoviruses (Moss & Steitz 2013). SisRNA-2 is less clearly defined, but it contains a thermodynamically stable RNA secondary structure, that includes a large (586-nt) hairpin, which is structurally conserved (but not in sequence) in diverse lymphocryptoviruses (Moss et al. 2014). The sisRNA-2 intron also contains a large uncharacterized open reading frame (BWRF1) with no standard (AUG) start codon (Baer et al. 1984), although the hairpin sits near the start of the ORF (in B95-8). There are no biological functions that have yet been ascribed to sisRNAs, although recent work in our lab has found that a point mutation in sisRNA-1 can lead to a transformation defect (Szymula et al. 2018).

Within IR1 units there are additional direct repeats (Cheung & Kieff 1982). For clarity, positions described herein are those of the first IR1 in the prototype EBV genome NC_007605. The 35-bp sequence "CCAGGCCAGCCGGAGGGACCCCGGCAGCCCGGGCG" is found to start at 13012 and 13306. The second repeated sequence is the 16-bp "CCCGGGCTTCAGGCCC" exists at 14986 and 15025 (32-bp before the end of the intact IR1 unit). Also, there are two 14-bp repeats; one repeat "TCCCCGGGCTTCAG" exists at 12999 and 13291. The other 14-bp "GCCTCCCCGGGTCC" is at 13073 (largely overlapping one of the 16bp repeats) and at 14984. The 14-bp at 12999 has a 1 nt overlap with the 35-bp at 13012 while its counterpart at 13291 ends one 1bp before the 35 bp seqeunce at 13291. Cheung & Kieff suggested that the 14- and 35-bp could be considered as imperfect direct repeats of 51-bp (the repeat starting at 12999 having lost a CA dinucleotide relative to that starting at 13291). As for BWRF1, the importance of these repeated sequences is not known, as no role has been ascribed to them.

1.7.3 Terminal repeats (TR)

Terminal repeat (TR) units are approximately 538-bp in size and GC-rich regions, and the number of units at each end of the linear EBV genome can vary (Raab-Traub & Flynn 1986; Zimmermann & Hammerschmidt 1995). TR is essential for circularization and packaging of the EBV virion DNA (Hammerschmidt & Sugden 1989). Lytic replication, activated spontaneously or with exogenous stimuli, generates long head-to-tail concatemers from the EBV DNA, which are then cleaved and packed into viral capsids. The cleavage occurs by unknown mechanism somewhere within the TRs via unidentified protein or complex of proteins called the terminase (Zimmermann & Hammerschmidt 1995).

A detailed study of the TRs of the linear virion DNA of B95-8 has been conducted by Julie Zimmermann and Wolfgang Hammerschmidt (Zimmermann & Hammerschmidt 1995). Comparing the sequence of the cloned termini with the termini of native virion DNA showed that the right terminus has a 3' overhang of one nucleotide "G". This overhang was overlooked in the cloned EBV genome due to DNA preparation during the cloning. Comparing the sequences of linear termini with the circular TR revealed alterations introduced near the termini. Each TR unit starts with an 11-bp element "GGGGCATGGGG", and after the last TR unit there is a unique 9bp motif ("GTGTTGGCG") followed by another 11bp motif. Curiously, this 9bp element – found only at the right end of TR – is duplicated with the 11bp motif onto the right terminus of the linear genome, such that the linear genome has the 11-bp element at both ends (Figure 1.11). This sequence redundancy does not exist in other herpesvirus TRs (such as varicella-zoster virus, herpes simplex virus and human cytomegalovirus) (Zimmermann & Hammerschmidt 1995).

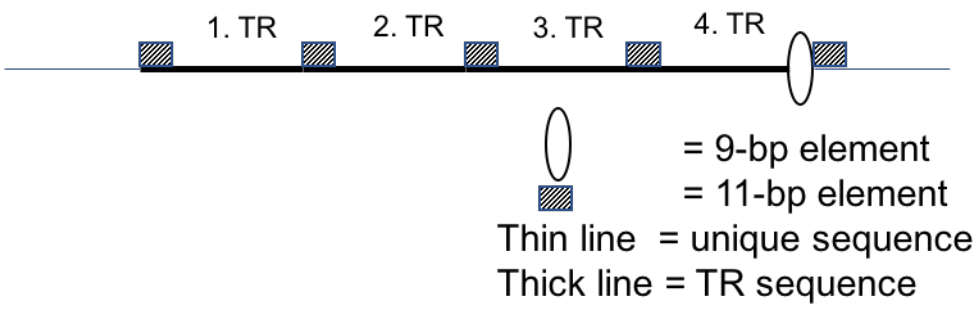


Figure 1.11 Terminal repeat structure.
A schematic shows the EBV termini with four TR units. The 11-bp elements exists in every TR units (black boxes) while the 9-bp elements with the 11-bp elements exist once on the right termini outside the TR region. (Modified from Zimmermann & Hammerschmidt 1995).

Using a transient packaging assay (4 days post lytic induction) showed that one TR is essential and sufficient for packaging while the orientation of TRs relative to ori-lyt has no effect on the packaging process (Zimmermann & Hammerschmidt 1995). The minimum cis-acting signal for cleavage and packaging is located within a region of 159-bp including the 11-bp element and the EBV equivalents of packaging signals; pac1 and pac2 which have homologues near the termini of diverse herpesviruses (Deiss et al. 1986). pac1 is just to the left of the right terminus (6Cs – AGGAAAGA – 5Cs) while pac2 is to the right of the left terminus (AAAAA ... GCGGCCC) (Zimmermann & Hammerschmidt 1995).

The number of TR units in the EBV episomes does not change in non-permissive cells for the lytic cycle. Thus, EBV episomes in daughter infected-cells will usually have the same number of repeat units as the parent cells. The number of TR units, therefore, is used as a marker to determine whether latently-infected cells are from the same progenitor or from multiple ones (Raab-Traub & Flynn 1986).

It has been reported that transcription of LMP1 in epithelial cells is initiated from promoter L1-TR in the first TR unit (on the left terminus in the EBV episome) (Sadler & Raab-Traub 1995). However, LMP1 expression is correlated inversely with TR copy number (Sun et al. 1994). Similarly, there is a reverse relationship between TR copy number and LMP2A transcripts which are transcribed across the TR region (Moody et al. 2003).

1.8 Aims of the thesis

Many efforts have illuminated the importance of studying EBV variation/mutations (genotype) as candidates for EBV-induced diseases. Ensuring the accuracy of the sequencing process will allow the reliable identification of mutations in virus genes against the background of sequence diversity, and the identification of genuinely wild-type viruses as candidates for study *in vivo* and *in vitro* model systems.

Most, if not all, of the EBV genome studies have ignored studying variation in the larger repeat regions, internal repeat 1 (IR1) and terminal repeats (TR), because of ambiguities that these repetitive sequences cause in sequence alignment and analysis. This strategy of ignorance may result in missing important biological features of these regions.

Also, current EBV gene function studies use EBV-BACs isolated from cancer-derived strains, which may carry mutations, and perhaps also more subtle, uncharacterised ones. Although revertants of designed BAC mutants can be used as a control to link a phenotype with specific genotype, it is still unclear to what extent variants and deletions intrinsic to the various strains may contribute to observed phenotypes.

There is a need for getting a better understanding of the nature of EBV diversity and clarifying how sequence variation alters the biology of EBV. Therefore, the aims of my thesis are:

1. To develop strategies to reliably sequence the repetitive regions of EBV, with a view to confidently identify non-mutant EBVs.
2. To develop a strategy to clone these non-mutant EBV genomes. Thus, they can be used to produce virus for phenotyping assays and genetic manipulation.

2 Material and methods

2.1 Sequence data analysis

2.1.1 Sequence data source

Sequence data used in this thesis was generated previously in Palser et al. 2015. Data are available online at the European Nucleotide archive (http://www.ebi.ac.uk/ena) under accession number: ERP001026). Briefly, EBV strains were obtained from a collection of cell lines and primary tumours from different geographical regions mainly Australia, Kenya, China, UK and USA (Appendix 1). Samples were sequenced using an Illumina HiSeq 2000 sequencer and sequence reads were 76-bp paired-end reads. All generated paired-reads were subjected to quality control pipeline to remove duplicated read pairs, paired reads if whether read had a raw median Phred quality score below 32. Reads shorter than 50-bp after trimming were eliminated. B95-8-cell line sequence data were generated by MiSeq (150-bp paired end) and provided by Dr Alan Chiang (Kwok et al. 2014).

For purpose of analysis, spontaneous LCLs (sLCLs) are considered herein as non-tumour-strains, while all long-established cell lines are considered as tumour-derived strains. All strains were included in multiple-sequence analysis (MSA) and minor variant analysis (MVs). However, strains with common origins were combined in analyses requiring distinct strains such as calculating proportions and diversity analysis, as follows: B95-8, LCL B95-8 delEBER, X50-7, and BL36 were combined into a single B95-8-cell line; TRL1-pre and TRL1-post were isolated from the same patient and had identical sequences so were described as a single strain, TRL1; P3HR1-HH514-C16 is originated from Jijoye, thus, they were combined into a single origin (Jijoye).

2.1.2 Constructing IR1 theoretical templates

IR1 region analysis was conducted by constructing a theoretical template consisting of 716-bp upstream of IR1 region (BamHI C region), 1.6 repeat units of IR1, and 456-bp downstream of IR1 region (BamHI Y region). This was achieved by using in-house scripts in Python to perform the following assembly (DNA sequence with annotations of B95-8 IR1 template is shown in appendix 2). For each strain, the initial sequence of the IR1 unit was obtained by mapping reads to first IR1 unit of the prototype EBV genome B95-8 (positions 12001 to 15072 of NC_007605.1 sequence) by use of the Burrows-Wheeler aligner (BWA) (v0.7.9a), using the default BWA-MEM algorithm (Li 2013). The consensus sequences of templates were obtained using SAMtools (version 0.1.19) (H. Li et al. 2009). To get the sequence of the last IR1 unit, the multiple-sequence alignment of IR1 sequences showed that the last 15nt "AGGCCCAGCCCCCTC" of the last IR1 unit in NC_007605.1 was conserved across all strains under studying. Thus, IR1 unit was concatemerized to give one intact IR1 unit followed by short repeat whose end defined by the conserved 15-bp. The flanking unique sequences were obtained from the published genomes (Palser et al. 2015). Finally, reads were mapped to the constructed template using BWA and the consensus sequence was generated again by use of SAMtools. The generated consensus was used as a reference for a second alignment to avoid bias that might occur due to reference-based read mapping.

Where the entire (or part of) BamHI-C or BamHI-Y region was missing, alignment to the whole genome was performed using a genome reference (NC_007605 for type 1, or NC_009334 for type 2). To predict the potential flanking regions, in-house Read Depth Analysis (RDA) (see section 3.3) was applied across the whole genome. At the potential flanks predicted by RDA, mapped reads generated by SAMtools were visualized carefully using Geneious (version 7.1.8) to verify the exact flanking sequence supported by partially trimmed

reads. If trimmed sequence in reads mapping to unique region matches sequence within IR1, it may indicate the other end of the deletion provided that the trimmed sequence in reads mapping to IR1 matches the unique sequence and vice versa.

2.1.3 Constructing TR unit template

To collect all terminal repeat reads, TR sequence was removed from the EBV prototype reference (NC_007605) by removing sequence from 1 to 1000 and from 168,824 to 171,823 of the genome. For each strain, reads were mapped to the trimmed reference using BWA aligner with default BWA-MEM algorithm (Li 2013). Unmapped reads (paired and orphan reads) were then obtained in a BAM file by using SAMtools (version 0.1.19) (Li et al. 2009) and then converted into .fastq format using picard-tools (version 1.72, http://broadinstitute.github.io/picard/). Collected reads were taken for de novo assembly using SPAdes 3.1.1 Genome Assembler (Bankevich et al. 2012). Generated contigs were visualized by using Geneious to identify overlapping contigs to get one TR unit flanked by unique sequences (DNA sequence with annotations of B95-8 TR template is shown in appendix 8).

2.1.4 Analysis of IR1 and TR sequences

2.1.4.1 Single nucleotide polymorphism (SNP) density

Multiple-sequence alignments (MSAs) were generated using Geneious alignment with default parameters. Sequences were then visualized in Geneious to fix misalignments within deleted regions. An in-house script was used to count single nucleotide polymorphisms (SNPs) (excluding indels) between strains relative to a consensus sequence generated by MSA.

2.1.4.2 Phylogenetic tree

Using Geneious (version 7.1.8), a phylogenetic tree was built using the neighbor-joining method and the Jukes-Cantor model for genetic distance.

2.1.4.3 BWRF1 ORF prediction

Since BWRF1 lacks the start codon (ATG), ORF length was calculated at it's the maximum possible length between in-frame stop codons. All ORF exceeding 600-bp (200 AA) were reported and shown in (figure 3.10).

2.1.4.4 Variant calling

After mapping reads to the corresponding template (i.e. reference), pileup files were generated from SAMtools (version 0.1.19) (Li et al. 2009) with a minimum Phred quality score of 20 considered minor variants. The mpileup2SNP algorithm in VarScan v2.3.7 (Koboldt et al. 2012), a platform-independent mutation caller, was used to retrieve variants mapped on both the forward (+) and reverse (-) strands of the reference templates from pileup files, using default parameters: minimum number of reads to call [8], minimum number of reads to set variant [2], minimum quality score to count in minor variant [15] and minimum variant allele frequency threshold [0.01]. Reported variants, then, underwent in-house python scripts for more filtration and analysis (see section 3.9). All MVs data were visualised by plotting read bias against MV frequency, using RStudio (v0.98.1103).

2.1.4.5 Statistics

Fisher's Exact test was used to assess whether recombination-like MVs occurred more often than expected in tumour-drived strains. The Benjamini–Hochberg correction was applied to adjust for multiple comparisons. Results were considered significant if a corrected p-value was <0.05. The statistical test was conducted using R programming language.

2.1.4.6 Detecting potential recombination events

The possibility of recombination between the sequences was examined by comparing a virus with a panel of other viruses using T-REC tools (Tsimpidis et al. 2017) and Bootscan analysis. T-REC software is capable of dealing with many comparator strains, but the similarity between strains can be interrupted by random mutation. Bootscan analysis was performed using the manual bootscan methods in RDP4 (v4.82) (Martin et al. 2015). It is worth noting that during working on this part, we noticed that where the sliding window contained no SNPs that distinguished the comparators, Bootscan algorithm favoured the comparator at the bottom of the list, and disfavoured the sample at the top, rather than giving equal odds of identifying any one of the identical comparators. Fortunately, the bug was fixed by RDP4 developer team. However, using different or additional strains changed the output substantially. So, this approach was useful for visualization but not for detecting recombination events in our data set.

2.2 Solutions and buffers

All solutions were prepared in Milli-Q-H_2O, unless stated otherwise.

Luria broth (LB)

1% Yeast extract, 1% BactoTryptone, 0.5% NaCl, made up in dH_2O and autoclaved.

Superbroth (SB)

2.4% Yeast extract, 1.2% BactoTryptone, 0.4% glycerol, autoclave and store at RT. Add 20X KPB to 5% final concentration before use.

KPB (20X stock)

4.6% KH_2PO_4, 24.3% K_2HPO_4. Autoclave and store at RT.

LB Agar

1% Yeast extract, 1% BactoTrypton, 0.5% NaCl, 1.5% agar, autoclave and store at RT.

STET

8% sucrose, 5% Triton X-100, 50 mM EDTA , 50 mM Tris-HCl (pH 8).

TE buffer

10 mM Tris-HCl (pH 8), 1 mM EDTA (pH 8)

$T_{10}E_{25}$

25 mM EDTA, 10 mM Tris (pH 8), autoclave and store at RT.

20x TBE buffer

2 M Tris, 2 M boric acid, 40 mM EDTA.

GTE

50 mM glucose, 50 mM Tris-HCl pH 8, 50 mM EDTA pH 8.

Alkaline SDS (P2)

0.2 M NaOH, 1% SDS.

Ficoll DNA loading buffer (6x)

25% ficoll-400, 25 mM EDTA, xylene cyanol and bromophenol blue.

RIPA lysis buffer

1x RIPA buffer, 4% protease inhibitor cocktail, 1 mM PMSF, autoclave and store at RT.

SDS-PAGE sample loading buffer (5x)

50 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 0.1% bromophenol blue, 1% 2-mercaptoethanol.

SDS-PAGE running buffer (10x)

0.25 M Tris, 192 mM glycine, 0.1% SDS.

SDS-PAGE transfer buffer

10x SDS running buffer, 23.3% ethanol

2.3 Antibiotics

Table 2.1 Antibiotics

Antibiotic	Working concentration
Ampicillin	0.1 mg/ml
Puromycin	1 μg/ml
Chloramphenicol	12.5 μg/ml
Hygromycin B	200 μg/ml

2.4 Chemical inducers of EBV lytic cycle

Table 2.2 chemical inducers of EBV lytic induction

chemical	Working concentration
2-O-tetradecanoylphorbol-13-acetate (TPA) (Sigma)	20 nM
Sodium butyrate (NaB) (Sigma)	5 mM

2.5 Cloning and DNA analysis

2.5.1 Restriction enzyme digestion and ligation

Digestion was executed by using restriction endonuclease enzymes from New England Biolabs (NEB) or Promega, and the appropriate buffers and incubation temperature were selected according to the manufacturer's recommendations. In diagnostic digestions, 1 µl plasmid DNA or 5 µl BAC miniprep were used in a 15 µl reaction, wherease preparative digestions were performed in 30 µl reactions using 1-5 µl of plasmid DNA to correspond to approximately 500-2000 ng of DNA. Where using a single restriction enzyme to digest a vector fragment (whose self-ligation could generate a viable plasmid), the phosphate groups were removed from the digests using Antarctic Phosphatase (NEB).

Digested DNA fragments were excised from 1% agarose gels and purified using QIAquick gel extraction kit (QIAGEN). Purified DNA fragments were ligated using T4 DNA ligase in 10X T4 ligase buffer (NEB). To insert a fragment, the ratio of added DNA was 1:3 vector to insert in total volume of 10 µl and the ligation mixture was incubated at room temperature for 3 h. To anneal oligos (table 2.3), the resuspended oligos were mixed 1:1 in a PCR tube with 1X T4 DNA ligase buffer (NEB). The tube was left in the PCR machine at 95℃ for 2 min and then gradually cooled to 25℃ over 45 min. Constructed plasmids were verified by diagnostic digestions and sequencing.

Table 2.3 Oligonucleotides

oligos	Sequence 5'-3'
pCAG linker-top	TCGAGACCGGTAGCGCTAGCGTCGACGAATTCACCTGCATGCAGCTGCAGGCGCGCC
pCAG linker-bottom	CTGGCCATCGCGATCGCTGCTGCTTAAGTGGACGTACGTCGACGTCCGCGCGGTTAA
TR-sgRNA-1-top	TAGGGGAGCCTGCACGCCGTTGGA
TR-sgRNA-1-bottom	AAACTCCAACGGCGTGCAGGCTCC
TR-sgRNA-2-top	TAGGATTCCTGGAAAAAGTGGAGG
TR-sgRNA-2-bottom	AAACCCTCCACTTTTTCCAGGAAT
TR-sgRNA-3-top	TAGGGAGCGTGACGAAGGGCCCCA
TR-sgRNA-3-bottom	AAACTGGGGCCCTTCGTCACGCTC
sgRNA-1_oligo_UP_Px330	CACCGGAGCCTGCACGCCGTTGGA
sgRNA-1_oligo_Down_Px330	AAACTCCAACGGCGTGCAGGCTCC
sgRNA-3_oligo_UP_Px330	CACCGAGCGTGACGAAGGGCCCCA
sgRNA-3_oligo_Down_Px330	AAACTGGGGCCCTTCGTCACGCTC
PpoI-site-BV -top	TCGAGGACTCTCTTAAGGTAGCCCG
PpoI-site-BV -bottom	TCGACGGGCTACCTTAAGAGAGTCC
PpoI-site-TR-top	GTACCGACTCTCTTAAGGTAGCCG
PpoI-site-TR-bottom	TCGACGGCTACCTTAAGAGAGTCG

2.5.2 Transformation

Pre-cooled DNA plasmid (typically the whole ligation reaction) was added into 50 µl of chemically competent cells DH5α *E. coli* and incubated on ice for 20 min. A heat chock was then executed by placing the transformation mixture into water bath at 42℃ for 45-50 sec and immediately transferring it into ice. Within 1-2 min, 900 µl of LB was added and the bacteria were incubated for 1 hour at 37 ℃ with shaking (200 rpm) to allow antibiotics resistance expression. Either 100 µl of the transformation mixture or a pellet of the transformed bacteria (spun down at 3k rpm for 5 min and resuspended in 100 µl) was spread on an agar plate prepared with an appropriate antibiotic for selection (see table 2.1) and incubated at 37 ℃ overnight or at 30 ℃ for > 24 hours for unstable plasmids.

2.5.3 Plasmid miniprep

A freshly grown bacterial colony was picked and grown overnight in 1.5 ml of Super Broth SB at 37 ℃ (or at 30 ℃ for unstable plasmids) with shaking (200 rpm). The bacteria were pelleted by centrifugation at 4000 rpm for 5 min in microcentrifuge. The pellet was resuspended in 70 µl of STET by vortexing, lysed by adding 200 µl of alkaline SDS while vortexing, neutralised by adding 150 µl of 7.5 M ammonium acetate, and then incubated on ice for 5 minutes. Debris was spun down by centrifugation at 13,000 rpm for 15 minutes at 4°C. The supernatant was poured into a new tube and DNA was precipitated by adding 240 µl isopropanol and vortexing. DNA was pelleted by centrifugation at 10,000 rpm for 5 min at RT. The DNA pellet was washed with 200 µl of 70% ethanol and centrifuged at 10,000 in a microcentrifuge for 3 min. The DNA pellet was left to dry at RT for 5-15 minutes and

resuspended in 50 µl TE containing 5 µg/ml RNase A (or in 120 µl for high copy number plasmids).

2.5.4 Plasmid maxiprep

For transfecting B cells, plasmid DNA was isolated and purified by using EndoFree plasmid maxi kit (Qiagen) according to the manufacturer's protocol.

2.5.5 CsCl maxiprep

To isolate larger quantities of purified plasmid suitable for transfection, a freshly grown bacterial colony was picked and grown in 1.5 ml of Super Broth SB at 37 ℃ with shaking (200 rpm) for 8 h. the culture was transferred into 500 ml of SB with antibiotics and grown in a shaking incubator at 37 ℃ overnight (aiming for ~ 16 h). Next day, cells were spun down for 15 min at 4000 rpm in SLA-1500 rotor. The pellet was resuspended in 20 GTE, lysed by adding 30 ml alkaline SDS (quick mixing), and neutralized by immediately adding 30 ml of 7.5 M ammonium acetate. After 15 min of incubation on ice, the lysate was centrifuged for 30 min at 6000 rpm in a Sorvall SLA-1500 rotor at 4 ℃. The supernatant was filtered using a 70 µm cell strainer, and the filtrate was centrifuged for 15 min at 7000 rpm at 4 ℃ and filtered again. DNA was precipitated by adding 0.6 – 0.7 volume (typically ~118 ml) isopropanol, mixing by gentle inversion and incubation at RT for 5-10 min (max 30 min). DNA was centrifuged for 25 min at 6500 rpm in Sorvall SLA-1500 rotor at RT. The pellet was washed with 5 ml of 70% ethanol, allowed to dry for approximately 10 min at RT, and then resuspended in 4 ml $T_{10}E_{25}$. The DNA solution (weight 4 g) was added into 4.3 g CsCl. After completing dissolving the CsCl by gentle

inversion, 200 µl of EtBr (10mg/ml). the DNA/CsCl was transferred into a Beckman ultracentrifuge tube (5.1 ml Quick-seal tubes ref #342412) using an open syringe and blunt needle (as a funnel). The tube was filled to the top with TE and sealed using a Beckman tube topper. The tubes were centrifuged in ultracentrifuge at 45,000 rpm, max acceleration, coast deceleration (no break), 20 ℃ for >12 h using a Beckman NVT165.2 rotor. After centrifugation, BAC DNA (the lower band of two visible bands) was extracted using a 19G needle and 1 ml syringe. CsCl mix was added to the extracted DNA to a final volume of 5.2 ml (CsCl mix: 36.9g CsCl + 37.8 ml $T_{10}E_{25}$ in 50 ml Falcon tube) and 70 µl EtBr was added. The DNA/CsCl was transferred into fresh Beckman ultracentrifuge tubes and centrifuged again at 60,000 rpm for 16 h at 20 ℃ in a Beckman NVT165.2 rotor. Next day, the visible band was extracted using a 19G needle and 1 ml syringe, and transferred into a 2 ml eppendorf tube. EtBr was removed (typically six times) by mixing with 1 ml TE saturated ethanol and extracting the top layer. DNA was precipitated by first adding 0.5 vol 7.5 M ammonium acetate (for low abundance or BAC DNA, 1 µl GenElute was added after mixing NH4Ac), and then 2.5 vol 100% EtOH. DNA was spun down at 13000 rpm at RT using microcentrifuge for 10 min, and the pellet was washed in 200 µl 70% EtOH. The pallet was allowed to dry (for maximum 10 min) and resuspended in TE (0.2-0.5 ml for small pellet, 1 ml for larger).

2.5.6 Agarose gel electrophoresis

Based on DNA sizes, DNA fragments were separated by migration in 0.8 – 1.5 % agarose gels containing SybrSafe stain (Invitrogen) in 1x TBA at 70-90 V for 1-2.5 h in a Hoefer™HE 33 Mini Submarine electrophoresis unit.

2.5.7 Gel extraction

DNA gel extraction was performed by using either a QIAquick gel extraction kit (Qiagen) or ZYMOCLEAN™ gel DNA recovery kit according to the manufacturers' protocols.

2.5.8 Pulsed-field gel electrophoresis

To separate larger DNA fragments (> 10 kb), the CHEF-DR II system (BioRad) was used to run pulsed-field 1% agarose gel electrophoresis. The gel was prepared by dissolving agarose in 0.5x TBE. The run was carried out in 0.5x TBE for 12 h at 14 ℃ (using the chiller system), at 6V/cm and initial switch = 1 sec and final switch = 10 sec. After running, DNA was stained by soaking the gel in 0.5x TBE with 0.5 μg/ml ethidium bromide (EtBr) for 30 min with gentle agitation. The gel was visualized on the Gel Dox XR+ system (BioRad).

2.6 Cell culture

293T cells were cultured in DMEM (Gibco), and B cells (sLCLs) were cultured in RPMI 1640 medium (Gibco). Media were supplemented with 10% fetal calf serum (FCS) (Gibco), 100 U/ml of penicillin and 100 μg/ml of streptomycin. Cell lines were incubated at 37°C in either a 10% CO_2 incubator for routine culturing or a 5% CO_2 for inducing lytic cycle experiments.

EBV-infected PBLs were cultured at 37°C, 5% CO2 in RPMI 1640 medium (Gibco) supplemented with 15% FCS (tested for suitability for EBV transformation of primary B cells: PAA or Gibco), penicillin, streptomycin.

2.7 Lentivirus production and sLCLs infection

2.7.1 Lentivirus production

293T cells were seeded at 70% confluence one day before transfection in 10 cm dishes. Next day, 10 μg of lentivector plasmid (i.e. plasmid of interest), 4 μg of packaging plasmid pMD2G and 8ug of envelope protein plasmid pPAX2 were added into 2.2ml Opti-MEM (Invitrogen). 16.5μl of lipofectin (0.75μl/1μg DNA) (Invitrogen) and 88 μg of peptide6 (4μg/1μg DNA) in 880ul Opti-MEM were mixed together and incubated for 5min at RT. The lentiviral plasmid mixture was added in wise into lipofectin-peptide mixture while mixing gently. Following 20 min standing at RT, the mixture was filled with Opti-MEM up to 5ml and added into the 293T cell dish from which all media had been removed. The dish was incubated at 37°C in a 10% CO_2 incubator for 6h. Then, all media were replaced with 10ml antibiotic-free RPMI with 10% FCS, and the dish was incubated at 37°C in a 5% CO_2 incubator. Packaged lentivirus-containing supernatant was harvested 48h later and filtered through a 0.45μm syringe filter.

2.7.2 Infecting sLCLs with lentivirus

Cell were fed one day before infection. On the following day, $3x10^5$ cells in 1ml were mixed with 1ml of lentivirus supernatant. To enhance the infection, the mixture was centrifuged at 2000g for 2h at RT. 1ml of supernatant was discarded and the pellet was

resuspended with the remainder. Cells were transferred to a well of a 24 well plate and 0.5 ml growth media was added to nhance a final volume of 1.5ml/well. Following 48h incubation, 1 μg/ml of puromycin was added for cell selection. The cells were moved to larger containers as they proliferated, or the media were replaced every 3 days. The ZHT protein was activated by adding 4HT to a final concentration of 0.4μM.

2.8 DNA extraction

DNA was isolated by using a DNeasy blood and tissue kit (Qiagen) according to the manufacturer's protocol.

2.9 Polymerase chain reaction PCR

All PCR experiments were conducted by using Q5 polymerase (NEB) according to the manufacturer's protocol and the primers in table 2.3. Cycling conditions were as follow:

For amplifying the EBNA-LP transcript (figure 3.25):
98°C for 30sec, (98°C-10sec, 65°C-30sec, 72°C-55sec) x 30 cycles, and 72°C for 2 min

For amplifying the IR1 repeats transcript (gradient PCR) (figure 3.26):
98°C for 30sec, and then 30 cycles of 98°C for 10sec, (56.7, 57.8, 59.3, and 61.0 °C) for 30sec, 72°C for 10min and 72°C for 2min.

For creating a homologous sequence in pR6K (section 5.2.2.2):
98℃ for 10min, (98℃-30sec, 62℃-30sec, 72℃-30sec) x 5 cycles, (98℃-30sec, 66℃-30sec, 72℃-30sec) x 5 cycles, (98℃-30sec, 70℃-30sec, 72℃-30sec) x 5 cycles, and 72℃ for 2min.

For amplifying a terminal repeat unit (section 5.2.4.1):

98℃ for 10min, (98℃-30sec, 70℃-30sec, 72℃-1min) x 5 cycles, (98℃-30sec, 67℃-30sec, 72℃-1min) x 5 cycles, (98℃-30sec, 65℃-30sec, 72℃-1min) x 20 cycles, and 72℃ for 2min.

For verifying homologous recombination (figure 5.25):

98°C for 1min, and then 30 cycles of 98°C for 30sec, 65°C for 30sec, 72°C for 30sec and 72°C for 2min.

Table 2.4 PCR primers

primers	sequence 5'-3'
R6K-AG876-sg1-fwd	GTCGACCTAGGCGCGCCGGAGGGTAGAATGACAG GGGGGCGGGGACAGAGAGGCGGTCGCGCCCCCCT ACACCTACAGGATGATGGACAC
R6K-AG876-sg1-rev	GGCGCGCCTAGGTCGACAACGGCGTGCAGGCTCCC GCGCCGCCCCCCGCTCCCGCCCCCACCAGACATGT AGACGACGACGAACAGAG
R6K-AG876-sg2-fwd	GTCGACCTAGGCGCGCCAGGGGGCGTGGCCTGCCC CCCGCGGCCCCCTAGCCCCCCCGCACAGAGCGCTA CACCTACAGGATGATGGACAC
R6K-AG876-sg2-5'rev	GGCGCGCCTAGGTCGACCACTTTTTCCAGGAATGC GCGGCCCATGGCCCCACCCCCATGGCCCCACTGTA GACGACGACGAACAGAG
R6K-AG876-sg3-3'fwd	GTCGACCTAGGCGCGCCAGGGCTGACCCCGGCAAA CGTGACCCGGGGCTCCGGGGTGACCCAGCCTACAC CTACAGGATGATGGACAC
R6K-AG876-sg3-5'rev	GGCGCGCCTAGGTCGACGGCCCTTCGTCACGCTCC GCTGCCCGCGCCCCGCCCCCCTTGGGGGCTTGTAG ACGACGACGAACAGAG
TR_unit_fwd	ATTCCTGGAAAAAGTGGAGG
TR_unit_rev	CCTGGGGGGCCTTTGTCAG
TR_Unit_BamHINotI_fwd	CGCGGATCCGCGGCCGCATTCCTGGAAAAAGTGGA GG
TR_Unit_EcoRISacI_rev	CGCGAATTCGAGCTCCCTGGGGGGCCTTTGTCAG
rt-recom-for	TATCCAGTCCAGACGCTTTTC
BAC-vec-down	TTGTATGCCTGCTGTGGATT
lt-recom-for	CGGGCCATTTACCGTAAGTTA
lt-recom-rev	TAAAGAACTGGCCGCTTTG
Cp-C2 forward	AAGAAACAGCCTCCTGCACG
W2 reverse	GGTGCCTTCTTAGGAGCTGTCC

2.10 Real-Time PCR

Real-Time PCR (also known as quantitative PCR) was conducted using QuantStudio 7 Flex Real-Time PCR System (Applied Biosystem), using Takyon Low ROX Probe mastermix dTTP blue (Eurogentec). The amplification was carried out for EBV DNA polymerase gene (BALF5), a highly conserved unique region of the EBV genome, and the human β-2 microglobulin (B2M) gene, which serves as an endogenous control for cellular DNA. DNA standards were prepared from Namalwa BL, which contains two integrated EBV genomes per cell (table 2.3). Regular cycling conditions were followed according to the manufacturer: 95°C for 3min, then 40 cycles of 90°C for 10sec and 60°C for 1min. Quantification of viral gene levels was carried out using the Delta-Delta Ct method.

Table 2.5 Primers for Real-Time polymerase chain reaction (qPCR).

EBV Pol (BALF5)	
Forward primer	5' CTTTGGCGCGGATCCTC 3'
Reverse primer	5' AGTCCTTCTTGGCTAGTCTGTTGAC 3'
EBV Pol probe	5'-FAM-CATCAAGAAGCTGCTGGCGGCC-TAMRA 3'
β2m	
Forward primer	5' GGAATTGATTTGGGAGAGCATC 3'
Reverse primer	5' CAGGTCCTGGCTCTACAATTTACTAA 3'
β2m probe (VIC)	5' AGTGTGACTGGGCAGATCATCCACCTTC 3'

2.11 Protein analysis

2.11.1 Protein extraction

Cells (usually 10^5-10^6 cells) were centrifuged at 1100 rpm in a microcentrifuge for 5 min. the cell pellet was washed with 1x PBS and lysed for 5 min on ice in 1-2x volume of cell pellet of RIPA lysis buffer (typically 70 µl). Samples were centrifuged at 14000 rpm for 10 min to pellet cell debris, and the supernatant was harvested. Protein concentrations were measured using the Bio-Rad RC DC™ Protein Assay against bovine serum albumin as standard.

2.11.2 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

Protein separation was conducted using(SDS-PAGE). The gel consists of two layers: resolving gel at the bottom (10%-12% acrylamide (Sigma), Tris pH=8.8, 0.1% SDS) and stacking gel at the top (5% acrylamide (Sigma), Tris pH=6.8, 0.1% SDS). Gels were polymerized by adding 24 µl of 10% ammonium persulphate (APS) and 10 µl of tetramethylethylenediamine (TEMED) per 1.5 mm gel. Prior to loading, equal amount of protein samples (20–30 µg) were mixed with 5x SDS-PAGE sample loading buffer and heated to 100 ℃ for 5 min. Samples were loaded, and the gel was run at 150 V for 75 min in SDS running buffer using a Mini-PROTEAN® II electrophoresis cell (Bio-Rad).

2.11.3 Transferring proteins from SDS-PAGE gel to a membrane

Separated proteins were electrotransferred onto Whatman nitrocellulose membranes (GE Healthcare) at 100 V for 1 h in SDS-PAGE transfer buffer using a Mini Trans-Blot module (Bio-Rad).

2.11.4 Western blotting

The blotted membrane was blocked in 5% milk in 0.1% PBST (PBS, 0.1% Tween) solution for 1 h on a shaker at RT. The blocking solution was discarded and primary antibodies were diluted (1/1000) in 10 ml fresh blocking solution and added. The membrane was incubated with the primary antibodies (JF186 or BZ-1) overnight at 4 ℃ with a gentle shaking (. The following day, the membrane was washed 3 times with 0.1% PBST for 10 min each. Secondary antibodies conjugated to horseradish peroxidase (HRP) were diluted (1:2000) in 10 ml of blocking solution, added to the membrane and incubated for 1 hour at room temperature with shaking. Again, the membrane was washed 3 times with 0.1% PBST for 10 min each. The membrane was incubated briefly (~5 min) in Amersham ECL western blotting detection reagents. The membrane, then, was exposed to Amersham Hyperfilm MP (GE Healthcare) and developed using an automatic X-ray film processor JP-33 (JPI).

2.12 Cell transfection

2.12.1 Transfection of 293T cells

293T cells were transfected using GeneJuice transfection reagent (Novagen) according to the manufacturer's protocol. Briefly, $3x10^5$ cells were seeded in 6 wells plate in 1 ml media, one day prior the transfection. For each 1 μg DNA, 3 μl of GeneJuice was added dropwise into 100 μl Opti-MEM. GeneJuice-Opti-MEM mixture was mixed thoroughly and incubated at RT for 5 min. DNA (1 μg) was added into the GeneJuice-Opti-MEM mixture and incubated at RT for 5-15 min. The cell media were replaced with fresh, and the entire DNA-GeneJuice mixture was added dropwise to the cells. the cells were analysed (e.g. GFP assay) 48-72 h post transfection.

2.12.2 Electroporation of LCLs

LCLs were electroporated using the Neon® Transfection System according to the manufacturer's protocol. For choosing the best electroporation conditions, the 24-well optimization protocol (designed by the manufacturer) was followed. Briefly, prior to electroporation, cells were split with fresh media such that they were around 70% confluent on the following day. Next day, $2x10^6$ cells/well x 24 wells were prepared and resuspended in buffer R. DNA plasmid (1 μg/well) was added into cells. Electroporation was conducted using the parameters (i.e. conditions) listed in table 2.2. For electroporation experiments, cells were electroporated with donor BAC (from 1-10 μg/well) and Cas9/sgRNA plasmid (0.5-5 μg/well) using the best two conditions. Transfection efficiency was estimated by doing GFP assay.

Table 2.6 Electroporation parameters.

Protocol	Pulse Voltage	Pulse width	Pulse no.
1	0	0	0
2	1400	20	1
3	1500	20	1
4	1600	20	1
5	1700	20	1
6	1100	30	1
7	1200	30	1
8	1300	30	1
9	1400	30	1
10	1000	40	1
11	1100	40	1
12	1200	40	1
13	1100	20	2
14	1200	20	2
15	1300	20	2
16	1400	20	2
17	850	30	2
18	950	30	2
19	1050	30	2
20	1150	30	3
21	1300	10	3
22	1400	10	3
23	1500	10	3
24	1600	10	3

2.12.3 GFP assay

To determine the transfection efficiency, 96 µl of each sample was transferred into a 1.5 µl Eppendorf tube 48-72 h post transfection. Cells were stained with 2 µl Hoechst 33342 (500 µg/ml) mixed by pipetting. The sample was incubated at 37 ℃ for 15 min. After incubation, 2 µl propidium iodine (PI) (500 µg/ml) was added into the sample and mixed by pipetting. The sample was loaded into NC-Slid A8™ chamber (~10) and a GFP assay was performed using a NucleoCounter® NC-3000™ machine and the "GFP Transfection Assay-Hoechst and PI" programme.

2.12.4 Low molecular weight DNA extraction

To obtain EBV-BAC for episome rescue, transfected cells were pelleted, washed in PBS and centrifuged at 3000 rpm for 5 min in a microcentrifuge. The pellet was resuspended in 60 µl STET and lysed by adding 130 µl alkaline SDS. This was then neutralized by addition of 110 µl 7.5 M ammonium acetate and incubated on ice for 5 min. After the incubation, cells were centrifuged at 13000 rpm in a microcentrifuge for 30 min at 4 ℃. The clear lysate was transferred into a new Eppendorf tube, and DNA was extracted with phenol:chloroform:isoamyl alcohol (25:24:1) once and then twice with chloroform using MaXtract gel tubes (Qiagen) to separate the organic and aqueous phases. The DNA was precipitated by addition of 670 µl of 100% EtOH and spun down at 13000 rpm at RT using a microcentrifuge for 10 min. The pellet was washed with 70% EtOH, allowed to dry for maximum 10 min and resuspended in 10-50 µl TE containing 5 µl/ml RNase A.

2.12.5 Episome rescue

Up to 2 µl of low molecular weight DNA isolated from hygromycin-resistant cells was electroporated into 20 µl of electrocompetent *E. coli* DH10B (NEB and Invitrogen) in a 1 mm electroporation cuvette (Bio-Rad). The electroporation was conducted using a Bio-Rad Gene Pulser electroporator at 1.8 kV, 200 Ω and a capacitance of 25 µF. Immediately, 400 µl of SOC medium was added after electroporation. Bacteria were incubated at 37°C for 1 h with shaking. After incubation, the whole volume was plated on a 14 cm agar plate containing chloramphenicol. DNA was isolated from growing colonies by miniprep and tested by restriction digestion endonucleases.

2.13 Generating LCLs

2.13.1 Virus titration

In 24-wells plate, 0.5 ml Raji cells ($2x10^5$ cells/ml) were infected with serial 10-fold dilutions of virus in a total volume of 1.5 ml/well. At 48 h post infection, 0.5 ml media supplemented with 20 nM TPA (Sigma) and 5 mM N-butyrate (Sigma) was added to infected cells to activate the CMV promoter-driven GFP. The next day, cells expressing GFP were counted per well under the fluorescent microscope.

2.13.2 Infection of peripheral blood leukocytes (PBLs)

Frozen PBLs were recovered and seeded at 2×10^6 cells/well (in a 24-well plate) in RPMI supplemented with penicillin-streptomycin, L-glutamine, 20% fetal calf serum, and cyclosporine (100 ng/ml). The cells were infected with 2×10^5 Raji infectious units of either HB9 or WT^{wt} EBV (Szymula et al. 2018). Media were replaced on the following day and every 3 days thereafter. After two weeks, growing cells (LCLs) were maintain in the regular RPMI medium.

2.14 sLCL proliferation assay

2.14.1 CellTrace Violet staining

Primary cells were suspended in PBS at final density of 10^6 cells/ml with 0.5 μl (50% of the recommended amount) of 5 μM CellTrace (Life Technology) in DMSO (Invitrogen) and incubated for 20 min at 37℃ in the dark. Cells with CellTrace mixture were then diluted 5

times in complete B cell media (RPMI) and incubated for 5 min at RT in the dark to remove any remaining free dye. Cells were pelleted by centrifugation and re-suspended in fresh pre-warmed complete B cell media with 100ng/ml cyclosporine in 6-wells plate.

2.14.2 CellTrace intensity-dilution-based proliferation assay

Cells in a volume equivalent to 10^6 cells in the initial infection were harvested and washed with cold PBS. The pellet was resuspended with 25 µl of PBS + 1 µl human IgG (sigma) and incubated for 10 min. The cells were incubated with (25 µl of PBS + 5 µl CD19 TexasRed (Invitrogen)) for 30 min dark. The cells were centrifuged and washed with 500 µl PBS. The pellet was resuspended in 500 µl PBS + 5 µl live/dead cell stain DRAQ7 (BioStatus). Cells were analysed on a Fluorescence-activated cell sorting (FACS) machine (BD LSR II) and cell proliferation visualised for live CD19+ singlet cells using the FlowJo software.

2.14.3 Flow cytometry procedure

Control samples were prepared, and compensation was conducted using Diva software (BD). The same total number of events was selected (using the Diva software). Gating for intact cells was done to exclude events with extreme of forward scatter area (FSC-A) and side scatter area (SSC-A) to eliminate debris, dead cells and cell clumps. Then, gating for singlets was defined using FSC width (FSC-W) vs FSC height (FSC-H) and in SSC-W vs SSC-H. Live cells were gated by using a live/dead control that was solely stained with DRAQ7.

3 Internal repeat 1 (IR1) analysis

3.1 Introduction

EBV has several regions contain tandem repeat sequences. The biggest repeat region is IR1 which is an approximately 20kb array of 3kb repeat units. IR1 is of interest as each repeat contains the early latency promoter Wp and two exons (W1 and W2) that are together encode the repeat domain of EBNA-LP (see section 1.7.2 for more details). Despite the increasing number of EBV genome sequences, repeat region analysis has been avoided due to the complexity of sequence alignment and assembly. Dealing with such regions creates ambiguities in mapping reads, as sequence reads generated from repetitive regions (called multi-mapped reads, or multi-reads) align naturally to multiple locations, since each repeat matches equally to a given read during alignment. As a result, the confidence of mapping multi-reads is not as high as mapping reads to a unique region.

If repetitive regions are longer than the read length (the read length in NGS platforms ranges from 50 bp to 700 bp), data analysis tasks become even more challenge. Alignment tools such as BWA and Bowtie (Li & Durbin 2009; Langmead et al. 2009) will map multi-reads randomly (reviewed in Treangen & Salzberg 2011), which means that reads with a unique SNP (i.e. SNP that does not exist in the reference sequence) will be scattered across repetitive regions as they align equally to repeat units (figure 3.1). Consequently, in downstream analysis (e.g. generating a consensus sequence and calling SNPs), tools assign the nucleotide supported by the majority of reads as consensus while attributing other nucleotides to sequencing error because of their low frequency compared to consensus supporting reads, resulting in losing true positive SNPs.

My first informatics challenge of this project, was to refine the sequence of the repetitive regions starting with the IR1 in order to address four questions: (1) To what extent is there heterogeneity in the W repeats? (2) Are the unique sequences flanking the W repeat region conserved across strains? If not, what are the correct junctions? (3) If there are variations in the W repeats, do they disrupt established features? (4) Can we develop a systematic method to analyse repeat regions?

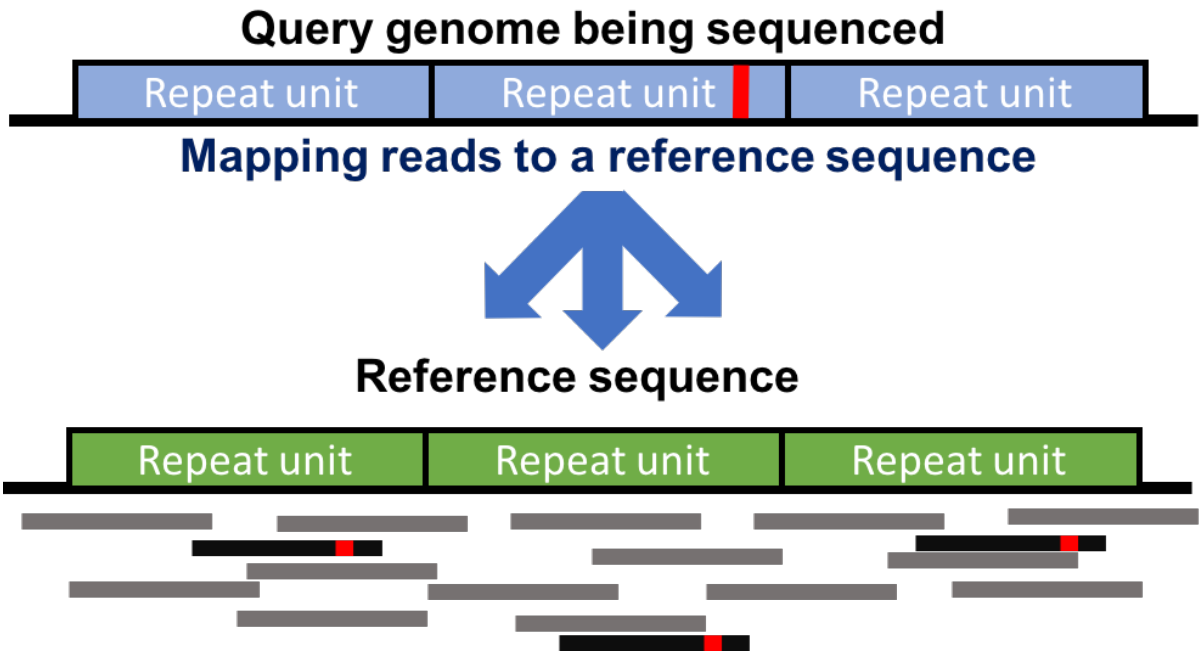


Figure 3.1 Mapping multi-reads.
Schematic diagram showing how alignment tools such as BWA and Bowtie deal with reads from a repetitive sequence. If this sequenced genome has an SNP in one repeat unit (represented by red line), SNP's supporting reads will map randomly and be scattered throughout repeat units as they align equally to all repeats. As a result, downstream analysis (e.g. SNP calling) may consider these SNP reads to be a technical error due to low frequency.

3.2 Constructing a theoretical template to simplify repeat sequence analysis

To simplify analysis of the repetitive region, a theoretical reference (Figure 3.2) was constructed for each EBV strain sequenced in Palser et al. (2015) (see appendix 1 for strain names and details). The theoretical reference is called herein a theoretical template, and was designed to reflect the whole IR1 region in the EBV genome, including the flanking unique regions. It consists of 716 nucleotides of the unique sequence upstream IR1 region (i.e. BamHI-C fragment); one full copy of the IR1 repeat unit (~3072 bp) followed by a partial copy (around 1852 bp) of IR1 unit (designated as 'last IR1' – this is because the IR1 repeat region starts and finishes within different parts of the repeat unit, such that there are 7.6 copies in the EBV type 1 and type 2 prototype genome sequences NC_007605 and NC_009334.1); and 456 nucleotides downstream of the IR1 region (i.e. BamHI-Y fragment) (figure 3.2 and appendix 2). The rationale for using a minimum number of IR1copies in the theoretical template is to force all multi-reads to map to a restricted area. This allows tools for calling variants to analyse a maximal read depth, whereas, if supporting reads were mapped to many copies, this would reduce the percentage of variant frequency. Also, including unique sequences flanking IR1 in the theoretical template would be useful for highlighting the SNPs that seem to be specific to the first or last IR1 repeat unit in the real genome by mapping reads to the template and examining mapped reads that span the junctions between the unique sequences and the IR1 termini.

For each strain, sequences of the unique regions flanking IR1 were obtained from the contigs assembly "continuous sequences generated by *de novo* assembly of smaller overlapping fragments" provided by Anne Palser and used to produce the genome sequences published in Palser et al 2015, while a preliminary IR1 sequence was generated by mapping

reads to the first IR1 repeat unit of the standard EBV reference genome NC_007605. From the multiple sequence alignments (MSA) of EBV genomes generated in Palser et al 2015, there were eight strains reported to have a deletion in the unique sequence upstream and/or downstream of IR1 (table 3.1). Since these strains had non-standard junctions, they needed more analysis to identify the junctions between IR1 and the unique regions, and produce a template (see section 3.4). To get a final consensus sequence for IR1, sequence reads were mapped to their corresponding template. The resultant consensus sequences were used as references for re-alignment, to generate pile-up files of reads aligned to consensus. These consensus and pile-up files were the main files used for subsequent analyses.

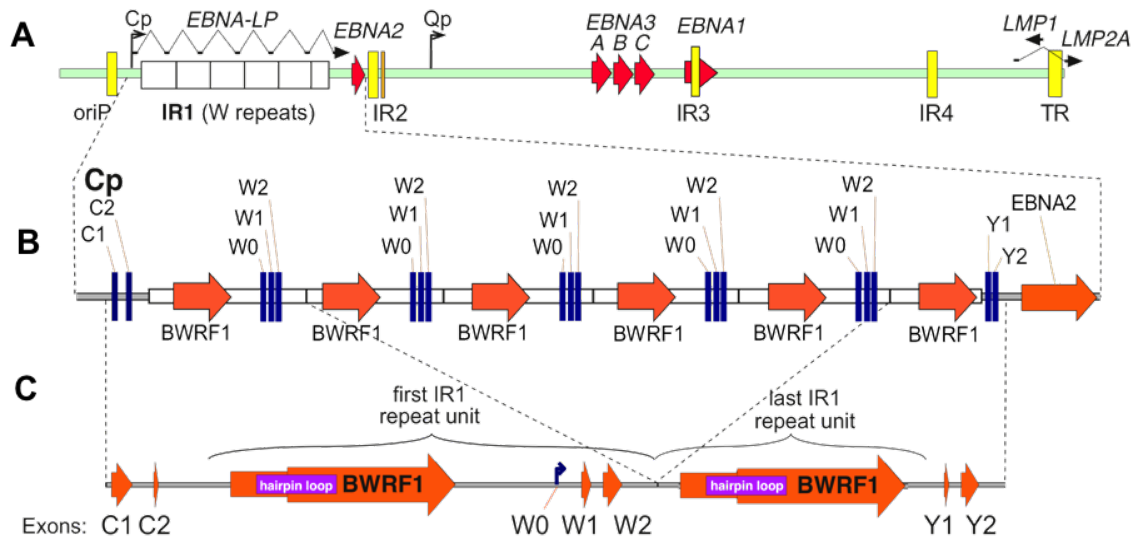


Figure 3. 2 Structure of the theoretical template.
(A) Schematic representing linear EBV genome showing the main EBV latency genes shown as ORFs (EBNA1 - 3 represented in red arrows) or spliced transcripts (EBNA-LP, LMP1 and LMP2A represented in black arrows) with the major internal repeat region IR1 (white boxes) and other repetitive region (internal repeat IR-2 to -4, family repeat (FR) of oriP and terminal repeat (TR); yellow bars) and the major latency promoters (Cp and Qp) which are represented as bent arrows. **(B)** Zoomed in view of IR1 region (white boxes) and the flanking sequences. The IR1 region is composed of the typical number of repeats (5.6 units). Exons are shown in blue bars while big red arrows represent the ORF BWRF1. **(C)** Structure of the theoretical template (with 1.6 IR1 repeat units) against which sequence reads were mapped (DNA sequence with annotations of B95-8 theoretical template is shown in appendix 2). The ORF BWRF1 is presented with two start points showing the main two lengths of BWRF1 (described in section 3.6) the purple box shows the hairpin sequence in the IR1 unit. Latency exons (C1, C2, W1, W2, Y1, and Y2) are indicated with exon W0 initiating at the Wp promoter (bent arrow). (Figure was produced by Dr White).

3.3 Read depth analysis (RDA) verifies the integrity of constructed templates

Since the theoretical templates were designed based on the general structure of the IR1 region in type 1 and type 2 EBV representative genomes NC_007605 and NC_009334.1 (figure 3.2), and it is not formally known that this structure is conserved across all virus strains, it was necessary to find a systematic technique to verify the fidelity of the constructed templates before conducting further analysis. I found that if a strain of query has an inserted or deleted sequence, mapping reads to a reference will show a sudden change in the number of reads mapped to adjacent nucleotides at both ends of defective region. Thus, read depth analysis (RDA) was devised to verify that the constructed templates that reflect the IR1 region in the EBV genomes.

The read depth is the number of sequence reads aligned to a nucleotide in the reference sequence. Read depth has been used as the basis in many tools for detecting the copy number of variants (CNV) in the human genome (Yoon et al. 2009; Yao et al. 2017). The divergence in read depth across sequenced samples is used to discover duplications and deletions in individuals by examining higher and reduced read depth compared to known diploid regions.

Unlike the algorithm used for counting read depth in CNV detection, which counts mapped reads in non-overlapping windows of a specified width (Yoon et al. 2009), RDA here checks the similarity between the number of reads supporting adjacent nucleotides across the template. It calculates the ratio of read depth at a nucleotide to the read depth of the previous (adjacent) nucleotide in a sliding window of 2 bp. The basic idea is that any unidentified deletion or insertion in a sample creates a big difference in the read depths supporting adjacent

nucleotides at both ends of defective region. The bigger the difference between the number of reads, the more the ratio will deviate from 1, indicating a sudden change in read depth. Positions with log2 ratio exceed ±0.5 were reported. This threshold was based on studying several strains individually to find the most reasonable limit for variation in the read depths of each of two adjacent nucleotides.

To test RDA, first the EBV strain TRL1 sequence data were used, as this strain has a deletion in EBNA3B region identified by PCR (Gottschalk et al. 2001). Reads were aligned to the standard EBV genome sequence NC_007605, and then RDA was applied to the alignment output. RDA was able to predict the known deletion in EBNA3B as shown in figure 3.3A. The region of the deletion in TRL1 was identified by PCR (Gottschalk et al. 2001) and as shown in figure 3.3 A and B, RDA was able to detect precisely the edges of the deletion.

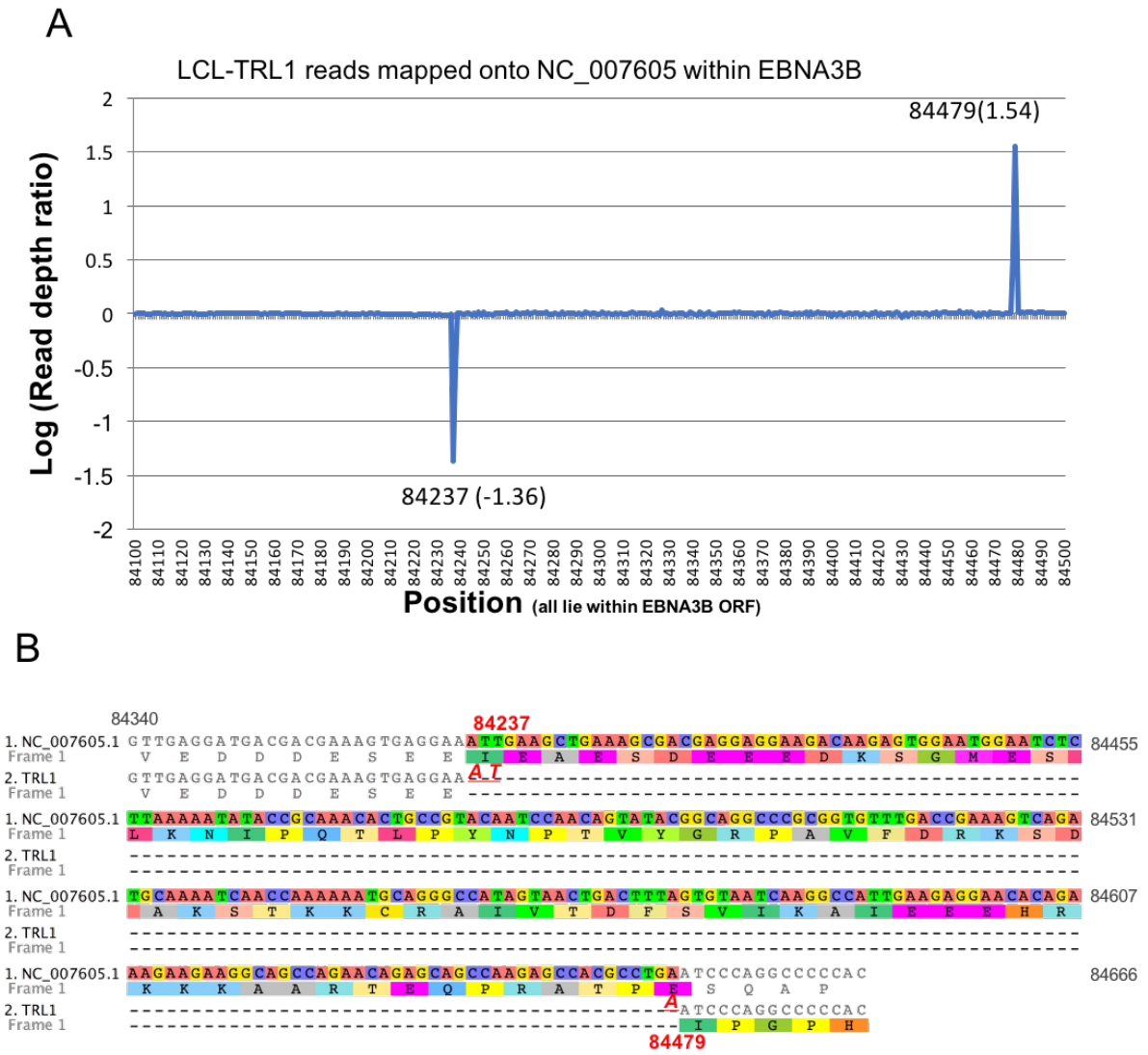


Figure 3.3 Testing the ability of RDA to predict deletions in sequences.
(A) The graph shows the ratio of read depths at every adjacent nucleotide in a sliding window manner of 2 bp. The Y-axis shows the read depth ratio in log space while the X-axis shows the position in the prototype EBV genome (accession number NC_007605). The TRL1 sequence reads were mapped against the EBV genome (NC_007605). Numbers on RDA signal peaks show the reported positions with extreme read depth ratio (between brackets), indicating the start and end of the deletion. It is worth noting that this figure is for visualization purposes while in practice, the flagging position is performed computationally based on the specified threshold (log2 of ratio exceeds ±0.5). **(B)** Sequence of start and end points of the deletion in EBNA-3B of TRL1 reported by Gottschalk et al. 2001 and highlighted by RDA in comparison to the prototype EBV genome (NC_007605). Sequence positions are presented on the right, numbers in red are positions of the start and end of the deletion flagged by RDA (shown in fig 3.3A). Nucleotides in red at the edges of the deletion are not shown in Gottschalk et al. 2001 which might be because Gottschalk et al. showed the sequence of the amino acids and ignored these extra nucleotides.

The integrity of all constructed templates was verified by RDA. Namalwa (a type 1 EBV isolated from an African Burkitt's lymphoma cell line) had a reduction in the read depth at the beginning of Y region, highlighting an issue with BamHI Y region in this strain. Namalwa is known to express EBNA-LP (low level compared to other tested strains though) and EBNA-2 (Cao et al. 2015). As a result, Namalwa sequence reads were mapped to the standard EBV genome type 1 sequence NC_007605, and RDA showed a dramatic increase in read depth ratio at position 40158. The region upstream of this peak exhibited fluctuating read depth ratios (due to low read depth), suggesting that the Namalwa sample had a deletion downstream of position 40158 (figure 3.4). It was easy to confirm the RDA output by looking at the numbers of mapped reads from the position of the end of IR1 region 35356 to 40158. The number of mapped reads covering the query region was from 1 to 350, while the read depth at the end of the questionable region was 2562. Notably, this deletion was identical to that of P3HR1, suggesting that the ‘Namalwa' DNA was in fact from a mis-labelled P3HR1 cell line. This suspicion fits in with a matrix of SNP differences between all available EBV genome sequences (Nick Grayson, Sanger Institute, unpublished) that showed that Namalwa was very similar to Jijoye and P3HR1. Sequence analysis of Rael raised the likelihood that it was also mislabelled P3HR1. As a result, ‘Namalwa' and ‘Rael' were removed from my project and from Palser et al. 2015. Such false positive reads arise through sample cross-contamination during DNA isolation. Alternatively, barcode exchange and ‘jumping PCR' during library preparation and sequencing, can cause misassignment of sequence but can be avoided by double barcoding (Kircher et al. 2012). Therefore, specific attention is obviously needed to avoid missing deletions in virus strains during alignment.

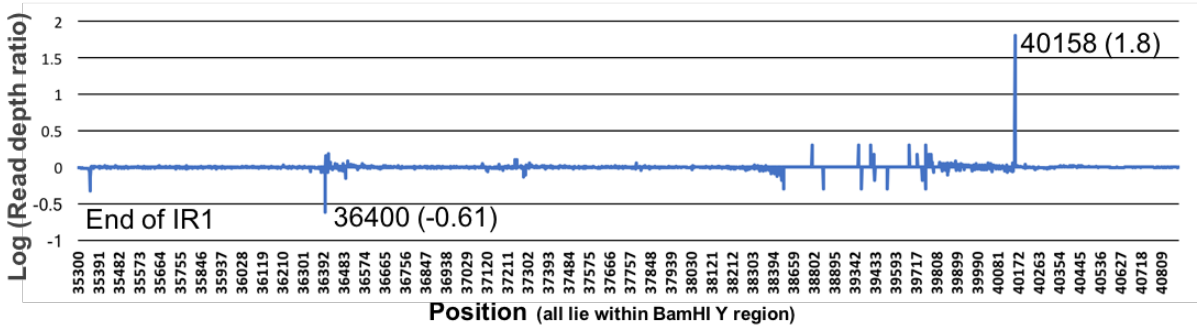


Figure 3.4 Read depth analysis predicted low read depth within BamHI Y in Namalwa. The Namalwa sequence reads (mis-labelled P3HR1) were mapped against the prototype EBV genome (accession number NC_007605). The graph shows the ratio of read depths at every adjacent nucleotides in a sliding window manner of 2 bp between position 35300 and 40809. Y-axis shows the read depth ratio in log space X-axis shows the position in EBV genome (NC_007605). Numbers on RDA signal peaks show the reported positions with extreme read depth ratio (between brackets). This figure is for visualisation purposes while in practice, flagging position is performed computationally based on the specified threshold (log2 of ratio exceeds ±0.5).

3.4 Read depth analysis identifies the junctions between unique sequences and IR1

Verifying the integrity of the constructed theoretical templates (section 3.2. and 3.3.) showed that, in general, the junctions between IR1 and flanking unique regions are structurally conserved, except where there are large deletions of the flanks (table 3.1). Some strains have previously been reported to have a deletion incorporating part of IR1. For instance, P3HR1 and Daudi are known to have a deletion downstream of IR1, including EBNA2 (Rabson et al. 1982; Jones et al. 1984), and X50-7 has a deletion of Cp upstream of IR1 (Yandava & Speck 1992). In some cases where strains had genuine deletions, I found that the deletions were supported by the reads, allowing contig assembly and alignment to produce an intact sequence. For instance, contig assembly showed reads supporting EBNA2 in P3HR1-HH514- C16 (figure 3.5).

Table 3. 1 Strains with deletions in unique sequences flanking IR1.

Strain name	Wild type junction	
	BamHI C-IR1	IR1-BamHI Y
X50-7	Complete deletion	Intact
L591	Partial deletion in C2	Intact
HL01	Complete deletion	Intact
Cheptages	Complete deletion	Complete deletion
Daudi	Intact	Complete deletion
Rael	Intact	Complete deletion
P3HR1_c16	Intact	Complete deletion
AFB1	Complete deletion	Intact

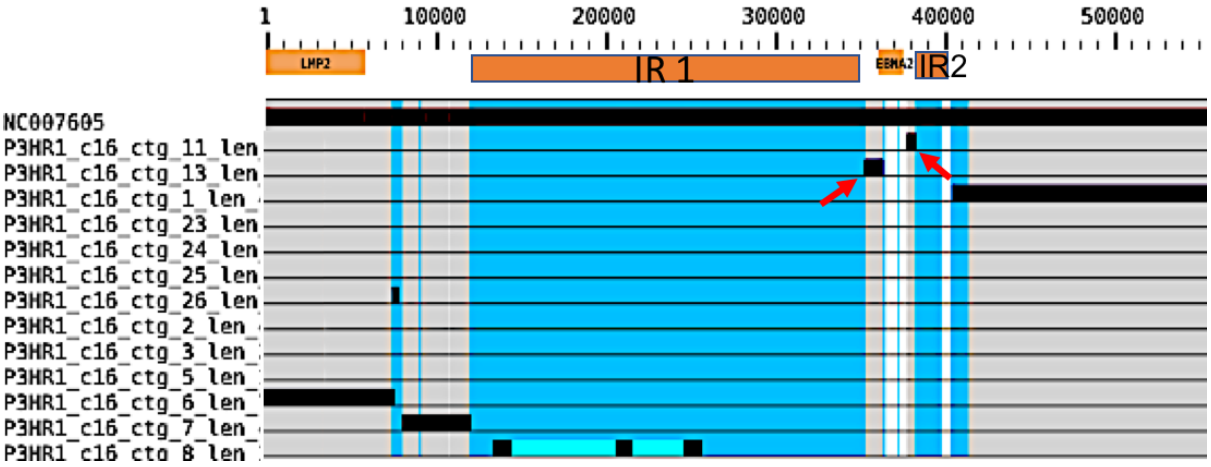


Figure 3.5 Contig assembly of P3HR1.
A diagram of contigs assembly of P3HR1 showing contigs in the IR1 region. The top black bar is the reference NC-007605. Blue areas are repeat sequences. Contigs mapping to the reference are in shown in black bars. The contigs mapping to BamHI Y are indicated by red arrows. (Figure was produced by Nicholas Grayson).

In order to validate deletions and identify the defective junctions, in which complete templates could not be generated from contigs assemblies, reads were aligned to the whole genome reference (NC_007605 for type 1, or NC_009334 for type 2) and then RDA was conducted. RDA was able to identify known deletions in non-repetitive regions efficiently, but

it had less sensitivity within repetitive regions due to the abundance of multi-reads. In addition, it always showed a disruption in read depth at the edge of the repeat unit (figure 3.6). Therefore, to identify the real unique/IR1 junctions, the mapped reads at positions where the log2 of read depth ratios exceeded ± 0.5 were visualised using Geneious to identify partially trimmed reads (Figure 3.7). A trimming process is applied before genome assembly to keep the highest quality sequence within reads and discard low-quality sequences (Del Fabbro et al. 2013). Using the standard EBV genome as a reference for alignment, reads from the mutant junction will be mapped partially to either IR1 region or unique sequence. By scanning X50-7 reads, for instance, the trimmed sequence of reads mapped to the unique sequence match the aligned sequence to a region within IR1 and vice versa (Figure 3.7A and B) which defines the actual junction between the unique sequence and IR1.

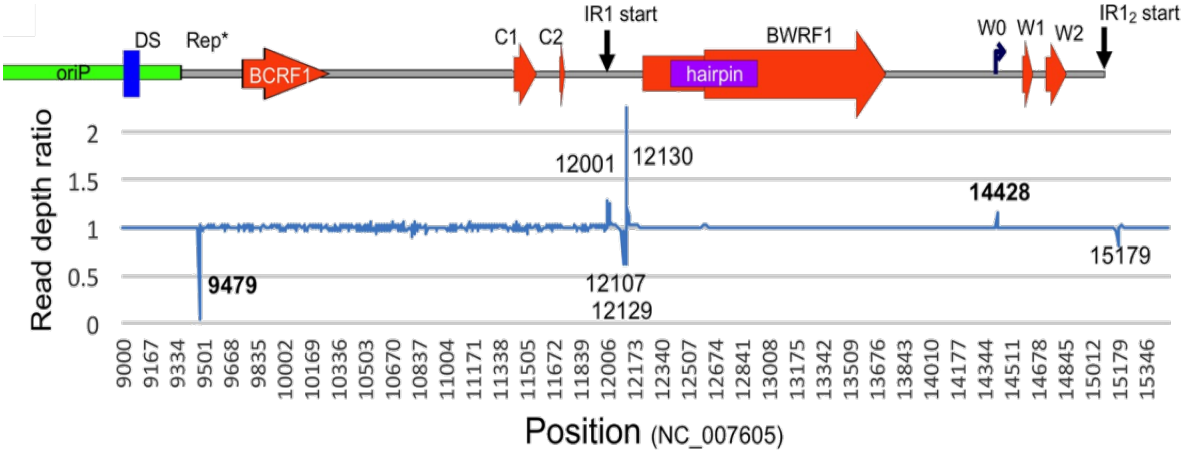


Figure 3.6 Read depth analysis to identify defective junctions.
The graph visualizes the ratio of read depth at every adjacent nucleotide from X50-7 reads mapped to the prototype EBV genome (NC-007605) on the Y-axis, with the position on the X-axis. The schematic above the graph depicts the features of the locations. Numbers on RDA signal peaks are positions on the genome, and those in bold are the edges of the X50-7 deletion. RDA is capable in non-repetitive regions as it can be seen in signal strength and noise in read depth upstream of IR1 while within IR1 peaks are smaller and the noise becames even due to dilution of the signal by the abundant of reads from other repeat units, which explains the high peak at the beginning of IR1 region.

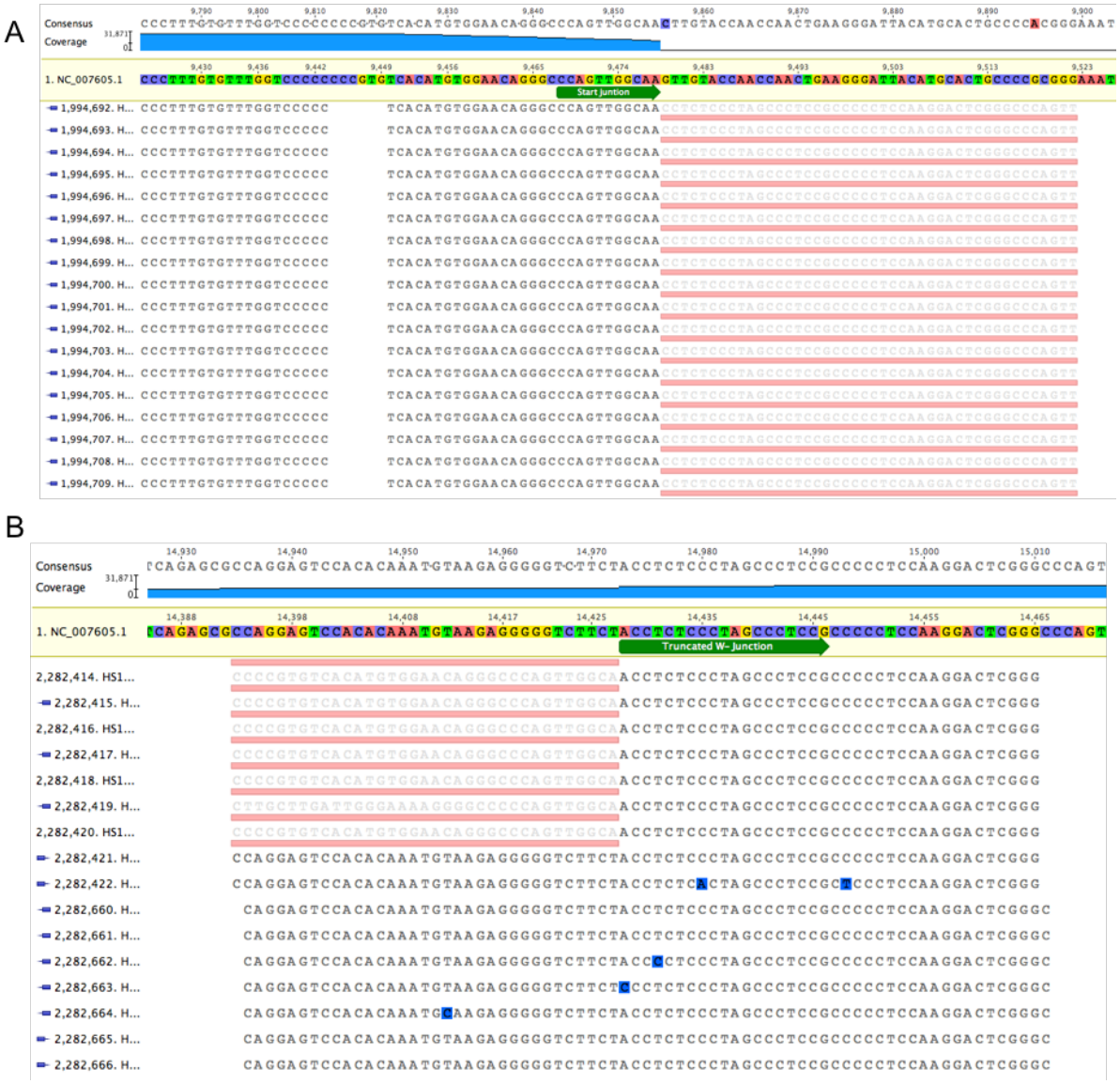


Figure 3.7 Visualisation of a deleted junction within IR1 by scanning mapped reads. **(A)** Partially trimmed reads of the X50-7 strain aligned to whole genome reference (NC_007605.1) around the RDA peak at position 9479 (upstream of the IR1 repeats – see fig 3.6). The trimmed sequences are actually from the W repeat. **(B)** Partially trimmed reads of X50-7 strain aligned to whole genome reference (NC_007605.1) within IR1 repeats region from 14393 to 14465. The trimmed sequence perfectly matched the untrimmed sequenced showed in (A) indicating the sites of the junction between IR1 repeat and unique sequence. Trimmed sequences are underlined in red. The blue bar above the reference sequence shows the read depth of each nucleotide. Note the sudden drop in read depth at the point of trimming in [A], and the smaller drop in [B] are due to the abundance of intact reads from other repeat units diluting the effect of the junction on read depth.

This approach corrected the erroneous deletion in the HL01 sequence, which was likely due to improper contig assembly rather than a genuine deletion. By scanning positions indicated by RDA, it was possible to identify the defective junction in all strains but Cheptages. RDA could spot the region of junctions but not the precise positions as the deletions on both sides of IR are in the repeat region. The upstream junction occurred between the family of repeat of *ori*-P and truncated IR1 unit while the downstream junction occurred between truncated IR1 unit and IR2. The precise positions were identified by scanning sequence reads at reported regions by RDA. The AFB1 deletion could not be solved as it had a human sequence in sequence reads, that may represent incorporation with the human genome.

The strains that were tested and deletion junctions are summarised (according to the equivalent genome positions in the consensus type 1 strain NC_007605) in (Table 3.2) and depicted in (figure 3.8). The deletion in X50-7 is upstream of the W0 exon, while the other 4 strains have deletion junctions <1kb upstream of the hairpin, and downstream of the W0 exon. No strain had interruptions in the hairpin, in the BWRF1 ORF or in Wp. However, since there is no known function for the hairpin or the BWRF1 region, it is not clear whether a biological function makes interruption of these regions un-tolerated.

Table 3.2 Deletions overlapping the junctions of IR1.

Strain	Upstream (BamC) deletions[a]		Downstream (BamY) deletions[a]	
	Upstream	IR1	IR1	Downstream
X50-7	9479 (in oriP rep region)	14428 (after exon W0)		
L591	11632 (in exon C2)	12193 (50 bp before hairpin)		
Cheptages	7777 (in oriP FR)	14669 (in W1-W2 intron)	33604 (between W2 and hairpin)	38301† (within IR2)
Daudi			33127 (W1-W2 intron)	40536 (oriLyt upstream element)
P3HR1			33351 (after exon W2)	40157 (nr BHLF1 start)
HL01	Intact (it was misassembled)		Intact	

a Strain NC007602 equivalents of first/last base of the deletion, above position relative to EBV features

† First position within repeat: sequence exists every 125bp, so actual junction is uncertain

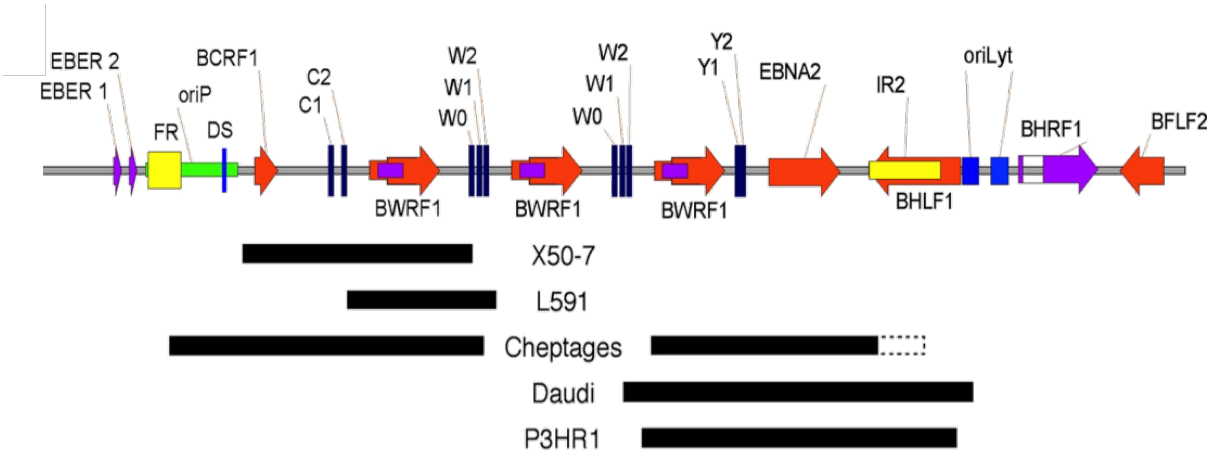


Figure 3.8 Schematic representation of deleted regions in strains with defective junctions upstream or downstream of the IR1 region.
The schematic of EBV genome features includes 1,6 copies of IR1 to simplify the presentation. Black boxes show the deleted regions; the dotted box in Cheptages indicates uncertainty over the position of the end of the deletion as the junction lies within a region where the sequence recurs every 125 bp. Precise positions of the deletions are in table 3.2. (Figure was produced by Dr Rob White).

3.5 IR1 has conserved regions including Wp, W2 and regions of ORF BWRF1

To check how the IR1 sequence differs among EBV strains, and to assess which parts of IR1 are the most and least diverse, consensus sequences were aligned to generate an MSAof generated templates from the recently assembled strains in Palser et al. 2015 (76 isolates) along with 9 previously published strains (85 strains in total, Appendix 1).

Indels in the repeat units are primarily restricted to two regions. First, is within the BWRF1 ORF (NC_007605, position 13210-13240), where Makau and sLCL1.19 have >11bp deletions while several other strains have 1-2 bp deletions. The second position with apparent indels is between exon W0 and W1 (NC_007605, position 14440-14500) where there are two distinct base indels; a deletion at 14447 exists in 7 strains (all of them are sLCL-type) whereas the other deletion at 14479 is in 9 cancer-derived samples (including 3 B95-8-derived samples) and one sLCL-IS type 1 sample.

The degree of diversity between strains within IR1 region was assessed within a 100 nt sliding window. The distribution of single nucleotide polymorphisms (SNPs) in IR1 (figure 3.9) shows a considerable variation density across the repeat. The highest SNP density within IR1 is 12 SNPs/100 bp. Most varieties are located within the BWRF1 ORF. Conversely, perfectly conserved regions include sequence immediately upstream IR1's hairpin structure (position 12225-12350) as well as within the last third of the BWRF1 ORF (13290-13450), immediately downstream of the deletions in Makau and sLCL1.19 described above. Also, the Wp promoter region (including exon W0) is 100% conserved. The W2 exon and the intron between W1 exon and W2 exon (which produces a stable RNA – EBV-sisRNA-1) are highly conserved (only one strain has one SNP). The conserved sequence of EBV-sisRNA-1 across

studied strains supports the suggestion that this intron is evolutionary conserved and might be of functional importance (Moss & Steitz 2013).

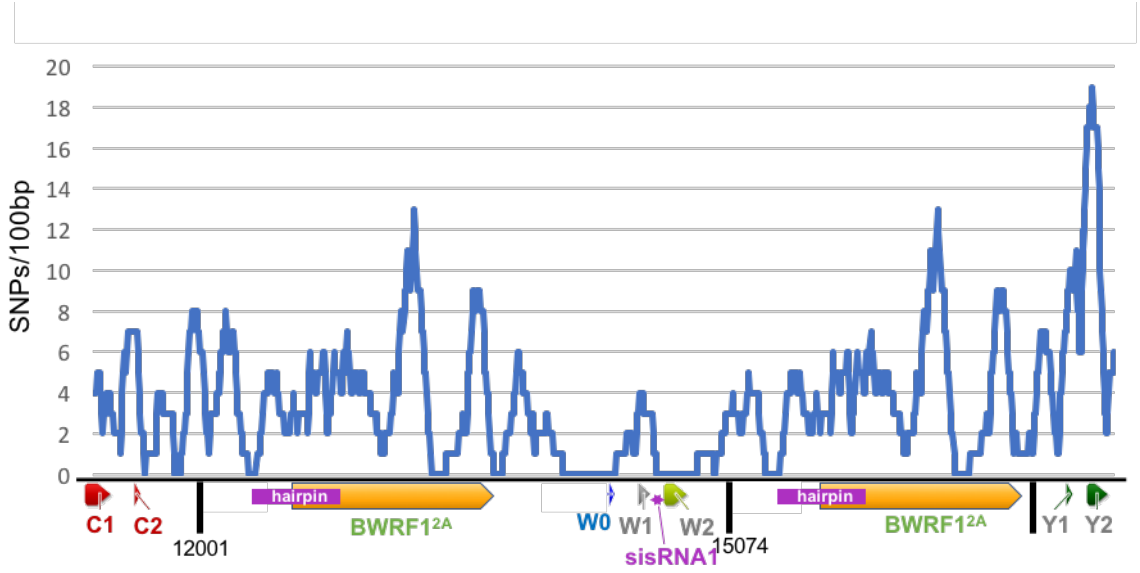


Figure 3.9 . Polymorphism density in IR1 region across 85 strains.
The line graph is plotted across the theoretical template showing the number of positions with SNPs found in more than one strain in 100-bp sliding window. Arrows underneath the plot indicate positions of exons/ORFs, and dark vertical lines represent edges of IR1 repeat units. (list of strain names is available in Appendix 1, while MSA data are available in SD3.1 file at http://www.epstein-barrvirus.org.uk/MBaData.php).

3.6 The length of the BWRF1 ORF is not the same in all EBV strains

In general, ORF identification is a way of predicting protein coding genes. IR1 units include a long open reading frame BWRF1 of unknown function. In B95-8, the prototype lab strain in which BWRF1 was identified (Cheung & Kieff 1982; Jones & Griffin 1983; Baer et al. 1984), it is a hypothetical ORF as It has no start codon and it has never been found on an mRNA (only as intron from Cp and Wp transcripts). There is no evidence of a protein being made. Since conservation of the ORF may highlight a potential importance, the question of this part: how broadly conserved is BWRF1?

As described in the previous section, BWRF1 is the most variable region of IR1 (figure 3.9). Because there is no start codon "ATG" at the beginning of BWRF1 in B95-8, I defined BWRF1 ORFs by calculating the maximum possible distance between in-frame stop codons within the IR1 unit for each strain using Geneious. These ORFs fall into two main lengths (models 1 and 2) based on an SNP that produces either TAC or TAA (stop codon) at position 12540 (NC_007605). This means that the ORF starts either after the TAG at NC_007605 position 12151 creating a long ORF, or after the TAA at position 12540, making a 390 nt shorter ORF, as in B95-8 (Jones & Griffin 1983) (figure 3.10, Appendix 3). The BWRF1 ORF of model 1 spans the hairpin while model 2 starts in its 5' side. There is a considerable gap in the frequency of BWRF1 models between the first and second most frequent models, which are labelled as model 1A and 2A respectively (figure 3.10). The smaller group (model 2A), with a 384 codon long ORF, exits in only 12% of studied strains, including the prototype lab strain B95-8. Model 2B is 3 nt shorter than Model 2A, resulting in 8 mismatched codons (no frameshift), and seems to be restricted to EBV type 2. The most frequent BWRF1 model 1A, which is 514 codons long, was found in around 60% of the strains including the one sequence direct from saliva. C666 and TRL1 have model 1A and 2B respectively, and they have a minor variant at position 12540 (NC_007605) TAM (where M is either C or A – see section 3.8), indicating that number of IR1 units (most probably only one copy) in C666 and TRL1 have BWRF1 model 2A and 1A respectively.

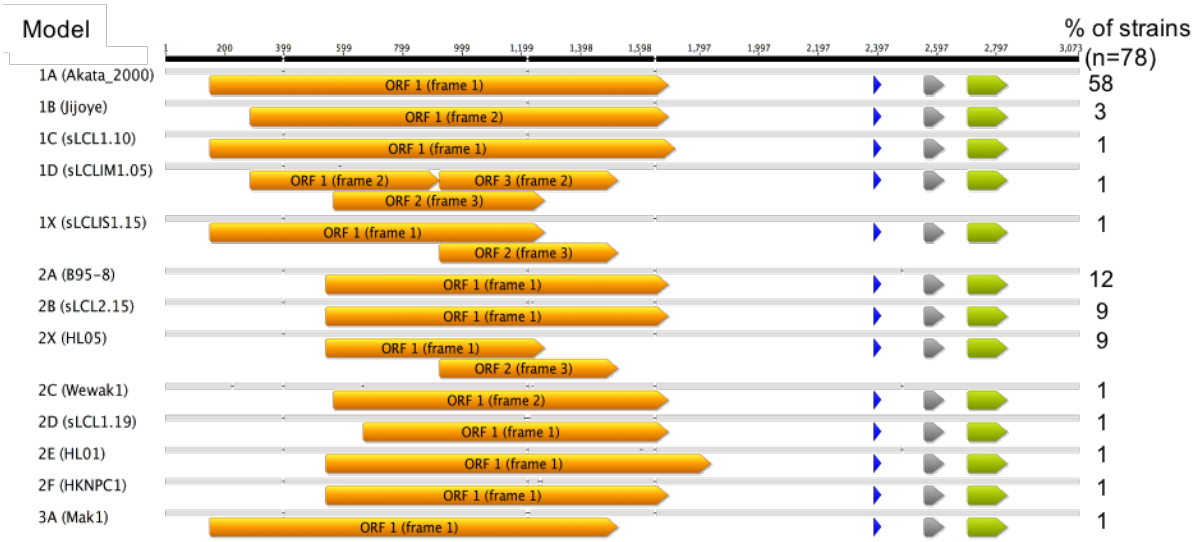


Figure 3.10 BWRF1 models.
Because of the lack of a canonical start codon in BWRF1, ORFs are defined at their maximum possible length between in-frame stop codons. ORFs over 600 bp are shown. BWRF1 model names with examples are on the left while the percentage of frequency of each model is shown to the right. Model of BWRF1 in all strains are listed in Appendix 3.

Among the 78 distinct strains there are nine strains (models 1X, 2X, and 1D) that have an insertion at position 13219 causing a frameshift that disrupts the BWRF1 ORF (Appendix 3). This disruption suggests that the presence of an intact full-length BWRF1 ORF is not essential, at least for the lifecycle of these strains. These strains with disrupted BWRF1 ORFs (except sLCL-IS1.15) form a distinct clade on the phylogenetic tree of the IR1 template sequence (figure 3.11), implying that this is a genuine circulating variant.

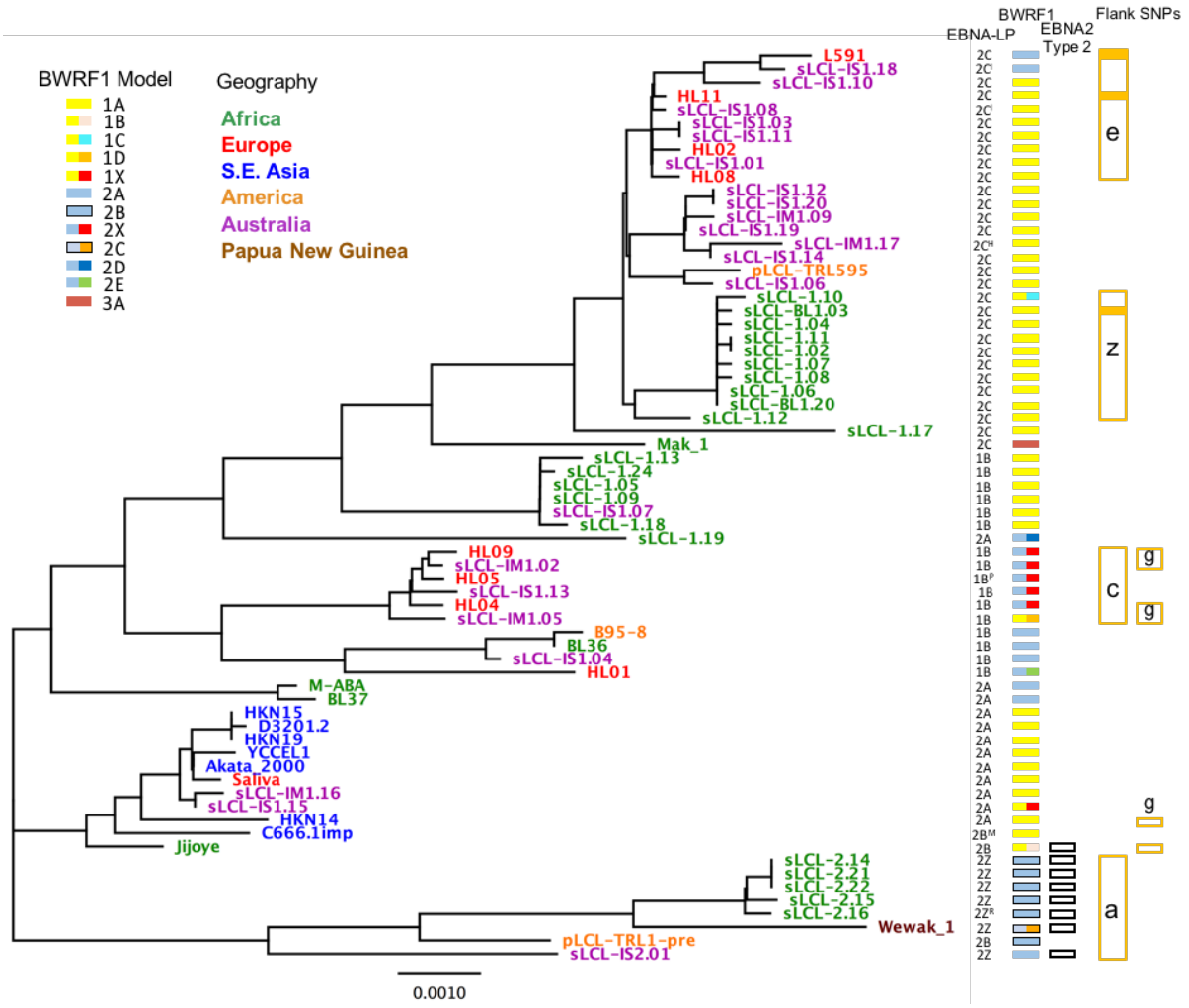


Figure 3.11 Phylogenetic tree classified by IR1 diversity.
The distance-methods phylogenetic tree (calculates pairwise distances between sequences) of the constructed theoretical template shows the degree of differences between strains via branch length. Strains that are descendents of others and those with defective IR1 junctions were excluded from this analysis. The scale bar represents the branch distance corresponding to 1 nucleotide substitution/kb. The geographic origins of samples are shown by the colours of the strain names (colour coding is presented at the top of the tree). To the right of the tree, for each strain, EBNA-LP subtype is shown by an alphanumeric designation (figure 3.14, table 3.3). BWRF1 subtypes are shown in coloured boxes. The left side of each BWRF1 box is colour coded for the three major groups (1, 2, and 3) (see Appendix 3), and strains with the indels characteristic of type 2 are indicated by a black border. The right part of the BWRF1 box indicates BWRF1 subgroups (red to indicate the disrupted BWRF1 ORF in 1X and 2X). EBV type 2 strains are marked with empty black boxes. Orange boxes show strains with common SNPs in the flanks and the letter designates which flank SNP group (see section 3.10 and Appendix 5). The solid orange boxes represent strains in which the flanking SNP has propagated throughout IR1.

The other models with less frequency or found in one strain that carries indels that change the potential start and/or end positions of ORF, but the length of BWRF1 ORF remains > 1kb. However, it remains unclear whether they represent genuine circulating variants, or are peculiar to the cell line in question. It is worth noting that all full-length BWRF1 models (except Model 3A which is only in Mak1) end at the same stop codon TGA at position 13692 (according to the first IR1 in NC_007605) and all studied strains have a stop codon (either TAA or TAG) at position 12150 (in NC_007605) suggesting that the wild-type BWRF1 could be Model 1A.

Since BWRF1 is considered to be ORF, I wanted to see to what extent it is conserved at the amino acid sequence level. For this, I tested the diversity within BWRF1 across all analysed strains (excluding those with insertions or deletions that are not divisible by three as such indels create a frameshift) at nucleotide and amino acid level. A multiple sequence alignment of BWRF1 protein sequences and diverse density analysis were performed. Differences in BWRF1 model lengths leading to missing N or C termini in the MSA were filled with gaps ignored in the SNP density analysis. To keep the size of the sliding window the same for both nucleotides and amino acids sequence analysis, sliding windows of 90 nt and 30 aa were used respectively (figure 3.12). The conserved region within the final third of the BWRF1 ORF is longer at the amino acid level as some diversity is due to silent mutations. The conservation of this region may highlight the need for further studies on BWRF1 ORF.

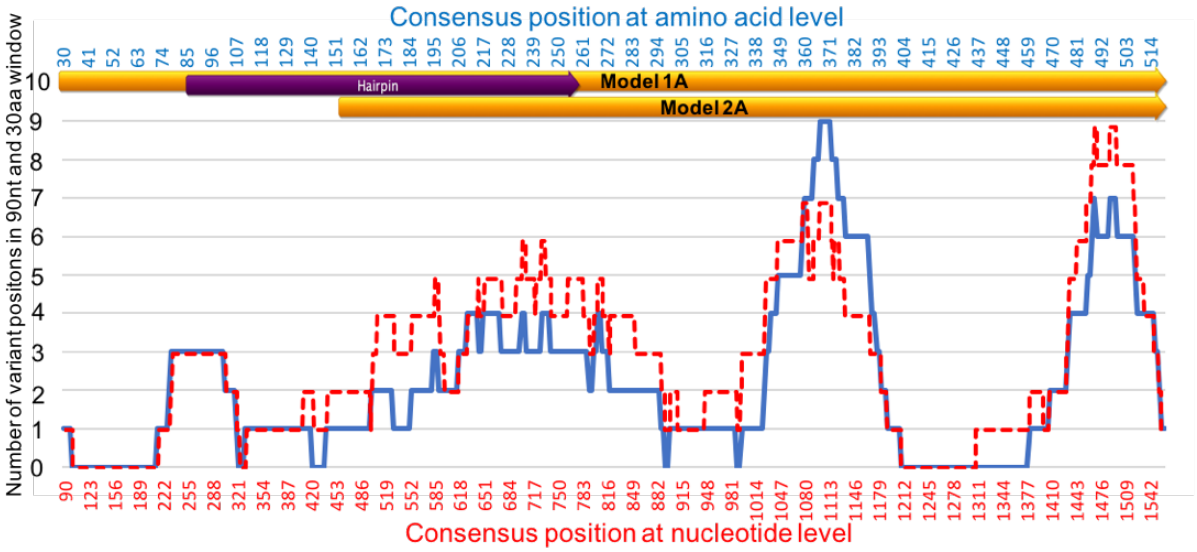


Figure 3.12 Diversity within BWRF1 at the nucleotide and amino acid sequence level. Dotted and solid lines are plotted across BWRF1 (model 1A from 13290-13450) for 62 strains (strains with insertions or deletions that are not divisible by three were excluded as this indel created a frameshift). The red dotted line shows diversity at the nucleotide level, while the blue solid line shows the difference at the aa level in a sliding window of 90 nt and 30 aa, respectively, to make the two assays comparable. Positions are indicated in red for nucleotides at the bottom and in blue for aa at the top. The schematic shows the start and end of BWRF1 model 1A and 2A. Although diversity in aa is supposed to be like or less than that in nucleotides, the diversity in aa is higher between 340 and 370 because of three single nucleotides deleted in sLCL type 2 strains (model 2B), which created a different aa sequence (no frameshift).

3.7 The hairpin is retained within IR1

The hairpin region (a 512-nt with a repeated CGGGCACCC motif at termini) appears to be undisrupted and retained in all analysed strains (see appendix 4 for the secondary structure of the hairpin within B95-8 IR1 units). SNP density within the hairpin region is moderate, ranging from 2 to 5 SNPs/100bp, although the sequence immediately upstream of the hairpin is highly conserved (figure 3.9 and 3.12). To assess the conservation of the hairpin, the sequence of the hairpin was aligned to its reverse complement (Supporting data file SD3.3 at http://www.epstein-barrvirus.org.uk/MBaData.php). There were some polymorphisms that disrupted theoretical base pairs within the hairpin, but there were no examples of

polymorphisms on one strand matches complementing polymorphism on the other strand (figure 3.13 and appendix 4). In general, the hairpin is retained within IR1 units in all studied strains there are minor variations across strains. The conserved structure of the hairpin may offer a selective advantage for EBV. However, the importance of co-conservation of bases is still unknown.

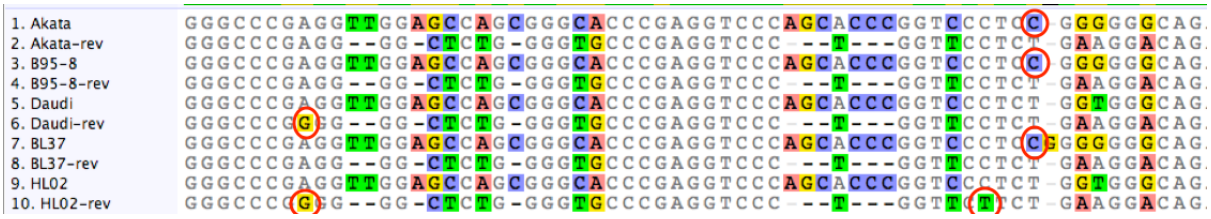


Figure 3.13 Disruption in the hairpin within the IR1.
Visualizing part of a sequence of the hairpin for several strains after aligning the sequence of the hairpin to its reverse complement. There are polymorphisms disruption theoretical base pairs within the hairpin (coloured nucleotides), and there is no polymorphism on one strand that has a complementary on the other strand (shown in red circles). (Sequences are available in file SD3.3 at http://www.epstein-barrvirus.org.uk/MBaData.php and the secondary structure of the hairpin within the B95-8 IR1 is shown in appendix 4).

3.8 EBNA-LP exons are more variant than were previously reported and have restricted combinations

The repetitive domain of EBNA-LP is encoded across the IR1 region, in which a number of the repeating exon pairs W1 and W2 (typically 2-7 copies) at the N-terminus are combined with unique exons Y1 and Y2 at a non-repetitive C terminal domain (see figure 1.7). EBNA-LP has been identified as two distinct isoforms called type 1 and type 2 variants, based on the presence of G8/T12 or V8/A12 in exon W1 respectively (Finke et al. 1987). However, many

EBV type 1 strains have the type 2 variant (Supporting data file SD3.1 at http://www.epstein-barrvirus.org.uk/MBaData.php).

Using the MSA for 74 distinct strains, the diversity of EBNA-LP sequence was examined and is summarised in figure 3.14 with a new nomenclature for EBNA-LP subtypes proposed in table 3.3. Our analysis confirmed that only two amino acids varied in the W1 exon, although four of the seventeen G8/T12 strains encode glycine with a GGT codon, rather than the usual GGC. Across the 74 strains, only one has an SNP in exon W2: an African EBV type 2 sLCL2.16 with a Q54R change. Arginine at this position is also found in the EBNA-LP sequence of chimpanzee lymphocryptovirus (homologue of EBV chimpanzee) (Ling et al. 2009), so it may represent a genuine circulating variant.

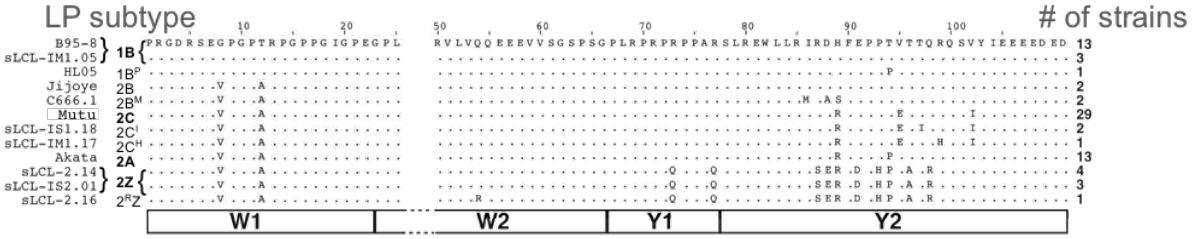


Figure 3.14 Variants in the amino acid sequence of the EBNA-LP protein.
Amino acid sequences of the EBNA-LP protein encoded by the viruses are presented relative to the prototype lab strain B95-8 (dots indicate no difference). The protein subtypes according to the proposed nomenclature in table 3 with an example strain are shown to the left (with the commonest subtypes in bold). The number of distinct strains (out of 74 strains) encoding each LP subtype is indicated to the right. The EBNA-LP subtype of each strain is shown in Appendix 1. Numbers at the top of the consensus sequence indicate the position according to a W1W2Y1Y2 exon structure which is shown as boxes below. The gap in the sequences is omitted amino acids from 26 to 49 as they are identical in all strains. (Figure was produced by Dr Rob White).

Table 3. 3 Subtypes of EBNA-LP.

Example strain	Subtype for region				Overall subtype	# of strains (n=74)
	W1	W2	Y1	Y2		
B95-8	1.1	1	1.1	B	1B	13
sLCL-IM1.05	1.2	1	1.1	B	1B	3
HL05	1.2	1	1.1	B-P	$1B^{P}$	1
GD2	1.1	1	1.1	B-M	$1B^{M}$	1
Jijoye	2	1	1.1	B	2B	2
C666.1[b]	2	1	1.1	B-M	$2B^{M}$	2
Mutu	2	1	1.1	C	2C	29
sLCL-IS1.18	2	1	1.1	C-I	$2C^{I}$	2
sLCL-IM1.17	2	1	1.1	C-H	$2C^{H}$	1
Akata	2	1	1.2	A	2A	13
sLCL-2.14[c]	2	1	2.1	Z	2Z	4
sLCL-IS2.01	2	1	2.2	Z	2Z	3
sLCL-2.16	2	1^{R}	2.2	Z	$2^{R}Z$	1

Adjacent highlighted rows differ in nucleotide sequence but are identical at the amino acid sequence.

[b] includes the M81 strain.

[c] Includes prototype type 2 EBV strain AG876.

In contrast to the W exons, the Y exons of EBNA-LP exhibit more diversity than has previously been described (Figure 3.14, table 3.3). There are four major sequence subgroups of the Y exons defined mainly by exon Y2 as it is the most variable part of EBNA-LP, whereas the only Y1 amino acid variation is the two aa in type 2 EBV. Subgroups; A, B, and C are in type 1 viruses, whereas subtype Z was found in type 2 EBV. Subgroup A was previously found in Akata (McCann et al. 2001) and contains a synonymous polymorphism in exon Y1. Subgroup B is found in the prototype B95-8 strain, and group Z is found in type 2 EBVs (Dolan et al. 2006). Surprisingly, subgroup C, whose V95E and V102I changes distinguish it from the other subtypes, is novel and is the most prevalent subtype, being found in both African and European strains, including the Mutu cell line (Lin et al. 2013). Several of these four main subtypes have rare variants. In particular, three NPC-derived samples (M81, C666.1 and GD2)

share a common IRDH to MRAS change at amino acids 86 to 89. The published GD2 is reported to have a type 1 W1 exon, whereas M81 and C666 both have type 2 W exons. However, while our C666 sequence agrees with that of M81, a previously published partial C666 sequence reported MRDH at this position (Tso et al. 2013), while another C666 sequence (Accession AB828190) contains the MRAS motif, although the genome region containing the Y exons is inverted in this sequence (Xiao et al. 2016). Other less common variants are indicated in figure 3.14, and the designations for all strains are given in appendix 1.

Looking at table 3.3, there are combinations between exon subtypes. For instance, the type 1 W1 exon is only found with a B subtype of Y exons, whereas type 2 appears to be compatible with all the Y exon subtypes. Conversely, Y exon subtypes A, C and Z require type 2 W1, whereas subtype B tolerates both type 1 and type 2 W exons. These relationships remain to be assessed whether they reflect historic recombination or some sort of functional incompatibility.

All the variations and combinations above were visualised against a phylogenetic tree of IR1 template sequences (figure 3.11). Some anomalies were obvious, such as pLCL-TRL1 which has EBNA-LP 2B, but it lies on the type 2 branch of the phylogenetic tree. It is also curious that the type 2 EBV strain Jijoye has a 2B EBNA-LP genotype that has otherwise been seen only in type 1 strains, rather than the 2Z that might be expected with a type 2 EBNA2 and EBNA3s. The IR1 of Jijoye is closest to that of the Akata branch of the phylogenetic tree (Figure 3.11). If this combination were found in circulating (i.e. non-disease associated) EBVs, it would suggest that EBNA-LP and EBNA2 types do not exhibit sequence co-dependency. However, this mixture may represent a disease-associated or lab-acquired mixture of two strains rather than a genuine example of a circulating virus strain.

Based on these observations, a new nomenclature for EBNA-LP was proposed, to classify the different EBNA-LP sequences into their major subtypes. This nomenclature system comprises an indicator for the W1 polymorphism (1 or 2) combined with a polymorphism for the Y exons (A, B and C for type 1 virus variants, and Z for type 2 (figure 3.14 table 3.3 and Appendix 1 for all strains). Further variants within these main subtypes can then be indicated by superscripted letters indicating a changed amino acid that is characteristic of their subtype. For instance, the EBNA-LP variant found in the NPC-derived C666 and M81 virus strains is type $2B^{M}$, with M representing the MRAS motif in exon Y2. Comparing both the BWRF1 and EBNA-LP variants in the phylogenetic tree (figure 3.11), it is apparent that certain EBNA-LP and BWRF1 subtypes usually co-segregate in a manner that defines the genetic variation of this region. Although there is not always true linkage between EBNA-LP and BWRF1, most strains with EBNA-LP subtypes 2A and 2C have the longer BWRF1 ORF (model 1), while subtype 2Z (type 2 EBV) is associated with the shorter BWRF1 ORF (model 2). In contrast, EBNA-LP subtype 1B is found in combinations with BWRF1 models (1A,2A and 2X) with equal frequency. Since there is no known function for BWRF1 ORF, the correlation between BWRF1 and EBNA-LP types can currently be considered only as reflecting broader sequence conservation/linkage, rather than highlighting a functional interdependence.

3.9 Developing a methodology to distinguish heterogeneity within repeats from technical errors in sequencing

NGS platforms generate a huge amount of data, but these data are not error-free. Errors in NGS are either due to the principle (i.e. the technology) used in the NGS platforms, which is inevitable, but the resultant errors can be predicted, or the samples preparation issues (e.g. contamination or low DNA quantity input) (Robasky et al. 2014; Guo et al. 2012). In general,

reads undergo several quality control steps to exclude all low-quality reads before mapping read to a reference or producing de novo contig assemblies. Downstream analysis such as SNP calling uses read depth and read bias as factors to discriminate genuine SNPs from false-positive SNPs.

In this project, raw data had already undergone quality control analysis (Palser et al. 2015). RDA (section 3.3) was able to predict contaminated samples through a sudden change in read depth within non-repetitive regions. However, one of the challenges of attempting to sequence large repeats with short read data is that low-frequency variations within repeats are missed by simple consensus alignment or assembly of contigs. Using a reference with a minimum number of repeat units (i.e. the theoretical template) forces reads from repeat units with different degrees of similarity to map to the same position. Thus, variations within repeats (i.e. differences between repeat units within a single strain – see schematic figure 3.15) will be considered by variant calling tools as nonconsensus nucleotides at a frequency proportional to the percentage of repeat units containing the alternative nucleotide. For clarity, lower frequency base calls within IR1 are described herein as “minor variants” (MVs). In other words, MVs are differences in sequence between repeat units within a given strain while SNPs are differences in sequence between different strains.

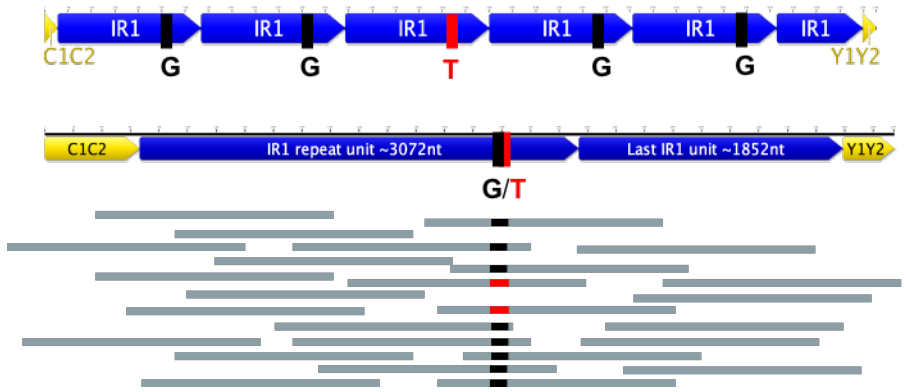


Figure 3.15 Mapping reads to the theoretical template.
The schematic shows IR1 in which 1 of 5 repeat units contains the minor variant T while the other have G at the same position. If the reads (gray bars) from sequencing of this repeat are mapped to the consensus template (the representation of IR1 containing 1.6 repeats), 80% of reads will be G (black dots) and 20% will be the minor variant T (red dots).

3.9.1 The quality of minor variants can be defined by their frequency and read bias

Obviously, it was necessary to distinguish genuine MVs from the various potential sources of sequencing error or sample contamination. After mapping reads to consensus templates, VarScan (mpileup2snp command) (Koboldt et al. 2012) was executed to report minor variants with minimum variant allele frequency threshold [1%]. VarScan calls minor variants from pileup files with consideration for read bias (i.e. checking for equivalent frequencies of the SNPs on + and - strand sequences) and ignores variants if >90% of supporting reads are mapped to one strand (+ or -). However, Varscan does not test whether +/- ratios of consensus and alternative are quite the same. This ratio matters to clarify whether the read bias was due to the nature of the sequenced DNA or it was a technical error. As a result, Varscan was used to report the numbers of reads mapped to forward and reverse strands for both consensus and variant bases. Then I wrote a script to calculate the ratio of (+) and (-) reads supporting consensus and variant. This is based on the assumption that variants caused by technical errors in the sequencing will be due to sequence features that are specific to one strand, whereas genuine diversity within the repeats will be at the same frequency on both strands (Meacham et al. 2011).

The calculation I used was:

If VR(+)> VR(-) strand bias score = $\frac{(DR+/VR+)}{(DR-/VR-)}$

Otherwise strand bias score = 1/ $\frac{(DR+/VR+)}{(DR-/VR-)}$ (to keep the ratio <= 1)

where DR(+), DR(-) represent the number of consensus reads mapped to positive and negative strands respectively, and VR(+), VR(-) represent numbers of variant reads mapped to positive and negative strands respectively. The higher the strand bias score is, the more the ratio will deviate from 1 towards 0, indicating a greater difference between the number of reads mapped to forward and reverse strands.

3.9.2 Minor variants at error-prone sites can be detected by the coefficient of variation

Another kind of possible error occurs in different samples at the same position of the genome, and I refer to this as error-prone sites. I found that such errors cannot be validated using solely read-bias score or frequency as they may occur with high frequency or unbiased-read score or even both. The key feature of error-prone sites is that they manifest as variants in many samples at the same genomic position at a fairly consistent frequency (table 3.4). To identify these particularly error-prone nucleotides, the coefficient of variation (the ratio of the standard deviation to the mean) was calculated at each position within the repeat. Error-prone sites would have a raised mean MV frequency in all strains, so its standard deviation would be small, whereas MVs that arise in a subset of strains (most strains having near 0 MV frequency, and a few having a high MV frequency) would have a high standard deviation and small mean. Therefore, positions where the mean MV frequency > SD were defined as potentially error-prone and MVs were considered as technical errors and excluded.

Table 3.4 Example of reported MVs at error-prone sites within IR1 unit.

Positions	StrainsWithVar*	EBV_Saliva	B958_cell_line	HL02	HL09	Sitoko	LCL_delEber2	sLCL-IS1.04
90	70	A(C,6.08%,0.23)	A	A(C,3.62%,0.47)	A(C,5.03%,0.44)	A(C,3.51%,0.79)	A(C,3.87%,0.99)	A(C,6.78%,0.66)
95	70	A(C,9.32%,0.02)	A	A(C,15.71%,0.01)	A(C,4.81%,0.02)	A(C,8.35%,0.0)	A(C,7.29%,0.0)	A(C,7.23%,0.01)
99	70	T(C,13.21%,0.19)	T	T(C,16.17%,0.03)	T(C,7.93%,0.05)	T(C,12.02%,0.02)	T(C,10.85%,0.03)	T(C,10.21%,0.09)
107	71	T(C,18.09%,0.02)	T(C,3.35%,0.1)	T(C,15.15%,0.01)	T(C,4.17%,0.02)	T(C,13.24%,0.01)	T(C,11.03%,0.01)	T(C,12.48%,0.02)

*total number of strains: 71
MV details are presented as following: Dominant nucleotide (MV, MV-frequency, MV-read bias score)

3.9.3 SNPs towards the edges of IR1 can be identified by read bias and frequency

Using the theoretical template as a reference, I found that several positions near the flanks of IR1 were reported to exhibit minor variants with high frequency (from 40 to 60 %) but with extreme read bias score (close to 0) (figure 3.16A). These extreme read bias scores were because of the big difference between the number of read spanning the junction and number of read mapping to the ends of junctions within the repeat unit (figure 3.16B). To define theses MVs as flank-SNPs, the frequency of MVs located close to the flanks was compared systematically to that of internal MVs at equivalent positions. Then, reads with potential flank-SNPs were visualized and make sure that reads span the junction between IR1 unit and the unique sequence or make sure that their paired mate reads map to the unique sequence. Positions in IR1 with verified flank-SNPs were replaced, and sequences were re-aligned. flank-restricted SNPs are listed in Appendix 5.

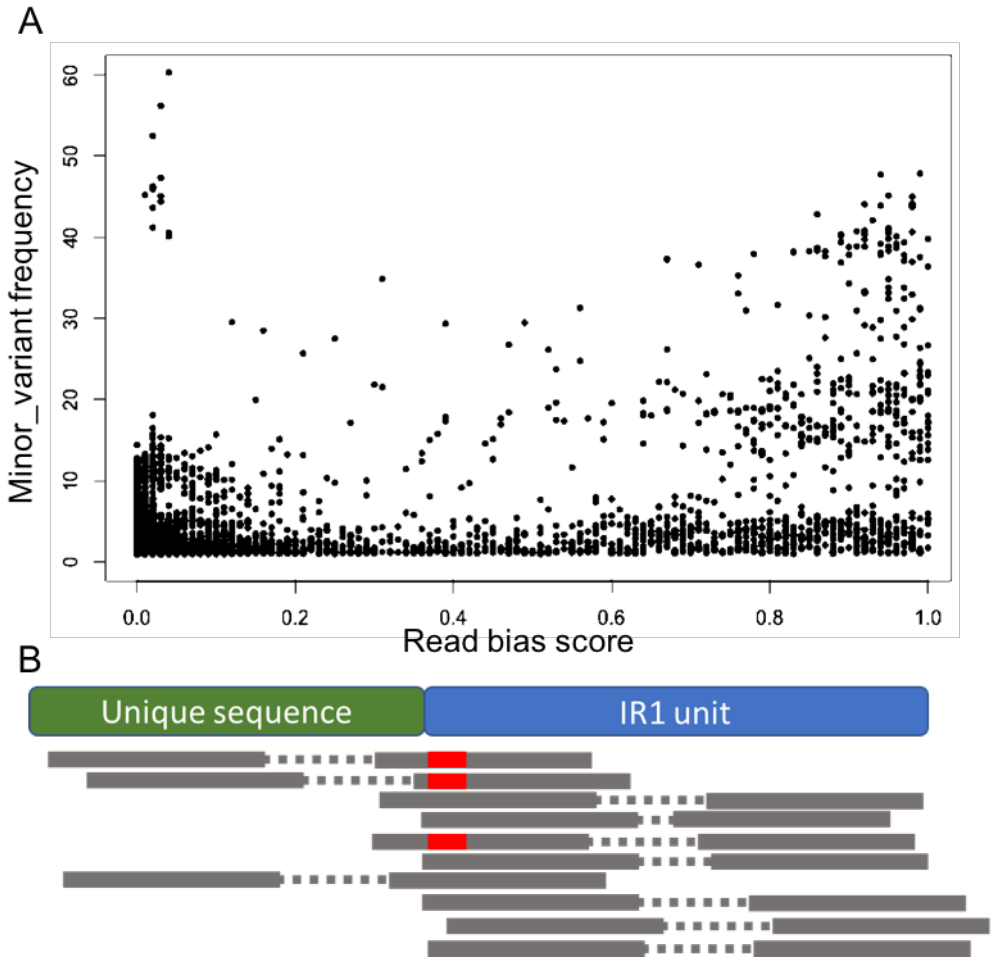


Figure 3.16 Identifying flank-SNPs.
(A) The graph plots the read bias (x-axis) against the MV frequency (y-axis) for every potential MV with a frequency of over 1% across 76 strains. Dots with high frequency and severe read bias score are potential Flank-SNPs. **(B)** Schematic shows the cause of biased read score for flank-SNPs when reads are mapped to the template. The difference between the number of read with a SNP (red line) that span the junction and the number of reads that support the consensus nucleotide and map to the ends of junctions within the repeat unit results in MV with extreme read bias.

3.9.4 Defining a strategy for calling MVs.

To lose as few genuine SNPs as possible, a cut-off algorithm was developed to consider the frequency of variants and read bias scores at the same time (instead of applying read bias and variant frequency independently):

$\sqrt{Read}_{bias} \times Variant_{frequency} \geq threshold$. Threshold in this study was 7% and it was selected based on the best fit for the data that distinguished likely technical errors from likely heterogeneity as described below.

A number of factors were considered in defining a strategy to separate the MV signal from noise. Since the IRI copy number is typically around 6 (Palser et al. 2015) but is known to be 11 in B95-8 (Baer et al. 1984), the frequency of MVs that are caused by heterogeneity in a single repeat unit would be expected to range from 8-15%. Our dataset shows many MVs at frequencies of under 7% (figure 3.16A). This is too low to be accounted for repeat heterogeneity, as the most repeats in a strain are found in B95-8, which has 11.6 repeats. This would give an 8% MV frequency, while the more typical repeat number (6.6) would give around 14-17% frequency. These low-frequency MVs probably arose from DNA cross-contamination, or low-level random mutation in the cell lines, but do not reflect genuine strain-specific heterogeneity within the repeats. The chosen threshold was able to predict minor variants that are verified in B95-8 as described in the following sections, but it is worth applying this assumption to heterogeneous repeat units with known variants to test whether it is straightforward to link the frequency of a MV with the number of repeat units or whether there are factors that skew this simple relation. All analysed MVs, within two BamHI W sites, for each strain are summarised in table ST3.1 at http://www.epstein-barrvirus.org.uk/MBaData.php.

3.9.5 Origin-based classification of minor variants

To further investigate the origin of MVs, they were classified into mutation-like MVs, recombination-like variants, and error-like MVs. The classification is based on the assumption

that mutation-like MVs will be randomly distributed across IR1. Therefore, where MVs are seen at positions that were otherwise conserved across strains, these were assumed to be mutations. In contrast, where strains contained variants at positions where other strains have SNPs, this was most likely caused by inter-strain recombination. Error-like MVs are those arising at error-prone sites (see section 3.9.2), and were filtered out of the analysis. To assess the quality of the MV data through visualisation, I plotted all the strains' MVs on a graph of read bias (i.e. relative frequency of MVs on top strand versus the bottom strand) against the MV frequency (figure 3.17), coloured by whether they were credible recombination-like MVs (green), likely errors (red), and mutation-like MVs (blue).

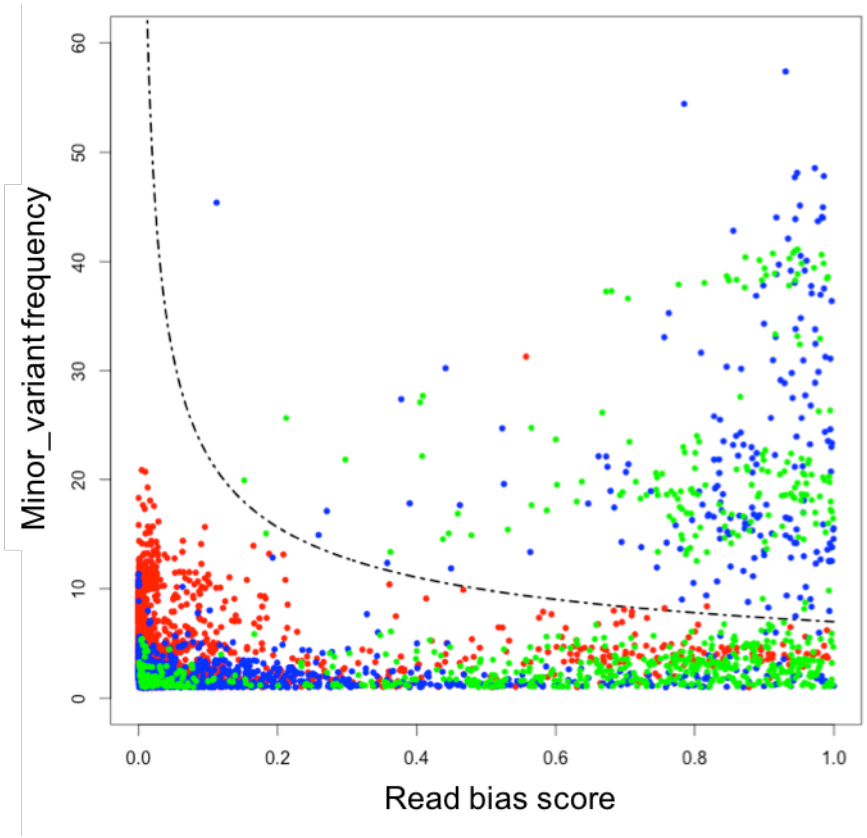


Figure 3.17 Visualizing MVs and defining a threshold for calling genuine MVs.
The graph plots the read bias (x-axis) against the MV frequency (y-axis) for every potential MV with a frequency of over 1% across 76 strains. Red points are potential MVs at error-prone positions. Green points are MVs at positions where at least one strain has SNP. All other MVs are indicated by blue points. The dashed line represents the cut-off above which the MVs are believed to be genuine, based on the equation: $\sqrt{Read}_{bias} \times Variant_{frequency} \geq 7$.

3.10 Variations in terminal IR1 units are protected from gene conversion

As described in section 3.9.3, I noticed that some positions showed a high MV frequency, but an extreme read bias score. These data points all arose from nucleotides close to the edges of the IR1, and MVs were not observed at the equivalent position in the middle of the repeat. The extreme read bias arose because the reads supporting the MV were paired to reads overlapping the unique region adjacent to IR1. This implied that these MVs existed only in the first or last repeat. By correcting the consensus sequence of the template to accommodate the variant at the flank and then remapping the reads, these MVs could be definitively identified as SNPs in IR1 adjacent to the unique sequence, as they no longer produced MVs. To find other SNPs near the ends of IR1 that may exhibit less strong read bias, or have been incorporated into the consensus templates, I identified all positions where equivalent positions were assigned different nucleotides in the two repeat units in the templates. I found 23 positions showing a difference between internal and edge nucleotides, 22 of them within 200 bp of either edge of IR1 (Appendix 5). The other (in B95-8) had been erroneously assigned different base calls in the two repeat units, and when this was corrected, the same MV frequency (close to 50%) was found in each repeat. Since these flank-SNPs have been definitively placed, they were corrected in the template, and as a result not included in subsequent MV frequency analyses.

Eight positions were shared by more than one distinct strain (names: a, b, c, d, e, f, g and z – Appendix 5). To understand the phytogenic relationship of the viruses with shared flank SNPs, they were visually represented adjacent to the phylogenetic tree (figure 3.11). Two of these flank variants on the Cp side of IR1, one of which has reported previously (Jones & Griffin 1983), were shared by all B95-8 sequences and BL36 (an intertypic recombinant with type 1 EBNA2 and type 2 EBNA3s (Palser et al. 2015)). In BL36 and B95-8 (both the BAC

and parental virus) and its related strain X50-7, the flank contained the same base at position 24 as all other EBV strains, whereas the equivalent position in the internal sequence contained a SNP throughout the repeat. Similarly, position 4 of the first repeat is G in all strains, while the equivalent position elsewhere in IR1 is a SNP (T) in all analysed type 2 strains with an intact Cp flank. In contrast, there is only one position on the right flank (Y exon side) that is shared by more than one (group Z). All member of his group contain G at internal positions within IR1 and A at the equivalent position in the last IR1. Notably, the only other strain on the same branch of the phylogenetic tree, sLCL-BL1.03, has the same A in the flank, but this is also in all IR1 repeat units. All members of group Z retain G in internal IR units but sLCL-1.07 has A as a MV within IR1.

Looking at strains sharing flank-variants on the phylogenetic tree, it was clear that strains with a common variant shared phylogeny, except group "g" members suggesting that the further these flank-variants are from the ends of IR, the less closer they were in the phylogenetic tree. There are two possible scenarios by which this may have arisen. First, a SNP arising in the repeat could rapidly progress to purity within the repeat by a mechanism that is inefficient near the edges of the repeat. This leaves only the edge position with a different sequence. Second, inter-strain recombination at one flank of IR1 could result in a different sequence near the flank, compared with that in the heart of the repeat. The type 2 variant mentioned above probably arose through the former mechanism, but the latter cannot be excluded. Since I am quite confident that these flank-variants are restricted to the first or last IR1 unit, these variants were therefore considered as SNPs (not MVs), and templates were corrected accordingly.

3.11 IR1 units are typically heterogeneous, and diversity is more common in tumour and tumour-derived cell lines

To test whether there is a relation between the source of virus and heterogeneity in IR1, strains were sorted into two groups based on whether the virus originated from tumours (or tumour-derived cell lines) or blood (i.e. sLCL) or saliva. Looking at the number of reported MVs in each strain, only 21 of the 71 distinct strains analysed were homogeneous (i.e. there were no MVs in their repeat). In other words, each of these 21 had identical IR1 units (unless the strain had flank-SNPs) (table 3.5).

Table 3.5 Strains with homogeneous IR1

sLCL-IS1.03	sLCL-IS1.07	sLCL-2.15
sLCL-IS1.08	sLCL-IS1.20	sLCL-IS1.14
sLCL-1.13	sLCL-BL1.03	sLCL-1.18
sLCL-1.24	sLCL-1.06	YCCEL1
sLCL-1.05	sLCL-1.11	Akata_2000
sLCL-IS1.12	sLCL-1.02	D3201.2
sLCL-IS1.19	sLCL-IM1.02	Jijoye

To avoid double counting, strains with common origins were excluded from the calculations. Based on strain sorting criteria, 17 strains of 21 were sLCLs (81%). Despite the fact that the number of strains with a homogeneous IR1 was limited, this clear difference in homogeneity rates between briefly-cultured spontaneous LCLs and EBV strains derived from long cultured and tumour derived cell lines or sequenced directly from tumours suggest that IR1 in circulating virus strains is more likely to be homogeneous than in tumours. Analysis of more EBVs isolated from healthy carriers, which is believed to represent circulating viruses, would be very important to support this hypothesis.

Strains with MVs were grouped according to the number of MVs in IR1 (figure 3.18). Looking at strains with MVs from 0-2, equal proportions of sLCL samples had one or no MVs in the IR1 region (72% in total) whereas the proportion of tumour-derived strains increased with increasing MVs. Nine strains (7 of which are tumour-derived strains) had from 5 to 21 minor variants in IR1.

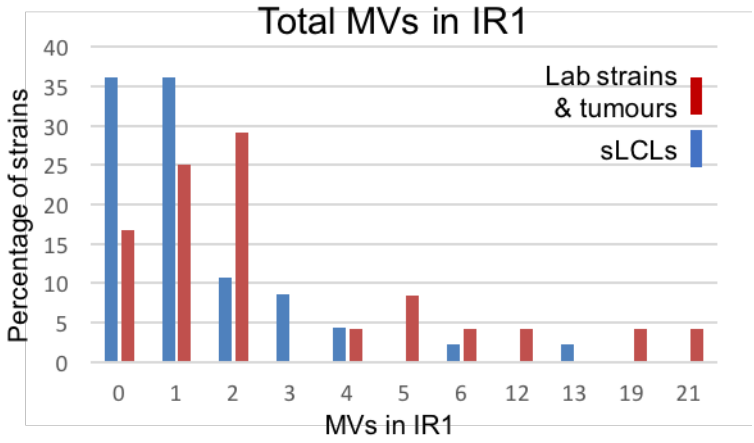


Figure 3.18 Overall distribution of MVs.
The number of MVs is shown on the x axis, and the y axis shows the percentage of strains with each number of MVs.

3.12 Strains with high number of MVs show evidence of interstrain recombination

To further characterize the origins of MVs, they were sorted into mutation-like MVs and recombination-like MVs (see section 3.9.5). The analysis was limited because there are probably SNPs that have not yet been identified (owing to the relatively small number of strains sequenced). These unidentified SNPs will lead to SNPs being classed as mutations. Also, there may be mutagenic processes within the cell that target certain sequences with higher frequencies. However, the analysis in this study suggests that viruses with few MVs mostly

exhibit sporadic mutations (figure 3.19 A, B and C). Cheptages exhibits the most mutation-associated MVs that are not accompanied by an example of recombination (table ST3.1 at http://www.epstein-barrvirus.org.uk/MBaData.php). Because Cheptages has deletions at both ends of IR1 (deletions in unique sequences flanking IR1 region and in IR1 edges- see table 3.2), this is perhaps further evidence that Cheptages has experienced a mutagenic environment.

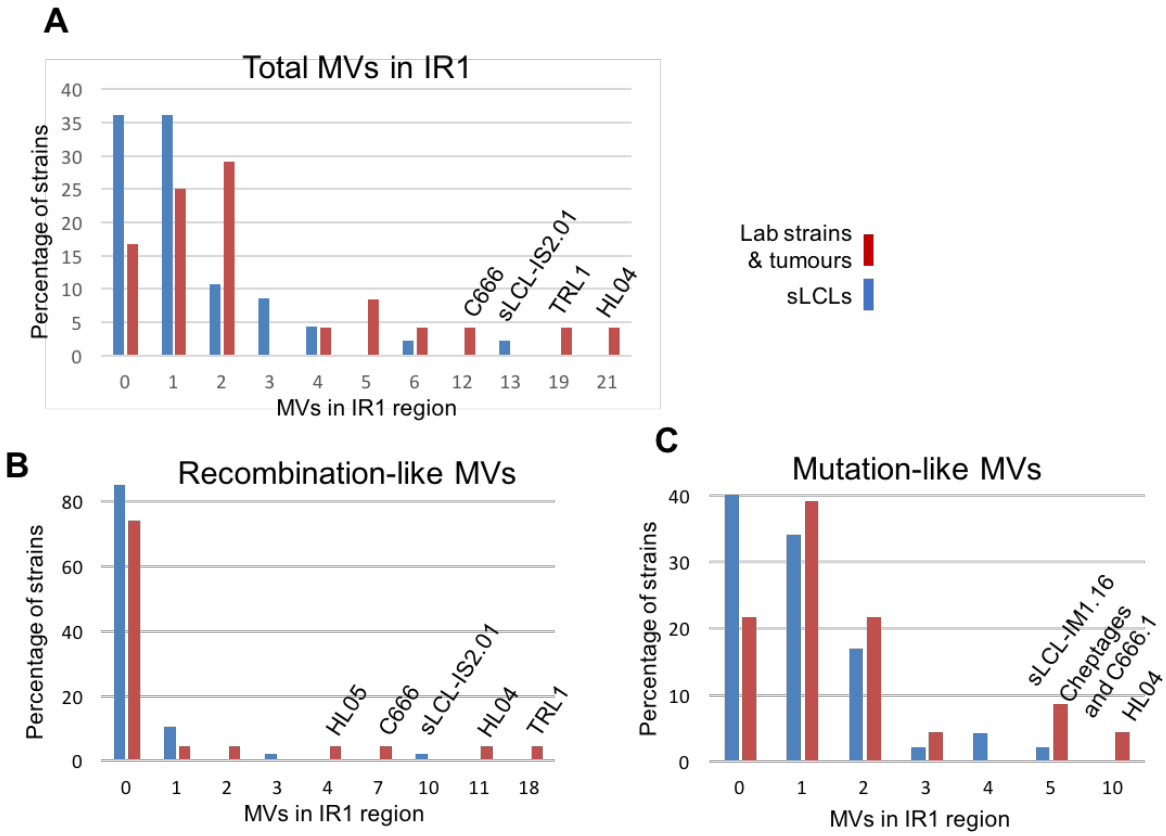


Figure 3.19 IR1 diversity in spontaneous LCLs and tumour-derived samples.
The Graphical representations show the number of MVs in the IR1 repeat region in each distinct strain. Samples are sorted into two groups based on whether the virus was isolated from tumours (or tumour-derived cell lines) (red bar) or blood (i.e. sLCL) or saliva (blue bar). The X-axis shows the number of MVs while Y-axis shows the percentage of strains with that number of MVs. **(A)** total MVs, **(B)** MVs that resemble SNPs (recombination-like MVs); and **(C)** MVs that resemble spontaneous mutations (mutation-like MVs). (All MVs in each strains are summarised in ST3.1 and labelled in sequences (using IUPAC code) in SD3.2 at http://www.epstein-barrvirus.org.uk/MBaData.php).

In contrast, much higher numbers of MVs are seen in strains that show evidence of interstrain recombination (figure 3.19B). Most of these strains also contain MVs that appear to be mutations, but we cannot conclude that these strains underwent both processes: it remains possible that we are a long way from having characterised all of the EBV diversity, and that these apparent mutations represent previously unseen SNPs. Strains, therefore, were classified into three classes: homogeneous, mutation-only and recombination-like, where any example of a SNP-like MV placed the sample into the recombination group. The distribution of these classes within spontaneous LCLs group and within the tumour and cell line groups is shown in (figure 3.20). This suggests that inter-strain recombination is more prevalent in cancer-associated viruses than in natural strains, whereas the sLCLs were more likely to have a homogeneous IR1. To assess whether recombination-like MVs occurred more often than expected in tumour-drived strains compared to sLSLs-drived strains, Fisher's Exact test was used. No statistically significant difference was found (p-value >0.05). This might be due to the limited number of strains (71 strains). Therefore, more data (from cancerous and non-cancerous strains) are required to verify this observation.

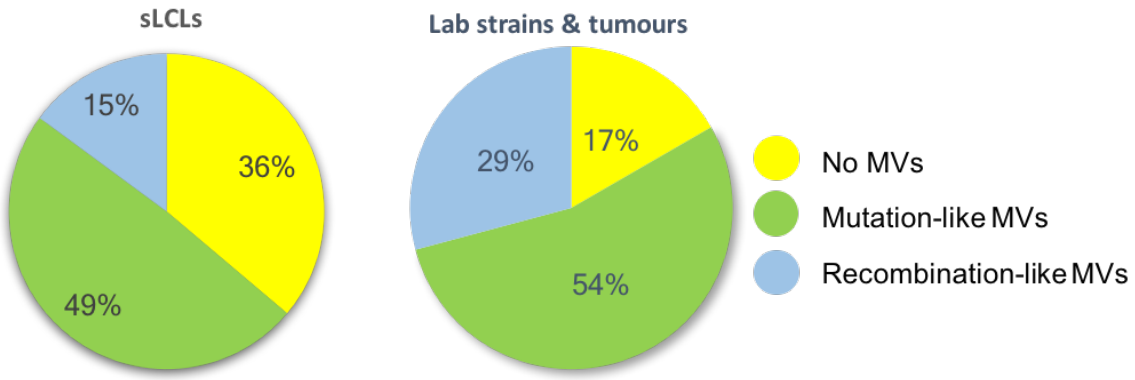


Figure 3.20 Minor variant accumulation in spontaneous LCLs and tumour-derived samples.
Pie charts show the types of genetic variations found in EBV IR1 from the sLCL and tumour groups. The type of variation was classified as either homogeneous repeats (yellow), at least one recombination-like MV (blue), or only mutation-like MVs (green).

3.13 There is no correlation between number of MVs and IR1 copy number

To find out whether the number of MVs is influenced by the number of repeats (i.e. more repeats gives more opportunity to contain mutations), the number of IR1 units were estimated based on the read depth. After mapping reads to the corresponding template, I calculated the average of the read depth of sequences within IR1 unit (from 2674 to 3074 and from 3231 to 3625, where there is no duplication of these regions in the 1.6 IR1 copy in the template) using SAMtools and an in-house script. The average depth of coverage of the non-repetitive region from 200 to 610 in the template (Cp region) was calculated for normalization. The repeat copy number for each strain was estimated by comparing the average read depth of the two regions in IR1 to the depth of coverage of the unique region in the templates (Appendix 1). Although it is theoretically possible to use read depth to estimate the copy number, using different locations (within and out of IR1) and using the theoretical template and the whole genome with 1.6 copies of IR1 gave different values for IR1 repeat number. Therefore, I tried to find the best locations that give the closest estimation to what I expected based on results showed by Palser et al. 2015 and based on IR1 copy number in YCCEL1 estimated by Kanda et al. 2016 (used PacBio for sequencing), in addition to the estimated number of IR1 in P3HR1 that I cloned as a BAC (see section 5.2.3 for P3HR1 cloning).

Strains were stratified into three approximately equally sized groups based on the estimated copy number of IR1, which were plotted to show the proportion of strains with different numbers of MVs (figure 3.21). There is no correlation between the number of MVs and the estimated length of the IR1 region.

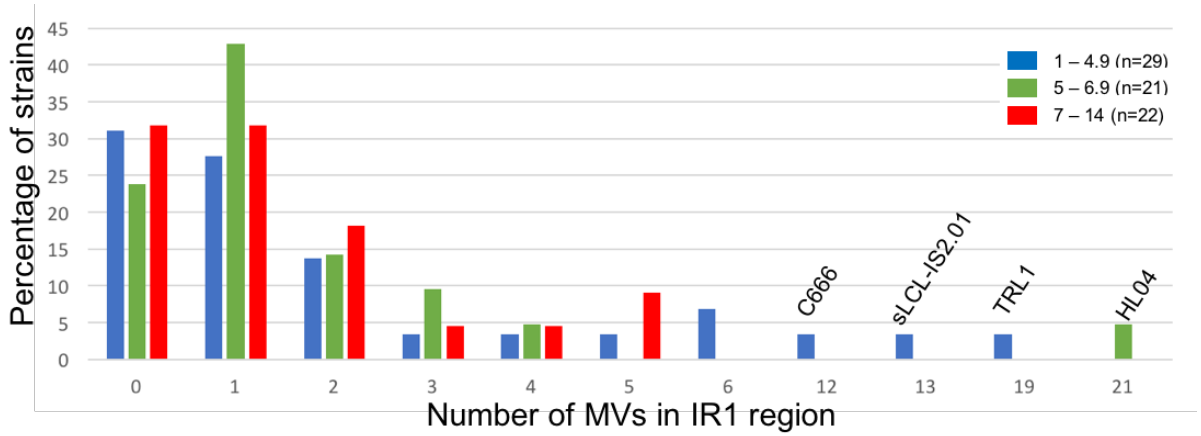


Figure 3.21 There is no correlation between IR1 size and number of minor variants. Strains were split into three groups (key colour is at top left), based on the number of IR1 repeats calculated from relative read depth in the Wp/W exon region of IR1 compared to Cp. The size of each group (n) is indicated. They are plotted to show the proportion of strains having each different number of MVs.

3.14 Diversity in IR1 is not caused by a mixture of strains in the sample

Although there were no MVs identified within the unique sequences flanking IR1 in the templates, I wanted to exclude the possibility that the MVs detected in IR1 were as a result of samples containing a mixture of virus strains. The consensus sequences of 1.6 copy of IR1 were assembled into the whole virus genome sequences (published by Palser et al. 2015), and the reads were mapped to this consensus. Then, the whole virus genomes were analysed for MVs using the same threshold of 7 (see section 3.9.4). MVs within repetitive regions were excluded. Five strains with MVs within IR1 had one MV across the genome, except IS1.06 and HL01 had 2 MVs while HL04 had 3 MVs (table ST3.2 at http://www.epstein-barrvirus.org.uk/MBaData.php). Other strains, even those with a high number of recombination-like MVs, had no reported MVs across their genomes. However, the reported MVs outside the IR1 region cannot be considered as a decisive sign of contamination, and they could be artificial as there were three strains with homogeneous IR1 (Jijoye, sLCL-1.08 and

D3201.2) had one MVs as well. Thus, the rate of MVs outside IR1 is not related to the rate in IR1. Only EBV-Saliva reported more MVs than others. It is most likely due to the low number of supporting reads, or it could be due to contamination in the sample or the person being infected with two closely related viruses.

3.15 Heterogeneity in IR1 arises regionally and may undergo a process of purification to homogeneity

Looking in more depth at recombination-like MVs to see if there are regions that are more susceptible to inter-strain recombination within IR1, B95-8 (and its related strains) has three MVs (two of which are recombination-like) within 126bp (discussed in more detailed in section 3.17) while other strains have more distributed recombination-like MVs, suggesting a wider region of inter-strain recombination. In Appendix 6, strains with the greatest number of recombination-like MVs are listed to highlight the clustering of the minor variants. HL04 and HL05 (Hodgkin lymphoma biopsies from different patients) have a cluster of 4 recombination-like MVs within 200bp. Intriguingly, two strains (TRL1 and sLCL-IS2.01) have MVs which likely arose from a recombination between type 1 and type 2. TRL1 is a type 1 EBV sequenced from two independent cell lines (TRL1-pre and -post) grown from an aggressive post-transplant DLBCL where an oligoclonal lymphoma (TRL1-pre) developed into a monoclonal tumour (TRL1-post) (Gottschalk et al. 2001). It is noteworthy that TRL1-pre cell line was not oligoclonal for EBV, as it seemed no different from –post and had identical sequences and the same MVs (with very similar frequencies). sLCL-IS2.01 is a spontaneous LCL containing a type 2 EBV established from the blood of a PTLD patient (Palser et al. 2015). These strains also show evidence of clustering of MVs: TRL1 has a cluster of six recombination-like MVs with a frequency of 38-41% between positions 649 and 802 in IR1 unit, and the remaining

thirteen MVs at a frequency of 13 to 21% spread across the repeat. sLCL-IS2.01 has most of its MVs clustered between position 1177 and 1252 of IR1, at a frequency of 18-20%, with three elsewhere found at similar frequency (16-23%) (Appendix 6).

To figure out the potential recombination events in TRL1 and sLCL-IS2.01, the consensus sequences for TRL1 and sLCL-IS2.01 were modified to instead contain all minor variant bases. Then, T-RECs (a pairwise Blastn alignment-based tool of sliding windows of query sequences against a user-defined sequence data) (Tsimpidis et al. 2017) was used to plot the consensus and the altered sequence strains against four strains chosen from different clades to represent the main IR1 sequence groups described by the phylogenetic tree (figure 3.11). Both TRL1 and sLCL-IS2.01 had regions of intertypic recombination. The consensus sequence of sLCL-IS2.01 (type 2) had acquired type 1 sequence (most similar to Akata 2000) within BWRF1, whereas TRL1 (type 1) had predominantly type 2 IR1 consensus sequence (resembles AG876) (figure 3.22).

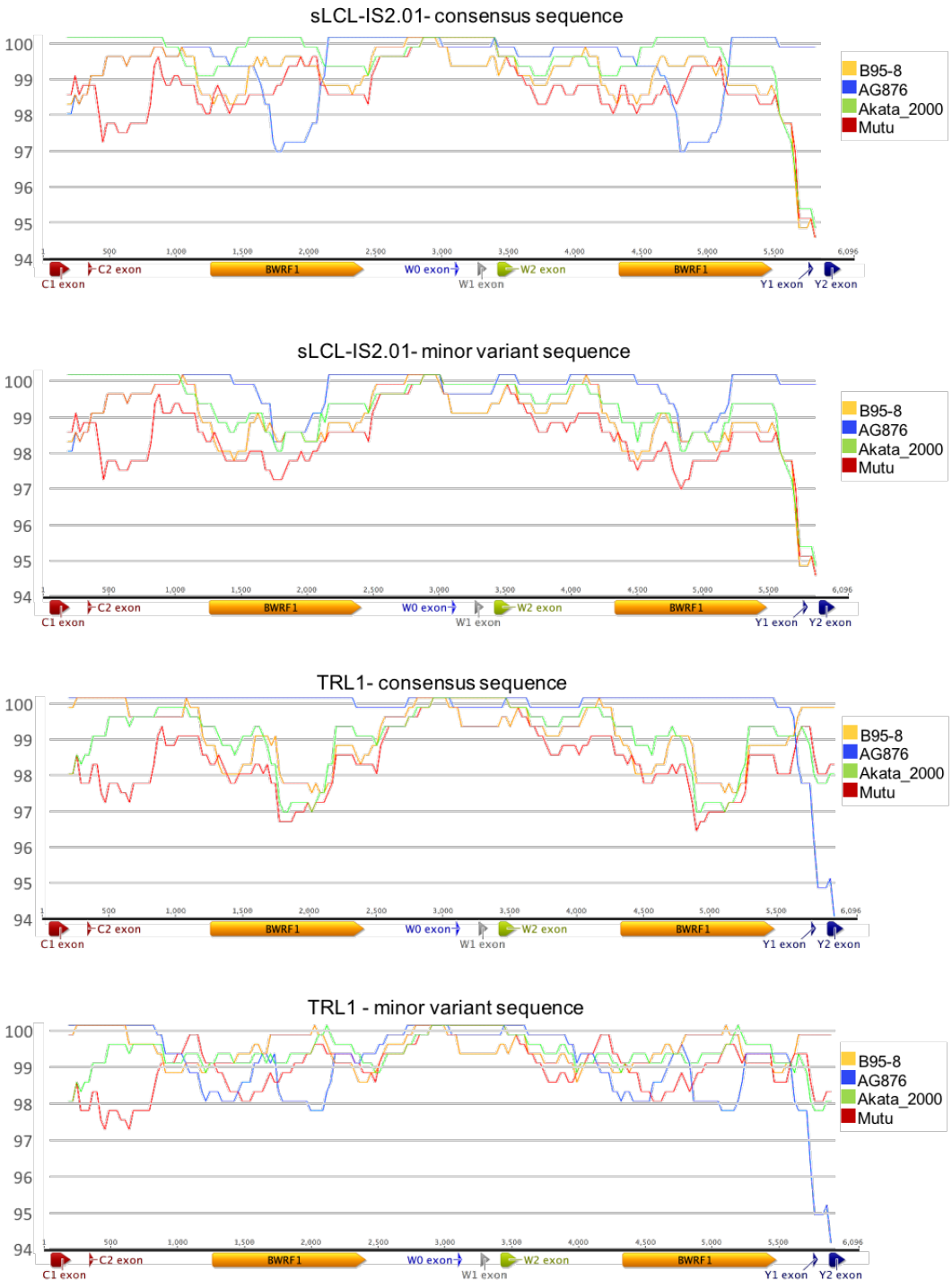


Figure 3.22 T-RECs analysis of genetic exchange in the IR1 regions of sLCL-IS2.01 and TRL1.

The consensus sequences of the theoretical template for sLCL-IS2.01 (A) and TRL1 (C) were modified to contain all the minor variant bases (B and D). Using T-RECs, these two sequences were compared with a set of four viruses from different clades on the phylogenetic tree, to identify regions of similarity (shown to the right of the plot). The similarity plots on Y-axis shows the percentage identity of the test strain to each comparator strain within a 200 bp scanning window and a step size of 20 bp**.** The x-axis indicates the positions which are illustrated by the schematic below each plot.

To avoid the noise in the T-RECs analysis created by spontaneous mutations, bootscan analysis (an analysis involves constructing bootstrapped neighbour-joining tree in a sliding window of defined size) was used to visualise the transitions from being most similar to the type 1, to being most similar to the type 2, and vice versa (figure 3.23). This analysis showed that both LCL-IS2.01 (type 2) and TRL1 (type 1) has acquired the region of the other type in the Cp region upstream of IR1. The region from Wp to exon W2 in both strains is highly conserved and has no SNPs that distinguished the comparators. sLCL-IS2.10 retains a type 2 sequence in the Y exons, across Wp and the W exons and the hairpin region of IR1. However, it has an Akata-like sequence predominates across much of the BWRF1 ORF in the consensus sequence and type 2 sequence in the same region with MVs. The frequency of the clustered minor variants in sLCL-IS2.01 are from 18% to 22% (Appendix 6) suggesting that only one unit retains the type 2 sequence since the estimated copy number of IR1 units is 4.1 (Appendix 1). Conversely, the type 1 virus TRL1 is most closely related to type 2 EBV across IR1, despite having a type 1 sequence in the Bam Y region. However, the MVs are most closely associated with type 1 viruses across the first two-thirds of BWRF1 region in the two repeat units. The estimated number of IR1 units in TRL1 is 3.6 and the frequency of the clustered minor variants are from 18% to 41% which may indicate that these MVs exist in one copy.

It is not possible to conclusively define the mechanism by which this combination of type 1 and type 2 arose. With this distribution of variants, it is likely that both of these strains have acquired sequence from another virus type but this acquisition (or part of it) could be fixed across the repeat and some region retained the original sequence.

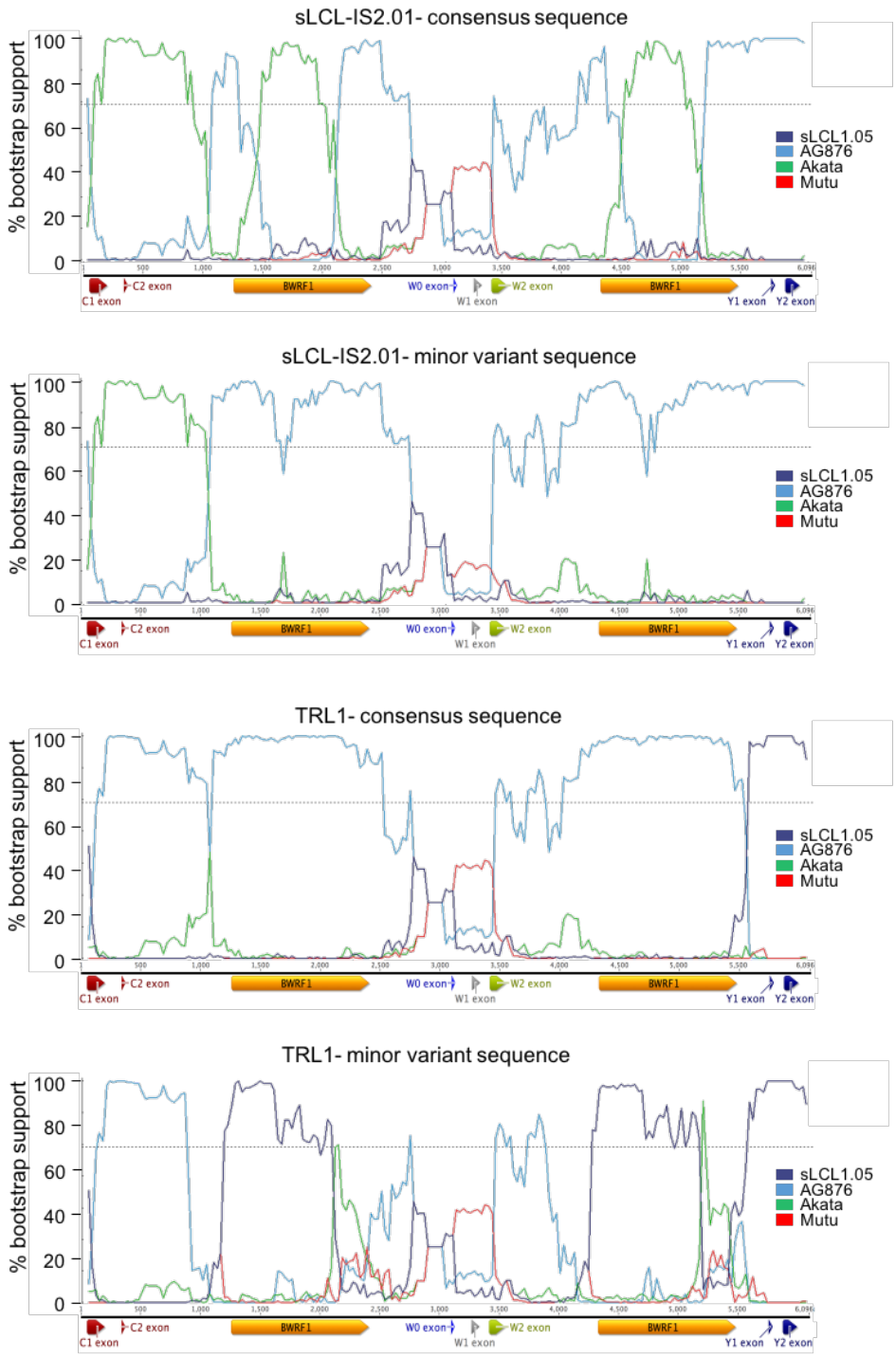


Figure 3.23 Bootsscan analysis of genetic exchange in the IR1 regions of sLCL-IS2.01 and TRL1.

The consensus sequences of the theoretical template for sLCL-IS2.01 (A) and TRL1 (C) were modified to instead contain all the minor variant bases (B and D). Using Bootscan, these two sequences were compared with a set of four viruses from different clades on the phylogenetic tree, to identify regions of similarity (shown to the right of the plot). This analysis simulates phylogenetic tree assembly for a 400-bp sliding window and a step size of 30. The plot indicates the events of clustering and the occasions that each strain is the nearest neighbour on the tree. The x-axis indicates the positions which are illustrated by the schematic below each plot.

3.16 BL36 is probably a lab-associated intertypic recombinant strains

BL36 (African Burkitt lymphoma cell line) has type 1 EBNA2 and type 2 EBNA3 sequences, and was thus reported as intertypic recombinant strain (Palser et al. 2015). BL36 contains the same MVs as B95-8, supporting the observation that BL36 is a hybrid strain (showed in figure 3.24B and summarised in table ST3.1 at http://www.epstein-barrvirus.org.uk/MBaData.php). Further examination by scanning systematically the MSA of the whole genome sequences (74 strains) showed that the BL36 and B95-8 genomes are identical between position 10024 (just after an insertion of a 14-bp palindromic sequence in BL36) and position 43611. There are five SNPs in this region (three in IR1 and two in BamHI H) in B95-8 and BL36, but no other strains (Appendix 7). This suggests that BL36 does not represent a natural strain, but instead arose from a recombination between a type 2 BL and B95-8 – presumably in a lab environment, where B95-8 is often prevalent – to produce the resultant hybrid virus.

3.17 The prototype strain B95-8 contains clustered of low frequency minor variants that include a stop codon mutation in EBNA-LP

The prototype EBV strain B95-8 contains five MVs (figure 3.24), two of which are recombination-like MVs. These MVs are found in several cell lines related to B95-8: a B95-8-transformed cell line from our stocks that was mislabelled as the Sitoko cell line, in sequence data generated from the parental B95-8 cell line by the Chiang laboratory (Kwok et al. 2014) and in B95-8-delEBER (Gregorovic et al. 2011), an EBER2-knockout recombinant produced from the HB9 subclone of the B95-8 BAC, p2089 (Delecluse et al. 1998). Also, BL36, an African BL which is believed to be an intertypic recombination stain (see section 3.16), had

the identical MVs. Two of the MVs were also found in X50-7, an LCL established by co-culture of umbilical cord B cells with X-ray-irradiated B95-8 cells (Wilson & Miller 1979). Four of these MVs share the same frequency in each strain (around 8% in parental B95-8 and around 20% in the B95-8 BAC and other lines), which most likely indicates that these MVs occur once per IR1 (i.e. each sequence variant occurs in one repeat unit). This assumption was obtained by comparing the percentage prevalence with the repeat copy number of each sequence (e.g. IR1 copy numbers in parental B95-8 (11.6) and in B95-8 BAC (6.6) (figure 3.24). The higher frequency MV is in the hairpin region of B95-8 (position 12647). This MV was reported previously as SNP in the first IR1 unit (Jones & Griffin 1983), whereas our data shows that it is fairly evenly split (around 50:50 in the B95-8 cell line and around 70:30 and 30:70 in related lines) predominating either the normal G or the variant T (figure 3.24 A). Of the three low-frequency variants clustered towards the end of BWRF1 (which are not MVs in X50-7), two are found as SNPs in at least one other strain (green in figure 3.24 B), while the third lies near the start of a relatively conserved region. These four BWRF1 MVs would all cause changes in the coding region of BWRF1 (R->L; S->R; V->L and V->A respectively).

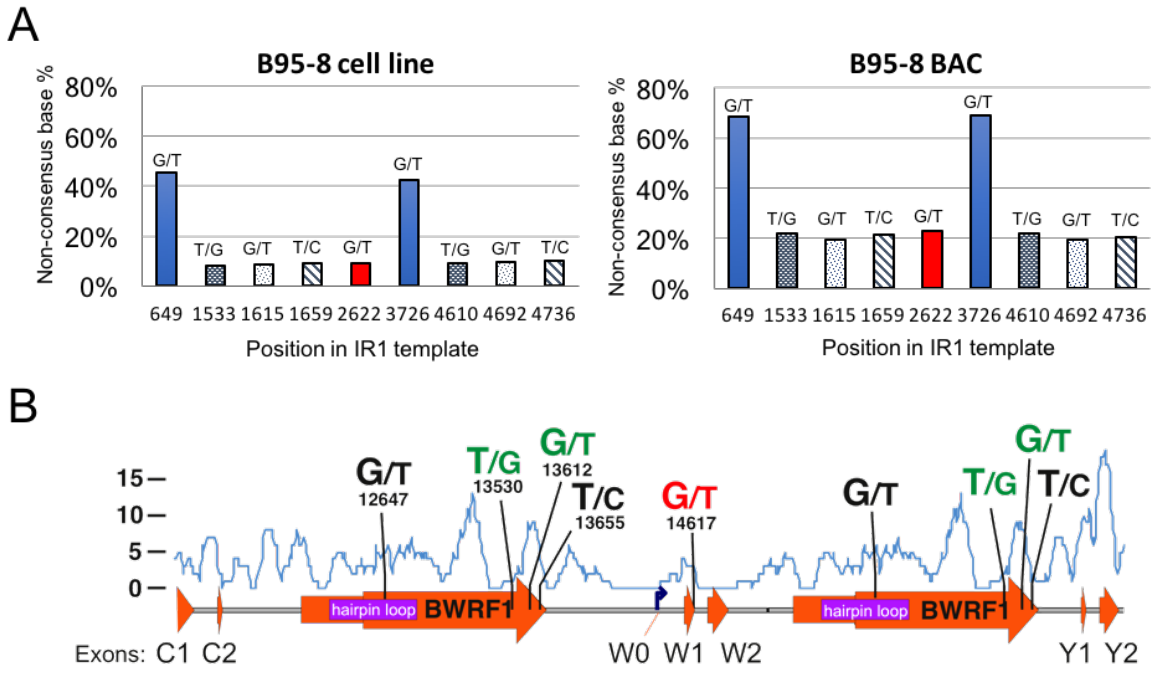


Figure 3.24 Minor variants in IR1 of the prototype lab EBV strain B95-8.
(A) Graphs showing the MV frequencies across the template sequence of IR1 of the B95-8-cell line sequence left, and B95-8-BAC (B95-8-delEBER) right. The Y-axis shows the frequency of the non-consensus nucleotide while The X-axis shows the position of those variants in the IR1 template in the multiple sequence alignment output (supporting data SD3.2 at http://www.epstein-barrvirus.org.uk/MBaData.php). Notably, for the HB9 BAC-based clone (right), the alternative nucleotide at position 649/3726 is the dominant allele. For other positions, the difference in MVs frequencies is attributed to different numbers of IR1 copies between the B95-8 cell line (11.6 repeat units) and the B95-8 BAC (6.6 repeats). **(B)** Schematic showing the position of MVs in IR1 of B95-8 relative main features of IR1 units. For each MV shown in (A), the left base is the published sequence, the right indicates the alternative variant with the equivalent position in the first repeat unit of the prototype EBV sequence (NC_007605). Bases in green indicate those changes that have been observed as SNPs in other strains (recombination-like MVs). The G/T in red indicates minor variant that creates the STOP codon in W1 where GAG becomes TAG in the last codon. The blue plot and Y-axis indicate the SNP density across the repeat (from figure 3.9).

Most strikingly, however, the fifth MV (G14617T) (highlighted in red in figure 3.24B and in table ST3.1 at http://www.epstein-barrvirus.org.uk/MBaData.php) falls within the last codon of the W1 exon of EBNA-LP and changes the GAG (glutamate) codon to a TAG Stop codon. To validate these mutations, BamHI W fragments from the HB9 B95-8 BAC were cloned and sequenced by Dr Richard Palermo. Two BamW fragments were cloned (from

independent DNA preparations of the B95-8 BAC) and found to contain the TAG STOP codon. Both clones also contained the three BWRF1 MVs that were found at about the same frequency. The hairpin DNA structure prevented reliable Sanger sequencing of the other variant, although it was found in one clone that lacked the other MVs. Overall, this shows that the B95-8 MVs that exhibit equivalent frequencies persist in the same IR1 repeat unit, and exist at one copy per virus genome in B95-8.

It was not possible to identify which IR1 repeat unit has the stop codon using the available short read sequence (76 bp). Therefore, an EBNA-LP transcript was reverse transcribed into cDNA (by Agnieszka Szymula) and amplified by PCR using forward primer (Cp-fwd) spans C1/C2 exon and reverse primer binds to Yh exon (Yh-rev). Because transcripts across IR1 are spliced in a manner that skips various of the IR1 exons, the longest PCR amplicon (which should contain all W1 exons) was extracted and sent for Sanger sequencing (figure 3.25). Unfortunately, only one W1W2 domain was sequenced from the Yh side, and then the sequence signal dropped. Also, I could not get sequence signal from the Cp side. However, this result shows that the last IR1 copy does not have the stop mutation.

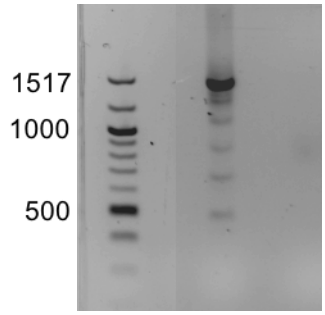


Figure 3.25 Gel electrophoresis of the RT-PCR-amplified EBNA-LP transcript (W1W2 exon) obtained from HB9.
PCR amplification of cDNA from EBNA-LP transcripts initiated at the Cp promoter. The forward primer spans the C1-C2 exon while the reverse binds to the Yh exon. Because of the splicing pattern in EBNA-LP transcription the expected sizes for bands containing from 1 to 6 W1W2 domains: 1530, 1332, 1134, 936, 738, 540 bp. Sizes of bands in the ladder (100bp ladder, NEB) are indicated to the left. (for PCR condition see section 2.9).

As an alternative strategy to identify which repeat unit contains the STOP codon, a pair of reverse primers were designed of which the last 3' nucleotide is specific for either the mutant W1 exon (Mut-rev) or the wild-type (WT^{w}-rev) (figure 3.26 A). These were used with a forward primer that binds to the unique region upstream IR1 in viral genome (C_region_fwd). Using gradient PCR with HB9 genomic DNA (B95-8 BAC) and a newly constructed B95-8 BAC (WT^{w}), which includes 6.6 identical copies of consensus IR1 (no MVs) (Szymula et al. 2018), showed that the first IR1 repeat might be the one that has the stop mutation (Figure 3.26 B): In HB9, the expected band obtained from the first IR1 copy using (WT^{w}-rev) became fainter at the highest annealing temperature than the band produced by using (Mut-rev) (figure 3.26 B). The same experiment with B95-8-WT^{w} showed the opposite results (figure 3.20 B). However, I cannot be sure that the first IR1 unit in HB9 has stop mutation without sequencing the region. Third generation sequencing (e.g. PacBio) could be used to verify this conclusion.

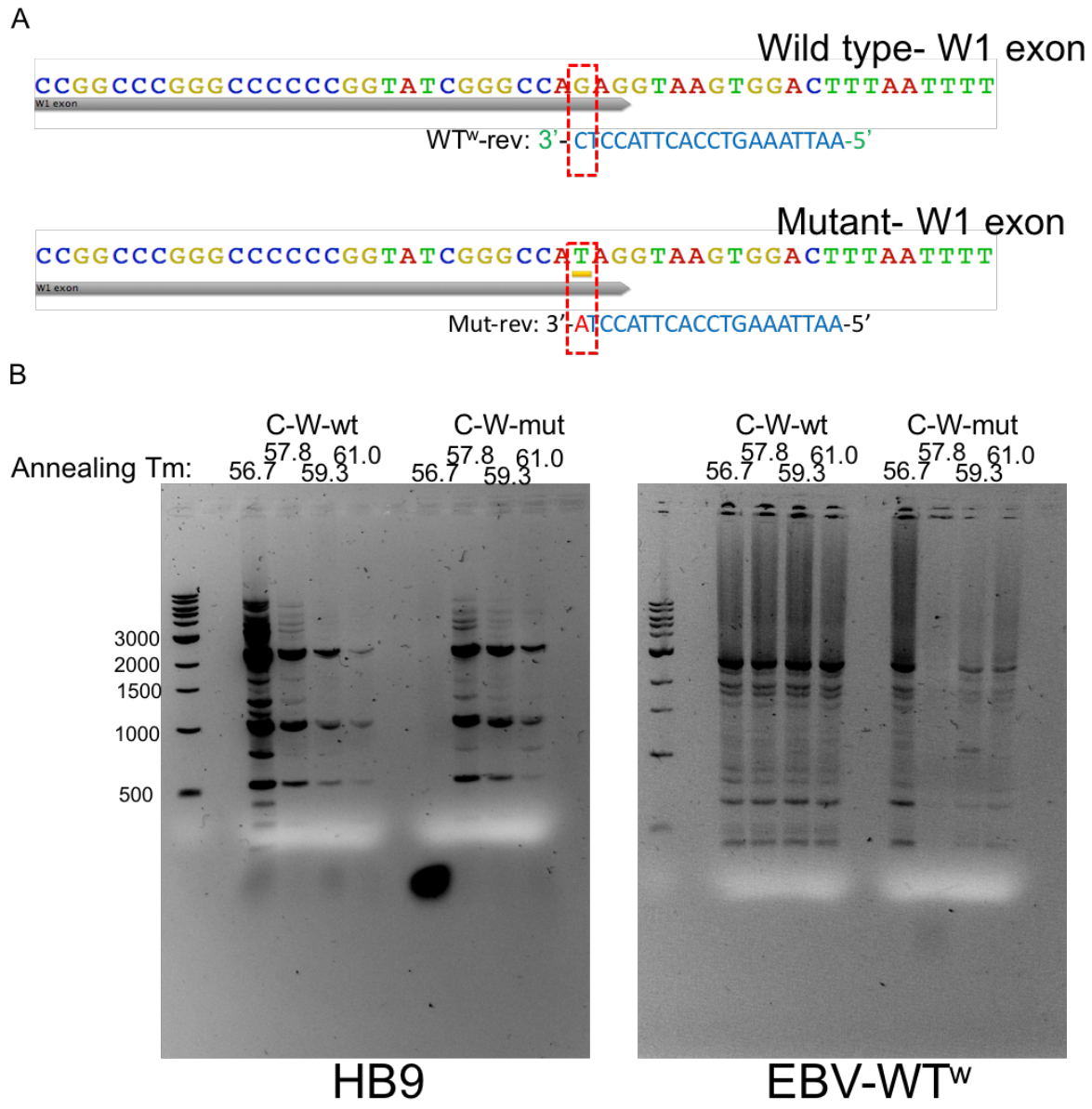


Figure 3.26 PCR experiment to identify the IR1 copy with the stop mutation.
(A) The schematics show where the reverse primer binds to in the wild-type exon W1 (top) or mutant exon W1 (bottom). The red box indicates the different nucleotide between the two revers primers. **(B)** Gel electrophoresis of PCR-amplified IR1 repeat units obtained from HB9 left or B95-8-WT^{w} right using forward primer (c_region_fwd) and reverse primer (either Mut_rev or WT^{w}_rev). Annealing Tm is indicated at the top of lanes with the primer pairs. The expected size of DNA band from one IR1 unit is 2678 bp, the smaller bands are non-specific bands. Sizes of bands in the ladder (1 kb ladder, NEB) are indicated to the left. (For PCR condition see section 2.9).

3.18 The stop codon in a single W1 exon in B95-8 interrupts EBNA-LP isoforms expression and slightly attenuates virus transformation ability

Since EBNA-LP is generated from transcripts that are spliced to skip various of the IR1 repeat-derived exons in a manner that remains uncharacterized (Rogers et al. 1990), the stop codon would not prevent the generation of full-length EBNA-LP, but could potentially result in the generation of both complete and truncated EBNA-LP proteins. Thus, it may reduce the average length and quantity of EBNA-LP proteins generated by B95-8 EBV.

To test the effect of the stop codon mutation on EBNA-LP expression, I compared WT^{w}, the B95-8 BAC with a repaired IR1 (No MVs) (Szymula et al. 2018), with HB9 (B95-8-BAC) by infecting frozen PBLs from 4 different donors. Western blotting for EBNA-LP in the LCLs established from these infections (figure 3.27) showed that EBNA-LP produced from the WT^{w} was generally larger, and there was much less variation in the range of sizes compared to those isolated from HB9-LCLs.

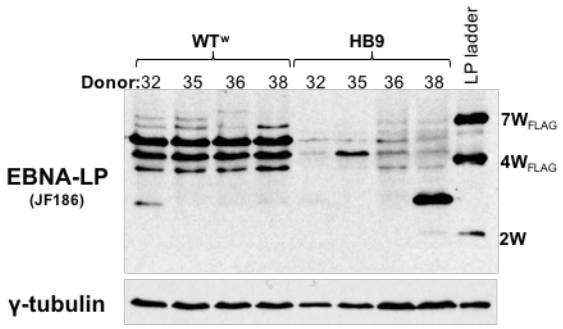


Figure 3.27 Western blot analysis of EBNA-LP extracted from LCLs 30 days post infection.
PBL samples from four independent donors were infected with either the original B95-8 EBV BAC (HB9) or a recombinant BAC from which all MVs were removed (WT^{w}). Proteins extracted from cell lysates were separated by SDS-PAGE (12.5% polyacrylamide gel) and EBNA-LP was detected by using JF186 antibodies. The EBNA-LP ladder is from 293 cells transfected with three different EBNA-LP sizes (from Dr Richard Palmero): the number of W1W2 domains in each band plus tag (if any) is indicated to the right. The blot was stripped and re-probed with a gamma-tubulin antibody as a loading control.

Additionally, the effect of the stop codon mutation on the ability of EBV to transform B cells was investigated. Uninfected B cells die after a few days in culture, whereas infecting B cells with EBV establishes immortal cells (LCLs). Therefore, I used CellTrace violet as a marker of cell proliferation. However, I found that the toxicity of CellTrace varied from tube to tube even from the same batch (Dr Agnieszka Szymula, a former PhD student in the White lab, reported the same observation). Therefore, the experiment was conducted three times until I found the non-toxic tube and I could not notice a difference in tolerating the toxicity between frozen and fresh PBLs. Ultimately, with fresh PBL samples and halving the recommended amount of the stain, a proliferation assay on day 8 showed that WT^{w}-infected cells expanded through more cycles of cell proliferation than HB9-infected cells (figure 3.28A). In culture, I observed that establishing LCLs with WT^{w} was more rapid than with HB9 (media acidified earlier). EBNA-LP was extracted from the same established LCLs, and a Western blot confirmed the effect of the stop codon on the abundance and size of EBNA-LP protein isoforms (figure 3.28 B).

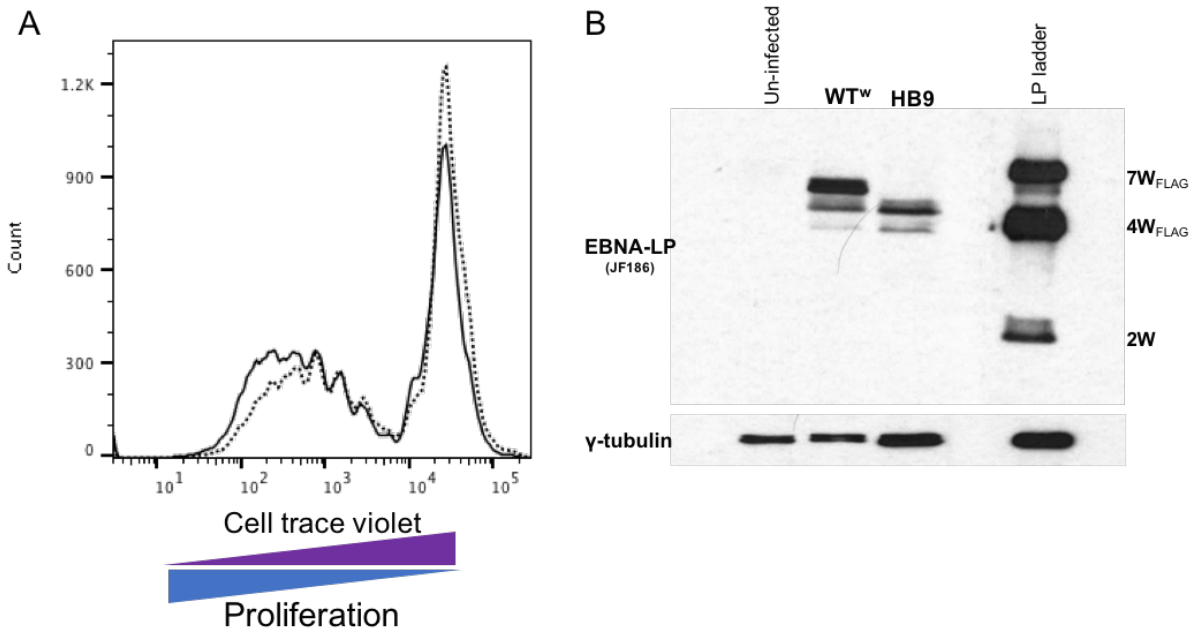


Figure 3.28 Effect of the stop codon mutation on the ability of EBV to transform B cells. **(A)** Flow cytometry traces the dilution of the CellTrace Violet stain on cells as an indication of cell proliferation. Fresh PBLs from one donor were stained with CellTrace Violet just before infection with either B95-8 EBV BAC (HB9) or a recombinant BAC from which all MVs were removed (WT^W). On day 8, CD19+ live cells were analysed using flow cytometry comparing equal cell numbers (50683 events) for the WT^W (solid line) and parental HB9 (dotted line) strains. The Y-axis shows the number of cells, and the X-axis shows the log10 CellTrace Violet fluorescence intensity (the lower CellTrace intensity in a cell, the more cell proliferation it has undergone as shown below the figure). **(B)** Western blot analysis of EBNA-LP extracted from the same samples (in A) 9 days post infection. Proteins extracted from cell lysates were separated by SDS-PAGE (12.5% polyacrylamide gel), and EBNA-LP was detected by using JF186 antibodies. The EBNA-LP ladder is from 293 cells transfected with three different EBNA-LP sizes (from Dr Richard Palmero): the number of W1W2 domains in each band plus tag (if any) is indicated to the right. The blot was stripped and re-probed with a gamma-tubulin antibody as a loading control.

3.19 Summary

In this chapter, I have described an approach to assemble and analyse the sequence of the largest internal repeat of EBV (IR1 or Bam W repeats). I showed how the devised read depth analysis (RDA) can detect low levels of contamination in a sample and predict flanks between IR1 and unique sequences in strains with defective junctions. Analysing more than 70 strains showed that IR1 has conserved regions around the W promoter and two regions within the putative open reading frame BWRF1. I identified two major models for the BWRF1 ORF. The largest BWRF1 exists in around 60% of strains, including strain isolated from the saliva of healthy person whereas BWRF1 model in B95-8 is found in only 12% of analysed strains. BWRF1 may have a role in EBV biology but the intact full-length BWRF1 ORF may be not essential as it is intact in around 80% of strains. EBNA-LP diversity analysis revealed a new subtype of EBNA-LP which is the most common one in the samples. IR1 is heterogeneous in the majority of strains. I showed that the heterogeneity arises though spontaneous mutation as well as interstrain recombination. It is likely that interstrain recombination contributes to pathogenic potential of EBV as we found that recombination-like MVs are more common in tumour-derived strains. Most striking I found a previously unreported stop codon mutation in one copy of the W1 exon of the prototype lab strain B95-8, and I showed how this mutation affects EBNA-LP expression and EBV efficiency in transforming B cells.

4 Terminal repeat analysis

4.1 Introduction

As described in section 1.7.3, the terminal repeat (TR) region of EBV consists of a variable number of direct, tandemly repeated sequences which are GC-rich (77%). The TR region is involved in EBV linearization and circularization during EBV lytic cycle (Zimmermann & Hammerschmidt 1995). Like IR1, TR has only been sequenced in a few strains that were analysed by conventional Sanger sequencing like; B95-8 (Baer et al. 1984), AG876 (Dolan et al. 2006) and GD1 (Zeng et al. 2005). The TRs of more recent (NGS-sequenced) strains have not been studied because of the difficulties in repetitive sequence alignment and analysis. In this chapter, I applied the repeat analysis strategy described in chapter 3 on TR in an attempt to fill the gap in EBV genome sequencing and get a better understanding of TR region diversity.

4.2 Diversity of the TR hindered making a theoretical template to simplify the analysis

Using a similar approach as for IR1 for the analysis of TR, first a theoretical template was constructed based on the EBV prototype sequence (B95-8/Raji, accession number: NC_007605) to be used as a reference for aligning sequence reads, in order to generate a reference template for each strain under study. In B95-8, each TR repeat unit starts with identical 11-bp element while the last unit starts with the 11-bp elements and ends with a 9-bp element plus the 11-bp element (Zimmermann & Hammerschmidt 1995) (figure 1.11). Therefore, to construct a theoretical template representing all feature of TR units, the most left TR unit was joined to the most right TR (the one that includes the 11-elements on both termini

and the 9-elements). Then the TR-TR unit was flanked with 500-bp of the unique sequences upstream and downstream of the TR region.

However, mapping reads against the constructed theoretical template raised two substantial issues, suggesting that using a single reference for mapping TR reads was not an ideal strategy. First, read depth analysis (RDA) (section 3.3) reported many strains with sudden changes in the read depth, especially close to the edges of the repeat unit, which means that there were either inserted or deleted chunk in the studied strains compared to the used reference. Second, the obtained consensus sequences seemed to be influenced by NC_007605 sequence resulting in losing genuine SNPs and indels. This influence was observed by looking at the TR unit sequence obtained by either mapping read to the constructed template or from de novo contig assembly for several strains (data not shown).

Since there was no ideal consensus reference for mapping reads from TR sequences, a strategy was developed to identify the sequence of at least one TR unit. Reads for 76 strains were mapped to the EBV prototype reference (NC_007605) lacking the first 1 Kb of the genome sequence and the last 3 kb where TR region is located. The unmapped reads (paired and orphan reads) from this analysis were then used for de novo assembly using the SPAdes genome assembler (Bankevich et al. 2012) (see section 2.1.3 in material and methods). Eight strains produced contigs containing an intact repeat unit in a single contig flanked by the unique sequences. In most cases, however, the resulting contigs were not long enough to include an intact TR unit and were in inverted directions. As a result, Geneious was unable to build longer contigs. To ensure that the multiple contigs that contain TR regions were in the same orientation, the contigs were visualised to find those that contain the unique sequences upstream and downstream of the TR region and to put them in the same orientation. Then, overlapping contigs that contain the TR sequence were identified in a stepwise approach, by

looking first for the TR sequence adjacent to the unique sequence and then start to build up one TR unit flanked by the unique sequences. This step was most time consuming and tedious, I found that it would not be suitable for scaling up to large numbers of strains. I was able to assemble a sequence of one TR unit plus flanks for only 31 of the 76 strains (table 4.1). Reads were then mapped to the constructed references using BWA. Sequences are available in SD4.1 at http://www.epstein-barrvirus.org.uk/MBaData.php and appendix 8 shows DNA sequence with annotations of the B95-8 TR theoretical template.

Table 4. 1 List of used strains for terminal repeats analysis.

Strain	Strain	Strain	Strain
B958_cell_line	sLCL-1.04	sLCL-IM1.16	HL01
C666.1imp	sLCL-1.05	sLCL-IS1.01	HL02
Daudi	sLCL-1.06	sLCL-IS1.03	HL04
EBV_Saliva	sLCL-1.11	sLCL-IS1.07	HL09
sLCL-1.02	sLCL-1.12	sLCL-IS1.11	Mak_1
sLCL-2.16	sLCL-1.13	sLCL-IS1.12	Jijoye
sLCL-BL1.03	sLCL-1.24	Sitoko (B95-8)	HKN15
sLCL-IM1.05	sLCL-2.15	sLCL-IM1.09	

4.3 The TR region in EBV genomes is highly diverse

From MSA of one TR unit for 31 strains (supplement data SD4.1 at http://www.epstein-barrvirus.org.uk/MBaData.php), the length of the TR unit in each EBV strain varied considerably ranging from 489-bp in Mak to 749-bp in sLCL-IS1.15_2120, compared to a TR length of B95-8 557-bp (Figure 4.1). Noticeably, two strains (sLCL-IS1.19 and sLCL-IM1.09) have an insertion at a position equivalent to 169967 in NC_007605. sLCL-IS1.19 has an insertion of 34-nt whereas sLCL-IM1.09 has an insertion of 76-nt. There are two mismatches in the first 34-nt of the inserted sequence in the two strains, and it seems that the sequence

originated from within the TR unit as the same sequence with one mismatch existed 39-bp upstream the insertion. It is unlikely that the duplication occurred during contig assembly as both sequences were on the same contigs suggesting that the insertion arose by a duplication event within TR in both strains. The variation in TR length across strains may help to explain why it was not ideal to map TR reads against one reference and why RDA reported strains with sudden changes in read depth.

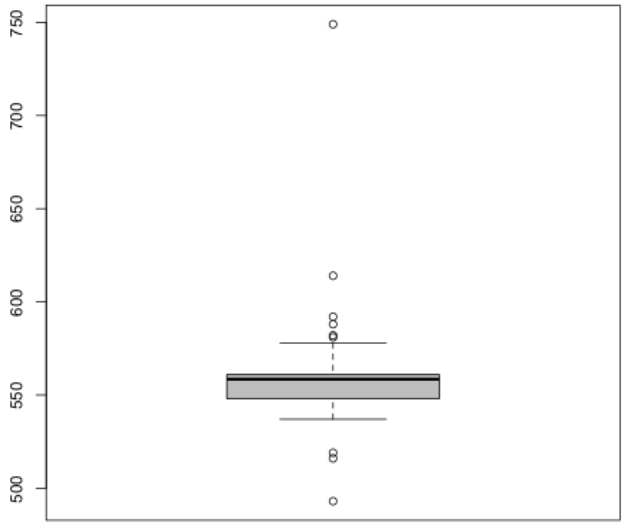


Figure 4.1 Variability of TR length.
Box-and-whisker plot depicts the mean and range of TR length in the studied strains. TR lengths are presented on the y-axis. The box extends from the 25th to the 75th percentile. The line inside the box denotes the median value, and the lines extending from either end of the box denote the extent of the data beyond the 25th and 75th percentiles and outliers.

To assess the degree of the diversity between strains within TR region, single nucleotide polymorphisms (SNP) density analysis was conducted with a 25-nt sliding window. The distribution of SNPs in the TR repeat unit in figure 4.2 A shows SNP density across the repeat unit as high as 12, which means that half of the bases in that window have SNPs. Even the 11- and 9-bp elements defined in B95-8 TR (Zimmermann & Hammerschmidt 1995) are not

conserved as might have been expected. Only two short regions (35 bp from 169902 to 169936 and 32 bp from 170040 to 170071 in NC_007605) are conserved. There is no known functional importance for the first conserved region while the second one has been reported to be essential for cleavage and packaging, as described below (Zimmermann & Hammerschmidt 1995). It is worth noting that the other apparently conserved regions in figure 4.2A are due to gaps in the MSA that are filled by unique sequences.

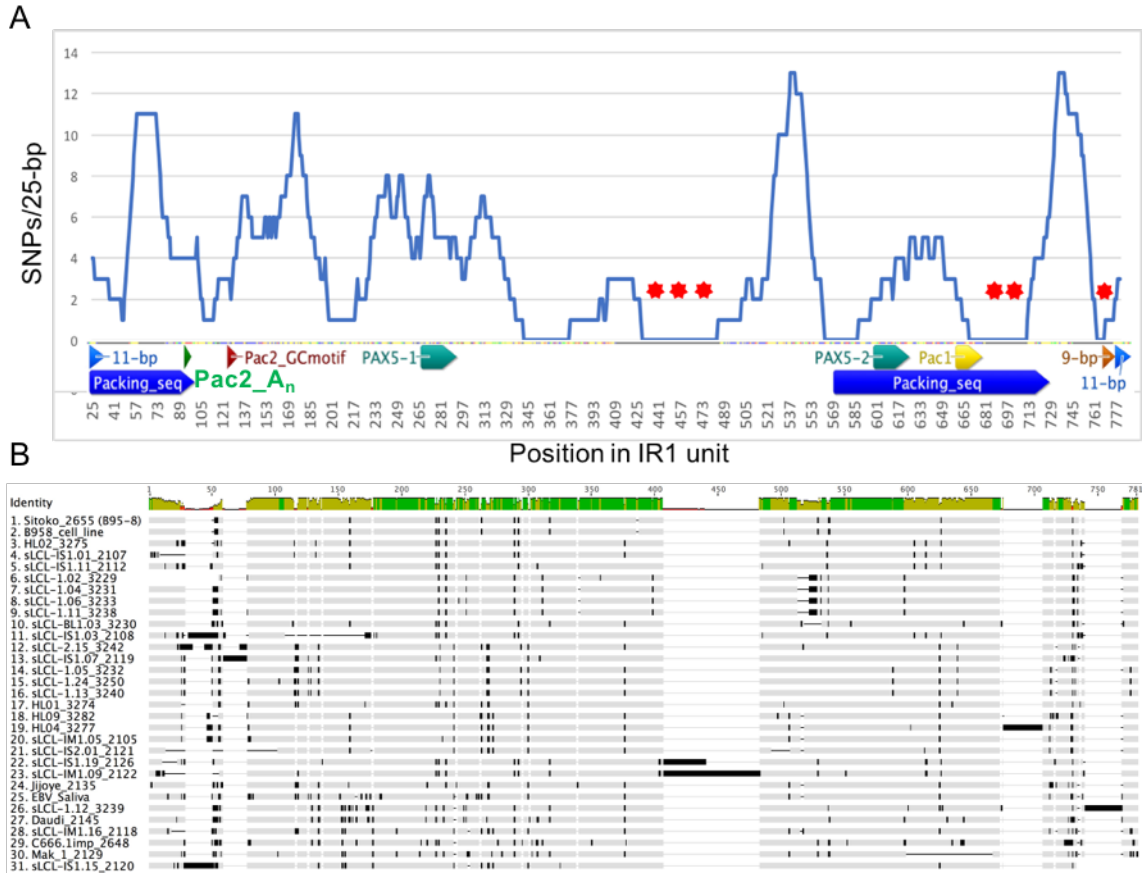


Figure 4. 2 TRs regions are highly variable.
(A) Polymorphism density in the TR region across 31 strains. The line graph is plotted across the TR unit showing the number of positions with SNPs found in more than one strain in a 25-bp sliding window. Arrows underneath the plot indicate positions of the 9- and 11-bp elements. Red asterisks indicate that apparently conserved regions are due to gaps in the MSA. Positions of functional sequence are shown underneath the graph. (B) A screen shot of Geneous showing MSA of the TR unit in 31 strains. The top line shows the depth of supporting read and the colour represents the identity: green, 100% identity; and green/brown,30-less than 100% identity. Black dots and bars show the disagreement to the consensus. (MSA available in SD4.1 at http://www.epstein-barrvirus.org.uk/MBaData.php).

To identify whether the 9 and 11-bp elements identified in B95-8 (figure 1.11) are present in other strains, the MSA highlighted a motif (N_4-CAT-N_4) that resembles the 11-bp elements in B95-8 (G_4-CAT-G_4), existed on both termini unless there is a deletion. For the 9-elements, there was motif on the right terminus (GTGT-N_5) in 30 strains (out of 31) resembles the 9-bp elements of B95-8 (**GTGT**TGGCG) and is followed by an 11bp element motif (Table 4.2). However, only seven strains have identical 11bp elements on both termini while the others have different 11bp element sequences. Table 4.3 shows the 11bp element-like motifs (designated A, B, C, to T) in both termini of each strain with the potential 9-bp elements (designated 9A, 9B, 9C, to 9H). Since I am working on one TR unit, it is not clear which 11-bp elements style is dominant in internal repeat units within each strain.

Table 4.2 potential 9- and 11-bp element sequences

11-elements style	Sequence	Length	No. of examples
A	GGGGCATGGGG	11	18
B	GGAGCATGGGG	11	6
C	GGGGCAGGGGG	11	2
D	AGGGCATGGGG	11	4
E	GGGCGTCGGGG	11	1
F	GGGCATCGGGG	11	1
G	GGCGCATGGGG	11	1
H	GAGGCATGGGG	11	2#
I	GGCATCGGGGG	11	3
J	GGGG--TGGGG	9*	2
K	GGAGCAGGGGG	11	1
L	GGGTCATGGGG	11	2
M	GGGGCGGGGGG or GGGGC--GGGGGG	11 or 9	1
N	GGGGCATGGGT	11	1
O	GGGCCATGGGG	11	3
P	GGGCCATGGGT	11	1
Q	GGGGCATGGGC	11	2
R	GGGCCATGGTG	11	2
S	GGGATAGATGG	11	1
T	GCGGGAGCGGG	11	1
U	GGGGGATGGGG	11	1
9-elements style	Sequence	length	No. of examples
9A	GTGTTGGCG	9	8
9B	GTGTTGGGA	9	5
9C	GTGTTGGAG	9	5
9D	GTAGTGGAG	9	1
9E	GTGTCGGAG	9	3
9F	GTGTTGGGG	9	1
9G	GTGTCGAGA	9	1
9H	GTGTTGATG	9	5
# - both in the same strain			
* missing "CA" from CAT motif			

Table 4.3 Potential 9- and 11-bp element styles in studied strains

Strain	11 bp element style	11 bp element		9 bp element style
		left side	right side	
Sitoko (B95-8	AA	A	A	9A
B958_cell_line	AA	A	A	9A
sLCL-IS1.19_2126	-C	deletion	C	9A
sLCL-IM1.09_2122	BC	B	C	9A
HL02_3275	LJ	L	J	9A
sLCL-IS1.01_2107	LJ	L	J	9A
sLCL-IS1.11_2112	MI	M	I	9A
sLCL-IS1.03_2108	II	I	I	9A
HL01_3274	AA	A	A	9A
sLCL-1.02_3229	NB	N	B	9B
sLCL-1.04_3231	OB	O	B	9B
sLCL-1.06_3233	OB	O	B	9B
sLCL-1.11_3238	PB	P	B	9B
sLCL-BL1.03_3230	OK	O	K	9B
sLCL-1.05_3232	QD	Q	D	9C
sLCL-1.24_3250	AD	A	D	9C
sLCL-1.13_3240	QD	Q	D	9C
sLCL-1.12_3239*	RA	R	A	9C
C666.1imp_2648	SU	S	U	9C
sLCL-2.15_3242	HH	H	H	9C
sLCL-IS1.07_2119	AD	A	D	9D
sLCL-IM1.16_2118	-B	Deletion	B	9E
Jijoye_2135	AA	A	A	9E
Mak_1_2129	-G	deletion	G	9E
sLCL-IS1.15_2120	T-	T	deletion	9F
Daudi_2145	RF	R	F	9G
HL09_3282	-A	"GGGC-deletion"	A	9H
HL04_3277	AA	A	A	9H
sLCL-IM1.05_2105	AA	A	A	9H
sLCL-IS2.01_2121	-A	deletion	A	9H
EBV_Saliva	AE	A	E	9H
C666.1imp_2648	SU	S	U	No motif

* has insertion of 29-bp on the right terminus includes another 11 style R and A

4.4 Regions of functional importance in TR are highly conserved

Since the TR region is the place where linearization and circularization of the genome occurs, Zimmermann & Hammerschmidt 1995, conducted a transient packaging assay using serial TR deletion mutants to identify the critical region(s) for packaging. They identified a region of 159-bp essential for packaging, from 170050 to 170208 of the first TR in NC_007605. It includes the 11-bp elements and two hypothesised packaging signals (pac 1 and pac 2) that resemble the packaging signals of alpha- and betaherpesviruses. However, they did not further narrow the region to formally identify the required sequences for packaging.

From the MSA of 31 strains, the first 31-bp of the essential 159-bp are conserved across all strains, but three (sLCL) strains have SNPs at the same position. Two regions have been reported to be conserved at herpesvirus termini called packaging signal 1 (pac1) and packaging signal 2 (pac2) (Deiss et al. 1986; Zimmermann & Hammerschmidt 1995). Our data showed that pac1 (CCCCCCAGGAAAGACCCCC) and pac2-An (AAAAA) are conserved in all strains except Mak1 and sLCL-IS2.01, which lack pac1 and pac2-An respectively due to deletions. The second component of pac2, pac2-GC-motif (GCGGCCC) is less conserved as it exists in 26 strains out of 31. It is worth noting that there is a deletion at the end of one TR unit in the B95-8 sequence, which means the 159-bp element is found in only two internal TR units while the conserved 31-bp and pac1 exist in every repeat unit of B95-8.

Also, it has been reported that Pax5 (Paired box protein 5) binds with high occupancy to two sites within TRs; Pax5-site 1 (193-218 in TR unit) and Pax5-site 2 (443-468 in TR unit) (Arvey et al. 2012). Pax5 is a B cell-specific transcription factor that regulates B cell development by enhancing B-cell specific gene expression and is also involved in EBV latency by activating the Wp promoter (Medvedovic et al. 2011; Tierney et al. 2007). However, the

sequences of Pax5-site 1 and Pax5-site 2 reported in Arvey et al. 2012 (based on B95-8) are not highly conserved as they were found in 12 and 15 strains out of 31 respectively. Interestingly, only B95-8 has the two Pax5 binding sites, while the other strains either have one binding site (Pax5-site 1 or –site 2) or do not have it at all (6 strains). However, there is a conserved sequence within the Pax5-site 1 and 2 "GCGTGAC" that exists in all strains, suggesting that this sequence is a core binding site.

4.5 The read bias score threshold was deactivated in defining minor variant within TRs

Defining the minor variants was conducted by using the same strategy applied in section 3.9 but with a minimum threshold of 8%. There was no study tested the number of TR copies across different EBV strains. The minimum threshold was chosen as the highest TR copy number reported in a sequenced EBV genome was 12 copies in the YCCEL1 BAC (accession number AP015016), which was sequenced using the PacBio sequencing platform (Kanda, Furuse, Oshitani, et al. 2016). Because of the relatively short length of the TR unit (~500-bp), variants with a high read bias score were expected due to the big difference between the number of reads mapping close to the edges of TR units and reads mapping to the junctions between the TR unit and the flanking sequences (figure 4.3). However, unlike the IR1 analysis in section 3.9.3, it was not possible to distinguish flank-SNPs from internal ones since there was one repeat unit in the used reference (the template). Since the edges of TR units are not identical (i.e. 11-bp element) in most studied strains (table 4.3), it was not known which 11-bp element style is in the internal TR units. Therefore, constructing template with 2 TR units would create ambiguities regarding the sequence of the internal junction during read mapping.

As a result, filtering heterogeneity from technical errors was conducted as described in (section 3.9) but the threshold was set on the minor-variant frequency only, ignoring the read bias to avoid losing potential genuine minor-variants. (All MVs within the TR unit, for each strain are summarised in table ST4.1 at http://www.epstein-barrvirus.org.uk/MBaData.php).

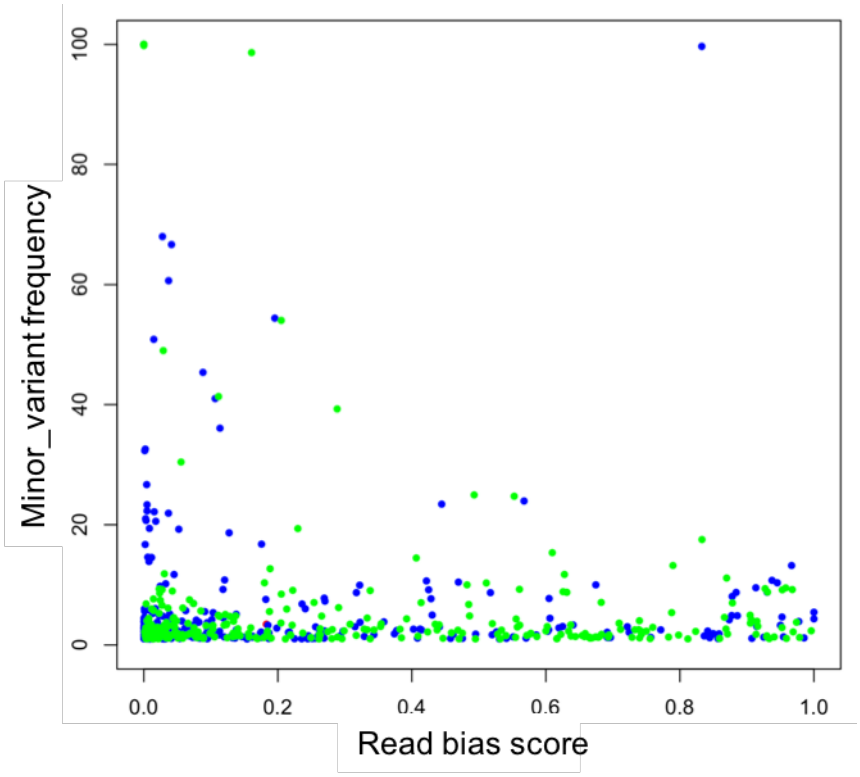


Figure 4.3 Visualizing MVs in the TRs region.
The graph plots the read bias (x-axis) against the MV frequency (y-axis) for every potential MV with a frequency of over 1% across 31 strains. Green points are MVs at positions where at least one strain has an SNP (recombination-like MVs). Blue points are randomly distributed MVs (mutation-like MVs).

4.6 The B95-8-cell line isolate from Hong Kong does not have the short TR unit

In the prototype EBV sequence (NC_007605), one repeat unit (starting at 170175 and ending at 170697) is 15-bp shorter than the others. This deletion created a mismatch when TR units were aligned to each other (figure 4.4A), which in turn is supposed to result in a minor variant at the position of the mismatch when reads map to one repeat unit. I managed to verify the presence of the short TR unit in B95-8-BAC (HB9) by Sanger sequencing (figure 4.4B). However, TR analysis showed that B95-8-cell line (sequence data from Dr Alan Chiang) seemed to have homogeneous repeat units which did not agree with the available data (NC_007605).

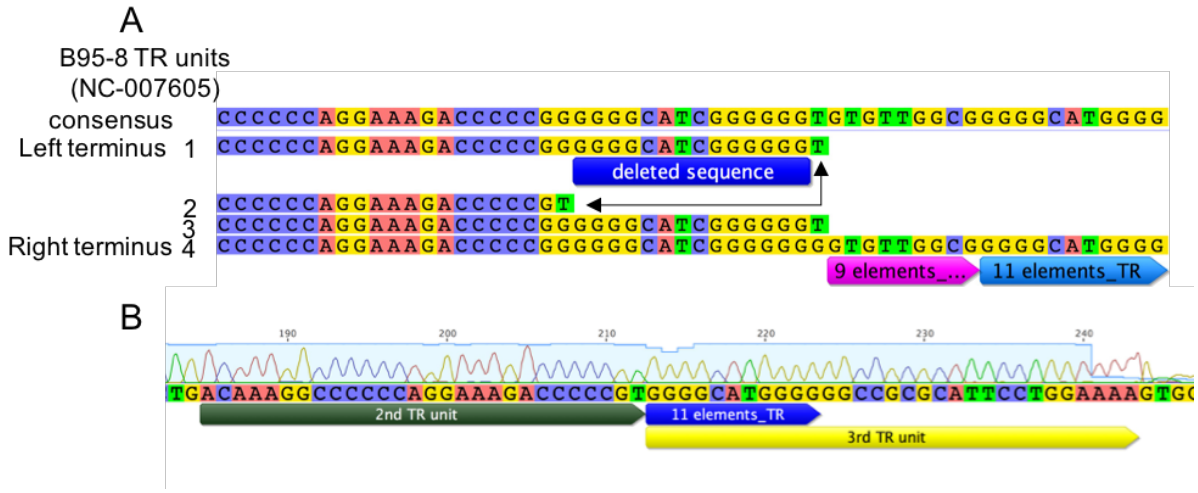


Figure 4.4 B95-8 terminal repeat alignment.
A. The four terminal repeat units of the prototype EBV genome B95-8 (accession no. NC_007605) were aligned and the presented part shows the end of each repeat unit. The consensus sequence is presented on the top. The deleted sequence in the 2nd unit is shown and the arrows show the new position of "T" after the deletion, which creates a mismatch. The 9bp-element and 11bp-element are shown below the 4th unit. **B.** Sanger sequencing of TR in B95-8-BAC (HB9) verifying the presence of the short TR unit.

To verify the output, a template was constructed containing an intact TR unit, the truncated TR unit, and the last intact TR unit which finishes with the 9-bp and 11-bp elements, flanked by 300-bp of unique sequences upstream and downstream of the TR region. Reads from the three B95-8 samples (del EBER, Sitoko and B95-8-cell line) were mapped to the constructed template to force reads from the short unit to map to the corresponding location. The consensus output of the alignment verified that the short TR unit is missing in the B95-8-cell line, as there were no reads mapping to the short unit (figure 4.5). However, the Sitoko sample (whose sequence is indistinguishable from B95-8) showed a minor-variant "T" at the position of the mismatch with a frequency of 23% which matches the copy number of TR units in B95-8 (4 units).



Figure 4.5 B95-8 cell line is missing the known short TR unit.
A screen shot of Geneous showing B95-8-cell line sequence reads (black bars) mapping to a template includes the short TR unit flanked by two intact TR units. The generated sequence is shown at the top and the black region in the consensus indicates un-supported region as there are no reads mapping to the junction between the short TR and the following TR unit. Orange lines are trimmed sequence in sequence reads.

4.7 Spontaneous mutations are common in the TR region, but recombination-derived diversity is more common in tumour and tumour-derived cell lines

In this thesis, I have proposed that sequence heterogeneity in repeats can arise from two sources: 1. recombination (sequence variants also exist as inter-strain SNPs) or 2. mutation (does not exist as SNPs).

All identified minor variants in all trains are summarized in table ST4.1 at http://www.epstein-barrvirus.org.uk/MBaData.php. Strains were sorted into two groups according to whether the virus was isolated from blood or saliva or isolated from a tumour biopsy or tumour-derived cell line. To get a general view of the nature of the diversity in TR of the two types of sample (sLCLs and tumour-derived sample), strains were classified into three groups based on the type of MVs (regardless of the number of MVs): homogeneous, in which strains have identical repeat units; recombination-like MVs, in which at least one MV is also a SNP elsewhere; and mutation-like MVs, where strains have at least one spontaneous mutation, but no recombination-like MV. The pie charts in figure 4.6 clearly show that 70% of tumour-derived strains that were analysed have recombination-like MVs while only 10% have homogeneous repeat units. In contrast, sLCLs strains are split evenly into the three groups suggesting that heterogeneity in the TR region is common in circulating EBV.

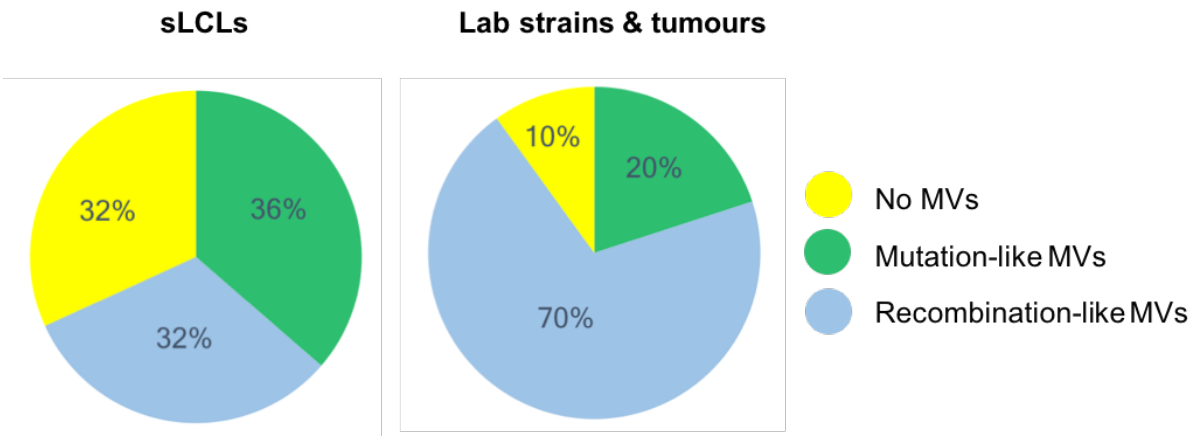


Figure 4.6 Minor variant accumulation in TR region in spontaneous LCLs and tumour-derived strains.
Pie charts show the types of genetic variations found in TR region from sLCL and tumour groups. The type of variation was classified as either homogeneous repeats (yellow), at least one recombination-like MV (blue), or only mutation-like MVs (green).

Strains were further analysed and grouped according to MV types and numbers. Looking at total MVs in the strains (figure 4.7 A), there is no clear difference between tumour-derived strains and sLCLs strains in the number of MVs that they harbour. This relative similarity is attributed to spontaneous-like MVs (figure 4.7 B). The high number of MVs in the saliva-derived EBV's TR is in agreement with the high number MVs throughout its genome compared to other strains (see section 3.14). In contrast, the majority of sLCL strains (70%) lack recombination-like MVs whereas the same percentage of tumour-derived strains have a minimum of one recombination-like MV (figure 4.7 C). Unlike IR1, there are no obvious examples of larger numbers of MVs in tumour-derived viruses. Five strains (out of 10) with homogeneous TR units have homogenous IR1 regions (sLCL-1.02, sLCL-1.06, sLCL-1.11, sLCL-BL1.03, and sLCL-IS1.19). EBV strain C666.1imp showed a relatively high number of recombination like-MV in IR1 and TR regions. The presented data cannot lead to a clear conclusion because of the limited number of strains. Nevertheless, it is quite striking that

recombination-like MVs are more common in both IR1 and TR regions in tumour-derived strains (see section 3.11).

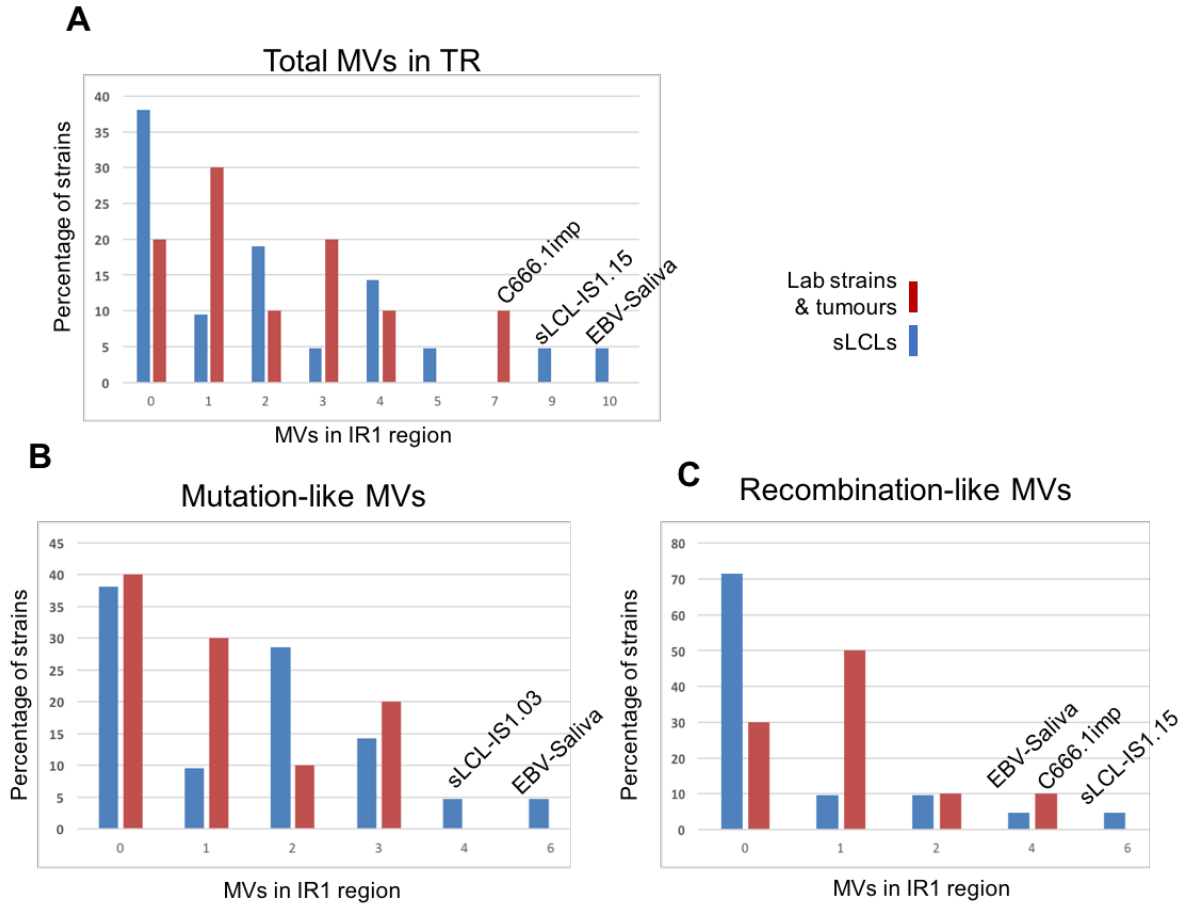


Figure 4.7 TR region diversity in spontaneous LCL and tumour-derived samples.
The graphical representations show the number of MVs in the TR region in each strain (31 sprains). Samples are grouped as being either spontaneous LCLs (including one saliva-derived sequence -blue bars) or tumour-derived (red bars). The X-axis shows the number of MVs while the Y-axis shows the percentage of strains with that number of MVs. Figures show: (A) total MVs, (B) MVs that resemble spontaneous mutations (mutation-like MVs), and (C) MVs that resemble SNPs (recombination-like MVs).

4.8 Summary

In this chapter, I applied the strategy described in chapter 3 to analyse the TR region. TR was hard to assemble by mapping reads to a universal reference due to a high level of diversity. Therefore, I managed to construct one TR unit for only 31 strains by de novo contig assembly. However, these sequences are preliminary versions, as they are dependent on the edge-based sequences (not internal).

Based on the analysed strains, the 11-bp element is surprisingly poorly conserved across strains, whereas the two packaging signal sites pac 1 and pac 2 are relatively conserved. I found that some B95-8 cell lines do not have the truncated TR unit reported in the original sequence, suggesting that it is not biologically significant. Heterogeneity in the TR region is common but, as for IR1, TR repeat recombination-like MVs are much more common in tumour-derived cell lines.

5 Producing circulating EBV viruses

5.1 Attempts to obtain viruses from cell lines by inducing the lytic cycle.

Studying the biological properties of viruses requires producing large amounts of virus in vitro. In LCLs, EBV is latent (latency stage III), and induction of EBV latently infected cells is a challenging task. There have been no reports of cells that are fully permissive for the EBV lytic cycle in vitro. Only a few cell lines respond well to induction and produce viruses in relatively large amounts such as B95-8, P3HR1 (derived from African Burkitt's lymphoma cells) and Akata (derived from Japanese Burkitt's lymphoma tissue) (Hutt-Fletcher & Turk 2001).

As described in section 1.6.1, lytic replication can be induced either by treatment with chemical inducers or by transfection of the BZLF1 or BRLF1 expression vector. Since I am seeking wild-type EBVs, I intentionally used lymphoblastoid LCLs (sLCLs) as they are most likely to host the virus closest to circulating EBV (non-disease associated EBV). sLCLs can sometimes be obtained by culturing peripheral blood lymphocytes from an EBV seropositive person for around a month in the absence of T-lymphocyte activity by using cyclosporine A (Nilsson 1971; Rickinson et al. 1984). Since they arise without any exogenous infection, relying on virus from the lymphocyte donor, these are called spontaneous LCLs.

Despite the fact that inducing the lytic cycle is challenging, I present in the following sections my attempts to induce the lytic cycle in sLCLs by inducing conditional BZLF1 expression and by treating cells with chemical inducers of the lytic cycle to find out how permissive sLCLs are for lytic cycle induction. Two cell lines were used as a pilot study, sLCL-IM1.16 (type 1 EBV) and sLCLsLCL-IS2.01 (type 2 EBV). These two cells lines were chosen

randomly, but my analyses of IR1 in sLCLsLCL-IS2.01 now suggests that this may contain a mixture of type 1 and type 2 properties (see section 3.15).

5.1.1 BZLF1 induction

The lytic cycle has been induced by inducing ZEBRA production (BZLF1 product), which in turn activates EBV early genes. The delivery of the BZLF1 protein into latently EBV-infected cell lines has been conducted by using gene delivery methods such as transfecting cells with a plasmid carrying the BZLF1 gene (Jenkins & Binne 2000) or infecting cells with replication-deficient adenovirus vectors expressing BZLF1 (Westphal et al. 1999). Owing to the low efficiency of transfecting many B cell lines (Dr Rob White's observation), BZLF1 was delivered by constructing lentivirus vectors expressing BZLF1. Lentivirus surpasses adenovirus at several points. For instance, lentivirus vectors are easier to deal with because of the small genome size, and are suitable for making stable cell lines as they have long lasting transgene expression due to integration.

5.1.1.1 Infecting cells with a CMV promoter-based lentiviral vector expressing conditional BZLF1

I used a lentiviral vector expressing BZLF1 conditionally to get stable cell lines and available for lytic induction if they respond to BZLF1 induction method. Dr White has previously generated two CMV promoter-based lentivectors encoding either mCherry (p8270) or a conditional BZLF1 (p8072). The conditional BZLF1 is encoded by a BZLF1 gene fused to the 4-hydroxytamoxifen (4HT)-responsive mutated estrogen receptor hormone-binding

domain (ZHT) to make a putative inducer of EBV replication with 4HT treatment (Johannsen et al. 2004) (figure 5.1). sLCL-IS2.01 and sLCL-IM1.16 cells were used for infection along with B95-8 as a control. The cell lines were infected with lentivectors expressing conditional BZLF1-ZHT or mCherry to verify the infectivity of constructed vectors under the fluorescent microscope (for lentivirus production and infection, see section 2.7).

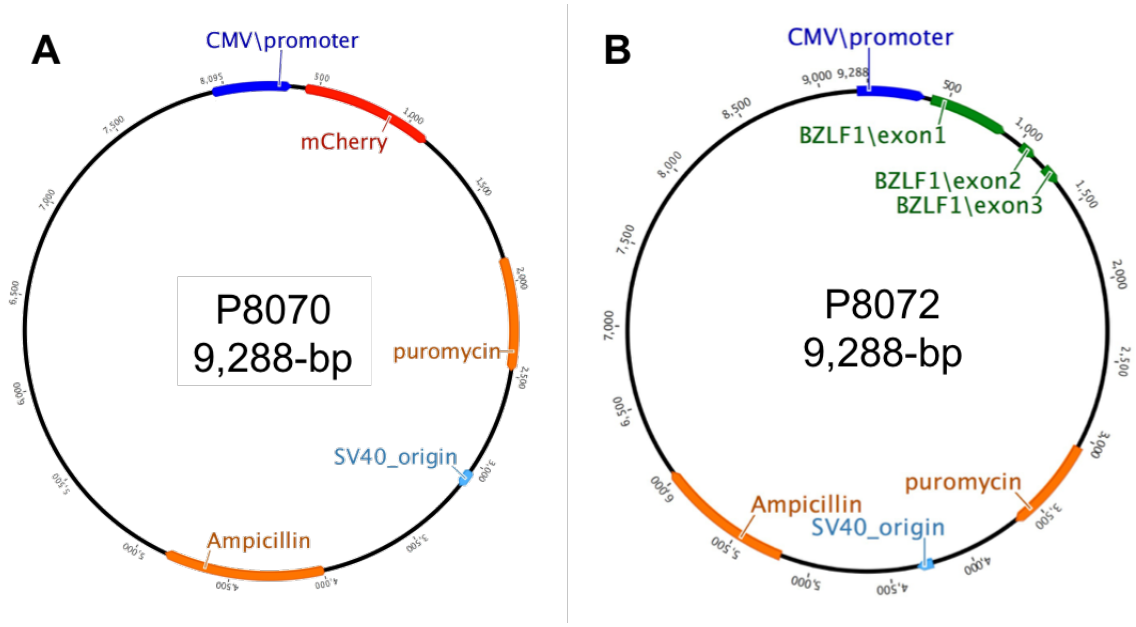


Figure 5.1 Structure of lentivirus vectors.
(A) p8270 is a CMV promoter-based lentivector encoding mCherry and was used as an indicator of the efficiency of infection. (B) p8072 is CMV promoter-based lentivector encoding a conditional BZLF1 (ZHT) which is a BZLF1 gene fused to the 4-hydroxytamoxifen (4HT)-responsive mutated estrogen receptor hormone-binding domain and becomes active in the presence of 4HT.

5.1.1.1.1 BZLF1 induction did not enhance viral proliferation but induced endogenous BZLF1 expression

To investigate whether exogenous BZLF1 induces viral replication in the studied strains, ZHT-transfected cells were harvested 96 h post treatment with 4HT and viral proliferation was evaluated by measuring the relative quantity of the EBV DNA polymerase gene (BALF5), which is a highly conserved nonrepetitive part of the EBV genome (Palser et al. 2015), by qPCR (for details see section 2.10). In figure 5.2, the readout of EBV proliferation was the appearance of BALF5 amplicon as detected by qPCR. It was clear that BZLF1 induction by adding 4HT could barely enhance the viral replication in LCL2.01 and LCL-IM1.16 compared to the control B95-8-ZHT which showed a proliferation 260-fold more than that of untreated B95-8-ZHT cells.

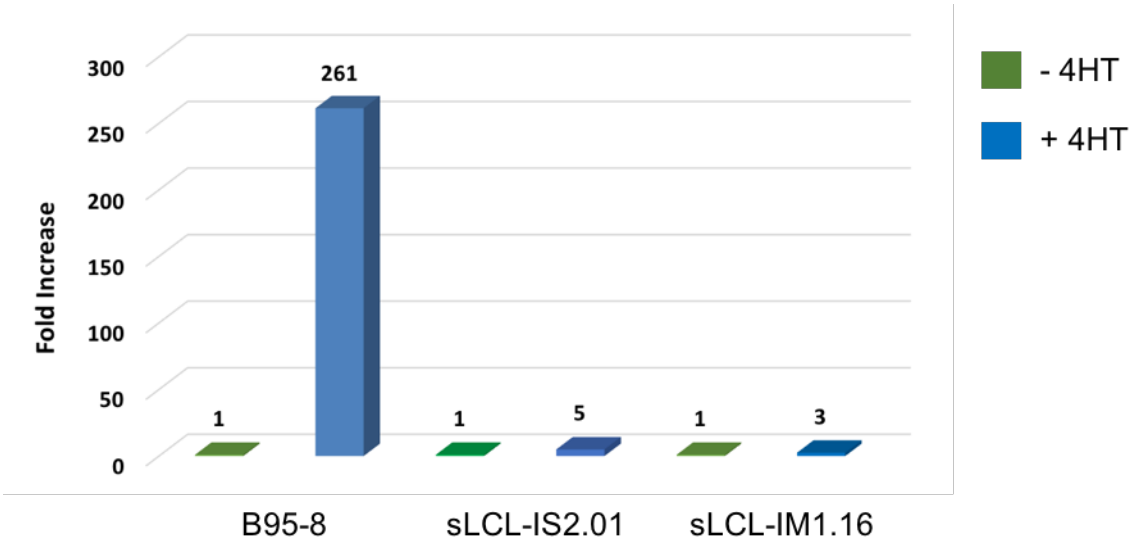


Figure 5.2 Evaluating the lytic cycle induction by activating pCMV-derived BZLF1. Cells harbouring pCMV-derived ZHT were harvested 96 h post treatment with 4HT, and viral proliferation was evaluated by calculating the relative expression of EBV DNA polymerase gene (BALF5) measured by qPCR in triplicate and calculated using delta delta Ct methods and normalized to an endogenous control, the human β-2microglobulin gene. (For qPCR conditions see section 2.10).

The expression of ZHT in infected cells was verified by Western blot. Since the ZHT protein is larger than BZLF1, it can be seen in figure 5.3 that activating ZHT (upper band) by 4HT induced the expression of endogenous BZLF1 (lower band) in all three studied samples.

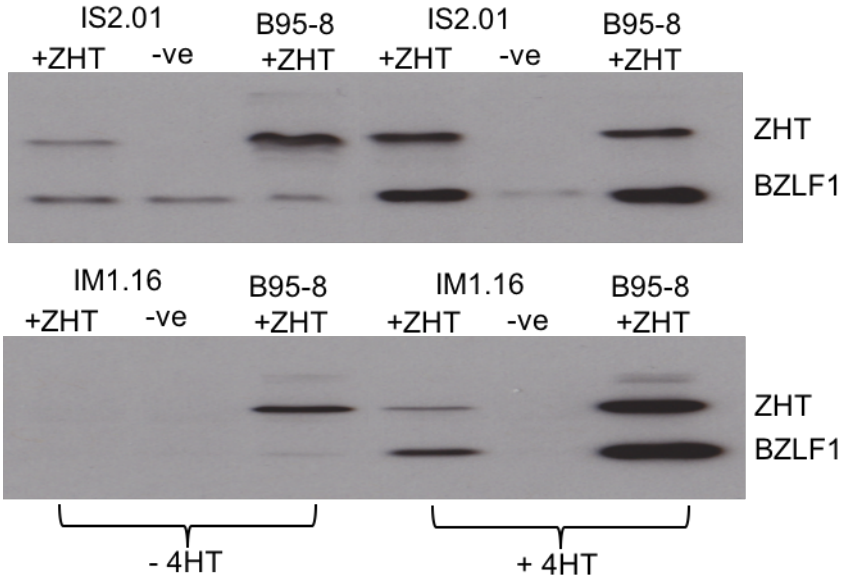


Figure 5.3 Western blot analysis of BZLF1 expression.
ZHT-transfected cell lines were uninduced (-4HT) or induced with 4HT (+4HT) and harvested 96 h post-treatment. Proteins extracted from cell lysates were separated by SDS-PAGE (12.5% polyacrylamide gel) and BZLF1 was detected by using BZ-1 antibodies. Negative samples were un-transfected cells.

5.1.1.2 Infecting cells with a CAG promoter-based lentiviral vector expressing conditional BZLF1

It has been previously observed that expression from the CMV promoter is unreliable in LCLs (Dr Rob White, unpublished data). Although the Western blot results (figure 5.3) showed that the CMV promoter-based lentivector expressed ZHT in transfected cells, I replaced the CMV promoter in the lentivectors (p8270 and p8072) with the CAG promoter

(pCAG) in an attempt to get a greater BZLF1 expression and consequently better lytic induction lytic induction. pCAG is a hybrid construct consisting of the CMV-immediate early enhancer, the chicken β-actin promoter, and rabbit β-globin intron (Niwa et al. 1991) and has shown to be highly active in many cell types, including cultured cells (Bornkamm et al. 2005; Chung et al. 2009).

5.1.1.2.1 Constructing CAG promoter-based lentivectors

The cloning scheme of constructing CAG promoter-based lentivirus vectors are shown in figure 5.4. The EcoRI site in the BlueScript-derived plasmid p4908 was removed to avoid undesired digestion in downstream cloning steps in the resultant lentivirus vectors pMBA01, and an oligonucleotide linker (pCAG linker) was inserted in plasmid p4908 to add required restriction sites in the resultant pMBA03. The pCAG promoter and enhancer were subcloned from p8306.6 – a construct based on pRTS (Bornkamm et al. 2005) – in a series of steps into pMB04 (figure 5.4A). The pCMV promoter in p8270 (the CMV-mCherry lentivector) was replaced with the CAG promoter from pMBA04 to obtain pMBA07 (figure 5.4B). To generate the CAG promoter-based lentivector expressing ZHT pMB09, the ZHT gene was derived from p8023 and subcloned in place of the mCherry gene in pMB07 (figure 5.4B).

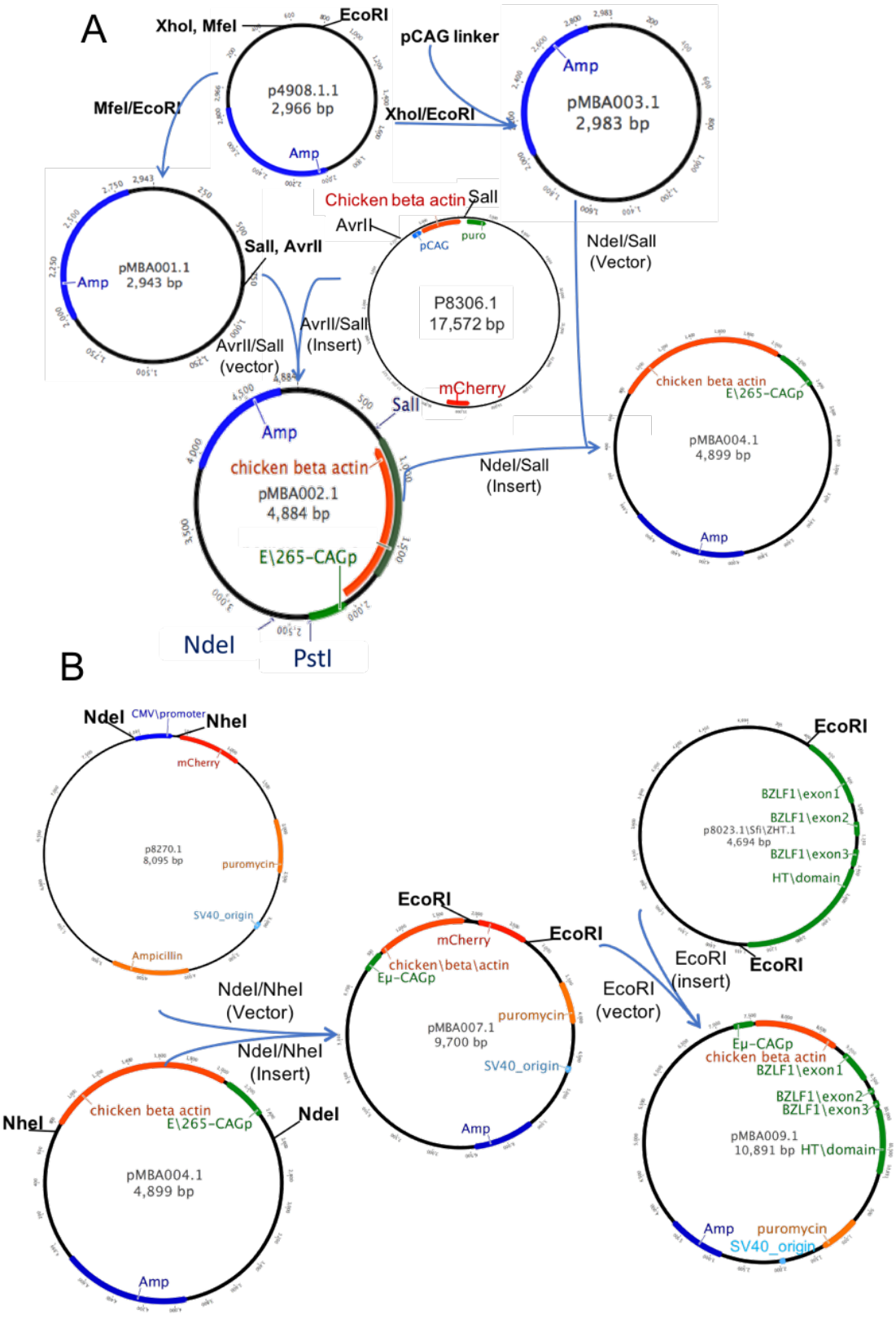


Figure 5.4 CAG promoter-based lentivector construction strategy.
The construction strategy is described in section 5.1.1.2. Plasmids and restriction enzymes are indicated.The final product pMBA007.1 is CAG promoter-based lentivector expressing mCherry, while pMBA009.1 is a CAG promoter-based lentivector expressing conditional BZLF1.

5.1.1.2.2 CAG promoter-derived BZLF1 induction has no effect on lytic cycle activation.

To test the ability of the pCAG-based ZHT lentivector to induce the lytic cycle, ZHT-lentivector infected cells of the three strains B95-8, LCLIS2.01 and LCLIM1.16 were harvested 96 h post treatment with 4HT. qPCR was used to measure viral proliferation by calculating the relative quantity of the EBV DNA polymerase gene (BALF5). As can be seen in figure 5.5, viral proliferation was even less than that in cells infected with the pCMV based ZHT lentivector. The limited lytic cycle induction in sLCLIS2.01 and sLCLIM1.16 might be attributed to the poor response of these cell lines to BZLF1 induction. It seems likely that the very limited viral proliferation in pCAG-ZHT infected B95-8 cells was due to the low efficiency of the CAG promoter at least in B95-8 which could be attributed to the cell type (Niwa et al. 1991; Bornkamm et al. 2005). The weakness of pCAG based expression was also noticed in cells infected with pCAG-based lentivectors expressing mCherry pMBA07 under the fluorescent microscope, compared to cells infected with pCMV-based lentivectors expressing mCherry p8270 (data not shown). Since the aim of this project was to produce viruses, I stopped at this stage without doing further investigation on expressing ZHT from various promoters, and moved forward with alternative methods.

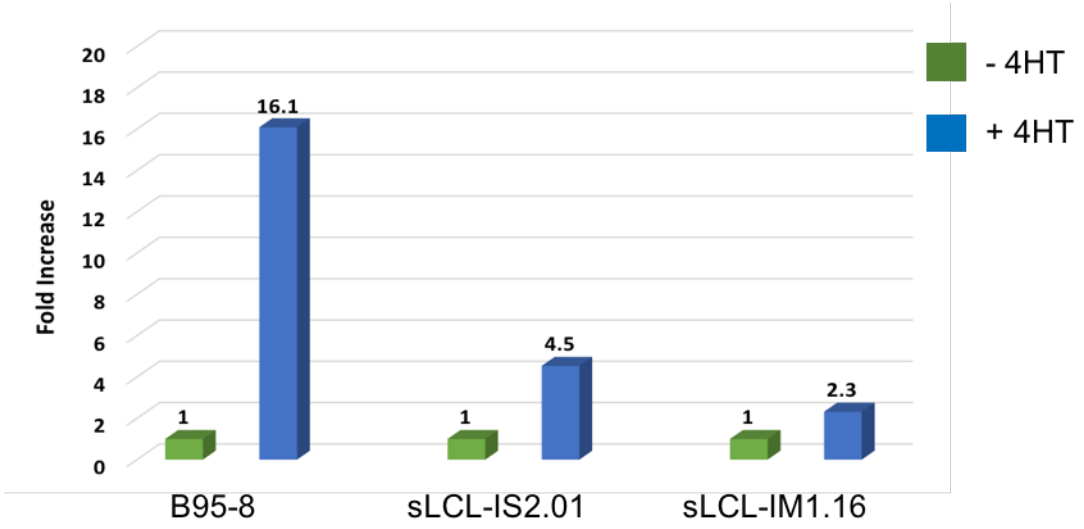


Figure 5.5 Evaluating lytic cycle induction by pCAG-derived BZLF1.
Cells harboring pCAG-derived ZHT were harvested 168 h post treatment with 4HT, and viral proliferation was evaluated by calculating the relative expression of EBV DNA polymerase gene (BALF5) measured by qPCR in triplicate and calculated using delta delta Ct methods and normalized to an endogenous control the human β-2microglobulin gene. (For qPCR conditions see section 2.10).

5.1.2 Inducing the lytic cycle by treating cells with chemical inducers

Besides transfecting EBV-infected cells with the BZLF1 gene to activate the lytic cycle in cell culture, the second conventional method is using chemicals agents (see section 1.6.1). No one chemical stimulus induces all cell backgrounds. For instance, it has been shown that the histone deacetylase (HDAC) inhibitors sodium butyrate (NaB) induces the expression of ZEBRA in P3HR1 but not B95-8. Conversely, tetradecanoyl phorbol acetate (TPA), which is a protein kinase C (PKC) activator, stimulates BALF1expression in B95-8 but not in P3HR1 (Gradoville et al. 2002). Treating B95-8 and P3HR1 with a combination of TPA and NaB increases the number of cell entering lytic cycle in P3HR1, while NaB has no effect on B95-8 (Gradoville et al. 2002). Thus, I used a combination of NaB and TPA to treat LCL-IS2.01 and LCL-IM1.16, along with B95-8 as a control. Also, to test if activating ZHT alongside chemical

stimuli can induce the lytic cycle in LCL-IS2.01 and LCL-IM1.16, I used the infected cells with the pCMV-based lentivector expressing ZHT p8072 for treatment with 4HT and a combination of NaB and TPA. To evaluate lytic activation, cells were harvested seven days post-treatment. In figure 5.6, IM1.16 showed a slight induction of viral proliferation after treatment NaB and TPA and BZLF1 activation (19.4 fold compared to untreated cells). The lytic cycle in B95-8 seemed to be more responsive to ZHT activation than to chemical treatment. Although the results in figure 5.2 and figure 5.6 were produced from independent experiments at different harvesting times (96 h and 168h), the similarity in viral proliferation in B95-8-ZHT after treatment with 4HT only (figure 5.2) and with NaB, TPA and 4HT (figure 5.6) suggests that ZHT activation is strong enough to achieve high viral proliferation in B95-8. It is likely that IS2.01 is resistant to chemical inducers. Since sLCL-IS2.01 showed a limited response to ZHT activation (figure 5.2 and 5.5), it is likely TPA and NaB did not improve the induction (figure 5.6).

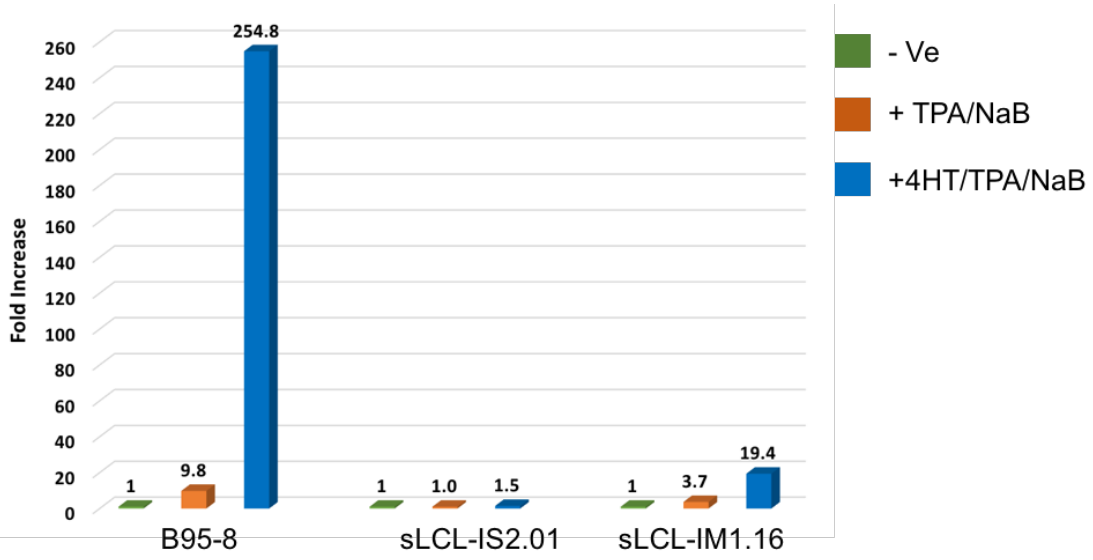


Figure 5.6 Evaluating lytic cycle induction using chemical inducers.
Cells harbouring pCMV-derived ZHT were harvested 96 h post treatment with 4HT only or with a combination of TPA, NaB and 4HT. Viral proliferation was evaluated by calculating the relative expression of the EBV DNA polymerase gene (BALF5) measured by qPCR in triplicate and calculated using delta delta Ct methods and normalized to an endogenous control the human β-2microglobulin gene. (For qPCR conditions see section 2.10).

In general, it is clear that the choice of lytic induction method depends on the nature of the cell line. The choice should be determined empirically, and it is not possible to assume based on an experiment applied on B95-8 or P3HR1 for instance. Viral replication induction levels were very low, and not sufficient to produce substantial quantities of virus. Therefore, I sought an alternative strategy to produce EBVs in reasonable amounts.

5.2 Cloning EBV genomes as BACs

As described in the previous section, it was difficult to induce the EBV lytic cycle in sLCLs to produce a significant amount of virus to perform infection experiments. Thus, an alternative way to study the biological properties of EBV viruses from source as close as possible to wild-type is to clone EBV genomes as BACs. Once an EBV genome is cloned as a BAC, it can be transferred into cells permissive for lytic cycle induction to conduct phenotypic studies. This also has the advantage of making the genome easier to manipulate than relying on viruses in cell lines.

The ultimate aim of this project was to develop a cloning technique which can be applied on EBVs from as many sources as possible. At the time of writing this section, six recombinant EBVs had been constructed; B95-8 (Delecluse et al. 1998), B95-8(ΔEBNA3B) (Chen et al. 2005), Akata (Kanda et al. 2004), M81 (Tsai et al. 2013), and SNU719 and YCCEL1 (Kanda, Furuse & Oshitani 2016). These recombinants were constructed by insertion of a prokaryotic F-plasmid (F-factor) into different sites in the EBV genome (see section 1.6.3.). All existing cloned EBV strains are type 1 and cloned from cancer cell lines, and they may carry mutations relative to naturally circulating EBV.

The presence of the BAC vector sequence in the viral genomes often causes alterations that would prevent the recombinant virus genome from producing virus particles due to the restricted size requirement of viral genome for packaging (Bloss & Sugden 1994). Also, the BAC vector may alter the expression pattern of neighbouring genes. So, to produce an EBV model as close as possible to wild-type, it is desirable to excise this extra foreign genetic material from the virus genome before any phenotypic characterization. Therefore, the aim of this project was to insert the BAC cassette within EBV terminal repeats (TRs) in EBV-infected B cells. This strategy has been done first on HCMV (Adler et al. 2000), and it has been done for rhesus-rhadinovirus RRV (a non-human model of EBV) and M81 EBV (an EBV strain maintained in epithelial cells) (Zhou et al. 2010; Tsai et al. 2013). Inserting BAC elements within the TRs does not (theoretically) disrupt coding regions and the BAC cassette is excised from EBV genome upon induction of the lytic cycle as a consequence of recombination between the two flanking terminal repeats (Feederle et al. 2010).

Previous approaches have relied on the incorporation of the BAC into the viral genome during lytic replication. The success of this method relies on viral DNA replication in infected cells. When virus undergoes lytic replication, it linearises its DNA for replication which in turn stimulates DNA recombination. In contrast to α- and β-herpesviruses, ɣ-herpesviruses are not highly replicative in cells, which makes homologous recombination between the plasmid and the resident EBV genome a real challenge. Furthermore, inducing lytic cycle in EBV-infected cells cannot be reliably achieved for all viruses, especially in sLCLs (Tsai et al. 2017).

To boost the chance of homologous recombination between the EBV genome and the donor BAC plasmid, I aimed to enhance the efficiency of homologous recombination by introducing double-strand DNA breaks in the EBV genome using a CRISPR (clustered regularly interspaced short palindromic repeat) type II system (Cong et al. 2013). The CRISPR

system has been used in EBV research fields as a tool to cleave DNA specifically for various purposes (see section 1.6.4). It has been used to construct a recombinant EBV that does not express BART (BamHI A rightward transcript) viral transcripts present in EBV-latently infected cells (Yuen et al. 2015). As a potential therapeutic strategy, CRISPR/Cas9 was designed to target several regions in the EBV genome in Raji cells to eradicate latent infection (Wang & Quake 2014). Since my aim was to insert a BAC vector into a site within the terminal repeats (TRs), there is a high likelihood that Cas9 would cut several repeat units. So, a controllable Cas9 was considered to reduce its strand breaking activity so as not to remove all of the TRs. There is a big chance to loose all TR units, but compensating the missed sequence could be easier in BAC. In this chapter, I present my efforts to clone EBV genomes as BACs and to simplify the cloning procedure.

5.2.1 Identifying potential insertion sites in the TR to be targeted by Cas9

Streptococcus pyogenes Cas9 (SpCas9) requires 20 nt identity followed by the protospacer adjacent motif (PAM) 5′NGG3′, which serves as a binding signal for Cas9. sgRNA binds to the target sequence, and Cas9 cuts 3 nucleotides upstream the PAM sequence. I tried to find consensus Cas9 targets across EBV strains to approach the ultimate aim which was to develop a cloning technique can be applied to as many as possible EBV-infected cell lines. The TR region was chosen as a target site for cloning to enable the EBV genome to excise the foreign DNA during the lytic cycle. As shown in section 4, the TR region is highly variable across EBV strains: only three entirely conserved regions of 20-bp followed by a PAM motif (5'NGG'3) were identified across the 31 sequenced strains (table 4.1) as potential targets for SpCas9 (figure 5.7).

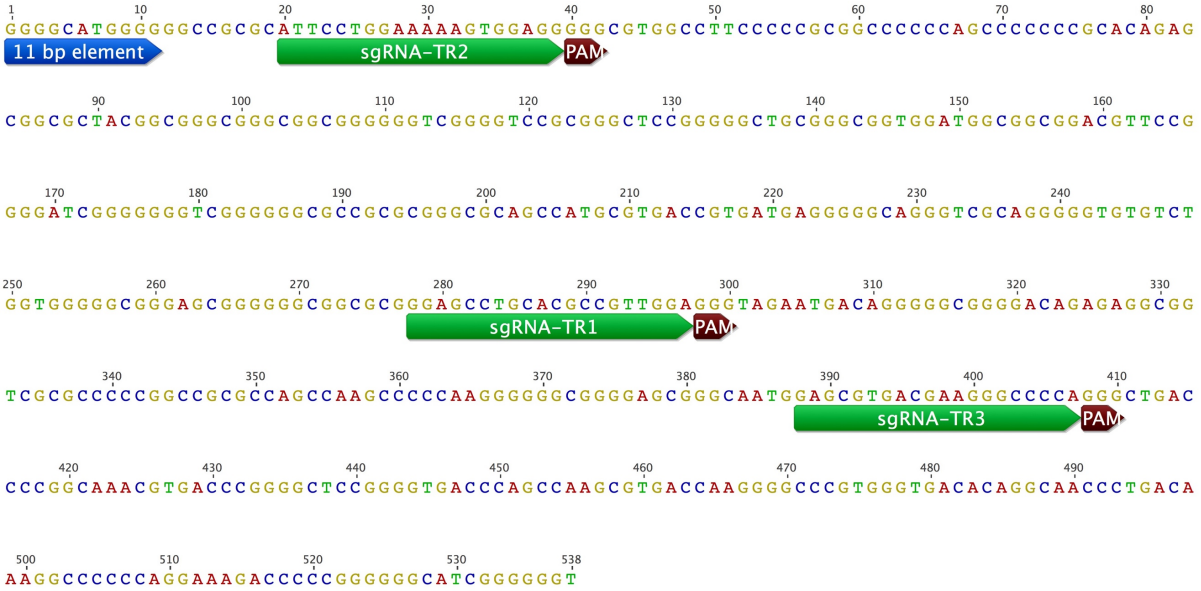


Figure 5.7 Positions of Cas9 targets in TR.
Nucleotide sequence of one unit of TR in B95-8: the sequence of one terminal repeat was assembled (see section 4.2). Across analyzed strains (31 strains in table 4.1) the conserved sequences of the three candidate CRISPR/Cas9 targets followed by PAM sequence are indicated with green and purple arrows respectively. The blue arrow indicates the sequence of the 11-bp element (Zimmermann & Hammerschmidt 1995).

5.2.2 Targeting the EBV TR region using a donor BAC plasmid with short homology sequences and conditional Cas9

5.2.2.1 Constructing CRISPR components

Since the intended targets sites for Cas9 are in a repetitive region, it is possible that Cas9 could cut every repeat unit, removing all of TR. As a strategy in which Cas9 protein does not last in cells for long, pKMD106e (Addgene 64190) that expresses a conditionally active Cas9 (intein-Cas9 (S219)) was used. The intein (an internal segment of amino acids) keeps Cas9 inactive until it is excised by the presence of 4HT (Davis et al. 2015). Three sgRNA producer plasmids were constructed by inserting an annealed oligo (TR-sgRNA_#_PUC57,

see table 2.4) for each of the candidate target sites into BsaI site of pUC57-sgRNA expression vector (Addgene 51132) to obtain the pUC57-sgRNA-TR# plasmids (where # is region 1, 2 or 3 from figure 5.7).

To test the functionality of CRISPR components, HEK-293 cells harbouring the B95-8 BAC (HB9) were transfected with pKMD106e and pUC57-sgRNA-TR3 vector and treated with 4HT (0.4 μM). Transfected cells were harvested at day 5 post-treatment with 4HT and DNA was isolated and transformed into DH10B to obtain chloramphenicol resistant colonies. Only one colony out of 24 showed a reduction in the size of the band that covers terminal repeat region, suggesting that one or more of the TRs has been deleted. (figure 5.8). This may highlight the low efficiency of the proposed procedure as only one colony showed the effect of Cas9, but it suggested that the Cas9 could cut TR units.

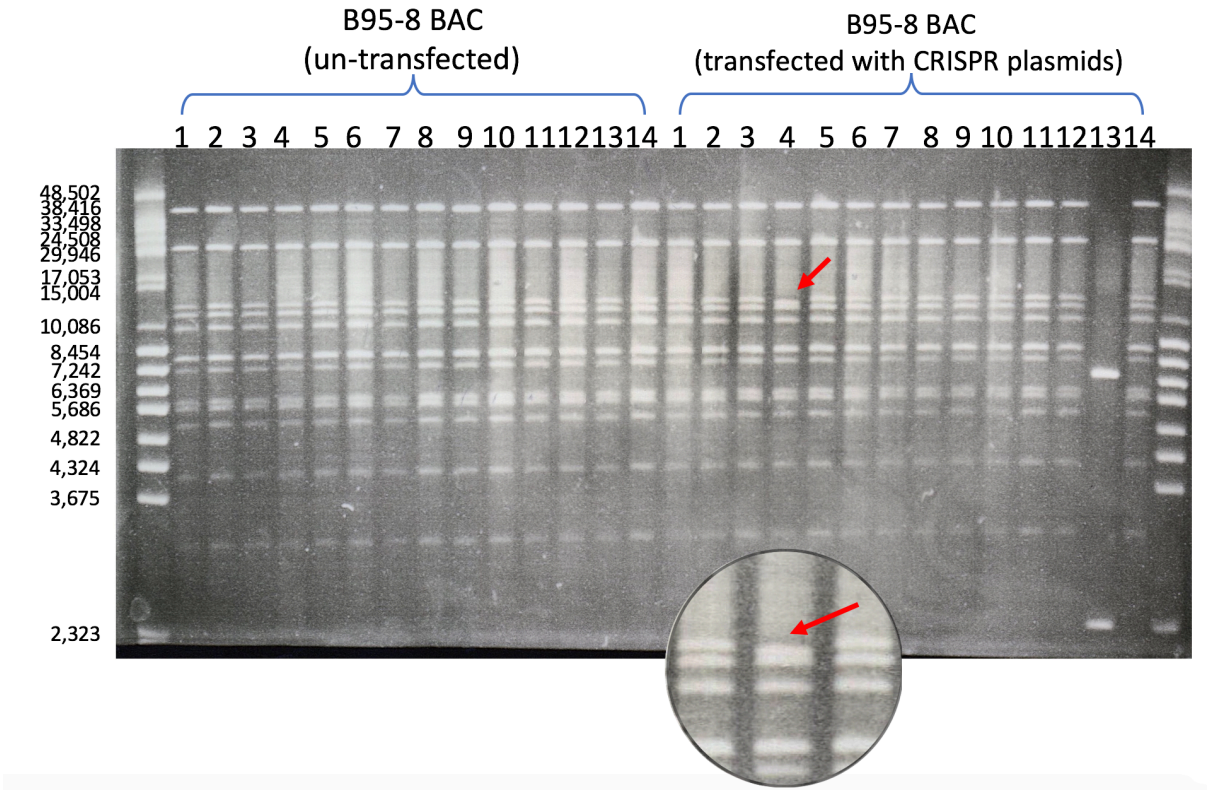


Figure 5. 8 Pulsed-field analysis of the functionality of CRISPR components in targeting TRs in B95-8 BAC.
293-B95-8-BAC cells were transfected with Cas9 and sgRNA plasmids and harvested 5 day post-treatment with 4HT. Miniprep DNAs were digested with EcoRI, and only one colony showed a missing band (12434 bp) that covers the TR region. The red arrows on the gel and the magnified picture indicate the missing band. The same number of colonies was tested from the sample and negative control to verify that the missing band in lane 4 of transfected samples was not a false positive.

5.2.2.2 Constructing donor BACs

The aim was to construct a BAC vector can be inserted into as many EBV genomes as possible by homologous recombination. However, because of the high variability in the TR region across EBV strains (see section 4.3), it was not possible to identify universally homologous regions large enough. Therefore, every donor BAC has to be built with homologous sequences derived from the corresponding strain.

Since there is not currently a type 2 EBV strain cloned as a BAC, the EBV type 2 strain AG876 was the first strain to be tested for cloning. First, an R6K plasmid was used in constructing donor BAC plasmid to enhance DNA yield. Since the intended BAC insertion site is at the Cas9 cutting site, the homologous sequences used were 50 nt upstream and downstream of the Cas9 cutting site. Three plasmids were constructed, each with sequences homologous to one of the three Cas9 target sites. Homologous sequences were added to the R6K plasmid by PCR using primers R6K-AG876-sgTR#-fwd and R6K-AG876-sgTR#-rev) to obtain the R6K-AG876-TR# plasmid (where # is 1, 2 or 3) (for PCR condition see section 2.9). The forward and reverse primers include sequences upstream and downstream from the Cas9 target site respectively. These primers consist of: 5'- restriction enzyme sites (AscI-AvrII-SalI; 17 nt), homologous sequence (50 nt) and then PCR primer sequence that binds to the R6K plasmid (24nt for forward and 21nt for reverse)-3'. The PCR products were digested with AvrII, ligated and transformed to produce a circular plasmid (pR6K-AG876-TR# , where # is 1, 2 or 3). The pR6K-AG876-TR# plasmids were digested with AscI and sub-cloned into a BAC vector (p4348.2) carrying dsRed and hygromycin expression cassettes, to obtain pBACR6K-AG876-sg# (figure 5.9). BAC construct DNA was maxiprepped by CsCl density gradient centrifugation.

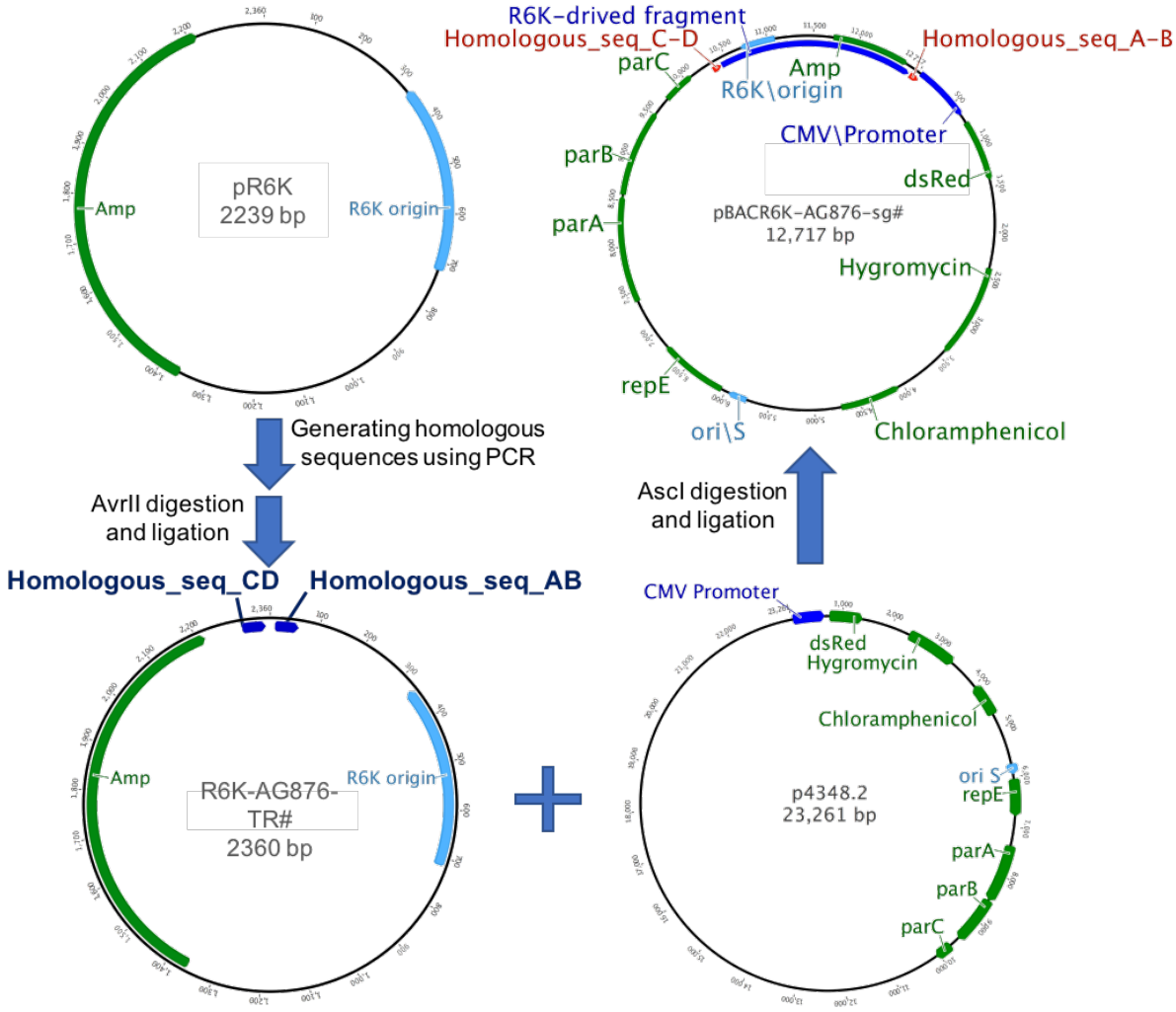


Figure 5.9 Donor BAC (pBACR6K-AG876-sg#) construction strategy.
To generate a donor plasmid with homology sequences to the Cas9 target site, 50 bp of upstream homology arm (indicated as A-B) and downstream homology arm (indicated as C-D) were generated in the high copy number plasmid (R6K) using PCR. The PCR product was circularized and minipreped. The donor BAC was generated by sub-cloning R6K-AG876-TR# plasmid into AscI sites of the BAC vector p4348.2 (# is the Cas9 target 1,2 or 3).

5.2.2.3 Targeting EBV using a donor BAC plasmid with short homology sequences of the TR region and conditional CRISPR/Cas9 components

I used 24 different programmes to identify the best conditions for electroporating AG876 cells, and identified programme 16 (pulse voltage: 1400, pulse width 20, pulse No.2) as the best (figure 5.10) (see section 2.12.2 for optimization protocol).

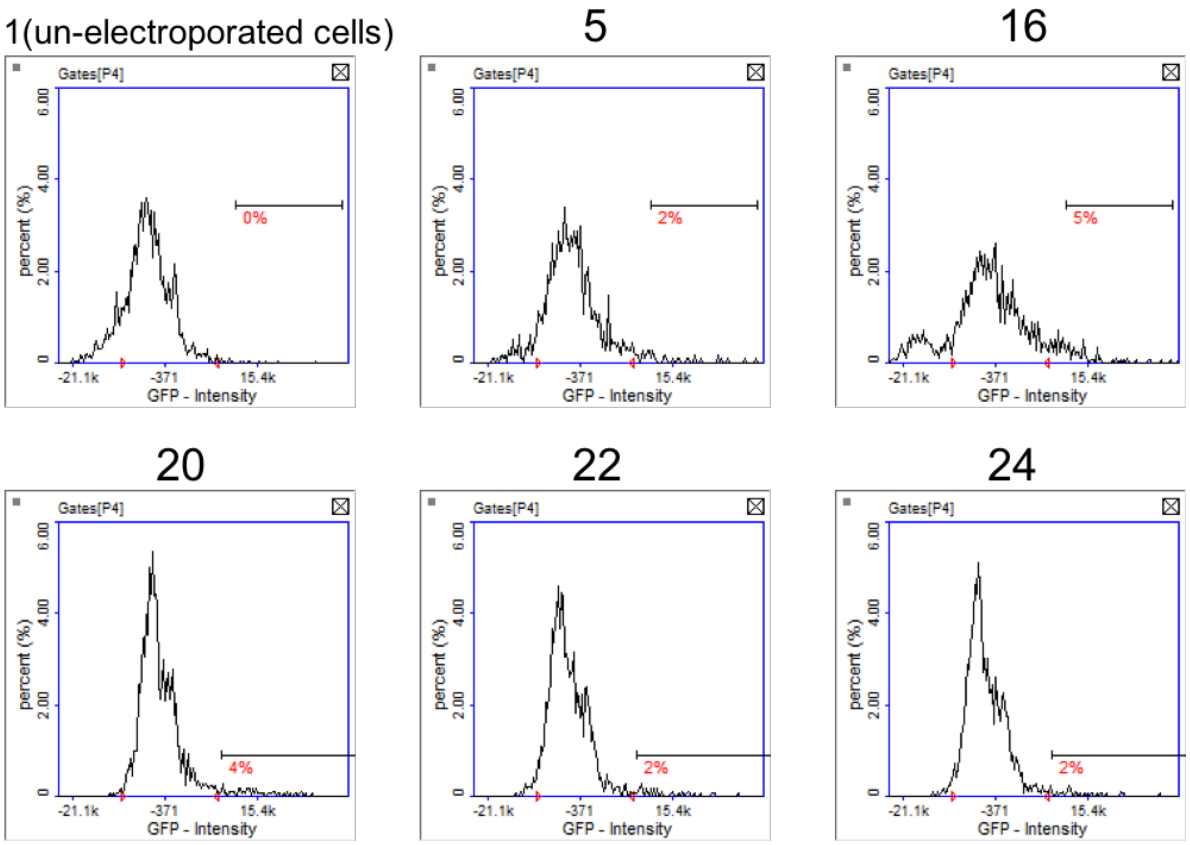


Figure 5.10 Electroporation optimization.
AG876 cells were transfected with the donor BAC using 24 different conditions (see material and methods) to get the best electroporation protocol based on GFP expression. Cells were harvested 72 h post transfection and stained with Hoechst and PI and analysed using the NucleoCounter NC 3000. The histograms show the best GFP intensity of the cell population obtained from the optimisation. Protocol numbers are labelled above each histogram. Protocol 1 is a mock and the conditions of all protocols are listed in table 2.6.

Three independent transfections were performed in parallel on AG876 cells using pBACR6K-AG876-sg#, pUC57-sgRNA-TR# and pKMD106e (where # is 1, 2, or 3). 4HT was added to the transfected cells 36 h post transfection, and hygromycin (200 ug/ml) was added 3 days post-transfection. After five days, the media were replaced with fresh media supplemented with hygromycin. Cells were grown under selection for 15 days. During this time, only cells transfected with the CRISPR component targeting TR3 kept growing whereas the TR1 and TR2-targeted samples died; so, the experiment proceeded with TR3-transfected cells. dsRed-expressing cells were sorted by flow cytometry (using the NIHR flow cytometry service) and maintained in media with hygromycin for a further six days, after which episomal DNA was isolated by the low molecular weight DNA extraction protocol (section 2.12.4). Episomal DNA was transformed into electrocompetent *E. coli* (DH10B) (section 2.12.5) to get chloramphenicol resistant bacterial colonies, which were miniprepped and analysed by restriction digestion. However, none of tested colonies had EBV-BAC. Some colonies showed extra bands that could be due to incomplete digestion, while the others seemed to be derived from BAC-plasmid only and did not contain any EBV genome (figure 5.11). It was surprising that the BAC plasmid survived for a month in dividing cells, as it contains no eukaryotic/episomal replication elements to support its maintenance.

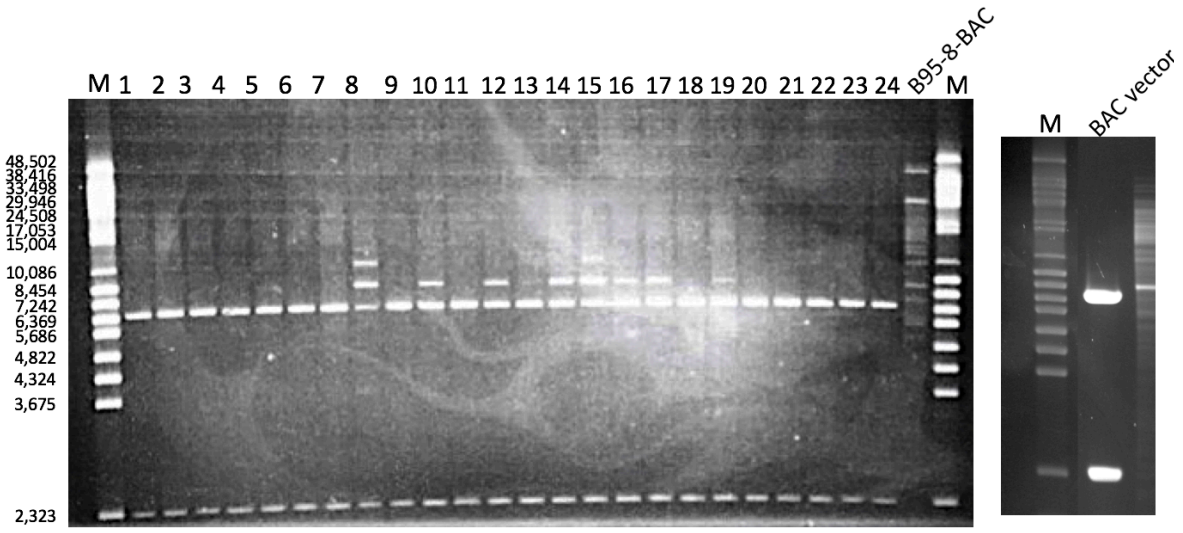


Figure 5.11 Pulsed-field analysis of miniprep DNA obtained from transfected AG876 cells.
Transfected cells were sorted and kept growing in a selective media for a month. The EcoRI digestion shows that all tested colonies were derived from the donor BAC, as it is shown compared to the positive control B95-8 BAC and negative control BAC-vector.

5.2.3 Targeting a unique region to check the feasibility of cloning selected EBV strains

To address whether the previous strategy did not manage to insert the BAC vector into EBV genome because of the targeting of a repetitive region (TR) or because the homologous recombination was inefficient, AG876 and other two EBV type 2 cell lines (Jijoye and P3HR1) were used to insert a BAC cassette into a unique region in the EBV genome.

During my work, a very similar cloning strategy (applied to epithelial cells) was published (Kanda, Furuse & Oshitani 2016). Therefore, this system was used to optimise the simpler task of incorporating a BAC cassette into a unique region in the virus genome, aiming to then use these optimised conditions to target the TR. Kanda et al. successfully cloned two EBV strains from gastric cancer cell lines (SNU19 and YCCEL1) using CRISPR components,

targeting a sequence within the BVRF1 open reading frame (encodes virion protein BVRF1). The sequences of the homology arms of the BVRF1-SNU19 construct and Cas9 target site were identical to those of AG876, Jijoye and P3HR1. Therefore, I used the donor BAC plasmid (pBS-SNU719-BVRF1(+)-PacI) and CRISPR plasmid (pX330-sgEBV) that were constructed and kindly provided by Dr Teru Kanda (figure 5.12).

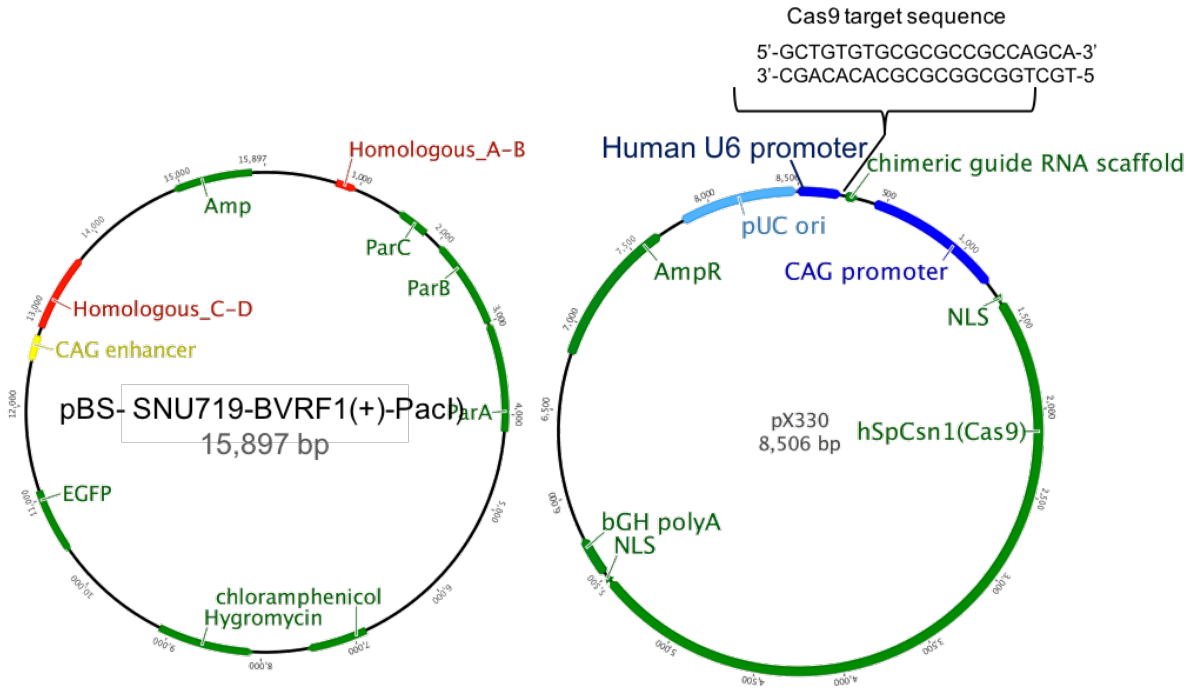


Figure 5.12 Structure of donor plasmid pBS_SNU719_BVRF1 and Cas9-sgRNA plasmid pX330.
To incorporating a BAC into a unique region in studied EBV strains (AG876, Jijoye and P3HR1), the donor plasmid pBS- SNU719-BVRF1(+)-PacI) and CRISPR plasmid (pX330-sgEBV) were used to insert a BAC vector within the BVRF1 ORF region. The plasmids were constructed and kindly provided by Dr Teru Kanda.

AG876, Jijoye and P3HR1 cells were transfected using the best two conditions (Protocols 16 and 20, see table 2.6) and were kept growing in media with hygromycin. Cells were harvested 3 weeks post-transfection and rescued episomes (section 2.12.4) were

transformed into DH10B (section 2.12.5). Only P3HR1 gave chloramphenicol-resistant colonies: 3 colonies from cells transfected using protocol 16 and 1 colony from cells transfected using protocol 20. Only the one colony obtained by using protocol 20 showed an intact EBV genome BAC (named P3HR1-BAC), whereas the other three colonies had defective EBV-BAC (figure 5.13).

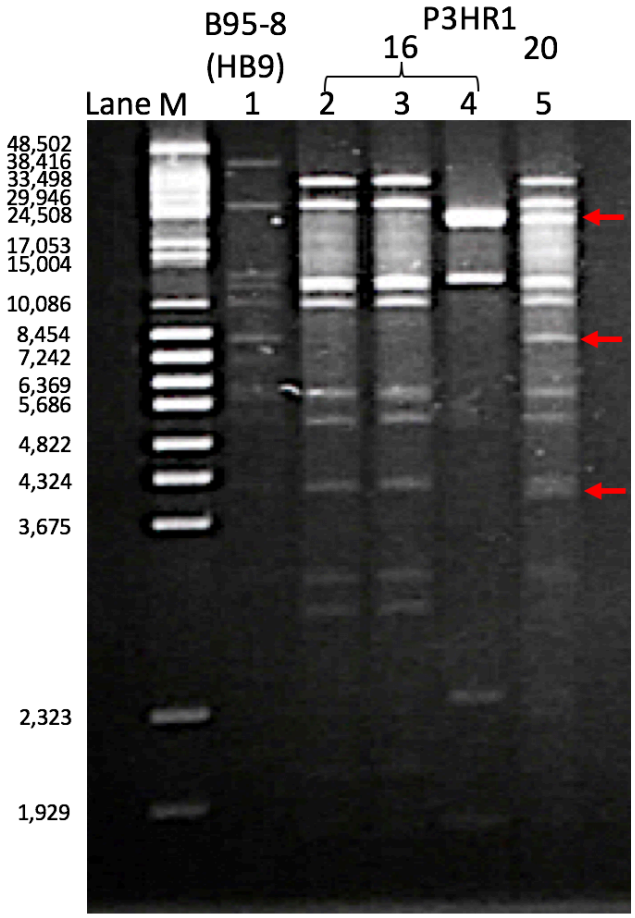


Figure 5.13 Pulsed-field analysis of miniprep DNA obtained from cloning P3HR1 as BAC.
Only four colonies were obtained from transforming DH10B with rescued DNAs obtained from transfected P3HR1 cells using two different electroporation conditions (protocol 16 and 20). EcoRI digest shows that only one colony has an intact EBV-BAC (lane 5), while lane 2 and 3 has missing bands indicted with red arrows.

Since the P3HR1 genome sequence did not resolved the size or sequence of many of the viruses' repeat regions, it was not possible to predict the size of every band generated by one restriction enzyme. Thus, P3HR1-BAC was digested with three different restriction enzymes to verify its integrity as shown in (figure 5.14).

It is worth noting that using P3HR1 genome sequence published by (Palser et al. 2015) (GeneBank accession no. LN827548) to predict band sizes on cloning software does not give the correct bands as the IR1 region is filled with unidentified nucleotides "N" that cover around nine repeat units. This number contradicts the estimated numbers of IR1 units (see section 3.13), and does not match with pulsed field electrophoresis results. The estimated number of IR1 units in P3HR1 based on read depth is 5.2. However, it seems that this number is slightly smaller than what P3HR1 actually has in pulsed field electrophoresis results. For instance, in (figure 5.14), Band "A" in EcoRI digestion is around 38416 bp, and was generated by cutting at 7321 (according to NC_007605), 4687 nt downstream unanalysed IR1 region in P3HR1 and at 56832 (according to NC_007605), 16358nt upstream unanalysed IR1 region in P3HR1. Thus, the rest of the sequence is around 17371 bp, which equals around 5.6 repeats.

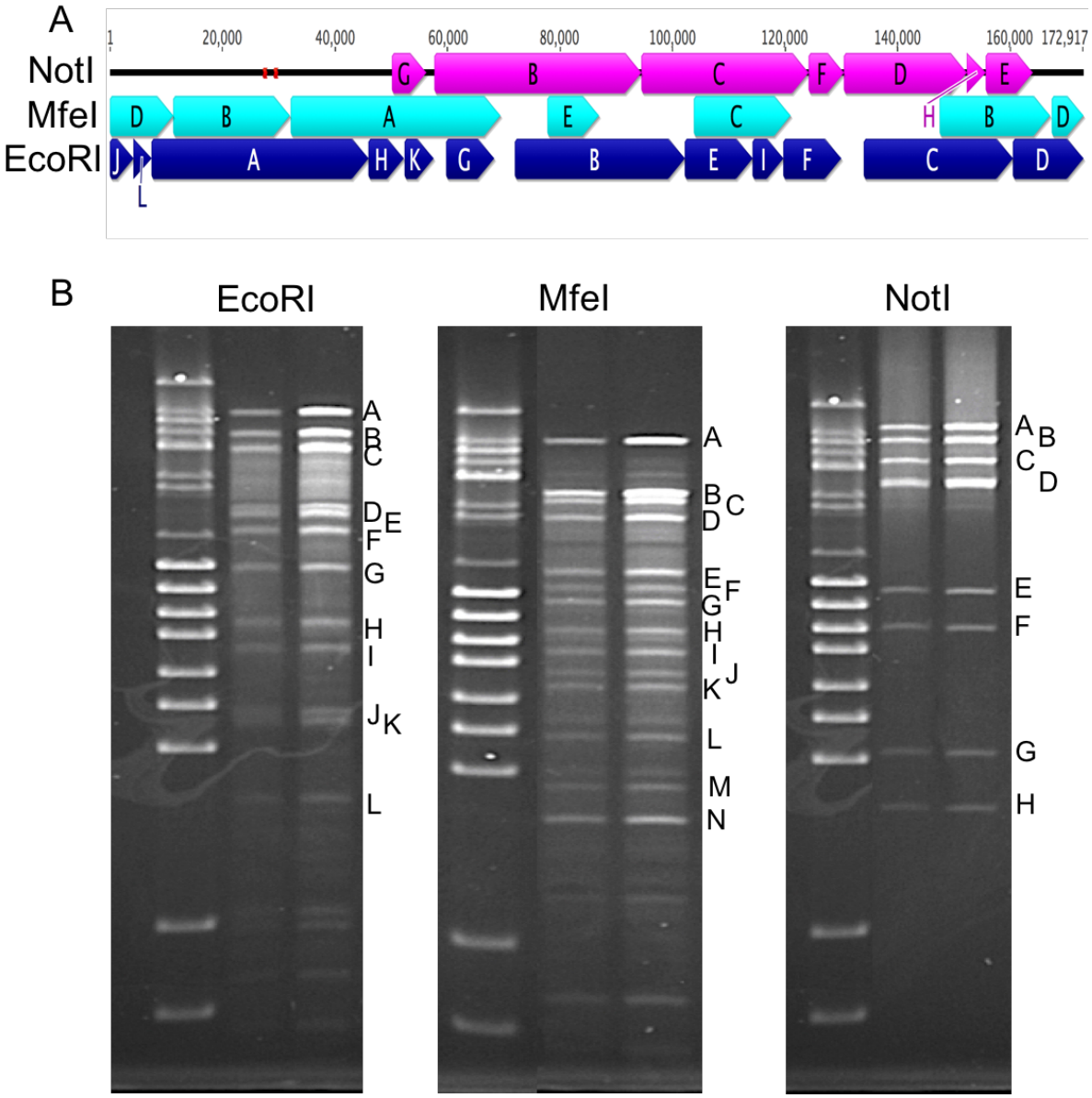


Figure 5.14 Pulsed filed analysis of P3HR1-BAC.
(A) Linear EcoRI, MfeI and NotI digestion maps of the P3HR1 (with the BAC cassette) showing the positions of produced bands on the Pulsed filed gel. The sequence of P3HR1 genome was obtained from Gene Bank (accession no. LN827548), and the two red dots show the region of the extra sequence within the IR1 region that was deleted (see text). (B) The clone with intact P3HR1-BAC construct was digested with three different restriction enzymes to verify the integrity of the construct. Both lanes in each gel are for the same sample with a different amount of loaded DNA to get a better resolution. Letters on the right side of each gel indicate the positions of bands on the P3HR1-BAC map in figure A.

5.2.4 Targeting the EBV TR region using a donor BAC plasmid with long homology sequences and an unconditional Cas9/sgRNA plasmid

The strategy published by Kanda et al. (2016) for constructing BAC and CRISPR plasmids was applied with modifications to construct plasmids for inserting BAC cassette within TR. The AG876, Jijoye and P3HR1 cell lines were used for cloning. The main differences between Kanda's strategy and my first cloning attempt (section 5.2.3.) are summarized in (table 5.1).

Table 5. 1 Comparison between Kanda lab's strategy and my strategy.

Feature	Kanda lab's strategy	My strategy
Cell line	Epithelial cells	B cells
Target site	BVRF1 ORF	Terminal Repeat
BAC vector	Hygromycin + EGFP	Hygromycin + dsRED
Homologous recombination	230 bp upstream and 851 bp downstream	50 bp upstream and down stream
Cas9 and sgRNA source	One plasmid (px330)	Two independent plasmids intein-Cas9 and pUC57-sgRNA
Transfection method	ViaFectTM Transfection Reagent	Neon® Transfection System (Electroporation Pipette Tip)

5.2.4.1 Constructing donor BACs

The lengths of the upstream and downstream homologous sequences used in Kanda's donor BAC were 230 bp and 851 bp respectively, whereas in my first attempt was 50-bp on both sides. To use the longest possible sequences of the TR on either side of the CRISPR cleavage site, I tried to amplify the whole TR region from AG876 and Jijoye using primers that

bind to the flanking unique sequences, but all attempts using various PCR conditions, polymerases and additives (e.g. glycerol, DMSO and betaine) were unsuccessful. Therefore, primers that bind to conserved regions within TR (TR_unit_fwd and TR_unit_rev) were designed based on TR sequence analysis and used for touchdown PCR (see section 2.9). The PCR product covers around 92% (492 bp) of one TR unit and includes target sites TR1 and TR3. Since the primers bind to the conserved sequences in TR, they can also be used to sequence TR in all analysed EBV strains (table 4.1). The PCR worked with type 2 AG876 as well as B95-8 and verified the sequences in the B95-8 and AG876 cell lines.

The TR unit was amplified using the TR_Unit_BamHINotI_fwd and TR_Unit_EcoRISacI_rev primers (table 2.4). The PCR fragment was cloned into the SacI and BamHI sites in p4908 (a pBluescript-based plasmid) to obtain pTR-4982. PCR was used to generated a unique PacI site in p-4908-TR at the Cas9 cleavage site at either the TR1 or the TR3 target independently, using inverse primers carry PacI site (TR-sg#-PacI_fwd and TR-sg#-PacI_rev) (table 2.4). The PCR products were circularised at PacI sites to obtain the TR1-4982-PacI and TR3-4982-PacI plasmids. Miniprepped DNA plasmids (TR1-4982-PacI and TR3-4982-PacI) were sub-cloned into the PacI site of the BAC vector used by Kanda et al. 2016 (which carries a CMV-promoter driven GFP gene, a hygromycin phosphotransferase gene driven by SV40 early promoter-enhancer and a chloramphenicol resistance gene (Kanda et al. 2011)) to make donor plasmids pBAC-4908-TR1 and pBAC-4908-TR3 for targeting the EBV genome at the TR1 or TR3 site, respectively (figure 5.15).

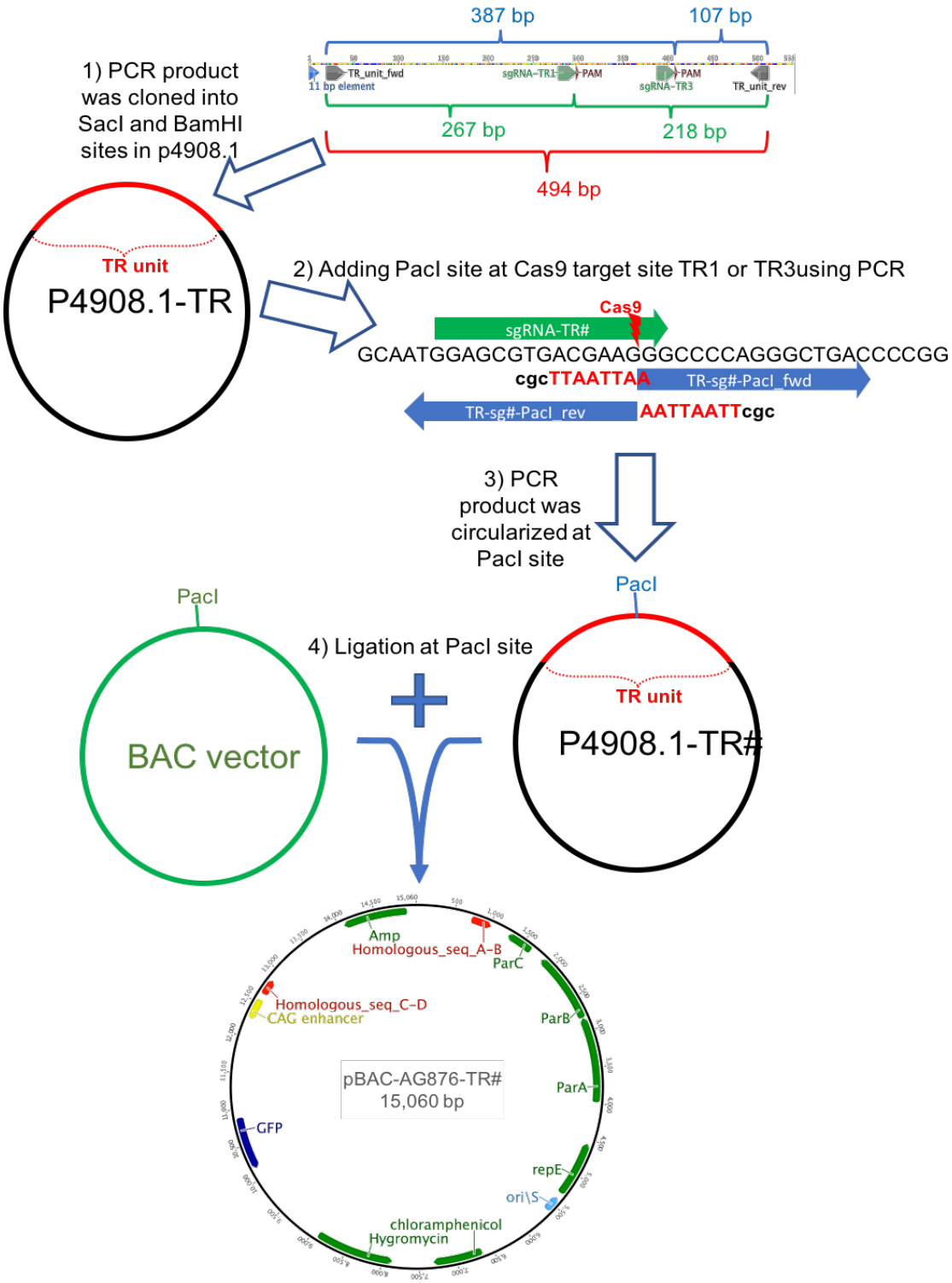


Figure 5.15 Donor plasmid BAC-AG876-TR# construction strategy.
The sequence of the AG876 TR unit (494 bp) covering target site TR1 and TR3 was amplified and cloned into Bluescript based plasmid p4908.1. A PacI recognition site was created at the Cas9 cutting site in either TR1 or TR3 target site using PCR. The PCR product was circularised at PacI site to get the high copy number plasmid p4908.1-TR# and was cloned into the PacI site in the BAC vector to obtain pBAC-AG876-TR#, where the length of upstream and downstream arms are 267 bp and 218 bp for TR1 and 387 bp and 107 bp for TR3. # indicates 1 and 3.

5.2.4.2 Construction and validation of CRISPR components.

The Cas9/sgRNA plasmids were constructed by inserting annealed oligos (sgRNA-#_oligo_UP_Px330) targeting either TR1 or TR3 into the BbsI site of pX330 (Addgene 42230) (which contains expression cassettes of both humanized Cas9 and sgRNA) to make pX330-TR1 and pX330-TR3 (figure 5.16A). The functionality of these constructed pX330 plasmids was validated using a reporter plasmid pCAG-EGxxFP (Addgene #50716), which contains 5' and 3' EGFP fragments that share 482 bp overlap, under a CAG promoter, split by stop codons and a multiple cloning site. The Cas9 target site can be inserted in the multicloning site between EGFP fragments, and when it is digested by sgRNA-guided Cas9 endonuclease, homology-dependent repair (HR; homologous recombination or SSA, single strand annealing) can take place and reconstitute the EGFP open reading frame (figure 5.16B). The reporter plasmid pCAG-EG-TR-FP was constructed by inserting a PCR-amplified TR unit into pCAG-EGxxFP (figure 5.16B).

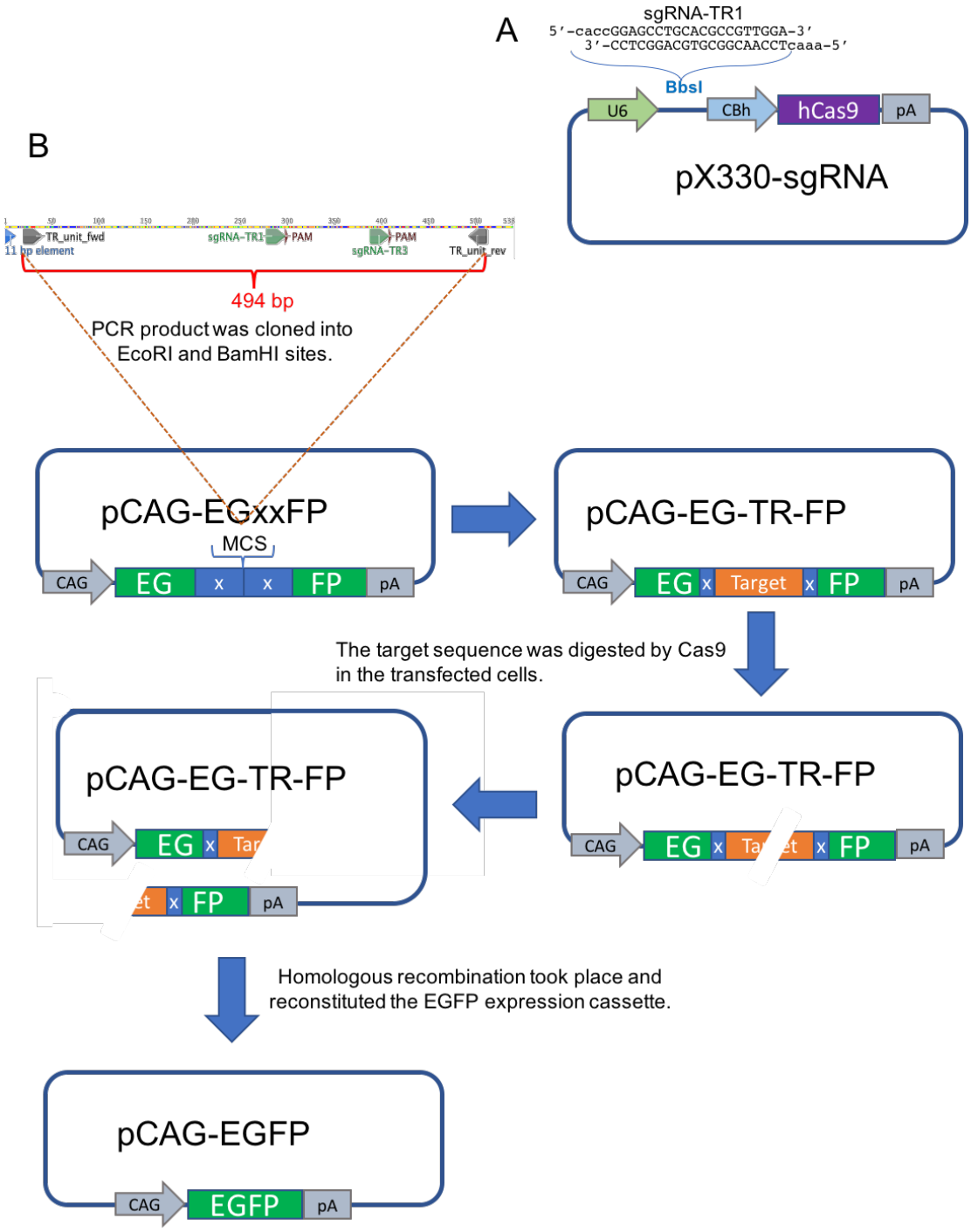


Figure 5.16 Schematic of constructing pX330 and GFP reporter plasmid pCAG-EG-TR-FP.

(**A**) Oligos of the Cas9 target sequences (TR1 as an example) were designed with complementary sequences (small letters), and placed at BbsI site in pX330. (**B**) pCAG-EGxxFP has 5' and 3' EGFP fragments that share 482 bp under the CAG promoter. The 494 bp PCR amplicon containing sgRNA target sequences (sgRNA TR1 and sgRNA-TR2) was placed between EGFP fragments to obtain pCAG-EG-TR-FP. The resultant plasmid was co-transfected with pX330 expressing Cas9 and sgRNA (targeting either TR1 or TR3) into 293T cells. When Cas9 created a DSB at the target site, the homology recombination took place and restored the GFP coding sequence. MCS; multiple cloning site.

To conduct CRISPR component validation, 293T cells were co-transfected with the reporter plasmid pCAG-EG-TR-FP and either pX330-TR1 or pX330-TR3. After 48 h incubation, cells were harvested, and GFP intensity assay showed that both pX330-TR vectors efficiently cleaved the reporter (figure 5.17). It is worth noting that GFP reporter plasmid pCAG-EGxxFP was useful with a potent Cas9 plasmid such as pX330 because the negative control (i.e. transfected with pCAG-EG-TR-FP and pX330 lacking an sgRNA) produced faint GFP-positive cells (observed by Kanda et al. 2016). As a result, using pCAG-EGxxFP to evaluate the efficiency of CRISPR components intein-Cas9(S219) and pUC57-sgRNA was tricky as intein-Cas9(S219) and pUC57-sgRNA did not produce bright GFP-positive cells making the evaluation of CRISPR efficiency unreliable.

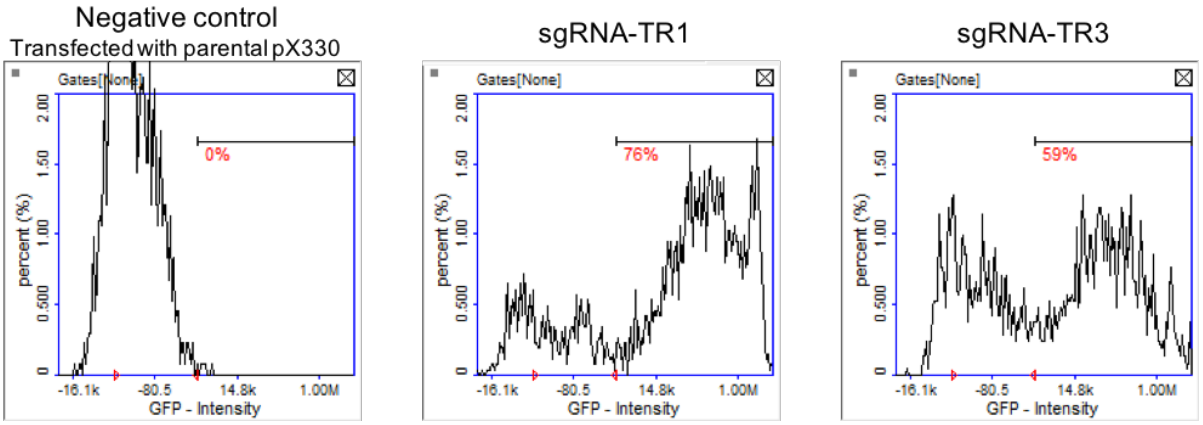


Figure 5.17 Verifying the functionality of the Cas9 plasmid.
2993T cells were co-transfected with the GFP-reporter (pCAG-EG-TR-FP) and Cas9/sgRNA (px330-TR#). After 72 h of incubation, cells were harvested, and efficiency of Cas9 in cutting intended targets was verified by measuring GFP intensity. Harvested cells were stained with Hoechst and PI and analyzed using the NucleoCounter NC 3000. The histograms show GFP intensity in samples compared to the negative control which transfected with GFP reporter pCAG-EG-TR-FP and parental pX330 (no designed sgRNA) (see section 2.12.1 and 2.12.3 for transfection and GFP assay, respectively).

5.2.4.3 EBV genome targeting and screening of bacterial colonies

AG876, Jijoye and P3HR1 cells were transfected with pBAC-4982-TR# and pX330-TR#, and grown in selective media for a month, before episomal DNA was extracted and transformed into bacteria. All screened colonies showed either the bands derived from the donor BAC or one bright band of unexpected size (figure 5.18).

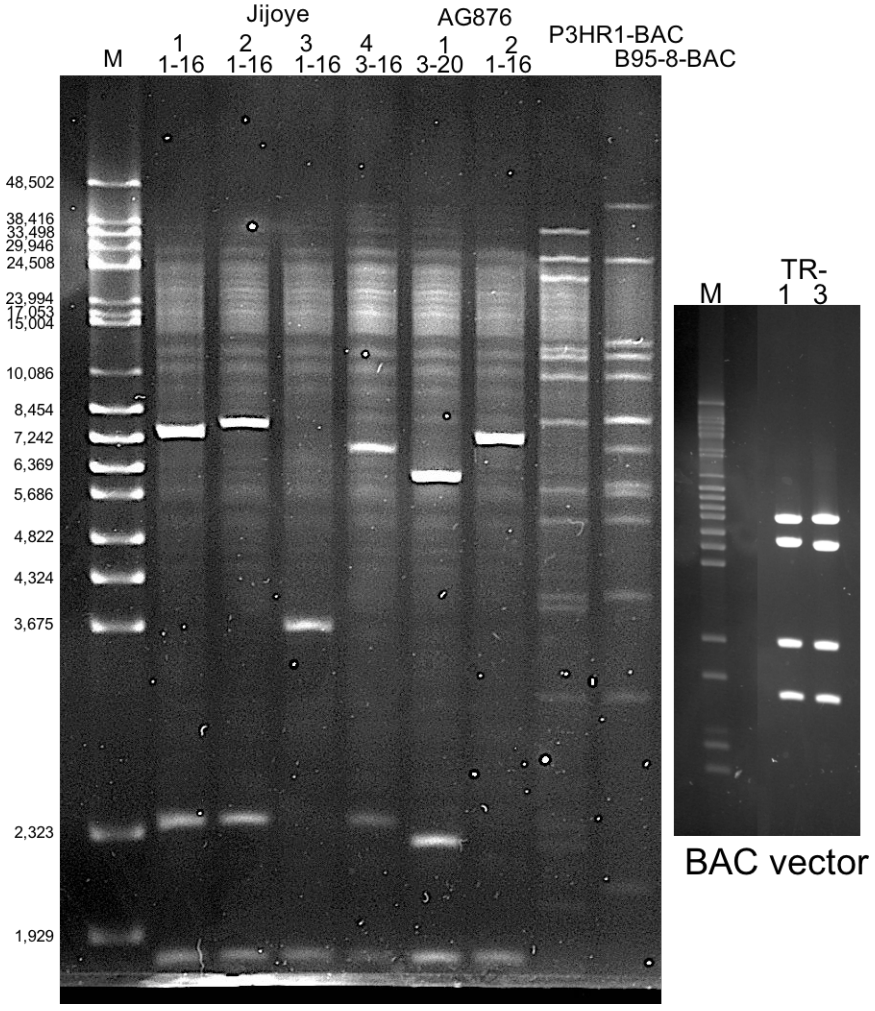


Figure 5.18 Pulsed-field analysis of miniprep DNA obtained from transfected Jijoye and AG876.
Cells were transfected with donor BAC, and CRISPR components were targeted with TR1 or TR3 using electroporation conditions (protocol 16 and 20). The target site and electroporation protocol are labelled above each lane. The EcoRI digest shows that no colony has EBV-BAC or intact donor BAC. AG876 lane 1 showed 3 bands that could be derived from the donor BAC. The band that cover high copy number plasmid (including Amp cassette) is missing in all samples while the smallest donor BAC-derived band (1801) is visible. The brighter band in all lanes (except AG876-1) are unexpected.

There was no evidence of the EBV genome compared to the positive control B95-8-BAC and P3HR1-BAC. Also, colony PCR to amplify several parts of EBV genome (e.g. IR1 region and EBNA2 region) did not show any sign of the EBV genome. By dissecting the donor BAC map (figure 5.19) and matching sizes of the observed band in (figure 5.18), the common feature of all defective rescued BACs was that all of them seemed to lose the fragment of the high copy number plasmid that harbours the ampicillin cassette (band size is around 4.7 kb) and keep the chloramphenicol/hygromycin fragment (band size is around 1.8 kb). The other two fragments that are supposed to contain GFP (around 2.2 kb) and Par A, B, and C and ori/S (around 6.1kb) were observed with slight size differences (unless they were absent). Since one EcoRI site is located upstream the homologous sequence_A_B, losing this site will result in change in the fragment with an original size of 6.1 kb (figure 5.19).

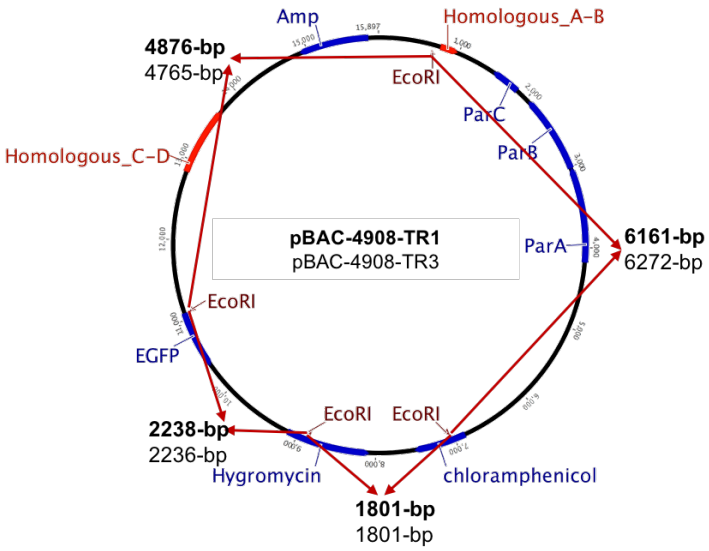


Figure 5.19 Structure of pBAC-4908-TR1 and 3.
The positions of the EcoRI site are labeled. Arrows indicate the edges of the EcoRI fragments and point at the expected size of each fragment in pBS_SNU719_BVRF1 (top) and pBS_SNU719_BVRF3 (bottom).

To verify that the ampicillin cassette was removed, miniprep DNA was digested with PacI restriction enzyme to separate high copy number fragments (including the homology sequences) from BAC. The pulsed field showed that all tested colonies (that did not grow on ampicillin plate) had lost the ampicillin cassette (figure 5.20), but none of them had the EBV genome.

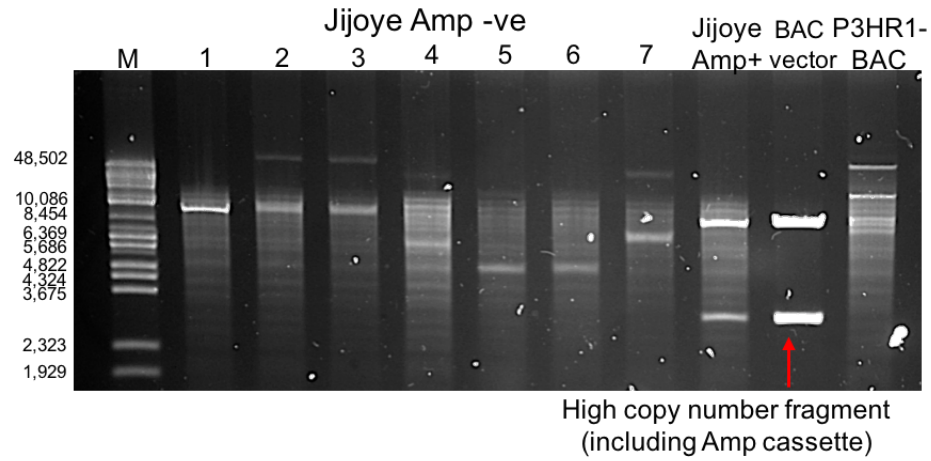


Figure 5.20 Pulsed-field analysis of recovered BACs that lacked the EBV genome. Chloramphenicol-resistant colonies were grown on a plate with chloramphenicol only and others with chloramphenicol plus Amp. Amp-negative colonies and one Amp-positive colonies were minipreped and digested with PacI to separate high copy number from BAC vector. The digest shows that all Amp-negative colonies have defective BACs, while the Amp-positive colony bands match those of the negative control (BAC vector).

The experiment described in section 2.3 that successfully generated P3HR1 BAC was repeated several times, but none of the tested colonies showed the EBV genome, and all attempts ended up with phenomena as can be seen in figure 5.21, where every colony has a bright band of different size and only one colony (lane 3) showed bands that seem to be derived from the donor BAC only. It is likely that the efficiency of cloning EBV maintained in LCLs was extremely low and that some modifications might be needed to enhance it.

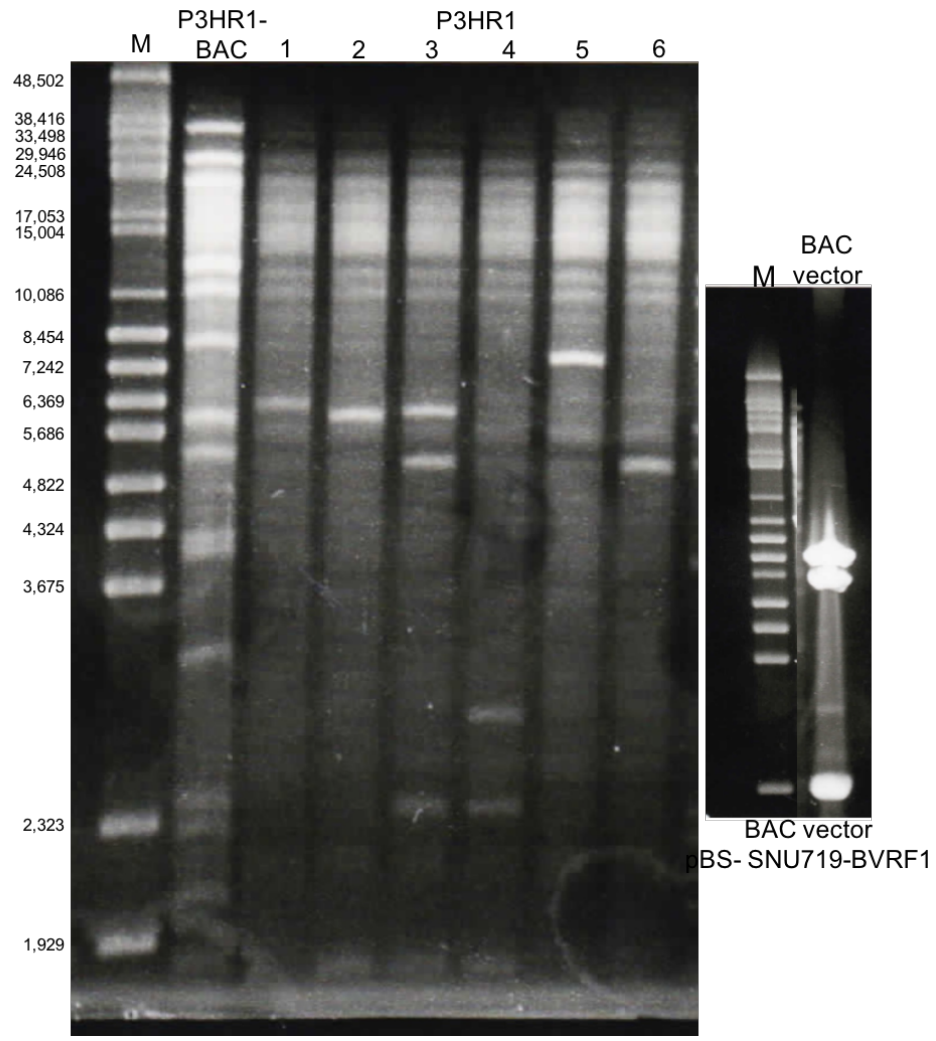


Figure 5. 21 Pulsed-field analysis of miniprep DNA obtained from transfected P3HR1 cells.

The P3HR1 cell line was transfected with the donor BAC, and CRISPR components targetinf BZLF1 using electroporation conditions (protocol 20- see table 2.6). EcoRI digest shows that no colony has EBV-BAC compared to the positive sample P3HR1-BAC. Only lane 3 seems to be derived from intact donor BAC, while others show defective BACs.

5.2.5 Modifications performed to improve cloning strategy

Since I had one success in cloning P3HR1 as a BAC but could not reproduce it, I aimed to make the cloning strategy more reliable by making some improvements. One of the challenges in cloning EBV genome as BACs is keeping transfected cells growing for a relatively long time, a month for instance. There are two reasons for maintaining cells in a selective media for a long time; first, the selection and cell division should increase the numbers of cloned EBV genomes. Thus, increasing the chance of transforming the genome into bacteria. Second, the number of intact BAC molecules decline with time, meaning that we need to screen fewer colonies for Amp resistance. This incubation time is quite long as sometimes transfected cells stop growing and die within two weeks (especially AG876). One possible reason for cells dying could be targeting TR in EBV as a recombination site, as this may interrupt LMP 2 transcription which in turn would cause cell death (Moody et al. 2003). Another possible reason is that long-term culture may select against integrated BACs (Zhou et al. 2010).

The aim was to avoid the need for long-term selection and find a way to remove (or even reduce) the excess of targeting vector (the donor BAC) without removing cloned EBV BACs. For this reason, an I-PpoI restriction site (which is ~15bp in length and does not exist in the EBV genome) was generated in the donor BACs by inserting annealed oligo for I-PpoI into the SalI site of the high copy number plasmid p4908 (for TR constructs) or pBS-SNU719 (for BVRF1 construct) using oligos PpoI-site-TR and PpoI-site-BV respectively (see oligo table 2.3). Then the resultant plasmid was sub-cloned into the same BAC vector used in section 5.2.3 (figure 5.22). The high copy number derived part (Bluescript plasmid) in the donor BAC is supposed to be removed when the expected homologous recombination occurs between the EBV genome and the donor BAC (figure 5.23). Using donor BAC–PpoI, it was possible to harvest cells 3 days post-transfection (instead of a month), and to digest isolated DNA to

eliminate non-integrated BAC. This modification was effective in saving time by reducing the proportion of ampicillin-resistant colonies.

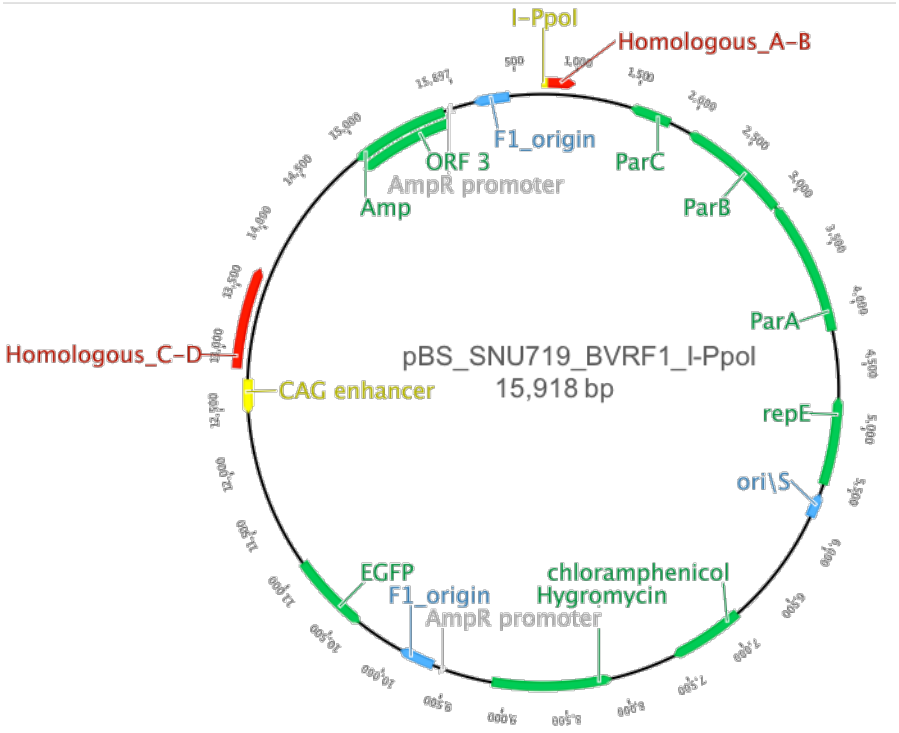


Figure 5.22 Structure of donor BAC with I-PpoI site.
An I-PpoI recognition sequence was added into the Bluescript plasmid, and the resultant plasmid was sub-cloned into the BAC vector (see the text). Adding the I-PpoI site was to reduce the excess of un-integrated vectors (the donor BAC) without removing cloned EBV BACs.

Conventionally, high throughput screening of chloramphenicol-resistant colonies can be done directly by colony PCR. However, for scanning many colonies, a PCR test would be chemicals consuming, and optimisation might become another challenge. To simplify scanning chloramphenicol-resistance colonies, colonies were grewn on two agar plates, one including chloramphenicol only while the other including chloramphenicol and ampicillin. EBV-BAC plasmid is expected to grow only on the chloramphenicol plate as it loses the ampicillin cassette

(and the high copy plasmid; R6K or Bluescript) after homologous recombination (figure 5.23). This test reduced the number of picked non-integrated chloramphenicol-resistant colonies for further analysis.

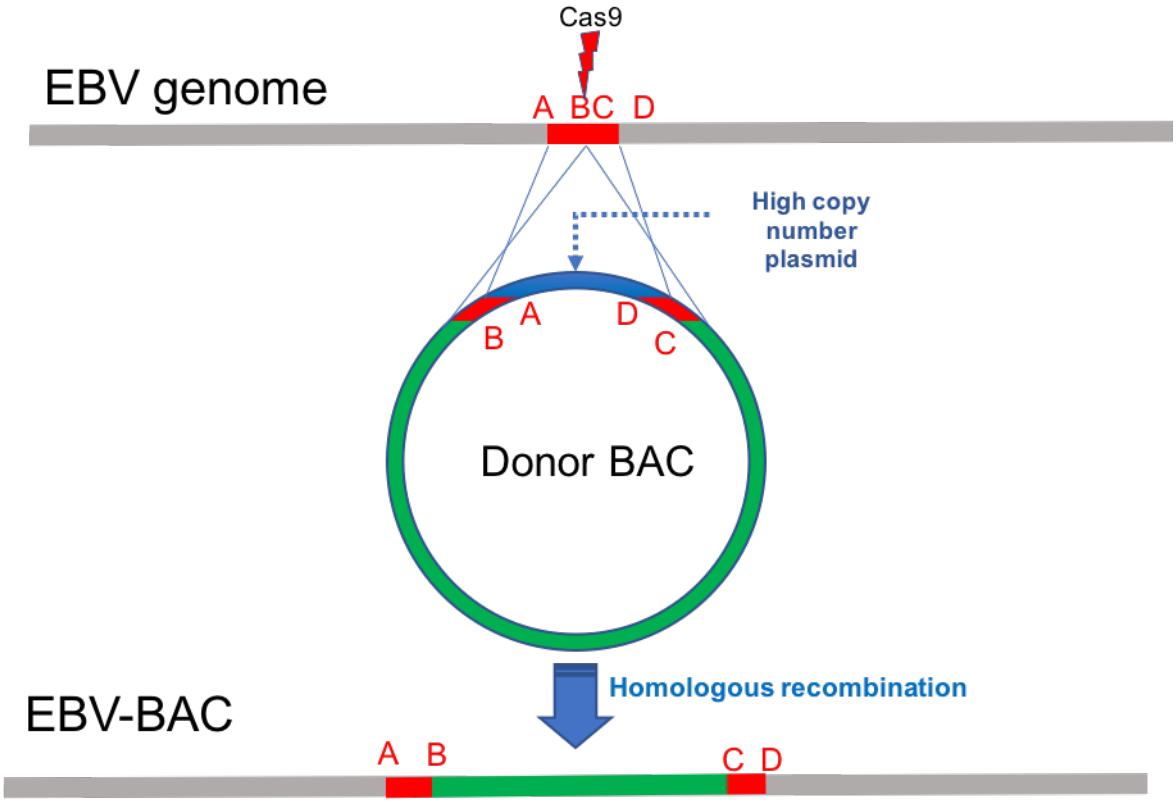


Figure 5.23 The expected structure of the EBV-BAC after homologous recombination. The homologous sequences are upstream and downstream of the Cas9 cutting site (indicated by A-B and C-D respectively). If homologous recombination between EBV genome and the donor BAC happens properly, the fragment derived from high copy number plasmid and including Amp cassette (blue part) will be excised.

5.2.6 Troubleshooting to understand reasons for obtained results showed in previous attempts to clone EBV as BACs

I repeatedly obtained the same outcome of getting defective BAC as described in section 2.4. With all candidate colonies although all steps prior to pulsed field electrophoresis showed promising preliminary results. In other words, a subset of the transfected cells proliferated and expressed GFP as expected. Also, I obtained Cm-resistant and Amp-susceptible bacterial colonies after episome rescue, with colonies obtained from DNA digested with I-PpoI being far lower than those obtained from un-digested DNA, and candidate colonies grew in chloramphenicol but not ampicillin. In pulsed field gel results (figure 5.18 and 5.21), all tested colonies (except those apparently derived from the donor BAC only) lost the fragment of the high copy number plasmid, which is supposed to be excised upon the homologous recombination between the EBV genome and the donor BAC. These results led me to speculate that any initial recombination between the EBV genome and the BAC vector could either be partial (i.e. only one side of the target site recombined with the homologous sequence in the BAC vector) or that complete recombination was followed by subsequent loss of EBV sequences from the construct.

5.2.6.1 High copy number plasmid does not affect cloning process

If homologous recombination uses only one of the homology regions to integrate with the targeting BAC, the high copy plasmid will remain in the resultant EBV BAC. Such a large high-copy construct would be unstable in bacteria, and would likely persist only by losing large amounts of sequence. Such instability could be attributed to the high replication rate of the resultant EBV-BAC construct powered by the high copy number fragment dived from the

Bluescript plasmid. To test this hypothesis, the Bluescript region in the donor BAC (pBS-SNU719-BVRF1(+)-PacI) was replaced with an ampicillin-resistant R6K plasmid modified to include an I-PpoI restriction site, to produce a new construct pR6K-donor-BAC-BVRF1_IPpoI (figure 5.24). The R6K replication origin requires the pir gene for its replication (Rakowski & Filutowicz 2013), and is inactive in the DH10B *E. coli* used for episome rescue, so its remaining in the BAC should have no impact on EBV-BAC replication in bacteria.

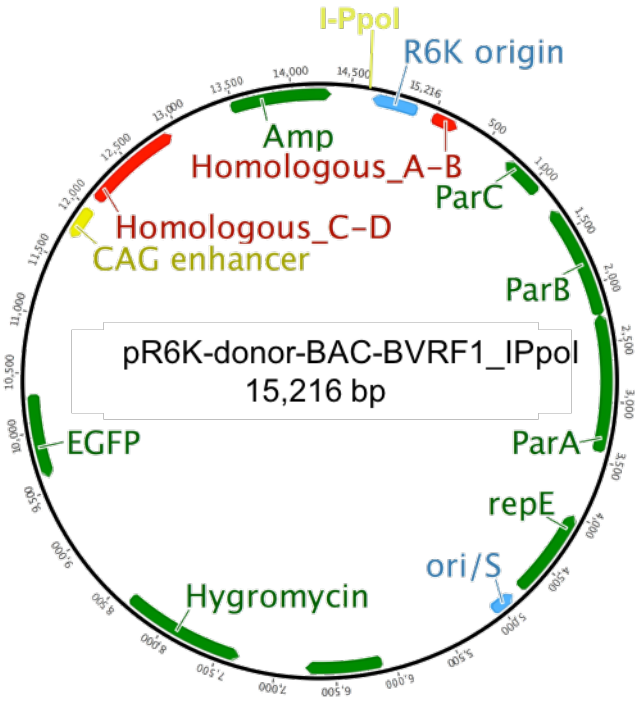


Figure 5.24 Structure of the donor BAC with I-PpoI site.
An I-PpoI recognition sequence was added into the Bluescript plasmid, and the resultant plasmid was sub-cloned into the BAC vector (see text). Adding I-PpoI site was to reduce the excess of not-integrated vectors (the donor BAC) without removing cloned EBV BACs.

AG876, Jijoye and P3HR1 cells were transfected with R6K-donor-BAC-BVRF1 and pX330-BVRF1. To see whether growth time affects the output, episomes were recovered both 3 days after transformation and after selection had established a hygromycin resistant cell line (30 days post-transfection), and the whole cloning procedure was as it was meant to be. Even

though the tested colonies resembled those shown in figure 5.18 and 5.21. Thus, it is not the high copy number fragment in the donor BAC that prevents the recovery of EBV-BAC construct into bacteria. Another possible explanation is the BAC cassette inserted into the EBV genome, but at some point, was excised.

To test whether the source of electro-competent cells influences the stability of EBV-BAC construct, episomal DNA samples were transformed into electrocompetent cells from two companies (NEB and Invitrogen). The P3HR1 episomal DNA source of P3HR1-BAC was used as a positive control. There was no difference between the two competent cells as pulse fields showed the same altered BAC vectors.

5.2.6.2 The BAC cassette may get into the EBV genome and then be excised

Given that most screened colonies had lost the band covering the high copy number plasmid (including the Amp cassette), as seen in figures (5.18, 5.20 and 5.21), it may be the BAC plasmid had lost the high copy number fragment as consequence of homologous recombination (figure 5.23).

The second place where potential EBV-BAC may lose its integrity was in the B cells. To test the stability of EBV-BAC construct in LCLs, AG876, Jijoye and P3HR1 cells were transfected with pBS- SNU719-BVRF1(+)-PacI) and CRISPR plasmid (pX330-sgEBV). PCR was conducted on episomal DNA isolated from transfected cells at 5 days and 30 days post-transfection using primers (rt-recom-for and BAC-vec-down) and (lt-recom-for and lt-recom-rev) (reported by Kanda et al. 2016) (table 2.4) to detect whether or not the homologous recombination between the EBV genome and the BAC had occurred successfully and whether recombination was still detectable at a specific point (figure 5.25A). As shown in figure 5.25B,

on day 5 all cell lines showed the expected amplicon generated from the intact junctions while on day 30 Jijoye sample seemed to have lost integrity as one band was missing and the other band was larger than the expected size. P3HR1 seemed to be less unstable, as one amplicon was missing. Only AG876 showed greater stability as both amplicons from the junctions were generated. The difference in PCR product between two time points may suggest that there was a rearrangement in BAC plasmid. This did not necessarily explain the main cause of the observed defective BACs in the pulsed field gel, but it did show that BAC construct could be unstable in at least the tested strains, and harvesting cells in short time post-transfection might be needed to get an intact construct. However, this PCR can produce false-positive result, as I was also able to generate the expected bands using a mixture of EBV DNA and vector in a tube, suggesting that this PCR is able to generate a product due to homology and is not specific for detecting cloned EBV.

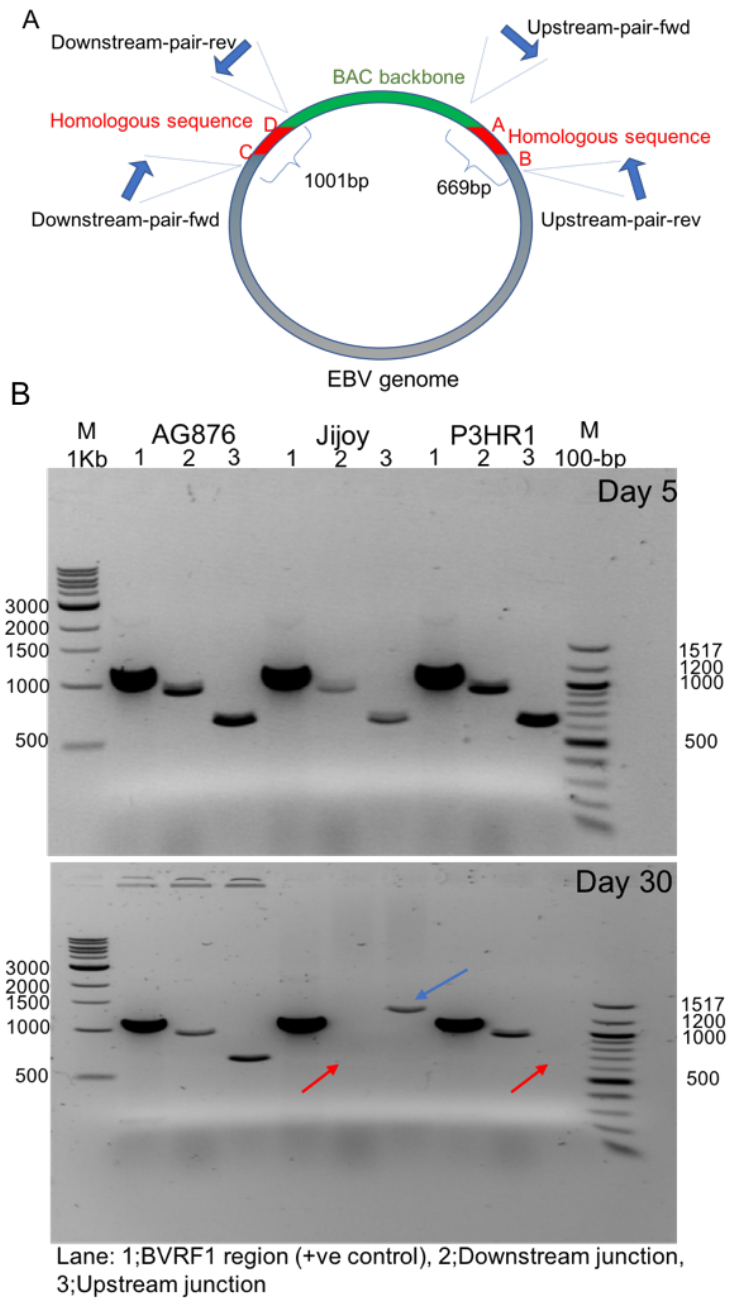


Figure 5.25 PCR analysis of homologous recombination between EBV genome and BAC cassette in two time points.

A: A diagram showing the positions of the primer binding sites and the expected amplicon sizes for verifying integrity of upstream and downstream junctions. B: Agarose gel electrophoresis of the PCR-amplified upstream and downstream junction of EBV-BAC in episomal DNA isolated from AG876, Jijoye and P3HR1 5 days and 30 days post-transfection. Lane 1 is presented as a positive control for PCR and was generated using the primers (Downstream_pair_fwd and Upstream_pair_rev) that bind to EBV genome and amplify BVRF1 region. The expected size of the positive control is 1153 bp while the expected size of the DNA band amplified from the upstream and downstream junctions are 669 bp and 1001 bp, respectively. Red arrows show the positions of missing bands while the blue arrow indicates a DNA band with unexpected size. (For PCR condition see section 2.9).

5.3 Summary

In this section, I showed attempts to find a reliable way to produce viruses from spontaneous LCLs. Neither using conditional expression of BZLF1 nor treating cells with known chemical stimuli (TPA and NaB) was able to induce the lytic cycle: sLCLs seem to be refractory to lytic induction. Then I tried to develop a strategy to clone EBV genomes as BACs using CRISPR/Cas9 to create a specific double-strand break in the genome in order to enhance the homologous recombination between the BAC vector and the EBV genome. I targeted the TR region for inserting the BAC vector, and the BVRF1 region using a strategy reported by Kanda et al. 2016. I successfully cloned P3HR1 as BAC by targeting the BVRF1 region and called it P3FR1-BAC. However, I could not do this reproducibly, or clone other strains, as all my attempts ended up with BACs carrying small unidentified inserts. I, also, showed that homologous recombination between the EBV genome and the BAC cassette could be not stable in B cells for long time.

6 Discussion

6.1 Assembling and analysing repeat regions is challenging

One of the major challenges of assembling genomes from high throughput sequence data is establishing the correct sequence for repetitive regions. Computationally, conventional assembly approaches align reads depending on user-selected penalty scores for gaps and mismatches. Alignment tools deal with multi-reads (reads that maps to multiple positions in the reference) in one of two different ways. First, alignment tools like BWA and Bowtie map multi-reads randomly (Li & Durbin 2009; Langmead et al. 2009). This multi-reads with variations are scattered among repeat copies which in turn conceals their credibility in downstream SNP calling. Other alignment tools like SOAPAligner and BFAST report all potential locations (Li et al. 2009; Homer et al. 2009). Despite the increasing number of sequenced the EBV genomes, analysis of large repeat regions in EBV genome, IR1 and TR region, have always been ignored.

In this thesis, a strategy was developed to assemble and analyse repeat regions from short-read data. The strategy relies on using the minimum number of repeat units flanked by unique sequences (see figure 3.2 as an example). This strategy may have broader applicability and I have applied it to look critically at the EBV repeats, starting with IR1 and TR by using short-read sequence data (76-bp paired-end) generated from the Illumina HiSeq 2000 by Palser et al. 2015.

During this work, I have noticed that using aligners (e.g. BWA) and contig assemblers (e.g. Velvet) to assemble genomes led to fill in genuine deletions if there are sufficient cross-contaminating reads to span the deletion. As a result, I devised read depth analysis (RDA) as a

quality control step to predict sudden changes in read depth (section 3.3). RDA identified strains with a falsely filled gap in the BamHI-Y region produced by Velvet. The strains were labelled as Namalwa and Rael, and a deeper analysis revealed that they were probably mislabelled P3HR1 strains, as the deletion and other sequence variations were identical to P3HR1. Therefore, Namalwa and Rael were excluded from Palser et al. 2015 and this thesis. Such erroneous reads likely arise through sample cross-contamination or from barcode exchange and PCR chimeras (jumping PCR) (Kircher et al. 2012). RDA, also, predicted a falsely identified deletion in HL01, which actually has an intact junction between the IR1 region and the upstream region, so can be used to help test gaps that are filled in on the end of contigs.

One of the useful features of RDA was identifying and verifying the junctions between unique sequences and IR1. Using RDA, known deletions in P3HR1, X50-7 and Daudi (Rabson et al. 1982; Jones et al. 1984; Yandava & Speck 1992) were confirmed, and a novel deletion in L591 was identified (figure 3.8). Cheptages has deletions at both ends of the IR1 region. In both cases, the deletion joined IR1 to a repetitive region, making identification of potential junctions even more challenging. RDA spotted the junctions in repeat regions; the family of repeat of *ori*-P and IR2 before and after IR1 respectively. Thus, the precise junctions were identified by looking at sequence reads at reported positions. This shows how RDA is less effective in predicting junctions within repeat regions, as any sudden change in read depth is diluted by the redundancy of reads mapping from other regions of the repeat region (figure 3.6).

Applying RDA on the YCCEL1 strain showed there was no extra truncated W repeat as reported in Kanda et al. 2016. They sequenced the YCCEL1 strain cloned into a BAC using the PacBio sequencing platform, and their sequence contained one internally truncated IR1 unit

(starts at 33781 and ends at 35244 in YCCEL1-BAC genome, accession number: AP015016). As a result of our observation form the YCCEL1 sequence data, further digestion experiments and PCR analysis were conducted on the YCCEL1-BAC, and there was no sign of the aberrant repeat unit (T. Kanda, personal communication). This observation suggests that my strategy is reliable and that analysing repeat regions, especially large ones like IR1, demands manual curation even with long sequence reads.

In the terminal repeat region, the high diversity in TRs across EBVs meant that mapping sequence reads to one reference did not generate accurate consensus sequences. Some strains had several locations with sudden changes in read depth indicating that they had insertion or deletion that was not present in the consensus sequences. Also, the sequences produced by BWA alignment were influenced by the initial B95-8 sequence, resulting in miscalling of SNPs and gaps. Therefore, de novo contig assembly was used to get the sequence for at least one TR unit flanked by the upstream and downstream unique sequences (see section 4.2). I managed to get the sequences of the TR unit for 31 of the 76 strains attempted, and this involved time-consuming manual curation of overlapping contigs.

6.2 IR1 and TR have conserved regions of known and unknown functions across EBV strains

The two largest repeat units in the EBV genome are within IR1 and TR. IR1 is a tandem array of 3-kb repeat unit. The biological importance of the IR1 region is derived from the content of the IR1 repeat units. Each unit includes: the Wp promoter, which is the first promoter to be used upon infection until transcription switches to the upstream Cp promoter (Woisetschlaeger et al. 1990), and two exons (W1 and W2), which together encode the

repetitive region of EBNA-LP (for more details see section 1.7.2). TR contains a variable number of 500-bp repeat units on each terminus of the linear genome (Raab-Traub & Flynn 1986; Zimmermann & Hammerschmidt 1995). TR in involved in circularization and packaging of EBV virion DNA (Hammerschmidt & Sugden 1989) (see section 1.7.3 for more details).

6.2.1 EBNA-LP exons

The data analysis in this thesis showed a high level of conservation in the Wp and W2 exon which supports the biological functions of these regions (figure 3.9). In W2 where X50-7 has a synonymous MV, sLCL1.12 has an MV creating R26C change. sLCL2.16 (an African type 2 sLCL strain) is the only sample that has an SNP in the W2 exon, resulting in a Q32R change (SD3.2 file at http://www.epstein-barrvirus.org.uk/MBaData.php). Such substitution was observed in the EBNA-LP sequences of chimpanzee lymphocryptoviruses (McCann et al. 2001; Ling et al. 2009). This observation suggests that Q32R substitution is genuine and might represent a naturally circulating variant, despite being seen in only one virus strain.

Diversity in the W1 exon has been described previously, and was used to distinguish two isoforms of EBNA-LP based on the presence of amino acid G8/T12 (named as the type 1 variant) or V8/A12 (named as the type 2 variant) (Finke et al. 1987). In the lab, the diversity in W1 exons can be distinguished by using the monoclonal antibody JF186, While binds only to the so-called type 1 variant (G8/T12) (Finke et al. 1987) compared the monoclonal antibody LP4D3 whose epitope is located in the W2 exon (Shaku et al. 2005). However, analyzed data shows that many type 1 strains have the so-called type 2 variants and, thus, the W1 exon variation does not solely correlate to the type1/type 2 split in EBNA-LP. I found that EBNA-LP could be better defined by Y exon variation.

The repeat domain of EBNA-LP, which is translated from IR1 transcripts, is highly conserved. The complex alternative splicing and the large number of interaction partners that have been reported for EBNA-LP may reflect the need for a high level of conservation in the W exons. However, this level of conservation is not seen in rhesus lymphocryptovirus (non-human counterpart of EBV), as the last W1-W2 pair (in the cDNA) differs from the first three copies, resulting in 14 amino acid changes out of 66 (Peng et al. 2000). On the other hand, with a large number of repeats, the consequence of individual repeat variants could be trivial (Davison 2007).

In this thesis, an analysis of the EBNA-LP exons (W1, W2, Y1 and Y2) revealed more isoforms than previously reported. Diversity in EBNA-LP is derived mainly from the Y exons as summarized in (figure 3.14). In order to clarify the complexity of the variations in the different parts of the EBNA-LP protein, a new nomenclature was proposed for EBNA-LP subtypes (table-3.3). The subtypes that have named A, B, and Z were published previously for Akata, B95-8 and AG876, respectively (McCann et al. 2001; Baer et al. 1984; Dolan et al. 2006). Interestingly, the new identified subtype C is the most common subtype in samples. Subtype B^M was found in only three strains, GD2, M81 and C666.1, all of which were isolated from Chinese nasopharyngeal carcinoma samples (out of the 5 Chinese NPC-derived strains we sequenced), suggesting that the B^M subtype might be common in the southern region of China where NPC is widespread, or even that it is associated with NPC, or perhaps the ability to grow NPC-derived cell lines. It was found with both W1 exon isoforms suggesting that B^M is more linked to the type of sample than to the type of W1 exon. Sequencing of more non-tumour associated circulating strains may address whether type B^M EBNA-LP is present in non-pathogenic strains with the same frequency as in NPC-derived strains.

The proposed nomenclature for EBNA-LP subtypes represents a simple way of describing the diversity and leaves space for additional variants to be added. Dissection of EBNA-LP exon subtypes showed a kind of restriction on the relationship between exons. For instance, the type 1 exon W1 existed only with the subtype B of exon Y2, while the subtype B of exon Y2 was seen with either type 1 or 2 exon W1. In contrast, the exon Y2 subtypes A, C and Z were found only with the type 2 exon W1. It is not clear whether these restricted associations have functional and/or structural importance for EBNA-LP, just as there are no indications of different functions for any EBNA-LP variants.

6.2.2 Stable non-coding RNAs and the BWRF1 ORF

Recently, RNAseq and computational prediction programme have detected a conserved and structurally stable non-coding RNA (sisRNA-1) generated from the short intron (81-nt) between the W1 and W2 exons (Moss & Steitz 2013). Our data showed that this short intron is highly conserved and that only two strains (sLCL1.04 and HL01) have a MV (SD3.2 at http://www.epstein-barrvirus.org.uk/MBaData.php). This conservation is consistent with the suggestion that the region between the W1 and W2 exons is evolutionary conserved and of functional importance (Moss & Steitz 2013).

The diversity in the IR1 region increases outside the exons and the most diverse regions localise within BWRF1, (a hypothetical ORF of unknown function, that has no canonical start codon). Since BWRF1 has no ATG to define a start codon, the maximum ORF length was defined by the distance between in-frame stop codons within the IR1 unit. This strategy revealed two main models (models 1 and 2) (figure 3.10). Intriguingly, the shorter BWRF1 (model 2), which consists of 384 codons, was found in the prototype EBV strains B9-58 and

only 12% of analysed strains while, the longer BWRF1 (model 1) consists of 514 codons, and was found in around 60% of strains including the saliva-isolated EBV (Appendix 3). Because BWRF1 has no ascribed function, it is not possible to predict whether there is any consequence of the variable lengths of BWRF1 models on EBV biology. Furthermore, 9 strains have disrupted BWRF1 (model 1D, and X) due to indels causing frameshifts which may suggest that BWRF1 is not essential for EBV life cycle, at least for these strains, or even BWRF1 is not required as an intact entity. Also, the fact that most of these 9 are on a specific branch evolutionary tree (figure 3.11) suggests that this disrupted BWRF variant is in common circulation, and that the intact ORF is not crucial.

The conserved sequences within BWRF1 and upstream in the hairpin region may support the notion that parts of BWRF1 could play a role in EBV life cycle. The reportedly stable EBV-sisRNA2 is generated from the BWRF1 region (Moss et al. 2014). Although the nature of EBV-sisRNA2 is unknown (sisRNA2 RNAseq reads did not seem to be evenly distributed across the repeats), it is difficult to interpret the significance of the conserved regions. Conceptually, sisRNA-2 resembles the latency-associated transcript (LAT) of HSV. LAT is a capped, and polyadenylated intron-derived persistent RNA but is processed (spliced) beyond simple excision of the intron and has a role in the maintenance of latent infection by inhibiting lytic reactivation and apoptosis (Farrell et al. 1991; Peng et al. 2008; Umbach et al. 2008). It, therefore, remains possible that some parts of the BWRF1 ORF might be used to generate translation products if sisRNA-2 is processed similarly to excise the less conserved regions. However, analysis of sisRNA-2 is still required to test these hypotheses.

The other indication that the BWRF1 region has a conserved function is the clustering of the breakpoints found in strains with deletions spanning the junctions of IR1 (figure 3.8). All six deletions in the five strains preserve the 2kb region containing the hairpin and BWRF1

intact, suggesting a selective pressure for the latent virus not to disrupt the hairpin-BWRF1 region. This preservation suggests not only that the integrity of this region is extremely important for some aspect of virus function, but also that partial copies of this region would in some way behave in a dominant negative manner, overriding the function of the intact IR1 repeat unit. An alternative explanation is that the first Wp (and perhaps the BWRF1 region downstream of the last Wp) is more important for B cell transformation and/or establishment of latency than the internal repeat units of IR1. A study in which deleted the regulatory region upstream of Wp observed that this deletion was tolerated in internal IR1 repeat units, but that mutation of the first repeat unit had a detrimental effect on B cell transformation (Yoo et al. 2002). This deletion had a negative impact on Cp usage and was not compensated for by the downstream Wp promoters, suggesting that the integrity of the BWRF1 region may be important for the proper activity of the Wp and Cp promoters. However, it remains a mystery what role this region plays and whether it is important for infection, latency or virus production.

6.2.3 Conserved regions in the terminal repeats

MSA of TR showed that the sequence of TR is highly variable between strains (figure 4.2A) and big indels are quite common (figure 4.2B). Nevertheless, conserved sequence of 35 bp (from 169902 to 169936 - NC_007605) was found but that has no functional importance assigned to it. Also, regions of known functional importance showed high conservation. For instance, Zimmermann & Hammerschmidt (1995), identified a region of 159 bp that was sufficient for packaging amplicons. This 159 bp includes the 11-elements and the EBV equivalents of packaging signals pac1 and pac2-An. However, Zimmermann & Hammerschmidt did not further analyse this region to identify the actual signal(s) required for processing and packaging. MSA of the TR unit for 31 strains (SD4.1 at http://www.epstein-

barrvirus.org.uk/MBaData.php) showed that the first 31-bp is conserved across all studied strains, but three sLCLs have an SNP at the same position, highlighting the potential importance of this 31-bp region (figure 4.2A). Also, our data verified the previously known conserved packaging signals pac1 and pac2 that are reportedly conserved at the termini of herpesviruses (Deiss et al. 1986; Zimmermann & Hammerschmidt 1995). We showed that pac1 and pac2-An (AAAAA) are conserved in all studied strains, but Mak1 and sLCL-IS2.01. Mak1 lacks pac1 and sLCL-IS2.01 is missing pac2-An due to deletions. The pac2-GC-motif (GCGGCCC) is perfectly conserved in only 26 strains out of 31. The missing pac1 and pac2 in Mak1 and sLCL-IS2.01, respectively, could suggest that the presence of pac1 and pac2 together may be important but not essential for packaging. However, Mak1 is a BL and I showed that sLCL-IS2.01 is unnatural virus with weird type 1/2 recombinant (see section 3.15). Therefore, the biological features (e.g. packaging) of these two strains need to be tested to verify whether the presence of both pac1 and pac2 is essential or not. Since one TR unit is sufficient and essential for DNA packaging (Zimmermann & Hammerschmidt 1995), it is unlikely that the whole 159-bp is required, as it crosses the 11-bp elements (covering the flank between two TR units. Thus, the conserved 31-bp mentioned above and pac1 and pac2 could be functional regions.

In the B95-8 strain, each TR unit starts with the 11 elements and to the right of the last unit there is an otherwise unique 9 elements followed by the 11-element. The 11 bp element has been suggested to be important because it is at both termini of the linear genome (Zimmermann & Hammerschmidt 1995. Also, a highly conserved cellular protein or protein complex (called terminal repeat binding protein, TRBP) has been found to bind to a sequence in TR includes the 11 elements (Sun et al. 1994). The same research group identified the cellular protein SP1 as a component of TRBP and hypothesized that SP1 drives the

recombination machinery to cleavage at recombination sites (Sun et al. 1997). However, our MSA data showed that the 11 bp elements and the 9 bp elements observed in B95-8 are not conserved in the analysed strains. However, a sequence motif (N_4-CAT-N_4), resembling the 11 bp elements in B95-8 (G_4-CAT-G_4), was found on both termini of most analysed strains (28 strains out of 31) and motif on the right terminus (GTGT-N5) resembling the 9 bp elements of B95-8 (**GTGT**TGGCG) was found in 29 of the strains, suggesting the 11 and 9 bp elements can tolerate a certain level of sequence diversity. Unlike B95-8, most of the analysed strains (25/31) have different 11-bp element motifs at each end of TR (table 4.11). This difference may either reduce the importance of the 11 bp elements or suggest that the identical sequence is not required in every unit as long the motif is present. However, it is not known the internal 11 bp elements and identifying this will help the 11 bp elements diversity to be better understood. PCR experiments would be useful to verify the 9 and 11 bp elements isoforms on the termini as there can be conducted using universal primers binding to a conserved region within the TR unit or to the unique sequences flanking TR.

The two Pax5 (Paired box protein 5) binding sites in B95-8 TR (identified by Arvey et al. 2012), were found not to be conserved. Pax5-site 1 and Pax5-site 2 were found in 12 and 15 strains out of 31 respectively. Only B95-8 had the two binding sites, whereas the others had one only biding site (Pax5-site 1 or Pax5-site 2). There is conserved sequence within a Pax5-site 1 and 2 (GCGTGAC) that exists in all strains suggesting that this sequence is core binding site. Conducting ChIP-seq and qPCR experiments on several strains along with B95-8 could test this hypothesis and estimate the effect of SNPs in the equivalent Pax5 binding sites in the strains that do not have Pax5-site 1 or Pax5–site 2.

6.3 Heterogeneity in IR1 and TR regions

After studying the diversity of the consensus sequences of the IR1 and TR regions across EBV strains, my analysis went deeper to investigate the diversity between repeat units within a given strain (i.e. heterogeneity). Nucleotides in repeat units that differ from the consensus sequence are called minor variants (MVs).

The heterogeneity in TR in B95-8 was described by sequencing the EcoRI Dhet fragment of B95-8 that contains the TR sequence (Baer et al. 1984). Chiara et al. 2016, reported variants in IR1 of 27 strains (obtained from multiple sclerosis) patients and healthy controls) after mapping reads to the prototype EBV genome reference (NC_007605). However, they presented variations between consensus IR1 sequences, and they did not go further to investigate diversity within each strain.

In this thesis, heterogeneity in repeats was assessed by developing an approach whose principle resembles those applied to detect the presence of multiple strains within an organism. For instance, genetic variations in HCMV populations were identified within individuals by detecting minor variants (Renzette et al. 2011; Gorzer et al. 2010). However, those specific studies focussed only on the presence of multiple genotypes and did not analyse diversity in depth.

The presented strategy of analysing repeat regions utilises the repetitive nature of the sequence to infer the origin(s) of genetic change by comparing MVs with wider virus diversity while using the majority consensus as an internal control. Filters based on read depth are applied (mostly in screening indels and copy number of variations), but systematic errors can still be consider as SNPs (Meacham et al. 2011). Therefore, I applied three criteria to filter MVs; frequency and read biased score were used to eliminate technical errors as many as

possible (see section 3.9.4). Furthermore, since I worked on many strains, I could predict MVs that occur at the same position in most strains to be due to an error-prone site rather than genuine MVs (see section 3.9.2). All of these features provide an internally controlled system that allowed me to distinguish MVs probably arising from technical error from genuine MVs.

In the few cases where the presence of MVs or IR1 seqeunce has been independently defined by others, my approach has held up. B95-8-associated MVs were verified by sub-cloning and Sanger sequencing BamHI fragments of IR1 (see section 3.17). Also, the presented analysis indicated that YCCEL1 has homogeneous IR1 units (no reported MVs), which was supported by the PacBio sequence of YCCE1-BAC (Kanda, Furuse & Oshitani 2016). Interestingly, I even identified an aberrant IR1 unit in the YCCEL1 sequence generated by the PacBio (accession number: AP015016) (Kanda, Furuse & Oshitani 2016) (see section 6.1). Furthermore, the B95-8-cell line was supposed to show one minor variant as a consequence of a deletion in one TR unit. Minor variant analysis showed that the B95-8-cell line has homogenous TR units and that does not have the known short TR unit (see section 4.6). All these observations support the reliability of the presented analysis output.

Mapping reads to a template with minimum repeats the contain all the junctions enables prediction of SNPs that are on edges of repeat regions but not in internal repeat units (see section 3.9.3 and 3.10). SNPs at 22 positions within 200 bp of either edge of IR1 were found on the flanks of IR1 that differ from the corresponding internal positions (Appendix 5). These SNPs were identified initially by investigating MVs with relatively high frequencies and severe read bias scores (figure 3.16). The absence of such SNPs in internal repeats could be attributed to a purification mechanism which is inefficient near the edges of the repeat and leaves only the edge position with a different sequence. Another possibility is inter-strain recombination at one flank of IR1 results in a different sequence near the flank. I could not define flank-SNPs

in the TR region, although there were MVs with high frequencies and severe read bias scores (figure 4.3). Such MVs in TR were considered as MVs since I worked on one TR unit (constructed by contig assembly), and thus it was not possible to compare a potential flank-SNP to its equivalent position in an internal unit. Since many straines have different 11 elements types on both termini (table-4.3), it is not known which type is dominant in the internal units. Therefore, to identify flank-SNPs in the TR, it would be necessary to map reads to a template with 2 TR units and all possible types of 11 bp element in the internal flank.

Generally, heterogeneity is common in both IR1 and TR. Strains derived from spontaneous LCLs tend to have more homogeneous IR1 than tumour-derived strains (figure 3.20). On the other hand, both sLCL strains and tumour strains showed a similar tendency toward having MV in TR (figure 4.7A). By looking at the type of MVs (either spontaneous mutation or recombination), viruses that contain recombination-like heterogeneity in IR1 appear to be more prevalent in tumour-derived viruses than in sLCLs (figure 3.19B and 4.7C). However, due to the limited number of studied strains (76 strains in IR1 analysis and 31 strains in TR analysis), statistical comparisons are not informative. There is a high chance that I missed false-positive and/or used false-negative MVs as there are no comprehensive references. Also, since most of the strains were propagated in cell culture, it remains possible that some of reported MVs were generated by unintentional selection pressure (Renner & Szpara 2017).

One remaining limitation of the current approach is in predicting indels (insertion/deletion polymorphisms) between repeat units within a given strain by using short read data. For instance, if a repeat unit has a deletion within poly-bases region, alignment program creates a random gap in the reads. Also, if a read maps to two regions, A and B, where the read can either map to A with a mismatch or to B with a deletion, the alignment program

will make a decision based on the penalty score which can be the wrong choice if the source DNA has a genuine deletion in location B.

6.4 Polymorphisms in the prototype strain B95-8

6.4.1 IR1 diversity

Although B95-8 is the main strain used in EBV research, detailed information about IR1 region has been unclear due to repeat sequence limitations. Previous limited attempts investigated heterogeneity within IR1 by using restriction enzymes and found no evidence of heterogeneity in the B95-8 or JHU-1 EBV strains (Jones & Griffin 1983; Hayward et al. 1980). It is clear that using restriction enzyme analysis will not reveal variations between restriction sites. Here, five minor variants were found in the parental B95-8 cell line sequenced in Hong Kong (Kwok et al. 2014) and in a recombinant virus B95-8-delEBER that was sub-cloned from B95-8 BAC (known as HB9) (Delecluse et al. 1998; Gregorovic et al. 2011) (figure 3.24). The first of these 5 MVs has been reported previously (Jones & Griffin 1983), but was not incorporated into the published B95-8 sequence (NC_007505) perhaps due to the disagreement over whether position 12647 in B95-8 was G or T (Jones & Griffin 1983; Baer et al. 1984). The MV at position 12647 was found to be about 50:50 across the studied viruses. The other 4 MVs are at the same frequency, and were cloned from B95-8 in a single BamW fragment, showing that they lie in the same repeat unit, possibly the first repeat.

Most interestingly, one of the four equal frequency MVs at 14617 is located in the last codon of the W1 exon of EBNA-LP and replaces the GAG (glutamate) with a TAG stop codon (figure 3.24). The frequency of this stop codon MV (and of the other three MVs that are in the same repeat unit) suggests that the stop codon occurs in one IR1 unit. However, it was not

possible to predict which IR1 unit has the stop codon MV by using the available short reads. Unfortunately, Sanger sequencing of the longest EBNA-LP transcript could only verify that that last W1-W2 exon pair does not contain the stop codon mutation. PCR of IR1 DNA using reverse primers binding either to the mutant W1 or to the wild-type W1 imply that the first repeat has the mutant exon W1. Using the third-generation sequencing platform (eg minION or PacBio) would prove or rebut the this result.

6.4.2 Effect of the stop codon mutation on EBNA-LP production and viral transformation ability

Since EBNA-LP is produced by alternative splicing and the exon skipping pattern has not been clearly defined, since variations in the transcribed EBNA-LP have been considered normal. I showed the consequence of the mutated exon W1 on EBNA-LP expression. EBNA-LP expressed from homogeneous IR1 units (6.6 units) in B95-8-WT^{W} seems to have three dominant sizes, while the key difference in the wild-type IR1 (represented by HB9 BAC) lacks the largest LP variants and the low abundance of protein (figure 3.27 and 3.28). Curiously, WT^{w} EBNA-LP bands were observed with sizes suggest they contain more repeat domains that the IR1 repeat itself. This suggests that another mechanism of producing EBNA-LP may exist other than simple splicing of a linear transcript. These long transcripts could arise where splicing occurs between transcripts rather than within them – called trans-splicing (Iwasaki et al. 2009). The low level of EBNA-LP protein could also suggest that the truncated EBNA-LP is not stable. It has been shown that deleting the Y exon from the B95-8 BAC (which includes a heterogeneous IR1) results in low protein levels (Szymula et al. 2018).

I also showed that the stop codon mutation reduces the efficiency with which EBV transforms B cells. The outgrowth efficiency of LCL established by infecting PBLs or purified naive B cells with B95-8 WT^{W} was higher than with HB9 (not shown). In the same context, cell proliferation as measured based on CellTrace dye dilution assay showed that WT^{w}-infected cells proliferated more than HB9-infected cells by 8 days post infection (figure 3.28 and Szymula et al 2018). However, the stop codon mutation did not measurably reduce the transformation rate (as measured by clone outgrowth in virus dilution experiments – Szymula et al 2018), but, since the HB9 B95-8 BAC has been used as a basis for many EBV gene manipulation studies, it would be better to consider using intact EBNA-LP region.

6.4.3 The short TR is missing in at least one clone of the B95-8 cell line

In the sequence of B95-8 (NC_007605), one TR unit is 15 bp shorter than the other three. This deletion was verified in B95-8-BAC by Sanger sequencing (figure 4.4B). From a computational perspective, the short TR unit creates a mismatch when B95-8 TR units are aligned to each other (figure4.4A) and consequently, the diversity analysis used in this thesis is expected to raise this mismatch as a MV. However, our analysis showed unexpectedly that the B95-8 cell line that was sequenced in Hong Kong (Kwok et al. 2014) has homogeneous TR units (no minor variants). Data from other B95-8 samples (B95-8 labelled mistakenly Sitoko, and B95-8-delEBER) verified the presence of the deletion, and the variation analysis reported a MV as expected, which support the analysis of the B95-8 cell line. To force any reads arising from the short TR copy to be obvious (if they exist) reads were mapped to a template consisting of the short TR unit flanked by intact TR units, in turn flanked by the unique genome either side of TR. This verified that there were no reads supporting the junction between the short TR

and the following one, indicating that the short TR unit is missing in B95-8 and all TR copies are intact (figure 4.5).

The absence of the truncated TR could be attributed to either the mechanism of virus replication where the episomes are amplified by generating concatenated DNA and a cleavage takes place between TR units resulting in a progeny with intact units, or the virus could compensate the deletion by homologous recombination during circularization (Sato et al. 1990; Zimmermann & Hammerschmidt 1995).

It is unlikely that the presence or absence of the short TR has any selectable consequence, as all sequences of potential importance mentioned above are upstream of the deletion. On the other hand, studying the diversity of repeat regions in this thesis has shed light on how the structure of the terminal repeats may differ between isolates. This example also shows the complexity of TR analysis and explain why mapping TR reads to one reference was not an ideal strategy (section 4.2).

6.5 Interstrain recombination in IR1

Whole genome analysis studies showed that interstrain recombination is a major component of diversity in HSV (Szpara et al. 2014) and EBV (Palser et al. 2015). Here, we showed that BL36 genome (previously described as an intertypic recombinant) has a large region (from 10024 to 43611) which most likely originated from B95-8 by interstrain recombination. Therefore, BL36 seems to be a lab artefact and it is not representative of circulating strains.

Analysing recombination-like MVs showed that TRL1 (type 1) and sLCL-IS2.01 (type 2) have MVs in the IR1 region that likely arose from recombination between type 1 and type 2 EBV strains (figure 3.22 and 3.23). However, the recombination-like MVs are located in a small region (from 150 to 1000-bp) (Appendix 6). This recombination was also reflected by the type of BWRF1 model, where the mainly type 1 strain TRL1 has BWRF1 model 2B, which was observed in only type 2 sLCLs and AG876, while sLCL-IS2.01 has model 2A, which seems to be the type 1 specific model, indicating that sLCL-IS2.01 acquired type 1 sequence. Visualising pairwise alignments indicated that both strains has the sequence of the other type in the Cp region and the first half of the IR1 unit, which could support the mechanism that rises flank-SNP in IR1 (see section 3.10). It is not possible to predict the preferred sites of recombination in the whole genome of all sequenced EBVs due to its geographic diversity. In future, creating a map of haplotype blocks generated from geographically grouped samples would be helpful to understand the nature of recombination in EBV genomes.

6.6 EBV lytic cycle induction

The ultimate aim of this work was to develop a strategy to produce virus reliably from strains that are as close as possible to circulating wild-type viruses, in order to enable study of the true biological properties of circulating EBVs rather than lab- or cancer-adapted strains. One of the major challenges in producing viruses is how to induce the lytic cycle in vitro. The stimuli that trigger switching from latent state to lytic cycle in vivo are still not clear (Murata & Tsurumi 2014). In vitro, all lytic induction stimuli lead to activation of the viral immediate early genes, BZLF1 and BRLF1, whose proteins trigger a lytic transcription cascade of early genes that lead to DNA replication and late gene expression (reviewed in Kenney & Mertz 2014). However, cell lines differ dramatically in response to different activation agents

(Countryman et al. 2009). Furthermore, only a small proportion of cell lines respond to triggers of lytic induction and produce enough virus to use in experiments. Therefore, the first question was whether sLCLs are permissive to lytic induction.

As a pilot study, two cell lines, sLCL-IM1.16 (type 1 EBV) and sLCL-IS2.01 (type 2 EBV), were used, along with the prototype EBV strain B95-8 as it is known to respond to various triggers of lytic induction. I used lentivectors expressing conditional BZLF1 (ZHT) (Johannsen et al. 2004) to induce lytic cycle under two different promoters (CMV and CAG promoters, the latter of which produced less protein). However, both attempts did not induce genome replication. Even in the positive control (B95-8), the induction with pCMV-ZHT was much higher than with pCAG-ZHT (figure 5.2 and 5.5), indicating that cell lines response differently to different stimuli.

Activating pCMV-ZHT enhanced the expression of endogenous BZLF1 (figure 5.3). These observations contrasted with the results observed in Raji cells, in which transfecting a BZLF1 expressing vector resulted in the activation of early genes but did not induce the expression of the endogenous BZLF1 (Chevallier-Greco et al. 1986). The results observed in Raji cells were attributed to the repressive chromatin structure at the Zp promoter, from which BZLF1 is expressed (McDonald et al. 2010). In case of LCL2.01 and LCL-IM1.16, expression of the endogenous BZLF1 with no viral proliferation suggest that BZLF1 alone is necessary, but it might not be sufficient, at least in these two cell lines, to activate lytic replication.

The reduced influence of pCAG-ZHT could be attributed to the low efficiency of gene expression, at least in the tested cells. The weakness in expression was also noticed in the same cell lines harbouring a plasmid expressing pCAG-driven mCherry compared to those expressing pCMV-derived mCherry. The activity of pCAG could be linked to the cell line, as

pCMV showed higher activity in eGFP expression than pCAG in Chinese hamster ovary (CHO) cells (Wang et al. 2017). In contrast, pCAG showed robust activity in mouse embryonic stem (ES) cells whereas pCMV showed marginal activity. Bornkamm et al. (2005) used pCAG to construct a plasmid (pRTS-1) for conditional expression of genes of interest. The used BL30-B95-8 cells for the experiments, but did not compare the activity of pCAG to pCMV. However, if the observed weakness of pCAG was verified (by doing WB assay to analyse expression activity of pCMV and pCAG for instance), pCAG could be useful for deriving limited expression of toxic/lethal viral gene such as LMP1 (Hammerschmidt et al. 1989).

Since inducing BZLF1 expression did not induce lytic replication in the tested cell lines, I also tried chemical inducers (TPA and NaB) with and without BZLF1 induction. BZLF1 alone in B95-8 could cause maximum induction without TPA/NaB (fig 5.2), but the chemicals produced only very modest (10 fold) (figure 5.6). By using FISH, Gradoville et al. (2002) showed that 30% of B95-8 entered the lytic cycle after treatment with TPA and NaB. In contrast, only the chemical stimuli with BZLF1 induced replication in sLCL-IM1.16 (19 fold). sLCL-IS2.01 showed high resistance levels to all stimuli. It is noteworthy that IR1 analysis showed that sLCL-IS2.01 appeared to be an unnatural virus with weird recombinant (see section 3.15), and this may be why it was even more resistant to induction than sLCL-IM1.16.

Generally, the lytic replication attempts presented in this work are in agreement with previous studies showing that few EBV strains are susceptible to lytic induction (Gradoville et al. 2002; Countryman et al. 2009). The choice of the best lytic replication stimulus or stimuli in one line cannot be extrapolated based on another cell line, and needs to be determined empirically. The differences in the level of response to lytic induction and the type of stimulus could be due to a lack of the functional target for the inducer. Another possible reason could be the presence of a cellular factor to prevent activation of the lytic cycle or a viral protein

block the entry into lytic cycle. For instance, it has been shown that LMP1 suppresses BZLF1 at the transcriptional level (Vrzalikova et al. 2011). Also, LMP2 blocks activation of the promoter for BZLF1 in LCLs, whereas in BL cell lines such as Akata, it is either not expressed or expressed at a low level (Jenkins et al. 2000). This may explain why Akata is one of the permissive cell lines for lytic induction. So far, there are no published methods for inducing lytic cycle in LCLs, and most strains used for producing virus are BLs. This may suggest that latency state determines the possibility of lytic induction, as BLs are latency I while LCLs are latency III.

6.7 Cloning EBV as BAC

Because of the difficulty of inducing the lytic cycle in sLCLs to produce viruses for studying the biological properties of EBVs, I tried to develop a strategy and tools to clone EBV genomes as BACs. For this approach, the BAC cassette is typically introduced into the genome via homologous recombination which takes place during viral DNA replication. Unlike α- and β-herpesviruses, ɣ-herpesviruses is not highly replicative in cells which reduces the chance of homologous recombination occur. Homologous recombination is also more efficient where there is a double strand break (DSB) at the site of recombination. Hence, I used CRISPR to create a DSB at the insertion site of the BAC cassette. Since the presence of the BAC vector sequence in the viral genomes could affect the virus phenotype (Zhou et al. 2010; Tsai et al. 2017), such as changing the expression of nearby genes, it is desirable to excise this extra genetic material from virus genome before any phenotypic characterization. One approach to this is to use the TR region as an insertion site for the BAC cassette. Inserting the BAC cassette within a TR will not (theoretically) disrupt coding regions (Tsai et al 2017). Upon lytic cycle induction, the BAC cassette will be excised from EBV genome as a consequence of

recombination between the two flanking terminal repeats (Feederle et al. 2010). Two EBV strains maintained in epithelial cells, M81 (Zhou et al. 2010; Tsai et al. 2013) and YCCEL1 (Tsai et al 2017), have been cloned by integrating BACs in the TR.

From the TR data analysis in section 4, only three conserved sequences were identified to be Cas9 targets (figure 5.7). Using conserved sequences would allow a universal cloning strategy to be developed. However, targeting the repeat sequence risks deleting all the repeats if the CRISPR is too efficient. However, with such challenging task, replacing missing TR sequences could be achieved if the genome was cloned successfully.

I started experiments with EBV type 2 strains; AG876, Jijoye and P3HR1. However, I could not clone any of them, despite trying different size targeting regions and different high copy vectors. To go one step back and test the visibility of cloning selected strains, I used published constructs to target the BAC to a unique sequence (the BVRF1 region) (Kanda et al. 2016). This was successful once, and I got one complete P3HR1 genome from one transformation (figure 5.13 lane 5 and 5.14). However, I was unable to repeat this cloning or replicate it in another cell line for reasons that remain unclear.

All tested colonies obtained from targeting TR with a short homology sequence (50-bp on each side of the insertion site) experiments seemed to be derived from the donor BAC only (figure 5.11). This observation rised several possibilities: 1, using B cell line for cloning EBV is not feasible; 2, integrating the BAC cassette into a repeat region complicates the recombination process; 3, since LMP2 is transcribed across the TR region (Moody et al. 2003) inserting the BAC cassette in TR may interrupt LMP2 expression which could be essential for cell survival; 4, homologous recombination was not efficient between EBV genome and the donor BAC possibly due to the short homology sequences. To test these possibilities, I used

published constructs to integrate the BAC into the BVRF1 region (Kanda et al 2016).

The differences between their strategy and ours are summarised in (table 5.1). Mainly Kanda et al. 2016 targeted EBV strains (SNU19 and YCCEL1) that are maintained in epithelial cells, which are easier to transfect than B cells. Also, they targeted a unique sequence (within the BVRF1 ORF) rather than the repeat region (i.e. TR), and the homologous sequences were long (230 bp and 851 bp). Their strategy, therefore, was applied, and only P3HR1 produced chloramphenicol resistant colonies, one containing an intact P3HR1 genome-BAC (figure 5.13 lane 5) and two other colonies contained defective EBV-BACs (figure 5.13 lane 2 and 3). The two EBV-BACs with missing bands may highlight that cloning EBV as a BAC could be susceptible to rearrangement processes. It has been reported that the family of repeats (FR) in Akata–derived EBV-BAC cannot be maintained at its original size possibly due to the secondary structure of the FR region (Kanda et al. 2011). Given that only 7 EBV BACs from 5 strains (two of which are from B cells) have been generated, it is pleasing to add P3HR1-BAC to this list and make it available for EBV research.

Although I got only one colony harboring an intact EBV-BAC, this showed that the cloning strategy was feasible in B cells, at least for P3HR1. So, I modified our TR-targeting system to use longer homology regions. Thus, the sequence of almost one TR unit was used as a homology sequence (the whole TR was too large to clone), and a unique restriction site was added to remove any background plasmid that remained after transfection. The cloning procedure was applied several times. None of the tested colonies had intact or even defective EBV-BACs. Instead, they showed a wide range of defective BAC constructs (figure 5.18). The common feature of defective rescued BACs was that they lost the high copy number plasmid that included the Amp cassette, and kept the chloramphenicol/hygromycin fragments. The other fragments included the BAC vector components (i.e. Par A, B, and C and ori/S) and GFP

were shown with slight different sizes (figure 5.19).

It was intriguing, since the donor BAC was designed to lose the high copy number fragment (including Amp) upon successful homologous recombination (figure 5.23). As a result, I speculated that the homologous recombination happened but the BAC cassette was excised with an extra DNA. However, I could not verify this hypothesis, as PCR experiments using primers bind upstream and downstream of the recombination sites (figure 5.25A) either produced no bands or bands whose sequences was derived from *E. coli* (not shown). Alternatively, because the targeting plasmid was circular, homologous recombination could have occurred at one recombination target site so the resultant EBV-BAC also contained a high copy number fragment that could make the BAC unstable in bacteria. This was excluded by using the conditional R6K origin (Rakowski & Filutowicz 2013) in place of Bluescript, but did not change the outcome. Furthermore, repeating the BAC cloning into the BVRF1 region produced the same defective BACs observed with targeting TR even with P3HR1(figure 5.21). Thus, it seemed that the target site is not the issue as targeting TR and BVRF1 ended up with the same result. It is most likely that the efficiency of cloning EBV in B cells is extremely low and I could not persuade it to happen again.

To find out where the EBV-BAC lost its integrity in culture, episomal DNA was isolated 5 and 30 days after electroporation. The PCR assay (as used by Kanda et al 2016) showed that the potential Jijoye-BAC and P3HR1-BAC lost the integrity of the construct after a long passage in culture, while AG876-BAC seemed more stable (figure 5.25). This suggests that the EBV-BAC is not maintained stably in B cells and highlights the effect of prolonged culturing during EBV cloning. Therefore, I introduced a restriction site (I-PpoI) in the BAC vector to reduce background from parental plasmid (see section 5.2.5 and figure 5.22). There was no reliable way to verify the occurrence of the homologous recombination after the electro-

transformation step as the PCR assay to verify the homologous recombination, can produce false positive results. Because the donor BAC includes homology sequence to EBV episom, PCR can produce product it from the homology in the tube from non-integrated DNA. Since epithelial cells are more reliable than B cells especially in transfection efficiency, applying the proposed cloning strategy to EBV-infected epithelial cells could clarify the visibility of targeting the TR region for cloning EBV as BAC.

6.8 Conclusion

This thesis tackles a problem that most EBV genomics researchers avoid: trying to resolve and extract sequence for the two largest repeat regions (IR1 and TR) from short-read sequencing data. Our approach has revealed a number of features that could enhance our understanding of EBV biology. We presented new observations about EBNA-LP diversity and showed conserved regions within BWRF1 the highlight a possible role for the intronic sequence from this region in EBV biology. We also defined a conserved sequence within the packaging signal region that could be essential for EBV genome circularisation. It is likely that interstrain recombination diversity raises the possibility EBV pathogenic potential. More interestingly, a novel stop codon was detected within the IR1-encoded EBNA-LP ORF of the prototype lab strain of EBV, B95.8. This mutation a effects the EBNA-LP level and transformation efficiency. Also, TR analysis revealed that the known short TR unit is missing in at least one clone of the B95-8 cell line. Determining commonalities and variabilities in IR1 and TR features (e.g. EBNA-LP variant BWRF1 model, and 11-bp element) indicated that B95-8 does not represent the majority of strains.

Since sLCLs – as a source of circulating (i.e. wild-type) virus strains - are resistant to producing infectious virus, I tried to develop a cloning strategy for cloneing EBV from these cells and producing them in an unmodified form from other cells as the best way to study the biology of natural viruses. However, I was unable to clone a BAC into the TR. I succeeded in cloning a BAC into BVRF1 ORF in one cancer-associated strain (P3HR1), but I was unable to establish this system as a routine strategy. Further attempts to optimise the cloning approach are required, such as establishing the TR-cloning system in easier epithelial cell lines,

incorporating the induction of the lytic cycle, or developing an alternative approach to capturing EBV sequences from sLCLs.

There is still a demand for an EBV model representing circulating EBVs. It is worth noting that Prof Paul Farrell and Dr Rob White have a new batch of sequences of EBV isolated from the saliva of more than 100 donors (read length: 150-bp paired-end), whose analysis will refine the understanding of EBV diversity. As future work, applying the described strategy to analyse the IR1 and TR region will enrich the study of EBV diversity. Having complete sequence of EBVs will contribute further to improve definition of wiled-type EBV.

7 References

Adler, H. et al., 2000. Cloning and Mutagenesis of the Murine Gammaherpesvirus 68 Genome as an Infectious Bacterial Artificial Chromosome. *Journal of Virology*, 74(15), pp.6964–6974. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.74.15.6964-6974.2000.

Alfieri, C., Birkenbach, M. & Kieff, E., 1991. Early events in Epstein-Barr virus infection of human B lymphocytes. *Virology*, 185(2), p.946. Available at: http://linkinghub.elsevier.com/retrieve/pii/004268229190579Z.

Ali, A.K.M. et al., 2009. Distinctive effects of the Epstein-Barr virus family of repeats on viral latent gene promoter activity and B-lymphocyte transformation. *Journal of virology*, 83(18), pp.9163–74. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2738248&tool=pmcentrez&rendertype=abstract.

Allday, M.J., Bazot, Q. & White, R.E., 2015. The EBNA3 Family: Two Oncoproteins and a Tumour Suppressor that Are Central to the Biology of EBV in B Cells. In C. Münz, ed. *Epstein Barr Virus Volume 2: One Herpes Virus: Many Diseases*. Cham: Springer International Publishing, pp. 61–117. Available at: https://doi.org/10.1007/978-3-319-22834-1_3.

Allday, M.J., Crawford, D.H. & Griffin, B.E., 1989. Epstein-Barr virus latent gene expression during the initiation of B cell immortalization. *Journal of General Virology*, 70(7), pp.1755–1764.

Ambinder, R.F. et al., 1990. Definition of the Sequence Requirements for Binding of the EBNA-1 Protein to Its Palindromic Target Sites in Epstein-Barr Virus DNA. *Journal of virology*, 64(5), pp.2369–2379.

Ambinder, R.F. et al., 1990. Definition of the sequence requirements for binding of the EBNA-1 protein to its palindromic target sites in Epstein-Barr virus DNA. *Journal of virology*, 64(5), pp.2369–2379.

Anderton, E. et al., 2008. Two Epstein-Barr virus (EBV) oncoproteins cooperate to repress expression of the proapoptotic tumour-suppressor Bim: Clues to the pathogenesis of Burkitt's lymphoma. *Oncogene*, 27(4), pp.421–433.

Baer, R., Bankier, A.T., et al., 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature*, 310(5974), pp.207–211.

Bahr, U. & Darai, G., 2001. Analysis and Characterization of the Complete Genome of Tupaia (Tree Shrew) Herpesvirus Analysis and Characterization of the Complete Genome of Tupaia (Tree Shrew) Herpesvirus. *Journal of virology*, 75(10), pp.4854–

4870.

Bankevich, A. et al., 2012. and Its Applications to Single-Cell Sequencing. *JOURNAL OF COMPUTATIONAL BIOLOGY*, 19(5), pp.455–477.

Bell, A. et al., 1998. Characterisation of Regulatory Sequences at the Epstein–Barr VirusBamHI W Promoter. *Virology*, 252(1), pp.149–161. Available at: http://www.sciencedirect.com/science/article/pii/S0042682298994405.

Bhaduri-McIntosh, S. & Miller, G., 2006. Cells lytically infected with Epstein-Barr virus are detected and separable by immunoglobulins from EBV-seropositive individuals. *Journal of Virological Methods*, 137(1), pp.103–114.

Bloss, T.A. & Sugden, B., 1994. Optimal lengths for DNAs encapsidated by Epstein-Barr virus. *J Virol*, 68(12), pp.8217–8222. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7966614.

Borestrom, C. et al., 2003. Functional interaction of nuclear factor y and sp1 is required for activation of the epstein-barr virus C promoter. *J Virol*, 77(2), pp.821–829. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12502798.

Bornkamm, G.W. et al., 2005. Stringent doxycycline-dependent control of gene activities using an episomal one-vector system. *Nucleic acids research*, 33(16), pp.1–11.

Borza, C.M. & Hutt-Fletcher, L.M., 2002. Alternate replication in B cells and epithelial cells switches tropism of Epstein-Barr virus. *Nature medicine*, 8(6), pp.594–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12042810.

Bryant, H. & Farrell, P.J., 2002. Signal Transduction and Transcription Factor Modification during Reactivation of Epstein-Barr Virus from Latency. *Journal of virology*, 76(20), pp.10290–10298.

Burkitt, D., 1958. A sarcoma involving the jaws in African children. *The British journal of surgery*, 46(197), pp.218–223.

Cao, S. et al., 2015. High-throughput RNA sequencing-based virome analysis of 50 lymphoma cell lines from the cancer cell line encyclopedia project. *Journal of virology*, 89(1), pp.713–29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25355872.

Chang, C.M. et al., 2009. The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns : A need for reappraisal. *Virus research*, 143, pp.209–221.

Chen, A. et al., 2005. Epstein-Barr virus with the latent infection nuclear antigen 3B

completely deleted is still competent for B-cell growth transformation in vitro. *Journal of virology*, 79(7), pp.4506–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1061580&tool=pmcentrez&rendertype=abstract.

Chen, H. et al., 1999. The Epstein-Barr virus latency BamHI-Q promoter is positively regulated by STATs and Zta interference with JAK/STAT activation leads to loss of BamHI-Q promoter activity. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16), pp.9339–44. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=17784&tool=pmcentrez&rendertype=abstract.

Cheung, A. & Kieff, E., 1982. Long internal direct repeat in Epstein-Barr virus DNA. *Journal of virology*, 44(1), pp.286–294.

Chevallier-Greco, A. et al., 1986. Both Epstein-Barr virus (EBV)-encoded trans-acting factors, EB1 and EB2, are required to activate transcription from an EBV early promoter. *The EMBO Journal*, 5(12), pp.3243–3249.

Chiara, M. et al., 2016. Geographic population structure in Epstein-Barr virus revealed by comparative genomics. *Genome Biology and Evolution*, 8(11), pp.3284–3291.

Chung, S. et al., 2009. NIH Public Access. , pp.41–56.

Coleman, C.B. et al., 2015. Epstein-Barr virus type 2 latently infects T cells, inducing an atypical activation characterized by expression of lymphotactic cytokines. *Journal of virology*, 89(4), pp.2301–2312.

Cong, L. et al., 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, 339(6121), pp.819–23. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3795411&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Contreras-Brodin, B. et al., 1996. B cell-specific activation of the Epstein-Barr virus-encoded C promoter compared with the wide-range activation of the W promoter. *Journal of General Virology*, 77(6), pp.1159–1162.

Countryman, J., Gradoville, L., Bhaduri-McIntosh, S., et al., 2009. Stimulus duration and response time independently influence the kinetics of lytic cycle reactivation of Epstein-Barr virus. *Journal of virology*, 83(20), pp.10694–10709.

Countryman, J., Gradoville, L., Ye, J., et al., 2009. Stimulus Duration and Response Time Independently Influence the Kinetics of Lytic Cycle Reactivation of Epstein-Barr Virus Stimulus Duration and Response Time Independently Influence the Kinetics of Lytic Cycle Reactivation of Epstein-Barr Virus □.

Crawford, D.H. et al., 2006. A Cohort Study among University Students: Identification of

Risk Factors for Epstein-Barr Virus Seroconversion and Infectious Mononucleosis. *Clinical Infectious Diseases*, 43(3), pp.276–282. Available at: http://dx.doi.org/10.1086/505400.

Davis, K.M. et al., 2015. Small molecule–triggered Cas9 protein with improved genome-editing specificity. *Nature Chemical Biology*, 11(5), pp.316–318. Available at: http://www.nature.com/doifinder/10.1038/nchembio.1793.

Davison, A.J., 2007. Chapter 02 Comparative analysis of the genomes. *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, 2, p.1408. Available at: http://www.ncbi.nlm.nih.gov/books/NBK47376/.

Deiss, L.P., Chou, J. & Frenkel, N., 1986. Functional domains within the a sequence involved in the cleavage-packaging of herpes simplex virus DNA. *Journal of virology*, 59(3), pp.605–618.

Delecluse, H.J. et al., 1998. Propagation and recovery of intact, infectious Epstein-Barr virus from prokaryotic to human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14), pp.8245–8250.

Delecluse, H.J. & Hammerschmidt, W., 2000. The genetic approach to the Epstein-Barr virus: From basic virology to gene therapy. *Journal of Clinical Pathology - Molecular Pathology*, 53(5), pp.270–279.

van Diemen, F.R. et al., 2016. CRISPR/Cas9-Mediated Genome Editing of Herpesviruses Limits Productive and Latent Infections. *PLoS Pathogens*, 12(6), pp.1–29. Available at: http://dx.doi.org/10.1371/journal.ppat.1005701.

Dirmeier, U. et al., 2003. Latent Membrane Protein 1 Is Critical for Efficient Growth Transformation of Human B Cells by Epstein-Barr Virus 1. *Cancer Research*, pp.2982–2989.

Dolan, A. et al., 2006. The genome of Epstein-Barr virus type 2 strain AG876. *Virology*, 350(1), pp.164–170.

Draborg, A.H., Duus, K. & Houen, G., 2013. Review Article Epstein-Barr Virus in Systemic Autoimmune Diseases. *Hindawi*, 2013.

Epstein, M.A. et al., 1965. MORPHOLOGICAL AND BIOLOGICAL STUDIES ON A VIRUS IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *The Journal of Experimental Medicine*, 121(5), pp.761–770. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2138004/.

Del Fabbro, C. et al., 2013. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS one*, 8(12), p.e85024. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871669&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Fåhraeus, R. et al., 1990. Epstein-Barr virus-encoded nuclear antigen 2 activates the viral latent membrane protein promoter by modulating the activity of a negative regulatory element. *Proceedings of the National Academy of Sciences of the United States of America*, 87(19), pp.7390–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=54752&tool=pmcentrez&rendertype=abstract.

Farrell, M.J., Dobson, A.T. & Feldman, L.T., 1991. Herpes simplex virus latency-associated transcript is a stable intron. *Proceedings of the National Academy of Sciences of the United States of America*, 88(3), pp.790–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=50899&tool=pmcentrez&rendertype=abstract.

Farrell, P.J., 2015. Epstein--Barr Virus Strain Variation. In C. Münz, ed. *Epstein Barr Virus Volume 1: One Herpes Virus: Many Diseases*. Cham: Springer International Publishing, pp. 45–69. Available at: https://doi.org/10.1007/978-3-319-22822-8_4.

Feederle, R., Bartlett, E.J. & Delecluse, H.-J., 2010. Epstein-Barr virus genetics: talking about the BAC generation. *Herpesviridae*, 1(1), p.6. Available at: http://herpesviridae.biomedcentral.com/articles/10.1186/2042-4280-1-6.

Finke, J. et al., 1987. Monoclonal and polyclonal antibodies against Epstein-Barr virus nuclear antigen 5 (EBNA-5) detect multiple protein species in Burkitt's lymphoma and lymphoblastoid cell lines. *Journal of virology*, 61(12), pp.3870–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=256005&tool=pmcentrez&rendertype=abstract.

Frappier, L., 2015. EBNA1. *Current topics in microbiology and immunology*, 391, pp.3–34.

Fruehling, S. et al., 1998. Tyrosine 112 of latent membrane protein 2A is essential for protein tyrosine kinase loading and regulation of Epstein-Barr virus latency. *Journal of virology*, 72(10), pp.7796–806. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=110092&tool=pmcentrez&rendertype=abstract.

Fruscalzo, A. et al., 2001. DNA sequence heterogeneity within the Epstein-Barr virus family of repeats in the latent origin of replication. *Gene*, 265(1–2), pp.165–173.

Fuentes-Pananá, E.M. & Ling, P.D., 1998. Characterization of the CBF2 binding site within the Epstein-Barr virus latency C promoter and its role in modulating EBNA2-mediated transactivation. *Journal of virology*, 72(1), pp.693–700. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=109424&tool=pmcentrez&rendertype=abstract.

Gahn, T. a & Sugden, B., 1995. An EBNA-1-dependent enhancer acts from a distance of 10 kilobase pairs to increase expression of the Epstein-Barr virus LMP gene. *Journal of virology*, 69(4), pp.2633–2636.

Glaser, S.L. et al., 1997. Epstein-Barr virus-associated Hodgkin's disease: Epidemiologic characteristics in international data. *International Journal of Cancer*, 70(4), pp.375–382. Available at: http://dx.doi.org/10.1002/(SICI)1097-0215(19970207)70:4%3C375::AID-IJC1%3E3.0.CO.

Goodwin, S., Mcpherson, J.D. & Mccombie, W.R., 2016. Coming of age : ten years of next-generation sequencing technologies. *Nature Publishing Group*, 17(6), pp.333–351. Available at: http://dx.doi.org/10.1038/nrg.2016.49.

Gorres, K.L. et al., 2014. Activation and repression of Epstein-Barr Virus and Kaposi's sarcoma-associated herpesvirus lytic cycles by short- and medium-chain fatty acids. *Journal of virology*, 88(14), pp.8028–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24807711 [Accessed August 16, 2014].

Gorzer, I. et al., 2010. Deep Sequencing Reveals Highly Complex Dynamics of Human Cytomegalovirus Genotypes in Transplant Patients over Time. *Journal of Virology*, 84(14), pp.7195–7203. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.00475-10.

Gottschalk, S. et al., 2001. An Epstein-Barr virus deletion mutant associated with fatal lymphoproliferative disease unresponsive to therapy with vires-specific CTLs. *Blood*, 97(4), pp.835–843.

Gradoville, L. et al., 2002. Protein Kinase C-Independent Activation of the Epstein-Barr Virus Lytic Cycle Protein Kinase C-Independent Activation of the Epstein-Barr Virus Lytic Cycle. *Journal of virology*, 76(11), pp.5612–5626.

Gregorovic, G. et al., 2011. Cellular Gene Expression That Correlates with EBER Expression in Epstein-Barr Virus-Infected Lymphoblastoid Cell Lines. *Journal of Virology*, 85(7), pp.3535–3545. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.02086-10.

Guo, Y. et al., 2012. The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, 13(1), p.666. Available at: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-666.

Hammerschmidt, W. & Sugden, B., 1989. Genetic analysis of immortalizing functions of Epstein-Barr virus in human B lymphocytes. *Nature*, 340(6232), pp.393–397.

Hammerschmidt, W., Sugden, B. & Baichwal, V., 1989. The transforming domain alone of the latent membrane protein of Epstein- Barr virus is toxic to cells when expressed at high levels. *Journal of virology*, 63(6), pp.2469–2475.

zur Hausen, H., Schulte-Holthausen, H. & G, K., 1970. EBV DNA in biopsies of Burkitt tumours and anaplastic carcinomas of the nasopharynx. *Nature*, 227, pp.680–685.

Hayward, S.D., Nogee, L. & Hayward, G.S., 1980. Organization of repeated regions within the Epstein-Barr virus DNA molecule. *J Virol*, 33(1), pp.507–521.

Heather, J.M. & Chain, B., 2016. The sequence of sequencers : The history of sequencing DNA. *Genomics*, 107(1), pp.1–8. Available at: http://dx.doi.org/10.1016/j.ygeno.2015.11.003.

Heller, M., van Santen, V. & Kieff, E., 1982. Simple repeat sequence in Epstein-Barr virus DNA is transcribed in latent and productive infections. *J Virol*, 44(1), pp.311–320.

Henle, G. & Henle, W., 1966. Studies on cell lines derived from Burkitt's lymphoma. *Transactions of the New York Academy of Sciences*, 29(1), pp.71–79.

Henle, W., 1968. Evidence for viruses in acute leukemia and Burkitt's tumor. *Cancer*, 21, pp.580–586.

Hertle, M.L. et al., 2009. Differential gene expression patterns of EBV infected EBNA-3A positive and negative human B lymphocytes. *PLoS Pathogens*, 5(7).

Hjalgrim, H. et al., 2007. Infectious mononucleosis, childhood social environment, and risk of Hodgkin lymphoma. *Cancer Research*, 67(5), pp.2382–2388.

Hofelmayr, H. et al., 1999. Activated mouse Notch1 transactivates Epstein-Barr virus nuclear antigen 2-regulated viral promoters. *J Virol*, 73(4), pp.2770–2780. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10074124.

Homer, N., Merriman, B. & Nelson, S.F., 2009. BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE*, 4(11).

Honess, R.W. et al., 1989. Deviations from expected frequencies of CpG dinucleotides in herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. *Journal of General Virology*, 70(4), pp.837–855.

Howe, J.G. & Shu, M. Di, 1989. Epstein-Barr virus small RNA (EBER) genes: Unique transcription units that combine RNA polymerase II and III promoter elements. *Cell*, 57(5), pp.825–834.

Hsu, J.L. & Glaser, S.L., 2000. Epstein-Barr virus-associated malignancies: Epidemiologic patterns and etiologic implications. *Critical Reviews in Oncology/Hematology*, 34(1), pp.27–53.

Hui, P., 2014. Next generation sequencing: chemistry, technology and applications. *Topics in current chemistry*, 336, pp.1–18.

Humme, S. et al., 2003. The EBV nuclear antigen 1 (EBNA1) enhances B cell immortalization several thousandfold. *Proceedings of the National Academy of Sciences of the United States of America*, 100(19), pp.10989–94. Available at: http://www.pnas.org/content/100/19/10989.full.

Hutt-Fletcher, L.M. & Turk, S.M., 2001. Virus Isolation . *Epstein-Barr Virus Protocols* , 174, pp.119–123. Available at: http://www.springerprotocols.com/Abstract/doi/10.1385/1-59259-227-9:119.

Iwakiri, D., 2014. Epstein-Barr virus-encoded RNAs: Key molecules in viral pathogenesis. *Cancers*, 6(3), pp.1615–1630.

Iwasaki, R. et al., 2009. Trans-splicing as a novel method to rapidly produce antibody fusion proteins. *Biochemical and Biophysical Research Communications*, 384(3), pp.316–321. Available at: http://dx.doi.org/10.1016/j.bbrc.2009.04.122.

Jenkins, F.J. & Hoffman, L.J., 2000. Overview of Herpesviruses. *Infectious Causes of Cancer*, (7), pp.33–50.

Jenkins, P.J., Binne, U.K. & Farrell, P.J., 2000. Histone acetylation and reactivation of Epstein-Barr virus from latency. *Journal of virology*, 74(2), pp.710–720.

de Jesus, O. et al., 2003. Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the BART (CST, BARFO) RNAs of EBV. *Journal of General Virology*, 84(6), pp.1443–1450.

Johannsen, E. et al., 1995. Epstein-Barr virus nuclear protein 2 transactivation of the latent membrane protein 1 promoter is mediated by J kappa and PU.1. *Journal of virology*, 69(1), pp.253–262.

Johannsen, E. et al., 2004. Proteins of purified Epstein-Barr virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46), pp.16286–16291.

Jones, M.. & Griffin, B.E., 1983. Clustered repeat sequences in the genome of Epstein Barr virus Nucleic. *Nucleic Acids Research*, 11(5), pp.3919–3937.

Jones, M.D. et al., 1984. The EB virus genome in Daudi Burkitt's lymphoma cells has a deletion similar to that observed in a non-transforming strain (P3HR-1) of the virus. *The EMBO journal*, 3(4), pp.813–21.

Kaiser, C. et al., 1999. The Proto-Oncogene c-myc Is a Direct Target Gene of Epstein-Barr Virus Nuclear Antigen 2. *Journal of Virology*, 73(5), pp.4481–4484. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC104340/.

Kalchschmidt, J.S. et al., 2016. Epstein-Barr virus nuclear protein EBNA3C directly induces expression of AID and somatic mutations in B cells. *The Journal of Experimental Medicine*, 213(6), pp.921–928. Available at: http://www.jem.org/lookup/doi/10.1084/jem.20160120.

Kanda, T., Furuse, Y., Oshitani, H., et al., 2016. Highly Efficient CRISPR/Cas9-Mediated

Cloning and Functional Characterization of Gastric Cancer-Derived Epstein-Barr Virus Strains. *Journal of virology*, 90(9), pp.4383–4393.

Kanda, T. et al., 2004. Production of High-Titer Epstein-Barr Virus Recombinants Derived from Akata Cells by Using a Bacterial Artificial Chromosome System Production of High-Titer Epstein-Barr Virus Recombinants Derived from Akata Cells by Using a Bacterial Artificial Chromosom. *Journal of Virology*, 78(13), pp.7004–7015.

Kanda, T., Shibata, S., Saito, S., Murata, T., Isomura, H., Yoshiyama, H., et al., 2011. Unexpected instability of family of repeats (FR), the critical cis-acting sequence required for EBV latent infection, in EBV-BAC systems. *PLoS ONE*, 6(11), pp.1–12.

Kang, M.-S. & Kieff, E., 2015. Epstein-Barr virus latent genes. *Experimental & molecular medicine*, 47(1), p.e131. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25613728.

Kelly, G., Bell, A. & Rickinson, A., 2002. Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. *Nature medicine*, 8(10), pp.1098–104. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12219084.

Kelly, G.L. et al., 2009. An epstein-barr virus anti-apoptotic protein constitutively expressed in transformed cells and implicated in burkitt lymphomagenesis: The Wp/BHRF1 link. *PLoS Pathogens*, 5(3).

Kelly, G.L. et al., 2013. Different Patterns of Epstein-Barr Virus Latency in Endemic Burkitt Lymphoma (BL) Lead to Distinct Variants within the BL-Associated Gene Expression Signature. *Journal of Virology*, 87(5), pp.2882–2894. Available at: http://jvi.asm.org/content/87/5/2882%5Cnhttp://jvi.asm.org/content/87/5/2882.full%5Cnhttp://jvi.asm.org/content/87/5/2882.full.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/23269792.

Kelly, G.L. et al., 2005. Epstein-Barr Virus Nuclear Antigen 2 (EBNA2) Gene Deletion Is Consistently Expression in Burkitt ' s Lymphoma Cells and with Increased Resistance to Apoptosis Epstein-Barr Virus Nuclear Antigen 2 (EBNA2) Gene Deletion Is Consistently Linked with EBNA3. *Journal of Virology*, 2(16), pp.10709–10717.

Kempkes, B. & Ling, P.D., 2015. EBNA2 and Its Coactivator EBNA-LP. *Current topics in microbiology and immunology*, 391, pp.35–59.

Kenney, S.C. & Mertz, J.E., 2014. Regulation of the latent-lytic switch in Epstein-Barr virus. *Seminars in cancer biology*, 26, pp.60–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24457012 [Accessed August 16, 2014].

Kieff, E.D. & Rickinson, A.B., 2007. Epstein-Barr Virus and Its Replication. *Fields Virology. 5th ed*, II(c), pp.1–90.

Kirby, H., Rickinson, A. & Bell, A., 2000. The activity of the Epstein-Barr virus BamHI W

promoter in B cells is dependent on the binding of CREB/ATF factors. *Journal of General Virology*, 81(4), pp.1057–1066.

Kircher, M., Sawyer, S. & Meyer, M., 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40(1), pp.1–8.

Koboldt, D.C. et al., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3), pp.568–76. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3290792&tool=pmcentrez&rendertype=abstract.

Kwok, H. et al., 2014. Genomic Diversity of Epstein-Barr Virus Genomes Isolated from Primary Nasopharyngeal Carcinoma Biopsy Samples. *Journal of Virology*, 88(18), pp.10662–10672. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.01665-14.

Lacoste, V. et al., 2010. Genetic diversity and molecular evolution of human and non-human primate Gammaherpesvirinae. *Infection, Genetics and Evolution*, 10(1), pp.1–13.

Langmead, B. et al., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), p.R25. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25.

Levitskaya, J. et al., 1995. Inhibition of antigen processing by the internal repeat region of the Epstein-Barr virus nuclear antigen-1. *Nature*, 375(6533), pp.685–688.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 0(0), p.3. Available at: http://arxiv.org/abs/1303.3997.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754–1760.

Li, R. et al., 2009. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15), pp.1966–1967.

Lin, Z. et al., 2013. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *Journal of virology*, 87(2), pp.1172–82. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3554088&tool=pmcentrez&rendertype=abstract [Accessed October 3, 2013].

Ling, P.D., Tan, J. & Peng, R., 2009. Nuclear-cytoplasmic shuttling is not required for the

Epstein-Barr virus EBNA-LP transcriptional coactivation function. *Journal of virology*, 83(April), pp.7109–7116.

Liu, L. et al., 2012. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012.

Longnecker, R. et al., 1993. The last seven transmembrane and carboxy-terminal cytoplasmic domains of Epstein-Barr virus latent membrane protein 2 (LMP2) are dispensable for lymphocyte infection and growth transformation in vitro. *Journal of virology*, 67(4), pp.2006–13. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=240269&tool=pmcentrez&rendertype=abstract.

Lucchesi, W. et al., 2008. Differential Gene Regulation by Epstein-Barr Virus Type 1 and Type 2 EBNA2. *Journal of Virology*, 82(15), pp.7456–7466. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.00223-08.

Ma, S.-D. et al., 2011. A New Model of Epstein-Barr Virus Infection Reveals an Important Role for Early Lytic Viral Protein Expression in the Development of Lymphomas. *Journal of Virology*, 85(1), pp.165–177. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.01512-10.

Mancao, C. & Hammerschmidt, W., 2011. Epstein-Barr virus latent membrane protein 2A is a B-cell receptor mimic and essential for B-cell survival Epstein-Barr virus Latent membrane protein 2A is a B-cell receptor mimic and essential for B-cell survival Christoph Mancao and Wolfgang Hammerschmi. *Hematology*, 110(10), pp.3715–3722.

Marquitz, A.R. & Raab-Traub, N., 2012. The role of miRNAs and EBV BARTs in NPC. *Seminars in Cancer Biology*, 22(2), pp.166–172. Available at: http://dx.doi.org/10.1016/j.semcancer.2011.12.001.

Martin, D.P. et al., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), pp.1–5. Available at: http://ve.oxfordjournals.org/cgi/doi/10.1093/ve/vev003.

Maruo, S. et al., 2006. Epstein–Barr virus nuclear protein EBNA3C is required for cell cycle progression and growth maintenance of lymphoblastoid cells. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51), pp.19500–19505.

McCann, E.M. et al., 2001. Genetic analysis of the Epstein-Barr virus-coded leader protein EBNA-LP as a co-activator of EBNA2 function. *Journal of General Virology*, 82(12), pp.3067–3079.

McDonald, C. et al., 2010. Regulation of the Epstein-Barr virus Zp promoter in B lymphocytes during reactivation from latency. *The Journal of general virology*, 91(Pt 3), pp.622–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19923262 [Accessed

September 25, 2014].

McGeoch, D.J., Rixon, F.J. & Davison, A.J., 2006. Topics in herpesvirus genomics and evolution. *Virus research*, 117(1), pp.90–104. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16490275 [Accessed October 3, 2013].

Meacham, F. et al., 2011. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12(1), p.451. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-451.

Mettenleiter, T.C., Klupp, B.G. & Granzow, H., 2009. Herpesvirus assembly: an update. *Virus research*, 143(2), pp.222–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19651457 [Accessed September 26, 2013].

Moghaddam, A., 1997. An Animal Model for Acute and Persistent Epstein-Barr Virus Infection. *Science*, 276(5321), pp.2030–2033. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.276.5321.2030.

Moody, C.A. et al., 2003. Length of Epstein-Barr virus termini as a determinant of epithelial cell clonal emergence. *Journal of virology*, 77(15), pp.8555–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165260&tool=pmcentrez&rendertype=abstract.

Moss, W.N. et al., 2014. RNA families in Epstein-Barr virus. *RNA Biology*, 11(1), pp.10–17.

Moss, W.N. & Steitz, J. a, 2013. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. *BMC genomics*, 14, p.543. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3751371&tool=pmcentrez&rendertype=abstract.

Murata, T. & Tsurumi, T., 2014. Switching of EBV cycles between latent and lytic states. , 1(December 2013), pp.142–153.

Nilsson, K., 1971. High-frequency establishment of human immunoglobulin-producing lymphoblastoid lines from normal and malignant lymphoid tissue and peripheral blood. *International Journal of Cancer*, 8(3), pp.432–442. Available at: http://dx.doi.org/10.1002/ijc.2910080311.

Niwa, H., Yamamura, K. & Miyazaki, J., 1991. Efficient selection for high-expression transfectants with a novel eukaryotic vector. *Gene*, 108(2), pp.193–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8662924.

Nonkwelo, C. et al., 1996. Transcription start sites downstream of the Epstein-Barr virus (EBV) Fp promoter in early-passage Burkitt lymphoma cells define a fourth promoter for expression of the EBV EBNA-1 protein. *Journal of virology*, 70(1), pp.623–627. Available at:

http://jvi.asm.org/content/70/1/623.short%5Cn/Users/Pouicky/Documents/ReadCube Media/623.pdf.

Palser, A.L. et al., 2015. Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection. *Journal of Virology*, 89(10), pp.5222–5237.

Paulson, E.J. & Speck, S.H., 1999. Differential methylation of Epstein-Barr virus latency promoters facilitates viral persistence in healthy seropositive individuals. *Journal of Virology*, 73(12), pp.9959–9968.

Pellett, P.E. & Roizman, B., 2013. Chapter 59 Herpesviridae. *Fields Virology*, 1, pp.1803–1822.

Peng, R. et al., 2000. Sequence and functional analysis of EBNA-LP and EBNA2 proteins from nonhuman primate lymphocryptoviruses. *Journal of virology*, 74(1), pp.379–389.

Peng, W. et al., 2008. Identification of two small RNAs within the first 1.5-kb of the herpes simplex virus type 1-encoded latency-associated transcript. *Journal of NeuroVirology*, 14(1), pp.41–52.

Raab-Traub, N. & Flynn, K., 1986. The structure of the termini of the Epstein-Barr virus as a marker of clonal cellular proliferation. *Cell*, 47(6), pp.883–889.

Rabson, M. et al., 1982. Non-immortalizing P3J-HR-1 Epstein-Barr virus: a deletion mutant of its transforming parent, Jijoye. *Journal of virology*, 44(3), pp.834–844.

Rakowski, S.A. & Filutowicz, M., 2013. Plasmid R6K replication control. *Plasmid*, 69(3), pp.231–242. Available at: http://dx.doi.org/10.1016/j.plasmid.2013.02.003.

Ran, F.A. et al., 2013. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, 8(11), pp.2281–2308. Available at: http://www.nature.com/doifinder/10.1038/nprot.2013.143.

Rechsteiner, M.P. et al., 2008. Latent membrane protein 2B regulates susceptibility to induction of lytic Epstein-Barr virus infection. *Journal of virology*, 82(4), pp.1739–1747.

Renner, D.W. & Szpara, M.L., 2017. The Impacts of Genome-wide Analyses on our Understanding of Human Herpesvirus Diversity and Evolution. *Journal of Virology*, p.JVI.00908-17. Available at: http://jvi.asm.org/lookup/doi/10.1128/JVI.00908-17.

Renzette, N. et al., 2011. Extensive Genome-Wide Variability of Human Cytomegalovirus in Congenitally Infected Infants. *PLoS Pathogens*, 7(5).

Reuter, J.A., Spacek, D. V & Snyder, M.P., 2015. High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4), pp.586–597. Available at:

http://dx.doi.org/10.1016/j.molcel.2015.05.004.

Rickinson, A.B. et al., 1984. T-cell-mediated regression of "spontaneous" and of Epstein-Barr virus-induced B-cell transformation in vitro: Studies with cyclosporin A. *Cellular Immunology*, 87(2), pp.646–658.

Rickinson, A.B. & Kieff, E.D., 2007. Epstein-Barr Virus. *Fields Virology. 5th ed*, II(155), pp.1–84.

Riddell, A., Jeffery-Smith, A. & Tong, C.Y.W., 2018. Herpesviruses. *Medicine*, 45(12), pp.767–771. Available at: http://dx.doi.org/10.1016/j.mpmed.2017.09.006.

Robasky, K., Lewis, N.E. & Church, G.M., 2014. The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1), pp.56–62. Available at: http://www.nature.com/doifinder/10.1038/nrg3655.

Rogers, R.P., Woisetschlaeger, M. & Speck, S.H., 1990. Alternative splicing dictates translational start in Epstein-Barr virus transcripts. *The EMBO journal*, 9(7), pp.2273–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=551952&tool=pmcentrez&rendertype=abstract.

Roschewski, M., 2012. EBV-associated lymphomas in adults. *Best practice & research. Clinical haematology*, 25(1), pp.75–89.

Sadler, R.H. & Raab-Traub, N., 1995. The Epstein-Barr virus 3.5-kilobase latent membrane protein 1 mRNA initiates from a TATA-Less promoter within the first terminal repeat. *Journal of virology*, 69(7), pp.4577–81. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=189209&tool=pmcentrez&rendertype=abstract.

Salamon, D. et al., 2003. High-resolution methylation analysis and in vivo protein-DNA binding at the promoter of the viral oncogene LMP2A in B cell lines carrying latent Epstein-Barr virus genomes. *Virus Genes*, 27(1), pp.57–66.

Sample, J. et al., 1990. Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. *Journal of virology*, 64(9), pp.4084–92. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=247870&tool=pmcentrez&rendertype=abstract.

Sample, J., Henson, E.B.D. & Sample, C., 1992. The Epstein-Barr Virus Nuclear Protein 1 Promoter Active in Type I Latency Is Autoregulated. *Journal of virology*, 66(8), pp.4654–4661.

Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–5467.

Sato, H. et al., 1990. Concatameric replication of Epstein-Barr virus: structure of the termini in virus-producer and newly transformed cell lines. *J Virol*, 64(11), pp.5295–5300. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2170673.

Schaefer, B.C. et al., 1997. Constitutive activation of Epstein-Barr virus (EBV) nuclear antigen 1 gene transcription by IRF1 and IRF2 during restricted EBV latency. *Molecular and cellular biology*, 17(2), pp.873–86. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=231814&tool=pmcentrez&rendertype=abstract.

Schaefer, B.C., Strominger, J.L. & Speck, S.H., 1995. Redefining the Epstein-Barr virus-encoded nuclear antigen EBNA-1 gene promoter and transcription initiation site in group I Burkitt lymphoma cell lines. *Proceedings of the National Academy of Sciences of the United States of America*, 92(23), pp.10565–10569.

Seto, E. et al., 2010. Micro RNAS of epstein-barr virus promote cell cycle progression and prevent apoptosis of primary human B cells. *PLoS Pathogens*, 6(8), pp.69–70.

Shaku, F. et al., 2005. Development of a monoclonal antibody against Epstein-Barr virus nuclear antigen leader protein (EBNA-LP) that can detect EBNA-LP expressed in P3HR1 cells. *Microbiology and Immunology*, 49(5), pp.477–483.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), pp.1135–1145.

Sinclair, a J. & Farrell, P.J., 1995. Host cell requirements for efficient infection of quiescent primary B lymphocytes by Epstein-Barr virus. *Journal of virology*, 69(9), pp.5461–5468.

Skalska, L. et al., 2010. Epigenetic repression of p16INK4A by latent epstein-barr virus requires the interaction of EBNA3A and EBNA3C with CtBP. *PLoS Pathogens*, 6(6).

Skalsky, R.L. & Cullen, B.R., 2015. EBV Noncoding RNAs. In C. Münz, ed. *Epstein Barr Virus Volume 2: One Herpes Virus: Many Diseases*. Cham: Springer International Publishing, pp. 181–217. Available at: https://doi.org/10.1007/978-3-319-22834-1_6.

Skalsky, R.L. & Cullen, B.R., 2010. Viruses, microRNAs, and Host Interactions. *Annual Review of Microbiology*, 64(1), pp.123–141. Available at: http://www.annualreviews.org/doi/10.1146/annurev.micro.112408.134243.

Skare, J. & Strominger, J.L., 1980. Cloning and mapping of BamHi endonuclease fragments of DNA from the transforming B95-8 strain of Epstein-Barr virus. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7), pp.3860–3864.

Sugano, N. et al., 1997. Epstein-Barr Virus Binding to CD21 Activates the Initial Viral

Promoter via NF-κB Induction . *The Journal of Experimental Medicine*, 186(5), pp.731–737. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2199015/.

Sun, R. et al., 1994. Autoantigenic proteins that bind recombinogenic sequences in Epstein-Barr virus and cellular DNA. *Medical Sciences Communicated by Dorothy M. Horstmann*, 91(August), pp.8646–8650.

Sun, R. et al., 1997. Sp1 binds to the precise locus of end processing within the terminal repeats of Epstein-Barr virus DNA. *J Virol*, 71(8), pp.6136–6143. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9223508%5Cnhttp://pubmedcentralcanada.ca/picrender.cgi?accid=PMC191874&blobtype=pdf.

Swaminathan, S., 1996. Characterization of Epstein – Barr Virus Recombinants with Deletions of the Bam HI C Promoter. , 541(217), pp.532–541.

Szpara, M.L. et al., 2014. Evolution and Diversity in Human Herpes Simplex Virus Genomes. *Journal of virology*.

Szymula, A. et al., 2018. Epstein-Barr virus nuclear antigen EBNA-LP is essential for transforming naive B cells, and facilitates recruitment of transcription factors to the viral genome. *PLoS pathogens*, 14(2), p.e1006890.

Takacs, M. et al., 2010. Epigenetic regulation of latent Epstein-Barr virus promoters. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1799(3–4), pp.228–235. Available at: http://dx.doi.org/10.1016/j.bbagrm.2009.10.005.

Takada, K. & Ono, Y., 1989. Synchronous ans sequentian activation of latently infected Epstein-Barr virus genomes. *J.Virol.*, 63(1), pp.445–449.

Thorley-lawson, D.A., 2015. *EBV Persistence—Introducing the Virus*, Available at: http://link.springer.com/10.1007/978-3-319-22822-8.

Thorley-Lawson, D.A., 2001. Epstein-Barr virus: exploiting the immune system. *Nature Reviews Immunology*, 1(1), pp.75–82. Available at: http://www.nature.com/doifinder/10.1038/35095584.

Thorley-Lawson, D.A. & Gross, A., 2004. Persistence of the Epstein–Barr Virus and the Origins of Associated Lymphomas. *New England Journal of Medicine*, 350(13), pp.1328–1337. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMra032015.

Thorley-Lawson, D. a & Allday, M.J., 2008. The curious case of the tumour virus: 50 years of Burkitt's lymphoma. *Nature reviews. Microbiology*, 6(12), pp.913–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19008891 [Accessed August 20, 2014].

Thorley-Lawson, D. & Babcock, G., 1999. A model for persistent infection with Epstein-Barr

virus: the stealth virus of human B cells. *Life sciences*, 65(14), pp.1433–1453. Available at: http://www.sciencedirect.com/science/article/pii/S0024320599002143.

Tierney, R. et al., 2000. The Epstein-Barr Virus Promoter Initiating B-Cell Transformation Is Activated by RFX Proteins and the B-Cell-Specific Activator Protein BSAP/Pax5. *Journal of Virology* , 74(22), pp.10458–10467. Available at: http://jvi.asm.org/content/74/22/10458.abstract.

Tierney, R.J. et al., 2011. Epstein-Barr Virus BamHI W Repeat Number Limits EBNA2 / EBNA-LP Coexpression in Newly Infected B Cells and the Efficiency of B-Cell Transformation : a Rationale for the Multiple W Repeats in Wild-Type Virus Strains Epstein-Barr Virus BamHI W Repeat Number. *Journal of Virology*, 85(23), pp.12362–12375.

Tomkinson, B. & Kieff, E., 1992. Use of Second-Site Homologous Recombination To. *Journal of virology*, 66(5), pp.2893–2903.

Tomkinson, B., Robertson, E. & Kieff, E., 1993. Epstein-Barr virus nuclear proteins EBNA-3A and EBNA-3C are essential for B-lymphocyte growth transformation. *Journal of virology*, 67(4), pp.2014–25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=240270&tool=pmcentrez&rendertype=abstract.

Toussirot, É. & Roudier, J., 2008. Epstein–Barr virus in autoimmune diseases. *Best Practice & Research Clinical Rheumatology*, 22(5), pp.883–896. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1521694208001046.

Treangen, T.J. & Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(November 2011). Available at: http://www.nature.com/doifinder/10.1038/nrg3117.

Tsai, C.N. et al., 1999. Additive effect of Sp1 and Sp3 in regulation of the ED-L1E promoter of the EBV LMP 1 gene in human epithelial cells. *Virology*, 261(2), pp.288–294.

Tsai, M.-H., Raykova, A., Klinke, O., Bernhardt, K., Gärtner, K., et al., 2013. Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. *Cell reports*, 5(2), pp.458–70. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24120866 [Accessed February 22, 2014].

Tsai, M.-H. et al., 2017. The biological properties of different Epstein-Barr virus strains explain their association with various types of cancers. *Oncotarget*, 8(6), pp.10238–10254.

Tsimpidis, M. et al., 2017. T-RECs: rapid and large-scale detection of recombination events among different evolutionary lineages of viral genomes. *BMC Bioinformatics*, 18(1), p.13. Available at: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1420-z.

Tso, K.K.-Y. et al., 2013. Complete genomic sequence of Epstein-Barr virus in nasopharyngeal carcinoma cell line C666-1. *Infectious agents and cancer*, 8(1), p.29. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3734220&tool=pmcentrez&rendertype=abstract.

Tzellos, S. et al., 2014. A single amino acid in EBNA-2 determines superior B lymphoblastoid cell line growth maintenance by Epstein-Barr virus type 1 EBNA-2. *Journal of virology*, 88(16), pp.8743–53. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24850736 [Accessed September 30, 2014].

Tzellos, S. & Farrell, P.J., 2012. Epstein-Barr Virus Sequence Variation — Biology and Disease. , pp.156–175.

Umbach, J.L. et al., 2008. MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. *Nature*, 454(7205), pp.780–783.

Vrzalikova, K. et al., 2011. Down-regulation of BLIMP1α by the EBV oncogene, LMP-1, disrupts the plasma cell differentiation program and prevents viral replication in B cells: implications for the pathogenesis of EBV-associated B-cell lymphomas. *Blood*, 117(22), pp.5907–5917.

Waltzer, L. et al., 1996. Epstein-Barr virus EBNA3A and EBNA3C proteins both repress RBP-J kappa-EBNA2-activated transcription by inhibiting the binding of RBP-J kappa to DNA. *Journal of virology*, 70(9), pp.5909–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=190609&tool=pmcentrez&rendertype=abstract.

Wang, J. & Quake, S.R., 2014. RNA-guided endonuclease provides a therapeutic strategy to cure latent herpesviridae infection. *Proceedings of the National Academy of Sciences*, 111(36), pp.13157–13162. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1410785111.

Wang, W. et al., 2017. Impact of different promoters, promoter mutation, and an enhancer on recombinant protein expression in CHO cells. *Scientific Reports*, 7(1), pp.1–10. Available at: http://dx.doi.org/10.1038/s41598-017-10966-y.

Wang, X. et al., 1998. Epstein-Barr virus uses different complexes of glycoproteins gH and gL to infect B lymphocytes and epithelial cells. *Journal of virology*, 72(7), pp.5552–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=110204&tool=pmcentrez&rendertype=abstract.

Weiss, E.R. et al., 2017. Early EBV genomic diversity and convergence towards the B95.8 genome in primary infection. *Journal of Virology* . Available at: http://jvi.asm.org/content/early/2017/10/26/JVI.01466-17.abstract.

Westphal, E. et al., 1999. Induction of Lytic Epstein-Barr Virus (EBV) Infection in EBV-associated Malignancies Using Adenovirus Vectors in Vitro and in Vivo Induction of Lytic Epstein-Barr Virus (EBV) Infection in EBV-associated Malignancies Using Adenovirus Vectors in Vitro a. *Cancer Research*, pp.1485–1491.

White, R.E. et al., 2012. EBNA3B-deficient EBV promotes B cell lymphomagenesis in humanized mice and is found in human tumors. *The Journal of Clinical Investigatio*, 122(4).

Wilson, G. & Miller, G., 1979. Recovery of Epstein-Barr virus from nonproducer neonatal human lymphoid cell transformants. *Virology*, 95(2), pp.351–358.

Woisetschlaeger, M. et al., 1990. Promoter switching in Epstein-Barr virus during the initial stages of infection of B lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 87(5), pp.1725–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=53555&tool=pmcentrez&rendertype=abstract.

Woisetschlaeger, M. et al., 1991. Role for the Epstein-Barr virus nuclear antigen 2 in viral promoter switching during initial stages of infection. *Proceedings of the National Academy of Sciences of the United States of America*, 88(9), pp.3942–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51569&tool=pmcentrez&rendertype=abstract.

Wu, Y. et al., 2007. Epstein-Barr virus (EBV)-encoded RNA 2 (EBER2) but not EBER1 plays a critical role in EBV-induced B-cell growth transformation. *Journal of virology*, 81(20), pp.11236–45. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2045559&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/17686859%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2045559.

Wysokenski, D. a & Yates, J.L., 1989. Multiple EBNA1-binding sites are required to form an EBNA1-dependent enhancer and to activate a minimal replicative origin within oriP of Epstein-Barr virus. *Journal of virology*, 63(6), pp.2657–2666.

Xiao, K. et al., 2016. Genome-wide analysis of Epstein-Barr Virus (EBV) integration and strain in C666-1 and Raji cells. *Journal of Cancer*, 7(2), p.214–224.pdf.

Yajima, M., Kanda, T. & Takada, K., 2005. Critical Role of Epstein-Barr Virus (EBV) -Encoded RNA in Efficient EBV-Induced B-Lymphocyte Growth Transformation Critical Role of Epstein-Barr Virus (EBV) -Encoded RNA in Efficient EBV-Induced B-Lymphocyte Growth Transformation. , 79(7), pp.4298–4307.

Yandava, C.N. & Speck, S.H., 1992. Characterization of the Deletion and Rearrangement in the BamHI C Region of the X50-7 Epstein-Barr Virus Genome , a Mutant Viral Strain Which Exhibits Constitutive BamHI W Promoter Activity. , 66(9), pp.5646–5650.

Yang, H.J. et al., 2013. Comprehensive profiling of Epstein-Barr virus-encoded miRNA species associated with specific latency types in tumor cells. *Virology Journal*, 10(1), p.1. Available at: Virology Journal.

Yao, R. et al., 2017. Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics*, 10(1), p.30. Available at: http://molecularcytogenetics.biomedcentral.com/articles/10.1186/s13039-017-0333-5.

Yoo, L.I. et al., 2002. Deletion of Epstein-Barr Virus Regulatory Sequences Upstream of the EBNA Gene Promoter Wp1 Is Unfavorable for B-Cell Immortalization Deletion of Epstein-Barr Virus Regulatory Sequences Upstream of the EBNA Gene Promoter Wp1 Is Unfavorable for B-Cell Immor. *Society*.

Yoon, S. et al., 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9), pp.1586–1592.

Young, L.S. & Murray, P.G., 2003. Epstein-Barr virus and oncogenesis: from latent genes to tumours. *Oncogene*, 22(33), pp.5108–21. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12910248 [Accessed September 25, 2013].

Young, L.S. & Rickinson, A.B., 2004. Epstein-Barr virus: 40 Years on. *Nature Reviews Cancer*, 4(10), pp.757–768.

Young, L.S., Sung, N.S. & Pagano, J.S., 2009. *Epstein-Barr Virus*. *Encyclopedia of Life Sciences*. Available at: http://doi.wiley.com/10.1002/9780470015902.a0001020.pub2.

Young, L.S., Yap, L.F. & Murray, P.G., 2016. Epstein – Barr virus : more than 50 years old and still providing surprises. *Nature Publishing Group*, 16(12), pp.789–802. Available at: http://dx.doi.org/10.1038/nrc.2016.92.

Yuen, K. et al., 2015. CRISPR / Cas9-mediated genome editing of Epstein – Barr virus in human cells. *Journal of General Virology*, (2015), pp.626–636.

Zalani, S., Holley-Guthrie, E. & Kenney, S., 1996. Epstein-Barr viral latency is disrupted by the immediate-early BRLF1 protein through a cell-specific mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 93(17), pp.9194–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=114329&tool=pmcentrez&rendertype=abstract.

Zeng, M. et al., 2005. Genomic Sequence Analysis of Epstein-Barr Virus Strain GD1 from a Nasopharyngeal Carcinoma Patient †. *Society*, 79(24), pp.15323–15330.

Zhao, B. & Sample, C.E., 2000. Epstein-barr virus nuclear antigen 3C activates the latent membrane protein 1 promoter in the presence of Epstein-Barr virus nuclear antigen 2 through sequences encompassing an spi-1/Spi-B binding site. *Journal of virology*, 74(11), pp.5151–60.

Zhou, F., Li, Q., Wong, S.W. & Gao, S.-J., 2010. Autoexcision of bacterial artificial chromosome facilitated by terminal repeat-mediated homologous recombination: a novel approach for generating traceless genetic mutants of herpesviruses. *Journal of virology*, 84(6), pp.2871–80.

Zimmermann, J. & Hammerschmidt, W., 1995. Structure and role of the terminal repeats of Epstein-Barr virus in processing and packaging of virion DNA. *Journal of virology*, 69(5), pp.3147–55.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research, 31(13), pp.3406–3415.

Appendix 1: EBNA-LP and BWRF1 subtypes of diverse viruses.

No.*	Sample name	Sample type	Geographic origin	EBNA2 type	Reference†	Accession number	Subtypes of each EBNA-LP exon				EBNA-LP subtype	BWRF1 subtype	Estimated #IR1
							W1	W2	Y1	Y2			
1	B95.8	Infectious mononucleosis	USA	Type 1	Baer et al 1983	V01555.2	1.1	1	1.1	B	1B	2A	13.9158879
	B95.8 (+Raji)				de Jesus et al. 2003	NC007605							
2	AG876	Burkitt's lymphoma	Ghana	Type 2	Dolan et al. 2006	NC009334	2	1	2.1	Z	2Z	2B	
3	GD1	NPC-saliva	China	Type 1	Zeng et al. 2005	AY961628	2	1	1.2	A	2A	1A	
4	GD2	NPC-tumour	China	Type 1	Liu et al. 2011	HQ020558	1.1	1	1.1	B-M	$1B^{M}$	2A	
5	HKNPC1	NPC-tumour	Hong Kong	Type 1	Kwok et al. 2012	JQ009376	2	1	ND	ND	N/A	2G	
6	Mutu	Burkitt's lymphoma	Kenya	Type 1	Lin et al. 2013	KC207814	2	1	1.1	C	2C	1A	
7	K4123-Mi	sLCL	USA	Type 1	Lei et al. 2013	KC440851	1.1	1	1.1	B	1B	1B	
8	K4413-Mi	sLCL	USA	Type 1	Lei et al. 2013	KC440852	2	1	1.1	C	2C	1A	
9	M81	NPC	Hong Kong	Type 1	Tsai et al, 2013	KF373730	2	1	1.1	B-M	$2B^{M}$	2A	
10	Raji	Burkitt's lymphoma	Nigeria	Type 1	Unpublished		1.1	1	1.1a	B	1B	2X	

11	Saliva1	Healthy saliva (wild type)	UK	Type 1		ERS074900	2	1	1.2	A	2A	1A	9.3
12	HL01	Hodgkin's lymphoma	UK	Type 1		ERS549508	1.1	1	1.1	B	1B	2F	10.4
13	HL02	Hodgkin's lymphoma	UK	Type 1		ERS549509	2	1	1.1	C	2C	1A	10.0
14	HL04	Hodgkin's lymphoma	UK	Type 1		ERS549511	1.1	1	1.1	B	1B	2X (17% 1X)	6.7
15	HL05	Hodgkin's lymphoma	UK	Type 1		ERS549512	1.2	1	1.1	B-P	$1B^{P}$	2X (27% 1X)	5.7
16	HL08	Hodgkin's lymphoma	UK	Type 1		ERS549513	2	1	1.1	C	2C	1A	10.3
17	HL09	Hodgkin's lymphoma	UK	Type 1		ERS549514	1.2	1	1.1	B	1B	2X	8.9
18	HL11	Hodgkin's lymphoma	UK	Type 1		ERS549515	2	1	1.1	C	2C	1A	10.9
19	L591	Hodgkin's lymphoma cell line	Germany	Type 1		ERS549506	2	1	1.1	C	2C	2A	3.2
20	YCCEL1	Gastric carcinoma cell line	Korea	Type 1		ERS549504	2	1	1.2	A	2A	1A	6.4
21	Wewak1	Burkitt's lymphoma	PNG	Type 2		ERS074939	2	1	2.2	Z	2Z	2D	3.7
22	HKN14	sLCL	Hong Kong	Type 1		ERS549501	2	1	1.2	A	2A	1A	7.0
23	HKN15	sLCL	Hong Kong	Type 1		ERS549502	2	1	1.2	A	2A	1A	6.9
24	HKN19	sLCL	Hong Kong	Type 1		ERS549503	2	1	1.2	A	2A	1A	7.2
25	D3201.2	NPC biopsy	China	Type 1		ERS549507	2	1	1.2	A	2A	1A	5.6

26	C666.1imp	NPC cell line	China	Type 1		ERS549471	2	1	1.1	B-M	$2B^{M}$	1A (15% 2A)	4.0
27	M-ABA	LCL from NPC virus	North Africa	Type 1		ERS074943	2	1	1.2	A	2A	2A	4.4
28	Akata 2000	Burkitt's lymphoma	Japan	Type 1		ERS074927	2	1	1.2	A	2A	1A	3.6
29	Jijoye	Burkitt's lymphoma	Nigeria	Type 2		ERS074941	2	1	1.1	B	2B	1B	3.6
30	Cheptages	Burkitt's lymphoma	Kenya	Type 2		ERS074945	1	1	del	del	N/A	2X	10.4
31	Daudi	Burkitt's lymphoma	Kenya	Type 1		ERS074946	2	1	del	del	N/A	1A	8.4
32	BL37	Burkitt's lymphoma	Africa	Type 1		ERS549473	2	1	1.2	A	2A	2A	8.3
33	Makau	Burkitt's lymphoma	Kenya	Type 1		ERS549472	2	1	1.1	C	2C	3A	5.1
34	sLCL-IM1.02	sLCL. IM.	Australia	Type 1		ERS074911	1.2	1	1.1	B	1B	2X	3.5
35	sLCL-IM1.05	sLCL. IM.	Australia	Type 1		ERS074912	1.2	1	1.1	B	1B	1D	3.4
36	sLCL-IM1.09	sLCL. IM.	Australia	Type 1		ERS074929	2	1	1.1	C	2C	1A	3.9
37	sLCL-IM1.16	sLCL. IM.	Australia	Type 1		ERS074925	2	1	1.2	A	2A	1A	3.6
38	sLCL-IM1.17	sLCL. IM	Australia	Type 1		ERS074932	2	1	1.1	C-H	$2C^{H}$	1A	3.8
39	sLCL-IS1.01	sLCL. PTLD	Australia	Type 1		ERS074914	2	1	1.1	C	2C	1A	4.0
40	sLCL-IS1.03	sLCL. PTLD	Australia	Type 1		ERS074915	2	1	1.1	C	2C	1A	4.8
41	sLCL-IS1.04	sLCL. PTLD	Australia	Type 1		ERS074923	1.1	1	1.1	B	1B	2A	3.7
42	sLCL-IS1.06	sLCL. PTLD	Australia	Type 1		ERS074916	2	1	1.1	C	2C	1A	4.8
43	sLCL-IS1.07	sLCL. PTLD	Australia	Type 1		ERS074926	1.1	1	1.1	B	1B	1A	3.2
44	sLCL-IS1.08	sLCL. PTLD	Australia	Type 1		ERS074917	2	1	1.1	C-I	$2C^{I}$	1A	3.8
45	sLCL-IS1.10	sLCL. PTLD	Australia	Type 1		ERS074918	2	1	1.1	C	2C	1A	4.5

46	sLCL-IS1.11	sLCL. PTLD	Australia	Type 1		ERS074919	2	1	1.1	C	2C	1A	5.2
47	sLCL-IS1.12	sLCL. PTLD	Australia	Type 1		ERS074930	2	1	1.1	C	2C	1A	4.5
48	sLCL-IS1.13	sLCL. PTLD	Australia	Type 1		ERS074913	1.1	1	1.1	B	1B	2X	3.8
49	sLCL-IS1.14	sLCL. PTLD	Australia	Type 1		ERS074931	2	1	1.1	C	2C	1A	5.0
50	sLCL-IS1.15	sLCL. PTLD	Australia	Type 1		ERS074927	2	1	1.2	A	2A	1X	4.1
51	sLCL-IS1.18	sLCL. PTLD	Australia	Type 1		ERS074920	2	1	1.1	C-I	2CI	2A	4.7
52	sLCL-IS1.19	sLCL. PTLD	Australia	Type 1		ERS074925	2	1	1.1	C	2C	1A	4.9
53	sLCL-IS1.20	sLCL. PTLD	Australia	Type 1		ERS074926	2	1	1.1	C	2C	1A	5.2
54	sLCL-IS2.01	sLCL. PTLD	Australia	Type 2		ERS074928	2	1	2.2	Z	2Z	1A	4.1
55	sLCL-1.02	sLCL	Kenya	Type 1		ERS549478	2	1	1.1	C	2C	1A	7.3
56	sLCL-BL1.03	sLCL	Kenya	Type 1		ERS549479	2	1	1.1	C	2C	1A	3.1
57	sLCL-1.04	sLCL	Kenya	Type 1		ERS549480	2	1	1.1	C	2C	1A	6.1
58	sLCL-1.05	sLCL	Kenya	Type 1		ERS549481	1.1	1	1.1	B	1B	1A	8.6
59	sLCL-1.06	sLCL	Kenya	Type 1		ERS549482	2	1	1.1	C	2C	1A	8.7
60	sLCL-1.07	sLCL	Kenya	Type 1		ERS549483	2	1	1.1	C	2C	1A	7.3
61	sLCL-1.08	sLCL	Kenya	Type 1		ERS549484	2	1	1.1	C	2C	1A	7.4
62	sLCL-1.09	sLCL	Kenya	Type 1		ERS549485	1.1	1	1.1	B	1B	1A	6.0
63	sLCL-1.10	sLCL	Kenya	Type 1		ERS549486	2	1	1.1	C	2C	1C	6.0
64	sLCL-1.11	sLCL	Kenya	Type 1		ERS549487	2	1	1.1	C	2C	1A	7.0
65	sLCL-1.12	sLCL	Kenya	Type 1		ERS549488	2	1	1.1	C	2C	1A	9.8
66	sLCL-1.13	sLCL	Kenya	Type 1		ERS549489	1.1	1	1.1	B	1B	1A	7.1
67	sLCL-2.14	sLCL	Kenya	Type 2		ERS549490	2	1	2.1	Z	2Z	2B	6.2
68	sLCL-2.15	sLCL	Kenya	Type 2		ERS549491	2	1	2.2	Z	2Z	2B	7.1
69	sLCL-2.16	sLCL	Kenya	Type 2		ERS549492	2	1^{R}	2.2	Z	$2Z^{R}$	2B	5.8
70	sLCL-1.17	sLCL	Kenya	Type 1		ERS549493	2	1	1.1	C	2C	1A	6.6
71	sLCL-1.18	sLCL	Kenya	Type 1		ERS549494	1.1	1	1.1	B	1B	1A	6.8
72	sLCL-1.19	sLCL	Kenya	Type 1		ERS549495	2	1	1.2	A	2A	2E	6.4

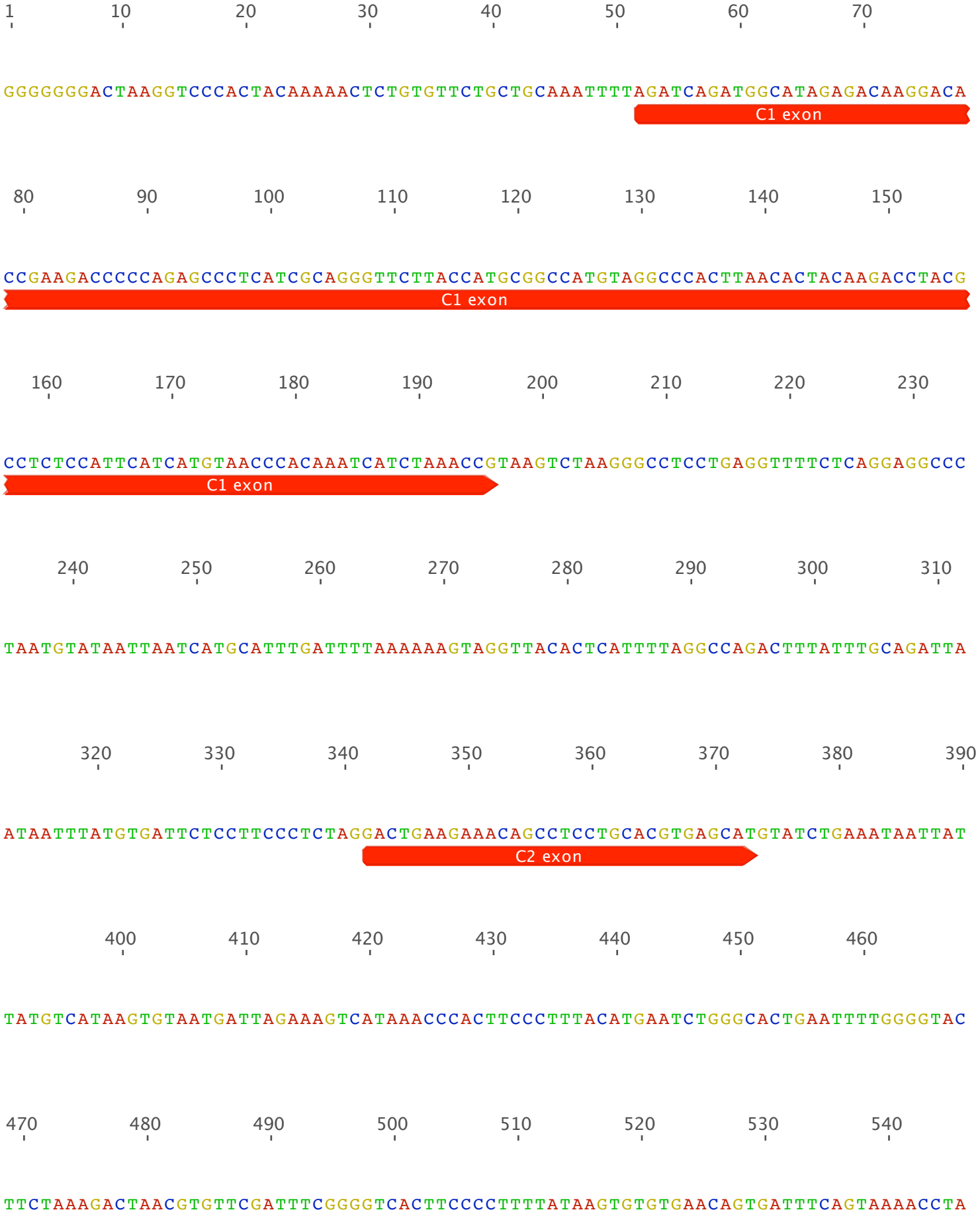
73	sLCL-BL1.20	sLCL	Kenya	Type 1		ERS549496	2	1	1.1	C	2C	1A	6.3
74	sLCL-2.21	sLCL	Kenya	Type 2		ERS549497	2	1	2.1	Z	2Z	2B	5.6
75	sLCL-2.22	sLCL	Kenya	Type 2		ERS549498	2	1	2.1	Z	2Z	2B	6.6
76	sLCL-1.24	sLCL	Kenya	Type 1		ERS549499	1.1	1	1.1	B	1B	1A	7.2
77	pLCL-TRL1-post	sLCL. PTLD fatal	USA	Type 1		ERS074922	2	1	1.1	B	2B	2B (15% 1A)	3.5
78	pLCL-TRL595	sLCL. PTLD	USA	Type 1		ERS074924	2	1	1.1	C	2C	1A	5.0
Strains excluded from quantitative analyses #													
79	Akata	Burkitt's lymphoma	Japan	Type 1	Lin et al. 2013	KC207813	see Akata 2000 (31)						
80	C666.1	NPC	China	Type 1	Tso et al. 2013	KC617875	See C666.1_imp (29)						
81	AFB1	LCL		Type 2		ERS549470	IR1 failed to assemble						
82	P3HR1 c16	Burkitt's lymphoma	Nigeria	Type 2		ERS074940	As Jijoye (32) but with deleted Y exons						5.2
83	BL36	Burkitt's lymphoma	North Africa	Type 1		ERS549475	As B95-8 (1)						4.6
84	Mak1 duplicate	Burkitt's lymphoma	Kenya	Type 1		ERS074936	as Makau (38)						3.3
85	pLCL-TRL1-pre	sLCL. PTLD early	USA	Type 1		ERS074921	as pLCL-TRL1-post						3.7
86	X50-7	LCL	USA	Type 1		ERS074942	As B95-8 (1)						3.4
87	LCL B95-8 del EBER2	LCL (from B95-8, deleted in EBER2)	USA	Type 1		ERS549474	As B95-8 (1)						4.0

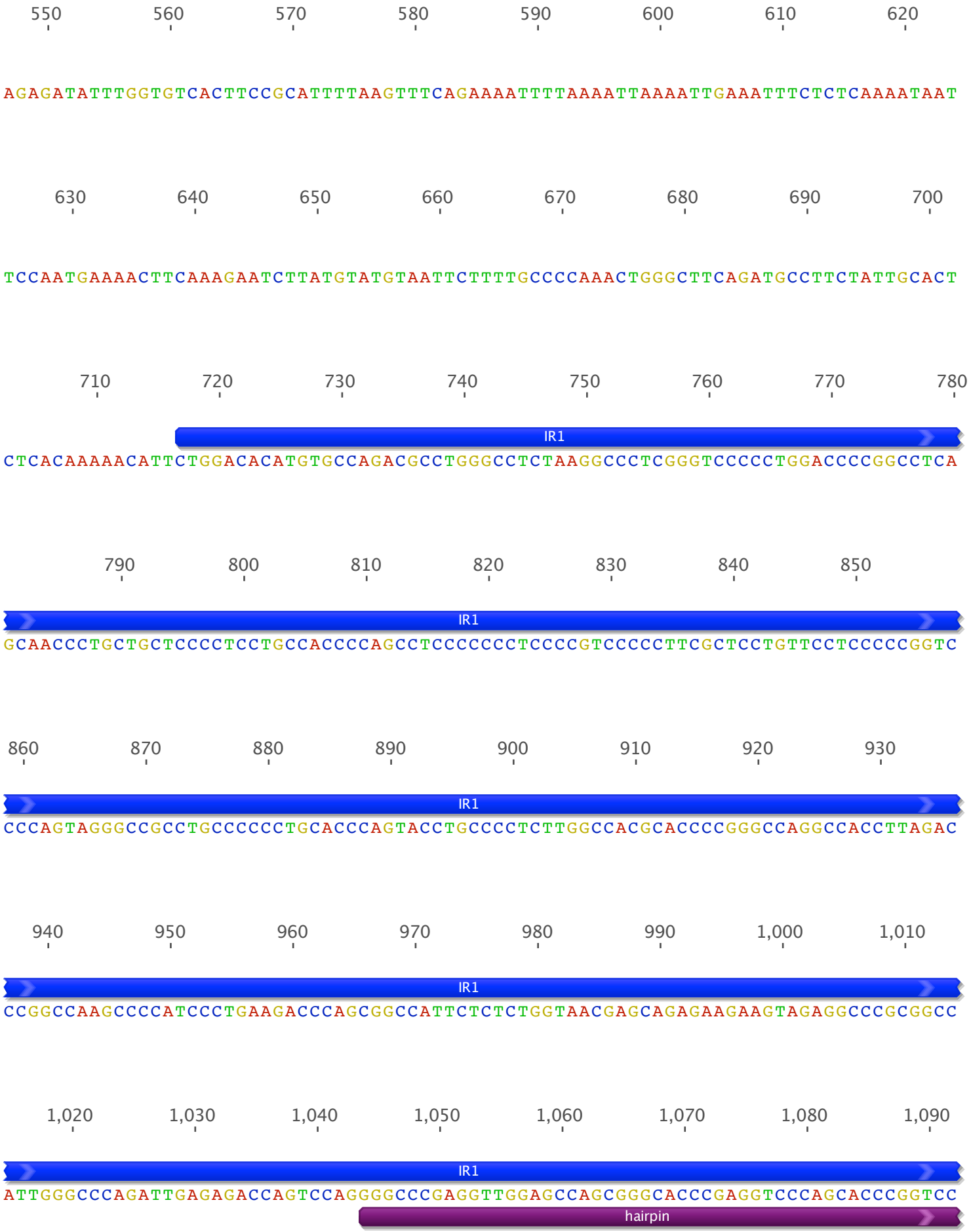
*** Sample numbers are the same as those generated in Palser *et al* (2015).**

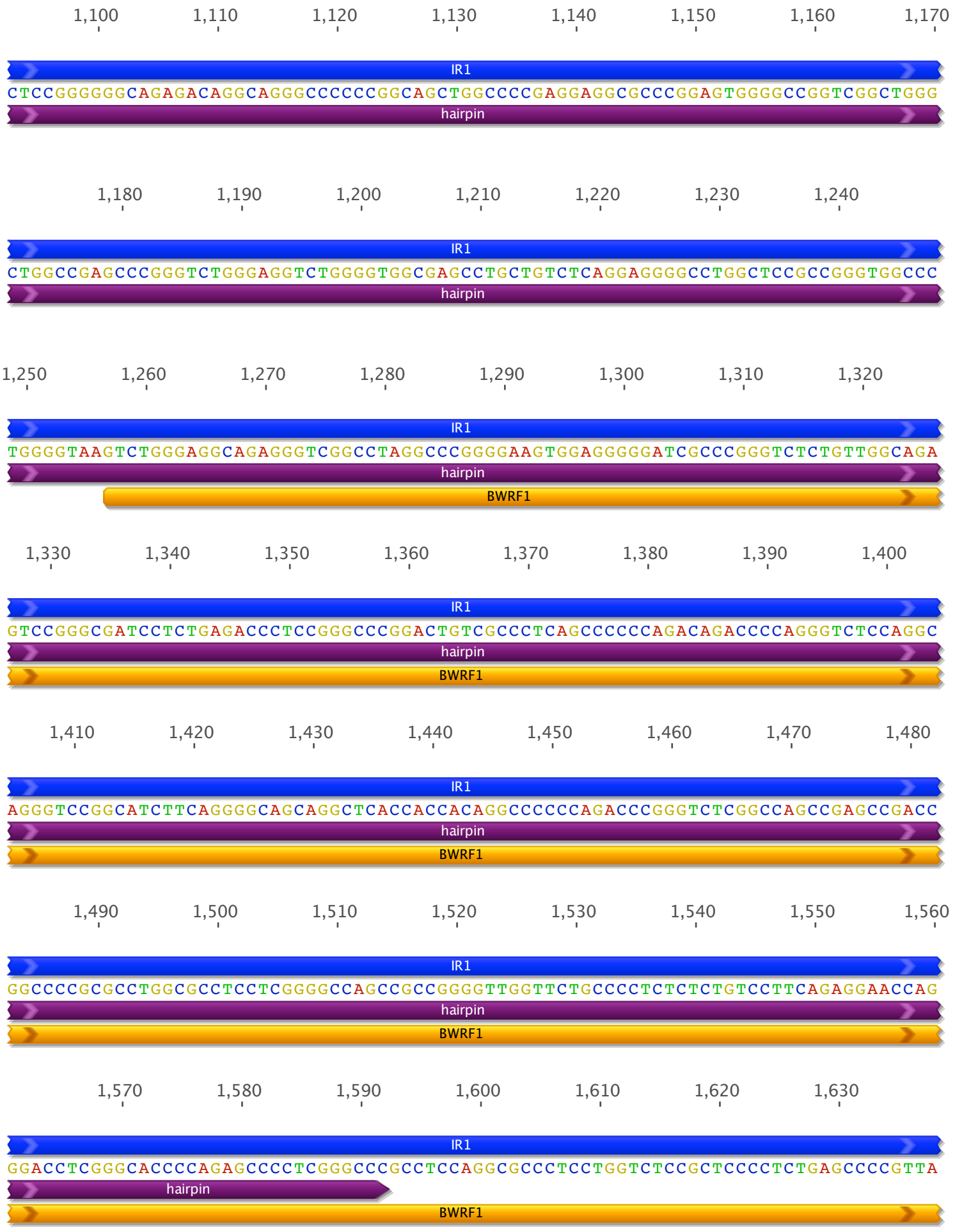
† Sequences first described in Palser *et al* (2015) unless otherwise stated.

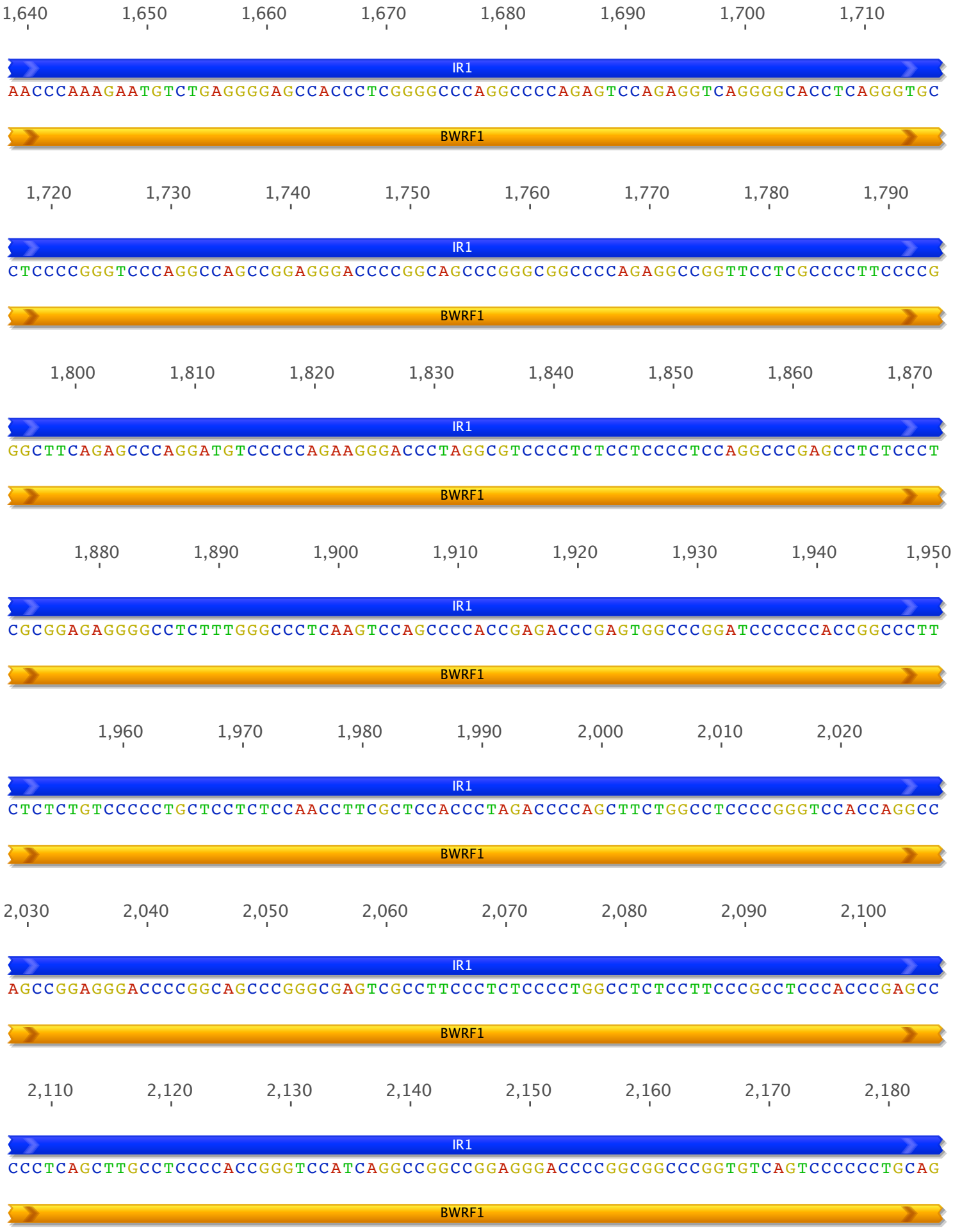
Seqeunces that were excluded from quantitative analyses (mainly because they shared a common origin with other strains) are shown along with the strain that represents them.

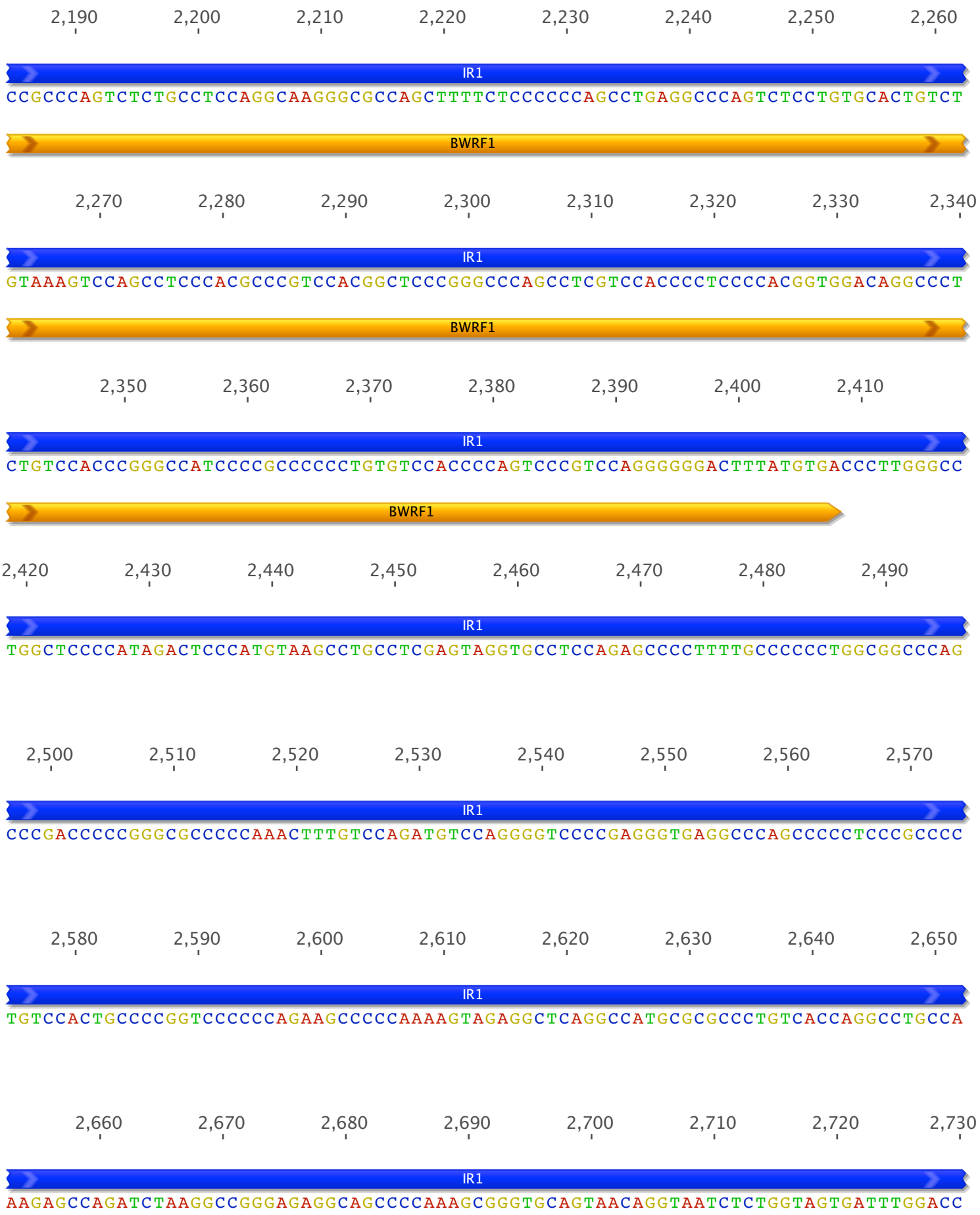
Appendix 2. DNA sequence with annotations of the B95-8 IR1 theoretical template

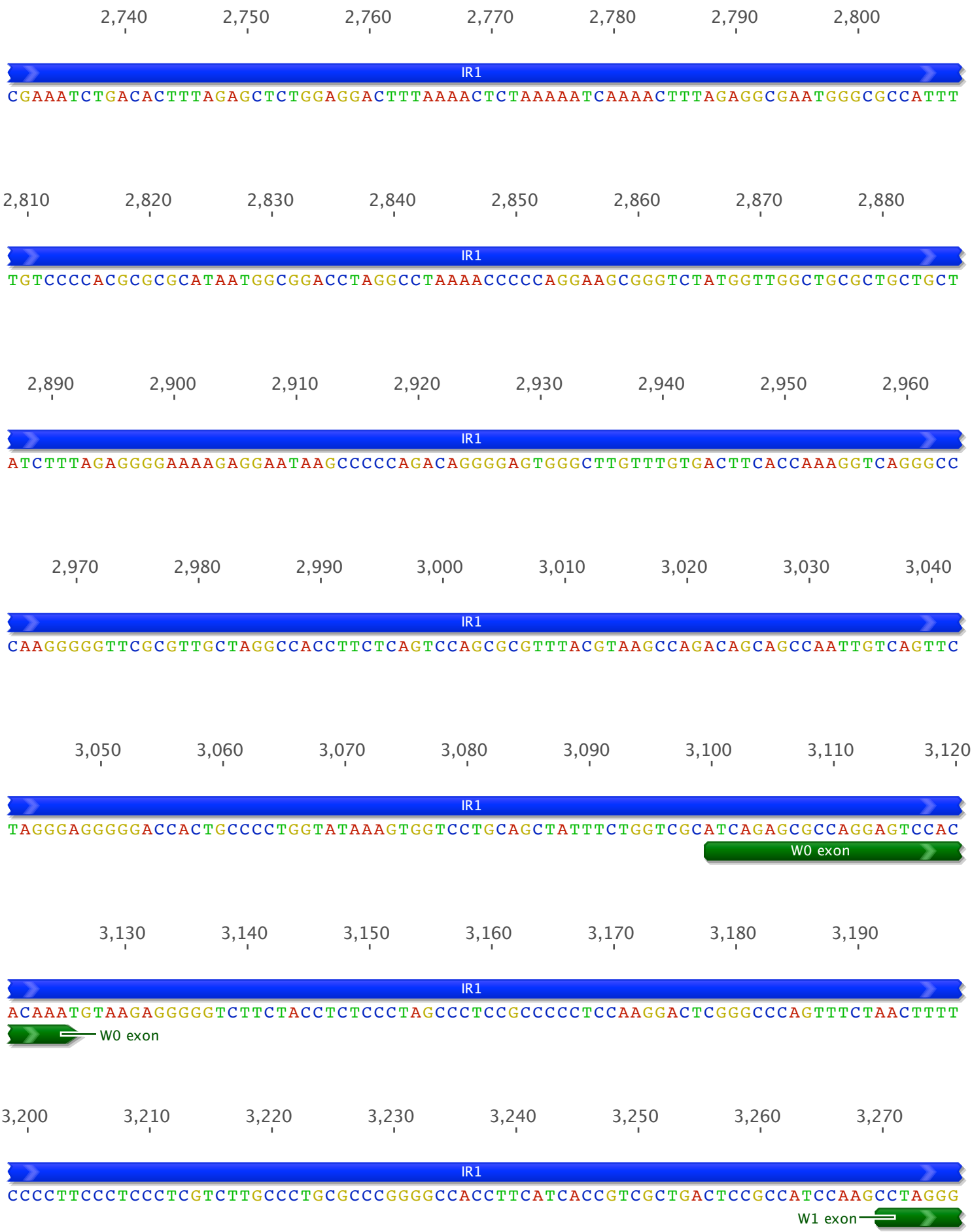


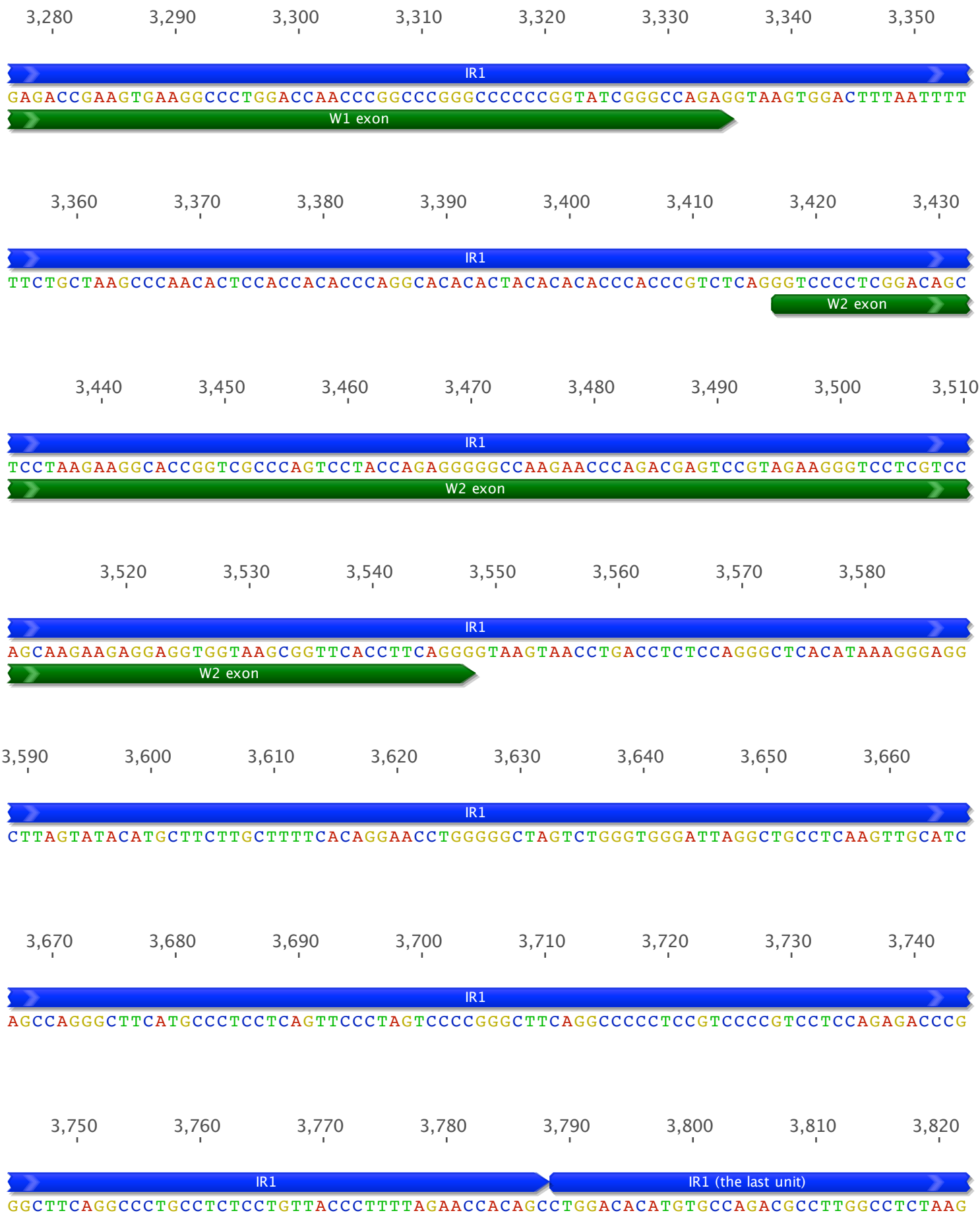

550
560
570
580
590
600
610
620
AGAGATATTTGGTGTCACTTCCGCATTTTAAGTTTCAGAAAATTTTAAAATTAAAATTGAAATTTCTCTCAAAATAAT
630
640
650
660
670
680
690
700
TCCAATGAAAACTTCAAAGAATCTTATGTATGTAATTCTTTTGCCCCAAACTGGGCTTCAGATGCCTTCTATTGCACT
710
720
730
740
750
760
770
780
IR1
CTCACAAAAACATTCTGGACACATGTGCCAGACGCCTGGGCCTCTAAGGCCCTCGGGTCCCCCTGGACCCCGGCCTCA
790
800
810
820
830
840
850
IR1
GCAACCCTGCTGCTCCCCTCCTGCCACCCCAGCCTCCCCCCCTCCCCGTCCCCCTTCGCTCCTGTTCCTCCCCCGGTC
860
870
880
890
900
910
920
930
IR1
CCCAGTAGGGCCGCCTGCCCCCCTGCACCCAGTACCTGCCCCTCTTGGCCACGCACCCCGGGCCAGGCCACCTTAGAC
940
950
960
970
980
990
1,000
1,010
IR1
CCGGCCAAGCCCCATCCCTGAAGACCCAGCGGCCATTCTCTCTGGTAACGAGCAGAGAAGAAGTAGAGGCCCGCGGCC
1,020
1,030
1,040
1,050
1,060
1,070
1,080
1,090
IR1
ATTGGGCCCAGATTGAGAGACCAGTCCAGGGGCCCGAGGTTGGAGCCAGCGGGCACCCGAGGTCCCAGCACCCGGTCC
hairpin

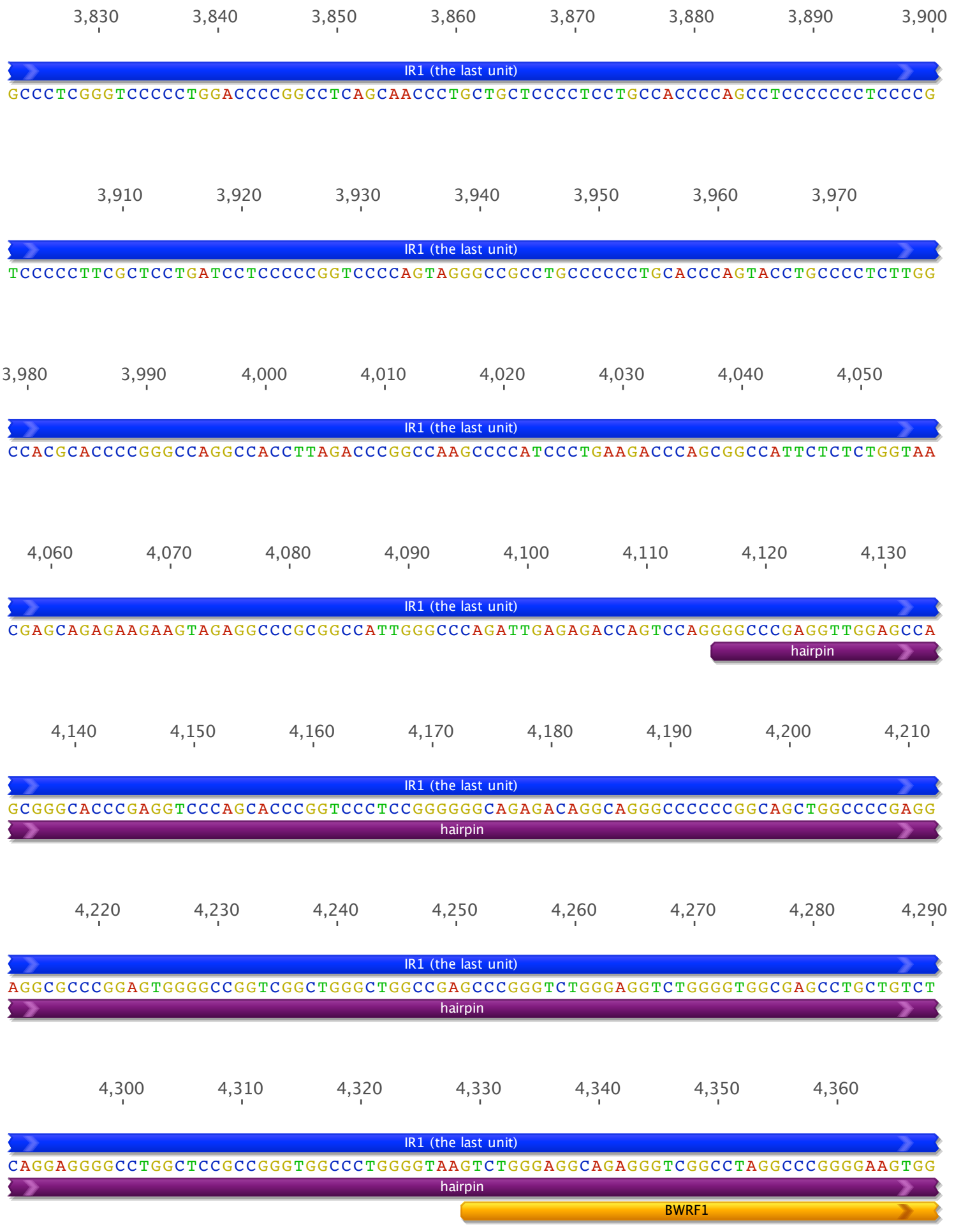

1,100
1,110
1,120
1,130
1,140
1,150
1,160
1,170
IR1
CTCCGGGGGGCAGAGACAGGCAGGGCCCCCCGGCAGCTGGCCCCGAGGAGGCGCCCGGAGTGGGGCCGGTCGGCTGGG
hairpin
1,180
1,190
1,200
1,210
1,220
1,230
1,240
IR1
CTGGCCGAGCCCGGGTCTGGGAGGTCTGGGGTGGCGAGCCTGCTGTCTCAGGAGGGGCCTGGCTCCGCCGGGTGGCCC
hairpin
1,250
1,260
1,270
1,280
1,290
1,300
1,310
1,320
IR1
TGGGGTAAGTCTGGGAGGCAGAGGGTCGGCCTAGGCCCGGGGAAGTGGAGGGGGATCGCCCGGGTCTCTGTTGGCAGA
hairpin
BWRF1
1,330
1,340
1,350
1,360
1,370
1,380
1,390
1,400
IR1
GTCCGGGCGATCCTCTGAGACCCTCCGGGCCCGGACTGTCGCCCTCAGCCCCCCAGACAGACCCCAGGGTCTCCAGGC
hairpin
BWRF1
1,410
1,420
1,430
1,440
1,450
1,460
1,470
1,480
IR1
AGGGTCCGGCATCTTCAGGGGCAGCAGGCTCACCACCACAGGCCCCCCAGACCCGGGTCTCGGCCAGCCGAGCCGACC
hairpin
BWRF1
1,490
1,500
1,510
1,520
1,530
1,540
1,550
1,560
IR1
GGCCCCGCGCCTGGCGCCTCCTCGGGGCCAGCCGCCGGGGTTGGTTCTGCCCCTCTCTCTGTCCTTCAGAGGAACCAG
hairpin
BWRF1
1,570
1,580
1,590
1,600
1,610
1,620
1,630
IR1
GGACCTCGGGCACCCCAGAGCCCCTCGGGCCCGCCTCCAGGCGCCCTCCTGGTCTCCGCTCCCCTCTGAGCCCCGTTA
hairpin
BWRF1

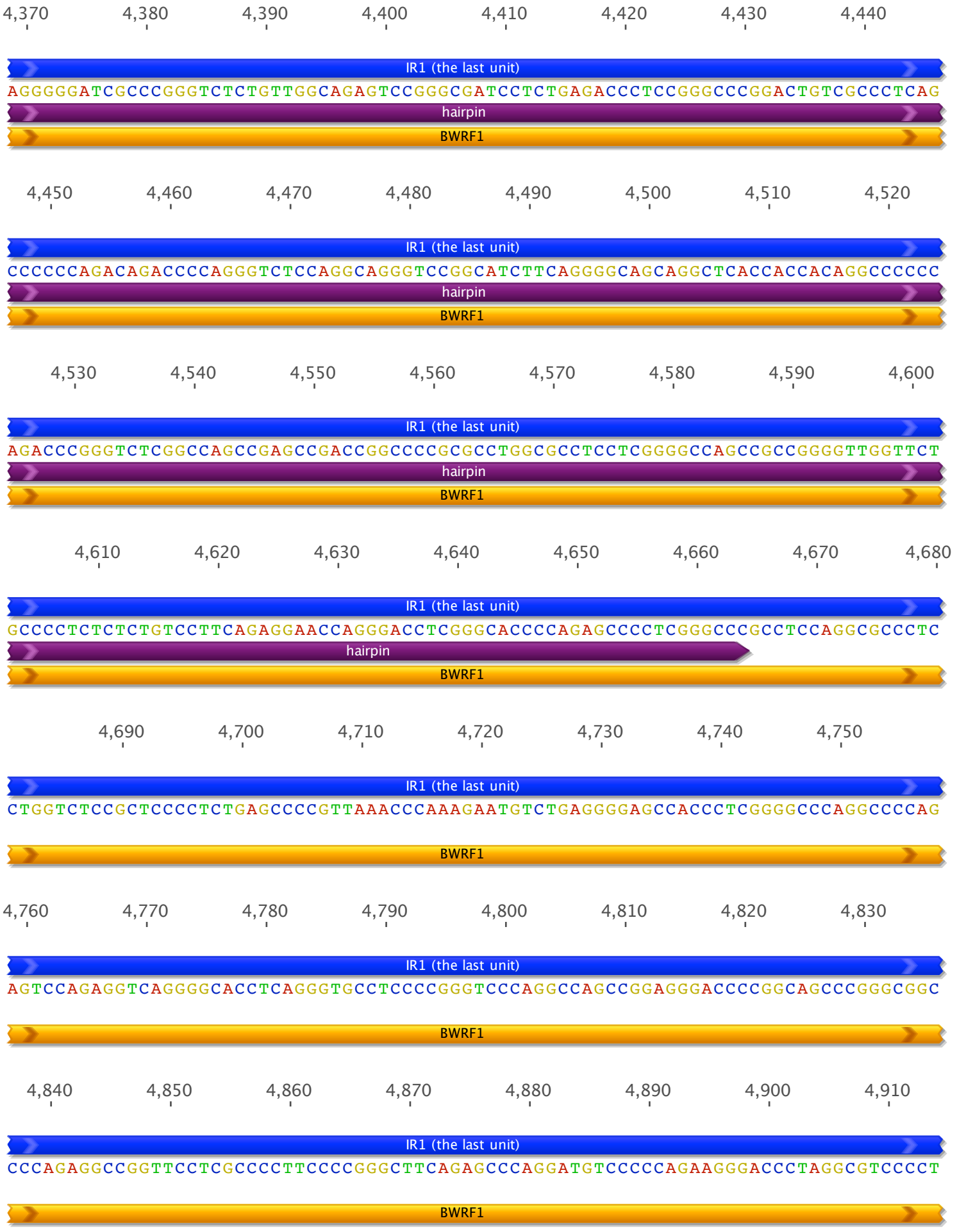

1,640 1,650 1,660 1,670 1,680 1,690 1,700 1,710
IR1
AACCCAAAGAATGTCTGAGGGGAGCCACCCTCGGGGCCCAGGCCCCAGAGTCCAGAGGTCAGGGGCACCTCAGGGTGC
BWRF1
1,720 1,730 1,740 1,750 1,760 1,770 1,780 1,790
IR1
CTCCCCGGGTCCCAGGCCAGCCGGAGGGACCCCGGCAGCCCGGGCGGCCCCAGAGGCCGGTTCCTCGCCCCTTCCCCG
BWRF1
1,800 1,810 1,820 1,830 1,840 1,850 1,860 1,870
IR1
GGCTTCAGAGCCCAGGATGTCCCCCAGAAGGGACCCTAGGCGTCCCCTCTCCTCCCCTCCAGGCCCGAGCCTCTCCCT
BWRF1
1,880 1,890 1,900 1,910 1,920 1,930 1,940 1,950
IR1
CGCGGAGAGGGGCCTCTTTGGGCCCTCAAGTCCAGCCCCACCGAGACCCGAGTGGCCCGGATCCCCCCACCGGCCCTT
BWRF1
1,960 1,970 1,980 1,990 2,000 2,010 2,020
IR1
CTCTCTGTCCCCCTGCTCCTCTCCAACCTTCGCTCCACCCTAGACCCCAGCTTCTGGCCTCCCCGGGTCCACCAGGCC
BWRF1
2,030 2,040 2,050 2,060 2,070 2,080 2,090 2,100
IR1
AGCCGGAGGGACCCCGGCAGCCCGGGCGAGTCGCCTTCCCTCTCCCCTGGCCTCTCCTTCCCGCCTCCCACCCGAGCC
BWRF1
2,110 2,120 2,130 2,140 2,150 2,160 2,170 2,180
IR1
CCCTCAGCTTGCCTCCCCACCGGGTCCATCAGGCCGGCCGGAGGGACCCCGGCGGCCCGGTGTCAGTCCCCCCTGCAG
BWRF1

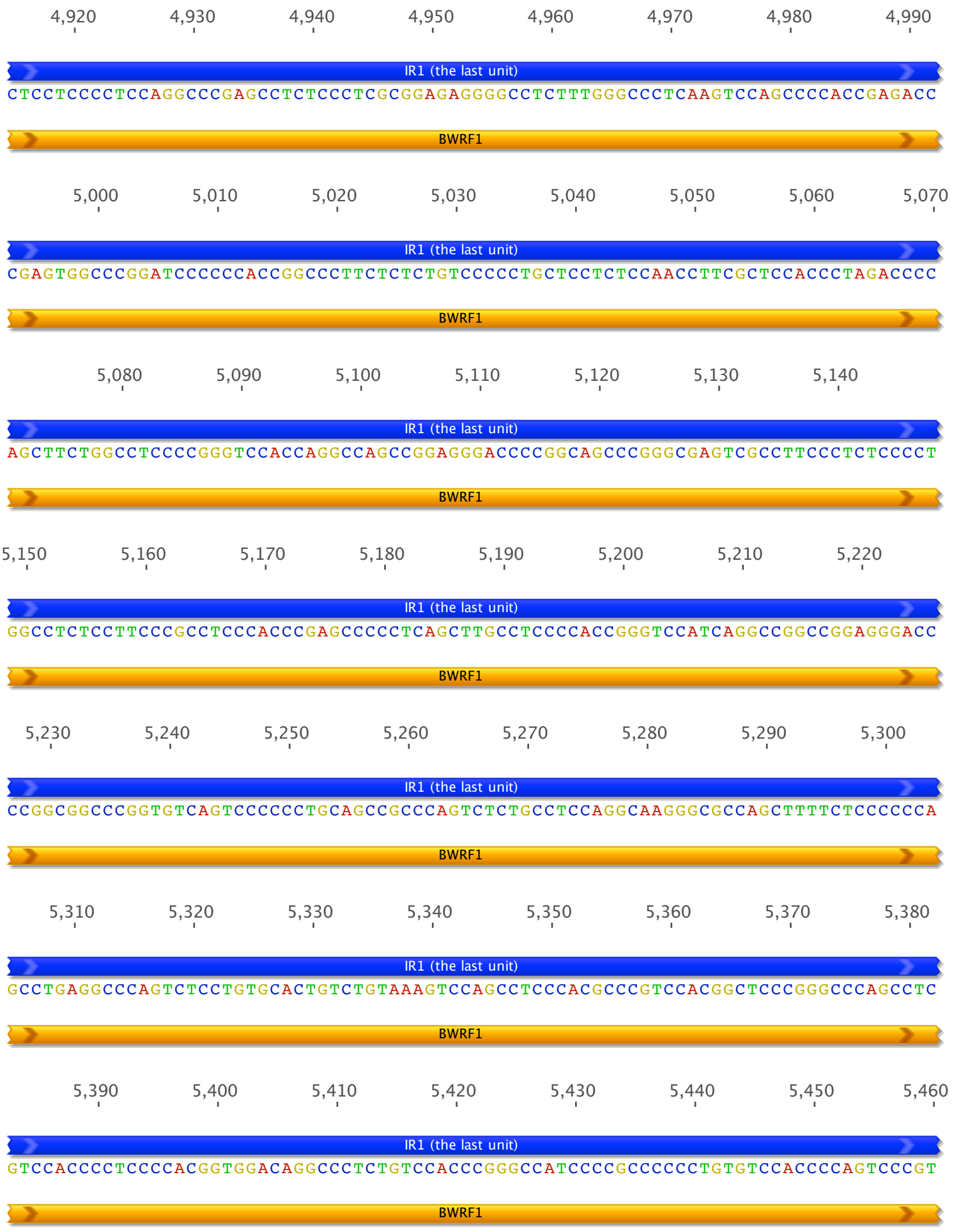

2,190
2,200
2,210
2,220
2,230
2,240
2,250
2,260
IR1
CCGCCCAGTCTCTGCCTCCAGGCAAGGGCGCCAGCTTTTCTCCCCCCAGCCTGAGGCCCAGTCTCCTGTGCACTGTCT
BWRF1
2,270
2,280
2,290
2,300
2,310
2,320
2,330
2,340
IR1
GTAAAGTCCAGCCTCCCACGCCCGTCCACGGCTCCCGGGCCCAGCCTCGTCCACCCCTCCCCACGGTGGACAGGCCCT
BWRF1
2,350
2,360
2,370
2,380
2,390
2,400
2,410
IR1
CTGTCCACCCGGGCCATCCCCGCCCCCCTGTGTCCACCCCAGTCCCGTCCAGGGGGGACTTTATGTGACCCTTGGGCC
BWRF1
2,420
2,430
2,440
2,450
2,460
2,470
2,480
2,490
IR1
TGGCTCCCCATAGACTCCCATGTAAGCCTGCCTCGAGTAGGTGCCTCCAGAGCCCCTTTTGCCCCCCTGGCGGCCCAG
2,500
2,510
2,520
2,530
2,540
2,550
2,560
2,570
IR1
CCCGACCCCCGGGCGCCCCCAAACTTTGTCCAGATGTCCAGGGGTCCCCGAGGGTGAGGCCCAGCCCCCTCCCGCCCC
2,580
2,590
2,600
2,610
2,620
2,630
2,640
2,650
IR1
TGTCCACTGCCCCGGTCCCCCCAGAAGCCCCCAAAAGTAGAGGCTCAGGCCATGCGCGCCCTGTCACCAGGCCTGCCA
2,660
2,670
2,680
2,690
2,700
2,710
2,720
2,730
IR1
AAGAGCCAGATCTAAGGCCGGGAGAGGCAGCCCCAAAGCGGGTGCAGTAACAGGTAATCTCTGGTAGTGATTTGGACC

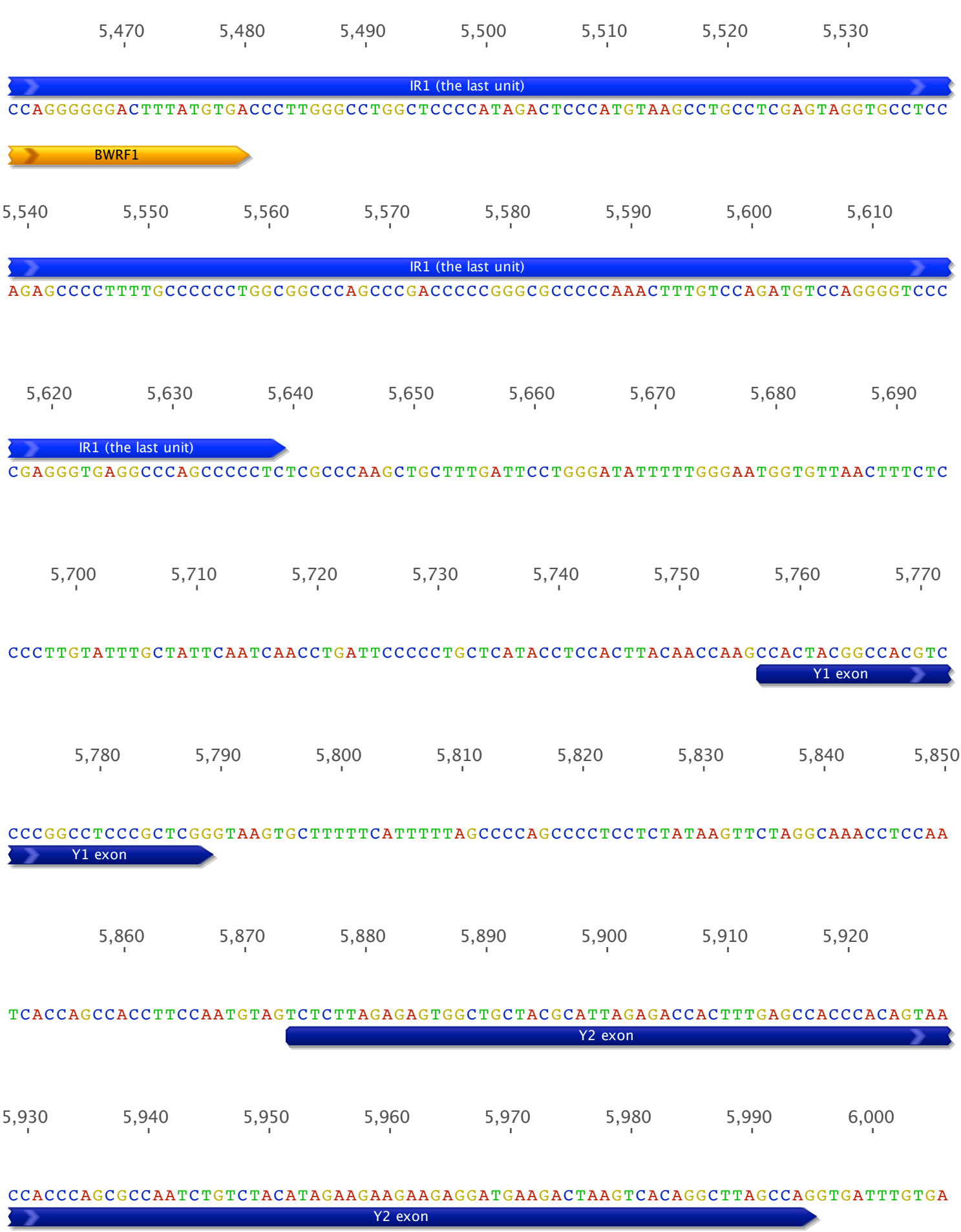

2,740
2,750
2,760
2,770
2,780
2,790
2,800
IR1
CGAAATCTGACACTTTAGAGCTCTGGAGGACTTTAAAACTCTAAAAATCAAAACTTTAGAGGCGAATGGGCGCCATTT
2,810
2,820
2,830
2,840
2,850
2,860
2,870
2,880
IR1
TGTCCCCACGCGCGCATAATGGCGGACCTAGGCCTAAAACCCCCAGGAAGCGGGTCTATGGTTGGCTGCGCTGCTGCT
2,890
2,900
2,910
2,920
2,930
2,940
2,950
2,960
IR1
ATCTTTAGAGGGGAAAAGAGGAATAAGCCCCCAGACAGGGGAGTGGGCTTGTTTGTGACTTCACCAAAGGTCAGGGCC
2,970
2,980
2,990
3,000
3,010
3,020
3,030
3,040
IR1
CAAGGGGGTTCGCGTTGCTAGGCCACCTTCTCAGTCCAGCGCGTTTACGTAAGCCAGACAGCAGCCAATTGTCAGTTC
3,050
3,060
3,070
3,080
3,090
3,100
3,110
3,120
IR1
TAGGGAGGGGGACCACTGCCCCTGGTATAAAGTGGTCCTGCAGCTATTTCTGGTCGCATCAGAGCGCCAGGAGTCCAC
W0 exon
3,130
3,140
3,150
3,160
3,170
3,180
3,190
IR1
ACAAATGTAAGAGGGGGTCTTCTACCTCTCCCTAGCCCTCCGCCCCCTCCAAGGACTCGGGCCCAGTTTCTAACTTTT
W0 exon
3,200
3,210
3,220
3,230
3,240
3,250
3,260
3,270
IR1
CCCCTTCCCTCCCTCGTCTTGCCCTGCGCCCGGGGCCACCTTCATCACCGTCGCTGACTCCGCCATCCAAGCCTAGGG
W1 exon

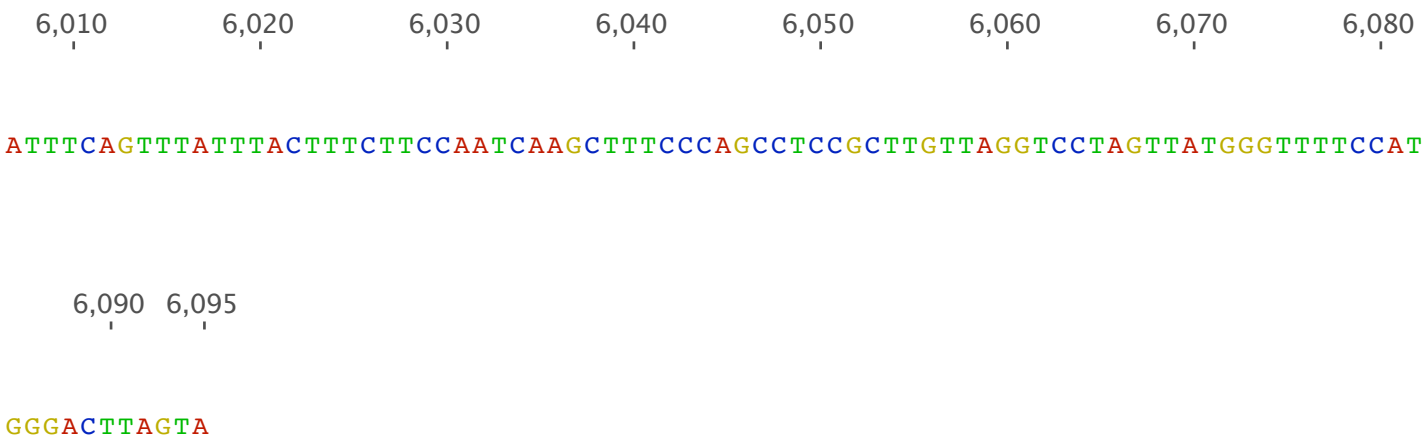

3,280
3,290
3,300
3,310
3,320
3,330
3,340
3,350
IR1
GAGACCGAAGTGAAGGCCCTGGACCAACCCGGCCCGGGCCCCCCGGTATCGGGCCAGAGGTAAGTGGACTTTAATTTT
W1 exon
3,360
3,370
3,380
3,390
3,400
3,410
3,420
3,430
IR1
TTCTGCTAAGCCCAACACTCCACCACACCCAGGCACACACTACACACACCCACCCGTCTCAGGGTCCCCTCGGACAGC
W2 exon
3,440
3,450
3,460
3,470
3,480
3,490
3,500
3,510
IR1
TCCTAAGAAGGCACCGGTCGCCCAGTCCTACCAGAGGGGGCCAAGAACCCAGACGAGTCCGTAGAAGGGTCCTCGTCC
W2 exon
3,520
3,530
3,540
3,550
3,560
3,570
3,580
IR1
AGCAAGAAGAGGAGGTGGTAAGCGGTTCACCTTCAGGGGTAAGTAACCTGACCTCTCCAGGGCTCACATAAAGGGAGG
W2 exon
3,590
3,600
3,610
3,620
3,630
3,640
3,650
3,660
IR1
CTTAGTATACATGCTTCTTGCTTTTCACAGGAACCTGGGGGCTAGTCTGGGTGGGATTAGGCTGCCTCAAGTTGCATC
3,670
3,680
3,690
3,700
3,710
3,720
3,730
3,740
IR1
AGCCAGGGCTTCATGCCCTCCTCAGTTCCCTAGTCCCCGGGCTTCAGGCCCCCTCCGTCCCCGTCCTCCAGAGACCCG
3,750
3,760
3,770
3,780
3,790
3,800
3,810
3,820
IR1
IR1 (the last unit)
GGCTTCAGGCCCTGCCTCTCCTGTTACCCTTTTAGAACCACAGCCTGGACACATGTGCCAGACGCCTTGGCCTCTAAG


3,830
3,840
3,850
3,860
3,870
3,880
3,890
3,900
IR1 (the last unit)
GCCCTCGGGTCCCCCTGGACCCCGGCCTCAGCAACCCTGCTGCTCCCCTCCTGCCACCCCAGCCTCCCCCCCTCCCCG
3,910
3,920
3,930
3,940
3,950
3,960
3,970
IR1 (the last unit)
TCCCCCTTCGCTCCTGATCCTCCCCCGGTCCCCAGTAGGGCCGCCTGCCCCCCTGCACCCAGTACCTGCCCCTCTTGG
3,980
3,990
4,000
4,010
4,020
4,030
4,040
4,050
IR1 (the last unit)
CCACGCACCCCGGGCCAGGCCACCTTAGACCCGGCCAAGCCCCATCCCTGAAGACCCAGCGGCCATTCTCTCTGGTAA
4,060
4,070
4,080
4,090
4,100
4,110
4,120
4,130
IR1 (the last unit)
CGAGCAGAGAAGAAGTAGAGGCCCGCGGCCATTGGGCCCAGATTGAGAGACCAGTCCAGGGGCCCGAGGTTGGAGCCA
hairpin
4,140
4,150
4,160
4,170
4,180
4,190
4,200
4,210
IR1 (the last unit)
GCGGGCACCCGAGGTCCCAGCACCCGGTCCCTCCGGGGGGCAGAGACAGGCAGGGCCCCCCGGCAGCTGGCCCCGAGG
hairpin
4,220
4,230
4,240
4,250
4,260
4,270
4,280
4,290
IR1 (the last unit)
AGGCGCCCGGAGTGGGGCCGGTCGGCTGGGCTGGCCGAGCCCGGGTCTGGGAGGTCTGGGGTGGCGAGCCTGCTGTCT
hairpin
4,300
4,310
4,320
4,330
4,340
4,350
4,360
IR1 (the last unit)
CAGGAGGGGCCTGGCTCCGCCGGGTGGCCCTGGGGTAAGTCTGGGAGGCAGAGGGTCGGCCTAGGCCCGGGGAAGTGG
hairpin
BWRF1


4,370
4,380
4,390
4,400
4,410
4,420
4,430
4,440
IR1 (the last unit)
AGGGGGATCGCCCGGGTCTCTGTTGGCAGAGTCCGGGCGATCCTCTGAGACCCTCCGGGCCCGGACTGTCGCCCTCAG
hairpin
BWRF1
4,450
4,460
4,470
4,480
4,490
4,500
4,510
4,520
IR1 (the last unit)
CCCCCCAGACAGACCCCAGGGTCTCCAGGCAGGGTCCGGCATCTTCAGGGGCAGCAGGCTCACCACCACAGGCCCCCC
hairpin
BWRF1
4,530
4,540
4,550
4,560
4,570
4,580
4,590
4,600
IR1 (the last unit)
AGACCCGGGTCTCGGCCAGCCGAGCCGACCGGCCCCGCGCCTGGCGCCTCCTCGGGGCCAGCCGCCGGGGTTGGTTCT
hairpin
BWRF1
4,610
4,620
4,630
4,640
4,650
4,660
4,670
4,680
IR1 (the last unit)
GCCCCTCTCTCTGTCCTTCAGAGGAACCAGGGACCTCGGGCACCCCAGAGCCCCTCGGGCCCGCCTCCAGGCGCCCTC
hairpin
BWRF1
4,690
4,700
4,710
4,720
4,730
4,740
4,750
IR1 (the last unit)
CTGGTCTCCGCTCCCCTCTGAGCCCCGTTAAACCCAAAGAATGTCTGAGGGGAGCCACCCTCGGGGCCCAGGCCCCAG
BWRF1
4,760
4,770
4,780
4,790
4,800
4,810
4,820
4,830
IR1 (the last unit)
AGTCCAGAGGTCAGGGGCACCTCAGGGTGCCTCCCCGGGTCCCAGGCCAGCCGGAGGGACCCCGGCAGCCCGGGCGGC
BWRF1
4,840
4,850
4,860
4,870
4,880
4,890
4,900
4,910
IR1 (the last unit)
CCCAGAGGCCGGTTCCTCGCCCCTTCCCCGGGCTTCAGAGCCCAGGATGTCCCCCAGAAGGGACCCTAGGCGTCCCCT
BWRF1


4,920
4,930
4,940
4,950
4,960
4,970
4,980
4,990
IR1 (the last unit)
CTCCTCCCCTCCAGGCCCGAGCCTCTCCCTCGCGGAGAGGGGCCTCTTTGGGCCCTCAAGTCCAGCCCCACCGAGACC
BWRF1
5,000
5,010
5,020
5,030
5,040
5,050
5,060
5,070
IR1 (the last unit)
CGAGTGGCCCGGATCCCCCCACCGGCCCTTCTCTCTGTCCCCCTGCTCCTCTCCAACCTTCGCTCCACCCTAGACCCC
BWRF1
5,080
5,090
5,100
5,110
5,120
5,130
5,140
IR1 (the last unit)
AGCTTCTGGCCTCCCCGGGTCCACCAGGCCAGCCGGAGGGACCCCGGCAGCCCGGGCGAGTCGCCTTCCCTCTCCCCT
BWRF1
5,150
5,160
5,170
5,180
5,190
5,200
5,210
5,220
IR1 (the last unit)
GGCCTCTCCTTCCCGCCTCCCACCCGAGCCCCCTCAGCTTGCCTCCCCACCGGGTCCATCAGGCCGGCCGGAGGGACC
BWRF1
5,230
5,240
5,250
5,260
5,270
5,280
5,290
5,300
IR1 (the last unit)
CCGGCGGCCCGGTGTCAGTCCCCCCTGCAGCCGCCCAGTCTCTGCCTCCAGGCAAGGGCGCCAGCTTTTCTCCCCCCA
BWRF1
5,310
5,320
5,330
5,340
5,350
5,360
5,370
5,380
IR1 (the last unit)
GCCTGAGGCCCAGTCTCCTGTGCACTGTCTGTAAAGTCCAGCCTCCCACGCCCGTCCACGGCTCCCGGGCCCAGCCTC
BWRF1
5,390
5,400
5,410
5,420
5,430
5,440
5,450
5,460
IR1 (the last unit)
GTCCACCCCTCCCCACGGTGGACAGGCCCTCTGTCCACCCGGGCCATCCCCGCCCCCCTGTGTCCACCCCAGTCCCGT
BWRF1


5,470
5,480
5,490
5,500
5,510
5,520
5,530
IR1 (the last unit)
CCAGGGGGGACTTTATGTGACCCTTGGGCCTGGCTCCCCATAGACTCCCATGTAAGCCTGCCTCGAGTAGGTGCCTCC
BWRF1
5,540
5,550
5,560
5,570
5,580
5,590
5,600
5,610
IR1 (the last unit)
AGAGCCCCTTTTGCCCCCCTGGCGGCCCAGCCCGACCCCCGGGCGCCCCCAAACTTTGTCCAGATGTCCAGGGGTCCC
5,620
5,630
5,640
5,650
5,660
5,670
5,680
5,690
IR1 (the last unit)
CGAGGGTGAGGCCCAGCCCCCTCTCGCCCAAGCTGCTTTGATTCCTGGGATATTTTTGGGAATGGTGTTAACTTTCTC
5,700
5,710
5,720
5,730
5,740
5,750
5,760
5,770
CCCTTGTATTTGCTATTCAATCAACCTGATTCCCCCTGCTCATACCTCCACTTACAACCAAGCCACTACGGCCACGTC
Y1 exon
5,780
5,790
5,800
5,810
5,820
5,830
5,840
5,850
CCCGGCCTCCCGCTCGGGTAAGTGCTTTTTCATTTTTAGCCCCAGCCCCTCCTCTATAAGTTCTAGGCAAACCTCCAA
Y1 exon
5,860
5,870
5,880
5,890
5,900
5,910
5,920
TCACCAGCCACCTTCCAATGTAGTCTCTTAGAGAGTGGCTGCTACGCATTAGAGACCACTTTGAGCCACCCACAGTAA
Y2 exon
5,930
5,940
5,950
5,960
5,970
5,980
5,990
6,000
CCACCCAGCGCCAATCTGTCTACATAGAAGAAGAAGAGGATGAAGACTAAGTCACAGGCTTAGCCAGGTGATTTGTGA
Y2 exon


6,010
6,020
6,030
6,040
6,050
6,060
6,070
6,080
ATTTCAGTTTATTTACTTTCTTCCAATCAAGCTTTCCCAGCCTCCGCTTGTTAGGTCCTAGTTATGGGTTTTCCATGG
6,090
6,095
GGGACTTAGTA

Appendix 3. BWRF1 models

Strain	Model	No of Strains	length (bp)	in IR1#		Defining polymorphism/mutation
				start position	end position	
EBV_Saliva	model 1A	45	1542	151	1692	Default sequence
HKN14						
HL02						
sLCL-IS1.06						
sLCL-IS1.03						
sLCL-IS1.11						
sLCL-IS1.10						
sLCL-IS1.08						
sLCL-IS1.01						
HL08						
pLCL-TRL595						
sLCL-1.13						
sLCL-1.24						
sLCL-1.05						
sLCL-1.09						
sLCL-1.17						
sLCL-IS1.19						
sLCL-IS1.12						
sLCL-IS1.07						
sLCL-IS1.14						
sLCL-IM1.17						
sLCL-IS1.20						
sLCL-1.12						
sLCL-1.04						
sLCL-1.06						
sLCL-BL1.03						
sLCL-IM1.09						
sLCL-1.07						
sLCL-1.11						
sLCL-1.02						
sLCL-BL1.20						
sLCL-1.08						
YCCEL1						
Akata 2000						
sLCL-IM1.16						
C666.1						
HKN19						
sLCL-1.18						
D3201.2						
HKN15						
HL11						
Daudi						
K4413-Mi						
GD1						
Mutu						
Jijoye	model 1B	2	1407	287	1693	Insertion at 400 (GAGACAGGCAGGGGCCCCCCGG).
P3HR1						
K4123-Mi				288	1694	
sLCL-1.10	model 1C	1	1566	151	1716	Insertion at 1647 (CCATCCCCGCCCCCCC)

sLCL-IM1.05	model 1D	1	636 , 711 and 600	287, 567 and 923	922, 1277 and 1522	Insertion at 1219 (GATCCCCCCCACC) – same as 1X – and deletion at 587 (GAGGGG-ATCGCCC)

sLCL-IS1.15	model 1X	1	1128 and 600	151 and 924	1278 and 1523	Insertion at 1219 (GATCCCCCCCACC) - same as 2X

LCL B95-8 delEber2	model 2A	9	1152	541	1692	All model 2 strains differ from model 1 strains due to G->A SNP at 540 that creates a STOP codon (GTGGCCCTGGGGTAAGTCTGGGAGGCAGAG.
sLCL-IS1.04						
sLCL-IS1.18						
M-ABA						
BL37						
sLCL-IS2.01						
L591						
BL36						
B95-8						
X50-7						
Sitoko						
M81						
GD2						

pLCL-TRL1-post	model 2B	7	1149	541	1689	Sequence for type 2 viruses: Three single nucleotide deletions relative to 2A shorten the ORF but do not create STOP codon.
pLCL-TRL1-pre						
sLCL-2.14						
sLCL-2.21						
sLCL-2.22						
sLCL-2.15						
sLCL-2.16						
AG876						

HL09	model 2X	7	738 and 600	541 and 924	1278 and 1523	Insertion at 1219 (GATCCCCCCCACC) - same as 1X
HL05						
sLCL-IS1.13						
sLCL-IM1.02						
HL04						
Cheptages						
Raji						

Wewak1	model 2C	1	1122	566	1687	deletion at 663 (GGCCCCC-AGACA) - otherwise like model 2B

sLCL-1.19	model 2D	1	1008	667	1674	SNP at 664 -> stop codon (CAGCCCCCTAGAC) and 18 nt replaced with A at 1211-1229 (no frameshift) (TGA----------CC).

HL01	model 2E	1	1296	541	1836	Deletion at 1601 (CCACCC-TCCCC).

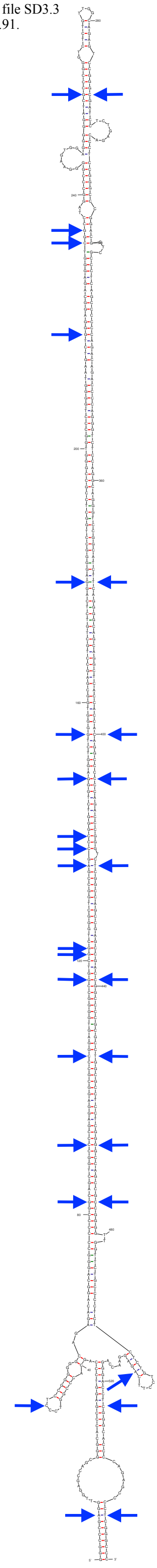
HKNPC1	model 2F	1	1140	541	1680	Deletion of 12 nt at position 1255-1266 (CCTGCTCCT----------CTCCACCCT)

Mak1 duplicate **Makau**	model 3A	1	1359	151	1509	An indel replacing 13nt at 1216-30 with A (GTGACCCGA----------CCCTTCTCT) causing a frameshift

*Strains in RED were excluded from quantitative analyses (mainly because they shared a common origin with other strains) are shown along with the strain that represents them.
*Strains in BLUE were reported by groups other than Palser et al. 2016 (see appendix 1)

Appendix 4. Secondary structure of the long hairpin within the IR1 of B95-8.

*Blue arrows shows positions where at least two strains have SNPs as shown in MSA file SD3.3
*Figure was produced by using the mfold web server (Zuker 2003),△G= -234.91.



Appendix 5. SNPs in the flanks of IR1.

No in group *	Flank SNP group #	strain	Position of Flank SNP in IR1-lastIR1 MSA (Data SD1)	Distance from IR1 start †	Distance from End of IR1 †	Flank Nucleotide (start of IR1) †	Internal Nucleotide
9(8)	a	pLCL-TRL1-post	4	4	1851	G	T
		pLCL-TRL1-pre					
		sLCL-2.14					
		sLCL-2.21					
		sLCL-2.22					
		sLCL-2.15					
		sLCL-2.16					
		sLCL-IS2.01					
		Wewak1					
4 (1)	b	B95-8	24	24	1831	G	T
		Sitoko					
		LCL B95-8 delEber2					
		BL36					
1		sLCL-IM1.09	38	38	1817	C	T
5	c	HL09	68	68	1787	C	A
		HL04					
		sLCL-IM1.02					
		sLCL-IM1.05					
		HL05					
2	d	sLCL-IM1.16	74	74	1781	T	C
		Wewak1					

1		Jijoye	82	82	1773	C	A
		P3HR1**					
		sLCL-IS1.06	112	112	1743	A	G
8	e	HL02	122	122	1733	G	A
		sLCL-IS1.18					
		sLCL-IS1.03					
		sLCL-IS1.11					
		sLCL-IS1.10					
		sLCL-IS1.08					
		sLCL-IS1.01					
		HL08					
1		sLCL-1.07	127	127	1728	T	C
1		sLCL-IS2.01	128	128	1727	G	T
4(1)	f	B95-8	129	129	1726	T	A
		Sitoko					
		LCL B95-8 delEber2					
		BL36					
		HKN14	150	150	1705	A	G
		C666.1	150	150	1705	G	A
7(6)	g	HL09	160	160	1695	T	C
		HL04					
		sLCL-IM1.02					
		sLCL-IM1.05					
		HKN14					
		Jijoye					
		P3HR1**					
1		C666.1	160	160	1695	C	T

2(1)		Jijoye P3HR1**	198	198	1657	T	C

		strain	Position of Flank SNP in IR1-lastIR1 MSA (Data SD1)	Distance from IR1 start	Distance from End of IR1	Flank Nucleotide (end of IR1) †	Internal Nucleotide
1		C666.1	1615	1615	240	T	G
2(1)		Mak_1 Makau	1725	1725	130	T	C
1		sLCL-1.08	1769	1769	86	T	C
2(1)		pLCL-TRL1-post pLCL-TRL1-pre	1790	1790	65	C	A
1		sLCL-IM1.02	1795	1795	60	G	T
1		sLCL-1.17	1806	1806	49	A	C
1		C666.1	1809	1809	46	G	T
2(1)		pLCL-TRL1-post pLCL-TRL1-pre	1839	1839	16	T	C
8	z	sLCL-1.04 sLCL-1.06 sLCL-1.11 sLCL-1.07 sLCL-BL1.20 sLCL-1.08 sLCL-1.02 sLCL-1.10	1840	1840	15	A	G

All positions are drawn from multiple sequence alignment SD1.

* Number of independent strains with this flank SNP, with parentheses around number of total samples sequenced showing the SNP.

Letter assigned where multiple strains share the same flank SNP. These isdentifiers are used in Figure 3.

† Distance from the edge of IR1 is shown according to whether the SNP is in the flank at the end of IR1 or the start of IR1. Flank base is that near the junction; internal nucleotide is the base call away from the junction.

** last IR1 does not represent the actual last unit in P3HR1.

Appendix 6. Distribution of MVs in viruses with the most interstrain recombination. †

Position #		Strain				
IR1 unit	SD1	pLCL-TRL1-pre	sLCL-IS2.01	C666.1	HL04	HL05
311	3388	T	T	T	T(A,11.88%,0.45)	T
326	3403	A(G,22.14%,0.66)	A	A	A	A
346	3423	A	A	A(G,35.03%,0.83)	A	A
380	3457	C(T,19.82%,0.64)	C	C	C	C
383	3460	G(T,19.54%,0.6)	G	G	G	G
411	3488	G	A(G,37.3%,0.68)	G	G	G
475	3552	G	G	G(A,13.22%,0.95)	G	G
503	3580	C	C	C	C(T,14.29%,0.69)	C
542	3619	A(C,15.73%,0.93)	A	C(A,14.55%,0.96)	A(C,13.67%,0.9)	A(C,22.3%,0.9)
560	3637	A(T,15.16%,0.85)	A	T	T	T
651	3728	C(T,38.85%,0.95)	C	C	T(C,15.09%,0.45)	T(C,23.48%,0.81)
653	3730	G	G	G	G(T,14.56%,0.44)	G(T,17.19%,0.59)
659	3736	G(A,40.99%,0.94)	G	A	A	A
705	3782	C(T,37.8%,0.94)	C	C	T(C,15.36%,0.88)	T(C,24.74%,0.56)
736	3813	C	C	C	C(A,13.38%,0.36)	C
768	3845	C(T,38.28%,0.85)	C	C	C	C
794	3871	A(G,40.73%,0.91)	G(A,21.61%,0.93)	G	G	G
802	3879	C(G,38.9%,0.91)	G(C,22.4%,0.94)	G	G	G
827	3904	C	C	C	C(T,16.56%,0.87)	C
840	3917	G	G	G	G(A,15.43%,0.84)	G
856	3933	G	G	G	G(T,13.82%,0.72)	G
857	3934	C	C	C	C(A,13.67%,0.78)	C
872	3949	C	C	C	C(T,16.45%,1.0)	C
881	3958	C	C	C	C(T,12.55%,0.99)	C
883	3960	C	C	C	C(T,12.32%,0.97)	C
884	3961	C	C	C	C(T,12.53%,0.99)	C
914	3991	G	T(G,16.65%,0.87)	T	G(A,16.75%,0.91)	G
929	4006	C	C(T,16.31%,0.8)	C	C	C
949	4026	C	C	C	C(T,17%,0.97)	C
954	4031	C	C	T(C,21.04%,0.8)	C	C
1051	4128	C	C	C(T,17.32%,0.78)	C(T,15.63%,0.81)	C
1145	4222	C	T	T(C,22.46%,0.8)	C	C
1177	4254	C(T,18.6%,0.73)	T(C,18.68%,0.95)	T	T	T
1179	4256	A(G,18.95%,0.77)	G(A,18.6%,0.93)	G	G	G
1187	4264	A	C(A,20.59%,0.96)	C(A,18.79%,0.8)	A	A
1194	4271	T(C,20.94%,0.86)	C(T,20.22%,0.98)	C	C	C
1228	1228	A	A	A	G(A,16.16%,0.88)	G

1229	1229	C	C	C(G,18.74%,0.89)	C	C
1244	1244	T(G,16.95%,0.79)	G(T,18.14%,0.97)	G	G	G
1252	1252	A(G,20.62%,0.74)	G(A,18.73%,0.94)	G	G	G
1269	1269	G	G	G	A(G,18.54%,0.93)	A
1398	1398	C	C	C(T,16.54%,0.93)	C	C
1457	1457	C	C	C	G(C,18.11%,0.96)	C
1461	1461	T(C,20.74%,0.91)	T	C(T,14.86%,0.8)	C	C
1586	1586	A(G,19.1%,0.74)	A	G	G	G
1640	1640	G(C,14.92%,0.48)	G	G	G	G
2530	2530	C	C(T,27.73%,0.96)	C	C	C
2622	2622	G	G	G(C,22.44%,0.89)	G	G
2951	2951	G	T(G,29.88%,0.98)	G	G	G
3049	3049	C	C	C(T,11.96%,0.75)	C	C
Groups of linked SNP-like MVs: positions in repeat unit (No of SNP MVs; MV frequency; nucleotides spanned)		380-560 (4; 13-18%; 181 nt)	794-1252 (9; 16-23%; 459 nt)	542 (1; 14%)	542-736 (5; 12-17%; 195 nt)	542-736 (4; 17-28%; 195 nt)
		651-802 (6; 38-41%; 152 nt)		954-1461 (7; 14-23%; 508 nt)	827-884 (7; 10-16%; 58 nt)	
		1177-1640 (8; 15-21%; 464 nt)			914-1269 (4; 16-19%; 356 nt)	
		1261	459	920	728	195

Mutation MVs:	1	3	5	10	0
SNP MVs	18	9	7	10	4
Rare SNP MVs	0	1	0	1	0
Total SNP MVs	18	10	7	11	4
Total MVs	**19**	**13**	**12**	**21**	**4**

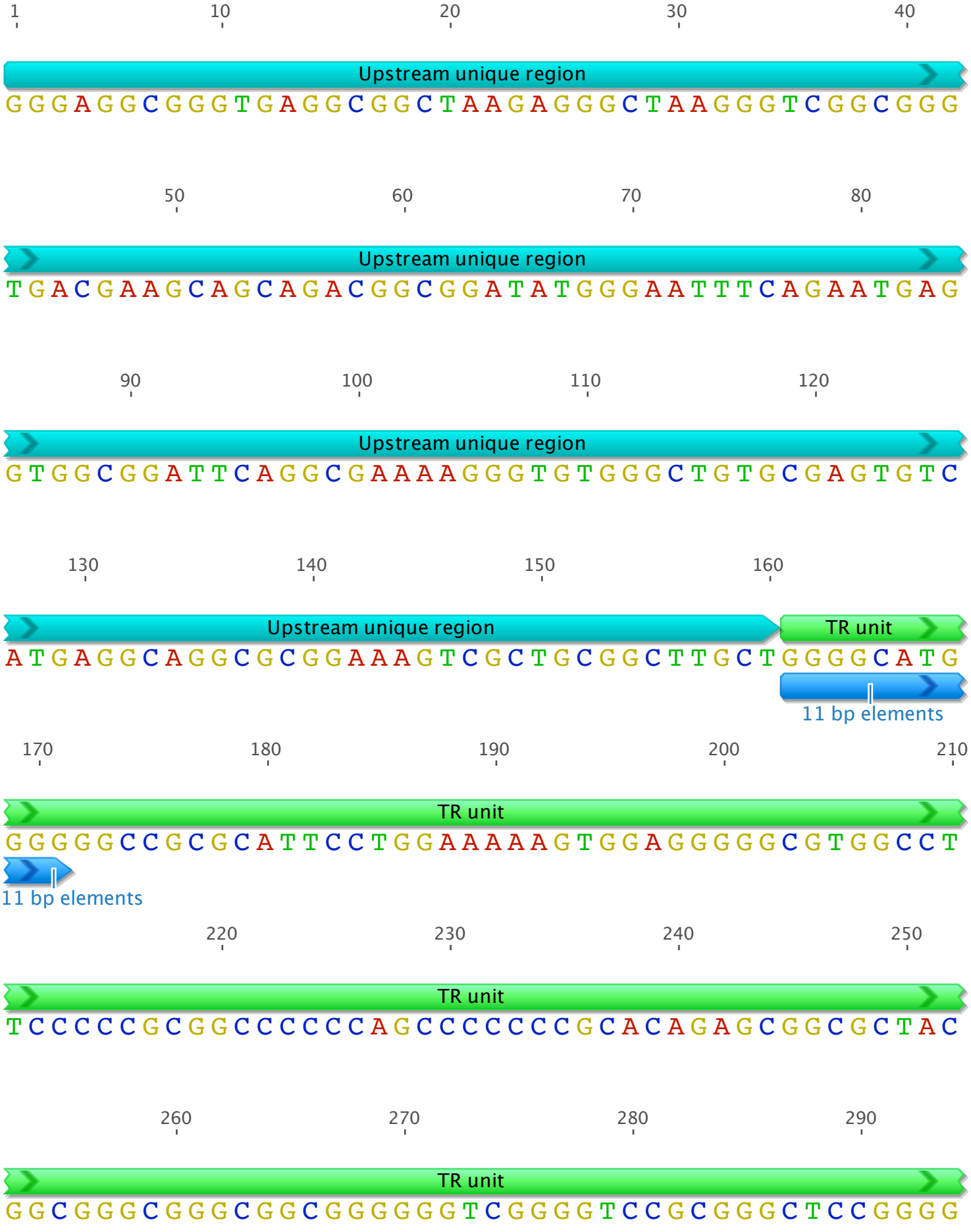
† Yellow coloured positions are those at which there are SNPs; pale yellow represents positions where there is a SNP in only one strain (Rare SNP). Uncoloured are putative spontaneous mutations.

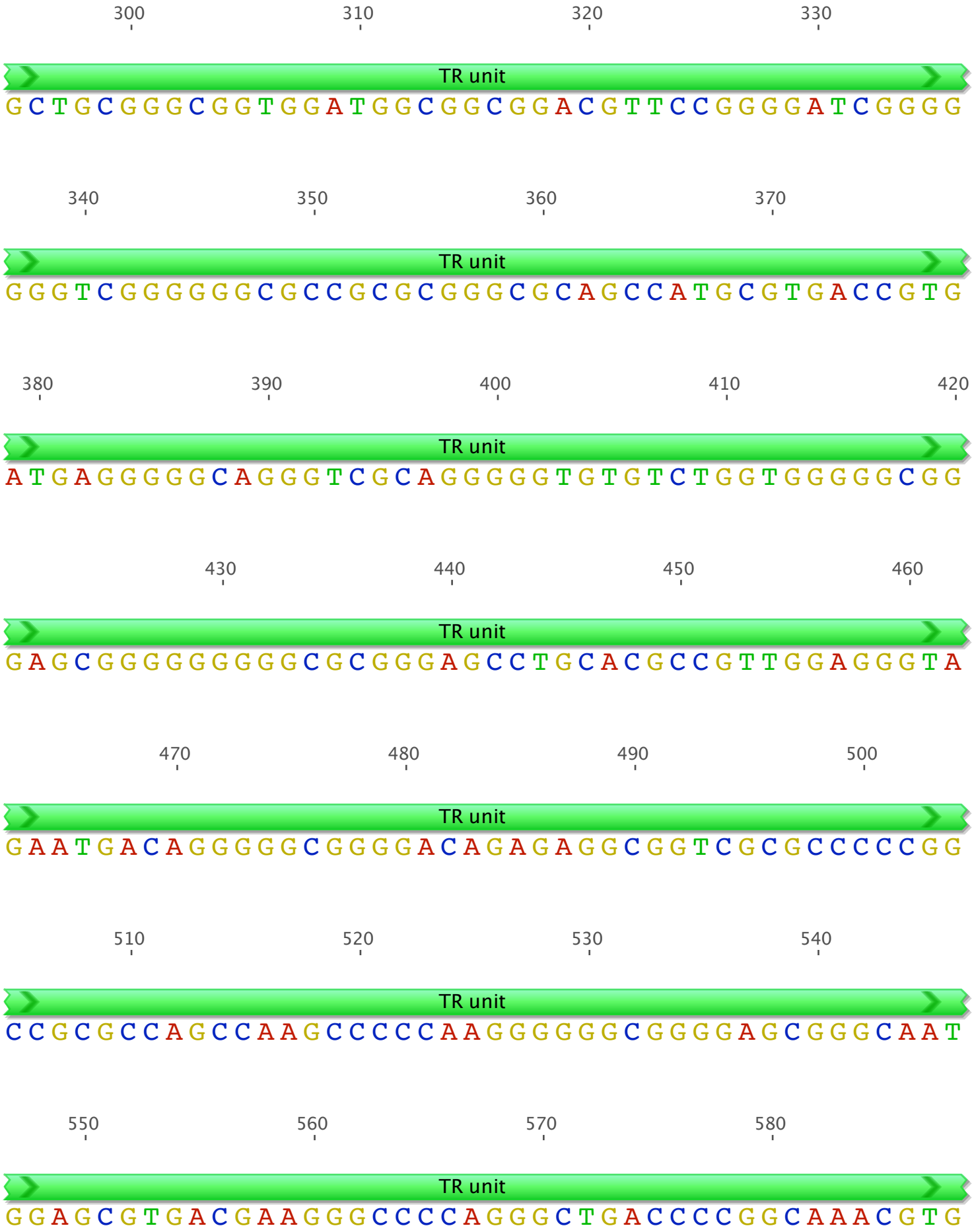
Position is shown in terms of both relative to a single IR1 repeat unit (left column), and the actual position within the multiple sequence alignment (Supplementary data file SD1) from which the MV data were extracted, as shown in supplementary table ST4 (right column).

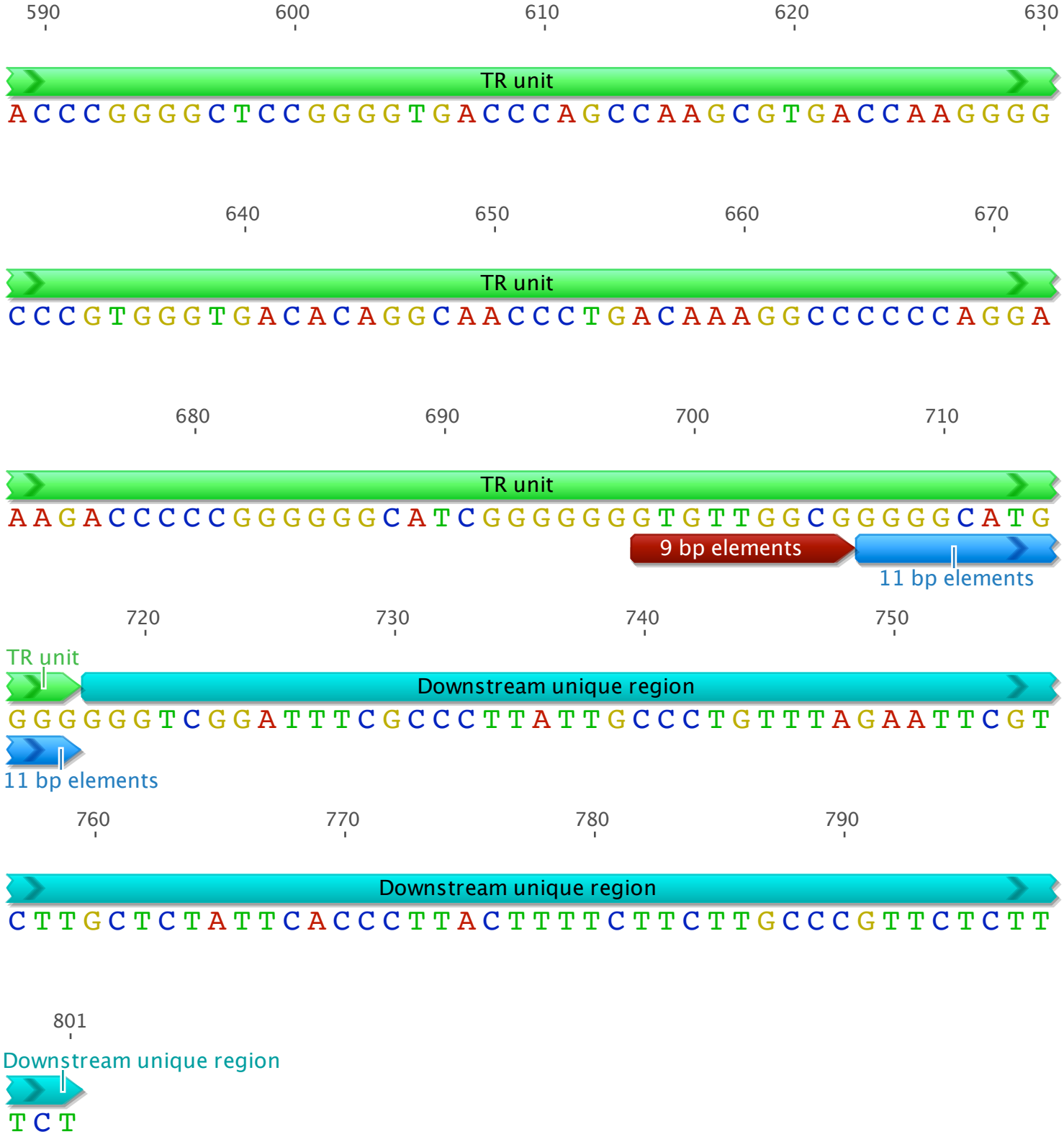
Appendix 7. Common SNPs in B95-8 and BL36

Notes	Nucleo -tide	Position in NC007605	#strains (+ B95-8, DelEbr2 and BL36)	Strain 1	Strain 2	Strain 3
there is no difference from 10024 up to IR1 region	A	4690	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	4823	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	5064	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	6182	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	A	9062	1	sLCLIS1.04_2116		
	T	10751	1	sLCLIS1.04_2116		
1.6 IR1 unit	T	12129	0			
	T	***12778***	1	sLCLIS1.04_2116		
	T	***13530***	1	sLCLIS1.04_2116		
	T	***13593***	0			
	T	12024	1	HL01_3274		
	T	***12778***	1	sLCLIS1.04_2116		
	T	***13530***	1	sLCLIS1.04_2116		
	T	***13593***	0			
There is no difference between B95-8 and BL36 at all from the end of IR1 region up to 43611 (in NC007605). If I exclude "Ns" and gaps the length will be 5955bp	T	35356	1	sLCLIS1.04_2116		
	G	36703	2	sLCLIS1.04_2116	HL01_3274	
	C	36710	2	sLCLIS1.04_2116	HL01_3274	
	T	37325	0			
	G	38034	0			
	G	38035	3	sLCLIS1.04_2116	sLCLIS1.18_2113	HL01_3274
	T	38083	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	C	40214	3	pLCLTRL595_2117	sLCL1.09_3236	sLCL1.18_3245
	T	41822	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	41823	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	41835	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	41855	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	41856	3	sLCLIS1.04_2116	sLCLIS1.18_2113	Jijoye_2135
	T	41857	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
	T	41904	2	sLCLIS1.04_2116	sLCLIS1.18_2113	
within 105 nt, there are 5 SNPs exist only in B95-8, DelEber2, sLCL-IS1.104 and BL36. their positions in NC007605: A:9062, A:9097, C:9137, A:9158, and T:9166. only the first SNP is mentioned in the table because other strains have one or two of the other SNPs but no strain has all five SNPs except these four strains.						
cluster (this region also includes SNP at 27893 but it exists in >15 strains)						

Appendix 8. DNA sequence with annotations of the B95-8 TR theoretical template




300
310
320
330
TR unit
GCTGCGGGCGGTGGATGGCGGCGGACGTTCCGGGGATCGGGG
340
350
360
370
TR unit
GGGTCGGGGGGCGCCGCGCGGGCGCAGCCATGCGTGACCGTG
380
390
400
410
420
TR unit
ATGAGGGGGCAGGGTCGCAGGGGGTGTGTCTGGTGGGGGCGG
430
440
450
460
TR unit
GAGCGGGGGGGGGCGCGGGAGCCTGCACGCCGTTGGAGGGTA
470
480
490
500
TR unit
GAATGACAGGGGGCGGGGACAGAGAGGCGGTCGCGCCCCCGG
510
520
530
540
TR unit
CCGCGCCAGCCAAGCCCCCAAGGGGGGCGGGGAGCGGGCAAT
550
560
570
580
TR unit
GGAGCGTGACGAAGGGCCCCAGGGCTGACCCCGGCAAACGTG


590
600
610
620
630
TR unit
ACCCGGGGCTCCGGGGTGACCCAGCCAAGCGTGACCAAGGGG
640
650
660
670
TR unit
CCCGTGGGTGACACAGGCAACCCTGACAAAGGCCCCCCAGGA
680
690
700
710
TR unit
AAGACCCCCGGGGGGCATCGGGGGGGTGTTGGCGGGGGCATG
9 bp elements
11 bp elements
720
730
740
750
TR unit
Downstream unique region
GGGGGTCGGATTTCGCCCTTATTGCCCTGTTTAGAATTCGT
11 bp elements
760
770
780
790
Downstream unique region
CTTGCTCTATTCACCCTTACTTTTCTTCTTGCCCGTTCTCTT
801
Downstream unique region
TCT

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology®

GENETIC DIVERSITY AND EVOLUTION

Heterogeneity of the Epstein-Barr Virus (EBV) Major Internal Repeat Reveals Evolutionary Mechanisms of EBV and a Functional Defect in the Prototype EBV Strain B95-8

Mohammed M. Ba abdullah,[a] Richard D. Palermo,[a] Anne L. Palser,[b]* Nicholas E. Grayson,[b] Paul Kellam,[a,b,c] Samantha Correia,[a] Agnieszka Szymula,[a]* Robert E. White[a]

Section of Virology, Imperial College Faculty of Medicine, St. Mary's Hospital, Norfolk Place, London, United Kingdom[a]; Wellcome Trust Sanger Institute, Hinxton, Cambridge, United Kingdom[b]; Kymab, Babraham Research Campus, Cambridge, United Kingdom[c]

ABSTRACT Epstein-Barr virus (EBV) is a ubiquitous pathogen of humans that can cause several types of lymphoma and carcinoma. Like other herpesviruses, EBV has diversified through both coevolution with its host and genetic exchange between virus strains. Sequence analysis of the EBV genome is unusually challenging because of the large number and lengths of repeat regions within the virus. Here we describe the sequence assembly and analysis of the large internal repeat 1 of EBV (IR1; also known as the BamW repeats) for more than 70 strains. The diversity of the latency protein EBV nuclear antigen leader protein (EBNA-LP) resides predominantly within the exons downstream of IR1. The integrity of the putative BWRF1 open reading frame (ORF) is retained in over 80% of strains, and deletions truncating IR1 always spare BWRF1. Conserved regions include the IR1 latency promoter (Wp) and one zone upstream of and two within BWRF1. IR1 is heterogeneous in 70% of strains, and this heterogeneity arises from sequence exchange between strains as well as from spontaneous mutation, with interstrain recombination being more common in tumor-derived viruses. This genetic exchange often incorporates regions of $<$1 kb, and allelic gene conversion changes the frequency of small regions within the repeat but not close to the flanks. These observations suggest that IR1—and, by extension, EBV—diversifies through both recombination and breakpoint repair, while concerted evolution of IR1 is driven by gene conversion of small regions. Finally, the prototype EBV strain B95-8 contains four nonconsensus variants within a single IR1 repeat unit, including a stop codon in the EBNA-LP gene. Repairing IR1 improves EBNA-LP levels and the quality of transformation by the B95-8 bacterial artificial chromosome (BAC).

IMPORTANCE Epstein-Barr virus (EBV) infects the majority of the world population but causes illness in only a small minority of people. Nevertheless, over 1% of cancers worldwide are attributable to EBV. Recent sequencing projects investigating virus diversity to see if different strains have different disease impacts have excluded regions of repeating sequence, as they are more technically challenging. Here we analyze the sequence of the largest repeat in EBV (IR1). We first characterized the variations in protein sequences encoded across IR1. In studying variations within the repeat of each strain, we identified a mutation in the main laboratory strain of EBV that impairs virus function, and we suggest that tumor-associated viruses may be more likely to contain DNA mixed from two strains. The patterns of this mixing sug-

Received 5 June 2017 **Accepted** 2 September 2017

Accepted manuscript posted online 13 September 2017

Citation Ba abdullah MM, Palermo RD, Palser AL, Grayson NE, Kellam P, Correia S, Szymula A, White RE. 2017. Heterogeneity of the Epstein-Barr virus (EBV) major internal repeat reveals evolutionary mechanisms of EBV and a functional defect in the prototype EBV strain B95-8. J Virol 91:e00920-17. https://doi.org/10.1128/JVI.00920-17.

Editor Richard M. Longnecker, Northwestern University

Copyright © 2017 Ba abdullah et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Robert E. White, robert.e.white@imperial.ac.uk.

* Present address: Anne L. Palser, Kymab, Babraham Research Campus, Cambridge, United Kingdom; Agnieszka Szymula, Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

gest that sequences can spread between strains (and also within the repeat) by copying sequence from another strain (or repeat unit) to repair DNA damage.

KEYWORDS B95-8, DNA sequencing, EBNA-LP, Epstein-Barr virus, viral evolution, genome analysis, human herpesviruses, internal repeat, virus mutation

Epstein-Barr virus (EBV) is a human herpesvirus that infects the vast majority of the human population. Usually, such infection is asymptomatic, but EBV can cause a number of malignancies, including both lymphomas—Burkitt's lymphoma (BL), Hodgkin's lymphoma (HL), and immunoblastic lymphomas, such as posttransplant lymphoproliferative disease (PTLD) or diffuse large B cell lymphoma (DLBCL)—and carcinomas, such as nasopharyngeal carcinoma (NPC) and gastric cancer (GC). While EBV-associated disease is relatively uncommon in the immunocompetent, the nearly ubiquitous nature of EBV infection makes it a considerable worldwide health burden (1).

Many of these cancers are geographically restricted: EBV-associated BL is prevalent predominantly in malarial regions of Africa and Asia/Australasia, the incidence of NPC is dramatically higher in southern China and nearby regions of Southeast Asia and unevenly distributed elsewhere around the world (2), and the prevalence of EBV-GC (compared to total GC) is higher in the Americas (3). While the global asymmetry in BL incidence is attributed to malarial coinfection, the reasons for other geographic variations in the incidence of EBV-associated diseases remain unclear. They may relate to environmental differences, behavioral or cultural factors, the genetics of the population, differences between virus strains, or some interaction between these possible causes. Sequences of EBV genes/genomes vary geographically around the world (4), likely reflecting the close host-pathogen relationship exemplified by the cospeciation typical of herpesviruses (5). Two major subgroups of EBV—type 1 and type 2—have been defined according to their EBNA2 and EBNA3 gene sequences. Biologically, type 2 viruses transform B cells less efficiently (6) and were recently shown to also transform T cells (7). However, the diversity of the main viral oncogenes (EBNA2, EBNA3s, and LMP1) has not yet revealed any clear associations with disease beyond supporting this geographic diversity (8). In recent years, viral genome sequencing has become a sufficiently cost-effective proposition to allow these ideas to be revisited.

When it was published in 1984, the sequence of the B95-8 strain of EBV was the longest contiguous DNA sequence known (9). This was updated to provide a more representative type 1 EBV genome (a B95-8/Raji hybrid) and a type 2 (AG876) virus genome (10, 11). However, the advent of short-read sequencing technology has resulted in the more widespread sequencing of various herpesvirus genomes, notably those of cytomegalovirus (CMV) (12, 13), herpes simplex viruses 1 and 2 (HSV-1 and HSV-2) (14–17), varicella-zoster virus (VZV) (18), and EBV (4, 19–22). The consensus approach for assembling herpesvirus genomes comprises *de novo* contig assembly followed by gap-filling approaches and genome assembly driven by known consensus genome structures. This assembly can then be annotated as a framework to address biological questions that arise from the genome sequence. This approach is exemplified by the VirGA protocol (applied to HSV-1) (15), but similar approaches have been followed for CMV (13) and EBV (4).

One of the major challenges for genome sequencing—particularly that using short-read libraries—is the accurate assembly of repetitive regions. Many viruses contain repetitive regions, particularly at their termini. Of the human herpesviruses, EBV contains perhaps the most repeat regions, yet sequencing the repeats is both troublesome and important, as many of these regions are replication origins or encode proteins (or parts of proteins) that play major roles in virus biology, particularly in viral latency and persistence (Fig. 1A). Accurately assembling these regions remains the largest barrier to producing complete EBV genomes: a recent publication of 71 virus genomes blanked out over 20 repeat regions to facilitate comparisons between the strains (4). Current viral genomes have often been obtained by use of Sanger sequencing to bridge these

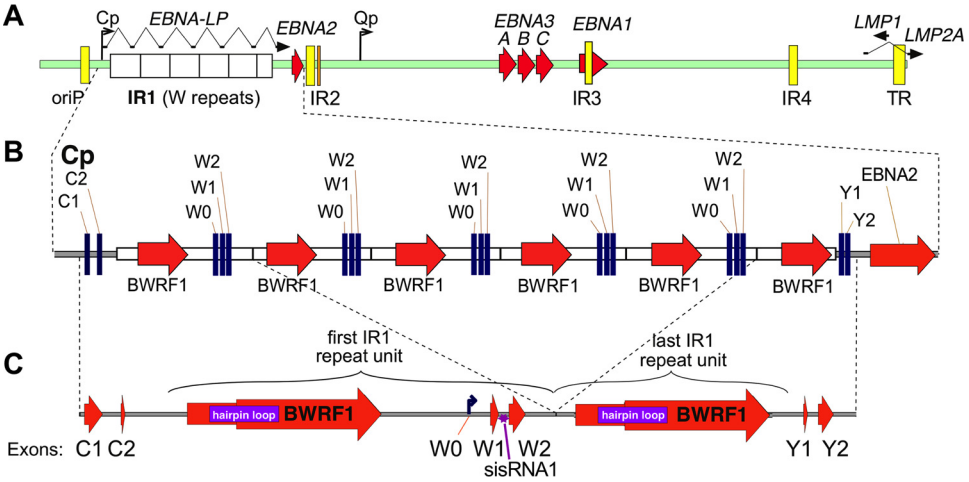


FIG 1 Schematic representations of the IR1 region of EBV. (A) Schematic representation of the EBV genome, showing IR1, composed of the typical 5.6 repeat units (white boxes), as well as the other major repeats of EBV (internal repeats [IR] 2 to 4, FR of *oriP*, and TR; yellow boxes), with latency genes shown as ORFs (red arrows) or spliced transcripts (black arrows) and the major latency promoters (Qp and Cp) shown as bent arrows. (B) Schematic view of the IR1 region with its flanks. (C) Structure of the simplified IR1 template (1.6 repeat units) against which reads were mapped. The symbol for BWRF1 (large red arrow) is stepped to represent the two major forms of the BWRF1 ORF: the left end of the arrow indicates the longer form (previously reported for AG876), and the widened arrow indicates the first position at which the shorter (B95-8) isoform might start. A purple box indicates the large hairpin sequence in IR1. Latent transcript exons (W1, W2, C1, C2, Y1, and Y2) are indicated, with exon W0 being the exon initiated at the Wp promoter, which lies within IR1.

gaps, and more recently, long-read technology (PacBio) was used to sequence across the EBV repeats in two bacterial artificial chromosome (BAC)-cloned viruses (23). However, even these methods struggle to resolve many of the EBV repeats due to their large size and complexity.

In order to develop a methodology for more systematically extracting the sequences of repetitive regions from short-read data, we chose to resolve and analyze internal repeat 1 (IR1; also known as the BamW repeats), the largest of EBV's repeat regions. IR1 is typically composed of 5.6 to 8.6 copies of a 3-kb repeat unit (24). Upstream of the repeat lies the major latency promoter (Cp) and the first two exons of the extensively spliced EBV nuclear antigen (EBNA) transcript (exons C1 and C2). Each repeat unit hosts a pair of exons (W1 and W2) after the immediate early latency promoter (Wp), which initiates transcription at exon W0. Exons W1 and W2 encode the repeat domain of EBNA leader protein (EBNA-LP), whose nonrepetitive C terminus is encoded by the Y exons immediately downstream of IR1. A complex and incompletely characterized pattern of alternative splicing and alternative promoter usage across IR1 and beyond dictates the transcription of the six EBNA genes during latency (25), as well as the expression of the antiapoptotic protein BHRF1 during the lytic cycle (26). In addition to the transcribed exons, B95-8 IR1 contains an extended open reading frame (ORF), BWRF1, that lacks a start codon and promoter (9, 27, 28). At the 5′ end of BWRF1 is an extended hairpin of approximately 500 bases (27). There is no direct evidence yet that either BWRF1 or the structured region is biologically significant. It is known that some component(s) of IR1 is essential for the transformation of cells by EBV, as viruses with 1.6 or fewer repeats are completely unable to transform B cells, while more than 4.6 repeats are required for maximal transformation efficiency (24). This deficiency is attributed to the reduced transcription of EBNA-LP and EBNA2 caused by having fewer Wp promoters.

Here we aimed to develop a strategy for both assembling and analyzing EBV repeat regions. In addition to identifying some idiosyncrasies of individual viruses, including a defect in the main lab strain B95-8, this analysis of the sequence characteristics and diversity of IR1 defined conserved regions and characterized the diversity of EBNA-LP and BWRF1 sequences. In addition, IR1 is often heterogeneous within each strain, and this heterogeneity probably arises through both spontaneous mutation and sequence

exchange between viruses. By treating the repeats as an internally controlled sequence, we can infer that interstrain genetic exchange may be more prevalent in tumor-derived strains, and we speculate that genetic exchange in IR1, both within and between viruses, can arise through the repair of DNA damage.

RESULTS

Templated assembly of IR1 shows that deletion breakpoints cluster in IR1. Our previous assembly of 71 EBV genomes produced a set of aligned genomes in which the repeat regions remain unresolved and are replaced by blank sequence (4). This approach was taken because the number and complexity resulting from variations in repeat structure and length, combined with relatively short reads (76-nucleotide paired-end reads), could not reliably be built by the available *de novo* assemblers, and the repeats disrupted genome alignment and subsequent ORF analysis. IR1 typically comprises 4 to 7 complete repeat units (Fig. 1B), containing BWRF1 and Wp regions, and a downstream partial repeat, i.e., the BWRF1 part of the repeat unit (4, 24). To simplify analysis of this region, a template was generated (based on a previous sequence and mapping of reads to a standardized IR1 repeat [see Materials and Methods]) to comprise one complete and one partial repeat unit flanked by the unique regions of the virus genome that exist either side of IR1 (Fig. 1). The reads were mapped to this template to confirm the consensus sequence of IR1 for each strain, and these consensuses were compared in a multiple-sequence alignment.

Overall, we found the junctions between IR1 and the unique region of the genome to be structurally conserved, except in a few strains which exhibit large deletions of the flanks of IR1. For instance, the P3HR1 and Daudi strains have previously been reported to contain deletions of the downstream junction between IR1 and the unique sequence (29–31), while X50-7 has a deletion at the C promoter end of IR1 (32). We found that these deletions are sometimes missed in cases where there is a large read depth and sufficient cross-contaminating reads to span the deletion. These erroneous reads may arise from sample cross-contamination or from barcode exchange and "jumping PCR" during library preparation and sequencing, which can be avoided by double barcoding (33). These hidden deletions are still characterized by a sudden change in read depth at each end, so we devised a read depth analysis (RDA; compares the numbers of reads mapped to adjacent nucleotides across the template to identify sudden changes in read depth) to identify the locations of large deletions. The read depth analysis for X50-7 (Fig. 2A) illustrates how this effectively identifies the junction of a mismapped deletion in unique regions of the genome and shows that this signal is much weaker in repetitive regions and that erroneous signals occur at the junctions between unique and repeat sequences. This approach corrected a falsely identified deletion in the original HL01 sequence and identified novel deletions in L591 and at one end of the Cheptages IR1, in addition to confirming the known deletions in P3HR1, Daudi, and X50-7. The deletion in the Cp end of IR1 in strain Cheptages occurred in two repeats (IR1 and the family of repeats of oriP) and thus could be resolved only by visually scanning the sorted mapped reads rather than by RDA, which loses the signal in repeats. The deletion junctions are summarized in Table 1 and shown schematically in Fig. 2B.

Curiously, all 6 of the breakpoints in the 5 strains with internal deletions lay within a $<$1-kb region upstream of the hairpin and downstream of the W0 exon. There were no strains with breakpoints in the hairpin, BWRF1, or Wp. The odds of all 6 independent breakpoints clustering within this 1-kb part of the repeat are around 1 in 700, and the odds of 6 breakpoints clustering in any 1-kb stretch can be approximated to 1 in 3^5, giving a *P* value of $<$0.005. Whether a biological process (such as different frequencies of breakpoint generation or selection for some function of the BWRF1 region that does not tolerate partial copies) influences this bias is not clear.

Conserved regions of IR1 include Wp, W2, and regions of BWRF1 of unknown function. After aligning the consensus sequences of IR1 for newly assembled and previously published EBV strains (85 strains in total) (see Supplemental Data SD2 in the

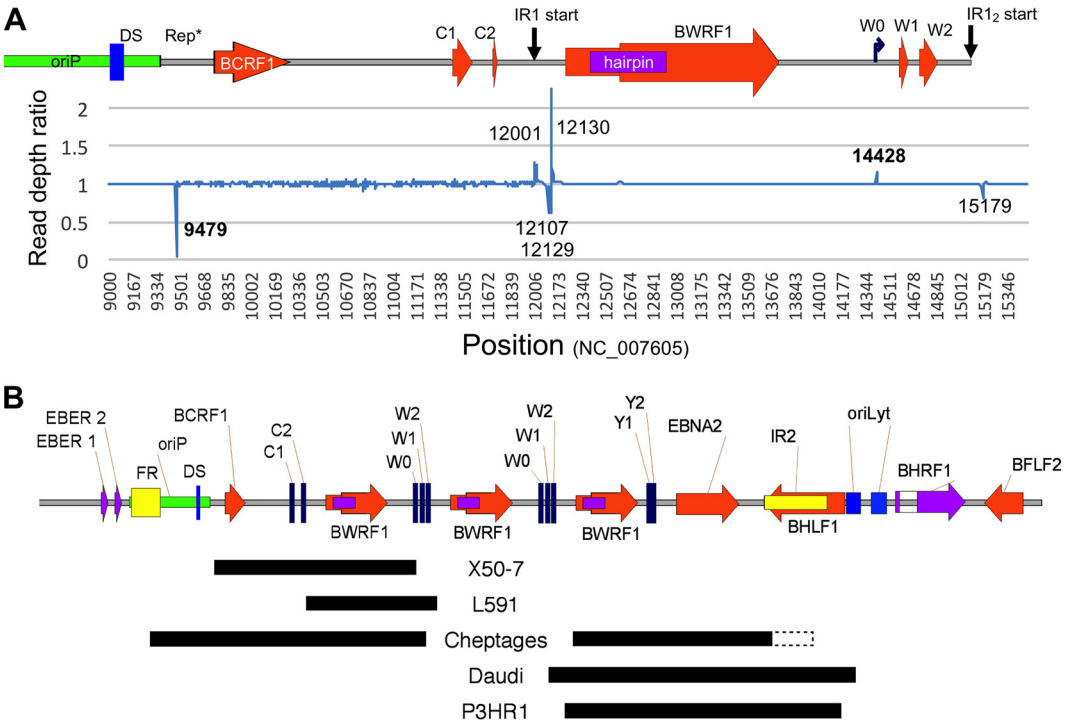


FIG 2 Characterization of strains with deletions affecting IR1. (A) Read depth analysis. The graph shows read depth ratios at adjacent nucleotides from X50-7 reads mapped against the prototype EBV genome (accession number NC_007605) on the *y* axis, with the EBV genome position shown on the *x* axis. The schematic above the graph shows positions of genomic features. Numbers in the graph show positions for the RDA signal peaks, and those in bold show the edges of the X50-7 deletion. (B) Schematic representation of deletions that include parts of IR1. A black box indicates the position of the deletion; a dotted box indicates uncertainty over the position of the deletion, as the junction lies within a repetitive region. Positional details for these deletions are found in Table 1. For scale, note that there are 3,072 bp between identical features in the repeat.

supplemental material), the degree of diversity between strains was assessed. Variations in the length of the repeat unit are largely restricted to the following two regions: positions 13210 to 13240 within BWRF1 (which has three distinct variant length genotypes and in which two strains [Makau and sLCL1.19] have independent >10-bp deletions) (accession number NC_007605.1) and the region between exons W0 and W1 (positions 14450 to 14500), which has two distinct single-base indels.

Analyzing the distribution of indels and single nucleotide polymorphisms (SNPs) showed considerable variation in the degree of sequence conservation across the repeats (Fig. 3A). The polymorphism densities in IR1 ranged from 0 to 12 SNPs/100 bp. The Wp promoter (upstream of and including exon W0) is highly conserved, as is a region encompassing the W2 exon and the adjacent intronic sequence, which is processed into a stable RNA (EBV-sisRNA1) (34). Other highly conserved regions were identified immediately upstream of the IR1 hairpin (positions 12225 to 12350) and

TABLE 1 Deletions overlapping the flanks of IR1

	Position of deletion (nearest EBV genome feature)[a]			
	Upstream (BamC)		Downstream (BamY)	
Strain	Upstream	IR1	IR1	Downstream
X50-7	9479 (in oriP rep* region)	14428 (after exon W0)		
L591	11632 (in exon C2)	12193 (50 bp before hairpin)		
Cheptages	7777 (in oriP FR)	14669 (in W1-W2 intron)	33604 (between W2 and hairpin)	38301[b] (within IR2)
Daudi			33127 (W1-W2 intron)	40536 (oriLyt upstream element)
P3HR1			33351 (after exon W2)	40157 (nr BHLF1 start)

[a]Positions in the prototype sequence NC_007605 that are equivalent to the first and last bases of the deletion are given.
[b]First position within the repeat; the sequence recurs every 125 bp, so the precise location of the junction is uncertain.

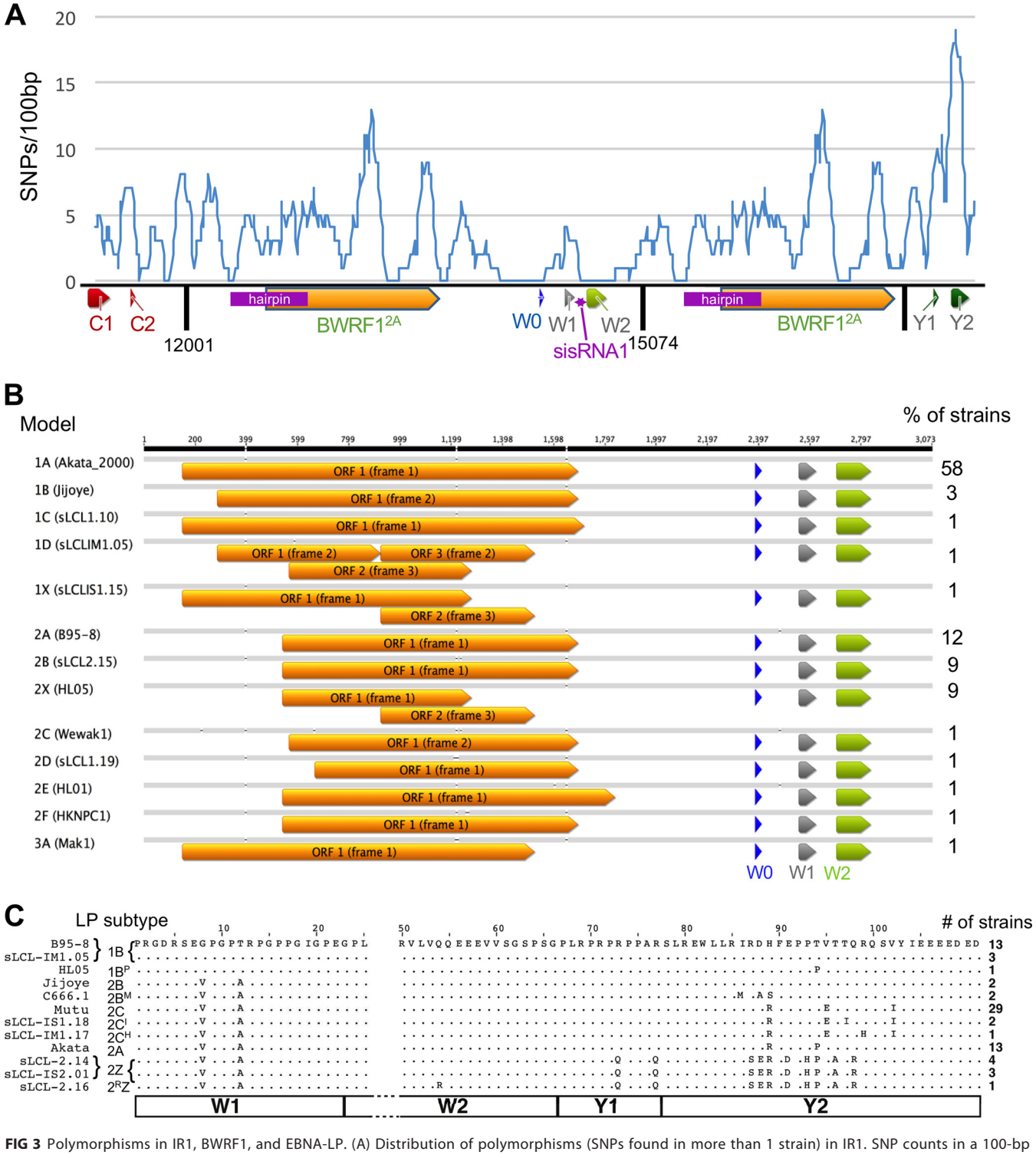


FIG 3 Polymorphisms in IR1, BWRF1, and EBNA-LP. (A) Distribution of polymorphisms (SNPs found in more than 1 strain) in IR1. SNP counts in a 100-bp sliding window are shown against their positions in the IR1 template. In the schematic, arrows indicate positions of exons/ORFs, and dark vertical lines represent edges of IR1 repeat units. (B) Visualization of all ORFs of over 600 nucleotides across the BWRF1 region of IR1. Since there are no AUG start codons in most of the BWRF1 region, ORFs are defined at their maximum possible length (i.e., maximum distance between stop codons). The frequency of each model (percentage) is shown to the right. (C) Amino acid sequence variants of the EBNA-LP protein encoded by the viruses. All sequences are shown relative to the B95.8 consensus (dots indicate no change). The protein subtype according to the nomenclature proposed here (with the commonest subtypes in bold) (Table 2) and the name of an example strain are given to the left. The number of unrelated strains (out of 74 total strains) encoding each variant is shown to the right (see Table ST1 in the supplemental material for strain-specific information). Numbers at the top indicate amino acid positions (according to a W1W2Y1Y2 exon structure), with amino acids 26 to 49 omitted (gap in sequence) because they are identical in all strains. The exon structure is shown by the boxes at the bottom of the panel.

within the final third of the BWRF1 ORF (positions 13290 to 13450), immediately downstream of the indels in Makau and sLCL1.19 described above.

The main annotated features of IR1 are fairly well conserved but are more diverse than previously reported. The 512-nucleotide hairpin region (whose ends are defined by a repeated CGGGCACCC motif [27]) appears to be retained in all of the virus strains analyzed. The SNP frequency across this region is moderate, and its conservation was assessed by aligning the hairpin to its own reverse complement (Supplemental Data SD3). There are no examples of polymorphisms on one strand of the hairpin being matched by a complementing polymorphism on the other strand. However, there are several SNPs that are predicted to change G-T into G-C base pairs on the transcribed strand, as well as changes in mismatched bases that retain the mismatch. Overall, the hairpin is retained within IR1 in all strains, with only slight variations, suggesting that it offers some sort of selective advantage for EBV. However, there is no indication of important structures based on coconservation of bases.

A second major feature of IR1 is BWRF1, the large, uncharacterized open reading frame with no canonical start codon that partly overlaps the hairpin in the second reportedly stable intron of IR1 (EBV-sisRNA-2) (34). Since BWRF1 has no ATG to define a start codon, we defined the ORF length as the distance between in-frame stop codons within IR1. These ORFs fall into two main groups (models 1 and 2) (Fig. 3B; Table ST2), dependent on a TAA-to-TAC polymorphism at position 12540 (accession number NC_007605.1) defining two different lengths of the ORF. The smaller group (model 2), with a shorter (384 codons) ORF, includes B95-8 (model 2A) and the type 2 viruses (model 2B), while about 60% of strains contain a much longer (514 codons) ORF that also spans the hairpin (model 1A). However, 9 of the 78 distinct strains contain an indel causing a frameshift mutation that truncates BWRF1 (models 1X, 2X, and 1D, where "X" indicates the disrupted ORF). This suggests that, at least for these strains, the presence of an intact full-length BWRF1 ORF is not essential for the virus life cycle. There are also other individual strains carrying indels that change the potential start and end positions of BWRF1, but all of these retain an ORF of at least 1,000 nucleotides. The high GC content (72.6%) of IR1 reduces the frequency of stop codons, meaning that extended ORFs are more likely to occur randomly than the case in a lower-GC environment. Nevertheless, the chance of an ORF of over 380 codons (seen in 86% of strains) appearing in this position by chance is still less than 1%, so it seems likely that there is some selection pressure to retain the BWRF1 ORF in most strains or to retain certain regions of the BWRF1 ORF in all strains.

EBNA-LP is encoded across the W repeat region, combining repeating exon pairs from IR1 (typically 2 to 7 copies of exons W1 and W2) with a nonrepetitive C-terminal domain (encoded by exons Y1 and Y2). It is among the first viral proteins translated after infection, during the establishment of latency. The diversity of EBNA-LP sequences is summarized in Fig. 3C, and a new nomenclature for EBNA-LP subtypes is proposed in Table 2.

Previous analyses of EBNA-LP sequences identified two distinct protein isoforms of EBNA-LP, based on the presence of G8/T12 or V8/A12 in exon W1 (35). These have been called type 1 and type 2 variants, respectively, even though the so-called type 2 variant is found in many type 1 strains. Our analysis confirms that these are the two W1 exon polymorphisms. Only one strain has a SNP in exon W2, namely, an African type 2 spontaneous lymphoblastoid cell line (sLCL) strain with a Q54R change. The same change is found in the EBNA-LP sequences of chimpanzee lymphocryptoviruses (36, 37), so it may represent a circulating variant.

The Y exons of EBNA-LP exhibit more diversity than was previously described (Fig. 3C; Table 2). This analysis identified four major sequence subgroups of the Y exons (defined mainly by exon Y2) that are herein designated A, B, C, and Z. Subgroup A was previously reported for strain Akata (36) and also contains a synonymous polymorphism in exon Y1. Group B is found in the prototype B95-8 strain, and group Z is found in type 2 EBVs (11). The C subtype is novel, characterized by V95E and V102I amino acid differences from B95-8, yet surprisingly is the most common one in the samples, being

TABLE 2 Subtypes of EBNA-LP[a]

Example strain	Exon subtype				Overall subtype	No. of strains (n = 74)
	W1	W2	Y1	Y2		
B95-8	1.1	1	1.1	B	1B	13
sLCL-IM1.05	1.2	1	1.1	B	1B	3
HL05	1.2	1	1.1	B-P	$1B^{P}$	1
GD2	1.1	1	1.1	B-M	$1B^{M}$	1
Jijoye	2	1	1.1	B	2B	2
C666.1[b]	2	1	1.1	B-M	$2B^{M}$	2
Mutu	2	1	1.1	C	2C	29
sLCL-IS1.18	2	1	1.1	C-I	$2C^{I}$	2
sLCL-IM1.17	2	1	1.1	C-H	$2C^{H}$	1
Akata	2	1	1.2	A	2A	13
sLCL-2.14[c]	2	1	2.1	Z	2Z	4
sLCL-IS2.01	2	1	2.2	Z	2Z	3
sLCL-2.16	2	1^{R}	2.2	Z	$2^{R}Z$	1

[a]Adjacent shaded rows differ in nucleotide sequence but are identical at the amino acid level.
[b]Includes the M81 strain.
[c]Includes prototype type 2 EBV strain AG876.

found in both African and European strains, including the Mutu cell line (19). Several of these four main subtypes have rare variants. Most intriguingly, three NPC-derived samples (M81, C666.1, and GD2) share an IRDH-to-MRAS change at amino acids 86 to 89. The new C666 sequence conflicts with a previously published partial C666 sequence, which reported MRDH at this position (38), while another C666 sequence (accession number AB828190) confirms the MRAS motif we observed at this position, although the genome region containing the Y exons is inverted (39). Other, less common Y exon subgroups are noted in Fig. 3C, and the designations for all strains are given in Table ST1 in the supplemental material. There also seem to be certain restrictions as to which exon combinations are compatible: most notably, the type 1 W1 exon seems to require the B subtype of Y exons, while Y exon subtypes A, C, and Z require type 2 W exons. Whether these relationships reflect historic opportunities for recombination or some sort of functional incompatibility remains to be assessed.

All of the variations described above were visualized against a phylogenetic tree of the IR1 template sequence (Fig. 4). Some anomalies were apparent, such as the mainly type 2 strain Jijoye, which has B95-8-like Y exons and segregates between the type 1 and type 2 strains. If this combination were found in circulating (i.e., non-disease-associated) virus types, this would suggest that EBNA-LP and EBNA2 genotypes do not exhibit sequence codependency. However, it remains possible that this mixture represents a disease-associated or lab-acquired mixture of two strains rather than a genuine example of a circulating virus strain.

Based on the observed variations, we propose a nomenclature for EBNA-LP that indicates both the W1 polymorphism (1 or 2) and the Y exon subgroup (A, B, C, or Z) (Fig. 3C and Table 2; Table ST1). Subtypes of these main variants are then indicated by superscripted letters indicating a characteristic amino acid change. For instance, the EBNA-LP variant found in the NPC-derived C666 and M81 virus strains is type $2B^{M}$, with M representing the MRAS motif in exon Y2.

From mapping of both the BWRF1 and EBNA-LP variants onto the phylogenetic tree (Fig. 4), it is apparent that certain EBNA-LP and BWRF1 subtypes usually cosegregate in a manner that defines the genetic variation of this region. While there is no linkage between the EBNA-LP and BWRF1 isoforms that is always true, EBNA-LP subtypes 2A and 2C are usually associated with the longer BWRF1 ORF, while subtype 2Z (type 2 EBV) is generally associated with the shorter BWRF1 ORF (model 2). In contrast, EBNA-LP subtype 1B can be found in combination with all the commonest BWRF1 models (1A, 2A, and 2X). Since no function has been ascribed to the BWRF1 ORF—certainly not one

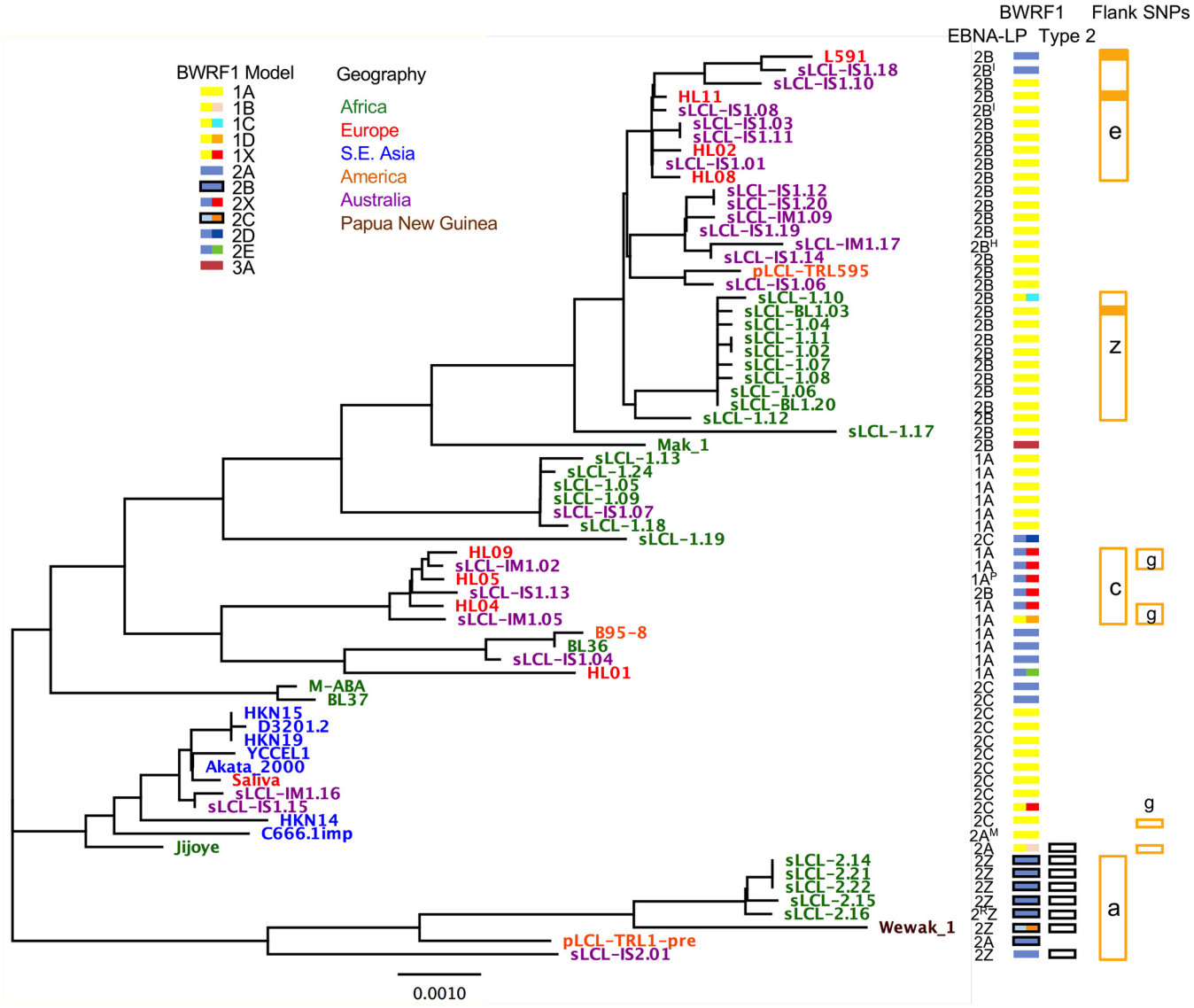


FIG 4 Phylogenetic tree showing the IR1 diversity between strains. The phylogenetic tree of the templated sequence shows branch lengths representative of degrees of difference. Nonindependent strains and strains without a complete IR1 region were excluded. The scale bar represents the tree distance corresponding to 1 nucleotide substitution/kb. The geographic origins of samples are shown by the colors of the strain names. To the right of the tree, for each strain, its EBNA-LP variant is shown by an alphanumeric designation (Table 2; Fig. 3C). BWRF1 subtypes are shown by colored boxes. The left side of each BWRF1 box is color coded for the 3 major groups (1, 2, and 3) (see Table ST2 in the supplemental material), with the presence of the indels characteristic of type 2 indicated by a dark border. The right part of the BWRF1 box is a different color in cases where subgroups are distinct from the major ones (red to indicate the disrupted BWRF1 ORF in 1X and 2X types). Strains with a type 2 EBNA2 are labeled with empty black boxes. Orange boxes indicate strains with common SNPs in the flanks, with the letter showing which SNP is present (Table ST3). A solid orange box represents a strain in which the flank SNP has propagated throughout IR1.

that is linked to EBNA-LP function—these correlations between BWRF1 and EBNA-LP types can currently be considered only to reflect wider sequence conservation/linkage rather than implying a functional interdependence.

IR1 repeats are frequently not homogeneous, but variations at the flanks of IR1 are protected from gene conversion. Attempting to sequence large repeats with short-read data means that variations within repeats are missed by simple consensus alignment or contig assembly methods. Repeats are often subject to concerted evolution, by which they evolve together as a result of ectopic gene conversion, a process by which the sequences are homogenized through DNA repair processes templated by other repeat units (40, 41). When reads are mapped to a consensus sequence, reads from repeat units with an alternative sequence will be mapped to the same position as

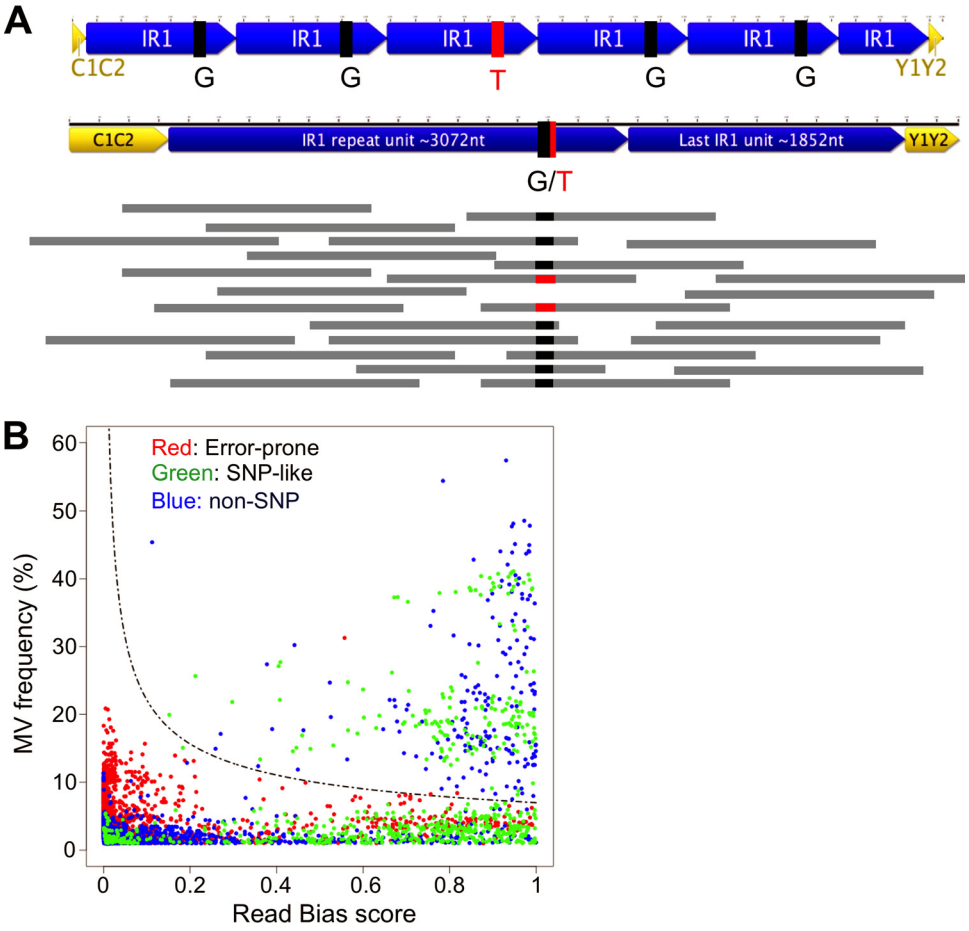


FIG 5 Defining minor variants. (A) Visual representation of the nature of minor variants (MVs), explaining their identification. The top section shows a repeat region in which 1 of 5 repeat units contains the minor variant T rather than the consensus G. If the reads (horizontal gray bars) from sequencing of this repeat are mapped against the consensus template (the representation of IR1 containing 1.6 repeats), 80% of reads will be G (black), but 20% will be the minor variant T (red). (B) Defining a threshold for calling minor variants. The graph plots the read bias (*x* axis) against the MV frequency (*y* axis) for every potential MV with a frequency of >1% across 76 strains. Red points are potential MVs at positions that show an elevated MV frequency in all strains (error-prone positions). Green points are MVs that are seen as SNPs in at least one strain. All other MVs are represented by blue points. The dashed line represents the cutoff above which we accepted the MVs as likely to be genuine, based on the following equation: $(\sqrt{\text{read bias}} \times \text{variant frequency}) \geq 7$. Potential MVs that were designated SNPs in the flanks of IR1 were excluded.

the consensus reads (Fig. 5A). Therefore, heterogeneity within repeats will be observable as a nonconsensus nucleotide at a frequency that is proportional to the percentage of repeats containing the alternative nucleotide. These "minor variants" (MVs) should be equally prevalent on both forward and reverse reads. False-positive MVs were eliminated by identifying and excluding error-prone sites, ensuring that MVs were represented on both strands, and setting an identification threshold of over 7% (see Materials and Methods) (Fig. 5B), since any MV occurring in at least one strain would have a frequency of 8 to 20%, based on the observation that the IR1 repeat number is typically 5.6 to 6.6, and no more than the ~11.6 copies found in B95-8 (4, 9, 42) (Table ST1).

In defining the MVs, several positions near the flanks of IR1 were identified that exhibited high MV frequencies (not shown) but also a high degree of read bias caused by read pairs that spanned the junctions between IR1 and the adjacent unique regions of the genome. After correcting these nucleotides in the template, it was apparent that these variants were found only at the edges of IR1, close to the unique sequence, and usually not as MVs elsewhere in IR1. There were 22 positions within 200 bp of either edge of IR1 that were different from the corresponding internal positions (Table ST3).

Two of these (one of which has been reported previously [28]) were found only in the B95-8-related strains and BL36 (previously described as an intertypic recombinant). By mapping the flank SNPs shared by more than two independent strains onto the phylogenetic tree (Fig. 4), it was clear that these generally cosegregate with the phylogeny. Notably, the further these edge-associated variants were from the ends of the repeat, the less tightly they were associated with the phylogeny. For instance, the SNP in group g is 160 bp from the edge of IR1 and is found in only a subset of related strains. In some cases, the edge-associated variants were propagated to purity throughout IR1 (e.g., sLCL-BL1.03). In contrast, there were no examples of groups having some members with SNPs near flanks reverted to the normal base calls. Thus, these edge variants are evolutionarily stable and not subject to gene conversion but are able to seed gene conversion events elsewhere in the repeats. These flank variants were therefore classified as SNPs rather than MVs, to prevent them from skewing our analyses of repeat heterogeneity elsewhere within IR1.

Heterogeneity in IR1 is more common in tumors and tumor-derived cell lines and arises though interstrain recombination as well as spontaneous mutation. Even after excluding the flank-associated SNPs, only 21 of the 70 distinct strains tested were homogeneous (i.e., had no MVs in their repeats), and most were sLCL strains. MV numbers were then compared according to whether the virus was isolated directly from tumors (or long-established cell lines) or, more recently, from blood (i.e., sLCLs) or saliva. While 70 to 80% of the samples had 2 or fewer heterogeneous positions in IR1, 9 viruses (7 of which were derived from tumors) had at least 5 (and up to 21) MVs in IR1 (Fig. 6A). To assess whether the number of MVs was simply a function of the number of repeats in IR1, the length of IR1 was estimated based on the read depth in the W exons relative to that for Cp (Table ST1). Strains with long IR1 regions were no more likely to contain (or lack) MVs. Conversely, the strains with the most MVs often had a smaller proportion of reads mapping to IR1 (i.e., fewer IR1 repeat units).

To further characterize the origins of these variants, the MVs were stratified into those arising from spontaneous mutation and those arising from some form of interstrain genetic exchange that we describe as "recombination-like." This stratification is based on the assumption that sporadic mutations will be randomly distributed throughout IR1, whereas MVs arising through interstrain recombination will also be observed in our data set as SNPs between strains. Even though this approximation is limited—there are undoubtedly SNPs that have not yet been identified, and there may be mutagenic processes within the cell that target certain sequences more often—this analysis suggested that viruses with few MVs exhibit mostly sporadic mutations, whereas those with many MVs appear to have arisen through genetic exchange between strains (Fig. 6B and C). Strain Cheptages exhibits the largest number of mutation-associated MVs that are not accompanied by any recombination-like changes. Since this strain also contains deletions at both ends of IR1, this is further evidence that Cheptages has experienced a highly mutagenic environment. In contrast, strains that show substantial evidence of interstrain recombination tend to also have large numbers of apparent mutations (Fig. 6C; Table ST4). This may mean that the circumstances that predispose viruses to genetic exchange also predispose them to mutational events. However, it also remains possible that these apparent mutations represent previously unseen SNPs.

Since a recombination-like signature was typically associated with mutations, the strains were split into three classes: the homogeneous, recombination-like, and mutation-only classes. The distribution of these classes within the "tumor" and "circulating" groups suggests that interstrain recombination is more prevalent in cancer-associated viruses than in circulating strains, whereas the latter are more likely to have a homogeneous IR1 region.

Changes in IR1 arise regionally and can efficiently be purified to homogeneity in IR1. Looking in more depth at the nature of interstrain recombination within IR1, we noted that HL05 (from a Hodgkin's lymphoma biopsy specimen) contains a cluster of 5 recombination-like MVs within a 200-bp region. We similarly noted a cluster of three

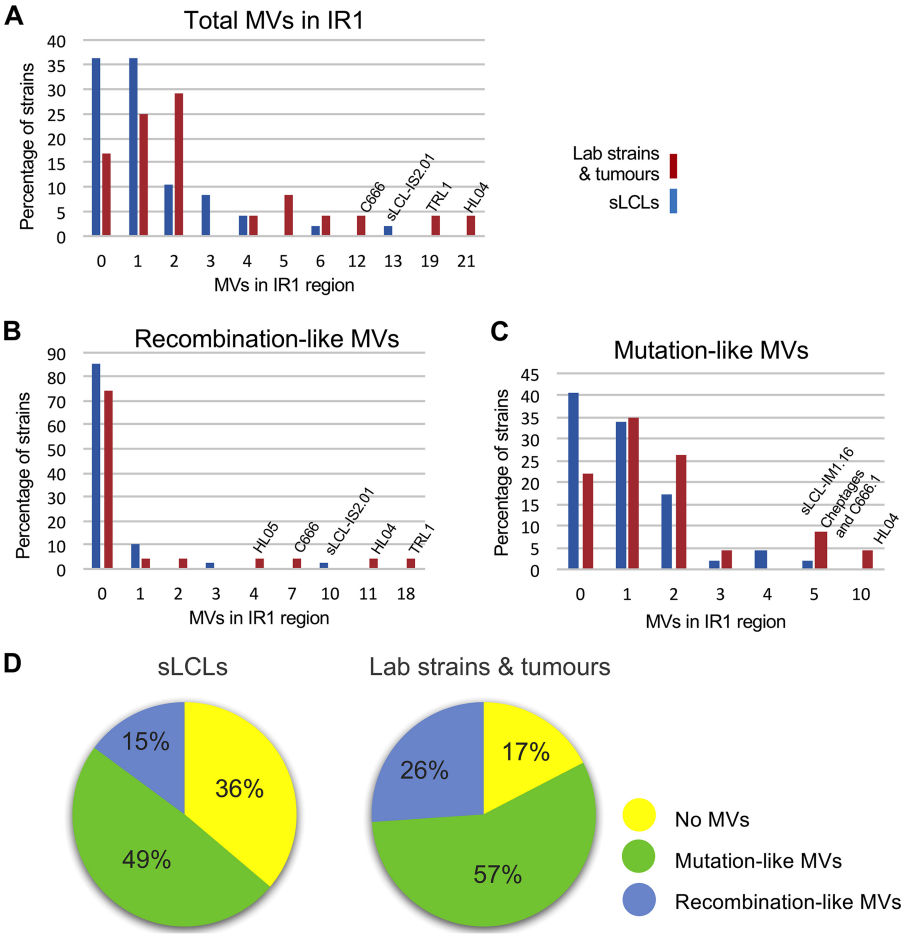


FIG 6 Minor variant diversity in spontaneous LCLs compared to that in tumor-derived samples. Graphical representations show the number of minor variants in the IR1 repeat region in each independent strain (defined in Materials and Methods). Samples were grouped as being either spontaneous LCLs (including one saliva-derived sequence) (blue bars) or tumor-derived samples (red bars). The number of MVs is shown on the *x* axis, and the *y* axis shows the percentage of strains with each number of MVs. Graphs display total MVs (A), MVs that resemble SNPs (recombination-like MVs) (B), and MVs that resemble spontaneous mutations (mutation-like MVs) (C). (D) Pie charts indicating the types of genetic variations found in EBV IR1 from the sLCL and tumor groups. The type of variation was classified as either homogeneous repeats (yellow), at least one recombination-like MV (blue), or only mutation-like MVs (green).

MVs in the prototype EBV strain B95-8 (two of which are recombination-like), separated by only 126 bp (discussed in more detail later). Other strains contained more widespread distribution of MVs, suggesting a larger region of genetic exchange. Intriguingly, two of these (TRL1 and sLCL-IS2.01) appeared to combine SNPs from type 1 and type 2 viruses. TRL1 is a type 1 EBV containing a deletion within EBNA3B, sequenced from two independent cell lines grown from an aggressive posttransplant DLBCL (43, 44), while sLCL-IS2.01 is a spontaneous LCL strain established from the blood of a PTLD patient containing a type 2 EBV (4). These strains also show evidence of clustering of the minor variants, with different groups of MVs exhibiting different MV frequencies (Table ST5).

In order to analyze these variations, we used the T-RECs tool to visualize the virus strains that most closely resembled both the consensus and minor variant versions of the two strains (Fig. 7). Both TRL1 and sLCL-IS2.01 showed evidence of having substantial regions of intertypic recombination, with TRL1 acquiring predominantly type 2 IR1 sequences in a type 1 background and sLCL-IS2.01 acquiring type 1 sequences in the type 2 background. To more clearly visualize the transitions between type 1 and

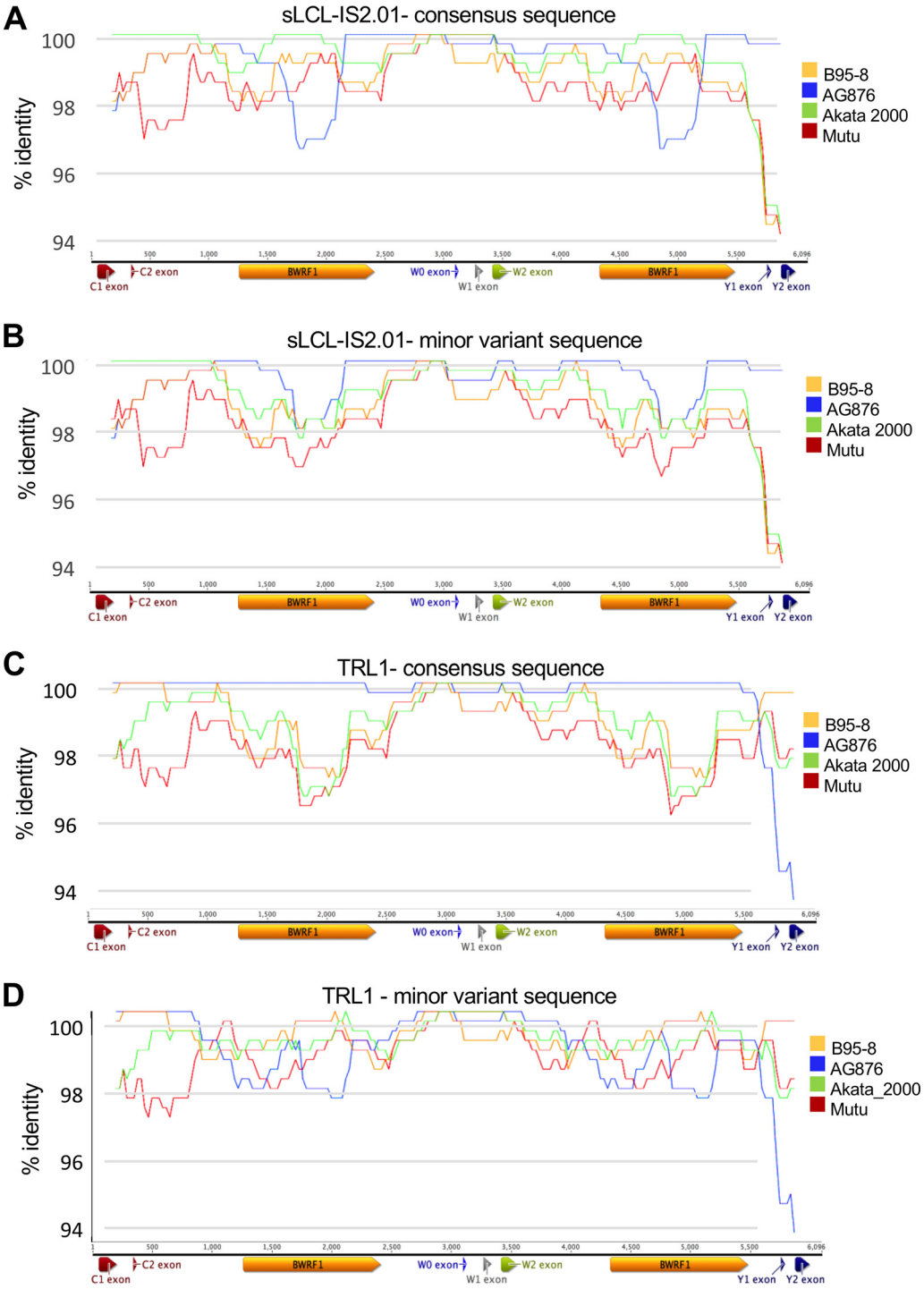


FIG 7 T-RECs analysis of genetic exchange in IR1 regions of sLCL-IS2.01 and TRL1. The consensus sequences for sLCL-IS2.01 (A) and TRL1 (C) were modified to instead contain all of the minor variant bases (B and D). These sequences were compared to those of a set of four viruses (see the key to the right of the plots) from different clades on the phylogenetic tree to identify regions of similarity, using T-RECs. The similarity plots show the percent identity of the test strain to each comparator strain within a 200-bp scanning window.

type 2 sequences without the signals caused by spontaneous mutations, we used a bootscan analysis to view the transitions between the most closely related type 1 and type 2 strains (Fig. 8). This analysis clearly showed that both TRL1 and sLCL-IS2.01 had acquired a region of the other type from Cp (the junction upstream of C1 is apparent

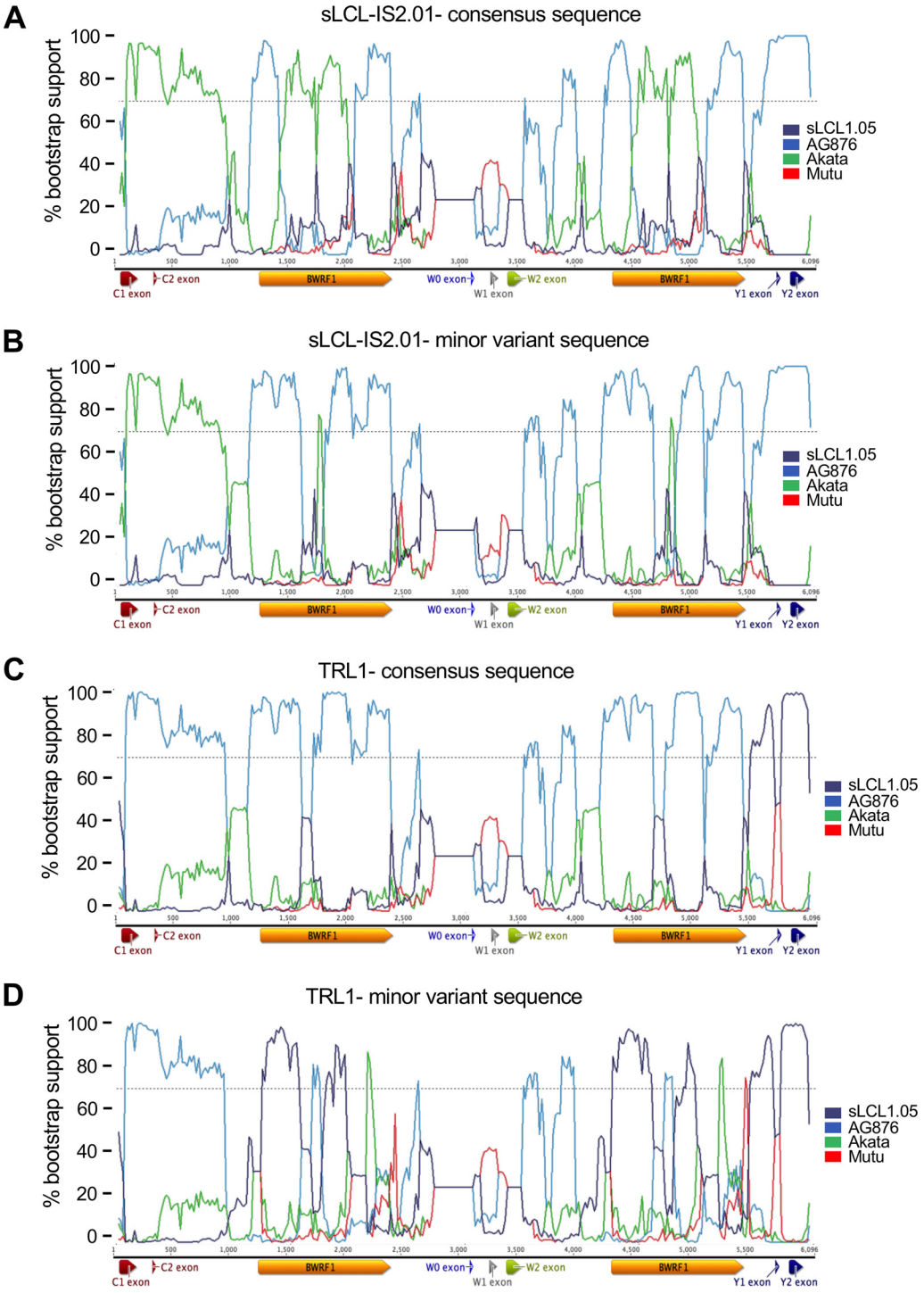


FIG 8 Bootstrap analysis of genetic exchange in IR1 regions of sLCL-IS2.01 and TRL1. The consensus sequences for sLCL-IS2.01 (A) and TRL1 (C) were modified to contain all of the minor variant bases (B and D). These two sequences were compared to those of a set of four viruses from different clades on the phylogenetic tree to identify regions of similarity, using Bootscan. This analysis simulates phylogenetic tree assembly for a 200-bp sliding window, with the *y* axis indicating the percentage of occasions that each strain is the nearest neighbor on the tree and the *x* axis marking position, as illustrated by the IR1 schematic below each plot.

in the bootscan plots) at a position in IR1. However, the cross-type acquisition of sequence was no longer represented by a single stretch of sequence. While the region from Wp to exon W2 is too highly conserved to discriminate between strains, it is clear that the region from exon W2 to the start of BWRF1 and from the end of BWRF1 into

Wp in sLCL-IS2.01 is entirely type 2, whereas the center of BWRF1 is almost entirely Akata-like, with an MV frequency of 17% for the type 2 sequence (Table ST5), suggesting that only a single repeat unit retains the type 2 sequence there. For TRL1, the situation is more complex, showing an almost entirely type 2 IR1 sequence, with the ends of BWRF1 having type 1 sequences in one repeat unit but the center of BWRF1 being of type 1 in two repeat units.

While it is not possible to conclusively establish how this combination of type 1 and type 2 sequences was acquired, in both cases the patchy nature of the type 1 and type 2 sequence transitions is most easily explained by a two-step process, with the virus first acquiring a large region of sequence (for TRL1, a region from Cp to a middle repeat unit in IR1; for sLCL-IS2.01, the region could be shorter), presumably by recombination, and the repeat subsequently undergoing ectopic gene conversion between repeat units, with some regions templated to the original sequence and others propagating the introduced region. Notably, in all five of the analyzed strains, the blocks with a common MV frequency are relatively short—less than 500 bp. Indeed, for HL05, the only part of IR1 with evidence of interstrain exchange is a small region which could have been acquired in a single allelic gene conversion event between different virus genomes rather than by recombination followed by gene conversion.

The prototype strain B95-8 contains low-frequency polymorphisms that include a stop codon in EBNA-LP. The prototype EBV strain (B95-8) contains five heterogeneous positions, two of which match SNPs in other virus genomes. The same set of MVs is found in one B95-8-infected cell line that we sequenced and one sequenced by the Chiang laboratory (21) and in cells containing B95-8-delEBER, a recombinant subclone of the B95-8 BAC originally produced in the Hammerschmidt laboratory (45, 46). In addition, the identical MVs were found in BL36—an African BL previously reported to be an intertypic recombinant (4, 47)—and two of these MVs were also found in X50-7, an umbilical cord LCL established by coculture with X-ray-irradiated B95-8 cells (48). For all of these cell lines, four of the five SNPs share the same frequency in each genome (around 8% in the parental B95-8 strain and around 20% in the other lines), which most likely corresponds to an occurrence of once per IR1 (i.e., each sequence variant occurs in only one repeat unit, as illustrated in Fig. 5A). The MV in the hairpin region of B95-8 (position 12647) occurs at a higher frequency (Fig. 9A).

These MVs within IR1 of B95-8 are distributed across the repeat unit (Fig. 9B), but the four MVs of equal frequency were twice (from independent digests) cloned from the EBV BAC in a single BamHI fragment, demonstrating that they were all in the same repeat unit. Of the three MVs that cluster toward the end of BWRF1, two are SNPs in at least one other strain, lying in a relatively polymorphic region of IR1. Most strikingly, however, the more distant MV (G14617T) (highlighted in red in Table ST4) changes the last codon of the W1 exon of EBNA-LP from GAG (glutamate) to a TAG stop codon. As described elsewhere (49), we generated a B95-8 BAC (WT^{w}) in which the original IR1 was replaced with 6.6 copies of the consensus IR1 repeat unit, removing all MVs. LCLs established with WT^{w} exhibit both higher and more consistent levels of EBNA-LP protein and a more consistent range of larger EBNA-LP isoforms (Fig. 9C), suggesting that the stop codon alters the character of EBNA-LP expression in LCLs. Additionally, LCLs are established more rapidly with WT^{w} than with the parental HB9 BAC (not shown). Consistent with this subjective observation, a dye dilution-based proliferation assay showed that after 8 days, WT^{w}-infected cells (those that had divided at least once) typically progressed through more cell divisions than those of HB9-infected cells (Fig. 9D).

DISCUSSION

Assembling repeat regions from short-read data. Repeats are challenging to assemble from short-read data, but we have shown here that they can offer considerable insights into the biology of the virus and perhaps of virus-associated disease. We have identified some strategies to improve the quality of repeat sequence assembly by considering flank and internal variations in sequence. Repeats also represent an inter-

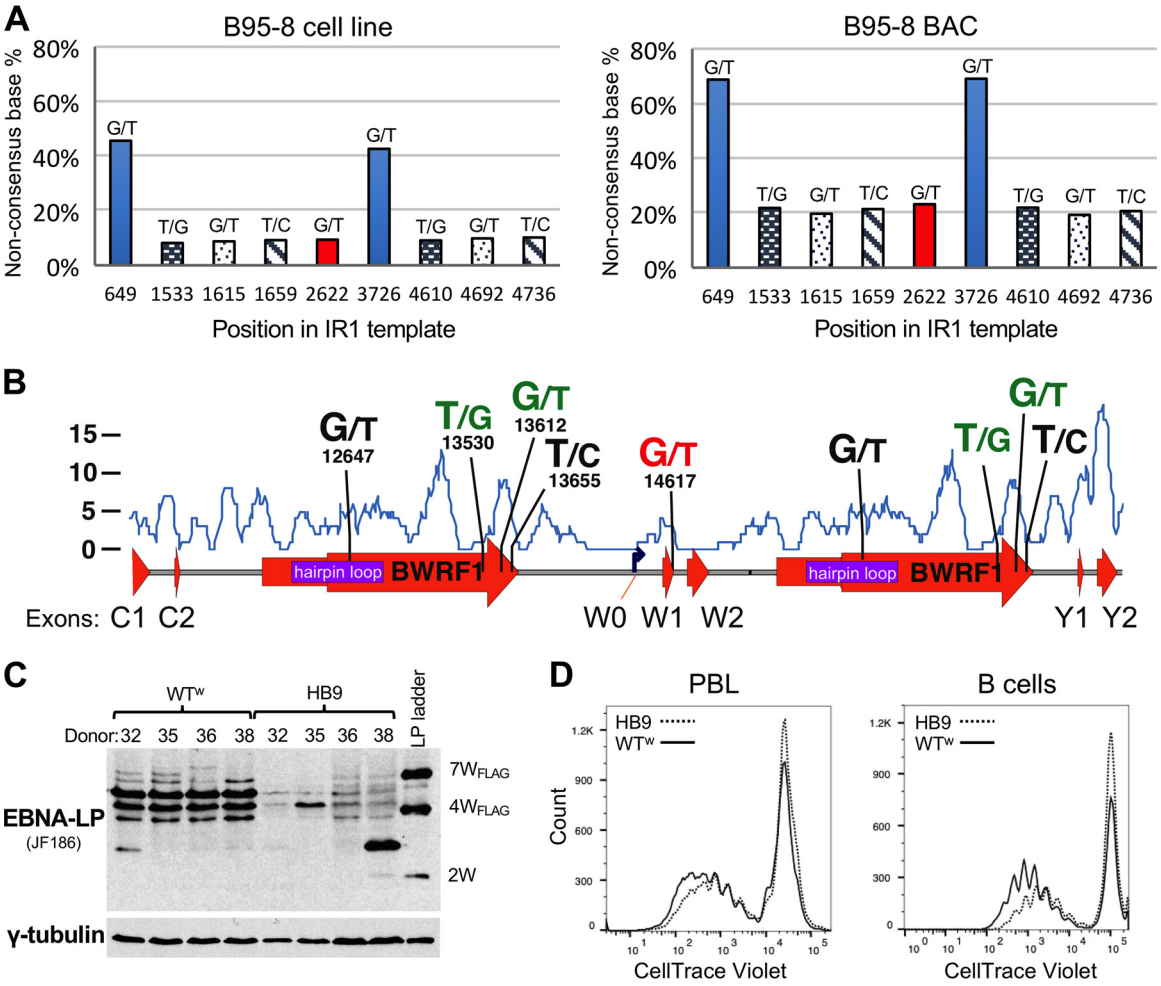


FIG 9 Minor variants in IR1 of the prototype EBV strain B95-8. (A) Graphs showing MV frequencies across the template sequence of IR1 of B95-8. The sequence variations are shown with the base of the published B95-8 sequence at left and that of the newly identified variant at right, with the *y* axis showing the frequency of the newly identified nucleotide. The *x* axis shows the position of that variant in the IR1 template multiple-sequence alignment (see Supplemental Data SD1 in the supplemental material). Note that this means that the alternative nucleotide at position 649/3726 is the dominant allele in the HB9 BAC-based genome (right). At other positions, the different frequencies of MVs are due to different numbers of IR1 repeat units between the B95-8 cell line (11.6 repeat units) and the HB9 clone of the B95-8 BAC (6.6 repeats). (B) Graphical representation of the B95-8 MVs relative to major features of IR1. For each MV, the left base indicates the published nucleotide and the right indicates the alternative variant, with the numerical position in the first repeat unit of the prototype EBV sequence (NC_007605) given below. Bases in green indicate changes that have been observed as SNPs in other strains. The "G/T" pair in red indicates the stop codon in W1 introduced by the MV T nucleotide at that position. The blue line and *y* axis indicate the SNP density (from Fig. 3A) across the repeat for comparison. (C) Western blot comparing EBNA-LP levels and sizes in LCLs 30 days after infection of four independent donors, using either the original B95-8 EBV BAC (HB9) or a recombinant BAC from which all MVs were removed (WT^w). The EBNA-LP ladder is from 293 cells transfected with three differently sized EBNA-LP variants: the number of W1W2 domains in each band plus any tag is indicated to the right. The blot was stripped and reprobed with a gamma tubulin antibody as a loading control (lower panel). (D) Flow cytometry traces showing dilution of the CellTrace Violet stain on B cells 8 days after infection, comparing equal cell numbers for the WT^w (solid line) and parental HB9 (dotted line) strains. The *y* axis shows the cell count, and the *x* axis shows the $\log_{10}$ CellTrace Violet fluorescence intensity. The left plot is for $CD19^+$ (gated) live cells infected as mixed lymphocytes, and the right plot is for CD19-purified B cells infected and grown in the absence of other cells, based on data also published elsewhere (49).

nally controlled sequence space where detection of minor variants offers insights into events occurring elsewhere in the virus genome. For instance, our data set—taken from clonal cell lines—exhibited widespread low-level variants, likely due to cross-contamination or index exchange facilitated by the high degree of multiplexing used in the sequencing reaction mixtures (33), and identified error-prone sequence positions, suggesting that considerable care is needed in searching for rare variants or quasispecies in virus populations.

Minor variant detection has only recently begun to be applied to herpesviruses at the whole-genome level. For instance, analysis of farmed carp showed infection by

multiple strains of cyprinid herpesvirus 3 to be normal in moribund fish (50). Similarly, CMV infection of humans is typified by infection with a combination of strains that is often stable over time (reviewed in reference 51) but can vary between different organs within individuals in both strain combination and degree of virus diversity (12, 52). However, these studies do not clearly distinguish between acquisition of new mutations and the presence of multiple diverse strains and do not make use of haplotype variation or shared MV frequency to explore the causes of these observations or provide analyses to estimate sample or technical cross-contamination that could confound them.

One of the challenges in assessing the accuracy of our study is the lack of read data for IR1 in other virus strains. We were able to confirm the B95-8-associated MVs by subcloning and Sanger sequencing repeat units from the B95-8 BAC and by analyzing B95-8 sequence data generated in another center (21). Additionally, YCCEL1 was recently cloned as a BAC, and this BAC was sequenced using PacBio long-read technology. In agreement with our observations for YCCEL1, the PacBio sequence did not find any MV SNPs within IR1. However, the assembled YCCEL1 BAC sequence did contain a 1,607-bp deletion in the 8th repeat unit of IR1 (accession number AP015016) (23). We found no evidence for this truncated repeat in our reads, and subsequent restriction and PCR analysis of the BAC were unable to detect this deleted repeat unit (T. Kanda, personal communication), suggesting that our assembly was accurate and, by extension, that the long-read sequence of YCCEL1 requires more careful manual curation, at least for IR1.

Biological implications of IR1 sequence diversity for EBV. The internal repeat regions of EBV are functionally poorly understood. Various numbers of W1-W2 exon pairs are known to comprise the upstream region of the EBNA transcripts (53). EBNA is alternatively spliced either to allow translation of EBNA-LP or to be retained as a 5′ untranslated region (UTR) upstream of EBNA2 or one of the other EBNAs (25, 54). The W0 exon (and its associated promoter, Wp) is at its most active immediately after infection, but its activity declines as transcription switches to the upstream Cp (55). Very recently, transcripts spliced between W1 exons were described that produce the antiapoptotic protein BHRF1 during the EBV lytic cycle (26). The W1-W2 intron is reported to be a persistent RNA (sisRNA1) in latency III cells, and there is some evidence that the larger W2-W1 exons may also produce a stable RNA, sisRNA2 (34).

The high degree of conservation observed in the Wp promoter and the W1-W2 intron supports the possibility that they have a specific conserved function. In contrast, the conservation of the BWRF1 ORF in only 80% of strains casts some doubt on its importance, at least in its current (hypothetical) context. Even so, it is unlikely that such a large ORF would arise by chance (and be conserved in so many strains). It is also striking that all six examples of genome deletions that delete the flanks of IR1 are in the small zone that leaves the hairpin-BWRF1 region of the genome intact (Fig. 2B), suggesting that this part of IR1 represents a single functional entity, either as a DNA structure or acting through sisRNA2. Because there are multiple copies of this region in IR1, it also implies that partial copies may act in some sort of dominant negative manner or that the first and last units are more important. A case for the first repeat unit being more important has previously been made, in that deleting the regulatory region upstream of Wp was found to be tolerated in internal IR1 repeat units, but mutation of the first repeat unit had a detrimental effect on B cell transformation and Cp activity (56), implying that the integrity of the BWRF1 region may be important for proper activity of the Wp and Cp promoters.

The BWRF1 region is the source of the reportedly stable intronic RNA sisRNA2 (34). Conceptually, sisRNA2 has similarities to the latency-associated transcript (LAT) of HSV, which is critical for HSV latency in neurons. HSV LAT is also an intron-derived persistent RNA that is processed by splicing events in addition to the simple excision of the intron (57, 58). Two regions of BWRF1 are highly conserved, while all of the sequence changes that disrupt BWRF1, including the larger indels found in a few strains, are in the central

polymorphic region. It therefore remains possible that some parts of the BWRF1 ORF may be used to generate translation products if sisRNA2 is processed to excise the less conserved regions. Additionally, some mechanism—perhaps mediated by an RNA structure associated with the hairpin—would be required to facilitate BWRF1 translation. The polypeptides encoded by HSV LAT are reported to be untranslated (58), but detailed analysis of sisRNA2 would be required to test whether this is also true for EBV.

In contrast to BWRF1, the repeat domain of EBNA-LP is highly conserved across IR1, with most of its diversity residing in the C-terminal Y exons. This is distinct from the case in rhesus lymphocryptovirus, whose EBNA-LP transcript contains a final W1/2 exon pair that differs substantially from the first two (59). Our MV analysis found no evidence of similar W exon diversity in EBV, and indeed, only two strains (X50-7 and sLCL1.12) contained an MV in exon W2, while the stop codon in B95-8-related strains was the only MV in exon W1 in any virus. Our analysis reinforces previous observations that the "type 2" W exon sequence is not confined to type 2 viruses. These are easily distinguished, as the monoclonal antibody JF186 binds only to the type 1 variant (35), whereas antibody 4D3 detects an epitope in exon W2 (60) and can thus be used to confirm this difference between strains (60; our unpublished observations). Using our new nomenclature to describe EBNA-LP variants highlights that there appear to be some restrictions on which W and Y exon variants are found together, which suggests possible functional codependence. Particularly intriguing is the B^M subgroup, which is found in only three NPC-derived viruses and appears to be compatible with both type 1 and type 2 W exons. However, functional assays and/or structural information will be required to establish whether different EBNA-LP W and Y exon pairs are cross-compatible.

Functional consequences of IR1 heterogeneity. Only limited attempts have previously been made to investigate the heterogeneity of IR1 or the other EBV repeats, although a recent study reported a relatively small proportion of positions containing MVs (61). While the genome sequence of B95-8 reports heterogeneity in the terminal repeats (9), contemporaneous investigations searching for heterogeneity relied on restriction analysis and found no evidence of IR1 heterogeneity in EBV strain B95-8 or JHU-1 (28, 62). However, the early sequences of B95-8 IR1 did disagree over whether B95-8 position 12647 was a G or a T (9, 27, 28). We found that this position shows an approximately 50-50 split in the parental B95-8 viruses (Fig. 9). It may appear more surprising that the stop codon in EBNA-LP has not previously been reported. However, because EBNA-LP transcription is characterized by a poorly characterized pattern of alternative splicing and exon skipping (25), variations in the predominant EBNA-LP sizes in LCLs have become regarded as normal, concealing this mutation. We have shown that an intact IR1 of 6.6 repeat units produces EBNA-LPs with predominantly 4, 5, and 6 repeats, whereas only the HB9 BAC seems to select single, smaller EBNA-LPs. Curiously, we even saw evidence of EBNA-LPs with more repeat domains than IR1 repeat units, which are most likely produced by splicing between (rather than within) transcripts. It is not formally possible to say whether the EBNA-LP produced in HB9 LCLs is truncated or not. However, we have found that deletion of the Y exons from the B95-8 BAC (in the context of heterogeneous IR1) results in very low protein levels (49), suggesting that any truncated EBNA-LP made may not be stable. Nevertheless, we have observed that the MVs in B95-8 do appear to have a detrimental impact on its transforming ability. Since these minor variants are retained in the B95-8 cell line, it is possible that the reduced levels of EBNA-LP are significant only during initial transformation or that they are favored in the marmoset background of the B95-8 cell line. Nevertheless, a great deal of research has been performed using the B95-8 BAC as a basis for genetic manipulation of the virus (63). It is therefore important to consider that these experiments were done in the background of a modestly defective EBNA-LP, and it may be advisable that future genetic studies use an EBV BAC with an intact EBNA-LP region.

The pattern of IR1 heterogeneity (including the B95-8-specific flank SNP) is retained in a number of B95-8-related strains. The commonly used X50-7 strain, whose genome

is integrated and has lost Cp (32), is missing three of the B95-8 MVs, but since the genome was UV irradiated prior to transformation (48) it seems likely that these MVs were lost during homology-driven DNA repair. The B95-8-type MV heterogeneity is completely reproduced in the intertypic recombinant BL36, which was not known to be B95-8 associated. Further examination of the BL36 sequence supports the hypothesis that its genome arose by superinfection of a type 2 BL with the B95-8 virus, resulting in intertypic recombination between the two: between position 10024 (where BL36 contains a 14-bp palindromic insertion disrupting the vIL10 gene—possibly a scar from the recombination) and position 43611, the BL36 and B95-8 sequences are identical. This includes five SNPs—three in IR1 and two in the Y exons—that are found only in B95-8. We therefore suggest that BL36 is a lab artifact and is not representative of a natural circulating strain. Nevertheless, additional examples of intertypic recombinants remain. They have been isolated reproducibly from both the blood and throat of T cell-immunodeficient Caucasian HIV patients (64), as well as from two NPC patients (one with mutant EBNA3A) and one healthy Chinese individual with type 1 EBNA2 and EBNA3A and type 2 EBNA3B and EBNA3C (65), and also include the African strain sLCL-1.18 from this data set, which has a typical type 1 IR1 and EBNA2 and type 2 EBNA3s. Since type 2 EBNA2 gives a reduced efficiency of LCL outgrowth (6), it remains possible that at least some of these examples may be artifactual strains produced as a result of the selection pressures associated with LCL outgrowth. However, despite our identification of BL36 as an artifactual intertypic recombinant, the evidence for the existence of such recombinants in the wild remains firm but would benefit from further corroboration.

Further corroboration is also required for our observation that tumor-derived viruses appear more likely to have evidence of interstrain genetic exchange in IR1. The nonrandom sampling of these strains make any statistical comparison of the two groups uninformative, but there are other examples of changes in virus genomes that are associated with disease. For instance, deletion of EBNA2 and the end of EBNA-LP is seen in approximately 10% of BLs (66–68), while EBNA3B mutation is associated with DLBCL and also found in HL and BL (44). Indeed, HL08 contains a small deletion in EBNA3B that has distinct haplotypes on either side (44), implicating an aberrant genetic exchange between viruses in its origin. The association of interstrain recombinants with disease may indicate that aberrations in the process of genetic exchange can produce pathogenic mutations. It is equally possible that the processes that promote oncogenesis also promote the exchange of DNA between distinct EBV strains. Nevertheless, this would still require the oncogenic cell to be coinfected by two distinct viruses, which—if it is random—seems likely to be a relatively rare event. Further studies, such as by comparing the sequences of EBV in tumors with those in the patient's blood and/or saliva, will be required to establish whether coinfection and genetic exchange between viruses may contribute to pathogenesis.

Mechanisms and implications of interstrain genetic exchange in IR1. Whole-genome analyses supported earlier suggestions that interstrain recombination is a major component of diversity in EBV (4), HSV-1 (15), and CMV (13), and by extension, this potentially makes evolutionary analysis of herpesvirus genomes enormously complicated. By using the repeat units as internal controls, we have been able to deduce patterns of accumulation of changes in virus genomes, classifying them as either spontaneous mutations or interstrain genetic exchanges. Our analysis is necessarily an oversimplification, as we do not yet have a comprehensive list of EBV SNPs, so some SNP-like changes will have been missed. However, the presence in a few strains of large clusters of exchanged SNPs has offered considerable insight into the processes by which EBV strains can exchange genetic information and thereby gain diversity.

Interstrain recombination would allow large regions of virus genomes to be exchanged, as seen in BL36, and probably TRL1 and sLCL-IS2.01 (Fig. 7 and 8). However, most of the regions of SNP-like MVs exchanged in IR1 are much smaller, between 150 and 1,000 bp, as best exemplified by the distinct region introduced in HL05. Small

interstrain exchanges of sequence have been reported elsewhere, such as a type 1 virus reported to contain a type 2 sequence at the end of EBNA3A that spanned 882 nucleotides (69). However, if these small exchanges of sequence are commonplace throughout the EBV genome (and those of other herpesviruses), it makes the development of a nomenclature describing herpesvirus strain variation and relationships between strains extremely complicated.

Classically, the diversification of repeats and their subsequent purification to a new sequence have been explained by the process of crossover fixation, i.e., the expansion and contraction of repeats by crossover recombination. This process at the gene level—the "gene accordion" model—has been reported to be a key mechanism in the diversification of poxvirus genes and requires a linear DNA intermediate (70). However, this model would predict the MV diversity to span whole repeat units, which we do not generally observe, with the possible exception of some parts of TRL1. Genetic repeats are also driven to homogeneity through allelic gene conversion, which is mechanistically based around the homologous repair of double-strand breaks (71). The small patches of linked MVs in IR1 are in the size range typical of synthesis-dependent, strand annealing-mediated break repair (71). Alternative methods of templated DNA break repair, such as single-strand annealing or DNA break-induced replication (BIR) (reviewed in references 41 and 72), may also cause these patterns. In particular, BIR generates an extended single-stranded DNA (ssDNA) intermediate that may be more prone to mutation, which could explain the cluster of apparently spontaneous mutations in HL04, and more generally the observation that strains with evidence of genetic exchange also contain spontaneous mutations.

Most analyses of herpesvirus recombination have been performed in HSV-1. These analyses have all used models of coinfection during the lytic virus life cycle (reviewed in reference 73). In contrast, EBV also undergoes a substantial proportion of its genome replication in latency in B cells, in which viral genome replication is similar to that of the chromosome, rather than during the lytic cycle. This difference is supported by sequence analysis of diverse herpesvirus genomes, which found that gammaherpesviruses—but not alphaherpesviruses—are enriched for distinct sequences reported to promote genetic exchange (74), including one also involved in the B cell-specific process of antibody class switch recombination (75). Thus, the apparently higher frequency of small repair-driven genetic exchange may be a reflection of the biological differences in latency between alpha- and gammaherpesviruses.

The sequence diversity of the repeat leads us to propose a model in which there is an initial introduction of genetic sequence into the repeat region followed by the purification through gene conversion of sequences across IR1. In the case of TRL1 and sLCL-IS2.01, the mixed genotype likely arose through a large recombination event that introduced Cp and one or more repeat units. In contrast, the small region in HL04 seems most likely to have been introduced through allelic gene conversion (i.e., repair in one virus genome templated by another genome), although a larger region could have been introduced by recombination and then much of it lost as the repeat mostly purified to its original sequence. It appears likely that ectopic gene conversion drives the repeat toward homogeneity, with repeat units in the same virus genome templating repairs.

Our analysis of IR1 also hints at the limits of gene conversion in this setting, as SNPs within 120 bp of the flanks of IR1 appear to be fixed mainly in their respective populations, while those a little further out are found in some, but not all, related strains. It suggests that the concerted evolution of IR1 typically relies on a considerable region of homology (perhaps 100 to 200 bp) on either side of any variant in the target region, fitting a gene conversion model. However, three strains (HL11, L591, and sLCL-BL1.03) have propagated a flank SNP to purity across IR1, suggesting that the slow conversion of SNPs in the flank by internal sequences contrasts with potentially rapid concerted evolution in the body of the repeat, as indicated by the overall low degree of IR1 heterogeneity.

In summary, our in-depth analysis of IR1 sequence variation has uncovered a number of surprising features of EBV biology. We identified a new variant of EBNA-LP and suggest that it is important to the virus to retain the BWRF1 region intact but not for the BWRF1 ORF to be retained in full. We observed that IR1 is often heterogeneous and that this heterogeneity offers considerable insights into viral evolution, suggesting that the exchange of genetic information between strains resembles gene conversion, and we demonstrated the subsequent concerted evolution of the repeats. We suggest that interstrain recombination may contribute to pathogenesis, and we identified a mutation in EBNA-LP of the prototype strain of EBV. We hope that the sequence analysis of further virus strains will help to clarify the validity (or otherwise) of these observations.

MATERIALS AND METHODS

DNA sequence and sample classification. The sequence data analyzed herein were described in detail in a previous report (4), and sequence data are available (accession number ERP001026) at the European Nucleotide archive (http://www.ebi.ac.uk/ena). Briefly, the EBV sequences came from a collection of cell lines and primary tumors from around the world, although predominantly from Europe, Australia, China, and East Africa. For the purposes of the analyses, we describe all spontaneous LCLs (sLCLs) as being "non-tumor-derived" cell lines, while all long-term cell lines are defined as tumor-derived cell lines. Since some of the sLCLs were from patients with EBV malignancies, this is necessarily an approximation, but there is no evidence that the virus strain in the patient's circulation (the one used to establish the sLCL) is the same as the one causing the malignancy. This information was published previously but is included in Table ST1 in the supplemental material, for convenience. For the purposes of analyses requiring unrelated or distinct strains, however, we combined certain strains because they have common origins, as follows: for diversity analyses, Jijoye and P3HR1-HH514-C16 were combined into a single origin (Jijoye); B95-8, LCL B95-8 delEBER, X50-7, and BL36 (see the text) were combined into a single B95-8 entry; and TRL1-pre and TRL1-post had identical sequences so were described as a single strain, TRL1 (Table ST1). For transparency, all sequences (including those of the very similar strains) were included in the multiple-sequence alignments and minor variant analyses. Additionally, we received Illumina HiSeq reads for the parental B95-8 cell line from Alan Chiang (processed as described in reference 21).

Assembly of an IR1 template sequence. Reads were mapped to template sequences by use of Burrows-Wheeler Aligner (v0.7.9a), using the BWA-MEM algorithm (76). The sequence of the IR1 region was determined by creating a theoretical template comprising 716 nucleotides upstream of IR1, 1.6 repeat units of IR1, and 456 nucleotides downstream of IR1. This was achieved using a Python script to undertake the following assembly. For each strain, reads were mapped to the first repeat unit of the B95-8 IR1 (positions 12001 to 15072 of the NC_007605.1 sequence), and the consensus of the repeat unit was extracted using SAMtools (v0.1.19) (77). This was then concatemerized to give one intact repeat and a second partial repeat whose end was defined by the sequence at the end of the B95-8 IR1 region (AGGCCCAGCCCCCTC). Flanking sequences were then added from the published genomes (4). Finally, reads were remapped to this template with BWA, and the consensus was extracted by use of SAMtools.

RDA for identification of deleted regions. Since deletions can be spanned by low levels of contaminating reads after BWA alignment, we devised a read depth analysis (RDA) to identify sudden changes in read depth that might indicate a missed deletion. RDA calculates the ratio of read depths at adjacent nucleotides for every position in the sequence, and this can be graphed across the sequence to visualize the data. Positions where the $\log_2$ of this ratio exceeded ± 0.5 were flagged for analysis. RDA was able to confirm known deletions in nonrepetitive regions (such as the EBNA3B deletion in TRL1 [not shown]) but gave a weak signal for the repeats and a strong signal at the junction between IR1 and the nonrepetitive part of the genome. Pileup files were visualized at the position of the RDA signal by using Geneious (version 7.1.8), which allows the identification of reads that span the sequence, and hence the identification of the real sequence at the deletion.

Analyses of IR1-templated multiple-sequence alignment. Single nucleotide polymorphism (SNP) densities (excluding indels) were counted from a multiple-sequence alignment of 85 strains (Supplemental Data SD2) by use of an in-house script. Phylogenetic analysis was conducted using Geneious, with a Jukes-Cantor genetic distance model and the neighbor-joining method. Identity plots comparing a virus with a panel of other viruses were generated using T-RECs (78). This approach is scalable to an unlimited number of comparator strains, but random mutations can cause a deviation from 100% similarity that may complicate interpretation. Bootscan analysis (79) was performed using the manual bootscan method in RDP4 (v4.82) (80). The four strains used as comparators in the bootscan analysis were carefully selected from different branches of the phylogenetic tree to maximize the discriminatory power of the bootscan analysis. Use of different or additional strains substantially affected the output, so this approach was usable for visualization but not for discovery in our data set. Identity plots and bootscan analyses were generated using scanning window sizes of 200 and 400 bp and step sizes of 20 and 30 bp, respectively.

The BWRF1 ORF was calculated at its maximum possible size, allowing any non-stop codon to initiate the ORF. All ORFs exceeding 600 nucleotides (200 amino acid) were identified and are presented in Table

ST1. To define the likelihood of obtaining an ORF as long as that of BWRF1, we used the following calculation. The distribution of bases in IR1 is 12.7% A, 44% C, 28.6% G, and 14.8% T. Based on this, the cumulative chance of the next codon after any position being a stop codon is 1.31%. Therefore, the odds of an ORF starting at a specified position having a minimum length of n codons is 0.9869^n.

To estimate the size of IR1 (i.e., number of repeat units), the read depth across the W exons was compared to that for unique regions of the template (Cp) and the genome (BALF5 gene) and corrected according to Southern blot-based estimates of IR1 size for a subset of samples (4). However, we are not confident that this method is consistently reliable, as samples with low read depth often gave higher IR1 repeat numbers (not shown).

Identification of MVs. Minor variants are base calls in a sample that occur less frequently than the consensus nucleotide but are more prevalent than the background for technical error. These might typically arise from samples containing multiple virus strains or quasispecies groups or—in the case of IR1—from differences in sequence between repeat units within IR1. Reads from the previously described sequencing analysis (4) were aligned to the consensus by use of BWA, and the resultant pileup files were generated from SAMtools, version 0.1.19 (77), with outputs with a minimum Phred quality score of 20 considered minor variants. The mpileup2SNP algorithm in VarScan v2.3.7 (81) was then used to output all variants mapped on both the forward (+) and reverse (−) strands of the reference templates, using default parameters. We then filtered MVs to remove likely false-negative results. Positions consistently exhibiting MVs in all strains were assumed to represent technical errors. Such positions were identified and excluded based on having a coefficient of variation of <1. Technical errors are also often more prevalent on one strand than the other, being exacerbated by a local template structure that differs between the strands. This read bias at each MV was scored by calculating [DR(+)/VR(+)]/[DR(−)/VR(−)], where DR(+) and DR(−) are numbers of consensus-supporting reads on the plus and minus strands, respectively, while VR(+) and VR(−) are numbers of variant-supporting reads on the plus and minus strands, respectively. All MV data for all strains were visualized by plotting read bias against MV frequency, using RStudio (v0.98.1103). This initially highlighted a substantial number of MVs with very high frequencies but a strong read bias that were identified as being close to the edges of IR1—the read bias derived from the predominant mapping of these variants to reads spanning the junction between IR1 and the adjacent unique part of the genome (not shown). These were used to correct the consensuses and are regarded as SNPs. Overall, a relatively high background level—up to around 7%—of underlying MVs were seen that corresponded to interstrain SNPs. This implied some degree of sample cross-contamination or barcode exchange during multiplexing or library generation. Since genuine MVs have minimum frequencies of between 8% and 20% (since strains have between 5 and 11 repeat units in IR1), we instituted a cutoff that escalated with increasing read bias, according to the following equation: $(\sqrt{\text{read bias} \times \text{variant frequency}}) \geq 7$.

Infection of PBLs with HB9 and EBV wild-type IR1. Peripheral blood leukocytes (PBLs) were isolated from blood by use of a Ficoll gradient, washed in RPMI medium supplemented with 1% fetal calf serum (FCS), and stored in RPMI-20% FCS-10% dimethyl sulfoxide (DMSO). Frozen PBLs from four donors were recovered, seeded at 2×10^6 cells per well in a 24-well plate, and infected with 2×10^5 Raji infectious units of either HB9 or WT^w EBV (49). Medium was replaced on the following day and twice weekly thereafter. Cells were kept growing in RPMI medium supplemented with penicillin-streptomycin, L-glutamine, and 10% fetal calf serum. Cyclosporine (100 ng/ml) was added to the medium for the first 2 weeks. Proliferation was assessed by staining cells—either CD19-purified B cells or peripheral blood leukocytes—with CellTrace Violet (Life Technologies) and analyzing them at 8 days postinfection by flow cytometry as described elsewhere (49).

Western blotting and cloning of B95-8 W repeats and LP. Lymphoblastoid cell lines were seeded at a defined density (3×10^5 cells per ml) and left for 24 h before harvest. To generate the EBNA-LP size ladder, three plasmids—pSG5-LP2W (36), a gift from Andy Bell, and the C-terminally FLAG-tagged constructs pSG5LP, with four W repeats (82), and pRSP693, with 7 W exon pairs (37), both provided by Paul Ling—were transfected into 293 cells, and cells were harvested after 48 h. Cells were washed once in phosphate-buffered saline (PBS) and the pellet resuspended in RIPA lysis buffer. Protein was quantitated by use of the DC protein assay (Bio-Rad), using bovine gamma globulin as a protein standard. Thirty micrograms of protein was loaded into a 0.75-mm-thick 12.5% polyacrylamide minigel, electrophoresed in a Protean2 minigel (Bio-Rad), and then electroblotted onto Amersham Protran nitrocellulose (GE Healthcare). The membrane was blocked with powdered milk in PBS-0.5% Tween 20 and hybridized with supernatant from the JF186 hybridoma cell line (1:100). The blot was washed and then reprobed with an anti-mouse antibody conjugated to horseradish peroxidase. The blot was washed, and the chemiluminescence signal was produced by use of enhanced chemiluminescence (ECL) reagents (GE Healthcare) and developed on Amersham Hyperfilm MP. The blot was stripped and reprobed with anti-gamma tubulin as a loading control.

IR1 repeat units were cloned by digestion with BamHI into a pBR322-based plasmid. The sequence was determined using fluorescent dideoxynucleotide DNA sequencing (CSC Genomics, Hammersmith Hospital, London, United Kingdom).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JVI.00920-17.

SUPPLEMENTAL FILE 1, XLSX file, 0.1 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.2 MB.

ACKNOWLEDGMENTS

We thank Paul Farrell for suggestions and a critical reading of the manuscript, James Abbot of the Imperial College bioinformatics department for technical advice on repeat sequencing, Alan Chiang for provision of his B95-8 short-read data, Teru Kanda for his prompt reanalysis of the YCCEL1 BAC, and Darren Martin for the rapid response to our bug report on RDP4.

This work was funded primarily by a PhD studentship to M.M.B.A. from the King Abdulaziz City for Science and Technology, Saudi Arabia, and the Medical Research Council (grants MR/ L008432/1 and MR/N010388/1) to R.E.W., which supported R.D.P. and A.S. R.E.W. and A.S. were also supported by the Imperial College Trust, S.C. was supported by a Medical Research Council PhD studentship, and A.L.P., N.E.G., and P.K. were supported by the Wellcome Trust (grant 098051).

REFERENCES

1. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, Plummer M. 2012. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol 13:607–615. https://doi.org/10.1016/S1470-2045(12)70137-7.
2. Chang ET, Adami H-O. 2006. The enigmatic epidemiology of nasopharyngeal carcinoma. Cancer Epidemiol Biomarkers Prev 15:1765–1777. https://doi.org/10.1158/1055-9965.EPI-06-0353.
3. Camargo MC, Murphy G, Koriyama C, Pfeiffer RM, Kim WH, Herrera-Goepfert R, Corvalan AH, Carrascal E, Abdirad A, Anwar M, Hao Z, Kattoor J, Yoshiwara-Wakabayashi E, Eizuru Y, Rabkin CS, Akiba S. 2011. Determinants of Epstein-Barr virus-positive gastric cancer: an international pooled analysis. Br J Cancer 105:38–43. https://doi.org/10.1038/bjc.2011.215.
4. Palser AL, Grayson NE, White RE, Corton C, Correia S, Ba abdullah MM, Watson SJ, Cotten M, Arrand JR, Murray PG, Allday MJ, Rickinson AB, Young LS, Farrell PJ, Kellam P. 2015. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. J Virol 89:5222–5237. https://doi.org/10.1128/JVI.03614-14.
5. McGeoch DJ, Dolan A, Ralph AC. 2000. Toward a comprehensive phylogeny for mammalian and avian herpesviruses. J Virol 74:10401–10406. https://doi.org/10.1128/JVI.74.22.10401-10406.2000.
6. Rickinson AB, Young LS, Rowe M. 1987. Influence of the Epstein-Barr virus nuclear antigen EBNA 2 on the growth phenotype of virus-transformed B cells. J Virol 61:1310–1317.
7. Coleman CB, Wohlford EM, Smith NA, King CA, Ritchie JA, Baresel PC, Kimura H, Rochford R. 2015. Epstein-Barr virus type 2 latently infects T cells, inducing an atypical activation characterized by expression of lymphotactic cytokines. J Virol 89:2301–2312. https://doi.org/10.1128/JVI.03001-14.
8. Tzellos S, Farrell PJ. 2012. Epstein-Barr virus sequence variation—biology and disease. Pathogens 1:156–174. https://doi.org/10.3390/pathogens1020156.
9. Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, Hudson GS, Satchwell SC, Séguin C, Tuffnell PS, Barrell BG. 1984. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310:207–211. https://doi.org/10.1038/310207a0.
10. de Jesus O, Smith PR, Spender LC, Elgueta Karstegl C, Niller HH, Huang D, Farrell PJ. 2003. Updated Epstein-Barr virus (EBV) DNA sequence and analysis of a promoter for the BART (CST, BARF0) RNAs of EBV. J Gen Virol 84:1443–1450. https://doi.org/10.1099/vir.0.19054-0.
11. Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ. 2006. The genome of Epstein-Barr virus type 2 strain AG876. Virology 350:164–170. https://doi.org/10.1016/j.virol.2006.01.015.
12. Renzette N, Pokalyuk C, Gibson L, Bhattacharjee B, Schleiss MR, Hamprecht K, Yamamoto AY, Mussi-Pinhata MM, Britt WJ, Jensen JD, Kowalik TF. 2015. Limits and patterns of cytomegalovirus genomic diversity in humans. Proc Natl Acad Sci U S A 112:E4120–E4128. https://doi.org/10.1073/pnas.1501880112.
13. Sijmons S, Thys K, Mbong Ngwese M, Van Damme E, Dvorak J, Van Loock M, Li G, Tachezy R, Busson L, Aerssens J, Van Ranst M, Maes P. 2015. High-throughput analysis of human cytomegalovirus genome diversity highlights the widespread occurrence of gene-disrupting mutations and pervasive recombination. J Virol 89:7673–7695. https://doi.org/10.1128/JVI.00578-15.
14. Parsons LR, Tafuri YR, Shreve JT, Bowen CD, Shipley MM, Enquist LW, Szpara ML. 2015. Rapid genome assembly and comparison decode intrastrain variation in human alphaherpesviruses. mBio 6:e02213–14. https://doi.org/10.1128/mBio.02213-14.
15. Szpara ML, Gatherer D, Ochoa A, Greenbaum B, Dolan A, Bowden RJ, Enquist LW, Legendre M, Davison AJ. 2014. Evolution and diversity in human herpes simplex virus genomes. J Virol 88:1209–1227. https://doi.org/10.1128/JVI.01987-13.
16. Newman RM, Lamers SL, Weiner B, Ray SC, Colgrove RC, Diaz F, Jing L, Wang K, Saif S, Young S, Henn M, Laeyendecker O, Tobian AAR, Cohen JI, Koelle DM, Quinn TC, Knipe DM. 2015. Genome sequencing and analysis of geographically diverse clinical isolates of herpes simplex virus 2. J Virol 89:8219–8232. https://doi.org/10.1128/JVI.01303-15.
17. Kolb AW, Larsen IV, Cuellar JA, Brandt CR. 2015. Genomic, phylogenetic, and recombinational characterization of herpes simplex virus 2 strains. J Virol 89:6427–6434. https://doi.org/10.1128/JVI.00416-15.
18. Zell R, Taudien S, Pfaff F, Wutzler P, Platzer M, Sauerbrei A. 2012. Sequencing of 21 varicella-zoster virus genomes reveals two novel genotypes and evidence of recombination. J Virol 86:1608–1622. https://doi.org/10.1128/JVI.06233-11.
19. Lin Z, Wang X, Strong MJ, Concha M, Baddoo M, Xu G, Baribault C, Fewell C, Hulme W, Hedges D, Taylor CM, Flemington EK. 2013. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. J Virol 87:1172–1182. https://doi.org/10.1128/JVI.02517-12.
20. Lei H, Li T, Hung G-C, Li B, Tsai S, Lo S-C. 2013. Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. BMC Genomics 14:804. https://doi.org/10.1186/1471-2164-14-804.
21. Kwok H, Wu CW, Palser AL, Kellam P, Sham PC, Kwong DLW, Chiang AKS. 2014. Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples. J Virol 88:10662–10672. https://doi.org/10.1128/JVI.01665-14.
22. Simbiri KO, Smith NA, Otieno R, Wohlford EEM, Daud II, Odada SP, Middleton F, Rochford R. 2015. Epstein-Barr virus genetic variation in lymphoblastoid cell lines derived from Kenyan pediatric population. PLoS One 10:e0125420. https://doi.org/10.1371/journal.pone.0125420.
23. Kanda T, Furuse Y, Oshitani H, Kiyono T. 2016. Highly efficient CRISPR/Cas9-mediated cloning and functional characterization of gastric cancer-derived Epstein-Barr virus strains. J Virol 90:4383–4393. https://doi.org/10.1128/JVI.00060-16.
24. Tierney RJ, Kao K-Y, Nagra JK, Rickinson AB. 2011. Epstein-Barr virus BamHI W repeat number limits EBNA2/EBNA-LP coexpression in newly infected B cells and the efficiency of B-cell transformation: a rationale for the multiple W repeats in wild-type virus strains. J Virol 85:12362–12375. https://doi.org/10.1128/JVI.06059-11.
25. Rogers RP, Woisetschlaeger M, Speck SH. 1990. Alternative splicing dictates translational start in Epstein-Barr virus transcripts. EMBO J 9:2273–2277.
26. O'Grady T, Wang X, Höner Zu Bentrup K, Baddoo M, Concha M, Flemington EK. 2016. Global transcript structure resolution of high gene density genomes through multiplatform data integration. Nucleic Acids Res 44:e145. https://doi.org/10.1093/nar/gkw629.
27. Cheung A, Kieff E. 1982. Long internal direct repeat in Epstein-Barr virus DNA. J Virol 44:286–294.

28. Jones MD, Griffin BE. 1983. Clustered repeat sequences in the genome of Epstein Barr virus. Nucleic Acids Res 11:3919–3937. https://doi.org/10.1093/nar/11.12.3919.
29. Rabson M, Gradoville L, Heston L, Miller G. 1982. Non-immortalizing P3J-HR-1 Epstein-Barr virus: a deletion mutant of its transforming parent, Jijoye. J Virol 44:834–844.
30. Jeang KT, Hayward SD. 1983. Organization of the Epstein-Barr virus DNA molecule. III. Location of the P3HR-1 deletion junction and characterization of the NotI repeat units that form part of the template for an abundant 12-O-tetradecanoylphorbol-13-acetate-induced mRNA transcript. J Virol 48:135–148.
31. Jones MD, Foster L, Sheedy T, Griffin BE. 1984. The EB virus genome in Daudi Burkitt's lymphoma cells has a deletion similar to that observed in a non-transforming strain (P3HR-1) of the virus. EMBO J 3:813–821.
32. Yandava CN, Speck SH. 1992. Characterization of the deletion and rearrangement in the BamHI C region of the X50-7 Epstein-Barr virus genome, a mutant viral strain which exhibits constitutive BamHI W promoter activity. J Virol 66:5646–5650.
33. Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40:e3. https://doi.org/10.1093/nar/gkr771.
34. Moss WN, Steitz JA. 2013. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. BMC Genomics 14:543. https://doi.org/10.1186/1471-2164-14-543.
35. Finke J, Rowe M, Kallin B, Ernberg I, Rosén A, Dillner J, Klein G. 1987. Monoclonal and polyclonal antibodies against Epstein-Barr virus nuclear antigen 5 (EBNA-5) detect multiple protein species in Burkitt's lymphoma and lymphoblastoid cell lines. J Virol 61:3870–3878.
36. McCann EM, Kelly GL, Rickinson AB, Bell AI. 2001. Genetic analysis of the Epstein-Barr virus-coded leader protein EBNA-LP as a co-activator of EBNA2 function. J Gen Virol 82:3067–3079. https://doi.org/10.1099/0022-1317-82-12-3067.
37. Ling PD, Tan J, Peng R. 2009. Nuclear-cytoplasmic shuttling is not required for the Epstein-Barr virus EBNA-LP transcriptional coactivation function. J Virol 83:7109–7116. https://doi.org/10.1128/JVI.00654-09.
38. Tso KK-Y, Yip KY-L, Mak CK-Y, Chung GT-Y, Lee S-D, Cheung S-T, To KF, Lo KW. 2013. Complete genomic sequence of Epstein-Barr virus in nasopharyngeal carcinoma cell line C666-1. Infect Agents Cancer 8:29. https://doi.org/10.1186/1750-9378-8-29.
39. Xiao K, Yu Z, Li X, Li X, Tang K, Tu C, Qi P, Liao Q, Chen P, Zeng Z, Li G, Xiong W. 2016. Genome-wide analysis of Epstein-Barr virus (EBV) integration and strain in C666-1 and Raji cells. J Cancer 7:214–224. https://doi.org/10.7150/jca.13150.
40. Hastings PJ. 2010. Mechanisms of ectopic gene conversion. Genes 1:427–439. https://doi.org/10.3390/genes1030427.
41. Verma P, Greenberg RA. 2016. Noncanonical views of homology-directed DNA repair. Genes Dev 30:1138–1154. https://doi.org/10.1101/gad.280545.116.
42. Given D, Kieff E. 1979. DNA of Epstein-Barr virus. VI. Mapping of the internal tandem reiteration. J Virol 31:315–324.
43. Gottschalk S, Ng CY, Perez M, Smith CA, Sample C, Brenner MK, Heslop HE, Rooney CM. 2001. An Epstein-Barr virus deletion mutant associated with fatal lymphoproliferative disease unresponsive to therapy with virus-specific CTLs. Blood 97:835–843.
44. White RE, Rämer PC, Naresh KN, Meixlsperger S, Pinaud L, Rooney C, Savoldo B, Coutinho R, Bödör C, Gribben J, Ibrahim HA, Bower M, Nourse JP, Gandhi MK, Middeldorp J, Cader FZ, Murray P, Münz C, Allday MJ. 2012. EBNA3B-deficient EBV promotes B cell lymphomagenesis in humanized mice and is found in human tumors. J Clin Invest 122:1487–1502. https://doi.org/10.1172/JCI58092.
45. Delecluse HJ, Hilsendegen T, Pich D, Zeidler R, Hammerschmidt W. 1998. Propagation and recovery of intact, infectious Epstein-Barr virus from prokaryotic to human cells. Proc Natl Acad Sci U S A 95:8245–8250. https://doi.org/10.1073/pnas.95.14.8245.
46. Gregorovic G, Bosshard R, Karstegl CE, White RE, Pattle S, Chiang AKS, Dittrich-Breiholz O, Kracht M, Russ R, Farrell PJ. 2011. Cellular gene expression that correlates with EBER expression in Epstein-Barr virus-infected lymphoblastoid cell lines. J Virol 85:3535–3545. https://doi.org/10.1128/JVI.02086-10.
47. Lenoir GM, Vuillaume M, Bonnardel C. 1985. The use of lymphomatous and lymphoblastoid cell lines in the study of Burkitt's lymphoma. IARC Sci Publ 1985:309–318.
48. Wilson G, Miller G. 1979. Recovery of Epstein-Barr virus from nonproducer neonatal human lymphoid cell transformants. Virology 95:351–358. https://doi.org/10.1016/0042-6822(79)90490-2.
49. Szymula A, Palermo RD, Groves IJ, abdullah MB, Holder BS, White RE. 22 August 2017. Epstein-Barr virus nuclear antigen EBNA-LP is essential for transforming naive B cells, and facilitates recruitment of transcription factors to the viral genome. bioRxiv https://doi.org/10.1101/176099.
50. Hammoumi S, Vallaeys T, Santika A, Leleux P, Borzym E, Klopp C, Avarre J-C. 2016. Targeted genomic enrichment and sequencing of CyHV-3 from carp tissues confirms low nucleotide diversity and mixed genotype infections. PeerJ 4:e2516. https://doi.org/10.7717/peerj.2516.
51. Renzette N, Gibson L, Jensen JD, Kowalik TF. 2014. Human cytomegalovirus intrahost evolution—a new avenue for understanding and controlling herpesvirus infections. Curr Opin Virol 8:109–115. https://doi.org/10.1016/j.coviro.2014.08.001.
52. Ross SA, Novak Z, Pati S, Patro RK, Blumenthal J, Danthuluri VR, Ahmed A, Michaels MG, Sánchez PJ, Bernstein DI, Tolan RW, Palmer AL, Britt WJ, Fowler KB, Boppana SB. 2011. Mixed infection and strain diversity in congenital cytomegalovirus infection. J Infect Dis 204:1003–1007. https://doi.org/10.1093/infdis/jir457.
53. Bodescot M, Perricaudet M. 1986. Epstein-Barr virus mRNAs produced by alternative splicing. Nucleic Acids Res 14:7103–7114. https://doi.org/10.1093/nar/14.17.7103.
54. Sample J, Hummel M, Braun D, Birkenbach M, Kieff E. 1986. Nucleotide sequences of mRNAs encoding Epstein-Barr virus nuclear proteins: a probable transcriptional initiation site. Proc Natl Acad Sci U S A 83:5096–5100. https://doi.org/10.1073/pnas.83.14.5096.
55. Woisetschlaeger M, Yandava CN, Furmanski LA, Strominger JL, Speck SH. 1990. Promoter switching in Epstein-Barr virus during the initial stages of infection of B lymphocytes. Proc Natl Acad Sci U S A 87:1725–1729. https://doi.org/10.1073/pnas.87.5.1725.
56. Yoo LI, Woloszynek J, Templeton S, Speck SH. 2002. Deletion of Epstein-Barr virus regulatory sequences upstream of the EBNA gene promoter Wp1 is unfavorable for B-cell immortalization. J Virol 76:11763–11769. https://doi.org/10.1128/JVI.76.22.11763-11769.2002.
57. Farrell MJ, Dobson AT, Feldman LT. 1991. Herpes simplex virus latency-associated transcript is a stable intron. Proc Natl Acad Sci U S A 88:790–794. https://doi.org/10.1073/pnas.88.3.790.
58. Spivack JG, Woods GM, Fraser NW. 1991. Identification of a novel latency-specific splice donor signal within the herpes simplex virus type 1 2.0-kilobase latency-associated transcript (LAT): translation inhibition of LAT open reading frames by the intron within the 2.0-kilobase LAT. J Virol 65:6800–6810.
59. Peng R, Gordadze AV, Fuentes Pananá EM, Wang F, Zong J, Hayward GS, Tan J, Ling PD. 2000. Sequence and functional analysis of EBNA-LP and EBNA2 proteins from nonhuman primate lymphocryptoviruses. J Virol 74:379–389. https://doi.org/10.1128/JVI.74.1.379-389.2000.
60. Shaku F, Matsuda G, Furuya R, Kamagata C, Igarashi M, Tanaka M, Kanamori M, Nishiyama Y, Yamamoto N, Kawaguchi Y. 2005. Development of a monoclonal antibody against Epstein-Barr virus nuclear antigen leader protein (EBNA-LP) that can detect EBNA-LP expressed in P3HR1 cells. Microbiol Immunol 49:477–483. https://doi.org/10.1111/j.1348-0421.2005.tb03743.x.
61. Chiara M, Manzari C, Lionetti C, Mechelli R, Anastasiadou E, Chiara Buscarinu M, Ristori G, Salvetti M, Picardi E, D'Erchia AM, Pesole G, Horner DS. 2016. Geographic population structure in Epstein-Barr virus revealed by comparative genomics. Genome Biol Evol 8:3284–3291. https://doi.org/10.1093/gbe/evw226.
62. Hayward SD, Nogee L, Hayward GS. 1980. Organization of repeated regions within the Epstein-Barr virus DNA molecule. J Virol 33:507–521.
63. Feederle R, Bartlett EJ, Delecluse H-J. 2010. Epstein-Barr virus genetics: talking about the BAC generation. Herpesviridae 1:6. https://doi.org/10.1186/2042-4280-1-6.
64. Yao QY, Tierney RJ, Croom-Carter D, Cooper GM, Ellis CJ, Rowe M, Rickinson AB. 1996. Isolation of intertypic recombinants of Epstein-Barr virus from T-cell-immunocompromised individuals. J Virol 70:4895–4903.
65. Midgley RS, Blake NW, Yao QY, Croom-Carter D, Cheung ST, Leung SF, Chan AT, Johnson PJ, Huang D, Rickinson AB, Lee SP. 2000. Novel intertypic recombinants of Epstein-Barr virus in the Chinese population. J Virol 74:1544–1548. https://doi.org/10.1128/JVI.74.3.1544-1548.2000.
66. Kelly GL, Milner AE, Tierney RJ, Croom-Carter DSG, Altmann M, Hammerschmidt W, Bell AI, Rickinson AB. 2005. Epstein-Barr virus nuclear antigen 2 (EBNA2) gene deletion is consistently linked with EBNA3A, -3B, and -3C expression in Burkitt's lymphoma cells and with increased resistance to

apoptosis. J Virol 79:10709–10717. https://doi.org/10.1128/JVI.79.16.10709-10717.2005.
67. Kelly G, Bell A, Rickinson A. 2002. Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. Nat Med 8:1098–1104. https://doi.org/10.1038/nm758.
68. Kelly GL, Long HM, Stylianou J, Thomas WA, Leese A, Bell AI, Bornkamm GW, Mautner J, Rickinson AB, Rowe M. 2009. An Epstein-Barr virus anti-apoptotic protein constitutively expressed in transformed cells and implicated in Burkitt lymphomagenesis: the Wp/BHRF1 link. PLoS Pathog 5:e1000341. https://doi.org/10.1371/journal.ppat.1000341.
69. Görzer I, Niesters HGM, Cornelissen JJ, Puchhammer-Stöckl E. 2006. Characterization of Epstein-Barr virus type I variants based on linked polymorphism among EBNA3A, -3B, and -3C genes. Virus Res 118:105–114. https://doi.org/10.1016/j.virusres.2005.11.020.
70. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS. 2012. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150:831–841. https://doi.org/10.1016/j.cell.2012.05.049.
71. Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 15 September 2009. Gene conversion in evolution and disease. *In* Encyclopedia of life sciences (eLS). John Wiley & Sons Ltd, Chichester, United Kingdom. https://doi.org/10.1002/9780470015902.a0005100.pub2.
72. Chan K, Gordenin DA. 2015. Clusters of multiple mutations: incidence and molecular mechanisms. Annu Rev Genet 49:243–267. https://doi.org/10.1146/annurev-genet-112414-054714.
73. Pérez-Losada M, Arenas M, Galán JC, Palero F, González-Candelas F. 2015. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. Infect Genet Evol 30:296–307. https://doi.org/10.1016/j.meegid.2014.12.022.
74. Brown JC. 2014. The role of DNA repair in herpesvirus pathogenesis. Genomics 104:287–294. https://doi.org/10.1016/j.ygeno.2014.08.005.
75. Chuzhanova N, Chen J-M, Bacolla A, Patrinos GP, Férec C, Wells RD, Cooper DN. 2009. Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum Mutat 30:1189–1198. https://doi.org/10.1002/humu.21020.
76. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv arXiv:1303.3997 [q-bio.GN].
77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
78. Tsimpidis M, Bachoumis G, Mimouli K, Kyriakopoulou Z, Robertson DL, Markoulatos P, Amoutzias GD. 2017. T-RECs: rapid and large-scale detection of recombination events among different evolutionary lineages of viral genomes. BMC Bioinformatics 18:13. https://doi.org/10.1186/s12859-016-1420-z.
79. Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21:98–102. https://doi.org/10.1089/aid.2005.21.98.
80. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol 1:vev003. https://doi.org/10.1093/ve/vev003.
81. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. https://doi.org/10.1101/gr.129684.111.
82. Harada S, Kieff E. 1997. Epstein-Barr virus nuclear protein LP stimulates EBNA-2 acidic domain-mediated transcriptional activation. J Virol 71:6611–6618.