Diversity patterns in parasite populations capable for persistence and reinfection

Cornelia Pokalyuk[1] and Irene Görzer[2]

[1] *Mathematical Institute, Goethe-University Frankfurt, D-60629 Frankfurt am Main, Germany*
[2] *Department of Virology, Medical University of Vienna, A-1095 Vienna, Austria*

Abstract

Many parasites like the cytomegalovirus, HIV and E.coli are capable to persist in and reinfect its host. The evolutionary advantage of these (complicated) mechanisms have not been quantitatively analyzed so far. We make a first step into this direction by analyzing a hierarchical host-parasite model with two parasite types in which parasites persist in its hosts and reinfection of hosts is common. In [50] a scenario has been analyzed in which within hosts both parasite types are maintained due to balancing selection and parameter regimes have been identified in which persistence and reinfection can be effective mechanisms to maintain both parasite types in the parasite population. Here we analyze the neutral counterpart of the model with respect to the expected diversity patterns.

We evaluate the biological relevance of both scenarios at the example of the human cytomegalovirus (HCMV).We argue that under neutrality the coexistence of genotypes should be relatively rare in contrast to the observation in HCMV and the observed diversity patterns in HCMV-infected host require under neutrality reinfection rates which might be unrealistically high for HCMV.

1 Introduction

The evolution of a host-parasite system is often interpreted in spite of an arms-race strategy, claiming that host and parasite have to adapt constantly to guarantee their own survival. However, eventually such a behaviour should lead to the ultimate extinction of the parasite or the host (while in the latter case also the parasite dies, at least if the parasite does not adapt to a new host species before). To guarantee survival the parasite rather has to evolve a strategy, which serves two conflicting forces, on the one side avoiding host impairment to prevent one's own death and on the other side using host resources to guarantee parasite spread before the host dies, see [25]. In this case an equilibrium state can be achieved at which both species coexist.

A certain amount of genetic diversity, especially in coding regions, seems to be necessary for the long-term survival of a parasite, since otherwise the host can adapt to the parasite which eventually should lead to its extinction. Random reproduction in finite populations induces a permanent decrease in diversity, see [68]. Mutation is a mechanism that can restore this diversity. However, there is evidence that the bulk of the mutations happening are either neutral or deleterious for an individual, see [8]. Hence, it should be of advantage if parasites manage to evolve strategies that counteract the permanent decrease in diversity in regions where diversity is beneficial.

If a parasite is capable to persist in its host, via reinfections the diversity of the surrounding parasite population can be introduced step by step. A concrete model of such kind of parasite evolution has been introduced in [50]. In this model two parasite types exist and each host is infected with a fixed finite number of parasites. The parasite population evolves by host replacement, i.e. the death of hosts (and simultaneous death of the parasite population of this hosts) and (instantaneous) primary infection of so far uninfected hosts, reinfection of hosts, parasite reproduction within hosts and mutation of parasites from type A to type B and vice versa. In the model in [50]

within host parasite reproduction is driven by balancing selection, i.e. there is a tendency that once a host is infected with both parasite types both parasites types coexist in this host for a long time. For this model parameter regimes have been identified, for which persistence and reinfection are effective mechanism to maintain both parasite types in the parasite population.

Here we analyze the neutral counterpart of the model, in which parasite reproduction within hosts is neutral. In particular, we are interested in the expected diversity as well in the proportion of hosts expected to be infected with both parasite types.

We evaluate the biological relevance of the two scenarios at the example of the human cytomegalovirus. HCMV is an old herpesvirus which is wide-spread in the human population and leads in the typical host to an asymptomatic infection. Hence, HCMV seems to be a parasite which evolved a strategy that guarantees long-term coexistence with its host. For HCMV complex mechanisms guarantee that also in the healthy host the virus persists, reactivates and reinfects other hosts, see [17, 18, 33, 36].

For many open reading frames (ORF) of the HCMV genome, especially coding for immune evasion strategies and cell tropism, one finds a clustering into a few genotypes. Furthermore when these ORFs have been analyzed further, one observed that hosts are often infected with several genotypes. We argue that under neutrality the coexistence of genotypes might be relatively rare and a large proportion of hosts is infected with several genotypes only, if reinfection rates are relatively large (in contrast to a scenario with balancing selection). Such reinfection rates might be unrealistically high for HCMV.

The manuscript is organized as follows: First we explain the model for the evolution of a parasite population with persistence and reinfection and then summarize the results from [50], where the scenario with balancing selection has been analyzed. In the results section we analyze the neutral variant of the model and evaluate the biological relevance of the model on the basis of diversity patterns in DNA data of HCMV.

2 A model for parasite evolution with reinfection and persistence

We consider the following two variants of a model for the evolution of a parasite population capable for persistence and reinfection. In variant a) parasites reproduce according to a neutral reproduction mechanism and in variant b) parasite reproduction is affected by balancing selection. Variant b) has been analyzed in [50].

Consider a population of parasites distributed over M infected hosts, each host is infected with N parasites. We distinguish two types of parasites, type A and type B. Denote by $\mathbf{X}^{N,M} = (X_t^1, ..., X_t^M)_{t\geq 0}$ the relative frequencies of type A in the M different hosts.

Evolution is driven by the following four factors:

- (Parasite reproduction within hosts)
  - Variant a) neutral reproduction:

    Parasites reproduce at rate g_N. The evolution of the parasite population within each host follow dynamics which both are modifications of the classical Moran dynamics, see [23], Chapter 3.4. At reproduction a parasite splits into two and replaces with its offspring a randomly chosen other parasite. A change in frequency only occurs if the type of the reproducing parasite differs from the type of the parasite which is replaced. Hence, if the frequency of type A in host i is x (for some $x \in \{0, 1/N, ..., (N-1)/N, 1\}$) jumps due to parasite reproduction to $x + \frac{1}{N}$ occur at rate $g_N N x(1-x)$ and to $x - \frac{1}{N}$ at rate $g_N N x(1-x)$.
  - b) reproduction with balancing selection We assume there exists some equilibrium frequency η (with $\eta \in (0,1)$, at which type A and B are optimally balanced in a host. If the frequency of type A in host i is x each parasites of type A reproduces in this host at rate $g_N(1 + s_N(\eta - x))$ and parasites of type B at rate $g_N(1 + s_N(1 - \eta - (1-x))) =$

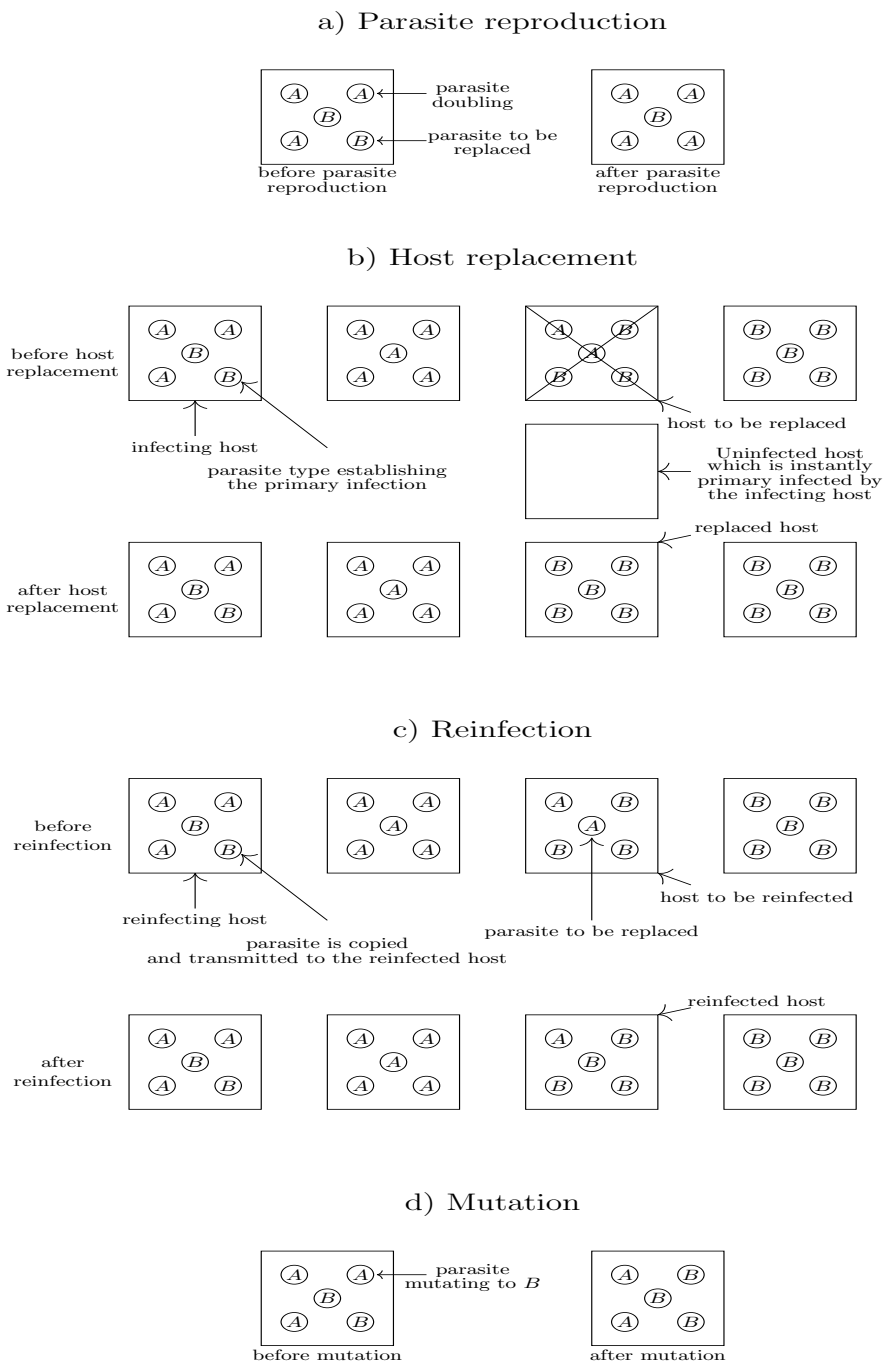

Figure 1: The parasite population evolves according to the four mechanisms a) parasite reproduction, b) host replacement, c) reinfection and d) mutation, see also the Definition in section 2: a) Parasites reproduce within hosts by cloning a parasite and simultaneously substituting a parasite. The rates at which parasites reproduce depend on the type of the parasite and the frequency of the parasite type in the host. b) Hosts die at a constant rate. When a host dies it is instantly replaced by a so far uninfected host, which is directly infected by a randomly chosen host. At primary infection only a single parasite type is transmitted. c) Hosts reinfect other (randomly chosen) hosts at a constant rate. At reinfection a single parasite is cloned and transmitted to the reinfected host. The transmitted parasite replaces a randomly chosen parasite in the reinfected host. d) Each parasite mutates at a (small) constant rate to type A and type B.

$g_N(1 - s_N(\eta - x))$, where $s_N > 0$ denotes the selection strength. That is the rate of reproduction of A-type parasites is larger than that of B-type parasites, if the frequency of type A is below the equilibrium frequency η, at which type A and type B are optimally balanced, and vice versa if the frequency is above the optimal frequency. If a parasite reproduces it doubles and replaces with its progeny a randomly chosen parasite in this host. A change in frequency only occurs if the type of the reproducing parasite differs from the type of the parasite which is replaced. Hence, if the frequency of type A in host i is x (for some $x \in \{0, 1/N, ..., (N-1)/N, 1\}$) jumps due to parasite reproduction to $x + \frac{1}{N}$ at rate $g_N(1 + s_N(1 + s_N(\eta - x)))Nx(1-x)$ and to $x - \frac{1}{N}$ at rate $g_N(1 + s_N(x - \eta))Nx(1-x)$.

- (Host replacement) Hosts die at rate 1. Whenever a host dies it is replaced by a so far uninfected host, which instantly suffers from primary infection. At primary infection the host is infected with a single type chosen randomly according to the type frequencies in the infecting host. This leads to a jump to frequency 0 or 1 in the primary infected host. Hence, if the frequency of type A in hosts $1, ..., M$ is $x_1, ..., x_M$ in host i jumps to state 1 occur at rate $\bar{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M} x_i$ and to state 0 at rate $1 - \bar{\mathbf{x}}$.

- (Reinfection) In addition to parasite reproduction, hosts get reinfected at rate r_N. At reinfection a single parasite, which is chosen randomly in the reinfected host, is replaced by a randomly chosen parasite of the reinfecting host. Consequently, changes in frequencies are only observed, if the replaced parasite has another type as the transmitted parasite. Hence, if the frequency of type A in hosts $1, ..., M$ is $x_1, ..., x_M$ in host i jumps due to reinfection to $x_i + \frac{1}{N}$ occur at rate $r_N \frac{1}{M}\sum_{j=1}^{M} x_j(1-x_i)$ and at rate $r_N \frac{1}{M}\sum_{j=1}^{M}(1-x_j)x_i$ to $x_i - \frac{1}{N}$.

- (Mutation) Each parasite mutates at rate μ_N to type A and at the same rate to type B.

To summarize jumps from state

$$\mathbf{x} = (x_1, ..., x_M) \text{ to}$$

$\mathbf{x} + \frac{1}{N}e_i$	occur at rate	$g_N(1 + s_N(\eta - x_i))Nx_i(1-x_i) + r_N\frac{1}{M}\sum_{j=1}^{M} x_j(1-x_i) + \mu_N(1-x)$
$\mathbf{x} - \frac{1}{N}e_i$	occur at rate	$g_N(1 + s_N(x_i - \eta))Nx_i(1-x_i) + r_N\frac{1}{M}\sum_{j=1}^{M}(1-x_j)x_i + \mu_N x$
$\mathbf{x} + (1-x_i)e_i$	occur at rate	$\bar{\mathbf{x}}$
$\mathbf{x} - x_i e_i$	occur at rate	$(1 - \bar{\mathbf{x}})$

for $i = 1, ..., M$ with $x_i \in \{0, 1/N, ..., 1\}$, $\bar{x} := \frac{1}{M}\sum_{j=1}^{M} x_j$ and $e_i = (0, ..., 1, ..., 0)$ the i-th unit vector of length M.

2.1 Parasite reproduction with balancing selection

The reproduction mechanism b) of parasites has been analyzed in detail in [50]. We give here the upshots of the results concerning the maintenance of diversity.

Assume the host as well as the parasite population size per host M and N are large and the following conditions are fulfilled:

Assume

1. (moderate selection)
$$s_N = N^{-b}$$
for some $0 < b < 1$.

2. (frequent reinfection)
$$\lim_{N\to\infty} r_N s_N = r$$
for some $r > 0$.

3. (fast parasite reproduction)

$$g_N \gg N^{3b+\epsilon}$$
$$g_N \ll \exp(N^{1-b(1+\epsilon)})$$

4. (rare mutation)

$$\mu_N N M_N \ll r_N$$

Then for large N host type frequencies are concentrated around 0, η and 1. Denote by v^ℓ the proportion of hosts which frequencies are concentrated around ℓ for $\ell = 0, \eta, 1$. Then in the limit $N \to \infty$ the evolution of the frequencies (v^0, v^η, v^1) is driven by the dynamical system

$$\dot{v}^0 = (1-\eta)v^\eta - 2r\eta v^0(v^1 + \eta v^\eta)$$
$$\dot{v}^\eta = -v^\eta + 2r(\eta^2 v^0 v^\eta + (1-\eta)^2 v^1 v^\eta + v^0 v^1) \quad (1)$$
$$\dot{v}^1 = \eta v^\eta - 2r(1-\eta)v^1(v^0 + (1-\eta)v^\eta). \quad (2)$$

For

$$r > \max\{\frac{2\eta - 1}{2\eta(1-\eta)^2}, \frac{1-2\eta}{2(1-\eta)\eta^2}\} \quad (3)$$

the dynamical system $\mathbf{v} = (v^0, v^\eta, v^1)$ has a unique stable equilibrium in the point $\mathbf{u} = (u^0, u^\eta, u^1)$ with

$$u^0 = \frac{2r\eta(1-\eta)^2 - (2\eta - 1)}{2r\eta^2 + 4r^2\eta^3(1-\eta)}$$
$$u^\eta = \frac{4r^2\eta^3(1-\eta)^3 - (2\eta-1)^2(2r\eta(1-\eta)+1)}{2r\eta^2(1-\eta)^2(1+2r\eta(1-\eta)}$$
$$u^1 = \frac{2r(1-\eta)\eta^2 + 2\eta - 1}{2r(1-\eta)^2 + 4r^2\eta(1-\eta)^3}$$

(and saddle points in $(1,0,0)$ and $(0,0,1)$).

The stability of the dynamical system is a key ingredient to show, that under Assumptions 1-4 and $r > \max\{\frac{\eta}{2(1-\eta)^2}, \frac{1-\eta}{2\eta^2}\}$ (which is a slightly stronger assumption on r than (3)) most of the time both types are present in the parasite population at non-trivial frequencies. Indeed in Theorem 3 in [50] it is shown that under these assumptions the time until an initially monomorphic parasite population (that is a parasite population for which all parasites are either of type A or all parasites are of type B) turns into a parasite population with host type frequency proportions close to the equilibrium state $\mathbf{u}$ is of the order $1/(\mu_N M N s_N) + M^\gamma$ for any $\gamma > 0$ and large N and M and vice-versa the time until a parasite population with host type frequency proportions close to the equilibrium state $\mathbf{u}$ is turned into a monomorphic population is exponentially large in M, more precisely this time is bounded from below by $\exp(M^{1-\gamma})$ for any $\gamma > 0$. Hence, if $\mu_N N M \gg \frac{1}{\exp(M^{1-\gamma})s_N}$ (which is a very weak lower bound on the mutation rate) most of the time both parasite types coexist in the parasite population.

In Figure 2 a simulation of the model is depicted. Except for the host population size, which has been chosen for feasibility of the simulations, the parameter values could be realistic for HCMV.

3 Results

3.1 Diversity patterns under neutrality

In this section we analyze the neutral setting. We consider two cases. In the first case the two parasite types are constantly generated by mutation, in the second case the two types are is assumed to be present in the population due to standing genetic variation, i.e. due to an equilibrium of genetic drift and mutation. We will evaluate the role of the neutral scenario in the case of HCMV by assessing genotyping data of HCMV. We make the following assumptions

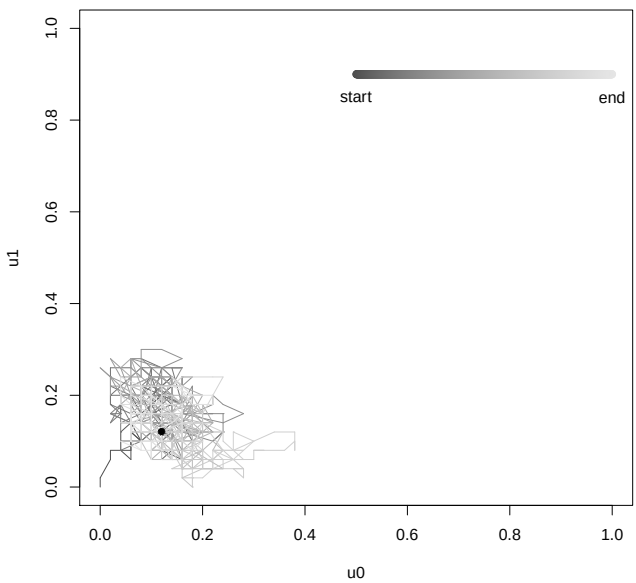

Figure 2: A simulation of the model with balancing selection for $\eta = 0.5$, $r = 6$, $g_N = 1000$, $M = 50, N = 1000, b = 0.3$. The thick black dot indicates the location of the equilibrium point $\mathbf{u}$.

- In both cases we assume as before that the host parasite populations as well as the host population is large, that is we are interested in the case when N and M is large. In precise mathematical terms we will consider the limiting case $N \to \infty$ and $M = M_N \to \infty$ for $N \to \infty$.
- As before parasite reproduction is frequent. More precisely, we assume $g_N \gg 1$.
- We consider a regime in which the reinfection rate is at most of the same order as the reproduction rate, i.e. either $r_N \ll g_N$ or $r_N \approx cg_N$ for some $c > 0$. The biological motivation for this assumption is that before another host can be reinfected, sufficiently parasites have to generated in the transmitting host first. This presumably requires (many) reproduction events within the persistent parasite population. In particular, this should be true, if the virulence of a virus is low, for further discussion see Section 3.2.

3.1.1 Diversity due to mutation

Consider the model from Section 2 with neutral parasite reproduction and assume that the just stated assumptions are fulfilled.

We show next that the length of the total genealogy of a sample of size n drawn randomly from the parasite population is at most of order $\log(n)\min\{M, NM/r_N\}$ with high probability, that is with a probability that tends to 1, when $N \to \infty$, the length of the total genealogy can be estimated from above by $clog(n)\min\{M, NM/r_N\}$ for some constant $c > 0$.

Draw randomly a sample of size n from the parasite population and follow back its ancestry. As the parasites are sampled randomly and the host population is large the parasites are drawn with high probability from n different hosts.

Since parasite reproduction is restricted to hosts, parasites have to infect a common host first to join ancestry. This happens via reinfection or host replacement (i.e. primary infection). Each pair of lines is assembled in a single host due to reinfection at rate $2r_N/(NM)$, because both lines reinfect other hosts at rate r_N/N and the probability that at a reinfection event the other host of

the pair is hit is $1/M$. Hence, the rate at which two out of n lines assemble in a single host due to reinfection is $n(n-1)r_N/(NM)$.

Once two lines infect the same host by reinfection, they coalesce with some positive probability (due to parasite reproduction or host replacement) before a reinfection event can separate lines again into different hosts. Indeed, each host is hit by an host replacement event at rate 1 and each pair of lines is hit by an parasite reproduction event at rate g_N. Hence, two lines coalesce at rate $2g_N/N + 1$, since if one of the lines is hit by a reproduction event the probability that the other lines is chosen to be replaced is $1/N$. Furthermore, each line is hit by a reinfection event at rate r_N/N. Since by assumption $r_N \lesssim g_N$ the probability that the two lines coalesce before they are hit by reinfection event is strictly larger than 0. Furthermore, the time till coalescence is in expectation N/g_N, if $g_N \gg N$, and of order 1 otherwise, because in this case lines coalesce due to host replacement.

Furthermore, each host is hit by an host replacement event at rate 1. Hence, due to host replacement pairs of lines assemble in a single host at rate $2/M$, i.e. the overall rate is $n(n-1)/M$.

Consequently, the time scale of the genealogy is of order $\min\{M, MN/r_N\}$. The structure of the genealogy is that of a Kingman coalescent. For a Kingman coalescent of size n it is well known, that the total tree length has expectation $\sim 2\log(n)$ and a finite variance, see [66], Chapter 3.

Hence, the length of the total genealogy is of order

$$O(log(n)\min\{M, NM/r_N\}).$$

Remark 1 (Transmission of multiple types at primary infection)**.** *In both variants of our model we assume that at primary infection only a single type is transmitted. If the primary transmission process lasts for a long time, like in the case of HCMV for which many hosts are primary infected by their mothers via breast feeding, multiple types could be transmitted, see also the point "primary infection with a single type" in Section 3.2.1.*

In a more general model of primary infection host replacement may just have a weaker effect and therefore in the general case the length of the total genealogy is at most of order $O(log(n)NM/r_N)$.

3.1.2 Diversity due to standing genetic variation

A certain locus might also be diverse due to standing genetic variation. Assume the frequency of type A in the parasite population is $0 < x < 1$. We are interested in the distribution of the frequency W^N of type A in a randomly chosen host. We distinguish the following cases:

a) Assume $r_N \ll g_N$. Then W^N converges (in distribution) for $N \to \infty$ to a Bernoulli-distributed random variable with parameter x, $Ber(x)$. That is for large N the proportion of hosts, for which the frequency of type A is close to 0, is almost x and the proportion of hosts, for which the frequency of type A is close to 1, is almost $1-x$.

b) Assume $r_N \sim cg_N$ for some $c > 0$ and assume $\frac{g_N}{N} \sim \frac{\gamma}{2}$ and $\frac{r_N}{N} \sim \rho$ for some $0 \leq \gamma, \rho \leq \infty$.

   i) If $\rho = 0$, then W^N converges again to a $Ber(x)$-distributed random variable.

   ii) If $\rho = \infty$, then W^N converges in distribution to a $Beta(xc, (1-x)c)$ distribution. That is for large N the proportion of hosts for which the frequency of type A lies between $0 < q_1 < q_2 < 1$ is $\int_{q_1}^{q_2} \frac{y^{xc-1}(1-y)^{(1-x)c-1}}{B(xc,(1-x)c)} dy$ with $B(xc, (1-x)c) = \frac{\Gamma(xc)\Gamma((1-x)c)}{\Gamma(c)}$ und Γ denotes the Gamma function.

   iii) If $0 < \rho, \gamma < \infty$, W^N converges weakly to a mixture of Beta distributions. This mixture has more mass close to the boundaries 0 and 1 then the corresponding Beta distribution in case (ii) (with the same parameter c). The concrete mixture is specified below.

To see the above statements consider a random sample of parasites of size n taken from a randomly chosen host and follow back its ancestry. To arrive at the frequency of type A in the chosen host, we need $1 \ll n$, but we are free to choose $n \ll N$.

By the same arguments as in Section 3.1.1 we have that the n lines coalescence (due to parasite reproduction) at rate $n(n-1)g_N/N$ and the n lines are hit by reinfection events at rate nr_N/N. Finally hosts are replaced at rate 1, in this event all lines are involved simultaneously. Consequently the relationship of coalescence : reinfection : host replacement is

$$\frac{n(n-1)g_N}{N} : \frac{nr_N}{N} : 1.$$

Case a): In a standard Kingman coalescent with a pair coalescence rate of 1 the time until n lines coalesce has expectation $2(1-1/n)$ and variance $8\sum_{i=2}^{n}\frac{1}{i^2} - 4(1-\frac{1}{n})^2 < \frac{8\pi^2}{6}$, see [66]. In our setting pairs of parasites coalesce at rate $2g_N/N$ due to parasite reproduction and all lines coalesce due to host replacement at rate 1. Hence, in expectation all lines coalesce till time $\approx 4N/g_N$, if $g_N \gg N$, and accordingly the variance is at most proportional to $(N/g_N)^2$, and otherwise all lines coalesce within a time frame of order 1.

Every parasite is hit by a reinfection event at rate r_N/N. By assumption $r_N \ll g_N$ hence, by Chebychev's inequality with high probability all lines coalesce (due to parasite reproduction or host replacement) before a reinfection kicks in.

The type of the MRCA (most recent common ancestor) is a random sample from the parasite population. Host replacement occurs at rate 1 per host and the time to the MRCA of the total host population is of order M_N. Hence, the time to the MRCA within a host is small in comparison to the MRCA of the total parasite population and therefore the frequency of type A in the total parasite population is still approximately x when the MRCA of the sample is reached. Hence the probability that the MRCA of the sample is of type A is approximately x.

Case b): In Case b) parasite reproduction and reinfection act on the same time scale. The diversity pattern depends on the relationship of the rate of parasite reproduction and reinfection to the host replacement rate.

In Case b)i) $\rho = 0$, i.e. $r_N \ll N$ as well as $g_N \ll N$. Consequently, a host replacement event hits the ancestral lines of the n parasites much earlier than host replacement and reinfection. At a host replacement event all lines coalesce into a single type. Consequently, similar as in Case a) only purely infected hosts are expected. By the same arguments as in Case a) with probability x the hosts carries only type A parasites, with probability $1-x$ only type B parasites.

In Case b)ii), host replacement is negligible in the limit $N \to \infty$, but parasite reproduction and reinfection are acting on the same time scale. The relationship of coalescence: reinfection reads

$$n(n-1)g_N : ncg_N.$$

Typically in this scenario some lines are hit by reinfection events before the MRCA of the n lines is reached. If an ancestral lineage is hit by a reinfection event the line is followed further in a randomly chosen host. Since $M = M_N \to \infty$ the probability, that two lines which are hit by a reinfection event infect the same host before the MRCA of the remaining non-reinfecting lines is reached, converges to 0 for $N \to \infty$. Furthermore, since the MRCA of the remaining lines is found within a time frame of order 1, no recognizable change of the frequency of type A occurs in the limit $N \to \infty$. Hence, each line hit by reinfection is, independently of the others, with probability x of type A and with probability $1-x$ of type B. This scenario can also be interpreted as the genealogy of a Wright-Fisher diffusion with parent independent mutation to type A at rate xc and to type B at rate $(1-x)c$. In this case the distribution of the frequency of type A is given by a Beta-distribution with parameters cx and $(1-c)x$, see [23], Chapter 5.6, or Chapter 1.3.4 in [9]. Indeed, forward in times the genealogy of the sample of parasites can be generated via a Yule process with immigration, see Chapter 1.3.4 in [9] or for similar results [20]. Yule-trees of type A (of type B, respectively) immigrate at rate xc (at rate $(1-x)c$) and each branch splits at rate 1. Let $Y_t^{(xc)}$ and $Y_t^{((1-x)c)}$ be the number of branches at time t in the two Yule process with immigration started at time 0. $Y_t^{(xc)}/t$ converges to a Gamma-distributed random variable with parameter cx and 1, see proof of Theorem 4 in [9]. Due to the properties of sums and quotients of

independent Gamma-distributed random variables, for $n \to \infty$ the proportion of type A branches $\frac{Y_t^{(xc)}}{Y_t^{(xc)}+Y_t^{((1-x)c)}}$ is Beta$(xc, (1-x)c)$ distributed.

In Case b)iii) host replacement, reinfection and parasite reproduction are acting on the same time scale. The relationship of parasite reproduction: reinfection: host replacement reads

$$\frac{n(n-1)}{2}\gamma : n\rho : 1.$$

Coalescence and reinfection alone create the Beta distributions as in case b) ii), but host replacement ends up the genealogy earlier giving reinfection less chance to introduce diversity. Consequently, more lines are in general of the same type and the distributions of the frequencies have a heavier mass close to 0 and 1.

Indeed, we can construct the genealogy of a sample of size n as follows:

Consider first a coalescent started with n lines which lines are killed at rate ρ and with pairs of lines coalescing at rate γ. Let K^n be the (random) number of lines left in such a coalescent when a clock with exponentially(1) distributed waiting time rings. These K^n lines are left in the genealogy, when the host replacement event occurs. Hence, these lines must have all the same type. With probability x they are of type A and with probability $1-x$ they are of type B.

To arrive at a sample of size n we can now successively add lines in a Yule-process with immigration. Start with K^n lines either all of type A or all of type B. Now add lines until n lines are in the sample as follows: Lines branch at rate 1 and lines of type A immigrate at rate cx and lines of type B immigrate at rate $c(1-x)$.

Let K^∞ be the (random) number of lines left, if the coalescent above is started with ∞-ly many lines and construct in the same manner as in the paragraph before a sample of size ∞. Then as in case b)ii) the limiting distribution of the proportion of type A branches is with probability x $Beta(K^\infty + cx, c(1-x))$ and with probability $1-x$ $Beta(cx, K^\infty + c(1-x))$. This can be seen as follows: Let Y_t be the number of branches in a Yule process, for which every branch splits at rate 1, at time t. Then Y_t/e^t converges for $t \to \infty$ to an exponentially distributed random variable with parameter 1, see Chapter 1.3.4, Lemma 2 in [9]. Furthermore, let $Y_t^{(cx)}$ the number of branches in a Yule process with immigration, for which every branch splits at rate 1 and branches immigrate at rate cx, at time t. Then $Y_t^{(cx)}/t$ converges to a Gamma-distributed random variable with parameter cx and 1 as in Case b)ii).

If $K^\infty = K$ and initially all lines are of type A the proportion of lines of type A is

$$Z := \frac{Y_t^1 + ... + Y_t^K + Y_t^{(cx)}}{Y_t^1 + ... + Y_t^K + Y_t^{(cx)} + Y_t^{(c(1-x))}},$$

since lines of type A immigrate at rate cx and lines of type B immigrate at rate $c(1-x)$. Z converges for $t \to \infty$ to Beta$(K + cx, (1-x)c)$, due to the properties of sums and quotients of independent Gamma-distributed random variables. In the same manner the proportion of lines of type B is obtained.

Remark 2 (Transmission of multiple types at primary infection)**.** *As in Section 3.1.1 we also want to discuss more general models of primary infection. In Section 3.2.2 we will apply the results in the scenario b) ii), because under this scenario the highest frequency of hosts infected with more than one genotype is expected.*

The asymptotics in this case are not changed, if at primary infection also more than one type can be transmitted. Indeed, in this case host replacement events (at which primary infection occurs) are not relevant, since the ancestry is (asymptotically) already determined before a host replacement occurs.

In case a) host replacement may occur before the MRCA of the sample is found. However if multiple virions may be found the persistent population at primary infection, from a backwards view this just means that these lines have to be followed further in another, but identical host. Since in case a) reinfection is negligible all lines will coalesce before reinfection can separate the lines into

different hosts. Consequently, in case a) host type frequencies are also Bernoulli-distributed if at primary infection multiple types can be transmitted.

In case b) i) lines merge predominantely due to host replacement, if at primary infection only a single genotype is transmitted. If at host replacement only a small number of virions found the persistent population after a few host replacement events all lines merge into one, so that reinfection cannot separate lines into different hosts and we end up at a Bernoulli-distribution again. If the effective number of founding virions at primary infection is larger reinfection might shift the distribution of host type frequencies towards a distribution with some mass at intermediate host type frequencies, because backwards in time reinfection could separate whole families of virions which coalesced already at host replacement events closer to the present time.

In case b) iii) the role of host replacement is in the model with primary infection with a single type to shift the mass of host type frequencies closer to the boundaries 0 and 1. In a general model of primary infection this effect may be just weakened.

3.2 Biological relevance

The model is mainly motivated by observations made for the human cytomegalovirus. However, the model described above and the parameter regimes we consider in this study might be also relevant for other parasites. In the following subsection we assemble some information about HCMV from the literature, which is relevant for the biological justification of the model.

3.2.1 Applicability to HCMV

HCMV is a herpesvirus which persists in its hosts after primary infection as well as after reinfection, that is once infected a host carries a HCMV population, established at possibly several infection events, till the end of its life. During persistence the viral population switches between phases of reactivation and latency. In this manner the virus manages to hide from the host immune system in phases of latency, but can spread at phases of reactivation. Complex mechanisms control the switches, see [21].

In DNA regions encoding proteins important for cell-to-cell virus transmission and host cell entry as well as encoding strategies of immune evasion and immune modulation (that is viral mechanisms to infer with the immune response of the host) and regions encoding proteins permitting the HCMV cell tropism (that is allowing HCMV to replicate in different cell types of the host) one finds a clustering into a few phylogenetically distinct genotypes, see [30] for a review. These genotypes have been shown to evoke different immune responses of the host (for UL75 (encoding glycoprotein gH), see [65]), or modified protein functions (for UL73, UL74 and UL55 (encoding for glycoprotein gN, gO and gB), see). Considering single open reading frames (ORFs) there are non-trivial proportions of hosts infected only with a single genotype and non-trivial proportions of hosts infected with multiple types, see e.g. [47] for the ORF UL75, [7, 46, 57] for UL55 , [45] for UL55, UL144, UL146 and UL09 . When an infected host is reinfected with a virus carrying a genotype that is absent in this host, this genotype may evade preexisting immune response to establish persistence and enhance reactivation of the virus due to a missing immune response on the host's part [12]. Furthermore, a certain diversity present in an infected host may act as a permanent stimulator for reactivation, as a diverse viral population might be more difficult to be controlled than a monotone one. In this study we are interested how these infection patterns with hosts infected with a single genotype and host infected with several genotypes may arise in a parasite population.

Our model is designed to be in line with the observed evolution of HCMV. Next we give the motivation for our model design from a perspective of HCMV. In the literature most studies investigated HCMV infections in transplantation patients and congenitally infected children in whom viral reproduction is often more enhanced than in the healthy immunocompetent host. On an evolutionary time scale these patients are probably of low relevance, therefore here we primarily refer to the literature analyzing the infection dynamics in healthy hosts.

- (Carrying capacity of within host parasite populations) We assume in our model that there exists some carrying capacity that the viral population maintains in the host also at phases of latency. For simplicity we assume, that this carrying capacity does not change over time and does not differ between hosts. It is currently unknown, if these model assumptions apply for HCMV. The persistent population may be increasing as long as the immune system cannot control HCMV sufficiently (the controllability of HCMV depends on the age of the host: in young, infected children (with and without symptoms) high HCMV DNA loads can be detected for more than a year, [3], in adults the virus sporadically reactivates, see [39, 69]). It is possible that the viral population size is decreasing over time when the virus does not manage to reactivate within a host or reinfect a host sufficiently often, since cells in which the virus persists are replaced. The host immune response on HCMV certainly differs between host. This might also effect the carrying capacities. However, the number of latently infected monocytes did not vary a lot in eight different, presumably immunocompetent hosts, see [61], which might be an indication for rather uniform carrying capacities among hosts.

  For parameter regimes of the carrying capacity of within host parasite populations see the separate point "parameter regimes".

- (well-mixed host population and well-mixed parasite populations within hosts)

  To the best of our knowledge, so far neither estimates of the host population size, nor an analysis of the host population structure is available. So far only a few studies analyze the geographic structure of the HCMV population. In [60] 124 full genome sequences have been analyzed with 122 sequences from Europe, one from the USA and one from South Korea. The estimated phylogeny exhibits a star-like form, the two sequences outside of Europe did not cluster separately. [49] analyzed the worldwide distribution for UL 73 (for samples from Europe, China, Australia and Northern America) and [11] for UL146 und UL 139 (for samples form Africa, Asia, Australia and Europe). No significant differences were detected in the genotype distributions. Therefore the simplistic assumption assume of a panmictic, i.e. well-mixed, host population might be realistic.

  Migration rates between compartments have been estimated to be small in HCMV samples from congenitally infected infants, [53]. However, we model the persistent parasite population. This population is thought to be located primarily in CD34+ bone marrow cells, i.e. in a single compartment, [21]. Consequently, also parasite populations within host might be pretty well-mixed.

- (viral reproduction) We model a reproducing "persistent viral population". On the real time scale the reproduction rate obviously is not constant. There are phases at which the immune system essentially silences viral reproduction and phases at which the virus reactivates. In phases of reactivation virus is produced and shed into bodily fluids to infect other hosts, but also cells are infected in which the virus eventually will persist. In our model tacitly the time for each host is rescaled accordingly to the speed of viral reproduction. There is increasing evidence, that HCMV reproduction is not restricted to early childhood, but continues at a lower frequency in (the immunocompetent) adulthood, especially in the eldery, see [39, 69]. In particular, van Boven et al. hypothezise that "infectious reactivation in adults is an important driver of transmission of CMV".

  For estimates of the viral reproduction rates see the separate point "parameter regimes".

- (primary infection with a single type) On a long evolutionary timescale most hosts should be primary infected with HCMV via breast milk or in early childhood via saliva, because in developing countries the HCMV seroprevalence is higher and acquired earlier, see [5] and references therein. The low virulence of HCMV (see also the point "relationship of the reinfection rate r_N and the parasite reproduction rate g_N") suggests that at primary infection most frequently only a single type can establish the infection in the new host. However, it has also been observed that during lactation several CMV genotypes have been transmitted to a preterm infant, see [38].

The results about the maintenance of diversity in the setting with balancing selection also hold, if at a primary infection more than a single type is transmitted. Only the dynamical system driving the evolution of the host type frequencies might change in favour of a higher proportion of hosts being infected with both types.

The case of neutral parasite reproduction we discuss in Remark 1 and Remark 2. We argue that the results of Section 3.1.1 and 3.1.2 also hold for a more general models of primary infection. The only exception is Case b) ii) in Section 3.1.2 for primary infections at which many parasites found the persistent parasite population. However, in this case typically a host is hit by a host replacement before a reinfection event happens. For HCMV this scenario seems not to be relevant, since diversity seems to be accumulating over the life of a host, see also the point reinfection.

- (reinfection) It has been shown that HCMV can superinfect persistently infected hosts despite CMV-specific humoral and cellular immunity, see [36].

  Women who have been infected with CMV recently have been shown to be infected with a single CMV genotype only at the ORF UL55, UL144, UL146 and UL09, see [45]. On the other hand in the majority (15/16) of random samples from CMV-seropositive women (with the same demographic origin, but most likely with past primary infections (as the women have been randomly sampled) several genotypes at the ORF UL55 and UL73 have been detected, see [46]. In [29] genotyping was performed for gB (UL55), gH (UL75) and UL10 for 36 immunocompetent patients during primary infection and in 14 patients experiencing HCMV reactivation/reinfection. In all cases only one gB-gH-UL10 genotype was detected in patients experiencing a primary infection, in contrast in 4 of 14 immunocompotent patients a mixed genotype infection was detected. The frequency of mixed infection thus differed significantly between primary and past infection (p-value =0.004, Fishers exact test).

  These findings indicate that also virions that are transmitted to a host at superinfection may persist in a host and therefore at least part of the diversity observed in hosts may have been introduced at (several) reinfection events.

- (balancing selection within hosts) In variant b) of the model we assume that both genotypes are actively maintained in host populations. Various mechanisms may lead to such an effect.

  For example HCMV exhibits a broad cell-tropism, HCMV virions can infect many different cell types of the human host, see [59]. The reproduction rate in the different cell types varies strongly between HCMV samples.

  For example for virus entry intro fibroblasts only the complex gH/gL/gO is necessary, but for virus entry into epithelial cells both complexes gH/gL/UL128-UL131 and gH/gL/gO are necessary. Zhou et al. suggest that the different genotypes of gO might affect the ratio gH/gL/gO to gH/gL/UL128-131, see [72].

  In a host several gO genotypes might be maintained in the persistent viral population, so that both virus entry into epithelial as well as virus entry into fibroblasts may be tweaked.

  Another example are immune evasion strategies or mechanism for immune modulation. Different antibody responses exist against the antigenes of genotypes gH1 and gH2 of the glycoprotein gH (UL75) and against two different antigenes of gB (UL55), see [44, 65]. Both genotypes may be controlled according to some predator-prey-dynamics with the immune system of the host on the one hand, on the other hand they may compete with each other due to a restricted amount of cells which can be latently/persistently infected.

- (parameter regimes)

  - (parasite population size N within hosts) Currently, there are no estimates of the effective size of the persistent viral population in the healthy host available. Only the population histories of congenitally infected infants have been inferred. The effective

compartmental population sizes of the HCMV populations were estimated to be of order 1000, see [53]. However, in these patient latent phases have not been reached yet.

It is assumed that HCMV latently infects CD34+ bone marrow cells, in the myeloid progenitor and monocytes (contained in the blood, differentiated from CD34+ cells), see [21]. (Experimental data suggest that endothelial and neuronal progenitor cells are latently infected as well as well as, however no data from natural latency currently exists for these cells, [21].) In [61] the number of latently infected monocytes in eight CMV-positive supposed to be CMV-immunocompetent adults has been identified. Assuming in an average adult 4.5 to 5.5 liters of blood are circulating and each liter blood contains $0.2 - 0.8 \cdot 10^9$ monocytes, one extrapolates from the study of Slobedman and Mocarski a total number of CMV particles contained in monocytes of 36.000 to 5.720.000. In general the effective population size is of some orders smaller than the natural population size. Hence, an estimate of 1000 for the effective population size within hosts might be not too bad.

- (host population size M_N) The effective population size of the human population is estimated to be 10^4, see [48].
- (parasite reproduction rate γ_N) The replication cycle of HCMV in HCMV-naive and HCMV-experienced hosts has been analyzed by [22], see also the Review [32]. HCMV replication dynamics were investigated in a population of 30 liver transplant recipients with active HCMV replication. The mean growth rate of virus during infection of HCMV-naive individuals was 1.82 units/day (95% confidence interval [CI], 1.44 2.56 units/day), corresponding to a viral doubling time of 0.38 units/day (9 h; 95% CI, 0.27 0.48 units/day). Infection of patients with preexisting HCMV immunity was associated with a significantly slower viral growth rate (0.62 unit/day; 95% CI, 0.54 0.71 unit/day) and viral doubling time (1.12 days; 95% CI, 0.99 1.25 days; P:0001).

  Assuming that a typical host is infected for 50 years (by assuming that on an evolutionary time scale a typical host is most of the time infected in early childhood) we arrive at an upper bound of $(356/0.38) \cdot 50 \approx 48000$ viral generations within a host.

  There is evidence, that in early childhood the virus replicates over a long period, [14]. Assuming that in early childhood the virus replicates for two years (not necessarily at a stretch, within the first 5 years of infection), that is 365·2 days, we arrive at a lower bound of $365 \cdot 2/1.12 \approx 651$ viral generations during this time period. In the eldery host the virus regularly spreads, see [64] in this study 57% of the investigated elderly hosts (which were older than 65 years old) were shedding CMV. On an evolutionary time scale these old hosts might not play a role. Therefore, for a lower bound on the number of viral generations we ignore this time period. Between early childhood and eldery the immune system is capable to control the virus in the healthy host pretty well. In healthy non-pregnant women the averaged shedding prevalence was estimated to be 4.7 %, see [39]. Hence, during this period we assume, that the virus is active each year at $0.047 \cdot 365 \approx 18$ days, which leads to 9 viral generations per year. Hence, assuming again that a typical host is infected for 50 years, we arrive at a lower bound of $651 + 45 \cdot 18/1.12 = 1374$ viral generations per host life.
- (mutation rate μ_N) The average mutation rate per site per replication cycle for HCMV has been estimated to be $2 \cdot 10^{-7}$, see [54]. In this study we are not interested in the mutation rate per site, but the mutation rate of entire reading frames, which genotype landscape clusters into a few types, see [30] for an overview.

  As an example we consider the ORF UL75 encoding for the glycoprotein gH. This glycoprotein forms together with the glycoproteine gL (UL 115) and either with the proteins encoded by UL 128, UL130 and UL 131 a pentameric complex or with the glycoprotein gO encoded by UL74 a trimeric complex, both of which are essential for the viral entry into epithelial/endothelial cells and fibroblasts, see [15, 58, 73]. UL128, UL130 and UL131 are highly conserved regions, but the DNA sequences of UL74, UL75

and UL115 cluster into eight, two and four genotypes, respectively, see [30] and the references therein. The genotypes gH1 and gH2 are distinguished by 8 amino-acid changes (corresponding to 8 non-synonymous point mutations) and 1 amino-acid deletion, these can be inferred from the 124 consensus sequences from [60]. (A consensus sequence is determined as follows: The HCMV DNA from e.g. a urine sample of a HCMV-positive patient is sequenced and the reads are aligned to a reference sequence. Then the consensus sequence is built from this alignment by choosing the most frequent variant at each site.) The 8 amino-acid changes and the amino-acid deletion are all located in a small N-terminal region of 40 amino-acids, encoding a linear neutralizing antibody site. The two genotypes obey functional differences, at least in the sense that they trigger genotype specific antibody responses (with a binding site of the neutralizing human antibody which involves 2 of the amino-acid changes and the amino-acid deletion), see [65], Epithelial/endothelial cells and fibroblasts are all cells types in which the virus can reproduce efficiently, see [59] and from which virus is released into bodily fluids. Hence, in general one should be infected and one should sample virus with an intact ORF UL75. As in samples (from [60]) only the two genotypes gH1 and gH2 are detected (at least on the basis of a consensus sequence) we assume that these two genotypes are the only variants which allow a proper virus entry into these cells. Furthermore, we assume in the following that the rate at which amino acids are deleted is the same as the mutation rate per site.

As (by our assumption) only genotypes gH1 and gH2 can be transmitted to other hosts, all 8 amino acid changes and the amino acid deletion have to happen on a parasite line within a host life. The rate at which this happens is $(g_N \cdot 2 \cdot 10^{-7})^9$, because within a host life on average g_N viral reproduction events occur. The leads to an upper estimate of $(4.8 \cdot 10^4 \cdot 2 \cdot 10^{-7})^9 \approx 10^{-18}$ for the mutation rate μ_N per parasite, see the point "viral reproduction rate g_N".

– (reinfection rates and the relationship of the reinfection rate r_N and the parasite reproduction rate g_N)

  For simplicity, we do not assume, that the reinfection rate depends on the host state in our model. Currently, there is no data available, how the reinfection rate is influenced by diversity. We suggest that diversity accumulating (with respect to the whole genome) may antagonize the host's immune system, and hence increases the reactivation rate and therefore increases also the reinfection rate. However, in this study we limit ourselves to a single locus and hence act on the assumption that the reinfection rate is homogeneous for all host types.

  There have been used two different methods to assess the frequency of reinfection. In both cases longitudinal samples from patients are taken. In method a) the samples (from the different time points) were sequenced at specific ORFs and the genotypes present in the different samples were compared. If a new genotype appeared this was considered as evidence for reinfection. In method b) the reactivity to antibodies specific for certain genotypes at the ORFs UL55 and UL75 have been analyzed in the different samples. If a new antibody specifity appeared that was considered at evidence for reinfection, [55].

  Both methods have certain drawbacks. It has been shown in methods using deep-sequencing that often genotypes are present only at very low frequencies and that the frequencies of the genotypes vary strongly over time. A reason for these findings might be that at different time points different virions reactivate and therefore the viral population detected in the samples changes over time. Hence, with methods a) it well might be that the genotypes detected at the different time points were present in the host all the time. This would lead to an overestimation of the reinfection rate.

  A problem with using method b) is that a large proportion of hosts (roughly 30%) does not show any antibody reactivity. At least for ORF UL75 the reason cannot be that the hosts react to a different epitope, because in this region the population clusters

only into two genotypes. In [55] the authors claim that reactivity to strain specific epitopes was shown to persist on average for 21 months. Hence, it well might be that if a specific genotype is not reactivated the antibody response to that genotype decreases over time, although virions carrying this genotype still persist in the host. A lacking antibody response on the other hand may facilitate the reactivation of the virions with that genotype and renew the antibody response which then is detected in the samples. A reinfection event however must not have occurred to generate this response. This would lead to an overestimation of the reinfection rate with method b).

In [55] an annual reinfection rate has been estimated by first determining the proportions of hosts reinfected in a certain time interval and the dividing this number by the length of the time interval (in time units of years). Using this algorithm the reinfection rate may be underestimated, since hosts may be reinfected, but might not be infected with one of the genotypes (under consideration) they are still missing. This is especially like to happen if only few genotypes are considered. If for example 20% of the CMV-infected persons are infected only with genotype gH1, 30% of the persons are infected with both genotypes gH1 and gH2 and 50% of the persons are infected only with genotype gH2, a person (so far only infected with genotype gH1) is only in roughly $30 \cdot 0.5$ % + 50% of all reinfections infected with genotype gH2 (if one assumes that a host infected with both genotypes transmits in 50% of the cases genotype gH1 and in 50% of the cases genotype gH2). By determining the proportion of the different genotypes in the population and transmission probabilities of the different genotypes from host infected with several genotypes, one should be able to estimate the reinfection rate quite well also when considering only a few (or a single) ORF (of course provided that one is able to distinguish the occurrence of a new genotype from the reactivation of an already present genotype).

With method a) in 7 (19%) of 37 young children attending day-cares evidence for reinfection has been shown within a time frame of sampling between 11 weeks to 38 months with a median of 11 months, see [7].

In [55] 205 CMV seropositve women and in [70] 149 CMV pregnant women (40 of these with congenitally infected infants and 109 with uninfected newborns) were followed prospectively with method b). In the first study an annual reinfection rate of 10% has been estimated and in the second study an annual reinfection rate of 34% for mothers with congenitally infected infants and an annual rate of 9% has been estimated for mothers with uninfected newborns.

In the same population as in [55] a similar primary infection rate has been estimated. It is not clear, why the primary infection rate is not higher than the reinfection rate, because preexisting immune response should prevent at least partially recurrent infections.

We assume that the reinfection rate is smaller than the viral reproduction rate. The reasoning is that before a host can reinfect another host the virus has to replicate first within this host. Furthermore the virulence of HCMV is pretty low. Transmission over breast milk is relatively ineffective. Nearly every (96 %) HCMV-positive mother reactivates HCMV when nursing, see [37]. However, transmission often (in approximately 80% of the cases) results only in local oral infections not leading to a persistent infection with HCMV, see [43]. In the study of Hamprecht et al. (2001) a mother-to-infant transmission rate of 37% was estimated and viral DNA or infectious virus was detectable in the infant with a mean of 42 days (95% CI [28,69]) after the estimated onset of transmission risk in the mother. Similarly, the transmission over saliva requires time to be effective. The annual seroconversion rate of women with infected children was estimated to be 30% and for mothers with infected children younger than 20 months of age the interval between identification of her child's infection and maternal infection ranged from 1 to 26 months (8 +/- 6 (SD) months), see [4].

3.2.2 HCMV data

Materials Apart from data from the literature we base our study on genotype data for the open reading frame UL 75 from 17 immunocompetent patients.

From all patients serum or saliva samples were sent to the Center for Virology, Medical University of Vienna between 2001 and 2008 and tested for HCMV-DNA since they had clinical symptomes as fever, hepatosplenomegaly, elevated transaminases, and/or lymphadenopathy, data from 14 of these 17 patients have already been included in the previously published study [29]. The youngest patient was 2 years old, the oldest 84.1 years, the age mean was 55.15 years with a standard deviation of 25.35 years. For two patients saliva samples were taken, for 15 patients serumsamples. In each sample the genotypes gH1 and gH2 of the region UL 75 were detected using genotype-specific qPCR. The detection threshold for each of the two genotype gH1 and gH2 was 100 copies/ml, see [28]. We excluded 4 (from initially 21 patients) for which only a single genotype was detected and the number of viral particles was below 1000 copies/ml, because in this cases the frequency of the non-detected type can be as large as 0.1. Patients for which the frequency of one of the two types was below 0.1 is treated as purely infected. Furthermore all patients had a high HCMV IgG antibody avidity as tested by a previously described avidity assay [62], indicative for a past HCMV infection, see [51]. We excluded hosts with a low HCMV IgG antibody avidity, indicating an ongoing primary infection with CMV, since these hosts were significantly more frequently infected only with a single genotype. In a sample of 35 immunocompetent patients (from the same hospitel and same symptomes as the 17 patients) with low IgG avidity a single genotype (with at least 1000 copies/ml) we detected in 34 patients and both genotypes were detected in a single host with the minor type at frequency 0.013. The number of hosts infected with both types is in the hosts with low IgG avidity significantly lower than in the other 17 hosts with high IgG avidity (p-value = 0.0036 using Mann-Whitney's U-test)). Since also in random samples from immunocompetent children a higher proportion of children was infected with both genotypes, see [47], we excluded the samples with low IgG avidity from the study.

3.2.3 Coexistence of both parasite types due to mutations at the example of HCMV

We are interested in the probability to find two parasite types in an ORF for HCMV. We base again our analysis on the data available for the ORF UL75. We already discussed in the point "mutation rate μ_N" that on the host time scale we estimate the mutation rate from genotype gH1 to genotype gH2 by 10^{-18}.

According to the results from section 3.1.1 the total tree length of a sample of size n taken from the entire parasite population is at most of order $\log(n)\min\{M, MN/r_N\}$ on the host time scale.

Furthermore, according to our assumptions, the most recent common ancestor of the sample is either of genotype gH1 or gH2. Hence at least least one mutation from genotype gH1 to gH2 (occuring during a host life) or vice-versa has to fall onto the genealogy. The probability that this happens can be estimated from above by $\log(n) \cdot M \cdot 10^{-18}$. With $n = 124$ and an estimate of the effective host population size of 10^4, we arive at an upper estimate of $5 \cdot 10^{-14}$ to observe both genotypes simultaneously in the population. To observe even more genotypes is even less likely.

However, one might argue that also intermediate variants of the genotypes might be as fit as the two genotypes gH1 and gH2 which we distinguished by the 8 amino-acid changes and one amino-acid deletion and just by chance the genotypes gH1 and gH2 are present at the observed frequencies in the population. This scenario we discuss in the next section.

3.2.4 Diversity due to standing genetic variation at the example of HCMV

The proportion of hosts infected with both genotypes is in our sample 3/17= 0.176 (with a 95%-confidence interval of (1/17,5/17)). These estimates accord with the estimates from [47], were the frequency of neonates infected with both types were 14.3%, of infants was 23.7% and of adults was 36.4%. In this paper neonates were congenitally infected, infants were infected postnatally (or by

an unproven congenitally CMV infection) and adults were HIV patients or kidney recipients. Since these adults are immunocompromised it is likely that they are more susceptible for reinfections, which may explain the higher rate (in comparison to our estimate) of mixed infections in adults.

Furthermore, three patients (17.6%) are purely infected with type gH1 and 11 (64.7%) patients only with type gH2. In [47] the proportion of hosts only infected with type gH2 was also in all patients groups larger than the proportion of hosts only infected with type gH1. The discrepancy was largest in adults (40.5% to 45.2% in newborns, 33.3% to 43% in infants and 25.4% to 38.2% in adults).

The average frequency of type gH1 is 0.251. Hence we estimate $\hat{x} = 0.251$.

In the neutral case the highest proportion of hosts infected with both types is expected, if $lim_{N\to\infty} r_N/N = \infty$ and $\lim_{N\to\infty} \frac{r_N}{g_N} \to c$ for some $c > 0$. According to the results from Section 3.1.2 in this case the proportion of hosts infected with both types is Beta distributed with parameters $0.251c$ and $0.251(1-c)$. As pointed out already only hosts with a frequency of type gH1 between 0.1 and 0.9 are treated as mixed (i.e. infected with both types). This leads to an estimate of $c = 0.26$ (with a 95% confidence interval of (1/15, 0.49)). That means that $g_N \approx 0.26 r_N$. So, in the neutral case roughly every fourth viral generation a reinfection event should happen.

Assuming that within a host life on average 1347 viral generations occur, we arrive at an average number of 350 reinfections per host life, that is 7 reinfection per year, which is 70 times higher than the estimated reinfection rates of [55].

Note that these estimates also hold true in more general models for primary infection as we discussed under the point "primary infection with a single type".

4 Discussion

In this study we were interested in the diversity patterns in parasite populations which are capable for persistence and reinfection. In [50] a model for the evolution of a population of parasites has been introduced. In this model parasites are distributed over a large number M of (infected) hosts. Each host is infected with a large number N of parasites and there exist two parasite types. Evolution is driven by host replacement, reinfection and parasite reproduction. In [50] it is assumed that parasite reproduction is driven by balancing selection, that is there is a tendency that both parasite types are maintained in host (until a host dies and is replaced by a primary infected new host). In this case it has been shown in [50] that persistence and reinfection are in certain parameter regimes effective mechanisms two maintain diversity in a parasite population.

Mechanisms allowing the maintenance of diversity (especially in coding regions) in a parasite population might be essential for its survival. Only if the parasite population is sufficiently diverse it can manage to survive in its host, otherwise the host can adapt its host defense to easily against the parasite. We propose that persistence and reinfection might be essential mechanisms enabling a parasite to accumulate sufficient diversity within hosts without generating diversity constantly de novo by mutation.

The motivation behind the model was the observation of certain diversity patterns in samples of the human cytomegalovirus. In coding regions on the whole HCMV genome the observed diversity is rather high. The HCMV genome possess roughly 171 canonical, protein coding ORF, see [19, 27]. At least 23 of these cluster into more than one genotype, see [30], that is more than 13 %. [1] Furthermore, when these ORFs have been analyzed further, a large proportion of hosts was infected with several genotypes.

In this study we analyzed the neutral counterpart of the model in which parasite reproduction within hosts is neutral. In the case of neutral evolution reinfection may allow the spread of

[1] In [63] in total 751 protein coding ORFs have been annotated. However most of the novel identified ORFs are very short (245 are less than 20 codons long and 239 ORFs are between 21-80 codons long), so that most of the underlying fitness landscape presumably do not allow for several genotypes. Furthermore these codons are found upstream of longer ORFs. 120 novel identified ORF were longer than 120 codons long. These are primarily ORFs contain slice junctions or alternative 5' ends of previous annotations. Hence the observed diversity might coincide for these ORFs with that of the previously annotated ORFs.

parasite types in a population, however reinfection and persistence do not enhance the long-term coexistence of the two types in the parasite population.

In the neutral model we give asymptotic upper bounds for the length of the total genealogy of a random sample taken from the parasite population. Furthermore, we analyzed the distribution of the proportions of parasite types in single hosts.

We used these results to evaluate the relevance of this scenario for HCMV. We argued on the basis of the literature and genotyping data at the ORF UL75 (clustering into two genotypes) from 17 immunocompetent patients that the coexistence of genotypes just due to mutations is unlikely for the ORF UL75 and the high proportions of host infected with both genotypes of UL75 requires a reinfection rate which seems to be unrealistically high for HCMV. Hence, our analysis is a first indication that a model with balancing selection might fit better the diversity patterns observed for HCMV than a neutral model and hence an evolutionary advantage of persistence and reinfection could be their capability to enhance the maintenance of in the parasite population. A larger set of ORFs should be analyzed to achieve a more precise picture. The estimation of the probability for observing two genotypes at the ORF UL75 is based on the number of amino-acid changes and deletions distinguishing the two genotypes (for at least 124 consensus sequences). We assumed that intermediate variants cannot be transmitted between hosts (in the sense that they may contribute to the persistent parasite population in the infected host). Empirical studies are necessary to verify this hypothesis. Furthermore, for other open reading frames two genotypes might be distinguished at less amino-acids and therefore the coexistence might be more likely just due to mutation. This investigation goes beyond the scope of this paper, because it involves the study of multi-locus models with some loci clustering into more than two types.

From a point of view of applications we concentrated in this study on the human cytomegalovirus. However, our parasite model is quite general and may apply also for other parasites, like for HIV and related retroviruses or E.coli. The evolutionary advantage of their capability to persist and reinfect in their hosts may also be related to the maintenance of diversity in these populations.

Acknowledgements

CP thanks Anton Wakolbinger for fruitful discussions. CP was supported by the DFG priority programme 1590.

References

[1] Arcangeletti M-C, Simone RV, Rodighiero I, de Conto F Medici M-C, Maccari C, Chezzi C, Calderaro A (2016). *Human cytomegalovirus reactivation from latency: validation of a switch model in vitro.* Virol. J., Volume 13: 179.

[2] Adland E, Klenerman P, Goulder P, Matthews PC (2015). *Ongoing burden of disease and mortality from HIV/CMV coinfection in Africa in the antiretroviral therapy era.* Front Microbiol, Volume 6, pp. 202–213.

[3] Adler SP (1991a). *Molecular epidemiology of cytomegalovirus: a study of factors affecting transmission among children at three day-care center.* Pediatr Infect Dis J Volume 10 no. 8, pp. 584–590.

[4] Adler SP (1991b). *Cytomegalovirus and child day care: risk factors for maternal infection.* Pediatr Infect Dis J Volume 10 no. 8, pp. 590–594.

[5] Adland E, Klenerman P, Goulder P, Matthews PC (2015). *Ongoing burden of disease and mortality from HIV/CMV coinfection in Africa in the antiretroviral therapy era.* Front Microbiol. Volume 6: 1016.

[6] Balfe P, Simmonds P, Ludlam CA, et al. (1990). *Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations.* J. Virol. Volume 64: pp. 6221–6233.

[7] Bale JF Jr, Petheram SJ, Souza IE, Murph JR (1996). *Cytomegalovirus reinfection in young children.* J Pediatr. Volume 128 no. 3: pp. 347–352.

[8] Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD (2014). *A Bayesian MCMC Approach to Assess the Complete Distribution of Fitness Effects of New Mutations: Uncovering the Potential for Adaptive Walks in Challenging Environments* Genetics Volume 196: pp. 841–852.

[9] Birkner M (2005). *Lectures note for the course "Stochastic models from population biology"* http://www.staff.uni-mainz.de/birkner/lehre_archiv/smpb-30.6.05.pdf

[10] Bovier A, Coquille L, Smadi C (2018). *Crossing a fitness valley as a metastable transition in a stochastic population model* arXiv:1801.06473v2

[11] Bradley AJ, Kovács IJ, Gatherer D, Dargan DJ, Alkharsah KR, Chan PK, Carman WF, Dedicoat M, Emery VC, Geddes CC, Gerna G, Ben-Ismaeil B, Kaye S, McGregor A, Moss PA, Pusztai R, Rawlinson WD, Scott GM, Wilkinson GW, Schulz TF, Davison AJ (2008). *Genotypic analysis of two hypervariable human cytomegalovirus genes.* Journal of medical virology, Volume 80 no. 9, pp. 1615–1623

[12] Burkhardt C, Himmelein S, Britt W, Winkler T, Mach M. (2009) *Glycoprotein N subtypes of human cytomegalovirus induce a strain-specific antibody response during natural infection.* J Gen Virol, Volume 90(Pt8), pp. 1951–1961.

[13] Cannon MJ, Schmid DS, Hyde TB (2010). *Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection.* Rev Med Virol, Volume 20, Article 1016.

[14] Cannon MJ, Schmid DS, Hyde TB (2011). *Review of cytomegalovirus shedding in bodily fluids and relevance to congenital cytomegalovirus infection.* Rev Med Virol, Volume 21 no. 4, pp. 240–255.

[15] Ciferri C, Chandramouli S, Donnarumma D, Nikitin PA, Cianfrocco MA, Gerrein R, Feire AL, Barnett SW, Lilja AE, Rappuoli R, Norais N, Settembre EC, Carfi A (2015). *Structural and biochemical studies of HCMV gH/gL/gO and Pentamer reveal mutually exclusive cell entry complexes.* Proc Natl Acad Sci U S A. Volume 112 no. 6, pp. 1767-1772.

[16] Chou SW (1986). *Acquisition of donor strains of cytomegalovirus by renal-transplant recipients.* N Engl J Med, Volume 314 no. 22, pp. 1418–1423.

[17] Collins-McMillen D, Buehler J, Peppenelli M, Goodrum F. (2018a) *Molecular Determinants and the Regulation of Human Cytomegalovirus Latency and Reactivation.* Viruses, Volume 10, no. 8: 444.

[18] Collins-McMillen D, Chesnokova L, Lee B-J, Fulkerson HL, Brooks R, Mosher BS, Yurochko AD. (2018b) *HCMV Infection and Apoptosis: How Do Monocytes Survive HCMV Infection?* Viruses, Volume 10, no. 10: 533.

[19] Dolan A, Cunningmahm C, Hector RD, Hassan-Walker AF, Lee L, Addison C, Dargan DJ, McGeoch DJ, Gatherer D, Emery VC, Griffiths PD, Sinzger C, McSharry BP, Wilkinson GW, Davison AJ. (2004) *Genetic content of wild-type human cytomegalovirus.* J Gen Virol., Volume 85(pt 5), pp. 1301–1312.

[20] Donnelly P, Kurtz TG (1996). *The asymptotic behaviour of an urn model arising in population genetics.* Stoch Proc Appl., Volume 64, pp. 1–16.

[21] Dupont L, Reeves MB (2016). *Cytomegalovirus latency and reactivation: Recent insights into an age old problem.* Rev Med Virol, Volume 26(2): pp. 75–89.

[22] Emery VC, Hassan-Walker AF, Burroughs AK, Griffiths PD. (2002) *Human cytomegalovirus (HCMV) replication dynamics in HCMV-naive and -experienced immunocompromised hosts.* J Infect Dis., Volume 185, no. 12, pp. 1723–1728.

[23] Ewens WJ. (2004) *Mathematical Population Genetics: I. Theoretical Introduction, Second Edition* Springer Science + Business Media New York.

[24] Fishman JA, Emery VC, Freeman R, Pascual M, Rostaing L, Schlitt HJ, Sgarabotto D, Torre-Cisneros J, Uknis ME (2007). *Cytomegalovirus in transplantation - challenging the status quo.* Clin. Transplant., Volume 21 no. 2, pp. 149–158.

[25] Frank SA (1996). *Models of parasite virulence.* Q. Rev. Biol., Volume 71 no. 1, pp. 37–78.

[26] Funk GA, Gosert R, Hirsch HH (2007). *Viral dynamics in transplant patients: implications for disease.* Lancet Infect. Dis., Volume 7 no. 7, pp. 460–472.

[27] Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, Baluchova K, et al. (2011). *High-resolution human cytomegalovirus transcriptome.* Proc. Natl. Acad. Sci. U.S.A., Volume 108, pp. 19755–19760.

[28] Görzer I, Kerschner H, Jaksch P, Bauer C, Seebacher G, Klepetko W, Puchhammer-Stöckl E (2008). *Virus load dynamics of individual CMV-genotypes in lung transplant recipients with mixed-genotype infections.* J Med Virol. Volume 80 no. 8, pp. 1405-1414.

[29] Görzer I, Kerschner H, Redlberger-Fritz M, Puchhammer-Stöckl E (2010). *Human cytomegalovirus (HCMV) genotype populations in immunocompetent individuals during primary HCMV infection.* J Clin Virol., Volume 48 no. 2, pp. 100-103.

[30] Puchhamer-Stöckl E, Görzer I (2011) *Human cytomegalovirus: an enormous variety of strains and their possible clinical significance in the human host.* Future Medicinee, Volume 6 no. 2.

[31] Griffiths PD, Baboonian C (1984). *A prospective study of primary cytomegalovirus infection during pregnangcy: final report.* BJOG, Volume 91 no. 4, pp. 307–315.

[32] Griffiths P, Baraniak I, Reeves M. (2015) *The pathogenesis of human cytomegalovirus.* J Pathol., Volume 235 no. 2, pp. 288-297.

[33] Goodrom F (2016). *HUMAN CYTOMEGALOVIRUS LATENCY: Approaching the Gordian Knot* Annu Rev Virol., Volume 3 no.1, pp. 333–357.

[34] Hassan J, Connell J (2007). *Translational Mini-Review Series on Infectious Disease: Congenital cytomegalovirs infection: 50 years on.* Clin. Exp. Immunol., Volume 149 pp. 205–210.

[35] Hahn BH, Shaw GM, Taylor ME, et al. (1986). *Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS.* Science; Volume 232 pp. 1548–53

[36] Hansen SG, Powers CJ, Richards R, et al. (2010). *Evasion of CD8+ T cells is critical for superinfection by cytomegalovirus.* Science., Volume 328(5974), pp. 102–106.

[37] Hamprecht K, Maschmann J, Vochem M, Dietz K, Speer CP, Jahn G (2001). *Epidemiology of transmission of cytomegalovirus from mother to preterm infant by breastfeeding.* Lancet; Volume 357(9255) pp. 513–518

[38] Hamprecht K, Maschmann J, Vochem M, Speer CP, Jahn G (2003). *Transmission of Cytomegalovirus to Preterm Infants by Breast-Feeding.* in Prösch S, Cinatly J, Scholz M (eds) *New Aspects of CMV-Related Immunopathology.* Monogr. Virol. Basel, Karger, Volume 24 pp. 43–52

[39] Huang Y, Guo X, Song Q, Wang H, Yu H, Zhang Y, Qiao E, Xue W, Li X, Zhuang S, Wei F, Li T, Ge S, Wu T, Xia N, Zhang J, (2018). *Cytomegalovirus Shedding in Healthy Seropositive Female College Students: A 6-Month Longitudinal Study* J Infect Dis.; 217(7) pp. 1069–1073

[40] Kenneson A, Cannon MJ (2007). *Review and meta-analysis of the epidemiology of congenital cytomegalovirus (CMV) infection.* Rev. Med. Virol., Volume 17 no. 4 pp. 253–276.

[41] Korber BT, Allen EE, Farmer AD, Myers GL (1995). *Heterogeneity of HIV-1 and HIV-2.* AIDS, Volume 9 pp. S5–S18

[42] Liu X, Wang X, Shixian Y, Zhang Z, Abecassis M, Hummel M (2013). *Epigenetic Control of Cytomegalovirus Latency and Reactivationn.* Viruses., Volume 5 no. 5 pp. 1325–1345.

[43] Mayer BT, Krantz EM, Swan D, Ferrenberg J, Simmons K, Selke S, Huang M-L, Casper C, Corey L, Wald A, Schiffer JT, Gantt S (2017). *Transient Oral Human Cytomegalovirus Infections Indicate Inefficient Viral Spread from Very Few Initially Infected Cells.* Journal of Virology, Volume 91 no. 12, e00380-17.

[44] Meyer H, Sundqvist VA, Pereira L, Mach M (1992). *Glycoprotein gp116 of human cytomegalovirus contains epitopes for strain-common and strain-specific antibodies.* J Gen Virol, Volume 73(Pt 9), pp. 2375–2383

[45] Murthy S, Hayward GS, Wheelan S, Forman MS, Ahn J-H, Pass RF, Arav-Boger R (2011). *Detection of a Single Identical Cytomegalovirus (CMV) Strain in Recently Seroconverted Young Women.* PLoS ONE, Volume 6 no. 1: e15949.

[46] Novak Z, Ross SA, Patro KP, SPati SK, Kumbla RA, Brice S, Boppana SB (2008). *Cytomegalovirus Strain Diversity in Seropositive Women.* J. Clin. Microbiol., Volume 46 no. 3, pp. 882–886.

[47] Paradowska E, Jabłońska A, Studzińska M, Kasztelewicz B, Zawilinńska B, Wiśniewsk-Ligier M, Dzierżanowska-Fangrat K, Woźniakowska-Gęsicka T, Kosz-Vnenchak M, Leśnikoski ZJ (2014). *Cytomegalovirus glycoprotein H genotype distribution and the relationship with hearing loss in children* J Med Virol., Volume 86 no. 8, p. 1421–1427.

[48] Park L (2011). *Effective population size of current human population.* Genetics Research, Volume 93 no.2, p. 105–114.

[49] Pignatelli S, Dal Monte P, Rossini G, Chou S, Gojobori T, Hanada K, Guo JJ, Rawlinson W, Britt W, Mach M, Landini MP (2003). *Human cytomegalovirus glycoprotein N (gpUL73-gN) genomic variants: identification of a novel subgroup, geographical distribution and evidence of positive selective pressure.* J Gen Virol., Volume 84(pt 3), p. 647–655.

[50] Pokalyuk C, Wakolbinger A (2018). *Maintenance of diversity in a hierarchical host-parasite model with balancing selection and reinfection* https://arxiv.org/abs/1802.02429

[51] Prince, HE, Lap-Nixon M (2014). *Role of cytomegalovirus (CMV) IgG avidity testing in diagnosing primary CMV infection during pregnancy* Clinical and vaccine immunology : CVI, Volume 21, no. 10, p. 1377–1384.

[52] Renzette N, Bhattacharjee B, Jensen JD, Gibson L, Kowalik TF (2011) *Extensive Genome-Wide Variability of Human Cytomegalovirus in Congenitally Infected Infants.* PLoS Path., Volume 7 no. 5: e1001344.

[53] Renzette N, Gibson L, Bhattacharjee B, Fisher D, Schliess MR, Jensen JD, Kowalik TF (2013) *Rapid Intrahost Evolution of Human Cytomegalovirus Is Shaped by Demography and Positive Selection.* PLoS Genet, Volume 9 no. 9: e1003735.

[54] Renzette N, Pokalyuk C, Gibson L, Bhattacharjee B, Schleiss MR, Hamrecht K, Yamamoto AY, Mussi-Pinhata MM, Britt WJ, Jensen JD, Kowalik TF (2011) *Limits and patterns of cytomegalovirus genomic diversity in humans.* PNAS, Volume 112 no. 30: E4120-E4128.

[55] Ross SA, Arora N, Novak Z, Fowler KB, Britt WJ, Boppana SB. (2010) *Cytomegalovirus reinfections in healthy seroimmune women.* J Infect Dis., Volume 201 no.3, pp. 386-389.

[56] Ross SA, Novak Z, Pati S, Patro RK, Blumenthal J, Danthuluri VR, Ahmed A, Michaels MG, Snchez PJ, Bernstein DI, Tolan RW, Palmer AL, Britt WJ, Fowler KB, Boppana SB (2011) *Mixed Infection and Strain Diversity in Congenital Cytomegalovirus Infection.* The Journal of Infectious Diseases, Volume 204 no. 7 , pp. 1003–1007

[57] Rycel M, Wujcicka W, Zawilińska B, Paradowska E, Suski P, Gaj Z, Wilczyński J, Leśnikowski Z, Nowakowska D (2015) *Mixed infections with distinct cytomegalovirus glycoprotein B genotypes in Polish pregnant women, fetuses, and newborns.* Eur J Clin Microbiol Infect Dis., Volume 34 no. 3 , p. 585–591

[58] Ryckman, BJ, Rainish BL, Chase MC, Borton JA, Nelson JA, Jarvis MA, Johnson DC (2008) *Characterization of the Human Cytomegalovirus gH/gL/UL128-131 Complex That Mediates Entry into Epithelial and Endothelial Cells.* J Virol, Volume 82 no. 1, pp. 60 – 70.

[59] Sinzger C, Digel M, Jahn G. (2008) *Cytomegalovirus cell tropism* Corr Top Microbiol Immunol, Volume 325, pp. 63–83.

[60] Sijmons S, Thys K, Mbong Ngwese M, Van Damme E, Dvorak J, Van Loock M, Li G, Tachezy R, Busson L, Aerssens J, Van Ranst M, Maes P. (2015) *High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination.* J Virol, Volume 89, pp. 7673 – 7695.

[61] Slobedman B, Mocarski ES. (1999) *Quantitative Analysis of Latent Human Cytomegalovirus* J Virol, Volume 73, no. 6, pp. 4806–4812.

[62] Steininger C, Kundi M, Kletzmayr J, Aberle SW, Popow-Kraupp T. (2004) *Antibody maturation and viremia after primary cytomegalovirus infection, in immunocompetent patients and kidney-transplant patients.* J Infect Dis., Volume 190, no. 11, pp. 1908-12.

[63] Stern-Ginossar N, Weisburd B, Michalski A, Le VT, Hein MY, Huang SX, et al. (2012). *Decoding human cytomegalovirus.* Science, Volume 338, p. 1088–1093

[64] Stowe RP, Kozlova EV, Yetman DL, Walling DM, Goodwin JS, Glaser R. (2007) *Chronic herpesvirus reactivation occurs in aging.* Exp Gerontol, Volume 42 no. 6: 563–570.

[65] Urban M, Britt W, Mach M. (1992) *The dominant linear neutralizing antibody-binding site of glycoprotein gp86 of human cytomegalovirus is strain specific.* J Virol 1992, Volume 66, p. 1303–1311

[66] Wakeley J. (2009) *Coalescent Theory: An Introduction.* Greenwood Village: Roberts & Company Publishers.

[67] Wang GC S, Dash P, Mc Cullers JA, Doherty PC, Thomas PG (2013). *T-cell receptor diversity inversely correlates with pathogen-specific antibody levels in human cytomegalovirus infection.* Sci. Transl. Med., Volume 4, 128.

[68] Wright (1929). *The evolution of dominance* The American Naturalist, Volume 63 no. 689, 556-561.

[69] van Boven M, van de Kassteele J, Korndewal MJ, van Dorp CH, Kretschmar M, van der Klis F, de Melker HE, Vossen AC, van Baarle D (2017). *Infectious reactivation of cytomegalovirus explaining age- and sex- specific patterns of seroprevalence* PLoS Comput Biol 13(9): e1005719.

[70] Yamamoto AY, Mussi-Pinhata MM, Boppana SB, Novak Z, Wagatsuma VM, de Frizzo Oliviera P, Duarte G, Britt WJ (2010). *Human cytomegalovirus reinfection is associated with intrauterine transmission in a highly cytomegalovirus-immune maternal population.* Am J Obstet Gynecol, Volume 202 no. 3, pp. 297.e1–297.28.

[71] Zanghellini F, Boppana SB, Emery VC, Griffiths PD, Pass RF (1999). *Asymptomatic primary cytomegalovirus infection: virologic and immunologic features.* J. Infect. Dis., Volume 180 no. 3, pp. 702–707.

[72] Zhou M, Yu Q, Wechsler A, Ryckman BJ (2013). *Comparative analysis of gO isoforms reveals that strains of human cytomegalovirus differ in the ratio of gH/gL/gO and gH/gL/UL128-131 in the virion envelope.* J Virol., Volume 87 no. 17, pp. 9680–9690.

[73] Zhou M, Lanchy JM, Ryckman BJ (2015). *Human Cytomegalovirus gH/gL/gO Promotes the Fusion Step of Entry into All Cell Types, whereas gH/gL/UL128-131 Broadens Virus Tropism through a Distinct Mechanism.* J Virol., Volume 89 no.17, pp. 8999–9009.