5 Chapter 5: Characterizing the Genetic Diversity of KSHV in The Uganda GPC

5.1 Introduction

The understanding of sequence diversity of KSHV at the whole-genome level and its relation to pathogenesis and disease is limited. In particular, if and how genetic variation in virus genes contributes to the control of KSHV infection and the potential development of disease is not very well understood. The first KSHV genomes to be sequenced used Sanger sequencing technology and were all derived from tumours or cell lines using cosmid clones or DNA inserted into bacterial artificial chromosome (BAC) technology. The large dsDNA ~165 kb genome sequence of KSHV was first determined by Russo and colleagues in 1996[138]. Using phage and cosmid libraries from the KSHV infected primary effusion lymphoma (PEL) cell line, BC-1, they revealed a long unique coding region (LUR) of ~140kb with 81 ORFs some of which had functional homologues in *herpesvirus samiri* (HVS), a related non-human primate gamma-herpesvirus, and 801 bp long terminal repeats (TR) at the ends of the genome[138]. Following this, a nearly complete genome was sequenced using shotgun sequencing of fragments following partial digestion of DNA isolated from AIDS-associated-KS biopsies[559]. The first complete and most extensively annotated KSHV genome sequence, GK18 was isolated from a Greek individual with classic KS and sequenced from a cosmid clone[560]. Two additional genomes were generated from KSHV-infected PEL cell lines isolated in the USA: JSC-1 and BCBL-1 were sequenced from DNA cloned into a BAC16 and BAC36 cassette, respectively[138,561,562]. A more recently sequenced whole-genome, DG-1, was sequenced using Illumina sequencing technology (unlike the five genomes described above) and was the first genome to be generated from actively replicating virus isolated from blood plasma of a patient infected with KSHV inflammatory cytokine syndrome (KICS) and also co-infected with HHV-6[563]. A very recent study, conducted by Olp and colleagues, used Illumina paired-end SureSelect deep-sequencing, and generated 16 whole-genomes

directly from skin lesions of Zambian KS patients, actively enriching for viral DNA, this represents the first non-Western genomes to be generated from sub-Saharan Africa, and a KSHV and KS endemic region[564].

The KSHV genome map has changed little since its discovery, with the annotation of the GK18 sequence revealing 86 genes of which 22 encode putative immunomodulatory proteins (see chapter 1 Fig. 1.5). Extensive transcriptional and proteomic profiling has increased our understanding of KSHV gene expression at different stages in its lifecycle and improved annotations of non-coding regions[204,565,566].

The KSHV genome displays high conservation with 99% sequence similarity between viral strains, however both 5' and 3' ends of the genome have displayed high sequence variability and as such have been used to characterize viral strains[187]. ORF K1 located at the 5' termini of the genome (see chapter 1, Fig. 1.5) encodes highly glycosylated transmembrane proteins, with hypervariable regions (V1 and V2) that display up to 30% amino acid variability, such that K1 variants are used for viral genotyping[186]. Seven major KSHV subtypes, A-E and more recently F and Z have been identified by K1 subtyping (and phylogenetics) and display considerable geographic variation[187,567,568]. The P (Predominant), M (Minor) and N alleles from the K15 locus at the 3' termini of the genome (see chapter 1, Fig. 1.5), that encodes an integral membrane protein with up to 70% inter-allele divergence at the amino acid level, has also been used to characterize viral variants[192,569,570]. While the central region of the KSHV genome is highly conserved, nine discrete loci: K12, K2, K3 ORF18/19, ORF26, K8, ORF73 and two loci within ORF75 (which make up ~5.6% of the genome) with low level variation have also been used in a number of phylogenetic studies for characterisation of subtypes[192]. For example, ORF26, encodes a minor capsid protein and has been used for variant characterisation to identify KSHV subtypes A/C, J, K/M, D/E, B, Q, R or N which are also heterogeneously distributed amongst different populations although distribution is found to parallel K1 subtypes[571]. The remaining >90% of the genome has not been taken into account due to lack of high

coverage whole-genome sequencing data. While the early sequencing studies provided insights into KSHV genome architecture, the variability of the K1 and K15 genes and the coding capacity of its genome, the Zambian KS whole-genome study is the first to show that low level genetic variation in the central conserved region contributes to a unique phylogenetic structure; showing distinct genomic variants of Zambian isolates compared to Western (USA and Greece) isolates, and therefore, are indeed important for accurate viral characterization[564]. With a small number of genomes from only three countries (USA, Greece and Zambia) and all from diseased individuals, whether genomic diversity contributes to the distribution of KSHV seroprevalence and incidence of associated diseases remains unclear.

Uganda presents a good candidate to study KSHV molecular epidemiology and phylogeography as it is inhabited by different ethno-linguistic groups with divergent historic origins as a result of migration over several hundred years from surrounding regions[391,572,573], and in addition the population sustains the highest seroprevalence of KSHV in the world[150,574,575]. In the GPC, the seroprevalence of KSHV is >90% (see chapters 2 and 4). Several studies conducted in Uganda have provided invaluable insights into KSHV seroepidemiology and transmission[145,146,378,382,385,576], therefore, characterising genetic diversity on a whole-genome scale will bridge gaps in the understanding of the co-evolution of host and virus and its implications in disease pathogenesis.

5.1.1 Chapter Aims

With only 21 full genome sequences of KSHV published to date from three different countries and all isolated from KSHV associated diseases, the understanding of KSHV genomic diversity in relation to disease pathogenesis is not clear. No genomes have been isolated from asymptomatic persistently infected individuals and thus the 'wild-type', non-tumour associated KSHV genome has never been characterised. Therefore, the aims of this chapter are to:

I. Generate whole-genome sequences of KSHV isolated from saliva of KSHV disease free individuals and assess the variability between KSHV genomes isolated from different sources and of diseased individuals.
II. Assess the population structure of KSHV strains within the Uganda GPC and in a global context.

Contributions

Sample collection, storage and shipment was conducted by the GPC team in Uganda. RNA bait libraries for target enrichment were developed by the Virus genomics team. Whole-genome sequencing was conducted by the Illumina high-throughput sequencing team at Sanger. All other analyses unless otherwise stated were performed by myself.

5.2 Methods

5.2.1 Sample Selection and Collection

The Uganda GPC and ethics are described in detail in chapter 2. Briefly, the GPC is a population-based cohort in rural south west Uganda consisting of 25 neighbouring villages mainly inhabited by peasant farmers who grow bananas as a subsistence crop, cultivate coffee for trade and also raise livestock[388,389]. Household are scattered with some concentrated in the trading centres. The Baganda are the predominant tribal group constituting ~70% of the population with a substantial number of migrants who settled from neighbouring Rwanda. To assess the determinants of viral shedding and diversity of KSHV whole-genomes, 2036 saliva samples were collected from individuals during medical survey round 24 between January to July 2015. 2ml of saliva was collected with the Oragene® DNA self-collection kit, OMNIgene®.ORAL OM-505 (DNA Genotek Inc., ON, Canada) following manufacturer's instructions by the GPC team in Uganda and stored at -80^{o}C prior to shipment on dry ice to the Sanger Institute.

5.2.2 DNA Extraction, Purification and Quantification

I conducted all sample preparation in class II biosafety cabinets using aseptic techniques. Saliva samples were transferred to Corning™ Costar™ 96 well plates (ThermoFisher scientific, UK) in 1 ml aliquots, and lysed. RNA was removed with proteinase K (600mAU/ml) Buffer VXL solution and RNase A (100mg/ml) treatment (Qiagen™, UK). 200 ul aliquots of lysates were then aliquoted into 96-well S-blocks (Qiagen™, UK) for DNA extraction using the QIAamp 96 DNA QIAcube®HT robot following the manufacturer's protocol, and the remainder stored at -80^{o}C. Briefly, 96 samples were processed simultaneously, optimal DNA binding and filtering of contaminants were achieved through buffering steps followed by ethanol wash steps to remove residual contamination and enzyme inhibitors. DNA was eluted in buffer AE under vacuum and then enhanced by overlaying elution buffer TE fluid to achieve a final volume of 80 ul of

DNA per sample. Quantification of total genomic DNA was performed using the Quanti-iT™ PicoGreen® dsDNA assay kit (ThermoFisher scientific, UK).

5.2.3 Quantitative PCR for Viral DNA Detection

I used quantitative PCR (qPCR) for viral genome detection and determination of viral genome load measured by determining the viral copy number relative to a control PEL DNA sample. Out of the 2000 samples, 746 were processed in duplicates using the QuantiTect Muliplex PCR kit (Qiagen, UK) on a Stratagene Mx3005P (Agilent Technologies, UK) owing to time-constraints. Primers and probes targeting KSHV *ORF73* were designed for viral detection using sequences from Lallemand *et al*.,[577] (Table 5.1). GAPDH was used as a normalizing assay with sequences from Pardieu *et al*,[578] (Table 5.1). All primers and probes were synthesised by Metabion international AG, Germany. Primer-probe mixes were diluted to a 20X solution, for KSHV this consisted of 10 pmol/ul of each primer and 1.25pmol/ul for the probe; for GAPDH 2.5pmol/ul for each primer and 1.25 pmol/ul for the probe for each reaction. The master mix for qPCR in a 25 ul/reaction was as follows: 1.25 ul of KSHV primer-probe 20X mix, 2 ul GAPDH primer-probe 20X mix, 12.5 ul QuantiTect multiplex mastermix, 4.25 ul nuclease-free water and 5 ul of DNA. DNA from the BCBL-1 cell line was used as a positive control and used to generate a standard curve with 10-fold serial dilutions from $3x10^6$ to 30. The qPCR conditions were as follows: Initial denaturation at 95^oC for 15 mins and 45 cycles of denaturation at 95^oC for 15s and annealing at 60^oC for 1min. The fluorescence data was captured during the annealing step. The Ct values were compared to the standard curve to assign a copy number per ml. Data analysis was performed using MxPro v4.10 qPCR software (Agilent Technologies).

Table 5.1 qPCR Primer and probe sequences

Primer/Probe	Sequence (5' -> 3')
ORF73 - Forward	TTGCCACCCACGCAGTCT
ORF73 - Reverse	GGACGCATAGGTGTTGAAGAGTCT
ORF73 - Probe	6-FAM-TCTTCTCAAAGGCCACCGCTTTCAAGTC-TAMRA
GAPDH- Forward	GGCTGAGAACGGGAAGCTT
GAPDH - Reverse	AGGGATCTCGCTCCTGGAA
GAPDH - Probe	HEX-TCATCAATGGAAATCCCATCACCA-BHQ-2

5.2.4 KSHV Whole-Genome Sequencing

I selected 240 samples for whole-genome sequencing based on viral DNA detection (Ct values <36) following qPCR. For KSHV target enrichment, overlapping 120mer RNA baits spanning the length of KSHV GK18 and BC1 reference sequences (Accession numbers: NC_00933 and NC_003409, KSHV Type P and Type M respectively) were designed by Matt Cotten using eArray software (Agilent Technologies). The Illumina High Throughput sequencing team at the Wellcome Trust Sanger Institute performed target enrichment and genome sequencing following the SureSelect™ protocol (version 1.1). Briefly, 1-3ug of each DNA sample was sheared to 200-500bp fragments followed by end-repair, non-template addition of 3'-A, adaptor ligation, hybridisation, enrichment PCR, index tagging and sample pooling. Samples were multiplexed on an 8 lane flow cell with 24 samples per lane, cluster generation and sequencing was performed on an Illumina HiSeq 2000 sequencer. Sequencing reads were 250bp paired-ends in FASTQ format with per base Phred quality scores.

5.2.5 Guided Assembly of KSHV Whole-Genomes

I used the QUASR QC pipeline (http://sourceforge.net/projects/quasr)[579] to retain high quality full length reads. Duplicate reads and paired reads with a raw median Phred quality score Q<32 were either filtered out or trimmed from the 3' end until Q>32. Any

reads less then 100bp in length post trimming were also excluded. High quality paired-end reads post-QC were then mapped back to GK18 and BC1 reference sequences using Burrows-Wheeler Aligner (BWA)[580] and the average depth and coverage calculated using SAMTools[581] with an in-house script written by Anne Palser. To investigate whether viral load influenced sequencing quality, I generated a pairwise-correlation matrix for qPCR viral load, KSHV mapped reads (%) and sequencing depth of coverage using Pearson's correlation in R.

5.2.6 Comparative and Phylogenetic Sequence Analysis

For comparative and phylogenetic analysis I selected 62 consensus sequences with an average sequencing depth of at least 10x and coverage of >90% across the genome, generated following BWA mapping and aligned them with 21 publically available KSHV Genomes from Greece, USA and Zambia using MAFFT[582] (v7.0) and viewed using AliView software. I masked repeat regions across the alignment with coordinates retrieved from the GK18 reference sequence annotation in Genbank (). SNPs between genomes were counted relative to consensus sequence generated from the multiple sequence alignment and genome-wide mutations were visualised in a 1000 nucleotide scanning window and the number of codon changes per gene (synonymous and non-synonymous) were calculated using an in-house script written by Simon Watson. I then used the multiple sequence alignment to generate whole-genome trees using maximum-likelihood methods implemented in RAxML (v8) with the general time reversible (GTR) model of nucleotide substitution including a Gamma distribution for among site rate variation[583]. Tree topology was assessed using 1000 bootstrap replicates in RAxML. I also generated whole-genome trees removing the K1 and K15 variable genes. To investigate genotypic diversity, I performed phylogenetic analysis following alignment of the coding sequences of the K15 gene and K1 gene along with representative sequences for the following genotypes (Genbank accession number): A1 (AF133038), A2 (AF130305), A3 (U86667), A4 (AF133039), A5 (AF178823), B1 (AF133040), B2 (AY042947), B3

(AY042941), B4 (DQ309754), C1 (AF133041), C3 (AF133042), D1 (AF133043), D2 (AF133043), E (AF220292) and F (FJ884616). All trees were midpoint rooted.

An overview of the workflow used in this chapter from sample collection to analysis is presented in Fig. 5.1.

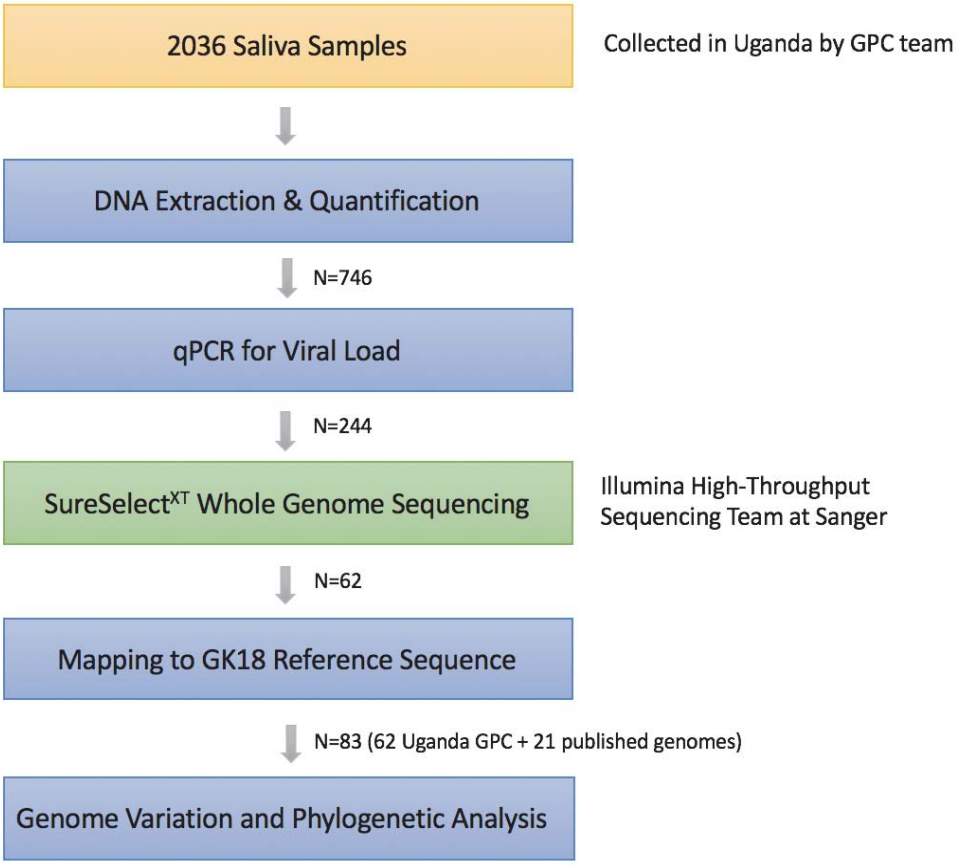

Fig. 5.1 KSHV genome analysis workflow. N represents the number of samples processed at each stage of the analyses.

5.3 Results

5.3.1 KSHV Shedding and Viral Load in the GPC

In this study, 2036 saliva samples were collected from asymptomatic individuals in a cluster of neighbouring villages in the GPC (previously described in chapter 2, Fig. 2.1) from January to July 2015 to characterise and assess the genetic diversity of wild-type KSHV whole-genome sequence. Following DNA extraction, I screened 746 randomly selected samples for KSHV positivity and viral load based on qPCR targeting the ORF73 gene and using a ten-fold dilution of BCBL-1 DNA to ascertain viral copy numbers against a standard curve with a detection range of $3x10^6$ – 10 copies/ml (C.t of 15 to 43) (Fig. 5.2). While most of the individuals in the GPC (>90%) are seropositive to KSHV (see chapter 2 and 4), qPCR positivity reflects individuals shedding virus and therefore with viral DNA in saliva, given that viral DNA is only detectable during the lytic stage of infection. Following qPCR, 244 (32.8%) individuals had detectable KSHV viral DNA above the qPCR threshold with C.t values ranging from 21.55 to 41.5 representing a viral load from $5.35x10^5$ to 1.5 copies/ml (mean=9186 copies/ml), respectively. Out of the 244 samples, 80 (32.7%) samples had viral loads from $>10^4$ – 10^5 (C.t <31), 70 (28.6 %) samples had viral loads from 10^2 – 10^4 (Ct>31-35) and 94 samples (38.7%) had viral loads from <10 – 10^2 (Ct >35).

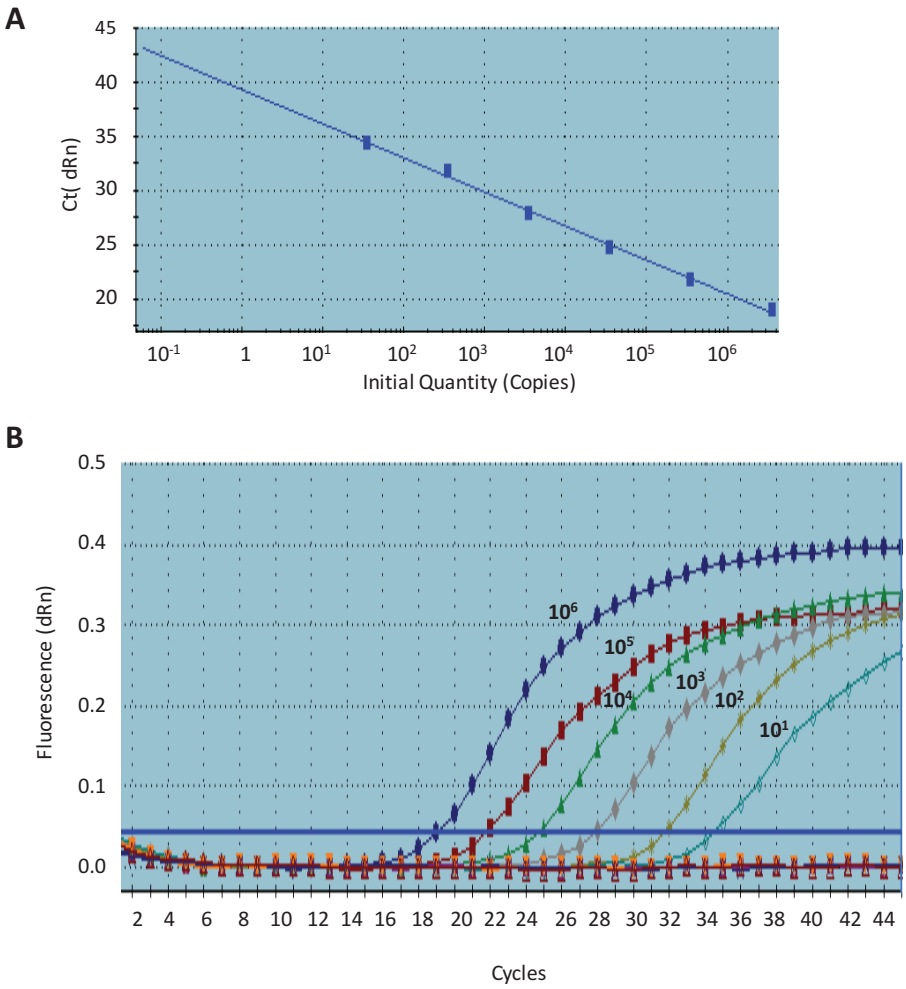

Fig. 5.2 KSHV ORF73 gene qPCR for BCBL-1 DNA dilution series from 30 to 3x10^6 viral copies/ml. A. Standard Curve. B. Amplification plot. Cycles represent cycles of PCR amplification and the blue solid line is the detection threshold for the FAM fluorescence channel. The amplification of standards is represented accordingly, all negative control samples (H_20) are below the threshold.

5.3.2 KSHV Viral Load Correlates with Whole-Genome Sequencing Quality

The low abundance of viral DNA compared to the host DNA in addition to the large KSHV genome, makes sequencing of KSHV quite the challenge, thus, to maximise the chances of success all 244 samples with detectable viral DNA were submitted for whole-genome sequencing using the SureSelect^{XT} method to allow selective capture of KSHV DNA using custom biotinylated RNA baits, the success of this approach has been demonstrated using cell lines and clinical samples by other studies[584]. Following this, all samples were mapped to the KSHV GK18 reference sequences for type P and type M KSHV strains (NC_009333 and NC_003409 respectively) using BWA. The GK18 sequence was selected as it was used for the SureSelect RNA bait library design and also represents the most comprehensively annotated whole-genome sequence.

To assess whether viral load influenced the number of KSHV reads that were mapped and the mean sequencing depth of coverage >90% and thus generated better quality reads, I generated a correlation matrix calculated using Pearson's correlation in R. This confirmed that qPCR viral load were strongly positively correlated with the percentage of mapped reads (r^2=0.84) and also showed that a high viral load was positively correlated with achieving good mean sequencing depth with at least 90% coverage across the genome (Fig. 5.3). Out of the 244 samples, 62 (25.4%) had at least a 10x mean sequencing depth of >90% coverage across the genome. Of these 62 samples, depth ranged between 10x-1000x and 55% of samples had an average 25x depth across the genome. As these 62 represent the samples that also have a higher percentage of KSHV mapped reads with greater accuracy (up to ~77% mapped reads) and thus, better genome quality, I used this sample set for downstream analysis. The 62 samples corresponded to the samples with the highest viral loads (10^4 – 10^5 copies/ml) and were collected from 8 neighbouring villages (12-19) in the GPC (Fig. 5.4); they consisted of 32 males and 30 females between the ages of 16-91 (Mean ± S.D = 41.65 ± 20.69), 5 individuals were also HIV positive.

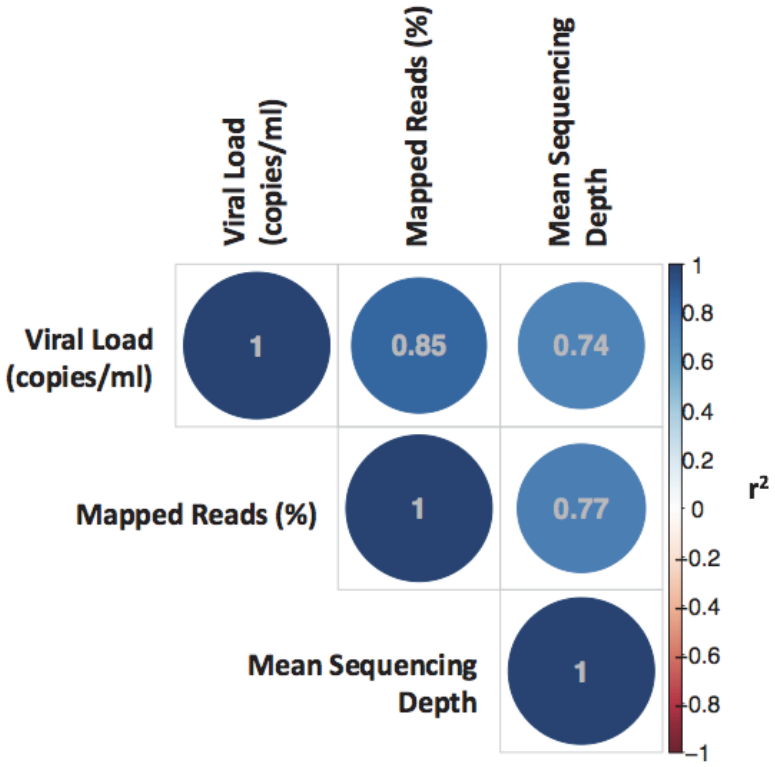

Fig. 5.3 Correlation matrix of Viral load (copies/ml), KSHV mapped reads (%) and mean sequencing depth of 200x. Correlating viral load as estimated by qPCR in copies/ml with % KSHV mapped reads and with a mean depth of 200X at >=90% coverage. Positive correlations are in blue, negative correlations are in red, intensity and size of the circle are proportional to the correlation coefficients (r^2) labelled in the circles and indicated on the right hand side of the correlogram. All tests meet Pearson's significance threshold of $p<0.01$.

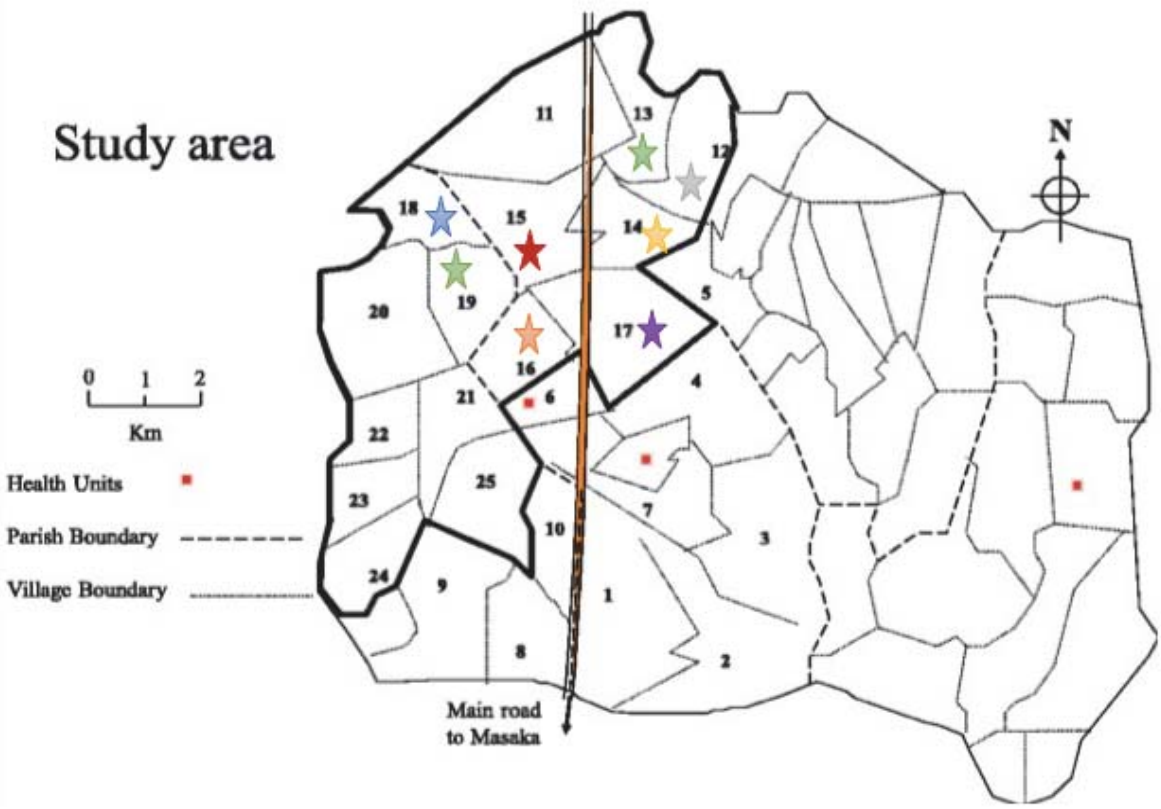

Fig. 5.4 Map showing GPC the study area in Uganda. The stars correspond to the 8 villages (12-19) where the 62 samples were collected from and the colours represent the number of samples in each village (Grey=2, blue=3, purple = 5, green=6, yellow=9, orange=13 and red=19).

5.3.3 KSHV Genome Variability

To determine how variable the 62 new saliva (wild-type) KSHV genomes were and explore which parts of the genome were contributing to the most variation, I performed a multiple sequence alignment including the 21 previously published KSHV genome sequences from Greece, USA and Zambia (Table 5.2) and used the consensus sequence generated from the alignment as reference for single nucleotide polymorphism (SNP) calling. Genes were annotated, and gaps and repeat regions were masked based on the GK18 sequence coordinates retrieved from Genbank (Table 5.3).

Table 5.2 Summary of KSHV samples used in this study

Sample Name	No.	Geographic Origin	Clinical Presentation	Source
GK18[560]	1	Greece	KS	KS Tumour
BCBL1[562]	1	USA	PEL	B Cell line
JSC1[561]	1	USA	PEL (EBV+)	B Cell line
BC1[138]	1	USA	PEL (EBV+)	B Cell line
DG1[563]	1	USA	KICS (HHV-6+)	Blood
ZM*[564]	16	Zambia	KS	KS Tumour
UG*	62	Uganda	Asymptomatic carrier	Saliva

KS-Kaposi's Sarcoma, PEL – Primary effusion lymphoma, KICS- KSHV inflammatory cytokine syndrome

Table 5.3 Eighty-four annotated KSHV genes based on the GK18 sequence

Gene	Start-Stop[a]	Timing[b]	Cycle[b]	Putative Function[b]
K1	105-944	Latent	Latent	Glycoprotein
ORF4	1112-2764	24h	Lytic	Complement binding protein
ORF6	3179-6577	24-48h	Lytic	ssDNA Binding Protein
ORF7	6594-8681	N.R	N.R	Virion Protein
ORF8	8665-11202	48-72h	Lytic	Glycoprotein B
ORF9	11329-14367	48-72h	Lytic	DNA Polymerase
ORF10	14485-15741	48-72h	Lytic	Regulator of Interferon Function
ORF11	15756-16979	8h	Lytic	Predicted UTPase
K2	17227-17841	Latent	Latent	vIL6 homolog
ORF2	17887-18519	N.R	N.R	Dihydrofolate Reductase
K3	18574-19542	24h	Lytic	Immune Modulator
ORF70	20023-21036	N.R	N.R	
K4	21480-21764	8h	Lytic	vMIP-II
K5	25865-26635	8h	Lytic	RING-CH E3 Ubiquitin Ligase
K6	27289-27576	8h	Lytic	vMIP-IA
K7	28774-29154	8h	Lytic	Late gene expression (overlaps with PAN)
ORF16	30242-30769	8h	Lytic	Bcl2 Homolog
ORF17	30920-32524	48h	Lytic	Protease
ORF18	32523-33296	24h	Lytic	Late gene regulation
ORF19	33293-34942			
ORF20	34710-35483			
ORF21	35482-37224	48-72h	Lytic	Thymidine Kinase
ORF22	37212-39404	48-72h	Lytic	Glycoprotein H
ORF23	39401-40615	48-72h	Lytic	Glycoprotein(predicted)
ORF24	40619-42877	48-72h	Lytic	Essential for replication (MHV68)
ORF25	42876-47006	48-72h	Lytic	Major Capsid Protein
ORF26	47032-47949	48-72h	Lytic	Minor Capsid Protein
ORF27	47973-48845	48-72h	Lytic	Glycoprotein (MHV68)
ORF28	49091-49399	48-72h	Lytic	BDLF3 EBV Homolog
ORF29	49462-50604, 53855-54775	72h	Lytic	Packaging Protein
ORF30	50723-50956	48-72h	Lytic	Late Gene Regulation (MHV68)
ORF31	50953-51537	48-72h	Lytic	Nuclear and Cytoplasmic (MHV68)
ORF32	51504-52868	48-72h	Lytic	Tegument Protein
ORF33	52861-53865	48-72h	Lytic	Tegument Protein (MHV68)
ORF34	54774-55757	24-48h	Lytic	N/A
ORF35	55738-56190	24-48h	Lytic	N/A
ORF36	56075-57409	24-48h	Lytic	Serine Protein Kinase
ORF37	57372-58832	24-48h	Lytic	Sox
ORF38	58787-58972	24-48h	Lytic	Myristylated Protein
ORF39	59072-60274	24-48h	Lytic	Glycoprotein M
ORF40	60407-61756, 61884-62543	48-72h	Lytic	Helicase Primase
ORF42	62535-63371	48-72h	Lytic	Tegument Protein
ORF43	63235-65052	48-72h	Lytic	Portal Protein (capsid)
ORF44	64991-67357	48-72h	Lytic	Helicase
ORF45	67452-68675	8h	Lytic	RSK activator
ORF46	68736-69503	24h	Lytic	Uracil deglycosylase

ORF47	69511-70014	24h	Lytic	Glycoprotein L
ORF48	70272-71480			N/A
ORF50	71695-71712, 72671-74728	8-24h	Lytic	RTA
ORF49	71729-72637			Activates JNK/p38
K8	74949-75662, 75744-75890	8-24h	Lytic	bZIP
K8.1	76014-76437, 76532-76794	48h	Lytic	Glycoprotein
ORF52	76901-77296	48-72h	Lytic	Tegument protein
ORF53	77432-77764	48-72h	Lytic	Glycoprotein N
ORF54	77835-78722	48-72h	Lytic	dUTPase/Immunomodulator
ORF55	78864-79547	48-72h	Lytic	Tegument Protein
ORF56	79535-82066	48-72h	Lytic	DNA Replication
ORF57	82169-82217, 82326-83644	8h	Lytic	mRNA Export/Splicing
vIRF-1	83960-85309	48-72h	Lytic	K9
vIRF-4	86174-88442, 88544-89010	48-72h	Lytic	
vIRF-3	89700-90945, 91042-91496	48-72h	Lytic	
vIRF-2	92066-93620, 93742-94229	48-72h	Lytic	
ORF58	94577-95650	24h	Lytic	N/A
ORF59	95655-96845	24h	Lytic	Processivity Factor
ORF60	96976-97893	24-48h	Lytic	Ribonucleoprotein Reductase
ORF61	97922-100300	24-48h	Lytic	Ribonucleoprotein Reductase
ORF62	100305-101300	72h	Lytic	N/A
ORF63	101314-104100	N.R	N.R	NLR Homolog
ORF64	104106-112013	N.R	N.R	Deubiquitinase
ORF65	112037-112549	48-72h	Lytic	Capsid
ORF66	112576-113865	48-72h	Lytic	Capsid
ORF67	113799-114614	48-72h	Lytic	Nuclear Egress Complex
ORF67A	114669-114911	48-72h	Lytic	N/A
ORF68	115108-116511	48-72h	Lytic	Glycoprotein
ORF69	116544-117452	48-72h	Lytic	BRLF2 Nuclear Egress
K12	118025-118207	Latent	Latent	Kaposin
ORF71	122393-122959	Latent	Latent	vFLIP
ORF72	123042-123815	Latent	Latent	vCyclin
ORF73	124057-127446	Latent	Latent	LANA
K14	128264-129079	24-48h	Lytic	vOX2
ORF74	129520-130548	24-48h	Lytic	vGPCR
ORF75	130699-134589	48-72h	Lytic	FGARAT
K15	134824- 136899	N.R	N.R	LAMP

[a] Genomic position from Genbank (GK18)
[b] Annotation from Arias et al, 2014[566]
N.R= Not reported

To get an overview of genomic variation across all 83 genomes, the proportion of variants to the consensus sequence were determined within a 1000 nucleotide sliding window. This showed a high proportion of nucleotide changes at the left termini of the genome (~35%), which corresponds to the K1 gene and at the right termini of the genome (~40%) corresponding to the K15 gene with modest variation observed across the central regions of the genome (Fig. 5.5). To further resolve the variation in the 84 individual genes across the genome, I calculated the number of synonymous and non-synonymous changes for each KSHV gene normalising for gene length (Fig. 5.6). This confirmed the presence of the highest variation in the K1 and K15 genes with a total of 39.1% and 39.6% base changes respectively. Other genes with high levels of non-synonymous changes include ORF73 (13.7%) and K12 (11.5%). The K1 gene had the highest ratio of non-synonymous to synonymous (dN/dS) changes of 7.8 and the K15 gene had a dN/dS ratio of 3.6. Other genes with a high dN/dS ratio (>=2) include, ORF73, K4.2, K8.1, ORF34, v-IRF2, ORF38, ORF67A with dN/dS ratios of 3.2, 2.8, 3.0, 2.2, 2.1, 2.0 and 2.0 respectively suggesting that these genes are evolving under selection (Fig. 5.7).

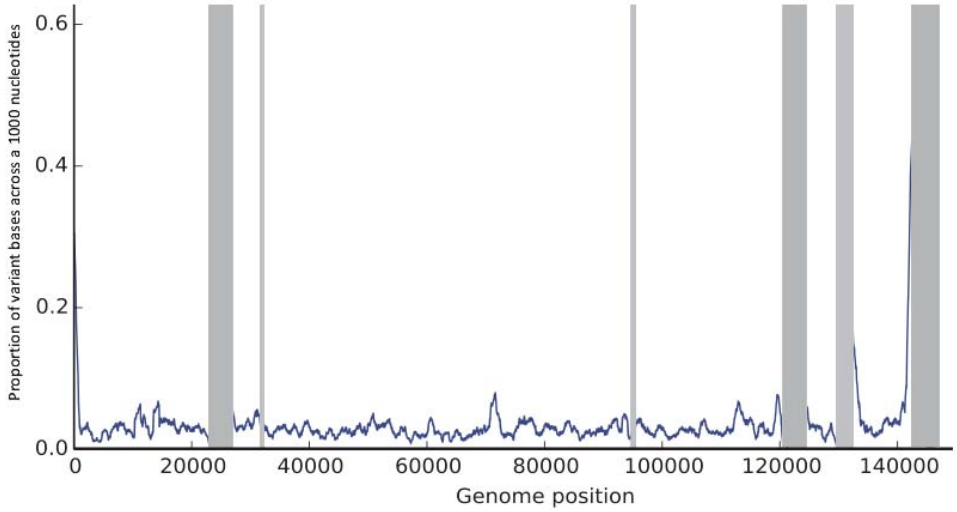

Fig. 5.5 Genome variability of 83 KSHV genomes. Line graph plotted across the genome showing the proportion of variant bases in a 1000 nucleotide sliding window where at least one KSHV genome sequence has a SNP relative to the consensus sequence generated from the alignment. Grey bars: Masked repeat regions.

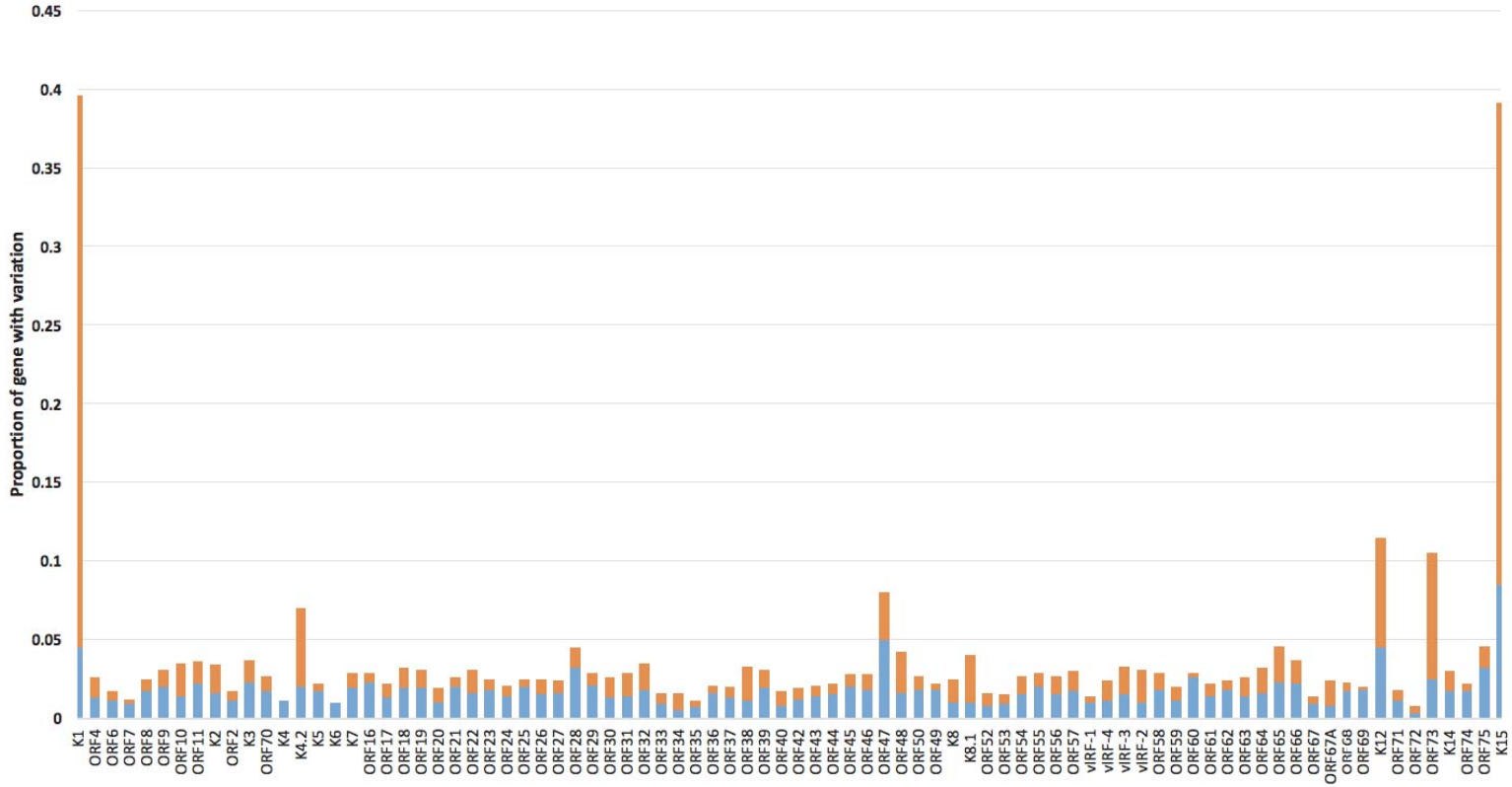

Fig. 5.6 SNP variation across coding region. Number of codon changes per gene across the genome relative to the consensus, presented as the proportion of the gene to normalise for gene length. Blue bars: Synonymous changes. Orange bars: Non-synonymous changes.

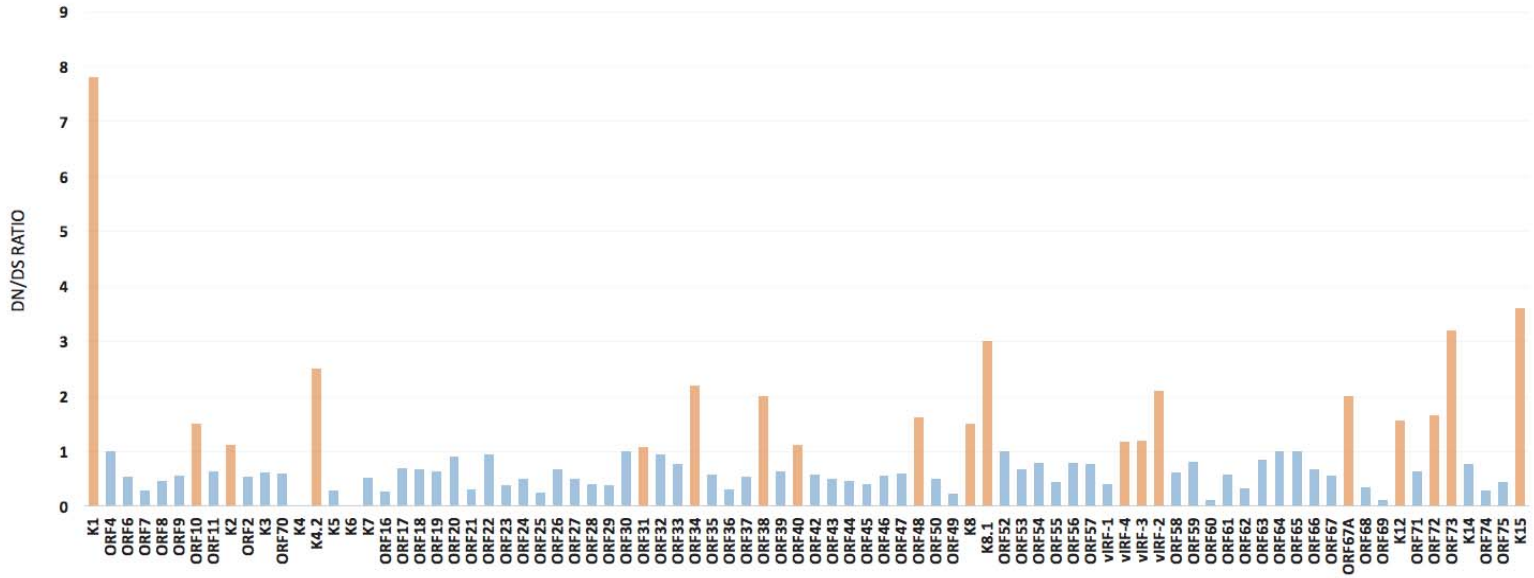

Fig. 5.7 Non-synonymous to synonymous change (dN/dS) analysis across KSHV coding region. Number of non-synonymous to synonymous mutations per gene across the genome relative to the consensus sequence generated from the alignment. Orange bars: dN/dS ratio >1. Blue bars: dN/dS ratio <=1

5.3.4 Virus Population Structure and Geographic Variability

To investigate the population structure of our 62 new wild-type KSHV whole-genomes from Uganda in a global context, I generated phylogenetic trees using a maximum likelihood method following a multiple sequence alignment of all 83 whole-genomes. The whole-genome analysis showed two distinct clades (Fig. 5.8), which have been previously classified as the type P and type M strains based on variation in the K15 gene. Two distinct sub-clades are also observed for both type P and type M strains. In addition, within each type, the Western samples (i.e. Greece and USA) cluster separately from the African samples (i.e. Zambia and Uganda) which show no distinct separation by country.

I also explored whether genomes had any patterns of distribution within villages in the GPC and observed no distinct clustering of samples by strain in the respective villages (Fig. 5.9). In addition, based on the tree there were no major differences between genomes isolated from saliva compared to other sources; and also no major differences between samples from asymptomatic vs. diseased individuals or cell lines (Fig. 5.8).

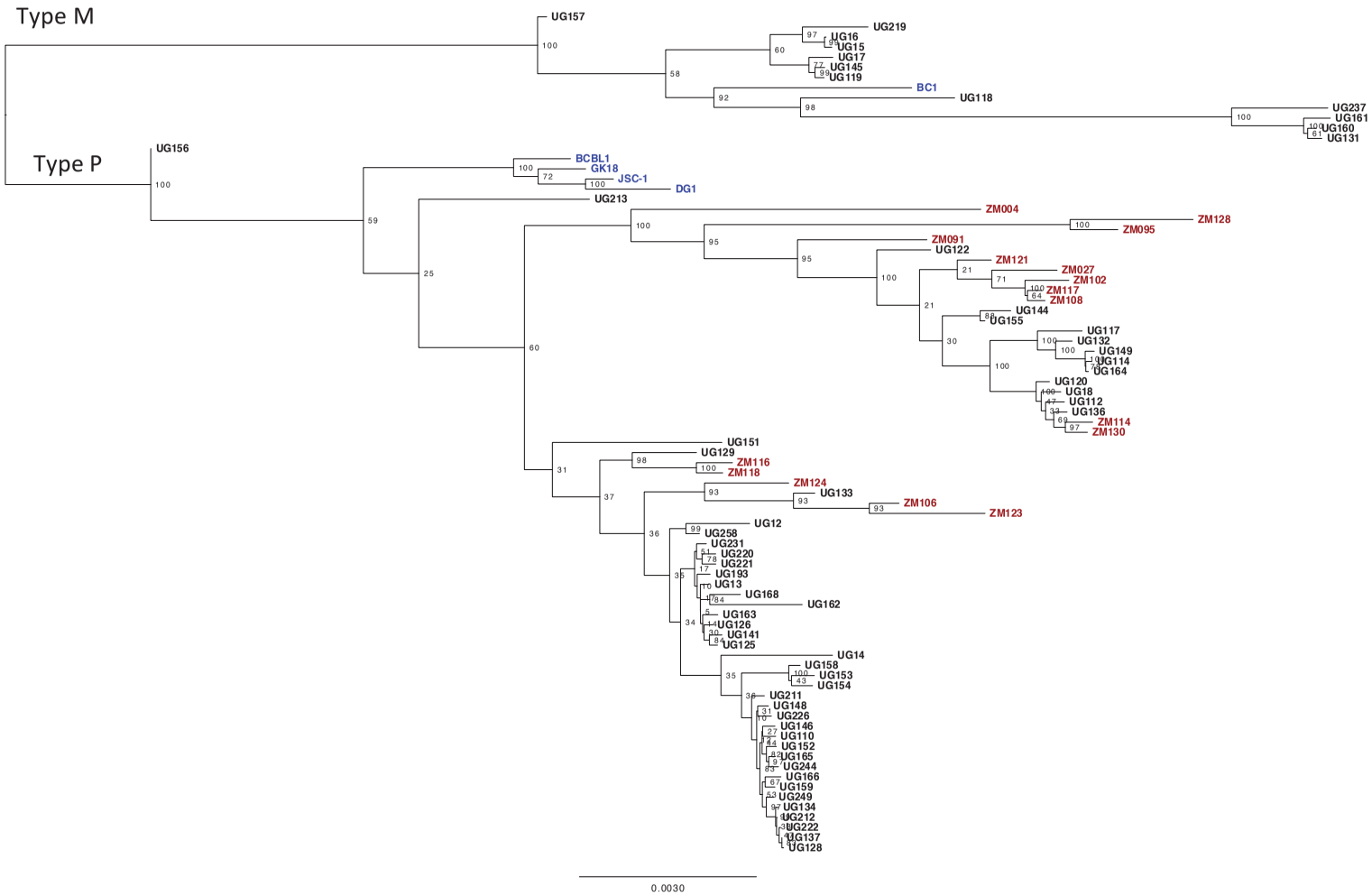

Fig. 5.8 KSHV whole-genome phylogenetic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names labelled in blue (Western), red (Zambian) and black (Ugandan GPC). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.003 nucleotide substitutions per site.

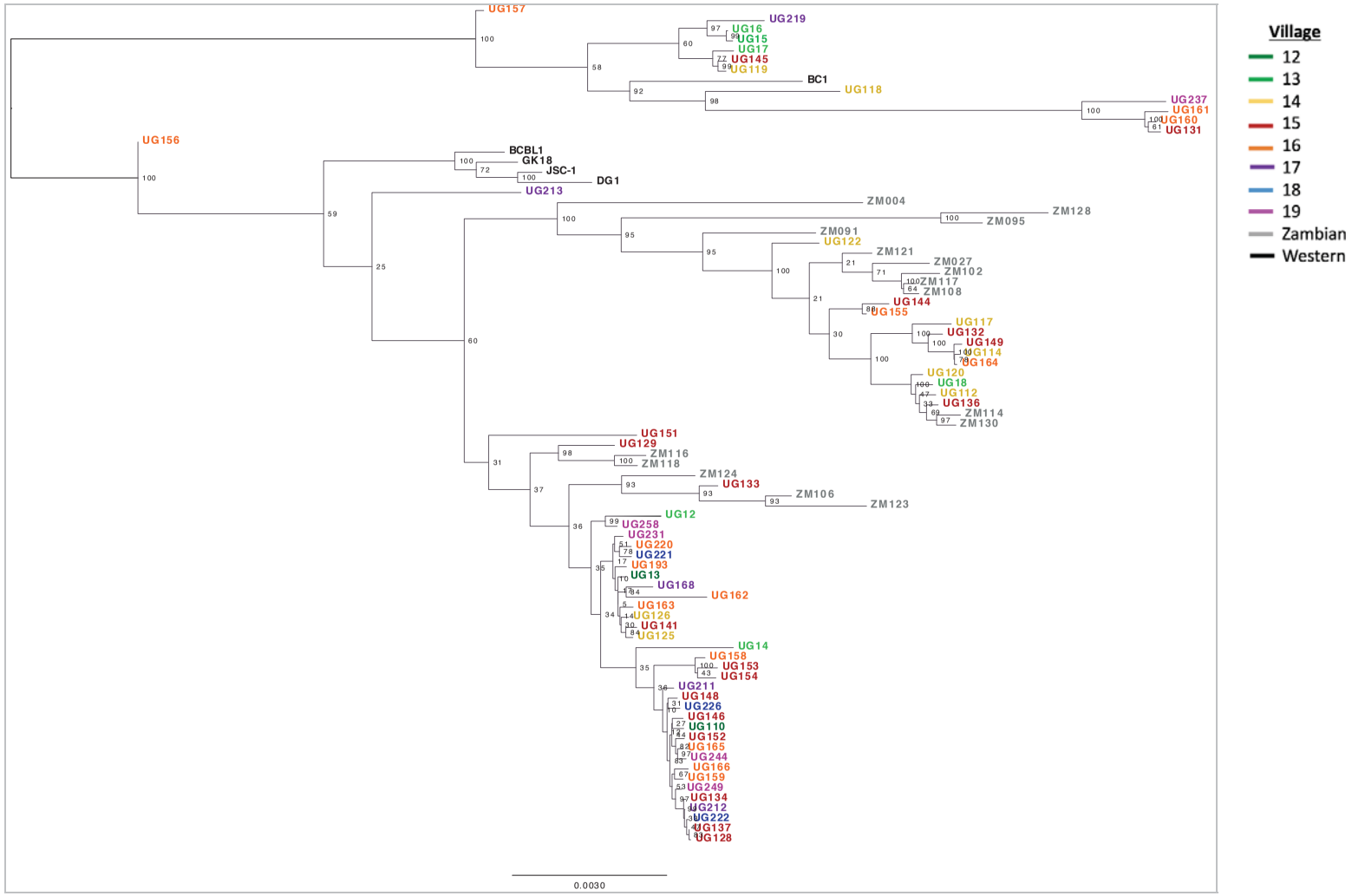

Fig. 5.9 KSHV whole-genome phylogeographic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names are coloured by village. Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.003 nucleotide substitutions per site.

To assess whether the tree topology was driven by the most variable genes, K1 and K15, I realigned the genomes of all the samples removing the K1 and K15 genes and generated a new tree. While the clustering by Type (P vs M) was lost, two distinct clades still remained and the Western isolates all remained clustered (Fig. 5.10). This substantiated that most of the variation was driven by K1 and K15, however, suggest that genes in the central region are also contributing to the diversity of genomes and thus the geographical clustering, this is consistent with the SNP analysis conducted previously (Fig. 5.5 and Fig. 5.6).

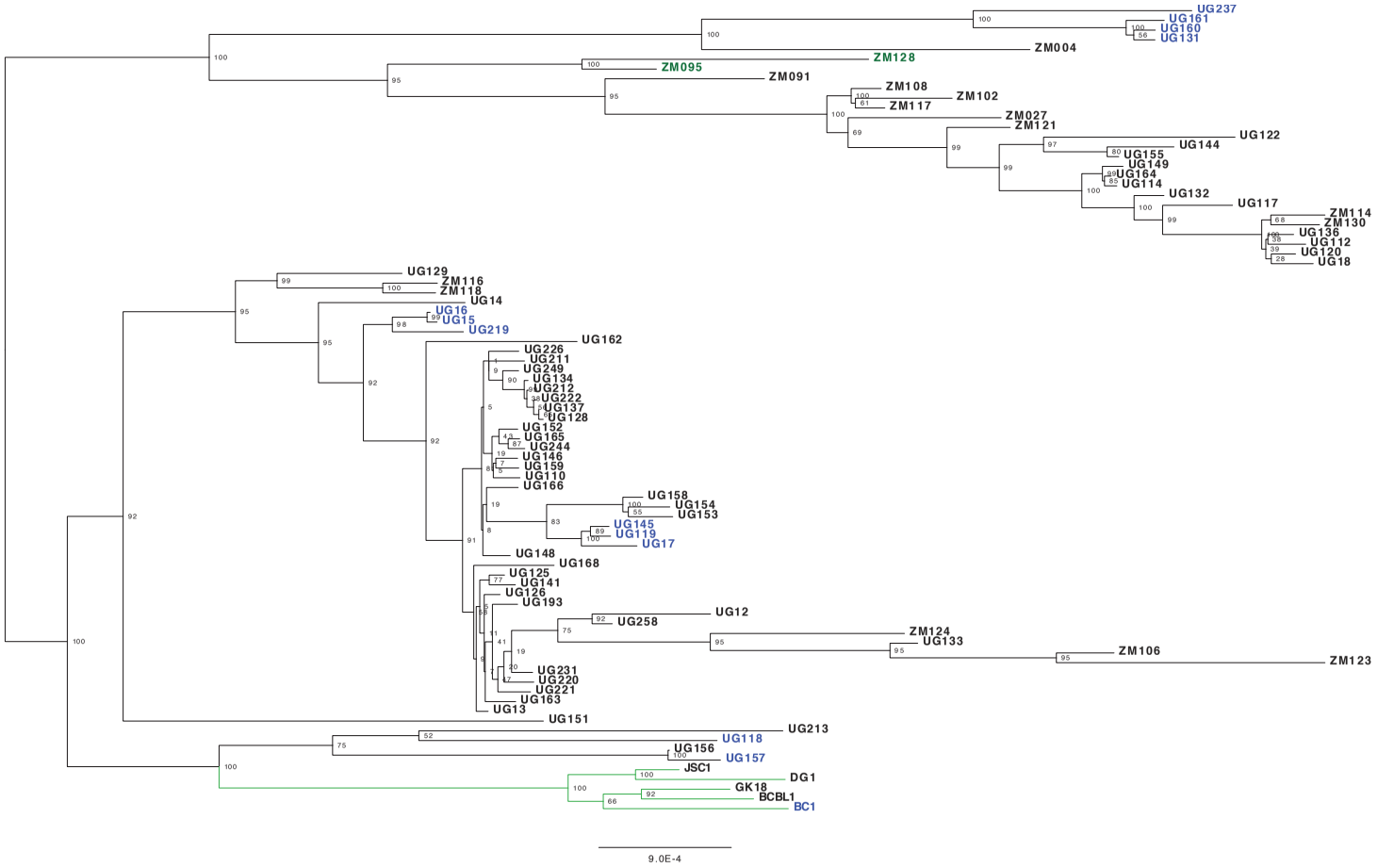

Fig. 5.10 KSHV genome phylogenetic analysis of central region minus K1 and K15 genes in 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names are coloured by Type: in black (P), blue (M) and green (N). Branches are also coloured by region: Black (Africa) and Green (Western). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.0009 nucleotide substitutions per site.

5.3.5 Genotypic Diversity of Strains in the GPC

Since the K1 and K15 genes are the most variable in the genome in this analysis and consistently with previous findings, I generated trees for each gene to determine the genotypes circulating in the 62 Ugandan samples and confirm the genotypes of the 21 samples that have been previously published. For the K15 phylogenetic analysis clear separation was observed between the strains types P Vs M, the majority of the Ugandan samples (50, 80%) remained clustered with the Type P strain and 12 (20%) samples cluster with the type M and none of the GPC samples belonged to the type N strain (Fig. 5.11). To date, these 12 samples from Uganda are the only type M genomes sequenced aside from the only published BC-1 sample which was sequenced from a PEL cell line isolated in the USA (Table 5.2). The major clades observed in the K15 phylogenetic tree are also consistent with that identified in the whole-genome tree (Fig. 5.8).

For the K1 phylogenetic analysis, I aligned the 83 genomes with representative K1 genes for the following genotypes: A1, A2, A3, A4, A5, B1, B2, B3, B4, C1, C3, D1, D2, E and F. Of the 62 Ugandan GPC samples, 30 (48.4%) clustered with B genotypes, 28 (45.1%) clustered with the A genotype and 4 (6.5%) samples clustered with the C genotype (Fig. 5.12). While the B genotypes displayed heterogeneity in subtypes, clustering mainly with B1 and B3, all the A genotypes clustered with the A5 subtype. The C genotypes clustered with the C1 subtype. All the Zambian samples but one, which belonged to the A5 genotype, clustered with the B genotypes and the Western samples clustered with the C and A genotypes as expected. No samples in this study clustered with the D, E or F genotypes. In the GPC, no clustering of genotypes by village was observed but rather a mixed distribution of types across all villages; only two samples were from the same family and living in the same household, UG129 and UG131, both belonged to the B genotype, but were K15 Type P and M respectively. A characterisation of all 62 samples in the GPC is presented in Table 5.4.

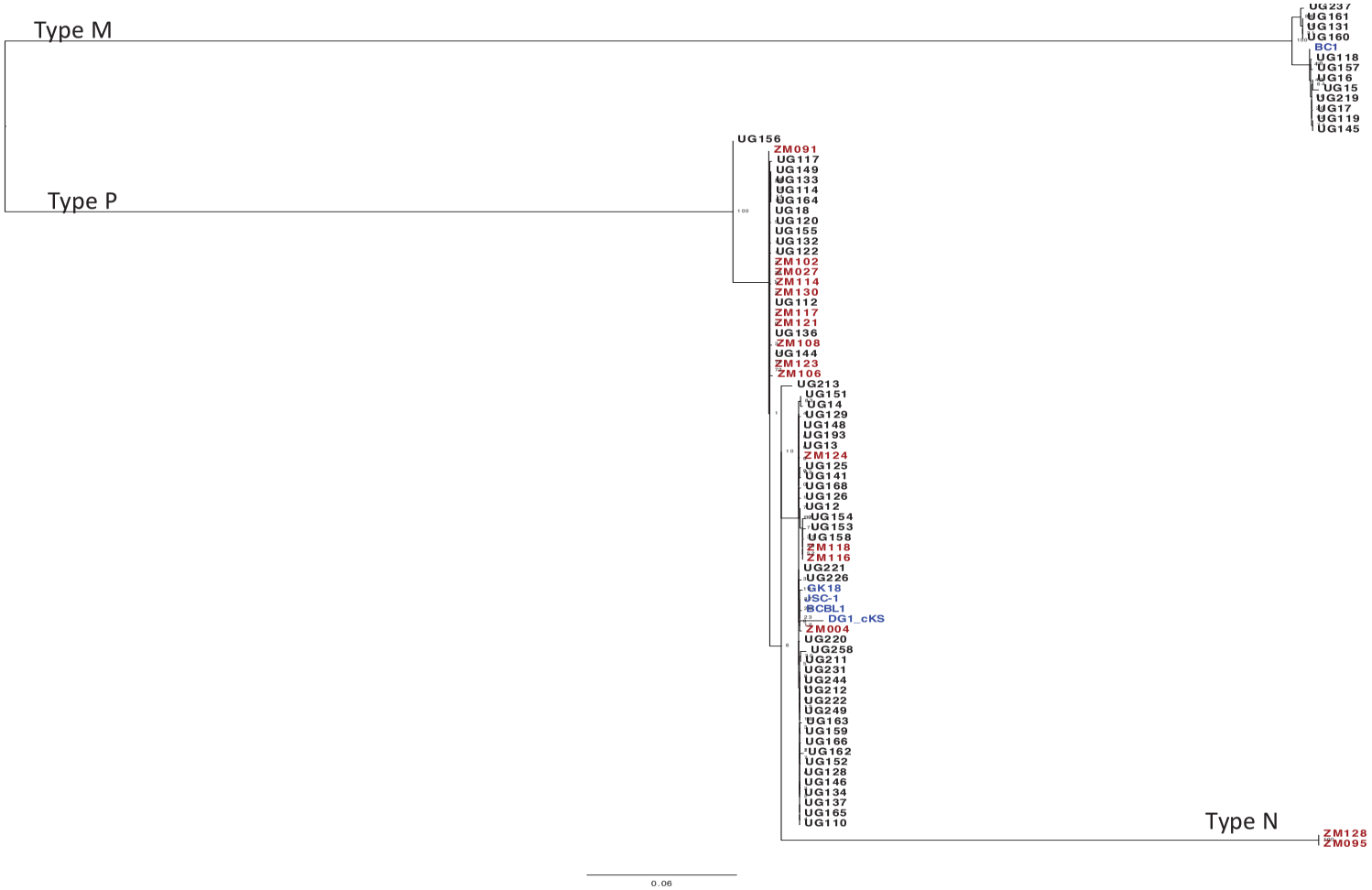

Fig. 5.11 KSHV K15 gene phylogenetic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of 21 published and 62 new KSHV genomes generated with 1000 bootstrap replicates. Sample names a labelled in blue (Western), red (Zambian) and black (Ugandan GPC). Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.09 nucleotide substitutions per site.

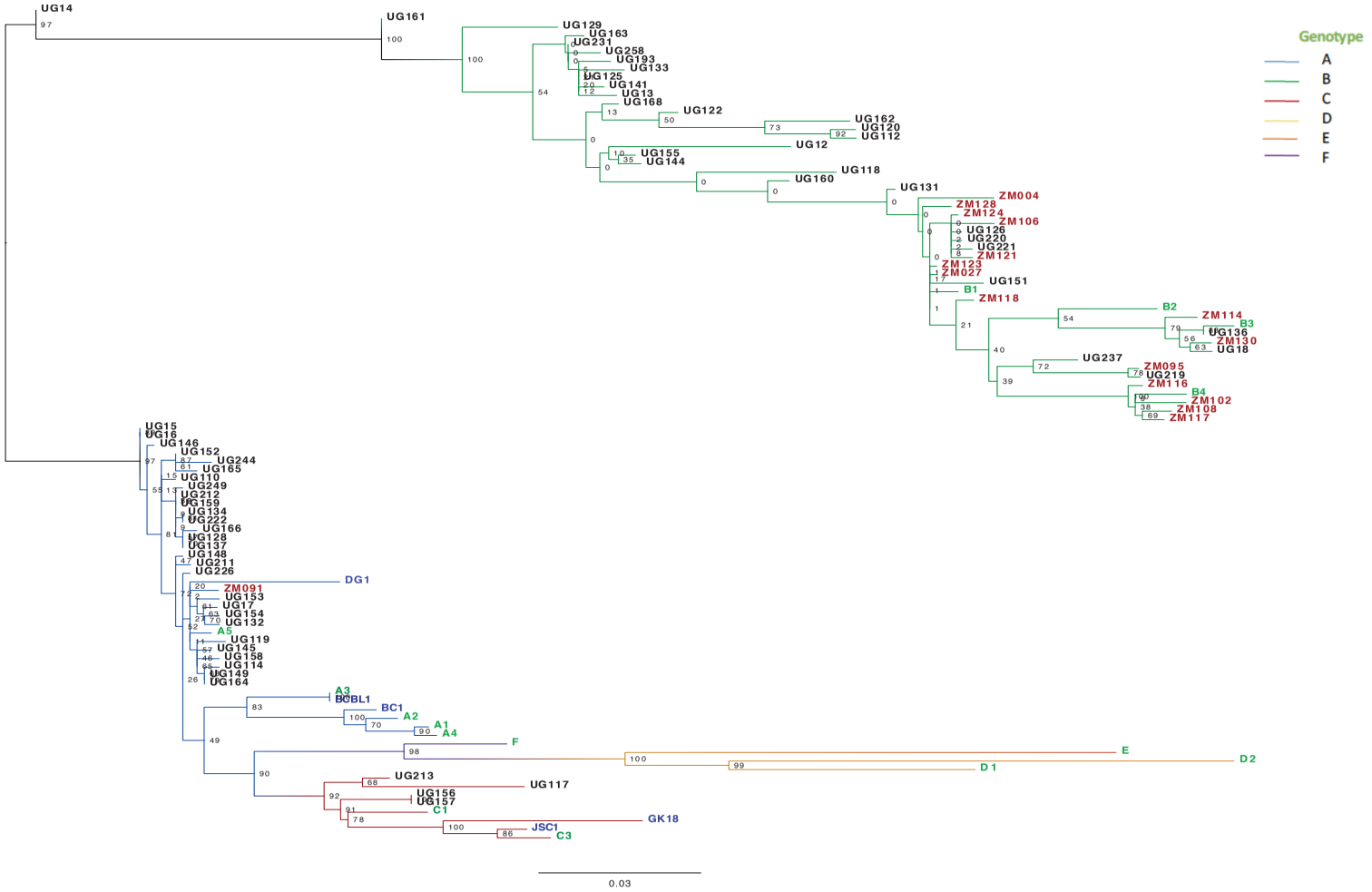

Fig. 5.12 KSHV K1 gene phylogeographic analysis of 83 samples. Midpoint-rooted maximum-likelihood tree of KSHV genomes generated with 1000 bootstrap replicates. Sample names are labelled in blue (Western), red (Zambian), black (Ugandan GPC) and green (representative K1 genotypes). The branches are coloured by genotype as labelled in the figure. Numbers on the nodes correspond to bootstrap values. The scale bar represents 0.03 nucleotide substitutions per site.

Table 5.4 Characteristics of Ugandan GPC Samples

Name	Village No.	Sex	Age	HIV status	Viral Load (Copies/ml)	Mean Depth	Mapped Reads (%)	K15 Type	K1 Genotype
UG110	12	1	40	0	2980	25X	23.3	P	A
UG13	12	2	16	0	16000	200X	29.7	P	B
UG12	13	2	53	1	33600	400X	40.3	P	B
UG14	13	1	59	0	1540	10X	4.4	P	B
UG15	13	1	85	0	7040	25X	21.4	M	A
UG16	13	2	69	0	2150	25X	13.9	M	A
UG17	13	1	22	0	784	10X	7.5	M	A
UG18	13	2	31	0	1960	10X	5.2	P	B
UG112	14	1	91	0	3930	10X	9.3	P	B
UG114	14	1	18	0	45200	25X	33.5	P	A
UG117	14	2	43	0	24700	25X	35.1	P	C
UG118	14	1	21	0	6000	25X	8.6	M	B
UG119	14	2	36	0	11900	25X	19.6	M	A
UG120	14	1	77	0	15500	25X	15.6	P	B
UG122	14	1	32	0	3390	10X	7.1	P	B
UG125	14	2	20	0	15300	25X	35.2	P	B
UG126	14	1	18	0	535000	1000X	70.8	P	B
UG128	15	1	21	0	67700	500X	65.2	P	A
UG129*	15	2	50	0	23400	25X	42.6	P	B
UG131*	15	2	52	0	93700	500X	62.2	M	B
UG132	15	2	46	0	18600	25X	11.7	P	A
UG133	15	2	25	0	26600	25X	35.7	P	B
UG134	15	2	19	0	5730	25X	12.5	P	A
UG136	15	1	79	0	4730	25X	15.6	P	B
UG137	15	1	30	0	13000	25X	22.5	P	A
UG141	15	1	23	0	24200	25X	34.4	P	B
UG144	15	1	33	0	1090	10X	5.2	P	B
UG145	15	1	59	0	7260	25X	21.8	M	A
UG146	15	2	42	0	20200	25X	26.1	P	A
UG148	15	2	17	0	1560	25X	10.7	P	A
UG149	15	2	86	0	53500	25X	40.6	P	A
UG151	15	1	79	0	5630	25X	11.4	P	B
UG152	15	1	19	0	3720	25X	1.7	P	A
UG153	15	1	62	0	19100	10X	16.6	P	A
UG154	15	1	39	0	9200	10X	13.0	P	A
UG155	16	2	29	0	6410	25X	30.3	P	B
UG156	16	2	31	1	104000	1000X	36.6	P	C

Name	Village No.	Sex	Age	HIV status	Viral Load (Copies/ml)	Mean Depth	Mapped Reads (%)	K15 Type	K1 Genotype
UG157	16	2	30	1	37000	1000X	34.9	M	C
UG158	16	2	37	0	9140	20X	1.9	P	A
UG159	16	2	56	0	8960	25X	15.6	P	A
UG160	16	2	40	0	8670	25X	19.8	M	B
UG161	16	1	60	0	7920	10X	4.6	M	B
UG162	16	1	22	0	9590	25X	30.2	P	B
UG163	16	1	30	0	22000	25X	23.9	P	B
UG164	16	1	21	0	135000	750X	76.8	P	A
UG165	16	1	72	0	15200	25X	15.2	P	A
UG166	16	1	18	0	9140	25X	24.9	P	A
UG193	16	2	20	0	1324	10X	2.8	P	B
UG220	16	1	38	0	4250	10X	4.3	P	B
UG168	17	2	50	1	24900	25X	26.5	P	B
UG211	17	1	34	0	2230	10X	3.4	P	A
UG212	17	1	20	0	31900	25X	16.9	P	A
UG213	17	2	67	0	3050	10X	4.4	P	C
UG219	17	2	42	0	16800	25X	11.0	M	B
UG221	18	2	77	0	3300	10X	5.3	P	B
UG222	18	1	22	0	19900	25X	10.0	P	A
UG226	18	1	51	1	8220	20X	1.5	P	A
UG231	19	2	50	0	2210	10X	8.0	P	B
UG237	19	1	75	0	11700	25X	18.0	M	B
UG244	19	2	43	0	2910	20X	5.3	P	A
UG249	19	1	19	0	4830	10X	8.1	P	A
UG258	19	2	33	0	2060	10X	6.3	P	B

[a] Sex: 1=Male, 2=Female
[b] HIV Status: 0=Negative, 1=Positive
*Belong to the same household
Coloured by village number

5.4 Discussion

Whole-genome sequence analyses of viruses are crucial to enhance our understanding of viral phenotypes, define virus population structure, transmission chains, elucidate variants under selection and explore relationships between genomic diversity and disease pathogenesis, which remains largely unknown for KSHV. In this study, I present for the first time 62 new wild-type genomes isolated from saliva of non-diseased adults, and performed a genomic variation and phylogenetic analyses in combination with 21 previously published genomes from Greece, USA and Zambia, thus representing the largest KSHV whole-genome study to date. In addition, this study presents a unique dataset with adults from eight neighbouring villages, as the majority of KSHV molecular epidemiology studies have focused on children and/or looked at transmission dynamics between mother-child or within hospitals[145,146,158,170,172,174,175,585].

Previous genetic analyses of whole-genomes generated from KS, PEL and KICS samples have provided invaluable insights into KSHV genomic architecture and viral epidemiology[138,560,563], however, they may not be representative of those found in the general population, particularly in KSHV endemic regions such as Uganda where oral transmission is the most likely route of infection. Studies using saliva pose a significant challenge given the virus is difficult to detect particularly in asymptomatic individuals unless they're shedding virus i.e. during the lytic stage of infection, and viral levels are much lower in saliva and blood compared to in tumour biopsies or cell lines[563]. With such low quantity viral DNA within a greater pool of complex human DNA recent studies have demonstrated that target enrichment technology is better suited for capturing viral genomes[584,586]. Therefore, here, I isolated KSHV DNA from saliva of asymptomatic carriers in Uganda, assessed the prevalence of viral shedding by qPCR and successfully sequenced multiple whole-genomes from a population-based cohort using a target enrichment approach with Illumina paired-end sequencing technology. qPCR analysis of DNA isolated from 746 adults showed that in the GPC ~33% (244) of individuals are

actively shedding KSHV (i.e. detection of KSHV genome in saliva), with inter-individual variability in viral loads from 1.5 to 5.35x10^5 copies/ml which is within range of previous findings[146,316,587-590]. Following whole-genome sequencing of all 244 samples, I sought to investigate whether viral load influenced sequencing quality and found that viral load was highly correlated with achieving good sequencing depth and percentage of KSHV mapped reads, this is reflected by the 62 samples (25%) with viral loads of >10^4 copies/ml, which had a mean sequencing depth of at least 10x of coverage >90% across the genome (Fig. 5.3 and Table 5.4). In addition, viral load was also correlated with the number of reads that mapped to the KSHV genome (Fig. 5.3). This is an important observation for future studies that wish to sequence KSHV DNA directly from clinical samples such as saliva, these data suggest one should focus on samples with a viral load >10^4 copies/ml. As these 62 samples had good sequencing coverage, I retained them for further genetic variation and phylogenetic analyses with an additional 21 published whole-genomes from diseased individuals from Greece, USA and Zambia.

Multiple sequence alignment of the 83 genomes showed high levels of sequence conservation, with a higher level of genetic variation in the 5' and 3' genome ends, corresponding to the K1 and K15 genes, respectively (Fig. 5.5). A low level of genetic variation was found across the central region of the genome, consistent with previous findings[563,564]. This was confirmed by analysis of SNPs across the coding region of the genome (Fig. 5.6). While it has been found that the K1 gene has been evolving under strong host selective pressure in association with cytotoxic T-lymphocyte recognition[145,591,592], few data exist for other genes under selection driven by the host. Along with the K1 and K15 genes, other genes of interest from the SNP analysis across the KSHV coding region are ORF73 (encoding LANA), K12 (encoding Kaposin), K4.2, K8.1 and v-IRF2 in the central region (Fig. 5.7), which have a higher proportion of non-synonymous polymorphisms suggesting that they're under positive selection. It is interesting that these genes are all encoding immunomodulatory proteins[593] and thus this selective pressure may facilitate KSHV's evasion of the immune response. All the

genes above, except ORF73 were identified with a high number of polymorphisms in the Zambian study, highlighting the advantage of having more genomes to comprehensively identify genes under selection.

Similarly to the well-established Type 1/ Type 2 classification and whole-genome clustering of EBV based on the divergent EBNA-3 allele[586,594], for KSHV, the Type P / Type M based on variation in the K15 gene remains the major form of variation correlating with whole-genome clustering (Fig. 5.8). KSHV Types based on the multiple sequence alignment of the K15 gene confirmed the Type P/M split observed in the whole-genome tree with 50 of the GPC samples belonging to the Type P and 12 to the Type M (Fig. 5.11). Whole-genome clustering also showed similarity in genomes irrespective of derivation from different clinical presentations i.e. asymptomatic vs diseased or source of sample isolation i.e. saliva vs biopsy/cell line. Distinct phylogenetic clustering was observed between the African samples (Uganda and Zambia) and the Western samples, as previously observed in the Zambian KS whole-genome study[564]. Giving evidence to this, the Type P/M clustering was lost following removal of the K15 gene from the genome alignment, however, geographical clustering of samples still remained (Fig. 5.10). The addition of more samples to the Zambian study refined the phylogenetic relationships in the tree and two distinct sub-clades were observed for types P and M, potentially arising as a result of variation in the central region which would need to be further resolved using ancestral reconstruction methods.

Geographic association of K1 genotypes has been reported by several studies globally. Hayward hypothesised that KSHV is an old human virus and the distribution of its' subtypes arose as a result of ancient human migration >100,000 years ago out of Africa[144,192]. The A (particularly A1-A4) and C subtypes are found to predominate in Europe, USA, Australia, the Middle East and Asia; the A5 and B genotypes are typical for populations of African descent and more recently the F genotype was identified in Ugandans; the D and E genotypes are more common in the Pacific Islands and Brazilian

Amerindians, respectively[186,187,567,570,571,576,595-603]. Genotypic analysis of the 62 samples based on the K1 gene revealed a heterogeneous distribution of subtypes throughout the villages in the GPC (Fig. 5.12), consistent with previous studies the B and A5 subtypes predominate at 48.4% and 45.1% respectively with a few samples belonging to the C1 (6.5%) genotype. Interestingly, while the A genotypes all clustered with A5, the B genotypes were heterogenous with more subgroups compared to the A, suggesting the K1 genotyping is not fully capturing variation of these genomes. No subgroups based on villages were observed. Conflicting data exists on whether different genotypes are attributed to pathogenic or tumourigenic properties of KSHV, a very recent study conducted in a South African population reported that the A5 genotype is associated with extensive disease in AIDS-KS[604] and a Zambian study found it to be associated with childhood KS[596], however, an earlier study showed that the A5 genotype is more prevalent in African children than mothers and thus represents more efficient viral transmission[145]. Thus, genotypic diversity and its relation to pathogenesis remains unclear, however, it might be further resolved by taking the whole genome into account.

In summary, in this study I assessed the prevalence of KSHV shedding (i.e. detection of KSHV genome in saliva) in the Ugandan GPC and identified a viral load threshold for the successful sequencing of KSHV whole-genomes isolated from clinical samples by target enrichment. I present the largest KSHV whole-genome analyses to date with 62 new wild-type whole-genomes from Uganda which are the first to be generated from saliva of asymptomatic individuals and extend the analysis conducted recently with 16 Zambian KS genomes and including 5 previously published genomes from Greece and USA. This study confirmed the presence of high level variation at the 5' and 3' ends of the genome that drives major variation between KSHV strains, in addition to low level variation in the central conserved region, with genes involved in modulating host response and under selective pressure contributing to distinct phylogenetic clustering between Western and African samples. The heterogenous distribution of KSHV strains, with a variety of genotypes observed throughout all villages, suggesting cross-ethnic and cross-village

transmission is not surprising given how well connected the villages are, with relaxed administrative boundaries enabling ease of access and movement between the villages[388]. Two adults who belonged to the same family/household had different K15 genotypes (Table 5.4), suggesting transmission is not horizontal between these adults, this is consistent with previous findings that KSHV is predominantly transmitted vertically (i.e. from mother-child)[145,174,585,605]. However, to reliably identify transmission patterns in this study more familial and household samples across all age groups would be required. It is also worth noting that as result of mapping the 62 new genomes to the GK18 reference sequence a bias may exist, for example, missing out insertions and deletions, and thus, underestimating genetic diversity.

In conclusion, despite extensive genotypic characterization worldwide, how selection pressure is driving genomic variation and whether specific genotypes are linked to pathogenesis and disease remains unclear and thus will require further investigation. From this study and the previous study of Zambian KS patients, while a high level of similarity exists between genomes, it is evident that K1 and K15 genotyping insufficiently capture genetic variation and whole-genome variation is greater than previously appreciated. The addition of genomes from Uganda to the Zambian dataset identified additional genes under selection and refined phylogenetic relationships between genomes. Therefore, whole-genome sequences from other parts of the world providing a more comprehensive global dataset would be essential to substantiate these findings. Viral characterisation based on whole-genome diversity needs to be considered coupled with a revision of the nomenclature.