MOLECULAR ECOLOGY

Molecular Ecology (2017) **26**, 1980–1990 doi: 10.1111/mec.13953

MICROBIAL LOCAL ADAPTATION

# Characterizing human cytomegalovirus reinfection in congenitally infected infants: an evolutionary perspective

CORNELIA POKALYUK,*† ‡ § NICHOLAS RENZETTE,§¶ KRISTEN K. IRWIN,‡ § SUSANNE P. PFEIFER,‡ §** LAURA GIBSON,† † WILLIAM J. BRITT,‡ ‡ APARECIDA Y. YAMAMOTO,§§ MARISA M. MUSSI-PINHATA,§§ TIMOTHY F. KOWALIK¶ and JEFFREY D. JENSEN‡ §**

*Institute for Mathematics, Goethe Universität Frankfurt, Frankfurt am Main, Germany, †Faculty for Mathematics, Otto-von-Guericke University Magdeburg, Magdeburg, Germany, ‡School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, §Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, ¶Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, MA, USA, **School of Life Sciences, Arizona State University, Tempe, AZ, USA, ††Departments of Medicine and Pediatrics, Divisions of Infectious Diseases and Immunology, University of Massachusetts Medical School, Worcester, MA, USA, ‡‡Department of Pediatrics, University of Alabama Birmingham, School of Medicine, Birmingham, AL, USA, §§Department of Pediatrics, Ribeirão Preto Medical School, University of São Paulo, São Paulo, Brazil

## Abstract

**Given the strong selective pressures often faced by populations when colonizing a novel habitat, the level of variation present on which selection may act is an important indicator of adaptive potential. While often discussed in an ecological context, this notion is also highly relevant in our clinical understanding of viral infection, in which the novel habitat is a new host. Thus, quantifying the factors determining levels of variation is of considerable importance for the design of improved treatment strategies. Here, we focus on such a quantification of human cytomegalovirus (HCMV) – a virus which can be transmitted across the placenta, resulting in foetal infection that can potentially cause severe disease in multiple organs. Recent studies using genomewide sequencing data have demonstrated that viral populations in some congenitally infected infants diverge rapidly over time and between tissue compartments within individuals, while in other infants, the populations remain highly stable. Here, we investigate the underlying causes of these extreme differences in observed intrahost levels of variation by estimating the underlying demographic histories of infection. Importantly, reinfection (*i.e.* population admixture) appears to be an important, and previously unappreciated, player. We highlight illustrative examples likely to represent a single-population transmission from a mother during pregnancy and multiple-population transmissions during pregnancy and after birth.**

*Keywords*: evolutionary medicine, molecular evolution, population genetics – empirical

*Received 22 May 2016; accepted 12 December 2016*

## Introduction

Variation is the raw material on which selection may act. Thus, when faced with a novel selective pressure, variation increases the adaptive potential of the population. Although discussed at some length in the literature with regard to the preservation of variation in endangered species for example, this notion is equally important clinically when considering pathogenic infections. As such, understanding the underlying demographic histories characterizing observed levels of variation will be of value in dictating improved treatment strategies seeking to limit such within-host diversity to prevent viral success in the host. One of the most important targets of

Correspondence: Jeffrey D. Jensen, E-mail: jeffrey.d.jensen@asu.edu and Timothy F. Kowalik, E-mail: timothy.kowalik@umassmed.edu

such research is human cytomegalovirus (HCMV) – a β-herpesvirus with a large 235-kb genome (Kenneson & Cannon 2007), and the most common source of congenital (*i.e.* before birth) infection.

In HCMV, initial or 'primary' infection can occur at any age and usually causes no clinical symptoms; however, in foetuses, infection can be severe and may result in infant hearing loss as well as neurodevelopmental delay among other maladies (Hassan & Connell 2007). Reinfection or 'secondary' infection is possible, and as the connection between mother and foetus persists throughout pregnancy, transplacental viral transmission could happen repeatedly or even continuously (Enders *et al.* 2011). In addition, reinfection may also occur after birth through contact with other infected individuals or from the mother through breastfeeding (Numazaki 1997). However, there are currently no methods to prevent transmission from mother to foetus or to reduce severity of disease in the infant; thus, a better understanding of the evolutionary processes underlying infections themselves could be the key to future treatments (see review of Renzette *et al.* 2014).

Whatever the timing of infection or immune status of the host, the virus persists for life, during which time multiple genetically distinct viral populations can be established in separate organ systems through a phenomenon known as compartmentalization (Zárate *et al.* 2007). During primary infection, the viral population tends to be relatively stable, at least across commonly studied genomic regions (Murthy *et al.* 2011). If the host is later reinfected, a new viral population may admix with the existing population. As HCMV is highly diverse (Sijmons *et al.* 2015; Renzette *et al.* 2017), this new population will likely be significantly different on the sequence level from the viral population already harboured. If this newly introduced population survives, these multiple, distinct viral subpopulations may exist concurrently in reinfected hosts. In other words, the structure of the viral populations in the host may serve as a marker of reinfection. However, given the relatively high estimated rates of recombination in HCMV (see Renzette *et al.* 2016), such structure should be observed on the subgenomic rather than whole-genomic level.

The majority of studies on HCMV genetic diversity to date have focused on a limited number of viral genes, demonstrating that the viral population of a single host may contain regions of several genotypes – a finding typically described in the literature as 'mixed' infection. In particular, Ross *et al.* (2011) found evidence of mixed infections in 45% of congenitally infected infants. Similarly, using whole-genome sequencing, Renzette *et al.* (2013) demonstrated that viral populations in some congenitally infected infants diverge drastically over time and between tissue compartments in a single individual, while in other infants, the populations remain highly stable (see Fig. S1, Supporting information). However, to date, the primary causes of mixed infections and population divergence are not well understood. Here, the occurrence of stable populations in congenitally infected infants is explained through transmission of a viral population from a primary maternal infection or the transmission of only a single viral subpopulation of nonprimary maternal infection. In contrast, divergent viral populations in congenitally infected infants, which we also explore, may result from two possible scenarios: (i) transmission of multiple viral populations from a nonprimary maternal infection or (ii) reinfection of the infant after birth. Population genetic models were constructed to represent these distinct scenarios of infection histories, and simulation tools were applied to determine the best-fit model for genomewide viral sequencing data from congenitally infected infants. This study highlights the capacity of population genetic approaches to provide clinically relevant estimates of the dynamics of virus infection, the results of which have important implications for the future prevention and treatment of congenital HCMV infection.

## Materials and methods

### *Patient samples*

Samples from three patients (B103, M74 and 1254) were collected and analysed (Table 1). From patient B103 described in Renzette *et al.* (2013), plasma and urine samples were collected 1 week and 6 months after birth and denoted B103 P1W (plasma 1 week), B103 U1W (urine 1 week), B103 P6M (plasma 6 months) and B103 U6M (urine 6 months). From patient M74, saliva samples were collected at 1 week (M74 S1W) and 31 months (M74 S31M) and urine samples at 2.5 months (M74 U2.5M) and 12 months (M74 U12M) after birth. From patient 1254, a saliva sample was collected at 4.5 months (1254 S4.5M) and urine samples at 3 weeks (1254 U3W), 1 month (1254 U1M), 4 months (1254 U4M), 6 months (1254 U6M) and 24 months (1254 U24M) after birth.

### *Sequencing*

*High-throughput sequencing.* The samples from patient B103 were sequenced on a Illumina GAII platform as specified previously in Renzette *et al.* (2013). Samples collected from patient M74 and 1254 were sequenced on an Ion Torrent platform using the genomewide PCR amplification strategy as described in Renzette *et al.* (2015). Sequence reads have been deposited into the

**Table 1** Sample collection points

| Patient | Compartment | 1 week | 3 weeks | 1 month | 2.5 months | 4 months | 4.5 months | 6 months | 12 months | 24 months | 31 months |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1254 | Saliva | | | | | | X | | | | |
| | Urine | | XX | X | | X | | X | | XX | |
| B103 | Plasma | XX | | | | | | X | | | |
| | Urine | XX | | | | | | XX | | | |
| M74 | Saliva | XX | | | | | | | | | XX |
| | Urine | | | | X | | | | XX | | |

Xs indicate that a patient sample was available for the given compartment at the given time postbirth; a second X indicates samples for which *UL73* was specifically amplified and sequenced.

sequence read archive (SRA), with Accession numbers listed in Renzette *et al.* (2013) (for B103) and Renzette *et al.* 2015 (for M74 and 1254). Error controls were included in all sequencing reads, as described previously.

*Sanger sequencing.* The *UL73* region was amplified from patient samples (as indicated in Table 1), and clones were generated using the StrataClone Blunt PCR Cloning Kit (Agilent Technologies) following the manufacturer's recommended protocol. At least 16 clones from each patient sample were then Sanger sequenced by Genewiz, Inc. (www.genewiz.com). Genotypes were classified according to Pignatelli *et al.* (2003).

### *Alignment*

Characterizing the genetic variability in high-throughput sequencing data stemming from pooled samples is a challenging task. A commonly applied strategy involves an initial alignment of the reads to a reference genome, followed by a stepwise modification of the reference genome to a so-called consensus sequence, consisting of the most frequent nucleotide at each position of the sample. At positions sufficiently covered, this consensus sequence likely reveals the most common variants of the sample. However, as this approach may fail to recover the genetic variability stemming from distinct viral populations with very different underlying genomes (Renzette *et al.* 2017), this study used 26 minimally passaged whole-genome sequences from patient samples as reference genomes instead of creating a consensus sequence; GenBank numbers are given in the Supplementary Materials.

Reads were aligned to one of the 26 reference genomes (arbitrarily chosen) using Bowtie 2 (Langmead & Salzberg 2012). If this initial alignment failed, another reference was chosen for the read mapping until either the read was aligned or no more reference genomes were available. To infer single nucleotide polymorphisms (SNPs) in a first step, GenBank reference sequences were matched as follows: To reduce the introduction of gaps, which would lead to an underestimation of the number of SNPs, annotations of the GenBank sequences were used to align from one annotated region, present in all reference genomes, to the next using the program CLUSTALW (Larkin *et al.* 2007). These subgenomic alignments were then concatenated in to single alignments. As repeat regions are difficult to align in general, the repeat regions IRL, IRS, TRL and TRS of the HCMV genome (representing about 7700 bp) were masked. In a second step, SNP frequencies were calculated at each position of the aligned reference genomes, if at least 15 reads were aligned. If more than

two variants were present at a given position, only the first two variants were considered, as these other variants did not exceed the threshold of 1% and thus were not distinguishable from sequencing errors. In Fig. 1, a graphical representation of the used algorithm is presented, and the command line used can be found in the Supplementary Materials.

### *Demographic history estimation*

In this study, we were interested in distinguishing between demographic histories stemming from infections of either a single or multiple distinct viral populations (see schematic representation in Fig. 2). The program FASTSIMCOAL2 (Excoffier *et al*. 2013) was used to infer the demographic histories of viral populations (*e.g.* population growth or decline, structure, migration or admixture) by simulating joint site frequency spectra for population histories given by the model described below and comparing these spectra to those observed in the data (Fig. 3).

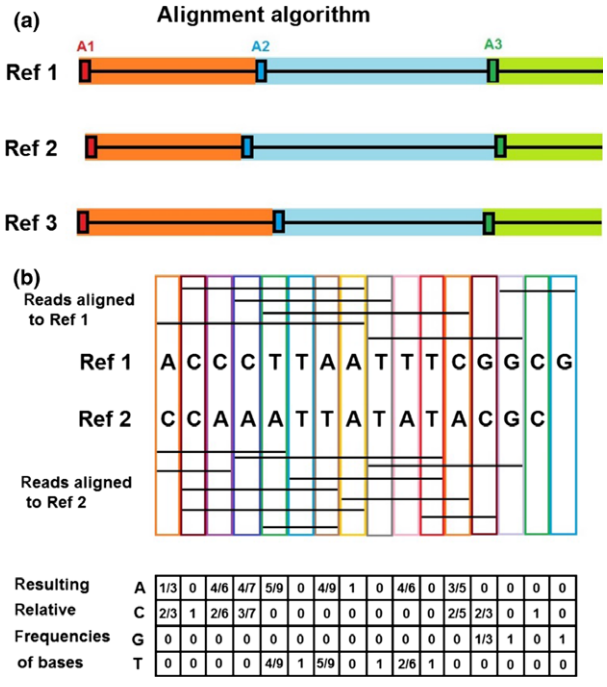



**Fig. 1** Schematic representation of the alignment algorithm. (a) Reference sequences are separated into segments reaching from the beginning of an annotated region (here A1, A2 and A3), present in all references, to the next. Corresponding segments (*e.g.* all orange segments) of the different reference sequences are then aligned to one another. (b) Read positions mapped to the same position of the aligned sequences are collected together to calculate relative frequencies of bases (in this example, the rule of a minimum read depth of 15 is not observed).

The maternal viral population was assumed to initially infect the circulatory system of the foetus (represented by the plasma samples). At this first infection event, allowance was made for a population bottleneck occurring no earlier than conception, following previous estimates (Renzette *et al.* 2013). From the circulatory system, the viral population was assumed to subsequently infect the salivary glands and kidneys (Smith *et al.* 2004; Bentz *et al.* 2006), the timing of which was inferred. After infection of these compartments, populations were allowed to decline or grow exponentially: given two (backward) time points $t$ and $s$ with $t < s$ (*i.e.* time $s$ lies further back in the past than time $t$), a growth rate $g$ (per individual per day) and an effective population size $N_e$ at time $t$ within a compartment, then the effective population size $N'_e$ at time $s$ in the same compartment is given by

$$N'_e = N_e * \exp((t - s) * g). \quad (1)$$

For example, if at day 150 after infection (time $t$), the effective population size ($N_e$) is 1000, and the growth rate ($g$) is 0.0001, then at day 100 postinfection (time $s$), the population size was 995. Finally, the population size at the final sampling time was estimated for each compartment. Note that together with the estimation of the growth rate, the effective population size can then be calculated at any time point before and after the final sampling.

Inference procedures like FASTSIMCOAL2 assume that the population is panmictic, or well mixed, and evolves in equilibrium with respect to a neutral Wright–Fisher model before the defined history begins. If a population consists of several viral subpopulations at sampling, the panmictic population assumed by FASTSIMCOAL2 would initially be split into subpopulations, which would evolve independently of one another for some time, accumulating genetic differences, and then admix within the mother and/or foetus during pregnancy, or the infant postnatally. In FASTSIMCOAL2, admixture between two populations can be realized as (i) discrete admixture events in which the second population replaces a part of the first population or (ii) continuous migration between the two populations. Here, the case of a single discrete admixture event is considered. Schematic representations of population histories with admixture events are given in Fig. 2 (demographic history reconstructions for patients B103 and M74).

Due to the high diversity of HCMV, it is often not clear which state is ancestral and which derived; thus, parameter inference is based on the minor (*i.e.* folded) joint site frequency spectra. In this case, for each position in each sample, the frequency of the base was calculated and summed up over all samples. As SNPs are

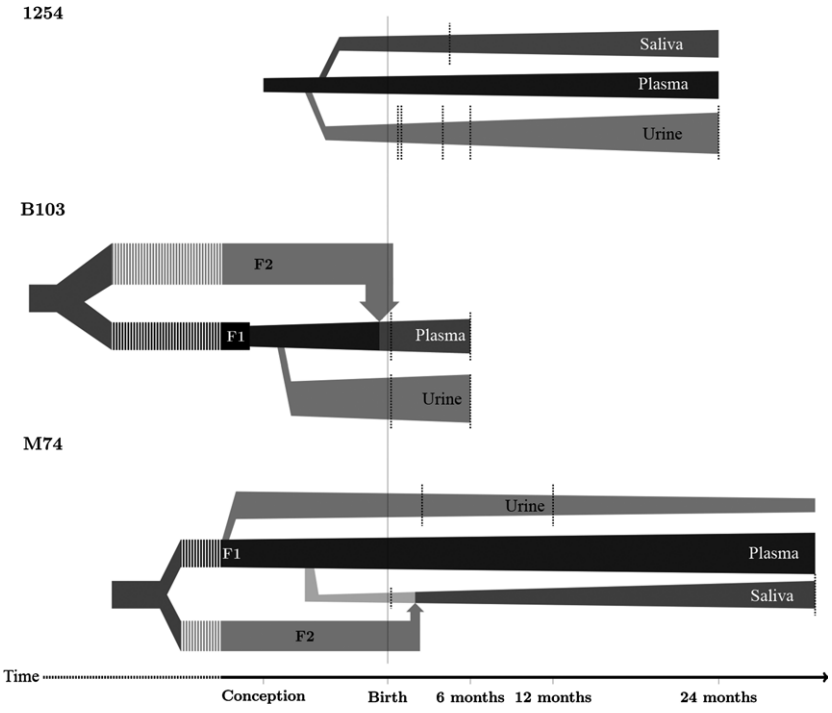



**Fig. 2** Inferred demographic histories of three patients congenitally infected with HCMV. Horizontally aligned shapes indicate the size (width) and age (length) of populations; age is either since a split or since the most recent bottleneck, where bottlenecks represent either infection of the foetus (1254 and B103) or of the mother (M74). Arrows indicate admixture events from outside populations (F2) into the initial infecting population (F1), where arrow width corresponds to the proportion of the new population replaced by the virions from the outside populations, and the subsequent populations represent the postadmixture presence of multiple populations. The solid black vertical line represents birth, and dashed black lines represent sampling time points. Dashed time line represents compressed time.

assumed to be biallelic, the base with the second highest frequency was taken as the minor variant in the population. If the highest and second highest frequency were equal, both bases contributed to the JSFS with a factor 0.5, as described in the fastsimcoal manual, version 2.5.2.21. The usage of $d$-dimensional SFS with $d > 2$ is possible in fastsimcoal, but requires a very large number of SNPs. In our case, there exist too few SNPs for such calculations. Frequencies were calculated from pooled sequencing data by projecting SNP frequencies down into 15 bins, the minimum depth restriction for alignment given the required coverage of 15 reads. Using HCMV-BAC resequencing data as a measure of sequencing error, SNPs called in this manner have a false-positive probability $\leq 0.00625$. To estimate parameters, 200 cycles of a conditional maximization algorithm (ECM, Meng & Rubin 1993) were performed, each based on 10,000 simulations. Monomorphic sites in observed JSFS were masked; consequently, parameter inference was only based on JSFS, ignoring mutation rates. Bias-corrected 99% confidence intervals were calculated for log-transformed data via the formula $[\exp(2\cdot\phi-\mu-2.576\cdot\sigma), \exp(2\cdot\phi-\mu+2.576\cdot\sigma)]$ using parametric bootstrapping, where $\mu$ is the mean of and $\sigma$ the standard deviation of the bootstrap results and $\phi$ is the fastsimcoal estimate obtained from the data (Efron & Tibshirani 1993). Time spans in Table 2 are given in backward calendar time assuming an HCMV doubling time of 1 day (Emery *et al.* 1999), starting from the latest sampling event.

An exemplary input file and summary outputs of program runs can be found in Table 2, in Tables S1–S3 (Supporting information) and in the Supplemental File 'Supp_code.txt'.

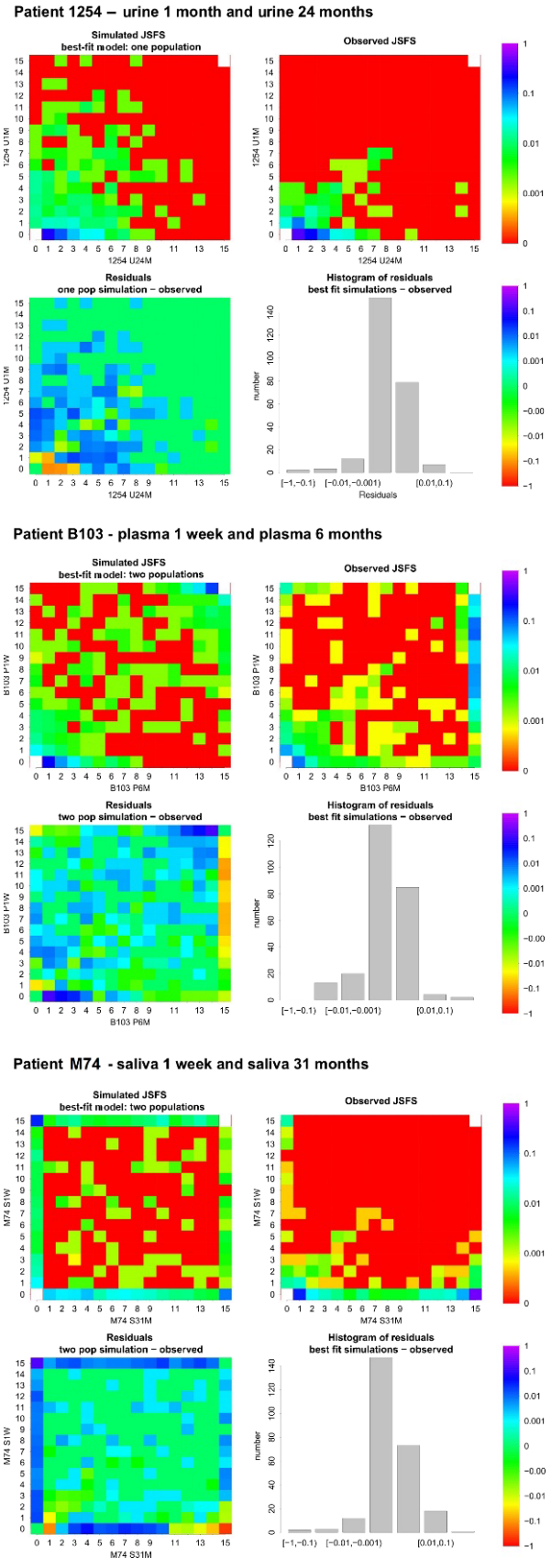
To compare the simulated and observed SFS in detail for each patient, the best-fitting models in each scenario were further simulated with FASTSIMCOAL2.

### *Comparison of best-fit models*

To assess both the best-fitting models and the best-fitting alternative models relative to the observed data, each model was simulated 100 times and the (composite) log-likelihood of the observed data was estimated using FASTSIMCOAL2, with the same parameters as used for the demographic history estimation. In this manner, a distribution of simulated log-likelihoods was obtained for the observed data given either the best-fit or the alternative best-fit model. As simulated log-likelihoods were approximately normally distributed, the probabilities to observe a smaller log-likelihood for the best-fit model relative to the alternative model were calculated.

## Results

High-throughput genomic sequencing data of HCMV sampled from urine, saliva and/or plasma of three patients (B103, M74 and 1254) were analysed. See Table 1 for sampling details.




### *Alignment using multiple references*

As described in the Methods section, instead of taking the traditional approach of creating a consensus sequence, several alignment references were used. To compare the two approaches, the SNP frequencies of the 1-week plasma samples of patient B103 were identified using multiple references or using a single reference from which a consensus was created. When using multiple reference genomes to identify SNPs, greater genomic variability was detected, because HCMV diversity was better represented by several reference sequences (Fig. S2, Supporting information). As a result, more intermediate frequencies ($f < 0.9$) were identified. Overall, when using a single reference, intermediate frequencies were detected at 806 positions, but when using multiple references, such frequencies were detected at 1266 positions, an increase of ~50%.

### *Detection of viral subpopulations*

Viral population histories were inferred for the three patients (Fig. 2 and Table 2), and the most likely scenario for each patient is summarized as follows. For patient 1254, the best-fitting model is that of a single-population infection of the foetus (AIC(1 population) = 22 128, AIC(2 subpopulations, admixed in urine) = 22 184, AIC(2 subpopulations, admixed in saliva) = 22 200; see also Table S1 (Supporting information) for parameter estimates of other models). That population then infects the kidneys (urine compartment) and salivary glands. For patient B103, the best-fitting model supports an infection with two subpopulations with admixture in plasma (AIC(2 subpopulations admixed in plasma) = 28 286, AIC(1 population) = 29 112, AIC(2 subpopulation admixed in urine) = 29 094; see also Table S2, Supporting information). Specifically, the foetus was infected with two distinct populations – the first during gestation and the second during gestation or perinatally during gestation. Finally, for patient M74, an infection with two subpopulations with admixture in saliva is the best-fitting model (AIC(2 subpopulations admixed in saliva) = 29 172, AIC(1 population) = 32 018 AIC(2 subpopulations

**Fig. 3** Joint site frequency spectra from temporally distant samples. For each patient, the two most temporally distant samples were compared, and the observed joint SFS between them as well as simulated joint SFS according to the best-fit model are plotted. The SFS predicted from simulation is also compared to the observed SFS in the adjacent panels. Comparisons among other sample pairs are available in the Supplementary Materials (Figs S3–S5, Supporting information).

**Table 2** Demographic estimates for all patients

| | 1254 | | | B103 | | | M74 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Parameter | Estimate | CI | Parameter | Estimate | CI | Parameter | Estimate | CI |
| Compartmental Population Size | NU24M | 293 | 108, 355 | NP6M | 530 | 119, 1611 | NS31M | 603 | 40, 16e3 |
| | NS4.5M | 226 | 1, 4309 | NU6M | 1304 | 705, 6101 | NU12M | 50 | 1, 196 |
| Size of admixture population F2 | | | | NF2 | 11 041 | 2451, 5e4 | NF2 | 436 | 44, 664 |
| Time since Population Splits | TSplitU_S | 918 | 441, 1502 | TSplitF1_F2 | 38 932 | 7636, 1.7e5 | TSplitF1_F2 | 14 932 | 2909, 2e5 |
| | | | | TSplitP_U | 449 | 350, 660 | TSplitU_S | 1461 | 1317, 2244 |
| Time since Admixture Event | | | | | | | | | |
| | | | | TAdmixF_2_P | 198 | 152, 240 | TAdmixF2_S | 877 | 205, 1.2e4 |
| Prop. of Ancestral Population Replaced | | | | PropF2_P | 0.97 | 0.92, 1 | PropF2_S | 0.35 | 0.21, 0.61 |
| Time and Size of Ancestral Pop. Bottleneck | TBotF1 | 1039 | 1014, 1769 | TBotF1 | 491 | 449, 571 | TBotF1 | 1476 | 1307, 2266 |
| | SizeBot | 430 | 108, 7286 | SizeBot | 433 | 168, 2414 | SizeBot | 258 | 71, 438 |
| Growth Rates by Compartment | GrU | 0.015 | 0.0035, 0.019 | GrP | 1.6e-4 | −0.012, 0.015 | GrS | 6.3e-4 | −0.003, 0.003 |
| | GrS | 0.0015 | −0.002, 0.0092 | GrU | 9.17e-3 | 0.003, 0.02 | GrU | −0.0034 | −0.012, −0.001 |

*NX* denotes the effective population size of sample *X*, and *NF2* is the effective population size of the admixture population *F2*. *GrU*, *GrS* and *GrP* denote the growth rates of the urine, saliva and plasma populations, respectively. All times (*T*) are listed in days and based on a 24-h generation time for HCMV (Emery *et al.* 1999). *TBot* is the (backward) time of a bottleneck in the founder population F1, which may correlate to infection of the mother or infection of the foetus, based on age of the bottleneck. *SizeBot* denotes the effective size of population F1 at the bottleneck. *TAdmixX_Y* denotes the (backward) time in days of the admixture events between populations X and Y. *PropX_Y* denotes the proportion of the population X being replaced during the admixture event with population Y. *TSplitX_Y* denotes the (backward) time the split between populations X and Y. The column labelled CI gives bias-corrected 99% confidence intervals.

admixed in urine) = 31 594; see also Table S3, Supporting information), characterized by an initial infection during pregnancy and later reinfection with another population after birth, as detected in the second saliva sample. Importantly, the splitting time of the two populations were both estimated to be more than 40 years in the past, indicating transmission of distinct populations.

### *Simulated JSFS and observed data*

To investigate the fit of the inferred population histories, the best-fitting models were simulated and compared to the observed JSFS. Importantly, population inference programs like FASTSIMCOAL2 make the assumption of independent sites. As recombination rates are estimated to be high in HCMV, this assumption is likely justified. In general, the inferred population histories well reproduce observation (Fig. 3).

### *Comparison of best-fit model and alternative models*

The model of a single population is a special case of the model with two subpopulations (with admixture size = 0). Hence, in the case of patient B103 and patient M74, we applied the log-likelihood ratio test based on (composite) log-likelihood approximation by FASTSIMCOAL2 to determine whether the more complex model fits significantly better than the simple one population model. In both cases, the *P*-values were smaller than $10^{-10}$.

In addition, we calculated the probabilities to observe smaller log-likelihoods under a simulation for the best-fit model than for the alternative best-fit model as described in the Materials and Methods section. For patient B103 and M74, these probabilities are less than 0.28% and 0.03%, respectively, and for patient 1254, this probability is less than 31%.

### *Mixed infections vs. infections with several viral subpopulations*

In the literature, HCMV infection is referred to as 'mixed' when several distinct genotypes are detected in specific genomic regions within a single individual. Frequently reported regions include *UL73*, *UL55*, *UL75* and *UL74,* which code for glycoproteins gN, gB, gO and gH, respectively (*e.g.* Pignatelli *et al.* 2003; Ross *et al.* 2011). To compare the whole-genomic analysis to the standard subgenomic version, the *UL73* region was sequenced. To avoid the loss of linkage information and subsequent difficulty inferring genotypes from pooled sequenced data, multiple full-length clones of the *UL73* (gN) ORF were sequenced from three samples of each patient. Results are summarized in Table 3. In patient 1254, the same genotype was observed in all three samples, consistent with our above inference of a single-population infection. In patient B103, the same genotype was observed in both urine samples, with a second genotype in the 6 month plasma sample. In patient M74, the same genotype was observed in the two early saliva and urine samples, with the late saliva sample being a mixture of genotypes 4a and 3b. Thus, with respect to *UL73,* patients B103 and M74 would be classified as mixed infections, also consistent with our inference of a multiple-population infection. However, we emphasize the importance of whole-genome data for inferring different infection scenarios.

## Discussion

HCMV populations of congenitally infected infants can range from being relatively stable to quite divergent, which can differ on a local scale, on a whole-genomic scale, or over time (see Ross *et al.* 2011; Renzette *et al.* 2013). Importantly, the presence of multiple variants during congenital HCMV infection has been correlated with lethal outcomes during gestation (Arav-Boger *et al.* 2002). Despite this clinical importance, the mechanisms underlying divergence in viral populations, and why diverse infections are found only in some infants, remain unclear. In this study, we analysed the role of reinfection in producing these patterns and discuss the impact of reinfection on within-host evolutionary dynamics.

There is increasing evidence that congenital HCMV infections are not characterized by single, short transmission events, but rather by transmission over a longer time period and/or in several distinct events during

**Table 3** Genotyping results of region *UL73*

| Patient | 1254 | 1254 | 1254 | B103 | B103 | B103 | M74 | M74 | M74 |
|---|---|---|---|---|---|---|---|---|---|
| Sample | S4.5M | U3W | U24M | U1W | U6M | P6M | S1W | S31M | U12M |
| *UL73* Genotype | 3b | 3b | 3b | 3a | 3a | 1 | 4a | 4a (87.5%), 3b (12.5%) | 4a |

Genotypes derived from clonal sequencing found in the genomic region UL73 in different samples from patients 1254, B103 and M74.

pregnancy. Studies demonstrating HCMV in the human placenta and cord blood (*e.g.* Pereira *et al.* 2014), and those demonstrating viral replication in the placenta in the presence of specific maternal antibodies in the guinea pig model (Griffith *et al.* 1985), suggest that the virus can persist in placental tissues and potentially transmit to the foetus long after HCMV has cleared from the maternal bloodstream. The inference of population histories reported here supports this hypothesis. In particular, we find evidence that reinfection of the foetus may lead to a host population consisting of multiple viral subpopulations, potentially explaining the rapid divergence observed over time and between tissue compartments in certain individuals. As transmission occurs frequently from mothers with both primary infections and reinfections (Manicklal *et al.* 2013), we hypothesize that reinfection of the foetus should also be common after maternal reinfection and/or reactivation of latent virus during pregnancy. Further, reinfection may also occur after birth, as inferred from a saliva sample of patient M74 – consistent with epithelial surfaces being the most common site for postnatal transmission of virus.

To detect multiple viral populations in infant samples, we aligned reads to multiple whole-genome reference sequences rather than a single consensus sequence, which may represent only the most common population contained in the sample. We have shown that the number of positions of intermediate frequencies is substantially higher when several reference sequences are used. However, the 26 sequences used as references may not be entirely representative of the HCMV diversity. Others have studied HCMV specieswide diversity through the analysis of 41 (Renzette *et al.* 2015) and 96 (Sijmons *et al.* 2015) genomic sequences. However, the total number of polymorphisms and nucleotide diversity was similar between these studies and our collection of 26 sequences, suggesting that inclusion of additional reference sequences would only minimally impact the results.

Our results extend the analysis of Ross *et al.* (2011) by demonstrating that the viral populations of some congenitally infected infants are structured on a whole-genome scale. Here, this structure is explained by the scenario of a primary infection followed by reinfection, leading to the presence of admixed subpopulations. In contrast, a single primary infection is expected to result in a relatively homogeneous viral population. We avoid using the terms 'single strain' or 'multiple strain' infection, as the term 'strain' is generally used to identify the genomic type of a single virion. For time-sampled data, as in our study, such a term would not be appropriate because the viral genomes change over time – particularly given the estimated high rate of recombination in HCMV. We also avoid referring to 'mixed' infections, as the term does not necessarily reflect the genomewide structure of viral populations. For example, as mutation can create distinct genotypes in specific regions of the genome over time, multiple genotypes may be detected within highly variable regions, which would imply a mixed-type infection based on the regions sequenced. However, the genome as a whole may be relatively stable, and a single founder virion may be responsible for the infection. Hence, depending on the regions sequenced, an infection might be erroneously typed 'mixed' when the viral population is actually stable on a genomewide scale, and infection with several viral subpopulations is unlikely. Similarly, single genotypes may be detected within low variable regions, especially if the number of regions sequenced was not sufficiently large, implying an 'unmixed' infection when the viral population is indeed variable on a genomewide scale.

Finally, our study suggests that the potentially long period of time during which HCMV transmission may occur during pregnancy should be taken into account when testing the optimal timing or efficacy of therapeutic interventions. It has been shown that infection with multiple genotypes of murine CMV may lead to functional complementation within coinfected cells, increasing viral fitness (Čičin-Šain *et al.* 2005). Moreover, infection from reinfected mothers may lead to long-term sequelae even when infants are asymptomatic at birth (Williamson *et al.* 1992). (Re)infection by multiple viral populations and the consequent transmission of broad genomic diversity may hence be factors determining severity of congenital HCMV infection. Reduction rather than the complete elimination of transmitted viral populations may be sufficiently effective in preventing or reducing disease severity. In particular, fewer transmission events and the transmission of a restricted number of viral subpopulations may effectively reduce genomic variability in the foetal host, limiting the potential for adaptation and hence increasing the efficacy of treatment strategies.

## 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 (WJB, TFK), AI109001 (TFK); F32AI084437 (NR)) and the National Center for Advancing Translational Sciences at the National Institutes of Health (UL1TR000161). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank the University of Massachusetts Deep Sequencing and Molecular Biology Core Laboratories for their contribution and services. CP received financial support in part

from Deutsche Forschungsgemeinschaft (Priority Programme SPP 1590 'Probabilistic Structures in Evolution'). We thank Matthieu Foll and Lisha Mathew for helpful discussion.

## References

Arav-Boger R, Willoughby RE, Pass RF *et al.* (2002) Polymorphism of the cytomegalovirus (CMV)-encoded tumor necrosis factor-α and β-chemokine receptors in congenital CMV disease. *Journal of Infectious Diseases*, **186**, 1057–1064.

Bentz GL, Jarquin-Pardo M, Chan G, Smith MS, Sinzger C, Yurochko AD (2006) Human cytomegalovirus (HCMV) infection of endothelial cells promotes naive monocyte extravasation and transfer of productive virus to enhance hematogenous dissemination of HCMV. *Journal of Virology*, **80**, 11539–11555.

Čičin-Šain L, Podlech J, Messerle M, Reddehase MJ, Koszinowski UH (2005) Frequent coinfection of cells explains functional in vivo complementation between cytomegalovirus variants in the multiply infected host. *Journal of Virology*, **79**, 9492–9502.

Efron B, Tibshirani R (1993) *An Introduction to the Bootstrap*. Chapman & Hall, Boca Raton, Florida.

Emery VC, Cope AV, Bowen EF, Gor D, Griffiths PD (1999) The dynamics of human cytomegalovirus replication in vivo. *Journal of Experimental Medicine*, **190**, 177–182.

Enders G, Daiminger A, Bäder U, Exler S, Enders M (2011) Intrauterine transmission and clinical outcome of 248 pregnancies with primary cytomegalovirus infection in relation to gestational age. *Journal of Clinical Virology*, **52**, 244–246.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLOS Genetics*, **9**, e1003905.

Griffith BP, McCormick SR, Fong CK, Lavallee JT, Lucia HL, Goff E (1985) The placenta as a site of cytomegalovirus infection in guinea pigs. *Journal of Virology*, **55**, 402–409.

Hassan J, Connell J (2007) Translational mini-review series on infectious disease: congenital cytomegalovirus infection: 50 years on. *Clinical & Experimental Immunology*, **149**, 205–210.

Kenneson A, Cannon MJ (2007) Review and meta-analysis of the epidemiology of congenital cytomegalovirus (CMV) infection. *Reviews in Medical Virology*, **17**, 253–276.

Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie2. *Nature Methods*, **9**, 357–359.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Manicklal S, Emery VC, Lazzarotto T, Boppana SB, Gupta RK (2013) The "Silent" global burden of congenital cytomegalovirus. *Clinical Microbiology Reviews*, **26**, 86–102.

Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Murthy S, Hayward GS, Wheelan S *et al.* (2011) Detection of a single identical cytomegalovirus (CMV) strain in recently seroconverted young women. *PLoS ONE*, **6**, e15949.

Numazaki K (1997) Human cytomegalovirus infection of breast milk. *FEMS Immunology & Medical Microbiology*, **18**, 91–98.

Pereira L, Petitt M, Fong A *et al.* (2014) Intrauterine growth restriction caused by underlying congenital cytomegalovirus infection. *Journal of Infectious Diseases*, **209**, 1573–1584.

Pignatelli S, Dal Monte P, Rossini G *et al.* (2003) Human cytomegalovirus glycoprotein N (gpUL73-gN) genomic variants: identification of a novel subgroup, geographical distribution and evidence of positive selective pressure. *Journal of General Virology*, **84**, 647–655.

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, Gibson L, Jensen JD, Kowalik TF (2014) Human cytomegalovirus intrahost evolution – a new avenue for understanding and controlling herpesvirus infection. *Current Opinions in Virology*, **8C**, 109–115.

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.

Renzette N, Kowalik TF, Jensen JD (2016) On the relative roles of background selection and genetic hitchhiking in shaping human cytomegalovirus genetics diversity. *Molecular Ecology*, **25**, 403–413.

Renzette N, Pfeifer SP, Matuszewski S, Kowalik TF, Jensen JD (2017) On the analysis of intra-host and inter-host viral populations: human cytomegalovirus as a case study of pitfalls and expectations. *Journal of Virology*.

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.

Sijmons S, Thys K, Ngwese MM *et al.* (2015) High-throughput analysis of human cytomegalovirus genome diversity highlight the widespread occurrence of gene-disrupting mutations and pervasive recombination. *Journal of Virology*, **89**, 7673–7695.

Smith MS, Bentz GL, Alexander JS, Yurochko AD (2004) Human cytomegalovirus induces monocyte differentiation and migration as a strategy for dissemination and persistence. *Journal of Virology*, **78**, 4444–4453.

Williamson WD, Demmler GJ, Percy AK, Catlin FI (1992) Progressive hearing loss in infants with asymptomatic congenital cytomegalovirus infection. *Pediatrics*, **90**, 862–866.

Zárate S, Kosakovsky Pond SL, Shapshak P, Frost SDW (2007) Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. *Journal of Virology*, **81**, 6643–6651.

## Data accessibility

The sequences and associated read accession numbers reported in this article are from Renzette *et al.* (2013, 2015).

C.P., N.R., T.F.K. and J.D.J. designed the research; C.P., N.R., T.F.K. and J.D.J. performed research; L.G., W.J.B, A.Y.Y. and M.M.M.-P. contributed new reagents/analytic tools; C.P., N.R., S.P.P., T.F.K. and J.D.J. analysed data; and C.P., N.R., K.K.I., L.G., T.F.K. and J.D.J. wrote the manuscript.

## Supporting information

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

**Table S1** Patient 1254, all models.

**Table S2** Patient B103, all models.

**Table S3** Patient M74, all models.

**Fig. S1.** Stable vs. unstable viral populations.

**Fig. S2.** Top panels: For each position in the alignment frequencies of the bases ACGT in the sample B103 P1W were determined and the largest frequency was plotted (for A using 26 references and B using a consensus sequence).

**Fig. S3.** Patient 1254.

**Fig. S4.** Patient B103.

**Fig. S5.** Patient M74.

**Appendix S1.** Example input files, and summary output files of the programs run.