

**GENOME-WIDE COMPARATIVE STUDY OF EPSTEIN-BARR VIRUS GENOMIC
RECOMBINATION IN ENDEMIC BURKITT LYMPHOMA AND HEALTHY
CHILDREN FROM JOOTRH, WESTERN KENYA**

BY

AGWATI EDDY OBALA

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE IN CELL AND MOLECULAR
BIOLOGY**


SCHOOL OF PHYSICAL AND BIOLOGICAL SCIENCES

MASENO UNIVERSITY

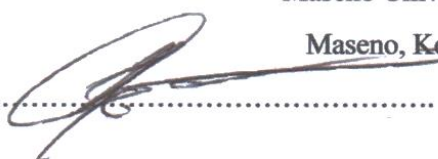
©2022


DECLARATION


I hereby declare that this thesis is mine, its contents are original and has not been submitted to any institution for any degree or academic qualification.

Agwati Eddy Obala
MSc/SC/00053/2018
Signature.......... Date.....4/2/2022.....

This thesis has been submitted for examination with our approval as supervisors:

Cyrus Ayieko, PhD
Department of Zoology,
School of Applied and Biological Sciences,
Maseno University,
Maseno, Kenya
Signature.......... Date.....18/2/2022.....

Cliff Oduor, PhD
Department of Pathology and Laboratory Medicine,
Warren Alpert Medical School,
Brown University, USA.
Signature.......... Date.....8/2/2022.....

Jeffrey Bailey, MD, PhD
Department of Pathology and Laboratory Medicine,
Warren Alpert Medical School,
Brown University, USA.
Signature.......... Date.....15/2/2022.....

ACKNOWLEDGEMENT

This thesis would have not been possible without the help and guidance of several individuals who extended their invaluable assistance in the preparation and completion of this study. My utmost gratitude goes to Dr. Cliff Oduor and Prof. Jeffrey Bailey for sponsoring my project work and for their guidance in the course of my project. I am also grateful to Dr. Cyrus Ayieko, who has been my academic supervisor and mentor. Equally, much gratitude goes to Dr. J. M. Ong'echa and Prof. Ann Moormann, who have also played vital roles as my mentors. Thanks to my colleagues at the KEMRI-UMMS lab; Erastus Kirwa, Titus K. Maina, Bonface Ariera, Joseph Nyagaya, and Sharon Akinyi for their support. Together they sacrificed much of their valuable time to discuss the project and thesis at various stages offering constructive criticism and helpful insights. Much appreciation also goes to the study participants who consented to this study, KEMRI-CGHR for housing me during this time, Maseno University, and precisely the Department of Zoology for the academic support they offered. Last but not least I thank my parents, my sisters, and one above all of us, God almighty, for giving me the strength to complete this study.

DEDICATION

This thesis is dedicated to Cayden Josiah Opollo for the joy and hopes he instills in our hearts.

ABSTRACT

Endemic Burkitt lymphoma (eBL), the most prevalent pediatric cancer in Western Kenya, is augmented by the interplay between Epstein-Barr Virus (EBV) and holoendemic malaria infection. Despite the relevance of genomic recombination on EBV genetic diversity, its genome-wide occurrence has not been characterized in genomic sequences from western Kenya hence the association with age, gender, and EBV type and eBL pathogenesis in children from western Kenya is unknown. This study, therefore, sought to: 1) characterize genome-wide occurrence of EBV genomic recombination events and breakpoints; 2) establish relationship between EBV genomic recombination events with age and gender; 3) establish relationship between genomic recombination events with EBV type and EBV genetic diversity; 4) determine the association of genomic recombination events and breakpoints with eBL. This study employed a case-control design using 86 archival samples involving 54 children diagnosed with eBL and 32 geographically matched healthy children previously collected from Jaramogi Oginga Odinga Teaching and Referral Hospital (JOORTH) that met the inclusion criteria. DNA was extracted and sequenced in the Illumina Sequencing Kit. Whole-genome multiple sequence alignment, recombination analyses, and phylogenetic inferencing were done using MAFFT, RDP4, and MEGA X software respectively. Wilcoxon rank test compared the occurrence of genomic recombination between genomic sequences of the; males and females, type 1 and type 2, and between the eBLs and the healthy. Univariate and multivariate logistic regression modeled eBL association with genomic recombination events and their breakpoints. This study identified 28 genomic recombination events present in 82.6% of the EBV genomes analyzed with most breakpoints reported in genes of the lytic phase. There was no significant difference across the age groups ($p=0.68$) and between males and females ($p=0.59$). Type 1 genomic sequences reported more genomic recombination events ($p=6.4e-06$). The EBV genomic sequences clustered on the neighbor-joining (NJ) phylogenetic tree by genomic recombination events suggesting an association with EBV genetic diversity. Genomic sequences from the eBLs reported more genomic recombination events compared to the healthy ($p=0.037$). Further, recombination breakpoints cutting through; *BRLF1*, *BZLF1*, *BDLF3.5*, *BDLF4*, *LMP2A*, *LMP2B*, and *EBNA2* genes were found enriched in genomic sequences from the eBLs. Evidence from this study suggests that there is minimal accrual of genomic recombination events with infections over time hence these genomic recombination events are most likely transferred vertically down EBV genome lineages. Type 1 EBV genomic sequences and those from the eBLs have accumulated more genomic recombination events pointing to the availability of factors that increases the propensity for genomic recombination in these set of genomic sequences. For future studies, long-read sequencing and improved EBV DNA enrichment methods should be employed to generate complete EBV sequences to allow the analysis of whole EBV genomes. In summary, this study addresses the complexities that underlie genomic recombination as a source of genetic variation in EBV, findings that contribute significantly to the pool of knowledge on EBV genetic diversity and its contribution to disease.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION	ii
ACKNOWLEDGEMENT	iii
DEDICATION	iv
ABSTRACT	v
TABLE OF CONTENTS	vi
ABBREVIATIONS AND ACRONYMS	ix
DEFINITION OF KEY TERMS	xii
LIST OF FIGURES	xiii
LIST OF TABLES	xiv
CHAPTER ONE INTRODUCTION	1
1.1 Background Information	1
1.2 Statement of the Problem	5
1.3 General Objective	6
1.3.1 Specific Objectives	6
1.3.2 Null Hypotheses	6
1.5 Significance of the Study	7
CHAPTER TWO:LITERATURE REVIEW	9
2.1 EBV Infection and its Life Cycle	9
2.2 Epstein-Barr Virus Genomic Variation	12
2.2.1 Genomic Recombination as Source of Genomic Variation	12
2.2.2 Molecular Mechanism of EBV Genomic Recombination	15
2.3 Age, Gender and EBV Genomic Recombination	20
2.4 EBV Type-Associated Genomic Recombination Events and EBV Diversity	22
2.5 Genomic Recombination Events and eBL Pathogenesis	24
2.5.1 <i>Plasmodium falciparum</i> and EBV Infection Augments eBL	24
2.5.2 EBV Genomic Recombination Events and eBL Pathogenesis	26
CHAPTER THREE:METHODOLOGY	28
3.1 Study Design and Site	28
3.2 Study Population	29
3.2.1 Inclusion Criteria	30
3.2.2 Exclusion Criteria	30
3.3 Sample Size Determination	31

3.4 Sample Processing and Storage	32
3.5 EBV Genotyping	32
3.6 Sequencing Library Preparation, EBV Specific Genome-wide Amplification, and EBV DNA Enrichment	33
3.7 Sequence Reads Pre-processing and de novo Genome Assembly	34
3.8 Multiple Sequence Alignment	35
3.9 Removal of Poorly Aligned Regions	35
3.10 Phylogenetic Analysis	36
3.11 Recombination Analyses	36
3.12 Genomic Feature Annotation	37
3.13 Data Analysis	37
3.14 Ethical Approval	38
CHAPTER FOUR:RESULTS	39
4.1 Demographic Characteristics of Study Participants	39
4.2 Genome-wide occurrence of Genomic Recombination Events and Breakpoints	40
4.3 Occurrence of Genomic Recombination across Age Groups, and between Males and Females	44
4.4 Association of Genomic Recombination Events with EBV Types and Diversity	47
4.5 Genomic Recombination Events in the Genomic Sequences from the eBLs and Healthy Participants	51
CHAPTER FIVE:DISCUSSION	56
5.1 General Introduction	56
5.2 Genome-wide occurrence of Genomic Recombination Events and Breakpoints	56
5.3 Occurrence of Genomic Recombination across Age Groups, and between Males and Females	60
5.4 Association of Genomic Recombination Events with EBV Types and Diversity	61
5. 5 Genomic Recombination Events in the Genomic Sequences from the eBLs and Healthy Participants	64
5.6 Study Limitations	66
CHAPTER SIX:SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS	68
6.1 Summary of Study Findings	68
6.2 Conclusions	68
6.3 Recommendations from this Study	69
6.4 Recommendations for Future Studies	70
REFERENCES	71
APPENDICES	86

Appendix I: KEMRI-SERU Approval	86
Appendix II: eBL Participant Consent Form	88
Appendix III: Maseno University School of Graduate Studies Approval	92
Appendix IV: Statistical Power Test	93
Appendix V: Genotyping Primers and Probe Sets	94
Appendix VI: EBV-Specific Genome Wide Amplification Primer Sets	95
Appendix VII: Genomes' Pre-processing Information and Sequencing Statistics	98
Appendix VIII: Genomic Recombination in Plasma-Tumor Replicates	103
Appendix IX: Summary of Key R Scripts used in the Data Analysis	104

ABBREVIATIONS AND ACRONYMS

ABACAS:	Algorithm Based Automatic Contiguation of Assembled Sequences
AID:	Activation Induced Deaminase
BART:	BamHI Rightward Transcript
BCL2L11:	B Cell Ligand 2 Like 11
BHRF1:	Bam HI-H Rightward Fragment 1
C terminal:	Carboxyl terminal
CCXR:	Coupled Chemokine Receptor
CD 21:	Cluster of Differentiation 21
CDKN2A:	Cyclin-Dependent Kinase Inhibitor 2A
CDS:	Coding Sequence
DNA:	Deoxyribonucleic Acid
EBER:	Epstein-Barr Virus Encoded RNA
eBL:	Endemic Burkitt Lymphoma
EBNA:	Epstein-Barr Nucleotide Antigen
EBNA-LP:	Epstein-Barr Nuclear Antigen-L Protein
EBV:	Epstein-Barr Virus
EDTA:	Ethylenediaminetetraacetic Acid
ENA:	European Nucleotide Archive
ERC:	Ethical Review Committee
GC:	Germinal Centre
GMRCA:	Grand Most Recent Common Ancestor
GP42/350:	Glycoprotein 42/350
HCMV:	Human Cytomegalovirus

HHV4:	Human Herpes Virus 4
HIV:	Human Immunodeficiency Virus
HSV1:	Human Simplex Virus 1
IgH/L:	Immunoglobulin Heavy/ Light
IGV:	Integrative Genome Viewer
IM:	Infectious Mononucleosis
IMAGE:	Iterative Mapping and Assembly for Gap Elimination
IR:	Internal Repeat
JOORTH:	Jaramogi Oginga Odinga Teaching and Referral Hospital
Kbp:	Kilobase pairs
KEMRI:	Kenya Medical Research Institute
LCL:	Lymphoblastic Cell Line
LCV:	Lymphocryptovirus
LMP:	Latent Membrane Protein
MAFFT:	Multiple Alignment by Fast Fourier Transform
MCMV:	Murine Cytomegalovirus
MEGA:	Molecular Evolutionary Genetics Analysis
MHC:	Major Histocompatibility Complex
miRNA:	Micro RNA
MSA:	Multiple Sequence Alignment
NEB:	New England Biology
NJ:	Neighbour Joining
NPC:	Nasopharyngeal Carcinoma
OR:	Odds Ratio

ORFs:	Open Reading Frames
Pf:	<i>Plasmodium falciparum</i>
RBP-JK:	Recombination Binding Protein-Janus Kinase
RNA:	Ribonucleic Acid
RPD4:	Recombination Detection Program 4
SERU:	Scientific and Ethical Review Committee
SGS:	School of Graduate Studies
SNP:	Single Nucleotide Polymorphism
SSA:	Sub-Saharan Africa
sWGA:	Specific Genome Wide Amplification
TP 53:	Tumor Protein 53
TR:	Terminal Repeat
UL:	Unique Region
UMMS:	University of Massachusetts Medical School
USA:	United States of America

DEFINITION OF KEY TERMS

EBV Type:	A classification of EBV genomic sequences based on the variations in the EBV typing genes
Genetic Diversity:	The proportion of polymorphic loci across the genome
Genomic Recombination:	The exchange of genomic segments between genomes that co-infect a cell
Intertypic Recombination:	Exchange of genetic segments between EBV genes used to classify EBV into type 1 and type 2
Phylogenetic Clade:	A group of isolates descending from a common ancestor or ancestral node
Recombination Breakpoints:	Sites along the genomic sequence that are cut to allow the incorporation of a genetic fragment during a genomic recombination event
Recombination Event:	A distinct incorporation of a unique genomic segment into a genome

LIST OF FIGURES

Figure 2.1: Epstein–Barr virus (EBV) life cycle in healthy carriers.	12
Figure 2.2: Replication and Genomic Recombination Steps in Herpesvirus	17
Figure 3.1: Distribution of eBL patients in western Kenya	29
Figure 4.1: Frequency of Genomic Recombination Events	41
Figure 4.2: Positions of Recombination Breakpoints along EBV Genome	42
Figure 4.3: Genomic Recombination Breakpoints Distribution in CDS	44
Figure 4.4: Occurrence of Genomic Recombination across Age Groups and between Males and Females	46
Figure 4.5: Recombination Patterns between Type 1 and Type 2	48
Figure 4.6: Phylogenetic Tree of EBV Genomic Sequences showing Diversity related to Genomic Recombination Events	50
Figure 4.7: Genomic Recombination Events in the Genomic Sequences from the eBLs and Healthy Participants	52

LIST OF TABLES

Table 4.1: Demographic Characteristics of Study Participants.	40
Table 4.2: Participant's Characteristics associated with Genomic Recombination	45
Table 4.3: Genomic Recombination Events association with EBV Types	47
Table 4.4: eBL Association with two Genomic Recombination Events	55

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Epstein-Barr Virus (EBV), also called human herpesvirus 4 (HHV4), is a ubiquitous gamma-herpesvirus in the family of primate lymphocryptovirus (LCVs) (Moukassa *et al.*, 2018). This virus assumes a biphasic life cycle shuttling between the lytic and the latent phase (Murata & Tsurumi, 2014). EBV infects the human epithelial cells initiating the lytic phase which is characterized by the sequential expression of lytic-associated genes, lytic replication of the EBV genome, and release of infectious viral particles into saliva, the main route of transmission (Rosemarie & Sugden, 2020). To activate the latent phase, the virus infects the B cells leading to the expression of the latency-associated genes, migration into the germinal centers (GC), and transit into peripheral blood where they persist lifelong in a latent state (Kanda, 2018; Kang & Kieff, 2015). Occasionally, the latent virus can intermittently reactivate to re-initiate epithelial cell infection and increase saliva viral shedding (Li *et al.*, 2016; McKenzie & El-Guindy, 2015). While over 90% of the global human population carry the virus lifelong asymptotically (Smatti *et al.*, 2017; Thorley-Lawson *et al.*, 2013), the infection is associated with over 1% of global human cancers. In sub-Saharan Africa (SSA), EBV along with chronic *Plasmodium falciparum* (*Pf*) malaria infections, is associated with an increased prevalence of a pediatric cancer of the B cell known as endemic Burkitt Lymphoma (eBL) (Hämmerl *et al.*, 2019; Stefan *et al.*, 2017). Although studies have provided critical insights into key concepts that underlie EBV virology, and oncology, its pattern of genetic variation and the influence of these patterns on EBV-associated malignancies such as eBL largely remain to be studied.

The EBV genome measures approximately 172kb and has at least 86 open reading frames (ORFs) (Tzellos & Farrell, 2012). Nine ORFs encode the key latent proteins including Epstein-Barr Nuclear Antigen (EBNA) -1, EBNA-2, EBNA 3A, -3B, -3C, EBNA-LP, Latent Membrane Protein (LMP) -1, LMP2A, and -2B (Kanda, 2018). Other ORFs encode capsid proteins, transcriptional factors, lytic proteins as well as non-coding Ribonucleic acids (RNAs) (Kang & Kieff, 2015). EBV genomic variation results from the joint processes of point mutation and genomic recombination (Telford *et al.*, 2020; Tzellos & Farrell, 2012). Point mutation is an alteration in a single nucleotide in the Deoxyribonucleic acid (DNA) molecule that makes up the genome of an organism leading to single nucleotide polymorphism (SNPs) (Sanjuán & Domingo-Calap, 2016). Genomic recombination on the flipside is the exchange of genomic segments between genomes that co-infect a cell (Pérez-Losada *et al.*, 2015). As a consequence of genomic variation, EBV is genotyped as type 1 and type 2 based on deep-seated divergence in variations in the *EBNA 2* gene and EBNA 3 family of genes (Tzellos & Farrell, 2012). Recombination has the potential to combine genetic variations that existed separately in different genomes and this may have a dramatic effect on the diversity of the resulting recombinant genome (Pérez-Losada *et al.*, 2015; Sijmons *et al.*, 2015). Further, genomic recombination in EBV is 2.5-fold more likely to occur compared to mutations hence over time the virus accumulates more genomic recombination events compared to mutations (Santpere *et al.*, 2014a; Zanella *et al.*, 2019). Despite the relevance of genomic recombination on EBV genetic diversity, the occurrence has not been characterized in genomic sequences from western Kenya.

The occurrence of genomic recombination in DNA viruses such as EBV may be influenced by demographic factors including age and gender (Martin, 2015). This assertion can be attributed to the fact that age and gender may influence human immune responses to both EBV and *Pf* due to the differences in sex-related chromosomes and hormones (van Lunzen &

Altfeld, 2014). Repeated exposure to uncomplicated asymptomatic *Pf* infections causes intermittent reactivation of EBV infected B cells in peripheral blood circulation, leading to their replication, increase in number, and consequently elevated EBV loads (Chattopadhyay *et al.*, 2013; Reynaldi, Schlub, Chelimo, *et al.*, 2016). High viral loads raise the chances of viral genomes exchanging genetic fragments between genomes of viruses (Martin, 2015; Prata *et al.*, 2015). Younger children below 5 years in western Kenya report high *Pf* densities, as well as high morbidity and mortality to parasitic infection (Redmond *et al.*, 2020; Snider *et al.*, 2012). Consequently, these younger children have high EBV loads compared to their older counterparts (Njie *et al.*, 2009) hence genomic fragments from their EBV isolates are more likely to recombine. Similarly, the females mount better immune responses to viral infections (Ballesteros-Zebadúa *et al.*, 2013; Domínguez-Rodríguez *et al.*, 2021) hence they have lower viral loads. In tandem with eBL occurrence where EBV is a key aetiological agent, the females are less likely to develop eBL compared to males (Mwanda, 2004; Rainey *et al.*, 2007) and this can be explained by the differences in how their immune system control EBV. Since the host immunity to EBV and *Pf* may impact the exchange of genetic fragments between genomes and age as well as gender influence immune responses to EBV and *Pf*, it would be interesting to find out if the occurrence of genomic recombination in EBV differs across different age groups and between the males and the females.

EBV type 1 or type 2 classification is a major feature of EBV genomic variation defined exclusively by the variations in the EBV typing genes i.e. *EBNA 2* gene and *EBNA 3* family of genes (*EBNA 3A*, *3B*, and *3C*) (Tzellos & Farrell, 2012). EBV genomes from western Kenya are characterized by the extensive presence of both EBV type 1 and type 2 (Kaymaz *et al.*, 2020) unlike other geographical regions where only one EBV type dominates (Neves *et al.*, 2017). This offers a good opportunity for multiple EBV-type infections, a factor that may augment the occurrence of genomic recombination. Comparison of the tumorigenicity of EBV

type 1 and type 2 show that EBV type 1 is better at immortalizing B cells since its EBNA2 carboxyl (C) terminal region greatly induces the expression of LMP 1 and Coupled Chemokine Receptor (CCXR) genes required for uncontrolled B cell proliferation (Lucchesi *et al.*, 2008; Wang *et al.*, 2012). Recently, EBV type 1 was associated with the development of eBL (Kaymaz *et al.*, 2020). This differential tumorigenicity sparks the interest to find out if the occurrence of genomic recombination differs between EBV type 1 and type 2 genomic sequences. The construction of the phylogenetic relationship of EBV genomes normally clusters the genomes distinctly as either EBV type 1 or type 2 representing the underlying EBV diversity (Chen *et al.*, 2018; Santpere *et al.*, 2014). Besides the first split in EBV phylogeny which is based on EBV types, the phylogeny also forms clusters of genomes further picturing the nature of diversity seated within these genomes (Kaymaz *et al.*, 2020; Zanella *et al.*, 2019). It is however not known if such clustering genomes on the phylogenetic tree are influenced by genomic recombination.

EBV infection of the B cells results in uncontrollably proliferating B cells, with potential oncogenicity which should be eliminated by the host immune response before they cause morbidity and even mortality (Moormann & Bailey, 2016). As the host applies innate and adaptive immune mechanisms to clear the virus, the virus adapts ways to evade the host immune surveillance (Thorley-Lawson *et al.*, 2013). Through genomic recombination, the virus may acquire beneficial traits and fitness giving it an advantageous edge over the host and augmenting the risk of host progression into disease (Combela *et al.*, 2011; Sijmons *et al.*, 2015). However, to the best of our knowledge, no specific genomic recombination events have been associated with the risk to eBL. Genomic recombination is able to change the immunogenic determinants of the viral proteins, especially when they affect the protein-coding sequences providing a common immune route for escape from the host (Zanella *et al.*, 2019). A study in Human Simplex Virus 1 (HSV1) revealed recombination breakpoints in latency-

associated genes which were associated with better capabilities to evade host immune surveillance (Lee *et al.*, 2015). Further, Berenstein *et al.* (2018) reported a highly variable landscape of recombination breakpoints in EBV genomes from other geographical regions but could not associate these patterns with any disease outcome. Despite the available evidence of genomic recombination in EBV, no patterns have been associated with EBV's biology of B cell transformation and pathogenesis of eBL.

1.2 Statement of the Problem

EBV is a class I carcinogen with known association with over 1% of the global human cancer cases. The virus has been isolated in almost all eBL tumors from western Kenya suggesting a necessary contribution which still needs to be clearly established. EBV and *Pf* infections work synergistically to augment eBL development necessitating case-control studies that investigate the two infections as co-etiological agents. Efforts have been made to understand the virology and oncology of EBV but there still missing links on the exact contribution of EBV genetic variation in eBL. With a focus on genomic recombination as a source of genetic variation, no studies have characterized their patterns of occurrence in genomes from western Kenya. Consequently, is not known whether genomic recombination patterns differ across age groups or between males and females. This gap in knowledge may significantly affect the management of EBV-associated risk to eBL in western Kenya since age and gender are key demographic factors in such epidemiological surveillance. Despite the observation that EBV type 1 better immortalizes B cells and is associated with eBL, no studies have compared genomic recombination between EBV types. This implies an inadequate understanding of how genomic recombination influences the tumorigenicity and the pathogenic potential of the EBV types. Further, genomic recombination can create novel genetic variations

which may augment eBL, and this needs to be investigated since no studies have provided evidence for or against this hypothesis. In a nutshell, eBL is still the most prevalent pediatric cancer in western Kenya causing significant morbidity and mortality in children hence research efforts are required to improve the understanding of genomic recombination as a source of genetic diversity in EBV and as a possible aetiological factor in eBL pathogenesis.

1.3 General Objective

To investigate the genome-wide occurrence of EBV genomic recombination, their relationship with age, gender, EBV type, and diversity, and establish their association with eBL pathogenesis.

1.3.1 Specific Objectives

1. To characterize genome-wide occurrence of EBV genomic recombination events and breakpoints.
2. To establish relationship between EBV genomic recombination events with age and gender.
3. To establish relationship between genomic recombination events with EBV type and EBV genetic diversity.
4. To determine the association of genomic recombination events and breakpoints with eBL.

1.3.2 Null Hypotheses

1. Genome-wide occurrence of genomic recombination events and breakpoints in EBV is homogenous.
2. Genome-wide occurrence of genomic recombination events in EBV do not across age groups and between males and females.

3. Genome-wide occurrence of genomic recombination events is the same in EBV type 1 and type 2 and has no impact on EBV diversity.

There is no genome-wide association of genomic recombination events and breakpoints with eBL pathogenesis

1.5 Significance of the Study

eBL causes significant morbidity and mortality in children from western Kenya, with the high incidences augmented by the interplay of *Pf* malaria and EBV infection. Despite EBV genetic variation bearing the potential to contribute to disease, genome-wide examination of EBV genomes in case-control of studies are few due to the low EBV loads in the healthy individuals preventing direct sequencing of the virus. To bridge this gap, this study used EBV-specific genome amplification to enrich for viral DNA and generate EBV genomic sequences from eBL patients and geographically matched healthy controls. This approach allowed for a properly controlled investigation of EBV genomic recombination. Further, the EBV genomes generated from western Kenya were both type 1 and type 2 giving an opportunity to study type-associated genomic recombination and the association with overall EBV genetic diversity. To the best of our knowledge, this is the first genome-wide comparative study of genomic recombination in EBV type 1 and EBV type 2 genomes and in the genomes of eBLs and healthy controls.

This study has characterized the genome-wide occurrence of genomic recombination and reported a heterogeneous landscape of recombination with some genomic regions being more prone to genomic recombination breakpoints. It also demonstrates different patterns of genomic recombination in EBV type 1 and types 2 genomes, thereby shedding more light on the question of differential tumorigenicity of EBV types. It further identifies genomic

recombination events associated with EBV types and enriched in the eBLs. The genomic recombination breakpoints enriched in the eBLs were mapped to cut genes i.e. *BRLF1*, *BZLF1*, *BDLF3.5*, *BDLF4*, *LMP2A*, *LMP2B*, and *EBNA2* which play critical roles in EBV's biology of B-cell transformation and eBL pathogenesis. This information adds to the pool of knowledge about the patterns of EBV genetic variation which may be linked to disease and are also important in the epidemiological surveillance of factors that augment the risk of developing EBV-associated cancers. Lastly, this study provided a bioinformatics workflow that may be used in future investigations of genomic recombination in other genomes.

CHAPTER TWO

LITERATURE REVIEW

2.1 EBV Infection and its Life Cycle

EBV belongs to the family of the gammaherpesvirus (Tzellos & Farrell, 2012) and is ubiquitous with human beings as the only natural hosts (Farrell, 2019). Human beings normally contract EBV through infectious saliva (Murray & Young, 2002), though there is the possibility of spreading through organ transplantation (Lau *et al.*, 2017; Le *et al.*, 2017), blood tissue transfusion (Trottier *et al.*, 2010), breast milk (Perera *et al.*, 2010), and cervical secretions (Berntsson *et al.*, 2013). While more than 90% of the world's adult population harbor EBV infection asymptomatically (Rochford, 2009), this virus is associated with over 1% of the world's human cancers (Farrell, 2019; Moukassa *et al.*, 2018) as well as other non-malignant conditions such as infectious mononucleosis (IM) (Thorley-Lawson *et al.*, 2013). Children in SSA contract EBV before they are 3 years of age with 35% of children in lowland parts of western Kenya contacting EBV before they are 6 months (Piriou *et al.*, 2012). Early exposure to EBV among children from western Kenya may predispose them to the risk of developing EBV-associated malignancies such as eBL (Piriou *et al.*, 2012; Reynaldi, Schlub, Piriou, *et al.*, 2016). This virus assumes a biphasic life cycle where it shuttles between lytic and latent phases (Murata & Tsurumi, 2014). The life cycle of EBV is summarized in Figure 2.1.

During primary EBV infection, EBV contacts the tonsillar epithelium cells initiating the lytic phase of the infection (Rosemarie & Sugden, 2020). The EBV lytic phase involves sequential expression of the lytic-associated genes resulting in three phases i.e. immediate-early phase, early phase, and late phase (Swaminathan & Kenney, 2008). The expression of immediate early genes i.e. *BRLF1* and *BZLF1* are necessary for the initiation of the EBV lytic phase (McKenzie & El-Guindy, 2015). Further, *BRLF1* and *BZLF1* genes also encode

transcriptional factors Rta and Zta respectively which are required for the transcription of the genes of the early lytic phase (Li *et al.*, 2016). The transcription and expression of early lytic genes result in proteins such as the viral DNA polymerase which make up the EBV replication machinery (Murata & Tsurumi, 2014). Further, the early lytic genes activate the expression of late lytic genes which yield proteins that make up the mature lytic virions (Swaminathan & Kenney, 2008). These infectious viral particles are released into saliva leading to infection of new cells vital for intra and inter-host propagation (Rosemarie & Sugden, 2020). Moreover, the lytic phase is critical for replication of the EBV genome occasioned by DNA double-strand (ds) opening and disruption and subsequent amplification of the viral genome within the host (Rosemarie & Sugden, 2020). Although all EBV-associated cancers involve EBV latent phase, the viral lytic phase contributes to the development and maintenance of these malignancies for instance through the induction of growth factors and the production of oncogenic cytokines (Li *et al.*, 2016).

EBV latent phase is activated when the virus infects the naïve B cells of the tonsils via the major protein envelope glycoprotein (gp) 350 binding to Cluster of Differentiation 21 (CD21) molecule found on the surface of these B cells (Thorley-Lawson *et al.*, 2013). The infected B cells migrate to the lymph node follicles where they initiate a germinal center reaction (GC) through the expression of EBV latency-associated genes (Kempkes & Robertson, 2015). The latency-associated genes include six EBV nuclear antigens (EBNA1, 2, 3A, 3B, 3C and the leader protein), three latent membrane proteins (LMP1, 2A, and 2B), two small EBV encoded RNAs (EBER1 and 2), and microRNA transcripts from the BamHI A rightward transcript (BART) (Kang & Kieff, 2015). The expression of these genes characterizes the “latency III” program which is required for the proliferation of the infected B cells (Murata *et al.*, 2021). After this comes the “latency II” program in which only *EBNA-1*, *EBERs*, *BARTs*, *LMP-1*, and *LMP-2A* genes are expressed and provide survival signals for the

infected B cells to migrate out of the GC into the memory B cell pool (Kempkes & Robertson, 2015). Their transit through GC is followed by a gradual shutdown of viral gene expression leading to either the “latency I program” where only the *EBNA1* gene is expressed or the “latency 0” program where no viral antigen is expressed. Shut down of genes expression is responsible for the persistence of infected B blasts in the memory B cell population (Thorley-Lawson *et al.*, 2013). Viral latent gene products are key contributors to EBV-mediated B cell transformation and are likely to play a role in lymphomagenesis (Kanda, 2018).

Occasionally the latently infected memory B lymphocytes can be triggered to migrate back to the tonsils where they initiate a new viral cycle in a process called reactivation (McKenzie & El-Guindy, 2015). Latently infected B cells are likely to reactivate to lytic replicating form in response to factors such as *Pf* infection (Daud *et al.*, 2015) and new viral infections (Murata *et al.*, 2021). EBV reactivation triggers lytic replication which is necessary for EBV genomic recombination (Pérez-Losada *et al.*, 2015). Further, it promotes genomic instability leading to genomic breaks which are efficient facilitators of genomic recombination (Wang *et al.*, 2016; Wu *et al.*, 2010). Moreover, reactivation encourages invasiveness which is essential for tumorigenesis (Li *et al.*, 2016). Although studies of EBV have provided critical insights into key concepts that underlie its virology and oncology, to date there is still an incomplete picture of the pattern and nature of viral variation and its impact on the risk of EBV-associated diseases.

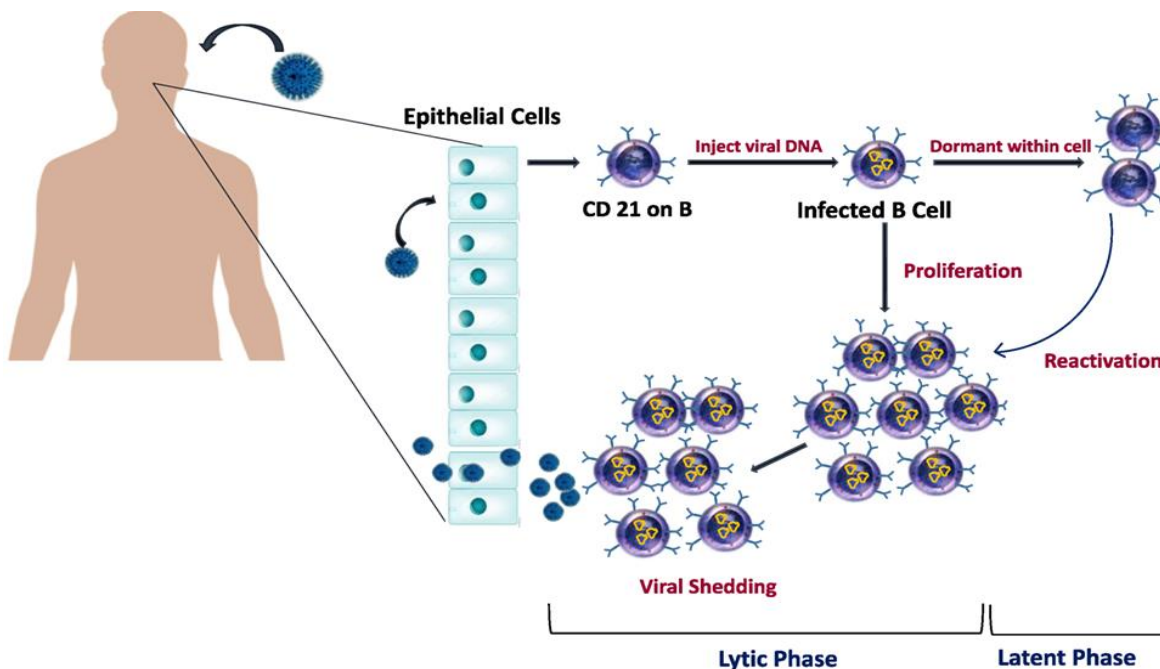


Figure 2.1: Epstein–Barr virus (EBV) life cycle in healthy carriers (Smatti *et al.*, 2018). EBV infection begins when it infects the epithelial cells and naïve B cells of the oropharynx initiating the latent and the lytic phases respectively. It integrates its genome into the B cell nucleus where it will replicate and result in the proliferation of B cells. Later, as cells recirculate between peripheral and oral compartments, resting B cells may be reactivated to induce viral shedding.

2.2 Epstein-Barr Virus Genomic Variation

2.2.1 Genomic Recombination as Source of Genomic Variation

EBV genome is approximately 172kb and has at least 86 protein-coding genes. EBV genome, though linear, assumes a circular episome within the host cell nucleus (Tzellos & Farrell, 2012). The genome is made up of a long unique region that consists of four major internal repeats (IR1 to IR4) and terminal repeats (TR) (Sample *et al.*, 2009). The genes encode 9 latent proteins including Epstein-Barr nuclear antigen 1 (EBNA-1), EBNA-2, EBNA-3A, -3B, -3C, EBNA-LP and Latent membrane protein 1 (LMP-1) and LMP2A, -2B (Kanda, 2018). Other genes encode capsid proteins, transcriptional factors as well as lytic proteins that play various roles in the lytic phase of the virus (Swaminathan & Kenney, 2008). Besides the protein-coding genes, the EBV genome also encodes non-poly-adenylated hence non-coding EBV RNAs such as the Epstein-Barr virus-encoded small RNA1 (EBER1) and 2 (EBER2),

microRNA transcripts from the BamHI A region (miRNAs-BARTs), and the Bam HI-H rightward fragment 1-derived miRNAs (miRNAs-BHRF1) (Tzellos & Farrell, 2012).

Genomic diversity in EBV is defined as the proportion of polymorphic loci across the virus genome (Kanda *et al.*, 2019). Genomic variation in EBV occurs majorly via point mutations and genomic recombination (Kanda *et al.*, 2019). Point mutation is defined as an alteration in a single nucleotide pair in the DNA molecule that makes up the genome of an organism (Sanjuán & Domingo-Calap, 2016). Such alteration in nucleotide states may result from a deletion, an insertion, or a substitution of a nucleotide leading to single nucleotide polymorphisms (SNPs) (Kupczok *et al.*, 2018). Genomic recombination, however, is the exchange of genetic segments between two genomes co-infecting the same cell resulting in a recombinant or a chimeric genome (Froissart *et al.*, 2005; Pérez-Losada *et al.*, 2015). By exchanging genetic segments the virus gains a novel variant profile which is a combination of the inherent genetic variations drawn from the parental genomes (Froissart *et al.*, 2005). The recombinant genome, therefore, has a greater genetic diversity than the parental genomes (Pérez-Losada *et al.*, 2015). Over time, genomic recombination provides a means through which the virus accumulates beneficial genetic traits to improve its fitness and to outwit the human host (Berenstein *et al.*, 2018; Lee *et al.*, 2015; Sijmons *et al.*, 2015). Sometimes, genomic recombination may dramatically combine SNPs that existed separately in different genomes (Pérez-Losada *et al.*, 2015). A single genomic recombination event, therefore, bears a greater effect on the overall genetic diversity of EBV when compared to a single mutational event. Zanella *et al.*, (2019) and Santpere *et al.*, (2014) both show that the EBV genome is 2 fold more likely to experience genomic recombination events compared to mutational events. These findings emphasize the relevance of EBV genomic recombination of overall genetic diversity.

A few studies have investigated the occurrence of genomic recombination in EBV. Palser *et al* (2015) analyzed a total of 83 strains and reported two intertypic recombinants. Kaymaz *et al.*, (2020) also analyzed 98 EBV genomes from eBL cases and healthy controls residing in western Kenya and reported 3 intertypic recombinants with similar patterns to those detected by Palser *et al.*, (2015). This study focussed on the exchange of genetic segments between EBV genes used to classify the virus into type 1 and type 2 i.e. EBNA2 and EBNA 3 family of genes (3A, 3B, 3C) (Neves *et al.*, 2017). The intertypic recombinants presented with a combination of type 1 EBNA2 sequence segments and type 2 EBNA3 sequence segments (Palser *et al.*, 2015). These recombinant genomes therefore could not be classified as type 1 or as type 2 but as intertypic recombinants. These studies despite showing that EBV genomes drawn from western Kenya (Kaymaz *et al.*, 2020) and from other geographical regions (Palser *et al.*, 2015) can exchange genetic segments, did not characterize the genome-wide occurrence of genomic recombination in the virus. Berenstein *et al.*, (2019) characterized the rates of genomic recombination along the entire EBV genome and reported a highly variable landscape. This study confirmed that diversity in EBV is impacted by mechanisms such as recombination, which extend beyond the usual consideration of point mutations as the only source of genetic diversity. It further emphasizes the need to understand the genome-wide occurrence of genomic recombination in the context of EBV-associated conditions such as eBL (Berenstein *et al.*, 2018). In this case, clinical and geographical records of study participants should be considered to investigate the association of genomic recombination with different clinical conditions, an approach which to the best of the study knowledge has not been utilized. As a first attempt to bridge this gap, this study characterized the genome-wide occurrence of genomic recombination in EBV genomic sequences obtained from eBL cases and healthy controls from western Kenya.

2.2.2 Molecular Mechanism of EBV Genomic Recombination

Genomic Recombination in EBV is intimately linked to the replication of its genome in the human host nucleus (Pérez-Losada *et al.*, 2015). The replication and recombination steps in herpesviruses are summarized in figure 2.2. Pre-replication steps begin during the lytic phase of EBV infection as soon as the virus enters the tonsillar epithelium cells (Hammerschmidt & Sugden, 2013). After infection, the linear viral DNA integrates into the host cell nucleus and is rapidly converted into a circular form by human DNA Ligase IV (Kenney, 2007). The replication process is initiated by a Unique region (UL9) binding to the origin of replication (*ori*) disrupting the double-stranded (ds) DNA structure (Hammerschmidt & Sugden, 2013). The single-strand binding protein, UL29 is then recruited leading to the formation of the UL9-UL29 complex (Weller & Coen, 2012). Next, the primase complex (UL5/UL18/UL52), in the presence of UL29, unwinds the DNA duplex and synthesizes the short RNA primers fundamental for DNA replication. The action of DNA polymerase (UL30/UL42) completes the replicative complex (Weller & Coen, 2012).

Replication begins via a bi-directional theta-type replication model which is required for the initial amplification of the viral DNA (Weller & Coen, 2012). This model involves the origin of replication and the DNA is replicated in two directions away from the origin of replication (Ueda, 2018). The replication model then switches to switches to a rolling circle replication model (Weller & Coen, 2012). In rolling circle replication, the circular DNA is replicated in one direction to generate multiple copies of circular DNA (Ueda, 2018). Rolling circle replication at the lytic phase generates sequences of larger-than-unit length which are called concatemers (Weller & Coen, 2012). Further, during replication, multiple random double-strands breaks are produced (Hammerschmidt & Sugden, 2013). These concatemers and strand breaks are efficient initiators and facilitators of the exchange of genetic segments between genomes co-infecting the same cell (Pérez-Losada *et al.*, 2015). Within the DNA

replication machinery are proteins that serve as recombinases and commandeering cellular proteins required in the genetic recombination steps (Pérez-Losada *et al.*, 2015). The UL29 protein has recombinase activity, promoting DNA strand breaks while the helicase/primase complex (UL5/UL8 and UL52) promotes strand exchange during replication (Weller & Coen, 2012). Genomic recombination involves these concatemers and strand breaks produced during DNA replication, two-component recombinases, and commandeering cellular proteins with studies showing that genomic recombination occurs as soon as the new replicated DNA is detected (Pérez-Losada *et al.*, 2015).

Genomic recombination in DNA viruses such as EBV occurs when linked genes or alleles associate through a break-join mechanism resulting in a homologous form of recombination (Pérez-Losada *et al.*, 2015). In this case, genomic fragments are exchanged between the same sites in both parental strands (Sijmons *et al.*, 2015). Recombination in EBV is largely homologous to maintain the integrity of the genome hence there is normally no change in the genomic structure of the recombinant genomes unlike in the case of non-homologous recombination that results in aberrant structures Froissart *et al.*, 2005). The resulting recombinant genome, therefore, carries the genetic variations drawn from different parental genome segments hence considered more genetically diverse than any of the parental genomes (Lee *et al.*, 2015; Sijmons *et al.*, 2015).

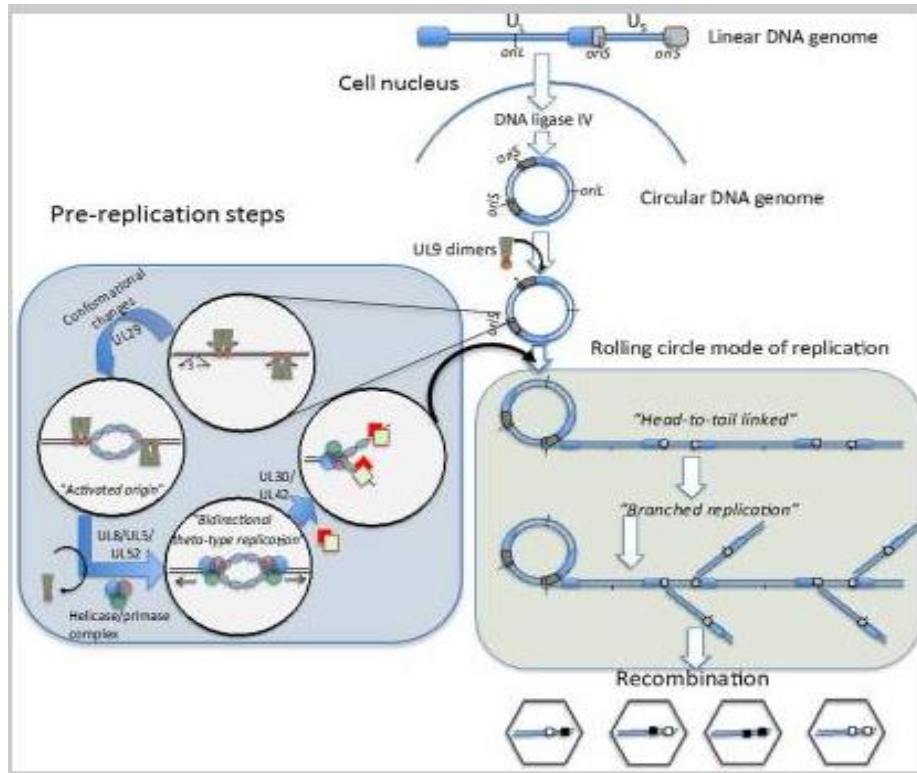


Figure 2.2: Replication and Genomic Recombination Steps in Herpesvirus (Pérez-Losada *et al.*, 2015a). Replication is intimately linked to genomic recombination. Before, replication is the pre-replication steps needed to recruit the replication machinery. Replication then proceeds via the bi-directional theta model before switching to the rolling circle model. Concatemers and strands are produced during replication and are efficient initiators and facilitators of genomic recombination.

Since genomic recombination is intimately linked to DNA replication, the positions of the genes that encode viral proteins necessary to form the replication and genomic recombination machinery may influence the sites where genomic recombination breakpoints would preferably cut along the EBV genome (Berenstein *et al.*, 2018; Lee *et al.*, 2015). EBV DNA replication is normally triggered by the lytic reactivation of the latently infected B cells (Hammerschmidt & Sugden, 2013; Kenney, 2007). This switch from latent to lytic infection requires the expression of the immediate early genes (*BZLF1* and *BRLF1*) (Li *et al.*, 2016). Further, the genes that encode the replication machinery must be expressed as well as those involved in the packaging of the lytic virions (Weller & Coen, 2012). The genes whose protein products are involved in replication and subsequently recombination are therefore more likely

to be affected by genomic recombination breakpoints (Berenstein *et al.*, 2018). Studies of EBV genomes from Nasopharyngeal Carcinoma (NPC) patients found that recurrent EBV reactivation promotes genome instability leading to breaks which could influence the exchange of genomic segments (Huang *et al.*, 2010; Wu *et al.*, 2010). It has been shown that homologous genomic recombination is important for EBV DNA damage repair to maintain and improve the genetic fitness of the virus to enhance its survival in the host (Huang *et al.*, 2010). Moreover, Berenstein *et al.* (2018) reported evidence of a heterogeneous and highly variable landscape of genomic recombination breakpoints rate along EBV genomes. Whether these patterns are similar to EBV genomes from western Kenya is not yet established. Further, the implication of these patterns in diseases such as eBL has not been investigated.

2.2.3 Detection and Analysis of Genomic Recombination

Based on the relevance of genomic recombination on viral diversity and possibly disease, tools are available for use in detecting and analyzing genomic recombination. Genetic diversity in EBV genomic sequences results from the joint processes of mutations and reshuffling or recombination (Chang *et al.*, 2009). Mutations results in alteration at the single nucleotide levels resulting in SNPs while genomic recombination involves rearrangement of large pieces of DNA sequences, involving several nucleotides hence the primary effect of genomic recombination is not site-specific or reliably measurable at the nucleotide level (Pérez-Losada *et al.*, 2015). Bioinformatics tools such as Recombination Detection Program 4 (RDP4) are available to exclusively detect genomic recombination within sequences without relying on the presence or absence of mutations (Martin *et al.*, 2015). This is a windows computer program that implements an array of methods to detect genomic recombination signals, identify genomic recombination breakpoints and identify genomic recombination

events responsible for the recombination signals (Martin *et al.*, 2015). RDP4 analyses the multiple sequence alignment (MSA) using a set of phylogenetic methods; Bootscan (Beiko & Hamilton, 2006) and RDP (Beiko & Hamilton, 2006) and substitution methods; Chimaera (Martin *et al.*, 2011), GENECONV (Martin *et al.*, 2011), MaxChi (Martin *et al.*, 2015), Siscan (Martin *et al.*, 2011), and 3Seq (Boni *et al.*, 2007) to detect genomic recombination signals. After the detection of the “genomic recombination signal” by these methods, RDP4 determines the breakpoint positions using a hidden Markov model, BURT (De Fonzo *et al.*, 2007), and identifies the recombinant sequences using PHYLPRO (Beiko & Ragan, 2008), VISRD (Lemey *et al.*, 2009) and EEP methods (Beiko & Hamilton, 2006). RDP4 then proceeds to determine the number of recombination events responsible for the recombination signals (Martin *et al.*, 2015). Finally, it outputs the identities of parental sequences. It is important to note that these “parental” sequences are not the actual parents of the recombinant sequence. They are, however, simply sequences within the analyzed dataset that were used to infer the existence of the actual parents (Martin *et al.*, 2015).

The program uses a set of pre-aligned nucleotide sequences, to identify and characterize genomic recombination since RPD4 cannot align the sequences for the user (Martin *et al.*, 2015). For this purpose, MSA tools such as IMPALE, MUSCLE, and CLUSTALW are normally distributed with RDP4. RDP4 however is usable with most MSA made using programs outside this distribution. This offers the user the advantage of using any MSA tool depending on the nature of the sequences to be aligned. For example, EBV sequences normally bear large gaps and a lot of ambiguous sequences hence hard to align. The user, therefore, has the liberty to use tools as Multiple Alignment using Fast Fourier Transform (MAFFT) to properly align these sequences to produce MSA which are also efficiently usable with RDP4. Further, RDP4 is able to perform genomic recombination analysis without the need for any predefined sets of any non-recombinant reference sequences (Martin *et al.*, 2015).

Additionally, combines a range of powerful heuristic recombination detection methods that sequentially test every possible combination of three sequences in an input alignment. The output is therefore reliable, accurate, and reproducible (Martin *et al.*, 2015). The full exploratory approach in RDP4 allows the program to characterize complex patterns of recombination including those arising when recombination events occur between parental sequences that are themselves recombinants (Martin *et al.*, 2015).

RDP4 is also capable of performing phylogenetic-based analyses using tools such as FastTree 2 (Price *et al.*, 2010) that account for genomic recombination. During such analyses, the MSA is stripped off of all detectable evidence of genomic recombination events (Martin *et al.*, 2015). Such analyses, therefore, mask the genomic recombination events hence the phylogeny constructed does not reflect the deep-seated diversity within the genomic sequences that result from genomic recombination (Rieux & Balloux, 2016; Zanella *et al.*, 2019). Unbiased phylogenetic analyses i.e. with the genomic recombination events can better infer the influence of genomic recombination on diversity.

2.3 Age, Gender and EBV Genomic Recombination

Genomic recombination events in a virus genome may be inherited from the parents, acquired early during primary infection, or later as the virus persists within the human host (Zanella *et al.*, 2019). Consequently, host and viral biological, as well as demographic factors, may influence the genome-wide occurrence of genomic recombination in viruses (Martin, 2015; Pérez-Losada *et al.*, 2015). Since genomic recombination in EBV is dependent on DNA replication, demographic factors such as age and gender may influence the rates of EBV replication which may significantly influence the exchange of genetic segments between genomes (Martin, 2015). Early in age exposure to EBV in children from malaria-endemic

regions of SSA is not only associated with high incidences of eBL (Piriou *et al.*, 2012; Reynaldi, Schlub, Piriou, *et al.*, 2016) but may also increase the chances of multiple EBV infections making available the genetic propensity to experience genomic recombination (Pérez-Losada *et al.*, 2015). Moreover, malaria in western Kenya is holoendemic i.e. is characterized by high rates of infection across the population with the highest parasite densities, morbidity, and mortality in children below the age of 5 years (Chattopadhyay *et al.*, 2013; Moormann & Bailey, 2016). The peak age for the development of eBL which is 4-9 years comes after this early exposure to both *Pf* and EBV infections (Rainey *et al.*, 2007; Redmond *et al.*, 2020). Comparison of the occurrence of genomic recombination across the age groups i.e. 0-4, 5-9, and 10-14 should help elucidate the influence of repeated *Pf* and EBV infections on genomic recombination.

Immunity to *Pf* is dependent on age with older children being able to mount better immune responses to *Pf* compared to the younger cohorts (Griffin *et al.*, 2015). Repeated infection with *Pf* is associated with EBV expansion and subsequent peaks of EBV viral load in the peripheral blood circulation suggestive of recurrent episodes of lytic reactivation (Daud *et al.*, 2015; Njie *et al.*, 2009). Viral reactivation normally leads to the lytic phase of infection where the virus replication (Kenney, 2007; Li *et al.*, 2016). Increased episodes of lytic replication may promote the exchange of genomic segments between genomes. Also, the greater the viral population measurable by high viral loads the more the chances for genomic recombination (Pérez-Losada *et al.*, 2015). The current *Pf* malaria status of a participant, however, may not bear an immediate impact on genomic recombination since the occurrence of genomic recombination and incorporation within a population is an evolutionary process that takes a long period of time (Pérez-Losada *et al.*, 2015). EBV genomes obtained from the human population from malaria holoendemic western Kenya allow this study to characterize the association of genomic recombination with age. Associating genomic recombination with

age helps elucidate whether these events are distant in evolutionary time i.e. acquired from previous hosts or whether they occur in the current hosts during primary infection or during persistence within the host.

Males and females have differences in their susceptibility and immune response to diseases. Such differences in gender are common though are highly neglected (van Lunzen & Altfeld, 2014). Such sex-based differences have been linked to immunological pathways affected by sex hormones such as estrogen (Klein, 2012). Estrogen levels, for example, have the ability to affect immune cell populations (Taneja, 2018). Males and females also have differential expression of X-chromosome-encoded genes on immune responses to pathogens (Klein, 2012). For instance, the X chromosome contains 10% of all microRNAs (miRNAs) in the genome, whereas, the Y chromosome has no miRNAs (van Lunzen & Altfeld, 2014). There are shreds of evidence that these miRNAs may infer differences in the immune responses of males and females (Migliore *et al.*, 2021). Studies that have compared immune responses between males and females show that females are better at mounting immune responses to viral infection such as Human Immunodeficiency Virus (HIV) (Ballesteros-Zebadúa *et al.*, 2013) and Coronavirus (Pradhan & Olsson, 2020). It is however not known if the observed differences can be reported in immune responses to EBV. Other studies show that males are more likely to develop eBL compared to females (Rainey *et al.*, 2007; Torgbor *et al.*, 2014). Since host immunity to EBV has the capacity to influence the occurrence of genomic recombination (Pérez-Losada *et al.*, 2015), the genomic recombination may differ between males and females.

2.4 EBV Type-Associated Genomic Recombination Events and EBV Diversity

Epstein-Barr virus (EBV) genomes are classified as either type 1 or type 2 (Sample *et al.*, 2009). This classification is based on the variations in the *EBNA 2* gene and *EBNA 3* genes

(*EBNA 3A, 3B, and 3C*) (Tzellos & Farrell, 2012). *EBNA 2* genes bear the main differences between the two types, with only 70% identity at the nucleotide level and 54% identity in the protein sequence (Kanda *et al.*, 2019). *EBNA 3A, 3B, and 3C* genes confer lesser base pair differences of 10%, 12%, and 19% respectively between the two types (Neves *et al.*, 2017). For this reason, the differences in the variations in the *EBNA 2* gene have been exploited using methods such as Polymerase Chain Reaction (PCR) to genotype the viral genomes (Habibian *et al.*, 2018; Palma *et al.*, 2013; Robaina *et al.*, 2008). Type 1 and type 2 viruses differ in their abilities to transform B cells with type 1 readily transforming B lymphocytes into lymphoblastic cell lines (LCLs) *in vitro*. The type 1 *EBNA2* Carboxyl (C) terminal region greatly induces the expression of LMP-1 and Coupled Chemokine Receptor 7 (CXCR7) genes (Lucchesi *et al.*, 2008). In Lucchesi *et al.*, (2008), the differential expression of CXCR7 and LMP1 genes conferred a strong growth and survival advantage to the cells. In a recent study, type 1 EBV genomes drawn from western Kenya were more associated with eBL (Kaymaz *et al.*, 2020). The differences in the ability of type 1 and type 2 to transform B cells and augment EBV-associated tumors such as eBL may also result from the differences in their genetic diversity. Such genetic differences may be linked to unique genomic recombination events creating new variant profiles in a given viral type and not the other hence there's a need to compare the occurrence of genomic recombination between the two EBV types.

Genomic diversity in EBV can be defined as the proportion of polymorphic loci across the virus genome. Type 1 and type 2 classification is a major feature of EBV genomic diversity (Sample *et al.*, 2009). This feature is normally defined by the first split in the phylogeny of EBV genomic sequences with *EBNA 2* and *EBNA 3* gene regions (Chiara *et al.*, 2016; Palser *et al.*, 2015; Santpere *et al.*, 2014b). Multiple genomic recombinations throughout the genome, therefore, have the potential to affect the phylogenetic tree topology by influencing the positioning and clustering of isolates into phylogenetic clades (Rieux & Balloux, 2016).

Consequently, the phylogenetic clades with evidence of genomic recombination normally have low node supports since the different parts of the sequences may have different phylogenetic histories (Zanella *et al.*, 2019). This may imply that the sequences under study are not related by a single phylogenetic tree, but rather a set of correlated trees over the sequence (Rieux and Balloux 2016). Therefore, most phylogenetic methods output the most suitable phylogenetic tree with the best node supports (Yoshida & Nei, 2016). The position of isolates in the most suitable phylogenetic tree is therefore under the influence of genomic recombination events (Rieux & Balloux, 2016). Further, the dynamics of such a phylogenetic tree can be studied to understand the divergence deeply seated in the genomic sequences being studied (Yoshida & Nei, 2016).

2.5 Genomic Recombination Events and eBL Pathogenesis

2.5.1 *Plasmodium falciparum* and EBV Infection Augments eBL

EBV was first isolated from an eBL tumor and the clonal presence of the virus in almost all eBL tumors suggests a necessary role (Redmond *et al.*, 2020). To establish persistence, EBV infects resting B cells and drives their proliferation as activated B cells (Thorley-Lawson *et al.*, 2013). Uncomplicated *Pf* infection causes polyclonal expansion of memory B cells leading to recurrent EBV lytic reactivations and frequent episodes of measurable viremia (Daud *et al.*, 2015; Njie *et al.*, 2009; Westmoreland *et al.*, 2017) both of which exhaust host T cell responses (Chattopadhyay *et al.*, 2013; Moormann *et al.*, 2007). Downregulation of EBV-specific T cell responses impairs the host's restriction of viral replication and control of the number of infected B blasts with potential oncogenicity, therefore, increasing chances of progression to eBL (Torgbor *et al.*, 2014).

The expansion of EBV-infected B blasts into memory B cells involves passage through the GC (Moormann & Bailey, 2016). Similarly, recurrent *Pf* malaria infection induces increased GC transition of EBV infected B cells hence they express elevated levels of the highly mutagenic Activation Induced Deaminase (AID) (Torgbor *et al.*, 2014). Further, *Pf* appears to augment the expression of AID in human tonsillar B cells even outside the GC (Moormann & Bailey, 2016). AID normally causes deamination of critical cytidine residues at the Ig loci which when repaired by the error-prone DNA repair mechanisms results in point mutations and strand breaks responsible for somatic hyper-mutations and Immunoglobulin (Ig) class switch recombination respectively (Hwang *et al.*, 2015). AID also induces off-target lesions at non-Ig loci causing mutations and translocations that have been associated with the development of several cancers (Love *et al.*, 2012). There is overwhelming evidence that activated AID by *Pf* can contribute to eBL development by directly mediating the translocation involving *myc* in GC B cells (Moormann & Bailey, 2016). This chromosome translocation between the *c-myc* and IgH or IgL causes constitutive activation and expression of the *c-MYC* gene transforming it from a proto-oncogene to an oncogene with abilities to cause lymphomagenesis (Love *et al.*, 2012; Panea *et al.*, 2019).

C-Myc overexpression, however, can trigger rapid apoptosis in a fail-safe mechanism hence apoptotic pathways must be disabled for the oncogene to promote cell transformation and cancer (Panea *et al.*, 2019). Latency-associated gene products such as EBNA1, LMP-1, EBERs, EBNA2, EBNA3A, EBNA 3C, etc. can inhibit a variety of pathways responsible for apoptosis and senescence (Kang & Kieff, 2015; Kempkes & Robertson, 2015). EBV, therefore, counteracts the proliferation-restricting activities of deregulated *myc* and so facilitates the development of eBL (Thorley-Lawson *et al.*, 2013). The high mutagenic activity in the infected B cells has also been shown to induce mutations in the tumor suppressor genes such as Tumor protein (*TP53*), Cyclin-Dependent Kinase Inhibitor 2A (*CDKN2A*), or B Cell Ligand 2 Like

11 (*BCL2L11*) and anti-apoptotic genes both of which precede the development of eBL (Love *et al.*, 2012; Panea *et al.*, 2019).

2.5.2 EBV Genomic Recombination Events and eBL Pathogenesis

Genomic recombination events can introduce novel phenotypes which can potentiate the pathogenesis of eBL. Zanella *et al.*, (2019) detected genomic recombination events in EBNA 3 genes (one in EBNA3A, another in EBNA 3B, and two more in EBNA 3C gene) that could be associated with changes in their immunogenic determinants providing a common route for EBV immune escape. Studies in Human Simplex Virus (HSV1) revealed genomic recombination breakpoints in latency-associated genes which could infer better capabilities to evade host immune surveillance (Lee *et al.*, 2015). During EBV infection, host T cell responses eliminate the newly infected B cells with potential oncogenicity before they can cause morbidity and mortality (Brooks *et al.*, 2016). T cell responses are normally directed against T cell epitopes hence variability in these epitopes can affect the MHC binding and subsequent recognition by T cell receptors (Thorley-Lawson *et al.*, 2013). This can facilitate viral escape from the host immune surveillance though it is not known if this is the case with eBL development.

Genomic recombination has the potential to affect the property and functionality of genes and their protein products thereby contributing to disease (de Been *et al.*, 2013; González-Candelas *et al.*, 2011; Sijmons *et al.*, 2015). Berenstein *et al.* (2018) observed recombination signals higher than the average genome in EBNA3C and EBNA3B both of which participate in attenuating DNA damage responses during EBV infection and transformation of naïve B cells. Based on the study, other gene products that also had high recombination rates were: BRRF2; a tegument protein, BBLF2-BBLF3; an accessory protein

to the viral helicase-primase complex, BZLF2; a glycoprotein gp42, BKRF2, a virion glycoprotein, BFRF2, early lytic protein, and late promoter activator as well as BFLF1; crucial for cleavage and package of viral particles (Berenstein *et al.*, 2018). BZLF1 and BZRF1 which are viral master regulator proteins involved in latent to lytic switch also reported high recombination density (Berenstein *et al.*, 2018). Latent to lytic switch is crucial for EBV genomes amplification through replication, formation of the mature virions, infection of new cells, and the release of infectious viral particles into saliva (Li *et al.*, 2016; Rosemarie & Sugden, 2020). The release of infectious particles not only facilitates EBV transmission but also propagates infections of new cells both of which are crucial for disease progression (Swaminathan & Kenney, 2008). Moreover, EBV genomes replication creates an environment favorable for genomic recombination (Kenney, 2007). Further, the genes affected by genomic recombination breakpoints are important to predict the molecular mechanisms involved in the genomic recombination (Pérez-Losada *et al.*, 2015). Despite the known implications of genomic recombination breakpoints on EBV genes, their association with eBL development has not been investigated. A comparison between eBL cases and the geographically matched healthy counterparts is important to identify genomic recombination events which may potentiate eBL.

CHAPTER THREE

METHODOLOGY

3.1 Study Design and Site

This was a case-control study that utilized randomly picked archival samples of healthy and eBL diagnosed children aged 2-14 years recruited between 2009 and 2012. The samples were collected at Jaramogi Oginga Odinga Teaching and Referral Hospital (JOOTRH), a public regional level five hospital that serves as the referral center for children diagnosed with cancer in western Kenya (Buckle *et al.*, 2016). The hospital is located in Kisumu (005°2' N 34°46'17" E Kenya), a region in western Kenya, with one of the highest incidences of *Pf* malaria in Kenya (Rainey *et al.*, 2007). The catchment of JOOTRH spans western Kenya and displays the expected geographic overlap with *Pf* malaria transmission (Rainey *et al.*, 2007) (Figure 4A). The spatial distribution and catchment area of pediatric eBL patients admitted at JOOTRH spans western Kenya (Buckle *et al.*, 2016) (Figure 4B).

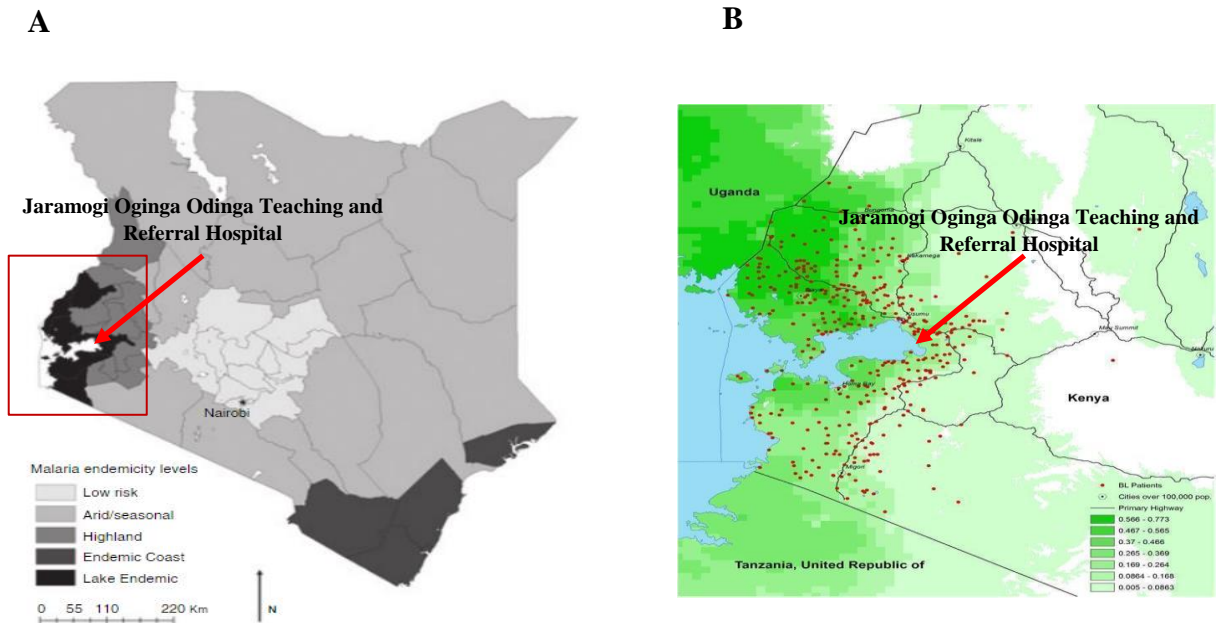


Figure 3.1: Distribution of eBL patients in western Kenya A) The map of Kenya showing *Pf* malaria endemicity levels across the country, with western Kenya and coast leading in prevalence (Rainey *et al.*, 2007) B) Spatial Distribution and catchment area of pediatric eBL patients admitted at JOOTRH (Buckle *et al.*, 2016). Red dots indicate the home village of over 600 eBL patients enrolled from 2003 to 2011. Shades of green illustrate malaria transmission intensity; Light green and dark green illustrate low and high transmission respectively.

3.2 Study Population

Only archival samples from participants who were residents of western Kenya at least 4 months before the recruitment, based on participants' records, were utilized in the study. This ensured that the archival samples used were strictly from participants who were residents of western Kenya. The demographic characteristics of the study participants including age and gender were gleaned from the participants' records. Since children from western Kenya contract EBV before they are 2 years (Piriou *et al.*, 2012), the healthy participants aged 4-6 were EBV positive just like the eBL cases. Western Kenya is holoendemic for *Pf* malaria meaning children and even adults from these regions experience several episodes of *Pf* malaria annually with high morbidity, mortality, and parasite densities in children below 5 years

(Chattopadhyay *et al.*, 2013). This population also had higher EBV viral loads that were required for direct sequencing (Njie *et al.*, 2009).

3.2.1 Inclusion Criteria

The eBLs archival samples used in this study were from children who were confirmed positive for eBL based on local histological and cytology reports. Only archival samples from participants who had not been initiated on eBL treatment, whose parents or guardians had consented for use of child samples for future studies, were 2-14 years in age and were residing in western Kenya for at least 4 months before the onset of eBL-related symptoms were used in this study. The information herein used was based on the stored records of the participants.

The controls archival samples were from participants determined as healthy based on medical history, physical examination as well as screening. Only participants aged 2-10 years, whose parents or guardians had consented for use of child samples for future studies, with no history of cancer or any other chronic illnesses and were residing in western Kenya for at least 4 months before the onset of the sample collection had their archival samples used. The information was gleaned from the stored sample records.

3.2.2 Exclusion Criteria

A confirmed eBL positive archival sample was excluded from the study if the participant was not residing in the pre-defined geographical region area for at least 4 months before the onset of eBL-related symptoms, was clinically unstable, had initiated eBL treatment and the parent or guardian had refused or were not able to consent for use of child sample for future studies. The information herein used was based on stored records of the participants.

An archival from a healthy control was excluded if the participant had the significant disease or illness determined by history or physical exam as well as screening if the parent or guardian did not provide informed consent for the use of samples for future studies if the child

had been exposed to immuno-modulatory therapeutics such as steroids in the past two months before recruitment, if he had any cancer or chronic illness and if the child was not a usual resident of western Kenya. The information was gleaned from the stored sample records.

3.3 Sample Size Determination

Based on the inclusion and the exclusion criteria, a total of 95 archival samples; 40 blood samples, 14 plasma samples, and 41 tumor samples were available for use in this study. A power test was done to confirm if this number of archival samples had enough statistical power to answer the study objectives. Using the statistical power analysis tool in R (Yatani, 2016) the sample size of 95 i.e. 55 eBL archival samples and 40 archival samples had enough statistical power to answer the study objectives (statistical power=97%). After sequencing and reads assembly, 9 archival samples; 8 from the healthy and 1 from an eBL participant had very poor sequence coverages (Appendix V) were therefore eliminated. Eliminating the 9 archival samples gave a new sample size of 86. A power test was carried out for this sample size (N=86) using the statistical analysis tool in R (Yatani, 2016). Of the 86 archival samples: 54 samples were from the eBLs and 32 were from the healthy participants; 28 were from females and 58 were males and finally 39 were from children aged 0 to 4 years, 34 from children 5 to 9 years and 13 from children 10 to 14 years. The formula below was used to determine the statistical power:

pwr.t2n.test (N1 =54, N2=32, d =0.8, sig.level =0.05, alternative="two.sided")

Where;

N1 and **N2**= are unequal sample sizes for each of the participants' characteristics. For example, 54 eBLs and 32 healthy archival samples for the eBL status of the participants.

d= Effect size i.e. 0.8 for a large effect size

sig.level= the significant level of 0.05

Alternative=two sided to signify a two-tailed test.

The statistical power was calculated for all the participants' characteristics as were gleaned from the participants' records i.e. eBL status (94%), Gender (93%), and Age group (92%) (Appendix IV)

3.4 Sample Processing and Storage

The archived primary tumor biopsies were collected using a biopsy gun and transferred into RNAlater at the bedside, before the induction of chemotherapy. Additionally, 2ml peripheral blood samples were collected in Ethylenediaminetetraacetic Acid (EDTA) tubes and fractionated by density centrifugation at a speed of 10,000g before freezing into plasma and cell pellets (Microcentrifuge 5424R, Eppendorf) before archival at -80°C. All prep DNA mini kits (QIAGEN Sciences, Germantown, MD, USA) were used for DNA extraction from tumor biopsies, blood, and plasma archival samples. The DNA was then stored at -80°C until further processing.

3.5 EBV Genotyping

EBV was genotyped as either type 1 or type 2 by Quantitative real-time Polymerase Chain Reaction (q-rPCR) targeting the EBNA2 gene (CFX Connect Real-Time PCR, Bio-Rad). The technique made use of 2 forward primers; one targeting the type 1 EBNA 2 region and the other the type 2 EBNA 2 gene region. A common reverse primer was used with a dye probe for detection (Genotyping Primer and Probe sets in Appendix V). The PCR run was set at 50°C

for 2 minutes, 95°C for 10 minutes, 40 cycles of 95°C for 15 seconds, and 60°C for 1 minute. The run cycle was followed by a melt curve set at 95°C for 15 seconds, 60°C for 1 minute, and 95°C for 15 seconds. All the 95 archival samples were genotyped as either type 1 or type 2.

3.6 Sequencing Library Preparation, EBV Specific Genome-wide Amplification, and EBV DNA Enrichment

Before sequencing, the Illumina sequencing library was prepared according to the steps outlined in Kaymaz *et al.*, (2020). The extracted DNA was sheared into fragments of sizes of 75 base pairs (bps) to 150bps usable by the Illumina Sequencer (DNA Shear Kit, NEB). Since the shearing resulted in 5' and 3' overhangs, the sticky ends were repaired into blunt-ends to give uniform 3'hydroxyl ends and 5'phosphate ends (Quick-Blunting kit, NEB). This was followed by 3'-adenylation to avoid the formation of dimers and to provide a complimentary 3' oligo AAA tail for the binding oligo dT tail of the sequencing adapters (Klenow Fragment 3' to 5' exo-, NEB). The indexed sequencing adapters (Quick Ligation kit, NEB) were then added to the fragments. The sequencing libraries were PCR amplified to a final concentration using KAPA HiFiHotStartReadyMix before quantification and quality check using a bioanalyzer (Agilent Bioanalyzer, DNA Technologies Core). The sample libraries were thereafter pooled into 2 pools according to their EBV viral loads read from the bioanalyzer. The pools were subjected to EBV-specific genome-wide amplification (sWGA) (Appendix VI) followed by EBV DNA enrichment using custom EBV biotinylated RNA probes (MyBaits, Arbor Biosciences). The libraries were then sequenced using Illumina sequencing instruments; Illumina MiSeq, HiSeq 2000, and NextSeq 500 platforms with various read lengths of 100bps, 200bps, and 300bps respectively. The sequencing was done at the University of Massachusetts Medical School (UMMS) Deep sequencing core facility in the United States. All the 95

archival samples were sequenced. Additional six plasma samples replicates of tumor samples from the same participants were sequenced. The Pre-processing Information and Sequencing Statistics including the Illumina paired-end read length, the total sequence reads, the assembled genome size, and the average depth of coverage over assembly for all the archival samples that were sequenced (Appendix VII).

3.7 Sequence Reads Pre-processing and de novo Genome Assembly

The residual sequencing adapters and low-quality bases were trimmed using Cutadapt (v 1.7.1) (Martin, 2011) and Prinseq (v 0.20.4) (Schmieder & Edwards, 2011) respectively. The sequences were thereafter checked for quality using FastQC (v0.10.1) (Trivedi *et al.*, 2014) before *de novo* assembly into a contiguous length of genomic sequence (contigs) using VelvetOptimiser (v 2.2.5) (Zerbino & Birney, 2008). The contigs were then ordered and oriented guided by EBV type 1 (Genbank accession: NC_007605) and type 2 references (Genbank accession: NC_009334) using Algorithm-Based Automatic Contiguation of Assembled Sequences (ABACAS) (Assefa *et al.*, 2009), extended with read support using Iterative Mapping and Assembly for Gap Elimination (IMAGE) (Tsai *et al.*, 2010), and merged into overlapping contigs to form larger scaffolds (Using in house scripts). To assess contigs' quality, the reads were aligned to the assembled scaffolds using Bowtie 2 Aligner (Langmead & Salzberg, 2012). The genomes were finally created by demarcating repetitive and missing regions due to low coverage with sequential ambiguous "N" nucleotides. A total of 95 genomic sequences were assembled together with 6 plasma replicates of tumor samples. The number of reads for each genome assembled is provided in Appendix VII. These sequences are currently available in the European Nucleotide Archive (ENA) (<http://www.ebi.ac.uk/ena>) under the study accession no. ERP122181.

3.8 Multiple Sequence Alignment

For the subsequent analysis, the study only utilized EBV genome sequences that had good coverage (Genomes coverage in Appendix VII). Consequently, 9 genomes; 8 from the healthy and 1 from an eBL patient were eliminated reducing the number of genomes from 95 to 86. The 86 genomic sequences had enough statistical power to answer all the study objectives. The alignment of the 86 genomic sequences was performed using Multiple Alignment using Fast Fourier Transform (MAFFT) software version 6 (Kato *et al.*, 2019). Compared to the MSA tools such as IMPALE, MUSCLE, and CLUSTALW distributed with Recombination Detection Program 4 (RDP4), MAFFT is best suited for the hard-to-align EBV sequences known for excessive gaps and ambiguous nucleotides (Kato *et al.*, 2019). The alignment was manually inspected for gaps and ambiguous Ns using PhyloSuite v1.2.2 (Zhang *et al.*, 2020).

3.9 Removal of Poorly Aligned Regions

Poorly aligned regions, with excessive gap alignments and considerable divergent regions, were trimmed out by G-blocks (Talavera & Castresana, 2007). This program selected blocks of conserved regions and ensured that phylogenetic analysis was performed on genomes with reliable genomic content and good coverage (Talavera & Castresana, 2007). Phylogeny is normally improved after the removal of divergent and ambiguously aligned blocks from the MSA (Talavera & Castresana, 2007). After exclusion of poorly aligned sites, the study recovered a genome size of 88,000 base pairs (bps) from all the 86 genomic sequences studied. Out of the 172,000bp, this region represented 51.2% of the whole EBV genome. As visualized in the circular plot (Figure 6), the poorly aligned genomic regions that were trimmed by G

blocks mostly lie within the EBV repeats known to bear large gaps after Illumina short-read sequencing and de novo assembly (Sample *et al.*, 2009).

3.10 Phylogenetic Analysis

Phylogenetic analyses were performed using Molecular Evolutionary Genetics Analyses version 7 (MEGA 7) (Kumar *et al.*, 2016). The evolutionary history was inferred using the Neighbor-Joining (NJ) algorithm which computes evolutionary changes at ancestral nodes to compute a single phylogenetic tree with the best topology (Yoshida & Nei, 2016). Evolutionary distances were computed using the Jukes-cantor model (Erickson, 2010). Ambiguous nucleotides were removed using pairwise deletion to control for the effect of such nucleotides on phylogenetic accuracy (Lemmon *et al.*, 2009). Bootstrap analyses of 10000 replicates were performed on each tree to determine confidence. Finally, the tree was rooted in the midpoint branch.

3.11 Recombination Analyses

Recombination analysis was thereafter performed using the Recombination Detection Program (RDP) version 4 (Martin *et al.*, 2015). RDP4 analyzed the MSA using a set of phylogenetic methods; Bootscan (Beiko & Hamilton, 2006) and RDP (Beiko & Hamilton, 2006) and nucleotide substitution methods; Chimaera (Martin *et al.*, 2011), GENECONV (Martin *et al.*, 2011), MaxChi (Martin *et al.*, 2015), Siscan (Martin *et al.*, 2011), and 3Seq (Boni *et al.*, 2007) to detect genomic recombination signals and provided a detailed output of the aligned sequences that were recombinants and their corresponding breakpoints. Each genomic recombination event corresponded to two genomic recombination breakpoints.

Genomic recombination events were only considered significant when all the six algorithms had a threshold p-value of 0.05 , using Bonferroni correction (Martin *et al.*, 2010). Using more than one algorithm ensured that the genomic recombination events used in the analysis were accurate and reliable without any false positives. The Bonferroni correction was done to correct for any errors in the hypothesis testing by each of these algorithms (Armstrong, 2014). Further, the precision of genomic recombination detection was validated by characterizing the occurrence of recombination in 6 plasma and tumor replicates. The study hypothesized that the EBV in the tumor is representative of the EBV in the plasma hence the plasma-tumor replicates were supposed to report similar evidence of genomic recombination events. (Appendix X).

3.12 Genomic Feature Annotation

The coordinates of the genomic recombination events and their breakpoints were mapped to the EBV type 1 (Genbank accession: NC_007605) and type 2 references (Genbank accession: NC_009334). Annotated genomic features including gene positions, coding regions, introns, as well as regulatory regions corresponding to the genomic recombination events were extracted from the reference genomes and used to generate a BED file for visualization using Integrative Genome Viewer (IGV) (Thorvaldsdóttir *et al.*, 2013)

3.13 Data Analysis

All statistical analyses were performed using R statistical software Version 4.1.2 (Hackenberger, 2020) setting 2-tailed alpha to reject the null hypothesis at 0.05 . A summary of key R scripts used in the analysis is in Appendix IX. For objective 1: Frequency and breakpoint distribution plots were used to describe and represent the genome-wide occurrence of EBV

genomic recombination events and their recombination breakpoints respectively. For objective 2: the Wilcoxon rank test was used to compare the occurrence of genomic recombination events between males and females while the Kruskal-Wallis test was utilized to compare the occurrence of genomic recombination events across the three age groups. For objective 3: the Wilcoxon rank test was used to compare the occurrence of genomic recombination events between viral types. Fisher exact test was used to determine EBV type association with unique genomic recombination events and their breakpoints while a Neighbour Joining (NJ) phylogeny predicted the association of genomic recombination events with EBV diversity. For objective 4: the Wilcoxon rank test was used to compare the occurrence of genomic recombination events between eBL cases and healthy controls. Univariate and multivariate logistic regression modeled eBL association with genomic recombination events and their breakpoints. Pearson's Chi-square and Fisher's exact tests were used to compare eBL status, age groups, viral type, and gender proportions.

3.14 Ethical Approval

The ethical approval of the mother study protocol was obtained from the Scientific and Ethical Review Unit (SERU) at the Kenya Medical Research Institute (KEMRI) and University of Massachusetts Medical School (UMMS) (Appendix I). The participants' consent form is provided in Appendix II. The mother study protocol covered all future studies that required the use of all archival samples where the participants consented for use of the archival samples for future use. Further, an approval to carry study this study that stemmed from the mother study was obtained from the School of Graduate Studies (SGS) Maseno University, Maseno, Kenya (Appendix III).

CHAPTER FOUR

RESULTS

4.1 Demographic Characteristics of Study Participants

The study analyzed 86 genomic sequences (Appendix VII) generated from archival samples of 86 participants for evidence of genomic recombination to associate the genomic recombination events with age, gender, EBV viral type, and eBL status. The general characteristics of the study participants as gleaned from participants' records are summarized in Table 4.1. Of the 86 genomic sequences, 54 (62.8%) were from confirmed eBL cases and 32 (37.2%) were from the healthy controls. Among the 54 eBL participants recruited, 40 (74.1%) were males. More BL-positive children were aged 5-9 years (57.4%), while more healthy controls were aged 0-4 years (90.6%) and none above 10 years (0%). More eBL positive participants had type 1 EBV genomes (72.2%). The group proportions were compared to check if they would affect downstream comparisons between the groups. The group proportions were comparable except age groups ($p=5.735e-11$) that would be controlled for where appropriate in the downstream comparisons.

Table 4.1: Demographic Characteristics of Study Participants

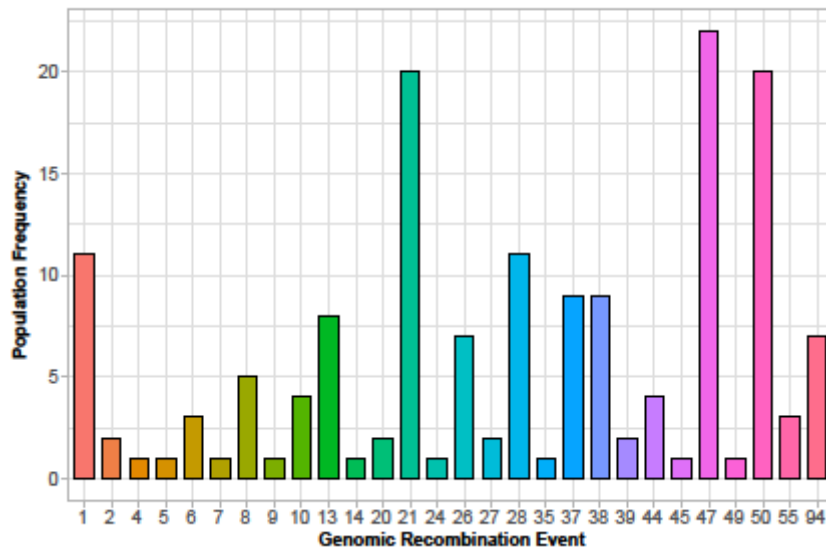
Characteristic		Total	eBL (%)	Healthy (%)	<i>P-Value</i>
Participants		86	54 (62.8)	32 (37.2)	
Gender	Female	28	14 (35.9)	14 (43.8)	0.1424^a
	Male	58	40 (74.1)	18 (58.2)	
Age Group	0-4	39	10 (18.5)	29 (90.6)	5.735e-11^b
	5-9	34	31 (57.4)	3 (9.4)	
	10-14	13	13 (24.1)	0 (0)	
Viral Type	Type 1	56	39 (72.2)	17 (53.1)	0.1183^a
	Type 2	30	15 (27.8)	15 (46.9)	

Abbreviation: eBL, endemic Burkitt lymphoma. Bold text indicates a statistically significant difference with a *P-value*<0.05. Groups' proportions were compared using ^aPearson's Chi-square and ^bFischer exact tests

4.2 Genome-wide occurrence of Genomic Recombination Events and Breakpoints

Recombination Detection Program 4 (RDP4) identified 28 genomic recombination events. Genomic recombination event was defined as the distinct incorporation of a unique genomic segment into a genome. The program identified genomic sequences with evidence of each of the genomic recombination events identified. Of the 86 sequences that were analyzed for genomic recombination, 71 (82.6%) reported evidence of one or more distinct genomic recombination events and were generally classified as recombinant genomes. The average number of recombinant fragments in each genome was 2 (*mean*=2, *range*=1-4). RDP4 gave each of the 28 genomic recombination events a unique identifying number; 1, 2, 4, 6, 7, 8, 9, 10, 13, 14, 20, 21, 24, 26, 27, 28, 35, 37, 38, 39, 44, 45, 47, 49, 50, 55 & 94. The counts of each distinct genomic recombination event within the dataset of the 86 genomic sequences were used to construct the genomic recombination events' frequency plot (Figure 4.1).

Figure 4.1: Frequency of Genomic Recombination Events

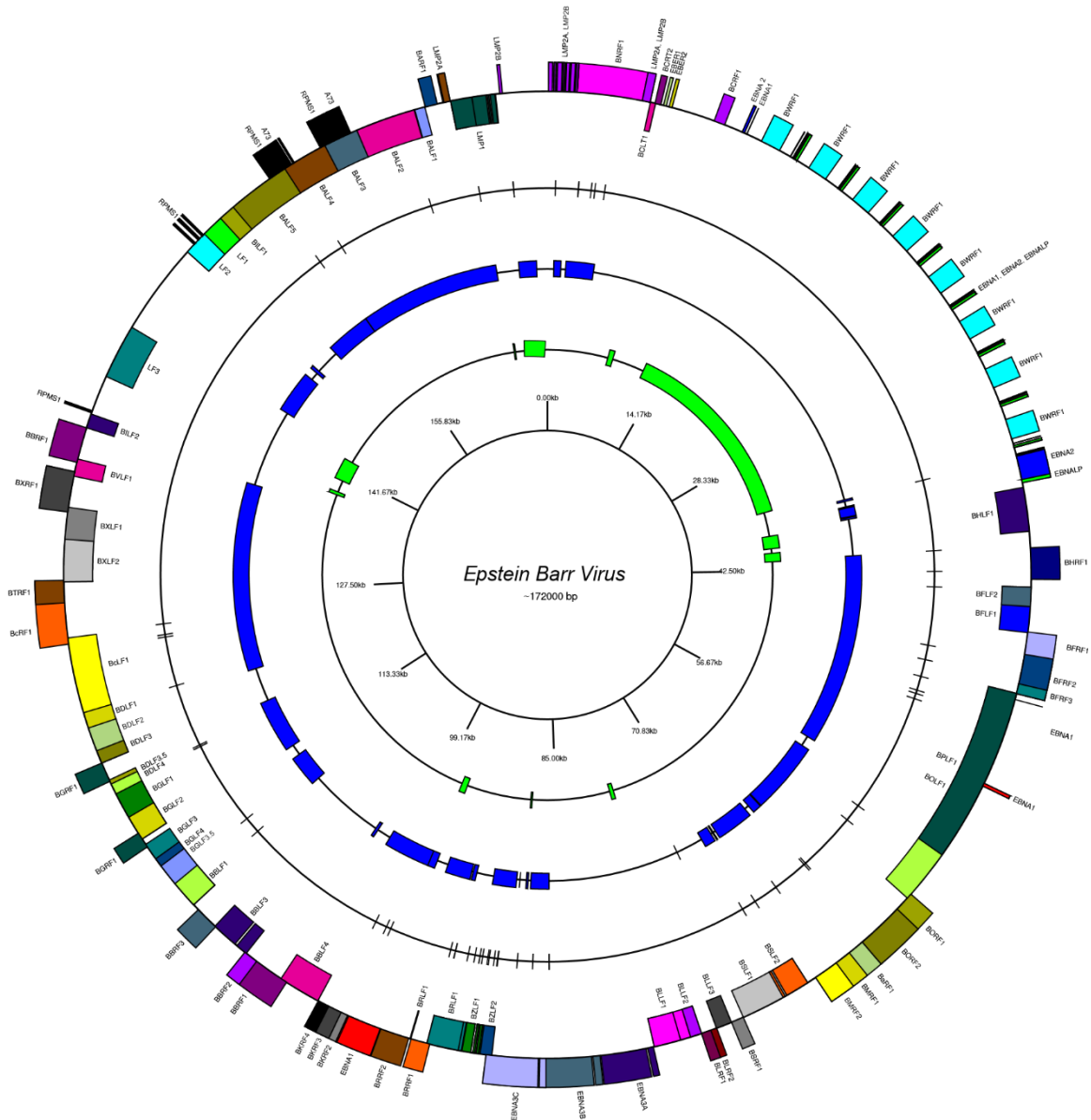


Each colored bar represents a distinct genomic recombination event as reported by RDP4. Each number on the x-axis is the name of each distinct genomic recombination event as coded by RDP4. The number of recombination events detected=28

Every distinct genomic recombination event was associated with two recombination breakpoints along the genome. Recombination breakpoints, therefore, were sites along with the genomic sequences that were cut to allow the incorporation of a genetic fragment during a genomic recombination event. One recombination breakpoint was at the beginning and was referred to as a “Start breakpoint” and the other at the end referred to as an “End breakpoint”. The position of each genomic recombination breakpoint in the MSA was mapped to the EBV reference genomes resulting in an EBV genome map displaying the position of every single genomic recombination breakpoint identified by RDP4 (Figure 4.2). Additionally, the map highlighted the coordinates of the MSA showing the good coverage regions (blue circle) as well as the EBV repeat regions (green circle) to display that the coordinates of the genomic recombination breakpoint were within the MSA analyzed. From the figure, the recombination breakpoints clustered in some genomic sites such as the region around the *BZLF1* and *BRLF1*

genes. Some genomic sites despite being within the MSA did not show any evidence of recombination breakpoints.

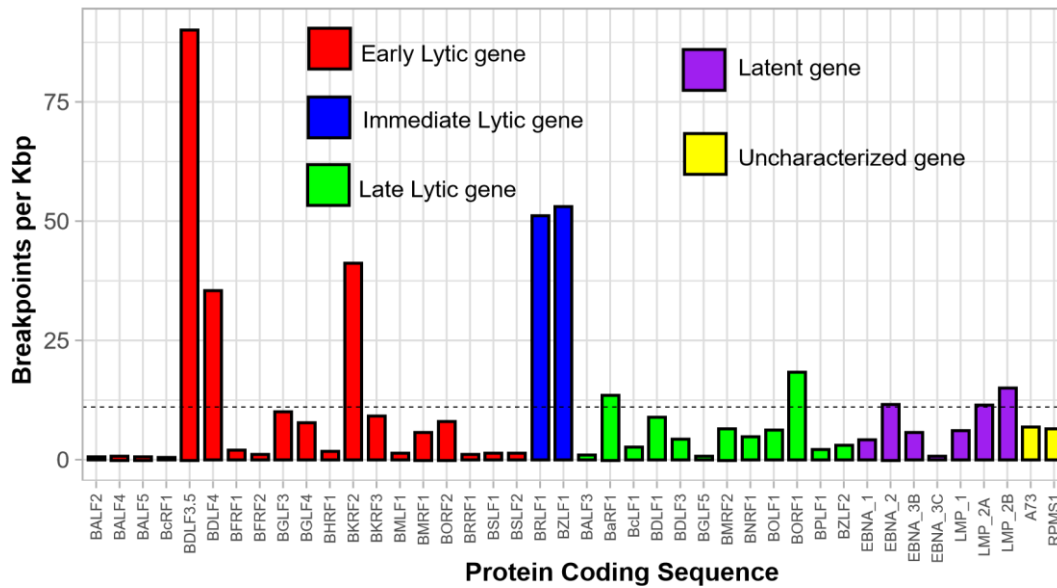
Figure 4.2: Positions of Recombination Breakpoints along EBV Genome



The innermost circle shows the EBV genome coordinates. The green circle displays the EBV repeats. The blue circle is the alignment showing the good coverage regions. The 4th circle highlight the position of each recombination breakpoint. The outermost circle displays the gene map where each colored bar corresponds to a gene exon. The gene names shown outside the circle are transcribed clockwise and the gene names shown inside the circle are transcribed counterclockwise. The figure was drawn using the GenomeVx program (Conant & Wolfe, 2008).

To investigate the effect of genomic recombination on EBV genes, the study identified genes that were cut by each recombination breakpoint. These breakpoints were found to cut through a total of 42 different protein-coding sequences (CDS) (Figure 4.3). Of the 42 genes, only 7 (16.67%) were genes of latent EBV cycle. Synthesizing further, 19 were early lytic genes, 12 were late lytic genes and the remaining 2 were immediate early genes i.e. *BZLF1* and *BRLF1*. Next, the study explored the recombination breakpoints per kilobase pair (Kbps) for each of the 42 coding CDS that reported evidence of genomic recombination breakpoints. Recombination breakpoint per Kbps allowed the study to compare the occurrence of recombination breakpoints among the 42 EBV CDS with varying lengths. The mean number of recombination breakpoints per Kbp for all the 42 CDS was 11.05. A total of 10 genes; *BZLF1*, *BRLF1*, *BDLF3.5*, *BDLF3*, *BORF1*, *BaRF1*, *BKRF2*, *BKRF2*, *LMP2A*, *LMP2B*, and *EBNA2* had recombination breakpoints per Kbp above this population mean i.e. 11.050. Of the 10 genes that reported high recombination breakpoints above the population mean, 7 were lytic genes (*BZLF1*, *BRLF1*, *BDLF3.5*, *BDLF3*, *BORF1*, *BaRF1*, *BKRF2* & *BKRF2*) while only 3 were latent genes (*LMP2A*, *LMP2B* & *EBNA2*).

Figure 4.3: Genomic Recombination Breakpoints Distribution in CDS.



Abbreviation: CDS, Coding sequence; Kbp, Kilobase pair. Each colored bar represents an EBV gene. The total number of CDS=42. Of the 42 CDS, 6 (14.3%) are latent genes and 36 (85.7%) are lytic genes. The bars are colored according to the classification of the genes in the EBV lytic cycle (Red; Early lytic genes, Blue; Immediate Early genes, Green; Late Lytic genes, Purple; Latent genes, Yellow; Uncharacterized gene. The black dotted strip denotes the mean number of recombination breakpoints per Kbp for all the genes (11.05).

4.3 Occurrence of Genomic Recombination across Age Groups, and between Males and Females

This study compared the occurrence of genomic recombination between EBV genomic sequences drawn from children of different age groups i.e. 0-4 years, 5-9 years, and 10-14 years and between males and females. The 86 genomes were classified as recombinant or non-recombinant based on the presence or absence of genomic recombination events. The study thereafter tested for the association between the recombination status of the genomes and the characteristics of the participants (Table 4.2). There was no association between recombination status and the age groups ($p=0.258$) just like gender ($p=1.000$). Figure 4.4A shows the distribution of unique genomic recombination events between age groups 0-4, 5-9, and 10-14 while Figure 4.4C shows their distribution between the males and the females. From the descriptive analysis (Figure 4.4A and 4.4C), the observable differences were noted in the

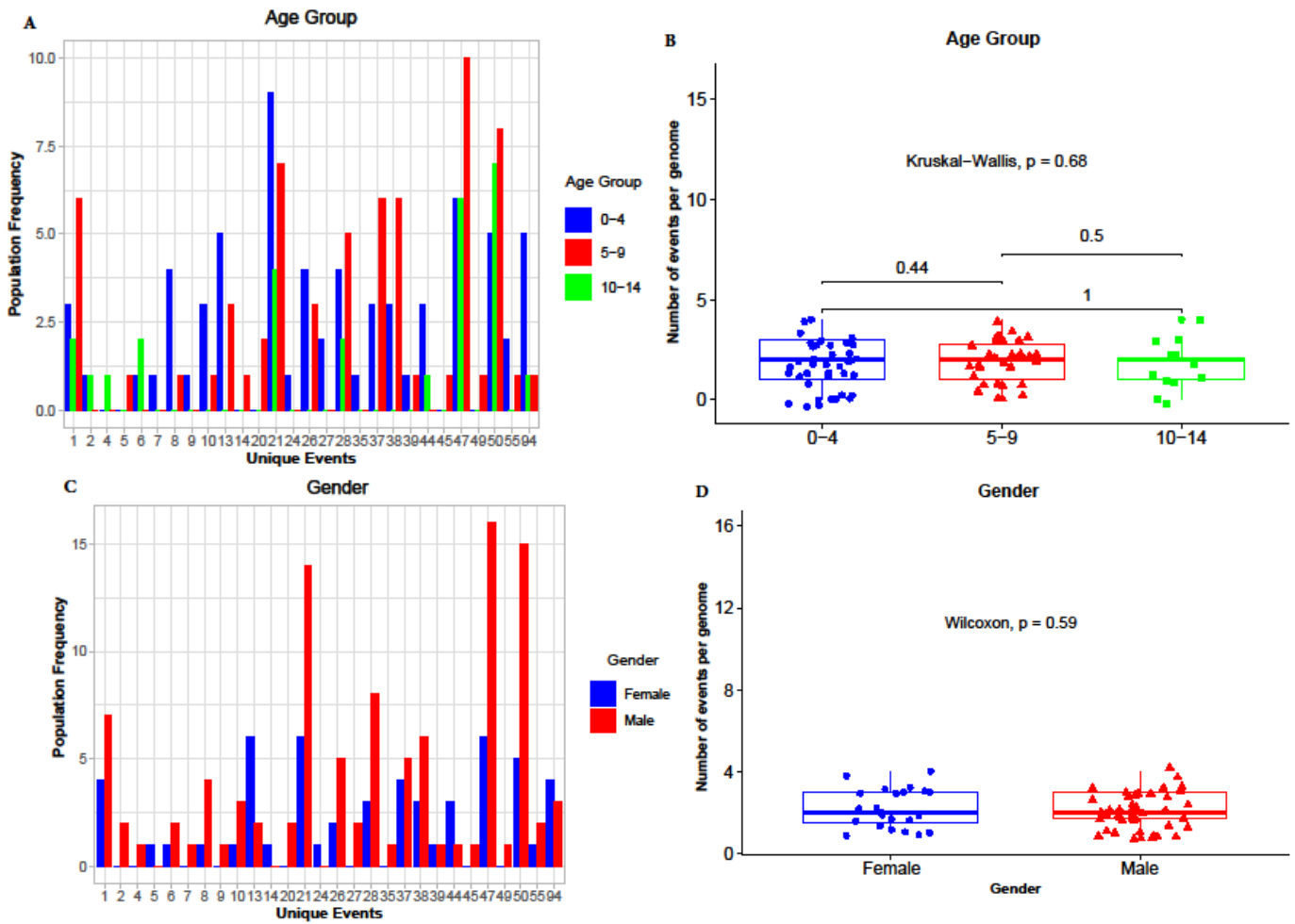
distribution of the unique recombination events. The number of genomic recombination events per genome was comparable across the age groups ($p=0.68$) (Figure 4.4B) and between males and females ($p=0.59$) (Figure 4.4D).

Table 4.2: Participant’s Characteristics associated with Genomic Recombination

Characteristic		Total	Recombinant (%)	Non-Recombinant (%)	<i>P-value</i>
	N	86	71 (82.6)	15 (17.5)	
Gender	Female	28	23 (82.1)	5 (17.9)	1.000 ^a
	Male	58	48 (82.3)	10 (17.2)	
Age Group	0-4	39	30 (76.9)	9 (23.9)	0.258 ^b
	5-9	36	30 (83.3)	6 (16.7)	
	10-14	11	11 (100)	0 (0)	
Viral Type	Type 1	56	51 (91.1)	5 (0.09)	0.011^a
	Type 2	30	20 (66.7)	10 (33.3)	
BL Status	eBL	54	48 (88.9)	6 (0.11)	0.086 ^a
	Healthy	32	23 (71.9)	9 (28.1)	

Abbreviation: eBL, endemic Burkitt Lymphoma. Recombinant and non-recombinant genomes were determined based on the presence or absence of recombinant segment/s respectively within it. Bold text indicates a statistically significant difference with a $P\text{-value} < 0.05$. Groups’ proportions were compared using ^aPearson’s Chi-square and ^bFisher exact tests.

Figure 4.4: Occurrence of Genomic Recombination across Age Groups and between Males and Females



A & C. Each bar plot represents the count of each unique event. B & D. Bold Lines represent medians, with lower and upper boundaries of the boxes representing first and third quartiles respectively. Wilcoxon (B) and Kruskal-Wallis (D) tests were performed and P -value < 0.05 was considered significant.

4.4 Association of Genomic Recombination Events with EBV Types and Diversity

As EBV type is the major classification of EBV diversity, the study sought to compare and contrast the occurrence of genomic recombination in type 1 genomic sequences and type 2 genomic sequences (Figure 4.5A). Statistical tests showed specific genomic recombination segments that were enriched among the type 1 genomic sequences and were corresponding to the distinct genomic recombination events; 28, 37, 38, 47, and 50 ($p=0.01$, 0.02 , 0.020 , 0.002 , and 0.0001 respectively). The recombinant segment corresponding to unique genomic recombination event 21 was highly enriched in Type 2 genomic sequences ($p=8.97e-10$) (Table 4.3). Each genomic recombination event had two recombination breakpoints that were cutting through specific CDS within the genomes. For instance, genomic recombination event 28 cutting through *BRLF1* at the start and through *BKRF2* at the end (Table 4.3).

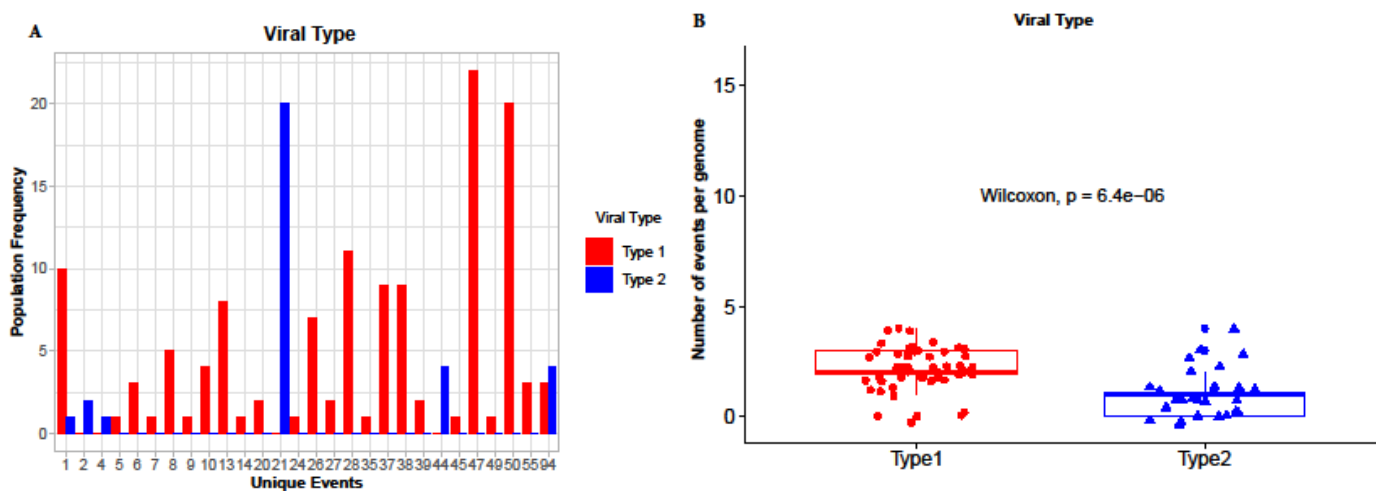
Table 4.3: Genomic Recombination Events association with EBV Types

Recombination Event	CDS cut by the Start Breakpoint	CDS cut by the end Breakpoint	Frequency in Type 1 (%)	Frequency in Type 2 (%)	P-value
21	BORF1, BORF2	BRLF1	0/56 (0)	20/30 (63)	8.97e-10
28	BRLF1	BKRF2	11/56 (20)	0/30 (0)	0.01
37	EBNA3B	BGLF1, BGLF4	9/56 (16)	0/30 (0)	0.02
38	BNRF1	BOLF1, BPLF1	9/56 (16)	0/30 (0)	0.02
47	BRLF1, BZLF1	BDLF3.5, BDLF4	22/56 (39)	0/30 (0)	0.002
50	LMP2A, LMP2B	EBNA2	20/56 (36)	0/30 (0)	0.0001

Abbreviation: eBL, endemic Burkitt Lymphoma, CDS, Coding Sequence. Bold text indicates a statistically significant difference with a $P\text{-value} < 0.05$. All groups' proportions were compared using Fisher's exact test.

The viral type was significantly associated with the recombination status of the genomic sequences ($p=0.011$) with more recombinant genomes being type 1 (71.8%) (Table 4.2). The study then compared the number of recombinant portions per genome between genomes that were type 1 and those that were type 2 (Figure 4.5B). Type 1 genomes had an average of 2.16 events per genome while type 2s had 1.03 events per genome. Consequently, type 1 genomes reported significantly more genomic recombination events per genome ($p=6.4e-06$).

Figure 4.5: Recombination Patterns between Type 1 and Type 2

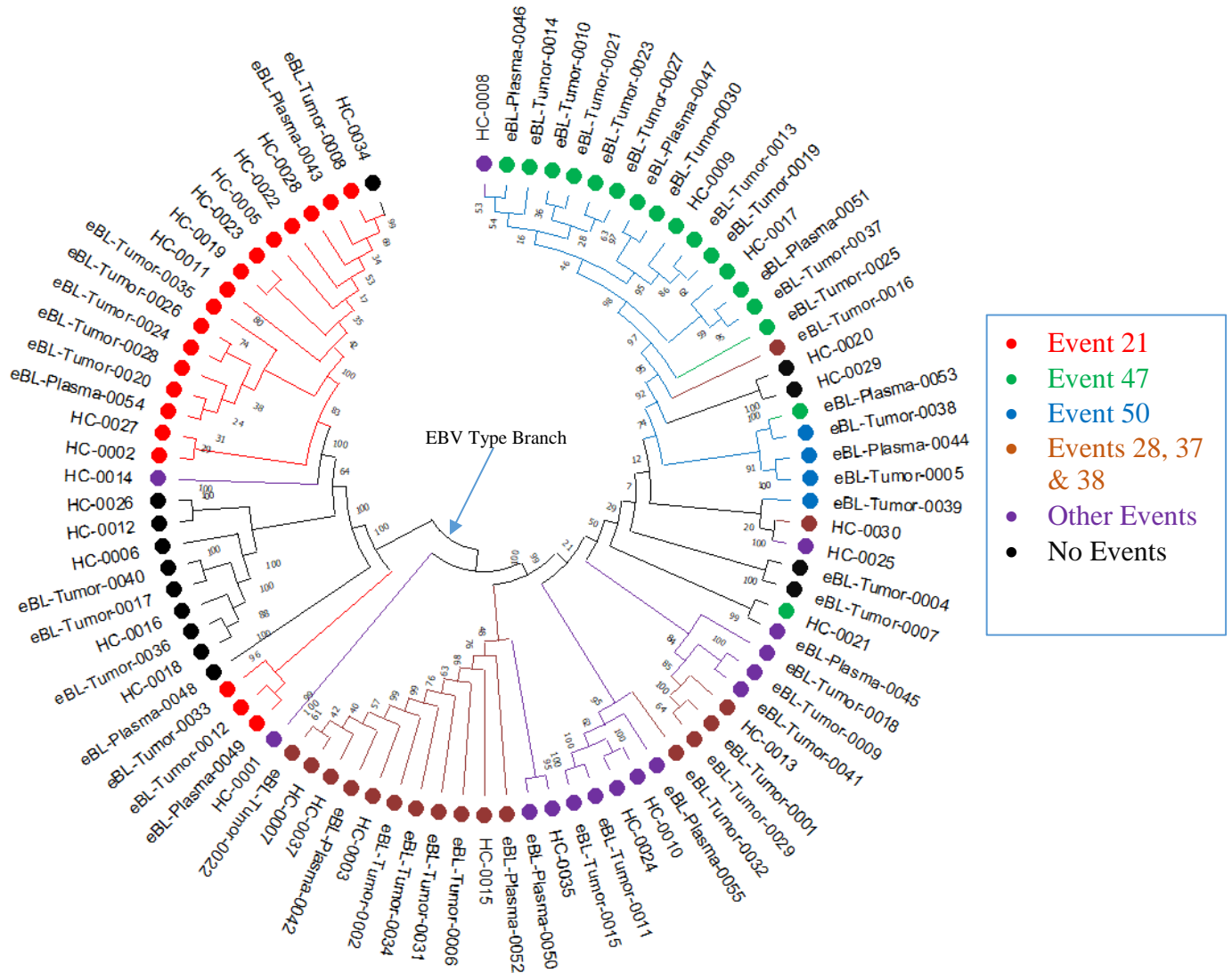


A. Each bar plot represents the count of each unique event. B. Bold Lines represent medians, with lower and upper boundaries of the boxes representing first and third quartiles respectively. Wilcoxon test was performed and P -value < 0.05 was considered significant.

Further, a phylogenetic tree of the EBV isolates was constructed and the isolates were colored by the occurrence of genomic recombination events (Figure 4.6). The first major division in the phylogenetic tree was between type 1 and type 2 genomic sequences. Based on the annotation, 19 genomic recombination events (67.9%) occur in multiple isolates and 9 events (32.1%) were exclusive to one genome. The events shared between multiple isolates, clustered by phylogenetic clades. Also to notice between the Type 1 and type 2 branches is that more type 2 genomes (10/30) had no evidence of recombinant portions compared to the type 1

genomes (5/56) (Table 4.2). The clustering of isolates in the type 2 branch was distinct to give 2 phylogenetic groups. The first phylogroup consisting of 11 isolates (Figure 4.6, annotated in black) had no evidence of recombination signatures and was much closer to the typing branch. The second phylogroup consisted of 17 isolates with evidence of recombination event 21 (Figure 4.6, annotated in red) convened distinctly away from the isolates of the first phylogroup. Strikingly, recombinant segments were hardly shared between EBV type 1 and type 2 genomes (Figure 4.5A & Figure 4.6); For instance, event 21 was only reported among the type 2 genomes and events; 28, 37, 38, 47, and 50 were exclusively in the type 1 genomes. Type 1 genomes reported 5 distinct genomic recombination events compared to the type 2 genomes which only reported 1. The values on the phylogenetic tree represent bootstrap support for each of the phylogenetic nodes. The bootstrap supports for the different phylogenetic nodes were ranging from 12, the lowest to 100, the highest. Most bootstrap supports were above 70%. Low bootstrap supports were reported among the recombinant nodes for instance around the nodes with genomic recombination events 21, 47, and 50. The node with no evidence of recombination (Figure 4.6, annotated in black) represented the best bootstrap supports of 100%.

Figure 4.6: Phylogenetic Tree of EBV Genomic Sequences showing Diversity related to Genomic Recombination Events



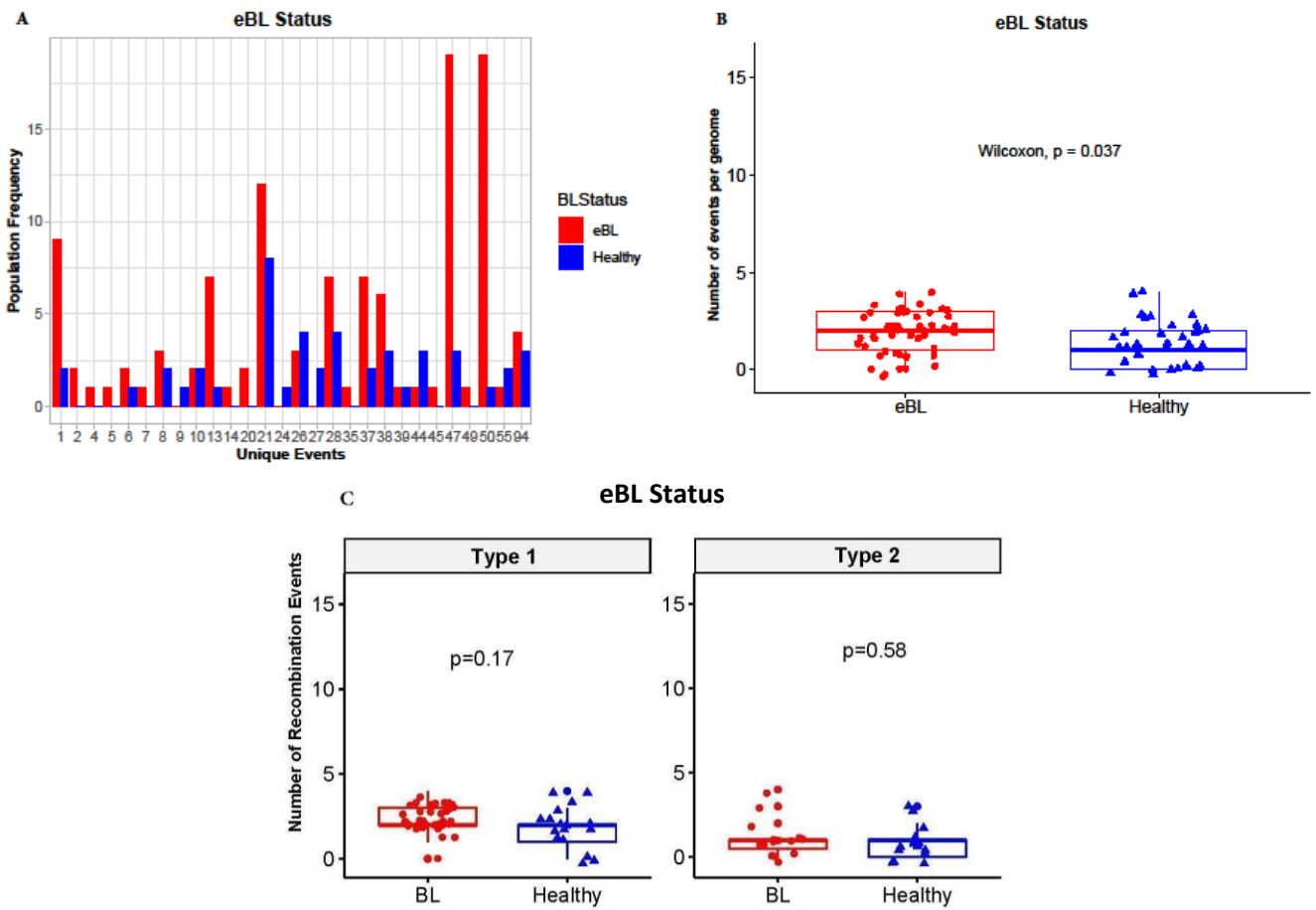
The analysis involved 86 genomic sequences. The evolutionary history was inferred using the NJ method. Evolutionary distances were computed using the Jukes-cantor model. Ambiguous nucleotides were removed using pairwise deletion. Bootstrap analysis of 5000 replicates was performed. The isolates were colored according to the type-associated recombination events detected. The phylogenetic tree is rooted to the midpoint branch i.e. EBV type branch.

4.5 Genomic Recombination Events in the Genomic Sequences from the eBLs and

Healthy Participants

The study compared the proportions of recombinant and non-recombinant genomes drawn from the healthy and the eBLs (Figure 4.7A). There was no association between recombination status and eBL status ($p=0.086$) (Table 4.2). The study thereafter compared the number of genomic recombination events per genome between the genomic sequences of the eBLs and healthy children. The eBLs reported more genomic recombination events per genome ($p=0.037$) (Figure 4.7B). Since the study had already shown differences between type 1 and type 2 genomes, it assessed the occurrence of genomic recombination between eBLs and the healthy in each EBV type. There was no significant difference in the number of genomic recombination events per genome in eBLs and healthy when EBV type 1 and 2 genomic sequences were compared separately (Among EBV type 1 genome; $p=0.17$ & Among EBV type 2 genomes; $p=0.58$) (Figure 4.7C). The mean and interquartile values were higher in the eBLs ($mean=2.282$, $range=2.00-3.00$) compared to the healthy ($mean=1.882$, $range=1.00-2.00$), particularly among the type1 genomic sequences.

Figure 4.7: Genomic Recombination Events in the Genomic Sequences from the eBLs and Healthy Participants



Each bar plot represents the count of each unique event. B & C. Bold Lines represent medians, with lower and upper boundaries of the boxes representing first and third quartiles respectively. Wilcoxon test was performed and P -value < 0.05 was considered significant.

It may also be possible that specific genomic recombination events are associated with eBL risk so the study probed eBL association with distinct genomic recombination events (Table 4.4). Two genomic recombination events were significantly enriched in the eBLs by univariate logistic regression; event 47 ($OR=4.07$, $p=0.038$) and 50 ($OR=14.24$, $p=0.012$). The coordinates of the breakpoints associated with these events may have biological significance that can inform their association with disease. Events 47 breakpoints cut through; *BRLF1*, *BZLF1*, *BDLF3.5*, *BDLF4* while event 50 associated breakpoints cut through; *LMP2A*, *LMP2B*, *EBNA2*. Controlling EBV viral type using multivariate logistic regression, only event 50 was statistically significantly enriched in the eBLs (event 50; $OR=12.36$, $p=0.020$). Event 50 showed a trend towards eBL association when the study controlled for EBV type in a similar multivariate logistic regression (event 47; $p=0.089$, $OR=3.31$).

Table 4.4: eBL Association with two Genomic Recombination Events

Recombination Event	CDS cut by the start Breakpoint	CDS cut by the End Breakpoint	Frequency in BLs (%)	Frequency in Healthy (%)	Without Controlling for Viral Type		Controlling for Viral Type	
					OR ^a (95% CI)	<i>P</i> -value	OR ^b (95% CI)	<i>P</i> -value
47	BRLF1, BZLF1	BDLF3.5, BDLF4	16/54 (29.6)	3/32 (9.4)	4.07 (1.21-1.87)	0.038^a	3.31 (0.99-1.58)	0.089 ^b
50	LMP2A, LMP2B	EBNA2	17/54 (31.5)	1/32 (3.1)	14.24 (2.69-2.84)	0.012^a	12.36 (2.18-2.34)	0.020^b

Abbreviation: CDS, Coding Sequence, eBL; endemic Burkitt Lymphoma; OR, Odds Ratio; Ref, Reference. Bold text indicates a statistical significance with a *P*-value < 0.05. ^aUnivariate and ^bMultivariate logistic regression were used to compute the Odds Ratios and *P* values non-significant *P*-value by univariate analysis.

CHAPTER FIVE

DISCUSSION

5.1 General Introduction

This study characterized the genome-wide occurrence of genomic recombination events and investigated their association with age, gender, viral type, diversity, and eBL. Despite, the sample size reducing from 95 to 86, the study still had enough statistical power to detect differences across these groups. As a result, the study found 28 genomic recombination events in 82.6% of the genomes analyzed with most genomic recombination breakpoints in genes of the EBV lytic cycle. The occurrence of genomic recombination events is not significantly associated with age and gender, but are associated with EBV type with more genomic recombination events in EBV type 1 genomic sequences. More genomic recombination events are reported in the EBV genomic sequences from the eBLs. Overall, this study has investigated genomic recombination as a source of genetic diversity in EBV and has adduced evidence linking these genomic recombination events and their breakpoints to EBV's biology of B cell transformation and eBL pathogenesis.

5.2 Genome-wide occurrence of Genomic Recombination Events and Breakpoints

This study reports 28 different genomic recombination events shared between EBV genomes from western Kenya with 82.6% of the genomic sequences analyzed bearing one or more recombinant segments. This evidence of genomic recombination in EBV genomes suggests that the EBV genome is prone to genomic recombination. The recombinant nature of the EBV genome has been demonstrated in genomic sequences of EBV isolates from other geographical regions other than western Kenya (Chiara *et al.*, 2016; Palser *et al.*, 2015; Zanella *et al.*, 2019). Overall, the high propensity of EBV to experience genomic recombination may

be attributed to the inability of the human host immune response to sufficiently prevent new EBV infections. Genomic recombination requires two or more EBV genomes to co-infect a host cell and exchange genetic segments (Pérez-Losada *et al.*, 2015) thus multiple EBV infections may increase the chances of EBV genomes exchanging genomic segments. Zero in on western Kenya, these EBV isolates are highly likely to experience genomic recombination because of the repeated exposure to *Pf* infection and early age infection with EBV. Repeated *Pf* exposure activates the polyclonal expansion of the B cells, causing EBV reactivation and a hike in peripheral blood viral loads both of which are known promoters of genomic recombination (Daud *et al.*, 2015). The study reported varying numbers of recombinant fragments within these genomic sequences providing a hint that the rates of occurrence of genomic recombination differ among these genomic sequences despite being drawn from the same geographical area. Characteristics other than the geographical variation may therefore be critical drivers of the occurrence of genomic recombination. Such characteristics may include the gender of the participant, the clinical condition, the age, or the EBV type of the genomic sequence being analyzed. In future studies, it will be necessary to make global comparisons of EBV genomic recombination as this may better elucidate the factors implicated in the occurrence of EBV genomic recombination. The evidence of genomic recombination adduced from this study was obtained after analyzing 51 % of the whole EBV genome. There is the possibility of missing genomic recombination events in the genomic regions not analyzed hence the genomic recombination events detectable across the whole EBV genome may be higher than what is represented here. This however may not affect the comparisons between groups since all the genomic sequences analyzed were 51% of the whole genome.

Each of the 28 genomic recombination events detected was associated with 2 genomic recombination breakpoints i.e. a start breakpoint and an end breakpoint. The breakpoints were found to cut through 42 distinct CDS along the EBV genome. These genes affected by

recombination breakpoints may be a good link to the biological implications of each unique genomic recombination. Genomic recombination breakpoints can affect the property of genes and in turn, their functionality (Berenstein *et al.*, 2018) hence identifying the genes affected by these breakpoints was vital to understand the possible contribution of these genomic recombination events in disease. EBV genome has more than 80 coding regions whose protein products play different roles in the life cycle of the virus (Sample *et al.*, 2009). This study performed a genome-wide study of genomic recombination without targeting any of these coding regions to limit biases thus providing a reliable account of genomic recombination along the genome. This genome-wide approach identified 42 different coding sequences that were cut by one or more genomic recombination breakpoints. Recombination breakpoints per kilobase base pair (bp) were computed for each of the 42 CDS to allow for the comparison of the rates of occurrence of recombination breakpoints across the various genes with different lengths. The mean recombination breakpoint per kbp for all the 86 genomic sequences analyzed was used as a cut-off to determine the CDS with elevated counts of recombination breakpoint per kbp. The *BZLF1* and *BRLF1* genes, which are immediate-early lytic genes encoding transcriptional factors Zta and RTA respectively (Li *et al.*, 2016), reported recombination per kbp above the mean recombination breakpoints per kbp for the population. Other CDS that also had recombination breakpoints per kbp above the population mean were *BKRF2*, *BDL3.5* and *BDLF4*.

According to the findings of this study, genes involved in the lytic EBV cycle are more prone to suffer recombination breakpoints compared to the genes of the latent phase. The site of occurrence of genomic recombination breakpoints along the genome has been shown in HSV to be a good indicator of the molecular mechanisms implicated in genomic recombination (Lee *et al.*, 2015). Going by this observation in HSV, the higher likelihood of recombination breakpoints in EBV lytic genes can be adduced to the fact that EBV genomic recombination is

intimately linked to EBV lytic phase which involves lytic reactivation and replication as key processes. For instance, the high-level expression of *BZLF1* and *BRLF1* is sufficient to induce the switch from the latent to the lytic form of EBV infection (Murata *et al.*, 2021); *BKRF2* encodes virion glycoprotein gL required for cell-to-cell spread (Swaminathan & Kenney, 2008) while *BDLF 3.5* and *BDLF4* are expressed in Early (E) kinetics both which are required for efficient expression of late lytic genes (Rosemarie & Sugden, 2020). During EBV lytic phase, the ds DNA opens, creating a high propensity for the production of strand breaks and concatemers required to facilitate the exchange of genomic segments between genomes (Hammerschmidt & Sugden, 2013).

Mapping of genomic recombination breakpoints against the EBV genome map confirms that the EBV genome displays a heterogeneous landscape of genomic recombination breakpoints. This heterogeneity is characterized by clusters of recombination breakpoints in some genomic regions and no recombination breakpoints in other genomic regions. This study reports the biasness of genomic recombination breakpoints towards the genomic and genic regions implicated in EBV lytic reactivation and replication. Besides the positions of the recombination breakpoints, the circular map also displays the coordinates of the MSA with the good coverages as well as the EBV repeats regions. Illumina being a short reads sequencer, the reads covering the EBV repeats are not efficiently assembled because of the high nucleotide similarity and ambiguity in the regions spanning these repeats (Zanella *et al.*, 2019). These regions are normally poorly aligned and may cause artificial genomic diversity hence were preferably trimmed out to ensure the phylogenetic and recombination analysis was based on reliable genomic content. All EBV genomes have a similar genomic and genic structure (Sample *et al.*, 2009) hence the gene map used was representative of all the 86 genomic sequences that were analyzed. Further, recombination in EBV is homologous i.e. genomic recombination occurs in the same site in both parental strands hence there is normally no

change in the genomic or genic structure of the EBV recombinant genomes (Pérez-Losada *et al.*, 2015). However, on EBV diversity, the recombinant genome carries the genetic variations drawn from different parental genomes hence considered more diverse than any of the parental genomes.

5.3 Occurrence of Genomic Recombination across Age Groups, and between Males and Females

Age and gender are critical demographic factors that may influence how individuals within a population respond to EBV infection. The participants were stratified as age 0-4, 5-9, and 10-14 years based on the temporal association between EBV, *Pf* infection, and occurrence of eBL similar to a previous study (Oluoch *et al.*, 2020). Several biological differences have been observed across these age groups hence this study hypothesized that the occurrence of genomic recombination would be different across the age groups. For example, at the age of 4 years, children from western Kenya already harbor EBV infection (Piriou *et al.*, 2012) and have experienced multiple *Pf* infections (Chattopadhyay *et al.*, 2013). Age 5-9 is the peak of eBL occurrence (Rainey *et al.*, 2007) while age 10-14 has acquired immune response to *Pf* and EBV infections (Griffin *et al.*, 2015). Despite these known differences, this study reports comparable patterns of recombination in children between 0-4 years, 5-9 years, and 10-14 years. This study suggests that there is minimal accrual of genomic recombination events over the period of 10 or 14 years. Based on the study findings, it requires a longer period of time of more than 10 or 14 years for genomic recombination events to be incorporated into a population and this does not imply that the virus genome remains static the whole period of time. Permanent incorporation of a genomic recombination event into a genome and spread within a population is influenced by selective forces exerted on the virus genome therefore some genomic

recombination events may occur but may be cleared out. EBV genome being relatively stable these genomic recombination events are likely inherited down the lineages from parental strains. It may also imply that genomic recombination is acquired early during primary EBV infection as a result of the first one-to-one battle between the virus and host hence the genomes isolated from the younger children reflect the genomes of the older children.

Gender is also an important cofactor in the control of viral infection with studies have shown that females unlike males are better at mounting immune responses to viral infections (Klein, 2012; Pradhan & Olsson, 2020). This study, therefore, sought to investigate if this parity in viral immune response between males and females would influence differences in the genome-wide occurrence of genomic recombination. This study, however, reports no differences in the genome-wide occurrence of genomic recombination between males and females. This observation supports the premise that the occurrence of genomic recombination EBV may be a snapshot of early adaptation of EBV. Such genomic recombination events are then handed down from parental lineages so that they persist within the population. Most of these genomic recombination events are not acquired temporarily like in the case of gene plasticity but are permanently incorporated within the genomes and significantly spread within the population. These processes take a long period of time and are not dependent on the gender of the participant from which the EBV was isolated.

5.4 Association of Genomic Recombination Events with EBV Types and Diversity

EBV type is the major diversity in EBV genomes (Tzellos & Farrell, 2012). Since EBV type 1 and type 2 differ in their capacity to immortalize B cells and cause disease, efforts have been channeled to understand any underlying genetic variation between EBV types that may inform this difference. In this regard, the study compared the genome-wide occurrence of genomic recombination between EBV type 1 and EBV type 2 genomes. According to the

findings, EBV types have differences in their genome-wide occurrence of genomic recombination with EBV type 1 genomes reporting more genomic recombinations. This suggests that EBV type 1 has accumulated more genomic recombination events in their evolutionary journey as compared to EBV type 2. Moreover, EBV type 1 genomes have displayed greater diversity in earlier studies (Kaymaz *et al.*, 2020; Panea *et al.*, 2019) and these accrued genomic recombination events may be a key contributing factor to EBV type 1 diversity. EBV type differences therefore may extend beyond the obvious divergence in the variations of *EBNA2*, *EBNA 3A*, *3B*, and *3C* genes warranting more studies to demonstrate their genetic differences and their implications in disease. To further demonstrate the EBV type 1 and type 2 differences, an association test was done on each of the 28 genomic recombination events detected within the population. Five genomic recombination events were associated with EBV type 1; Event 21, 28, 38, 47, and 50 and one genomic recombination event; Event 21 was associated with EBV type 2. This observation still emphasizes that EBV type 1 has accumulated more genomic recombination events in the course of evolution. Besides, each type-associated recombination event just like other genomic recombination events had its 2 breakpoints pervading the CDS of genes. The affected genes have vital roles in the EBV biology of B cell transformation and could significantly influence the oncogenic potential of the EBV isolates. These findings demonstrate differential patterns of genomic recombination between EBV types and contribute critical answers to the question of the relative tumorigenicity of EBV types.

To further understand the contribution of genomic recombination events on EBV diversity, an NJ phylogeny was constructed and annotated according to genomic recombination events detected. The positions and clustering of the isolates on the phylogenetic tree were studied to explain the genetic divergence seated within the 86 genomic sequences that were analyzed. The first observable split on the phylogenetic tree is the EBV type split which is the

main genetic diversity of EBV genomes. This first observation points to the significance of phylogeny in representing the diversity of genomic sequences. According to the annotation of the phylogenetic tree, more genomic recombination events were shared between genomic sequences within the dataset analyzed. This finding confirms that genomic recombination events are old enough to be shared and spread within the population and are not merely the case of temporary gene plasticity. It further emphasizes the initial findings of this study that most genomic recombination events are either vertically acquired from the parents or acquired early during primary infection. Moreover, most of these genomic recombination events were shared by EBV isolates belonging to a phylogenetic clade. A phylogenetic clade in this case was defined as a group of isolates descending from a common ancestor or ancestral node (Rieux & Balloux, 2016). Type 2 EBV genomes are separated into two 2 distinct phylogroups as proposed in a previous study (Kaymaz *et al.*, 2020). This finding of this study demonstrates a possible contribution of genomic recombination event 21 in the occurrence of these novel substructures in type 2 EBV genomes.

Bootstrapping is important to determine the support given to a phylogenetic node adduced as evidence of evolutionary history on a phylogenetic tree (Beiko & Hamilton, 2006). The phylogenetic tree constructed in this study had varying values of bootstrap supports ranging from 12% to 100% and this is because the dataset contained a mixer of non-recombinant and recombinant genomic sequences. Recombination bars the phylogenetic inference of a single evolutionary history since the recombinant genomic sequences contain genomic segments possibly acquired from different parents with different ancestral lineages (Rieux & Balloux, 2016). However, phylogenetic reconstruction methods such as NJ outputs the most suitable phylogenetic tree with the grand most recent common ancestors (GMRCA) supported at different nodes (Simonsen *et al.*, 2011; Yoshida & Nei, 2016) and therefore the study was able to reconstruct a single phylogenetic tree despite the evidence of recombination.

Because of this, the phylogenetic nodes with evidence of genomic recombination events reported low bootstrap values, and the non-recombinant reported 100% bootstrap values. Since recombinant genomic sequences contain genomic segments with different ancestral lineages, the confidence of the node supports is affected leading to low bootstrap values as observed in this study. A high bootstrap value of 10000 was used to ensure that no false results were supported. This observation still shows that EBV has evidence of genomic recombination across its genomic sequences and most of the genomic recombination events reported in this study are occurring at ancestral nodes hence shaping the positioning and clustering of isolates on the phylogenetic tree.

5. 5 Genomic Recombination Events in the Genomic Sequences from the eBLs and Healthy Participants

This study presents the first comparison of recombination patterns between viral genomes generated from the archival of eBL positive children and geographically matched healthy controls. The study hypothesized that the EBV isolates from the eBL positive participants bear genomic recombination events which gives them an advantageous age over the human host augmenting the progression to disease. The genomes from the eBLs report higher genomic recombination events compared to the healthy, suggesting the availability of factors that increase the propensity of the virus genomes to recombine in the eBLs. The eBL positive children have a lower EBV-specific immune response compared to the healthy children (Chattopadhyay *et al.*, 2013; Njie *et al.*, 2009). Based on these previous findings, eBLs positive children report high viral loads (Westmoreland *et al.*, 2017), higher chances of lytic replication as well as reactivation (Snider *et al.*, 2012), and high chances of multiple EBV

infections. These factors are possible potentiates of genomic recombination between EBV genomes.

In the primary analysis, the genomic recombination events between eBLs and healthy cohorts were statistically different, however, when type 1 and type 2 genomic sequences were studied separately, no statistically significant difference was observed. In this secondary analysis, the number of genomic sequences analyzed were 56 and 30 for type 1 and type 2 genomic sequences respectively representing a reduction in sample size. The statistical p values are normally subject to the statistical power which is also dependent on the sample size (Murphy *et al.*, 2014) hence these lower sample sizes used in this secondary analysis may not bear sufficient statistical power to demonstrate the differences between the eBLs and the healthy. In this scenario, a number of case studies recommend that an observable trend can be discussed (Murphy *et al.*, 2014). These secondary comparisons show higher means and interquartile values for the number of genomic recombination events among the eBLs, particularly among EBV type 1 genomic sequences. Future studies may require more type 1 and type 2 genomic sequences to efficiently compare genomic recombination between the eBLs and healthy cohorts when type 1 and type 2 genomic sequences are studied separately. Based on these findings, genomic recombination events may be a risk factor to eBL in a manner that still calls for more scrutiny.

The identified differences between genomic sequences from the eBLs and healthy children sparked the interest to identify potential associations between recombination breakpoints and the occurrence of eBL. Previous studies have identified breakpoints cutting through genes that had biological significance and may significantly contribute to disease (Berenstein *et al.*, 2018; Zanella *et al.*, 2019). The identified 7 gene pervading recombination breakpoints enriched in the eBLs that were cutting through; *BRLF1*, *BZLF1*, *BDLF3.5*, *BDLF4*, *LMP2A*, *LMP2B*, and *EBNA2* genes. The known roles of these genes in EBV's biology of B

cell transformation point to their contribution to eBL onset and progression. Future mechanistic studies however are required to determine the actual roles of these recombination events and their breakpoint in disease. Breakpoints in *LMP2A*, *LMP2B*, and *EBNA2* may change their immunogenic determinants affecting their MHC binding and subsequent recognition by T cell receptors providing a route for EBV immune escape similar to previous findings in HSV (Lee *et al.*, 2015). Pervasive recombination disrupting genes may also affect gene property and functionality to influence the virulence and pathogenicity of EBV isolates (de Been *et al.*, 2013; Sijmons *et al.*, 2015). Other genes include; *BDLF3.5* and *BDLF4* genes whose products are required for the expression of late lytic genes (McKenzie & El-Guindy, 2015); *LMP2A* is known to induce the expression of genes involved in cell-cycle induction, inhibition of apoptosis, and suppression of cell-mediated immunity (Kanda, 2018); *LMP2B* is a potential oncogene (Tzellos & Farrell, 2012); *EBNA2* which interacts with sequence-specific DNA binding protein, Janus kinase recombination binding protein (RBP-JK), to transcriptionally activate cellular genes such as *CD23* and the key viral genes including *LMP1* and *LMP2A* (Kang & Kieff, 2015). From this study, specific recombination events and their breakpoints cut genes that play critical functions in viruses and this may influence the virulence and pathogenicity of EBV isolates. Consequently, this may augment the development of eBL in children bearing these EBV isolates.

5.6 Study Limitations

Even with DNA enrichment methods, capturing viral DNA from healthy individuals remains a challenge (Depledge *et al.*, 2011). Therefore 9 genomic sequences that were not efficiently covered were struck from the analysis. However, the new sample size of 86 still had enough statistical power to answer the objectives of the study (Appendix IV). Further, our analysis used about 50% of the whole genome since poorly aligned sites were excluded from

the analysis. This offered the advantage of using only the regions that had reliable genomic content (Talavera & Castresana, 2007) however, there is a possibility of missing genomic recombination events occurring by chance in the genomic regions that were not analyzed in this study. Further, the study was limited to associations between genomic recombination and eBL occurrence since mechanistic studies were beyond the scope of this study design.

CHAPTER SIX

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

6.1 Summary of Study Findings

In the current study, use is made of tools in RDP4 to infer recombination patterns across the EBV genome, their relationship with age, gender, EBV type and diversity, and their role in eBL pathogenesis. In summary, EBV genome architecture has evidence of genomic recombination that seems to have been acquired over years and transferred vertically down the EBV lineages. Such recombination displays a heterogeneous landscape across the EBV genome with lytic genes having a higher propensity to harbor recombination events. Recombination is also described as an evolutionary force that impacts EBV diversity as predicted from how it influences the clustering and position of EBV isolates on the phylogenetic tree. Further, the age and gender of participants from each EBV are isolated do not influence the occurrence of recombination in the virus. EBV type 1 genomic sequences have accrued more genomic recombination events, making them more diverse than type 2 genomic sequences, and are highly likely to harbor risk variants that may potentiate eBL oncogenesis. Overall, these findings address the complexities of genomic recombination in EBV, its association with EBV genetic diversity, and provide novel insight into viral variation which has the potential to influence EBV's biology and eBL pathogenesis.

6.2 Conclusions

1. EBV genome has is a high likelihood to experience genomic recombination with a heterogeneous landscape of genomic recombination around it. Genes of the lytic EBV phase are more prone to genomic recombination breaks compared to the genes of the latent phase.

2. Age and sex do not influence the occurrence of recombination and this implies that there are minimal accrual genomic recombination events with repeated infections over time.
3. EBV type 1 genomic sequences have accumulated more genomic recombination events in their evolutionary journey compared to EBV type 2 genomic sequences hence are more genetically diverse. Further, genomic recombination significantly influences the genetic diversity of EBV genomes.
4. Genomic sequences from the eBL cohorts have higher genomic recombination events compared to the genomic sequences from the healthy cohorts. High genomic recombination rates among the genomic sequences from the eBLs suggest the availability of factors that increases the propensity for genomic recombination.

6.3 Recommendations from this Study

1. EBV genome is from western Kenya is highly likely to experience genomic recombination. Comparisons should be made with EBV genomic sequences from other geographical regions to examine the full extent of EBV genomic recombination on a global scale.
2. Age and gender may not be important factors in the epidemiological surveillance of children from western Kenya who are highly likely to develop eBL risk variants resulting from recombination.
3. Since children harboring EBV type 1 genomic sequences are have accumulated more recombination events they need to be monitored as having a greater likelihood of developing eBL risk variants.

4. Genomic sequences from the eBL cohorts have higher genomic recombination events compared to the genomic sequences from the healthy cohorts therefore this study recommends that these children should be monitored in order to determine the possible causes of the high genomic recombination rates and find ways to mitigate this.

6.4 Recommendations for Future Studies

1. In the future, long-read sequencing methods should be used to generate complete EBV genomes that will allow genomic recombination analysis along the entire 172 kbp genome.
2. Improvement in EBV DNA capture methods is needed to be able to better enrich EBV DNA and allow for complete sequencing of EBV genomes from healthy participants.
3. Mechanistic study designs should be adopted to probe the actual roles of genomic recombination events and their breakpoints in diseases such as eBL.
4. EBV genomic sequences from individuals with different clinical manifestations should be added to better decipher the role of recombination in disease.
5. Further, gene-based analysis of recombination should also be carried out to understand the rates of recombination between genes.

REFERENCES

- Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, 34(5), 502–508
- Assefa, S., Keane, T. M., Otto, T. D., Newbold, C., & Berriman, M. (2009). ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, 25(15), 1968–1969
- Ballesteros-Zebadúa, P., Villarreal, C., Cocho, G., Huerta, L., & Estrada, J. L. (2013). Differences in HIV-1 Viral Loads Between Male and Female Antiretroviral-untreated Mexican Patients. *Archives of Medical Research*, 44(4), 296–301
- Beiko, R. G., & Ragan, M. A. (2008). Detecting Lateral Genetic Transfer. In J. M. Keith (Ed.), *Bioinformatics: Data, Sequence Analysis and Evolution* (pp. 457–469)
- Beiko, R., & Hamilton, N. (2006). *Phylogenetic identification of lateral genetic transfer events*. *Genetics*, 2(15), 56–66
- Berenstein, A. J., Lorenzetti, M. A., & Preciado, M. V. (2018). Recombination rates along the entire Epstein Barr virus genome display a highly heterogeneous landscape. *Infection, Genetics and Evolution*, 65, 96–103
- Berntsson, M., Dubicanac, L., Tunbäck, P., Ellström, A., Löwhagen, G.-B., & Bergström, T. (2013). Frequent detection of cytomegalovirus and Epstein-Barr virus in cervical secretions from healthy young women. *Acta Obstetrica Et Gynecologica Scandinavica*, 92(6), 706–710
- Boni, M. F., Posada, D., & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, 176(2), 1035–1047
- Brooks, J. M., Long, H. M., Tierney, R. J., Shannon-Lowe, C., Leese, A. M., Fitzpatrick, M., Taylor, G. S., & Rickinson, A. B. (2016). Early T Cell Recognition of B Cells following

- Epstein-Barr Virus Infection: Identifying Potential Targets for Prophylactic Vaccination. *PLOS Pathogens*, 12(4), e1005549.
<https://doi.org/10.1371/journal.ppat.1005549>
- Buckle, G., Maranda, L., Skiles, J., Ong'echa, J. M., Foley, J., Epstein, M., Vik, T. A., Schroeder, A., Lemberger, J., Rosmarin, A., Remick, S. C., Bailey, J. A., Vulule, J., Otieno, J. A., & Moormann, A. M. (2016). Factors influencing survival among Kenyan children diagnosed with endemic Burkitt lymphoma between 2003 and 2011: A historical cohort study. *International Journal of Cancer*, 139(6), 1231–1240
- Chang, C. M., Yu, K. J., Mbulaiteye, S. M., Hildesheim, A., & Bhatia, K. (2009). The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: A need for reappraisal. *Virus Res.*, 143(2), 209–221.
- Chattopadhyay, P. K., Chelimo, K., Embury, P. B., Mulama, D. H., Sumba, P. O., Gostick, E., Ladell, K., Brodie, T. M., Vulule, J., Roederer, M., Moormann, A. M., & Price, D. A. (2013). Holoendemic Malaria Exposure Is Associated with Altered Epstein-Barr Virus-Specific CD8 T-Cell Differentiation 87 (10), 143-516
- Chen, J., Zhou, L., Qiu, X., Yang, R., Liang, J., Pan, Y., Li, H., Peng, G., & Shao, C. (2018). Determination and genome-wide analysis of Epstein-Barr virus (EBV) sequences in EBV-associated gastric carcinoma from Guangdong, an endemic area of nasopharyngeal carcinoma. *Journal of Medical Microbiology*, 67(11), 1614–1627
- Chiara, M., Manzari, C., Lionetti, C., Mechelli, R., Anastasiadou, E., Chiara Buscarinu, M., Ristori, G., Salvetti, M., Picardi, E., D'Erchia, A. M., Pesole, G., & Horner, D. S. (2016). Geographic Population Structure in Epstein-Barr Virus Revealed by Comparative Genomics. *Genome Biology and Evolution*, 8(11), 3284–3291

- Combelas, N., Holmblat, B., Joffret, M.-L., Colbère-Garapin, F., & Delpeyroux, F. (2011). Recombination between Poliovirus and Coxsackie A Viruses of Species C: A Model of Viral Genetic Plasticity and Emergence. *Viruses*, 3(8), 1460–1484
- Conant, G. C., & Wolfe, K. H. (2008). GenomeVx: Simple web-based creation of editable circular chromosome maps. *Bioinformatics*, 24(6), 861–862
- Daud, I. I., Ogolla, S., Amolo, A. S., Namuyenga, E., Simbiri, K., Bukusi, E. A., Ng'ang'a, Z. W., Ploutz-Snyder, R., Sumba, P. O., Dent, A., & Rochford, R. (2015). Plasmodium falciparum infection is associated with Epstein-Barr virus reactivation in pregnant women living in malaria holoendemic area of Western Kenya. *Matern. Child Health J.*, 19(3), 606–614.
- de Been, M., van Schaik, W., Cheng, L., Corander, J., & Willems, R. J. (2013). Recent Recombination Events in the Core Genome Are Associated with Adaptive Evolution in *Enterococcus faecium*. *Genome Biology and Evolution*, 5(8), 1524–1535
- De Fonzo, V., Aluffi-Pentini, F., & Parisi, V. (2007). Hidden Markov Models in Bioinformatics. *Current Bioinformatics*, 2(1), 49–61
- Domínguez-Rodríguez, S., Serna-Pascual, M., Foster, C., Palma, P., Nastouli, E., De Rossi, A., Seoane, J., Rossi, P., Giaquinto, C., Tagarro, A., Rojo, P., & EPIICAL Project. (2021). Faster Initial Viral Decay in Female Children Living With HIV. *Journal of the Pediatric Infectious Diseases Society*, 10(5), 674–676
- Erickson, K. (2010). The Jukes-Cantor Model of Molecular Evolution. *PRIMUS*, 20(5), 438–445
- Farrell, P. J. (2019). Epstein-Barr Virus and Cancer. *Annual Review of Pathology*, 14, 29–53
- Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S., & Michalakis, Y. (2005). Recombination Every Day: Abundant Recombination in a Virus during a Single Multi-Cellular Host Infection. *PLOS Biology*, 3(3), e89

- González-Candelas, F., López-Labrador, F. X., & Bracho, M. A. (2011). Recombination in Hepatitis C Virus. *Viruses*, 3(10), 2006–2024
- Griffin, J. T., Hollingsworth, T. D., Reyburn, H., Drakeley, C. J., Riley, E. M., & Ghani, A. C. (2015). Gradual acquisition of immunity to severe malaria with increasing exposure. *Proceedings of the Royal Society B: Biological Sciences*, 282(1801)
- Habibian, A., Makvandi, M., Samarbaf-Zadeh, A., Neisi, N., Soleimani-Jelodar, R., Makvandi, K., & Izadi, S. (2018). Detection and Genotyping of Epstein-Bar Virus Among Paraffin Embedded Tissues of Hodgkin and Non-Hodgkin's Lymphoma Patients in Ahvaz, Iran. *Acta Medica Iranica*, 434–440.
- Hackenberger, B. K. (2020). R software: Unfriendly but probably the best. *Croatian Medical Journal*, 61(1), 66–68
- Hämmerl, L., Colombet, M., Rochford, R., Ogowang, D. M., & Parkin, D. M. (2019). The burden of Burkitt lymphoma in Africa. *Infectious Agents and Cancer*, 14, 17
- Hammerschmidt, W., & Sugden, B. (2013). Replication of Epstein-Barr viral DNA. *Cold Spring Harbor Perspectives in Biology*, 5(1), a013029
- Huang, S.-Y., Fang, C.-Y., Tsai, C.-H., Chang, Y., Takada, K., Hsu, T.-Y., & Chen, J.-Y. (2010). N-methyl-N'-nitro-N-nitrosoguanidine induces and cooperates with 12-O-tetradecanoylphorbol-1,3-acetate/sodium butyrate to enhance Epstein-Barr virus reactivation and genome instability in nasopharyngeal carcinoma cells. *Chemico-Biological Interactions*, 188(3), 623–634
- Hwang, J. K., Alt, F. W., & Yeap, L.-S. (2015). Related Mechanisms of Antibody Somatic Hypermutation and Class Switch Recombination. *Microbiol Spectr*, 3(1), MDNA3-0037–2014
- Kanda, T. (2018). EBV-Encoded Latent Genes. In Y. Kawaguchi, Y. Mori, & H. Kimura (Eds.), *Human Herpesviruses* (pp. 377–394)

- Kanda, T., Yajima, M., & Ikuta, K. (2019). Epstein-Barr virus strain variation and cancer. *Cancer Science*, *110*(4), 1132–1139
- Kang, M.-S., & Kieff, E. (2015). Epstein–Barr virus latent genes. *Experimental & Molecular Medicine*, *47*(1), e131
- Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166
- Kaymaz, Y., Oduor, C. I., Aydemir, O., Luftig, M. A., Otieno, J. A., Ong’echa, J. M., Bailey, J. A., & Moormann, A. M. (2020). Epstein-Barr Virus Genomes Reveal Population Structure and Type 1 Association with Endemic Burkitt Lymphoma. *Journal of Virology*, *94*(17)
- Kempkes, B., & Robertson, E. S. (2015). Epstein-Barr virus latency: Current and future perspectives. *Curr. Opin. Virol.*, *14*, 138–144
- Kenney, S. C. (2007). Reactivation and lytic replication of EBV. In A. Arvin, G. Campadelli-Fiume, E. Mocarski, P. S. Moore, B. Roizman, R. Whitley, & K. Yamanishi, *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis*, *22*(15), 745–746
- Klein, S. L. (2012). Sex influences immune responses to viruses, and efficacy of prophylaxis and therapeutic treatments for viral diseases. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(12), 1050
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874
- Kupczok, A., Neve, H., Huang, K. D., Hoepfner, M. P., Heller, K. J., Franz, C. M. A. P., & Dagan, T. (2018). Rates of Mutation and Recombination in Siphoviridae Phage

- Genome Evolution over Three Decades. *Molecular Biology and Evolution*, 35(5), 1147–1159
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359
- Lau, J. S. Y., Low, Z. M., Abbott, I., Shochet, L., Kanellis, J., Kitching, A. R., & Korman, T. M. (2017). Epstein-Barr virus encephalitis in solid organ transplantation. *The New Microbiologica*, 40(3), 212–217
- Le, J., Durand, C. M., Agha, I., & Brennan, D. C. (2017). Epstein-Barr virus and renal transplantation. *Transplantation Reviews (Orlando, Fla.)*, 31(1), 55–60
- Lee, K., Kolb, A. W., Sverchkov, Y., Cuellar, J. A., Craven, M., & Brandt, C. R. (2015). Recombination Analysis of Herpes Simplex Virus 1 Reveals a Bias toward GC Content and the Inverted Repeat Regions. *Journal of Virology*, 89(14), 7214–7223
- Lemey, P., Lott, M., Martin, D. P., & Moulton, V. (2009). Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. *BMC Bioinformatics*, 10(1), 126
- Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Lemmon, E. M. (2009). The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Systematic Biology*, 58(1), 130–145
- Li, H., Liu, S., Hu, J., Luo, X., Li, N., M.Bode, A., & Cao, Y. (2016). Epstein-Barr virus lytic reactivation regulation and its pathogenic role in carcinogenesis. *International Journal of Biological Sciences*, 12(11), 1309–1318
- Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., Richards, K. L., Dunphy, C. H., Choi, W. W. L., Srivastava, G., Lugar, P. L., Rizzieri, D. A., Lagoo, A. S., Bernal-Mizrachi, L., Mann, K. P., Flowers, C. R., Naresh, K. N., Evens, A. M., Chadburn, A., ... Dave,

- S. S. (2012). The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.*, *44*(12), 1321–1325.
- Lucchesi, W., Brady, G., Dittrich-Breiholz, O., Kracht, M., Russ, R., & Farrell, P. J. (2008). Differential Gene Regulation by Epstein-Barr Virus Type 1 and Type 2 EBNA2. *Journal of Virology*, *82*(15), 7456–7466
- Martin, D. P., Lemey, P., & Posada, D. (2011). Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*, *11*(6), 943–955
- Martin, D. P., Murrell, B., Golden, M., Khoosal, A., & Muhire, B. (2015). RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, *1*(1)
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12
- Martin, W. J. (2015). Stealth Adapted Viruses: A Bridge between Molecular Virology and Clinical Psychiatry. *Open Journal of Psychiatry*, *5*, 311–319
- McKenzie, J., & El-Guindy, A. (2015). Epstein-Barr Virus Lytic Cycle Reactivation. In C. Münz (Ed.), *Epstein Barr Virus Volume 2: One Herpes Virus: Many Diseases*, *12*, 237–261
- Migliore, L., Nicolì, V., & Stoccoro, A. (2021). Gender Specific Differences in Disease Susceptibility: The Role of Epigenetics. *Biomedicines*, *9*(6), 652
- Moormann, A. M., & Bailey, J. A. (2016). Malaria—How this parasitic infection aids and abets EBV-associated Burkitt lymphomagenesis. *Curr. Opin. Virol.*, *20*, 78–84.
- Moormann, A. M., Chelimo, K., Sumba, P. O., Tisch, D. J., Rochford, R., & Kazura, J. W. (2007). Exposure to holoendemic malaria results in suppression of Epstein-Barr virus-specific T cell immunosurveillance in Kenyan children. *J. Infect. Dis.*, *195*(6), 799–808

- Moukassa, D., Boumba, A. M., Ngatali, C. F., Ebatetou, A., Mbon, J. B. N., & Ibara, J.-R. (2018). Virus-Induced Cancers in Africa: Epidemiology and Carcinogenesis Mechanisms. *OJPathology*, *08*(01), 1–14
- Murata, T., Sugimoto, A., Inagaki, T., Yanagi, Y., Watanabe, T., Sato, Y., & Kimura, H. (2021). Molecular Basis of Epstein–Barr Virus Latency Establishment and Lytic Reactivation. *Viruses*, *13*(12), 2344
- Murata, T., & Tsurumi, T. (2014). Switching of EBV cycles between latent and lytic states. *Reviews in Medical Virology*, *24*(3), 142–153. <https://doi.org/10.1002/rmv.1780>
- Murphy, K. R., Myers, B., & Wolach, A. (2014). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests*, *15*(12), 7466
- Mwanda, O. W. (2004). Clinical characteristics of Burkitt’s lymphoma seen in Kenyan patients. *East African Medical Journal*, *8 Suppl*, S78-89
- Neves, M., Marinho-Dias, J., Ribeiro, J., & Sousa, H. (2017). Epstein-Barr virus strains and variations: Geographic or disease-specific variants? *Journal of Medical Virology*, *89*(3), 373–387
- Njie, R., Bell, A. I., Jia, H., Croom-Carter, D., Chaganti, S., Hislop, A. D., Whittle, H., & Rickinson, A. B. (2009). The effects of acute malaria on Epstein-Barr virus (EBV) load and EBV-specific T cell immunity in Gambian children. *J. Infect. Dis.*, *199*(1), 31–38.
- Oluoch, P. O., Oduor, C. I., Forconi, C. S., Ong’echa, J. M., Münz, C., Dittmer, D. P., Bailey, J. A., & Moormann, A. M. (2020). Kaposi Sarcoma-Associated Herpesvirus Infection and Endemic Burkitt Lymphoma. *The Journal of Infectious Diseases*, *222*(1), 111–120
- Palma, I., Sánchez, A. E., Jiménez-Hernández, E., Alvarez-Rodríguez, F., Nava-Frias, M., Valencia-Mayoral, P., Salinas-Lara, C., Velazquez-Guadarrama, N., Portilla-Aguilar, J., Pena, R. Y., Ramos-Salazar, P., Contreras, A., Alfaro, A., Espinosa, A. M., Nájera, N., Gutierrez, G., Mejia-Arangure, J. M., & Arellano-Galindo, J. (2013). Detection of

- Epstein-Barr Virus and Genotyping Based on EBNA2 Protein in Mexican Patients With Hodgkin Lymphoma: A Comparative Study in Children and Adults. *Clinical Lymphoma Myeloma and Leukemia*, 13(3), 266–272
- Palser, A. L., Grayson, N. E., White, R. E., Corton, C., Correia, S., Ba Abdullah, M. M., Watson, S. J., Cotten, M., Arrand, J. R., Murray, P. G., Allday, M. J., Rickinson, A. B., Young, L. S., Farrell, P. J., & Kellam, P. (2015). Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. *J. Virol.*, 89(10), 5222–5237.
- Panea, R. I., Love, C. L., Shingleton, J. R., Reddy, A., Bailey, J. A., Moormann, A. M., Otieno, J. A., Ong'echa, J. M., Oduor, C. I., Schroeder, K. M. S., Masalu, N., Chao, N. J., Agajanian, M., Major, M. B., Fedoriw, Y., Richards, K. L., Rymkiewicz, G., Miles, R. R., Alobeid, B., ... Dave, S. S. (2019). The whole-genome landscape of Burkitt lymphoma subtypes. *Blood*, 134(19), 1598–1607
- Perera, R. A. P. M., Samaranayake, L. P., & Tsang, C. S. P. (2010). Shedding dynamics of Epstein-Barr virus: A type 1 carcinogen. *Archives of Oral Biology*, 55(9), 639–647
- Pérez-Losada, M., Arenas, M., Galán, J. C., Palero, F., & González-Candelas, F. (2015). Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, 30, 296–307
- Piriou, E., Asito, A. S., Sumba, P. O., Fiore, N., Middeldorp, J. M., Moormann, A. M., Ploutz-Snyder, R., & Rochford, R. (2012). Early age at time of primary Epstein-Barr virus infection results in poorly controlled viral infection in infants from Western Kenya: Clues to the etiology of endemic Burkitt lymphoma. *J. Infect. Dis.*, 205(6), 906–913.
- Pradhan, A., & Olsson, P.-E. (2020). Sex differences in severity and mortality from COVID-19: Are males more vulnerable? *Biology of Sex Differences*, 11, 53
- Prata, T. T. M., Bonin, C. M., Ferreira, A. M. T., Padovani, C. T. J., Fernandes, C. E. dos S., Machado, A. P., & Tozetti, I. A. (2015). Local immunosuppression induced by high

- viral load of human papillomavirus: Characterization of cellular phenotypes producing interleukin-10 in cervical neoplastic lesions. *Immunology*, 146(1), 113–121
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3), e9490
- Rainey, J. J., Mwanda, W. O., Wairiumu, P., Moormann, A. M., Wilson, M. L., & Rochford, R. (2007). Spatial distribution of Burkitt's lymphoma in Kenya and association with malaria risk. *Trop. Med. Int. Health*, 12(8), 936–943
- Redmond, L. S., Ogwang, M. D., Kerchan, P., Reynolds, S. J., Tenge, C. N., Were, P. A., Kuremu, R. T., Masalu, N., Kawira, E., Otim, I., Legason, I. D., Dhudha, H., Ayers, L. W., Bhatia, K., Goedert, J. J., & Mbulaiteye, S. M. (2020). Endemic Burkitt lymphoma: A complication of asymptomatic malaria in sub-Saharan Africa based on published literature and primary data from Uganda, Tanzania, and Kenya. *Malaria Journal*, 19(1), 239
- Reynaldi, A., Schlub, T. E., Chelimo, K., Sumba, P. O., Piriou, E., Ogolla, S., Moormann, A. M., Rochford, R., & Davenport, M. P. (2016). Impact of Plasmodium falciparum Coinfection on Longitudinal Epstein-Barr Virus Kinetics in Kenyan Children. *J. Infect. Dis.*, 213(6), 985–991
- Reynaldi, A., Schlub, T. E., Piriou, E., Ogolla, S., Sumba, O. P., Moormann, A. M., Rochford, R., & Davenport, M. P. (2016). Modeling of EBV Infection and Antibody Responses in Kenyan Infants With Different Levels of Malaria Exposure Shows Maternal Antibody Decay is a Major Determinant of Early EBV Infection. *J. Infect. Dis.*, 214(9), 1390–1398.
- Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: A review and a practical guide. *Molecular Ecology*, 25(9), 1911–1924

- Robaina, T., Valladares, C., Tavares, D., Napolitano, W., Silva, L., Dias, E., & Leite, J. (2008). Polymerase chain reaction genotyping of Epstein-Barr virus in scraping samples of the tongue lateral border in HIV-1 seropositive patients. *Memórias Do Instituto Oswaldo Cruz*, *103*(4), 326–331
- Rochford, R. (2009). Epidemiology of EBV Infection. In B. Damania & J. M. Pipas (Eds.), *DNA Tumor Viruses* (pp. 205–215)
- Rosemarie, Q., & Sugden, B. (2020). Epstein–Barr Virus: How Its Lytic Phase Contributes to Oncogenesis. *Microorganisms*, *8*(11), 1824
- Sample, J. T., Marendy, E. M., Hughes, D. J., & Sample, C. E. (2009). The Epstein–Barr Virus Genome. In B. Damania & J. M. Pipas (Eds.), *DNA Tumor Viruses* (pp. 241–258)
- Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433–4448
- Santpere, G., Darre, F., Blanco, S., Alcamí, A., Villoslada, P., Mar Albà, M., & Navarro, A. (2014a). Genome-Wide Analysis of Wild-Type Epstein–Barr Virus Genomes Derived from Healthy Individuals of the 1000 Genomes Project. *Genome Biology and Evolution*, *6*(4), 846–860
- Santpere, G., Darre, F., Blanco, S., Alcamí, A., Villoslada, P., Mar Albà, M., & Navarro, A. (2014b). Genome-Wide Analysis of Wild-Type Epstein–Barr Virus Genomes Derived from Healthy Individuals of the 1000 Genomes Project. *Genome Biology and Evolution*, *6*(4), 846–860
- Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, *27*(6), 863–864
- 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. *Journal of Virology*, 89(15), 7673–7695
- Simonsen, M., Mailund, T., & Pedersen, C. N. S. (2011). Inference of Large Phylogenies Using Neighbour-Joining. In A. Fred, J. Filipe, & H. Gamboa (Eds.), *Biomedical Engineering Systems and Technologies* (pp. 334–344)
- Smatti, M. K., Yassine, H. M., AbuOdeh, R., AlMarawani, A., Taleb, S. A., Althani, A. A., & Nasrallah, G. K. (2017). Prevalence and molecular profiling of Epstein Barr virus (EBV) among healthy blood donors from different nationalities in Qatar. *PLoS One*, 12(12), e0189033.
- Snider, C. J., Cole, S. R., Chelimo, K., Sumba, P. O., Macdonald, P. D. M., John, C. C., Meshnick, S. R., & Moormann, A. M. (2012). Recurrent Plasmodium falciparum malaria infections in Kenyan children diminish T-cell immunity to Epstein Barr virus lytic but not latent antigens. *PLoS One*, 7(3), e31753.
- Stefan, C., Bray, F., Ferlay, J., Liu, B., & Maxwell Parkin, D. (2017). Cancer of childhood in sub-Saharan Africa. *Ecancermedicalscience*, 11, 755.
- Swaminathan, S., & Kenney, S. (2008). The Epstein–Barr Virus Lytic Life Cycle. In *DNA Tumor Viruses* (pp. 285–315). https://doi.org/10.1007/978-0-387-68945-6_13
- Talavera, G., & Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology*, 56(4), 564–577
- Taneja, V. (2018). Sex Hormones Determine Immune Response. *Frontiers in Immunology*, 9, 1931
- Telford, M., Hughes, D. A., Juan, D., Stoneking, M., Navarro, A., & Santpere, G. (2020). Expanding the Geographic Characterisation of Epstein–Barr Virus Variation through Gene-Based Approaches. *Microorganisms*, 8(11), 1686

- Thorley-Lawson, D. A., Hawkins, J. B., Tracy, S. I., & Shapiro, M. (2013). The pathogenesis of Epstein-Barr virus persistent infection. *Curr. Opin. Virol.*, *3*(3), 227–232.
- Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192
- Torgbor, C., Awuah, P., Deitsch, K., Kalantari, P., Duca, K. A., & Thorley-Lawson, D. A. (2014). A multifactorial role for *P. falciparum* malaria in endemic Burkitt's lymphoma pathogenesis. *PLoS Pathog.*, *10*(5), e1004170
- Trivedi, U. H., Cézard, T., Bridgett, S., Montazam, A., Nichols, J., Blaxter, M., & Gharbi, K. (2014). Quality control of next-generation sequencing data without a reference. *Frontiers in Genetics*, *5*, 111
- Trottier, H., Alfieri, C., Robitaille, N., Duval, M., Buteau, C., Tucci, M., & Lacroix, J. (2010). Transfusion-Related Epstein-Barr Virus (EBV) Infection Among Stem Cell Transplant Recipients: A Retrospective Cohort Study In Children. *Blood*, *116*(21), 3340
- Tsai, I. J., Otto, T. D., & Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, *11*(4), R41
- Tzellos, S., & Farrell, P. J. (2012). Epstein-Barr virus sequence variation-biology and disease. *Pathogens (Basel, Switzerland)*, *1*(2), 156–174
- Ueda, K. (2018). KSHV Genome Replication and Maintenance in Latency. In Y. Kawaguchi, Y. Mori, & H. Kimura (Eds.), *Human Herpesviruses* (pp. 299–320). Springer
- van Lunzen, J., & Altfeld, M. (2014). Sex Differences in Infectious Diseases—Common but Neglected. *The Journal of Infectious Diseases*, *209*(suppl_3), S79–S80
- Wang, S., Xiong, H., Yan, S., Wu, N., & Lu, Z. (2016). Identification and Characterization of Epstein-Barr Virus Genomes in Lung Carcinoma Biopsy Samples by Next-Generation Sequencing Technology. *Scientific Reports*, *6*(1), 26156

- Wang, X., Wang, Y., Wu, G., Chao, Y., Sun, Z., & Luo, B. (2012). Sequence analysis of Epstein-Barr virus EBNA-2 gene coding amino acid 148-487 in nasopharyngeal and gastric carcinomas. *Virology*, 9, 49
- Weller, S. K., & Coen, D. M. (2012). Herpes Simplex Viruses: Mechanisms of DNA Replication. *Cold Spring Harbor Perspectives in Biology*, 4(9), a013011
- Westmoreland, K. D., Montgomery, N. D., Stanley, C. C., El-Mallawany, N. K., Wasswa, P., van der Gronde, T., Mtete, I., Butia, M., Itimu, S., Chasela, M., Mtunda, M., Kampani, C., Liomba, N. G., Tomoka, T., Dhungel, B. M., Sanders, M. K., Krysiak, R., Kazembe, P., Dittmer, D. P., ... Gopal, S. (2017). Plasma Epstein-Barr virus DNA for pediatric Burkitt lymphoma diagnosis, prognosis and response assessment in Malawi. *International Journal of Cancer*, 140(11), 2509–2516
- Wu, C.-C., Liu, M.-T., Chang, Y.-T., Fang, C.-Y., Chou, S.-P., Liao, H.-W., Kuo, K.-L., Hsu, S.-L., Chen, Y.-R., Wang, P.-W., Chen, Y.-L., Chuang, H.-Y., Lee, C.-H., Chen, M., Wayne Chang, W.-S., & Chen, J.-Y. (2010). Epstein–Barr Virus DNase (BGLF5) induces genomic instability in human epithelial cells. *Nucleic Acids Research*, 38(6), 1932–1949
- Yatani, K. (2016). Effect Sizes and Power Analysis in HCI. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI* (pp. 87–110)
- Yoshida, R., & Nei, M. (2016). Efficiencies of the NJp, Maximum Likelihood, and Bayesian Methods of Phylogenetic Construction for Compositional and Noncompositional Genes. *Molecular Biology and Evolution*, 33(6), 1618–1624
- Zanella, L., Riquelme, I., Buchegger, K., Abanto, M., Ili, C., & Brebi, P. (2019). A reliable Epstein-Barr Virus classification based on phylogenomic and population analyses. *Scientific Reports*, 9(1), 9829

- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829
- Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W. X., & Wang, G. T. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular Ecology Resources*, *20*(1), 348–355

APPENDICES

Appendix I: KEMRI-SERU Approval



KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840 - 00200 NAIROBI, Kenya
Tel: (254) (020) 2722541, 2713349, 0722-205901, 0733-400003; Fax: (254) (020) 2720030
E-mail: director@kemri.org; info@kemri.org Website: www.kemri.org

KEMRI/RES/7/3/1

MAY 23, 2008

FROM: SECRETARY, KEMRI/National Ethical Review Committee

THRO': Dr. J Vulule,
CENTRE DIRECTOR, CGHR,
KISUMU

TO: Dr. Ann M Moormann (Case Western Reserve
University) (Principal Investigator)

RE: SSC No.1381 – The effect of Plasmodium falciparum malaria on T cell
immunity and Endemic Burkitts lymphoma

Dear Madam,

This is to inform you that the abovementioned protocol has undergone expedited review.

We acknowledge receipt of the following documents:

1. The study proposal version 10 March 2008
2. The Informed Consent Document (ICD) form 1 for healthy Kenyan children in English
3. Form 1.1: Healthy child enrollment questionnaire and venous blood sample collection
4. The ICD form 2: for children diagnosed with Burkitts lymphoma
5. Form 2.1 BL patient enrollment questionnaire and venous blood sample collection
6. The ICD form 3: for children who have been diagnosed with BL but are in remission for a repeat venous blood sample
7. Form 2.2 BL patient discharge summary and venous blood sample collection
8. Form 2.3: BL patient follow-up and venous blood sample collection
9. The ICD form 4: for healthy Kenyan adults to optimize laboratory assays
10. Form 3.1 Healthy adult enrollment questionnaire and venous blood sample collection
11. The ICD form 5: for healthy US adults never exposed to malaria
12. Form 4.1 Healthy US adult enrollment questionnaire and venous blood collection

Thank you for your informative study proposal that aims to investigate the mechanisms of malaria-induced dysregulation of EBV-specific T cell immunity and its relationship to eBL. This will be a prospective study of healthy Kenyan children with divergent malaria exposure histories and by examining children with eBL compared to healthy Kenyan and US adults who have robust immune responses to EBV.

In Search of Better Health

We note that you have requested a waiver of assent for subjects 7 to 14 years of age for this study citing the Kenyan culture where parents or guardians decide whether a child should participate or not. This is generally true however if the children express any interest in knowing the procedures of the study, efforts should be made to explain to the children.

Due consideration has been given to ethical issues and the study is granted approval from today the 23rd of May 2008 to 22nd May 2009.

Please note that any changes to the research study must be reported to the Scientific Steering Committee and to the Ethical Review Committee prior to implementation. This includes changes to research design, equipment, personnel, funding or procedures that could introduce new or more than minimum risk to research participants.

Respectfully,

R.C. Kithinji

R. C. Kithinji,
For: Secretary,
KEMRI/NATIONAL ETHICAL REVIEW COMMITTEE



21 OCT 2014

F.O. 2014/10/21

KENYA MEDICAL RESEARCH INSTITUTE

P.O. Box 54840-00200, NAIROBI, Kenya
Tel (254) (020) 2722541, 2713349, 0722-205901, 0733-400003; Fax (254) (020) 2720030
E-mail: director@kemri.org info@kemri.org Website:www.kemri.org

KEMRI/RES/7/3/1

TO: **DR. JOHN ONG'ECHA,**
PRINCIPAL INVESTIGATOR

THROUGH: **DR. STEPHEN MUNGA**
THE DIRECTOR, CGHR
KISUMU

Dear Sir,

RE: **SSC PROTOCOL NO. 2844 (INITIAL SUBMISSION): IMPACT OF MALARIA ON SHAPING IMMUNITY TO EBV AND ENDEMIC BURKITT LYMPHOMA**

This is to inform you that during the 231st meeting of the KEMRI/ERC held on September 16, 2014, the above study was reviewed.

The Committee noted that the above referenced study aims to determine how malaria infection influences the differentiation and survival of EBV-specific T cell responses and loss of t-cell control over EBV.

Due consideration has been given to ethical issues, this study is therefore granted approval for implementation effective this day **October 13, 2014**. Please note that authorization to conduct this study will automatically expire on **October 12, 2015**.

If you plan to continue with data collection or analysis beyond this date please submit an application for continuing approval to the ERC secretariat by **August 31, 2015**. You are also required to submit any proposed changes to this protocol to the SSC and ERC prior to initiation and advise the ERC when the study is completed or discontinued.

You may embark on the study.

Yours faithfully,

EAB
PROF. ELIZABETH BUKUSI,
ACTING SECRETARY,
KEMRI/ETHICS REVIEW COMMITTEE

October 13, 2014



Appendix II: eBL Participant Consent Form

KEMRI/UMMS

CONSENT FOR INVESTIGATIONAL STUDIES

Project Title: Impact of Malaria on Shaping Immunity to EBV and Endemic Burkitt Lymphoma

Principal Investigators: Ann Moormann, Ph.D., MPH and John M. Ong'echa, Ph.D.

Consent 4: BL biopsy for research request

POSSIBLE RISKS OF STUDY PARTICIPATION

There are no added risks to your child if you consent for a biopsy to be used for research.

However, normal side effects of the biopsy procedure may include pain (stinging and burning), bleeding and infection. The biopsy areas will be covered with a bandage and may be tender for several days. It should heal within one or two weeks.

ALTERNATIVES TO TAKING PART IN THIS RESEARCH STUDY

You do not have to sign this consent form. You do not have to participate in this study.

Your child will still get regular care at JOOTRH, even if you decide not to participate in this study.

POSSIBLE BENEFITS OF STUDY PARTICIPATION

Your child will not benefit directly from results of this study. The diagnosis and treatment will be decided by the doctors at JOOTRH caring for your child. This study will in no way interfere with the care of your child. It is possible that what we learn from this study will help doctors and scientists learn how to prevent BL in other children and to improve our ability to cure this cancer.

SUMMARY OF YOUR RIGHTS AS A PARTICIPANT IN A RESEARCH STUDY

Your participation in this research study is voluntary. Refusing to participate will not alter your usual health care or involve any penalty or loss of benefits to which you are otherwise entitled. If you decide to join the study, you may withdraw at any time and for any reason without penalty or loss of benefits. If information generated from this study is published or presented, your identity or your child's identity will not be revealed. In the event new information becomes available that may affect the risks or benefits associated with this study or your willingness to participate in it, you will be notified so that you can decide whether or not to continue participating.

CONFIDENTIALITY OF RECORDS

Your identity, your child's identity and his or her records will remain confidential. The biopsy specimen will be identified by a code number. Any information about your child to be used in this study will be marked by this code number. If you agree to participate in this study we will ask for some basic information to be recorded about your child. This information will include your child's age, sex, site of tumor(s). During the course of your child's illness, we would also like permission to know if your child is responding well to treatment and which treatment regimen was used. This will be determined by Dr. Otieno or another doctor at JOOTRH caring for your child. This information can be found in your child's medical records. Signing this consent form gives the study staff permission to look at your child's chart and record this information for the purpose of the study. This information will be helpful to learn more about BL.

The study code number assigned to your child will be linked to your child's hospital record number. The log book that will match the code number with your child's hospital number will be kept in a locked file cabinet at Center for Global Health Research, KEMRI and on a password protected file. Links between the study code and information that could identify your child will be

<p>KEMRI/UMMS CONSENT FOR INVESTIGATIONAL STUDIES Project Title: Impact of Malaria on Shaping Immunity to EBV and Endemic Burkitt Lymphoma Principal Investigators: Ann Moormann, Ph.D., MPH and John M. Ong'echa, Ph.D. Consent 4: BL biopsy for research request</p>

kept for five years unless you provide consent that they may be kept indefinitely. Research data is kept indefinitely but will only be labeled by the study code number.

We will do everything we can to protect your privacy. We cannot guarantee absolute privacy, however. Your personal information may be disclosed if required by law. If your confidential information is released, you will be told. Some people may need to review the study records. They may include the UMMS or KEMRI Ethical Review Boards and the study staff from UMMS-KEMRI. The results of this study will be published. Any publication will not use your name, your child's name or identify you or your child personally.

SPECIMEN STORAGE

Specimens collected from your child will be tested by laboratories associated with the University of Massachusetts, USA. It is anticipated that the entire specimen will be used for the studies proposed by this protocol. If you consent, any remaining specimens will be kept indefinitely. Any future testing of your child's specimens will not be done unless written approval has been given from the ethical review committees in Kenya and UMMS. If you circle "yes" on this consent form below, you are allowing your child's specimen to be used for future research, if it is related to cancer, without seeking re-approval from you individually.

If you change your mind in the future, you may contact Dr. John Ong'echa (0733-447920) or the KEMRI/National Ethical Review Committee (ERC), PO Box 54840, Nairobi 00200 at (020) 272-2541 or the Director of KEMRI, PO Box 54840, Nairobi at (020)272-2541.

Consent for use of child's samples for future studies yes no
(Please circle parent's response)

REASONS FOR REMOVING YOUR CHILD FROM THE STUDY WITHOUT YOUR CONSENT

If the doctor cancels the request for a biopsy then your child will no longer be eligible to be in this study. This may occur if your child's doctor receives other information that means your child does not have BL.

COSTS TO YOU

There is no cost to you for being in this study. There are no added tests or procedures being done only for research purposes. You will be asked to pay for medical procedures requested by the doctor in order to find out why your child is ill, the same as for children not in the study. All other medical costs outside of this study will be paid by you or your health insurance carrier (if you have insurance).

IF YOU THINK YOUR CHILD HAS A RESEARCH-RELATED INJURY

There are no procedures being done for study purposes only. All procedures are for the care and diagnosis of your child. However, you may call Dr. Otieno with any question you may have regarding the study or your child's participation in this study.

**KEMRI/UMMS
CONSENT FOR INVESTIGATIONAL STUDIES**

Project Title: Impact of Malaria on Shaping Immunity to EBV and Endemic Burkitt Lymphoma

Principal Investigators: Ann Moormann, Ph.D., MPH and John M. Ong'echa, Ph.D.

Consent 4: BL biopsy for research request

No funds are available to provide compensation for non-physical injury such as lost work or pain and suffering. You and/or your health insurance carrier will continue to be responsible for costs for your child's medical care or for medical expenses determined not directly related to study procedures. You will not be giving up any of your legal rights by signing this consent form.

IF YOU SIGN THIS CONSENT FORM, YOU DO NOT NEED TO DO ANYTHING ELSE.

Once you sign this consent form, the study coordinator will show the nurse-in-charge and the doctor taking care of your child that your child is enrolled in this study. This will allow the nurse to tell the study coordinator when the biopsy procedure is scheduled. Then the surgeon will be notified that during the procedure, a small piece of the tumor can be put in a tube for research purposes. Project staff will be there during the procedure to collect the specimen and will bring it to the UMMS-KEMRI lab.

PAYMENT

You will not be paid to participate in this research study. However, in appreciation for your participation the project can pay for transport home when your child is discharged from hospital. We will also assist with transport reimbursement for out-patient clinic visits to make sure your child is recovering well. If your child is experiencing symptoms that indicate that the cancer may be coming back, we will transport you and your child to JOOTRH for evaluation. Please contact Mrs. Pamela Omolo for assistance with at 0722890318.

Contact information

One of our team members named _____ has described to you what is going to be done, the risks, hazards, and benefits involved. Further information with respect to illness or injury resulting from a research procedure as well as a research subjects' rights is available from KEMRI/National Ethical Review Committee (ERC), PO Box 54840, Nairobi 00200 at (020) 272-2541 or the director of KEMRI, PO Box 54840, Nairobi at (020)272-2541, or the Coordinator for the Committee for the Protection of Human Subjects in Research in the United States at (508) 856-4261 or write to Committee for the Protection of Human Subjects in Research, University of Massachusetts Medical Center, 55 Lake Avenue North, Worcester, MA 01655 U.S.A.

For study coordinator:

Study assigned unique identification number: BL - _____
(number given at time of enrollment on Form 2.1)

KEMRI/UMMS

CONSENT FOR INVESTIGATIONAL STUDIES

Project Title: Impact of Malaria on Shaping Immunity to EBV and Endemic Burkitt Lymphoma

Principal Investigators: Ann Moormann, Ph.D., MPH and John M. Ong'echa, Ph.D.

Consent 4: BL biopsy for research request

SIGNATURE PAGE

Study staff conducting consent discussion (print name)

Study staff signature

Date (dd/mm/yyyy)

PARTICIPANT'S STATEMENT

The study described above has been explained to me. I agree to volunteer to participate in the study. I have had a chance to ask questions. I have been told that if I have future questions about the research I can ask one of the contacts listed above or the study staff named above. I give permission to the researchers to use my child's medical records as described in this consent form and to collect a biopsy specimen for research purposes. I will receive a copy of this consent form. If you have read this consent form (or had it explained to you), understand it and agree for your child to take part in this study, please sign your name below.

Child's name (print name)

Study Code number: UMMS-KEMRI BLB-00_____

(Study code assigned in order of enrollment, first number is UMMS-KEMRI BLB-0001, etc).

Parent or legal guardian's name (print name)

Parent or legal guardian's signature

Or mark (right thumb unless otherwise indicated)

Date (dd/mm/yyyy)

Cell phone contact of parent or nearest neighbor/relative

Witness's name, if necessary (print – this name should be a different UCI staff member than the person who conducted the consent discussion)

Witness's signature

Date (dd/mm/yyyy)

Original to: Participant's study file.

Copy to: Participant. If participant declines a copy of this CF, then check this box and initial: []

Appendix III: Maseno University School of Graduate Studies Approval



MASENO UNIVERSITY
SCHOOL OF GRADUATE STUDIES

Office of the Dean

Our Ref: MSC/SC/00053/018

Private Bag, MASENO, KENYA
Tel:(057)351 22/351008/351011
FAX: 254-057-351153/351221
Email: sgs@maseno.ac.ke

Date: 16th February, 2021

TO WHOM IT MAY CONCERN

RE: PROPOSAL APPROVAL FOR AGWATI EDDY —MSC/SC/00053/018

The above named is registered in the Master of Science in Cell and Molecular Biology degree programme in the School of Physical and Biological Sciences, Maseno University. This is to confirm that his research proposal titled “Impact of Genome-Wide Recombination Events on Epstein Barr Virus Diversity, Genome Population Structure and Endemic Burkitt Lymphoma Pathogenesis among Children from Western Kenya” has been approved for conduct of research subject to obtaining all other permissions/clearances that may be required beforehand.

Prof. J.O. Agure
DEAN, SCHOOL OF GRADUATE STUDIES



Appendix IV: Statistical Power Test

Characteristic		Number of Participants N=86	Effect size	Significant Level	Statistical Power (%)
eBL Status	eBL	54	0.8	0.05	94
	Healthy	32			
Gender	Males	58	0.8	0.05	93
	Females	28			
Age Group	0-4	39	0.8	0.05	91.5
	5-9	34			
	10-14	13			

Abbreviation: eBL, endemic Burkitt Lymphoma, N, Number of participants. A large effect size of 0.8 was used to determine if the number of participants in each group proportion had enough statistical power to detect group differences at a 95% confidence level. From the power calculations, all the proportions had enough statistical power to answer the study objectives.

Appendix V: Genotyping Primers and Probe Sets

Primer Sequences Used to determine the genomic subtype of EBV (Type 1 and Type 2)

Type 1 EBNA 2 Forward Primer	TTGTGACAGAGAGGTGGACAAAA
------------------------------	-------------------------

Type EBNA 2 Forward Primer	TGGAAGAGTATGTTCCCTAGG
----------------------------	-----------------------

Type 1 and Type common Reverse Primer	AGGGAATGCCTGGACACAGGA
---------------------------------------	-----------------------

Appendix VI: EBV-Specific Genome Wide Amplification Primer Sets

Pool 1	Primer sequence 1	Pool 2	Primer sequence 2
Fwd1-1	TTCTGGTGATGCTTGTGCTC	Fwd2-1	CTGTTTATGAGACGCCAGC A
Rev1-1	TGCTGGCGTCTCATAAACAG	Rev2-1	TTTTCGCTGCTTGTCTTTT
Fwd1-2	AAAAGGACAAGCAGCGAAA A	Fwd2-2	TTATGGTTCAGTGCCTCGA G
Rev1-2	GTGCAGGAGGCTGTTTCTTC	Rev2-2	GAACTGAGGAGGGCATGA AG
Fwd1-3	ATGCCTACATTCTATCTTGC GTTAC	Fwd2-3	AGGGATGCCTGGACACAA GA
Rev1-3	TTACTGGATGGAGGGGCGA GGTCTT	Rev2-3	AACATGGACTGGGAGTGG AG
Fwd1-4	CTAGAGGTCCGCGAGATTTG	Fwd2-4	GCAGGCAGTACGAGATGT CA
Rev1-4	AGAAGGCAAGCGAAAATTG A	Rev2-4	TCCCTTCACATCCCAGAGA C
Fwd1-5	CGACATTGACAGCCTTCTCA	Fwd2-5	TGCTCCTGATGTTTCTGAG GTGGA
Rev1-5	AAACACGAATGCCAAGAAC C	Rev2-5	AGGTAACTTCTTTGAGCCT CCCGA
Fwd1-6	TTGCTCCATCTGTCAGCAAC	Fwd2-6	GGTGACCACTGAGGGAGT GT
Rev1-6	CACAAGCCTCCTCTCAGGAC	Rev2-6	ATTCAGGACTACCTGCGC GACTT
Fwd1-7	GGACATCTCTGGCTCGAAAG	Fwd2-7	TCAGGAGGTCGTCAAAATC C
Rev1-7	AGGAGGAGAACCCGAGGAT A	Rev2-7	TTTCACATCCGACTCATTC CCTGC
Rev1-7-t2	AGGAGGAGAACCCGAGGAT C	Fwd2-8	CCAGTCGCCGTTACTCATC T
Fwd1-8	TCCAGGCTGTTGGAGAACAC TTCA	Rev2-8	ACCTTTCATCCGAACTCCT CAGGT
Rev1-8	ATCACAGTCACCCCCAGAA G	Fwd2-9	GCCTCTATGTCGCTCTGAC C
Fwd1-9	CAGACGGTGGCGTATATGA G	Rev2-9	CGGAGGCGTGGTTAAATA AA
Rev1-9	CAAAGAGCCCCGTAAAGAT G	Fwd2-10	CTCGCGTGTTAGGAAGGAA G
Fwd1-10	GCGAGCCATAAAGCAGTTTC	Rev2-10	AGGCAAAGCTGGTCAAAG AA
Rev1-10	TCTCCCGAACTAGCAGCATT	Fwd2-11- t2	ATACATAGGAGCCTCACGA A
Fwd1-11	GCCTTCTTTGACCAGCTTTG	Fwd2-11	GGTGAAACGCGAGAAGAA AG
Rev1-11	GACGGGTTCTACTGGCATGT	Rev2-11	TTAGCAGTTCCTCCGCAC T

Rev1-11-t2	GACGGGTTCTACTGGCATGG	Fwd2-12	CCCACCACGTCTTCAACTT T
Fwd1-12	AGTGCGGAGGAACTGCTAA A	Rev2-12	CCATACCAGGTGCCTTTTG T
Rev1-12	TGCAGAGGATGAGACCAGT G	Fwd2-12	ACTCCCGGCTGTAAATTCC T
Fwd1-13	TCCAAGGTGACCCCTGTTAG	Rev2-12	TGGCCAGAAATACACCAA CA
Rev1-13	TGATGCAGAGTCGCCTAATG	Fwd2-13	ACAGACCATCTACGCCAAC C
Fwd1-14	CCCATGTTGTCACGTCACTC	Rev2-13	CCACCACAAGAAGGTGTCC T
Rev1-14	CACCGTGTGGAGACCTTTT	Fwd2-14	GATGTTGCTGGGGCTAATG T
Fwd1-15	TACGGGGCACTTAACCTGAC	Rev2-14	AGAGAGGGAGTTTCGCTTC C
Rev1-15	TGACGGAGCTGTATCACGA G	Fwd2-15	CGTTGGAAGTTGTTGGGAC T
Fwd1-16	GGCACCATAGCATGTCACAC	Rev2-15	CATTTTACCAGGGACGAGG A
Rev1-16	AGTCCCAACAACCTCCAACG	Fwd2-16	GGTCTCAACGTGTCCTGGT T
Fwd1-17	CCCGTTCACCAAACAGTCT	Rev2-16	GTGAAGGTATGTGCCGGTC T
Rev1-17	AACCAGGACACGTTGAGAC C	Fwd2-17	CCTGAGAACGCTCCAGGTA G
Fwd1-18	ACCTCCCATAGCAACACCAG	Rev2-17	CCTGGTGAGAAGTTGGTGG T
Rev1-18	CCCGTGCGATGAGTTTATTT	Fwd2-18	TTTGGGATGCATCACTTTG A
Fwd1-19	CCAGACATACCCCAAACCA C	Rev2-18	CCTCAAAGGTGTGGTTCGTT T
Rev1-19	CTCCAGAGGGCAGACGTTA G	Fwd2-19	TCGTGGCTCGTACAGACGA TTGTT
Fwd1-20	GCCCGTTGGGTTACATTAAG GTGT	Rev2-19	ACCTGGTACATTGTGCCCA TCAGA
Rev1-20	CATGCAGTGGTGTCAGACA GGAAA	Fwd2-20	CCCACACCTTCACTCCTTG T
Fwd1-21	CTTTGGGTTCCATTGTGTGC CCTT	Rev2-20	CAGAGCCAGGCACATCTAC A
Rev1-21	TTTGCGCCTTCTCCTGGTTTA TGC	Fwd2-21	TGGAAGAAGGCGTAGAGC AT
Fwd1-22	ACGCCATACCCAAGTGAGTC	Rev2-21	GCAAGGCTGACTCACCTGT TTGA
Rev1-23	TCAAGAACCTGACGGAGCTT	Fwd2-22	AGGTTGCACACCACATCAA A
Fwd1-24	ACGCCGAGTCATCTCTCATT TGGA	Rev2-22	GACTCGCTCACCCAAGAAA G

Rev1-24	CGTGACTACCCCCACGTACT	Fwd2-23	CACGGGGTTTATGTTTCTG G
Fwd1-25	GTGCAGAGCCTTGACATTGA	Rev2-23	CCCCCTCCACTTTTTCCA
Rev1-25	TGAACACCACCACGATGACT	EBNA2_ Fw_Extra	TGGGAATGGTGTAACTTT C
EBNA2_Fw_ Extra	TGGGAATGGTGTAACTTTC	EBNA2_ Rev_Extra	ATGTGTTGTGTGTGGTTTT G
EBNA2_Rev_ Extra	ATGTGTTGTGTGTGGTTTTG		

Appendix VII: Genomes' Pre-processing Information and Sequencing Statistics

Sample ID	Amplification	Capture Method	Illumina Paired-end Read Length	Total Sequence reads	Assembled Genome Size (bases)	Average Depth of Coverage over Assembly
HC-0001	mlrPCR-sWGA	EBV specific MyBait oligos	300	1989031	92740	6434.2
HC-0002	mlrPCR-sWGA	EBV specific MyBait oligos	300	2573344	135457	5699.2
HC-0003	mlrPCR-sWGA	EBV specific MyBait oligos	300	3256865	101378	9637.8
HC-0004	mlrPCR-sWGA	EBV specific MyBait oligos	300	6191128	130528	14229.4
HC-0005	mlrPCR-sWGA	EBV specific MyBait oligos	300	2230574	130100	5143.5
HC-0006	mlrPCR-sWGA	EBV specific MyBait oligos	300	14005054	113876	36895.5
HC-0007	mlrPCR-sWGA	EBV specific MyBait oligos	300	13554411	125967	32280.9
HC-0008	mlrPCR-sWGA	EBV specific MyBait oligos	300	4038485	112717	10748.6
HC-0009	mlrPCR-sWGA	EBV specific MyBait oligos	300	5264476	132675	11903.8
HC-0010	mlrPCR-sWGA	EBV specific MyBait oligos	300	5080095	91925	16579
HC-0011	mlrPCR-sWGA	EBV specific MyBait oligos	300	2361855	114610	6182.3
HC-0012	mlrPCR-sWGA	EBV specific MyBait oligos	300	4609922	114898	12036.6
HC-0013	mlrPCR-sWGA	EBV specific MyBait oligos	300	399047	104101	1150
HC-0014	mlrPCR-sWGA	EBV specific MyBait oligos	300	1373086	123167	3344.4
HC-0015	mlrPCR-sWGA	EBV specific MyBait oligos	300	3517550	128859	8189.3
HC-0016	mlrPCR-sWGA	EBV specific MyBait oligos	300	5312268	131404	12128.1
HC-0017	mlrPCR-sWGA	EBV specific MyBait oligos	300	1194255	89724	3993.1
HC-0018	mlrPCR-sWGA	EBV specific MyBait oligos	300	3443411	85526	12078.5
HC-0019	mlrPCR-sWGA	EBV specific MyBait oligos	300	1401820	137320	3062.5
HC-0020	mlrPCR-sWGA	EBV specific MyBait oligos	300	3297486	97564	10139.5
HC-0021	mlrPCR-sWGA	EBV specific MyBait oligos	300	1691945	110989	4573.3

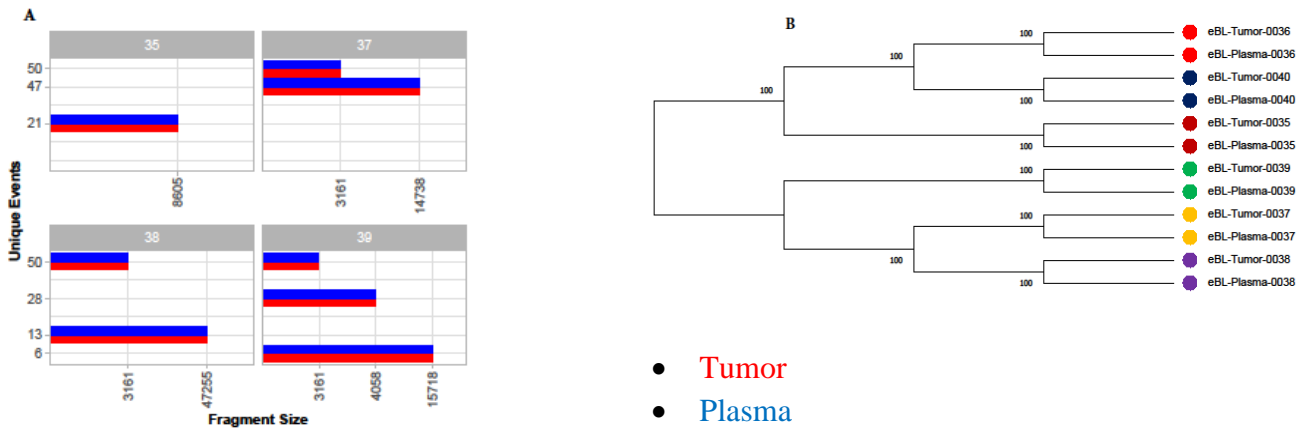
HC-0022	mlrPCR-sWGA	EBV specific MyBait oligos	300	986623	120580	2454.7
HC-0023	mlrPCR-sWGA	EBV specific MyBait oligos	300	149295	132533	337.9
HC-0024	mlrPCR-sWGA	EBV specific MyBait oligos	300	1248198	128373	2917
HC-0025	mlrPCR-sWGA	EBV specific MyBait oligos	300	1179837	82870	4271.2
HC-0026	mlrPCR-sWGA	EBV specific MyBait oligos	300	555072	137682	1209.5
HC-0027	mlrPCR-sWGA	EBV specific MyBait oligos	300	457153	136835	1002.3
HC-0028	mlrPCR-sWGA	EBV specific MyBait oligos	300	1910804	129668	4420.8
HC-0029	mlrPCR-sWGA	EBV specific MyBait oligos	300	8132282	138631	17598.4
HC-0030	mlrPCR-sWGA	EBV specific MyBait oligos	150	514390	89660	860.6
HC-0031	mlrPCR-sWGA	EBV specific MyBait oligos	150	896398	15578	8631.4
HC-0032	mlrPCR-sWGA	EBV specific MyBait oligos	150	1230192	59837	3083.9
HC-0033	mlrPCR-sWGA	EBV specific MyBait oligos	150	331520	16212	3067.4
HC-0034	mlrPCR-sWGA	EBV specific MyBait oligos	150	11241210	97685	17261.4
HC-0035	mlrPCR-sWGA	EBV specific MyBait oligos	150	24810	77012	48.3
HC-0036	mlrPCR-sWGA	EBV specific MyBait oligos	150	3822180	57343	9998.2
HC-0037	mlrPCR-sWGA	EBV specific MyBait oligos	150	41852	108638	57.8
HC-0038	mlrPCR-sWGA	EBV specific MyBait oligos	150	107808	3440	4700.9
HC-0039	mlrPCR-sWGA	EBV specific MyBait oligos	150	1028620	47534	3246
HC-0040	mlrPCR-sWGA	EBV specific MyBait oligos	150	9429614	111687	12664.3
eBL-Tumor-0001	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	4222227	142384	7413.5
eBL-Tumor-0002	mlrPCR-Swga	EBV specific MyBait oligos	200 & 300	6479721	142052	11403.8
eBL-Tumor-0003	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	8625903	60909	35404.9
eBL-Tumor-0004	mlrPCR-sWGA	EBV specific MyBait oligos	300	1123872	137537	2451.4
eBL-Tumor-0005	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	4648123	142405	8160
eBL-Tumor-0006	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	6448397	143731	11216.1
eBL-Tumor-0007	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	5448856	142866	9534.9
eBL-Tumor-0008	mlrPCR-sWGA	EBV specific MyBait oligos	300	10438253	141842	22077.2

eBL-Tumor-0009	mlrPCR-sWGA	EBV specific MyBait oligos	300	926150	136703	2032.5
eBL-Tumor-0010	mlrPCR-sWGA	EBV specific MyBait oligos	300	21509972	135211	47725.3
eBL-Tumor-0011	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	14949732	144063	25943
eBL-Tumor-0012	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	3673227	141985	6467.6
eBL-Tumor-0013	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	7861933	144125	13637.4
eBL-Tumor-0014	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	8701944	143606	15149
eBL-Tumor-0015	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	5825147	144044	10110
eBL-Tumor-0016	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	1161216	139661	2078.6
eBL-Tumor-0017	No amplification	EBV specific MyBait oligos	200	10699217	145418	14715.1
eBL-Tumor-0018	No amplification	EBV specific MyBait oligos	200	11258545	143781	15660.7
eBL-Tumor-0019	No amplification	EBV specific MyBait oligos	200	16192472	143175	22619.1
eBL-Tumor-0020	No amplification	EBV specific MyBait oligos	200	14633495	146920	19920.4
eBL-Tumor-0021	No amplification	EBV specific MyBait oligos	200	14430586	143942	20050.6
eBL-Tumor-0022	No amplification	EBV specific MyBait oligos	200	7687200	143947	10680.6
eBL-Tumor-0023	No amplification	EBV specific MyBait oligos	200	4481944	140842	6364.5
eBL-Tumor-0024	No amplification	EBV specific MyBait oligos	200	2585251	143612	3600.3
eBL-Tumor-0025	No amplification	EBV specific MyBait oligos	200	9775107	143175	13654.8
eBL-Tumor-0026	No amplification	EBV specific MyBait oligos	200	4070074	142949	5694.4
eBL-Tumor-0027	No amplification	EBV specific MyBait oligos	200	2473450	142313	3476.1
eBL-Tumor-0028	No amplification	EBV specific MyBait oligos	200	2785996	142423	3912.3
eBL-Tumor-0029	No amplification	EBV specific MyBait oligos	200	7602074	142738	10651.8
eBL-Tumor-0030	No amplification	EBV specific MyBait oligos	200	6079669	145541	8354.6
eBL-Tumor-0031	No amplification	EBV specific MyBait oligos	200	3972963	126974	6257.9
eBL-Tumor-0032	No amplification	EBV specific MyBait oligos	200	7118274	144616	9844.4
eBL-Tumor-0033	No amplification	EBV specific MyBait oligos	200	18395334	146197	25165.1
eBL-Tumor-0034	No amplification	EBV specific MyBait oligos	200	5679398	144524	7859.5
eBL-Tumor-0035	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	6133367	142874	10732.1

eBL-Tumor-0036	mlrPCR-sWGA	EBV specific MyBait oligos	300	1873678	137887	4076.6
eBL-Tumor-0037	No amplification	EBV specific MyBait oligos	200	4781768	143452	6666.7
eBL-Tumor-0038	mlrPCR-sWGA	EBV specific MyBait oligos	200 & 300	14668183	144437	25388.5
eBL-Tumor-0039	No amplification	EBV specific MyBait oligos	200	3484138	145453	4790.7
eBL-Tumor-0040	No amplification	EBV specific MyBait oligos	200	3758609	145880	5153
eBL-Tumor-0041	No amplification	EBV specific MyBait oligos	200	6158191	145025	8492.6
eBL-Plasma-0035	mlrPCR-sWGA	EBV specific MyBait oligos	300	385303	131409	879.6
eBL-Plasma-0036	mlrPCR-sWGA	EBV specific MyBait oligos	300	727755	139446	1565.7
eBL-Plasma-0037	mlrPCR-sWGA	EBV specific MyBait oligos	300	52615528	125217	126058.4
eBL-Plasma-0038	mlrPCR-sWGA	EBV specific MyBait oligos	300	2982299	135347	6610.3
eBL-Plasma-0039	mlrPCR-sWGA	EBV specific MyBait oligos	300	2834853	135720	6266.3
eBL-Plasma-0040	mlrPCR-sWGA	EBV specific MyBait oligos	300	336725	138187	731
eBL-Plasma-0042	mlrPCR-sWGA	EBV specific MyBait oligos	300	1226736	138256	2661.9
eBL-Plasma-0043	mlrPCR-sWGA	EBV specific MyBait oligos	300	643566	137788	1401.2
eBL-Plasma-0044	mlrPCR-sWGA	EBV specific MyBait oligos	300	992336	137048	2172.2
eBL-Plasma-0045	mlrPCR-sWGA	EBV specific MyBait oligos	300	9600176	138091	20856.2
eBL-Plasma-0046	mlrPCR-sWGA	EBV specific MyBait oligos	300	7416764	112949	19699.4
eBL-Plasma-0047	No amplification	EBV specific MyBait oligos	300	2209552	141331	4690.2
eBL-Plasma-0048	No amplification	EBV specific MyBait oligos	300	117683	125628	281
eBL-Plasma-0049	No amplification	EBV specific MyBait oligos	300	352588	136504	774.9
eBL-Plasma-0050	No amplification	EBV specific MyBait oligos	300	528192	138955	1140.4
eBL-Plasma-0051	No amplification	EBV specific MyBait oligos	300	6039246	142313	12730.9
eBL-Plasma-0052	mlrPCR-sWGA	EBV specific MyBait oligos	300	1682856	138645	3641.4
eBL-Plasma-0053	mlrPCR-sWGA	EBV specific MyBait oligos	300	9207274	135925	20321.4
eBL-Plasma-0054	No amplification	EBV specific MyBait oligos	300	1063382	142801	2234
eBL-Plasma-0055	No amplification	EBV specific MyBait oligos	300	1514583	142029	3199.2

Multiple long range Polymerase chain reaction (mlrPCR) was utilized for EBV specific genome-wide amplification (sWGA). Samples with higher viral loads were not subjected to sGWA. EBV-specific bait oligos (EBV-specific biotinylated RNA probes) were utilized to enrich EBV DNA. Illumina paired-end read length was ranging from 150 to 200 and 300. Provided as sequencing results are total sequence reads, genome size in bases, and the average depth of coverage over assembly for each sample sequenced. A total of 95 archival samples plus 6 plasma replicates of tumor samples (Colored in green) were sequenced. 9 samples (colored in red) were eliminated from phylogenetic and recombination analysis.

Appendix VIII: Genomic Recombination in Plasma-Tumor Replicates



A. The figure illustrates a comparison of recombination patterns of 4 plasma-tumor replicates (35, 37, 38, and 39). Each side-by-side bar represents a unique event in a plasma and tumor isolate. B. Abbreviation: eBL, endemic Burkitt lymphoma. Phylogenetic Tree of 6 plasma and tumor replicates. Each plasma and tumor replicate has a unique color e.g. eBL-Tumor-0036 and eBL-Plasma-0036 are colored in red.

To validate the precision of the recombination detection, the study characterized the occurrence of recombination in 6 plasma and tumor replicates. As expected RDP4 identified the same recombination events in 4 plasma samples as well as their tumor replicates (A). A total of 2 replicates had no evidence of recombination. Further, a phylogenetic tree was constructed for the 6 plasma tumor replicates to confirm the output from RDP4. The replicates clustered together perfectly in the same node (B).

Appendix IX: Summary of Key R Scripts used in the Data Analysis

#Importing the data for analysis

```
Eddy=read.csv(file.choose(),header=T)
```

#Genomic recombination breakpoints per kilobase pair distribution in coding sequences

```
ggbarplot(Eddy, x = "Coding.Sequence", y = "Breakpoint.Kbp", fill = "Phase",  
  palette=c("red","blue","green","purple","yellow")) + theme_light() +  
  ggtitle("Distribution of Recombination Breakpoints") + xlab("Coding Sequence") +  
  theme_light() + ylab("Number of Recombination Breakpoints") +  
  theme(plot.title = element_text(size=12, face="bold", hjust = 0.5),  
  axis.title.y = element_text( size=10, face="bold"), axis.title.x =element_text( size=10,  
  face="bold")) +  
  theme(axis.text.x = element_text(size=6, angle = 90,vjust = 0.5, hjust = 1)) +  
  theme(legend.position = "top") + geom_hline(yintercept=11.05, linetype="dashed",  
  color="black", size=0.2)
```

#Plotting the frequency of genomic recombination events stratified study characteristics

```
ggplot(Eddy, aes(fill=study.characteristic, y=Frequency, x=Unique.Event)) +  
  geom_bar(position="dodge", stat="identity") + theme_light() +  
  scale_fill_manual(values = c("blue","red","green")) + ggtitle("study.characteristic") +  
  xlab("Recombination Events") + ylab("Population Frequency") +  
  theme(plot.title = element_text(size=12, face="bold", hjust = 0.5),  
  axis.title.y = element_text( size=10, face="bold"),  
  axis.title.x =element_text( size=10, face="bold"))
```

#A univariate logistic regression model for the association of genomic recombination events with eBL

```
logistic=glm(relevel(eBl.Status, ref = "BL") ~ relevel(Recombination.event, ref = "Present"),  
  data = Eddy,family = "binomial")  
summary(logistic)
```

```
exp(coef(logistic))
exp(cbind(OR = coef(logistic), confint(logistic, level = 0.95)))
```

#A multivariate logistic regression model for the association of genomic recombination events with eBL

```
logistic=glm(relevel(eBl.Status, ref = "BL") ~ relevel(Recombination.event, ref = "Present")
+ Viral.Type ,data = Eddy,family = "binomial")
```

```
summary(logistic)
```

```
exp(coef(logistic))
```

```
exp(cbind(OR = coef(logistic), confint(logistic, level = 0.95)))
```

#Plotting genomic recombination events between plasma tumor replicates

```
ggplot(Eddy, aes(fill=Sample.source, y=Fragment_Size, x=Unique.events)) +
  geom_bar(position="dodge", stat="identity") + theme_light() +
  scale_fill_manual(values = c("red","blue")) + theme_light() +
  ggtitle(" Plasma Tumor Replicates") + xlab("Unique Events") +
  theme_light() + ylab("Fragment Size") + theme(plot.title = element_text(size=12,
face="bold", hjust = 0.5),
  axis.title.y = element_text( size=10, face="bold"), axis.title.x =element_text( size=10,
face="bold")) +
  facet_wrap(~Eddy$Replicates, scales = "free_x") + coord_flip() +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5, hjust = 1))
```