MOLECULAR ECOLOGY

Molecular Ecology (2015) doi: 10.1111/mec.13331

SPECIAL ISSUE: DETECTING SELECTION IN NATURAL POPULATIONS: MAKING SENSE OF GENOME SCANS AND TOWARDS ALTERNATIVE SOLUTIONS

On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetic diversity

NICHOLAS RENZETTE,* TIMOTHY F. KOWALIK*† and JEFFREY D. JENSEN‡§

*Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01655, USA, †Immunology and Microbiology Program, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01655, USA, ‡Swiss Institute of Bioinformatics (SIB), Lausanne CH-1015, Switzerland, §School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne CH-1015, Switzerland

Abstract

A central focus of population genetics has been examining the contribution of selective and neutral processes in shaping patterns of intraspecies diversity. In terms of selection specifically, surveys of higher organisms have shown considerable variation in the relative contributions of background selection and genetic hitchhiking in shaping the distribution of polymorphisms, although these analyses have rarely been extended to bacteria and viruses. Here, we study the evolution of a ubiquitous, viral pathogen, human cytomegalovirus (HCMV), by analysing the relationship among intraspecies diversity, interspecies divergence and rates of recombination. We show that there is a strong correlation between diversity and divergence, consistent with expectations of neutral evolution. However, after correcting for divergence, there remains a significant correlation between intraspecies diversity and recombination rates, with additional analyses suggesting that this correlation is largely due to the effects of background selection. In addition, a small number of loci, centred on long noncoding RNAs, also show evidence of selective sweeps. These data suggest that HCMV evolution is dominated by neutral mechanisms as well as background selection, expanding our understanding of linked selection to a novel class of organisms.

Keywords: background selection, genetic hitchhiking, human cytomegalovirus, intraspecies diversity

Received 24 April 2015; revision received 20 July 2015; accepted 21 July 2015

Introduction

Genetic diversity varies across the genomes of a wide variety of organisms, and considerable theoretical and empirical work has been dedicated to deciphering the relative roles of neutral and selective processes in shaping this variation. The correlation between diversity and recombination rates in many organisms has provided evidence consistent with two competing models, namely background selection and genetic hitchhiking (Begun & Aquadro 1992; Charlesworth *et al.* 1993), with background selection decreasing the fixation probabilities of neutral mutations linked to a deleterious mutation (Charlesworth *et al.* 1993, 1995; Charlesworth 2013), and genetic hitchhiking increasing the fixation probabilities of neutral mutations linked to a beneficial mutation (Maynard-Smith & Haigh 1974; Kaplan *et al.* 1989). Both models predict a reduction in diversity around the regions targeted by selection, where the extent of the reduction of diversity is correlated with the rate of recombination—as recombination will dictate the genomic scale of the effect. Thus, a correlation between diversity and recombination itself does not distinguish between models of background selection and genetic hitchhiking. However, the effect on the distribution of polymorphism frequencies, summarized in the site frequency spectrum (SFS), is expected to differ between

Correspondence: Nicholas Renzette, Fax: +1 508 856 5920; E-mail: nicholas.renzette@umassmed.edu

the two models of selection (Stephan 2010). Specifically, single hitchhiking events are expected to result in a larger proportion of high frequency-derived alleles (Fay & Wu 2000) and an increased proportion of low frequency alleles as compared to a neutral model due to an increased proportion of new (i.e. young) alleles near the site of selection that have arisen since the fixation (Stephan 2010). In contrast to single hitchhiking models, recurrent hitchhiking models predict an excess of low frequency alleles but not an excess of high frequency-derived alleles (Przeworski 2002; Kim 2006), as subsequent beneficial fixations erase this pattern. Lastly, a skewed SFS is not expected under a strong background selection model in which the effect is simply a rescaling of the effective population size (Charlesworth *et al.* 1995); however, under moderate levels of background selection, the situation is more complex (see Ewing & Jensen in this issue). Thus, the SFS is a powerful tool to distinguish between the competing models.

Neutral explanations for variations in patterns of diversity also exist and are consistent with data sampled from a subset of organisms. Mutation rates are shown to vary across the genome of many organisms (Gao & Xu 2008; Lynch *et al.* 2008; Lynch 2010; Ananda *et al.* 2011), which could alter local levels of polymorphism as well as the level of interspecies divergence. The relative contribution of selection or neutral processes in shaping patterns of diversity can thus be explored through the analysis of intraspecies diversity, interspecies divergence and rates of recombination (e.g. Cutter & Payseur 2013). The influence of these mechanisms can vary drastically between species. For example, recurrent hitchhiking has been argued as the dominant evolutionary force in the *Drosophila* species (Begun & Aquadro 1992, 1994; Andolfatto & Przeworski 2001; Innan & Stephan 2003) and *Caenorhabditis elegans* (Andersen *et al.* 2012), and to a lesser degree in *Saccharomyces cerevisia* (Cutter & Moses 2011), while background selection against a multitude of deleterious alleles appears to be important in human evolution (Lohmueller *et al.* 2011). In other species, such as *Lycopersicon*, the evidence of selection is weak, favouring neutral explanations for the variation in diversity (Baudry *et al.* 2001; Roselius *et al.* 2005). It is important to note, however, that all of the above claims remain contentious and the estimation of background selection in particular has been under-utilized.

Unfortunately, a full survey of the dominance of these mechanisms has not been completed across a wide range of species. In particular, there has been a dearth of analysis of bacterial and viral systems. Viruses, due to the short generation times, large effective population sizes, high mutation rates and compact, coding sequence dense genomes, provide a novel class of organisms in which to study these processes. The few studies available addressing the roles of background selection and genetic hitchhiking in viral evolution have focused on RNA viruses (Ramachandran *et al.* 2011; Strelkowa & Lässig 2012) with low rates of intrasegment recombination (Worobey *et al.* 2002; Shi *et al.* 2012) and therefore have highlighted the influence of clonal interference in this context (Strelkowa & Lässig 2012). In contrast, recombination is common in DNA viruses (Fleischmann 1996) which should lessen the influence of clonal interference and increase rates of adaptation (Miralles *et al.* 1999) relative to comparably sized populations without recombination.

Human cytomegalovirus (HCMV), a large dsDNA virus, is an ideal viral system in which to study the roles of these processes in shaping levels of diversity in natural populations. HCMV contains the largest genome of any human viral pathogen and was first sequenced over 20 years ago (Chee *et al.* 1990). The approximately 235 kilobase (kb) double-stranded DNA genome encodes at least 160 genes (Dolan *et al.* 2004), with recent estimates suggesting >500 unique protein products (Stern-Ginossar *et al.* 2012). HCMV is an opportunistic pathogen that rarely causes disease in healthy hosts, but can lead to severe pathology in the immunosuppressed or immunonaive, such as foetuses and neonates (Griffiths *et al.* 2015). Indeed, HCMV is the leading cause of birth defects associated with an infectious agent (Alford *et al.* 1990). The cytomegaloviruses are an ancient family of viruses that appear to infect all mammals (McGeoch *et al.* 1995), but also exhibit a remarkable level of host specificity (Mocarski *et al.* 2007), such that each host species is infected with a cognate species of cytomegalovirus. Recent studies have investigated the intrahost evolution of the virus on relatively short timescales, showing that the viral populations can evolve rapidly due to the joint effects of positive selection and demography (Renzette *et al.* 2013). It is currently unclear, however, how these same mechanisms alter the longer term evolution and diversity of the HCMV species.

Here, we evaluate the genomic patterns of HCMV intraspecies diversity, showing that the levels can vary by two orders of magnitude across the genome. Comparison of intraspecies diversity and interspecies divergence suggest that a portion, though not all, of the variation can be explained by neutral mechanisms. Background selection appears to be the dominant effect altering these patterns of diversity, although a genomewide scan of selection does suggest that hitchhiking may have reduced diversity in regions encoding long noncoding RNAs (lncRNAs). These regions of adaptation may provide insight in the phenotypic differences between human and chimpanzee cytomegaloviruses (CCMVs).

Materials and methods

Data

The HCMV intraspecies data set contains all available whole genomes of minimally passaged HCMV isolates from GenBank (n = 41). Whole genomes of 1 strain of Panine herpesvirus 2 (i.e. CCMV) and two strains of Macacine herpesvirus 3 (i.e. rhesus macaque cytomegalovirus) were also obtained from GenBank and were used to infer derived alleles in the HCMV data set. Information relevant to the strains and Accession nos used for the analyses can be found in Table S1 (Supporting information).

Data analysis

Alignment of the cytomegalovirus genomes was performed with MAFFT (Katoh *et al.* 2002) and the alignment was manually investigated for quality and accuracy, particularly in repeat regions. Genomewide and locus-specific summary statistics were calculated with the POPGENOME package (Pfeifer *et al.* 2014) for R (R Core Team 2014) and were estimated in 500-bp sliding windows and reported as diversity per-site in Table 1. Effective population size (N_e) was estimated using the equation $\Theta_w = 2N_e\mu$ where μ is the mutation rate per-site per-generation. The estimate of Θ_w from this study (Table 1) was used along with a previously reported estimate of μ (Renzette *et al.* 2015). All statistical analyses were performed in R (R Core Team 2014). For this study, polymorphisms were not divided into putatively neutral or selected sites due to the difficulty in assigning these classes to the HCMV genome. HCMV transcription is complex, with nearly the whole genome being transcribed in the sense and antisense directions, and many sites being functional at the RNA and protein levels (Gatherer *et al.* 2011). Further, alternative transcriptional start sites lead to production of different proteins from the same locus and further complicate the assignment of silent and nonsynonymous sites (Stern-Ginossar *et al.* 2012).

Physical estimates of rates of crossing over have not been reported for HCMV. Alignments of the HCMV genomes used in this study were analysed with the *interval* algorithm from the LDHAT program (McVean *et al.* 2004) to estimate the population recombination rates in 500-bp sliding windows (Fig. S1, Supporting information). All values of nucleotide diversity (π), divergence (D_{xy}) and population recombination rates ($2N_er$, where N_e is effective population size and r is per-site per-generation recombination rate) were log-transformed prior to plotting and during calculations of correlation coefficients to improve the fit of the statistics to a normal distribution (Figs S2–S4, Supporting information). The SWEEPFINDER program, which compares local increases in high frequency-derived alleles against the genomewide distribution (Nielsen *et al.* 2005), was used to identify putative selective sweeps in the HCMV data set. Significance of the SWEEPFINDER result was determined through two sets of simulations. The first set consisted of 1000 neutral simulations using the *ms* program (Hudson 2002) with values of theta and rates of recombination set at the genomewide averages from the HCMV data set and constant population size (Fig. S5A, Supporting information). The second set of simulations (Fig. S5B, Supporting information) was similar to the first, but included a demographic model based on previous estimates of HCMV population histories (Renzette *et al.* 2013). This model included a 99% reduction of population size 500 generations in the past followed by an exponential growth to the current size. The second set of simulations, due to the presence of a recent and strong bottleneck, serves as a more stringent threshold than the first set of simulations.

Table 1 Summary of human cytomegalovirus genomewide intraspecies diversity

Number of sites analysed*	Segregating sites	Θ_w†	π‡	N_e§
210 730	28 101	0.0315	0.0197	78 750

*Sites with insertions and deletions were excluded from the analysis.
†Watterson's estimator of intraspecies diversity.
‡Intraspecies nucleotide diversity.
§Effective population size estimated by $N_e = \Theta_w/2\mu$.

Results

An analysis of the patterns of HCMV intraspecies diversity has not been previously reported, although considerable attention has been applied to loci of clinical importance, such as the virally encoded glycoproteins (Puchhammer-Stöckl & Görzer 2011). Thus, the level of genomewide variation was here quantified through analysis of 41 publicly available whole genomes of HCMV. Because extensive in vitro culturing of HCMV has been shown to result in considerable genomic changes (Dolan *et al.* 2004; Bradley *et al.* 2009; Cunningham *et al.* 2009; Stanton *et al.* 2010), only low passage number strains, or so-called clinical strains, were included in this analysis. The total level of variation across the genome was reported as Watterson's estimator of diversity (Θ_w) and nucleotide diversity (π), the average nucleotide difference between sequences, with

insertions and deletions excluded from the analysis (Table 1). The estimates of HCMV intraspecies diversity were approximately 10-fold higher than values reported for intrahost populations of HCMV (Renzette *et al.* 2011), a result that is not unexpected given the strong bottlenecks that intrahost populations probably experienced prior to sampling (Renzette *et al.* 2013). Furthermore, the diversity varied by approximately two orders of magnitude across the genome, with elevated levels of diversity near the termini of the genome and the internal repeat region (Fig. 1).

Variation in genomewide diversity has been reported in many organisms previously, with both neutral and selective models invoked to explain the patterns. To explore the mechanisms influencing variation in HCMV intraspecies diversity (π), the level of interspecies divergence (D_{xy}) and genomewide population recombination rates ($2N_e r$) were estimated. Intraspecies diversity was shown to positively correlate with the rates of recombination (Fig. 2A, Pearson's $R = 0.35$, $P = 1.9 \times 10^{-11}$), a result consistent with models of linked selection. However, levels of diversity also positively correlated with the level of divergence (Fig. 2B, Pearson's $R = 0.38$, $P = 1.2 \times 10^{-13}$), and in fact, divergence and rates of recombination are also positively correlated (Fig. 2C, Pearson's $R = 0.28$, $P = 5.4 \times 10^{-8}$). In addition, a linear model was fit to the data in which intraspecies diversity was the response variable and interspecies divergence and recombination rates were the predictor variables. Both divergence and recombination rates were shown to be significant predictors of intraspecies diversity (Table S2, Supporting information). These results can be explained by neutral processes, such as variation in ancestral population diversity or fluctuations in mutation rate across the genome. If the correlation between diversity and recombination rates results from linked selection, it is expected that the correlation will remain when intraspecies diversity is corrected by interspecies divergence (Roselius *et al.* 2005); and indeed, corrected intraspecies diversity (π/D_{xy}) was positively correlated with recombination rates (Fig. 2D, Pearson's $R = 0.19$, $P = 3.8 \times 10^{-4}$). Lastly, because the sequential sliding windows may not represent independent observations, the data set was thinned to only include windows every 5000 bp (Fig. S6, Supporting information). The results from the thinned data set were largely consistent with those from the whole data set (Fig. S6, Supporting information). In sum, these results suggest that the observed variation in HCMV genomewide intraspecies diversity can be explained by neutral processes along with some mode of linked selection.

Human cytomegalovirus infects a wide range of host cells and organs, and viral intrahost diversity can vary among organs (Ross *et al.* 2011; Frange *et al.* 2013; Renzette *et al.* 2013). This phenomenon has been referred to as compartmentalization and is influenced by neutral mechanisms, including demography, and by local adaptation (Renzette *et al.* 2013). Thus, HCMV sampled from various organs of a single host can be viewed as subdivided populations. To test whether the species-level correlations between diversity, divergence and recombination rates are also observed in a potentially subdivided HCMV population, only genomes of HCMV collected from host urine were analysed (i.e. the kidney compartment). Urine isolates were selected because the largest number of genomes ($n = 17$) has been sequenced from this host fluid. The overall patterns observed in the urine isolate subset was similar to that observed in

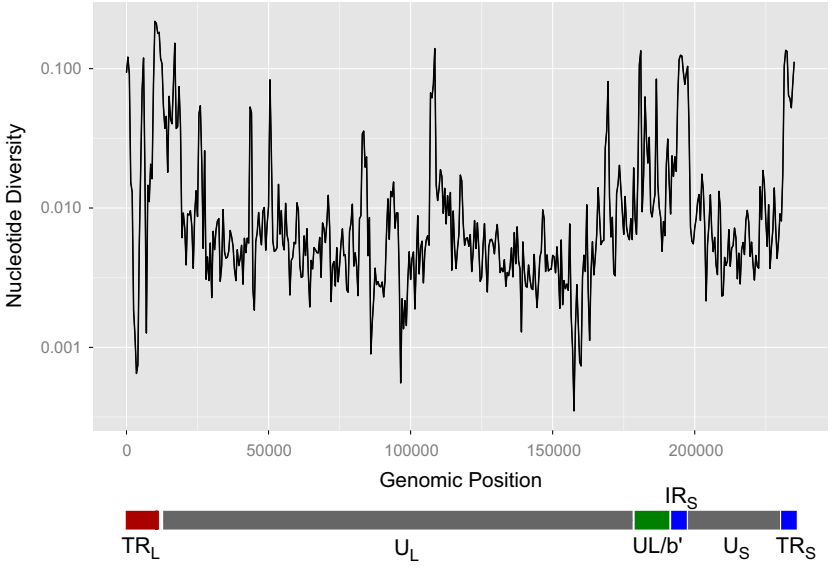

Fig. 1 Human cytomegalovirus (HCMV) genomewide diversity: Intraspecies nucleotide diversity (π) was calculated in 500-bp windows and plotted across the genome on a log scale. A diagram of the main regions of the HCMV genome is shown below the plot. TR_L, Terminal Repeat Long; U_L, Unique Long; U_L/b', Unique Long of clinical strains; U_S, Unique Short; IR_s, Internal Repeat Short; TR_S, Terminal Repeat Short. Terminal Repeat Long is a misnomer as clinical strains of HCMV encode a single copy of the region, although the region is repeated in laboratory strains of HCMV, which were excluded from the current study.

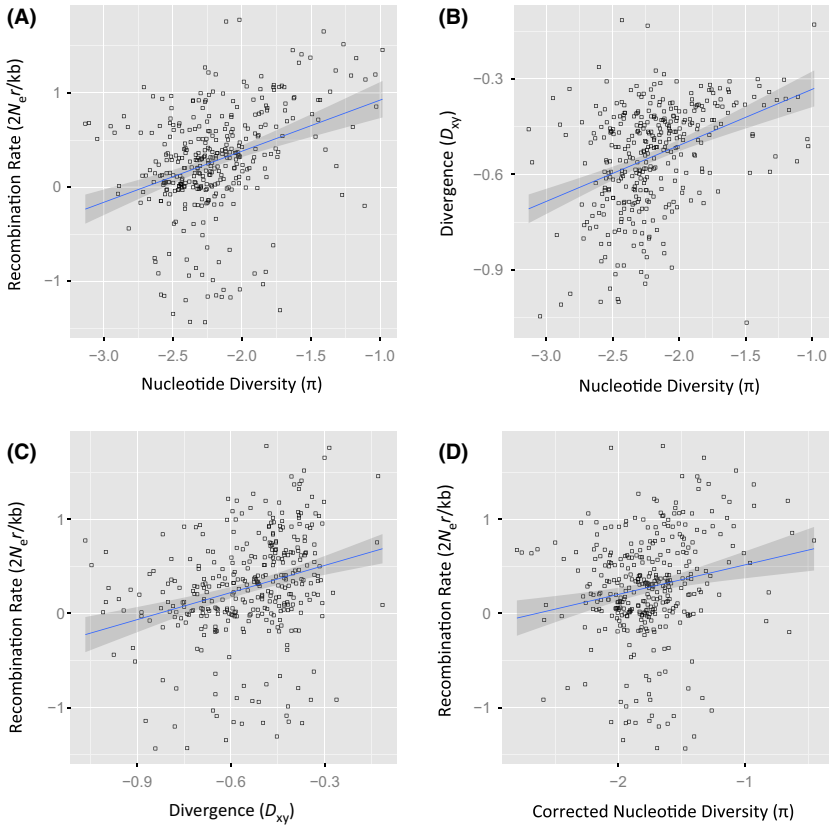

Fig. 2 Correlation of human cytomegalovirus intraspecies diversity, interspecies divergence and recombination rates: Scatter plots of (A) population recombination rates ($2N_e r$) and intraspecies diversity (π) (Pearson's $R = 0.35$, $P = 1.9 \times 10^{-11}$) (B) interspecies divergence (D_{xy}) and intraspecies diversity (π) (Pearson's $R = 0.38$, $P = 1.2 \times 10^{-13}$) (C) population recombination rates ($2N_e r$) and interspecies divergence (D_{xy}) (Pearson's $R = 0.28$, $P = 5.4 \times 10^{-8}$) and (D) the corrected levels of intraspecies diversity (π/D_{xy}) and population recombination rates ($2N_e r$) (Pearson's $R = 0.19$, $P = 3.8 \times 10^{-4}$). Blue lines represent linear regressions of the data, and grey shading indicates the 95% confidence intervals.

the data set of all isolates (Fig. 3A–E). Specifically, diversity (π) and divergence (D_{xy}) were positively correlated with each other and with recombination rates ($2N_e r$) (Fig. 3B–D). Further, divergence-corrected diversity (π/D_{xy}) remains significantly correlated with recombination rates (Fig. 3E, Pearson's $R = 0.31$, $P = 3.1 \times 10^{-10}$) although the correlation was stronger than that observed in the whole data set. The similarity of the two analyses suggests that the relative contribution of neutral processes and linked selection to HCMV evolution is not influenced by host compartment, although further surveying of different patient cohorts from a larger number of host organs would be highly informative.

The two competing models of background selection (Charlesworth *et al.* 1993, 1995) and genetic hitchhiking (Maynard-Smith & Haigh 1974; Kaplan *et al.* 1989) could each possibly explain the observed correlation between diversity and recombination rates reported in Figs 2 and 3. Thus, further analysis was performed to distinguish between these models. Innan & Stephan (2003) demonstrated that in regions of low (but non-zero) rates of recombination the relationship between Θ and Θ/*r* (per-site per-generation recombination rate) is markedly different under the two models. Hitchhiking models result in a negative correlation between these values, while background selection leads to a positive correlation. In the HCMV whole species data set, a strong positive correlation is observed between Θ and Θ/*r* in regions of low recombination ($r < 3 \times 10^{-9}$ crossovers per-site per-generation), consistent with a model of background selection (Fig. 4A, Pearson's $R = 0.95$, $P < 2.2 \times 10^{-16}$). The strong correlation remained even when three regions of low recombination rates but high diversity were excluded from the analysis (Fig. 4B, Pearson's $R = 0.86$, $P = 3.8 \times 10^{-11}$) and was also observed in the isolates collected only from urine (Fig. 3F, Pearson's $R = 0.97$, $P > 2.2 \times 10^{-16}$). Thus, the results suggest that background selection has had a globally larger effect than genetic hitchhiking in shaping patterns of HCMV intraspecies diversity.

Genetic hitchhiking and background selection models are also expected to leave different imprints on the distribution of polymorphism frequencies, or the SFS. Specifically, for large populations with biologically reasonable deleterious mutation rates, background selection is expected to cause little skewing of the SFS as compared to the neutral expectation (Charlesworth *et al.* 1995). In contrast, single hitchhiking models predict an excess of high frequency-derived alleles (Fay & Wu 2000) and low frequency mutations in the SFS, with the effect expected to be more pronounced in regions of low rates of recombination (Kim & Stephan 2000). An

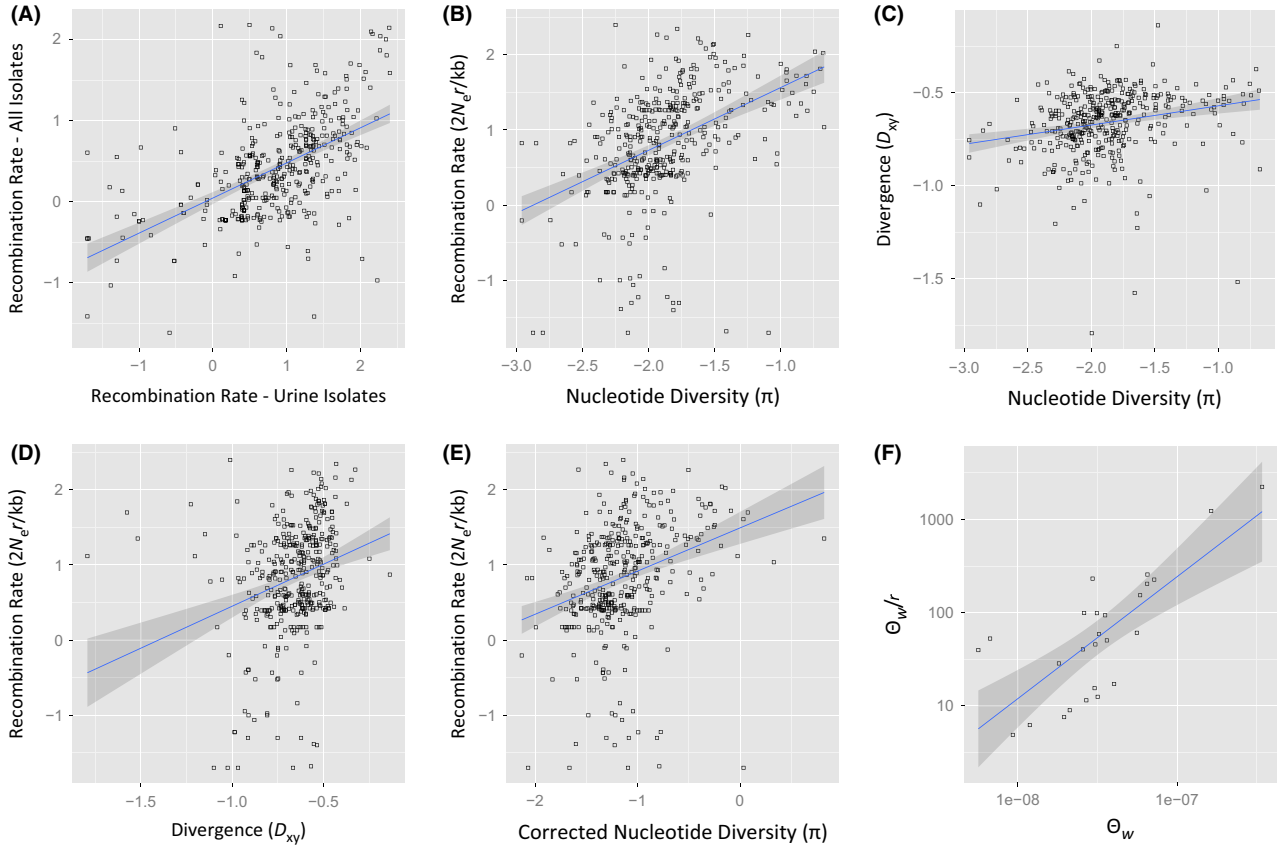

Fig. 3 Correlation of human cytomegalovirus (HCMV) intraspecies diversity, interspecies divergence and recombination rates in urine Isolates: (A) Scatter plots of population recombination rates of all HCMV isolates ($2N_e r$) and recombination rates of only urine isolates ($2N_e r$) (Pearson's $R = 0.54$, $P < 2.2 \times 10^{-16}$). Scatter plots for data from urine isolates of (B) population recombination rates ($2N_e r$) and intraspecies diversity (Pearson's $R = 0.44$, $P < 2.2 \times 10^{-16}$), (C) interspecies divergence and intraspecies diversity (Pearson's $R = 0.23$, $P = 3.6 \times 10^{-6}$), (D) population recombination rates ($2N_e r$) and interspecies divergence (Pearson's $R = 0.27$, $P = 5.6 \times 10^{-8}$), (E) the corrected levels of intraspecies diversity (π/D_{xy}) and population recombination rates ($2N_e r$) (Pearson's $R = 0.31$, $P = 3.1 \times 10^{-10}$), and (F) Θ_w and Θ_w/r for all loci with recombination rates $<3 \times 10^{-9}$ (Pearson's $R = 0.97$, $P < 2.2 \times 10^{-16}$). Blue lines represent linear regressions of the data, and grey shading indicates 95% confidence intervals.

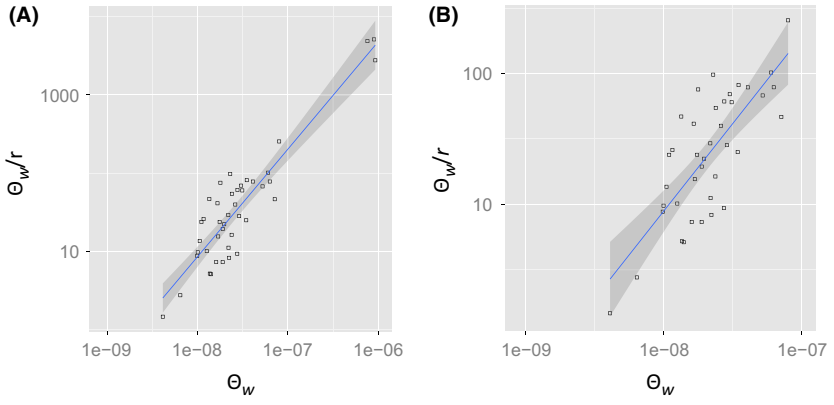

Fig. 4 Scatter plots of Θ_w and Θ_w/r in regions of low recombination rates. (A) Scatter plots of Θ_w and Θ_w/r for all regions with recombination rates $<1 \times 10^{-8}$ (Pearson's $R = 0.94$, $P = 1.3 \times 10^{-8}$), (B) Scatter plots of Θ_w and Θ_w/r with three regions of elevated Θ_w removed (Pearson's $R = 0.88$, $P = 3.1 \times 10^{-6}$). Blue lines represent linear regressions of the data, and grey shading indicating 95% confidence intervals.

increased proportion of low frequency alleles is also expected in recurrent hitchhiking models (Braverman *et al.* 1995), and the proportion should be negatively correlated with recombination rates (Andolfatto & Przeworski 2001). However, recurrent hitchhiking models do not predict an excess of high frequency-derived alleles (Przeworski 2002; Kim 2006). The predictions of the models were tested with the HCMV species data by

calculating unfolded SFS for the entire genome as well as the regions with the lowest and highest recombination rates (5th and 95th percentiles, respectively, where $n = 23$ windows for both groupings). The SFS across the different regions were statistically indistinguishable ($P = 0.98$, Kruskal–Wallis, Fig. 5A–C). Furthermore, no significant correlation was observed between the proportion of low frequency alleles ($f < 0.05$) and rates of recombination (Fig. 5E, Pearson's $R = 0.07$, $P = 0.19$). Thus, predictions of both single hitchhiking and recurrent hitchhiking models, including a skewing of the SFS and an increased proportion of low frequency alleles with decreasing rates of recombination, were not supported by the data. These results, in combination with those presented in Fig. 4, are more consistent with predictions of a background selection model.

The previous results suggest that background selection is the dominant mode of linked selection influencing variation in HCMV diversity, but do not exclude the possibility that a fraction of the HCMV genomewide diversity has been altered by genetic hitchhiking. To test for evidence of hitchhiking within the HCMV species, two statistics were calculated across the genome. The first is a SFS -based approach (SWEEPFINDER; Nielsen *et al.* 2005), and the second, a linkage disequilibrium (LD)-based approach (ZnS; Kelly 1997). The SWEEPFINDER analysis searches for genomic regions containing a SFS consistent with the genetic hitchhiking model, with the test statistic being the ratio of the composite likelihood (CLR) under a selective sweep model compared to the composite likelihood under a null model derived from the genome wide data. The combination of the CLR-based approach with LD data has been shown to be a powerful method for confidently identifying selective sweeps (Pavlidis *et al.* 2010), and thus, both analyses were applied to the HCMV data set, as the CLR-based approach alone has been shown to suffer from high false-positive rates under a range of nonequilibrium models (Crisci *et al.* 2013).

Elevated CLR values were calculated at many loci across the genome (Fig. 6A). In contrast, marked increases in LD were only observed in three genomic regions, at approximately genomic positions 4, 96 and 158 kb, which coincided with regions of elevated CLR values. Interestingly, the high LD regions are devoid of protein coding sequence but are nearly perfectly centred on the three lncRNAs, specifically lncRNA2.7, lncRNA4.9 and lncRNA5. Haplotype maps of these three regions show that there is strong conservation of the loci across all HCMV strains (Fig. S7, Supporting information), and the regions also exhibited the lowest levels of intraspecies diversity (Fig. 1) and patterns expected after a recent selective sweep. Characterizing the functions of the lncRNAs is an active area of

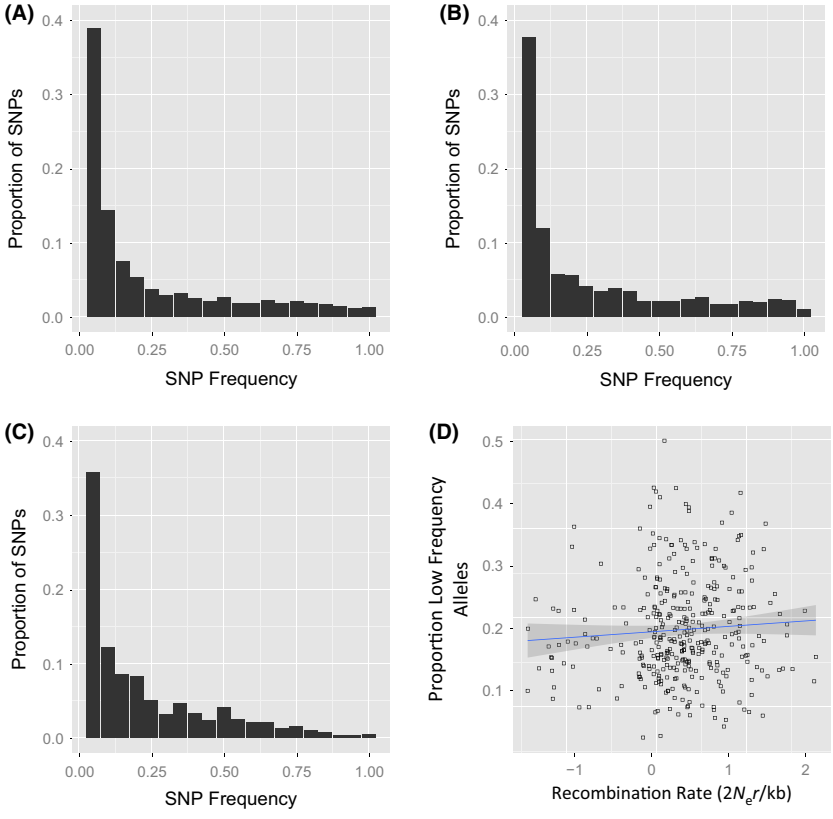

Fig. 5 Relationship between recombination rates and site frequency spectra. Unfolded site frequency spectra were estimated for (A) the whole genome and regions in the 5th (B) and 95th (C) percentile of recombination rates. The 5th and 95th percentile regions each include 23 windows. SNP frequencies were binned into 20 groups and plotted. The last bin corresponds to SNPs of $0.95 < f \leq 0.975$ (i.e. frequency of a SNP present in 40 of 41 analysed genomes) and does not included fixed derived alleles. The proportion of low frequency alleles ($f < 0.05$) within the site frequency spectrum were calculated in 500-bp windows across the genome and compared to local recombination rates. (D) Scatter plot of the proportion of low frequency alleles ($f < 0.05$) and recombination rates (Pearson's $R = 0.07$, $P = 0.19$).

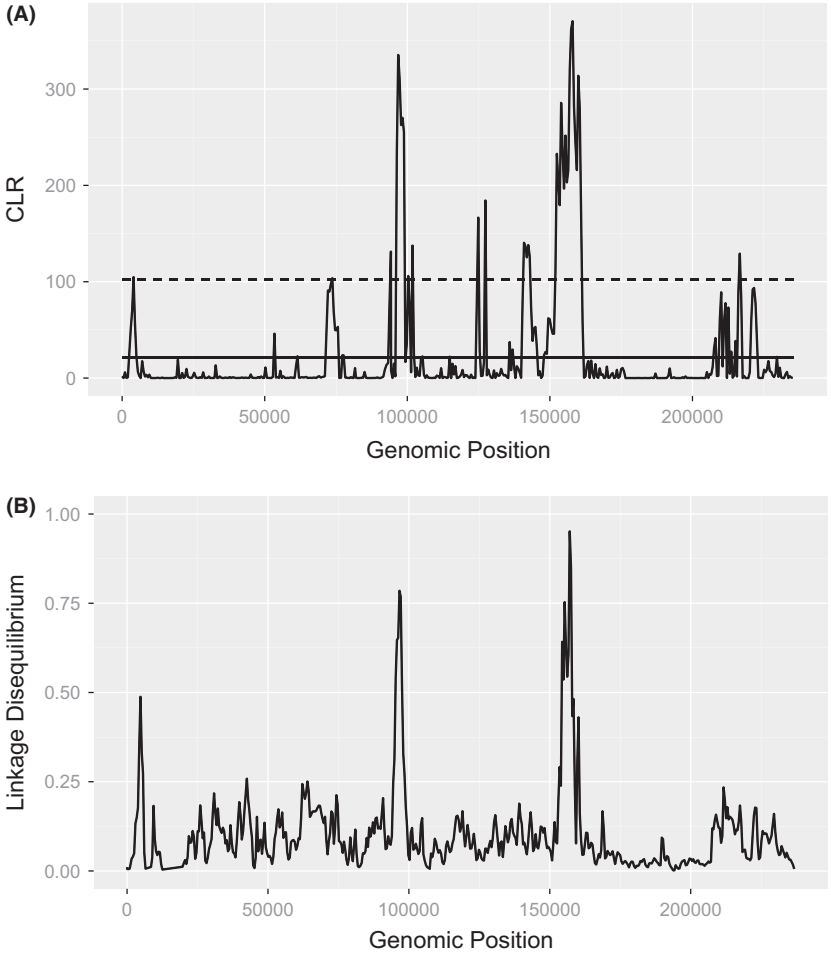

Fig. 6 Genomewide scans of selective sweeps in the human cytomegalovirus (HCMV) genome. (A) Composite likelihood ratio test (i.e. SWEEPFINDER analysis) of selective sweeps in the HCMV genome. Solid horizontal line represents 99% significance threshold as obtained from neutral simulations with constant population size, and dashed horizontal line represents 99% significance threshold from simulations with a recent, strong bottleneck. (B) Genomewide patterns in linkage disequilibrium (LD). LD was estimated with Kelly's ZnS statistic calculated in 500-bp sliding windows (Kelly 1997).

research, although it is clear that the encoded RNAs are highly transcribed during in vitro and in vivo HCMV infections (Stark *et al.* 2012; DeBoever *et al.* 2013). Thus, the data suggest that positive selection is here targeting RNA, rather than protein, function.

Discussion

Studies of variation in natural populations have increased our understanding of the relative contribution of neutral processes and selection in evolution. Here, analysis of HCMV diversity, divergence and recombination has shown that viral evolution is shaped by both neutral and selective processes. Purifying selection against deleterious alleles appears to have an important influence on HCMV diversity, although evidence of selective sweeps of advantageous alleles is also observed at lncRNAs.

Our observations suggest that no single evolutionary model can alone explain HCMV evolution. Neutral processes clearly influence patterns of intraspecies diversity. These processes could include variation of mutation rates across the genome, particularly near sequence motifs that also influence recombination rates (Hellmann *et al.* 2003). However, our data are not inconsistent with a model in which recombination itself is mutagenic (Magni 1964), although the validity of this model has been questioned in other organisms (Cutter & Moses 2011). These models are highly tractable to in vitro experimental studies and can be disentangled in future work. In addition, purifying selection against recurrent deleterious mutations appears to be the dominant selective mechanism shaping patterns of diversity, particularly in regions of low recombination rates. Lastly, evidence of genetic hitchhiking was observed at three loci, which were all centred on lncRNAs. The function of the lncRNAs has been the focus of recent research. All three are highly transcribed (Gatherer *et al.* 2011; Stark *et al.* 2012) and easily detectable in human hosts during natural infection (DeBoever *et al.* 2013). lncRNA2.7 has been shown to interact with host enzymes to prevent host cell apoptosis and alter metabolism in favour of viral replication (Reeves *et al.* 2007). A homologue of lncRNA5 is critical for transitioning from acute to persistent viral infection, thereby influencing virulence (Kulesza & Shenk 2006). Little has been

reported about lncRNA4.9. The underlying biological mechanisms that cause these loci to be targets of positive selection are unclear. However, it is tempting to speculate that the divergence of these loci between CCMV and HCMV, as well as the strong conservation within HCMV, may play a role in the strong host specificity of CCMV and HCMV (Mocarski *et al.* 2007). For example, other herpesviruses are able to infect cells from highly diverged species (Whitley 1996), and the reason that HCMV and CCMV can efficiently infect solely their cognate hosts is unclear, although both viral and host factors probably play a role (Jurak & Brune 2006; Child *et al.* 2012). The identification of the lncRNAs in this genomewide screen suggests that these genes have played an important role in the fitness of HCMV in human hosts and may also play a part in the phenotypic divergence between CCMV and HCMV.

The current study extends our knowledge of the correlation between diversity, divergence and recombination in viral populations. Previous studies in higher organisms have shown that the relationship between these measures can vary drastically between species. For example, in Drosophila, a seminal study by Begun & Aquadro (1992) showed that diversity and recombination rates are strongly correlated, but divergence does not scale with recombination. The authors concluded that that these patterns are consistent with genetic hitchhiking models, a conclusion supported by other studies (Begun & Aquadro 1994; Andolfatto & Przeworski 2001; Innan & Stephan 2003). In *Caenorhabditis elegans* and *Saccharomyces cerevisiae*, hitchhiking has reduced diversity in multiple chromosomes, although the signature is weaker than that observed in Drosophila (Cutter & Moses 2011; Andersen *et al.* 2012). In contrast, studies of wild and domesticated tomato species (*Lycopersicon* spp.) have shown patterns largely consistent with neutral evolution (Baudry *et al.* 2001), although weaker evidence of background selection is also observed (Roselius *et al.* 2005). Similarly, background selection is an important contributor to the patterns of diversity observed in humans (Lohmueller *et al.* 2011) and wild and domesticated rice (Flowers *et al.* 2012). In comparison, HCMV evolution is consistent with a weak signature of linked selection, likely driven by background selection. The explanation for this pattern is not clear, but a few aspects of virus biology may contribute. Perhaps most importantly, the smaller more gene-dense genomes of viruses probably increase the relative proportion of deleterious alleles, while the higher mutation rates increase the absolute input of deleterious mutations. Similar studies of other viruses can be used to test the potential connection between the particulars of viral biology and the dominant underlying selection mechanism.

A confounding factor missing from the current analysis is the role of demography in shaping the observed patterns of diversity. Studies of other human pathogens have shown that host and pathogen demographic histories are similar, including origins in Africa (Linz *et al.* 2007; Montano *et al.* 2015) and recent population expansions (Wirth *et al.* 2008). Cytomegaloviruses are proposed to be an ancient virus family that have speciated with their cognate hosts (McGeoch *et al.* 1995), and thus, HCMV has likely experienced a similar demographic history to its human hosts. Furthermore, HCMV, such as humans, could be geographically subdivided, which would inhibit the ability to detect hitchhiking in species-level data (Cutter & Payseur 2013). The majority of publically available HCMV sequences were sampled from Europe and the United States, although increasing numbers have recently been sampled from Asia. By studying the global distribution of HCMV diversity, future work can be directed at understanding whether HCMV is geographically subdivided and thus whether patterns of local subpopulation-specific adaptation may be more dominant than species-wide patterns.

Acknowledgements

This work was supported by grants from the Swiss National Science Foundation and a European Research Council (ERC) Starting Grant to JDJ, and by the National Institutes of Health (HD061959 (TFK), AI109001 (TFK); F32AI084437 (NR)) and the National Center for Advancing Translational Sciences of the NIH (UL1TR000161). We would like to thank Kristen Irwin for a careful reading of the manuscript and Susanne Pfeifer for helpful suggestions regarding recombination rate estimation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Alford CA, Stagno S, Pass RF, Britt WJ (1990) Congenital and perinatal cytomegalovirus infections. *Review of Infectious Diseases*, **12**, S745–S753.

Ananda G, Chiaromonte F, Makova KD (2011) A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biology*, **12**, R27.

Andersen EC, Gerke JP, Shapiro JA *et al.* (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics*, **44**, 285–290.

Andolfatto P, Przeworski M (2001) Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics*, **158**, 657–665.

Baudry E, Kerdelhué C, Innan H, Stephan W (2001) Species and recombination effects on DNA variability in the tomato genus. *Genetics*, **158**, 1725–1735.

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Begun DJ, Aquadro CF (1994) Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of drosophila: selection and geographic differentiation. *Genetics*, **136**, 155–171.

Bradley AJ, Lurain NS, Ghazal P *et al.* (2009) High-throughput sequence analysis of variants of human cytomegalovirus strains Towne and AD169. *Journal of General Virology*, **90**, 2375–2380.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, **140**, 783–796.

Charlesworth B (2013) Background selection 20 years on: the Wilhelmine E. Key 2012 Invitational Lecture. *Journal of Heredity*, **104**, 161–171.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth D, Charlesworth B, Morgan MT (1995) The pattern of neutral molecular variation under the background selection model. *Genetics*, **141**, 1619–1632.

Chee MS, Bankier AT, Beck S *et al.* (1990) Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. *Current Topics in Microbiology and Immunology*, **154**, 125–169.

Child SJ, Brennan G, Braggin JE, Geballe AP (2012) Species specificity of protein kinase R antagonism by cytomegalovirus TRS1 genes. *Journal of Virology*, **86**, 3880–3889.

Crisci JL, Poh YP, Mahajan S, Jensen JD (2013) The impact of equilibrium assumptions on tests of selection. *Frontiers in Genetics*, **4**, 235.

Cunningham C, Gatherer D, Hilfrich B *et al.* (2009) Sequences of complete human cytomegalovirus genomes from infected cell cultures and clinical specimens. *Journal of General Virology*, **91**, 605–615.

Cutter AD, Moses AM (2011) Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. *Molecular Biology and Evolution*, **28**, 1745–1754.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262–274.

DeBoever C, Reid EG, Smith EN *et al.* (2013) Whole transcriptome sequencing enables discovery and analysis of viruses in archived primary central nervous system lymphomas. *PLoS One*, **8**, e73956.

Dolan A, Cunningham C, Hector RD *et al.* (2004) Genetic content of wild-type human cytomegalovirus. *Journal of General Virology*, **85**, 1301–1312.

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics*, **155**, 1405–1413.

Fleischmann WRJ (1996) Chapter 43: Viral genetics. In: *Medical Microbiology* (ed. Baron S), pp. 421–426. University of Texas Medical Branch, Galveston, Texas.

Flowers JM, Molina J, Rubinstein S *et al.* (2012) Natural selection in gene-dense regions shapes the genomic pattern of polymorphism in wild and domesticated rice. *Molecular Biology and Evolution*, **29**, 675–687.

Frange P, Boutolleau D, Leruez-Ville M *et al.* (2013) Temporal and spatial compartmentalization of drug-resistant cytomegalovirus (CMV) in a child with CMV meningoencephalitis: implications for sampling in molecular diagnosis. *Journal of Clinical Microbiology*, **51**, 4266–4269.

Gao LZ, Xu H (2008) Comparisons of mutation rate variation at genome-wide microsatellites: evolutionary insights from two cultivated rice and their wild relatives. *BMC Evolutionary Biology*, **8**, 11.

Gatherer D, Seirafian S, Cunningham C *et al.* (2011) High-resolution human cytomegalovirus transcriptome. *Proceedings of the National Academy of Sciences, USA*, **108**, 19755–19760.

Griffiths P, Baraniak I, Reeves M (2015) The pathogenesis of human cytomegalovirus. *The Journal of Pathology*, **235**, 288–297.

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *The American Journal of Human Genetics*, **72**, 1527–1535.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Innan H, Stephan W (2003) Distinguishing the hitchhiking and background selection models. *Genetics*, **165**, 2307–2312.

Jurak I, Brune W (2006) Induction of apoptosis limits cytomegalovirus cross-species infection. *EMBO Journal*, **25**, 2634–2642.

Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. *Genetics*, **123**, 887–899.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics*, **146**, 1197–1206.

Kim Y (2006) Allele frequency distribution under recurrent selective sweeps. *Genetics*, **172**, 1967–1978.

Kim Y, Stephan W (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, **155**, 1415–1427.

Kulesza CA, Shenk T (2006) Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. *Proceedings of the National Academy of Sciences*, **103**, 18302–18307.

Linz B, Balloux F, Moodley Y *et al.* (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, **445**, 915–918.

Lohmueller KE, Albrechtsen A, Li Y *et al.* (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genetics*, **7**, e1002326.

Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, **107**, 961–968.

Lynch M, Sung W, Morris K *et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences*, **105**, 9272–9277.

Magni GE (1964) Origin and nature of spontaneous mutations in meiotic organisms. *Journal of Cellular and Comparative Physiology*, **64** (Suppl 1), 165–171.

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EAR (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *Journal of Molecular Biology*, **247**, 443–458.

McVean GAT, Myers SR, Hunt S *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**, 581–584.

Miralles R, Gerrish PJ, Moya A, Elena SF (1999) Clonal interference and the evolution of RNA viruses. *Science*, **285**, 1745–1747.

Mocarski ES, Shenk T, Pass RF (2007) Cytomegaloviruses. In: *Fields Virology* (eds Knipe DM, Howley PM), pp. 1960–2015. Lippincott Williams & Wilkins, Philadelphia, Pennsylvania.

Montano V, Didelot X, Foll M *et al.* (2015) Worldwide population structure, long term demography, and local adaptation of *Helicobacter pylori*. *Genetics*, **200**, 947–963.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics*, **185**, 907–922.

Pfeifer B, Wittelsbürger U, Ramos Onsins SE, Lercher MJ (2014) PopGenome: an efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, **31**, 1929–1936.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics*, **160**, 1179–1189.

Puchhammer-Stöckl E, Görzer I (2011) Human cytomegalovirus: an enormous variety of strains and their possible clinical significance in the human host. *Future Virology*, **6**, 259–271.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran S, Campo DS, Dimitrova ZE *et al.* (2011) Temporal variations in the hepatitis C virus intrahost population during chronic infection. *Journal of Virology*, **85**, 6369–6380.

Reeves MB, Davies AA, McSharry BP, Wilkinson GW, Sinclair JH (2007) Complex I binding by a virally encoded RNA regulates mitochondria-induced cell death. *Science*, **316**, 1345–1348.

Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF (2011) Extensive genome-wide variability of human cytomegalovirus in congenitally infected infants. *PLoS Pathogens*, **7**, e1001344.

Renzette N, Gibson L, Bhattacharjee B *et al.* (2013) Rapid intrahost evolution of human cytomegalovirus is shaped by demography and positive selection. *PLoS Genetics*, **9**, e1003735.

Renzette N, Pokalyuk C, Gibson L *et al.* (2015) Limits and patterns of cytomegalovirus genomic diversity in humans. *Proceedings of the National Academy of Sciences, USA*, **112**, E4120–E4128.

Roselius K, Stephan W, Städler T (2005) The relationship of nucleotide polymorphism, recombination rate and selection in wild tomato species. *Genetics*, **171**, 753–763.

Ross SA, Novak Z, Pati S *et al.* (2011) Mixed infection and strain diversity in congenital cytomegalovirus infection. *Journal of Infectious Diseases*, **204**, 1003–1007.

Shi W, Freitas IT, Zhu C *et al.* (2012) Recombination in hepatitis C virus: identification of four novel naturally occurring inter-subtype recombinants. *PLoS One*, **7**, e41997.

Stanton RJ, Baluchova K, Dargan DJ *et al.* (2010) Reconstruction of the complete human cytomegalovirus genome in a BAC reveals RL13 to be a potent inhibitor of replication. *Journal of Clinical Investigation*, **120**, 3191–3208.

Stark TJ, Arnold JD, Spector DH, Yeo GW (2012) High-resolution profiling and analysis of viral and host small RNAs during human cytomegalovirus infection. *Journal of Virology*, **86**, 226–235.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 1245–1253.

Stern-Ginossar N, Weisburd B, Michalski A *et al.* (2012) Decoding human cytomegalovirus. *Science*, **338**, 1088–1093.

Strelkowa N, Lässig M (2012) Clonal interference in the evolution of influenza. *Genetics*, **192**, 671–682.

Whitley RJ (1996) Chapter 68: Herpesviruses. In: *Medical Microbiology*, 4th edn (ed. Baron S), pp. 679–688. University of Texas Medical Branch, Galveston, Texas.

Wirth T, Hildebrand F, Allix-Béguec C *et al.* (2008) Origin, spread and demography of the *Mycobacterium tuberculosis* complex. *PLoS Pathogens*, **4**, e1000160.

Worobey M, Rambaut A, Pybus OG, Robertson DL (2002) Questioning the evidence for genetic recombination in the 1918 "Spanish flu" virus. *Science*, **296**, 211, discussion 211.

N.R., T.F.K. and J.D.J. designed and performed the research and wrote the manuscript.

Data accessibility

All nucleotide sequences used in this study are publically available and Accession nos are provided in Table S1 (Supporting information). DNA sequence alignments used in this study have been uploaded to Dryad with identifier doi: 10.5061/dryad.88755.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Strains and sequence data used in this study.

Table S2 Linear model fit of intraspecies diversity model.

Fig. S1 HCMV genome-wide recombination rates.

Fig. S2 Log transformation of intraspecies diversity (π).

Fig. S3 Log transformation of population-scaled recombination rates ($2N_e r$).

Fig. S4 Log transformation of interspecies divergence (D_{xy}).

Fig. S5 CLR values from neutral simulations.

Fig. S6 Correlation of HCMV intraspecies diversity, interspecies divergence and recombination rates from a thinned dataset.