Within-Host Diversity of Kaposi Sarcoma-associated Herpesvirus

Jan Clement A. Santiago

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

James I. Mullins, Chair

Michael Lagunoff

Alexander Greninger

Roger Bumgarner

Program Authorized to Offer Degree:

Microbiology

University of Washington

Abstract

Within-Host Diversity of Kaposi Sarcoma-associated Herpesvirus

Jan Clement Ang Santiago

Chair of the Supervisory Committee:

James I. Mullins

Department of Microbiology

Kaposi sarcoma (KS) is a progressive, incurable soft tissue disease that is the most common cancer of men in regions of sub-Saharan African and the second most common among people with HIV. Kaposi Sarcoma-associated Herpesvirus (KSHV) is the infectious cause of KS, but why only a small fraction of those infected develop KS is not understood. Immune status, viral co-infection, host genetics and environmental factors all are likely to play a role. My thesis sought to identify a possible viral genetic component of KS, through an analysis of the diversity of whole KSHV genomes *within* individuals afflicted by KS. Strain variation or *de novo* mutations occurring in other oncoviruses have been associated with variations in disease risk, disease manifestation and clinical course. These variations have the potential to become diagnostic biomarkers and help reveal insights into the pathogenicity of the virus. To identify potential tumor-associated mutations in KSHV I analyzed the ~131 kb unique sequence regions of 43 whole KSHV genomes from tumors and oral swabs from 22 individuals and then screened additional tumors by targeted sequencing and ddPCR to better assess the frequency of identified mutations. In total, 65 KS

tumors and 18 oral swabs were evaluated from a cohort of 30 study participants with KS, all from the Uganda Cancer Institute. In addition, the major internal repeat regions of the KSHV genome were examined from 16 individuals. The newly developed technologies of duplex Unique Molecular Identifier (dUMI) and single molecule real time sequencing with UMI (SMRT-UMI) were employed to provide an unprecedented accuracy and depth of study of herpesvirus populations within individuals. These studies revealed recurring tumor-associated mutations, specifically an overrepresentation of a specific region encompassing the K5 and K6 genes and inactivating mutations in the K8.1 gene, both of which were found in at least a third of the cohort and had associations with late-stage tumor characteristics. In contrast, the translation potential of full-length Kaposin proteins from IR2 was lost in a majority of individuals, but no tumor association or clinical correlate was identified. Rearrangement breakpoints were sometimes shared between different tumors from the same person, indicating spread by metastases, helper viruses or residual infectivity. In summary, these studies revealed the presence of selection for tumor- and clinical-stage specific changes the KSHV genome. Hypotheses on whether these changes play a role in KS tumor formation, immune evasion and driving persistence in the host are put forth along with proposed experiments to test them. Among them, findings from my thesis suggest that targeted screening of the K5-K6 and K8.1 gene regions can be helpful in tumor classification.

CONTENTS

CHAPTER 1: Introduction 10
Kaposi Sarcoma and KSHV associated diseases 10
Gaps in the Literature to be Addressed 12
Kaposi Sarcoma-associated Herpesvirus 13
Viral Genetic Heterogeneity in Human Tumor Viruses 15
Genetic Heterogeneity of KSHV 16
Whole Genome Heterogeneity of KSHV 17
KSHV Increases Both Host and Virus Genome Instability 18
Defective Viral Genomes 20
Significance of Repeat Sequence Heterogeneity 21
Sequencing Challenges and Innovations 22
Scope of Thesis Work 23
CHAPTER 2: Methods 26
Ethics statement 26
Study Cohort and Specimen collection 26
DNA preparation 27
PCR 27
Copy number quantification 27
UMI-addition and Illumina library preparation 31
Library enrichment & sequencing 31
De novo assembly of sample-consensus genomes 33
Variant identification from dUMI-consensus reads 35
Structural variation and integration breakpoint detection 36
Analysis of DNA structures and motifs 37
Single-molecule-real-time unique-molecular-identifier (SMRT-UMI) sequencing of major internal repeat regions of KSHV 39
Demultiplexing and UMI consensus processing 40
Statistical Analysis 40
CHAPTER 3: Changes in KSHV genomes in Ugandan adults with Kaposi sarcoma 41
Introduction 41

Results 42

Assessment of the dUMI sequencing protocol with a KSHV infected cell line 42

KSHV sequence derivation from tumor tissues and oral swabs 45

KSHV genomes were virtually identical at the point mutational level between tumors and oral swabs from the same individual 49

Aberrant KSHV genome structures in tumors 49

The same aberrant KSHV genomes are found in multiple lesions from the same individual 57

Mutations in sample-consensus KSHV genomes from tumors that impacted protein coding sequences 59

Lack of evidence for integration of KSHV sequences into human chromosomes 63

Co-infection with EBV detected predominantly in oral swabs 63

Chapter 4: Tumor-associated KSHV genome aberrations, mutations and their clinical correlates in Ugandan adults with KS 64

Introduction 64

Results 65

KSHV genomes in KS tumors often have significantly higher representation of a subgenomic region near IR1 65

A 2.2-kb segment region encompassing K5 and K6 corresponds to the minimal region of overrepresentation. 69

KSHV rearrangement breakpoints are associated with G-quadruplexes 71

Potentially inactivating mutations are common in the K8.1 gene in tumors. 73

Polymorphisms in miR-K10, K4.2 and K11.2. 76

KSHV mutations associated with disease course and tumor characteristics 79

CHAPTER 5: Variation Within Major Internal Repeats of Kaposi Sarcoma Herpes Virus In Vivo 89

Introduction 89

Results 91

Intra-host diversity of KSHV major internal repeat regions 91

KSHV internal repeats had higher rates of intra-host variation than the rest of the genome 101

Direct repeats in IR1 were imperfect more often than in IR2 103

No full-length Kaposin isoform from IR2 in the majority of IR2 sequences from Africa 109

Clinical Phenotypes associated with IR1 and IR2 diversity 109

CHAPTER 6: Discussion & Future Directions 115

The low point mutational diversity of KSHV in adults with KS 115
KSHV genome aberrations are associated with KS tumors 116
Shared mutations in distinct tumors of the same individuals 117
Recurrent K5-K6 region overrepresentations in KS Tumors. 117
Significance of the IR1 region overrepresentation 118
Association of KSHV genome rearrangement breakpoints to G-quadruplexes 119
K8.1 inactivating mutations 120
Point mutations in microRNA-K10 121
Novel genotypes of K4.2 and K11.2 122
Clinical phenotypes of tumor-associated mutations and polymorphisms. 123
KSHV repeats variation as revealed by SMRT-UMI 124
Heterogeneity of LANAr sequences 125
IR1 and IR2 imperfect repeats 126
Loss of full-length Kaposin isoforms from IR2 128
CHAPTER 7: Conclusions and Future Directions 129
BIBLIOGRAPHY 131

TABLES

Table 1. Study participant characteristics *26*
Table 2. Primer list *28*
Table 3. Breakpoints identified in KSHV genomes *38*
Table 4. Origin and processing results from specimens for KSHV genome analysis *43*
Table 5. Gene copy numbers in tumor or oral swab DNA *54*
Table 6. Unique KSHV mutations observed in tumors compared to oral swabs from the same individual. *60*
Table 7. Summary of KSHV genome characteristics in all 65 KS tumor *67*
Table 8. P-values of non-B-DNA summed probabilities/scores and distances to breakpoints *73*
Table 9. Coding sequence mutations *77*
Table 10. PCR screening and sequencing in oral swabs *78*
Table 11. of KSHV genome types with K4.2 and K11.2 polymorphisms *82*
Table 12. Participant clinical traits *84*

Table 13. Individual KS tumor-level characteristics and morphotypes *85*
Table 14. Odds ratios of mutations to clinical or tumor traits *87*
Table 15. KSHV genetic polymorphisms and survival rates, Cox regression *88*
Table 16. Correlations between observed KSHV mutations *88*
Table 17. Sample counts, UMI counts and intra-host diversity of KSHV repeats in tumors and oral swabs of 16 individuals *92*
Table 18. UMI counts, circular consensus sequence (CCS) counts and minimum agreement ... *94*
Table 19. Summary of distinct populations of repeat sequences detected. *96*
Table 20. Clinical associations of repeat polymorphisms *113*
Table 21. Survival rate association of repeat genotypes *114*
Table 22. Correlation between degraded IR1 and loss Kaposin isoforms. *114*

FIGURES

Figure 1. dUMI-adaptors and primers for duplex UMI sequencing 32
Figure 2. Workflow for analyzing intra-host KSHV genome diversity from clinical samples 34
Figure 3. Workflow of genome assembly and variant analysis 35
Figure 4. KSHV genomes in BCBL-1 cells have low point mutational diversity 44
Figure 5. Raw, sUMI and dUMI read coverage across tumor- and oral swab-derived KSHV genomes 46
Figure 6. Potential sequencing artifacts 47
Figure 7. Point mutational diversity in KSHV genomes from tumors and oral swabs 48
Figure 8. KSHV phylogenetic relationships in variable genes K1 and K15 and whole genomes 50
Figure 9. KSHV genomes in the U003-C tumor 52
Figure 10. KSHV genomes in two tumors from participant U008 had a 14.8-kb region flanking IR1 duplicated and translocated into IR2 54
Figure 11. KSHV genomes in U020-B and U020-C have large, distinct deletions 57
Figure 12. U008-B and U008-D were from distinct lesions on the left leg. 58
Figure 13. Genome inversion found in multiple tumors of participant U003. 58
Figure 14. Mutations of KSHV genomes in tumors from participants 004 and 003 61
Figure 15. Predicted secondary structure of the stem loop precursor of miR-K10a 62
Figure 16. Raw read coverage of KSHV genomes detected in 20 tumors from 16 individuals ... 66
Figure 17. PCR and sequencing confirmation of rearrangement connecting IR1 to TR sequences in tumor U048-D 70

Figure 18. Overlap in KSHV IR1 overrepresentation regions.....71
Figure 19. Association of breakpoints to non-B-DNA.....75
Figure 20. Inactivating mutations observed in the K8.1 gene 76
Figure 21. Polymorphisms in K4.2 and K11.2 81
Figure 22. Diagram of KSHV internal repeats IR1 and IR2.....90
Figure 23. Representative sequence alignments.....97
Figure 24. Detectable intra-host diversity as a function of UMIs recovered. 102
Figure 25. KSHV tandem repeat counts and mismatches in lab strains and other published KSHV genomes. 110
Figure 26.Translation start sites of Kaposin isoforms in different KSHV strains.....112

CHAPTER 1: Introduction

Kaposi Sarcoma and KSHV associated diseases

Kaposi Sarcoma (KS) is a soft tissue cancer of endothelial cell origin, most commonly presenting as vascular, cutaneous lesions and discolorations on the skin. It is a multifocal angiogenic disease and can range from being indolent and localized to hyperproliferative and disseminated [1]. KS lesions can develop in surfaces of the oral cavity and involve lymph nodes, visceral organs and the lungs, and can often be a cause of mortality [2,3]. The lesions are characterized by pervasive angiogenesis, infiltration of inflammatory cells and atypical differentiation of spindle-like tumor cells [4]. The tumors typically progress from distinct patch, to plaque and then nodular stages, but a patient can simultaneously present tumors in various stages [5]. There is no cure for KS and current treatments are mostly palliative [3].

There are 4 epidemiological classes of KS --classic, endemic, iatrogenic and epidemic-- although the tumor lesions are indistinguishable by histopathology [6]. Classic KS was described in 1872 by an Austro-Hungarian doctor Moritz Kaposi, the disease's namesake [7]. Classic KS mostly presents as chronic, indolent tumors at the lower extremities of elderly men of Mediterranean and Eastern European descent [8]. Endemic KS is found in sub-Saharan Africa and has variable prognoses, with its most aggressive forms involving disseminated lymphadenopathy in children [8]. Iatrogenic KS can be found in immune-suppressed patients after solid organ transplantation [8]. Epidemic KS is rapidly progressive, multicentric and is HIV-associated; its incidence reflected the emergence of the HIV pandemic. While previously rare in the United States, KS became an AIDS-defining disease, developing in about 50% of men with both HIV and KSHV before the widespread use of combined antiretroviral therapy [9], now referred to simply as ART. Recently a fifth epidemiologic class, nonepidemic KS, was proposed to account for KS that develops in men who have sex with men but who are HIV-negative and have no immunodeficiency, and whose indolent clinical course resembles classic KS [8].

A key feature of KS lesions is the ubiquitous presence of endothelial spindle-like cells infected with Kaposi Sarcoma-associated Herpesvirus (KSHV). Identified in KS lesions in 1994 in the laboratory of Yuan Chang and Patrick Moore, DNA of the then novel herpesvirus was discovered by PCR and DNA subtractive hybridization to reveal genetic differences between KS lesions and

adjacent normal tissue from the same patients, followed by selective enrichment of this tumor-unique DNA from KS tissues [10]. KSHV was subsequently recognized as the etiologic agent of KS [11] and is one of seven known human oncoviruses (others being Epstein-Barr virus, human papillomavirus, hepatitis B and C viruses, human T-cell lymphotropic virus-1, and Merkel cell polyomavirus) [12].

KSHV as a cause of KS fulfills all Bradford Hill Criteria for causation [13,14], which were nine principles originally proposed when establishing cigarette smoking as a cause of lung cancer [15]. Epidemiological data and the detection of KSHV in all cases of KS showed an association that is (1) strong, (2) consistent, (3) coherent and (4) specific, correcting for false positives from healthy controls and unrelated diseases. The higher probability of detecting KSHV in tumor than non-tumor sites fulfills the (5) dose gradient criterion in the virology field [13]. Detection of KSHV by PCR and antibodies before onset of KS showed (6) a temporal sequence. Subsequent laboratory characterization of KSHV-encoded oncogenes provided (7) biological plausibility and (8) analogy. Finally, (9) experimental evidence was provided by an incidental finding in a randomized clinical trial of an anti-herpesvirus drug, which showed that it prevented new KS tumor formation, albeit being ineffective against established KS tumors [13].

In addition to KS, KSHV is the etiologic agent of the lymphoproliferative diseases primary effusion lymphoma (PEL) and a plasmablastic form of multicentric Castleman disease (KSHV-MCD) [3]. PEL is a clonal B-cell neoplasm, commonly presenting as an effusion in the pleura, peritoneum or pericardium, but sometimes as an extra-cavitary solid lymphoma [11]. KSHV-MCD is a polyclonal neoplasm of enlarged plasmablasts that occurs along with systemic inflammation, high KSHV viremia, and bouts of elevated cytokines, especially IL-6 [3]. These two B-cell malignancies often arise in conjunction with KS in HIV-infected individuals [3]. KSHV is additionally implicated in KSHV-associated inflammatory cytokine syndrome (KICS) [3,16]. Like MCD, KICS is characterized by systemic inflammation, widespread lytic KSHV replication and flares of high IL-6 and IL-10 cytokine levels, but without the enlarged lymph nodes and plasmablasts that define MCD [17]. Finally, immune reconstitution inflammatory syndrome (KS-IRIS) results in 10% of AIDS-KS cases upon initiation of ART [11].

KS is a non-trivial public health concern, it along with non-Hodgkin lymphoma and cervical cancer are the most frequent cancers [18,19] among the global population of 37 million HIV-infected individuals [20]. In sub-Saharan Africa, where both KSHV and HIV are prevalent, KS has

become the leading cancer in men and second in women [18,19]. The advent of ART has drastically reduced the rates of KS along with AIDS [21,22]. However, KS can arise even with restored CD4+ cell counts, and people taking ART comprise a third of AIDS-KS cases in the US [23–26]. The risk for KS in ART-controlled HIV infections remains 800-fold greater than in the general population [22]. Since old age is also a risk factor for KS [3,27], the drop in rates of KS worldwide may reverse as more people surviving with HIV due to ART are aging. ART has no effect independent of HIV on the risk of developing KS [28] and is not expected to impact rates of non-AIDS KS [8,18,29]. Standard chemotherapy does not induce even partial remission in ~30% of KS patients, and paradoxically, ~10% of AIDS-KS patient develop KICS upon initiation of ART [3]. Anti-herpesvirus drugs have inconsistent results against new KS tumors and established tumors are refractory to them [30]. For these reasons KS will continue to be a substantial disease burden for the foreseeable future.

Gaps in the Literature to be Addressed

While KSHV is a requirement for KS, infection alone is far from sufficient. The rate of KS among the KSHV-infected worldwide is estimated to be only 1 in 10,000 [9]. KS emerges only after years or decades of chronic infection. The precise factors essential for developing KS are poorly understood. HIV co-infection is a potent risk factor for KS; HIV-infected people before taking ART have a 2,800-fold higher risk of KS compared to the general population [22]. Aside HIV infection, KS is also common in organ transplant recipients after chemically-induced immunosuppression. However, KS can also develop with little evidence of immunodeficiency as evidenced by endemic and non-epidemic KS. Known host factors that contribute to higher KS rates are host genetics, old age and male sex [31]; KS occurs at significantly higher rates in men [18]. Environmental factors include hypoxia or low oxygen conditions, thought to possibly occur in the extremities, and oxidative damage from chronic infection [31,32].

Viral genetic differences arising from strain diversity or de novo mutations can influence the properties of oncogenic viruses and the diseases they cause, as will be discussed below. This thesis examines the role of viral genetic factors in the natural history of KS, and sheds light on genetic aspects of KSHV genome diversification, which for practical purposes may inform rational therapeutic targets. The work that will be described is the most comprehensive characterization

to date of KSHV intra-host genomic variation. Elucidation of this variation can be a powerful means of uncovering KSHV polymorphisms and mutations associated with various KS disease manifestations, clinical presentations and outcomes, which can potentially serve as predictive and treatment-selective biomarkers. While factors leading from KSHV infection to the development of KS are likely multifactorial – including host genetics and immune function, viral co-infections, and environmental factors – the central question this thesis addresses what role(s) KSHV genomic diversity might play in the natural history of KSHV infection in KS patients.

Kaposi Sarcoma-associated Herpesvirus

Alternatively known as Human Herpesvirus-8 (HHV-8), KSHV is a large, double-stranded DNA virus of the *Herpesviridae* family, in the lymphotropic or gamma (γ) *herpesvirinae* subfamily [9]. Gamma-herpesviruses are typically tropic to T or B lymphocytes [9]. KSHV can be further classified into the gamma-2 or rhadinovirus genera together with *Herpesvirus saimri* [33] and KSHV's closest animal virus homolog, *retroperitoneal fibromatosis-associated herpesvirus Macaca nemestrina* (RFHVMn) [34]. KSHV is one of only 2 gamma-herpesviruses known to infect humans, the other being Epstein-Barr Virus (EBV) of the gamma-1 or lymphocryptovirus subfamily. Like other gamma-2 herpesviruses, the genome of KSHV is composed of a central unique sequence region interspersed with 3 major internal repetitive regions (IR1, IR2 and LANAr), flanked by 20-40 copies of terminal repeat (TR) sequences. It contains at least 86 open reading frames (ORF) that were numbered serially from 5' to 3' (ORF4 to ORF75). The ORFs consist of genes for herpesvirus replication and virion structure shared with other gamma-herpesviruses, and 15 genes unique to KSHV and other gamma-2 herpesviruses are designated K1 to K15 instead of ORFxx. Protein coding sequences discovered after KSHV gene names had been established were appended with a decimal number (*e.g.,* K4.1, K8.1). Moreover, the KSHV genome encodes several transcripts for long non-coding, micro and circular RNAs [35,36].

KSHV is highly endemic (>50%) in parts of sub-Saharan Africa, moderately endemic (10%-25%) in the Mediterranean region and parts of Latin America, and rare (<10%) in most of the US, Europe and Asia, except for men who have sex with men [16]. KSHV is primarily transmitted via saliva during childhood and sexually, with some cases of infection via blood transfusion or intravenous drug use. In endemic areas, family members will often infect other family members

[37]. B cells are a major reservoir for KSHV persistence in the body, and human primary tonsillar B cells can be infected by KSHV [38]. Other than B-lymphocytes KSHV has a broad host cell type tropism encompassing monocytes, dendritic cells, fibroblasts, endothelial cells and epithelial cells [9]. EphA2R and EphA4R are among the receptors employed by KSHV for cell entry [38,39].

The infection cycle of KSHV, like other herpesviruses, comprises latent and lytic phases [9]. KSHV generally enters latency upon infection, after a brief abortive replication to create a few dozen KSHV genome copies [40]. KSHV genomes are linear while inside virion capsids, but during latency the TR on both ends are covalently joined to form circular episomes, which are tethered to host cell nucleosomes by the product of viral gene ORF73/LANA (Latency associated nuclear antigen), which also mediates viral DNA replication and propagation to daughter cells [16]. Only a few viral latency genes are expressed for viral genome maintenance and enhanced host cell survival [41]. Upon reactivation triggered by incompletely understood host or environmental factors, the lytic genes are expressed in ordered stages – immediate-early, delayed-early, and late [32]. Immediate-early genes are generally viral transcription factors. One key immediate-early gene, KSHV trans-activator ORF50/RTA, regulates the switch from latent to lytic phase and initiates transcription of most other lytic genes. Delayed-early genes encode enzymes and regulatory proteins for viral DNA replication and creating a favorable cellular environment for productive viral replication. Late genes encode most viral structural proteins including capsid, tegument and glycoproteins. Concatemers of multiple, full-length viral genomes connected head-to-tail are produced by rolling circle replication. Unit-length genomic DNA, cut between variable numbers of TR sequences, is packaged into preformed capsids in the nucleus. The intranuclear capsid then acquires its inner envelope from the nuclear membrane, tegument proteins in the cytoplasm, and its outer envelope by budding into Golgi vesicles. Mature virions are finally released extracellularly via fusion of vesicle and plasma membranes [32].

KSHV genes expressed in both lytic and latent phases are implicated in tumorgenicity. They have functions in promoting cell proliferation, angiogenesis, inflammation, anti-apoptosis, immune evasion, among other effects. KSHV latency genes that were found to be oncogenes *in vivo* are LANA, which maintains viral latency; ORF72/vCyclin, a homolog of cellular cyclin D, that regulates cell cycle progression; and K13/vFLIP, which inhibits apoptosis and activates NF-κB [42]. Among the oncogenic lytic genes are K1, a cell membrane protein that functions akin to a constitutively active B-cell receptor; K2/vIL-6, a homolog of inflammatory cytokine IL-6; ORF36/vPK, a viral serine-threonine kinase that alters cell signaling; and ORF74/vGPCR, a constitutively active G

protein-coupled receptor functionally homologous to human IL-8 receptor [42]. Additionally, KSHV miRNAs, by influencing and subverting cellular regulatory networks, contribute to viral oncogenicity [43].

Viral Genetic Heterogeneity in Human Tumor Viruses

Strain variation and *de novo* mutations in tumor viruses have been associated with varying disease risk and course. The Hepatitis C (HCV; an RNA virus) genotype 1b confers twice the risk for developing hepatocellular carcinoma compared to other HCV genotypes, and deep sequencing revealed that the proportions of disease-associated HCV core protein polymorphisms increased along with disease severity [44]. The various genotypes of Hepatitis B Virus (HBV; a DNA virus that replicates through an RNA intermediate) have different seroconversion rates, lifetime risk for hepatocellular carcinoma and responses to interferon-based therapy. There are HBV core promoter and pre-S protein coding sequence mutations that are predictive of disease progression [45]. Cancer-causing human papillomavirus (HPV; a small DNA virus) types are divided into high risk and low risk groups, with HPV types 16 and 18 being the most carcinogenic [46]. In both HPV and Merkel cell polyomavirus (MCPyV; another small DNA virus), integration into host chromosomes and disruptive mutations to viral genes E2 and Large T, respectively, are necessary drivers for the progression of their respective cancers [46].

The gamma-herpesvirus EBV is the most closely related human tumor virus to KSHV, and EBV strains have genetic differences among them that are associated with different biological properties and distinct clinical diseases. For example, EBV strains with a defective EBNA3B gene are uniquely linked to diffuse large B-cell lymphomas and were more tumorigenic in mice [47]. In another example, EBV strains derived from nasopharyngeal carcinomas (NPC) poorly immortalize B-cells and have higher tropism to epithelial cells, unlike other strains that are highly B-cell transforming [48,49]. Replacing one residue in the transactivator protein EBNA-2 of a weakly transforming strain with that of an efficiently transforming strain led to superior growth in a B-cell growth maintenance assay [50]. Also, independent contributors to the B-cell hyper-lytic phenotype of NPC-derived strains were mapped to a 7-nt segment of its DNA polymerase gene [48] and to NPC strain-specific polymorphisms in the DNA binding domain of EBV latency gene EBNA-1 [47].

EBV genomes reportedly have measurable intra-host diversity between saliva and peripheral B-cells during acute infection [51]. These differences were concentrated in EBV latency and glycoprotein encoding genes and diminished over 6-12 months as acute infection transitioned into chronic infection. At the last time point measured, EBV genomes were genetically closer to a B-cell transforming genotype, mostly due to nonsynonymous mutations in latency-associated genes [51]. Human cytomegalovirus (HCMV) herpesvirus genomes also exhibit compartment-specific mutations and intra-host diversity [52]. Single polymorphisms in specific herpesvirus genes can have phenotypes that impact disease course and management [53], including single amino acid changes in HCMV kinases or DNA polymerase that can confer drug resistance [54,55].

Occasionally, EBV genomes or genome fragments have been found integrated into chromosomes in gastric and nasopharyngeal carcinoma and lymphoma cells [56], even though integration is not a required part of its life cycle. The number of integration sites correlated with the amount of EBV structural variations [57], and these sites had microhomology between sequences in human and EBV genomes [58]. Intriguingly, EBV integrations were commonly found near or inside tumor suppressor and inflammation-related genes, some of which were shown to affect cellular gene expression levels [58].

Genetic Heterogeneity of KSHV

KSHV has genetic variability that could potentially contribute to disease risk and course. The hypervariable gene K1 is by convention used to classify KSHV isolates into at least 6 subtypes A to F, and sub-subtypes [59]. K1 is an oncogene encoding a single transmembrane glycoprotein with a hypervariable extracellular domain at its N-terminus [60]. The C-terminal cytoplasmic domain is highly conserved and contains a constitutively active immunoreceptor tyrosine-based activation motif (ITAM) thought to imitate signals used to promote the survival of host lymphocyte cells [60]. It is unclear what drives the hypervariability of K1 extracellular domains, but there are numerous reports of certain K1 subtypes having differing KS risk profiles and prognoses: Subtype A was correlated with more aggressive KS and higher plasma viral loads than subtype C [61–63]; subtype A5 was associated with more extensive KS (>10 KS lesions) among the sub-subtypes of both A and B [64], and subtype B was associated with better KS prognosis compared to A, C and F subtypes [65]. However, these findings have not been consistently observed [66–69]. K1

subtypes are predictive of host ethnicity and geographic origin and are thought to follow ancient human migration patterns [59]. Subtypes A and C predominate in Europe and the US, subtype B predominates in Africa, D is present in Pacific islands, E has been found among Amerindians, and F was discovered among Bantu tribes in Uganda [59].

Polymorphisms in the KSHV microRNA-dense region, found 3' of IR2 in the KSHV latency locus, are highly correlated with the occurrence of MCD [70]. This region has significantly more variability and divergence in MCD patients than in non-MCD infections. The polymorphisms were subsequently shown to alter processing and expression levels of the 10 encoded microRNAs *in vitro* [71] and in KS lesions [72]. Polymorphisms in these 22 - 23 nucleotide microRNAs are of consequence since they co-opt cellular processes that drive cell survival, differentiation and transformation [43].

Whole Genome Heterogeneity of KSHV

The first two decades following the discovery of KSHV saw only 6 whole KSHV genomes sequenced, owing to the difficulty of sequencing the entire 165 kb genome interspersed and flanked by GC-rich repetitive sequence regions. Furthermore, this number lagged far behind the sequencing of other human herpesviruses, and were restricted to isolates in the US and Europe [73]. Over the next 5 years, however, a total of 66 additional KSHV genomes were sequenced from Japan, Zambia and Uganda [74–77]. The growing use of second/next-generation sequencing fueled the increase in available KSHV genomes. The 72 whole KSHV genome sequences then available painted a more comprehensive picture of KSHV global diversity than single KSHV genes alone. Importantly, these studies showed that polymorphisms in the rest of the genome contribute in aggregate much more to KSHV diversity than within the 0.9 kb hyper-variable K1 gene alone. KSHV whole genome phylogenetic trees are markedly different from K1 gene trees of the same strains, and removing the hypervariable genes did not change the phylogenetic tree topology from the whole genome tree [75,76]. Unmistakable signatures of recombination were also seen from analyzing whole genomes [77], corroborating the mosaic genomes inferred previously by sequencing short genic segments and complicating disease risk associations solely based on K1 subtypes [59].

Whole viral genome sequencing has also revealed genome-level structural variations. The first KSHV genome published, from a biopsied KS tumor and sequenced using the Sanger method, had a 33-kb mid-genome section duplicated into the midst of KSHV terminal repeats [33]. A study of 16 whole KSHV genome sequences from KS tumors biopsied from Zambia was done by *de novo* assembly of short reads [76]. It reported 4 KSHV genomes with regions of up to 3 times the average read coverage, suggesting duplications, although the exact structures were not elucidated. In another example, a 9-kb region from IR1 to ORF19 was found duplicated into the TR region of BAC36 [78], a bacterial artificial chromosome constructed from the KSHV-infected PEL cell line BCBL1 [79]. Incidentally, this was where the BAC cassette was targeted for insertion, between ORF18 and ORF19 [79]. Selection markers inside the duplication would favor retention of the short sub-genomic fragment in *E. coli* during the BAC cloning process and the loss of the full-length KSHV genome via homologous recombination. It is thus unclear how BAC cloning could produce the observed duplication if it had not naturally existed already as a minor variant in the BCBL1 cell line.

KSHV Increases Both Host and Virus Genome Instability

As genomic instability and gene amplification are hallmarks of tumor cells [80], intra-genome copy number variation and rearrangements of resident herpesvirus genomes is not surprising in KS tumors. Host cell chromosomal aberrations in KSHV-infected tumor cells are frequent and have been suggested by several findings to be non-random. Recurring losses and translocations in specific regions of chromosomes 3, 7, 8, 11, 14, 16, 17, 21 and Y are associated with KSHV-infected primary KS tumors or cell lines [81]. KSHV-positive PEL cells have recurring abnormalities in specific regions of chromosomes 1, 4, 7, 8, 12 and X [82–84]. KSHV also induces mitotic abnormalities leading to chromosomal aberrations in human umbilical vein endothelial cells (HUVEC) [85]. Not only have integrations of EBV sequences into host cell chromosomes been observed [56], but host cell genome instability has been found to be a common result of repeated chemical reactivations from latency [86].

Double-stranded DNA (dsDNA) breaks are key intermediates to gene amplification and genomic instability [80]. Many components of the DNA damage response are hijacked during KSHV infection and replication, and KSHV lytic and latency proteins contribute directly to dsDNA

breaks [87]. KSHV processivity factor PF-8 increases dsDNA breaks and impedes non-homologous end-joining repair in a dose-dependent manner [88]. KSHV viral transcription factor ORF57 can induce dsDNA breaks by sequestering the human Transcription and Export complex (hTREX) at the expense of cellular transcription, increasing the formation of highly vulnerable RNA:DNA hybrids in the form of R-loops [89]. KSHV vGPCR downregulates several DNA damage and repair genes through microRNA-34a [90]. Latency gene vCyclin, and to an extent LANA, can independently activate DNA damage response pathways [87].

Large DNA viruses typically have their own proof-reading enzymes, helicases and processivity factors, which results in high replication fidelity [9]. Recombination plays an outsized role in generating genetic diversity in all alpha-, beta-, and gamma-herpesvirus subfamilies within human prehistory and contemporary times [91–93]. As discussed below, the conserved replication mechanism of herpesviruses is inherently conducive to strand-transfer and genome segment inversions, independent of the herpesvirus genome template [64]. Additionally, genomes of gamma-herpesviruses, compared to other herpesvirus families, are enriched in motifs that are inducers of double-stranded breaks and recombination in B-cells, their primary host cells [94]. An analysis of recombination across 171 EBV genomes showed that recombination signals significantly clustered with CCCAG and TGGAG motifs [92].

Herpesviruses are commonly thought to replicate by rolling-circle replication (RCR) during the lytic phase [9]. However, RCR alone yields linear amplification, and most dsDNA viruses first undergo an exponential circle-to-circle, or theta (θ), replication before switching to RCR to produce linear genome concatemers, which are in turn cut into single genome units and packaged into capsids [95,96]. Recombinases are integral in the switch from theta replication to RCR [96]. Multi-branched and entangled DNA structures associated with theta-like replication had been found in HSV-1 and EBV [95–97]. Additionally, double-rolling circle replication (DRCR), as when there are 2 replication forks continuously going around a circular genome in the same direction, has been suggested to explain the frequent inversions of HSV-1 genome isomers [98]. DRCR can be induced from template switching events and are thought to accumulate repeating DNA structures in close conformations where they can more frequently recombine. In yeast and mammalian cell systems, inducing DRCR can elevate deletions, inversions and amplifications of a marker gene by 2 orders of magnitude [98,99]. Aside from viral infections, RCR and DRCR can occur with so-called double minute chromosomes and extra-chromosomal circular DNA. Tumor

cells are prone to produce these acentric, circular DNA bodies endogenously, which oftentimes serve to amplify oncogenes [100,101].

Defective Viral Genomes

Deleted or defective KSHV genomes have been detected in primary KS tumors, and helper virus activity has been reported in cell culture. For example, defective KSHV genomes developed during long term culture of BCBL1 suffered an 82 kb deletion from the 5' end [102]. Cells exclusively harboring these defective genomes had more malignant phenotypes than the parental BCBL-1 line but were lytic replication incompetent. The defective genomes still include TR sequences, and they could be packaged into infectious virions in cells containing both intact and defective genomes [102]. Similar genomic deletions were detectable by PCR in a few primary KS tumors and other KSHV-infected cell lines [102].

There are known examples of herpesviruses generating defective genomes in animal model systems and in clinical disease [103–105], and specific DNA sequences in herpesvirus genomes can promote their generation [106,107]. While generation of defective viral genomes is common among RNA and DNA viruses, their role in a natural infection is unclear. They can have large deletions or other major mutations that render them defective for independent replication, but their smaller sizes can contribute to a replicative advantage over full-length genomes [108,109]. They typically compete and interfere with the propagation of full-length viral genomes but cannot completely outcompete and eliminate the full-length genomes. Defective viral genomes modulating viral replication is increasingly thought to facilitate persistence of some viruses in vivo [108]. Cooperative networks of mutants, or quasispecies, have been thought to be integral to the life cycle of some viruses in their host organisms [110]. Detection of defective viral genomes has been associated with distinct clinical outcomes in respiratory syncytial virus [111]. In EBV, rearranged viral DNA can been readily detected in hairy oral leukoplakia and Hodgkin's disease [104,105], where rearranged genomes can remain detectable even in the absence of the EBV diagnostic marker EBER [105]. The biological significance of defective EBV genomes is unclear, but intriguingly, there are defective EBV genomes that, instead of interfering, transactivate latent EBV in cell culture [112,113]. Such defective EBV genomes have rearrangements that enhance the expression of EBV transactivator protein ZEBRA [114].

Significance of Repeat Sequence Heterogeneity

The large internal (IR1, IR2 and LANAr) and terminal (TR) repeat sequences can be sites of rapid evolution [115], yet they are rarely sequenced in KSHV genomes because of the presence of long, GC-rich repeats which are largely insurmountable by Sanger and next generation, short-read sequencing. The IR1 and IR2 tandem repeats have up to 84% GC content and can span more than 1 kb, while paired-end short reads can map erroneously to homologous sequences in IR1 or IR2. LANAr is composed of GC-rich trinucleotide repeats of up to 2 kb. It has 3 subdomains of imperfectly repeating strings of the amino acids DED, QED and QEL, including short runs of Qs and Es. The major internal repeats are masked in 65 of 72 KSHV genome sequences published to date. The few studies on KSHV major internal repeats show them to be heterogeneous across individuals in length and sequence [116–119]. They could be overlooked sources of both genetic and functional diversity in KSHV, since they encode DNA elements, transcripts and protein domains essential throughout the viral infection cycle [32,120,121].

Heterogeneity of repeat sequences has been shown to be consequential in other DNA viruses. EBV IR1 was observed to have minor sequence variants and repeat defects in tumor-derived and lab strains more frequently than in saliva-derived and non-tumor strains [122]. Repairing defects in IR1 of the EBV lab strain B-95 resulted in increased production of large LP isoforms and higher transformation efficiency [122,123]. Mutations in repetitive sequences may also disrupt structures important for their functions. For example, GC-rich repeats in EBV origins of lytic replication adopt triple helix DNA structures *in vivo* that when interrupted with point mutations abrogates Origin of Lytic replication (Ori-Lyt)-dependent replication [124]. The KSHV GC-rich tandem units consist of sequences similar to the JC polyomavirus Ori-Lyt pentanucleotide repeat AGGGA [125], which in JC regulates DNA replication and transcription that depends on host cell type [126,127]. Secondary structures adopted by this sequence impedes and facilitates transcription of early and late genes, respectively. Furthermore, mutations that delete or disrupt Ori-Lyt repeat sequences in BK and SV40 polyomaviruses result in viruses unable to fully replicate but which are highly transforming [128,129].

The variation in tandem repeat numbers within repetitive regions adds another level of diversity. Numbers of tandem repeats expand and contract at 10^{-3} to 10^{-6} per generation, a rate

up to 10 orders of magnitude higher than nucleotide substitutions (10^{-9} to 10^{-12}) [115]. Tandem repeats can be important contributors to variation in gene expression and protein function, and there are numerous examples of phenotypic diversity dependent on tandem repeat unit counts (e.g., onset of several human degenerative diseases, regulation of circadian clocks, dog breed morphologies) [115]. Excessive structure-forming repeats can modulate and stall transcription through the formation of RNA-DNA hybrid R-loops, and it can strain replication machinery [130]. Repeats have functional significance during tumor progression as well, since repeat instability is markedly increased in early stage tumor genomes compared to those of matched healthy normal tissues and late stage tumors, and this transient increase is thought to allow for adaptive evolution to maintain fitness while driver mutations have not yet accumulated [131].

Sequencing Challenges and Innovations

Clinical samples with low viral loads pose challenges for accurate sequencing, hence amplifying DNA is a requisite step. Previous studies that assessed KSHV single nucleotide polymorphisms, variants or multi-strain infection in KS tumors, mucosa and plasma may have had bias introduced during sample preparation and PCR [132][25]. For whole genome sequencing, conventional next generation sequencing (NGS) can have error rates 1 – 0.1%, due to PCR misincorporation, PCR chimeras and DNA damage from long, repeated high-temperature incubations [133–136]. Both Sanger sequencing and NGS can miss genomic and minor sequence variants. As well, KSHV has rather high GC content averaging 54%, peaking at 84% at the GC-rich tandem repeats. GC content affects the efficiencies of PCR amplification, sequencing [137] and enrichment specificity.

Augmenting NGS with duplex Unique Molecular Identifiers (dUMI), which are double-stranded oligonucleotides that include a stretch of random base pairs to barcode both ends of individual DNA molecules before PCR amplification, can direct removal of errors due to PCR [138]. During bioinformatic processing, the DNA sequences as they were before all PCR amplification are reconstructed by taking the consensus of reads with identical dUMI. This removes nearly all PCR misincorporation errors since they cannot be found in 100% of all PCR products derived from a template molecule. Point mutations in the original template sequence are left behind, and the error rate is reduced to ~10^{-9}. Incorporating dUMI is especially advantageous when multiple

rounds of bait enrichment are employed [139], which necessitates more rounds of PCR but is often required for analyzing the typically less than 0.1% KSHV DNA present in human clinical specimens. Moreover, since reads are tagged on both ends, dUMI pairing information is used to identify and remove PCR-related recombination, identified by a given dUMI pairing with another dUMI in a minority of sequences. A chimeric read relative to a reference but supported by a majority of dUMI pairings would indicate a real genomic rearrangement breakpoint.

While typical NGS sequencing can identify mutations shared by a substantial proportion of the genomes within a population, it is fundamentally limited in its ability to identify rare variants in heterogeneous mixtures, regardless of read depth [140]. Given that herpesviruses have a lower error rate than RNA viruses [141], NGS errors will correspond to a greater proportion of suspected mutations. The use of dUMI, which reduces the error rate approximately 5 orders of magnitude below that of typical NGS, is therefore essential for accurate low-level variant discovery and confirmation. Finally, since each dUMI tags a unique DNA molecule before PCR, the depth of unique dUMI-consensus reads per base allows accurate quantification of the number of unique DNA templates sequenced per base [142,143].

UMI-consensus sequencing can also be applied to 3rd generation sequencing like Pacific Biosciences single-molecule long read sequencing (PacBio SMRT), making it truly effective for analyzing intra-host diversity of GC-rich KSHV internal repeats. PacBio SMRT is by far more cost-effective and higher throughput than Sanger sequencing and has no GC-content bias unlike NGS [144]. However, with longer DNA templates polymerases during PCR have more opportunity for template switching and indel errors that destroy reading frames in the rest of the read, especially within homopolymer runs. These errors can accumulate to become debilitating, given that PacBio SMRT has lower throughput than NGS. Tagging DNA templates with UMI before PCR amplification allows for error correction by UMI-consensus, resulting in essentially error-free 1.5 – 4kb reads, as was shown in our lab (manuscript in preparation).

Scope of Thesis Work

Upon starting my graduate studies, the genomic variability of KSHV was not well understood, with the exception of K1 and other small segments that together constitute ~4% of the 165-kb

KSHV genome. Since then, many more whole KSHV genomes have been added to the literature, particularly from Africa. Nevertheless, the viral genetic contributions to KS has not been fully addressed, and much remains unknown of the natural history and evolution of KSHV genomes. My thesis focuses on some outstanding questions about possible KSHV genomic diversity *in vivo* in the individual level: Do KSHV whole genome sequences in KS tumors have characteristic mutations that make them different from "circulating" strains that would be found in saliva of the same person? Are there intra-host genetic variants of KSHV, be it from quasispecies, compartmentalization, spontaneous integrations, or defective viral genomes, that can arise in people with advanced KS? Are there regions of underestimated genetic variability in the KSHV genomes outside of K1? If any novel polymorphisms or mutations are found, do they have the potential to explain the variability of KS manifestations, tumor morphology, staging, progression, treatment response, and clinical outcomes, as well as offer insights into KS pathogenesis?

With the novel sequencing methodologies described above, I had set out to explore these questions in detail from clinical samples provided by a cohort of 30 individuals with KS in Uganda enrolled in the HIPPOS study, administered by Dr. Warren Phipps of the Uganda Cancer Institute - Fred Hutchinson Cancer Research Center consortium. The study participants included men and women with advanced KS from Uganda, most of which have HIV co-infections. Importantly, many clinical specimens examined were from matching KS tumor biopsies and oral swabs taken from the same individuals. In the course of my graduate work, I will publish 43 new whole KSHV genomes that were enriched and sequenced from tumors and oral swabs of 22 individuals. I also obtained informative viral genetic data from a total of 65 tumors and 18 orals swabs from 30 individuals, as well as fully sequenced KSHV internal major repeats in matching samples from 16 individuals.

Chapter 3 showcases the utility of dUMI technology when deep sequencing low viral load clinical samples, including oral swabs, from 9 individuals. This study highlights the remarkably little KSHV diversity existing within hosts at the nucleotide sequence level within a sample and between oral swabs and tumors. However, whole genome sequencing and *de novo* assembly revealed tumor-associated KSHV genome aberrations that frequently retain the IR1 region, inactivating mutations of K8.1, and shared viral genome rearrangement breakpoints, which when results in defective viruses, may be indicative of metastatic spread of clonal tumor cells or propagation by helper viruses. Chapter 4 follows up on these observations in a larger cohort of 30 individuals and examines their associations with clinical phenotypes. I found that the IR1 region

read over-coverage, minimally encompassing the K5 and K6 genes, was a significantly common and specific occurrence (10/30 individuals), among all possible KSHV genome regions. Nine unique K8.1-inactivating mutations were found in 8/30 individuals, a significant finding given that no other intra-host mutation recurred in more than one individual in all the other 85 KSHV ORFs. Polymorphisms in K4.2, K11.2 and miR-K10 across individuals were also characterized. In Chapter 5, I found via PacBio SMRT-UMI sequencing that KSHV internal repeat regions have higher variability within hosts compared to the genome-wide variability I reported in Chapter 3. IR1 encoded imperfect repeats more often than IR2 in persons with KS, and polymorphisms that lead to loss of the translation potential of all full-length Kaposin isoforms from IR2 is frequent among KSHV isolates from Africa. Chapter 6 discusses each of the viral genomic aberrations and mutations I observed from the preceding 3 chapters, their biological and clinical implications, and the means of assessing whether such changes could be causal or incidental to tumor formation. To conclude, Chapter 7 summarizes my body of work and future studies that could be undertaken to extend my findings to clinical settings.

CHAPTER 2: Methods

Ethics statement

All participants provided written informed consent. This protocol was approved by the Fred Hutchinson Cancer Research Center Institutional Review Board, the Makerere University School of Medicine Research and Ethics Committee (SOMREC), and the Uganda National Council on Science and Technology (UNCST).

Study Cohort and Specimen collection

Specimens were obtained from participants enrolled in the "HIPPOS" Study, an ongoing prospective cohort study, begun in 2012, of KS patients initiating treatment at the Uganda Cancer Institute (UCI) in Kampala, Uganda (**Table 1**). Participants were eligible for the HIPPOS study if they were >18 years of age with biopsy-proven KS, and naïve to antiretroviral therapy (ART) and chemotherapy at enrollment. Participants attended 12 study visits over a one-year period and received treatment for KS consisting of ART and chemotherapy (a combination of bleomycin and vincristine or paclitaxel) per standard protocols by UCI physicians. At each visit, participants received a detailed physical exam to assess clinical response using the ACTG KS response criteria [145].

Table 1. Study participant characteristics

Number of participants	30
Gender: Male / female	23 / 7
Age in years, median (range)	32 (23 - 78)
ACTG stage:	
Tumor extent (T1)	26
Immune Status (I1) (CD4 <200)	13
Systemic symptoms (S1)	21
HIV positive individuals:	25
Median CD4 T-cell cells/mm^3 (IQR)	215 (97, 385)
Median plasma HIV RNA, copies/ml (IQR)	2.1 x 10^5 (1.2 - 4.1x10^5)

Participants provided plasma samples at each visit for KSHV, CD4 and HIV viral load testing, and in addition, provided up to 12 biopsies of KS lesions before, during, and after KS treatment. KS tumor biopsies were obtained using 4mm punch biopsy tools after cleaning the skin with alcohol, and either snap-frozen at the clinic site and stored in liquid nitrogen (LN2) or placed in

RNAlater (Sigma-Aldrich, Cat. # R0901) and stored at -80^{0}C. Study clinicians collected swabs of the oral mucosa at each study visit and participants self-collected oral swabs at home for 1 week after the visit after education on the sample collection technique, as previously validated by our group in Uganda [146]. Briefly, a Dacron swab is inserted into the mouth and vigorously rubbed along the buccal mucosa, gums, and hard palate. The swab is then placed in 1 mL of filter-sterilized digestion buffer [147] and stored at ambient temperature [148] before being placed at -20^{o}C for storage.

DNA preparation

DNA was extracted from 300µL homogenized tumor lysates using the AllPrep DNA/RNA Mini Kit (QIAGEN, Cat. # 80204) and eluted into 100µL EB Buffer. For oral swab specimens, DNA was extracted from the swab tip eluate using the QIAamp Mini Kit (Qiagen, Cat. # 51304) following the manufacturer's protocol. Purification of DNA from saliva stabilized in RNAprotect Saliva Reagent (Qiagen) was performed following the manufacturer's protocol with the following modifications: There was no initial pelleting or PBS wash, 20 µL proteinase K was used per 200 µL specimen, and DNA was eluted in 50 µL water. DNA was quantified using a NanoDrop 1000 Spectrophotometer (ThermoScientific).

PCR

All PCR reactions for genetic marker screening and breakpoint junction confirmation were set up in a PCR-clean room, except for the addition of control templates. PCR was conducted using the PrimeSTAR GXL kit (Takara, Cat. # R050B) with ThermaStop (Thermagenix) added. Cycling conditions were: 98°C for 2 mins; 35 cycles of 98˚C for 10 secs, 60-65˚C (depending on primer) for 15 secs, 68˚C for 1min/kb; 68˚C for 3 mins and then hold at 4˚C. All primer sequences used throughout are listed in **Table 2**.

Copy number quantification

KSHV DNA copy number across the genome were quantified by droplet digital PCR (ddPCR) using the QX200 Droplet Generator and Reader (Bio-Rad), with ddPCR SuperMix for Probes (No dUTP) (Bio-Rad, Cat. # 186-3024). Primers and probes (**Table 2**) were designed to detect segments of 4 KSHV- unique genes K2/vIL-6, ORF16/vBCL-2, ORF50/RTA and ORF73/LANA,

Table 2. Primer list

PRIMERS	SEQUENCE	NOTES
For PCR confirmation of breakpoint junctions and other polymorphisms		
K8.1intron_innerF	GTCTGGAGAAGATGCAATAGATGAAT	
K8.1intron_innerR	AAGTATAAGGACAAGCCCCAGCAAT	
K8.1intron_outerF	AATGTCTCCGTATCTGTTGAAGATA	
K8.1intron_outerR	GGCATCGTGGAACGCACAGGTAA	
U030C-K15gapF	GAAAGGCATGAATGGGAAACGCAAA	
U030C-K15gapR	AACATAAGTGTTTACTACCAACTTT	
U008B-IR2int-F	GGACCTAGCATCCCGCTCCCATT	Breakpoint junction PCR of 14.8 kb IR1-region insertion into IR2
U008B-K3int-R	CAGATATGTGGCGTCGTATACAACA	Breakpoint junction PCR of 14.8 kb IR1-region insertion into IR2
K3-F	GAGTTTGGTATCAACCGCAACTA	
ORF19internal-F	CCAAGAACCATAAAGTACGCTCTAT	
ORF19internal-R	ATGGACAGAGGCGGGAAGATTCT	
003CgapTR-R	GGGCTAGGCCACGCCTACTTT	
003CgapTR-F	TATTTTACCCCCCTCATATTCATCT	
ORF11int-F	GCAAAGTCACAGCCATCGTGTCAA	
ORF11-R	CTACAGTCACTATGTCTTTGTGTGT	
ORF19int-R	GTGTCCTTCATCTTCCCTGGCATA	
ORF18-F_alt2	TTTCGGTCTAAATGTGTACCTGTTAA	
ORF4-F	GTCGCCCTAACGTGTGGACTGAT	
K1out-R	ATGTCCAACAAGGTCTGAAAGACA	
ORF25in-F	TTGCTCTACGACCATTCAATACCA	
ORF25in-R	TACACCACCTTGTTTCCGAGCTT	
K12in-R	GGATAGAGGCTTAACGGTGTTTGT	
IR2-R	GGTGCCCAAATTGCTCAATAATGAA	
K8.1-F1	GCTGTCTTTGTAGGCATTAGAAGA	
K8.1-F2	CTATGTTTCCTAGCGGCGAGATT	
K8.1-F3	GGAGGCACATAAGGAACAGATTAT	
K8.1-R1	ACCGCTAAACCGCCTCCTGGT	No PCR product with other K8.1 primers
K8.1-R2	ACGAAGAGGAGCCTCGACAACA	
K8.1-R3	GTACCACTCTTTATCATGTGAACA	
IRcomplexF	ATCCCTAGAACTCCCAAGCTGATT	
K5-F	GCGTCACGTCACATATCTCTGT	
K6-F	GCGTCATCACTAGTTATGAGAGAA	
K6-R	CATATAAGGAACTCGGCGTTACAT	
ORF9in-F	GTTATGGATCTTCTGCTACGGTT	
ORF9in-R	AATAACACAGTAGTTTCCCACCTT	
ORF8in-F	ATTATCGTTATAGCAATCATCCTGA	

ORF10in-R	CCAAATGCCGATGGGTTCTGAA	
ORF9in-F2	ACGACAGAATCCCCTACGTGTT	
ORF10in-R2	TATGTGCTCGGACAGTCACGATT	
TR-F	GAGCCCTGGACACTAMGTGAACA	no PCR product with TR-R, or any other primers
TR-F_alt1	AGATAAGATGAAGAGTATCCTGGAA	no PCR product with TR-R, or any other primers
TRout_alleleP_F	CCGACACAAGGCTCATAAATATTC	no PCR product with TR-R, or any other primers
TR-F_alt3	GCGGCAGCGGAGCGCGAGC	no PCR product with TR-R, or any other primers
TR-F_alt4	TCCTCAGTGCTTGCTACGTGGA	no PCR product with TR-R, or any other primers
TR-R	GGCTCCACGTAGCAAGCACTGA	
TR-R_alt1	GCGGGAGAAAACGAAAGCAAGC	no PCR product with TR-F, or any other primers
TR-R_alt2	CGCATCCCCCCCCCTATTTTAC	no PCR product with TR-F, or any other primers
TR-R_alt3	GCGCTCGCGCTCCGCTGC	no PCR product with TR-F, or any other primers
K4.1inF1	TGGACAGGGAATTCTCGCAACAA	
K4.1inF2	GCTCAATCACGGCCACCCAGA	no PCR product with IRcomplexR1 or R2
vIRF2r_F1	CCGTTAAATCAAAGTGGGTCTTCA	
vIRF2r_R1	TGTAGGGAGGGATATGCACAGTT	
vIRF2r_F2	CGAAATAATACACTTCCACCACTA	
vIRF2r_R2	TCATCTGGTCAGTCATCGAGCTT	
For ddPCR		
vIL6-F	TTGGATGCTATGGGTGATCGA	
vIL6-R	TCAGTATCGTTGATGGCTGGTAG	
vIL6-probe (HEX)	CGTACCGGCATCTGCAAGGGTATTCTAGA	
vBCL2-F	GCCTGTGGATTAAACGAACCTG	
vBCL2-R	GTCTCGCATTAAGCCTGTGATG	
vBCL-2_probe (FAM)	CCTGTACCATCCTTTGCTCAGCCCTATTAAGC	
RTA-F	GACGAACTGAAGGCCCAACTCTA	
RTA-R	ATGCACACATCTTCCACCACTCTA	
RTA-probe (HEX)	CGCATACGAAACAATCTACGATCCCAGTGAC	
LANA-F	CCAGGAAGTCCCACAGTG TTC	
LANA-R	GCCACCGGTAAA GTAGGACTAGAC	
LANA-probe (FAM)	CATCCGGGCTGCCAGCATTTG	
T.07-K12-F	TCCCCCACCGAGYGCTT	
T.07-K12-R	GCACGCGGTGTCAACCA	
For Illumina library amplification		
mws13	AATGATACGGCGACCACCGAG	forward primer

mws20	GTGACTGGAGTTCAGACGTGTGC	reverse primer for pre-enrichment amplification
mws21	CAAGCAGAAGACGGCATACGAGATXXXXXX GTGACTGGAGTTCAGACGTGTGC	reverse primer with sample index for final amplification
mws15	CAAGCAGAAGACGGCATACGAGAT	reverse primer without sample index for pre-enrichment and final amplification
For SMRT-UMI PCR		
UMI-tagging		
illu-IR1-R_alt1	AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGAC(N:25252525)(N)(N)(N)(N)(N)(N)(N)GGTTGTGTGCTTGTGACTGATA CAA	(N) = random bases (hand-mixed), Integrated DNA Technologies
illu-IR2-R	AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGAC(N:25252525)(N)(N)(N)(N)(N)(N)(N)GGTGCCCAAATTGCTCAATAAT GAA	(N) = random bases (hand-mixed), Integrated DNA Technologies
illu-LANAr-R	AATGATACGGCGACCACCGAGATCTACACTC TTTCCCTACACGAC(N:25252525)(N)(N)(N)(N)(N)(N)(N)GAAAATAATCAGGCTGGCGAG GAT	(N) = random bases (hand-mixed), Integrated DNA Technologies
1st Round		
IR1-F	TGGCAAGGTGACTGAAAAGGTCATA	forward
IR2-F_alt1	GGGGACAACACTAATCGCCAACA	forward
LANAr-F_alt0	ACTTCCATTTCGTCCTCGGATGA	forward
illu_F1	AATGATACGGCGACCACCGA	reverse
2nd Round		
IR1-F_alt1	GGTCCCATTTCCCACGGTCCAAA	forward
IR2-F_alt2	AACACCGTYAAGCCTCTATCCAT	forward
LANAr-F	CCCGTGCAAGATTATGGGCTCTT	forward
illu-adapt	GATCTACACTCTTTCCCTACACG	reverse, contains adapter sequences
3rd Round		
IR1-F_alt1	GGTCCCATTTCCCACGGTCCAAA	forward
IR2-F_alt1	GGGGACAACACTAATCGCCAACA	forward
LANAr-F	CCCGTGCAAGATTATGGGCTCTT	forward

with the copy number reported as the average of the 4 measures. 420 ng BCBL-1 cell line DNA diluted 1:10,000 (~475 genome copies) was used as positive control, 1 ng human genomic DNA (Bioline, Cat. # BIO-35025) as negative control, and water as no template control. Cycling conditions were: 95°C for 10 mins; 40 cycles of: 94°C for 30 secs, 56°C for 30 secs, 60°C for 1 min; one cycle at 98°C 10 for mins, and then hold at 12°C. The KSHV on-target percent was calculated using the copy number quantification by ddPCR normalized to the total nucleic acid concentration.

UMI-addition and Illumina library preparation

To obtain ~500-bp DNA fragments, 10-20 ng/µL of DNA extract in 100 µL chilled TLE buffer (10mM Tris, pH8.0, 0.1mM EDTA) was sheared using a Bioruptor (Diagenode) on high power for up to 15 min, or Covaris S2 Sonicator set to duty cycle 5%, 200 cycles per burst for a total of 30 seconds. Fragment sizes were assessed on 1.5% agarose gels. Sheared DNA was purified using 1.2X volume of Agencourt AMPure XP Beads (Beckman Coulter Cat. # A63880) and eluted in 50 µL water. Library preparation (end repair, A-tailing and adapter-ligation) was performed using the KAPA HyperPrep Library Preparation Kit (Cat. # KR0961/KK8503). For dUMI sequencing, double-stranded DNA adapters containing a random 12-bp dUMI sequence and a defined 5-bp spacer sequence were added to Illumina TruSeq adaptor sequences [138] (**Fig 1**). Subsequently, DNA was bead-purified with 1X volume of beads and eluted in 50 µL water. Samples were then divided into aliquots of up to a maximum of 240 ng, prior to the next PCR step, to be pooled afterwards. For dUMI sequencing DNA libraries were subjected to pre-enrichment amplification with primers mws13 & mws20 (**Table 1**, **Fig 1**). For libraries that were sequenced without dUMI, KAPA Hyperprep adapter primers (Cat# KR1317) were used. PCR conditions were: 95°C 4 mins; 5-8 cycles of 98°C 20 sec, 60°C 45 sec, 72°C 45 sec; 72°C 3 mins, 4°C hold. PCR products were then bead-purified as above with 1.2X volume beads and elution in 100 µL water, quantified using a Nanodrop spectrophotometer, and their sizes assessed using a Bio-analyzer (Agilent DNA 7500) or Qiaxcel (QIAGEN AM420).

Library enrichment & sequencing

Biotinylated RNA baits used for KSHV DNA enrichment from Illumina libraries were those designed in [149] and were obtained from Agilent, Inc. (Santa Clara, CA). The design was a 120-

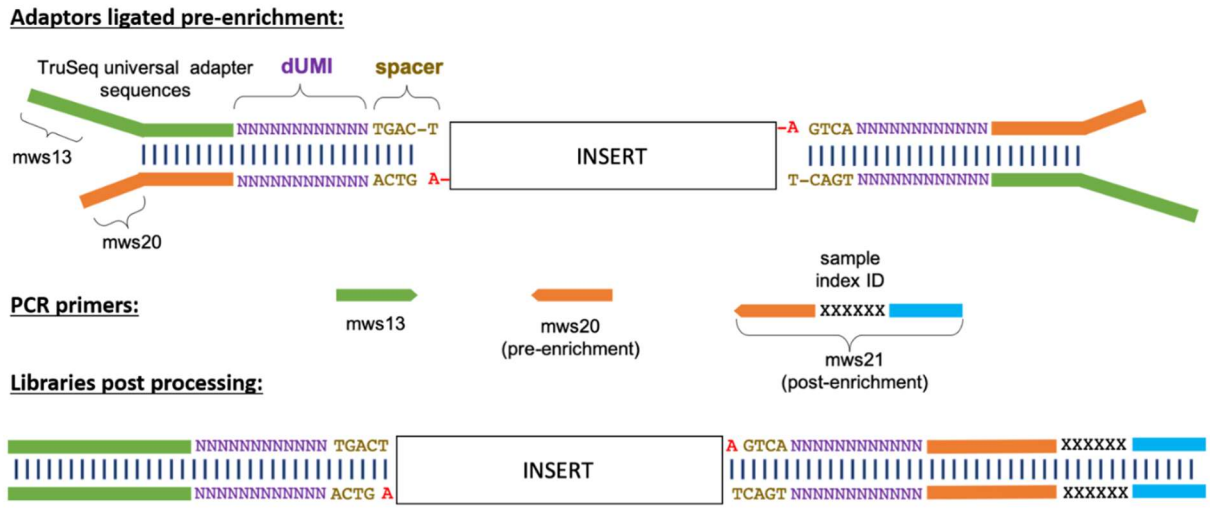


Figure 1. dUMI-adaptors and primers for duplex UMI sequencing

During library preparation, sheared DNA fragments were A-tailed and ligated with forked, double-stranded oligonucleotides containing Illumina TruSeq universal adaptor sequences, 12-random base pairs (the dUMI) and spacer sequences. DNA libraries were then PCR amplified before enrichment with primers mws13 and mws20, which bind to Illumina Truseq adaptors. Primer mws21, containing sample index ID for multiplex sequencing, was used instead of mws20 for PCR following enrichment. DNA library fragments produced post processing are shown below.

bp, 12X tiling of the genome of KSHV isolate GK18 (Genbank ID: AF148805.2). The diversity of the bait library was further increased by including K1, ORF75, K15, ORF26 and TR sequences of strains JSC-1 (Genbank ID: GQ994935.1), DG1 (Genbank ID: JQ619843.1), BC-1 (Genbank ID: U75698.1), BCBL-1 (Genbank ID: HQ404500.1), Sau3A (Genbank ID: U93872.2), and of all Western and African isolates in [66,76] (Genbank ID: AF130259, AF130266, AF130267, AF130281, AF130305, AF133039, AF133040, AF133043, AF133044, AF151687, AF171057, AF178780, AF178810, AF220292, AF220293, AY329032, KT271453-KT271468).

Target enrichment was performed using SureSelect Target Enrichment Kit v1.7 (Agilent) with all suggested volumes reduced by half. DNA hybridized to biotinylated-RNA baits was captured with streptavidin beads (Dynabeads myOne Streptavidin T1, Invitrogen) and resuspended in 20µL water. The DNA-streptavidin bead mixture was split into two and used directly in post-enrichment PCR amplification. For dUMI libraries, mws13 and mws21 were used, the latter of which includes a sample index sequence (**Table 1**, **Fig 1**). When omitting dUMI, PCR primers were mws13 and

mws15. The PCR cycle number ranged from 10-16, with products monitored every 2 to 3 cycles on a TapeStation (Agilent) to ensure correct fragment sizes (~500bp). When over-amplification resulted in library fragment sizes larger than expected dur to heteroduplex formation, a single "reconditioning" PCR cycle with fresh reagents was done, which had the effect of reducing fragment size [150]. PCR products were cleaned using 1.2X volume AMPure XP beads and the eluted DNA library was sequenced using Illumina HiSeq 2500 with 100-bp paired end (PE) reads or Illumina HiSeqX with 150-bp paired-end reads. For some tumor samples with low KSHV copy numbers and all oral swab samples, a second library enrichment was performed.

De novo assembly of sample-consensus genomes

Initially, a sample-consensus KSHV genome (**Fig 2**) was generated *de novo* from paired-end (PE) reads of each sample using custom scripts as follows (**Fig 3**, https://github.com/MullinsLab/HHV8-assembly-SPAdes). First, the first 17-bp from read ends were trimmed to remove dUMI sequences if present. Next, reads were subjected to windowed quality filtering using *sickle pe* [151] with a quality threshold of 30 and a window size 10% of read length. Filtered reads were aligned to a human genome (GRCh38 p12, GenBank GCA_000001405.27) using *bwa mem* [152]. Unmapped reads were used as input into the de novo assembler SPAdes v3.11.1 [153], with the setting *-k 21,35,55,71,81* for 100 PE reads and *-k 21,35,55,71,81,127* for 150 PE reads. This often yielded 3 to 4 scaffolds that together encompassed the entire 131-kb unique sequence regions of KSHV, bounded by the major repeat regions: IR1, IR2, LANAr and the TR. Next, all scaffolds over 500 bp were aligned using *bwa mem* to the genome of reference KSHV isolate GK18. From the aligned scaffolds a draft genome was generated in Geneious (Biomatters, Ltd) with manual correction as needed. To finish the assembly, GapFiller v1.1 [154] was used, setting *bwa* as the aligner and filtered paired-end reads as the input library. The genomes were annotated in Geneious based on the GK18 reference, also adding the annotation for long non-coding RNA T1.4 based on [155]. All K8 and some K1 and K15 annotations were transferred from the most homologous sequence among KSHV genomes GK18, Japan1 (Genbank LC200589) and ZM095 (Genbank KT271456). The major repeat regions were masked by the insertion of N residues since they were poorly resolved by assembly of short reads that can map to multiple locations within the repeat regions.

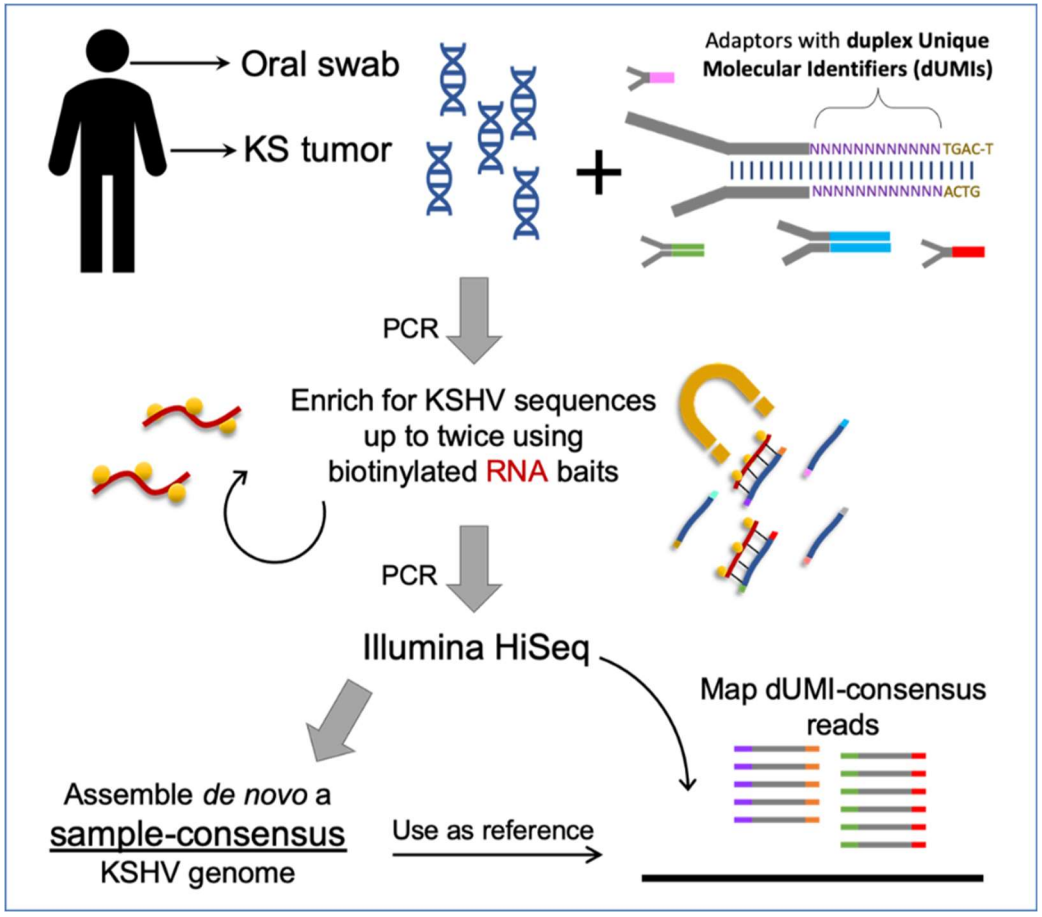


Figure 2. Workflow for analyzing intra-host KSHV genome diversity from clinical samples

*Each study participant contributed multiple KS tumors and oral swabs, but only those samples with the highest viral loads are reported here. Sequencing libraries were prepared from DNA extracts from each sample with adaptors containing duplex Unique Molecular Identifiers (dUMIs, see **Fig 1**). Adaptor-labelled DNA libraries were enriched using biotinylated RNA baits homologous to KSHV sequences. Captured DNA was PCR-amplified to levels sufficient for Illumina HiSeq sequencing. For most samples, libraries were subjected to a second round of enrichment and PCR amplification. Upon sequencing, whole KSHV genomes were first assembled de novo from each sample without the use of dUMIs. The sample-specific genomes generated (sample-consensus) were then used as reference to map the consensus of reads with identical dUMI-tags (i.e., dUMI-consensus reads).*

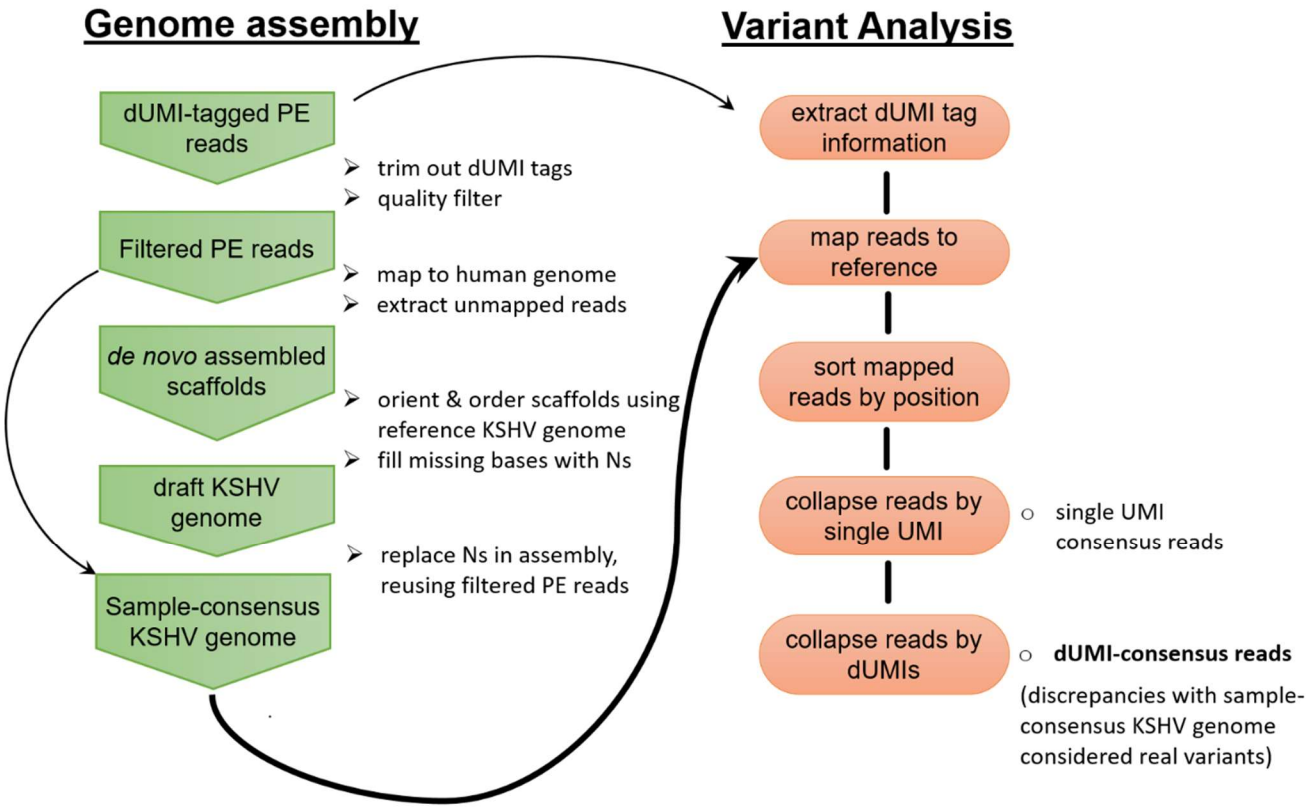


Figure 3. Workflow of genome assembly and variant analysis

A series of python and bash scripts were used to process Illumina dUMI library sequences (described in the text). KSHV genomes were first assembled de novo from sequence reads of each sample, before being used as reference for mapping their respective dUMI-consensus reads (adapted from [138]*, see details below). Discrepancies between the sample-consensus genome and mapped dUMI-consensus reads were taken to be real intra-sample variants.*

Variant identification from dUMI-consensus reads

PE reads, including their dUMI sequence tags, were mapped using *bwa* to sample-consensus genomes (**Fig 2**) using a Makefile adapted from [138] (https://github.com/MullinsLab/ Duplex-Sequencing). Briefly, all reads mapping to the same genomic position were collapsed by single strand UMIs (sUMI) to make sUMI-consensus reads (**Fig 3**). Complementary UMI tags from opposing strands were matched to create dUMI-consensus reads, thus removing nearly all PCR polymerase misincorporation and chimera artifacts. Nine bases from both read ends were then trimmed to minimize read end artifacts. Discrepancies between mapped dUMI-consensus reads

and the sample-consensus genomes were manually inspected in Geneious and misalignments around homopolymer runs were corrected.

All genome and subgenome sequence alignments were generated using MAFFT [156] [algorithm FFT-NS-i x1000, scoring matrix 1PAM/k=2], and all phylogenetic trees were inferred using RAxML [157] (-f d, GTR gamma, N=100 starting trees), using a representative KSHV genome from each individual. The NeighborNet phylogenetic network was generated using SplitsTree5, excluding gap sites [158]. Consensus genome sequences were deposited in GenBank (Accession numbers MT510648 - MT510670, MT936340, with additional sequences to be submitted) annotated with the genomic rearrangements, when present.

Structural variation and integration breakpoint detection

To identify potential chimeric KSHV-human sequences, the KSHV reference genome GK18 was appended as an extra chromosome to the human genome reference GRCh38 p12. PE reads from each sample library were mapped to the appended human genome reference. Alignment files were generated using SpeedSeq [38]. All potential rearrangement breakpoint locations, along with the number of supporting split and discordant paired reads, were tabulated using LUMPY [39] and imported as VCF annotations of genome GK18 into Geneious Prime. Supporting reads of breakpoint positions in which the corresponding breakpoint had been identified in LUMPY were manually inspected in the Geneious Prime alignment viewer. For cases in which LUMPY did not output a breakpoint location for a region with an abrupt change in read coverage, split reads were identified using the Geneious Prime mapping tool. Human chromosomes linked to KSHV sequences were taken to be putative integration sites. Additionally, each library was searched using local BLASTN against both human and KSHV sequences and then using the Perl script SummonChimera [159] to extract coordinates of potential integration sites.

Potential breakpoints detected in LUMPY were accepted for analysis using the following criteria. Breakpoints had to have at least 6 supporting split reads for libraries that did not use dUMIs and 3 for dUMI libraries, and plus >1 supporting PE reads. For my earlier sequencing dataset that utilized duplex unique molecular identifiers [31], the threshold was lowered to 3 supporting split reads, the lowest that were validated by PCR. Breakpoints had to be at least 1 kb apart, approximately twice the library insert sizes. Identical breakpoint locations, sometimes found in tumors from the same person, were counted as one. Breakpoints of rearrangements wholly inside KSHV repeat regions were filtered out. The original coordinates of remaining breakpoints,

along with the number of supporting reads and other output information from LUMPY and Geneious Prime, were exported as a table .tsv file for analysis of the DNA sequence context in R (**Table 3**).

Analysis of DNA structures and motifs

The following were implemented with custom scripts in bash and in R with the Biostrings package [160] (https://github.com/MullinsLab/). Five kilobase (kb) windows with position 1 centered at the observed breakpoints (**Table 3**) were sliced from KSHV genome sequences. A randomized control dataset was generated from 5 kb windows centering on an equivalent number of random positions along the KSHV reference GK18 genome, minus the TR sequences. A permuted control dataset was generated by shuffling the sequences of the same 5 kb slices of the observed breakpoints. For each window, probabilities of cruciform, denaturation and Z-DNA formation at every base position were computed using SIST (Stress-Induced Structural Transitions), set to the algorithm that considers competition among the three possibilities [161]. Potential triplex and G-quadruplex formation was analyzed separately using R packages *triplex* [162] and *pqsfinder* [163], respectively. Both algorithms output a score at each position, which was normalized to a maximum of 300 [164] and 50, respectively. Probabilities and relative scores of only the central 1 kb were plotted for each observed breakpoint. Means of probabilities and relative scores were taken from equivalent positions across all breakpoint windows to create a plot of averages.

Summed probabilities or relative scores were obtained for each predicted non-B DNA stretch by adding their values at all positions $\pm$200 bp from the breakpoint. The shortest distance to each predicted non-B DNA site was taken from the minimum of 0 or the distance in either direction to the nearest position with non-B DNA probabilities or scores >0.20. The summed values and distances were also calculated for the set of random breakpoints, and the differences between means of observed and random sets were tested using the Student's T-test assuming unequal variance. To estimate the probabilities of attaining G-quadruplex structures higher summed scores and shortest distances equal to or less than the observed 31-breakpoint dataset, G-quadruplex summed scores and shortest distances were calculated from 1000 simulations of 31 randomly sampled points along the GK18 reference genome.

Table 3. Breakpoints identified in KSHV genomes

Count	Source	GenBank	Breakpoint coordinate	Number of split reads	Number of paired-end reads	Read Support	SV length	Reference	Random breakpoint dataset (GK18 coordinates)
1	BCBL1	MT936340	25,816	N/A	N/A	N/A	N/A	[78]	5,670
2	BCBL1	MT936340	34,943	N/A	N/A	N/A	N/A	[78]	106,870
3	BCBL1	MT936340	82,922	N/A	N/A	N/A	N/A	[102]	71,578
4	U003-C	MT510648	54,270	N/A	N/A	N/A	N/A	Chapter 4, Geneious Prime	1,828
5	U003-C	MT510648	77,054	163		163		Chapter 3, LUMPY	131,143
6	U004-D	MT510665	3,129	713	2808	3521	-134644	Chapter 3, LUMPY & Geneious Prime	22,956
7	U008-B	MT510656	19,691	203	386	589	100522	Chapter 3, LUMPY & PCR, sequencing	7,134
8	U008-B	MT510656	34,762	3	641	644	-85370	Chapter 3, LUMPY & PCR, sequencing	5,418
9	U008-B	MT510656	120,133	N/A	N/A	N/A	N/A	Geneious Prime	114,328
10	U020-B	MT510666	47,218	209	1336	1545	-91032	Chapter 3, LUMPY & Geneious Prime	47,221
11	U020-B	MT510666	17,135	163		163	120619	Chapter 3, LUMPY & PCR, sequencing	114,698
12	U020-B	MT510666	33,614	279		279	-104247	Chapter 3, LUMPY & PCR, sequencing	55,652
13	U034-B	MT510659	16,712	12	71	83	41549	Chapter 4, LUMPY & Geneious Prime	118,061
14	U048-E	U048-D	24,207	11	397	408		Chapter 4, LUMPY & PCR, sequencing	81,221
15	U048-E	U048-E	24,550	258		258		Chapter 4, LUMPY & PCR, sequencing	63,068
16	U048-E	U048-E	27,832	258		258		Chapter 4, LUMPY & PCR, sequencing	67,405
17	U099-D	U099-D	25,019	1064	375	1439	111925	Chapter 4, LUMPY	51,232
18	U099-D	U099-D	25,545	49	0	49	111367	Chapter 4, LUMPY	35,391
19	U099-D	U099-D	29,731	1744	0	1744	-107394	Chapter 4, LUMPY	27,081
20	U099-D	U099-D	81,699	5	14	19	7718	Chapter 4, LUMPY	2,453
21	U108-B	U108-B	24,340	N/A	N/A	N/A	N/A	Chapter 4, Geneious Prime	135,915
22	U108-B	U108-B	29,308	544	693	1237	-107523	Chapter 4, LUMPY & Geneious Prime	48,675
23	U108-B	U108-B	32,137	8	13	21	-10798	Chapter 4, LUMPY	57,360
24	U108-B	U108-B	120,972	556	546	1102		Chapter 4, LUMPY & Geneious Prime	13,034
25	U156-B	U156-B	12,695	8	1	9	-100965	Chapter 4, LUMPY	104,564
26	U156-B	U156-B	96,697	10	1	11		Chapter 4, LUMPY	93,964
27	U156-B	U156-D	23,052	6628	3660	10288	114878	Chapter 4, LUMPY & Geneious Prime	54,990
28	U156-B	U156-D	29,806	21	0	21	-107695	Chapter 4, LUMPY & Geneious Prime	71,235
29	U210-B	U210-B	22,208	2976	1584	4560	115458	Chapter 4, LUMPY & Geneious Prime	126,605
30	U210-B	U210-B	34,137	4795	2708	7503	-103203	Chapter 4, LUMPY & Geneious Prime	34,326
31	U215-D	U215-D	25,889	8	1	9	71772	Chapter 4, LUMPY	72,668

Single-molecule-real-time unique-molecular-identifier (SMRT-UMI) sequencing of major internal repeat regions of KSHV

UMI-tagged, single-stranded copies of IR1, IR2 and LANAr were simultaneously generated from DNA extracts of individual tumor biopsy and oral swabs in 25 µL linear extension reactions. UMI-tagged reverse primers were used for IR1, IR2 and LANAr and were were composed of conserved sequences downstream of the repeat regions, a random 8-nt UMI and 2 primer binding sites for nested PCR. All primers were synthesized by IDT (https://www.idtdna.com/pages) with "Hand-Mix" for the 8-nt UMI (**Table 2**). The PrimeSTAR GXL kit (Takara Cat. # R050B) was used for all PCR reactions. Sample DNA corresponding to a maximum of 1,500 KSHV genome copies, estimated from ddPCR from 4 genomic regions [165], was used per reaction. 16 µL was used for oral swab DNA extracts regardless of measured copy number. Conditions for the single-strand synthesis reaction were 98°C for 4 mins; 60˚C for 20 secs, 68˚C for 6 mins, then hold at 4˚C. To remove unincorporated primers the reaction was cleaned with 0.7X volume of AMpure XP magnetic beads then eluted in 20 µL water.

Next, IR1, IR2 and LANAr were amplified separately by nested PCR from the single stranded templates. Three rounds of PCR, including 2 nested rounds, were performed, with each round not exceeding 22 cycles to limit heteroduplex formation. ThermaStop (Thermagenix, Inc.) was added in all rounds to increase specificity of PCR primer binding [166]. Templates for 1st round PCR were limited to an estimated 100 - 200 copies per 25 µL reaction. Primer pairs corresponded to conserved sequences upstream of the repeats and the UMI-tagged primers. Conditions for 1st round PCRs were: 95˚C for 4 mins, 22 cycles of 98˚C for 20 secs, 61˚C for 15 secs and 68˚C for 5 mins, and then hold at 4˚C. 2 µL of 1st round products was used as template for the 25 µL second round PCR with inner primers. Conditions for 2nd round PCRs were: 95˚C for 4 mins; 15 cycles of 98˚C for 20 secs, 63˚C for 15 secs, and 68˚C for 4 mins, and then hold at 4˚C. To remove unincorporated primers, 2nd round PCR products were purified using 0.7X volume AMpure XP magnetic beads and eluted in 20 µL water. 10 µL of the eluate was used in a 3rd round of PCR to append sample IDs for multiplexed sequencing. Conditions for 3rd round PCRs were: 95˚C for 4 mins; 10 cycles of 98˚C for 20 secs, 55˚C for 15 secs and 65˚C for 4 mins, and then hold at 4˚C. Final PCR products were separated on a 0.8% agarose gel and bands extracted using NucleoSpin Clean-up Columns (Machery-Nagel, Cat #740609). If PCR products were not visible on the gel, amplification was repeated using the 2nd or 3rd round products as template.

The concentrations of purified PCR products were determined using the Qubit dsDNA HS

Assay Kit (ThermoFisher). PCR products were combined in equimolar amounts into pools of 5,000-6,000 UMI-tagged templates. Each pool was purified with 0.7X volume AMpure XP beads and quantitated. Library preparation was performed on each pool using the SMRTbell Express Template Prep Kit 2.0 or SMRTbell Template Prep Kit (Pacific Biosciences, Inc.). Each library received a different barcoded adapter from the Barcoded Overhang Adapter Kit–8A/8B (Pacific Biosciences, Inc.). Completed libraries were sequenced on either the Sequel I or Sequel II instruments (Pacific Biosciences, Inc.).

Demultiplexing and UMI consensus processing

The bioinformatic pipeline used to demultiplex SMRT-UMI-consensus reads from PacBio sequencing is available at https://github.com/MurrellGroup/PORPID.jl. Briefly, each read corresponded to one circular consensus sequence (CCS) produced from one sequencing well. All CCS reads with the same UMI, i.e., a "UMI family", were collapsed to form one UMI-consensus read, accepting a minimum agreement of 0.5. This removes the PCR errors that accumulate during amplification after the synthesis of the DNA copy of the repeat region Since each UMI barcodes one DNA molecule template before PCR amplification, the number of UMIs represents the number of original DNA molecules sequenced per sample. In turn, each UMI family member is composed of several CCS reads, referred to as read depth, per template DNA molecule. The minimum threshold of CCS to accept a UMI was set at 5.

Statistical Analysis

All clinical data, individual tumor characteristics and viral genetic polymorphisms observed were classified into binary categories. For intrahost mutations, individuals were classified into whether they have at least one tumor with the observed mutation. A chi-squared test was used to compare the odds ratio of genetic markers with clinical or tumor data. Cox regression was used to assess potential differences in survival. No multiple tests corrections were applied.

CHAPTER 3: Changes in KSHV genomes in Ugandan adults with Kaposi sarcoma

Results published in: Santiago JC, Goldman JD, Zhao H, Pankow AP, Okuku F, Schmitt MW, et al. ***Intra-host changes in Kaposi sarcoma-associated herpesvirus genomes in Ugandan adults with Kaposi sarcoma****.* PLoS Pathog*. 2021;17. PMID: 33465147*

Introduction

As noted in the first Chapter of this thesis, KSHV, is the etiologic agent of KS yet only a small fraction of KSHV infections progress to KS. The factors contributing to this progression are poorly understood often associated with HIV infection and immunosuppression [6], but others factors, including KSHV genome variation, may contribute to the differential outcomes of KSHV infection. Studies of other human oncogenic viruses have revealed that viral genetic variation or *de novo* mutations may be important to their pathogenicity, yet whether KSHV genetic variation can similarly influence KS pathogenesis or manifestation is unknown.

Studying whole KSHV genomes provides a far more comprehensive picture of diversity than the variable regions alone (most often the K1 gene), which until recently has comprised the majority of sequences available. For example, KSHV genomes have signatures of ancient recombination [59,77], resulting in mosaic genomes that could complicate disease risk association solely with K1 subtypes. This Chapter describes our initial efforts to evaluate the possible impact of KSHV strain variation on KS development and clinical outcomes.

Whether KSHV infection commonly displays intra-host diversity is also unclear. Some studies examining virus in different anatomic sites or multiple isolates from an individual have reported detection of KSHV quasispecies, multi-strain infections [167–172] and intra-host KSHV viral evolution [170,173,174], while some studies of individuals with AIDS-associated KS have only reported a single persisting strain [132,175–177]. Apparent recombination events in the evolutionary history of KSHV [59,77,175,178,179] do suggest that co-infection of divergent KSHV strains occur, at least sporadically. However, limitations of commonly employed PCR technologies can undermine reliable interpretation of observed intra-host KSHV variation.

To much more accurately assess minor intra-host KSHV sequence variation as well as tumor specific changes than has been done previously, we determined viral genome sequences in distinct anatomic compartments (oral and tumor sites), using a highly sensitive short-read

sequencing method termed “duplex sequencing” [143]. This method incorporates duplex unique molecular identifiers (dUMI), which are double-stranded strings of random base pairs used to barcode individual DNA molecules before PCR amplification and enrichment [143]. By utilizing dUMI-consensus reads of each DNA molecule in a sample library, PCR-associated errors are reduced to ~10^{-9}, revealing the original sequence variation within a sample [143] (**Fig 2**). In this chapter, we report the results of successfully enriching and duplex sequencing whole KSHV genomes from tumors and oral swabs from 9 Ugandan adults with HIV-associated KS.

Results

Assessment of the dUMI sequencing protocol with a KSHV infected cell line

As part of the optimization of the dUMI-sequencing protocol, KSHV genome sequences (**Fig 4A**) were first obtained from an early passage of BCBL-1, a KSHV-infected PEL cell line [79]. BCBL-1 cells were grown as previously described [180]. After DNA extraction, KSHV sequences corresponded to ~0.16% of the total DNA using a ddPCR assay for ORF73 and T0.7-K12, normalized by comparison to the human gene POLG (DNA polymerase subunit gamma). Following a single round of bait capture, the fraction of sequence reads corresponding to KSHV from BCBL-1 DNA extracts (i.e., the “on-target” level), was 15.6%, corresponding to 173-fold enrichment.

Sequencing of the BCBL-1 KSHV genome produced a median coverage of 16,805 reads per base excluding the repeat regions. Collapsing raw reads by identical sUMI to generate sUMI-consensus reads resulted in a median of 2,677 sUMI reads per base. When collapsed further into consensus sequences derived from both strands, a median of 302 dUMI reads per base was obtained that were essentially free of PCR errors (**Table 4; Fig 4B**). Since each dUMI tags a unique DNA molecule before PCR, the number of unique dUMI tags indicates the number of unique viral templates sequenced [142,143]. Using this measure, 302 also approximated the number of KSHV genomes sampled from BCBL-1.

Eighty-one base positions (0.06%) in the BCBL1 consensus KSHV genome had detectable variants in dUMI-consensus reads, and the average frequency of minor variants was 1.35%. No variant exceeded 14% of the total dUMI-consensus reads at any position (**Fig 4C)**. No doubling of read coverage within the 9-kb genomic region previously reported in the BCBL-1-derived KSHV recombinant clone BAC-36 was found in our study [78].

Table 4. Origin and processing results from specimens for KSHV genome analysis

Pt ID	Age	Sex	Plasma HIV RNA (copies/mL)	CD4 T cells /µL	Sample ID	Sample Type	% on-target		Median read coverage	Standard Deviation	Median dUMI-consensus read coverage	Standard Deviation	Genome length (excluding masked repeats)	# positions with variants	Mean frequency of variant base	Genbank Accession
							Pre-enrichment	*Post*-enrichment								
n/a	n/a	n/a	n/a	n/a	BCBL-1	cell line	0.0009%	15.61%	16,805	1,858	302	52.49	132,676	81	1.35%	MT936340
U003	25	M	759,635	45	**U003-C**	**Tumor**	**0.0037%**	**7.15%**	**16,265**	**1,715**	**270**	**30.47**	**131,292**	**21**	**1.80%**	**MT510648**
					U003-o1	*Oral swab*	*0.0000%*	*35.20%*	*49,598*	*26,446*	*7*	*6.22*	*131,102*	*12*	-	*MT510649*
					U003-o2	*Oral swab*	*0.0003%*	*31.60%*	*66,527*	*13,995*	*310*	*68.48*	*131,129*	*218*	*0.51%*	*MT510650*
					U003-o3	*Oral swab*	*0.0000%*	*41.90%*	*65,984*	*21,979*	*47*	*21.31*	*131,143*	*24*	-	*MT510651*
U004	37	M	277,655	85	**U004-C**	**Tumor**	**0.0016%**	**17.20%**	**19,582**	**3,078**	**34**	**7.41**	**131,277**	**12**	**-**	**MT510663**
					U004-D	**Tumor**	**0.0017%**	**75.10%**	**60,340**	**10,253**	**1,291**	**280.27**	**131,237**	**73**	**0.73%**	**MT510665**
					U004-o1	*Oral swab*	*0.0001%*	*18.90%*	*37,794*	*6,280*	*56*	*11.70*	*131,277*	*31*	-	*MT510664*
U007	26	M	91,096	136	**U007-B**	**Tumor**	**0.0006%**	**86.90%**	**75,001**	**19,771**	**1,172**	**268.44**	**131,352**	**156**	**0.17%**	**MT510654**
					U007-o1	*Oral swab*	*0.0001%*	*27.60%*	*43,707*	*6,258*	*206*	*25.54*	*131,126*	*61*	*1.93%*	*MT510655*
U008	56	M	860,937	488	**U008-B**	**Tumor**	**0.0049%**	**16.97%**	**19,309**	**4,168**	**637**	**148.42**	**131,142**	**31**	**0.93%**	**MT510656**
					U008-D	**Tumor**	**0.0040%**	**29.98%**	**26,209**	**4,674**	**195**	**42.95**	**131,102**	***24***	***0.70%***	**MT510657**
					U008-o1	*Oral swab*	*0.0000%*	*23.50%*	*48,339*	*6,869*	*114*	*16.62*	*131,116*	*76*	*2.11%*	*MT510658*
U020	27	M	118,191	370	**U020-B**	**Tumor**	**0.0070%**	**1.47%**	**2,913**	**14,971**	**24**	**196.60**	**131,102**	**10**	**2.40%**	**MT510666**
					U020-C	**Tumor**	**0.0003%**	**34.38%**	**420**	**3,674**	**3**	**2.51**	**131,471**	**30**	**-**	**MT510667**
					U020-o1	*Oral swab*	*0.0000%*	*77.30%*	*73,287*	*22,429*	*80*	*21.66*	*131,115*	*22*	-	*MT510668*
U023	33	F	338,285	191	*U023-o1*	Oral	0.0001%	*2.70%*	19,889	6,683	2	1.30	131,122	*2*	-	*MT510669*
U030	40	M	100,184	70	**U030-C**	**Tumor**	**0.0135%**	**15.68%**	**38,935**	**6,664**	**490**	**62.59**	**131,282**	**17**	**1.08%**	**MT510670**
U032	23	F	587,149	274	**U032-B**	**Tumor**	**0.0003%**	**43.50%**	**69,815**	**9,269**	**890**	**67.12**	**131,266**	**45**	**0.41%**	**MT510652**
					U032-o1	*Oral*	*0.0000%*	*7.70%*	*7,311*	*3,218*	*1*	*0.80*	*130,842*	-	-	*MT510653*
U034	47	F	130,375	237	**U034-B**	**Tumor**	**0.0014%**	**84.40%**	**74,927**	**13,969**	**1,747**	**189.51**	**131,248**	**107**	**0.11%**	**MT510659**
					U034-C	**Tumor**	**0.0013%**	**30.70%**	**17,968**	**2,505**	**133**	**24.36**	**131,088**	**11**	**1.46%**	**MT510660**
					U034-o1	*Oral swab*	*0.0000%*	*7.30%*	*7,851*	*2,818*	*2*	*1.22*	*130,754*	*2*	-	*MT510661*
					U034-o2	*Oral swab*	*0.0000%*	*6.00%*	*4,085*	*1,735*	*1*	*0.62*	*130,884*	-	-	*MT510662*

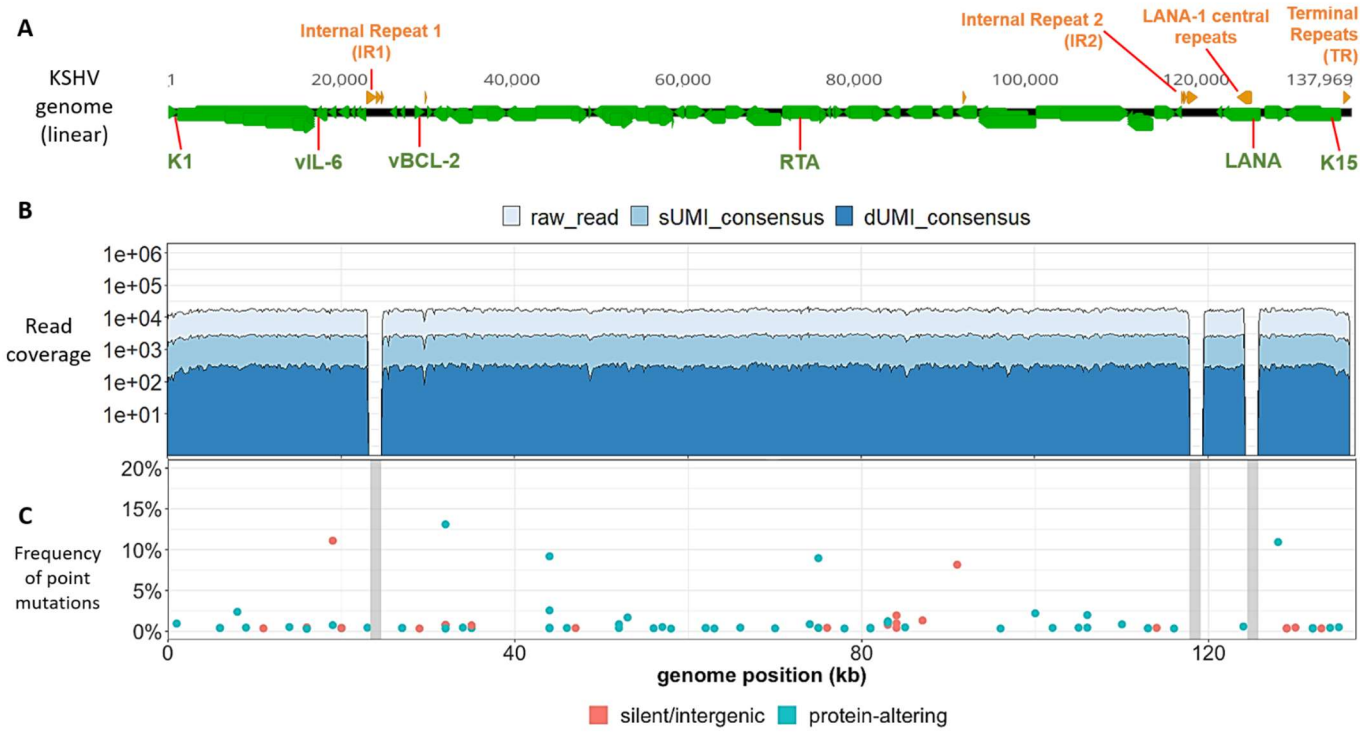


Figure 4. **KSHV genomes in BCBL-1 cells have low point mutational diversity**

(A) *Schematic representation of a linear KSHV genome, with genes colored in green and the major repeat regions in orange. The locations of the K1, vIL-6, vBCL-2, RTA, LANA and K15 genes used for genome quantitation are indicated in green type.* ***(B)*** *Raw (light blue), sUMI-consensus (blue) and dUMI-consensus (dark blue) read coverage along the de novo assembled, BCBL-1 KSHV genome. Reads from the 3 large internal repeat regions were masked out, with read coverage showing as zero.* ***(C)*** *Bubble plot of minor sequence variants. Each bubble represents a position within the genome at which a variant base or indel was detected, colored by whether they were predicted to be silent or protein-altering mutations. Mutations likely to be silent included synonymous and intergenic point mutations, while protein-altering mutations included non-synonymous, nonsense and frameshift mutations. Bubble height represents variant base frequency among dUMI-consensus reads at that position. Vertical grey columns represent masked internal repeat regions.*

The consensus *de novo*-assembled KSHV genome in BCBL-1 had 3 differences from the published BAC-36 sequence: a C→A change in the noncoding sequence before ORF K5 (BAC-36 position 24,630), 2 additional Gs in a homopolymer run at BAC-36 position 25,210), and a

synonymous T→C change in the K7 gene (BAC-36 position 28,409). No variant bases were found in dUMI-consensus reads at the equivalent positions, indicating that the 3 BAC-36 sequence variants were not present in this passage of the BCBL-1 line at detectable levels (i.e., <1 copy per 302 genomes).

KSHV sequence derivation from tumor tissues and oral swabs

KSHV genome sequences were successfully obtained from samples provided by 9 participants with HIV-associated KS, including 12 KS tumors and 11 oral swabs. (**Table 4**). The representation of KSHV DNA in a sample was determined by ddPCR analysis of segments of vIL-6, vBCL-2, RTA and LANA genes (**Fig 4A**) and provided as the percentage "on-target" KSHV DNA. These levels ranged from 0.03% to 1.35% (median 0.17%) in tumors, while most oral swab samples were below 0.01% on-target (**Table 4**). Following one enrichment with RNA baits, KSHV DNA corresponded to a median of 1.3% on-target, and after a second enrichment a median of 24.2% on-target, for a median final enrichment of 123,955-fold.

Median read coverage across KSHV genomes, excluding the major repeat regions, was 22,896 for tumors and 37,794 for oral swab samples. After collapsing mapped reads by dUMI, the median dUMI-consensus read coverage was 380 for tumors and 27 for oral swabs (**Table 4, Fig 5**). The lower dUMI-consensus read coverage of oral swab KSHV sequences, despite having higher raw read coverage than in tumors, was due to oral swab sample libraries having lower amounts of KSHV DNA and higher proportions of PCR duplicates. This resulted from low viral genome input requiring more rounds of enrichment and PCR cycles, and more significant DNA degradation during storage. Indeed, only 11 of 43 oral specimens attempted yielded acceptable library quality for whole KSHV genome sequencing, compared to 12 of 13 tumor samples. Since the median dUMI-consensus read coverage corresponds to the number of viral genomes sampled, tumor U032-B had the highest number of genomes analyzed at 1,653. The lowest number of genomes accepted for confident assignment of variant frequencies was set to 100 (**Fig 6A**)**;** below this number dUMI-consensus read coverage was judged to be too sparse. U020-B was an exception due to most genomes having a large deletion, to be discussed below. For other samples with genome counts below 80, dUMI-consensus reads generated were insufficient to cover the entire KSHV genome, even if whole KSHV genomes could be assembled

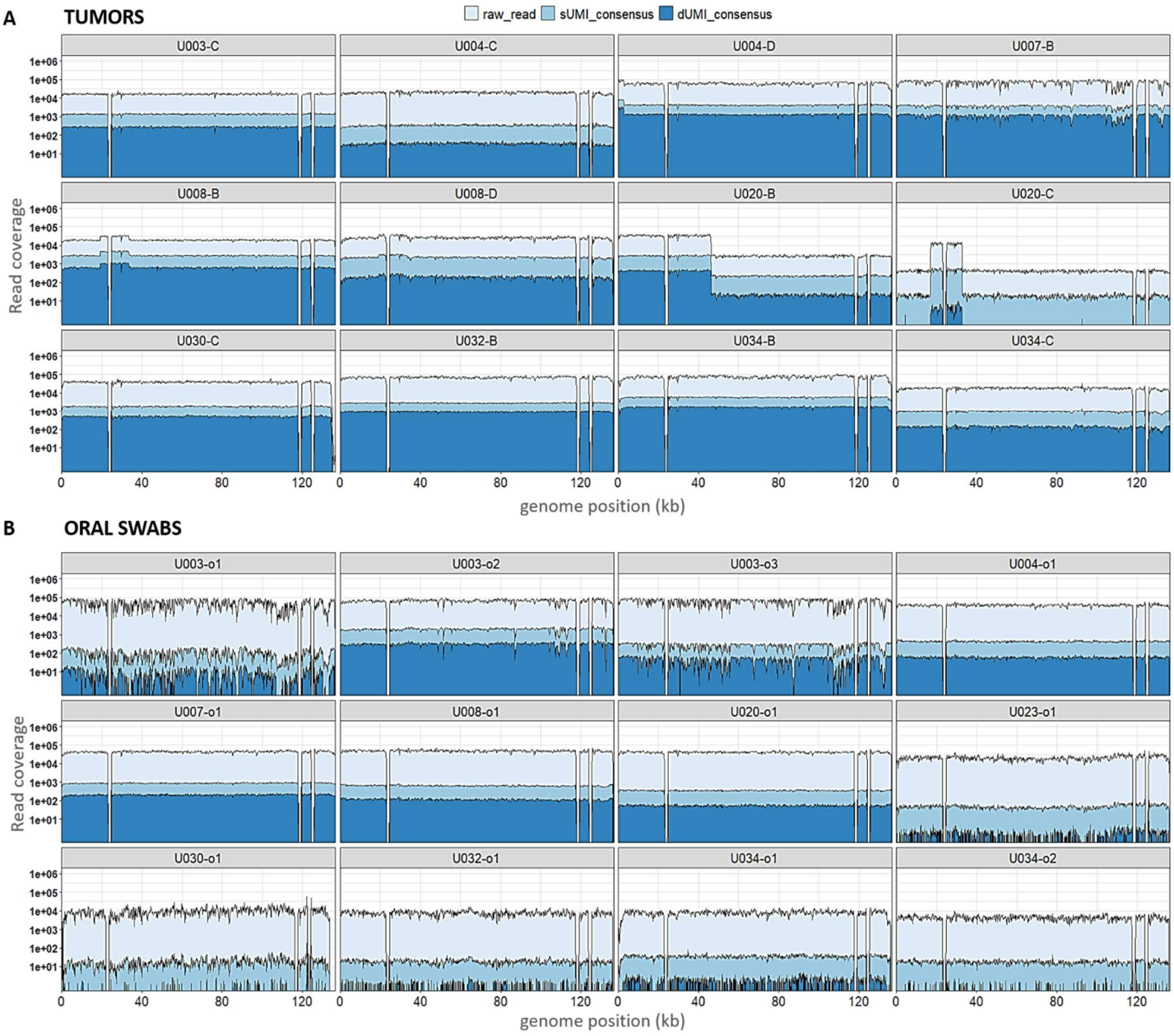


Figure 5. Raw, sUMI and dUMI read coverage across tumor- and oral swab-derived KSHV genomes

*Raw (light blue), sUMI (blue) and dUMI-consensus (dark blue) read coverage is shown on log scale along the de novo assembled, sample-consensus KSHV genomes from tumors **(A)** and oral swabs **(B)** reported in Chapter 3. Major repeat regions were masked and seen here as no coverage regions.*

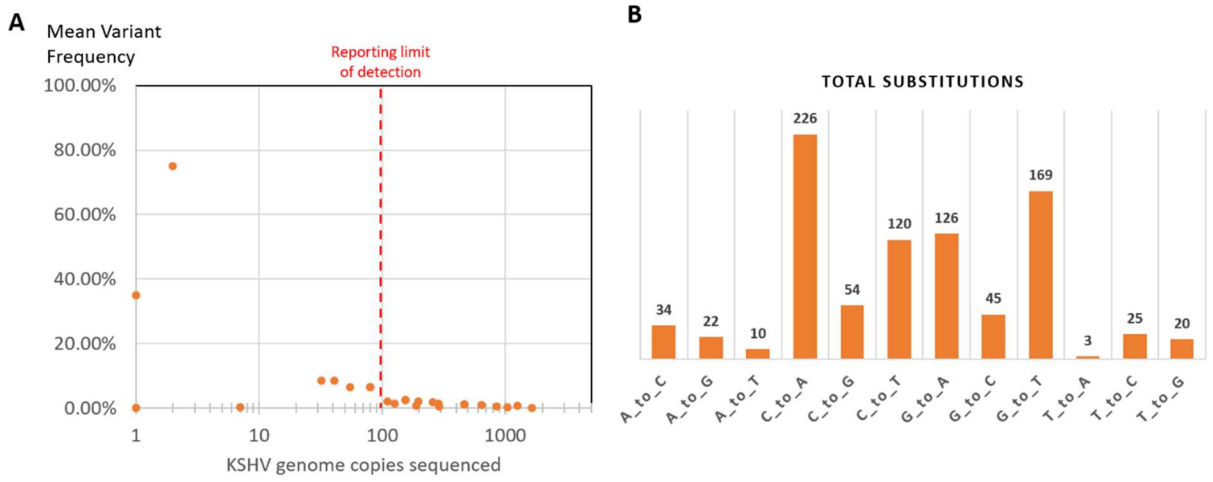


Figure 6. **Potential sequencing artifacts**

(A) *KSHV intrasample variant frequency as a function of read coverage. Sample variant frequencies were shown in Table 4 when at least 100 viral genomes were sampled, since below that level, minor variant frequencies were judged to be unreliable.* ***(B)*** *Minor variants detected in dUMI-consensus reads of all samples by type of base substitution.*

from raw reads. Overall, read coverage was relatively uniform along the KSHV genome for most tumors (**Fig 5A**) and all adequately sampled oral swabs (**Fig 5B**).

Very few point mutations were found in dUMI-consensus reads from either tumors (**Fig 7A**) or oral swabs (**Fig 7B**). Excluding the major repeat regions, the number of genome positions with a detectable intrasample variant base ranged from 2 – 218 (<0.01 – 0.17%) (**Table 4**). These frequencies were lower or comparable to those in the BCBL-1 cell line, although clinical samples had detectable variation in long homopolymer runs not observed in the BCBL-1 viruses. The sample-consensus genome was generally the only KSHV sequence present in each sample, hence, there was no evidence for the existence of quasispecies [181].

Artifacts resulting from the end-repair step in DNA library preparation, which precedes the application of dUMI tags, cannot be corrected by duplex sequencing [138,143,182]. Hence, 9 bases were trimmed from ends of dUMI-consensus reads before analyses, substantially reducing the variation observed in the raw data (not shown). The minor base variants remaining in all samples revealed a preponderance of C→A and G→T substitutions (**Fig 6B**) as well as

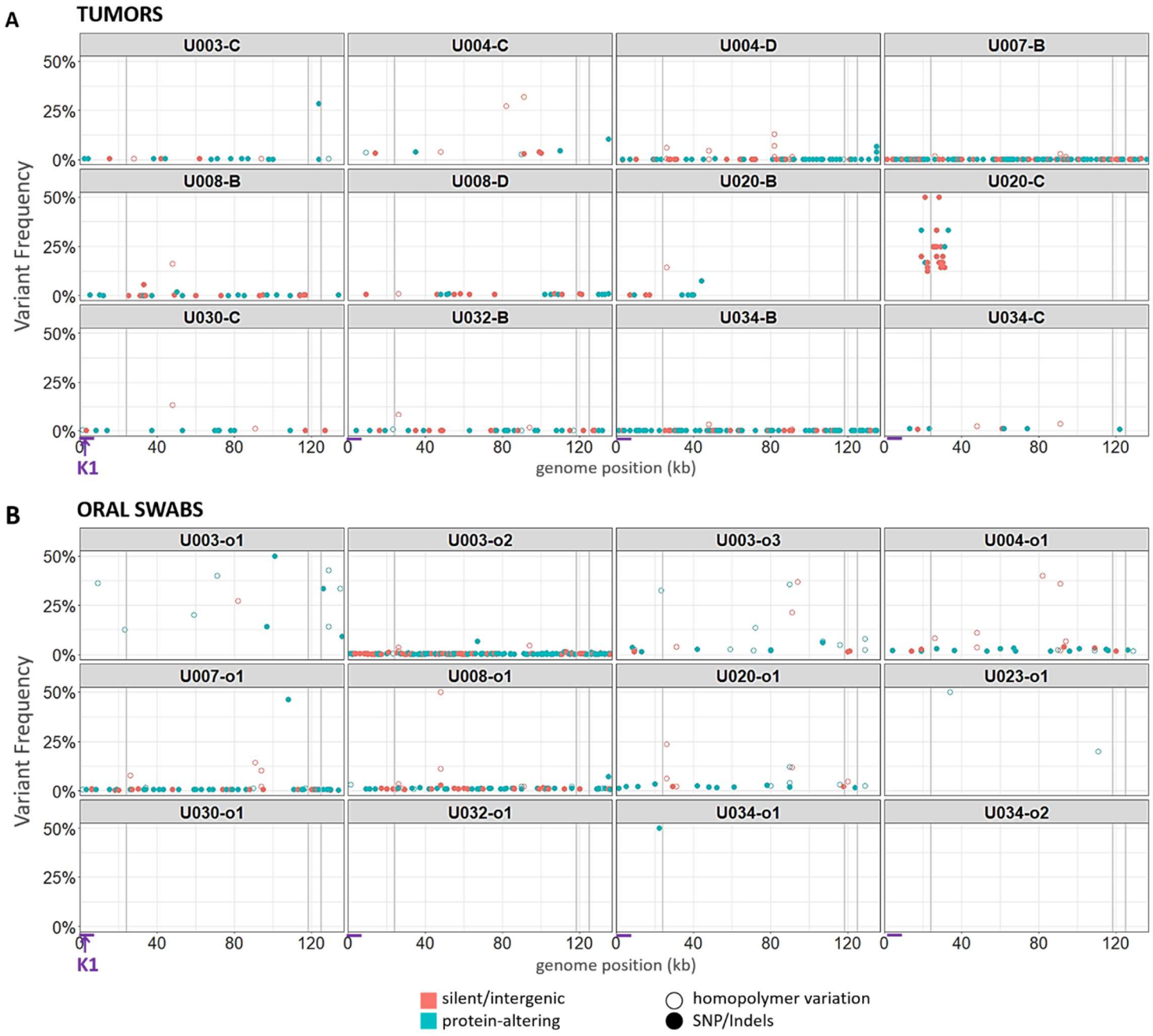


Figure 7. Point mutational diversity in KSHV genomes from tumors and oral swabs

Bubble plots of minor sequence variants remaining after removal of PCR errors, in KSHV genomes from tumors ***(A)*** *and oral swabs* ***(B)****. Each bubble represents a variant base or indel, colored by whether they were predicted to be silent or protein-altering mutations. Silent mutations include synonymous and intergenic point mutations, while protein-altering mutations included non-synonymous, nonsense and frameshift mutations. Hollow circles represent mutations occurring in homopolymer runs. Bubble heights represent the frequency of the variant base among dUMI-consensus reads at that position. Vertical gray columns represent the masked repeat regions. The region containing the K1 gene is indicated with arrows at the bottom of the figure.*

differences in homopolymer run lengths (**Fig 7A & B**). However, most minor variants were supported by only one dUMI-consensus read. Overall, mean variant frequency and median dUMI-consensus read coverage were inversely correlated (**Fig 6A**). Since the remaining variation cannot be distinguished from artifacts, true minor variant frequencies could be even lower than reported here.

KSHV genomes were virtually identical at the point mutational level between tumors and oral swabs from the same individual

Within-individual single nucleotide differences between tumor and oral swabs ranged in number from 0 – 2 across the entire ~131-kb genomes, not counting the major repeat regions. Notably, there were almost no polymorphisms in the KSHV hypervariable gene K1 (**Fig 7A & B**). Hence, no evidence for minor KSHV variants, quasispecies or multi-strain infections was found in these individuals.

KSHV genomes were distinct across the 9 participants, with between-individual differences ranging from 3.06-4.85%. They included K1 subtypes A5, B1 and C3 (**Fig 8A**) and K15 alleles P and M (**Fig 8B**). While K1 and K15 are the most variable KSHV genes, polymorphisms along the rest of the genome have been reported to contribute more in aggregate to the total diversity of KSHV [75–77]. Consistent with this, maximum-likelihood phylogenetic trees using entire KSHV genomes (**Fig 8C**) were topologically distinct from those of K1 or K15. Moreover, due to signatures of recombination in the evolutionary history of KSHV [59,77], differing phylogenies along sections of the KSHV genome may be better represented by a phylogenetic network (**Fig 8D)**, in which higher degrees of conflict result in a more web-like structure rather than a tree.

Aberrant KSHV genome structures in tumors

Among the 12 tumor-derived KSHV genomes examined, 7 had anomalous read coverage that shifted abruptly once or twice along the viral genome (**Fig 5A**). In contrast, oral swab KSHV genomes from the same individuals had uniform read coverage while being identical in sequence. This argues against preferential target capture by RNA baits or sequencing biases. Repeating the enrichment and sequencing of reproduced their distinctive read coverages in the 2 instances

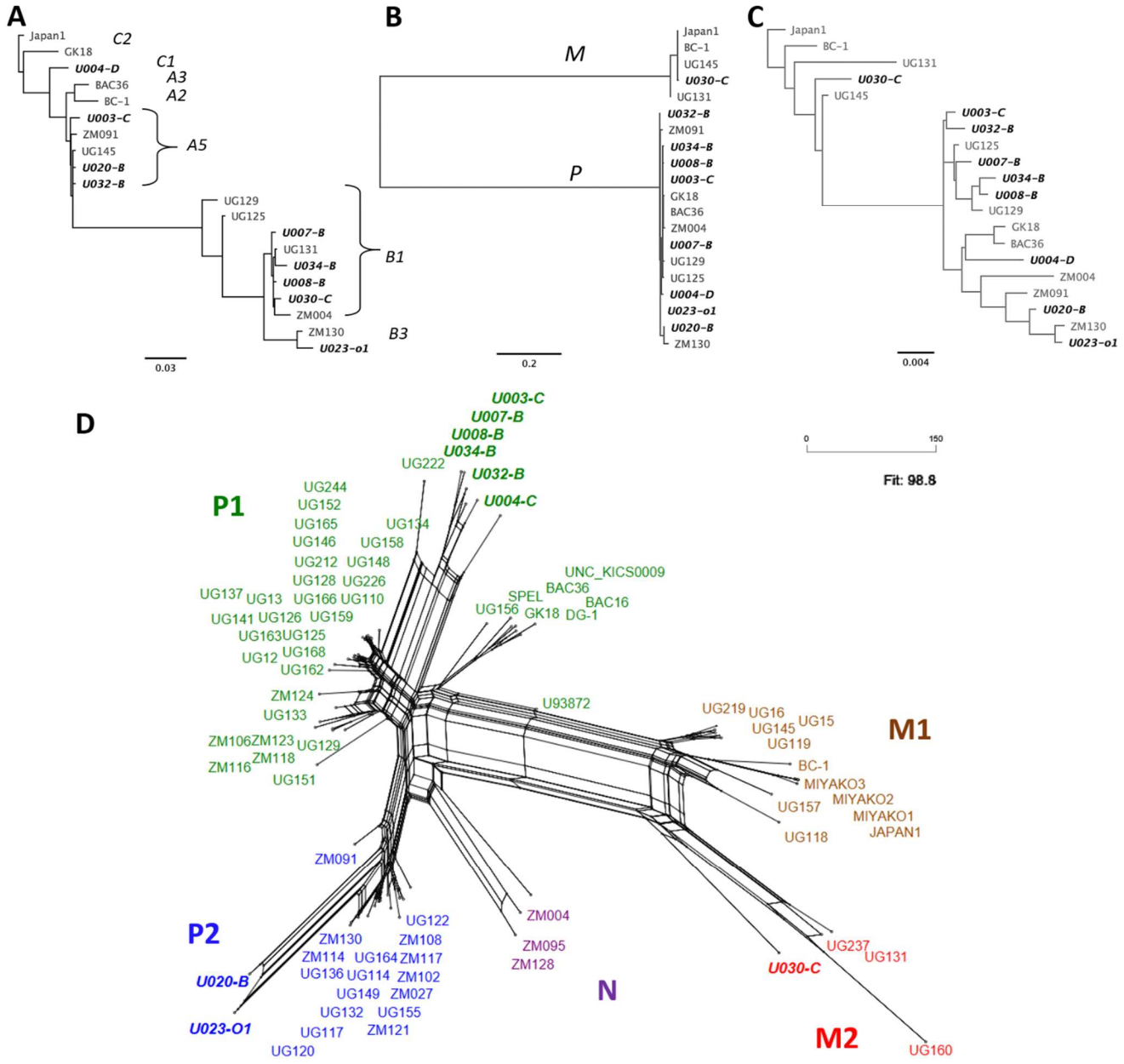


Figure 8. KSHV phylogenetic relationships in variable genes K1 and K15 and whole genomes

*Phylogenetic trees of **(A)** K1 genes, **(B)** K15 genes and **(C)** whole genomes from this study and of select genomes from other publications. K1 and K15 subtypes are indicated in the K1 **(A)** and K15 (B) trees. **(D)** A neighbor-net phylogenetic network of all published KSHV genomes to date, color-coded by genome types proposed in* [77]*: P1 in green, P2 in blue, N in purple, M1 in red and M2 in maroon. All de novo-assembled genomes from this study are in bolded italics.*

examined. Split reads accumulated at the points of abrupt shifts in read coverage, even after collapsing reads by their dUMI consensus, indicating that these were not PCR template switching artifacts but rather were breakpoints in the viral genomes in vivo. Specific viral genomic anomalies observed are detailed below:

Tumor U003-C. Read coverage from U003-C was high (average of 15,635) and uniform across the KSHV genome except for a 6-bp gap within the K8.1 gene intron extending to the first base of the second K8.1 exon (**Fig 9A**). No read documented a deletion in this region, nor was any read found with its mate pair located across the 6-bp gap. To investigate its structure, this region was PCR amplified from unsheared U003-C tumor DNA using conserved primers flanking the gap (**Fig 9B**) yet, no PCR product was detectable. In contrast, in contrast this region was intact when amplified and sequenced from an unrelated tumor (**Fig 9C**) and BCBL-1 (not shown). Rather, *de novo* assembly revealed that the reverse complement of TR sequences continued from the gap position (**Fig 9B**). K8.1-TR junctions were confirmed by PCR with primers flanking the junctions (**Fig 9D**) and Sanger sequencing. Inversion of the 60-kb 3' half of the U003-C genome, starting inside K8.1, is a parsimonious explanation for the breakpoints.

Tumor U004-D. The first 3kb, from K1 to the end of gene ORF4, had 1.5X read coverage compared to the rest of the KSHV genome (**Fig 5A**). However, no split reads or chimeric read pairs were found to explain this result from a genome rearrangement or deletion.

Tumor U008-B and D. U008-B had 1.7X greater read coverage over a 14.8-kb segment from inside K3 to inside ORF19 (GK18 reference positions 19,168 to 33,980, **Fig 10A**), including IR1 (masked). This was corroborated by ddPCR quantitation of vBCL-2, inside the 1.7X coverage region, with 1.7 – 1.9-fold higher gene copy number in the tumor compared to vIL-6, RTA and LANA (**Table 5**).

Inferring from split reads, the 14.8-kb segment was inside IR2 (to GK18 position 119,496, **Fig 10D**). This was confirmed in the unsheared tumor DNA extract by PCR and Sanger sequencing using primers spanning the breakpoint (**Fig 10D & E**, lanes 4 & 5). Other primer pair combinations were tested to see if there were DNA species with the 14.8-kb segment inverted, deleted in place, duplicated in tandem or rearranged in other ways. None generate detectable PCR products except for primer pairs showing that the 14.8-kb segment also exists in the native

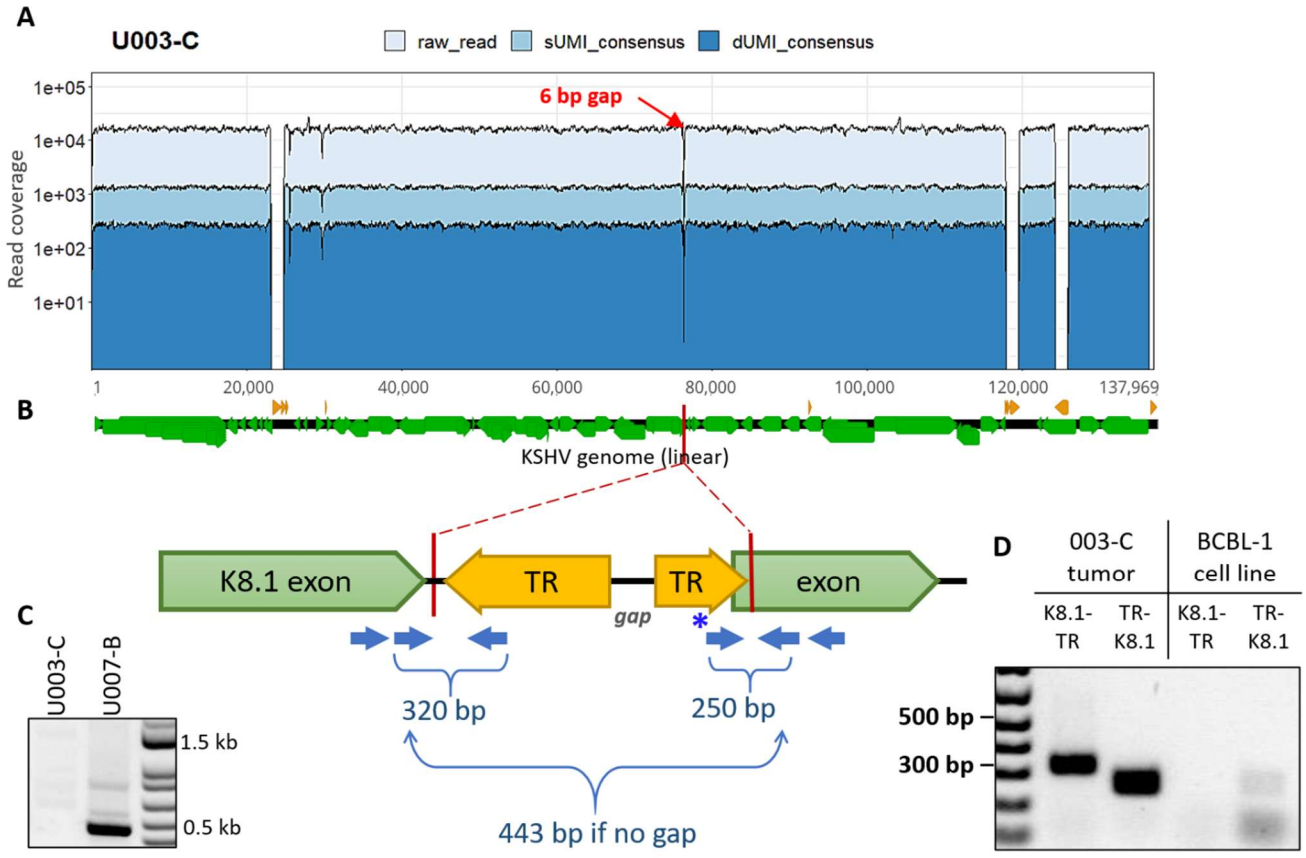


Figure 9. **KSHV genomes in the U003-C tumor**

(A) *Read coverage of the U003-C KSHV genome, showing a 6-bp gap (red arrow) where no read pairs were mapped.* ***(B)*** *Cartoon of the de novo-assembly sequences generated at either side of the gap, both ended within the K8.1 gene intron and continued into terminal repeat (TR) sequences. Green and yellow arrows show the directions of the K8.1 gene and terminal repeat sequences, respectively. Blue arrows show the position of PCR primers used to confirm breakpoint junctions, with the expected PCR product sizes.* ***(C)*** *PCR products generated from U003-C tumor DNA using primers flanking the gap. The 443-bp PCR product expected if the K8.1 gene intron was intact was not detected in U003-C (left column) and was detected in tumor U007-B (right column) from another person.* ***(D)*** *Hemi-nested PCR of U003-C tumor DNA for the K8.1-TR (left) and TR-K8.1 (right) junctions produced products of the predicted sizes. These structures were confirmed by Sanger sequencing. No K8.1-TR or TR-K8.1 junction fragment was produced from BCBL-1 DNA. The light bands at the TR-K8.1 lane under BCBL-1 were determined from Sanger sequencing to be amplicons generated from the forward primer sequence (indicated with * in panel B) overlapping with K8.1; this primer was used since the rest of the connected TR sequence assembled was GC-rich and unsuitable for primer design.*

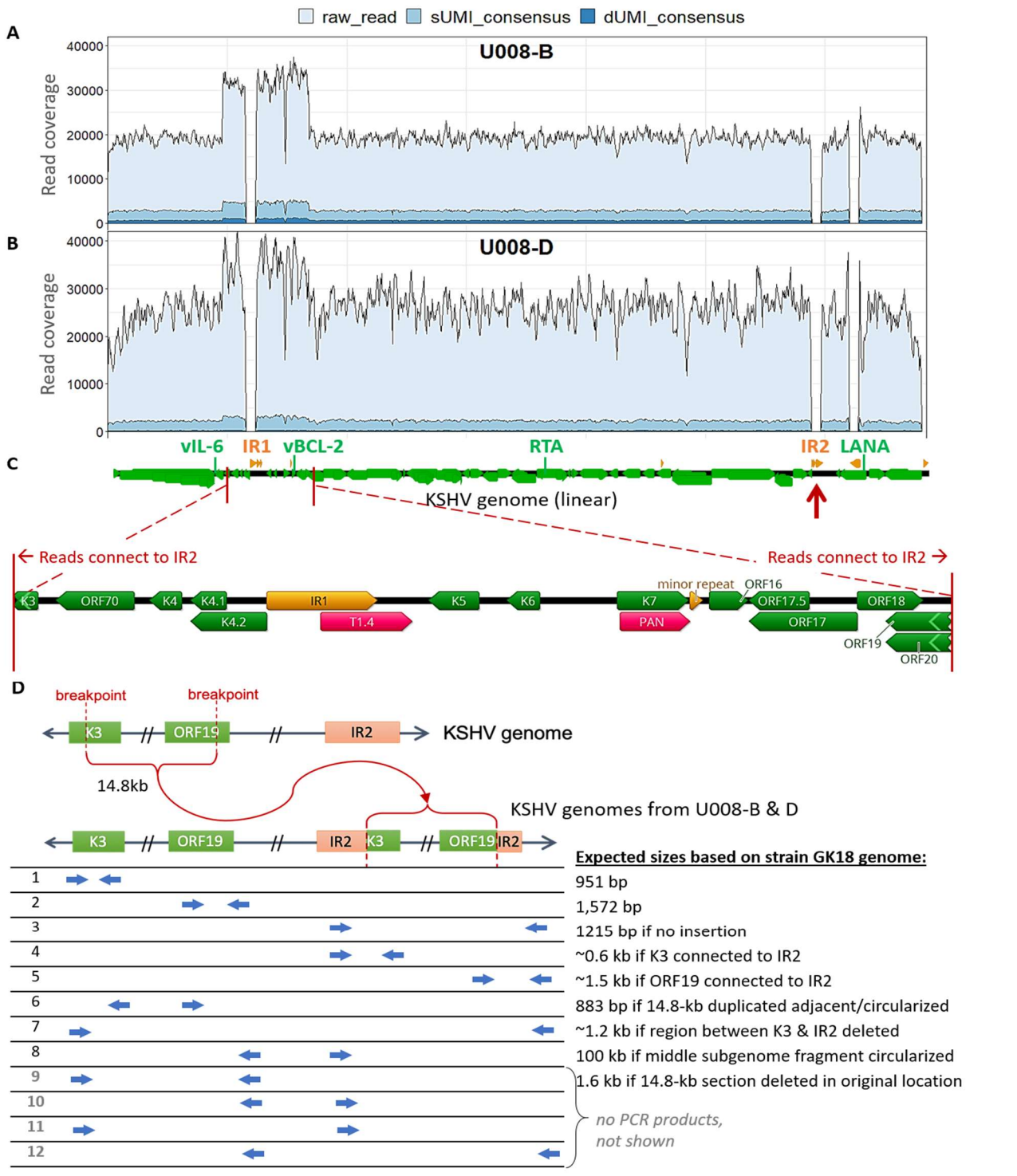

raw_read
sUMI_consensus
dUMI_consensus
A
U008-B
Read coverage
40000
30000
20000
10000
0
B
U008-D
C
vIL-6
IR1
vBCL-2
RTA
IR2
LANA
KSHV genome (linear)
Reads connect to IR2
Reads connect to IR2
K3
ORF70
K4
K4.1
IR1
K5
K6
K7
minor repeat
ORF16
ORF17.5
ORF18
K4.2
T1.4
PAN
ORF17
ORF19
ORF20
D
breakpoint
breakpoint
K3
ORF19
IR2
KSHV genome
14.8kb
KSHV genomes from U008-B & D
K3
ORF19
IR2
K3
ORF19
IR2
1
2
3
4
5
6
7
8
9
10
11
12
Expected sizes based on strain GK18 genome:
951 bp
1,572 bp
1215 bp if no insertion
~0.6 kb if K3 connected to IR2
~1.5 kb if ORF19 connected to IR2
883 bp if 14.8-kb duplicated adjacent/circularized
~1.2 kb if region between K3 & IR2 deleted
100 kb if middle subgenome fragment circularized
1.6 kb if 14.8-kb section deleted in original location
no PCR products,
not shown
E
U008-B
BCBL-1
1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8
1500 bp
850 bp
500 bp
F
U008-D
1 2 3 4 5 6 7 8
1500 bp
850 bp
500 bp

Figure 10. KSHV genomes in two tumors from participant U008 had a 14.8-kb region flanking IR1 duplicated and translocated into IR2

*Total, sUMI and dUMI-consensus read coverage of tumor B **(A)** and D **(B)** genomes from individual U008. **(C)** Annotations of the region with 1.5-2X read coverage, with genes in green, repeat regions in orange, and long non-coding RNAs in red. Many reads on the edge of this region continue into IR2 (red arrows). Annotations are from the KSHV reference isolate GK18. **(D)** Cartoon showing the duplication of the 14.8 kb region into IR2 and the relative positions of PCR primers used to examine the genomic rearrangement in unsheared tumor DNA extracts from tumors U008-B and U008-D. PCR products produced from primer pairs numbered in D from U008-B and BCBL-1 **(E)** and in U008-D **(F)**. All visible bands were excised from the agarose gel and sequenced, confirming the indicated junction sequences. Primer pairs # 9-12 produced no PCR products discernible on an agarose gel and are not shown here.*

Table 5. Gene copy numbers in tumor or oral swab DNA

Sample	vIL-6	vBCL-2	RTA	LANA
U003-C	N/A	N/A	N/A	9,664
U003-o1	127	100	120	135
U003-o2	1,298	1,238	1,452	1,628
U003-o3	191	184	169	199
U004-C	1,243	1,274	1,205	998
U004-D	4,433	4,466	4,543	4,290
U004-o1	117	128	120	112
U007-B	1,842	1,864	2,040	1,925
U007-o1	376	356	322	344
U008-B	19,140	33,629	19,910	19,195
U008-D	24,360	34,755	12,737	17,189
U008-o1	129	136	119	138
U020-B	49,600	55,850	4,550	5,500
U020-C	3,658	5,033	145	139
U020-o1	254	231	234	248
U020-o2	100	183	123	225
U023-o1	62	57	50	66
U030-C	N/A	N/A	N/A	59,730
U030-o1	2	5	5	7
U032-B	476	520	494	466
U032-o1	9	0	2	0

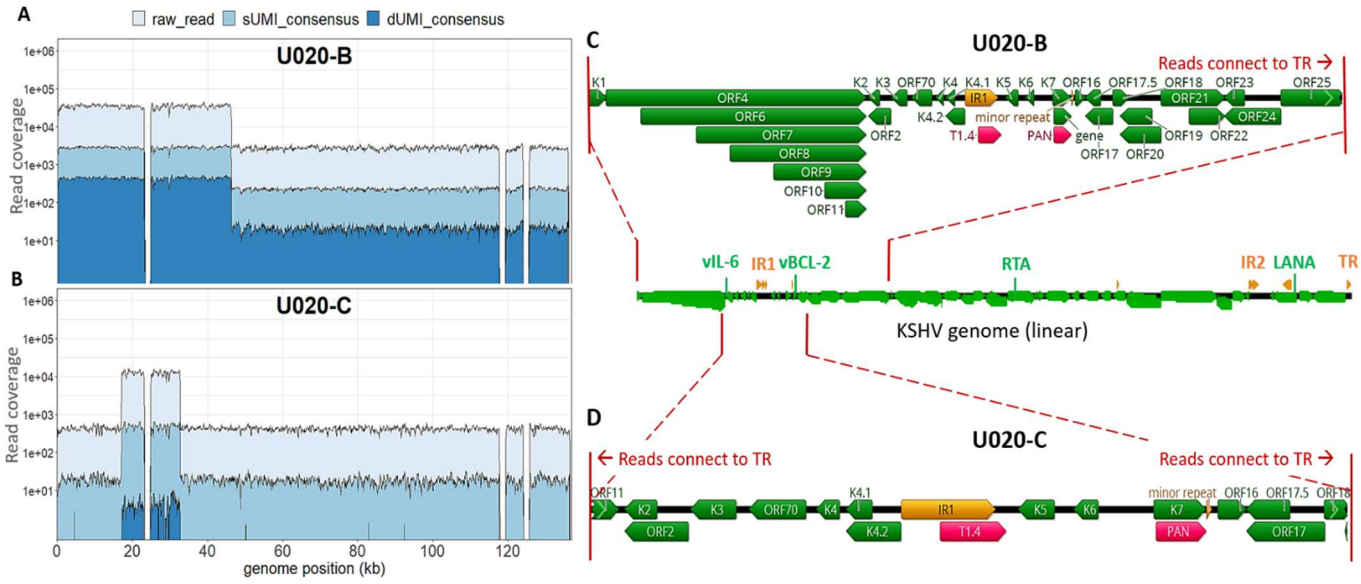
U034-B	1,920	1,887	1,870	1,793
U034-C	13,083	13,335	8,148	6,552
U034-o1	9	6	9	7
U034-o2	6	11	0	1

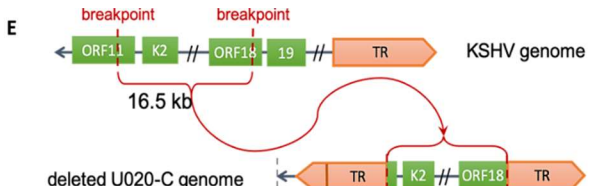
configuration (**Fig 10E**). Thus, the 14.8-kb segment was copied into IR2 but had not been deleted from its original location.

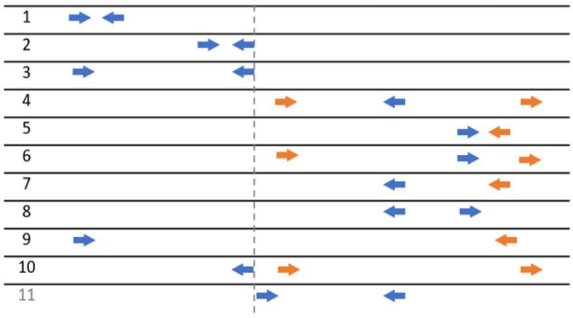
In a parallel study of viral transcriptomes [119], abundant expression of a chimeric Kaposin transcript fused to the 14.8-kb segment was found in tumor U008-B, consistent with the viral genome structure I observed. Another tumor from the same participant, **U008-D** (**Fig 10B**), had 100% nucleotide identity and was confirmed to have the same duplication and breakpoint junctions (**Fig 10F**).

Tumor U020-B. Read coverage abruptly dropped 12.8-fold over the 3' ~90 kb of the KSHV genome in this tumor (**Fig 11A**). This was consistent with ddPCR quantitation, with vIL-6 and vBCL-2 gene amplicons having 9.0 – 12.3-fold higher levels than RTA and LANA (**Table 5**). The coverage shifted before the end of ORF25 (GK18 position 46,615) and reads at this breakpoint continue into TR sequences ~90 kb downstream (**Fig 11C**). Thus, U020-B appeared to have KSHV genome variants with a ~90-kb deletion, or formally, a 12.8X amplification of a 46-kb subgenomic region. No U020-B tumor DNA remained to allow confirmation of this breakpoint.

Tumor U030-C. Greater than 30,000 reads/position were uniformly observed throughout most of the KSHV genome. However, coverage dropped or was missing within the K15 gene (**Fig 5A**). The remaining K15 sequences corresponded to the K15 M-allele, which is less common than the P allele but was included in the RNA bait design I used (GenBank U75698). PCR amplification and Sanger sequencing of this region showed that the U030-C tumor did contain some copies of the entire M-allele K15 sequence. The U030-C sample-consensus genome was finished with this Sanger sequence result, since no dUMI-consensus reads mapped to the gaps in K15. In the parallel RNAseq study of tumors of this same participant, U030-B and C, transcripts of K15 were also lacking, unlike tumors from all other participants [81].







Expected sizes based on strain GK18 genome:

478 bp if wild-type

969 bp if wild-type

892 bp if 16.5-kb section deleted from original location, 17 kb if not

at least ~1 kb if K2 connected to TR (will include 1-unit length TR)

240 bp if ORF18 connected to TR

at least ~1 kb if ORF18 connected to TR in reverse

277 bp if K2 connected to TR in reverse

392 bp if 16.5-kb region duplicated adjacent/circularized

328 bp if ORF11 connected to TR (end of ORF11 to TR deleted)

at least ~1 kb if ORF 18 end connected to TR

~1 kb if unique sequence upstream of TR connected to ORF11 fragment (no band, not shown)

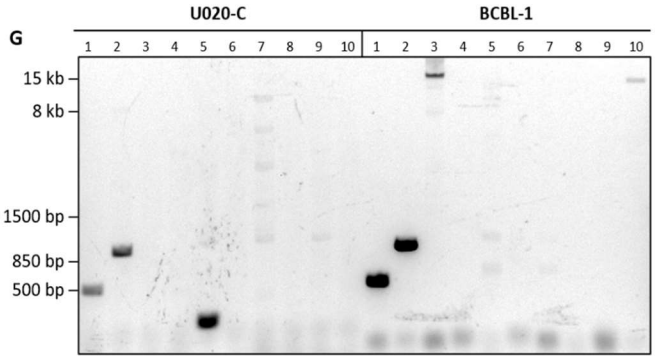


Figure 11. KSHV genomes in U020-B and U020-C have large, distinct deletions

*Total, sUMI and dUMI-consensus read coverages of U020-B **(A)** and U020-C **(B)** KSHV genomes. GK18 reference annotations in the high-coverage regions of U020-B **(C)** and U020-C **(D)**. **(E)** Cartoon showing the region encompassing the high coverage region of U020-C viral genomes, leaving a 16.5-kb region connected to TR. **(F)** Relative positions of PCR primers used to examine genome structure. Primers for unique genomic sequences are in blue, and primers for TR sequences are in orange. **(G)** PCR products produced from primer pairs numbered in E, with DNA from U020-C and BCBL-1 as templates. Bands from lanes 1, 2 and 5 were excised from the agarose gel and the DNA used for Sanger sequencing. Faint bands in lanes 7 (five bands) and 9 (one) under U020-C were extracted from the gel but did not yield enough DNA for sequencing, except for the light 1-kb band in lane 7. Top BLAST hits to this sequence were human phosphatidylserine synthase 2 gene sequences, likely amplified due of spurious primer homology. The ~16 kb band in lane 3 under BCBL-1 was confirmed by sequencing to be ORF11 and ORF18 sequences from its two ends, and lane 10 under BCBL-1 corresponded to sequences common to many cloning vectors such as pCMV-VEE-GFP. Row 11 primers in (F), in which the forward primer binds to unique genomic sequences preceding the TR, yielded no discernible product (not shown).*

The same aberrant KSHV genomes are found in multiple lesions from the same individual

In the case of U008-B and U008-D, two tumors biopsied from distinct lesions on the left leg (**Fig 12**), full-length genome sequencing showed that they had the same 14.8 kb KSHV sub-genomic sequence duplicated in IR2 (**Fig 11F**). PCR primers across those breakpoints were used to screen for the same structures in 6 other distinct KS lesions from this individual, and none had this duplication (not shown). In contrast, four additional tumors tested from participant U003 had the same inversion breakpoints, detected by PCR, as tumor U003-C (**Fig 13**). Moreover, no intact K8.1 sequences were detected in 2 of these 4 tumors by nested PCR of the region spanning the K8.1 intron gap (**Fig 13**). These biopsies came from distinct lesions in the left leg. Lastly, in participant U020, the ORF18-TR junction sequences found spanning the U020-C genomic deletion was not detected in the 2 other tumors tested.

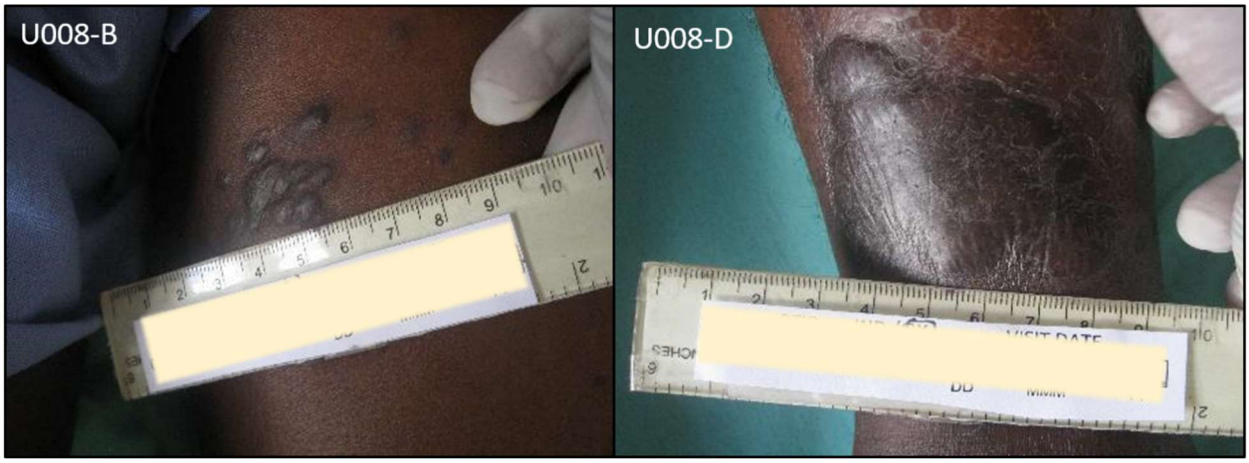


Figure 12. U008-B and U008-D were from distinct lesions on the left leg.

U008-B and U008-D were from distinct lesions on the left leg. The U008-B biopsy was obtained from lesions in the upper thigh, while the U008-D tumor tissue was biopsied from a large lesion on the knee.

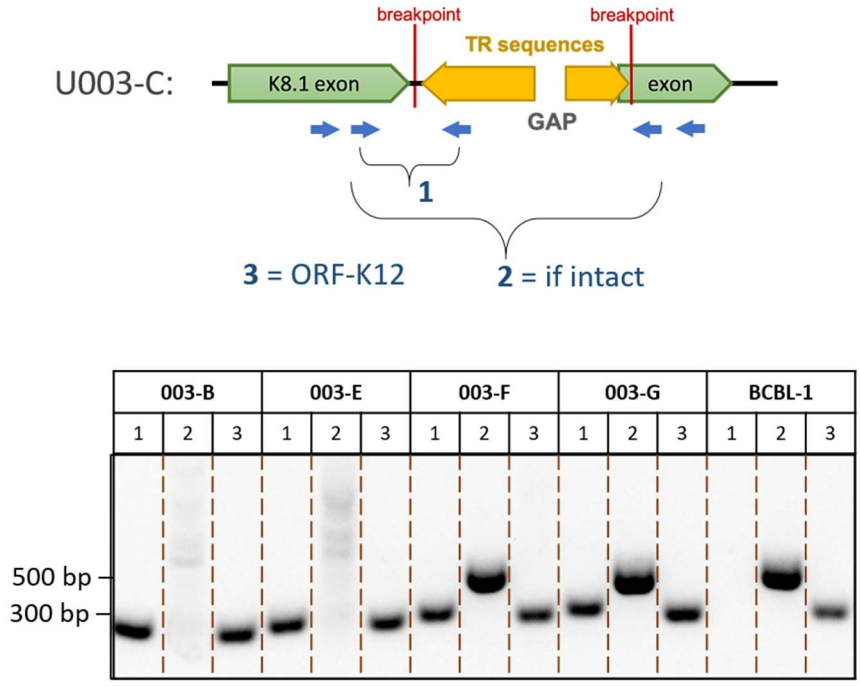


Figure 13. Genome inversion found in multiple tumors of participant U003.

Junction sequences marking the genomic aberration in U003-C were detected in all 4 other tumors tested, while intact K8.1 sequences were detected in 2 of the 4. The cartoon shows the breakpoints in the K8.1 intron of U003-C extending into TR sequences, along with PCR primers

used to confirm the genome structure. The numbers (1, 2 and 3) correspond to lanes in the bottom of the figure. A positive PCR result in lanes 1 indicates a product spanning the left-hand breakpoint in the rearranged genome and a positive PCR result in lanes 2 indicates the presence of an intact K8.1 sequence. The products in lanes 3 are positive controls showing amplification of the KSHV K12 gene. PCR products from other tumors of participant U003 and from the BCBL-1 cell line are shown at the bottom of the figure. All visible bands were excised from the agarose gel and their structures confirmed by Sanger sequencing.

Mutations in sample-consensus KSHV genomes from tumors that impacted protein coding sequences

Among the 7 participants with KSHV sequences from at least one oral swab and one tumor, sample-consensus KSHV genomes were identical in the oral and tumor samples from 2 participants with few differences in 4 others. In the remaining participant, U004, the sample-consensus KSHV genome in one tumor was identical to that in the oral swab but a second tumor had mutations. Tumor-unique mutations were typically nonsynonymous point mutations resulting in highly dissimilar residues or were mutations that would disrupt their expression, such as rearrangement breakpoints (**Table 6**).

Several of the mutations or genome aberrations observed in tumors occurred in structural genes (**Table 6**), and frequently involved the K8.1 gene, which encodes an envelope glycoprotein: The U003 inversion breakpoint cleaved the K8.1 gene; U004-D had an R56Q mutation in its ORF32 tegument protein coding sequence, as well as a 28-nt deletion in the promoter region of K8.1 (**Fig 14A**). The deletion was after the K8.1 core promoter sequence [183], but encompassed the K8.1 transcription start site [184]; the ORF25 major capsid protein in U020-B had a Q594K mutation, in addition to the U020-B genomic deletion that started downstream of ORF25; U020-C had a nonsense mutation at the beginning of the second K8.1 exon; and U032-B had a T848A mutation in ORF63, a tegument protein.

The only intra-host synonymous point mutation observed was in ORF K12 of U003-C (GK18 position 118,082). This C to T change occurred within the oncogenic microRNA K10 (miR-K10a-3p) sequence embedded in the K12 transcript. The three oral swab samples from this participant maintained the consensus C at this position (**Fig 14B**), whereas the 4 other tumors from this participant examined had T at this position, with tumor U003-G having a minor population of

viruses with the consensus C (**Fig 14C**). The change was outside the seed sequence of miR-K10a-3p and may have resulted in a slightly more stable stem loop precursor (ΔG -32.40 vs. -32.70, **Fig 15**).

Table 6. Unique KSHV mutations observed in tumors compared to oral swabs from the same individual.

Sample ID	Tumor-specific differences
U003-C	K12 synonymous mutation within miR-K10 genomic inversion starting at K8.1
U004-C	NONE
U004-D	ORF32 nonsynonymous mutation R56Q K15 nonsynonymous mutation A290P 28-nt deletion within the K8.1 promoter 3-kb segment duplication from before K1 to after ORF4
U007-B	NONE
U008-B	duplication of 14.8 kb segment around IR1 into IR2 - breakpoints inside K3 & ORF19 - genes duplicated: ORF70, K4.1, K4.2, K5, K6, K7, ORF16, ORF17, ORF17.5, ORF18
U008-D	same as U008-B
U020-B	ORF25 nonsynonymous mutation Q594K genomic deletion connecting end of ORF25 coding sequence to TR sequences, 47 kb remaining
U020-C	ORF11 nonsynonymous mutation T396P K3 nonsynonymous mutation F88L in transmembrane domain K8.1 nonsense mutation at start of 2nd exon genomic deletion leaving only 16 kb segment surrounding IR1 connected to TR sequences - breakpoints inside ORF11 and ORF18 - ~30X coverage for: K2, ORF2, K3, ORF70, K4, K4.1, K4.2, K5, K6, K7, ORF16, ORF7, ORF17.5
U032-B	ORF63 nonsynonymous mutation T848A
U034-B	NONE
U034-C	NONE

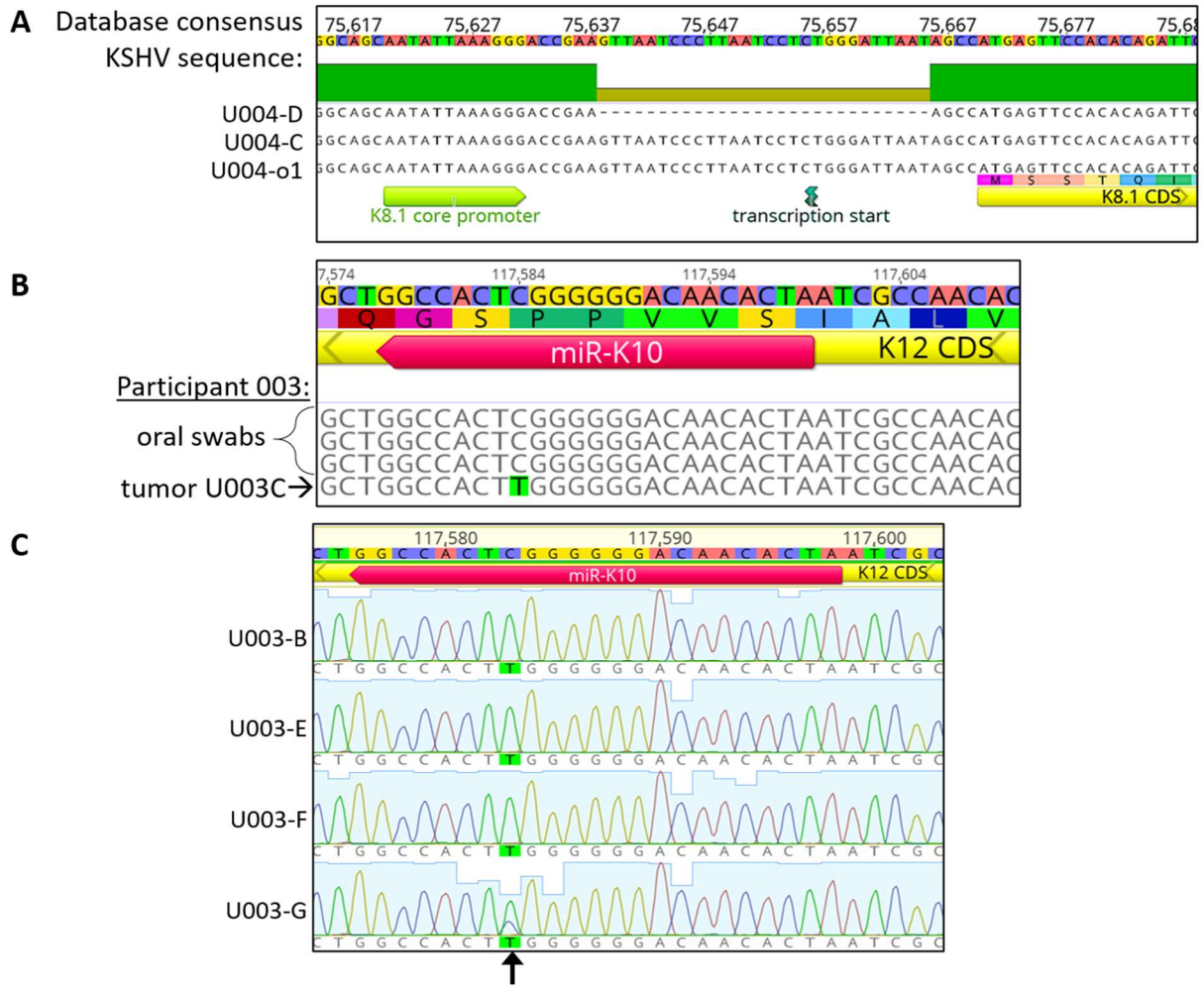


Figure 14. Mutations of KSHV genomes in tumors from participants 004 and 003

(A) *Alignment of KSHV genomes from Participant 004, showing a 28-bp deletion in the K8.1 promoter in U004-D. U004-D and U004-C are from tumors while U004-o1 is from an oral swab.* ***(B)*** *The only intra-host synonymous mutation found in this study, within miR-K10 in participant 003.* ***(C)*** *Sequence chromatograms of miR-K10 in other tumors of participant U003, with a T in all tumors and a mixture of T and the database consensus C in a minority of viruses in U003-G.*

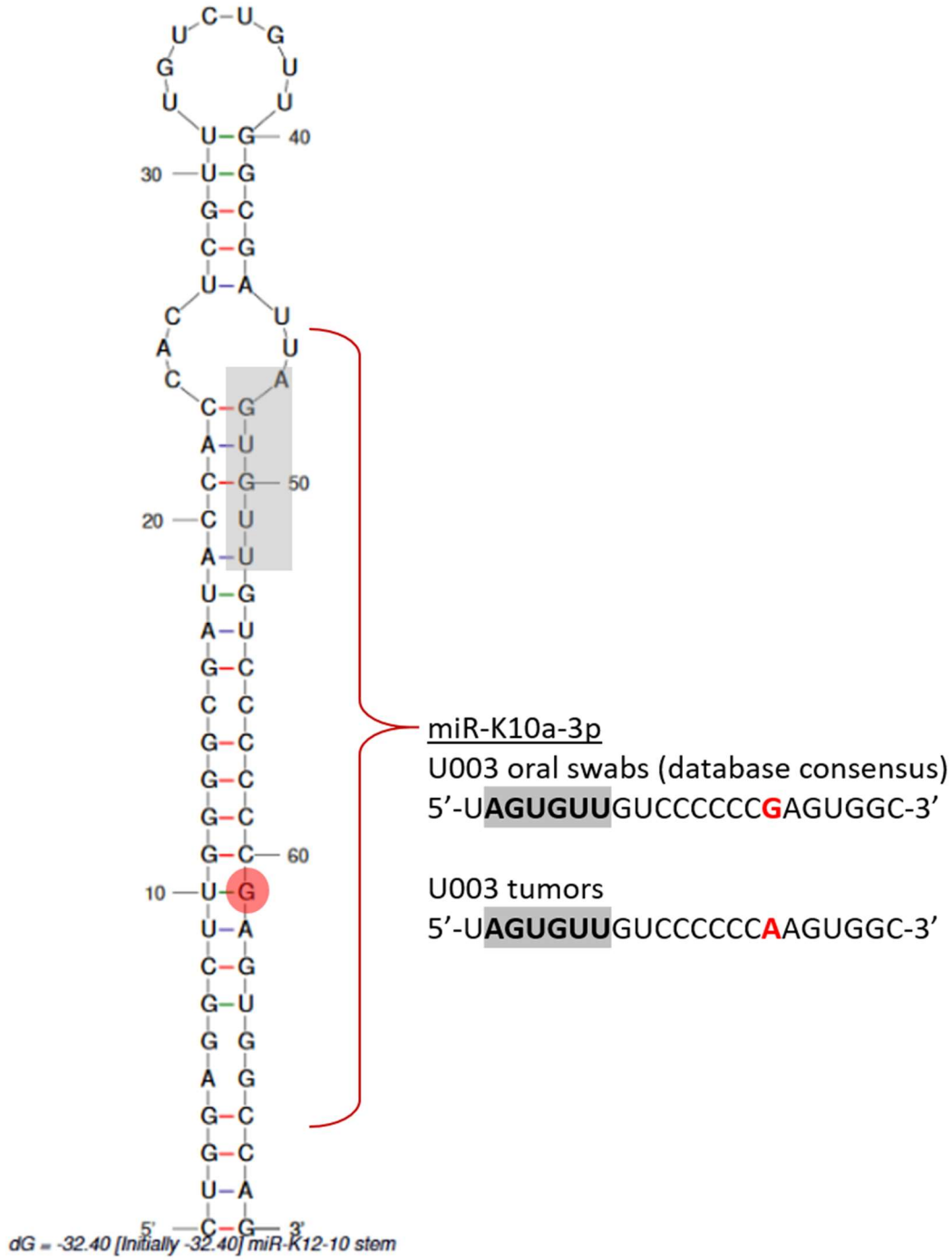


Figure 15. Predicted secondary structure of the stem loop precursor of miR-K10a

The structure of pre-miR-K12-10a was predicted using mfold (http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form), indicating the mature miRNA, seed sequence (grey) and the intra-host polymorphism (red) found in participant U003. The G to A change in RNA sequence resulted in a slightly more stable stem loop (∆G -32.40 à -32.70).

Lack of evidence for integration of KSHV sequences into human chromosomes

No *de novo*-assembled scaffolds, split reads or improperly-paired read mappings suggested an instance of KSHV sequences fused to human DNA. Nevertheless, attempts were made to systematically search for human-KSHV chimeric sequences. The methods employed were the same as those used to screen for all integrated herpesviruses sequences in public databases [185] and EBV integration sites in in primary gastric and nasopharyngeal carcinomas [58]. The KSHV genome inversions, duplications and deletions described above were detected with high confidence values. In contrast, putative breakpoints joining human and KSHV sequences were supported by about 2 orders of magnitude lower in numbers of reads (tens of reads), and often involved LANA repeats and low-complexity human repeat sequences. These were judged to be artifacts.

Co-infection with EBV detected predominantly in oral swabs

Some scaffolds generated during de novo assembly corresponded to EBV sequences. Nearly all oral swab samples yielded multiple EBV-mapping scaffolds up to 73 kb in length, with no region of the EBV genome over-represented. In contrast, EBV-sequences were detected in only 5 of 12 tumors, and in all cases were sequences flanking the EBNA-1 repeat. The proportion of reads mapping to EBV in oral swabs ranged from 0 - 33%, median 1.8%, whereas in tumors the range was from 0 - 0.5%, median 0.002% (not shown). No other eukaryote viruses were identified.

Chapter 4: Tumor-associated KSHV genome aberrations, mutations and their clinical correlates in Ugandan adults with KS

Introduction

Potentially consistent patterns of diversity across the ~165-kb KSHV genome are beginning to emerge. The first whole genome sequence of KSHV reported had a 33-kb region duplicated in the TR region [33]. KSHV genomes with major deletions have been found by PCR screening in some KS tumors and cell lines [102]. In a cohort of 16 individuals with KS in Zambia, 4 harbored KSHV genomes with uneven read coverages indicative of genomic aberrations [76]. A study of near full-length KSHV genome sequences from Zambia revealed frequent nonsense mutations in the K4.2 gene [76]. In Chapter 3 we extended these observations through detailed characterization of intra-host mutations that were tumor-specific, not found in the oral swabs of the same individuals. Identical mutations were found in multiple tumors from the same individual and similar mutations were found in multiple individuals. Most pronounced were mutations that appeared to inactivate the K8.1 gene in 3 individuals, large inversions, deletions and duplications present in tumors from 4 of 8 individuals [165], including sole persistence or over representation of a region near IR1.

The effect of KSHV sequence variation and *de novo* mutations is of potential significance to the clinical course of KS. KSHV encodes immunomodulatory, angiogenic and anti-apoptotic factors, and its gene expression and replication are tightly regulated [186]. Hence, KSHV polymorphisms observed *in vivo* coupled with detailed clinical data may reveal insights into the pathogenic process of KS. Extending the results from Chapter 3, I confirmed from a study of a total of 65 KS tumors from 30 individuals that K8.1 mutations and the over-representation of IR1 region (further localized to the K5-K6 gene region) do indeed occur at significantly high frequencies. Moreover, KSHV genes K4.2 and vIRF-2 were found to be unusually polymorphic between hosts and breakpoints in the genome were associated with sites of potential G-quadruplex formation. Finally, some genotypic variations were associated with clinical manifestations of KS.

Since we established through the use of dUMI-tagged sequences in the study reported in Chapter 3 that little point mutational variation exists within KSHV genomes found in vivo, the

following study did not include the use of dUMI as the goal was primarily to validate the larger scale or specific mutations found in Chapter 3.

Results

KSHV genomes in KS tumors often have significantly higher representation of a subgenomic region near IR1

The previous chapter described 5 unique tumor-associated KSHV genome aberrations, 3 of which involved a 2 to 30-fold greater read coverage of a sharply delineated region close to IR1 compared to the rest of the viral genome [165]. To better determine the frequency of discordant read representation in the IR1 region in tumors, a total of 67 tumors from 30 individuals, including the 12 from Chapter 3, were screened for IR1 region copy number elevation. Four gene segments of the viral genome, within K2, ORF16, ORF50 and ORF73 genes, were quantified by ddPCR. K2 and ORF16 are located near IR1, while ORF50 and ORF73 are located near the midpoint and 3' end of the KSHV genome, respectively (**Fig 16A**).

Of the 65 tumors with positive PCR results, 13 had ORF50 and ORF73 copy numbers that were a median of 6-fold less than K2 and ORF16, or below the limit of detection (**Table 7**). In contrast, 2 tumors had K2 and ORF16 that were 2-fold less than ORF50 and ORF73 levels or below the limit of detection. Two tumors sequenced in Chapter 3 had no DNA extracts left for ddPCR (U003-C and U030-C), but their whole genome sequencing results did not exhibit the IR1 region read over-coverage [165]. In the remaining 48 tumors, copy numbers of the 4 genes were within 50% of each other. Notably, when high copy number variation was detected, it did not apply to all tumors sampled from an individual.

Consensus KSHV genome sequences were obtained from 10 individuals with >2-fold higher copy numbers of K2 and/or ORF16 compared to ORF50 and/or ORF73, as well as 1 -2 additional tumors from the same individuals and 1 tumor each from the remaining group having the highest mean genome copy numbers. In total, 24 tumor DNA samples were prepared for whole KSHV genome sequencing. Greater than 99.9% of the KSHV genome (aside from the 3 repeat regions) was successfully sequenced from 20 tumors, including all with at least 769 genome copies per µL, as estimated by ddPCR (**Table 7**). Eight of the 20 tumors had elevated read coverage that encompassed a region between genome positions 26,000 and 32,000 (**Fig 1B**). In 5 of these (U099-D, U108-B, U156-D, U156-E, U210-B) the spike in read coverage was sharply confined to

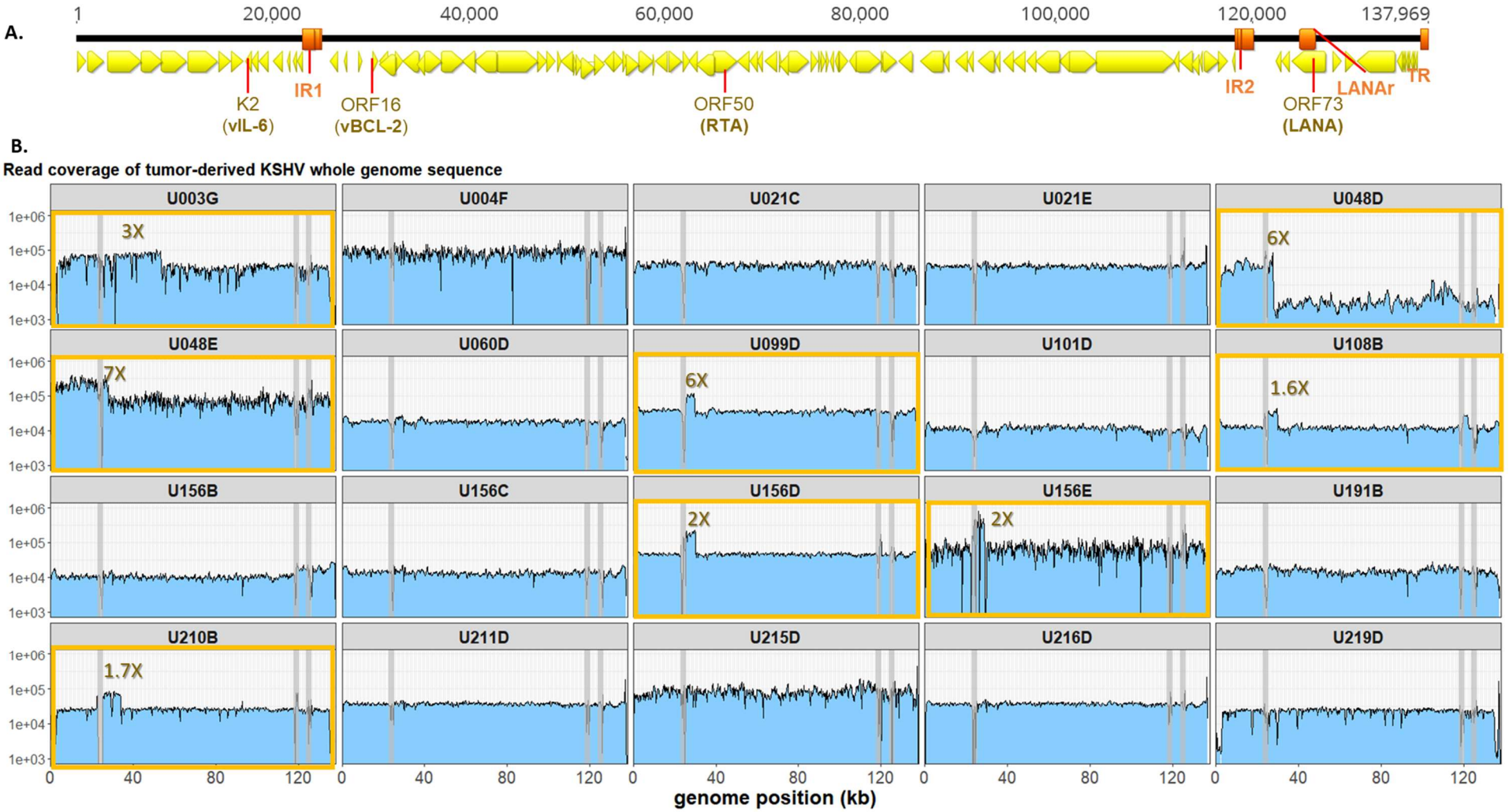


Figure 16. Raw read coverage of KSHV genomes detected in 20 tumors from 16 individuals

(A.) *Linear schematic of KSHV genome. Yellow arrows represent open reading frames, as annotated in KSHV reference strain GK18. The 4 gene regions used for copy number quantitation are labeled by gene name (protein product in parenthesis). Orange bars represent the major repeat regions Internal Repeat 1 (IR1), Internal Repeat 2 (IR2), LANA repeat domain (LANAr) and Terminal Repeats (TR).* ***(B.)*** *Read coverages of KSHV genome sequences from 20 tumor biopsies. Heights of the blue fill represent the raw read coverage along the KSHV genome. The yellow boxes highlight the 8 tumors with the KSHV IR1 region overrepresented in read coverage. The translucent grey vertical lines represent the major repeat regions where precise sequence mapping was unreliable.*

Table 7. Summary of KSHV genome characteristics in all 65 KS tumor

Sample count	Sample ID [a]	Amplicon copy number (by ddPCR) [b]				K2 + ORF16 / ORF50 + ORF73 [c]	KSHV genome (avg copies/µL)	Genome Sequenced	Overrepresented K5-K6 region [d]	K8.1	miR-K10	K4.2 size (bp)	K11.2 [e] 174-bp dup
		K2	ORF16	ORF50	ORF73								
1	U003-B	9,757	9,372	5,720	8,223	1.37	8,268			Breakpoint	G16A	549	no
2	U003-C	ND	ND	ND	9,664	ND	9,664	yes		Breakpoint	G16A	549	no
3	U003-E	2,541	3,971	2,750	5,132	0.83	3,599			Breakpoint	G16A	549	no
4	U003-F	3,702	2,063	2,618	3,119	1.00	2,876			Breakpoint	G16A	549	no
5	U003-G	31,735	31,185	10,896	12,595	**2.68**	21,603	yes	yes	Breakpoint	G16A	549	no
6	U004-C	1,243	1,274	1,205	998	1.14	1,180	yes		Intact	WT	549	yes
7	U004-D	4,433	4,466	4,543	4,290	1.01	4,433	yes		TSS deletion	WT	549	yes
8	U004-F	1,198	1,322	799	881	1.50	1,050	yes		Intact	WT	549	yes
9	U007-B	1,842	1,864	2,040	1,925	0.93	1,918	yes		Intact	WT	378	yes
10	U008-B	19,140	33,660	19,910	19,195	1.35	22,976	yes	yes	Intact	WT	378	no
11	U008-C	16,995	17,325	19,170	15,235	1.00	17,181			Intact	WT	378	no
12	U008-D	24,360	34,755	12,737	17,189	**1.98**	22,260	yes	yes	Intact	WT	378	no
13	U008-E	7,744	7,871	6,617	5,451	1.29	6,921			Intact	WT	378	no
14	U008-F	4,257	4,120	4,252	4,362	0.97	4,248			Intact	WT	378	no
15	U008-G	3,839	4,164	3,889	4,257	0.98	4,037			Intact	WT	378	no
16	U008-H	560	1	1,309	1,463	*0.20*	833			Intact	WT	378	no
17	U008-I	8,701	9,059	7,541	8,080	1.14	8,345			Intact	WT	378	no
18	U020-B	49,600	55,850	4,550	5,500	**10.49**	28,875	yes	yes	Intact	WT	237	no
19	U020-C	3,658	5,033	145	139	**30.62**	2,243	yes	yes	Stop codon	WT	237	no
20	U020-E	167	84	1	172	1.45	106			TSS deletion	WT	237	no
21	U020-F	11,479	13,640	8,943	11,105	1.25	11,292			NPD	ND	NPD	NPD
22	U021-C	5,691	5,943	6,867	6,101	0.90	6,151	yes		Intact	C15T	237	no
23	U021-D	1,190	1,134	1,348	1,680	0.77	1,338			Intact	C15T	237	no
24	U021-E	15,792	16,558	23,083	20,050	0.75	18,871	yes		Intact	C15T	237	no
25	U021-H	1	424	454	826	*0.33*	426	failed		Intact	C15T	237	no
26	U021-I	234	311	312	205	1.05	266			Intact	C15T	237	no
27	U030-C	ND	ND	ND	59,730	ND	59,730	yes		Intact	WT	369	yes
28	U032-B	476	520	494	466	1.04	489	yes		Intact	WT	549	yes
29	U034-B	1,920	1,887	1,870	1,793	1.04	1,868	yes		Intact	WT	378	yes
30	U034-C	13,083	13,335	8,148	6,552	**1.80**	10,280	yes		Intact	WT	378	yes
31	U039-B	5,313	5,502	4,946	4,799	1.11	5,140			Intact	WT	378	no
32	U048-B	2,791	2,353	2,285	2,524	1.07	2,488			Intact	WT	372	no
33	U048-C	549	84	79	71	**4.22**	196	failed	yes	Intact	WT	372	no
34	U048-D	237,350	22,470	20,276	21,000	**6.29**	75,274	yes	yes	Intact	WT	372	no
35	U048-E	7,142	717	602	592	**6.58**	2,263	yes	yes	Intact	WT	372	no
36	U060-C	26,460	25,830	24,990	25,725	1.03	25,751			Intact	G16A	369	yes
37	U060-D	24,885	25,305	26,040	24,360	1.00	25,148	yes		Intact	G16A	369	yes

38	U062-B	1,765	1,991	2,127	1,846	0.95	1,932			Intact	WT	369	no
39	U062-C	3,266	3,287	2,510	2,407	1.33	2,868			Intact	WT	369	no
40	U066-C	1,107	1,147	1,192	974	1.04	1,105			Intact	WT	378	yes
41	U094-B	1,829	1,846	1,855	1,419	1.12	1,737			Intact	WT	378	no
42	U099-D	999,999	999,999	140,910	195,090	**5.95**	584,000	yes	yes	Stop codon	WT	549	yes
43	U101-D	5,061	5,271	4,767	3,150	1.31	4,562	yes		Intact	C15T	549	no
44	U106-B	3,108	3,108	4,001	4,106	0.77	3,581			Intact	WT	549	no
45	U108-B	53,655	50,610	45,360	45,150	1.15	48,694	yes	yes	Stop codon	WT	549	no
46	U108-H	1	1,067	1	1	**534.00**	268	failed	yes	NPD	NPD	NPD	NPD
47	U146-C	3,003	2,982	2,489	2,128	1.30	2,651			Intact	WT	549	no
48	U156-B	3,045	3,119	2,877	4,862	0.80	3,476	yes		Intact	WT	378	no
49	U156-C	8,526	9,240	7,875	9,356	1.03	8,749	yes		Intact	WT	378	no
50	U156-D	163,083	159,750	164,500	171,583	0.96	164,729	yes	yes	TSS deletion	WT	378	no
51	U156-E	1,010	1,001	275	789	**1.89**	769	yes	yes	TSS deletion	G16A	378	no
52	U156-G	387	413	379	287	1.20	367			Intact	WT	378	no
53	U156-H	3,262	3,471	2,206	2,822	1.34	2,940			Intact	WT	378	no
54	U191-B	21,389	21,515	19,541	21,431	1.05	20,969	yes		Intact	WT	237	no
55	U191-C	9,261	9,534	9,597	9,807	0.97	9,550			Intact	WT	237	no
56	U191-D	623	662	455	676	1.14	604			Intact	WT	237	no
57	U191-E	324	599	372	411	1.18	427			Intact	WT	237	no
58	U191-F	39	1,091	43	26	**16.38**	300	failed	yes	NPD	NPD	NPD	NPD
59	U210-B	14,700	31,675	15,125	12,917	**1.65**	18,604	yes	yes	Stop codon	WT	237	no
60	U211-D	87,583	91,667	89,750	84,167	1.03	88,292	yes		Intact	WT	378	no
61	U215-D	36,752	51,538	4,100	4,150	**10.70**	4,135	yes	yes	Intact	WT	237	yes
62	U216-D	121,333	130,083	121,000	110,250	1.09	20,667	yes		Stop codon	C15T	237	no
63	U217-D	2,914	3,008	3,988	3,289	0.81	3,300			Intact	WT	375	no
64	U218-D	13,075	13,467	12,500	12,125	1.08	2,792			Intact	WT	378	no
65	U219-D	19,188	19,600	32,958	23,875	0.68	3,905	yes		Intact	WT	378	no

Key

ND = not determined; tumor extracts were exhausted in previous study [165]

WT = wildtype; database consensus

NPD = no product detected

Breakpoint = breakpoint of a genomic rearrangement in KSHV, interrupting the K8.1 coding sequence [165]

[a] Adults with KS contributed multiple tumor biopsies from distinct lesions to this study and were anonymized with a "U" number. The following letter is the tumor identifier

[b] 1 = below the limit of detection; 999,999 = over the limit of detection

[c] Ratio of ddPCR counts of (K2 and ORF16) over (ORF50 and ORF73). Ratios higher that 1.5 are bolded, ratios lower than 0.5 are italicized

[d] Incorporates both ddPCR screening data and whole genome sequencing read coverage data

[e] Duplication of the central 174-bp domains of K11.2 (vIRF-2)

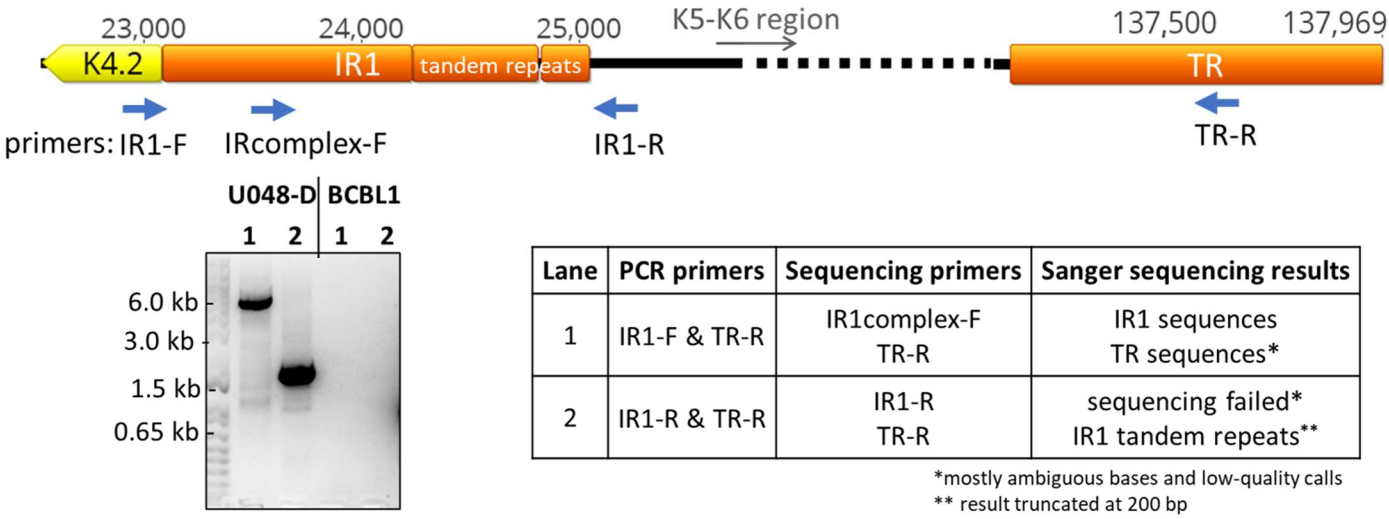
this region. Read over-coverage was also found in 2 tumors (U108-B and U156-D) in which copy numbers of the K2 and ORF16 regions were not elevated over the ddPCR amplicon regions (**Table 7**). Taking together from the ddPCR copy number variation and whole genome sequencing results, a total 16 of 65 tumors, from 10 individuals out of 30, were determined to have IR1 region over-representation.

The precise boundaries of the overrepresented regions, as well as any genomic aberrations not apparent from read coverage changes, were determined by mapping supporting split reads (see Methods). Two tumors from participant U048 (U048-D and U048-E) were found to have the same breakpoint. Two of 4 tumors from U156 (U156-D and U156-E) also had identical breakpoints, while 2 of his other tumors had relatively even read coverage across the genome (**Fig 1B**).

Putative rearrangements in the original tumor DNA from U048-D were chosen for validation by PCR across breakpoints and sequencing. PCR products were produced that confirmed breakpoint junctions directly connecting IR1 and TR sequences (**Fig 17**). More breakpoints junctions connecting IR1 and K6 sequences in the opposite orientation were found and confirmed by PCR and sequencing (not shown). The structures, which may involve multiple rearrangements of the K5-K6 region and at least one inversion, were not further characterized due to the length of the repeats and high GC content in IR1. Lastly, minor variants in U048-D containing an 86-bp inversion between ORF9 and ORF10 coding sequences were detected and confirmed by PCR and Sanger sequencing.

A 2.2-kb segment region encompassing K5 and K6 corresponds to the minimal region of overrepresentation.

In the 32 tumor-derived KSHV genomes sequenced between this and the previous chapter, nine unique KSHV genome aberrations had an abrupt, $\geq$1.5-fold read coverage over-representation in a subgenomic region near IR1. Strikingly, all these elevated read coverage regions minimally encompassed a 2.2kb segment downstream of IR1 and included the K5 and K6 genes (**Fig 18**). The recombinant KSHV strain BAC36 had been reported to exhibit read overrepresentation in the same region [78] (**Fig 18**). Four over-coverage regions partially or did not include the T1.4 long non-coding RNA, and the PAN long non-coding RNA was included in all but one over-coverage region.



Lane	PCR primers	Sequencing primers	Sanger sequencing results
1	IR1-F & TR-R	IR1complex-F TR-R	IR1 sequences TR sequences*
2	IR1-R & TR-R	IR1-R TR-R	sequencing failed* IR1 tandem repeats**

*mostly ambiguous bases and low-quality calls
** result truncated at 200 bp

Figure 17. PCR and sequencing confirmation of rearrangement connecting IR1 to TR sequences in tumor U048-D

A genome section of the KSHV isolate GK18 from K4.2 through IR1, and a TR unit sequence is shown. ORFs are in yellow, repetitive sequences in orange, and primer sequences in blue arrows. The K5 and K6 ORFs are further downstream. PCR products were separated on 0.8% agarose gel, and all visible bands were extracted and sequenced, with the results shown in the table. Cell line BCBL1 DNA was used as control. PCR conditions for GC-rich sequences: 30 cycles, 5 min extension time, annealing temperature 63°C

The average length of the genome segments with read coverage overrepresentation was 17.6 kb (**Fig 18**), or ~13% of the 137 kb KSHV genome excluding the terminal repeats (TR). Given ten random 17.6-kb windows of the KSHV genome sequence, that all 10 would include the 2.2-kb K5-K6 region is highly improbable (1 in 4.3 x 10^{37}). No other non-repeat region segment >3 kb in size had >1.5-fold read over-coverage in any of the 32 tumors with whole KSHV genomes sequenced in this cohort (U004-D had 1.5X read over-coverage on the first 3 kb from the 5' end; **Figs 5A & 16B**). The K5-K6 region overrepresentation, detected by ddPCR screening or whole viral genome sequencing, was found in at least one tumor from 10 individuals (**Table 7**), or 1/3 of this cohort of 30 individuals. Of note, participant U020 had two tumors with different breakpoints yet both contained the minimally overrepresented region (**Figs 5A & 18**).

There were 4 clusters of breakpoints defining the endpoints of each overrepresented region: encompassing K4.1-K4.2, IR1, the minor internal repeat between K7 to ORF16, or at ORF18-ORF19 (**Fig 18**). The 5' breakpoints from two individuals (U156-B and U210-B) were 0.2 kb apart within genes K4.1 and K4.2. Three 5' breakpoints were within the second GC-rich tandem repeat

family of IR1, with one more ~500 bp downstream. Three 3' breakpoints were within 0.3 kb of each other at or near the minor internal repeat. Four 3’ breakpoints were within OR18 or ORF19, two of which were only 3 bp apart (U008-B and U210-B).

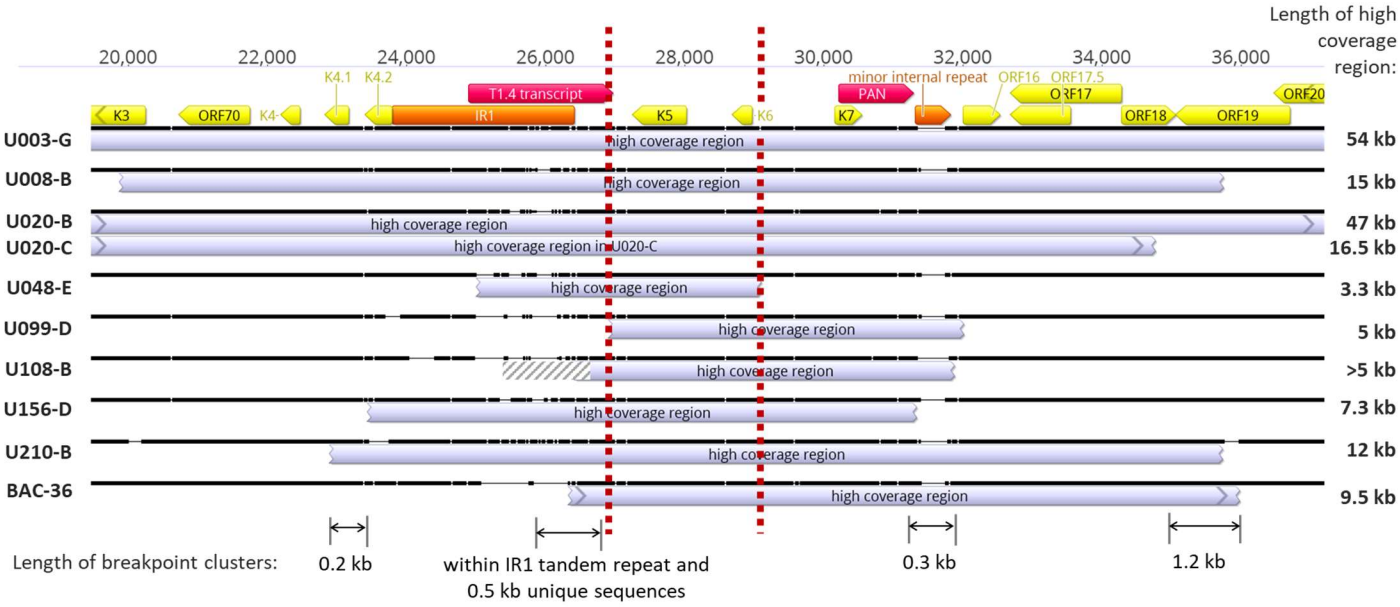


Figure 18. Overlap in KSHV IR1 overrepresentation regions

Sequence alignment of regions with excess read coverage near IR1 (showing those with identical breakpoints only once). Location and orientations of ORFs are in yellow, long non-coding RNAs in red, and repetitive sequences in orange. In addition to IR1, a region labelled 'minor internal repeat' is composed of the 13-bp sequence TGGGATGGGGGTG repeated 4 to 13 times. Vertical dashed red lines demarcate the overlap among all IR1 overrepresented regions. Below are indicated the maximum distances between breakpoints in the 3 identified breakpoint clusters, aside from those that could not be precisely mapped in IR1.

KSHV rearrangement breakpoints are associated with G-quadruplexes

Chromosomal rearrangement breakpoints are prone to cluster at fragile regions, typically sequences that form non-B DNA structures [187]. Genomes of gamma-herpesviruses are enriched, more than expected from nucleotide composition, with GC-rich DNA motifs known to induce double-stranded breaks that undergo repair and recombination in B-cells [92,94]. I therefore assessed whether KSHV rearrangement breakpoints were associated with these DNA features.

To determine any association between KSHV rearrangement breakpoints and potential non-B DNA sequences, sequences ±500 bp from all 31 breakpoints identified here and in other studies [78,165,188] (**Table 3**) were scanned for the following 5 sequence motifs associated with non-B DNA sequences: cruciforms, Z-form DNA, AT-rich local melt regions, mirror repeats or triplexes, and G-quadruplexes (G4). This window length was chosen because breakpoints have been associated with G4 DNA up to 500 bp distant [189]. Control windows were generated from a sequence permutation of the observed breakpoint windows and from the same number of random 500 bp windows of the KSHV genome. Cruciform DNA was not predicted to be formed in any of the breakpoint or control windows (not shown). However, breakpoints were most common at or near Z-DNA and G4 sequences (**Fig 19A and B**). The 2 breakpoints 3 bp apart at ORF19 were found within a high probability Z-DNA sequence. On average, Z-DNA probabilities had a local peak at the observed breakpoint although a nearby peak was also generated from the random control dataset (**Fig 19C**), while G4 scores peaked at the breakpoint position in contrast to controls (**Fig 19D**).

To quantify any association with breakpoints, total non-B DNA probabilities and relative scores within 200 bp in both directions from the observed or control breakpoints were summed. Also measured were distances from each breakpoint to the nearest position with >0.20 non-B DNA probability or relative score in either direction. The summed scores and the nearest distances of the 31 observed breakpoints were compared to those of 31 random breakpoints (**Table 3**) using a Welch 2-sample T-test, assuming unequal variance (**Table 8**). Among the summed scores of the 5 non-B DNA motifs analyzed, G4 had the largest difference between the means of the observed and random datasets, and the difference was statistically significant (p=0.0040). The closest distances to local melt regions and regions of relative G4 scores >0.20 were significantly smaller for observed breakpoints than for random dataset (p=0.024, p=0.0007, respectively). To get a more robust assessment of association with G4, 1000 simulations of 31 randomly sampled points along the GK18 genome were made in R, generating a normal distribution of means. The probability of attaining equal or higher means than the observed mean G4 summed scores was 1.4×10^{-6}, and the probability of attaining equal or lower than the observed mean distances to G4 was 0.0009.

Table 8. P-values of non-B-DNA summed probabilities/scores and distances to breakpoints

Sum total of probabilities or relative scores of non-B-DNA structures +/-200 bp from breakpoint

Non-B-DNA	Mean obs.	Mean random	95% CI lower	95% CI upper	P-value
melt	0.04	0.12	-0.26	0.10	0.380
Z-DNA	1.58	1.84	-2.16	1.63	0.782
cruciform	0	0	0	0	NA
triplex	0.93	0.3	-0.82	2.08	0.385
G4	16.09	2.64	4.6	22.31	0.004

Average distance to closest non-B-DNA structure with probability or relative score >0.2

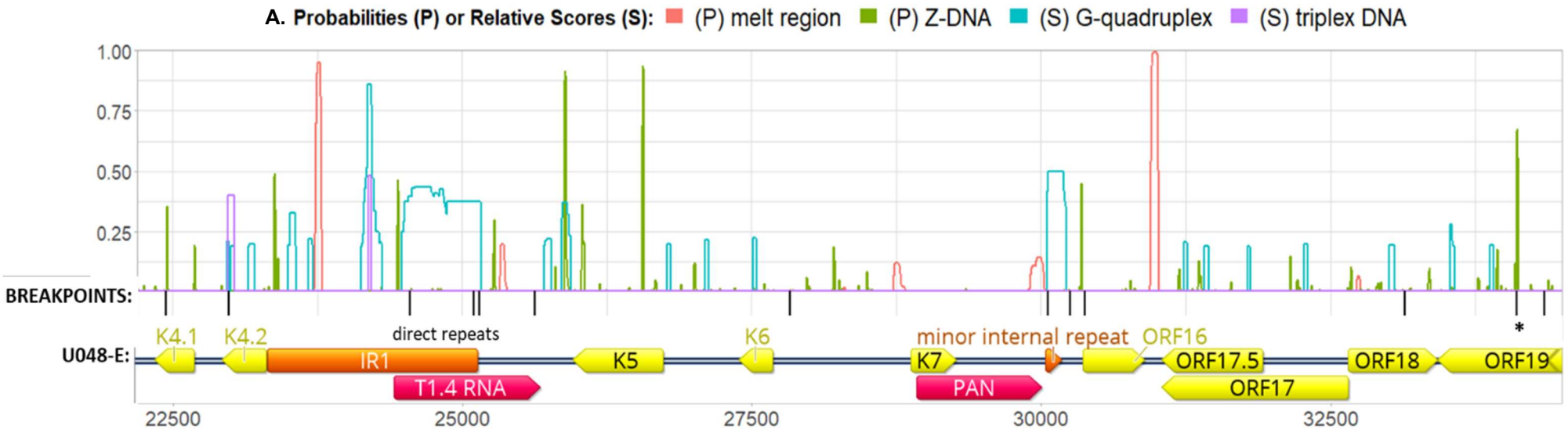
Non-B-DNA	Mean obs.	Mean random	95% CI lower	95% CI upper	P-value
melt	450.26	498.20	-89.11	-6.76	0.024
Z-DNA	278.52	327.56	-154.86	56.77	0.357
cruciform	>500	>500	NA	NA	NA
triplex	433.87	442.20	-77.06	60.40	0.809
G4	308.65	460.48	-235.30	-68.37	0.001

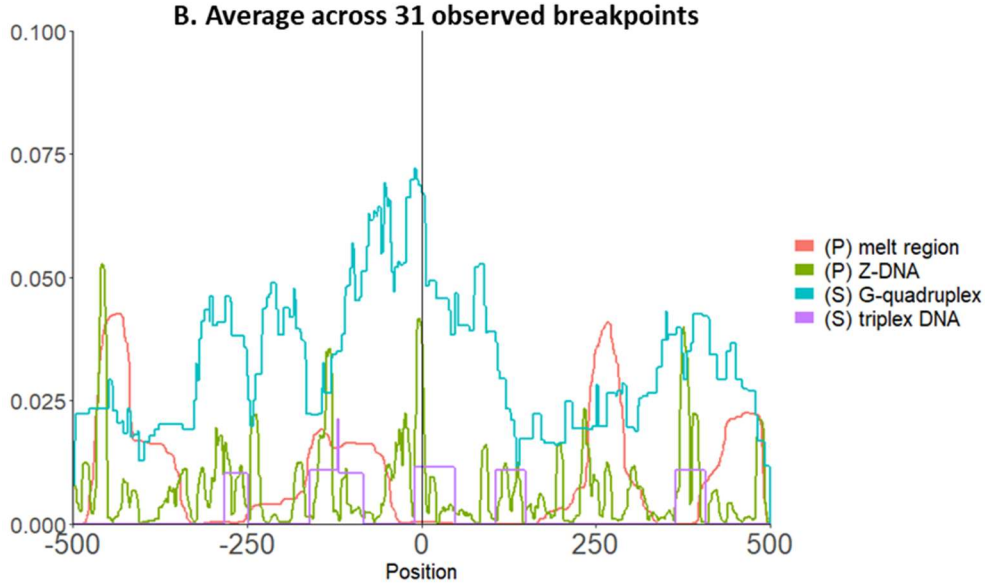
Note: probabilities for melt, Z-DNA and cruciform; relative scores for triplex and G4

Potentially inactivating mutations are common in the K8.1 gene in tumors.

I observed in Chapter 3 that mutations in the K8.1 gene were common in KS tumors, all putatively disrupting K8.1 gene expression due to large truncations, frameshifts, or promoter region deletions. As with the genomic alterations noted above, in individuals where K8.1 mutations were observed, not all their tumors had the K8.1 mutation, indicating that they most likely arose *de novo* in each individual. Remarkably, none of the other 85 KSHV coding sequences, three-quarters of which are longer than K8.1, had intra-host mutations in more than one person. To better assess the frequency of K8.1 mutations, a 1.4 kb region of the K8.1 coding sequence and its promoter region was determined from all 65 tumors from 30 individuals. Parallel analysis of a 250 bp K12 gene sequence served as sample control.

Nine unique K8.1 mutations were detected in 8 individuals (**Table 7**, **Fig 20A**). Five were nonsense mutations in the second exon, and three were 28 - 32 bp deletions between the K8.1 core promoter and the K8.1 coding sequence, spanning the transcription start site (**Fig 20A**). One of the deletions also included the first base of the K8.1 coding sequence (U020-E in **Fig 20B**). Finally, participant U003 had a genomic inversion interrupting the second K8.1 exon and extending to TR sequences [165]. U020 had tumors with different K8.1 mutations – a nonsense mutation in U020-C and a deletion in the promoter region of U020-E (**Fig 20A**).


A. Probabilities (P) or Relative Scores (S):
(P) melt region
(P) Z-DNA
(S) G-quadruplex
(S) triplex DNA
1.00
0.75
0.50
0.25
BREAKPOINTS:
U048-E:
K4.1
K4.2
direct repeats
IR1
T1.4 RNA
K5
K6
minor internal repeat
ORF16
K7
PAN
ORF17.5
ORF17
ORF18
ORF19
*
22500
25000
27500
30000
32500


B. Average across 31 observed breakpoints
0.100
0.075
0.050
0.025
0.000
-500
-250
0
250
500
Position
(P) melt region
(P) Z-DNA
(S) G-quadruplex
(S) triplex DNA

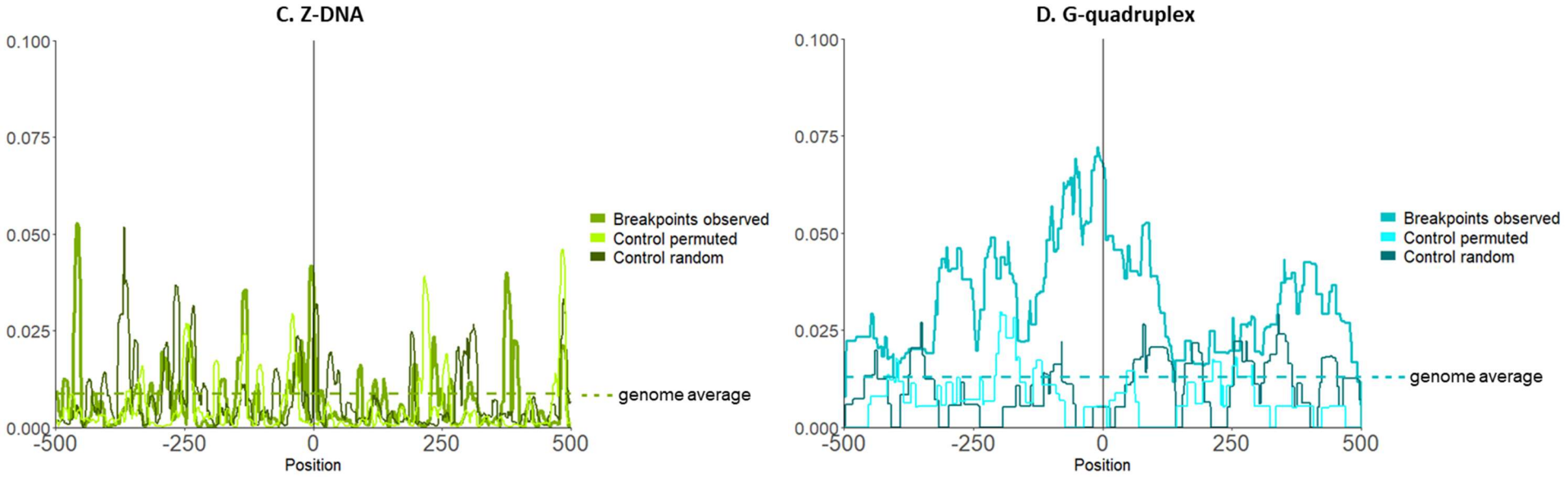


Figure 19. Association of breakpoints to non-B-DNA

(A) *The probabilities and relative scores of 5 non-B-DNA structures (local melt region, Z-DNA, G-quadruplex and triplex DNA) were graphed along the genome of KSHV isolate U048-E. Shown is a genome segment including the overrepresented region including K5 and K6. The black ticks mark the breakpoints observed, and below are the annotations for ORFs (yellow), repeat regions (orange) and long non-coding RNAs (red). The asterisk at ORF19 denotes 2 breakpoints 3 bp apart found in 2 different individuals. G-quadruplex scores were normalized to 300, near the human genome maximum (https://pqsfinder.fi.muni.cz/genomes), and triplex scores were normalized to a maximum of 50.* ***(B)*** *Relative scores were plotted in DNA sequences +500 bp of 31 observed breakpoints along the entire KSHV genome excluding the TR and averaged across all windows with score normalization as in panel A. Of note, all averages were below 0.10 since probabilities and scores were 0 at most breakpoint windows.* ***(C)*** *Z-DNA probabilities and* ***(D)*** *G-quadruplex scores in control permuted sequences, 31 random windows from the GK18 genome. The genome-wide averages are shown with the horizontal broken lines.*

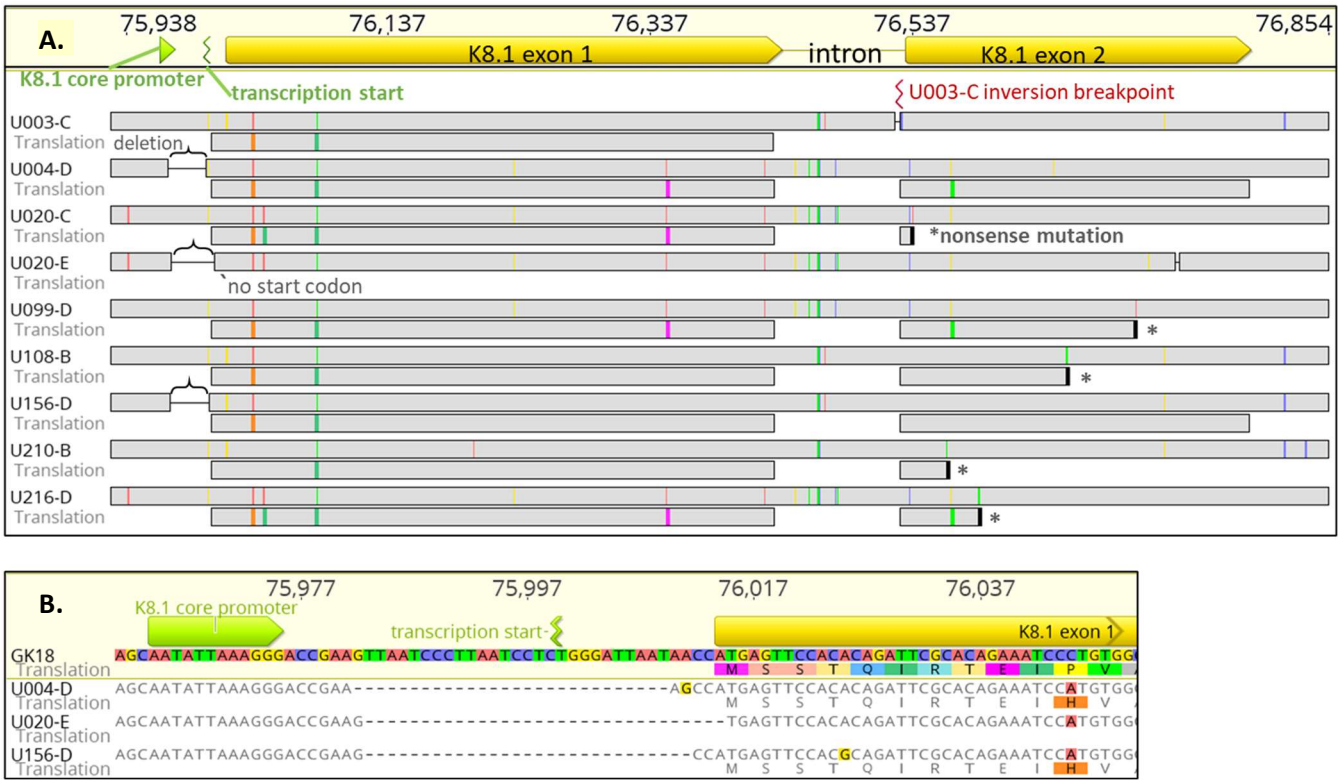


Figure 20. Inactivating mutations observed in the K8.1 gene

(A) *Alignment of the K8.1 gene and promoter mutations found in this study, showing identical mutations from the same person once. Indicated above are locations of the K8.1 core promoter, transcription start site and open reading frame. Sequence coordinates are for reference isolate GK18. Gray bars adjacent to tumor IDs represent their nucleotide sequences, below which are the representations of their translations. Colored ticks represent nucleotide or amino acid differences from the GK18 sequence, brackets indicate transcription start site deletions, asterisks indicate nonsense mutations, and red mark corresponds to the inversion breakpoint location in U003 tumors.* ***(B)*** *Alignment zoomed in on the transcription start site deletions.*

Polymorphisms in miR-K10, K4.2 and K11.2.

Promoter region deletions, indels and nonsense mutations in the 85 KSHV genes aside from K8.1 were compiled for all 32 KSHV genomes sequenced. No promoter region deletions or nonsense mutations were found in the other genes. There were seven different indel mutations, 4 of which were in-frame (**Table 9**). A single nucleotide extension of an A homopolymer run in K11.2 was found in only 1 of 4 tumors from U156 (**Table 9**).

Table 9. Coding sequence mutations*

In frame deletions				
ORF34	Related to HHV-5 UL95	3 nt	U004-C,D,F	3 of 3 tumors
ORF47	Envelope glycoprotein gL; herpesvirus core gene UL1 family	15 nt	U004-C,D,F	3 of 3 tumors
ORF50	Transactivator; RTA, induces switch from latent to lytic infection	24 nt	U099-D	1 of 1 tumor
K9	Interferon regulatory factor, vIRF-1	18 nt	U216-D	1 of 1 tumor
Frameshifting indels				
ORF4	Complement control protein; membrane protein; contains four SCR domains; KCP	23 nt deletion, truncated ORF	U048-D,E	2 of 2 tumors
K7	IAP-like inhibitor of apoptosis; contains hydrophobic domain; vIAP	1 nt homopolymer deletion, extended ORF	U216-D	1 of 1 tumor
K11.2	Interferon regulatory factor, vIRF-2	1 nt homopolymer insertion, truncated ORF	U156-D	1 of 3 tumors
Intra-host non-synonymous point mutations				
ORF11	herpesvirus dUTPase	T396P	U020-C	1 of 3 tumors and 1 oral swab
K3	E3 ubiquitin ligase, membrane protein MIR1, downregulation of MHC1; ORF12	F88L	U020-C	1 of 3 tumors and 1 oral swab
ORF25	major capsid protein; herpesvirus core gene UL19 family	Q594K	U020-B	1 of 3 tumors and 1 oral swab
ORF32	tegument protein; herpesvirus core gene UL17 family; DNA packaging	R56Q	U004-D	1 of 3 tumors and 1 oral swab
ORF63	tegument protein; herpesvirus core gene UL37 family	T848A	U032-B	1 of 1 tumor and 1 oral swab
K15	LAMP; signal transducing membrane protein	A290P	U004-D	1 of 3 tumors and 1 oral swab
Intra-host synonymous point mutations				
ORF8	envelope glycoprotein gB; herpesvirus core gene UL27 family	nucleotide: C762AT	U021-E	1 of 2 tumors
K12	Kaposin A, hydrophobic membrane protein; contains microRNA K10	nucleotide: G126A (K12); G16A (miR-K10)	U003-B,C,E,G; U156-D	4 of 4 tumors and 0 of 3 oral swabs; 1 of 4 tumors

* Not shown are mutations arising from rearrangement breakpoints, mutations in K8.1, K4.2 or K11.2, all of which are discussed separately.

The intra-host differences between matching sample-consensus genomes were determined within 10 individuals who had whole KSHV genomes sequenced from more than one sample, including the oral swabs also evaluated in the previous chapter. Six nonsynonymous point mutations were found in 6 different genes (**Table 9**). Only 3 instances of synonymous point mutations were detected, 2 of which were the same G126A nucleotide substitution in K12 from 2 different individuals. This mutation also impacted the overlapping microRNA K10 sequence (G16A mutation in miR-K10, or position C118082T in the GK18 genome). G16A was found in 1 of 4 tumors from U156, and in all of 4 tumors (**Table 9**) and none of 3 oral swabs (**Table 10**, **Fig 15**) in participant U003.

PCR screening was conducted on all 65 tumors (**Table 7**) and 18 oral swabs (**Table 10**) to determine the extent of miR-K10 diversity in this cohort. Six of 30 individuals had either a miR-K10 C15T or G16A mutation while the remainder had the miR-K10 database consensus nucleotide.

Table 10. PCR screening and sequencing in oral swabs

Sample ID	K8.1	miR-K10
U003-oral1	*intact*	*wt consensus*
U003-oral2	*intact*	*wt consensus*
U003-oral3	*intact*	*wt consensus*
U004-oral	*intact*	*wt consensus*
U007-oral	*intact*	*wt consensus*
U008-oral	*intact*	*wt consensus*
U020-oral1	*intact*	*wt consensus*
U020-oral2	*intact*	*wt consensus*
U023-oral	*intact*	*wt consensus*
U032-oral	*intact*	*wt consensus*
U034-oral1	*intact*	*wt consensus*
U034-oral2	*intact*	*wt consensus*
U191-oral1	no product	wt consensus
U191-oral2	intact	wt consensus
U191-oral3	intact	wt consensus
U191-oral4	intact	wt consensus
U211-oral1	intact	wt consensus
U211-oral1	intact	wt consensus

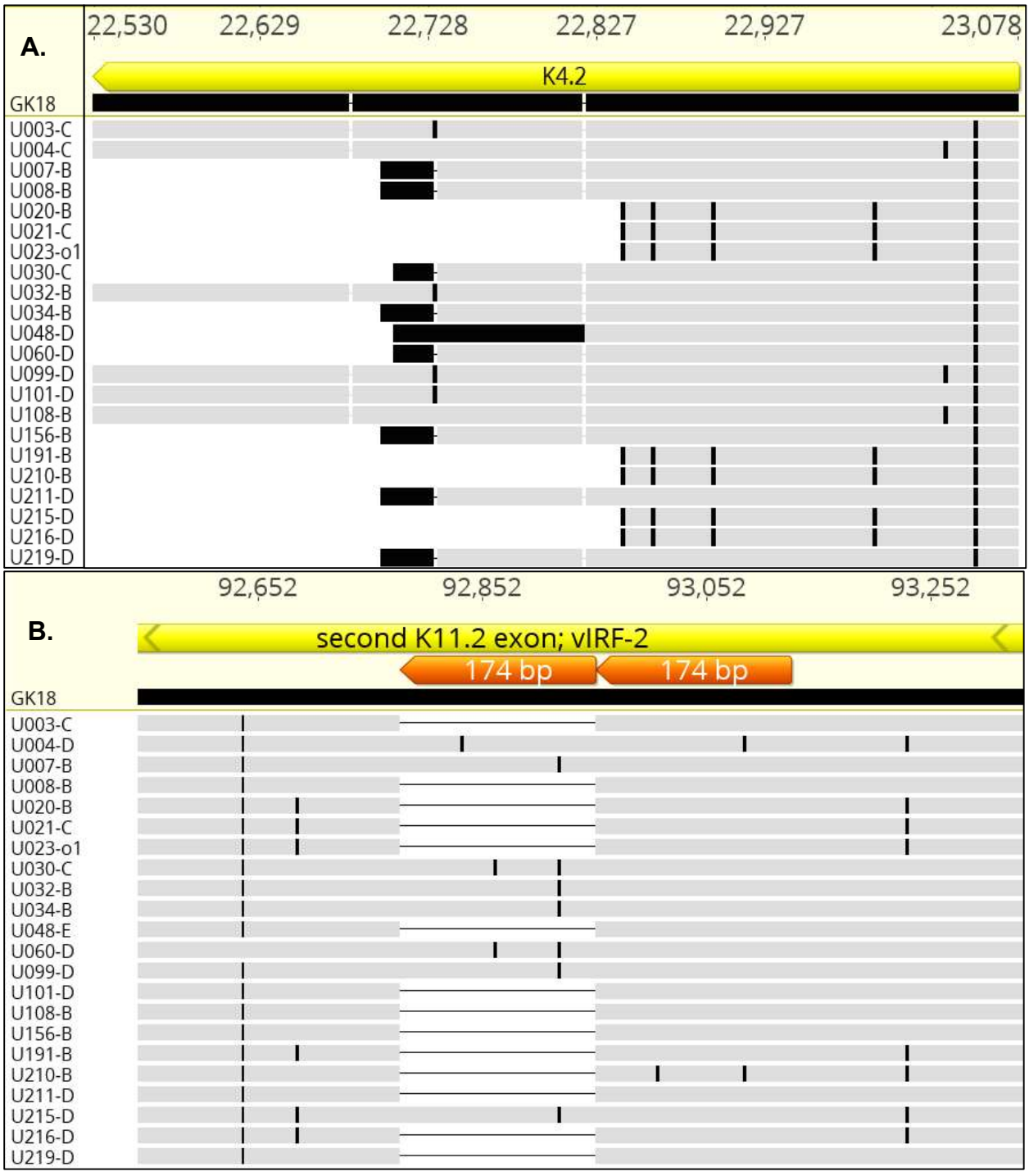
Note: Italicized samples were sequenced and reported in Chapter 3

Truncations of the K4.2 coding sequence, relative to the GK18 reference, were reported in 12 of 16 adults with KS in Zambia [76]. Here, 16 of 22 individuals that had whole KSHV genomes sequenced were found to have K4.2 truncations resulting from frameshifts and premature stop codons (**Fig 21A**). Another gene polymorphism was also found here, in a 174-bp sequence duplication in the central domain of K11.2 that was present 8 of the 22 individuals (**Fig 21B**). In inspecting the 16 genomes from Zambia, the duplication was found in 12 (a partially overlapping subset with those that had truncated K4.2). To determine the overall extent of K4.2 and K11.2 diversity in this cohort, both genes were sequenced following gene-specific PCR in all 65 tumors (**Table 7**). The K4.2 length polymorphisms observed were recurring between individuals, and no intra-host differences were found. K4.2 was truncated in 23 of the 30 total individuals in our cohort, while the central 174-bp duplication in K11.2 was absent in 11 of 30.

Among the 96 KSHV genomes sequenced to date, 3 predominant length polymorphisms of the K4.2 gene were found: 31 were full-length at 546 or 549 bp, 47 had a truncation to 369 or 378 bp (Truncation A and B), and 18 had a truncation to 237 bp or shorter (Truncation C) (**Table 11**). Of the 94 KSHV genomes with a complete K11.2 gene, 76 had the 174-bp duplication (**Table 11**), including 31 with 1-2 single nucleotide imperfections. These K4.2 and K11.2 polymorphisms did not coincide with major KSHV phylogenetic groupings [77] (**Fig 8 & 21C; Table 11**).

KSHV mutations associated with disease course and tumor characteristics

An exploratory statistical analysis was conducted for associations between the mutations and genetic polymorphisms observed and clinical traits (**Table 12**) and tumor characteristics (**Table 13**). The genetic markers evaluated were K5-K6 region read over-coverage, K8.1 inactivation, miR-K10 point mutations, K4.2 truncation, K11.2 central duplications, and marker co-occurrence. Clinical or individual traits evaluated were: gender, age, plasma KSHV load, HIV load, CD4+ cell counts, ACTG staging (tumors extensive or not [T], systemic symptoms [S] and immune status or CD4+ T-cell count below 200 cells/mL []; see Table 12 keys) [145], survival, treatment responses, prevalence (number of anatomic areas lesions were found), presence of any head, neck or oral lesions, and presence of lesions in the extremities. Tumor characteristics evaluated were morphotype (nodular, fungating or macular), size (greater or less than 1 cm diameter), anatomic area (head/neck, hard palate, oral excluding hard palate, back, chest/abdomen, groin/genitals, upper limbs, and lower limbs), and sampling time from first visit.


A.
22,530
22,629
22,728
22,827
22,927
23,078
K4.2
GK18
U003-C
U004-C
U007-B
U008-B
U020-B
U021-C
U023-o1
U030-C
U032-B
U034-B
U048-D
U060-D
U099-D
U101-D
U108-B
U156-B
U191-B
U210-B
U211-D
U215-D
U216-D
U219-D
B.
92,652
92,852
93,052
93,252
second K11.2 exon; vIRF-2
174 bp
174 bp
GK18
U003-C
U004-D
U007-B
U008-B
U020-B
U021-C
U023-o1
U030-C
U032-B
U034-B
U048-E
U060-D
U099-D
U101-D
U108-B
U156-B
U191-B
U210-B
U211-D
U215-D
U216-D
U219-D

C.

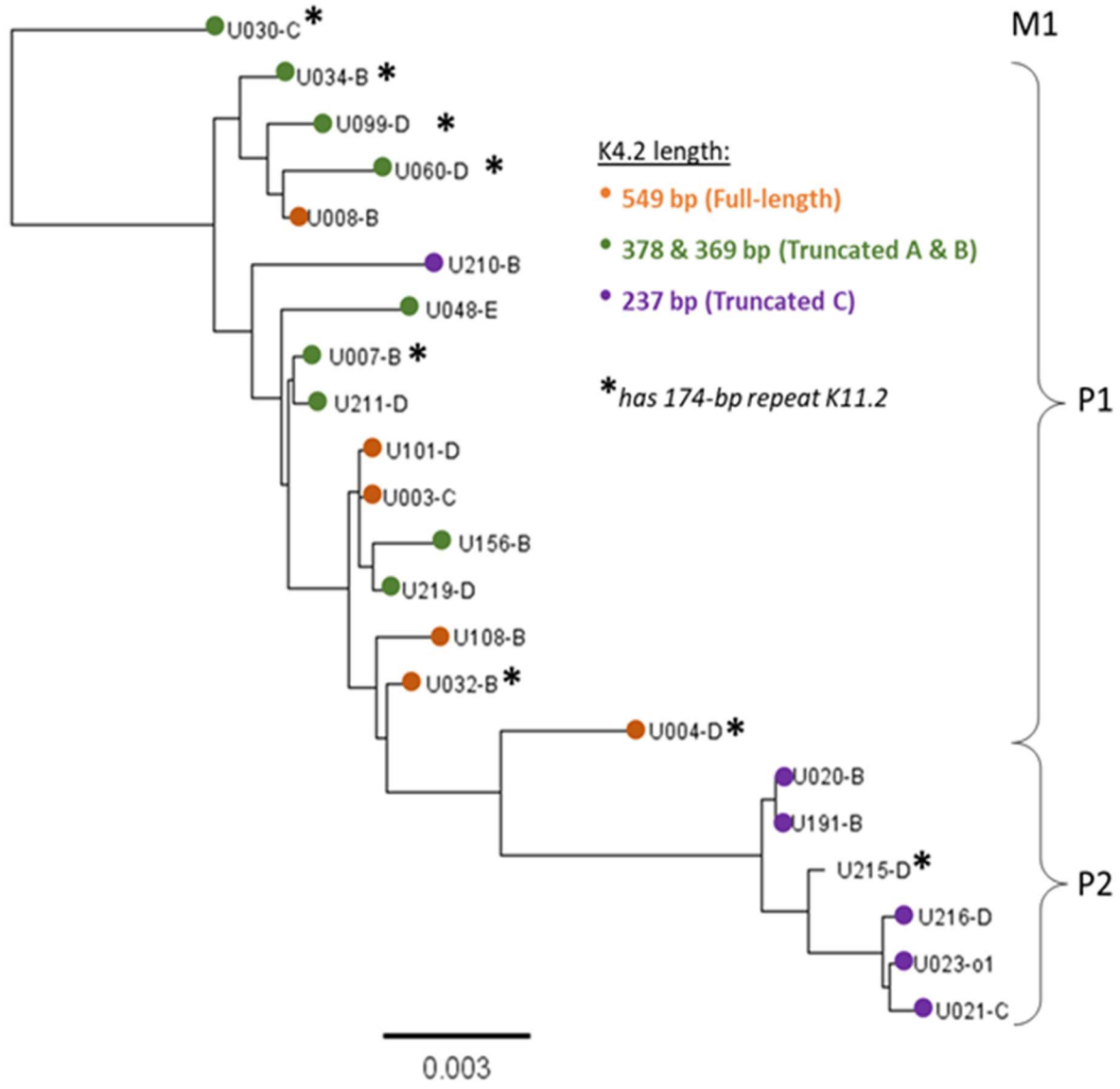


Figure 21. Polymorphisms in K4.2 and K11.2

Alignment of predicted K4.2 ***(A.)*** *and* ***(B.)*** *K11.2 amino acid sequences in whole KSHV genomes sequenced in this study aligned against the GK18 reference, showing 3 length genotypes. Black marks indicate amino acid differences from GK18. The 22 sequences in (A) are representative of K4.2 length polymorphisms found in the 98 KSHV genomes sequenced to date. The 174-bp repeats found in K11.2 are in orange.* ***(C.)*** *Phylogenetic tree of KSHV genomes in this study. Colored circles indicate K4.2 length genotypes, and asterisks indicate the presence of 174-bp duplication inside K11.2. Major genome types* [77] *are indicated on the right.*

Table 11. Linkages of KSHV genome types with K4.2 and K11.2 polymorphisms

NAME	GENOTYPE	K4.2 CDS	K11.2 REPEAT
Japan1	M1	FL	yes
Miyako1	M1	FL	yes
Miyako2	M1	FL	yes
Miyako3	M1	FL	yes
UG118	M1	FL	yes
UG119	M1	FL	yes
UG145	M1	FL	yes
UG157	M1	FL	yes
UG15	M1	Tr B	yes
UG16	M1	Tr B	yes
UG219	M1	Tr B	yes
BC-1	M1	Tr C	yes
UG237	M2	FL	yes
U030-C	M2	Tr B	yes
UG131	M2	Tr B	yes
UG160	M2	Tr B	yes
ZM095	N	FL	yes
ZM004	N	Tr A	no
ZM128	N	Tr A	no
BAC16	P1	FL	yes
BAC36	P1	FL	yes
SPEL	P1	FL	yes
U004-D	P1	FL	yes
U032-B	P1	FL	yes
U099-D	P1	FL	yes
U93872	P1	FL	yes
UG129	P1	FL	yes
UG151	P1	FL	yes
UG156	P1	FL	yes
UG158	P1	FL	yes
UNC_KICS0009	P1	FL	yes
ZM118	P1	FL	yes

NAME	GENOTYPE	K4.2 CDS	K11.2 REPEAT
U003-C	P1	FL	no
U101-D	P1	FL	no
U108-B	P1	FL	no
GK18	P1	Tr A	yes
U007-B	P1	Tr A	yes
U034-B	P1	Tr A	yes
UG110	P1	Tr A	yes
UG12	P1	Tr A	yes
UG125	P1	Tr A	yes
UG126	P1	Tr A	yes
UG128	P1	Tr A	yes
UG13	P1	Tr A	yes
UG133	P1	Tr A	yes
UG134	P1	Tr A	yes
UG137	P1	Tr A	yes
UG141	P1	Tr A	yes
UG148	P1	Tr A	yes
UG152	P1	Tr A	yes
UG159	P1	Tr A	yes
UG162	P1	Tr A	yes
UG163	P1	Tr A	yes
UG165	P1	Tr A	yes
UG166	P1	Tr A	yes
UG168	P1	Tr A	yes
UG212	P1	Tr A	yes
UG222	P1	Tr A	yes
UG226	P1	Tr A	yes
UG244	P1	Tr A	yes
ZM106	P1	Tr A	yes
ZM123	P1	Tr A	yes
ZM124	P1	Tr A	yes
U008-B	P1	Tr A	no

NAME	GENOTYPE	K4.2 CDS	K11.2REPEAT
U156-B	P1	Tr A	no
U211-D	P1	Tr A	no
U219-D	P1	Tr A	no
U060-D	P1	Tr B	yes
UG146	P1	Tr B	yes
ZM116	P1	Tr B	yes
U048-E	P1	Tr B	no
DG-1	P1	Tr C	yes
U210-B	P1	Tr C	no
UG122	P2	FL	yes
UG155	P2	FL	yes
ZM091	P2	FL	no
ZM121	P2	Tr A	no
ZM027	P2	Tr B	yes
ZM102	P2	Tr B	yes
ZM108	P2	Tr B	yes
ZM117	P2	Tr B	yes
U215-D	P2	Tr C	yes
UG114	P2	Tr C	yes
UG117	P2	Tr C	yes
UG120	P2	Tr C	yes
UG132	P2	Tr C	yes
UG136	P2	Tr C	yes
UG149	P2	Tr C	yes
UG164	P2	Tr C	yes
ZM114	P2	Tr C	yes
ZM130	P2	Tr C	yes
U020-B	P2	Tr C	no
U021-C	P2	Tr C	no
U023-o1	P2	Tr C	no
U191-B	P2	Tr C	no
U216-D	P2	Tr C	no

Key: Full-length (FL) = 546 or 549 bp, Truncation A (Tr A) = 369 bp, Truncation B (Tr B) = 369 bp, Truncation C (Tr C) = 237 bp or shorter

K5-K6 over-representation and K8.1 inactivation were significantly more common in nodular and fungating compared to macular lesions (OR=20.24, p<0.001 and OR=6.38, p=0.0005 respectively; **Table 14**). Both were also more common in participants that had limited KS lesions (found in ≤4 of 8 anatomic regions: OR=0.20, p=0.01 and OR=0.05, p=0.007, respectively). Since all K4.2 truncations observed removed its putative transmembrane region, K4.2 genotype was classified into having either a full-length or truncated coding sequence. A truncated K4.2 was less common in nodular compared to macular lesions (OR=0.28, p=0.028, **Table 14**). While barely significant, it can be noted that the K11.2 174-bp domain duplication was more common among female participants (OR=7.56, p=0.048), and was not found in any of the 4 HIV-negative participants. The miR-K10 mutations C15T and G16A were less prevalent in lesions >1 cm (OR= 0.20, p=0.01) (**Table 14**) and weakly associated with lower survival rates (HR=4.11, p=0.053, **Table 15**). No other statistically significant correlations were observed between KSHV mutations and clinical phenotypes, but among associations between mutations, a truncated K4.2 was inversely correlated with having K8.1 inactivating mutations (OR=0.12, p=0.008, **Table 16**).

Table 12. Participant clinical traits

PTID	Died	Days [a]	Best response [b]	Age	Sex	HIV	CD4 count	CD8 count	HIV RNA	ACTG Stage [c]			KSHV plasma VL [d]	Edema	Lesions in:								Lesion Sites (of 8)	Total Lesion Number
										T	I	S			Head/ Neck	Hard Palate	Oral, ex-Hard Palate	Back	Chest	Groin	Arms	Legs		
U003	Yes	114	PR	25	M	+	45	837	759,635	1	1	1	12,728	Yes	No	No	No	No	Yes	No	Yes	Yes	3	109
U004	No	669	PD	37	M	+	85	1,197	277,655	1	1	0	81,324	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	5	84
U007	No	444	PD	26	M	+	136	589	91,096	0	1	1	10,927	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	6	38
U008	Yes	188	PD	56	M	+	422	800	860,937	1	0	1	17,815	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	6	83
U020	Yes	294	PD	27	M	+	370	819	118,191	1	0	1	48,73	Yes	No	No	No	No	No	No	Yes	Yes	2	3
U021	Yes	214	PD	35	M	+	306	4,283	55,805	1	0	0	20,251	Yes	Yes	No	No	Yes	Yes	No	Yes	Yes	5	35
U023	No	445	CR	33	F	+	191	1,487	338,285	1	1	1	3,438	Yes	No	No	No	No	No	No	No	Yes	1	2
U030	No	453	CR	40	M	+	70	261	100,184	1	1	0	1,124	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	207
U032	No	433	PR	23	F	+	274	1,614	587,149	0	0	0	5,246	No	Yes	Yes	No	No	Yes	No	No	No	3	50
U034	No	432	PR	47	F	+	237	1,246	130,375	1	0	0	47,657	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	7	54
U039	Yes	162	PD	27	M	+	182	1,661	68,357	1	1	1	15,341	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	151
U048	No	441	PR	23	M	+	115	1,427	194,676	1	1	0	3,364	Yes	No	No	No	No	No	No	No	Yes	1	81
U060	Yes	145	PR	26	M	+	134	970	319,573	1	1	0	15,781	Yes	No	No	No	No	No	Yes	No	Yes	2	46
U062	No	441	PR	31	M	+	436	982	1,058,360	1	0	1	540,178	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	285
U066	No	792	PR	40	M	+	36	1,087	138,338	1	1	1	5,208	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	6	148
U094	No	426	PR	30	F	+	400	741	413,511	1	0	1	4,770	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	67
U099	No	746	NA	42	M	+	215	1,153	313,323	0	0	1	0	No	No	No	No	No	No	No	No	Yes	1	26
U101	No	476	PR	34	M	+	331	340	208,808	1	0	1	1,448	No	Yes	No	Yes	No	No	No	Yes	Yes	4	25
U106	No	630	SD	37	M	+	1	298	95,361	1	1	1	6,143	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	116
U108	No	593	PR	34	M	+	1,437	1,042	377,043	1	0	0	7,312	Yes	No	No	No	No	Yes	Yes	Yes	Yes	4	116
U146	No	479	SD	30	M	+	302	1,142	920,350	1	0	1	8,439	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	6	73
U156	Yes	320	PD	30	M	+	109	740	103,371	1	1	1	45,117	Yes	No	No	No	No	No	Yes	Yes	Yes	3	77
U191	No	333	CR	29	M	+	16	485	1,289,560	0	1	1	25,694	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	6	98
U210	No	540	CR	37	M	-	NA	NA	NA	1	NA	1	536	Yes	No	No	No	No	Yes	No	Yes	Yes	3	41
U211	No	790	PR	25	M	-	NA	NA	NA	1	NA	1	10,540	Yes	No	No	No	No	No	No	Yes	Yes	2	47
U215	No	650	PR	32	F	+	4	69	150,734	1	1	1	ND	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8	89
U216	No	436	PR	78	M	-	1,066	1,045	NA	1	0	1	8,642	Yes	No	No	No	No	Yes	No	Yes	Yes	3	82
U217	No	579	PR	26	F	+	NA	NA	NA	1	NA	0	1,648	Yes	Yes	Yes	No	No	No	No	Yes	Yes	4	143
U218	No	518	PR	32	M	+	612	1,578	155,437	1	0	1	10,86	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	7	260
U219	Yes	20	PD	45	F	-	479	585	NA	1	0	1	21,511	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	6	542

Key:
[a] Number of days from baseline to last known vital status
[b] PR = Partial Response; CR = Complete Response; SD = Stable Disease; PD = Progressive Disease; NA= Not Applicable
[c] T(1) = tumors not limited to skin, with extensive oral, gastrointestinal & visceral KS; I(1) = CD4+ T-cell count <150/uL, NA if HIV(-) ; S1 = systemic illness (fever, night sweats, ≥10% weight loss, diarrhea for >2 weeks) [145]
[d] ND = Not Determined

Table 13. Individual KS tumor-level characteristics and morphotypes

Tumor ID	Visit no.	Visit day	Biopsy Location	Lesion Type	Lesion appearance	Size 1st visit (cm)	Size 2nd visit (cm)	Treatment Response
U003-B	1	0	Left leg	Nodular	Raised	0.60	0.60	Baseline
U003-C	1	0	Left leg	Macular	Raised	1.20	1.00	Baseline
U003-E	5	90	Left leg	Macular	Raised	1.00	0.60	Regressing
U003-F	5	90	Left leg	Nodular	Raised	0.50	0.40	Regressing
U003-G	5	90	Left foot	Fungating	Raised	5.50	3.60	Regressing
U004-C	1	0	Left leg	Macular	Flat	3.20	2.10	Baseline
U004-D	1	0	Right leg	Nodular	Raised	0.50	0.50	Baseline
U004-F	5	87	Right leg	Macular	Flat	0.30	0.40	Regressing
U007-B	1	0	Right leg	Macular	Raised	1.80	0.70	Baseline
U008-B	1	0	Left leg	Nodular	Raised	2.00	1.00	Baseline
U008-C	1	0	Right leg	Macular	Flat	5.20	3.00	Baseline
U008-D	1	0	Left leg	Macular	Raised	7.50	4.00	Baseline
U008-E	5	84	Right leg	Macular	Raised	3.00	2.00	Regressing
U008-F	5	84	Left leg	Macular	Raised	5.90	3.30	Regressing
U008-G	5	84	Left leg	Macular	Raised	4.20	2.00	Regressing
U008-H	8	154	Right leg	Macular	Flat	1.00	0.80	Unable to assess
U008-I	8	154	Right leg	Macular	Raised	1.20	1.00	New/Progressing
U020-B	1	0	Left upper extremity	Nodular	Raised	1.90	0.90	Baseline
U020-C	1	0	Left foot	Fungating	Raised	21.00	13.00	Baseline
U020-E	9	282	Right upper extremity	Macular	Raised	3.00	2.00	New/Progressing
U020-F	9	282	Left leg	Macular	Raised	0.60	0.60	New/Progressing
U021-B	1	0	Chest/ Abdomen	Macular	Flat	0.70	0.40	Baseline
U021-C	1	0	Chest/ Abdomen	Macular	Flat	0.80	0.40	Baseline
U021-D	1	0	Right leg	Macular	Raised	0.50	0.50	Baseline
U021-E	5	94	Right upper extremity	Macular	Raised	1.20	1.10	Regressing
U021-H	8	175	Right upper extremity	Macular	Flat	0.60	0.40	Stable
U021-I	8	175	Right upper extremity	Macular	Flat	0.80	0.50	Regressing
U030-C	1	0	Right upper extremity	Nodular	Raised	1.00	0.70	Baseline
U032-B	1	0	Chest/ Abdomen	Macular	Flat	1.00	0.30	Baseline
U034-B	1	0	Right upper extremity	Macular	Flat	2.00	1.50	Baseline
U034-C	1	0	Left upper extremity	Macular	Flat	1.20	0.50	Baseline
U039-B	1	0	Right leg	Macular	Flat	1.10	0.70	Baseline
U048-B	1	0	Right leg	Macular	Flat	0.70	0.50	Baseline
U048-C	1	0	Right leg	Nodular	Raised	1.00	0.70	Baseline
U048-D	1	0	Right leg	Nodular	Raised	0.80	0.60	Baseline
U048-E	5	91	Right leg	Nodular	Raised	0.70	0.70	Regressing
U060-C	1	0	Right leg	Macular	Raised	0.50	0.60	Baseline

Tumor ID	Visit no.	Visit day	Biopsy Location	Lesion Type	Lesion appearance	Size 1st visit (cm)	Size 2nd visit (cm)	Treatment Response
U060-D	1	0	Right leg	Macular	Raised	0.50	0.50	Baseline
U062-B	1	0	Right upper extremity	Macular	Raised	0.60	0.70	Baseline
U062-C	1	0	Right leg	Macular	Raised	0.70	0.60	Baseline
U066-C	1	0	Right leg	Macular	Raised	1.10	0.70	Baseline
U094-B	1	0	Right upper extremity	Macular	Raised	1.20	0.60	Baseline
U099-D	1	0	Right foot (09)	Nodular	Raised	1.00	0.80	Baseline
U101-D	5	86	Left leg	Nodular	Raised	1.00	1.00	Regressing
U106-B	1	0	Back (06)	Macular	Raised	4.20	2.00	Baseline
U108-B	1	0	Right upper extremity	Nodular	Raised	1.00	0.80	Baseline
U108-H	8	280	Right leg	Macular	Raised	0.50	0.50	Regressing
U108-I	8	280	Right leg	Macular	Raised	0.60	0.50	Regressing
U146-C	1	0	Left leg	Macular	Flat	1.80	1.60	Baseline
U156-B	1	0	Left leg	Macular	Flat	0.60	0.50	Baseline
U156-C	1	0	Left leg	Macular	Flat	0.50	0.50	Baseline
U156-D	1	0	Right leg	Macular	Raised	1.20	0.80	Baseline
U156-E	5	84	Right leg	Nodular	Raised	0.30	0.30	Regressing
U156-G	8	167	Left leg	Macular	Flat	0.80	0.50	Stable
U156-H	8	167	Left leg	Macular	Flat	0.60	0.50	Stable
U191-B	1	0	Right upper extremity	Macular	Raised	1.60	1.10	Baseline
U191-C	1	0	Right upper extremity	Macular	Raised	1.30	1.30	Baseline
U191-D	1	0	Right upper extremity	Macular	Raised	1.00	0.70	Baseline
U191-E	5	87	Left upper extremity	Macular	Flat	2.20	1.70	Regressing
U191-F	5	87	Right leg	Macular	Flat	1.80	1.50	Regressing
U210-B	1	0	Right leg	Nodular	Raised	0.90	0.80	Baseline
U211-D	1	0	Left leg	Macular	Raised	0.50	1.00	Baseline
U215-D	1	0	Left leg	Macular	Raised	2.00	2.8	Stable
U216-D	1	0	Left leg	Macular	Raised	0.80	0.60	Baseline
U217-D	1	0	Right upper extremity	Macular	Raised	1.60	1.20	Baseline
U218-D	1	0	Left upper extremity	Macular	Raised	1.80	1.30	Baseline
U219-D	1	0	Left upper extremity	Macular	Raised	1.00	1.00	Baseline

Table 14. Odds ratios of mutations to clinical or tumor traits

	K5-K6			K8.1			miR-K10			K4.2 Length			vIRF2 174bp		
Personal/Clinical Traits	**OR**[a]	**(95%CI)**	**P**	**OR**[a]	**(95%CI)**	**P**	**OR**[a]	**(95%CI)**	**P**	**OR**[a]	**(95%CI)**	**P**	**OR**[a]	**(95%CI)**	**P**
Age, per 5 years	0.90	(0.69, 1.17)	0.434	0.96	(0.59, 1.57)	0.886	0.89	(0.52, 1.51)	0.657	1.25	(0.80, 1.96)	0.325	1.03	(0.72, 1.48)	0.882
Female v. Male	0.50	(0.05, 5.02)	0.556	0	[d]		0	[d]		2.00	(0.17, 23.66)	0.582	**7.56**	**(1.02, 56.09)**	**0.048**
HIV Positive v. Negative	0.94	(0.09, 10.33)	0.958	0.21	(0.02, 1.96)	0.170	0.95	(0.07, 12.72)	0.972	0	[e]		inf.	[f]	
Plasma KSHV (log10 copies/ml), per unit [b]	0.71	(0.16, 3.10)	0.650	0.81	(0.22, 2.98)	0.756	0.71	(0.26, 1.98)	0.516	1.82	(0.33, 10.00)	0.493	1.30	(0.20, 8.44)	0.783
Patient no. of sites w/lesions, per site	**0.61**	**(0.42, 0.90)**	**0.011**	**0.57**	**(0.35, 0.93)**	**0.024**	0.73	(0.47, 1.13)	0.163	1.19	(0.80, 1.76)	0.381	1.14	(0.74, 1.76)	0.544
More than 4 lesion sites v. 4 or fewer [j]	**0.20**	**(0.06, 0.71)**	**0.013**	**0.05**	**(0.00, 0.44)**	**0.007**	0.31	(0.03, 3.76)	0.358	3.41	(0.43, 27.17)	0.247	2.25	(0.33, 15.19)	0.405
Any head, neck, or oral lesions	**0.13**	**(0.03, 0.55)**	**0.006**	0	[g]		0.45	(0.04, 4.59)	0.502	5.05	(0.76, 33.59)	0.094	1.01	(0.15, 6.70)	0.990
Any lesions other than extremities	**0.08**	**(0.03, 0.22)**	**<0.001**	0.41	(0.05, 3.38)	0.407	inf.	[h]		0.31	(0.02, 4.26)	0.382	2.40	(0.18, 31.44)	0.505
KS T-stage 1	2.39	(0.18, 31.11)	0.507	1.50	(0.11, 20.76)	0.762	inf.	[i]		1.07	(0.09, 12.85)	0.957	0.34	(0.03, 4.32)	0.406
KS S-stage 1	1.20	(0.26, 5.50)	0.814	2.81	(0.45, 17.73)	0.271	0.44	(0.04, 4.25)	0.476	1.70	(0.21, 13.97)	0.621	0.16	(0.02, 1.05)	0.056
KS I-stage 1 [c]	1.24	(0.32, 4.80)	0.753	1.50	(0.21, 10.93)	0.689	1.33	(0.12, 15.01)	0.816	0.68	(0.09, 5.46)	0.721	2.86	(0.40, 20.55)	0.295
CD4+ cell count, per 100 /ul [c]	1.02	(0.92, 1.14)	0.704	1.08	(0.76, 1.55)	0.664	0.81	(0.43, 1.50)	0.502	0.88	(0.65, 1.19)	0.417	0.62	(0.35, 1.09)	0.099
HIV viral load (log10 copies/ml), per unit [c]	0.63	(0.15, 2.64)	0.524	1.06	(0.13, 8.47)	0.959	0.43	(0.02, 11.01)	0.609	0.27	(0.03, 2.43)	0.241	0.44	(0.07, 2.84)	0.392
Tumor Traits															
Visit, Followup v. Baseline	0.83	(0.24, 2.85)	0.771	0.71	(0.19, 2.62)	0.606	**3.74**	**(1.27, 11.02)**	**0.017**	0.59	(0.22, 1.59)	0.294	**0.12**	**(0.02, 0.76)**	**0.024**
Lesion Type, Nodular/Fungating v. Macular	**20.24**	**(4.97, 82.48)**	**<0.001**	**6.38**	**(1.74, 23.41)**	**0.005**	1.64	(0.28, 9.43)	0.582	**0.28**	**(0.09, 0.87)**	**0.028**	0.95	(0.23, 3.96)	0.940
Lesion Site, Legs v. Other	2.57	(0.47, 14.18)	0.278	3.06	(0.55, 17.09)	0.202	1.56	(0.16, 15.16)	0.702	0.39	(0.07, 2.09)	0.271	1.12	(0.19, 6.83)	0.898
Large Lesion (>1 cm diameter v. less)	1.32	(0.33, 5.30)	0.697	0.5	(0.10, 2.54)	0.404	**0.20**	**(0.06, 0.68)**	**0.01**	2.12	(0.62, 7.26)	0.23	1.07	(0.25, 4.54)	0.93
Large Lesion (>1 cm2 PD v. less)	1.43	(0.35, 5.76)	0.618	0.46	(0.10, 2.20)	0.331	0.31	(0.09, 1.10)	0.07	1.42	(0.38, 5.27)	0.60	0.70	(0.18, 2.70)	0.60

[a] univariable analysis
[b] among those with measured plasma KSHV RNA
[c] HIV+ only
[d] No non-intact K8.1 or non-wt miR-K10 among women
[e] All HIV-negative had truncated K4.2
[f] No vIRF2 or IR1 perfect repeats in HIV negative (all vIRF2 observed in HIV+)
[g] No K8.1 broken from pts with head/oral lesions
[h] All non-wt miR-K10 from pts with lesions outside of extremities
[i] All non-wt miR-K10 was from T=1 pts
inf. = infinite
Significant values are bolded (p<0.5)

Table 15. KSHV genetic polymorphisms and survival rates, Cox regression

Marker	HR	95%CI	P
IR1 over-representation v. normal	1.44	(0.38, 5.47)	0.593
K8.1 not intact v. intact	1.54	(0.40, 6.03)	0.531
miR-K10 non-WT v. WT	**4.11**	**(0.98, 17.15)**	**0.053**
K4.2 Truncated v. Full-length	2.8	(0.31, 25.06)	0.357
vIRF2 repeat v. none	0.29	(0.03, 2.50)	0.26

Significant values are bolded (p<0.5)
HR = Hazard ratio

Table 16. Correlations between observed KSHV mutations

	K5-K6 Region							K8.1							miR-K10							K4.2						
	Normal		Overrep.					Intact		Not Intact					Wt		Non-wt					Full		Trunc.				
	N	%	N	%	OR	(95% CI)	P	N	%	N	%	OR	(95% CI)	P	N	%	N	%	OR	(95% CI)	P	N	%	N	%	OR	(95% CI)	P
K8.1																												
Intact	40	87	10	62.5	ref.			50	100																			
Not intact	6	13	6	37.5	4	(0.56, 28.49)	0.166			12	100																	
miR-K10																												
wt consensus	33	71.7	14	87.5	ref			40	80	7	58.3	ref.			47	100												
non-wt	13	28.3	2	12.5	0.36	(0.07, 1.83)	0.219	10	20	5	41.7	2.86	(0.37, 22.30)	0.317			15	100										
K4.2																												
Full length	12	23.5	4	25	ref.			7	14	7	58.3	**ref.**			8	17	6	40	ref.			16	100					
Truncated	39	76.5	12	75	0.92	(0.32, 2.67)	0.883	43	86	5	41.7	**0.12**	**(0.02, 0.57)**	**0.008**	39	83	9	60	0.31	(0.03, 3.24)	0.326			51	100			
K11.2																												
No 174 bp dup.	41	80.4	13	81.3	ref.			39	78	10	83.3	ref.			36	76.6	13	86.7	ref.			11	69	43	84.3	ref.		
174 bp dup.	10	19.6	3	18.8	0.95	(0.23, 3.88)	0.939	11	22	2	16.7	0.71	(0.13, 3.98)	0.696	11	23.4	2	13.3	0.5	(0.04, 6.08)	0.589	5	31	8	15.7	0.41	(0.05, 3.14)	0.39

OR =Odds Ratio
Overrep. = Copy number or read coverage overrepresentation of K5-K6 region >1.5X over rest of genome (**Table 7**)
Wt = Wild-type (database consensus sequence)
Trunc. = All K4.2 truncation polymorphisms
Ref = Reference (=1), base ratio
Significant values are bolded (p<0.5)

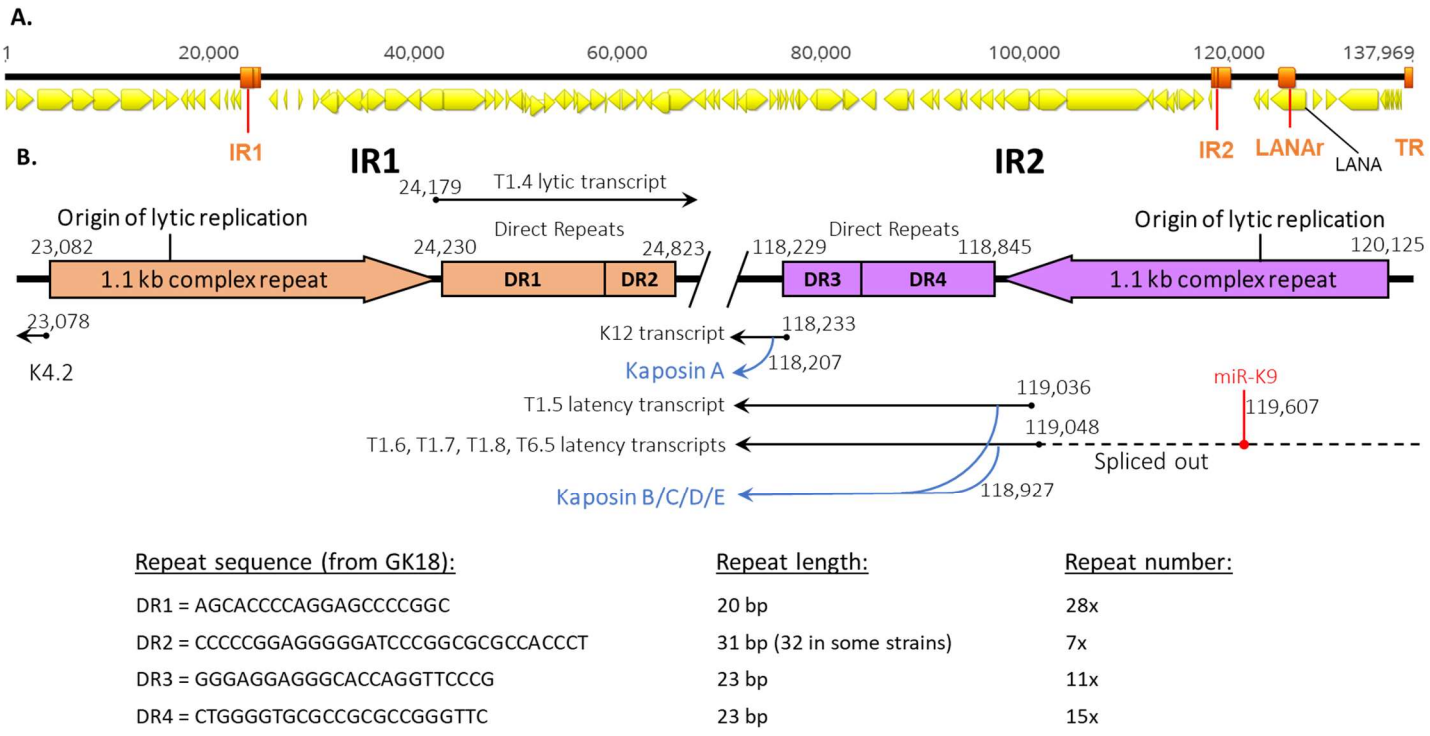
CHAPTER 5: Variation Within Major Internal Repeats of Kaposi Sarcoma Herpes Virus In Vivo

Introduction

The ~165-kb KSHV genome has 3 major repetitive sequence regions (**Fig 22A**): the two origins of lytic replication (Ori-Lyt) found in IR1 and IR2 [125,190], and the latency associated nuclear antigen (LANA) gene central repeat domain (LANAr). IR1 and IR2 are inverted homologs of each other, both having a complex 1.1 kb sequence that has 86.5% sequence identity and two 0.3 – 1.4 kb GC-rich direct repeat families – DR1 and DR2 in IR1, and DR3 and DR4 in IR2 (**Fig. 22B**). The four repeat families have different tandem repeat unit (TRU) sequences of length 20 bp in DR1, 23 bp in DR3 and DR 4, and 31 or 32 bp in DR2, with repeat numbers varying across KSHV strains [116,118]. The direct repeats are necessary for KSHV lytic replication, since progressive deletion of repeats eliminates KSHV Ori-Lyt mediated transactivation [125,190,191].

A non-coding RNA expressed from IR1 referred to as T1.4, comprised largely of DR1 and DR2 repeat families, is necessary for KSHV Ori-Lyt-dependent replication [155,191] and for interactions of KSHV viral co-transactivator bZIP with Ori-Lyt [192]. Transcripts through DR4 and DR3 in IR2 from the opposite strand are found in latently-infected cells and in KS tumors [117,119]. Isoforms of Kaposin proteins are translated from cryptic CUG start sites immediately upstream of DR4 sequences in some KSHV strains [117]. (**Fig 22B**).

Repeat sequence heterogeneity has been shown to be consequential in other DNA viruses. EBV IR1 was observed to have minor sequence variants and imperfect repeats in tumor-derived and lab strains more frequently than in saliva-derived and non-tumor derived lab strains [122]. Repairing defects in IR1 within the EBV lab strain B-95 resulted in higher production of large EBNA-LP isoforms and higher transformation efficiency [122,123]. GC-rich repeats in EBV Ori-Lyt's adopt triple helical DNA structures that when interrupted by point mutations abrogate Ori-Lyt-dependent replication [124]. The JC polyomavirus Ori-Lyt pentanucleotide repeat AGGGA, elements of which are present in KSHV GC-rich tandem repeats [125], regulates DNA replication and transcription, depending on host cell type [126,127]. In BK and SV40 polyomaviruses, mutations that delete or disrupt Ori-Lyt repeats result in viruses unable to fully replicate but are highly transforming [128,129].



Repeat sequence (from GK18):	Repeat length:	Repeat number:
DR1 = AGCACCCCAGGAGCCCCGGC	20 bp	28x
DR2 = CCCCCGGAGGGGGATCCCGGCGCGCCACCCT	31 bp (32 in some strains)	7x
DR3 = GGGAGGAGGGCACCAGGTTCCCG	23 bp	11x
DR4 = CTGGGGTGCGCCGCGCCGGGTTC	23 bp	15x

Figure 22. Diagram of KSHV internal repeats IR1 and IR2.

(A.) *Linear schematic of KSHV genome. Yellow arrows represent open reading frames, as annotated in KSHV reference strain GK18. Orange bars represent the major repeat regions IR1, IR2, LANAr and the TRs.* ***(B.)*** *Diagram of IR1 and IR2 regions of the KSHV genome. IR1 and IR2 are inverted, related copies of each other that serve as origins of lytic replication. Each is composed of a 1.1 kb complex sequence and 2 families of GC-rich direct tandem repeats. Shown in black arrows are transcripts through the repeats, including a T1.4 lytic transcript from IR1 (above cartoon) and latency transcripts from IR2 (below cartoon). Blue arrows show the Kaposin proteins translated from the latency transcripts. T1.6, T1.7, T1.8, T6.5 latency transcripts are transcribed from promoters further upstream, and the dashed line indicates the portion spliced out to generate KSHV microRNAs. Only miR-K9, in red, is included in the region depicted. Coordinate positions are for KSHV reference genome GK18. Below are the sequences of each family of direct tandem repeats and their counts in GK18.*

Despite being sites of rapid evolution [115], repeat regions are rarely sequenced in KSHV genomes because their long, GC-rich repeats are insurmountable by short-read sequencing. Short reads can also map erroneously to homologous sequences between IR1 and IR2. To overcome these constraints and accurately assess the intra-host diversity of the major internal repeats in KSHV, Pacific Biosciences' single-molecule real-time long read sequencing (SMRT) platform was used along with unique molecular identifier (UMI) tags, which were applied before PCR amplification (SMRT-UMI). Full-length sequences of IR1, IR2 and LANAr from 1-2 tumors and 0-1 oral swab were obtained from 16 Ugandans with KS. This study revealed higher intra-host diversity in internal repeats compared to the rest of the KSHV genome, frequent mismatches in IR1 repeat sequences, and a frequent loss of the coding potential of full-length Kaposin products due to polymorphisms in the translation initiation in IR2.

Results

Intra-host diversity of KSHV major internal repeat regions

UMI-consensus reads of the repeat regions IR1, IR2 and/or LANAr were obtained from tumors from 16 individuals – 1 tumor each from 8 individuals, and 2 tumors each from 8 others (**Table 17**). One, two or all KSHV internal repeats from matching oral swabs were successfully sequenced by SMRT-UMI from only 6 individuals, yielding in total 4 tumor-oral swab pairs of IR1, 4 pairs of IR2, and 4 pairs of LANAr. Altogether, repeat sequences from at least 2 samples were obtained from 11 individuals. A median of 21.5 UMI-consensus reads (UMIs), range 1 – 171, were obtained from tumors, while a median of 2.5 UMIs, range 1 – 13, were obtained from oral swabs (**Table 17**). The median read depth per DNA template was 29.17 CCS, range 1 - 4,843 (**Table 18**). By repeat region, the median UMIs obtained was 21.5 for IR1, 26.5 for IR2, and 38.5 for LANAr.

KSHV internal repeats were largely conserved within individuals. Intra-host variation, when present, derived from minor single nucleotide variants (SNV), homopolymer G runs and tandem repeat unit (TRU) counts. Minor SNVs were found in an average of 0.23% of base positions of the KSHV internal repeats (**Table 17**), and nearly all were present in only one UMI-consensus read. Homopolymer G runs were conserved within hosts when the total length was below 8 nucleotides and unstable when longer (**Table 17**). TRU counts were the largest contributor of diversity. Indels at tandem repeat regions were always by multiples of the TRU length. When TRU

Table 17. Sample counts, UMI counts and intra-host diversity of KSHV repeats in tumors and oral swabs of 16 individuals

		no. of samples			no. of UMI consensus						no. of columns with minor SNV		longest homopoly-G	Tandem Repeat Unit (TRU)			
PTID*	Region	tumor	oral	total	tumor	oral	total	avg UMIs per sample	consensus length (bp)	% pairwise identity **	no.	% of repeat length		DR1 consensus TRU count	consensus UMI over total UMI	DR2 consensus TRU count	consensus UMI over total UMI
U003	IR1	1	0	1	4	0	4	4	2,615	99.7%	1	0.04%	8	58	3/4	6	1
U004	IR1	2	0	2	58	0	58	29	2,354	99.7%	11	0.47%	8	48	48/58	4	1
U007	IR1	1	1	2	8	2	10	5	2,736	97.2%	4	0.15%	10 to 11	62	8/10	7	1
U008	IR1	2	1	3	13	2	15	5	2,364	100.0%	2	0.08%	10	47	1	5	1
U020	IR1	2	1	3	19	1	20	7	2,113	98.7%	3	0.14%	15 to 20	34	14/20	6	17/20
U021	IR1	2	0	2	119	0	119	60	1,876	98.9%	10	0.53%	11 to 14	20	68/119	6	1
U030	IR1	1	0	1	21	0	21	21	2,043	100.0%	4	0.20%	8	21	21/21	9	1
U032	IR1	1	0	1	3	0	3	3	2,279	100.0%	1	0.04%	8	45	1	4	1
U034	IR1	2	0	2	16	0	16	8	2,184	99.9%	0	0.00%	11 to 13	38	15/16	5	1
U048	IR1	2	0	2	90	0	90	45	1,897	99.9%	4	0.21%	16 to 19	17	1	9	1
U156	IR1	2	0	2	110	0	110	55	2,215	97.7%	11	0.50%	8	38	95/110	6	109/110
U191	IR1	1	0	1	15	0	15	15	2,383	99.8%	6	0.25%	22	46	1	7	1
U210	IR1	1	0	1	31	0	31	31	2,521	91.1%	3	0.12%	9	39	20/32	16	26/32
U215	IR1	1	0	1	48	0	48	48	1,898	99.7%	6	0.32%	11 to 13	25	1	5	44/47
U216	IR1	1	0	1	127	0	127	127	1,754	99.9%	4	0.23%	9 to 11	15	126/127	7	126/127
U217	IR1	1	1	2	20	2	22	11	1,603	90.7%	2	0.12%	9	11	12/23	5	19/23
	MEDIAN						21.50		2,200	98.3%							
U003	IR2	1	1	2	1	1	2	1	2,372	100.0%	0	0.00%	11	19	1	10	1
U004	IR2	2	1	3	36	3	39	13	2,455	98.4%	7	0.29%	10 to 11	21	37/40	13	30/40
U007	IR2	1	1	2	18	3	21	11	2,092	100.0%	1	0.05%	7	15	1	6	1
U008	IR2	1	1	2	2	3	5	3	2,324	100.0%	0	0.00%	8	18	1	8	1
U020	IR2	1	0	1	3	0	3	3	2,415	99.5%	1	0.04%	7	28	2/3	11	1
U021	IR2	2	0	2	39	0	39	20	2,227	96.7%	5	0.22%	7	23	26/39	8	26/39
U030	IR2	1	0	1	75	0	75	75	1,946	100.0%	10	0.51%	8	17	1	3	1
U032	IR2	1	0	1	4	0	4	4	2,100	100.0%	1	0.05%	9	18	1	8	1
U034	IR2	1	0	1	7	0	7	7	2,252	100.0%	0	0.00%	9	23	1	9	1
U048	IR2	2	0	2	139	0	139	70	2,224	99.9%	10	0.45%	9	17	132/139	13	1
U156	IR2	2	0	2	34	0	34	17	2,385	99.8%	3	0.13%	10 to 12	33	33/34	5	1
U191	IR2	1	0	1	6	0	6	6	2,206	99.7%	1	0.05%	7	19	5/6	12	1

		no. of samples			no. of UMI consensus						no. of columns with minor SNV		longest homopoly-G	Tandem Repeat Unit (TRU)			
PTID*	Region	tumor	oral	total	tumor	oral	total	avg UMIs per sample	consensus length (bp)	% pairwise identity **	no.	% of repeat length		DR1 consensus TRU count	consensus UMI over total UMI	DR2 consensus TRU count	consensus UMI over total UMI
U210	IR2	1	0	1	61	0	60	60	1,849	99.9%	1	0.05%	11 to 13	3	1	12	59/61
U215	IR2	1	0	1	19	0	19	19	1,896	99.3%	2	0.11%	7	8	1	9	1
U216	IR2	1	0	1	84	0	84	84	2,252	99.7%	11	0.49%	7	22	72/84	10	1
U217	IR2	1	0	1	32	0	32	32	2,084	99.9%	1	0.05%	7	13	31/32	8	1
	MEDIAN						26.50		2,226	99.6%							
U003	LANA	1	0	1	1	0	1	1	1,693	-	-	-	-				
U004	LANA	2	1	3	70	6	76	25	1,492	99.9%	9	0.60%	N/A				
U007	LANA	1	1	2	19	12	31	16	1,489	100.0%	1	0.07%	N/A				
U008	LANA	2	1	3	33	13	46	15	1,183	99.8%	3	0.25%	N/A				
U020	LANA	2	1	3	15	1	16	5	1,741	100.0%	2	0.11%	N/A				
U021	LANA	2	0	2	113	0	113	57	1,522	100.0%	6	0.39%	N/A				
U030	LANA	1	0	1	93	0	93	93	1,582	99.9%	6	0.38%	N/A				
U032	LANA	1	0	1	8	0	8	8	1,807	99.9%	1	0.06%	N/A				
U034	LANA	2	0	2	17	0	17	9	1,456	100.0%	1	0.07%	N/A				
U048	LANA	2	0	2	171	0	171	86	1,909	99.8%	24	1.26%	N/A				
U156	LANA	2	0	2	122	0	122	61	1,624	100.0%	9	0.55%	N/A				
U191	LANA	2	0	2	6	0	6	3	1,735	100.0%	1	0.06%	N/A				
U210	LANA	1	0	1	59	0	59	59	1,636	100.0%	4	0.24%	N/A				
U215	LANA	1	0	1	22	0	22	22	1,741	94.1%	2	0.11%	N/A				
U216	LANA	1	0	1	96	0	96	96	1,909	94.3%	6	0.31%	N/A				
U217	LANA	0	0	0	0	0	13	-	-	-	-	-	-				
	MEDIAN						38.50		1,636	99.1%							
TOTAL MEDIAN					**21.5**	**2.5**	**26.5**	**17**				**0.23%**					

* PTID - Participant ID

** % pairwise identity from all UMI-consensus sequences of given repeat in individual; including gap versus non-gap sites but excluding gap versus gap sites.

Table 18. UMI counts, circular consensus sequence (CCS) counts and minimum agreement

PTID	Region	no. of UMIc			average UMIc per sample	median CCS count of all UMIc	median minimum agreement*
		tumor	oral	total			
U003	IR1	4	0	4	4.0	325.0	0.90
U004	IR1	58	0	58	29.0	34.7	0.85
U007	IR1	8	2	10	5.0	192.8	0.76
U008	IR1	13	2	15	5.0	182.7	0.79
U020	IR1	19	1	20	6.7	41.1	0.72
U021	IR1	119	0	119	59.5	20.0	0.68
U030	IR1	21	0	21	21.0	1044.6	0.76
U032	IR1	3	0	3	3.0	45.0	0.91
U034	IR1	16	0	16	8.0	8.5	0.75
U048	IR1	90	0	90	45.0	49.1	0.69
U156	IR1	110	0	110	55.0	7	0.67
U191	IR1	15	0	15	15.0	16.0	0.67
U210	IR1	31	0	31	31.0	56.0	0.70
U215	IR1	48	0	48	48.0	9.0	0.73
U216	IR1	127	0	127	127.0	11.0	0.78
U217	IR1	20	2	22	11.0	11.9	0.61
MEDIAN:				21.5		37.9	
U003	IR2	1	1	2	1.0	598.0	0.76
U004	IR2	36	3	39	13.0	86.7	0.73
U007	IR2	18	3	21	10.5	48.1	0.88
U008	IR2	2	3	5	2.5	359.4	0.84
U020	IR2	3	0	3	3.0	53.0	0.85
U021	IR2	39	0	39	19.5	38.0	0.89
U030	IR2	75	0	75	75.0	24.9	0.82
U032	IR2	4	0	4	4.0	47.0	0.92
U034	IR2	7	0	7	7.0	11.0	0.82
U048	IR2	139	0	139	69.5	33.8	0.87
U156	IR2	34	0	34	17.0	6.0	0.67
U191	IR2	6	0	6	6.0	32.5	0.88
U210	IR2	61	0	60	60.0	37.0	0.71
U215	IR2	19	0	19	19.0	6.0	0.80
U216	IR2	84	0	84	84.0	12.0	0.78
U217	IR2	32	0	32	32.0	21.0	0.83
MEDIAN:				26.5		35.4	
U003	LANA	1	0	1	1.0	2537.0	0.94
U004	LANA	70	6	76	25.3	70.5	0.92
U007	LANA	19	12	31	15.5	69.5	0.87
U008	LANA	33	13	46	15.3	23.8	0.87
U020	LANA	15	1	16	5.3	21.6	0.75
U021	LANA	113	0	113	56.5	41.0	0.90
U030	LANA	93	0	93	93.0	13.6	0.80

PTID	Region	no. of UMIc			average UMIc per sample	median CCS count of all UMIc	median minimum agreement*
		tumor	oral	total			
U032	LANA	8	0	8	8.0	15.0	0.82
U034	LANA	17	0	17	8.5	6.3	0.77
U048	LANA	171	0	171	85.5	19.8	0.81
U156	LANA	122	0	122	61.0	9.0	0.82
U191	LANA	6	0	6	3.0	25.8	0.79
U210	LANA	59	0	59	59.0	20.0	0.78
U215	LANA	22	0	22	22.0	16.0	0.56
U216	LANA	96	0	96	96.0	15.0	0.53
U217	LANA	0	0	13	-	-	-
MEDIAN:				38.5		20.0	
Total Median:		**21.5**	**2.5**	**26.5**	**17**	**25.83**	**0.79**

Key:

CCS – Pacific Biosciences SMRT circular consensus sequences/reads

UMIc – UMI-consensus reads, the consensus of all CCS with the same UMI tag

CCS count – the number of CCS comprising a given UMIc

Minimum agreement – In an alignment of UMI reads comprising one UMI-consensus read, the proportion of the majority base in the alignment column that is least conserved among all alignment columns

Median minimum agreement – the median of minimum agreements from all UMI-consensus reads obtained from an individual

counts were variable within hosts, the UMI reads in the minority varied typically by 1 or 2 units from the intra-host consensus.

Distinct populations of repeat sequences, defined as shared non-consensus SNVs or TRU counts in at least 2 UMIs, were detected in 12 of 16 individuals (**Table 19A**). In 7 of 10 individuals with >5 UMIs from at least 2 samples, the sample-consensus in one or more of the 3 KSHV internal repeats were different (**Table 19B**). There was an SNV or a TRU count difference between 3 pairs of oral swab and tumor IR1 sequences (e.g., U020 in **Fig 23A**; U007 in **Fig 23B; Table 19B**), and an SNV and a TRU count difference between an oral swab-tumor pair of IR2 sequences (**Table 19B**). The TRU count differences between oral swab-tumor pairs of U007 and U217 were 7 and 19 units, respectively (U007 in **Fig 23B**), whereas all other intra-host TRU count differences were $\leq$2. LANAr had distinct intra-host populations of sequences as well (**Table 19A**), although none were between a tumor and an oral swab from an individual (**Table 19B**). Differences also existed between pairs of tumors from single individuals. For instance, 2 tumors

Table 19. Summary of distinct populations of repeat sequences detected.

A. Difference(s) between populations of sequences of repeat region within an individual

PTID	IR1 UMI counts		IR2 UMI counts		LANAr UMI counts	
U003	4	none	2	none	1	NA
U004	58	TRU	39	SNV, TRU	76	none
U007	10	SNV, TRU	21	SNV	31	none
U008	15	none	5	none	46	none
U020	20	none	3	none	16	SNV
U021	119	TRU	39	TRU	113	SNV
U030	21	none	75	SNV	93	SNV
U032	3	none	4	none	8	SNV
U034	16	none	7	none	17	none
U048	90	none	139	TRU	171	SNV
U156	110	none	34	none	122	none
U191	15	none	6	none	6	none
U210	31	TRU	61	none	59	none
U215	48	TRU	19	TRU	22	none
U216	127	none	84	TRU	96	TRU,SNV
U217	22	TRU	32	TRU	0	NA

*B. Difference(s) of repeat region between sample-consensus of given individual**

PTID	Total tumors	Total oral swabs	IR1	IR2	LANAr
U004	2	1	none	SNV	none
U007	1	1	SNV, TRU	none	none
U008	2	1	none	none	none
U020	2	1	none	NA	SNV
U021	2	0	TRU	TRU	none
U034	2	0	none	NA	none
U048	2	0	none	none	none
U156	2	0	none	none	none
U191	2	0	NA	NA	SNV
U217	1	1	TRU	NA	NA

Key:

* - Had at least 2 samples with KSHV repeat sequences obtained

PTID – Participant ID

TRU – repeat region has distinct populations of sequences intra-host due to tandem repeat unit counts differences; in LANAr they signify triplet nucleotide (codon) differences

SNV – repeat region has distinct populations of sequences intra-host due to single nucleotide variant(s)

NA – no UMI-consensus sequences available for comparison

none – repeat region has no intra-host difference detected

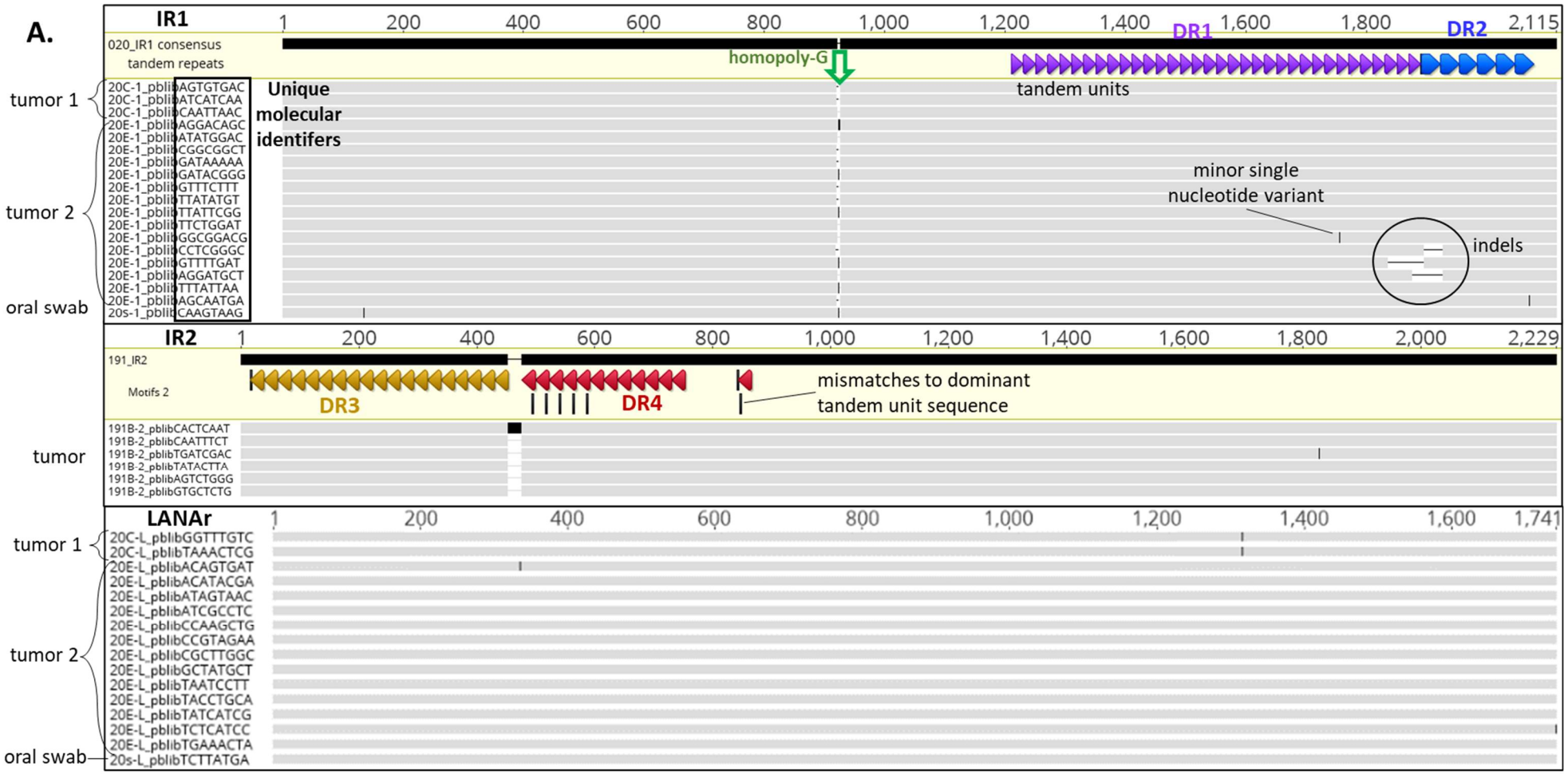
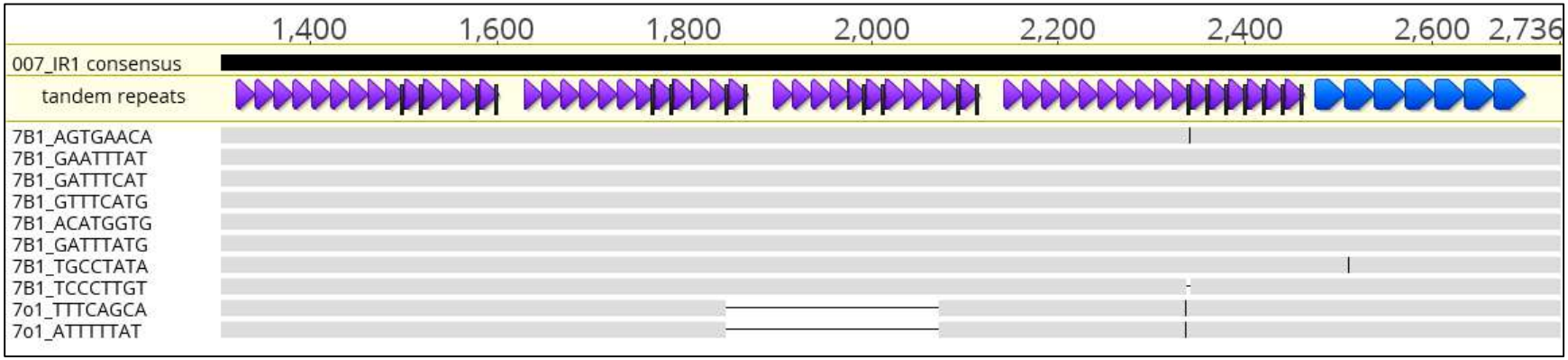
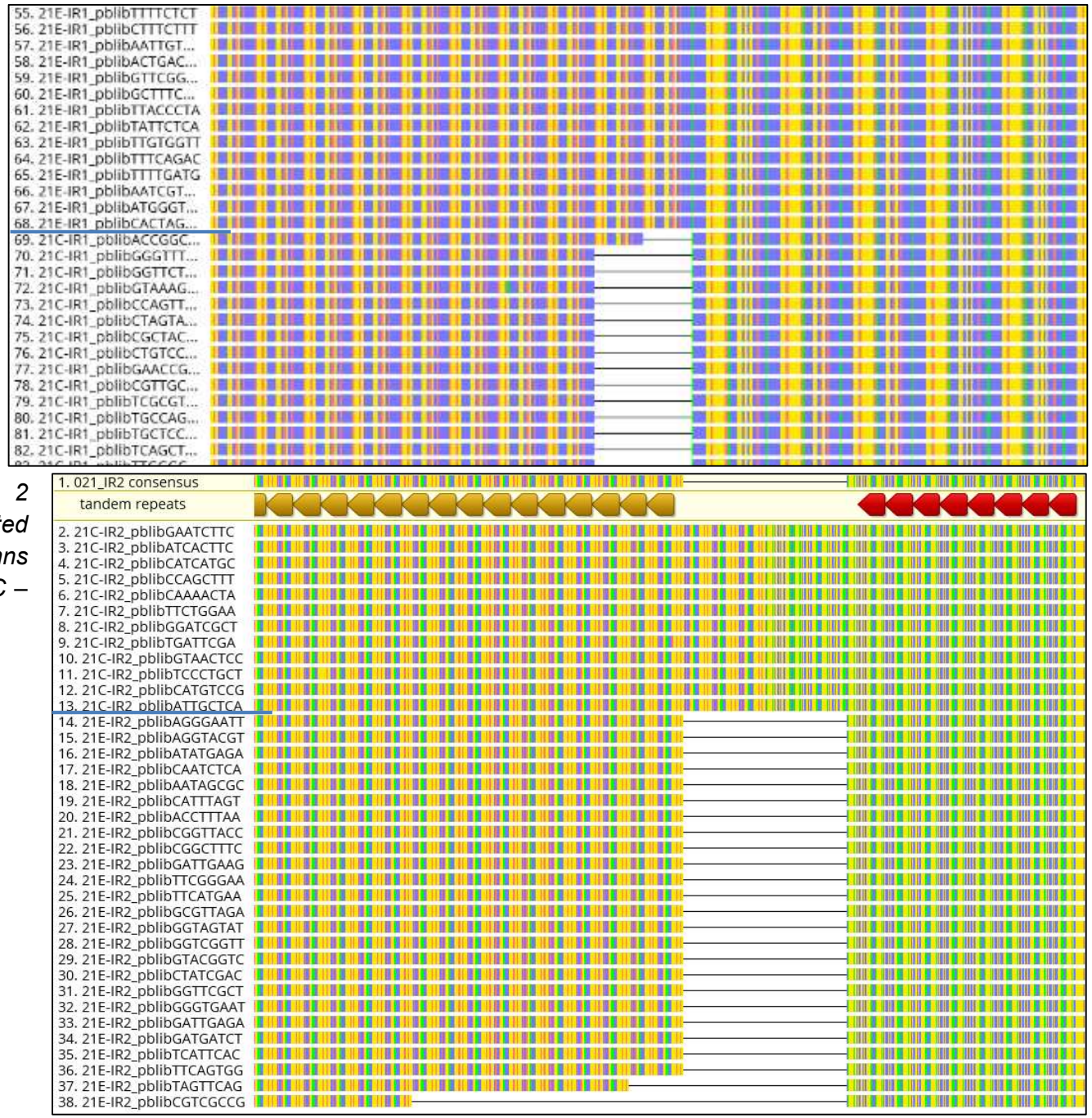


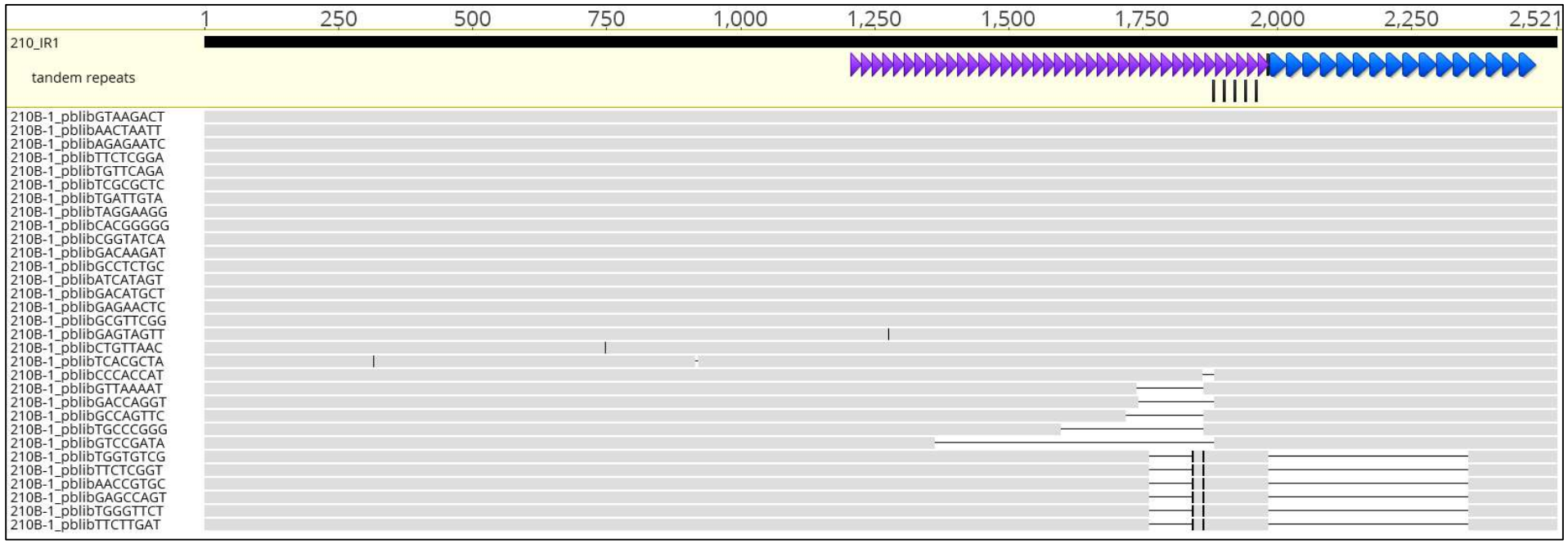
Figure 23. Representative sequence alignments.
Single molecule sequences of IR1, IR2 or LANAr derived from all tumor biopsies and oral swabs sequenced from an individual were aligned together. Above each alignment is the consensus sequence. For IR1 and IR2 sequences, locations of each tandem repeat unit were indicated by colored triangles. Black ticks below each represent mismatches to the dominant tandem unit sequence. Intra-host variation comprised of homo-poly G runs (hollow green arrow), minor single nucleotide variants (black ticks on alignment) and tandem unit count differences (alignment gaps under tandem units). Sequences were obtained from 1-2 tumors and sometimes an oral swab, as labeled on the left. Sequences are named by participant ID (number), tumor identifier (letter) and UMIs (unique molecular identifiers, after 'plib'). Alignment numbering are only of the displayed sequences. A. Representative alignments of highly conserved IR1, IR2 and LANAr sequences from participants U020 and U191.

B. Alignment of IR1 showing 2 distinct variants in tumor and oral swab in one individual





C. Alignments of IR1 and IR1 showing 2 distinct variants between 2 tumors (separated by a blue line) from one individual. Columns are colored by nucleotide base (A – red, C – blue, G – yellow, T – green).



D. Alignment of IR1 showing 2 distinct variants in 1 tumor

from U021 had different TRU counts in both IR1 and IR2 (**Fig 23C**). There was a distinguishing SNV in LANAr between 2 tumor pairs from participants U020 and U191 (**Table 19B**). Distinct sequence populations were also found in a single sample source; for example, a bimodal variation in IR1 TRU counts was found within one tumor, U210-B (**Fig 23D**). Only 3 individuals that had >5 UMIs from 2 - 3 of their samples had no differences detected IR1, IR2 or LANAr - U008, U034, U156 (**Table 19B**).

As a measure of total intra-host diversity within an individual, all UMI-consensus reads from all tumors and oral swabs within each person were aligned together for every repeat region. The percent pairwise identity was then calculated, including *gap versus non-gap* sites, but excluding *gap versus gap* sites. Hence, indels or TRU count differences are considered and detract from the total pairwise identity. IR1, IR2 and LANAr sequences had 100% intra-host pairwise identity in 3/16, 5/16 and 8/16 individuals, respectively (**Table 17**). The proportions increase to 6/16, 9/16 and 11/16 individuals respectively when counting ≥99.9% intra-host pairwise identity. The average % intra-host pairwise identity among individuals was 98.3% for IR1, 99.6% for IR2 and 98.9% for LANAr, and the lowest intra-host pairwise identity, i.e., most diverse, found in any one individual was 90.6% for IR1, 96.7% for IR2, 94.1% and LANAr. From these data, LANAr was homogeneous in the most number of individuals, while IR2 sequences consistently had the least amount of intra-host diversity (**Table 17**). The consistently lower intra-host % pairwise identity for IR1 reflects the frequent intra-host variation in TRU counts of IR1.

KSHV internal repeats had higher rates of intra-host variation than the rest of the genome

We hypothesized that KSHV internal repeats will be subjected to a higher rate of mutation due to high GC content and repetitive sequences. The 131 kb non-repeat regions of the KSHV genome was found in Chapter 3 to have on average 0.17% of base positions with a minor SNV detected [165]. Here, an average of 0.23% of base positions of the KSHV internal repeats (~6 kb altogether) had minor SNVs detected (**Table 17**).

In general, the depth of UMIs positively correlated with how much diversity can be detected (**Figs 6A**). This trend was observed here with SMRT-UMI on repeat regions (**Fig 25A**), although the median UMI-consensus read depth was lower in this study of KSHV internal repeats (21.5 for tumors and 2 for oral swabs) compared to unique KSHV genomic sequences in Chapter 3 (302

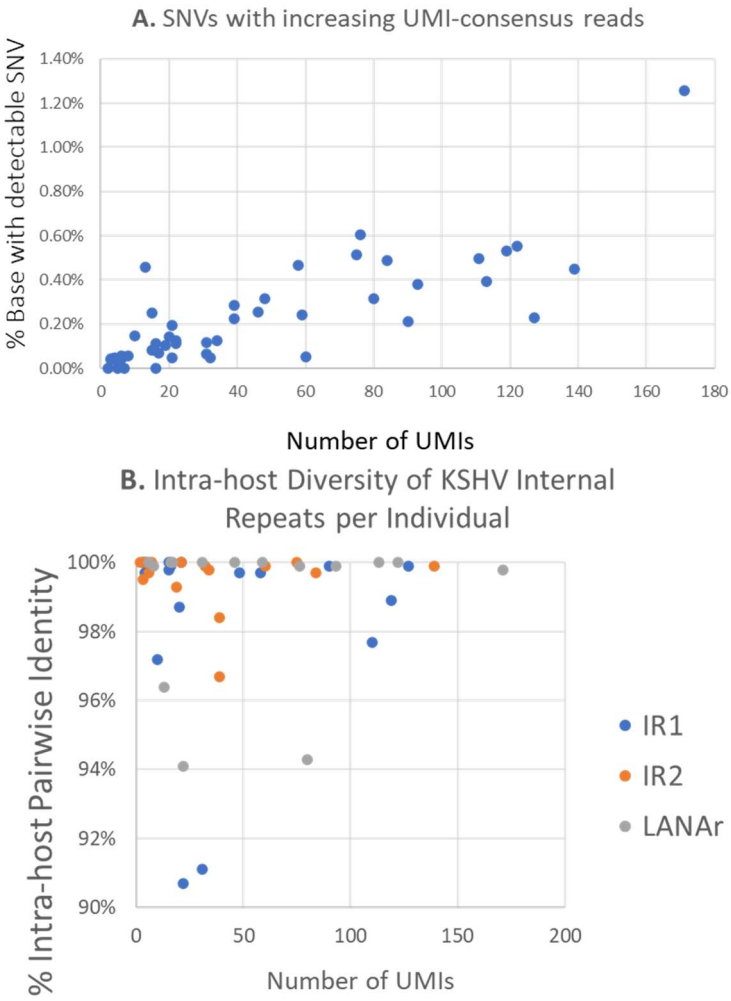


Figure 24. Detectable intra-host diversity as a function of UMIs recovered.

A. The percent of total bp positions with a detectable SNV in a sequenced repeat region was plotted against the number of all UMIs recovered from one individual for that repeat region. B. High sequence identity intra-host was maintained even with increasing number of UMIs recovered per individual. % intra-host pairwise identity is the percentage of pairwise identical sites in an alignment, including gap vs non-gap sites, but not gap vs gap sites. Number of UMI-consensus reads is equivalent to the number of unique DNA molecules sequenced per repeat region in all oral swab and tumor samples within an individual.

for tumors and 27 for oral swabs, **Table 4**). Nevertheless, there were samples with >100 UMI coverage that have 100% intra-host pairwise identity (**Fig 25B**). The discrepancy comes from *% bases with detectable SNV* considering the absolute counts of SNVs detected regardless of read

depth, while *% pairwise identity* takes into account the total number of sequences being evaluated. Hence % pairwise identity is relatively unaffected by very minor SNVs at high read depth, although it is greatly affected by large indels stemming from TRU count differences.

Lengths of internal repeats were highly variable between individuals. From the conserved primer binding sites I used (**Table 2**), the median length of IR1 in 16 Ugandans with KS was 2,200 bp, ranging from 1,603 to 2,736 bp (**Fig 26A, Table 17**), for IR2 it was 2,226 bp, ranging from 1,849 to 2,455 bp (**Fig 26B, Table 17**), and for LANAr it was 1,665 bp, ranging from 1,183 to 1,909 bp (**Fig 26C, Table 17**). TRU counts were the source of most of IR1 and IR2 *inter*-host length variation across individuals, even as the inverted complex repeats remained largely conserved. LANAr sequences had the most sequence variation across individuals, but the least length variation. In fact, there were 2 pairs of individuals (U020 and U215, U048 and U216) in which the consensus LANAr sequences were the same length, although the sequences of the same-length pairs were still different (**Fig 17C, Table 17**). LANAr can be divided into 3 subdomains of imperfectly repeating codons. While acidic residues were abundant, there were no repeating tandems or any discernible regularity in either the nucleotide or amino acid sequences **(Figs 26C** and **26D)**. All inter-host polymorphisms of consensus LANAr maintained the reading frame (**Fig 26C**).

Direct repeats in IR1 were imperfect more often than in IR2

We hypothesized that IR1 and IR2 sequences of viruses in oral swabs, having originated from host cells that completed reactivation, would have less minor SNVs and imperfect repeats compared to tumor viruses. However, imperfect tandem repeats, when present, were conserved throughout all an individual's tumor and oral swab viruses (**Fig 24A**). The sequence of the predominant TRU in a given repeat region, i.e., the 'master' TRU, was determined by finding a sequence that maximizes the TRU lengths and counts while minimizing imperfect TRU counts. A sequence is considered a related but imperfect copy of the dominant TRU only when it has mismatches less than 20% of a TRU length. Under these search criteria the dominant TRU sequence is not the same in all isolates. Mismatches to the dominant TRU sequence could be found propagated to more than 1 TRU and at irregular intervals (**Fig 26E**). The TRUs can also overlap or have a gap of a few bases between them that is present in a majority of the TRU family.

Considering TRU families with $\geq$2 imperfect TRU as degraded, IR1 DR1 was the most often degraded TRU family, observed in 12 of 16 individuals (**Fig 26A**). In 2 individuals, U003 and U007,

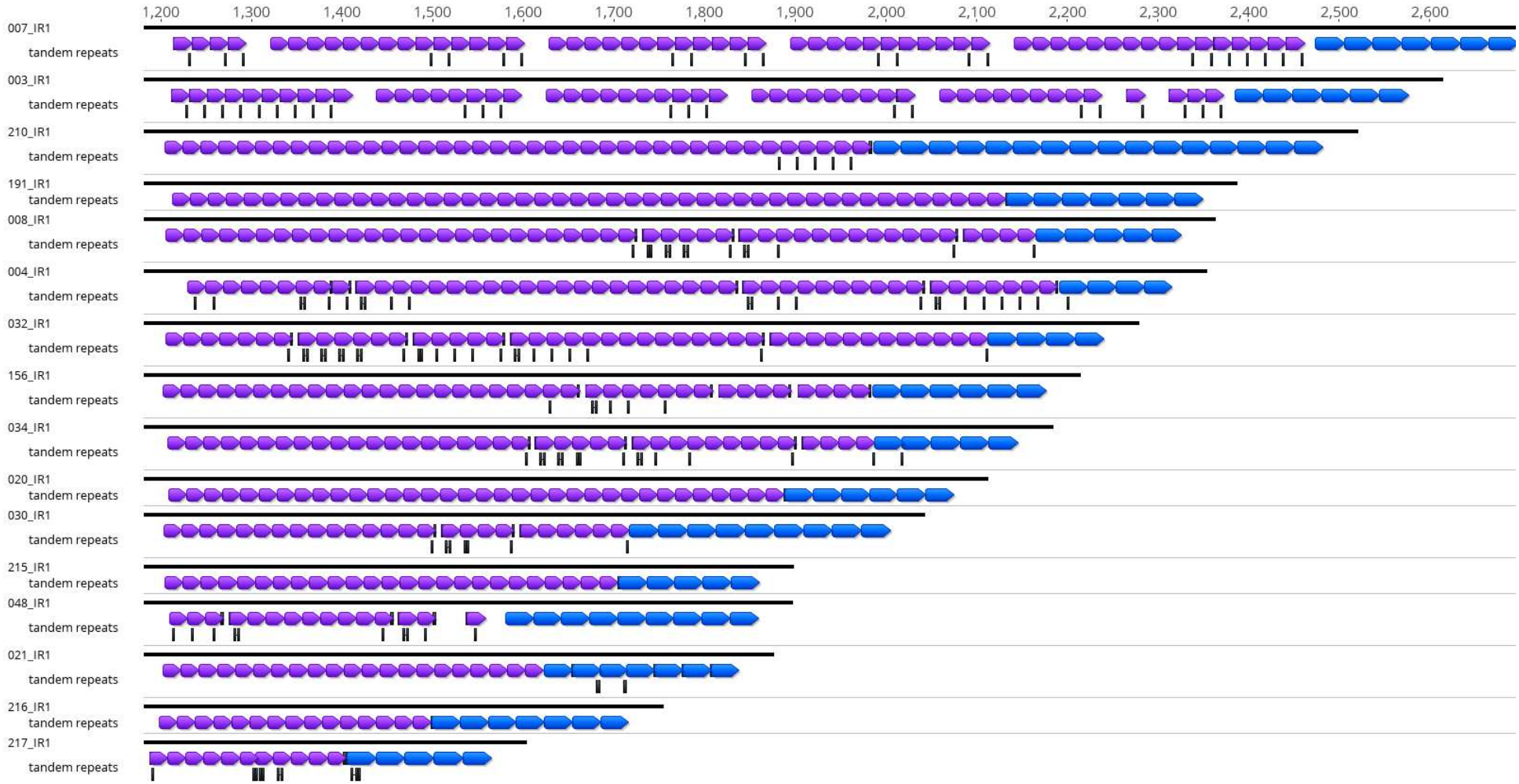
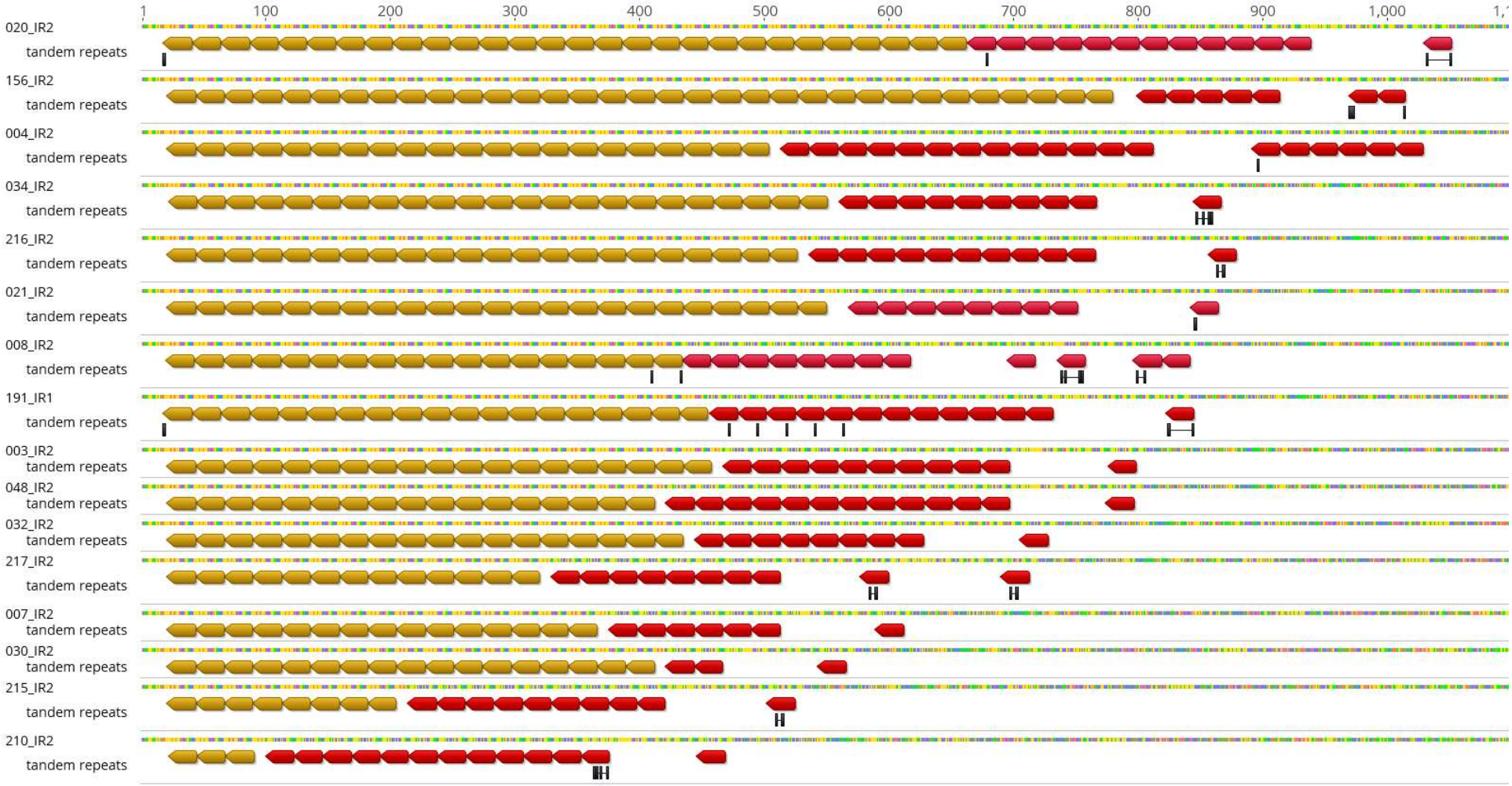
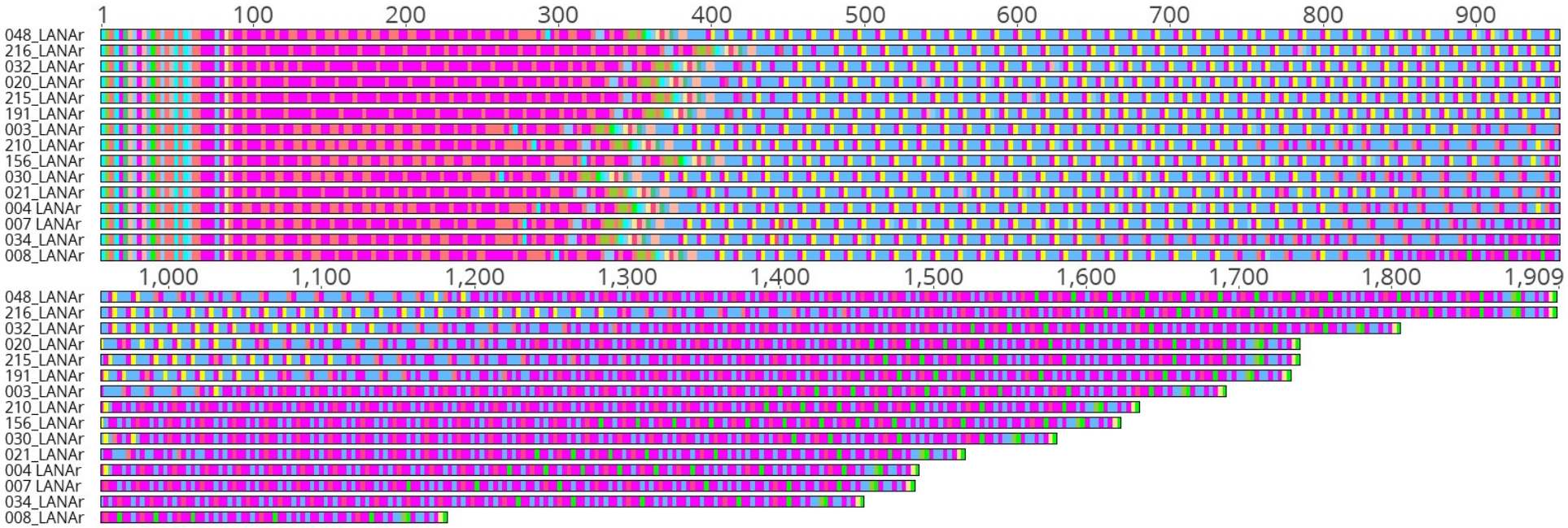


Figure 26. Diversity of KSHV Internal Repeats.

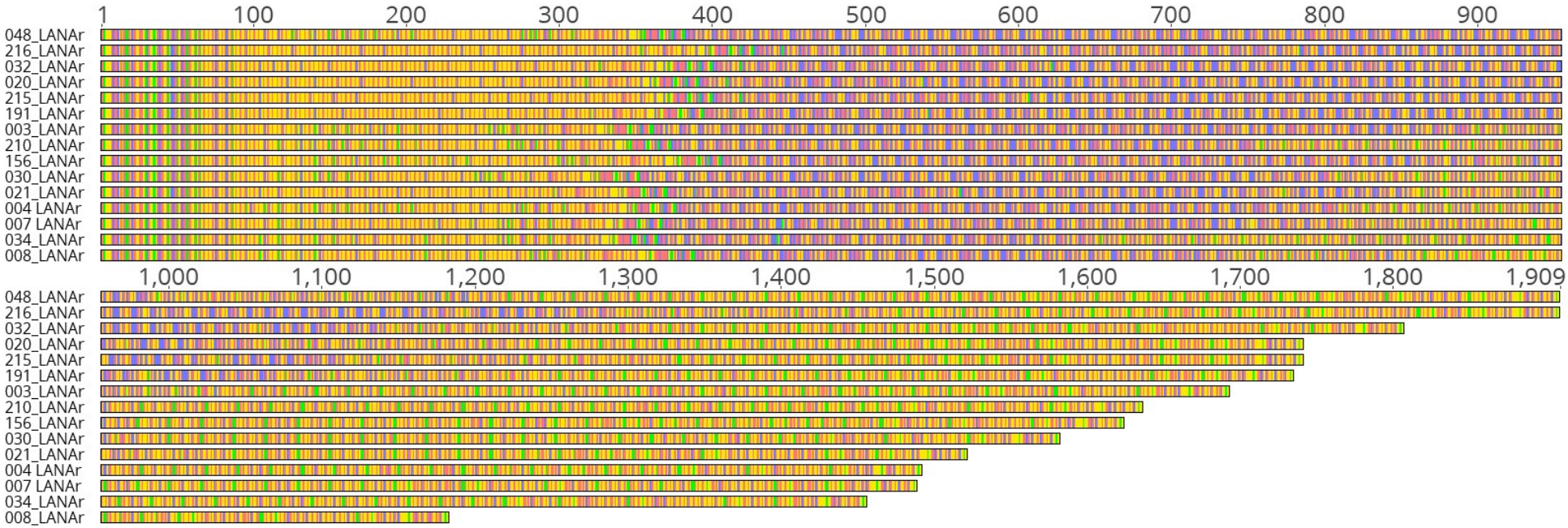
A. Individual-level consensus of IR1 sequences from participants in this study, sorted in descending length. DR1 tandem repeat units are in purple, DR2 tandem repeat units are in blue, and black ticks are mismatches to their respective dominant tandem unit sequence.



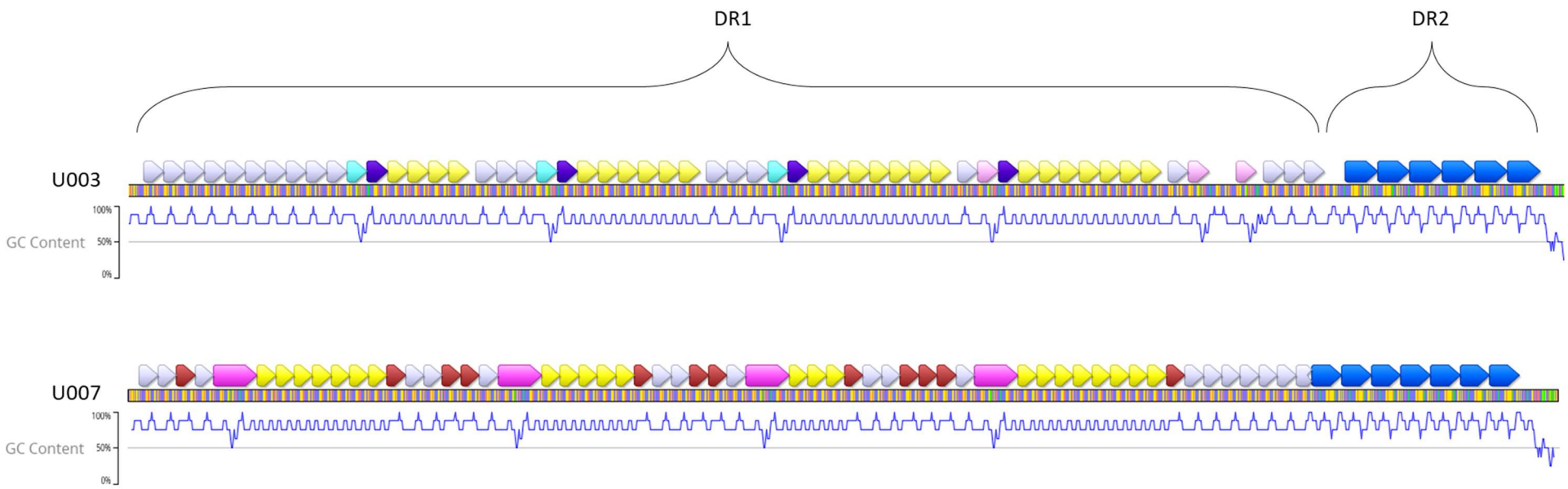
B. Individual-level consensus of IR2 sequences (reverse complemented) from participants in this study, sorted in descending length. DR1 tandem repeat units are in gold, DR2 tandem repeat units are in red, and black ticks are mismatches to their respective dominant tandem unit sequence.



C. Individual-level consensus of LANAr amino acid sequences (reverse complemented) from participants in this study, sorted in descending length. U048 and U216 LANAr have identical lengths, and U020 and U216 have identical lengths. Color-coded translations of the codons are shown (most common amino acids: Glu – magenta, Asp – orange, Gln – blue, Pro – yellow, Val – green).



D. Individual-level consensus of LANAr nucleotide sequences (reverse complemented) from participants in this study, sorted in descending length. U048 and U216 LANAr have identical lengths, and U020 and U216 have identical lengths. Color-coded nucleotides are shown (A – red, C – blue, G – yellow, T – green).



E. Mismatches to the DR1 dominant tandem repeat unit sequence in IR1 propagate to more than one unit at irregular intervals. Shown are IR1 direct repeat families from participants U003 and U007, with repeat sequences differentiated by color. DR2 is in blue. Nucleotides are colored: A – red, C – blue, G – yellow, T – green. Below each sequence is a graph of the GC content, with sliding window size of 8.

a few irregular DR1 TRUs were 27 bp instead of the standard 20 bp DR1 unit length (**Fig 26E**). DR1 TRU counts had the widest range (11 to 62), in addition to being the most variable within host (**Table 17**). In contrast, IR2 TRU families DR3 and DR4 were degraded in only 2 of 16 individuals (**Fig 26E**), and the mismatches propagated at regular intervals. DR4 TRU elements of up to 6 units (in U004) can be found downstream the DR4 region before the Kaposin translation start sites.

KSHV internal repeats in KSHV-infected lymphoma cell line BCBL1 were sequenced by SMRT-UMI as well, and no imperfect DR1 TRUs were observed (**Fig 27A & B**). Published KSHV genomes that had the internal repeats sequenced by Sanger were also included in **Fig 27A & B**.

No full-length Kaposin isoform from IR2 in the majority of IR2 sequences from Africa

Polymorphisms in IR2 sequences result in different sets of Kaposin family proteins being translated in different strains: Kaposins B and C, Kaposins D and E, Kaposins B and E, or none (**Fig 28**). The latter two genotypes were discovered in a previous study of viral transcriptomes in KS tumors from Uganda [119]. In the last genotype, one 'CTG' start codon has 'GCG' instead, and the other CUG codons are followed immediately by an in-frame TAG stop codon. I examined this region of the IR2 complex repeat just before DR4 in 21 individuals, including 5 with whole KSHV genomes sequenced in Chapter 4 that included this region. A third of the individuals (7/21) were revealed to encode intact Kaposins B and E reading frames, while the translation potential of full-length Kaposin isoforms was lost in others (**Fig 28**). These polymorphisms were conserved within hosts in 16 individuals, and there were no intra-sample minor SNVs detected at the 'GCG' and 'TAG' positions. Considering all 96 KSHV genomes sequenced to date, a majority of at least 55, all from Africa, had the 'CTG' to 'GCG' mutation, and 52 in addition had the 'TAG' mutation. This number excludes KSHV genomes with incomplete or ambiguous IR2 sequences.

Clinical Phenotypes associated with IR1 and IR2 diversity

Given the known biological activities of Kaposin B, I assessed whether its loss in many KSHV isolate sequences have discernible clinical phenotypes. Another genotype assessed was IR1 repeat degradation, as defined above. There were no intra-host differences detected in these genetic markers, and individuals were classified as harboring KSHV that have Kaposin isoform open reading frames or not, and IR1 either perfect or degraded. Clinical traits considered were

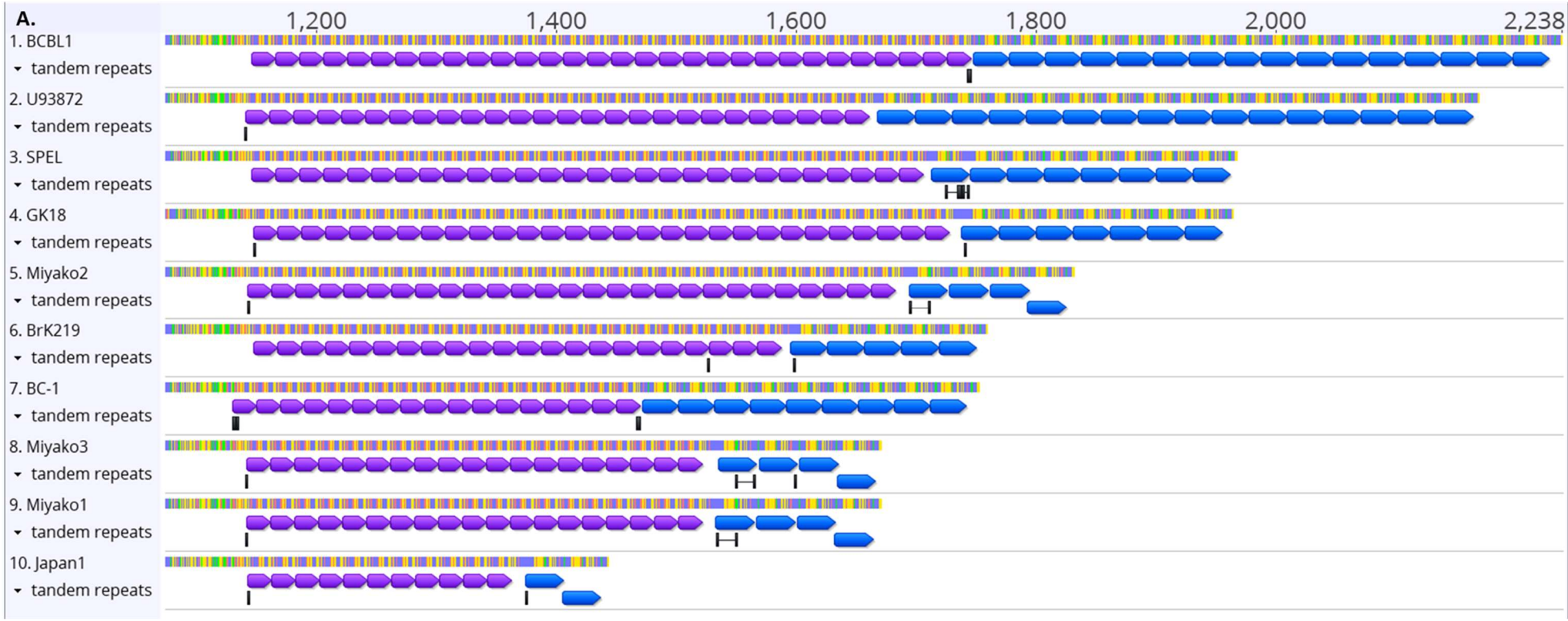
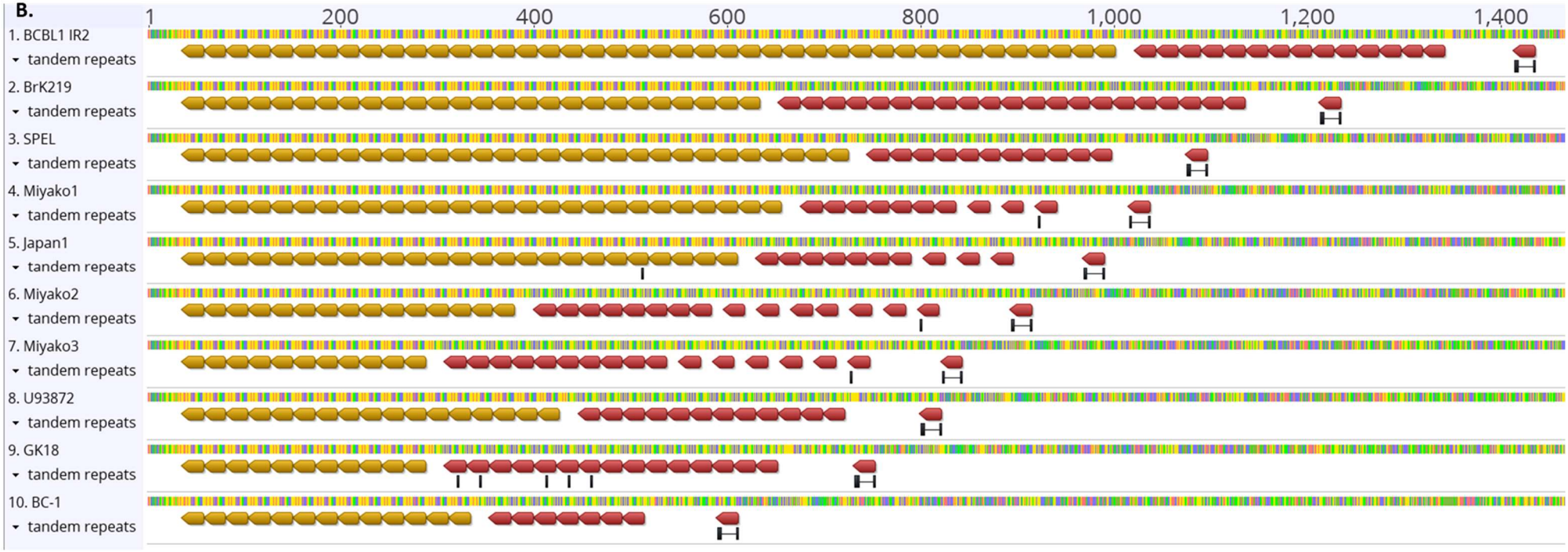


Figure 25. KSHV tandem repeat counts and mismatches in lab strains and other published KSHV genomes.

A. Individual-level consensus of IR1 sequences from BCBL1 and published genomes in other studies, sorted in descending length. SPEL, BC-1, BrK219 and BCBL-1 are derived from primary effusion lymphoma lines, while the rest are from KS biopsies. DR1 tandem repeat units are in purple, DR2 tandem repeat units are in blue, black ticks are mismatches to their respective dominant tandem unit sequence, and misaligned tandem units indicate repeat overlaps. Only BCBL1 internal repeats were sequenced by UMI and PacBio SMRT; others were sequenced by Sanger sequencing of cosmid libraries (U93872, GK18, BrK219, BC-1) or PCR products (SPEL, Japan1, Miyako1, Miyako2, Miyako3). DR2 unit counts in U93872 were only estimated from PCR products (annotation).



B. Individual-level consensus of IR2 sequences from BCBL1 and published genomes in other studies, sorted in descending length. SPEL, BC-1, BrK219 and BCBL-1 are derived from primary effusion lymphoma lines, while the rest are from KS biopsies. DR3 tandem repeat units are in gold, DR4 tandem repeat units are in red and black ticks are mismatches to their respective dominant tandem unit sequence. Only BCBL1 internal repeats were sequenced by UMI and PacBio SMRT; others were sequenced by Sanger sequencing of cosmid libraries (U93872, GK18, BrK219, BC-1) or PCR products (SPEL, Japan1, Miyako1, Miyako2, Miyako3).

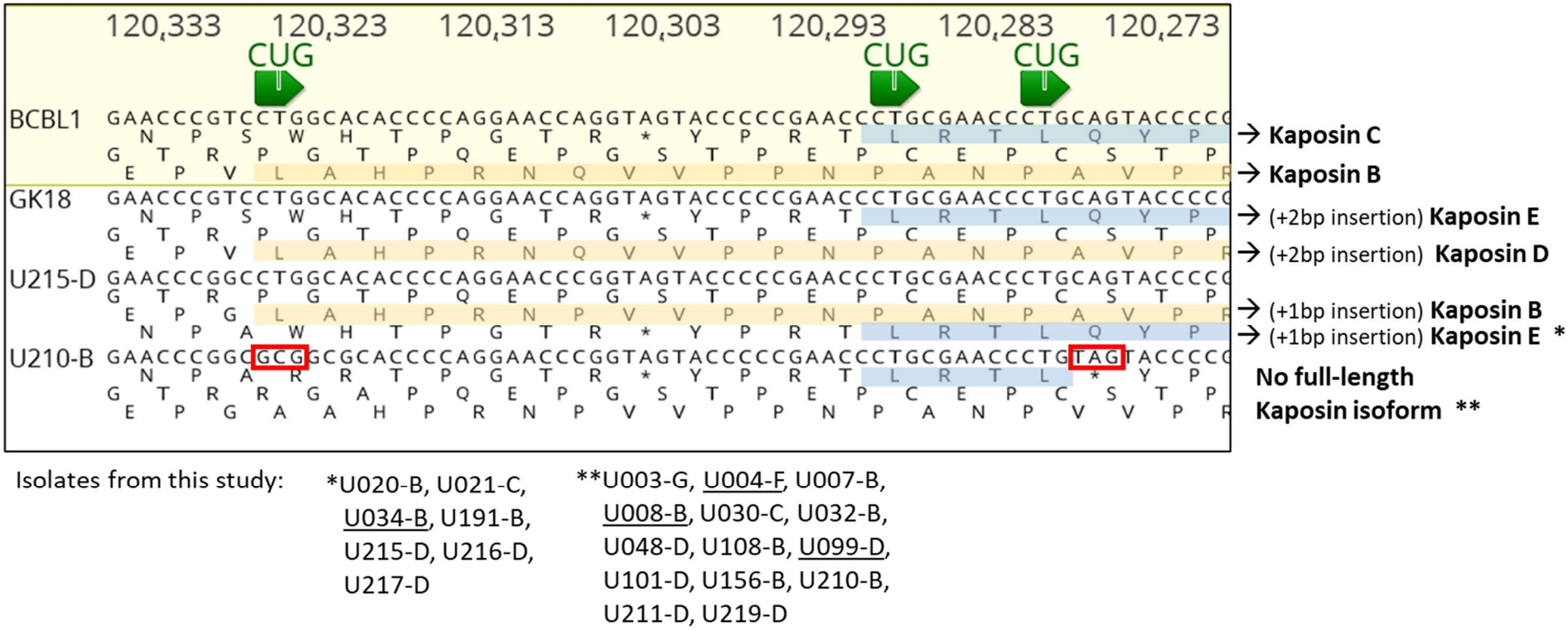


Figure 26. Translation start sites of Kaposin isoforms in different KSHV strains.

Kaposin translation start sequences of representative strains (BCBL1, GK18, U215-D and U210-B) encoding different sets of Kaposin isoforms are shown in reverse complement from their layout in a linear KSHV genome. Coordinate numbering is for BCBL1 strain. Amino acid sequences in 3 reading frames are below each nucleotide sequence. The cryptic 'CUG' translation start sites are indicated in green, and the putative products are highlighted in yellow and blue. Along with indels further downstream, the open reading frames result in different Kaposin isoforms, indicated on the right. Sequences boxed in red shows the mutations that result in loss of full-length Kaposin products from IR2. Written below are genotypes of the KSHV sequences in this study, which are largely conserved within each genotype. The underlined ones have further polymorphisms in the sequence region depicted but which do not change the translation potential. Only one isolate is shown from individuals with multiple samples sequenced.

HIV status, plasma KSHV load, CD4+ cell count, prevalence of tumors (≥4 anatomic areas), survival rates and treatment responses. The presence of an intact Kaposin coding sequence and IR1 perfect repeats were far less common in leg/foot lesions (odds ratio=0.05, p<0.001 and odds ratio=0.14, p=0.03, respectively, **Table 20**). IR1 perfect repeats additionally were not found in any HIV-positive participants. There was no other significant correlation observed in Kaposin and IR1 genotypes with survival rates, treatment response and other clinical traits considered (**Table 20, 21**). Strikingly, the presence of an intact Kaposin coding sequence was tightly correlated with perfect IR1 repeats (odds ratio=126, p=0.001, **Table 22**), even though these sequences are found in different regions of the genome.

Table 20. Clinical associations of repeat polymorphisms

	IR1 Perfect Repeat			Kaposin presence		
	OR[a]	**(95%CI)**	**P**	**OR**[a]	**(95%CI)**	**P**
Personal/Clinical Traits						
Age, per 5 years	0.97	(0.62, 1.50)	0.876	0.95	(0.61, 1.48)	0.832
Female v. Male	0.54	(0.03, 8.73)	0.663	4.37	(0.44, 43.75)	0.209
HIV Positive v. Negative	inf.	[d]		1.68	(0.12, 23.09)	0.699
plasma KSHV (log10 cp/mL), per unit [b]	1.36	(0.16, 11.55)	0.781	4.09	(0.37, 45.62)	0.252
Pt #sites(of 8) w/Lesions, per site	0.98	(0.53, 1.79)	0.945	1.33	(0.71, 2.50)	0.37
More than 4 lesion sites v. 4 or fewer	1.51	(0.11, 21.46)	0.76	3.83	(0.36, 41.15)	0.267
Any head, neck, or oral lesions	1.71	(0.12, 24.50)	0.691	5.54	(0.45, 67.84)	0.181
Any lesions other than extremities	0.36	(0.01, 9.23)	0.535	0.77	(0.05, 12.52)	0.857
KS T-stage 1	0.87	(0.05, 15.16)	0.922	0.26	(0.02, 3.74)	0.325
KS S-stage 1	0.8	(0.05, 12.55)	0.874	0.59	(0.06, 5.90)	0.651
KS I-stage 1 [c]	0.12	(0.01, 1.74)	0.12	0.32	(0.02, 4.39)	0.396
CD4+ cell count, per 100 /ul [c]	1.53	(0.55, 4.22)	0.414	0.9	(0.66, 1.23)	0.519
HIV viral load (log10 cp/mL), per unit [c]	0.02	(0.00, 3.60)	0.143	0.27	(0.01, 11.10)	0.493
Biopsy Lesion Traits						
Visit, Follow up v. Baseline	0.81	(0.29, 2.25)	0.689	0.71	(0.31, 1.63)	0.416
Lesion Type, Nodular/Fungating v. Macular	0.6	(0.09, 3.94)	0.595	0.18	(0.02, 1.57)	0.122
Lesion Site, Legs v. Other	**0.14**	**(0.02, 0.85)**	**0.033**	**0.05**	**(0.01, 0.23)**	**<0.001**
Large Lesion (>1 cm diameter v. less)	0.85	(0.15, 4.90)	0.856	3.12	(0.53, 18.39)	0.208
Large Lesion (>1 cm2 PD v. less)	1.1	(0.17, 6.98)	0.923	2.89	(0.48, 17.53)	0.249

[a] univariable analysis
[b] among those with measured plasma KSHV RNA
[c] HIV+ only
[d] No vIRF2 or IR1 perfect repeats in HIV negative (all vIRF2 observed in HIV+)
Significant values in bold

Table 21. Survival rate association of repeat genotypes

	Cox regression		
Marker	**HR**	**95%CI**	**P**
IR1 perfect repeat v. not	1.07	(0.19, 5.94)	0.94
Kaposin yes v. no	0.92	(0.18, 4.65)	0.924
	CR/PR v. SD/PD		
IR1 perfect repeat v. none	1.33	(0.15, 11.87)	0.796
Kaposin B/C yes v. no	1.56	(0.20, 11.97)	0.667

Marker is categorized by the presence of "Abnormal" sequence in any lesion sequenced at baseline
CR/PR = complete response or partial response
SD/PD = stable disease or progressive disease

Table 22. Correlation between degraded IR1 and loss Kaposin isoforms.

	IR1 Perfect Repeat						
	No		**Yes**				
	N	%	N	%	OR	(95% CI)	P
Kaposin							
No	29	90.6	1	7.1	ref.		
Yes	3	9.4	13	92.9	**126**	**(6.86, 2300.84)**	**0.001**

Significant values in bold

CHAPTER 6: Discussion & Future Directions

The low point mutational diversity of KSHV in adults with KS

My study is the first to explore KSHV diversity within individuals at the whole genome level, across tumor and non-tumor sites, and provides an unprecedented level of precision to KSHV genome sequence analysis definition in clinical specimens. In Chapter 3, KSHV genome sequences were obtained essentially error-free from 11 oral swabs and 12 KS tumor samples provided by 9 Ugandan adults with KS. More than 100 genome copies per sample were sequenced from most tumors, while fewer copies were sequenced from oral swabs due to lower input and sample quality. By incorporating dUMI, PCR misincorporation errors and template switching artifacts were substantially eliminated, permitting detection of variants as infrequent as 0.01% and a theoretical error rate of $1/10^{-9}$, approximately the DNA replication error rate in eukaryotic cells [143]. Given that herpesviruses have among the lowest replication error rates among viruses [141], next-generation sequencing errors would have corresponded to a greater proportion of suspected mutations and confounded discernment of any rare variants in a potentially heterogeneous mixture, regardless of read depth [140].

I did not observe evidence of KSHV quasispecies within individuals [181], consistent with the low mutation rates of large dsDNA viruses [141]. Less than 0.01% of all base positions in the 131-kb KSHV genomes (excluding the major repeat regions), or <1 site per genome, were found to have a detectable variant, and these were typically supported by only one dUMI-consensus read. While the possibility of artifacts cannot be totally excluded, the exceedingly low intra-sample variation observed is within the published resolution of dUMI sequencing [143]. There are reports of intra-host KSHV variability in certain KSHV-endemic populations [167], in children [173], in iatrogenic settings [169–171] and in blood of AIDS-KS patients [172], but these findings were arrived at by Sanger sequencing of PCR amplicon clones of hypervariable regions in K1 or other genes. Such protocols are more likely to detect errors that occurred during PCR. My study found virtually no intra-sample or intra-host diversity even at K1 in the 9 individuals examined. While second enrichment could conceivably bias sequencing results, comparing double-enriched oral swabs libraries to their matching, single-enriched tumor libraries resulted in identical sequences (page 47). Nevertheless, since recombination is evident in the evolutionary history of KSHV [59,77,175,178], co-infections by multiple KSHV strains have occurred, but this does not appear to be a common feature of KSHV infection.

KSHV genome aberrations are associated with KS tumors

A striking finding in Chapter 3 was the frequency of aberrant KSHV genomes in KS tumors. At least 4 of the 7 participants whose samples were examined had KSHV with major inversions, deletions or duplications in their tumors. In contrast, no aberrant genome structures were found in oral swabs from the same individuals. Moreover, aberrant genomes comprised the majority of the KSHV genomes in the tumors in which they were detected, from 1.7-fold (U008-B) to 30-fold (U020-C) more than full-length intact KSHV genomes. Further, no intact viral genomes were detected in 3 of 5 tumors examined from participant U003. I did not find evidence of integration of KSHV into human chromosomal DNA; all breakpoints detected in KSHV sequence reads, when mapped, connected to other regions of the KSHV genome.

The fact that aberrant KSHV genomes were only observed in tumor samples is intriguing yet their significance is unclear, since it is well known that tumor cells suffer substantial genomic instability [193]. KSHV genome aberrations have been reported previously: The first whole genome sequence of KSHV published reported a 33-kb portion of the KSHV unique central region duplicated into the TR region [33]. A study of 16 tumor-derived KSHV whole genomes from Zambia reported 4 that had regions with as much as 3-fold more coverage than sample average, although the regions were not specified [76]. A PCR screen for some KSHV genes showed that some KS tumors and KSHV-infected B-cell lines can harbor deleted KSHV genomes [102], and one such B-cell line proliferated faster than the parental BCBL-1 line [102]. The infecting KSHV virus in this line had an 82-kb deletion from the 5’ end of its genome, was lytic replication-incompetent, and could be packaged by a helper virus. The genomic deletion reported in that study is different from all the deleted KSHV genomes confirmed to be present in this cohort.

The genomic locations and character of tumor-associated mutations I observed in chapters 3 and 4 suggested that mutations that propagated to high copy numbers may not have been random. However, again, they could represent regions of particular susceptibility to genome instability, contribute to tumor formation, or are associated with tumor persistence due to loss of targets for immune elimination. All point mutations and structural variations I observed impacted protein coding sequences, and most rearrangement breakpoints truncated lytic gene products. Given that the KSHV genome densely encodes many immunomodulatory, angiogenic and anti-apoptotic factors [4,16,194], it is therefore possible that some mutations observed here could contribute to KS disease.

Shared mutations in distinct tumors of the same individuals

I found 4 persons to date with KSHV genome rearrangement breakpoints shared between 2-4 of their tumors. That the same viral mutational signatures were found in separate KS lesions provides strong evidence that tumor-associated and rearranged KSHV genomes can spread by metastasis, by helper virus activity, or because of residual infectivity of the aberrant genome. Identical KSHV genome rearrangements in distinct tumors implies that the viruses were clonally related, since independently acquiring the same breakpoint sites is highly improbable. One participant, U003, had a KSHV genome inversion from gene K8.1 that would likely disable production of infectious virions, yet the signature K8.1-TR breakpoint junction was present in all 5 tumors collected over 3 months. Only 3 of the 5 had intact K8.1 sequence detectable by PCR. This suggests that the genomic inversion and K8.1 loss of function was not too detrimental to KS tumor persistence and suggests a capacity for tumor-associated KSHV to continually seed tumors in other anatomic sites. Evidence of possible metastatic spread of KSHV genomes has not been previously reported, and these findings may have significant implications for understanding the progression of KS. However, consistent with the helper virus hypothesis, all deleted genomes retained the terminal repeat sequences, which contain the putative virion packaging (pac1) and cleavage sites.

Recurrent K5-K6 region overrepresentations in KS Tumors.

KSHV genome diversity was characterized in Chapter 4 from a total of 65 KS tumors from 30 individuals with advanced KS. Rearranged KSHV genomes involving read coverage over-representation of a 3.3 - 53 kb region were present in 10 individuals, in all cases including the K5 and K6 genes. Read over-coverage ranged from 1.5 to 6,000-fold with a median of 6-fold. The actual frequency of K5-K6 region overrepresentation could be higher, since only one or a minority of tumors from each participant were biopsied, and the ddPCR screen conducted quantified genomic segments outside this minimal high coverage region. There could be other rearranged KSHV sub-genomes sequenced, but low numbers of split reads connecting non-adjacent regions of the KSHV genome are indistinguishable from PCR chimera artifacts, hence they were not considered. It is striking that genome fragments containing K5 and K6 were amplified or retained in nearly all KSHV genome rearrangements that were detected, becoming the majority genome form in some tumors.

The high-coverage regions found in 10 individuals all encompass at minimum a 2.2-kb

sequence containing only the K5 and K6 genes. However, the 2 genes are not among the known KSHV oncogenes [16] and thus a direct role in tumorigenesis is not likely. K5/MIR2 is a viral ubiquitin E3 ligase best known to ubiquitinate and degrade MHC1 from the cell surface [195]. It is a membrane protein expressed upon primary infection and reactivation [196,197] and at low levels during latency [198]. Another KSHV ubiquitin E3 ligase, K3/MIR1, is far more efficient in degrading MHC1 [195], but K5 modulates numerous more immunoreceptors than K3, targeting also adherins, ligands and co-stimulators for cytotoxic T and NK cells, cytokine receptors, restriction factors and members of the SNARE, ephrin, plexin and other receptor families [195,199,200]. One downregulated ephrin receptor of note is EphA2 [199,200], one of the receptors employed by KSHV for entry into endothelial cells [201]. Amplification of K5 might suggest that cells may become less visible to cytotoxic T and NK cells and less susceptible to KSHV superinfection as EphA2 expression on host cell surface is depleted by K5. Lytic gene K6 encodes vMIP-I, a viral homolog of cellular cytokine MIP1α (CCL3). vMIP-I is a selective, high affinity agonist for chemokine receptor CCR8 and has angiogenic and chemotactic effects *in vitro* [202–204]. The inclusion of K6 in the minimum region is puzzling, given that vMIP-I is not readily detectable in KS lesions by immunohistochemistry [205].

Selection for the K5-K6 region could alternatively be acting on the DNA sequence level, with the overrepresentation of K5 and K6 simply being a passenger mutation. Nearly all K5-K6 sub-genome fragments are attached to IR1, a KSHV origin of lytic replication [125,190], and to TR sequences, which have cleaving signals for packaging [206]. These 2 DNA elements together could be sufficient for amplification *in situ* and for horizontal transmission if circularized and complemented by a 'helper' full-length KSHV genome.

To determine if selection for K5-K6 region is acting on the gene functions of K5 and K6, transcriptome sequencing will be needed to corroborate whether they are over-expressed in KS tumors that have K5-K6 copy number elevation, in comparison to other KS tumors without the mutation, as well as immunohistochemistry for vMIP-I and cell surface receptors downregulated by K5 such as Eph2A. Silencing RNA complementary to K5 and K6 can be used in primary KS tumors to directly assess the effects of inhibiting K5 or K6 expression.

Significance of IR1 region overrepresentation

While K5-K6 region overrepresentation minimally contain the K5 and K6 genes, other protein coding sequences, the IR1 region, as well as those for known non-coding RNA were also included

in most. For example, sequences encoding the T1.4 and PAN long non-coding RNAs are commonly included, with the latter found in all but one sample with the K5-K6 read over-coverage regions. These 2 transcripts are among the most highly expressed in KS tumors [119] and have indispensable roles during lytic reactivation of KSHV [155,191,192,207–209]. PAN has been shown to interact with promoters of cellular genes involved in inflammation, cell cycle regulation and metabolism, and exogenous expression of PAN alone enhanced cell growth phenotypes [210]. Additionally, virally-encoded circular RNAs encoded within PAN were abundant in clinical samples and inducible in KSHV-infected cell lines [35,36,211]. Other non-coding transcripts are potentially expressed from this region but their biological significance is unknown [209,212]. Finally, most ORFs encoded in the 14.8-kb retained region are lytic genes that have functions in subverting adaptive (K5/MIR2) [213,214] or innate immunity (K4/vCCL-2, K4.1/vCCL-3, K4.2 and K6/vCCL-1)[60,214–216], and apoptosis (K7 and ORF16/vBCL-2) [214,216], all of which may enhance the survivability of a host cell. There is evidence to suggest that KSHV lytic gene expression is crucial to KS pathogenesis [30], and that residual lytic gene expression plays a role in latent KSHV persistence *in vivo* [217]. These lytic genes were frequently and specifically amplified or retained in KS tumors at higher copy numbers than the rest of the viral genome, including the KSHV latency gene region at the 3' end. It is possible that the underrepresented regions were deleted in some genome aberrations. These findings support the idea that viral lytic genes or their expression may be necessary to KS tumor persistence.

Association of KSHV genome rearrangement breakpoints to G-quadruplexes

There was clustering of KSHV rearrangement breakpoints found in different individuals at K4.1-K4.2, near IR1, at the minor repeat region, and at ORF18-ORF19. In the latter cluster, 4 unique breakpoints were within 1.2 kb, with 2 breakpoints only 3 bases apart. IR1 and the minor repeat region are composed of tandem repeats with poly-G or C runs. Such repeats are prone to forming G-quadruplexes (G4), stable structures formed by 4 consecutive guanines folding into a tetrad and stacking with other guanine tetrads on the same strand [130,189,218]. ORF19 has a high-probability Z-DNA-forming sequence, precisely at the 2 breakpoints that were 3 bases apart. Z-DNA refers to a left-handed helix in a zigzag pattern formed by alternating purines and pyrimidines [218]. Z-DNA and G4 are alternative DNA structures that depart from canonical, right-handed B-form DNA helix. When left unresolved, they can interfere with the progression of replication forks, leading to double-stranded DNA breaks [218]. They can form during replication,

transcription and other events that generate negative supercoiling and expose single-stranded DNA [218].

Within ±500 bp from the 31 KSHV rearrangement breakpoints identified here and in other studies [78,102,165], G-quadruplex (G4) was the most common among the 5 non-B DNA forms I searched for, with others being denaturation regions, cruciforms, Z DNA and triplex DNA. No cruciform DNA was found. Compared to 1000 simulations of 31 random points on the GK18 genome, G4 had significantly higher summed scores and shorter distances to the observed breakpoints. While only 31 data points, this result is consistent with G4 being associated with recombination breakpoints in herpesviruses [219]. G4 in DNA and RNA is quite stable and require complexes of helicases and other enzymes to melt to allow unfettered access by polymerases [130]. Under replication stress and perturbation of the DNA damage response pathways during KSHV lytic replication [87], the G4-rich regions of KSHV may have become particular regions of susceptibility to double-stranded DNA breaks and non-homologous end joining.

K8.1 inactivating mutations

Eight of 30 individuals were found to have had 9 different inactivating mutations in the late lytic gene K8.1. Additionally, K8.1 was the only KSHV gene that had intra-host differences in more than two members of this cohort. All the K8.1 mutations observed were putatively inactivating -- nonsense mutations, transcription start site deletions and rearrangement breakpoints. The K8.1 mutations were typically local only to one tumor since other tumors sampled from the same individual still had intact K8.1. In fact, one person had 2 different K8.1 mutations. Another individual had K8.1 gene sequence interrupted by a genomic inversion breakpoint; however, this was detected by PCR in all 5 of his tumors tested. Full-length, uninterrupted K8.1 could only be detected in 3 of those five tumors. No K8.1 inactivating mutations were found in matching oral swab samples. K8.1 truncations had been reported previously, and all were from KS tumor isolates. The original GK18 isolate has a 74-bp deletion at the 3' end (GenBank ID AF148805 K8.1 annotation); the Zambian isolate ZM124 (GenBank ID: KT271466) has a 25-nt deletion resulting in a frameshift and premature stop [76], and the Japanese isolate Miyako1 has a stop codon early in its first exon (GenBank ID LC200586 miscellaneous annotation). The unique sequence mutations in every individual implies that they were acquired independently.

K8.1 encodes an envelope glycoprotein that interacts with heparin sulfate for attachment [220–223]. It is not required for entry into primary endothelial [221] or 293 cells [188], although it

had been shown to be necessary for infection of primary and cultured B-cells [224]. The K8.1 protein is among the most immunogenic KSHV proteins [225–227] and can be targeted by cellular immunity [228]. The truncations observed here and in other studies removed its C-terminal transmembrane anchor domain [227]. It is therefore conceivable that the preponderance of K8.1 mutations might be due to immune elimination of cells expressing K8.1 glycoproteins. Evading immune responses by impairing K8.1 expression may confer better survival of the host tumor cell. However, K8.1 sequences are highly conserved *across* individuals, unlike that of K1, a KSHV lytic membrane protein with hypervariable extra-cellular domains [4]. Given that (1) full-length K8.1 sequences are conserved across individuals and that (2) K8.1 inactivating mutations were often not present in all tumors of a given individual, the selection against K8.1 expression could be occurring only locally at the tumor site. Characterizing the immunoreactivity of individuals harboring tumors with inactivated K8.1 will be needed to substantiate immune selection.

Point mutations in microRNA-K10

An intra-host miR-K10 G16A mutation (GK18 position 118,082), was found in two individuals, out of 10 that had KSHV genomes sequenced from >1 sample. In participant U003, all his tumors had G16A in miR-K10, while the oral swab counterpart maintained the database consensus genotype G16. In the other participant, U156, miR-K10 G16A mutation was present in only 1 of his 4 tumors sequenced. The intra-host single nucleotide difference stands out given the absence of other synonymous intra-host mutations along the entire ~131 kb of non-repeat KSHV genome sequences. KSHV genomes from all matching pairs of samples in 10 individuals differed by at most two point mutations.

The 23-nt U<u>AGUGUU</u>GUCCCCC**CG**AGUGGCC sequence of miR-K10 is tightly conserved in the vast majority of KSHV genomes sequenced to date [229]. Among published KSHV genomes, only tumor-derived ZM106 (GenBank KT271458) also had miR-K10 G16A. Aside from this mutation, a non-consensus miR-K10 C15T mutation was also found in 2 other individuals in this cohort, although there were no intra-host differences detected at the C15 position. Both the C15T and G16A polymorphisms (bolded above) are outside the miR-K10 seed sequence (underlined) [209]. The KSHV microRNA miR-K10 is weakly transforming [230], and a single A to G change at position 2, inside the seed sequence, has been reported to alter its tumorigenicity [231]. The miR-K10 G16A mutation is predicted to result in a slightly more stable stem loop precursor (ΔG -32.40à-32.70). The biologic implications of this change are not clear, but polymorphisms within

and around KSHV miRNAs have been reported to affect viral microRNA levels in KS lesions [72]. Though the tumor-specificity of the miR-K10 G16A mutation is suggestive, there is insufficient evidence to say that it was selected for in tumors. If the miR-K10 G16A mutation can be observed to impact the processing, silencing activity, and expression level of miR-K10 *in vitro*, these differential effects can be a basis for selection. Finally, the effects of G16A and C15T on the tumorgenicity of miR-K10 can be tested in a transformation assay in comparison to the database consensus sequence to determine a biological impact.

Novel genotypes of K4.2 and K11.2

Genes K4.2 and K11.2 were revealed to be polymorphic across individuals, and they were not tumor-associated. Considering all 96 sequenced KSHV genomes to date, K4.2 has 3 coding sequence length clusters: the prototypic 549-bp full-length form, a 378 or 369-bp truncated form, and a 237-bp or shorter truncated form. A majority of KSHV whole genomes from Africa, now comprising most of the sequenced KSHV genomes, carry the truncated K4.2 forms. The K11.2 polymorphism relates to a 174-bp duplication in its central domain, which was not present in 18 of 94 published whole KSHV genomes. Both K4.2 and K11.2 polymorphisms do not wholly coincide with the phylogenetic groupings of KSHV genomes, although the 237-bp truncated K4.2 appears to largely be monophyletic to genome type P2.

Full-length K4.2 expresses an immediate-early protein that inhibits the endoplasmic reticulum chaperone protein pERP1, which enhances KSHV glycoprotein maturation, among other effects [232]. All K4.2 truncations delete its putative transmembrane domain, which abrogates the localization of K4.2 to the ER [232]. Completely deleting K4.2 diminishes the expression of KSHV glycoproteins K8.1 and gB, dampening infectious virus titer by 4-fold [232]. This effect on K8.1 production is noteworthy, because K4.2 truncation and K8.1 inactivation were inversely correlated i.e., tended to not occur together in this cohort ($p=0.013$). Suggestively, defective K4.2 may have the same effect as K8.1 inactivation in depressing K8.1 production. This can be tested by immunohistochemistry for surface expression of K8.1 in KS tumors with and without the K4.2 truncations but with intact K8.1 sequence. If K4.2 truncation and K8.1 inactivation have the same phenotypic effect, this finding will further support the hypothesis that K8.1 expression is highly selected against. Another prediction is that natural KSHV strains with K4.2 truncations will have lower levels of K8.1 glycoproteins on their virion surface.

K11.2 encodes a viral homolog of cellular interferon regulatory factor 2 (vIRF-2), which modulates the antiviral signaling of interferon [233,234] and regulates KSHV lytic replication by suppressing KSHV early lytic gene expression [235]. Endothelial cells infected with vIRF-2 deletion KSHV mutants had increased and prolonged expression of KSHV lytic proteins K-bZIP and ORF45 following induced reactivation and de novo infection [235]. While the C- and N-terminal domains of vIRF-2 have been characterized [235,236], the function of the central 174-bp sequence duplication has not been elucidated. Given that duplication was not common in our cohort (21 of 30 individuals) and was not associated with tumors or with an observable tumor characteristic, the duplicated sequence may be of little consequence in KS. However, its effect on the role of vIRF-2 in suppressing KSHV lytic expression can be tested in an infection assay, where K-bZIP and ORF45 expression levels are measured following chemical reactivation.

Clinical phenotypes of tumor-associated mutations and polymorphisms.

K5-K6 region overrepresentation and K8.1 inactivation were more likely in nodular rather than macular tumors ($p<0.001$ and $p=0.010$, respectively), while it was the opposite for K4.2 truncations ($p=0.028$). Nodular forms are typical of later stages of KS tumor development. Individuals with more widespread lesions (>4 anatomic areas) were less likely to have K5-K6 over-representation ($p=0.01$) and K8.1 inactivation ($p=0.005$), which suggests that these mutations limit the anatomic spread of KS tumors. Non-database consensus miR-K10 mutations were rarer in tumors <1 cm in diameter ($p=0.01$), suggesting that miR-K10 can be involved in the aggressiveness of tumor growth. When considering mortality, a trend for the miR-K10 mutations was found in individuals with lower survival rates ($p=0.053$), which has to be verified in a larger cohort.

Study participants entered the study at varying stages of their illness, so there likely is heterogeneity in both the tumor stages sampled at baseline as well as in clinical course following admission. Tumors of study participants were not exhaustively sampled, which implies a likely under-counting of intra-host mutations. While my findings may suggest that perhaps KS tumors have a tendency to develop K5-K6 region overrepresentation and K8.1 inactivation mutations, these 2 most common tumor-associated viral mutations observed were still in only a minority of KS tumors sampled. It follows that the tumor-associated KSHV mutations were not required for tumorigenesis, but they could still be influential or be a trend during KS tumor development. There in addition may still be a remote possibility of undiscovered but common and clinically relevant

KSHV mutations. Longitudinal screening targeted to K5-K6 copy number elevation, K8.1 inactivation and miR-K10 point mutations over time may be needed to verify the mutation trend suggested. Their potential as contributing driver mutations can be assessed by observing their source tumors following chemotherapy and following immune reconstitution upon initiation of ART for HIV-positive patients. If they are contributing driver mutations, one prediction is that tumors harboring such mutations would be less amenable to these treatments.

KSHV repeats variation as revealed by SMRT-UMI

The three KSHV major internal repeats, IR1, IR2 and LANAr, comprise transcripts and protein domains that play important roles throughout the KSHV infection cycle [32,121,237]. Yet their diversity is rarely analyzed because of long, repetitive sequences and >60% GC content. Additionally, PCR amplification, which is an obligate step when sequencing from low viral load clinical samples, is inherently error prone. These challenges were overcome with SMRT-UMI. PacBio SMRT sequencing is relatively insensitive to GC content [144], while UMIs remove misincorporation errors and frame-shifting indels. The consensus of reads with identical UMI converges on the pre-amplification DNA sequence template, allowing for single molecule resolution of the intra-sample diversity of KSHV internal repeats even from low viral load oral swabs. There are 2 important caveats. First, because the primers used to amplify the repeat regions target conserved flanking sequences, rearrangements involving other parts of the KSHV genome would not be detected. Second, because size selection was done before SMRT sequencing, and because smaller sizes are more efficiently PCR amplified and sequenced, DNA molecules far from the mean size ($\pm$400 bp) would less likely be detected. This was minimized by extracting and size selecting for DNA 1 kb greater than the visible band sizes.

Single molecule sequences of the 3 KSHV internal repeats were successfully obtained from 1 – 2 KS tumor biopsies and oral swabs of 16 individuals, such that 4 tumor-oral swab pairs each of IR1, IR2 and LANAr were sequenced. UMI-consensus reads per sample was only a median of 21.5, but the sequencing accuracy allowed for definitive observations on intra-host sequence diversity. Nearly all SNVs were found in only one UMI-consensus read. An average of 0.23% of base positions in the repeat regions had a detectable minor SNV, more than the 0.17% found for the rest of the KSHV genome [165]. These values are not directly comparable, however, because duplex UMIs on both ends of short reads were employed for whole KSHV genome dUMI-sequencing. With SMRT-UMI, only single-stranded UMI was used and only one end is tagged,

because each UMI-tagging step by PCR requires additional purification that can lead to more loss of template DNA. In contrast, both ends of sheared DNA are simultaneously ligated with adapters containing dUMI when doing whole-genome dUMI-sequencing.

Higher diversity in repeats may be expected, since repeat sequences diversify at rates much higher than nucleotides substitutions [115]. Furthermore, it has been reported that tumor genomes have increased repeat instability early during malignant transformation while point mutations accumulate only late in their transformation [131]. Although highly conserved overall, there were distinct populations of KSHV internal repeats detected within more than half the individuals examined, with minor sequence populations defined as being at least 2 UMIs with a shared non-consensus SNV or TRU count. Variation in TRU counts when present were typically only by one or 2 units. Differences between sample-consensus sequences as large as 7 and 19 TRU were found in only 2 tumor-oral swab pairs of IR1. Variation by one TRU can be explained by simple replication slippage, whereas larger changes can be attributed more to recombination and repair mechanisms [115,130].

Heterogeneity of LANAr sequences

LANAr encodes the highly acidic, repetitive central domain of LANA, the key viral protein in maintaining KSHV latency and episome propagation to daughter cells [41]. LANA interacts with a myriad of proteins, e.g., transcriptional activators, cell cycle regulators, tumor suppressors, antiviral sensors, heterochromatin modifiers, DNA repair proteins [121,238]. Deletion of LANAr eliminates activation of a viral promoter and results in diffuse intranuclear distribution of LANA [239]. LANAr also contributes to the extraordinarily long half-life of LANA; it retards translation, inhibits proteasomal degradation, and interferes with antigen processing for MHC1 peptide presentation [240–242], and it induces efficient ribosomal frameshifting to produce alternative LANA isoforms [243]. LANAr has >61% GC content, and the G-rich sequences of LANAr form G4 secondary structures in mRNA that negatively regulates translation [244].

The LANAr region is highly variable in length and sequence between strains [116]. In the cohort I studied, LANAr ranged from 1.2 to 1.9 kb in 16 individuals, but strikingly, had identical lengths in 2 pairs of individuals, though DNA sequences were still unique. LANAr sequence variations were always in-frame at the sample-consensus. LANAr was the most stable KSHV internal repeat intra-host.

Variation in LANAr could conceivably influence the affinities of LANA isoforms to its many protein targets. Amino acid repeats in mammals are overrepresented in transcription factors and proteins that have numerous interaction partners [245,246]. They are thought to serve as flexible linkers between domains and stabilize protein-protein interactions [245,246]. However, it may be challenging to discriminate the effects of LANAr variation *in vitro*, as repeat variations typically do not result in deleterious changes but rather, in small incremental phenotypic changes [115]. Additionally, the instability of amino acid repeats has been attributed more to GC content than to expression level or to adaptive evolution [245]. Hence, I would hypothesize that LANAr at the extreme length ranges observed will have no phenotypic differences, except in the amount of detectable G4 and retardation of translation.

IR1 and IR2 imperfect repeats

Gamma-herpesviruses have evolved two functional Ori-Lyts for optimal fitness in nature, with either one preferentially employed depending on host cell type and bound by different cell type-specific proteins [247]. Both IR1 and IR2 have Ori-Lyt activity in a plasmid transient-transfection replication assay [125,190], but it has been observed in the context of a full KSHV genome bacterial artificial chromosome (BAC) that IR1 alone but not IR2 alone is sufficient for propagation in Vero cells [248]. IR1 and IR2 GC-rich tandem repeats contain GRGGC motifs and similar elements of the polyomavirus non-coding control region (NCCR) [249]. These repeat elements serve as transcription factor binding motifs and regulate polyomavirus gene expression [249]. In BK and SV40 polyomaviruses, mutations that delete or disrupt Ori-Lyt repeats result in viruses unable to fully replicate but are highly transforming [128,129]. In JC polyomavirus the enhancer and promoter activities are cell type specific, exerting their effects by forming secondary structures that impede and facilitate transcription of early and late genes, respectively [250]. If the KSHV GC-rich tandem repeats function analogously, mutations in them may alter transactivation regulation and binding of cell-specific nuclear factors.

Tandem repeats add another level of diversity, which comes from mismatches within the same TRU family. Between individuals, the DR1 family in IR1 was the most variable in length and by far the most frequently degraded, defined here as having at least 2 TRU with <20% mismatches to the dominant TRU sequence. Degraded DR1 usually had TRU mismatches propagated to other TRU at irregular periodicities, and some TRUs were overlapping or were longer than the normal 20 bp. Altogether IR1 had degraded TRU families more often (12/16 individuals) than IR2 (2/16

individuals). The degraded repeats were also conserved in all tumor and oral swab KSHV sampled from within individuals. Degraded IR1 and IR2 can be found also in 3 and 1, respectively, of 10 published whole KSHV genome sequences that included the internal repeats.

GC-rich repeats are prone to forming G-quadruplex secondary structures, which are mutagenic at the single nucleotide level [251]. Imperfect repeats dramatically increase the stability of repeats [115,130]. Point mutations in EBV repeats have been suggested to be signatures of genetic exchange or DNA damage repair when TRUs are treated as internally controlled sequences [122]. In the KSHV internal repeats from this cohort, purifying selection could be maintaining the fidelity of IR2 TRU sequences in persons with KS, while selection was relaxed in IR1. Another possibility is that degraded IR1 can contribute to KS disease. It is one of the KSHV transactivation initiation sites for lytic gene expression, and inefficient, "leaky" or abortive lytic expression has been thought to be a possible factor in the pathogenesis of KS since full reactivation leads to lysis of the host cell [9,32]. In EBV an inverse relationship has been observed among EBV strains between their efficiencies in lysing or immortalizing B-cells [49]. One fairly transforming but poorly B-cell lytic EBV strain, B95-8, has among other mutations one of its two Ori-Lyts deleted [252]. JC virus isolates from the central nervous system of progressive multifocal leukoencephalopathy (PML) patients have an Ori-Lyt rearrangement differentiating them from urine isolates, the latter of which are thought to be the transmissible form [249]. Whether KSHV with heavily degraded IR1 can successfully establish infection in a new host organism is unknown.

The importance of secondary structures in Ori-Lyt activity has already been observed for EBV. Its Ori-Lyt sequences have conserved mirror repeats that adopt a triple helical DNA structure [124]. Its capacity to mediate lytic DNA replication was abrogated by introducing point mutations that disrupt the triple helix but was restored to a degree by compensatory mutations recreating the mirror repeats and triple helix properties [124].

IR1 repeat degradation was not significantly associated with high plasma KSHV load, high CD4+ cell count, prevalent tumors (>4 anatomic areas), lower survival rate and treatment responsiveness. Degraded IR1 was not found in the 2 HIV-negative participants in this cohort of 16. The IR1 mutations observed could have clinical effects more subtle than can be determined from this small cohort or could simply be stochastic or passenger mutations. In either case, a transient plasmid replication assay of IR1 and IR2 with divergent structures is one way to determine the biological impact of the imperfect repeats. As well, biophysical characterization [189] can be done to determine what non-B-DNA structures form in these GC-rich repeats and to

assess their roles in recruiting cell-specific transcription factors and in regulating lytic transcription. Finally, assaying their transactivation capacities [155,191,248] separately in different cell types in the context of a full-length KSHV genome BAC can give more insights into the cell types that are important for successful KSHV infection of an organism and, if degraded repeats were to be found deficient in their functions, provide insights also into the cell types that give rise to these mutations in KS tumors.

Loss of full-length Kaposin isoforms from IR2

In KSHV strains typified by viruses infecting BCBL-1, Kaposins B and C are translated from CUG start sites ~70 bp downstream of IR2 DR4, in the 3' to 5' orientation on the KSHV genome [117]. In strains typified by GK18, the reading frames are shifted due to a 2-bp insertion between DR3 and DR4, so Kaposins D and E are encoded instead [118,119]. There were polymorphisms in the Zambian and Ugandan isolates that engender either both Kaposins B and E reading frames, or loss of the translation potential of all full-length Kaposin isoforms from inside IR2 [119]. Kaposin B is the best characterized isoform among the Kaposin family proteins. It promotes stability of mRNA with AU-rich elements, particularly those encoding cytokines and lymphatic differentiation transcription factor PROX1 [253–255]. It also associates with the cell cycle regulator c-Myc to promote angiogenesis [256,257] and phosphorylates proto-oncogene STAT3, leading to increased activation of pro-inflammatory genes [258].

Considering all 96 whole KSHV genomes to date, including those derived from my thesis work, a majority of at least 55, all from Africa, have polymorphisms that lead to the loss of all full-length Kaposin isoform translation potential downstream of DR4. The Kaposin translation start genotype was conserved within hosts. It is noteworthy that there were no other open reading frames in that region that could supplant Kaposin B, given the significance suggested by its known biological activities. The nearest CUG or AUG start codons upstream are beyond the transcription start and splice sites of Kaposin transcripts. Downstream there is a 'CUG' within every DR3 TRU that could initiate translation [117,118], but the preceding DR4 sequence that would be missing encodes the protein-binding domain of Kaposin B [253]. It is also noteworthy that loss of all full-length Kaposin isoforms from IR2 resulted in no observable clinical phenotypic difference. Whether full-length Kaposin isoforms can be translated or not from IR2 had no detectable association with survival rates, treatment response or other clinical traits considered in our cohort of 16 individuals.

CHAPTER 7: Conclusions and Future Directions

How chronic KSHV infection leads to KS disease is still poorly understood. Herpesviruses, while known to have the lowest mutation rates among virus classes, can still exhibit strain differences and acquire mutations that affect their virulence. Developing and employing innovative sequencing methodologies, I comprehensively examined the viral genomic component of KS directly from clinical specimens at the whole viral genome level. I sought to discover viral genomic polymorphisms that could shed light on the pathogenesis and progression of KS. While KS is likely an end result from a combination of many factors, the central thrust of my thesis was to discover KSHV genomic changes likely to play a role in the natural history of KSHV infections.

In Chapter 3, highly accurate deep sequencing revealed that whole KSHV genomes in paired oral swabs and tumors from individuals with advanced KS were virtually identical at the point mutational level. The use of dUMI provides a proof of concept for utilizing this technique to study other DNA viruses such as human cytomegalovirus, which can exhibit substantial intra-host genome heterogeneity [122]. Where there were differences, the viruses detected in saliva had the database consensus genotype while viruses detected in some tumors had novel mutations. I found that KS tumors can harbor KSHV with genomic aberrations affecting coding regions, which raised the possibility that these mutations could alter the function of these genes and were of biologic significance. The most prominent mutations were the read over-coverage of a region around IR1 and K8.1 inactivating mutations. Furthermore, this study demonstrated that signature KSHV genome aberrations can be found in distinct tumors from the same individual, suggesting that infected tumor cells can seed KS lesions in other sites. Another interpretation suggested is that aberrant or deleted genomes can be propagated through helper viruses within a host.

Chapter 4 followed up on my observations of the frequent IR1 region read over-coverage and K8.1 inactivating mutations. I analyzed multiple tumors from the same individuals in a larger, 30-person cohort. Read over-representation specific to IR1 region and intra-host K8.1 inactivating mutations were common in persons afflicted with advanced KS. The genomic region of frequent read over-coverage was found to minimally encompass the K5 and K6 genes. This sub-genomic fragment differed in length across individuals, but there appears to be clustering of their breakpoints with G4-forming sequences in the KSHV genome. The K5-K6 and K8.1 mutations arose independently in a third of individuals in this cohort (different subsets), evidenced by unique breakpoint locations or sequence mutations, and they were often observed in only a subset of

tumors sequenced from the same person. Additionally, tumor-associated, intra-host miR-K10 point mutation was observed in 2 individuals. Aside from intrahost mutations, *inter*host polymorphisms of K4.2 and K11.2 were found but had no detectable changes within a host. Exploratory analyses revealed some potential association of these mutations and polymorphisms with clinical presentation of KS. Among the most significant, K5-K6 region overrepresentation was more likely in nodular rather than macular tumors, and it and K8.1 inactivation were much less likely in persons with widely disseminated KS lesions. These findings call for targeted surveys of these genes and more studies into their biological roles in an *in vivo* setting. Consistent viral polymorphisms may be indicative of or be influential to tumor progression and escape from immune selection.

Chapter 5 described both the inter-host and intra-host diversity in the often-overlooked major internal repeat sequences of KSHV, also highlighting the application of SMRT-UMI sequencing I helped develop. I found that IR1, IR2 and LANAr can reveal distinct variants within and between tumors and oral swabs from persons with advanced KS. As such they were found to have higher variability than the rest of the KSHV genome. LANAr was found to have the most stable internal repeat but was still quite heterogeneous across individuals. IR2 sequences that do not encode full-length Kaposin protein isoforms were found in a majority of KSHV isolates from Africa. IR1 had imperfect repeats much more often than IR2. Follow up work can be done to test the transactivation and Ori-Lyt activities of IR1 and IR2 separately in different cell types, identify cell-specific host proteins that associate with them, examine their secondary structures through biophysical means, and determine the impact of the degraded IR1 DR1 families observed here on the functionality of IR1.

To conclude, I uncovered several recurring KSHV genome mutations and mutation patterns that occur to KSHV genomes within persons afflicted with KS, showing that KS disease has a viral genetic component. The frequency and clinical associations of these tumor-associated KSHV mutations suggest that viral genetics play a role in KS development and presentation, if not pathogenesis. They suggest numerous hypotheses about the potential of KSHV mutations as contributing driver mutations that can explain aspects of KS and ultimately, inform clinical treatments and management of the disease.

BIBLIOGRAPHY

1. Hoffmann C, Sabranski M, Esser S. HIV-Associated Kaposi's Sarcoma. Oncol Res Treat. 2017;40: 94–98. PMID: 28259888

2. Bishop BN, Lynch DT. Cancer, Kaposi Sarcoma. StatPearls. StatPearls Publishing; 2018. PMID: 30521260

3. Gonçalves PH, Uldrick TS, Yarchoan R. HIV-associated Kaposi sarcoma and related diseases. Aids. 2017;31: 1903–1916. PMID: 28609402

4. Abere B, Schulz TF. KSHV non-structural membrane proteins involved in the activation of intracellular signaling pathways and the pathogenesis of Kaposi's sarcoma. Curr Opin Virol. 2016;20: 11–19. PMID: 27518127

5. Schneider JW, Dittmer DP. Diagnosis and Treatment of Kaposi Sarcoma. Am J Clin Dermatol. 2017;18: 529–539. PMID: 28324233

6. Mesri EA, Cesarman E, Boshoff C. Kaposi's sarcoma and its associated herpesvirus. Nature Reviews Cancer. NIH Public Access; 2010. pp. 707–719. PMID: 20865011

7. Kaposi. Idiopathisches multiples Pigmentsarkom der Haut. Arch für Dermatologie und Syph 1872 42. 1872;4: 265–273.

8. Vangipuram R, Tyring SK. Epidemiology of Kaposi sarcoma: review and description of the nonepidemic variant. Int J Dermatol. 2018. PMID: 29888407

9. Fields BN, Knipe DM (David M, Howley PM. Fields virology. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013.

10. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, et al. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science. 1994;266: 1865–9. PMID: 7997879

11. Paoli PDE, Carbone A. Kaposi ' s Sarcoma Herpesvirus : twenty years after its discovery. Eur Rev Med Pharmacol Sci. 2016;20: 1288–1294.

12. Mui UN, Haley C, Tyring SK. Viral Oncology: Molecular Biology and Pathogenesis. J Clin Med. 2017;6: 111.

13. Moore PS, Chang Y. The conundrum of causality in tumor virology: the cases of KSHV and MCV. Semin Cancer Biol. 2014;26: 4–12. PMID: 24304907

14. Sarid R, Gao S-J. Viruses and human cancer: from detection to causality. Cancer Lett. 2011;305: 218–27. PMID: 20971551

15. Hill AB. The Environment and Disease: Association or Causation? J R Soc Med. 1965;58: 295–300. PMID: 14283879

16. Yan L, Majerciak V, Zheng ZM, Lan K. Towards Better Understanding of KSHV Life Cycle: from Transcription and Posttranscriptional Regulations to Pathogenesis. Virologica Sinica. Science Press; 2019. pp. 135–161.

17. Polizzotto MN, Uldrick TS, Wyvill KM, Aleman K, Marshall V, Wang V, et al. Clinical

Features and Outcomes of Patients With Symptomatic Kaposi Sarcoma Herpesvirus (KSHV)-associated Inflammation: Prospective Characterization of KSHV Inflammatory Cytokine Syndrome (KICS). Clin Infect Dis An Off Publ Infect Dis Soc Am. 2016;62: 730. PMID: 26658701

18. Liu Z, Fang Q, Zuo J, Minhas V, Wood C, Zhang T. The world-wide incidence of Kaposi's sarcoma in the HIV/AIDS era. HIV Med. 2018;19: 355–364.

19. La Ferla L, Pinzone MR, Nunnari G, Martellotta F, Lleshi A, Tirelli U, et al. Kaposi' s sarcoma in HIV-positive patients: the state of art in the HAART-era. Eur Rev Med Pharmacol Sci. 2013;17: 2354–2365. PMID: 24065230

20. Global HIV and AIDS statistics | AVERT. [cited 3 Feb 2019].

21. Hernández-Ramírez RU, Shiels MS, Dubrow R, Engels EA. Cancer risk in HIV-infected people in the USA from 1996 to 2012: a population-based, registry-linkage study. Lancet HIV. 2017;4: e495–e504. PMID: 28803888

22. Shiels MS, Engels EA. Evolving epidemiology of HIV-associated malignancies. Curr Opin HIV AIDS. 2017;12: 6–11. PMID: 27749369

23. Graf L, Gillessen S, Korte W. HIV-Associated Kaposi ' s Sarcoma with a High CD4 Count and a Low Viral Load. N Engl J Med. 2007;357: 1352–1353. PMID: 17898112

24. Krown SE, Lee JY, Dittmer DP, AIDS Malignancy Consortium. More on HIV-associated Kaposi's sarcoma. N Engl J Med. 2008;358: 535–6; author reply 536. PMID: 18234764

25. von Braun A, Braun DL, Kamarachev J, Gunthard HF. New Onset of Kaposi Sarcoma in a Human Immunodeficiency Virus-1-Infected Homosexual Man, Despite Early Antiretroviral Treatment, Sustained Viral Suppression, and Immune Restoration. Open Forum Infect Dis. 2014;1: ofu005–ofu005. PMID: 25734079

26. Cao W, Vyboh K, Routy B, Chababi-Atallah M, Lemire B, Routy JP. Imatinib for highly chemoresistant Kaposi sarcoma in a patient with long-term HIV control: a case report and literature review. Curr Oncol. 2015;22: 395. PMID: 26628884

27. Marcoval J, Bonfill-Ortí M, Martínez-Molina L, Valentí-Medina F, Penín RM, Servitje O. Evolution of Kaposi sarcoma in the past 30 years in a tertiary hospital of the European Mediterranean basin. Clin Exp Dermatol. 2019;44: 32–39. PMID: 29934954

28. Dubrow R, Qin L, Lin H, Hernández-Ramírez RU, Neugebauer RS, Leyden W, et al. Association of CD4+ T-cell Count, HIV-1 RNA Viral Load, and Antiretroviral Therapy With Kaposi Sarcoma Risk Among HIV-infected Persons in the United States and Canada. J Acquir Immune Defic Syndr. 2017;75: 382–390. PMID: 28394855

29. Minhas V, Wood C. Epidemiology and transmission of kaposi's sarcoma-associated herpesvirus. Viruses. Multidisciplinary Digital Publishing Institute (MDPI); 2014. pp. 4178–4194. PMID: 25375883

30. Andrei G, Snoeck R. Kaposi's sarcoma-associated herpesvirus: The role of lytic replication in targeted therapy. Current Opinion in Infectious Diseases. 2015. pp. 611–624. PMID: 26524334

31. Mariggiò G, Koch S, Schulz TF. Kaposi sarcoma herpesvirus pathogenesis. Philos Trans

R Soc B Biol Sci. 2017;372: 20160275.

32. Aneja KK, Yuan Y. Reactivation and Lytic Replication of Kaposi's Sarcoma-Associated Herpesvirus: An Update. Front Microbiol. 2017;8. PMID: 28473805

33. Russo JJ, Bohenzky RA, Chien MC, Chen J, Yan M, Maddalena D, et al. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). Proc Natl Acad Sci U S A. 1996;93: 14862–7. PMID: 8962146

34. Bruce AG, Ryan JT, Thomas MJ, Peng X, Grundhoff A, Tsai C-C, et al. Next-generation sequence analysis of the genome of RFHVMn, the macaque homolog of Kaposi's sarcoma (KS)-associated herpesvirus, from a KS-like tumor of a pig-tailed macaque. J Virol. 2013;87: 13676–93. PMID: 24109218

35. Toptan T, Abere B, Nalesnik MA, Swerdlow SH, Ranganathan S, Lee N, et al. Circular DNA tumor viruses make circular RNAs. Proc Natl Acad Sci U S A. 2018;115: E8737–E8745. PMID: 30150410

36. Tagawa T, Gao S, Koparde VN, Gonzalez M, Spouge JL, Serquiña AP, et al. Discovery of Kaposi's sarcoma herpesvirus-encoded circular RNAs and a human antiviral circular RNA. Proc Natl Acad Sci U S A. 2018;115: 12805–12810. PMID: 30455306

37. Mariggiò G, Koch S, Schulz TF. Kaposi sarcoma herpesvirus pathogenesis. Philos Trans R Soc Lond B Biol Sci. 2017;372. PMID: 28893942

38. Kumar B, Chandran B. KSHV Entry and Trafficking in Target Cells—Hijacking of Cell Signal Pathways, Actin and Membrane Dynamics. Viruses. 2016;8. PMID: 27854239

39. Chen J, Zhang X, Schaller S, Jardetzky TS, Longnecker R. Ephrin Receptor A4 is a New Kaposi's Sarcoma-Associated Herpesvirus Virus Entry Receptor. MBio. 2019;10: e02892-18. PMID: 30782663

40. Purushothaman P, Thakker S, Verma SC. Transcriptome Analysis of Kaposi's Sarcoma-Associated Herpesvirus during De Novo Primary Infection of Human B and Endothelial Cells . J Virol. 2015;89: 3093–3111. PMID: 25552714

41. Purushothaman P, Dabral P, Gupta N, Sarkar R, Verma SC. KSHV Genome Replication and Maintenance. Front Microbiol. 2016;7: 54. PMID: 26870016

42. Bravo Cruz AG, Damania B. In vivo models of oncoproteins encoded by Kaposi's sarcoma-associated herpesvirus. J Virol. 2019; JVI.01053-18. PMID: 30867309

43. Hancock MH, Skalsky RL. Roles of Non-coding RNAs During Herpesvirus Infection. Curr Top Microbiol Immunol. 2018;419: 243–280. PMID: 28674945

44. Sedeño-Monge V, Vallejo-Ruiz V, Sosa-Jurado F, Santos-López G. Polymorphisms in the hepatitis C virus core and its association with development of hepatocellular carcinoma. J Biosci 2017 423. 2017;42: 509–521.

45. Lin CL, Kao JH. Hepatitis B virus genotypes and variants. Cold Spring Harb Perspect Med. 2015;5: a021436. PMID: 25934462

46. Haley CT, Mui UN, Vangipuram R, Rady PL, Tyring SK. Human Oncoviruses: Mucocutaneous Manifestations, Pathogenesis, Therapeutics, and Prevention (Part I:

Papillomaviruses and Merkel cell polyomavirus). J Am Acad Dermatol. 2018.

47. Dheekollu J, Malecka K, Wiedmer A, Delecluse H-J, Chiang AKS, Altieri DC, et al. Carcinoma-risk variant of EBNA1 deregulates Epstein-Barr Virus episomal latency. Oncotarget. 2017;8: 7248–7264. PMID: 28077791

48. Church TM, Verma D, Thompson J, Swaminathan S. Efficient Translation of Epstein-Barr Virus (EBV) DNA Polymerase Contributes to the Enhanced Lytic Replication Phenotype of M81 EBV. J Virol. 2018;92. PMID: 29263273

49. Tsai M-H, Lin X, Shumilov A, Bernhardt K, Feederle R, Poirey R, et al. The biological properties of different Epstein-Barr virus strains explain their association with various types of cancers. Oncotarget. 2017;8: 10238–10254. PMID: 28052012

50. Tzellos S, Correia PB, Karstegl CE, Cancian L, Cano-Flanagan J, McClellan MJ, et al. A Single Amino Acid in EBNA-2 Determines Superior B Lymphoblastoid Cell Line Growth Maintenance by Epstein-Barr Virus Type 1 EBNA-2. J Virol. 2014;88: 8743–8753. PMID: 24850736

51. Weiss ER, Lamers SL, Henderson JL, Melnikov A, Somasundaran M, Garber M, et al. Early Epstein-Barr Virus Genomic Diversity and Convergence toward the B95.8 Genome in Primary Infection. J Virol. 2017;92: JVI.01466-17. PMID: 29093087

52. Frange P, Boutolleau D, Leruez-Ville M, Touzot F, Cros G, Heritier S, et al. Temporal and spatial compartmentalization of drug-resistant cytomegalovirus (CMV) in a child with CMV meningoencephalitis: implications for sampling in molecular diagnosis. J Clin Microbiol. 2013;51: 4266–9. PMID: 24108608

53. Lurain NS, Chou S. Antiviral drug resistance of human cytomegalovirus. Clin Microbiol Rev. 2010;23: 689–712. PMID: 20930070

54. Chou S, Ercolani RJ, Vanarsdall AL. Differentiated Levels of Ganciclovir Resistance Conferred by Mutations at Codons 591 to 603 of the Cytomegalovirus UL97 Kinase Gene. Loeffelholz MJ, editor. J Clin Microbiol. 2017;55: 2098–2104. PMID: 28446569

55. Fischer L, Imrich E, Sampaio KL, Hofmann J, Jahn G, Hamprecht K, et al. Identification of resistance-associated HCMV UL97- and UL54-mutations and a UL97-polymporphism with impact on phenotypic drug-resistance. Antiviral Res. 2016;131: 1–8.

56. Morissette G, Flamand L. Herpesviruses and chromosomal integration. J Virol. 2010;84: 12100–9. PMID: 20844040

57. Xiao K, Yu Z, Li X, Li X, Tang K, Tu C, et al. Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in C666-1 and Raji Cells. J Cancer. 2016;7: 214–24. PMID: 26819646

58. Xu M, Zhang W-L, Zhu Q, Zhang S, Yao Y-Y, Xiang T, et al. Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. Theranostics. 2019;9: 1115–1124. PMID: 30867819

59. Hayward GS, Zong JC. Modern evolutionary history of the human KSHV genome. Curr Top Microbiol Immunol. 2007;312: 1–42. PMID: 17089792

60. Wei F, Zhu Q, Ding L, Liang Q, Cai Q. Manipulation of the host cell membrane by human

γ-herpesviruses EBV and KSHV for pathogenesis. Virol Sin. 2016;31: 395–405. PMID: 27624182

61. M L, P B, R M, MG F, G T. Human herpesvirus 8 strain variability in clinical conditions other than Kaposi's sarcoma. J Virol. 1997;71: 8082–8083. PMID: 9311909

62. Boralevi F, Masquelier B, Denayrolles M, Dupon M, Pellegrin JL, Ragnaud JM, et al. Study of human herpesvirus 8 (HHV-8) variants from Kaposi's sarcoma in France: is HHV-8 subtype A responsible for more agressive tumors? J Infect Dis. 1998;178: 1546–7. PMID: 9780286

63. Mancuso R, Biffi R, Valli M, Bellinvia M, Athanasia T, Ferrucci S, et al. HHV8 a subtype is associated with rapidly evolving classic Kaposi's sarcoma. J Med Virol. 2008;80: 2153–2160. PMID: 19040293

64. Isaacs T, Abera AB, Muloiwa R, Katz AA, Todd G. Genetic diversity of HHV8 subtypes in South Africa: A5 subtype is associated with extensive disease in AIDS-KS. J Med Virol. 2016;88: 292–303. PMID: 26174882

65. Tozetto-Mendoza TR, Ibrahim KY, Tateno AF, Oliveira CM de, Sumita LM, Sanchez MCA, et al. Genotypic distribution of HHV-8 in AIDS individuals without and with Kaposi sarcoma: Is genotype B associated with better prognosis of AIDS-KS? Medicine (Baltimore). 2016;95: e5291. PMID: 27902590

66. Cook PM, Whitby D, Calabro ML, Luppi M, Kakoola DN, Hjalgrim H, et al. Variability and evolution of Kaposi's sarcoma-associated herpesvirus in Europe and Africa. AIDS. 1999;13: 1165–1176. PMID: 10416519

67. Lacoste V, Judde JG, Brière J, Tulliez M, Garin B, Kassa-Kelembho E, et al. Molecular epidemiology of human herpesvirus 8 in Africa: Both B and A5 K1 genotypes, as well as the M and P genotypes of K14.1/K15 loci, are frequent and widespread. Virology. 2000;278: 60–74. PMID: 11112482

68. Kadyrova E, Lacoste V, Duprez R, Pozharissky K, Molochkov V, Huerre M, et al. Molecular epidemiology of Kaposi's sarcoma-associated herpesvirus/human herpesvirus 8 strains from Russian patients with classic, posttransplant, and AIDS-associated Kaposi's sarcoma. J Med Virol. 2003;71: 548–556. PMID: 14556268

69. Cordiali-Fei P, Trento E, Giovanetti M, Lo Presti A, Latini A, Giuliani M, et al. Analysis of the ORFK1 hypervariable regions reveal distinct HHV-8 clustering in Kaposi's sarcoma and non-Kaposi's cases. J Exp Clin Cancer Res. 2015;34: 1. PMID: 25592960

70. Ray A, Marshall V, Uldrick T, Leighty R, Labo N, Wyvill K, et al. Sequence analysis of Kaposi sarcoma-associated herpesvirus (KSHV) microRNAs in patients with multicentric Castleman disease and KSHV-associated inflammatory cytokine syndrome. J Infect Dis. 2012;205: 1665–76. PMID: 22448005

71. Han S-J, Marshall V, Barsov E, Quiñones O, Ray A, Labo N, et al. Kaposi's sarcoma-associated herpesvirus microRNA single-nucleotide polymorphisms identified in clinical samples can affect microRNA processing, level of expression, and silencing activity. J Virol. 2013;87: 12237–48. PMID: 24006441

72. Marshall VA, Labo N, Sztuba-Solinska J, Castro EMC, Aleman K, Wyvill KM, et al.

Polymorphisms in KSHV-encoded microRNA sequences affect levels of mature viral microRNA in Kaposi Sarcoma lesions. Oncotarget. 2018;9: 35856–35869. PMID: 30533200

73. Strahan R, Uppal T, Verma SC. Next-Generation Sequencing in the Understanding of Kaposi's Sarcoma-Associated Herpesvirus (KSHV) Biology. Viruses. 2016;8: 92. PMID: 27043613

74. Osawa M, Mine S, Ota S, Kato K, Sekizuka T, Kuroda M, et al. Establishing and characterizing a new primary effusion lymphoma cell line harboring Kaposi's sarcoma-associated herpesvirus. Infect Agent Cancer. 2016;11: 37.

75. Awazawa R, Utsumi D, Katano H, Awazawa T, Miyagi T, Hayashi K, et al. High Prevalence of Distinct Human Herpesvirus 8 Contributes to the High Incidence of Non-acquired Immune Deficiency Syndrome-Associated Kaposi's Sarcoma in Isolated Japanese Islands. J Infect Dis. 2017;216: 850–858. PMID: 28968717

76. Olp LN, Jeanniard A, Marimo C, West JT, Wood C. Whole-Genome Sequencing of Kaposi's Sarcoma-Associated Herpesvirus from Zambian Kaposi's Sarcoma Biopsy Specimens Reveals Unique Viral Diversity. J Virol. 2015;89: 12299–12308. PMID: 26423952

77. Sallah N, Palser AL, Watson SJ, Labo N, Asiki G, Marshall V, et al. Genome-Wide Sequence Analysis of Kaposi Sarcoma-Associated Herpesvirus Shows Diversification Driven by Recombination. J Infect Dis. 2018;218: 1700–1710. PMID: 30010810

78. Yakushko Y, Hackmann C, Günther T, Rückert J, Henke M, Koste L, et al. Kaposi's sarcoma-associated herpesvirus bacterial artificial chromosome contains a duplication of a long unique-region fragment within the terminal repeat region. J Virol. 2011;85: 4612–7. PMID: 21307197

79. Zhou F-C, Zhang Y-J, Deng J-H, Wang X-P, Pan H-Y, Hettler E, et al. Efficient infection by a recombinant Kaposi's sarcoma-associated herpesvirus cloned in a bacterial artificial chromosome: application for genetic analysis. J Virol. 2002;76: 6185–96. PMID: 12021352

80. Mondello C, Smirnova A, Giulotto E. Gene amplification, radiation sensitivity and DNA double-strand breaks. Mutation Research - Reviews in Mutation Research. Elsevier; 2010. pp. 29–37. PMID: 20093194

81. Pyakurel P, Pak F, Mwakigonja AR, Kaaya E, Biberfeld P. KSHV/HHV-8 and HIV infection in Kaposi's sarcoma development. Infect Agent Cancer. 2007;2: 4.

82. Gaidano G, Capello D, Fassone L, Gloghini A, Cilia AM, Ariatti C, et al. Molecular characterization of HHV-8 positive primary effusion lymphoma reveals pathogenetic and histogenetic features of the disease. J Clin Virol. 2000;16: 215–224. PMID: 10738140

83. Mullaney BP, Ng VL, Herndier BG, McGrath MS, Pallavicini MG. Comparative genomic analyses of primary effusion lymphoma. Arch Pathol Lab Med. 2000;124: 824–826. PMID: 10835513

84. Nair P, Pan H, Stallings RL, Gao SJ. Recurrent genomic imbalances in primary effusion lymphomas. Cancer Genet Cytogenet. 2006;171: 119–121. PMID: 17116491

85. Pan H, Zhou F, Gao SJ. Kaposi's sarcoma-associated herpesvirus induction of chromosome instability in primary human endothelial cells. Cancer Res. 2004;64: 4064–4068. PMID: 15205312

86. Fang C-Y, Lee C-H, Wu C-C, Chang Y-T, Yu S-L, Chou S-P, et al. Recurrent chemical reactivations of EBV promotes genome instability and enhances tumor progression of nasopharyngeal carcinoma cells. Int J Cancer. 2009;124: 2016–2025.

87. Ohsaki E, Ueda K. Interplay Between KSHV and the Host DNA Damage Response. Frontiers in Cellular and Infection Microbiology. Frontiers Media S.A.; 2020. PMID: 33425783

88. Xiao Y, Chen J, Liao Q, Wu Y, Peng C, Chen X. Lytic infection of Kaposi's sarcoma-associated herpesvirus induces DNA double-strand breaks and impairs non-homologous end joining. J Gen Virol. 2013;94: 1870–1875. PMID: 23677788

89. Jackson BR, Noerenberg M, Whitehouse A. A Novel Mechanism Inducing Genome Instability in Kaposi's Sarcoma-Associated Herpesvirus Infected Cells. PLoS Pathog. 2014;10. PMID: 24788796

90. Krause CJ, Popp O, Thirunarayanan N, Dittmar G, Lipp M, Müller G. MicroRNA-34a promotes genomic instability by a broad suppression of genome maintenance mechanisms downstream of the oncogene KSHV-vGPCR. Oncotarget. 2016;7: 10414–10432.

91. Renner DW, Szpara ML. Impacts of Genome-Wide Analyses on Our Understanding of Human Herpesvirus Diversity and Evolution. J Virol. 2018;92: 908–925.

92. Berenstein AJ, Lorenzetti MA, Preciado MV. Recombination rates along the entire Epstein Barr virus genome display a highly heterogeneous landscape. Infect Genet Evol. 2018;65: 96–103.

93. Cudini J, Roy S, Houldcroft CJ, Bryant JM, Depledge DP, Tutill H, et al. Human cytomegalovirus haplotype reconstruction reveals high diversity due to superinfection and evidence of within-host recombination. Proc Natl Acad Sci U S A. 2019;116: 5693–5698. PMID: 30819890

94. Brown JC. The role of DNA repair in herpesvirus pathogenesis. Genomics. 2014;104: 287–294.

95. Rennekamp AJ, Lieberman PM. Initiation of lytic DNA replication in Epstein-Barr virus: Search for a common family mechanism. Future Virology. NIH Public Access; 2010. pp. 65–83.

96. Lo Piano A, Martínez-Jiménez MI, Zecchi L, Ayora S. Recombination-dependent concatemeric viral DNA replication. Virus Research. Elsevier; 2011. pp. 1–14. PMID: 21708194

97. Pfu¨ller R, Hammerschmidt W. Plasmid-Like Replicative Intermediates of the Epstein-Barr Virus Lytic Origin of DNA Replication. J Virol. 1996.

98. Okamoto H, Watanabe TA, Horiuchi T. Double rolling circle replication (DRCR) is recombinogenic. Genes to Cells. 2011;16: 503–513. PMID: 21501343

99. Watanabe T, Tanabe H, Horiuchi T. Gene amplification system based on double rolling-circle replication as a model for oncogene-type amplification. Nucleic Acids Res. 2011;39: e106. PMID: 21653557

100. Paulsen T, Kumar P, Koseoglu MM, Dutta A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet. 2018;34: 270–278. PMID: 29329720

101. Verhaak RGW, Bafna V, Mischel PS. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nature Reviews Cancer. Nature Publishing Group; 2019. pp. 283–288. PMID: 30872802

102. Deng J-H, Zhang Y-J, Wang X-P, Gao S-J. Lytic replication-defective Kaposi's sarcoma-associated herpesvirus: potential role in infection and malignant transformation. J Virol. 2004;78: 11108–20. PMID: 15452231

103. Campbell DE, Kemp MC, Perdue ML, Randall CC, Gentry GA. Equine herpesvirus in vivo: Cyclic production of a DNA density variant with repetitive sequences. Virology. 1976;69: 737–750.

104. Patton DF, Shirley P, Raab-Traub N, Resnick L, Sixbey JW. Defective viral DNA in Epstein-Barr virus-associated oral hairy leukoplakia. J Virol. 1990;64: 397–400. PMID: 2152824

105. Gan YJ, Razzouk BI, Su T, Sixbey JW. A defective, rearranged Epstein-Barr virus genome in EBER-negative and EBER-positive Hodgkin's disease. Am J Pathol. 2002;160: 781–786. PMID: 11891176

106. Mocarski ES, Deiss LP, Frenkel N. Nucleotide sequence and structural features of a novel US-a junction present in a defective herpes simplex virus genome. J Virol. 1985;55: 140–146. PMID: 2989551

107. Charvat RA, Zhang Y, O'Callaghan DJ. Deletion of the UL4 gene sequence of equine herpesvirus 1 precludes the generation of defective interfering particles. Virus Genes. 2012;45: 295–303. PMID: 22752566

108. Manzoni TB, López CB. Defective (interfering) viral genomes re-explored: Impact on antiviral immunity and virus persistence. Future Virology. Future Medicine Ltd.; 2018. pp. 493–503.

109. Yang Y, Lyu T, Zhou R, He X, Ye K, Xie Q, et al. The antiviral and antitumor effects of defective interfering particles/genomes and their mechanisms. Front Microbiol. 2019;10.

110. Domingo E, Perales C. Quasispecies and virus. Eur Biophys J 2018 474. 2018;47: 443–457.

111. Felt SA, Sun Y, Jozwik A, Paras A, Habibi MS, Nickle D, et al. Detection of respiratory syncytial virus defective genomes in nasal secretions is associated with distinct clinical outcomes. Nat Microbiol. 2021;6: 672–681.

112. Rabson M, Heston L, Miller G. Identification of a rare Epstein-Barr virus variant that enhances early antigen expression in Raji cells. Proc Natl Acad Sci U S A. 1983;80: 2762–2766. PMID: 6302703

113. Miller G, Rabson M, Heston L. Epstein-Barr virus with heterogeneous DNA disrupts

latency. J Virol. 1984;50: 174–182. PMID: 6321789

114. Rooney C, Taylor N, Countryman J, Jenson H, Kolman J, Miller G. Genome rearrangements activate the Epstein-Barr virus gene whose product disrupts latency. Proc Natl Acad Sci U S A. 1988;85: 9801–9805. PMID: 2849118

115. Gemayel R, Cho J, Boeynaems S, Verstrepen KJ. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. Genes. Multidisciplinary Digital Publishing Institute (MDPI); 2012. pp. 461–480.

116. Gao S, Zhang Y, Deng J, Rabkin CS, Flore O, Jenson HB. Molecular Polymorphism of Kaposi's Sarcoma--Associated Herpesvirus (Human Herpesvirus 8) Latent Nuclear Antigen: Evidence for a Large Repertoire of Viral Genotypes and Dual Infection with Different Viral Genotypes. J Infect Dis. 1999;180: 1466–1476. PMID: 10515805

117. Sadler R, Wu L, Forghani B, Renne R, Zhong W, Herndier B, et al. A complex translational program generates multiple novel proteins from the latently expressed kaposin (K12) locus of Kaposi's sarcoma-associated herpesvirus. J Virol. 1999;73: 5722–30. PMID: 10364323

118. Li H, Komatsu T, Dezube BJ, Kaye KM. The Kaposi's sarcoma-associated herpesvirus K12 transcript from a primary effusion lymphoma contains complex repeat elements, is spliced, and initiates from a novel promoter. J Virol. 2002;76: 11880–11888. PMID: 12414930

119. Rose TM, Bruce AG, Barcy S, Fitzgibbon M, Matsumoto LR, Ikoma M, et al. Quantitative RNAseq analysis of Ugandan KS tumors reveals KSHV gene expression dominated by transcription from the LTd downstream latency promoter. Schulz TF, editor. PLoS Pathog. 2018;14: e1007441. PMID: 30557332

120. McCormick C, Ganem D. The kaposin B protein of KSHV activates the p38/MK2 pathway and stabilizes cytokine mRNAs. Science. 2005;307: 739–41. PMID: 15692053

121. Ballestas ME, Kaye KM. The latency-associated nuclear antigen, a multifunctional protein central to Kaposi's sarcoma-associated herpesvirus latency. Future Microbiol. 2011;6: 1399–413. PMID: 22122438

122. Ba Abdullah MM, Palermo RD, Palser AL, Grayson NE, Kellam P, Correia S, et al. Heterogeneity of the Epstein-Barr Virus (EBV) Major Internal Repeat Reveals Evolutionary Mechanisms of EBV and a Functional Defect in the Prototype EBV Strain B95-8. J Virol. 2017;91. PMID: 28904201

123. Szymula A, Palermo RD, Bayoumy A, Groves IJ, Ba Abdullah M, Holder B, et al. Epstein-Barr virus nuclear antigen EBNA-LP is essential for transforming naïve B cells, and facilitates recruitment of transcription factors to the viral genome. PLoS Pathog. 2018;14: e1006890. PMID: 29462212

124. Portes-Sentis S, Sergeant A, Gruffat H. A particular DNA structure is required for the function of a cis-acting component of the Epstein-Barr virus OriLyt origin of replication. Nucleic Acids Res. 1997;25: 1347–1354. PMID: 9060428

125. AuCoin DP, Colletti KS, Xu Y, Cei SA, Pari GS. Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) contains two functional lytic origins of DNA replication. J Virol.

2002;76: 7890–6. PMID: 12097603

126. Chun-Fan C, Hiroomi T, Khalili K. The role of a pentanucleotide repeat sequence, AGGGAAGGGA, in the regulation of JC virus DNA replication. Gene. 1994;148: 309–314. PMID: 7958960

127. Liu M, Kumar KU, Pater MM, Pater A. Dual NF1-requiring effect of human neurotropic JC virus composite pentanucleotide repeat elements on early and late viral gene expression. Virology. 1997;227: 7–12. PMID: 9007053

128. Gluzman Y. SV40-transformed simian cells support the replication of early SV40 mutants. Cell. 1981;23: 175–182.

129. Watanabe S, Yoshiike K. Decreasing the number of 68-base-pair tandem repeats in the BK virus transcriptional control region reduces plaque size and enhances transforming capacity. J Virol. 1985;55: 823–5. PMID: 2991597

130. Khristich AN, Mirkin SM. On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology Inc.; 2020. pp. 4134–4170. PMID: 32060097

131. Persi E, Prandi D, Wolf YI, Pozniak Y, Barnabas GD, Levanon K, et al. Proteomic and genomic signatures of repeat instability in cancer and adjacent normal tissues. Proc Natl Acad Sci U S A. 2019;116: 16987–16996. PMID: 31387980

132. Zong JC, Arav-Boger R, Alcendor DJ, Hayward GS. Reflections on the interpretation of heterogeneity and strain differences based on very limited PCR sequence data from Kaposi's sarcoma-associated herpesvirus genomes. Journal of Clinical Virology. NIH Public Access; 2007. pp. 1–8. PMID: 17698410

133. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. 2013;41: e67. PMID: 23303777

134. Chen G, Mosier S, Gocke CD, Lin M-T, Eshleman JR. Cytosine Deamination Is a Major Cause of Baseline Noise in Next-Generation Sequencing. Mol Diagn Ther. 2014;18: 587–593. PMID: 25091469

135. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019;20: 50.

136. Park G, Park JK, Shin S-H, Jeon H-J, Kim NKD, Kim YJ, et al. Characterization of background noise in capture-based targeted sequencing data. Genome Biol. 2017;18: 136. PMID: 28732520

137. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12: R18. PMID: 21338519

138. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc. 2014;9: 2586–2606. PMID: 25299156

139. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nat Methods. 2015;12: 423–425. PMID: 25849638

140. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nature Reviews Genetics. Nature Publishing Group; 2018. pp. 269–285.

141. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016;73: 4433–4448. PMID: 27392606

142. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012;9: 72–74. PMID: 22101854

143. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012;109: 14508–13. PMID: 22853953

144. Höijer I, Tsai Y-C, Clark TA, Kotturi P, Dahl N, Stattin E-L, et al. Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. Hum Mutat. 2018;39: 1262–1272. PMID: 29932473

145. Krown SE, Metroka C, Wernz JC. Kaposi's sarcoma in the acquired immune deficiency syndrome: a proposal for uniform evaluation, response, and staging criteria. AIDS Clinical Trials Group Oncology Committee. J Clin Oncol. 1989;7: 1201–1207. PMID: 2671281

146. Johnston C, Orem J, Okuku F, Kalinaki M, Saracino M, Katongole-Mbidde E, et al. Impact of HIV infection and Kaposi sarcoma on human herpesvirus-8 mucosal replication and dissemination in Uganda. PLoS One. 2009;4: e4222. PMID: 19156206

147. Ryncarz AJ, Goddard J, Wald A, Huang ML, Roizman B, Corey L. Development of a high-throughput quantitative assay for detecting herpes simplex virus DNA in clinical samples. J Clin Microbiol. 1999;37: 1941–7. PMID: 10325351

148. Zaniello B, Huang M-L, Cheng A, Selke S, Wald A, Jerome KR, et al. Consistent viral DNA quantification after prolonged storage at ambient temperature. J Virol Methods. 2016;228: 91–94.

149. Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, et al. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLoS One. 2011;6. PMID: 22125625

150. Thompson JR, Marcelino LA, Polz MF. Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by "reconditioning PCR". Nucleic Acids Res. 2002;30: 2083–8. PMID: 11972349

151. Joshi N, Fass J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle. 2011; 2011.

152. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

153. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a

new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–77. PMID: 22506599

154. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13: R56. PMID: 22731987

155. Wang Y, Li H, Chan MY, Zhu FX, Lukac DM, Yuan Y. Kaposi's Sarcoma-Associated Herpesvirus ori-Lyt-Dependent DNA Replication: cis-Acting Requirements for Replication and ori-Lyt-Associated RNA Transcription. J Virol. 2004;78: 8615–8629. PMID: 15280471

156. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–80. PMID: 23329690

157. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–3. PMID: 24451623

158. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution. Oxford University Press; 2006. pp. 254–267. PMID: 16221896

159. Katz JP, Pipas JM. SummonChimera infers integrated viral genomes with nucleotide precision from NGS data. BMC Bioinformatics. 2014;15: 348. PMID: 25331652

160. Aboyoun P, DebRoy S, Gentleman R, Pagès H. Biostrings: Efficient manipulation of biological strings. 2021.

161. Zhabinskaya D, Madden S, Benham CJ. SIST: stress-induced structural transitions in superhelical DNA. Bioinformatics. 2015;31: 421–422.

162. Hon J, Martínek T, Rajdl K, Lexa M. Triplex: An R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences. Bioinformatics. 2013;29: 1900–1901. PMID: 23709494

163. Hon J, Martínek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics. 2017;33: 3373–3379. PMID: 29077807

164. pqsfinder – imperfection-tolerant G-quadruplex prediction. [cited 3 Jul 2021].

165. Santiago JC, Goldman JD, Zhao H, Pankow AP, Okuku F, Schmitt MW, et al. Intra-host changes in Kaposi sarcoma-associated herpesvirus genomes in Ugandan adults with Kaposi sarcoma. PLoS Pathog. 2021;17. PMID: 33465147

166. Technology. [cited 18 May 2021].

167. Beyari MM, Hodgson TA, Cook RD, Kondowe W, Molyneux EM, Scully CM, et al. Multiple Human Herpesvirus–8 Infection. J Infect Dis. 2003;188: 678–689. PMID: 12934184

168. Beyari MM, Hodgson TA, Kondowe W, Molyneux EM, Scully CM, Porter SR, et al. Genotypic profile of human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus) in urine. J Clin Microbiol. 2004;42: 3313–6. PMID: 15243103

169. Al-Otaibi LM, Ngui SL, Scully CM, Porter SR, Teo CG. Salivary human herpesvirus 8

shedding in renal allograft recipients with Kaposi's sarcoma. J Med Virol. 2007;79: 1357–1365. PMID: 17607792

170. Al-Otaibi LM, Al-Sulaiman MH, Teo CG, Porter SR. Extensive oral shedding of human herpesvirus 8 in a renal allograft recipient. Oral Microbiol Immunol. 2009;24: 109–115. PMID: 19239637

171. Al-Otaibi LM, Moles DR, Porter SR, Teo CG. Human herpesvirus 8 shedding in the mouth and blood of hemodialysis patients. J Med Virol. 2012;84: 792–797. PMID: 22431028

172. Leao JC, de Faria ABS, Fonseca DDD, Gueiros LAM, Silva IHM, Porter SR. Intrahost genetic variability of human herpes virus-8. J Med Virol. 2013;85: 636–645. PMID: 26928661

173. Mbulaiteye S, Marshall V, Bagni RK, Wang C, Mbisa G, Bakaki PM, et al. Molecular Evidence for Mother-to-Child Transmission of Kaposi Sarcoma–Associated Herpesvirus in Uganda and K1 Gene Evolution within the Host. J Infect Dis. 2006;193: 1250–1257. PMID: 16586362

174. Caro-Vegas C, Sellers S, Host KM, Seltzer J, Landis J, Fischer WA, et al. Runaway Kaposi Sarcoma-associated herpesvirus replication correlates with systemic IL-10 levels. Virology. 2020;539: 18–25.

175. Zong J, Ciufo DM, Viscidi R, Alagiozoglou L, Tyring S, Rady P, et al. Genotypic analysis at multiple loci across Kaposi's sarcoma herpesvirus (KSHV) DNA molecules: Clustering patterns, novel variants and chimerism. J Clin Virol. 2002;23: 119–148. PMID: 11595592

176. Stebbing J, Wilder N, Ariad S, Abu-Shakra M. Lack of intra-patient strain variability during infection with Kaposi's sarcoma-associated herpesvirus. Am J Hematol. 2001;68: 133–4. PMID: 11559954

177. Meng YX, Spira TJ, Bhat GJ, Birch CJ, Druce JD, Edlin BR, et al. Individuals from North America, Australasia, and Africa are infected with four different genotypes of human herpesvirus 8. Virology. 1999;261: 106–119. PMID: 10441559

178. Poole LJ, Zong JC, Ciufo DM, Alcendor DJ, Cannon JS, Ambinder R, et al. Comparison of genetic variability at multiple loci across the genomes of the major subtypes of Kaposi's sarcoma-associated herpesvirus reveals evidence for recombination and for two distinct types of open reading frame K15 alleles at the right-hand end. J Virol. 1999;73: 6646–60. PMID: 10400762

179. Kakoola DN, Sheldon J, Byabazaire N, Bowden RJ, Katongole-Mbidde E, Schulz TF, et al. Recombination in human herpesvirus-8 strains from Uganda and evolution of the K15 gene. J Gen Virol. 2001;82: 2393–2404. PMID: 11562533

180. Renne R, Zhong W, Herndier B, McGrath M, Abbey N, Kedes D, et al. Lytic growth of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) in culture. Nat Med. 1996;2: 342–346.

181. Eigen M. On the nature of virus quasispecies. Trends Microbiol. 1996;4: 216–218. PMID: 8795155

182. Peng Q, Xu C, Kim D, Lewis M, DiCarlo J, Wang Y. Targeted Single Primer Enrichment

Sequencing with Single End Duplex-UMI. Sci Rep. 2019;9: 4810. PMID: 30886209

183. Tang S, Yamanegi K, Zheng Z-M. Requirement of a 12-base-pair TATT-containing sequence and viral lytic DNA replication in activation of the Kaposi's sarcoma-associated herpesvirus K8.1 late promoter. J Virol. 2004;78: 2609–14. PMID: 14963167

184. Tang S, Zheng Z-M. Kaposi's sarcoma-associated herpesvirus K8 exon 3 contains three 5'-splice sites and harbors a K8.1 transcription start site. J Biol Chem. 2002;277: 14547–56. PMID: 11832484

185. Cantalupo PG, Katz JP, Pipas JM. Viral sequences in human cancer. Virology. 2018;513: 208–216. PMID: 29107929

186. Aneja KK, Yuan Y. Reactivation and Lytic Replication of Kaposi's Sarcoma-Associated Herpesvirus: An Update. Front Microbiol. 2017;8: 613. PMID: 28473805

187. Bacolla A, Tainer JA, Vasquez KM, Cooper DN. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. Nucleic Acids Res. 2016;44: 5673–88. PMID: 27084947

188. Luna RE, Zhou F, Baghian A, Chouljenko V, Forghani B, Gao S-J, et al. Kaposi's Sarcoma-Associated Herpesvirus Glycoprotein K8.1 Is Dispensable for Virus Entry. J Virol. 2004;78: 6389–6398. PMID: 15163732

189. Wang G, Vasquez KM. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. DNA Repair (Amst). 2014;19: 143–151. PMID: 24767258

190. Lin CL, Li H, Wang Y, Zhu FX, Kudchodkar S, Yuan Y. Kaposi's sarcoma-associated herpesvirus lytic origin (ori-Lyt)-dependent DNA replication: identification of the ori-Lyt and association of K8 bZip protein with the origin. J Virol. 2003;77: 5578–88. PMID: 12719550

191. Wang Y, Tang Q, Maul GG, Yuan Y. Kaposi's sarcoma-associated herpesvirus ori-Lyt-dependent DNA replication: dual role of replication and transcription activator. J Virol. 2006;80: 12171–86. PMID: 17020951

192. Liu D, Wang Y, Yuan Y. Kaposi's Sarcoma-Associated Herpesvirus K8 Is an RNA Binding Protein That Regulates Viral DNA Replication in Coordination with a Noncoding RNA. J Virol. 2018;92. PMID: 29321307

193. Rangel N, Forero-Castro M, Rondón-Lagos M. New insights in the cytogenetic practice: Karyotypic chaos, non-clonal chromosomal alterations and chromosomal instability in human cancer and therapy response. Genes. MDPI AG; 2017. pp. 2–29.

194. Dittmer DP, Damania B, Silverberg M, Robbins H, Pfeiffer R, Shiels M, et al. Kaposi sarcoma–associated herpesvirus: immunobiology, oncogenesis, and therapy. J Clin Invest. 2016;126: 3165–3175.

195. Boname JM, Lehner PJ. What has the study of the K3 and K5 viral ubiquitin E3 ligases taught us about ubiquitin-mediated receptor regulation? Viruses. 2011;3: 118–131. PMID: 22049306

196. Krishnan HH, Naranatt PP, Smith MS, Zeng L, Bloomer C, Chandran B. Concurrent Expression of Latent and a Limited Number of Lytic Genes with Immune Modulation and

Antiapoptotic Function by Kaposi's Sarcoma-Associated Herpesvirus Early during Infection of Primary Endothelial and Fibroblast Cells and Subsequent Decline of Lytic Gene Expression. J Virol. 2004;78: 3601–3620. PMID: 15016882

197. Brulois K, Toth Z, Wong L-Y, Feng P, Gao S-J, Ensser A, et al. Kaposi's Sarcoma-Associated Herpesvirus K3 and K5 Ubiquitin E3 Ligases Have Stage-Specific Immune Evasion Roles during Lytic Replication. J Virol. 2014;88: 9335–9349. PMID: 24899205

198. Okuno T, Jiang YB, Ueda K, Nishimura K, Tamura T, Yamanishi K. Activation of human herpesvirus 8 open reading frame K5 independent of ORF50 expression. Virus Res. 2002;90: 77–89. PMID: 12457964

199. Timms RT, Duncan LM, Tchasovnikarova IA, Antrobus R, Smith DL, Dougan G, et al. Haploid Genetic Screens Identify an Essential Role for PLP2 in the Downregulation of Novel Plasma Membrane Targets by Viral E3 Ubiquitin Ligases. PLoS Pathog. 2013;9. PMID: 24278019

200. Gabaev I, Williamson JC, Crozier TWM, Schulz TF, Lehner PJ. Quantitative Proteomics Analysis of Lytic KSHV Infection in Human Endothelial Cells Reveals Targets of Viral Immune Modulation. Cell Rep. 2020;33. PMID: 33053346

201. Hahn AS, Kaufmann JK, Wies E, Naschberger E, Panteleev-Ivlev J, Schmidt K, et al. The ephrin receptor tyrosine kinase A2 is a cellular receptor for Kaposi's sarcoma-associated herpesvirus. Nat Med. 2012;18: 961–966. PMID: 22635007

202. Dairaghi DJ, Fan RA, McMaster BE, Hanley MR, Schall TJ. HHV8-encoded vMIP-I selectively engages chemokine receptor CCR8. Agonist and antagonist profiles of viral chemokines. J Biol Chem. 1999;274: 21569–21574. PMID: 10419462

203. Endres MJ, Garlisi CG, Xiao H, Shan LX, Hedrick JA. The Kaposi's sarcoma-related herpesvirus (KSHV)-encoded chemokine VMIP- I is a specific agonist for the CC chemokine receptor (CCR)8. J Exp Med. 1999;189: 1993–1998. PMID: 10377196

204. Boshoff C, Endo Y, Collins PD, Takeuchi Y, Reeves JD, Schweickart VL, et al. Angiogenic and HIV-inhibitory functions of KSHV-encoded chemokines. Science (80-). 1997;278: 290–294. PMID: 9323208

205. Nakano K, Katano H, Tadagaki K, Sato Y, Ohsaki E, Mori Y, et al. Novel monoclonal antibodies for identification of multicentric Castleman's disease; Kaposi's sarcoma-associated herpesvirus-encoded vMIP-I and vMIP-II. Virology. 2012;425: 95–102. PMID: 22297135

206. Lagunoff M, Ganem D. The Structure and Coding Organization of the Genomic Termini of Kaposi's Sarcoma-Associated Herpesvirus (Human Herpesvirus 8). Virology. 1997;236: 147–154.

207. Campbell M, Kung H-J, Izumiya Y. Long non-coding RNA and epigenetic gene regulation of KSHV. Viruses. 2014;6: 4165–77. PMID: 25375882

208. Conrad NK. New insights into the expression and functions of the Kaposi's sarcoma-associated herpesvirus long noncoding PAN RNA. Virus Res. 2016;212: 53–63. PMID: 26103097

209. Chavez-Calvillo G, Martin S, Hamm C, Sztuba-Solinska J. The Structure-To-Function Relationships of Gammaherpesvirus-Encoded Long Non-Coding RNAs and Their Contributions to Viral Pathogenesis. Non-coding RNA. 2018;4. PMID: 30261651

210. Rossetto CC, Tarrant-Elorza M, Verma S, Purushothaman P, Pari GS. Regulation of Viral and Cellular Gene Expression by Kaposi's Sarcoma-Associated Herpesvirus Polyadenylated Nuclear RNA. J Virol. 2013;87: 5540–5553.

211. Abere B, Li J, Zhou H, Toptan T, Moore PS, Chang Y. Kaposi's Sarcoma-Associated Herpesvirus-Encoded circRNAs Are Expressed in Infected Tumor Tissues and Are Incorporated into Virions. MBio. 2020;11.

212. Taylor JL, Bennett HN, Snyder BA, Moore PS, Chang Y. Transcriptional analysis of latent and inducible Kaposi's sarcoma-associated herpesvirus transcripts in the K4 to K7 region. J Virol. 2005;79: 15099–106. PMID: 16306581

213. Hu Z, Usherwood EJ. Immune escape of γ-herpesviruses from adaptive immunity. Rev Med Virol. 2014;24: 365–78. PMID: 24733560

214. Lee H-R, Lee S, Chaudhary PM, Gill P, Jung JU. Immune evasion by Kaposi's sarcoma-associated herpesvirus. Future Microbiol. 2010;5: 1349–65. PMID: 20860481

215. Wei X, Lan K. Activation and counteraction of antiviral innate immunity by KSHV: an Update. Sci Bull. 2018;63: 1223–1234. PMID: 30906617

216. Manners O, Murphy JC, Coleman A, Hughes DJ, Whitehouse A. Contribution of the KSHV and EBV lytic cycles to tumourigenesis. Curr Opin Virol. 2018;32: 60–70. PMID: 30268927

217. Grundhoff A, Ganem D. Inefficient establishment of KSHV latency suggests an additional role for continued lytic replication in Kaposi sarcoma pathogenesis. J Clin Invest. 2004;113: 124–36. PMID: 14702116

218. Wang G, Vasquez KM. Effects of replication and transcription on DNA Structure-Related genetic instability. Genes. MDPI AG; 2017.

219. Saranathan N, Biswas B, Patra A, Vivekanandan P. G-quadruplexes may determine the landscape of recombination in HSV-1. BMC Genomics. 2019;20: 382.

220. Wang FZ, Akula SM, Pramod NP, Zeng L, Chandran B. Human herpesvirus 8 envelope glycoprotein K8.1A interaction with the target cells involves heparan sulfate. J Virol. 2001;75: 7517–27. PMID: 11462024

221. Birkmann A, Mahr K, Ensser A, Yağuboğlu S, Titgemeyer F, Fleckenstein B, et al. Cell surface heparan sulfate is a receptor for human herpesvirus 8 and interacts with envelope glycoprotein K8.1. J Virol. 2001;75: 11583–93. PMID: 11689640

222. Akula SM, Pramod NP, Wang F-Z, Chandran B. Human Herpesvirus 8 Envelope-Associated Glycoprotein B Interacts with Heparan Sulfate-like Moieties. Virology. 2001;284: 235–249. PMID: 11384223

223. Akula SM, Wang F-Z, Vieira J, Chandran B. Human Herpesvirus 8 Interaction with Target Cells Involves Heparan Sulfate. Virology. 2001;282: 245–255. PMID: 11289807

224. Dollery SJ, Santiago-Crespo RJ, Chatterjee D, Berger EA. Glycoprotein K8.1A of Kaposi's Sarcoma-Associated Herpesvirus Is a Critical B Cell Tropism Determinant Independent of Its Heparan Sulfate Binding Activity. Longnecker RM, editor. J Virol. 2018;93. PMID: 30567992

225. Li M, MacKey J, Czajak SC, Desrosiers RC, Lackner AA, Jung JU. Identification and characterization of Kaposi's sarcoma-associated herpesvirus K8.1 virion glycoprotein. J Virol. 1999;73: 1341–9. PMID: 9882339

226. Raab MS, Albrecht JC, Birkmann A, Yağuboğlu S, Lang D, Fleckenstein B, et al. The immunogenic glycoprotein gp35-37 of human herpesvirus 8 is encoded by open reading frame K8.1. J Virol. 1998;72: 6725–31. PMID: 9658120

227. Chandran B, Bloomer C, Chan SR, Zhu L, Goldstein E, Horvat R. Human Herpesvirus-8 ORF K8.1 Gene Encodes Immunogenic Glycoproteins Generated by Spliced Transcripts. Virology. 1998;249: 140–149. PMID: 9740785

228. Hislop AD, Sabbah S. CD8+ T cell immunity to Epstein-Barr virus and Kaposi's sarcoma-associated herpes virus. Seminars in Cancer Biology. Academic Press; 2008. pp. 416–422.

229. Marshall V, Parks T, Bagni R, Wang CD, Samols MA, Hu J, et al. Conservation of Virally Encoded MicroRNAs in Kaposi Sarcoma–Associated Herpesvirus in Primary Effusion Lymphoma Cell Lines and in Patients with Kaposi Sarcoma or Multicentric Castleman Disease. J Infect Dis. 2007;195: 645–659. PMID: 17262705

230. Forte E, Raja AN, Shamulailatpam P, Manzano M, Schipma MJ, Casey JL, et al. MicroRNA-Mediated Transformation by the Kaposi's Sarcoma-Associated Herpesvirus Kaposin Locus. Hutt-Fletcher L, editor. J Virol. 2015;89: 2333–2341. PMID: 25505059

231. Gandy SZ, Linnstaedt SD, Muralidhar S, Cashman KA, Rosenthal LJ, Casey JL. RNA editing of the human herpesvirus 8 kaposin transcript eliminates its transforming activity and is induced during lytic replication. J Virol. 2007;81: 13544–51. PMID: 17913828

232. Wong L-Y, Brulois K, Toth Z, Inn K-S, Lee S-H, O'Brien K, et al. The Product of Kaposi's Sarcoma-Associated Herpesvirus Immediate Early Gene K4.2 Regulates Immunoglobulin Secretion and Calcium Homeostasis by Interacting with and Inhibiting pERP1. J Virol. 2013;87: 12069–12079. PMID: 23986581

233. Burysek L, Yeow WS, Pitha PM. Unique properties of a second human herpesvirus 8-encoded interferon regulatory factor (vIRF-2). J Hum Virol. 1999;2: 19–32. PMID: 10200596

234. Burýšek L, Pitha PM. Latently Expressed Human Herpesvirus 8-Encoded Interferon Regulatory Factor 2 Inhibits Double-Stranded RNA-Activated Protein Kinase. J Virol. 2001;75: 2345–2352. PMID: 11160738

235. Koch S, Damas M, Freise A, Hage E, Dhingra A, Ruckert J, et al. Kaposi's sarcoma-associated herpesvirus vIRF2 protein utilizes an IFN-dependent pathway to regulate viral early gene expression. PLoS Pathog. 2019;15. PMID: 31059555

236. Xiang Q, Ju H, Nicholas J. USP7-Dependent Regulation of TRAF Activation and Signaling by a Viral Interferon Regulatory Factor Homologue. J Virol. 2019;94. PMID:

31666375

237. Wu Y-H, Hu T-F, Chen Y-C, Tsai Y-N, Tsai Y-H, Cheng C-C, et al. The manipulation of miRNA-gene regulatory networks by KSHV induces endothelial cell motility. Blood. 2011;118: 2896–905. PMID: 21715310

238. Weidner-Glunde M, Mariggiò G, Schulz TF. Kaposi's Sarcoma-Associated Herpesvirus Latency-Associated Nuclear Antigen: Replicating and Shielding Viral DNA during Viral Persistence. J Virol. 2017;91. PMID: 28446671

239. Viejo-Borbolla A, Kati E, Sheldon JA, Nathan K, Mattsson K, Szekely L, et al. A Domain in the C-terminal region of latency-associated nuclear antigen 1 of Kaposi's sarcoma-associated Herpesvirus affects transcriptional activation and binding to nuclear heterochromatin. J Virol. 2003;77: 7093–100. PMID: 12768028

240. Zaldumbide A, Ossevoort M, Wiertz EJHJ, Hoeben RC. In cis inhibition of antigen processing by the latency-associated nuclear antigen I of Kaposi sarcoma Herpes virus. Mol Immunol. 2007;44: 1352–1360. PMID: 16828498

241. Kwun HJ, da Silva SR, Shah IM, Blake N, Moore PS, Chang Y. Kaposi's Sarcoma-Associated Herpesvirus Latency-Associated Nuclear Antigen 1 Mimics Epstein-Barr Virus EBNA1 Immune Evasion through Central Repeat Domain Effects on Protein Processing. J Virol. 2007;81: 8225–8235. PMID: 17522213

242. Kwun HJ, da Silva SR, Qin H, Ferris RL, Tan R, Chang Y, et al. The central repeat domain 1 of Kaposi's sarcoma-associated herpesvirus (KSHV) latency associated-nuclear antigen 1 (LANA1) prevents cis MHC class I peptide presentation. Virology. 2011;412: 357–365. PMID: 21324504

243. Kwun HJ, Toptan T, Ramos da Silva S, Atkins JF, Moore PS, Chang Y. Human DNA tumor viruses generate alternative reading frame proteins through repeat sequence recoding. Proc Natl Acad Sci. 2014;111: E4342–E4349. PMID: 25271323

244. Dabral P, Babu J, Zareie A, Verma SC. LANA and hnRNP A1 Regulate the Translation of LANA mRNA through G-Quadruplexes. J Virol. 2020;94. PMID: 31723020

245. Cruz F, Roux J, Robinson-Rechavi M. The expansion of amino-acid repeats is not associated to adaptive evolution in mammalian genes. BMC Genomics. 2009;10: 619. PMID: 20021652

246. Schaefer MH, Wanker EE, Andrade-Navarro MA. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res. 2012;40: 4273–4287. PMID: 22287626

247. Sattler C, Steer B, Adler H. Multiple Lytic Origins of Replication Are Required for Optimal Gammaherpesvirus Fitness In Vitro and In Vivo. PLoS Pathog. 2016;12: e1005510. PMID: 27007137

248. Xu Y, Rodriguez-Huete A, Pari GS. Evaluation of the Lytic Origins of Replication of Kaposi's Sarcoma-Associated Virus/Human Herpesvirus 8 in the Context of the Viral Genome. J Virol. 2006;80: 9905–9909. PMID: 16973596

249. Moens U, Prezioso C, Pietropaolo V. Genetic diversity of the noncoding control region of

the novel human polyomaviruses. Viruses. MDPI AG; 2020. PMID: 33297530

250. Ferenczy MW, Marshall LJ, Nelson CDS, Atwood WJ, Nath A, Khalili K, et al. Molecular biology, epidemiology, and pathogenesis of progressive multifocal leukoencephalopathy, the JC virus-induced demyelinating disease of the human brain. Clinical Microbiology Reviews. American Society for Microbiology (ASM); 2012. pp. 471–506. PMID: 22763635

251. Guiblet WM, Cremona MA, Harris RS, Chen D, Eckert KA, Chiaromonte F, et al. Non-B DNA: A major contributor to small-and large-scale variation in nucleotide substitution frequencies across the genome. Nucleic Acids Res. 2021;49: 1497–1516. PMID: 33450015

252. Hammerschmidt W, Sugden B. Identification and characterization of oriLyt, a lytic origin of DNA replication of Epstein-Barr virus. Cell. 1988;55: 427–433.

253. McCormick C, Ganem D. The kaposin B protein of KSHV activates the p38/MK2 pathway and stabilizes cytokine mRNAs. Science (80-). 2005;307: 739–741. PMID: 15692053

254. McCormick C, Ganem D. Phosphorylation and Function of the Kaposin B Direct Repeats of Kaposi's Sarcoma-Associated Herpesvirus. J Virol. 2006;80: 6165–6170. PMID: 16731955

255. Yoo J, Kang J, Lee HN, Aguilar B, Kafka D, Lee S, et al. Kaposin-B enhances the PROX1 mRNA stability during lymphatic reprogramming of vascular endothelial cells by Kaposi's sarcoma herpes virus. PLoS Pathog. 2010;6: 37–38. PMID: 20730087

256. Chang H-C, Hsieh T-H, Lee Y-W, Tsai C-F, Tsai Y-N, Cheng C-C, et al. c-Myc and viral cofactor Kaposin B co-operate to elicit angiogenesis through modulating miRNome traits of endothelial cells. BMC Syst Biol. 2016;10: S1.

257. Wu YH, Hu TF, Chen YC, Tsai YN, Tsai YH, Cheng CC, et al. The manipulation of miRNA-gene regulatory networks by KSHV induces endothelial cell motility. Blood. 2011;118: 2896–2905. PMID: 21715310

258. King CA. Kaposi's Sarcoma-Associated Herpesvirus Kaposin B Induces Unique Monophosphorylation of STAT3 at Serine 727 and MK2-Mediated Inactivation of the STAT3 Transcriptional Repressor TRIM28. J Virol. 2013;87: 8779–8791. PMID: 23740979