J Mol Evol (2004) 58:239–251
DOI: 10.1007/s00239-003-2543-1

JOURNAL OF MOLECULAR EVOLUTION

Model-Based Inference of Recombination Hotspots in a Highly, Variable Oncogene

G. Greenspan,[1] D. Geiger,[1] F. Gotch,[2] M. Bower,[2] S. Patterson,[2] M. Nelson,[2] B. Gazzard,[2] J. Stebbing[2]

[1] Computer Science Department, Technion, Technion City, Haifa 32000, Israel
[2] Department of Immunology, Division of Investigative Science, Faculty of Medicine, Imperial College of Science, Technology and Medicine, The Chelsea and Westminster Hospital, London, United Kingdom

Received: 14 August 2003 / Accepted: 30 August 2003

Abstract. An emergent problem in the study of pathogen evolution is our ability to determine the extent to which their rapidly evolving genomes recombine. Such information is necessary and essential for locating pathogenicity loci using association studies, and it also directs future screening, therapeutic and vaccination strategies. Recombination also complicates the use of phylogenetic approaches to infer evolutionary parameters including selection pressures. Reliable methods that identify the presence of regions of recombination are therefore vital. We illustrate the use of an integrated model-based approach to inferring recombination structure using all available sequences of the highly variable, transforming Kaposi's sarcoma-associated herpesviral gene, ORF-K1. This technique learns the parameters of a statistical model that takes recombination hotspots, population genetic effects, and variable rates of mutation into account. As there are no known mechanisms to explain the high mutation rate in this DNA viral gene, recombination may account for some of the variability observed. We infer recombination hotspots in conserved sites such as the tyrosine kinase signaling motif, referred to here as recombination drift, as well as in nonconserved sites, a process described as recombination shift.

Correspondence to: Dr. Justin Stebbing, Department of Immunology, The Chelsea and Westminster Hospital, 369 Fulham Road, London, SW109 NH, UK; *email:* j.stebbing@imperial.ac.uk

Key words: Recombination — Hotspot — Bayesian model — K1 — Oncogene

Introduction

Recombination occurs at substantial frequencies in the natural populations of many species and appears to play a significant role in determining selective pressures (Fisher 1958; Maynard-Smith 1982; Rouzine et al. 2003). Areas of high recombination allow adjacent genomic regions to have different evolutionary histories, although the precise loci of recombination in several species, specifically in viruses, remain unknown (Awadalla 2003).

Viruses, including HIV and herpesviruses, in which persistent infections are a well-known feature, provide an opportunity to test evolutionary theories by comparing model predictions against previously obtained data. Conversely, evolutionary theories can also be used to predict future properties of pathogenic populations providing valuable insights into coexistence and coevolution with their hosts, the so-called molecular arms race (Gilbert et al. 1998; Holmes 2001; Holub 2001; Stebbing and Gazzard 2003b). Forces that affect genetic variability within a species include the systematic pressures of natural selection and migration (Perrin et al. 2003; Rouzine et al. 2003). In situations where these are the only pressures, as may be the case in very large populations, deterministic models suffice.

As evolution occurs in populations of finite size, random genetic drift adds a stochastic element by causing the disappearance of genetic variants, a situation most marked when the population size is small (Kimura 1968, 1969, 1976a, b; Kimura and Ota 1971). The balance between these has been shown to result in an accumulation of deleterious mutations via drift, resulting in cycles in which the best-fit genotypes in a population are lost, events known as Muller's (1964) ratchet.

Recombination, on the other hand, allows species to recreate fit genotypes following the accumulation of mutations, thus resetting the ratchet (Maynard-Smith 1982; Rouzine et al. 2003). As recombination appears to be a critical feature in many viruses studied (in both their zoonotic and human hosts (Hahn et al. 2000), it appears reasonable to suggest that it has an active role in their life history and fitness (Guttman and Dykhuizen 1994; Robertson et al. 1995; Conway et al. 1999; Ochman et al. 2000; Twiddy and Holmes 2003). Indeed, recombination events can explain dynamics of endemicity and pathogenicity and predict the development of advantageous or deleterious mutations conferring drug resistance, virulence and immune evasion, and the spread of such mutations between genetically different hosts and populations (Stebbing and Gazzard 2002, 2003a). Recombination is necessary for genetic mapping to locate genes that underlie important phenotypes, but its presence often complicates phylogenetic reconstruction and methods used to infer population parameters (Schierup and Hein 2000a,b; Feil et al. 2001). However, while almost all organisms engage in some form of recombination, our understanding of why it occurs and how it is maintained remains controversial.

Members of the *Herpesviridae* family are important human and animal pathogens and have the largest DNA genomes among known mammalian viruses, with up to 200 potential open reading frames (ORFs). Functional characterization of these genes, accomplished by generating virus mutants and investigating resulting changes in phenotype, is important for understanding molecular aspects of herpesvirus replication and pathogenesis, and may also provide a basis for the rational development of new vaccines and chemotherapeutics (Epstein 2001; Gaschen et al. 2002; Ho and Huang 2002).

Kaposi's sarcoma-associated herpesvirus (KSHV; also known as human herpesvirus-8) is a γ-2-herpesvirus related to three other tumorigenic viruses: herpesvirus samirii (HVS), Epstein–Barr virus (EBV), and murine γ-herpesvirus-68 (Russo et al. 1996). KSHV is implicated in the pathogenesis of all epidemiologic forms of Kaposi's sarcoma (classic, African endemic, posttransplantation, and acquired immunodeficiency syndrome [AIDS]-associated) (Stebbing et al. 2003b, c). At the far left-hand end of its 140-kb double-stranded DNA episome lies a unique gene encoding a 46-kDa transmembrane type I glycoprotein, K1, containing a sequence that functions as an immunoreceptor tyrosine-based activation motif (ITAM) (Russo et al. 1996; Lee et al. 1998; Damania et al. 2000; Lagunoff et al. 2001; Zong et al. 2002). K1, expressed predominantly in the lytic viral life cycle, appears able to couple extracellular signals to multiple intracellular signalling pathways in response to ligand–receptor interactions, which in turn leads to cellular responses, including proliferation, differentiation, and death (Samaniego et al. 2000; Lagunoff et al. 2001; Bowser et al. 2002). K1 is also able to transduce signals in the absence of exogenous cross-linking ligands, and can orchestrate the expression of an array of transcription factors involved in cellular activation that may ultimately lead to growth dysregulation. As such, K1 appears capable *in vitro* and *in vivo* of inducing cellular transformation (Lee et al. 1998; Prakash et al. 2002).

Although the great majority of the KSHV genome is conserved, K1 is variable (Nicholas et al. 1998). The variability of K1, with a massive preponderance of amino acid altering (nonsynonymous) mutations, has been known for some time (Hayward 1999; Zong et al. 1999; McGeoch 2001). More recently, it has been shown that at an individual codon level, specific sites in K1 appear to undergo a considerably greater positive selective pressure than sites in other highly variable mammalian or viral genes, such as the 58 codons in exons 2 and 3 that comprise the antigen recognition site of major histocompatibility complex (MHC) class I and the 34 codons in the V3 loop of HIV-1 *env* (irrespective of evolutionary time scale) (Stebbing et al. 2003a). The presence of clusters of MHC class I-restricted epitopes within K1 have been cited as putative evidence that pressure from CD8-positive cytotoxic T lymphocytes may help drive this gene's extreme variability and positive selection (Stebbing et al. 2003a). However, in the absence of an error prone polymerase (as for HIV [Munoz et al. 1993]), there are no known mechanisms to explain this. DNA–virus DNA polymerases do not perform the template switches required to generate such diversity (An and Telesnitsky 2002). Furthermore, K1 does not appear to change over time within an individual, nor does it differ between different tumor sites within the same patient, quite unlike the situation observed with retroviruses (Stebbing et al. 2001; Walker and Korber 2001).

We therefore assessed the potential role of hotspots of recombination in generating the variability observed in K1 using a model-based inference technique. This technique has been previously validated by its application to the haplotype resolution problem for high density areas of the human chromosome

21 (extending over 21.7 Mb [Patil et al. 2001]) and the angiotensin converting enzyme gene on chromosome 17 (24 kb [Rieder et al. 1999]), where it obtained error rates between 3 and 200 times lower than previously published methods (Greenspan and Geiger 2003).

There are four key differences between our model-based approach and traditional phylogenetic tree construction. First, by inferring specific recombination points, our model divides sequences into contiguous stretches, examining the relationships separately within each. This is justified by the observation that a region of high recombination will result in the areas either side having different evolutionary histories. Second, we explicitly allow for the presence of mutation hotspots, inferring their presence as part of a model. This is consistent with the observation that mutation occurs in an uneven fashion within the K1 gene and appears clustered in two areas, termed variable regions 1 and 2 (VR1 and VR2). Third, within each inferred stretch, we do not seek to create a complete family tree, accepting instead that distant relationships between sequences are difficult to accurately ascertain and recover. This approach is justified by population genetic considerations which suggest that bottlenecks, genetic drift, and selection pressures will narrow a populations genepool losing the vast majority of ancient strands. This process leaves behind a few groups, each of which contains minor variations on a consensus sequence. The fourth and final difference is that within each such group, we do not attempt to infer relationships between the sequences, opting instead to consider them all as offspring of a single founding ancestor. As before, this is justified by population genetics—if a viral population grows rapidly from a few founders then the most recent common ancestor (MRCA) of any two contemporary sequences is likely to be very close to those founders.

Thus, whereas traditional phylogenetic analysis attempts to create a complete tree topology to relate the observed sequences, we infer a set of disconnected stars, each of which centers around a consensus sequence which may itself remain unobserved. By simplifying the phylogenetic model in this way, we reduce the calculation time for each candidate assignment of recombination hotspots. Furthermore, if our assumptions accurately reflect the underlying population processes, we will produce more accurate inferences regarding mutation rates and selection pressures.

An inferred model contains a lull description of the variation structure of a set of observed sequences. For our purposes, the most important parameters of the model are (a) the location of the recombination hotspots, (b) for each stretch between hotspots, the number of inferred clades, and (c) cumulative mutation rates for each site. Here, a clade is defined as a group of similar sequences, within a single nonrecombinant stretch. The cumulative mutation rates represent the probability that the allele observed in a sequence is different from the allele in the consensus sequence for its inferred clade. It should be noted that the full model also describes the linkage dependencies between stretches which are separated by recombination hotspots but we will not be using that information here. However, due to error rates a magnitude lower than those previously observed. By examining an inferred sample of suitable models, we identify recombination hotspots within conserved and nonconserved areas of a DNA viral oncogene (K1), the most positively selected mammalian or viral gene so far identified, thus postulating one mechanism by which it generates its remarkable variability. The hotspots in the conserved regions are postulated to generate long–term variability and evolutionary persistence (recombination drift), whereas those within the variable regions produce immediate changes, probably altering antigenicity in specific populations, a phenomenon denoted recombination shift.

Methods

Sequences and Alignment

Nucleotide sequences encoding the KSHV ORF-K1 were obtained from NCBI (Table 1); these were derived from nested PCR reactionts (Zong et al. 1999; Cook et al. 1999, 2002; Meng et al. 1999, 2001; Lampinen et al. 2000; Lacoste et al. 2000a, b; Biggar et al. 2000; Zhang et al. 2000, 2001). All available 269 K1 sequences were aligned successfully using CLUSTAL X, version 1.6. Unlike previous analyses of K1 variability and evolution (Stebbing et al. 2003a), no sequences were excluded and K1 was not divided into strains (approximately 30% of the sequences could not be discreetly placed in defined strains A–D). No two K1 sequences were identical and all K1 sequences were derived from different hosts. Treeview, version 1.5, was used to create a radial phenogram (Fig. 1).

The 269 K1 sequences were converted to amino acids and these were used to create a consensus sequence at the website http://prodes.toulouse.inra.fr/multalin/html (Corpet 1988), which was then in turn used to predict a likely secondary structure for this protein at the website http://bioinf.cs.ucl.ac.uk/psipred (Jones 1999; McGuffin et al. 2000; Marsden et al. 2002) (Fig. 1). We also examined the relationship between sites of recombination and phosphorylation (http://www.cbs.dtu.dk/services/NetPhos), N- and O-glycosylation (http://www.cbs.dtu.dk/services/NetNGlyc/ or NetOGlyc/), myristoylation (http://mendel.imp.univie.ac.at/myristate), sulfination (http://us.expasy.org/tools/sulfinator/), cleavage (http://www.cbs.dtu.dk/services/SignalP/), and changes in Kyte and Dolittle hydropathy values (http/us.expasy.org/cgi-bin/protscale.pl) (Hansen et al. 1995; Blom et al. 1999; Monigatti et al. 2002; Maurer-Stroh et al. 2002a, b).

Model-Based Inference of Recombination Hotspots

To model the complex multivariable distribution underlying viral sequences, a *Bayesian Network* was used. Such networks are useful, as they allow distributions to be represented, learned, and queried

Table 1. Accession numbers of ORF-K1 sequences (n = 269) obtained from http://www.ncbi.nlm.nih.gov

KSHV ORF-K1 accession number												
11138417	13377083	15193049	15281822	17980811	4589184	4589226	5326891	9454396	9587246	9886771	9886810	9886850
11138419	13377085	15193051	15281824	17980813	4589186	4589228	5670271	9454398	9587248	9886773	9886812	9886852
11192010	13377087	15281784	17980773	2065556	4589188	4589230	5833939	9454400	9587250	9886775	9886814	9886854
13377047	13377089	15281786	17980775	3047216	4589190	4589232	5833941	9454402	9587252	9886777	9886816	9886855
13377049	13377091	15281788	17980777	4589150	4589192	4589234	5833943	9454404	9587254	9886779	9886818	9886857
13377051	13377093	15281790	17980779	4589152	4589194	4589236	5833945	9454406	9587256	9886781	9886820	9886859
13377053	13377095	15281792	17980781	4589154	4589196	4589238	6636400	9587216	9587258	9886783	9886822	9886861
13377055	13377097	15281794	17980783	4589156	4589198	4589240	6636402	9587218	9587260	9886785	9886824	9886862
13377057	15193026	15281796	17980785	4589158	4589200	4589242	7274375	9587220	9587262	9886787	9886826	9886864
13377059	15193027	15281798	17980787	4589160	4589202	4589244	7274377	9587222	9587264	9886788	9886828	9886866
13377061	15193028	15281800	17980789	4589162	4589204	4836703	7861771	9587224	9587266	9886790	9886830	9886868
13377063	15193031	15281802	17980791	4589164	4589206	4836705	7861773	9587226	9587268	9886792	9886832	9886870
13377065	15193033	15281804	17980793	4589166	4589208	4836707	7861775	9587228	9587270	9886794	9886834	9886872
13377067	15193035	15281806	17980795	4589168	4589210	4836709	7861777	9587230	9587272	9886796	9886836	9886874
13377069	15193037	15281808	17980797	4589170	4589212	4836711	7861779	9587232	9587274	9886798	9886838	9886876
13377071	15193039	15281810	17980799	4589172	4589214	4836713	7861781	9587234	9587276	9886800	9886840	9886878
13377073	15193040	15281812	17980801	4589174	4589216	4836715	7861783	9587236	9587278	9886802	9886842	9886880
13377075	15193041	15281814	17980803	4589176	4589218	5326879	8926144	9587238	9587280	9886804	9886844	9886882
13377077	15193044	15281816	17980805	4589178	4589220	5326881	9454390	9587240	9886767	9886806	9886846	9886884
13377079	15193045	15281818	17980807	4589180	4589222	5326885	9454392	9587242	9886769	9886808	9886848	9886886
13377081	15193046	15281820	17980809	4589182	4589224	5326887	9454394	9587244				

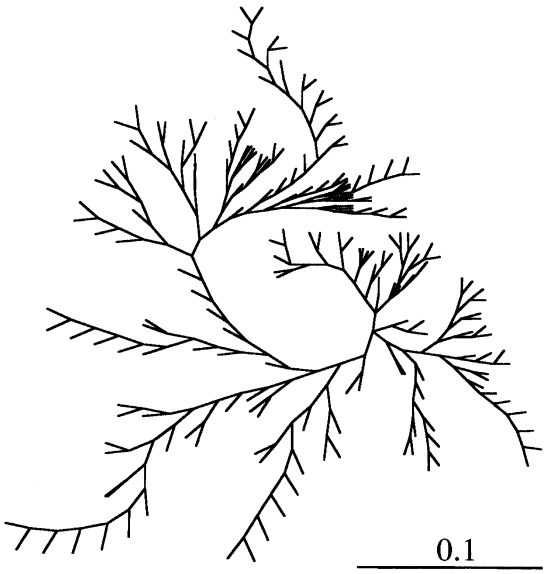

Fig. 1. A radial phenogram of 269 KSHV ORF-K1 sequences constructed using alignment (CLUSTAL X, version 1.6) and Treeview (version 1.5) programs. Phylogenetic distance between sequences is marked.

efficiently by making the independence relationships between variables explicit. Each Network has a natural graphical representation, in which a circle corresponds to a variable and the distribution for each variable is conditional only upon those variables that point to it (Pearl 1988; Jensen 1996). The probability of an assignment to the variables in a Bayesian Network can be calculated efficiently by bucket variable elimination (Dechter August 1–4, 1996). In addition, parameters for the distributions can be inferred from observed data sets by the expectation maximization (EM) algorithm, which are used at many stages during the search for a model to fit observations (Lauritzen 1995).

An example of the model used is shown by the Bayesian Network in Fig. 2. It consists of a variable C_k for each block k and two random variables A_j and H_j for each site j. A partition by recombination hotspots of the sites into blocks is defined by the groups of variables A_j pointed to by each C_k in the Bayesian Network. For example, the model in the diagram places hotspots between adjacent single nucleotide polymorphism (SNP) pairs 3–4 and 5–6.

A model of the variation in observed sequences is given by the distributions of these variables, while the etiology of a particular sequence is specified by an assignment to them. For each block k, the variable C_k represents the index of the clade for block k to which a viral sequence belongs. The Markov chain between variables C_k reflects the assumption that the probability of a viral sequence belonging to a particular clade for block k depends only on its clade for block $k-1$. Simulations and other analyses which demonstrate that this assumption is accurate to a high degree have been performed.

Given a value for C_k specifying a sequence's clade for block k, variables A_j which descend from C_k specify the consensus sequence of that clade. Since there is no variation within each consensus sequence, the value of A_j is fixed by C_k, thus the double border in the diagram (Fig. 2). Here, we begin to see the power of the Bayesian Network representation—although there is clearly a strong dependency between the variables A_j within the same block, this can be expressed solely in terms of their shared dependency on C_k. Given a value for A_j specifying the allele at site j of the consensus sequence for the appropriate clade, variable H_j specifies the allele actually observed at that site in a viral sequence. The distribution of H_j, conditioned on the value of A_j, is given by the cumulative mutation rates inferred for site j. Unlike the rest of the Bayesian Network, variables H_j are actually observed, as denoted by the small dot within each.

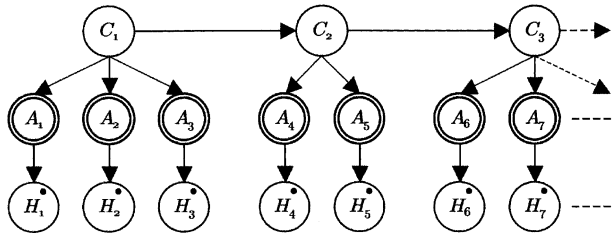

Fig. 2. Example Bayesian network used here to infer recombination hotspots.

The goal of this inference technique is to infer a sample of block partitions and parameters of this Bayesian Network which provide an accurate and concise description of the set of observed viral sequences. Although ideally one would like to infer the single correct model for our data, the uncertainty which is inevitable in these forms of inference is best addressed by generating a wider sample. Therefore, a search occurs for models according to the minimum description length (MDL) criterion, which simultaneously seeks to minimize the model complexity and maximize the probability of the observed data under that model (Rissanen 1978). Formally, if DL(M) bits are required to represent a model M for data D, then the description length for model M is DL(M)-$\log_2$ Pr(D|M). While Pr(D|M) is defined precisely by the Bayesian Network, DL(M) is highly dependent on our representation scheme. A scheme was chosen which represents the model parameters efficiently, based on known results in the data compression field (Shannon 1948; Rissanen 1983).

The search procedure is complex, since the space of possible models for a given data set is vast. To begin with, for l sites, there are 2^{l-1} different partitions into blocks, but our search must also cover different numbers of clades for each block, possible consensus sequences for each clade, and a wide range of cumulative mutation rates. Each aspect of the search is addressed differently, using a combination of local minimum search techniques and the EM algorithm (Greenspan and Geiger 2003). While developing our search algorithm, we confirmed that it is able to accurately reproduce the parameters of an artificial model from which simulated data were generated, given a few dozen haplotypes or more. Recombination hotspots in the artificial model were consistently inferred to within one or two SNPs, while false positives did not appear. Since the data set for this paper contains 269 haplotype sequences, we are confident in the accuracy of the inferences made.

The only parameter required by the model search is the maximum cumulative mutation rate, which constrains the distributions for variables H_j. In general, as we allow more mutation, less recombination will be inferred. We chose to use three maximum mutation rates, 0.5, 0.1, and 0.01, to allow different degrees of variation within each clade. Values greater than 0.5 are meaningless in the context of our technique, since we have no basis on which to infer that a particular allele belongs in a consensus sequence if a different allele usually appears in its place.

For each of the three maximum mutation rates, we chose to sample 100 models. Each iteration of the sampling algorithm took up to 3 h of processing time on a 2-GHz Pentium Xeon workstation, leading to a total running time of several weeks. The mean values were calculated for the summary statistics of interest separately for each of the three sets of models. Each model specifies the full allele-to-allele cumulative mutation matrix for each site, which is converted to an overall cumulative rate of mutation for that site using the clade distribution. We also report the average of the cumulative mutation matrix over all the sites, reflecting different mean substitution, insertion and deletion rates for the different nucleotides.

Fig. 3. Predicted amino acid secondary structure ORF-K1, where a line represents a strand, an arrow a helix, and a cylinder a coil. Corresponding nucleotide base pair positions are marked for the start and end of variable regions 1 and 2 (VR1 and VR2) and the immunoreceptor tyrosine kinase activation motif (ITAM).

To demonstrate the potential of this modeling approach, it was applied to the high-density haplotype resolution problem. The accuracy of haplotype resolution based on our model for several regions in chromosome 21 (Patil et al. 2001) was compared against that for four other previously published methods (see supplementary information): (i) Clark's (1990) algorithm, (ii) a variation of the EM algorithm (Excoffier and Slatkin 1995; Long et al. 1995), (iii) the PHASE algorithm (Stephens et al. 2001), and (iv) a beta version of the HAPLOTYPER algorithm (Niu et al. 2002). We obtained error rates which were 3–200 times lower than those observed for the above methods, suggesting that our model accurately captures the effects of uneven recombination and mutation.

Results

Sequences

Figure 1 shows the phylogenetic relationship between 269 different K1 sequences (Table 1), each of which was sequenced from an individual host (or cell line derived from such hosts). Traditionally, it has been considered that KSHV can be subdivided into strains according to the K1 sequence, which is thought in turn to correspond to geographical origins of the virus.

The KHSV A strain is found in northern Europe and America, the B strain (thought to be the most ancient) is from Africa and the C strain, often associated with classical KS, is found in Mediterranean countries (Cook et al. 1999). A KSHV D strain containing nucleotide insertions has also been recently described (mainly from South America and the Pacific Islands) (Zong et al. 1999, 2002; Meng et al. 1999; Poole et al. 1999; Biggar et al. 2000). The evolution and changes in these strains are thought to reflect patterns of migration commencing in Africa (Hayward 1999; Stebbing et al. 2003a).

However, Fig. 1 suggests that many sequences cluster between previously recognized strains and that the depicted phylogenetic distance between different strains is often closer than between two sequences of the same strain. This may reflect recent events where travelers contribute to the spread of viral diversity worldwide, an important contributing factor being world migration of rural populations due to poverty, famine, and wars (Quinn 1994; Malim and Emerman 2001; Perrin et al. 2003).

Figure 3 represents a secondary structural prediction from a consensus amino acid sequence created from the nucleotides. The variable regions (VR1 and VR2) are repeated strand–helix–strand motifs, while the ITAM and transmembrane domains are coils. Consensus sequences from K1 derived from each KSHV strain show no significant differences in their structure in spite of amino acid variability.

No putative myristoylation sites were detected in K1. The potential sulfination, cleavage, and O- and N-glycosylation sites did not occur at inferred recombination hotspots (see below), and changes in hydropathy values at recombination hotspots were not unique in comparison to other sites. Two serine phosphorylation sites (scores $\geq$0.9) were located at hotspots, however, K1 is heavily phosphorylated (at least 23 residues have a phosphorylation potential of >0.5). Here, as for Figs. 1 and 3, we are using consensus sequences based on the multiple alignment and not the model-based inferences described in the Methods.

Hotspot Strength

For each of the three maximum mutation rates and for each individual base pair, Fig. 4 demonstrates the proportion of the 100 sampled models which placed a recombination hotspot at that site. The midlines on each graph represent the point at which 50% of the inferred models have a hotspot, so any peak that reaches or is close to this point is a likely position of a recombination hotspot. High areas which are more spread out suggest a region in which there is a hotspot whose exact location is unclear. Low areas near the zero-line represent regions in which it appears that no recombination hotspots are present. As expected, the less mutation allowed in the model, the more recombination is inferred (top line, Fig. 4; maximum mutation rate, 0.01).

For the two higher mutation rates (0.1 and 0.5), codons 212 (base pair 616) and 230 (base pair 690) were identified as recombination hotspots. Base pair 616 is located 27 nucelotides upstream of the second variable region (VR2) of K1. While VR2 is an area characterized by insertions, deletions, and nonsynonymous mutations, base pair 616 is in a relatively conserved area of this gene. Basepair 690 is located midway within VR2 itself. At a maximum mutation rate of 0.1, a further site was identified at base pair 606 in the most conserved area of K1 between the

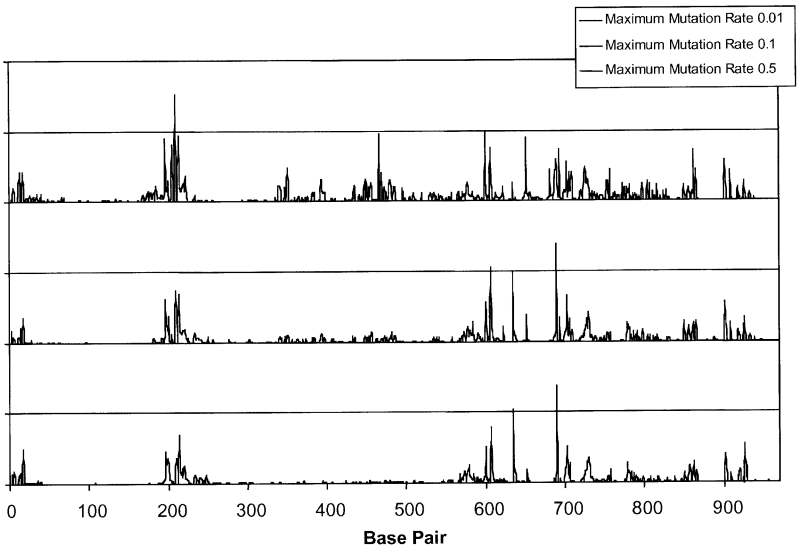

Fig. 4. Hotspot strength. One hundred models were sampled for each maximum mutation rate value. The height for each base pair shows the proportion of those models which had a recombination hotspot placed at that point.

variable regions. At these mutation rates, no sites in or around VR1 were identified as recombination hotspots in spite of the known positive selection occurring here (>85% of nucleotide substitutions in this region lead to amino acid changes). At the lowest mutation rate (0.01), base pair 606 (codon position 202) was also identified as a likely recombination hotspot. The highest likelihood of recombination was found with a maximum mutation rate of 0.01 at base pair 210 (codon 70), located within the hypervariable area of the first variable region (VR1). VR1 was not flanked by recombination hotspots at any mutation rate.

As groups of small peaks in one area suggest a likely hotspot without an exact site, our data also provide evidence of recombination occurring near the start codon, within VR1, following VR2 and, interestingly, within the cytoplasmic ITAM motif (base pairs 909 to 960). There were no recombination hotspots within the transmembrane region (Fig. 3).

Inferred Clades

Clades of shared ancestry can be visible because of bottlenecks and genetic drift, both of which serve to reduce the variability within each region of low recombination. Bottlenecks occur when a population is descended from a small group of individuals, for example, if founded by some pioneers from an original population. Genetic drift occurs particularly in small populations, when chance significantly alters allele frequencies, often causing some variants to disappear completely. Although we do not know when KSHV or ORF-K1 first appeared or underwent significant reductions in variation, we can assume that such formative processes have taken place in the past.

Full ancestral sequences among the K1 pool are unknown, and it is unlikely that any are present since they can be expected to have recombined out of recognition. However, for each block in between recombination hotspots, our model infers the number of ancestral sequences that appear to be present for that particular block. This is converted to a value for each base pair, by endowing each base pair with the same number of ancestors as the block which contains it. For each of the three mutation rates described above, Fig. 5 shows the number of ancestors inferred for each base pair, averaged over the 100 samples. As for recombination, the less mutation inferred, the fewer ancestors are required to explain the observations.

A reduction in clades within a certain region may reflect greater selection in that region. We observe this reduction at the 5′ end of K1, around base pairs 469 and 607 (between VR1 and VR2) and in a stretch of nucleotides between base pair 787 and base pair 909. Selective pressures here are negative, as they lead to conservation of nucleotides as opposed to variability.

Results were similar at all three mutation rates with variation around a mean of four ancestors. The highest number of putative ancestors ($n = 6$ to 7) are located in the hypervariable area of VR1. Other peaks are located within and beyond VR2 and the lowest number ($n = 3$) are located in the relatively conserved area between VR1 and VR2 and near the start codon.

Cumulative Mutation Rate

For each of the three mutation rates described above and for each pair, the average cumulative mutation rate over the 100 samples models is shown (Fig. 6).

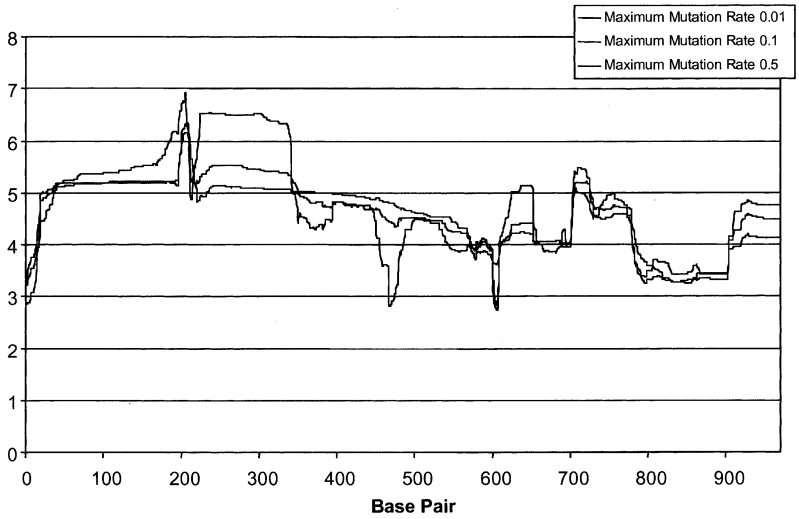

Fig. 5. One hundred models were sampled for each maximum mutation rate value. This graph shows the mean number (over the sampled models) of ancestral sequences inferred for each base pair's enclosing block.

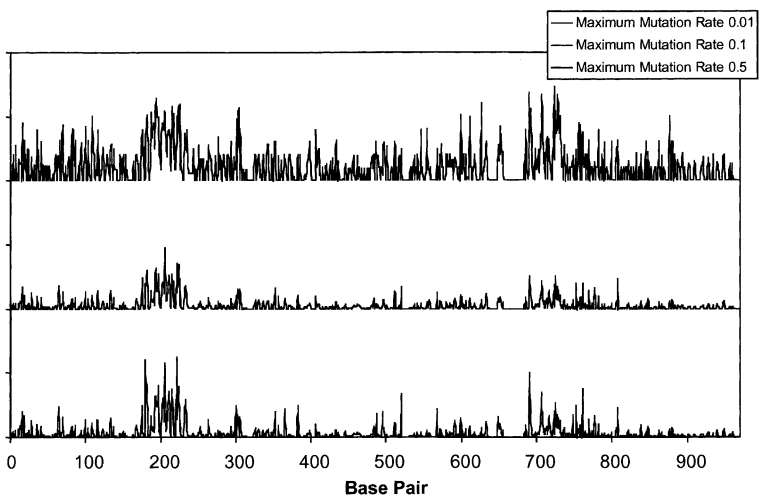

Fig. 6. Total mutation rate. One hundred models were sampled for each maximum mutation rate value. This graph shows the mean total mutation rate inferred for each base pair. The scale is relative, with the top line magnified ×10.

As expected, the highest probability of mutation is observed within VR1, less so within VR2. There are small peaks suggesting mutation within the conserved areas between VR1 and VR2, none within the ITAM motif.

Table 2 describes the average cumulative mutation rates inferred over the K1 region, for each of the three maximum mutation rates allowed. Each of these tables reflects a broadly similar pattern, with the substitutions A↔T, G↔T, and −↔G significantly rarer than the others. The effects of selection can be filtered out of this analysis by considering the average mutation rate from one allele to another only over those sites at which both are observed. In this case, the imbalances observed are substantially reduced and in some cases disappear. This in turn suggests that the mutation patterns inferred are more likely to reflect selection pressures than actual per generation substitution rates.

Discussion

This study of recombination within this highly variable viral gene enables comparisons to be drawn in genealogical history between different regions of K1. We observe inferred hotspots of recombination facilitating sequence shuffling leading to patterns of variation within KSHV. Recombination hotspots are identified in K1 at both conserved and nonconserved nucleotide positions. The mechanism of recombination at the different sites may therefore involve separate mechanisms akin to immunoglobulin gene recombination, with VDJ recombination leading to greater sequence variability on account of nucleotide insertion or deletion compared to isotype class switch recombination (Bassing et al. 2002). Interestingly, K1 is thought to be related to members of the immunoglobulin gene superfamily (Lee et al. 1998; Zong et al. 1999; Lagunoff et al. 200l;

Table 2. The average cumulative mutation rates inferred over the K1 region, for each of the three maximum mutation rates allowed

	Contemporary allele				
Ancestral allele	A	C	G	T	—
Cumulative mutation probabilities for maximum rate 0.01					
A		0.000699	0.000847	0.000362	0.000102
C	0.000664		0.000677	0.000775	0.000072
G	0.000676	0.000584		0.000336	0.000059
T	0.000324	0.000703	0.000439		0.000073
—	0.000129	0.000122	0.000050	0.000096	
Cumulative mutation probabilities for maximum rate 0.1					
A		0.002562	0.002414	0.001033	0.000191
C	0.002639		0.002563	0.002662	0.000162
G	0.002142	0.001879		0.001117	0.000076
T	0.001225	0.002169	0.001269		0.000177
—	0.000792	0.000621	0.000164	0.000551	
Cumulative mutation probabilities for maximum rate 0.5					
A		0.004203	0.003467	0.001773	0.000461
C	0.004760		0.004864	0.004185	0.000558
G	0.003215	0.002829		0.002084	0.000203
T	0.002270	0.003768	0.001912		0.000424
—	0.001511	0.001820	0.000405	0.001334	

Bowser et al. 2002), and it has also been shown to inhibit transport of immunoglobulins to B cell surfaces (Lee et al. 2000). The process of generating variability in this DNA viral gene is termed here recombination shift or drift based on the time scale in which they are postulated to affect the viral sequence (as per previous assumptions [Verhoeyen et al. 1980; Gething et al. 1980]).

Recombination shift involves changes that affect variable positively selected sites wherein immediate effects will be evident and drift is thought to occur when homologous recombination occurs in conserved regions, resulting in longer-term sequence changes over generations.

Among the challenges of the analysis of the role of recombination in evolution is the detection and estimation of recombination in genomes where the rate of substitution is sufficiently high that some sites have experienced multiple mutational events. Viral genes evolve at a high speed compared with genes of higher organisms and hence viral evolution provides interesting material for the study of molecular evolution by recombination. Although recurrent mutations in viruses can generate patterns of genetic variability that resemble the effects of recombination (McVean et al. 2002), the model inference technique shown here adopts the more suitable explanation for each region of the observed data (in this case K1 sequences). In addition, no assumptions are made about the number of mutations that may have occurred at an individual base pair site during the phylogenesis of such a gene (other models assume that a site will have mutated a maximum of once during their evolution).

We used different maximum mutation rates, allowing the results of allowing different degrees of mutation to be compared. Importantly, we observed almost-absolute consistency between the models, especially between mutation rates 0.1 and 0.5, suggesting that our conclusions are independent of whether KSHV is an ancient (Hayward 1999) or in fact a relatively recent pathogen (Russo et al. 1996; Antman and Chang 2000). We do not know, however, when KSHV was introduced, and it is likely that many of its genes have been acquired or pirated presumably by recombination from the host genome over time (Murphy 1997; Nicholas et al. 1998). Since then, these genes have apparently evolved to facilitate viral survival (Nicholas et al. 1998; Haig 2001). However, while most KSHV open reading frames have known homologues or at least suggested homologues, BLAST searches of nonrecombinant stretches of K1 (and K15; data not shown) reveal no sequences that suggest a pirated origin from human genes.

We and others have not observed any relationship between the two types of K15 alleles (predominant or minor) and the K1 sequence. It is, however, notable that both of these genes are located at the extreme ends of the KSHV episome on the right- and left-hand sides of the origin of replication, respectively. Thus changes in K1 and K15 appear to reflect significant recombination activity in this specific area. The absence of any known homologues of these two genes may in turn reflect this. This activity appears, however, to spare the remainder of the viral episome, perhaps due to differences in methylation and/or histone positioning. Supporting this hypothesis, the

transforming EBV oncogene, latent membrane protein-1, shares positional homology with KSHV K1 and induces the expression and activity of the DNA methyltransferases (Tsai et al. 2002).

PCR-based studies examining the predominant and minor forms of K15 have demonstrated evidence for recombination within KSHV. The first of these used classical linkage analysis and the criterion of lack of cosegregation at multiple genetic loci. This led to the hypothesis that an original recombination event occurred that introduced exogenous sequences from a related primate virus of unknown source and that subsequent mutations led to certain KSHV lineages (Poole et al. 1999). A conflicting study used network analysis to show that the proposed introduction of these exogenous sequences did not occur via a single recombination event (Kakoola et al. 2001). Overall, however, analysis of K15 sequences from individuals within the same family provides evidence for recombination in approximately 20–30% of cases.

Positive or negative selective pressures influence nucleotide changes within all genes that change or preserve them, respectively. The neutral theory of evolution predicts that the stronger the selective constraint against nucleotide changes, the lower the rate of base substitutions (Kimura 1968). This prediction is supported by a large number of observations at the DNA sequence level. For example, the rate of synonymous or silent substitutions that produce no alteration in translated proteins is usually much higher than the rate of nonsynonymous substitutions (Endo et al. 1996). However, that positive selection in K1 favors change is evident by the large number of nucleotide changes in the middle position of a codon triplet, a substitution always resulting in amino acid alterations. The recombination hotspots within this highly variable gene provide a possible mechanism by which positive selective pressures exert their effects over time. For this viral oncogene, data suggest that the selective pressure resides with the host cytotoxic T lymphocyte response, as first suggested by Hay ward et al. (Hayward 1999). The inferred hotspots of recombination within the unconserved areas (recombination shift) may be responsible for short-term antigenic variation and evasion or even recognition of host defenses, while those in conserved areas maintain successful long-term viral evolution, i.e., persistence in hosts, due to recombination drift.

The small number of putative ancestors that we infer could originally have been generated by bottleneck events and genetic drift. In this case, the troughs in Fig. 6 infer the presence of highly increased negative selective pressures. This model allows, for the first time, inference of the strength of negative selection, acting to maintain the same amino acids as ancestral sequences. In addition, as well as bottlenecks acting within the context of specific populations, the inferred results also provide evidence that diversifying selection on amino acid variants may be occurring (possibly at the same time). Thus, in K1 we observe recombination facilitating strong positive or negative selective pressures. The variability in K1 represents an example of a gene that is undergoing intense positive or negative selection and the complex mosaics appear to represent coordinated evolutionary drift at multiple loci.

This model is supported by recent preliminary experimental data in which K1 from the human primary effusion lymphoma cell line BCP-1 was cloned into γ-murine herpesvirus-68 (an animal model of a γ-herpesvirus) and then injected into mice. Sequence data from days 10, 14, and 21 postinfection in different animals demonstrates consistent changes occurring at nucleotide position 690, the site at which we infer the highest likelihood of recombination occurring (S. Talbot, personal commun.). In addition, on superimposing Fig. 6 (mutation rate) on Fig. 3 (recombination hotspots), it appears that hotspots are inferred where there appears to be increased nucleotide variation. As the number of segregating sites in these regions appears to be high, this suggests a spatial correlation between mutation rates and recombination rates, although we are unable to infer such rates from this model.

It is quite conceivable that recombination is itself mutagenic (Zhuang et al. 2002) and there are data to suggest that this may in fact be the case. In an attempt to evaluate the genetic diversity of HIV, Srinivasan et al. (1989) designed experiments to analyze recombination between retroviral DNAs by using DNA transfection in cell cultures. They reported the successful recombination between truncated HIV proviral DNAs with an overlap homology of 53 base pairs that leads to the formation of viable hybrid virus. Recombination was also seen between exogenous DNA introduced into cells and homologous HIV sequences resident in the cells. These results indicated that recombination among various HIV isolates may play a significant role in the generation of genetic diversity of HIV. It may also be possible that if a codon has two mutations within it that are segregating in a population, recombination within that codon creates a new amino acid by shuffling those particular variants.

Herpesviruses have evolved through cospeciation and coevolution with their hosts (McGeoch 2001; Stebbing et al. 2003a). Evasion from all host immune control mechanisms will lead to overwhelming viral infection, with subsequent death of the host and therefore the virus. For these viruses to persist as a latent infection without causing harm, an equilibrium between pathogen and host must be established.

Unlike the error-prone reverse transcriptase of retroviruses, herpesviruses do not have a mechanism that will result in rapid sequence variation. Previous data show a clustering of functional cytotoxic T lymphocyte (CTL) epitopes within the most positively selected sites of the whole viral genome (K1), indicating that the pressure causing this selection is partly to facilitate immune recognition (CTL capture as opposed to escape) (Stebbing et al. 2003a). As K1 is expressed predominantly in the early lytic cycle of viral replication, a certain level of viral replication occurs prior to immune recognition and the subsequent death of the infected cell. Recombination provides a mechanism here to generate diversity in response to selective pressures that lead to, we believe, the attraction of an immune response to this variable oncogene. This would ensure that the virus–host equilibrium is established and that latent infection may be achieved. This effect may be most important when the virus is introduced into a new population group containing, for example, new MHC alleles.

Although no homologues of K1 have been identified, it is possible that the K1 sequences represent divergent forms of key genes that evolved very rapidly, with all intermediate forms being lost as each subtype of the virus occupied a new biological niche. Alternatively, conserved areas within K1 may represent relics of older forms of the virus or of related viral species that persist as small areas of their original genomes by virtue of rare recombination events with more modern forms. The continuous expansion of viral diversity over time is influenced by social, behavioral, and biological forces (Perrin et al. 2003). Such biological forces are driven by host immune responses to K1, antiviral drugs, rapid turnover of virus, and mutation events. In retroviruses, the error-prone reverse transcriptase makes significant contributions to these mutation events and facilitates escape from cytotoxic T lymphocyte responses. We show that recombination contributes to the extreme diversity of a DNA viral gene.

Acknowledgments. We are grateful to Professor Chales Bangham for helpful suggestions and critical reading of the manuscript.

References

(1980) Antigenic shift and drift. Nature 283:524–525

An W, Telesnitsky A (2002) HIV-1 genetic recombination: experimental Approaches and observations. AIDS Rev 4:195–212

Antman K, Chang Y (2000) Kaposi's sarcoma. N Engl J Med 342:1027–1038

Awadalla P (2003) The evolutionary genomics of pathogen recombination. Nat Rev Genet 4:50–60

Bassing CH, Swat W, Alt FW (2002) The mechanism and regulation of chromosomal V (D) J recombination. Cell 109 (Suppl):S45–S55

Biggar RJ, Whitby D, Marshall V, Linhares AC, Black F (2000) Human herpesvirus 8 in Brazilian Amerindians: A hyperendemic population with a new subtype. J Infect Dis 181:1562–1568

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294:1351–1362

Bowser BS, DeWire SM, Damania B (2002) Transcriptional regulation of the K1 gene product of Kaposi's sarcoma-associated herpesvirus. J Virol 76:12574–12583

Clark AG (1990) Inference of haplotypes of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Conway DJ, Roper C, Oduola AM, Arnot DE, Kremsner PG, Grobusch MP, Curtis CF, Greenwood BM (1999) High recombination rate in natural populations of Plasmodium falciparum. Proc Natl Acad Sci USA 96:4506–4511

Cook PM, Whitby D, Calabro ML, Luppi M, Kakoola DN, Hjalgrim H, Ariyoshi K, Ensoli B, Davison AJ, Schulz TF (1999) Variability and evolution of Kaposi's sarcoma-associated herpesvirus in Europe and Africa. International Collaborative Group. AIDS 13:1165–1176

Cook RD, Hodgson TA, Waugh AC, Molyneux EM, Borgstein E, Sherry A, Teo CG, Porter SR (2002) Mixed patterns of transmission of human herpesvirus-8 (Kaposi's sarcoma-associated herpesvirus) in Malawian families. J Gen Virol 83:1613–1619

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881–10890

Damania B, Choi JK, Jung JU (2000) Signaling activities of gammaherpesvirus membrane proteins. J Virol 74:1593–1601

Dechter R (1996) Bucket elimination: A unifying framework for probabilistic inference. In: Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-96), August 1–4, pp 211–219

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13: 685–690

Epstein MA (2001) Reflections on Epstein-Barr virus: Some recently resolved old uncertainties. J Infect 43:111–115

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927

Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, Zhou J, Spratt BG (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci USA 98:182–187

Fisher RA (1958) The genetical theory of natural selection. Oxford University Press, London

Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B (2002) Diversity considerations in HIV-1 vaccine selection. Science 296:2354–2360

Gething MJ, Bye J, Skehel J, Water field M (1980) Cloning and DNA sequence of double-stranded copies of haemagglutinin genes from H2 and H3 strains elucidates antigenic shift and drift in human influenza virus. Nature 287:301–306

Gilbert SC, Plebanski M, Gupta S, Morris J, Cox M, Aidoo M, Kwiatkowski D, Greenwood BM, Whittle HC, Hill AV (1998) Association of malaria parasite population structure, HLA, and immunological antagonism. Science 279:1173–1177

Greenspan G, Geiger D (2003) Model-based inference of haplotype block variation. In: The Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin

Guttman DS, Dykhuizen DE (1994) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380–1383

Hahn BH, Shaw GM, De Cock KM, Sharp PM (2000) AIDS as a zoonosis: Scientific and public health implications. Science 287:607–614

Haig DM (2001) Subversion and piracy: DNA viruses and immune evasion. Res Vet Sci 70:205–219

Hansen JE, Lund O, Engelbrecht J, Bohr H, Nielsen JO (1995) Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase. Biochem J 308:801–813

Hayward GS (1999) KSHV strains: The origins and global spread of the virus. Semin Cancer Biol 9:187–199

Ho DD, Huang Y (2002) The HIV-1 vaccine race. Cell 110:135–138

Holmes EC (2001) On the origin and evolution of the human immunodeficiency virus (HIV). Biol Rev Camb Philos Soc 76:239–254

Holub EB (2001) The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet 2:516–527

Jensen FV (1996) An introduction to Bayesian networks. Springer Verlag, New York

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Kakoola DN, Sheldon J, Byabazaire N, Bowden RJ, Katongole-Mbidde E, Schulz TF, Davison AJ (2001) Recombination in human herpesvirus-8 strains from Uganda and evolution of the K15 gene. J Gen Virol 82:2393–2404

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci USA 63:1181–1188

Kimura M (1976a) How genes evolve; A population geneticist's view. Ann Genet 19:153–168

Kimura M (1976b) Population genetics and molecular evolution. Johns Hopkins Med J 138:253–261

Kimura M, Ota T (1971) On the rate of molecular evolution. J Mol Evol 1:1–17

Lacoste V, Judde JG, Briere J, Tulliez M, Garin B, Kassa-Kelembho E, Morvan J, Couppie P, Clyti E, Forteza Vila J, Rio B, Delmer A, Mauclere P, Gessain A (2000a) 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 278:60–74

Lacoste V, Kadyrova E, Chistiakova I, Gurtsevitch V, Judde JG, Gessain A (2000b) Molecular characterization of Kaposi's sarcoma-associated herpesvirus/human herpesvirus-8 strains from Russia. J Gen Virol 81:1217–1222

Lagunoff M, Lukac DM, Ganem D (2001) Immunoreceptor tyrosine-based activation motif-dependent signaling by Kaposi's sarcoma-associated herpesvirus K1 protein: Effects on lytic viral replication. J Virol 75:5891–5898

Lampinen TM, Kulasingam S, Min J, Borok M, Gwanzura L, Lamb J, Mahomed K, Woelk GB, Strand KB, Bosch ML, Edelman DC, Constantine NT, Katzenstein D, Williams MA (2000) Detection of Kaposi's sarcoma-associated herpesvirus in oral and genital secretions of Zimbabwean women. J Infect Dis 181:1785–1790

Lauritzen SL (1995) The EM algorithm for graphical association of missing data. Computational statistics and data analysis 19:191–201

Lee BS, Alvarez X, Ishido S, Lackner AA, Jung JU (2000) Inhibition of intracellular transport of B cell antigen receptor complexes by Kaposi's sarcoma-associated herpesvirus K1. J Exp Med 192:11–21

Lee H, Veazey R, Williams K, Li M, Guo J, Neipel F, Fleckenstein B, Lackner A, Desrosiers RC, Jung JU (1998) Deregulation of cell growth by the K1 gene of Kaposi's sarcoma-associated herpesvirus. Nat Med 4:435–440

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

Malim MH, Emerman M (2001) HIV-1 sequence variation: drift, shift, and attenuation. Cell 104:469–472

Marsden RL, McGuffin LJ, Jones DT (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. Protein Sci 11:2814–8224

Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002a) N-terminal N-myristoylation of proteins: Prediction of substrate proteins from amino acid sequence. J Mol Biol 317:54l–557

Maurer-Stroh S, Eisenhaber B, Eisenhaber F (2002b) N-terminal N-myristoylation of proteins: Refinement of the sequence motif and its taxon-specific differences. J Mol Biol 317:523–540

Maynard-Smith J (1982) The century since Darwin. Nature 296:599–601

McGeoch DJ (2001) Molecular evolution of the gamma-Herpesvirinae. Philos Trans R Soc Lond B Biol Sci 356:421–435

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Meng YX, Spira TJ, Bhat GJ, Birch CJ, Druce JD, Edlin BR, Edwards R, Gunthel C, Newton R, Stamey FR, Wood C, Pellett PE (1999) Individuals from North America, Australasia, and Africa are infected with four different genotypes of human herpesvirus 8. Virology 261:106–119

Meng YX, Sata T, Stamey FR, Voevodin A, Katano H, Koizumi H, Deleon M, De Cristofano MA, Galimberti R, Pellett PE (2001) Molecular characterization of strains of Human herpesvirus 8 from Japan, Argentina and Kuwait. J Gen Virol 82:495–506

Monigatti F, Gasteiger E, Bairoch A, Jung E (2002) The Sulfinator: predicting tyrosine sulfation sites in protein sequences. Bioinformatics 18:769–770

Muller HJ (1964) The relation of recombination to mutational advance. Mutat Res 1:2–9

Munoz JL, Parks WP, Wolinsky SM, Korber BT, Hutto C (1993) HIV-1 reverse transcriptase. A diversity generator and quasispecies regulator. Ann NY Acad Sci 693:65–70

Murphy PM (1997) AIDS-Pirated genes in Kaposi's sarcoma. Nature 385:296–297

Nicholas J, Zong JC, Alcendor DJ, Ciufo DM, Poole LJ, Sarisky RT, Chiou CJ, Zhang X, Wan X, Guo HG, Reitz MS, Hayward GS (1998) Novel organisational features, captured cellular genes, and strain variability within the genome of KSHV/HHV8. J Natl Cancer Inst Monogr 79–88

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157–169

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Pearl J (1988) Probabilistic reasoning in intelligent Systems. Morgan Kaufman, San Mateo, CA

Perrin L, Kaiser L, Yerly S (2003) Travel and the spread of HIV-1 genetic variants. Lancet Infect Dis 3:22–27

Poole LJ, Zong JC, Ciufo DM, Alcendor DJ, Cannon JS, Ambinder R, Orenstein JM, Reitz MS, Hayward GS (1999) 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 73:6646–6660

Prakash O, Tang ZY, Peng X, Coleman R, Gill J, Farr G, Samaniego F (2002) Tumorigenesis and aberrant signaling in transgenic mice expressing the human herpesvirus-8 K1 gene. J Natl Cancer Inst 94:926–935

Quinn TC (1994) Population migration and the spread of types 1 and 2 human immunodeficiency viruses. Proc Natl Acad Sci USA 91:2407–2414

Rieder MJ, Taylor SL, Clark AG, Nickerson DA (1999) Sequence variation in the human angiotensin converting enzyme. Nat Genet 22:59–62

Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471

Rissanen J (1983) A universal prior for integers and estimation by minimum description length. Ann Stat 11:416–431

Robertson DL, Sharp PM, McCutchan FE, Hahn BH (1995) Recombination in HIV-1. Nature 374:124–126

Rouzine IM, Wakeley J, Coffin JM (2003) The solitary wave of asexual evolution. Proc Natl Acad Sci USA 100:587–592

Russo JJ, Bohenzky RA, Chien MC, Chen J, Yan M, Maddalena D, Parry JP, Peruzzi D, Edelman IS, Chang Y, Moore PS (1996) Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). Proc Nat Acad Sci USA 93:14862–14867

Samaniego F, Pati S, Karp J, Prakash O, Bose D (2000) Human herpesvirus 8 k1-associated nuclear factor-kappa b-dependent promoter activity: Role in kaposi's sarcoma inflammation? J Natl Cancer Inst Monogr 28:15–23

Schierup MH, Hein J (2000a) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879–891

Schierup MH, Hein J (2000b) Recombination and the molecular clock. Mol Biol Evol 17:1578–1579

Shannon CE (1948) A mathematical theory of communication. Bell systems technical journal 27:379–423, 623–656

Srinivasan A, York D, Jannoun-Nasr R, Kalyanaraman S, Swan D, Benson J, Bohan C, Luciw PA, Schnoll S, Robinson RA (1989) Generation of hybrid human immunodeficiency virus by homologous recombination. Proc Natl Acad Sci USA 86:6388–6392

Stebbing J, Gazzard B (2002) The clinical utility of resistance testing. Journal of HIV Therapy 7:75–80

Stebbing J, Gazzard B (2003a) Stemming the epidemic: Prevention and therapy go hand-in-hand. J HIV Ther 8:51–55

Stebbing J, Gazzard B (2003b) Virus-host interactions. SAC Review Obstet Gynaecol 5:103–106

Stebbing J, Wilder N, Ariad S, Abu-Shakra M (2001) Lack of intrapatient strain variability during infection with Kaposi's sarcoma-associated herpesvirus. Am J Hematol 68:133–134

Stebbing J, Bourboulia D, Johnson M, Henderson S, Williams I, Wilder N, Tyrer M, Youle M, Imami N, Kobu T, Kuon W, Sieper J, Gotch F, Boshoff C (2003a) Kaposi's sarcoma-associated herpesvirus cytotoxic T lymphocytes recognize and target darwinian positively selected autologous K1 epitopes. J Virol 77:4306–4314

Stebbing J, Portsmouth S, Bower M (2003b) Insights into the molecular biology and sero-epidemiology of Kaposi's sarcoma. Curr Opin Infect Dis 16:25–31

Stebbing J, Portsmouth S, Gotch F, Gazzard B (2003c) Kaposi's sarcoma—An update. Int J STD AIDS 14:225–227

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Tsai CN, Tsai CL, Tse KP, Chang HY, Chang YS (2002) The Epstein-Barr virus oncogene product, latent membrane protein 1, induces the downregulation of E-cadherin gene expression via activation of DNA methyltransferases. Proc Natl Acad Sci USA 99:10084–10089

Twiddy SS, Holmes EC (2003) The extent of homologous recombination in members of the genus Flavivirus. J Gen Virol 84:429–440

Verhoeyen M, Fang R, Jou WM, Devos R, Huylebroeck D, Saman E, Fiers W (1980) Antigenic drift between the haemagglutinin of the Hong Kong influenza strains A/Aichi/2/68 and A/Victoria/3/75. Nature 286:771–776

Walker BD, Korber BT (2001) Immune control of HIV: The obstacles of HLA and viral diversity. Nat Immunol 2:473–475

Zhang Y, Davis T, Wang X, Deng J, Baillargeon J, Yeh T, Jenson H, Gao S (2000) Distinct distribution of rare us kshv genotypes in south texas. Implications for kshv epidemiology and evolution. Ann Epidemiol 10:470

Zhang YJ, Davis TL, Wang XP, Deng JH, Baillargeon J, Yeh IT, Jenson HB, Gao SJ (2001) Distinct distribution of rare US genotypes of Kaposi's sarcoma-associated herpesvirus (KSHV) in South Texas: Implications for KSHV epidemiology. J Infect Dis 183:125–129

Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD, Dougherty JP (2002) Human immunodeficiency virus type 1 recombination: Rate, fidelity, and putative hot spots. J Virol 76:11273–11282

Zong JC, Ciufo DM, Alcendor DJ, Wan X, Nicholas J, Browning PJ, Rady PL, Tyring SK, Orenstein JM, Rabkin CS, Su IJ, Powell KF, Croxson M, Foreman KE, Nickoloff BJ, Alkan S, Hayward GS (1999) High-level variability in the ORF-K1 membrane protein gene at the left end of the Kaposi's sarcoma-associated herpesvirus genome defines four major virus subtypes and multiple variants or clades in different human populations. J Virol 73:4156–4170

Zong J, Ciufo DM, Viscidi R, Alagiozoglou L, Tyring S, Rady P, Orenstein J, Boto W, Kalumbuja H, Romano N, Melbye M, Kang GH, Boshoff C, Hayward GS (2002) Genotypic analysis at multiple loci across Kaposi's sarcoma herpesvirus (KSHV) DNA molecules: clustering patterns, novel variants and chimerism. J Clin Virol 23:119–148