A genomic transmission graph for modelling parasite transmission dynamics and population genetics

Dominic Kwiatkowski

Draft version 2 March 2023. Confidential

Abstract

The genetic structure of a parasite population is shaped by its transmission dynamics but superinfection and recombination make this relationship complex and hard to analyse. This paper aims to simplify the problem by introducing the concept of a genomic transmission graph whose essential parameters are the effective number of hosts, the quantum of transmission and the crossing rate of transmission chains. This enables rapid simulation of coalescence times allowing for superinfection and recombination, and it also provides a mathematical framework for analysis of within-host variation. Taking malaria as an example, we use this theoretical model to examine how transmission dynamics and migration affect parasite genetic variation, including haplotypic metrics of recent common ancestry. We show how key parameters of this model can be inferred from deep sequencing data, and we discuss how these concepts could help in using genomic surveillance data for disease control and elimination.

Introduction

Quantitative models of infectious disease transmission are important in planning public health strategies for disease control [1, 2]. By analysing variation in the genome sequence of parasites sampled over space and time, it is in principle possible to derive information about the recent history and dynamics of host to host transmission that cannot readily be obtained by other means.

Current methods for inference of transmission dynamics from genomic surveillance data are mostly based on an approach known as phylodynamics [3, 4], whose starting point is to construct a phylogenetic tree representing the genetic relationship between isolates. This approach works well for viruses such as SARS-CoV-2 in epidemic scenarios where recombination can essentially be ignored. However it runs into problems when there is frequent recombination combined with *superinfection*, meaning that a host acquires infection from multiple independent sources. This is the case for many parasitic microorganisms including some viruses and bacteria as well as sexually-reproducing eukaryotic parasites.

The fundamental problem is that a superinfected host carries a mixture of parasite genotypes with different ancestral histories, which may then be *cotransmitted* to other hosts. Recombination within genetically mixed infections causes different regions of the genome to have different genealogies (figure 1). This presents an extremely complex problem for genealogical inference [5, 6] and conventional phylodynamic approaches do not work in this situation, because the parasite population cannot be represented by a single phylogenetic tree.

Malaria as an example. Malaria provides an interesting paradigm for exploring this problem because people living in some parts of the world are bitten by malaria-infected mosquitoes many times per year [7, 8]. In these areas of high transmission, it is common for an infected person to carry a mixture of parasite genotypes, either due to superinfection (because they have been bitten by multiple mosquitoes each carrying parasites from a different source) or due to cotransmission (because they have been bitten by a mosquito that is carrying parasites with mixed genotypes due to superinfection of a previous host) [9]. The malaria parasites undergo sexual reproduction in the mosquito, allowing cotransmitted genomes to recombine.

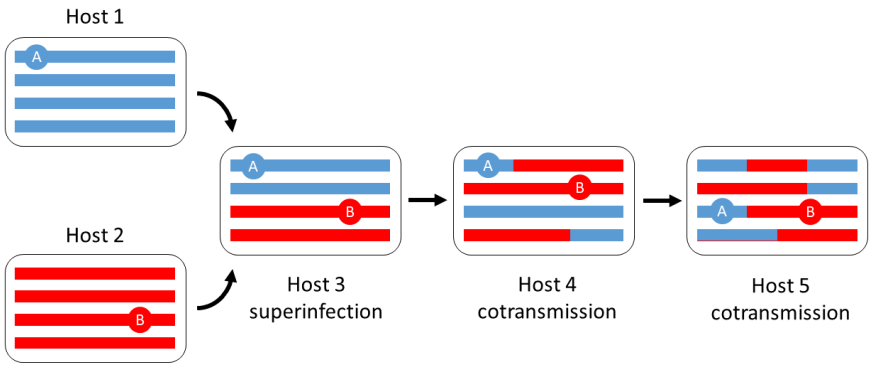

Figure 1: **Superinfection leads to cotransmission and recombination of distinct lineages.** Coloured bars represent parasite genomes within a host. Here we imagine that hosts 1 and 2 each carry a clonal population of parasites, and that the two clonal populations can be differentiated by genotyping, represented by blue and red respectively. Host 3 is superinfected and carries a mixture of the two genotypes. The two genotypes recombine when this mixture is cotransmitted from host 3 to host 4, and there is further recombination when the mixture is cotransmitted from host 4 to host 5. The circles marked A and B represent two different loci in a parasite genome carried by host 5: locus A is inherited from host 1 whereas locus B is inherited from host 2. Thus it is not possible to represent this parasite population by a single phlyogenetic tree - because the red/blue recombinant genomes cannot be mapped onto a single position on the tree.

Thus malaria infections have a broad spectrum of genetic complexity that reflects the local transmission dynamics. In regions of low transmission intensity, where superinfection is rare, most infections are essentially clonal. In regions of high transmission intensity, many infections comprise a mixture of parasites with distinct genotypes, and these can have complex pedigree structures due to repeated cycles of cotransmission and sexual recombination [9].

Modelling a parasite population with superinfection and recombination. One approach to this problem is to build an epidemiological model of malaria transmission coupled to a separate model that simulates the process of genetic variation [10–12]. This allows many biological and epidemiological features to be incorporated into the model, but it tends to be computationally laborious and requires considerable guesswork about parameter values.

Here we describe an alternative approach to modelling the relationship between transmission dynamics and population genetics, based on an idealised transmission graph that takes account of superinfection and cotransmission. We show how this naturally yields a Markov process for coalescent simulation of a recombining parasite population as well as providing a theoretical framework for estimation of transmission parameters from empirical genetic observations.

This paper is theoretical but is illustrated with empirical data relating to the human malaria parasite *Plasmodium falciparum*. *P. falciparum* parasites are single-celled haploid organisms that are transmitted from host to host by a mosquito vector. The parasites reproduce prolifically within the host and vector. Their mode of reproduction is asexual, except at a specific point of the transmission cycle within the mosquito vector, when sexual forms mate to produce recombinant offspring that are transmitted to the next human host (see glossary figure 27).

For clarity, in describing our model we shall use the terms parasite, host and vector to refer specifically to malaria. By *parasite* we mean a haploid individual of the species *P. falciparum*; a *host* is a person that is carrying parasites and is capable transmitting them to others; and a *vector* is a mosquito that transmits parasites from one host to another. However this does not mean that the model is applicable only to malaria, and with suitable modification most of the underlying concepts could equally be applied to other parasites and recombining populations in general.

A glossary of terminology is included before the Methods section. Analyses shown in figures

and tables use the open source Python package `coalestr`. Worked examples and tutorials on using `coalestr` are available at d-kwiat.github.io/gtg.

The concept of a genomic transmission graph

Imagine a directed acyclic graph in which the nodes represent hosts and edges represent vectors, as illustrated in figure 2. The graph is plotted on an axis of time and we make the simplifying assumption that a host exists at a discrete point in time. The directionality of the graph can be viewed in two ways. When thinking about the transmission of parasites from host to host, we are moving forward in time so we follow the edges of the graph from left to right. When thinking about the ancestry of a parasite, we are going back in time and therefore we follow the edges from right to left.

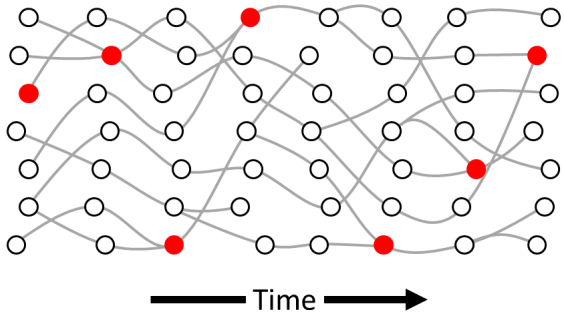

Figure 2: **An example of a transmission graph** representing all the transmission chains in some locality during some interval of time. Each node represents a host, i.e. a person that is carrying parasites and is capable of transmitting them to others. Each edge represents a vector, i.e. a mosquito that transmits parasites from one host to another, and is therefore directed forwards in time. Red marks a node where transmission chains cross, representing a host that is superinfected.

If we pick any node and trace a path forward in time along the edges to some other node, that is a *transmission chain*. Some transmission chains terminate in a host that does not transmit to the next generation. Transmission chains can *branch* when a host is the source of parasites for multiple other hosts. Transmission chains can also *cross* when a host acquires parasites from multiple sources, i.e when there is superinfection. If transmission chains branch but do not cross then the graph will have a tree-like topology. If there is both branching and crossing then the graph will have a reticulate structure as in figure 2.

Parasites reproduce as they flow along transmission chains, and parasites that are flowing along the same transmission chain can genetically recombine with each other. We can use the transmission graph to account for recombination with the aid of three basic concepts:

- A *locus* is a specific location in the genome. This could be anything from a single nucleotide position (which we call a *point locus*) to a whole chromosome.
- An *allele* is an instance of the parasite genome. We usually speak of an allele with reference to a particular locus, in which case it means the DNA sequence of that locus in an individual parasite genome.
- A *lineage* is a path through the transmission graph that we define by taking an allele at a point locus and tracing its ancestry back in time through the generations.

Our definition of a lineage specifically refers to a point locus because this is unaffected by recombination, so we can follow a lineage over many generations despite frequent recombination events. Note that this definition differs from common usage, e.g. in the SARS-CoV-2 literature the term lineage refers to the viral genome as a whole.

An individual parasite could have many different lineages each following a unique path through the graph. To understand how this is possible, imagine two point loci (A and B) in a

parasite's genome. If we trace the two corresponding lineages back in time, they are obliged to follow the same transmission chain until they reach a host that is superinfected, i.e. a node in the graph at which two transmission chains cross. At that point their paths through the graph can diverge, because in the presence of recombination it is possible for locus A to be inherited from one of the transmission chains and locus B from the other (figure 1). This state of affairs means that times to coalescence can vary across the genome (figure 3) as we shall discuss in later sections.

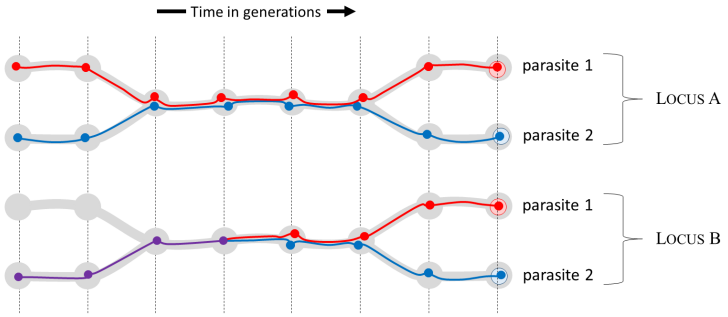

Figure 3: **Recombination causes coalescence times to vary across the genome**. We sample two parasites (red and blue circles at far right) and trace their lineages back in time at two genomic loci, A and B. Grey blobs represent hosts and broad grey lines represent transmission chains. At locus A the red and blue lineages meet in the same transmission chain but separate again before coalescing. At locus B the red and blue lineages meet in the same transmission chain and coalesce to form a single lineage marked in purple.

These simple concepts suggest a logical framework for thinking about how genetic variation is related to transmission dynamics in a recombining parasite population. Instead of attempting to construct a phylogenetic tree, we start by imagining a directed acyclic graph onto which we can map the lineages of different loci in the genome - we call this the genomic transmission graph. This allows for superinfection and recombination because lineages at different loci can take different pathways through the graph. We are left with two fundamental questions that are the main topic of this paper: what are the essential parameters of the genomic transmission graph and how are they mathematically related to parasite genetic variation?

Constructing an idealised genomic transmission graph A comprehensive model of parasite transmission dynamics would require consideration of many factors, e.g. hosts vary in their likelihood of getting infected, their duration of infection and their risk of infecting others, while vectors vary in their biting behaviour. This could be achieved by embedding the genomic transmission graph within an agent-based epidemiological model but it would require the inclusion of a large number of parameters whose values we would need to guess. As our aim is to estimate transmission parameters from genetic data, here we will construct an idealised model that makes a number of simplifying assumptions in order to minimise the number of parameters that need to be estimated.

When thinking about how genomes are transmitted through the generations, it is clear that some individuals have more progeny than others. Hundreds of millions of people around the world are infected with *P. falciparum* and an infected person can carry billions of parasites [8]. However the majority of infected people probably do not pass on parasites to anyone else, and a vector transmits only a small number of parasites from one host to the next [13]. These *transmission bottlenecks* are critical parameters of the genomic transmission graph.

Our idealised model imagines non-overlapping cycles of host to host transmission, and we refer to each cycle as a generation of the transmission graph (Figure 4). We specify that:

1. There are N_h hosts in each generation of the transmission graph: we refer to this as the **effective number of hosts**. We can think of N_h as the number of hosts that are in effect responsible for transmitting parasites from one generation to the next, which is likely to be much less than the total number of infected individuals, and represents a major popula-

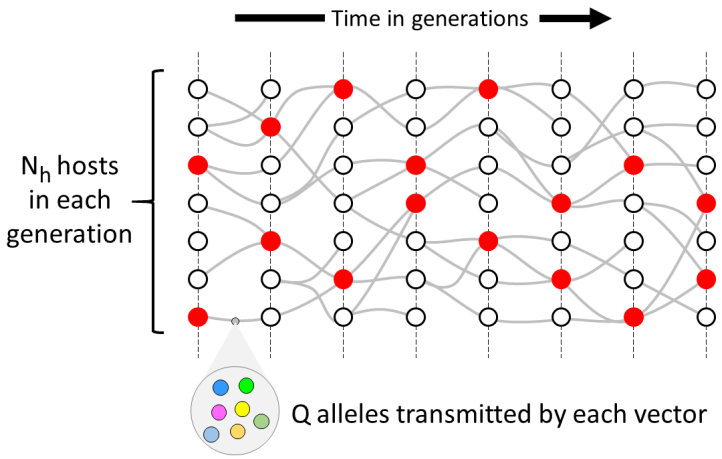

Figure 4: **An idealised model of the genomic transmission graph**. We imagine that transmission occurs in non-overlapping generations. In each generation there are N_h hosts. Each vector transmits Q alleles to the recipient host. χ is the crossing rate of transmission chains: this corresponds to the proportion of hosts that are superinfected from two sources, denoted in the figure by red nodes.

tion bottleneck. The source of infection of a host is determined by random sampling with replacement from the N_h hosts in the previous generation.

2. Each vector transmits Q alleles to the recipient host: we call this the **quantum of transmission**. We can think of Q as the number of parasites that are inoculated by a mosquito into the host, but this is an over-simplification because Q summarises a complex series of bottlenecks in host-vector and vector-host transmission occuring during one generation of the parasite life-cycle [13,14]. The alleles transmitted to the recipient host are copied by random sampling with replacement from the alleles carried by the source host.

3. Each host has either one or two sources of infection in the previous generation, i.e. they are either superinfected or not. The proportion of hosts that are superinfected is denoted χ. From the perspective of the genomic transmission graph, we refer to χ as the **crossing rate of transmission chains**.

This idealised model is obviously not an exact represention of the way things work in the real world. For example, we specify that a host exists at a discrete point in time, so if the same person was infected at multiple time points, we would have to treat these as separate instances of being a host. However our simplifying assumptions will be familiar to population geneticists as they are closely based on the Wright-Fisher model. Indeed if $\chi = 0$ and $Q = 1$ then the transmission graph is equivalent to a Wright-Fisher population of N_h haploid individuals and, as will become clear in the next section, the Wright-Fisher model can be viewed as a special case of the idealised transmission graph.

Serial interval of transmission τ. Let τ be the length of time corresponding to a generation of the transmission graph. From a epidemiological perspective, this is equivalent to the mean serial interval between parasites entering one host and the next on a transmission chain. Our model assumes that the serial interval is constant but in practice it ranges from approximately 6 weeks - the minimum time required for parasites to complete a full developmental cycle within the host and vector - to several years. The mean serial interval probably depends many local factors, including the intensity and seasonality of malaria transmission. At this stage we do not need to specify the value of τ but this will be necessary later when we are estimating rates of mutation and recombination, and for the purpose of illustration we shall suppose that τ is approximately 3 months.

Relationship of χ to incidence of infection. We have defined χ in terms of the transmission graph but what does it mean from an epidemiological perspective? If $\chi = 0$ then each host on the transmission graph acquires infection from exactly one source. We can therefore think of χ as the probability that a host acquires a new infection from another source during the same generation of the transmission graph. If we make the simplifying assumption that this is approximately the same as the probability of any random individual acquiring a new infection, then the incidence of infection (i.e. the rate of infection per unit of time) is given by

$$\text{Incidence of infection} \approx \frac{\chi}{\tau} \tag{1}$$

This theoretical statement should be interpreted with caution as it is based on a number of simplifying assumptions, but it provides a practical motivation for estimating χ from genetic data, as a potential tool for detecting local fluctations in the incidence of infection when this is difficult to monitor by other means.

A framework for analysing the coalescent process

The idealised genomic transmission graph naturally lends itself to coalescent analysis. To start with a simple example, consider the special case of a parasite population with no superinfection, i.e. $\chi = 0$. We shall sample two alleles at a point locus and follow their lineages back in time until they coalesce in a common ancestral allele, as illustrated in figure 5. Let **T** be a random variable representing time to coalescence of the two alleles.

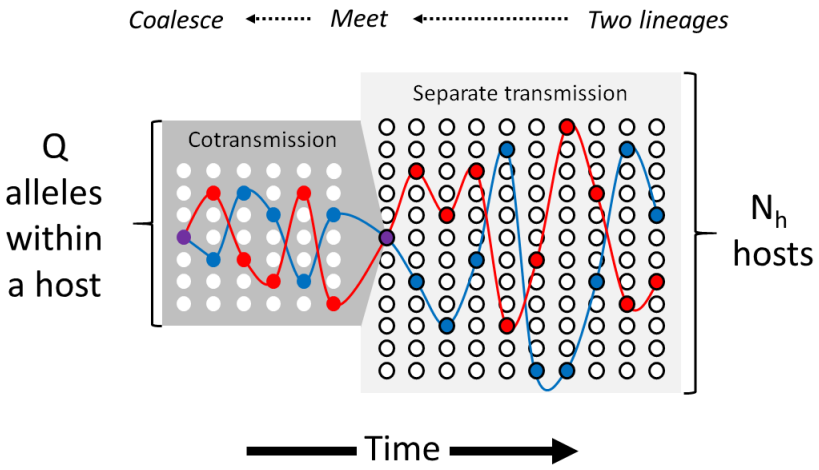

Figure 5: **Coalescence of two lineages in the absence of superinfection**. Each column represents one generation of the transmission graph: the light grey block on the right represents all hosts and the dark grey block on the left represents the within-host population of a single host. Imagine that we sample two alleles from different hosts (red and blue circles on the far right) and follow their lineages back in time. The two lineages remain in separate transmission chains for several generations until they meet in a common host. They are then cotransmitted until they coalesce in a common ancestral allele (purple circle on the far left).

If we sample two alleles from different hosts in the same generation, they are by definition on *separate transmission chains*, but if we trace their lineages back in time they will eventually *meet* in a common host. Once that has happened, the two lineages are *cotransmitted* along the same transmission chain until they eventually *coalesce* in a common ancestral allele.

If two lineages are on separate transmission chains at a particular point in time, there is a probability of $1/N_h$ that they meet in a common host when we go back one generation. If two lineages are cotransmitted, there is a probability of $1/Q$ that they coalesce when we go back one

generation. As described in Methods section 1.1 from this we can obtain the expectation of time to coalescence:

$$E\{\mathbf{T}\} = N_h + Q - 1$$

Thus if $\chi = 0$ and $Q = 1$ then two alleles have an expected coalescence time of N_h generations, equivalent to a Wright-Fisher population of N_h haploid parasites. Likewise, if $\chi = 0$ and $N_h = 1$ then two alleles have an expected coalescence time of Q generations, equivalent to a Wright-Fisher population of Q haploid parasites. Thus we can view the standard Wright-Fisher model as a special case of the genomic transmission graph.

Mapping the coalescent onto the transmission graph. We shall now consider the general case of a parasite population in which superinfection can occur, i.e. $\chi \geq 0$. As in the previous section, we sample two alleles at a point locus and follow their lineages back in time, but the journey to coalescence is more complicated. As illustrated in figure 6, this can be broken down into three stages:

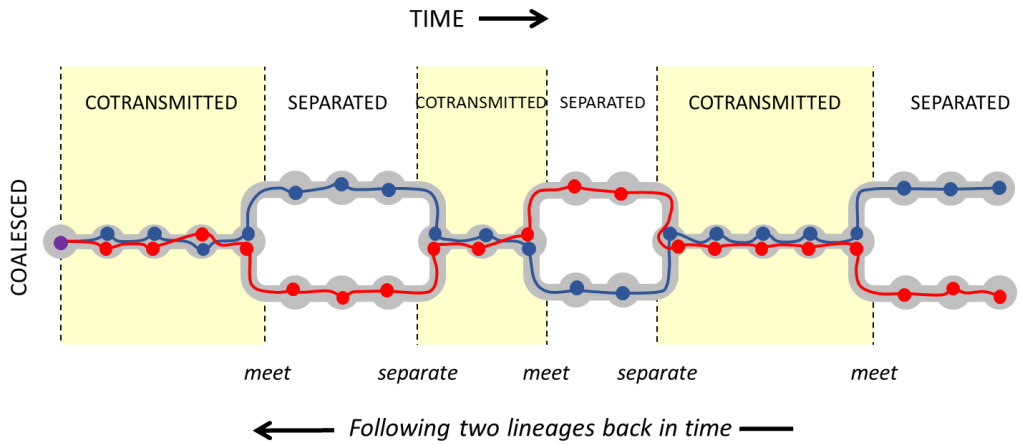

Figure 6: **A graph of coalescing lineages**. Imagine that we sample two alleles from different hosts, as depicted by the red and blue circles on the far right. The corresponding lineages (red and blue lines) can be mapped onto specific transmission chains (thick grey lines) and individual hosts (grey blobs). Proceeding back in time, the corresponding lineages occasionally meet in the same host and are cotransmitted for a period of time before separating again to different hosts. Eventually they meet and coalesce in a common ancestral allele represented by the purple circle on the far left.

1. When we sample two alleles from different hosts in the same generation, at that point in time they are on separate transmission chains, but if we trace their lineages back in time they will eventually meet in the same host.
2. Once the lineages have met in the same host, as we proceed further back in time they are *cotransmitted* along the same transmission chain until one of two events occurs:
   (a) The two lineages *coalesce* in a common ancestral allele.
   (b) The two lineages *separate* onto different transmission chains.
3. If the two lineages separate at stage 2, then we are effectively back at stage 1 and we must wait again for the two lineages to meet in the same transmission chain before they have the possibility of coalescing.

It will be evident that we are dealing with an iterative loop that might need to be repeated for multiple cycles before the two lineages eventually coalesce. If we consider all the possible ways in which two lineages could progress through the transmission graph, at any point in time the system must be in one of three states:

- SEPARATED - the two lineages are in different hosts
- COTRANSMITTED - the two lineages are in the same host
- COALESCED

We can write down the probability of transition between these three states if we follow two lineages back in time by one generation. For example, if two lineages are separated and we go back one generation, there is a probability of $1/N_h$ that they will meet in the same host and, if they do so, then there is a probability of $1/Q$ that they will coalesce in that host.

$$\Pr\{\text{SEPARATED} \rightarrow \text{COALESCED}\} = \frac{1}{N_h Q}$$

To give another example, if two lineages are cotransmitted and we go back one generation, they will separate onto different transmission chains if their current host is superinfected ($\Pr = \chi$) and they come from different source hosts ($\Pr = Q/(2Q-1)$).

$$\Pr\{\text{COTRANSMITTED} \rightarrow \text{SEPARATED}\} = \frac{Q\chi}{2Q-1}$$

In a similar manner we can define the transition probabilities for all possible states of two lineages when we go back in time by one generation, as described in Methods section 1.2, and the results are given in Table 1.

State	Separated	Cotransmitted	Coalesced
Separated	$1-\frac{1}{N_h}$	$\frac{1}{N_h}(1-\frac{1}{Q})$	$\frac{1}{N_h Q}$
Cotransmitted	$\frac{Q\chi}{2Q-1}$	$\frac{(Q-1)(2Q-Q\chi-1)}{Q(2Q-1)}$	$\frac{2Q-Q\chi-1}{Q(2Q-1)}$
Coalesced	0	0	1

Table 1: **Transition probabilities for the three possible states of two lineages**. At any point in time, two lineages must be (1) separated or (2) cotransmitted or (3) coalesced. Row i column j of the table gives the probability that lineages in state i will transition to state j if we go back a single generation. By definition, the probabilities in each row sum to 1.

Markov chain simulation of time to coalescence. Table 1 gives us a transition probability matrix that allows us to evaluate the state of two lineages at any point in time by Markov chain simulation (Methods section 1.3). We start the simulation by sampling two imaginary alleles at a point locus, and then following their lineages back in time through the generations. To study between-host variation we imagine that the two alleles are sampled from different hosts, i.e. the two lineages are separated at the start of the simulation. Alternatively, we can study within-host variation by imagining that the two alleles are sampled from the same host, i.e. the lineages are cotransmitted at the start of the simulation. From this we can calculate the probability distribution of coalescence time for two alleles, sampled either between-host or within-host, as illustrated in figure 7.

The probability distribution of coalescence times depends on the combination of transmission parameters, and it contrasts with the classic Wright-Fisher coalescent process which invariably gives a geometric distribution. In the case of between-host variation (figure 8 upper panels), when

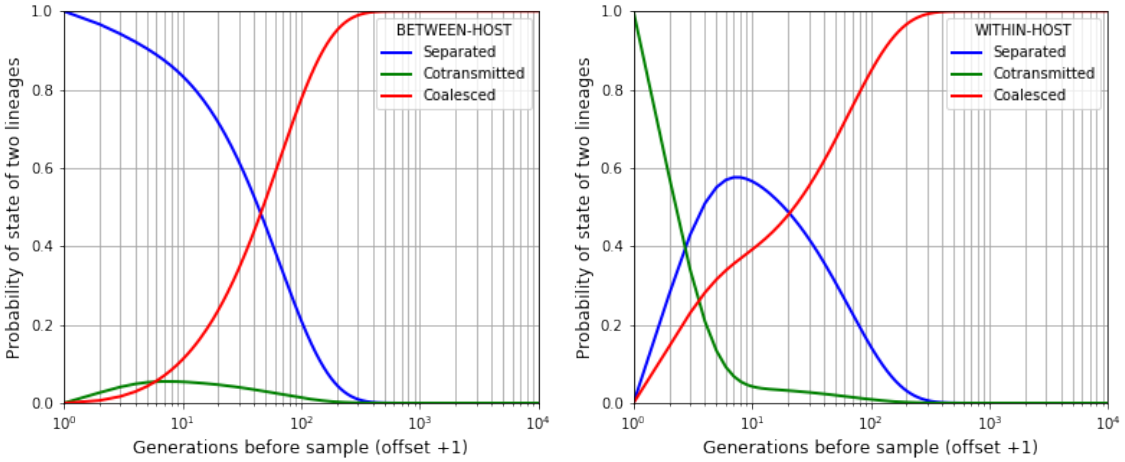

Figure 7: **State of two lineages as we proceed back in time after sampling two alleles.** If we sample the alleles from different hosts (**left panel**) then the lineages are initially separated (blue) but over time there is an increasing probability that they are cotransmitted (green) and eventually they coalesce (red). If we sample alleles from the same host (**right panel**) then the lineages are initially cotransmitted. In this example the transmission parameters are $N_h = 30$, $Q = 5$, $\chi = 0.5$. The horizontal axis is offset by +1 to allow the use of a log scale. See worked example.

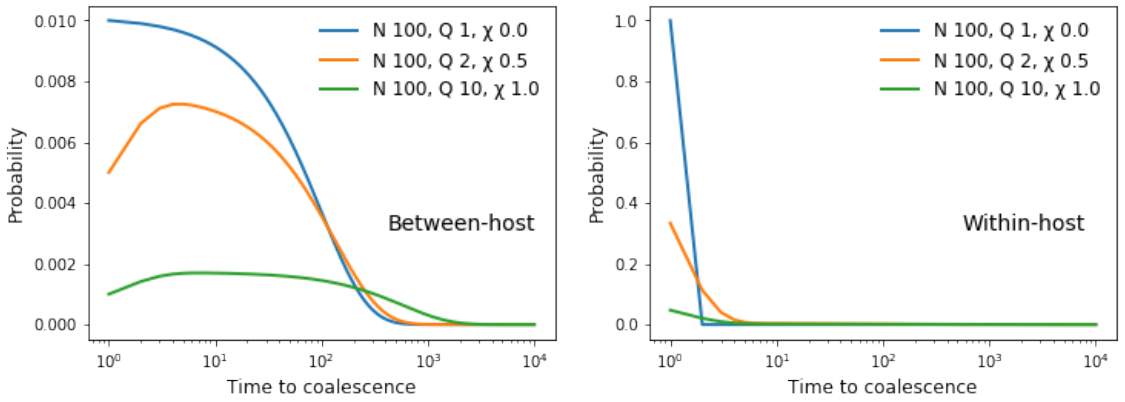

Figure 8: **Between-host and within-host times to coalescence.** We use a Markov process to compute the probability distribution of coalescence times for two alleles sampled from different hosts (**left panel**) or from the same host (**right panel**) for different combinations of transmission parameters. Note that the two panels have markedly **different scales**, and that two alleles sampled from the same host coalesce much more rapidly than two alleles sampled from different hosts when $\chi = 0$, whereas the difference is less marked when $\chi = 1$.

$Q > 1$ the rate of coalescence tends to increase over the first few generations, and depending on χ it may then stay fairly constant for some time before declining asymptotically to zero.

In general, two alleles sampled from the same host coalesce more rapidly than two alleles sampled from different hosts. The difference is marked when $\chi = 0$. As χ increases, the coalescence time of two alleles sampled from the same host starts to approach that of two alleles sampled from different hosts, and when $\chi = 1$ the difference becomes relatively small. An exception to this general rule is discussed in Methods section 1.4.

Genetic variation at a point locus

Now that we have a way to estimate time to coalescence, we can use this to estimate levels of genetic variation in the parasite population. We start by focusing on an imaginary locus in the

parasite genome. We refer to a single nucleotide position as a *point locus* and we define a *haplotype locus* as a sequence that extends over multiple nucleotide positions (see glossary figure 26). A haplotype locus can undergo recombination whereas a point locus cannot. We shall discuss haplotype loci in the next section but here we focus on point loci, so we can ignore recombination for the present.

We say that alleles are *homozygous* if they have the same DNA sequence, and that they are *heterozygous* if they have different DNA sequences. We define *homozygosity* as the probability that two randomly sampled alleles at a particular locus are homozygous, and *heterozygosity* as the probability that they are heterozygous. Following population genetics convention, we shall denote homozygosity by G and heterozygosity by H, where $H = 1 - G$.

Imagine that we sample two alleles at a point locus and trace their lineages back in time until they coalesce. Let u be the mutation rate per generation at this locus and let $\mathbf{T}$ be a random variable representing time to coalescence of the two lineages. The two alleles must have the same DNA sequence if neither lineage is affected by mutation from the time of sampling to the time of coalescence, so as described in Methods section 2 we can obtain the expectation of heterozygosity

$$E\{H\} = 1 - \sum_{i=1}^{\infty} \Pr\{\mathbf{T} = i\} \times (1-u)^{2i}$$

Let T_C be the mean time to coalescence measured in generations. If u is sufficiently small (say $< 10^{-5}$) we can safely ignore factors of u^2 and above to make the approximation

$$H \approx 2uT_C \tag{2}$$

Mechanisms of mutation at a point locus. There are various mechanisms of mutation, each causing a characteristic type of genetic varation. They include substitutions, insertions, deletions and structural rearrangements. Here we focus on single nucleotide substitution, the mutational process that causes a very common type of genetic variation known as a single nucleotide polymorphism (SNP). SNPs naturally correspond to point loci and are convenient for analysis because they are relatively easy to ascertain using current genome sequencing technologies. We can therefore use the rate of single nucleotide substitution as a form of molecular clock as we track lineages back in time.

Single nucleotide substitution rate μ. Let μ be the probability of single nucleotide substitution occuring at a point locus during one generation of the transmission graph. *In vitro* studies of *P. falciparum* clone trees have estimated the probability of a single nucleotide substitution to be in the region of 10^{-9} to 10^{-10} per nucleotide during each 48-hour cycle of replication within erythrocytes [15–17]. Here we shall use the rather conservative estimate of 1.2×10^{-10} per nucleotide per day based on the largest of these studies [17]. We do not know the rate of mutation at other stages of the life cycle, e.g. when parasites replicate within the mosquito or in the human liver, but let us assume that 1.2×10^{-10} per nucleotide per day is representative of the entire life cycle. If we also assume that the serial interval of transmission τ is 3 months, we obtain an estimate of $\mu \approx 1.1 \times 10^{-8}$ per generation.

Nucleotide diversity π. We define nucleotide diversity as the probability that two alleles are heterozygous at a random nucleotide position in the genome, and we denote this by π. The value of π will vary from population to population but we can measure it using genome sequence data, and it provides a direct estimate of the genome-wide average of heterozygosity for all point loci. This provides a starting point for analysis of parasite population history as it allows us to estimate the mean time to coalescence. If we take equation 2 and substitute π and μ respectively for H and u, we obtain

$$T_C \approx \frac{\pi}{2\mu} \tag{3}$$

Measuring π in a parasite population is straightforward in principle: we randomly sample parasites from the population, obtain their genome sequences, and analyse the number of pairwise differences between individual genome sequences. In practice, the definition of π needs to be modified slightly to make these empirical measurements consistent with our definition of μ, and to exclude potential sources of error and bias.

Nucleotide diversity of the global parasite population. From large genome sequencing studies of thousands of *P. falciparum* samples from around the world, we can get an estimate of global, regional and local levels of nucleotide diversity [18, 19]. There are some technical caveats about the precision of these estimates, as discussed in Methods section 2.1, but for present purposes we can use $\pi \approx 4 \times 10^{-4}$ as a first approximation for the global parasite population. This estimate is obtained by analysis of coding SNPs as opposed to other types of genetic variation. It is restricted to SNPs because our estimate of μ is based on the rate of single nucleotide substitution. It excludes SNPs in non-coding regions because, in the case of *P. falciparum*, these are error-prone due to many tandem repeat sequences.

If $\pi \approx 4 \times 10^{-4}$ and $\mu = 1.1 \times 10^{-8}$, then equation 3 tells us the mean time to coalescence for two alleles sampled at random from the global parasite population is approximately 18,000 generations. This is equivalent to 4,500 years since we are assuming that each generation has a duration τ of 3 months. If we specified a different value for τ this would change our estimates of μ and of T_C measured in generations, but it would still give us a mean coalescence time of approximately 4,500 years.

Transmission parameters compatible with global levels of nucleotide diversity. Knowing the single nucleotide substitution rate μ, we can use Markov chain simulation of coalescence times to explore what combinations of transmission parameters would be compatible with the observed levels of nucleotide diversity in the global parasite population. Undoubtedly the transmission parameters have varied considerably over the past few thousand years, but for the purpose of illustration we shall assume here that they are constant over time.

As an example, the combination of $N_h = 18764$, $Q = 1$, $\chi = 0$ would give $\pi = 4 \times 10^{-4}$. As we noted in the previous section, a transmission graph with $Q = 1$, $\chi = 0$ is equivalent to a Wright-Fisher population of N_h haploid individuals. In other words, if we applied the Wright-Fisher model to these data, we would obtain an effective population size of 18,764 haploid individuals.

Another possible combination is $N_h = 3269$, $Q = 10$, $\chi = 1$. This gives the same value of nucleotide diversity in the general parasite population but a much higher level of within-host diversity, both because transmission chains cross more frequently, and also because the transmission bottleneck is not so tight. Other examples that lie in between these two extremes are shown in table 2.

χ	Q	N_h	π_T	$\widehat{\pi}_W$
0	1	18,764	4.0×10^{-4}	2.2×10^{-8}
0	10	18,754	4.0×10^{-4}	2.2×10^{-7}
0.5	1	18,764	4.0×10^{-4}	2.0×10^{-4}
0.5	10	5,568	4.0×10^{-4}	3.1×10^{-4}
1	1	18,764	4.0×10^{-4}	4.0×10^{-4}
1	10	3,269	4.0×10^{-4}	3.7×10^{-4}

Table 2: **Examples of transmission parameters giving $\pi = 4 \times 10^{-4}$.** Here we use a simplistic model with constant population size to simulate the heterozygosity of a point locus. The input parameters are N_h, Q and χ. The results are π_T, the nucleotide diversity of the total parasite population, and $\widehat{\pi}_W$, the mean level of within-host nucleotide diversity. See worked example.

It is evident from these observations that current levels of nucleotide diversity in the global parasite population have built up over thousands of years. To put this in context, various lines of evidence indicate that *P. falciparum* originated in Africa and underwent a major population expansion somewhere in the region of 10 to 50 thousand years ago [20, 21]. Figure 9 illustrates this with a toy model in which nucleotide diversity gradually accumulates over time to reach a level of $\pi_T \approx 4 \times 10^{-4}$ in the total parasite population and of $\pi_W \approx 1 \times 10^{-4}$ in the within-host population.

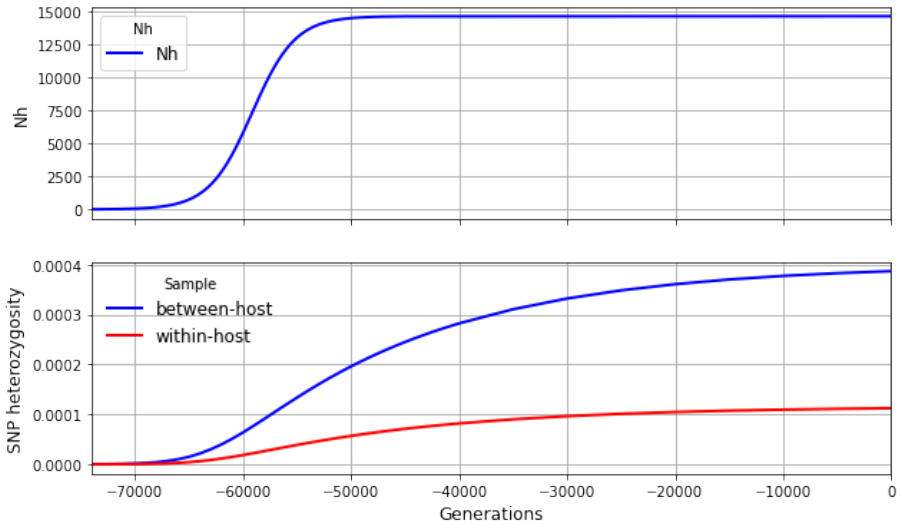

Figure 9: **A toy model of the global parasite population**. In this very simplistic scenario, there is a major population expansion approximately 15 thousand years before the present, eventually reaching a plateau with $N_h = 14,660$, $Q = 3$ and $\chi = 0.2$ which is maintained over the past few thousand years. Nucleotide diversity gradually increases to $\pi_T \approx 4 \times 10^{-4}$ in the total parasite population and $\pi_W \approx 10^{-4}$ in the within-host population. To explore other simple scenarios see this Jupyter notebook.

Local levels of nucleotide diversity. We might intuitively expect the nucleotide diversity of local parasite populations to be much lower than that of the global parasite population, but empirical measurements show that this is generally not the case. Throughout much of Africa, the nucleotide diversity of the parasite population in a large village can be almost as high as that in the global population. A modest reduction in nucleotide diversity is observed in parts of Southeast Asia where transmission intensity is much lower than Africa [18, 19].

As we shall discuss later, relatively low levels of ongoing migration across the global metapopulation can maintain high levels of nucleotide diversity in a local subpoulation. It is quite possible that human migration out of Africa was a major factor in dispersing *P. falciparum* around the world, but it is equally possible that the nucleotide diversity of local parasite populations has been maintained by human migration in the modern era, and these two possibilities are not mutually exclusive.

Therefore we might not be able to learn much about local patterns of malaria transmission from levels of nucleotide diversity in the local parasite population, because these are greatly influenced by historical patterns of global population expansion and long-range dispersal. However this background genetic diversity is extremely useful for other metrics that are more informative about recent local transmission, such as within-host heterozygosity, haplotype homozygosity and local population structure, as we shall discuss in the following sections.

Genetic variation at a haplotype locus

We need to take account of recombination as well as mutation at a haplotype locus (see glossary figure 26). For a locus spanning hundreds of kilobases, the rates of recombination and mutation

may be so high that it is rare for two alleles to have the same haplotype, i.e. the same DNA sequence. Let G_L be the probability that two alleles have the same DNA sequence at a haplotype locus of length L. We call this *haplotype homozygosity* and we would like to be able to estimate its value for any combination of transmission parameters.

Locus scaled recombination rate r. Let r be the rate of recombination at a haplotype locus, scaled by its length in kilobases. For a locus of length L kilobases, the probability of recombination occurring within the locus during one generation of transmission is rL.

Plasmodium parasites reproduce asexually for most of their life cycle but shortly after entering a mosquito vector they undergo sexual mating, with the result that recombination occurs exactly once per generation of host to host transmission. The rate of recombination between two point loci is conventionally expressed in centimorgans, where 1 centimorgan denotes 1% probability of recombination per generation. From experimental genetic crosses it has been estimated that 1 centimorgan is equivalent to 13.5 kilobases when averaged across the *P. falciparum* genome [22]. From this we obtain an estimate of $r = 7.4 \times 10^{-4}$ per kilobase per generation.

Locus scaled mutation rate v. Let v be the rate of mutation at a haplotype locus, scaled by its length in kilobases. For a locus of length L kilobases, the probability of a mutation occurring within the locus during one generation of transmission is vL.

To estimate v we must consider all types of mutation that might alter the DNA sequence of a haplotype locus. *P. falciparum* has a very high rate of indel mutation, estimated by laboratory studies *in vitro* to be $\sim 2 \times 10^{-9}$ per nucleotide per 48 hour erythrocytic growth cycle [17]. This is much greater than the single nucleotide substitution rate, but these indel mutations occur mainly within short tandem repeat sequences, so they probably include many recurrent mutations and reversions. Other types of mutation, such as large copy number variations and structural variations, are much less common. In principle we could assign different values to v depending on genomic location, since short tandem repeat sequences and indel mutations are concentrated largely in non-coding regions, but for present purposes we shall assume that the mutation rate is constant across the genome and across the life cycle. If we take the rate of indel mutations plus single nucleotide substitutions to be 10^{-9} per nucleotide per day, and if we assume a serial interval of $\tau = 3$ months, we obtain an estimate of $v = 9 \times 10^{-5}$ per kilobase per generation.

Inaccessible regions of the parasite genome. Here we focus on haplotype loci within the 'core' *P. falciparum* genome which excludes the sub-telomeres and a few other regions that are extremely difficult to sequence using short-read technologies because of their exceptionally complex patterns of polymorphism [22]. These hypervariable regions contain genes involved in immune evasion that undergo frequent structural rearrangements by means of a specialised mutational process called non-allelic homologous recombination [16]. Highly mutable genes could in theory be extremely informative about transmission dynamics, but for present purposes we treat them as inaccessible to haplotypic analysis because they cannot be reliably ascertained in field samples with current methodologies.

Determinants of haplotype homozygosity. What is the expected homozygosity of a haplotype locus of length L? To address this question, let us randomly sample two alleles from the parasite population and call them alleles 1 and 2. We randomly select a single nucleotide position within our haplotype locus and call it point A. Let A_1 and A_2 be the lineages corresponding to alleles 1 and 2 at point locus A. Let $\mathbf{T}$ be a random variable representing time to coalescence of the A_1 and A_2 lineages. Time to coalescence can vary across a haplotype locus (figure 3) but, as long as the probability distribution of time to coalescence is the same for every point locus, for the following calculations it does not matter where point A is situated within our haplotype locus.

By definition, the A_1 and A_2 lineages coalesce when they meet in the same ancestral parasite, whose DNA sequence we will call their *common ancestral haplotype* (figure 10). We are interested in what happens to this common ancestral haplotype over the course of $\mathbf{T}$ generations between the time of sampling our two alleles and the time of coalescence. If we track the haplotype associated

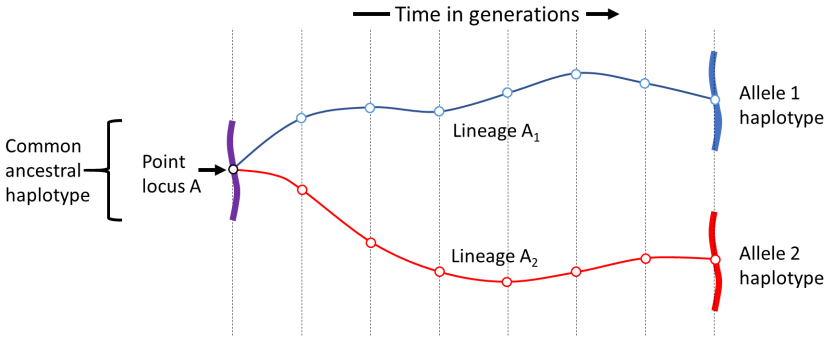

Figure 10: **Illustration of the principle of haplotype homozygosity.** We imagine a haplotype locus of length L kilobases and a point locus A (open circle) somewhere within this haplotype locus. We sample two alleles at point locus A and trace their lineages back in time until they coalesce in a common ancestral haplotype. Haplotype homozygosity is the probability that the haplotype locus will be unaffected by recombination or mutation in the time that it takes for the two lineages to coalesce.

with either the A_1 or the A_2 lineage over time, in each generation there is a probability of vL that it will be affected by mutation and of rL that it will experience recombination.

Recombination will change the DNA sequence of the haplotype in some circumstances but not others, e.g. it might not do so if the recombining parasites are siblings and have identical haplotypes at this locus. We are interested in the frequency of *effective recombination* which we define as recombination between genetically distinct alleles that acts to change the DNA sequence of the haplotype locus. Suppose that recombination occurs at time t and let ϕ_t be the probability that this recombination event results in a change to the DNA sequence of this locus. This means that the probability of effective recombination is $\phi_t rL$ and the probability that the haplotype remains unchanged over the course of one generation is $(1 - vL - \phi_t rL)$.

By following both lineages over $\mathbf{T}$ generations to their point of coalescence, we can obtain the probability that alleles 1 and 2 have retained the common ancestral haplotype, which gives us the probability G_L that the alleles are homozygous at this haplotype locus.

$$G_L = \prod_{t=1}^{\mathbf{T}} (1 - vL - \phi_t rL)^2$$

We obtain the expectation of haplotype homozygosity by summation across the probability distribution of coalescence times:

$$E\{G_L\} = \sum_{i=1}^{\infty} \Pr\{\mathbf{T} = i\} \prod_{t=1}^{i} (1 - vL - \phi_t rL)^2 \tag{4}$$

Effective recombination parameter ϕ_t. We are left with the question of how to estimate ϕ_t which we call the *effective recombination parameter*. In order for sexual recombination to change the DNA sequence of a haplotype, it is necessary for the mating parasites to be heterozygous at that locus. We could therefore say that ϕ_t is equivalent to the probability that the mating parasites are heterozygous, which is given by the value of within-host heterozygosity H_W at that locus at time t. However this assumes random mating (i.e. that a vector randomly samples parasites from a host, and that these randomly mate within the vector) whereas a number of empirical studies have found evidence of mating bias and a tendency to selfing. Another complication is that the recombination of two heterozygous haplotypes might not result in a new haplotype if their DNA sequences are similar, especially if the recombination breakpoint is away from the centre of the locus. Therefore we shall say that

$$\phi_t = f\widehat{H}_W \tag{5}$$

where $\widehat{H}_W$ is the mean level of within-host heterozygosity in the population at time t, and f is a factor that we use to correct for mating bias and other causes of non-effective recombination. The value of f is in the range [0,1] where $f = 1$ indicates that there is no mating bias and that the recombination of two heterozygous haplotypes always results in a new haplotype.

It will be evident from equations 4 and 5 that evaluation of haplotype homozygosity at a particular point in time requires knowledge of within-host heterozygosity at multiple previous time points, i.e. this is a non-Markovian process. We use a heuristic approach to solve this problem. We start by assuming some arbitrary value for $\widehat{H}_W$ in the distant past and then progressively construct a time series of $\widehat{H}_W$ values by forwards-in-time simulation, pausing to perform backwards-in-time Markovian simulation of coalescence times for each new timepoint.

Relationship of haplotype homozygosity to haplotype length. Using the above principles we can determine the expected haplotype homozygosity for a locus of any given length. Figure **??** shows how haplotype homozygosity falls away rapidly as haplotype length increases. Levels of haplotype homozygosity are much higher if we sample alleles from the same host compared to sampling from different hosts, as we would expect. Here we see that haplotype homozygosity declines as population size and the rate of superinfection increase, but there can be long stretches of haplotype homozygosity in within-host samples when there is no superinfection.

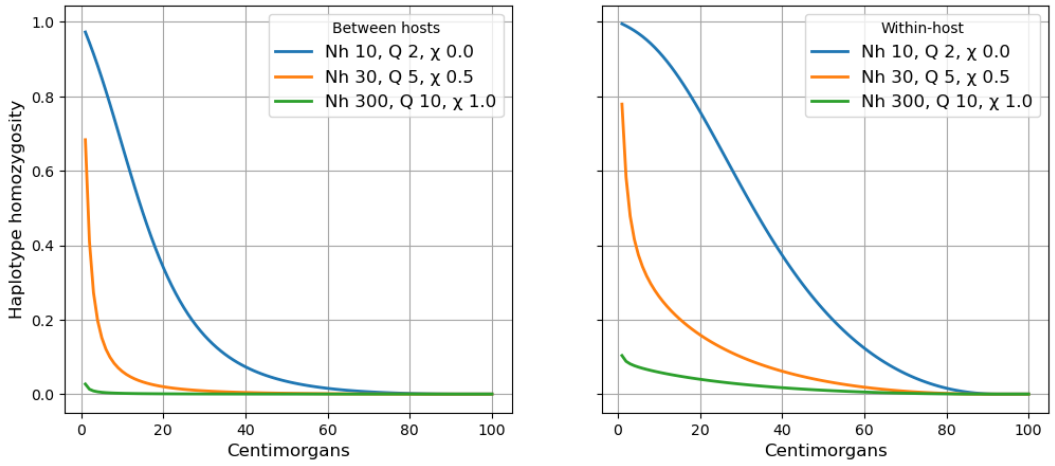

Figure 11: **Relationship between haplotype homozygosity and haplotype length.** Left panel shows between-host variation (i.e. two alleles sampled from different hosts), right panel shows within-host variation (i.e. two alleles sampled from the same host). See worked example.

Shared haplotype segments, recent common ancestry and identity by descent. If we compare the genome sequences of two parasites, we can identify segments of the genome where their haplotypes are identical. We call these *shared haplotype segments*. Unrelated parasites often have some shared haplotype segments that extend over a few kilobases, but if we observe a substantial number of shared haplotype segments that are hundreds of kilobases long, this suggests that the parasites share a recent common ancestor.

Identity by descent (IBD) is a population genetic term that refers to genome sequences that are identical between individuals due to recent common ancestry [23]. There are various methods to estimate levels of IBD for *P. falciparum* using whole genome sequence data [24, 25] or genetic barcodes such as SNP panels, microsatellites and microhaplotypes [26–29]. A commonly used metric of genetic relatedness between individuals is the proportion of the genome that is IBD.

A rather simplistic view of whole genome IBD methods is that they detect shared haplotype segments of above a certain size, typically around 2 centimorgans. In general we are more likely to observe recent common ancestry and high levels of IBD if the parasite population size is small. This raises the question of what is the expected proportion of the genome that is IBD for a given set of transmission parameters.

As shown in Methods section 3.1 the proportion of the genome occupied by shared haplotype segments of $> k$ centimorgans can be crudely approximated by $E\{G_k\}$, the expected haplotype homozygosity of a locus of k centimorgans. Thus if we define shared haplotype segments of > 2 centimorgans as IBD, then the proportion of the genome that is IBD is approximated by the mean homozygosity of a haplotype locus of 2 centimorgans.

Let γ be the mean haplotype homozygosity of a 2 centimorgan locus, which corresponds to 27 kilobases if we assume that 1 centimorgan is equivalent to approximately 13.5 kb on average. Figure 12 shows how γ varies with different transmission parameters, where γ_S represents the local subpopulation and γ_W the within-host population. γ_S declines rapidly with increasing levels of N_h, χ and Q. In the absence of superinfection, γ_W is high and independent of N_h.

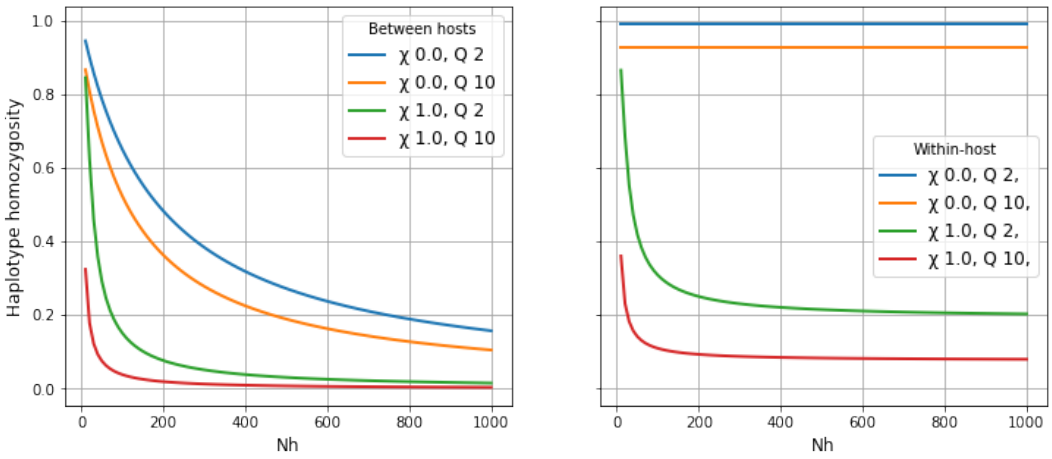

Figure 12: **Haplotype homozygosity of a 2 centimorgan locus.** This metric, denoted γ, is approximately equal to the proportion of the genome that is identical by descent (IBD) between individuals randomly sampled from a population, as discussed in the main text and Methods section 3.1. Left panel shows γ_S (comparing two alleles sampled from different hosts in the local subpopulation), right panel shows γ_W (comparing two alleles sampled from the same host). See worked example.

Population structure and migration

So far we have treated the parasite population as a homogenous entity but it is more like a set of interconnnected subpopulations each inhabiting a particular geographical area. There is spatial structure in transmission dynamics, e.g. neighbouring villages can vary in malaria prevalence due to differences in their mosquito breeding sites and other factors [30,31]. There is also a global population structure with genetic differentiation between continental regions, some of which reflects evolutionary adaptation to the resident vector and host populations [19,32,33].

Essentially we have many *local subpopulations* that are more or less loosely connected with each other and together make up a *metapopulation*. We could break this down into many different levels of spatial scale within a hierachical population structure, e.g. local subpopulations could be embedded within regional metapopulations which are themselves embedded within the global metapopulation, as illustrated in figure 13. By convention, we use the subscript $_S$ to denote a local subpopulation and $_T$ to denote the total (or global) metapopulation.

Effect of global parasite dispersal on local population genetics. Migration from the global metapopulation into a local subpopulation - either ongoing or due to historical patterns of global parasite dispersal - can have a profound effect on local population genetics. To illustrate this let us consider the simple scenario of a local subpopulation within a much larger metapopulation.

Let m be the probability that a host within the local subpopulation acquired their infection from the metapopulation, and let the number of such hosts per generation be $N_m = mN_h$. These migrant hosts could be either immigrants from the metapopulation or local residents who have

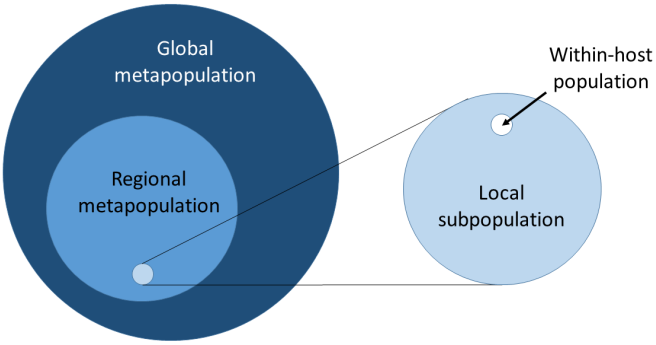

Figure 13: **Hierarchical population structure.** Here we imagine that local subpopulations (e.g. villages) are embedded within much larger regional metapopulations (e.g. West Africa or Southeast Asia) which themselves are embedded within the global metapopulation. The local subpopulation can itself be broken down into multiple within-host populations.

been travelling outside the local area. Methods section 4 describes how we can work out the transmission probability matrix for the subpopulation, shown in table 5, while the transition probability matrix of the metapopulation is essentially as described in table 1.

Table 3 illustrates the effects of migration on the nucleotide diversity of a local subpopulation ($N_h = 30$) embedded within a much larger metapopulation ($N_h = 3000$). In the absence of migration, the nucleotide diversity of the subpopulation ($\pi_S = 2.4 \times 10^{-6}$) is two orders of magnitude lower than that of the metapopulation ($\pi_T = 3.7 \times 10^{-4}$). With a migration rate of just one host every ten generations ($N_m = 0.1$) it increases dramatically ($\pi_S = 1.5 \times 10^{-4}$) and with a migration rate of one host per generation it is almost the same as the nucleotide diversity of the metapopulation.

N_m	π_T	π_S	γ_T	γ_S	F_{ST}
0	3.7×10^{-4}	2.4×10^{-6}	0.004	0.41	0.99
0.1	3.7×10^{-4}	1.5×10^{-4}	0.004	0.32	0.59
1	3.7×10^{-4}	3.3×10^{-4}	0.004	0.10	0.12
10	3.7×10^{-4}	3.7×10^{-4}	0.004	0.01	0.01

Table 3: **Effect of global dispersal on local genetic diversity.** N_m is the number of migrants per generation entering a local subpopulation ($N_h = 30, \chi = 0.5, Q = 10$) from a much larger global metapopulation ($N_h = 3000, \chi = 1, Q = 10$). The table shows the nucleotide diversity of the metapopulation (π_T) and the subpopulation (π_S); the haplotype homozygosity of a 2cM locus in the metapopulation (γ_T) and the subpopulation (γ_S); and F_{ST}, the fixation index of the subpopulation relative to the metapopulation. See worked example.

We saw in section that current levels of nucleotide diversity in the global parasite population have built up over thousands of years, and here we see how relatively low levels of migration from the global metapopulation can cause a local subpopulation to achieve relatively high levels of between-host nucleotide diversity.

Using fixation indices as a measure of hierarchical population structure. In describing the effects of migration on population structure, it is helpful to use Wright's fixation index:

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

where H_T is the heterozygosity of the total (or global) metapopulation and H_S is the heterozygosity of a local subpopulation. As shown in figure 14, F_{ST} is inversely related to the rate of migration and the size of the local population.

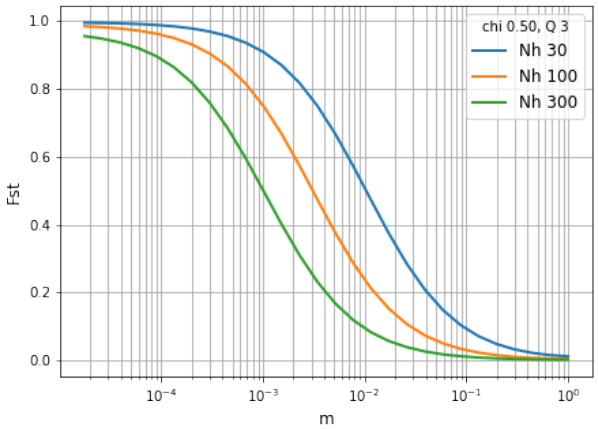

Figure 14: **Relationship between F_{ST} and the rate of migration in a hierarchical population structure.** We imagine a local subpopulation embedded within a metapopulation of $N_h = 3000, Q = 10, \chi = 1$. m is the probability that a host within the local subpopulation acquired their infection from the metapopulation. F_{ST} is inversely related to m and also to the size of the local population. View code

Estimating the quantum of transmission

In this section and the next we explore ways to infer local transmission parameters from measurements of within-host variation. With modern sequencing technologies it is feasible to measure within-host variation at millions of SNP loci. For each host, we can calculate the mean of within-host heterozygosity for all nucleotide positions across the genome, π_W, which is the within-host equivalent of nucleotide diversity. We will show how this can be used to estimate the quantum of transmission Q.

Unlike the coalescent approach used in previous sections, which proceeded backwards in time, we shall now imagine that we are following parasites as they flow along an individual transmission chain, and consider the effects of mutation, genetic drift and superinfection as we proceed forwards in time. Although this approach requires some approximation and is more cumbersome than the coalescent approach, it provides valuable insights into how the transmission parameters Q and χ could be estimated by deep sequencing of individual infections.

Within-host heterozygosity in the absence of superinfection. Consider a transmission chain that never crosses with another transmission chain, i.e. there is no superinfection. Imagine two hosts that are x generations apart on this transmission chain. Let H_W be the within-host heterozygosity of the first host and H'_W that of the second at some arbitrary point locus. As parasites flow from the first to the second host, genetic drift due to the transmission bottleneck will act to reduce heterozygosity, while mutation will act to increase heterozygosity. Here we will focus on SNPs so the relevant mutation rate is that of single nucleotide substitution $\mu \approx 1.1 \times 10^{-8}$ per generation. As described in Methods section 5.1, the Wright-Fisher model gives us the relationship

$$H'_W \approx H_W \alpha^x + 2\mu \sum_{i=0}^{x-1} \alpha^i \tag{6}$$

where

$$\alpha = \left(1 - \frac{1}{Q}\right)(1 - 2\mu) \tag{7}$$

If we follow this transmission chain over time, it will eventually reach an equilibrium level of within-host heterozygosity as long as it does not cross with another transmission chain. We can evaluate this equilibrium value by letting $x \to \infty$ in equation 6. We can get an empirical estimate of this value by using deep sequencing to determine π_W, the mean within-host heterozygosity at all nucleotide positions in the parasite genome. As shown in Methods section 5.1, in the absence of superinfection

$$Q \approx \frac{\pi_W(1 - 2\mu)}{2\mu(1 - \pi_W)} \approx \frac{\pi_W}{2\mu} \tag{8}$$

This is reminiscent of equation 3 which gave $T_C \approx \pi/2\mu$. Here we have a special case of the genomic transmission graph where $T_C = Q$ because we are sampling two alleles that are cotransmitted and because $\chi = 0$.

Inferring Q from measurements of within-host nucleotide diversity π_W. Equation 8 provides a way of estimating the quantum of transmission Q by deep sequencing of the parasite genome within individual hosts. It requires that we sample from hosts who lie on transmission chains that have not experienced superinfection at any time in the recent past. We would expect this to include a relatively high proportion of hosts in regions with low malaria transmission intensity, e.g. South America, but to be much less common in regions with high transmission intensity such as West Africa.

It is beyond the scope of this paper to carry out a sufficiently detailed analysis of empirical data to make a reliable estimate of the quantum of transmission, but we can make a crude preliminary estimate as proof of concept using genome variation data from a global sample of thousands of malaria-infected individuals produced by the MalariaGEN network [18]. The methods of this preliminary analysis are described in Methods section 5.2. As shown in figure 15 we find a striking bimodal distribution for π_W, with the first peak comprising hosts with low π_W ($\sim 4 \times 10^{-7}$) and the second peak comprising hosts with high π_W ($\sim 5 \times 10^{-5}$).

Here we postulate that the high π_W peak is caused by hosts with superinfection and cotransmission whereas the low π_W peak is caused by hosts that lie on transmission chains that have not experienced superinfection in the recent past. This interpretation of the data is supported by the further observation that the relative heights of the two peaks vary according to the population sampled, with the low π_W peak being more prominent in regions of low transmission and the high π_W peak more prominent in regions of high transmission, as outlined in Methods section 5.2. However we need to be aware of potential artefacts caused by genotyping errors, and in particular the phenomenon of hyper-heterozygosity due to incorrect alignment of short sequencing reads, as described in reference [32]. In the data shown in figure 15 we attempt to exclude such artefacts by analysing only biallelic coding SNPs with good data quality scores and with mean minor allele frequency of $< 1\%$ across all samples.

Taking the value of $\pi_W \approx 4 \times 10^{-7}$ together with a single nucleotide substitution rate of $\mu \approx 1.1 \times 10^{-8}$, equation 8 gives us

$$Q \approx \frac{\pi_W}{2\mu} \approx \frac{4 \times 10^{-7}}{2 \times 1.1 \times 10^{-8}} \approx 18 \tag{9}$$

Various experimental studies have quantified the number of sporozoites inoculated by an infectious mosquito: Rosenberg *et al* estimated a median of 15 with very wide range [34] while Beier *et al* estimated a geometric mean of 4.5 [35]; and a review of the literature by Graumans *et al* states that median inocula ranged between 8 and 39 sporozoites [13].

It is reassuring that our preliminary estimate of $Q \approx 18$ is consistent with these experimental data. However it would be wrong to treat Q as a simple estimate of the number of inoculated sporozoites, as it summarises a series of transmission bottlenecks that occur before and after an infectious mosquito bite. Other caveats about this estimate are discussed in Methods section 5.2.

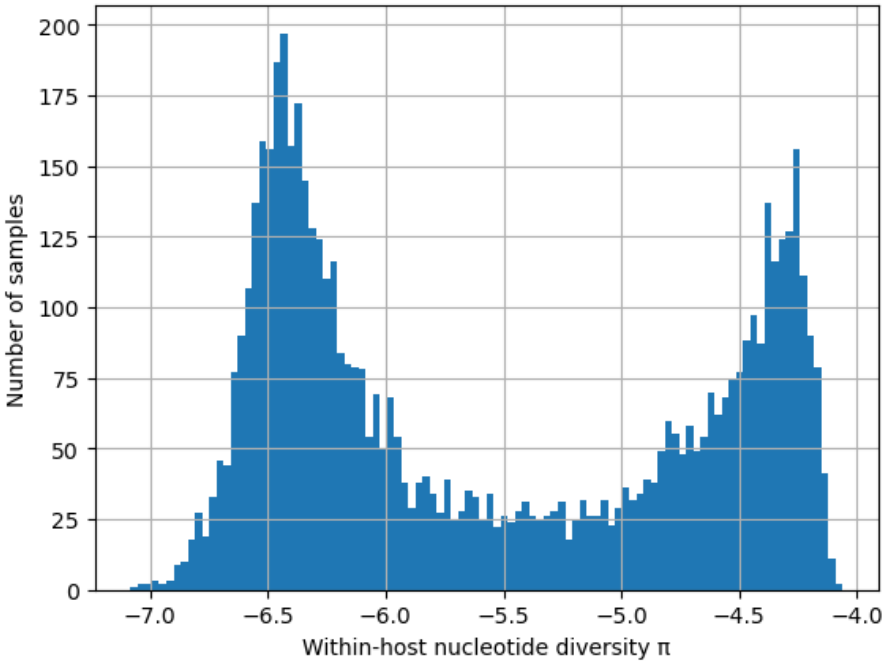

Figure 15: **Distribution of π_W in samples from around world.** Levels of within-host nucleotide diversity obtained from a preliminary analysis of 5970 samples from 30 countries in the MalariaGEN Pf6 dataset [18]. The methods are described in Methods section 5.2. The results show a striking bimodal distribution for π_W which is discussed in the text.

Understanding the relationship between H_W and H_S

When a very large number of SNP loci are analysed by deep genome sequencing of parasite samples from malaria-infected individuals, there is a striking linear correlation between H_W (the heterozygosity of a locus within an individual host) and H_S (the heterozygosity of that locus in the local subpopulation) [32, 36]. This relationship is not apparent if we examine a small number of SNPs in isolation, but it becomes highly statistically significant if we aggregate data on hundreds of thousands of SNPs.

Figure 16 is taken from the study where this phenomenon was first described [32]. SNPs are sorted into bins corresponding to different levels of H_S and this is plotted against the mean value of H_W observed for that set of SNPs in an individual host. The figure shows a series of lines of varying slope, each of which represents the linear relationship between H_W and H_S for an individual host.

The slope of this linear relationship varies between infected individuals but there is a pattern to this variation. At low levels of malaria transmission intensity, H_W tends to be very low and the slope of H_W/H_S is close to zero. At high levels of transmission intensity, there is a much wider range of H_W values and the slope of H_W/H_S varies considerably between infected individuals. If $\widehat{H}_W$ denotes the mean of H_W in the local subpopulation, we find that the slope of $\widehat{H}_W/H_S$ tends to increase with the malaria transmission intensity of the location.

This raises the question of why there is a linear relationship between $\widehat{H}_W$ and H_S, and what determines the slope of this relationship. Here we approach this question by imagining that we are following a transmission chain forward in time as it crosses with other transmission chains. As we shall see, this leads to insights into how measurements of within-host heterozygosity can be used to estimate χ.

An isolated episode of superinfection. Imagine an episode of superinfection in which host C acquires infection from host A and host B. Let the Q alleles acquired from host A have heterozygosity H'_A, and the Q alleles acquired from host B have heterozygosity H'_B. Note that H'_A is not

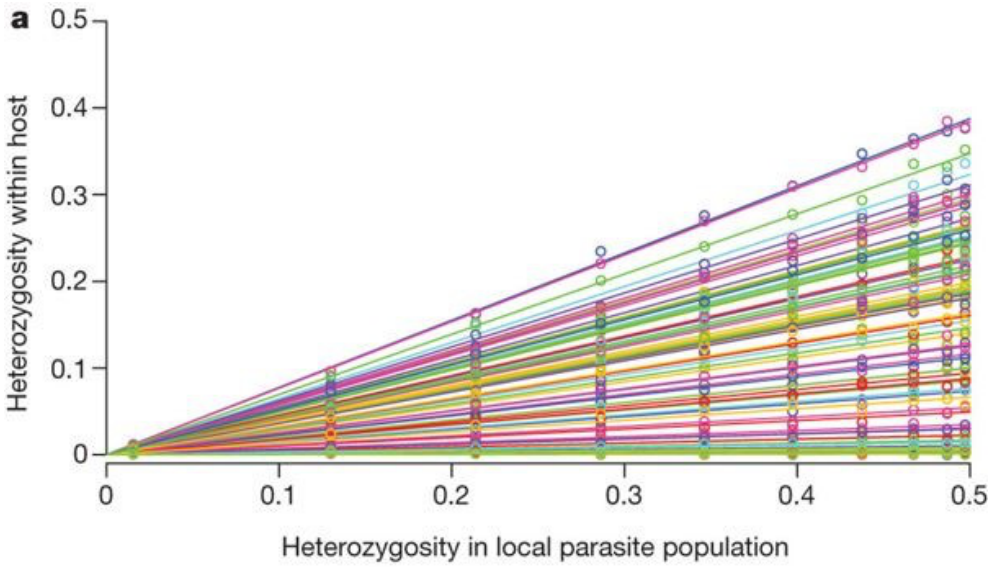

Figure 16: **Empirical relationship between parasite heterozygosity within individual hosts and within the local parasite subpopulation.** Based on genome sequencing of blood samples from patients with malaria. Data on 86,000 SNPs were aggregated by placing SNPs into frequency bins based on H_S (heterozygosity in the local parasite population) and then plotting the mean value of H_W (heterozygosity within an individual host). H_W shows a strong linear correlation with H_S for each sample, but the slope of this line varies greatly between samples. From reference [32]

exactly the same as the heterozygosity of host A as it allows for genetic drift and mutation that have occurred in the process of transmission from host A to host C (figure 17).

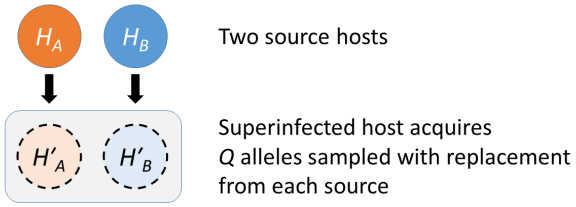

Figure 17: **An episode of superinfection.** In our idealised transmission graph, a superinfected host acquires Q alleles from each of two source hosts, who have heterozygosity values H_A and H_B. The acquired alleles are sampled with replacement from the source hosts and they are also subject to mutation. If we know H_A and H_B then we can obtain H'_A and H'_B from equation 6. We obtain the heterozygosity of the superinfected host by sampling without replacement from the pool of $2Q$ acquired alleles.

To obtain the heterozygosity of host C we must sample two alleles without replacement from the pool of $2Q$ acquired alleles, which themselves were sampled with replacement from host A and host B. As described in Methods section 5.3 the heterozygosity of the superinfected host is given by

$$H_C = \frac{(Q-1)(H'_A + H'_B)}{2(2Q-1)} + \frac{QH_S}{2Q-1} \tag{10}$$

Thus superinfection typically causes a considerable increase in the within-host heterozygosity of a transmission chain because H_S is generally much greater than H'_A or H'_B.

Recurrent episodes of superinfection along a transmission chain. Now imagine that we are following a transmission chain that crosses with other transmission chains with a probability of χ

per generation. Let **X** be a random variable representing the number of generations that separate two crossing events on this transmission chain:

$$Pr\{\mathbf{X} = i\} = \chi(1-\chi)^{i-1} \tag{11}$$

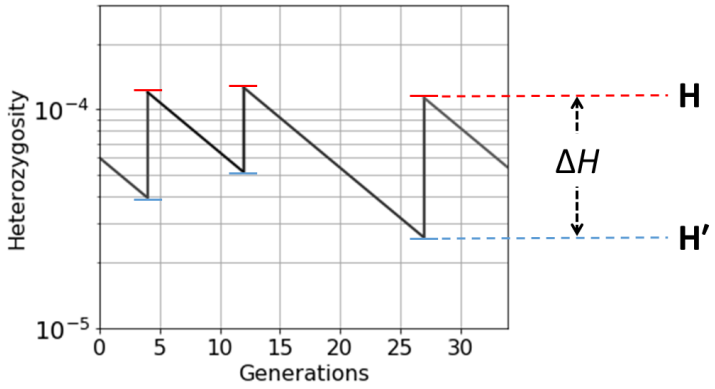

Figure 18: **A transmission chain that crosses at random intervals with other transmission chains.** There are large temporal fluctuations in within-host heterozygosity. Heterozygosity is boosted by each crossing event and then declines gradually due to genetic drift until it is boosted by the next crossing event. **H** and **H**′ are random variables representing the peaks and troughs, respectively, of within-host heterozygosity along our transmission chain. ΔH is the increase in within-host heterozygosity that occurs as a result of a crossing event.

Each crossing event causes within-host heterozygosity to rise abruptly to a peak, and then genetic drift causes it to decline gradually to a trough before it is boosted by another crossing event, as illustrated in figure 18. These peaks and troughs will vary in magnitude according to the number of generations that separate crossing events. Let **H** and **H**′ be random variables representing the peaks and troughs, respectively, of within-host heterozygosity along our transmission chain at some arbitrary locus. We can think of **H**′ and **H** as the states of our transmission chain immediately before and after crossing has occurred in a superinfected host, analogous to H'_A and H_C in equation 10.

Let ΔH be the increase in within-host heterozygosity that occurs as a result of a crossing event. If we assume that all transmission chains have the same probability distributions for **H** and **H**′ as our transmission chain, by applying equation 10 we obtain the expectation of ΔH:

$$E\{\mathbf{H} - \mathbf{H}'\} = \frac{Q}{2Q-1} E\{H_S - \mathbf{H}'\} \tag{12}$$

For the system to be in equilibrium, the expectation of ΔH must equal the expected decrease in heterozygosity that occurs due to genetic drift in the interval between two crossing events, which we can obtain from equations 6 and 11.

Let $\widehat{H}_W$ be the mean value of within-host heterozygosity across our transmission chain. We can evaluate $\widehat{H}_W$ by utilising equations 6, 11 and 12 and making some approximations, as described in Methods section 5.4, to obtain this linear relationship between $\widehat{H}_W$ and H_S:

$$\widehat{H}_W \approx \kappa H_S + \lambda \tag{13}$$

where

$$\kappa = \sum_{i=1}^{\infty} \frac{Q\chi(1-\chi)^{i-1}}{2Q-(Q-1)\alpha^i - 1} \times \sum_{j=0}^{\infty} \chi(1-\chi)^j \alpha^j$$

and

$$\lambda = 2u \sum_{j=0}^{\infty} \sum_{k=0}^{j-1} \chi(1-\chi)^j \alpha^k$$

Since our transmission chain is representative of all transmission chains, $\widehat{H}_W$ is the mean value of within-host heterozygosity for the population as a whole.

This provides a mathematical rationale for the empirically observed relationship between $\widehat{H}_W$ and H_S. The slope of this linear relationship κ is determined by χ and Q, and ranges between 0 and 1. The intercept λ is a very small value that represents the accumulation of mutations along a transmission chain in the interval between time of sampling and the most recent crossing event.

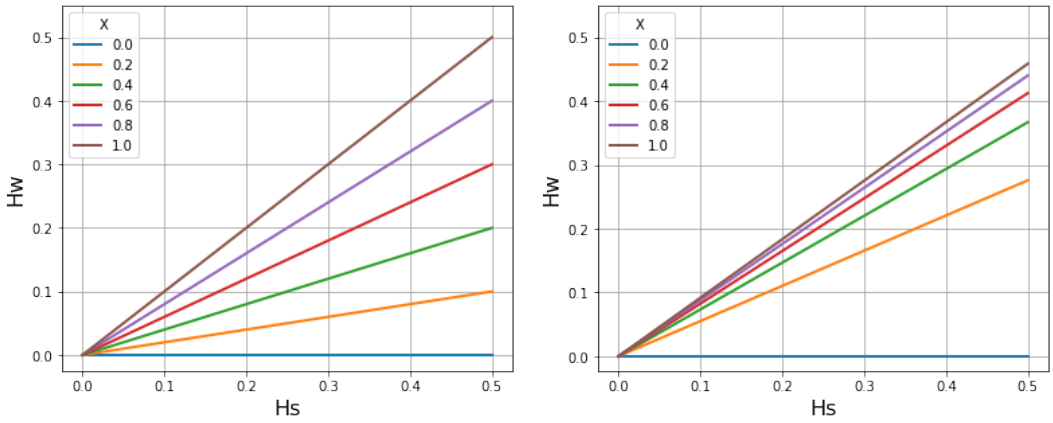

Figure 19: **Theoretical relationship between $\widehat{H}_W$ and H_S based on equation 13.** $\widehat{H}_W$ is the mean of within-host heterozygosity for the local population. Showing results for $Q = 1$ (left panel) and $Q = 10$ (right panel). Different lines represent different values of χ ranging from 0 to 1. $\widehat{H}_W$ shows a strong linear correlation with H_S and the slope depends on χ and Q. View code

Figure 19 illustrates the linear relationship between $\widehat{H}_W$ and H_S based on equation 13, showing how the slope of the line depends on the combination of χ and Q, being zero if there is no superinfection, i.e. if $\chi = 0$.

Using F_{WS} to estimate the rate of superinfection χ. We can measure the slope of $\widehat{H}_W$ versus H_S by deep genome sequencing of parasites in a sample of infected hosts drawn from the local population, as illustrated in figure 16. Equation 13 provides a way to use these empirical measurements to estimate χ, particularly if we are able to estimate Q independently using equation 8.

We previously discussed the use of Wright's fixation indices to describe hierarchical population structure and we can extend this concept to within host-diversity if we let $F_{WS} = 1 - \widehat{H}_W / H_S$. F_{WS} is analogous to an inbreeding coefficient that measures deviation from random mating. For parasite populations, the primary cause of non-random mating is compartmentalisation of the population into discrete transmission chains that do not cross, although there might be other contributory factors such as gametocyte mating bias. In the special case of $\chi = 1$ and $Q = 1$, the genomic transmission graph has properties similar to a randomly mating diploid population, with $\widehat{H}_W / H_S = 1$ and $F_{WS} = 0$, i.e. this is analogous to Hardy-Weinberg equilibrium.

It is arbitrary whether we use $\widehat{H}_W / H_S$ or $F_{WS} = 1 - \widehat{H}_W / H_S$ to summarise measurements of within-host variation by deep sequencing, but F_{WS} is now commonly used in the literature and we shall follow that practice here. In general F_{WS} is inversely related to transmission intensity [18, 32, 36] and is broadly correlated with complexity of infection, i.e. the number of distinct parasite haplotypes detected within a sample [36]. However there is the possibility that F_{WS} could be confounded by population structure as discussed in Methods section 5.6.

A key question is whether equation 13 gives the same result as Markov chain simulation in describing the relationship of F_{WS} to χ and Q. Figure 20 compares the two methods. This confirms that they give essentially the same results when the effective number of hosts is large, but the results deviate when the effective number of hosts is small. This is to be expected as the simplifying assumptions used to derive equation 13 depend on the number of transmission chains being relatively large.

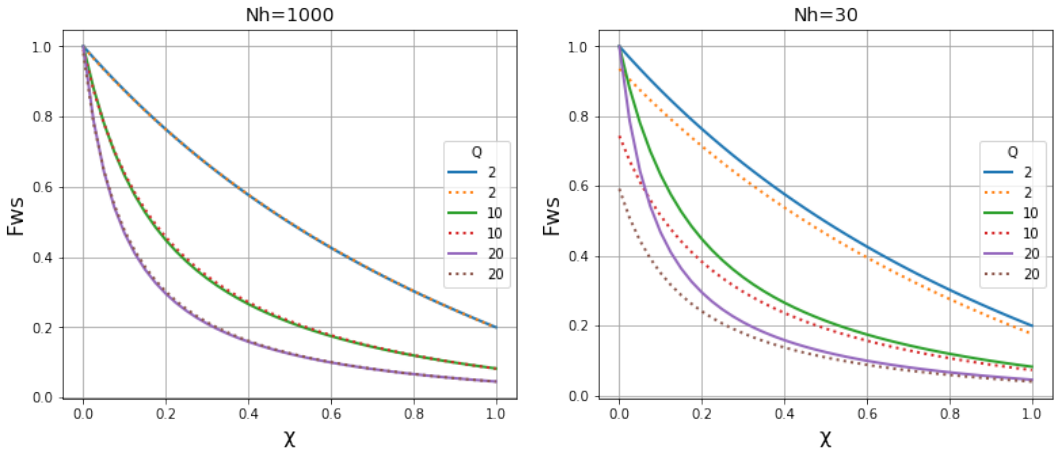

Figure 20: **The inbreeding coefficient F_{WS} is inversely related to χ.** Colours represent different values of Q. Solid lines show the results obtained from equation 13 and dotted lines show the results obtained by Markov chain simulation of coalescence times. When $N_h = 1000$ (left panel) the two methods give very similar results. When $N_h = 30$ (right panel) equation 13 tends to overestimate F_{WS} at low values of χ as compared with the results obtained by Markov chain simulation. Methods section 5.5 shows results for other values of N_h. View code

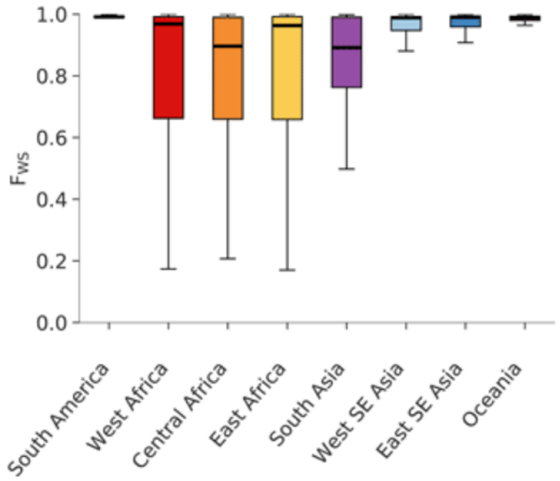

Figure 21: **Estimates of F_{WS} from deep genome sequencing of *P. falciparum* samples**. This figure is taken from the MalariaGEN Pf6 dataset (reference [18] fig. 2) which analysed 5,970 samples from locations around the world. Thick lines represent median values, boxes show the interquartile range, and whiskers represent the bulk of the distribution, discounting outliers.

Thus empirical measurements of F_{WS} allow us to estimate χ, particularly if we are also able to estimate Q using equation 8. Figure 21 shows typical measurements of F_{WS} in different malaria-endemic regions of the world. In South America, where levels of malaria transmission are relatively low, $F_{WS} > 0.98$ in the majority of samples, and from figure 20 this implies that $\chi < 0.02$.

In contrast, in Central Africa F_{WS} is much more variable between samples, ranging from 0.2 to 1 with a median value of ~ 0.9, i.e. the distribution is very assymetrical. If both Q and N_h are relatively large this implies that $\chi < 0.1$, whereas if Q and N_h are small this implies that $\chi \approx 0.1$.

This is a somewhat remarkable result. It implies that, even in regions of high malaria transmission, where at least half of all samples show evidence of multiclonal infections, only a small minority of samples (≤ 0.1) are actually superinfected. This means that the majority of multiclonal infections are due to cotransmission rather than superinfection. It is consistent with recent findings from single cell genome sequencing studies in Malawi that cotransmission of related parasites is much more common than superinfection, and that complex infections can undergo serial passage through multiple hosts without loss of diversity [9].

Modelling epidemiological scenarios

One of the most important potential applications of the genomic transmission graph is to assist in using genetic data to understand epidemiological changes over time. For example, if there is a sudden rise in the local prevalence of infection, we would like to know whether this is due to a local increase in transmission intensity, or to an influx of infections due to migration, or to other factors. In previous sections, we have glossed over the issue of temporal variation by assuming that the transmission parameters are constant over time.

In this section we incorporate temporal variation in N_h, Q, χ and N_m into our Markov chain simulations of the genomic transmission graph. In order to assess changes in the genetic state of the population over time, we need to sample the population at different points of time - we call these *observation times*. For each observation time we must launch a separate Markov chain simulation of the coalescent process going backwards in time.

The `coalestr` module. Here we use `coalestr`, a Python package with accompanying Jupyter notebooks, for running coalescent simulations and computing genetic variation based on the genomic transmission graph. This allows the user to specify a hierarchical population structure and for the transmission parameters to vary over time. It returns time series data for multiple observation times. Tutorials, worked examples and help on how to install and use `coalestr` are available at `d-kwiat.github.io/gtg`.

Effects of a step change in transmission parameters of a local subpopulation within the global metapopulation. In these examples we examine a small local subpopulation with $N_h = 10$, $Q = 5$ and $\chi = 0$ with an ongoing level of migration ($N_m = 1$) from a global metapopulation with $N_h = 14660$, $Q = 3$ and $\chi = 0.2$ as illustrated in figure 9. In each case we examine the effects of a step change in the transmission parameters at 100 to 50 generations before the present. We are looking at the nucleotide diversity of the subpopulation π_S, mean within-host nucleotide diversity π_W, haplotype homozygosity of the subpopulation at a 2cM locus γ_S, mean within-host haplotype homozygosity at a 2cM locus γ_W, the fixation index F_{ST} and the inbreeding index F_{WS}.

χ	N_h	N_m	π_S	π_W	γ_S	γ_W	F_{ST}	F_{WS}
↑	–	–	↑	↑↑	↓	↓↓	↓	↓↓
–	↑	–	(↓)	↓	(↑)	↑	↑	↑
–	–	↑	↑	↑	↓	↓	↓	↓

Table 4: **Effects of a step change in transmission parameters.** We examine three scenarios in a local subpopulation that is embedded within the global metapopulation: (i) a sharp transient increase in χ as in fig. 22; (ii) a sharp transient increase in N_h as in fig. 23; (iii) a sharp transient increase in N_m as in fig. 24. This table summarises the effect of these step changes on nucleotide diversity (π_S and π_W), haplotype homozygosity at a 2cM locus (γ_S and γ_W), F_{ST} and F_{WS}. View code

We consider three scenarios (table 4). In the first scenario, the level of χ in the subpopulation transiently increases from 0 to 1 during the period 100 to 50 generations before the present (figure 22). The result is a sharp rise in π_W and a sharp fall in γ_W, F_{ST} and F_{WS}. There is also a modest rise in π_S and a modest fall in γ_S.

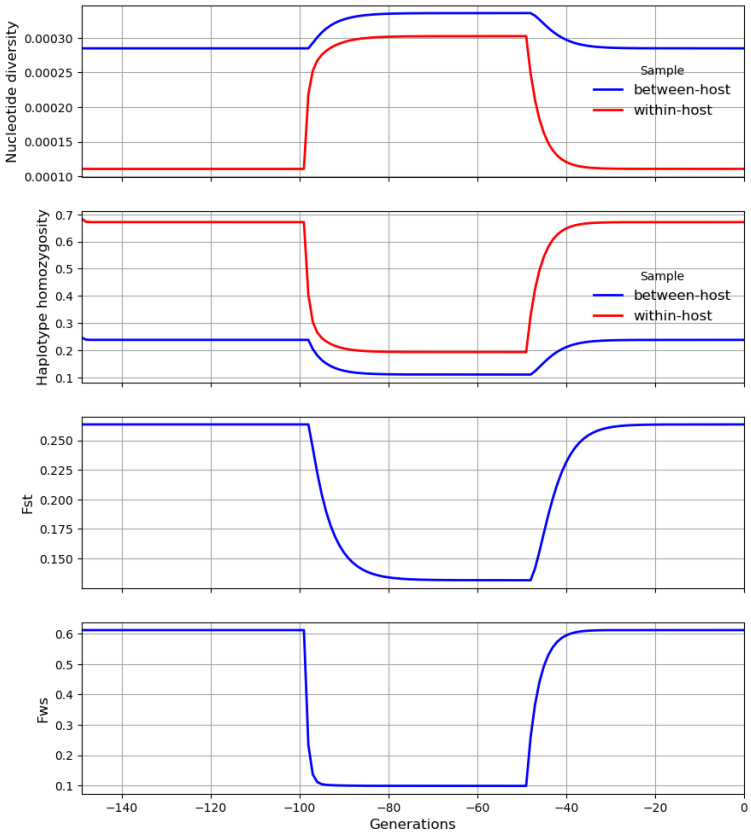

Figure 22: **Increase in the crossing rate of transmission chains.** In our first scenario χ transiently increases from 0 to 1 at 100 to 50 generations before the present. Nucleotide diversity rises, haplotype homozygosity falls View code

In the second scenario, the rate of migration N_m from the metapopulation into the subpopulation transiently increases from 1 to 5 during the period 100 to 50 generations before the present (figure 23). This causes a sharp rise in π_W, a more modest rise in π_S and a sharp fall in γ_W, γ_S, F_{ST} and F_{WS}. Thus the effects of an increase in N_m are rather similar to those of an increase in χ.

In the third scenario, the level of N_h in the subpopulation transiently increases from 10 to 30 during the period 100 to 50 generations before the present (figure 24). The result is a sharp rise in F_{WS}, a modest rise in F_{ST} and γ_W, a modest fall in π_W, and small reduction in π_S. These results appear paradoxical because we might expect an increase in N_h to cause π_S to rise whereas it falls slightly. The paradox can be explained by recalling that $N_m = mN_h$. Although N_m is constant, the rise in N_h causes m to decline, and this reduction in the proportion of hosts that have migrated from the metapopulation counterbalances the local increase in effective number of hosts, causing π_S to remain almost unchanged.

Discussion

Although the biology of parasite transmission dynamics is extremely complex, it is possible to summarise many of the fundamental processes in an idealised model with just three essential parameters: the quantum of transmission, the effective number of hosts and the crossing rate of transmission chains. We have shown how this model, which we call the genomic transmission

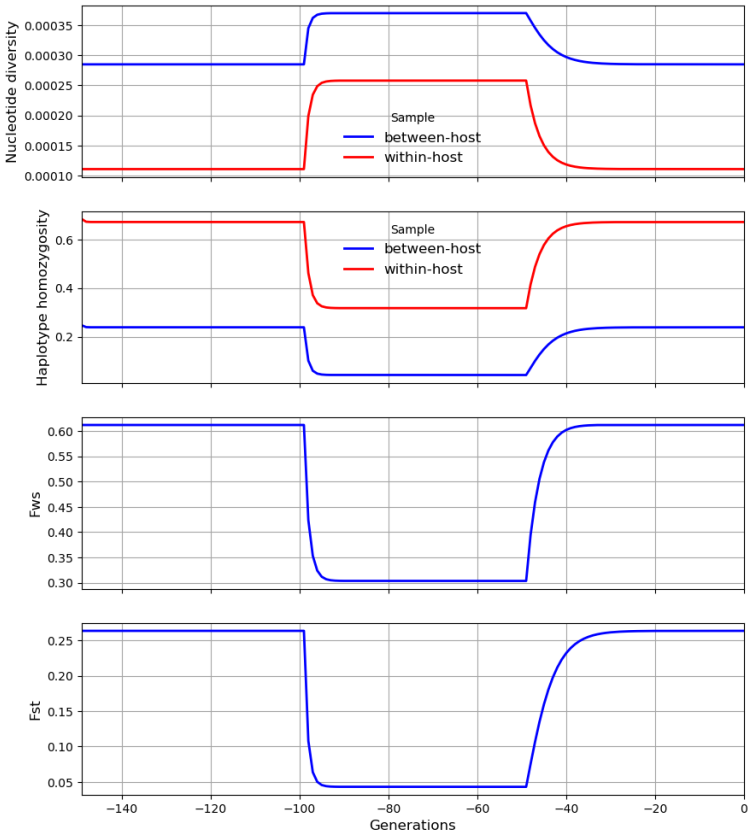

Figure 23: **Increase in the rate of migration from the global metapopulation.** In this scenario N_m transiently increases from 1 to 5 at 100 to 50 generations before the present. Nucleotide diversity rises, haplotype homozygosity falls View code

graph, lends itself to coalescent modelling and rapid simulation of different population genetic scenarios. It also provides a mathematical framework for analysing within-host variation, and we show how this allows key parameters to be inferred from deep sequencing data.

In this concluding section we discuss the practical relevance of these findings in the specific context of malaria biology and disease control. Finally we consider the broader applications of the genomic transmission graph for recombining populations in general.

The quantum of transmission Q. Malaria parasites make a complex and arduous journey to get from host to host via a mosquito vector. Only a tiny fraction of the billions of parasites carried by a human host finds its way into a mosquito vector, and fewer than 1% of the sporozoites that develop within an infected mosquito find their way into the next human host on the transmission chain. A wide range of transmission bottlenecks lie along this pathway including developmental roadblocks, specialised invasion mechanisms, host and vector immunity, and mosquito physiology and biting behaviour, as reviewed in reference [13].

In our model, the quantum of transmission Q represents the number of alleles that pass through this series of transmission bottlenecks in each cycle of host-to-host transmission. One way of estimating Q would be to count the number of sporozoites that are inoculated by a mosquito into a new host, which in various experiments has been estimated to have a median value of 8 to 39 [13] but this is technically challenging to quantify directly, and it does not take account of other bottlenecks, e.g. the number of gametocytes taken up by the mosquito from the previous host.

The value of Q is crucial for understanding parasite cotransmission. If $Q = 1$ this means that only one allele is transmitted from one host to the next, hence we would expect to see a predominance of clonal infections. Multiclonal infections arise due to superinfection and their

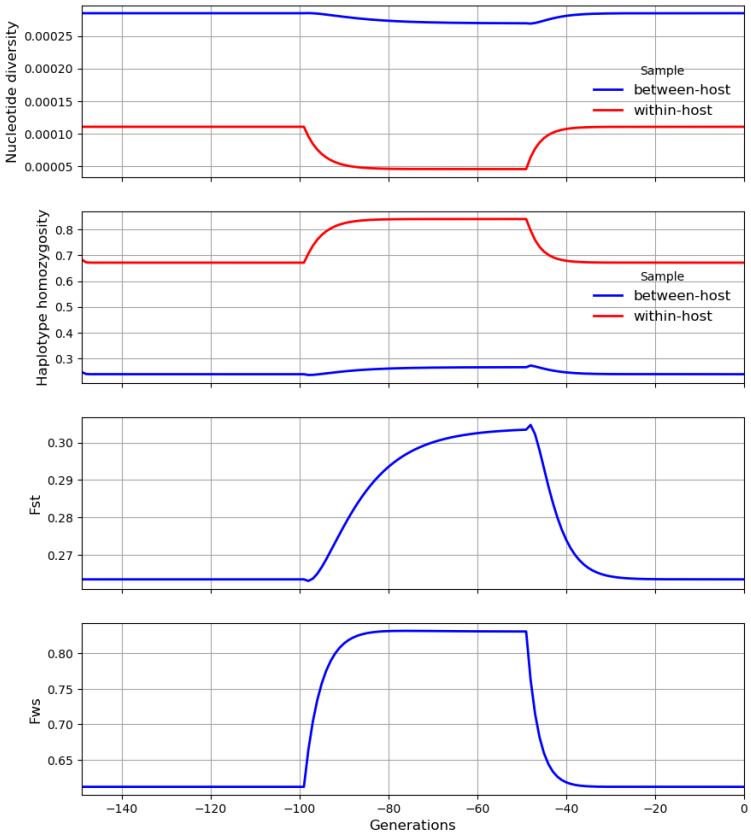

Figure 24: **Increase in the effective number of hosts.** In this scenario N_h transiently increases from 10 to 30 at 100 to 50 generations before the present. Paradoxically, nucleotide diversity falls, haplotype homozygosity rises. View code

recombinant progeny are propagated along the transmission chain. As Q increases it becomes increasingly likely that multiclonal infections will be cotransmitted from one host to the next. If Q is large, there could be many successive generations of cotransmission and recombination of multiclonal infections following a single episode of superinfection.

Here we describe a way of estimating Q from measurements of within-host nucleotide diversity π_W by deep genome sequencing. Analysis of thousands of *P. falciparum* samples from around the world reveals that the empirical distribution of π_W is strikingly bimodal, and we postulate that one of the peaks ($\pi_W \approx 4 \times 10^{-7}$) comprises transmission chains that have not experienced superinfection in the recent past. From equation 8 this implies that $Q \approx 18$, which is reassuringly close to experimental estimates of the median number of sporozoites inoculated by an infectious mosquito, but this is a very preliminary estimate that requires more detailed analysis as described in Methods section 5.2.

The crossing rate of transmission chains χ. In regions of high transmission, malaria infections are often multiclonal, i.e. they contain multiple genetically distinct forms of the parasite. Various methods have been developed to assess the number of distinct genetic forms of the parasite in a malaria-infected individual, known as the complexity of infection (COI) [37–40]. In the past it was widely assumed that the COI was a measure of how often an individual had been superinfected, but there is growing evidence - from single-cell sequencing [9], deep sequencing [41] and epidemiological modelling [11] - that multiclonal infections are often the result of cotransmission rather than superinfection.

In our model, the crossing rate of transmission chains χ represents the probability that a host is superinfected from two sources in one generation of the transmission graph. We do not explicitly model COI but it is implicit in our model that a relatively modest value of χ could lead to a high

value of COI as long as the value of Q is sufficiently high to allow cotransmission of multiclonal infections across multiple successive generations of the transmission graph.

From an epidemiological perspective, χ serves as a proxy measure of the incidence of infection. More specifically, equation 1 states that the incidence of infection is approximately equal to χ/τ, where τ is the serial interval of infection, although this involves many simplifying assumptions.

From a genetic perspective, χ determines the frequency of outcrossing in the population. If $\chi = 0$ there is no outcrossing and the parasite population may appear to be clonal in nature, even though sexual recombination occurs with each generation of transmission between closely related sibling parasites. If $\chi = 1$ there is a high rate of outcrossing, and in the special case of $\chi = 1$ and $Q = 1$ the genomic transmission graph is similar in many ways (but not identical) to a randomly mating diploid population.

Here we describe a method of estimating χ from measurements of F_{WS} by deep genome sequencing. F_{WS} is a metric that summarises the remarkably linear relationship that is observed between within-host heterozygosity ($\widehat{H}_W$) and the heterozygosity of the local subpopulation (H_S) when data are aggregated across hundreds of thousands of SNPs [32]. In equation 13 we derive a formula that describes the slope of this linear relationship, and thus the value of F_{WS}, as a function of χ and Q. Analysis of F_{WS} in thousands of samples indicates that $\chi \leq 0.1$ even in regions of high transmission where at least half of infections are multiclonal. This supports the view that superinfection is much less common than cotransmission.

The effective number of hosts N_h. A key question in malaria epidemiology is the nature and size of the human infectious reservoir. Only a small fraction of the hundreds of millions of people who are infected with *P. falciparum* malaria annually [8] go on to transmit parasites to a new host, and identifying those who are most likely to do so requires highly specialised and laborious epidemiological methods [42].

It would be a major advance if parasite genetic data could be used to estimate the size of the infectious reservoir, and to determine how this varies over space and time following malaria control interventions. In our model, N_h represents the number of individuals that effectively transmit parasites to the next generation, and this might serve as a proxy measure for the human infectious reservoir.

Effective population size (N_e) is a fundamental parameter of population genetics. In theory it is the number of individuals that effectively contribute progeny to the next generation. In practice it is the estimated number of individuals required for an idealised population to reproduce genetic features observed in the real population. N_e is usually much smaller than the census population size, e.g. the ancestral effective population size of humans is on the order of magnitude of 10,000 individuals [43].

Previous studies have estimated N_e for *P. falciparum* using a range of different approaches [20, 44–46]. The results are perplexing as they vary over several orders of magnitude depending on the method used, ranging from 10^2 to 10^6. This problem is elegantly reviewed in reference [46]. Some variation in N_e is to be expected, depending on whether the population sampled is local or global, whether the methodology is designed to assess short-term or long-term N_e, and whether we are looking at rates of genetic drift or of adaptive evolution. However the extreme variability observed in parasite N_e implies that there is some fundamental problem in the application of classical population genetic methods to malaria [14, 46, 47]. In our model we do not specify parasite N_e but we know the size of the parasite population bottleneck which is given by $N_h Q$.

There are a variety of ways by which N_h might be estimated from empirical data (as is the case for N_e) giving different perspectives on the effective population size, e.g. short-term versus long-term and local versus global. Here we illustrate the basic principles of a coalescent method of estimating long-term N_h from the levels of nucleotide diversity observed in the global parasite population ($\pi_T \approx 4 \times 10^{-4}$). Table 2 shows various combinations of transmission parameters that would give this value of π_T, with N_h ranging from 3,269 to 18,764 depending on the values of Q and χ.

Rate of migration N_m. Understanding patterns of migration of infected individuals is of basic importance in designing effective strategies for malaria elimination. One way of thinking about this is in terms of a hierarchical population structure in which local subpopulations of parasites are interconnected parts of a global metapopulation. In our model, the rate of migration N_m represents the number of hosts that migrate each generation into a local subpopulation from the global (or regional) metapopulation. Migration is also extremely important from a genetic perspective, as surprisingly low rates of N_m can cause a small local subpopulation to acquire very nearly the same level of genetic diversity as the global metapopulation. Here we present a simple model of migration across a hierarchical population structure that illustrates this point.

The effective reproduction number R. This key parameter of infectious disease transmission dynamics was conceived by Ronald Ross in his pioneering work on mathematical models of malaria over a hundred years ago [1,2]. In our model we do not specify R but it is straightforward to estimate a reproduction number for N_h. Caution is needed in equating this with conventional epidemiological estimates of R because N_h represents the effective number of hosts that transmit parasites to the next generation and is probably much less than the total number of infected individuals. As discussed above, we can view N_h as a proxy for the human infectious reservoir, which may have different dynamical properties from the total number of infected individuals.

The effective recombination parameter ϕ_t. In an insightful review, Camponovo and colleagues envisage how new statistical methods [5,6] will in future allow malaria transmission dynamics to be inferred from genome-wide ancestral recombination graphs that are much more informative on an epidemiological timescale than current mutation-based methods [48]. However they point out that this depends on the effective recombination rate which is affected by superinfection, cotransmission and population structure.

Here we provide a method of estimating the rate of effective recombination, which we define as recombination between heterozygous alleles that acts to change the DNA sequence of a haplotype locus. By focusing on a haplotype locus, we allow for the possibility that recombination may be effective in some regions of the genome and not in others, particularly in the case of mating between genetically distinct but closely related individuals.

In our model, the effective recombination parameter ϕ_t represents the probability that, if recombination occurs at a haplotype locus at time t, this will result in a change to the DNA sequence of that locus. Thus the effective recombination rate is given by $\phi_t r L$ where r is the locus-scaled recombination rate and L is the length of the haplotype locus.

The crucial insight is that ϕ_t is determined by the mean level of within-host heterozygosity in the population, and that this may vary over time. Mating occurs within the vector but we make the simplifying assumption that within-host heterozygosity determines the probability that two mating alleles are genetically distinct, i.e. that they have different DNA sequences at a particular haplotype locus. In equation 5, we let $\phi_t = f\widehat{H}_W$ where f is a correction factor to allow for the possibility of mating bias and other confounders of the relationship between ϕ_t and $\widehat{H}_W$.

Identity by descent and recent common ancestry. There is considerable interest in the use of IBD metrics to evaluate genetic relatedness between malaria parasites and to establish patterns of connectivity and recent migration betwen different geographical locations of malaria endemicity [24–29]. Conventional models of IBD count the number of meioses that separate two individuals, and estimate how segments of IBD are broken down by meiotic recombination assuming panmyxia, i.e. that mating occurs randomly across the population [23]. However malaria parasite populations are far from panmyctic because mating is rigidly compartmentalised into discrete within-host populations.

By modelling the within-host population structure that arises from the parasite life cycle, the genomic transmission graph allows a more accurate view of the effective recombination rate. In our model, shared haplotype segments of > 2 centimorgans are essentially equivalent to segments of IBD, and the proportion of the genome that is IBD between two parasites can be crudely approximated by γ, the mean haplotype homozygosity of 2 centimorgan locus (figure 12). This

provides a starting point for constructing a model of IBD that better reflects the parasite life cycle and thus allows more accurate inference of genetic relatedness.

Using the transmission graph to infer epidemiological processes from genetic data. The work of Anderson and colleagues on the evolution of antimalarial drug resistance in Southeast Asia provides an elegant example of how longitudinally sampled genetic data, accompanied by rich epidemiological data, can be used in time series analysis to infer the parameters of a population genetic model [46]. At the same time, their work highlights the limitations of current population genetic models as applied to malaria parasites, e.g. different types of genetic measurement give widely different estimates of effective population size.

The amount and the quality of time-series data on parasite genetic variation linked to rich epidemiological data will increase greatly over the next few years as genome sequencing becomes a routine part of malaria surveillance [49]. Epidemiological events - including fluctuations in population size, migration and transmission intensity - all shape the genetic architecture of the parasite population. Our challenge is to infer those epidemiological events from genetic data, by understanding the causal relationship between epidemiological variables and genetic variables (figure 25).

The genomic transmission graph provides a theoretical model of this causal process, giving the mathematical relationship between a set of idealised transmission parameters and population genetic variables. If we have rich epidemiological data coupled to genetic data from the same locations, ideally sampled over space and time, we can look for a correlation between the transmission parameters and the epidemiological variables. We do not expect the idealised transmission parameters to represent exact values of any epidemiological variables, but we hope to find a sensible and reliable set of correlations that allow us to estimate the epidemiological variables from the transmission parameters and hence from the genetic data.

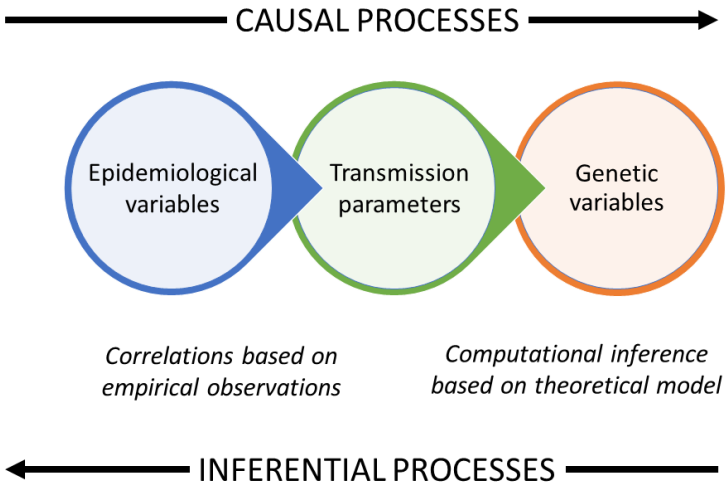

Figure 25: **How epidemiological events can be inferred from genetic data.** First the genetic data are used to make estimates of transmission parameters based on the mathematical relationships specified by the theoretical model of the transmission graph. Then the epidemiological variables are inferred from the transmission parameters using a set of correlations and mappings that have been built up by many empirical observations.

Natural selection. It is conventional to focus on selectively neutral loci when using genetics to infer population dynamics, where the key processes of interest are mutation and genetic drift. Therefore in our current model we ignore natural selection and adaptive evolution although these

play a major role in shaping parasite genetic variation, particularly at drug resistance loci but also more generally across the genome [33, 46, 50].

Natural selection can operate through many different mechanisms on a wide range of time scales. The malaria life-cycle involves repeated cycles of rapid population expansion - of clonally replicating parasites within the host and of sexually reproducing parasites within the vector - punctuated by tight transmission bottlenecks. This acts both to intensify natural selection and to obscure classic population genetic signatures of selection [14, 47].

Of particular interest for epidemiological monitoring is recent positive selection of new forms of drug resistance. With modern methods of genetic epidemiology, it is often possible to identify the causal mutation of drug resistance and to measure its rising frequency in the population, i.e. to observe the selective sweep [46, 50]. In principle it would be possible to extend the current model to describe the selective sweep of a drug resistance locus using a structured coalescent approach, but that is beyond the scope of the current paper.

Limitations of the model. The genomic transmission graph is a deliberately simplistic model that sets aside many details in order to gain a clear view of fundamental mechanisms. It seeks to elucidate how a small set of key transmission parameters - Q, χ, N_h and N_m - act individually and in combination to shape parasite genetic diversity and population structure. Its utility as a tool for learning about fundamental processes underlying genetic variation is greatly enhanced by its amenability to coalescent simulation which stems from the simple structure of the model.

This approach has obvious limitations, some of which would be relatively straightforward to address in modified versions of the model, e.g. we assume that parasites are transmitted from host to host in non-overlapping generations of fixed duration, which is clearly over-simplistic and might be improved by using a continuous-time Moran model [12]. As another example, the current model allows superinfection from only two sources, but in principle we could allow any number of sources of superinfection by making χ a random variable.

Other limitations are more complex to address, e.g. we take no account of acquired immunity, antimalarial drug usage, vector biting behaviour and many other sources of heterogeneity in the host, parasite and vector populations. In principle this could be addressed by incorporating the basic principles of the genomic transmission graph into agent-based epidemiological models that explictly deal with the details of malaria transmission biology, at the cost of introducing a large number of parameters that might be difficult to ascertain with confidence (and without overfitting the model) from available empirical data [10, 11, 51, 52].

Broader applications of the genomic transmission graph for recombining populations. Although this paper focuses on malaria, the transmission graph has wider implications for other parasites and for recombining populations in general. In the case of malaria, each node of the graph represents the parasite subpopulation carried by an individual host, but more generally we can think of this as a transient subpopulation, i.e. a discrete group of individuals that exists at a certain point in time. Framed in this general manner, each node of the graph represents a discrete group of individuals that undergo recombination before propagating along the edges of the graph to form one or more new groups. The transmission graph describes the effective number and size of these groups and the rate at which they form new groups, merge with other groups, migrate between locations, or disappear. With suitable modification, the genomic transmission graph might be useful for studying the evolutionary dynamics of other species that naturally cluster into many small groups that are continually propagating, merging and migrating, such as shoals of fish, flocks of birds or herds of animals. It could also possibly be used for analysis of viral and bacterial species that undergo horizontal gene transfer when they congregate within individual hosts or other transient ecological niches.

References

[1] R Ross. Some a Priori Pathometric Equations. *The British Medical Journal*, 1(2830):546–547, mar 1915.

[2] David L. Smith, Katherine E. Battle, Simon I. Hay, Christopher M. Barker, Thomas W. Scott, and F. Ellis McKenzie. Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. *PLoS Pathogens*, 8(4):e1002588, apr 2012.

[3] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James L N Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, 303(5656):327–32, jan 2004.

[4] Erik M. Volz, Katia Koelle, and Trevor Bedford. Viral Phylodynamics. *PLoS Computational Biology*, 9(3):e1002947, mar 2013.

[5] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329, sep 2019.

[6] Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338, sep 2019.

[7] D. L. Smith, J. Dushoff, R. W. Snow, and S. I. Hay. The entomological inoculation rate and Plasmodium falciparum infection in African children. *Nature*, 438(7067):492–495, nov 2005.

[8] WHO. World malaria report 2022. 2022.

[9] Standwell C. Nkhoma, Simon G. Trevino, Karla M. Gorena, Shalini Nair, Stanley Khoswe, Catherine Jett, Roy Garcia, Benjamin Daniel, Aliou Dia, Dianne J. Terlouw, Stephen A. Ward, Timothy J.C. Anderson, and Ian H. Cheeseman. Co-transmission of Related Malaria Parasite Lineages Shapes Within-Host Parasite Diversity. *Cell Host and Microbe*, 27(1):93–103.e4, jan 2020.

[10] Rachel F. Daniels, Stephen F. Schaffner, Edward a. Wenger, Joshua L. Proctor, Hsiao-Han Chang, Wesley Wong, Nicholas Baro, Daouda Ndiaye, Fatou Ba Fall, Medoune Ndiop, Mady Ba, Danny a. Milner, Terrie E. Taylor, Daniel E. Neafsey, Sarah K. Volkman, Philip a. Eckhoff, Daniel L. Hartl, and Dyann F. Wirth. Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proceedings of the National Academy of Sciences*, page 201505691, 2015.

[11] Oliver J Watson, Lucy C Okell, Joel Hellewell, Hannah C Slater, H Juliette T Unwin, Irene Omedo, Philip Bejon, Robert W Snow, Abdisalan M Noor, Kirk Rockett, Christina Hubbart, Joaniter I Nankabirwa, Bryan Greenhouse, Hsiao-Han Chang, Azra C Ghani, and Robert Verity. Evaluating the performance of malaria genetics for inferring changes in transmission intensity using transmission modelling. *Molecular Biology and Evolution*, sep 2020.

[12] Jason A. Hendry, Dominic Kwiatkowski, and Gil McVean. Elucidating relationships between P. falciparum prevalence and measures of genetic diversity with a combined geneticepidemiological model of malaria. *PLoS Computational Biology*, 17(8), aug 2021.

[13] Wouter Graumans, Ella Jacobs, Teun Bousema, and Photini Sinnis. When Is a Plasmodium-Infected Mosquito an Infectious Mosquito? *Trends in Parasitology*, 36(8):705–716, aug 2020.

[14] Chang H, Moss E, Park D, Ndiaye D, Mboup S, Volkman S, Sabeti P, Wirth D, Neafsey D, and Hartl D. Malaria life cycle intensifies both natural selection and random genetic drift. *PNAS*, 110(50):20129–20134, 2013.

[15] Selina E R Bopp, Micah J Manary, a Taylor Bright, Geoffrey L Johnston, Neekesh V Dharia, Fabio L Luna, Susan McCormack, David Plouffe, Case W McNamara, John R Walker, David a Fidock, Eros Lazzerini Denchi, and Elizabeth a Winzeler. Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS genetics*, 9(2):e1003293, jan 2013.

[16] Antoine Claessens, William L Hamilton, Mihir Kekre, Thomas D Otto, Adnan Faizullabhoy, Julian C Rayner, and Dominic Kwiatkowski. Generation of Antigenic Diversity in Plasmodium falciparum by Structured Rearrangement of Var Genes During Mitosis. *PLoS genetics*, 10(12):e1004812, dec 2014.

[17] William L. Hamilton, Antoine Claessens, Thomas D. Otto, Mihir Kekre, Rick M. Fairhurst, Julian C. Rayner, and Dominic Kwiatkowski. Extreme mutation bias and high AT content in Plasmodium falciparum. *Nucleic Acids Research*, 45(4):gkw1259, dec 2016.

[18] MalariaGEN, Ambroise Ahouidi, Mozam Ali, Jacob Almagro-Garcia, Alfred Amambua-Ngwa, Chanaki Amaratunga, Roberto Amato, Lucas Amenga-Etego, Ben Andagalu, Tim J.C. Anderson, Voahangy Andrianaranjaka, Tobias Apinjoh, Cristina Ariani, Elizabeth A. Ashley, Sarah Auburn, Gordon A. Awandare, Hampate Ba, Vito Baraka, Alyssa E. Barry, Philip Bejon, Gwladys I. Bertin, Maciej F. Boni, Steffen Borrmann, Teun Bousema, Oralee Branch, Peter C. Bull, George B.J. Busby, Thanat Chookajorn, Kesinee Chotivanich, Antoine Claessens, David Conway, Alister Craig, Umberto D'Alessandro, Souleymane Dama, Nicholas PJ Day, Brigitte Denis, Mahamadou Diakite, Abdoulaye Djimdé, Christiane Dolecek, Arjen M. Dondorp, Chris Drakeley, Eleanor Drury, Patrick Duffy, Diego F. Echeverry, Thomas G. Egwang, Berhanu Erko, Rick M. Fairhurst, Abdul Faiz, Caterina A. Fanello, Mark M. Fukuda, Dionicia Gamboa, Anita Ghansah, Lemu Golassa, Sonia Goncalves, William L. Hamilton, G. L.Abby Harrison, Lee Hart, Christa Henrichs, Tran Tinh Hien, Catherine A. Hill, Abraham Hodgson, Christina Hubbart, Mallika Imwong, Deus S. Ishengoma, Scott A. Jackson, Chris G. Jacob, Ben Jeffery, Anna E. Jeffreys, Kimberly J. Johnson, Dushyanth Jyothi, Claire Kamaliddin, Edwin Kamau, Mihir Kekre, Krzysztof Kluczynski, Theerarat Kochakarn, Abibatou Konaté, Dominic P. Kwiatkowski, Myat Phone Kyaw, Pharath Lim, Chanthap Lon, Kovana M. Loua, Oumou Maïga-Ascofaré, Cinzia Malangone, Magnus Manske, Jutta Marfurt, Kevin Marsh, Mayfong Mayxay, Alistair Miles, Olivo Miotto, Victor Mobegi, Olugbenga A. Mokuolu, Jacqui Montgomery, Ivo Mueller, Paul N. Newton, Thuy Nguyen, Thuy Nhien Nguyen, Harald Noedl, François Nosten, Rintis Noviyanti, Alexis Nzila, Lynette I. Ochola-Oyier, Harold Ocholla, Abraham Oduro, Irene Omedo, Marie A. Onyamboko, Jean Bosco Ouedraogo, Kolapo Oyebola, Richard D. Pearson, Norbert Peshu, Aung Pyae Phyo, Chris V. Plowe, Ric N. Price, Sasithon Pukrittayakamee, Milijaona Randrianarivelojosia, Julian C. Rayner, Pascal Ringwald, Kirk A. Rockett, Katherine Rowlands, Lastenia Ruiz, David Saunders, Alex Shayo, Peter Siba, Victoria J. Simpson, Jim Stalker, Xin zhuan Su, Colin Sutherland, Shannon Takala-Harrison, Livingstone Tavul, Vandana Thathy, Antoinette Tshefu, Federica Verra, Joseph Vinetz, Thomas E. Wellems, Jason Wendler, Nicholas J. White, Ian Wright, William Yavo, and Htut Ye. An open dataset of Plasmodium falciparum genome variation in 7,000 worldwide samples. *Wellcome Open Research*, 6:1–31, 2021.

[19] MalariaGEN, Muzamil Mahdi Abdel Hamid, Mohamed Hassan Abdelraheem, Desmond Omane Acheampong, Ambroise Ahouidi, Mozam Ali, Jacob Almagro-Garcia, Alfred Amambua-Ngwa, Chanaki Amaratunga, Lucas Amenga-Etego, Ben Andagalu, Tim Anderson, Voahangy Andrianaranjaka, Ifeyinwa Aniebo, Enoch Aninagyei, Felix Ansah, Patrick O Ansah, Tobias Apinjoh, Paulo Arnaldo, Elizabeth Ashley, Sarah Auburn, Gordon A Awandare, Hampate Ba, Vito Baraka, Alyssa Barry, Philip Bejon, Gwladys I Bertin, Maciej F Boni, Steffen Borrmann, Teun Bousema, Marielle Bouyou-Akotet, Oralee Branch, Peter C Bull, Huch Cheah, Keobouphaphone Chindavongsa, Thanat Chookajorn, Kesinee Chotivanich, Antoine Claessens, David J Conway, Vladimir Corredor, Erin Courtier, Alister Craig, Umberto D'Alessandro, Souleymane Dama, Nicholas Day, Brigitte Denis, Mehul Dhorda, Mahamadou Diakite, Abdoulaye Djimde, Christiane Dolecek, Arjen Dondorp, Seydou Doumbia, Chris Drakeley, Eleanor Drury, Patrick Duffy, Diego F Echeverry, Thomas G Egwang, Sonia Maria Mauricio Enosse, Berhanu Erko, Rick M. Fairhurst, Abdul Faiz, Caterina A Fanello, Mark Fleharty, Matthew Forbes, Mark Fukuda, Dionicia Gamboa, Anita Ghansah, Lemu Golassa, Sonia Goncalves, G L Abby Harrison, Sara Anne Healy, Jason A Hendry, Anastasia Hernandez-Koutoucheva, Tran Tinh Hien, Catherine A Hill, Francis Hombhanje, Amanda Hott, Ye Htut, Mazza Hussein, Mallika Imwong, Deus Ishengoma, Scott A Jackson, Chris G Jacob, Julia Jeans, Kimberly J Johnson, Claire Kamaliddin, Edwin

Kamau, Jon Keatley, Theerarat Kochakarn, Drissa S Konate, Abibatou Konaté, Aminatou Kone, Dominic P Kwiatkowski, Myat P Kyaw, Dennis Kyle, Mara Lawniczak, Samuel K Lee, Martha Lemnge, Pharath Lim, Chanthap Lon, Kovana M Loua, Celine I Mandara, Jutta Marfurt, Kevin Marsh, Richard James Maude, Mayfong Mayxay, Oumou Maïga-Ascofaré, Olivo Miotto, Toshihiro Mita, Victor Mobegi, Abdelrahim Osman Mohamed, Olugbenga A Mokuolu, Jaqui Montgomery, Collins Misita Morang'a, Ivo Mueller, Kathryn Murie, Paul N Newton, Thang Ngo Duc, Thuy Nguyen, Thuy-Nhien Nguyen, Tuyen Nguyen Thi Kim, Hong Nguyen Van, Harald Noedl, Francois Nosten, Rintis Noviyanti, Vincent Ntui-Njock Ntui, Alexis Nzila, Lynette Isabella Ochola-Oyier, Harold Ocholla, Abraham Oduro, Irene Omedo, Marie A Onyamboko, Jean-Bosco Ouedraogo, Kolapo Oyebola, Wellington Aghoghovwia Oyibo, Richard Pearson, Norbert Peshu, Aung P Phyo, Christopher V Plowe, Ric N Price, Sasithon Pukrittayakamee, Huynh Hong Quang, Milijaona Randrianarivelojosia, Julian C Rayner, Pascal Ringwald, Anna Rosanas-Urgell, Eduard Rovira-Vallbona, Valentin Ruano-Rubio, Lastenia Ruiz, David Saunders, Alex Shayo, Peter Siba, Victoria J Simpson, Mahamadou S. Sissoko, Christen Smith, Xin-zhuan Su, Colin Sutherland, Shannon Takala-Harrison, Arthur Talman, Livingstone Tavul, Ngo Viet Thanh, Vandana Thathy, Aung Myint Thu, Mahamoudou Toure, Antoinette Tshefu, Federica Verra, Joseph Vinetz, Thomas E Wellems, Jason Wendler, Nicholas J White, Georgia Whitton, William Yavo, and Rob W van der Pluijm. Pf7: an open dataset of Plasmodium falciparum genome variation in 20,000 worldwide samples. *Wellcome Open Research*, 8:22, jan 2023.

[20] Deirdre A. Joy, Xiaorong Feng, Jianbing Mu, Tetsuya Furuya, Kesinee Chotivanich, Antoniana U. Krettli, May Ho, Alex Wang, Nicholas J. White, Edward Suh, Peter Beerli, and Xin zhuan Su. Early origin and recent expansion of Plasmodium falciparum. *Science (New York, N.Y.)*, 300(5617):318–321, apr 2003.

[21] Kazuyuki Tanabe, Toshihiro Mita, Thibaut Jombart, Anders Eriksson, Shun Horibe, Nirianne Palacpac, Lisa Ranford-Cartwright, Hiromi Sawai, Naoko Sakihama, Hiroshi Ohmae, Masatoshi Nakamura, Marcelo U. Ferreira, Ananias A. Escalante, Franck Prugnolle, Anders Björkman, Anna Färnert, Akira Kaneko, Toshihiro Horii, Andrea Manica, Hirohisa Kishino, and Francois Balloux. Plasmodium falciparum Accompanied the Human Expansion out of Africa. *Current Biology*, 20(14):1283–1289, 2010.

[22] Alistair Miles, Zamin Iqbal, Paul Vauterin, Richard Pearson, Susana Campino, Michel Theron, Kelda Gould, Daniel Mead, Eleanor Drury, John O'Brien, Valentin Ruano Rubio, Bronwyn MacInnis, Jonathan Mwangi, Upeka Samarakoon, Lisa Ranford-Cartwright, Michael Ferdig, Karen Hayton, Xin-zhuan Su, Thomas Wellems, Julian Rayner, Gil McVean, and Dominic Kwiatkowski. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research*, 26(9):1288–1299, sep 2016.

[23] Sharon R Browning and Brian L Browning. Identity by descent between distant relatives: detection and applications. *Annual review of genetics*, 46:617–33, jan 2012.

[24] Stephen F. Schaffner, Aimee R. Taylor, Wesley Wong, Dyann F. Wirth, and Daniel E. Neafsey. hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal*, 17(1):196, dec 2018.

[25] Lyndal Henden, Stuart Lee, Ivo Mueller, Alyssa Barry, and Melanie Bahlo. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLOS Genetics*, 14(5):e1007279, may 2018.

[26] Aimee R. Taylor, Stephen F. Schaffner, Gustavo C. Cerqueira, Standwell C. Nkhoma, Timothy J. C. Anderson, Kanlaya Sriprawat, Aung Pyae Phyo, François Nosten, Daniel E. Neafsey, and Caroline O. Buckee. Quantifying connectivity between local Plasmodium falciparum malaria parasite populations using identity by descent. *PLOS Genetics*, 13(10):e1007065, oct 2017.

[27] Aimee R. Taylor, Pierre E. Jacob, Daniel E. Neafsey, and Caroline O. Buckee. Estimating Relatedness Between Malaria Parasites. *Genetics*, 212(4):1337–1351, aug 2019.

[28] Aimee R. Taylor, Diego F. Echeverry, Timothy J.C. Anderson, Daniel E. Neafsey, and Caroline O. Buckee. Identity-by-descent with uncertainty characterises connectivity of Plasmodium falciparum populations on the Colombian-Pacific coast. *PLOS Genetics*, 16(11):e1009101, nov 2020.

[29] Inna Gerlovina, Boris Gerlovin, Isabel Rodríguez-Barraquer, and Bryan Greenhouse. Dcifer: an IBD-based method to calculate genetic distance between polyclonal infections. *Genetics*, 222(2), sep 2022.

[30] Philip Bejon, Thomas N Williams, Anne Liljander, Abdisalan M Noor, Juliana Wambua, Edna Ogada, Ally Olotu, Faith H a Osier, Simon I Hay, Anna Färnert, and Kevin Marsh. Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS medicine*, 7(7):e1000304, jul 2010.

[31] I. Omedo, P. Mogeni, T. Bousema, K. Rockett, A. Amambua-Ngwa, I. Oyier, J.C. Stevenson, A.Y. Baidjoe, E.P. de Villiers, G. Fegan, A. Ross, C. Hubbart, A. Jeffreys, T.N. Williams, D. Kwiatkowski, and P. Bejon. Micro-epidemiological structuring of Plasmodium falciparum parasite populations in regions with varying transmission intensities in Africa. *Wellcome Open Research*, 2, 2017.

[32] Magnus Manske, Olivo Miotto, Susana Campino, Sarah Auburn, Jacob Almagro-Garcia, Gareth Maslen, Jack O'Brien, Abdoulaye Djimde, Ogobara Doumbo, Issaka Zongo, Jean-Bosco Ouedraogo, Pascal Michon, Ivo Mueller, Peter Siba, Alexis Nzila, Steffen Borrmann, Steven M Kiara, Kevin Marsh, Hongying Jiang, Xin-Zhuan Su, Chanaki Amaratunga, Rick Fairhurst, Duong Socheat, Francois Nosten, Mallika Imwong, Nicholas J White, Mandy Sanders, Elisa Anastasi, Dan Alcock, Eleanor Drury, Samuel Oyola, Michael a Quail, Daniel J Turner, Valentin Ruano-Rubio, Dushyanth Jyothi, Lucas Amenga-Etego, Christina Hubbart, Anna Jeffreys, Kate Rowlands, Colin Sutherland, Cally Roper, Valentina Mangano, David Modiano, John C Tan, Michael T Ferdig, Alfred Amambua-Ngwa, David J Conway, Shannon Takala-Harrison, Christopher V Plowe, Julian C Rayner, Kirk a Rockett, Taane G Clark, Chris I Newbold, Matthew Berriman, Bronwyn Macinnis, and Dominic P Kwiatkowski. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379, jun 2012.

[33] Gavin Band, Ellen M. Leffler, Muminatou Jallow, Fatoumatta Sisay-Joof, Carolyne M. Ndila, Alexander W. Macharia, Christina Hubbart, Anna E. Jeffreys, Kate Rowlands, Thuy Nguyen, Sónia Gonçalves, Cristina V. Ariani, Jim Stalker, Richard D. Pearson, Roberto Amato, Eleanor Drury, Giorgio Sirugo, Umberto D'Alessandro, Kalifa A. Bojang, Kevin Marsh, Norbert Peshu, Joseph W. Saelens, Mahamadou Diakité, Steve M. Taylor, David J. Conway, Thomas N. Williams, Kirk A. Rockett, and Dominic P. Kwiatkowski. Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature*, 602(7895):106–111, feb 2022.

[34] Ronald Rosenberg, Robert Burge, and Imogene Schneider. An estimation of the number of malaria sporozoites ejected by a feeding mosquito. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(2):209–212, 1990.

[35] J. C. Beier, J. R. Davis, J. A. Vaughan, B. H. Noden, and M. S. Beier. Quantitation of Plasmodium falciparum sporozoites transmitted in vitro by experimentally infected Anopheles gambiae and Anopheles stephensi. *The American journal of tropical medicine and hygiene*, 44(5):564–570, 1991.

[36] Sarah Auburn, Susana Campino, Olivo Miotto, Abdoulaye a Djimde, Issaka Zongo, Magnus Manske, Gareth Maslen, Valentina Mangano, Daniel Alcock, Bronwyn MacInnis, Kirk a Rockett, Taane G Clark, Ogobara K Doumbo, Jean Bosco Ouédraogo, and Dominic P Kwiatkowski. Characterization of within-host Plasmodium falciparum diversity using next-generation sequence data. *PloS one*, 7(2):e32891, jan 2012.

[37] S Thaithong, G H Beale, B Fenton, J McBride, V Rosario, A Walker, and D Walliker. Clonal diversity in a single isolate of the malaria parasite Plasmodium falciparum. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 78(2):242–5, 1984.

[38] S Viriyakosol, N Siripoon, C Petcharapirat, P Petcharapirat, W Jarra, S Thaithong, K N Brown, and G Snounou. Genotyping of Plasmodium falciparum isolates by the polymerase chain reaction and potential uses in epidemiological studies. *Bulletin of the World Health Organization*, 73(1):85–95, 1995.

[39] Kevin Galinsky, Clarissa Valim, Arielle Salmier, Benoit de Thoisy, Lise Musset, Eric Legrand, Aubrey Faust, Mary Lynn Baniecki, Daouda Ndiaye, Rachel F Daniels, Daniel L Hartl, Pardis C Sabeti, Dyann F Wirth, Sarah K Volkman, and Daniel E Neafsey. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria journal*, 14:4, 2015.

[40] Hsiao-Han Chang, Colin J. Worby, Adoke Yeka, Joaniter Nankabirwa, Moses R. Kamya, Sarah G. Staedke, Grant Dorsey, Maxwell Murphy, Daniel E. Neafsey, Anna E. Jeffreys, Christina Hubbart, Kirk A. Rockett, Roberto Amato, Dominic P. Kwiatkowski, Caroline O. Buckee, and Bryan Greenhouse. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. *PLOS Computational Biology*, 13(1):e1005348, jan 2017.

[41] Sha Joe Zhu, Jason A. Hendry, Jacob Almagro-Garcia, Richard D. Pearson, Roberto Amato, Alistair Miles, Daniel J. Weiss, Tim C.D. Lucas, Michele Nguyen, Peter W. Gething, Dominic Kwiatkowski, and Gil McVean. The origins and relatedness structure of mixed infections vary with local prevalence of P. Falciparum malaria. *eLife*, 8, jul 2019.

[42] Bronner P. Gonçalves, Melissa C. Kapulu, Patrick Sawa, Wamdaogo M. Guelbéogo, Alfred B. Tiono, Lynn Grignard, Will Stone, Joel Hellewell, Kjerstin Lanke, Guido J.H. Bastiaens, John Bradley, Issa Nébié, Joyce M. Ngoi, Robin Oriango, Dora Mkabili, Maureen Nyaurah, Janet Midega, Dyann F. Wirth, Kevin Marsh, Thomas S. Churcher, Philip Bejon, Sodiomon B. Sirima, Chris Drakeley, and Teun Bousema. Examining the human infectious reservoir for Plasmodium falciparum malaria in areas of differing transmission intensity. *Nature Communications 2017 8:1*, 8(1):1–11, oct 2017.

[43] Brenna M. Henn, L. L. Cavalli-Sforza, and Marcus W. Feldman. The great human expansion. *Proceedings of the National Academy of Sciences of the United States of America*, 109(44):17758–17764, oct 2012.

[44] Hsiao-Han Chang, Daniel J Park, Kevin J Galinsky, Stephen F Schaffner, Daouda Ndiaye, Omar Ndir, Souleymane Mboup, Roger C Wiegand, Sarah K Volkman, Pardis C Sabeti, Dyann F Wirth, Daniel E Neafsey, and Daniel L Hartl. Genomic sequencing of Plasmodium falciparum malaria parasites from Senegal reveals the demographic history of the population. *Molecular biology and evolution*, 29(11):3427–3439, nov 2012.

[45] Standwell C. Nkhoma, Shalini Nair, Salma Al-Saai, Elizabeth Ashley, Rose McGready, Aung P. Phyo, François Nosten, and Tim J. C. Anderson. Population genetic correlates of declining transmission in a human pathogen. *Molecular Ecology*, 22(2):273–285, jan 2013.

[46] Timothy J.C. Anderson, Shalini Nair, Marina McDew-White, Ian H. Cheeseman, Standwell Nkhoma, Fatma Bilgic, Rose McGready, Elizabeth Ashley, Aung Pyae Phyo, Nicholas J. White, and François Nosten. Population Parameters Underlying an Ongoing Soft Sweep in Southeast Asian Malaria Parasites. *Molecular Biology and Evolution*, 34(1):131–144, jan 2017.

[47] Hsiao Han Chang and Daniel L. Hartl. Recurrent bottlenecks in the malaria life cycle obscure signals of positive selection. *Parasitology*, 142(S1):S98–S107, feb 2015.

[48] Flavia Camponovo, Caroline O Buckee, and Aimee R Taylor. Measurably recombining malaria parasites. *Trends in parasitology*, 39(1), nov 2023.

[49] Seth C. Inzaule, Sofonias K. Tessema, Yenew Kebede, Ahmed E. Ogwell Ouma, and John N. Nkengasong. Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *The Lancet. Infectious diseases*, 21(9):e281–e289, sep 2021.

[50] Roberto Amato, Richard D Pearson, Jacob Almagro-Garcia, Chanaki Amaratunga, Pharath Lim, Seila Suon, Sokunthea Sreng, Eleanor Drury, Jim Stalker, Olivo Miotto, Rick M Fairhurst, and Dominic P Kwiatkowski. Origins of the current outbreak of multidrug-resistant malaria in southeast Asia: a retrospective genetic study. *The Lancet Infectious Diseases*, 18:337–345, feb 2018.

[51] Philip A. Eckhoff. Malaria parasite diversity and transmission intensity affect development of parasitological immunity in a mathematical model. *Malaria journal*, 11, 2012.

[52] Jamie T. Griffin, Samir Bhatt, Marianne E. Sinka, Peter W. Gething, Michael Lynch, Edith Patouillard, Erin Shutes, Robert D. Newman, Pedro Alonso, Richard E. Cibulskis, and Azra C. Ghani. Potential for reduction of burden and local elimination of malaria by reducing Plasmodium falciparum malaria transmission: a mathematical modelling study. *The Lancet. Infectious diseases*, 16(4):465–472, apr 2016.

[53] Malcolm J Gardner, Neil Hall, Eula Fung, Owen White, Matthew Berriman, Richard W Hyman, Jane M Carlton, Arnab Pain, Karen E Nelson, Sharen Bowman, Ian T Paulsen, Keith James, Jonathan A Eisen, Kim Rutherford, Steven L Salzberg, Alister Craig, Sue Kyes, Man-Suen S Chan, Vishvanath Nene, Shamira J Shallom, Bernard Suh, Jeremy Peterson, Sam Angiuoli, Mihaela Pertea, Jonathan Allen, Jeremy Selengut, Daniel Haft, Michael W Mather, Akhil B Vaidya, David M A Martin, Alan H Fairlamb, Martin J Fraunholz, David S Roos, Stuart A Ralph, Geoffrey I McFadden, Leda M Cummings, G Mani Subramanian, Chris Mungall, J Craig Venter, Daniel J Carucci, Stephen L Hoffman, Chris Newbold, Ronald W Davis, Claire M Fraser, and Bart Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, oct 2002.

Glossary

This glossary explains the terminology used in this paper, which sometimes differs from common usage, or is a specific interpretation of it.

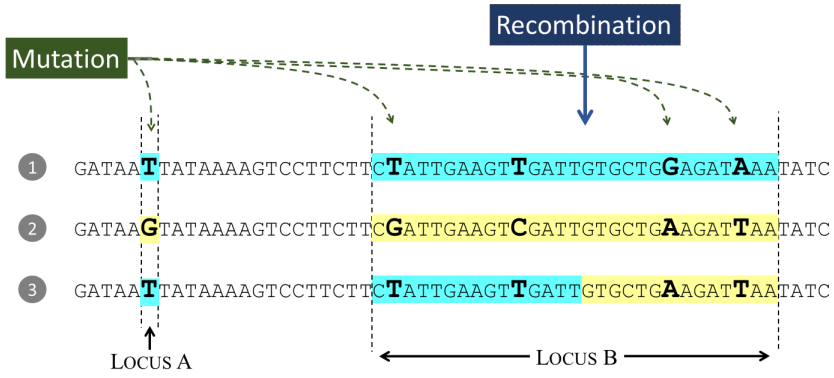

Figure 26: **Heterozygosity is measured by comparing alleles at a locus**. An allele is an instance of the parasite genome and a locus is a specific location in the genome. Here we see three alleles at two loci: locus A is a single nucleotide position (we call this a point locus) and locus B extends over multiple nucleotide positions (we call this a haplotype locus). Both loci have been affected by mutation, and locus B has also been affected by recombination.

Allele. An instance of the parasite genome. For example, n haploid individuals correspond to n alleles. We often speak of an allele with reference to a particular locus, in which case it means the DNA sequence of that locus in an individual parasite genome.

Coalescence. If two lineages are traced back in time, coalescence occurs when they meet in the same ancestral allele.

Cotransmission. Transmission from one host to another of a mixture of parasite alleles with different ancestral histories.

Crossing rate of transmission chains (χ). The proportion of hosts that acquire parasites from more than one transmission chain, i.e. from more than one host in the previous generation. This is equivalent to proportion of hosts that are superinfected.

Effective number of hosts (N_h). The number of hosts that effectively transmit parasites in each generation of the transmission graph. This is a form of population bottleneck.

Effective recombination. Recombination between genetically distinct alleles that acts to change the DNA sequence of a haplotype locus.

Effective recombination parameter ϕ_t. The probability that, if recombination occurs at a locus at time t, this will change the DNA sequence of the locus.

Haplotype. A specific DNA sequence observed at a haplotype locus. At a large haplotype locus there will typically be many different haplotypes.

Haplotype locus. A locus that extends over multiple nucleotide positions and that can therefore undergo recombination (figure 26).

Heterozygosity (H). The probability that two alleles sampled randomly from some population are heterozygous, i.e. that they have different DNA sequences.

Homozygosity (G). The probability that two alleles sampled randomly from some population are homozygous, i.e. that they have the same DNA sequence.

Host. A person that is carrying parasites and capable of transmitting them to others. Each host exists for a single generation of the genomic transmission graph.

Lineage. A path that traces the ancestry of an allele at a point locus, going backwards in time through the transmission graph. A point locus is not affected by recombination, so a lineage can be traced back over many generations despite frequent recombination events.

Locus. A specific location in the genome. This can be either a single nucleotide position (a point locus) or a sequence extending over multiple nucleotide positions (a haplotype locus).

Nucleotide diversity (π). The probability that two alleles are heterozygous at a random nucleotide position in the genome.

Parasite. A malaria parasite of the species *Plasmodium falciparum* that is transmitted from host to host by a mosquito vector. It is a single-celled organism that is haploid for most of its lifecycle. It reproduces asexually apart from a brief phase of sexual reproduction within the vector (see figure 27).

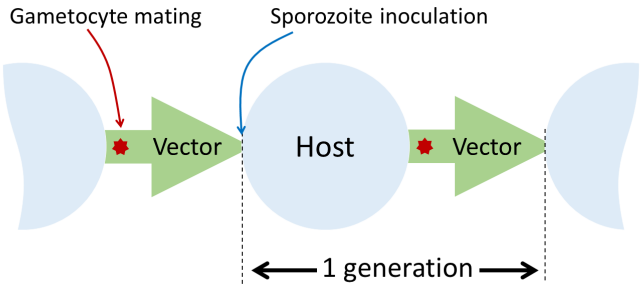

Figure 27: **One cycle of host-to-host transmission corresponds to one generation of sexual reproduction by the parasite**. *Plasmodium falciparum* parasites reproduce asexually and prolifically within the red blood cells of their human host. Gametocytes (sexual forms of the parasite) mate after being taken up by a blood-sucking *Anopheles* mosquito. Their progeny then go through a series of developmental stages to produce sporozoites (asexual forms of the parasite that are inoculated by the mosquito into a new host). In our model, we define the start of a new generation of transmission as the inoculation of sporozoites into a new host, noting that this occurs shortly after sexual reproduction within the vector.

Point locus. A specific single nucleotide position in the genome.

Polymorphism. Variation in the DNA sequence at some specified locus.

Quantum of transmission (Q). The number of parasite alleles transmitted from one host to the next via a vector. Q summarises a complex series of bottlenecks in host-vector and vector-host transmission occuring during one generation of the parasite life-cycle.

Single nucleotide polymorphism (SNP). A polymorphism that involves only a single nucleotide position. It usually arises from a type of mutation called a single nucleotide substitution.

Superinfection. Infection of a host with parasites from more than one source in the previous generation. In the genomic transmission graph this is equivalent to crossing of transmission chains.

Transmission bottleneck. A population bottleneck that affects the number of alleles that are passed from one generation of the genomic transmission graph to the next. From the perspective of the parasite population as a whole, the quantum of transmission Q and the effective number of hosts N_h can both be considered as transmission bottlenecks.

Transmission chain. A sequence of host-to-host transmission events. If we pick any node in the transmission graph, and trace a path forward in time along the edges to another node, that is a transmission chain.

Vector. An *Anopheles* mosquito that transmits malaria parasites from one host to another.

Methods

1 A framework for modelling the coalescent process

1.1 The coalescent process when $\chi = 0$

To understand the coalescent properties of the genomic transmission graph, it is instructive to start by considering a parasite population with no superinfection, i.e. $\chi = 0$.

Imagine that we sample two alleles from different hosts and, focusing on a point locus, we follow their lineages back in time until they coalesce in a common ancestral allele, as illustrated in figure 5. Let **T** be a random variable representing time to coalescence of the two alleles.

If we sample two alleles from different hosts in the same generation, they are by definition on *separate* transmission chains, but if we trace their lineages back in time they will eventually *meet* in the same host and thus in the same transmission chain. If two lineages are separated in generation t then the probability that they meet in generation $t-1$ is $1/N_h$. Let T_1 be the expectation of the time taken for two lineages to meet in the same host.

$$T_1 = \sum_{i=1}^{\infty} i\Big(\frac{1}{N_h}\Big)\Big(1-\frac{1}{N_h}\Big)^{i-1} = N_h$$

Once the two lineages have met in the same host, as we proceed back in time, the two lineages are *cotransmitted* along the same transmission chain until they eventually *coalesce* in a common ancestral allele. If two lineages are cotransmitted in generation t then the probability that they coalesce in generation $t-1$ is $1/Q$. Let T_2 be the expectation of the time taken for two lineages to coalesce after meeting in the same transmission chain.

$$T_2 = \sum_{j=0}^{\infty} j\Big(\frac{1}{Q}\Big)\Big(1-\frac{1}{Q}\Big)^{j} = Q-1$$

Here we are allowing for the possibility that coalescence could occur as soon as two lineages meet in the same host, as represented by the condition $j = 0$ in the above expression. We can now combine these two parts to get the expectation of time to coalescence:

$$E\{\mathbf{T}\} = T_1 + T_2 = N_h + Q - 1$$

1.2 The coalescent process when $\chi \geq 0$

Imagine that we sample two alleles at some point locus and follow the two lineages back in time until they coalesce, as illustrated in figure 6. At any point in time the system must be in one of three states:

- SEPARATED - the two lineages are in different hosts
- COTRANSMITTED - the two lineages are in the same host
- COALESCED - the two lineages have coalesced

If two lineages are separated and we go back a single generation, these are the possibilities:

1. the two lineages meet in the same host (Pr $= 1/N_h$) and
   (a) they coalesce (Pr $= 1/Q$)
   (b) they are cotransmitted (Pr $= 1 - 1/Q$)
2. or the two lineages stay separated (Pr $= 1 - 1/N_h$)

from which we obtain these transition probabilities:

$$\Pr\{\text{SEPARATED} \rightarrow \text{SEPARATED}\} = 1 - \frac{1}{N_h}$$

$$\Pr\{\text{SEPARATED} \rightarrow \text{COTRANSMIITTED}\} = \frac{1}{N_h}(1 - \frac{1}{Q})$$

$$\Pr\{\text{SEPARATED} \rightarrow \text{COALESCED}\} = \frac{1}{N_h Q}$$

If two lineages are cotransmitted and we go back a single generation, these are the possibilities:

1. there is one source of infection ($\Pr = 1 - \chi$)
   (a) the lineages coalesce ($\Pr = 1/Q$)
   (b) the lineages remain cotransmitted ($\Pr = 1 - 1/Q$)
2. there are two sources of infection ($\Pr = \chi$)
   (a) the lineages come from different sources ($\Pr = Q/(2Q-1)$)
   (b) the lineages come from the same source ($\Pr = (Q-1)/(2Q-1)$)
      i. they coalesce ($\Pr = 1/Q$)
      ii. they are cotransmitted ($\Pr = 1 - 1/Q$).

from which we obtain these transition probabilities:

$$\Pr\{\text{COTRANSMITTED} \rightarrow \text{SEPARATED}\} = \frac{Q\chi}{2Q-1}$$

$$\Pr\{\text{COTRANSMITTED} \rightarrow \text{COTRANSMITTED}\} = \frac{(Q-1)(2Q-Q\chi-1)}{Q(2Q-1)}$$

$$\Pr\{\text{COTRANSMITTED} \rightarrow \text{COALESCE}\} = \frac{2Q-Q\chi-1}{Q(2Q-1)}$$

Based on these observations we can construct a matrix of transition probabilities for the three possible states of two lineages as we proceed back in time through the transmission graph (table 1).

1.3 Markov chain simulation of time to coalescence

Using the matrix of transmission probabilities (table 1) we can calculate the probability distribution of time to coalescence for any combination of the transmission parameters N_h, Q and χ. As we have seen, if we follow two lineages back in time they can be in three possible states: (a) separated or (b) cotransmitted or (c) coalesced. Let $P_{a,t}$, $P_{b,t}$ and $P_{c,t}$ respectively denote the probabilities of these states at time t. We can represent the overall state of the system at time t by a probability vector $\mathbf{X}_t$ where

$$\mathbf{X}_t = \begin{bmatrix} P_{a,t} & P_{b,t} & P_{c,t} \end{bmatrix}$$

Let $\mathbf{Y}$ be a matrix of transition probabilities, where y_{ij} is the probability that state i will transition to state j if we go back a single generation, as represented in table 1. As we move back in time, i.e. as we proceed from time t to $t-1$,

$$\mathbf{X}_{t-1} = \mathbf{X}_t \mathbf{Y}$$

This allows us to compute the probability of each state at any given time by Markov chain simulation. In each simulation, we sample two imaginary alleles and follow their lineages back

in time, recalculating the probability of each state as we move from generation to generation, as shown in figure 7.

To study between-host variation we specify that two alleles are sampled from different hosts, i.e. at the start of the simulation $\mathbf{X}_0 = \begin{bmatrix}1 & 0 & 0\end{bmatrix}$. Alternatively, we can study within-host variation by specifying that two alleles are sampled from the same host, i.e. $\mathbf{X}_0 = \begin{bmatrix}0 & 1 & 0\end{bmatrix}$. By running the Markov chain simulation over many generations we obtain the probability distribution of coalescence time, as illustrated in figure 8.

1.4 Coalescence time when $\chi = 1$ and $Q = 1$

In general, two alleles sampled from the same host coalesce more rapidly than two alleles sampled from different hosts. This difference is marked when $\chi = 0$. As χ increases, the coalescence time of two alleles sampled from the same host starts to approach that of two alleles sampled from different hosts, and when $\chi = 1$ the difference becomes relatively small.

There is an exception to this general rule. When $\chi = 1$ and $Q = 1$, the mean coalescence time of two alleles sampled from the same host is exactly one generation greater than that of two alleles sampled from different hosts. This apparent anomaly arises because, in this situation, the two alleles acquired by a host must come from different source hosts so they cannot coalesce in the generation prior to that in which they are sampled.

2 Genetic variation at a point locus

Imagine that we sample two alleles at a point locus and trace their lineages back in time until they coalesce. The two alleles must have the same DNA sequence if neither lineage is affected by mutation. In principle the two alleles could have the same DNA sequence if both lineages are affected by mutation but we rule out this possibility by assuming an 'infinite alleles' model. Let u be the mutation rate per generation at this locus, let G be the homozygosity of the locus, and let $\mathbf{T}$ be a random variable representing the time to coalescence of two lineages measured in generations.

$$G = (1-u)^{2\mathbf{T}}$$

Let H be the heterozygosity of our point locus, where $H = 1 - G$. We obtain the expectation of H for any two alleles sampled at random from the population by summing over the probability distribution of time to coalescence.

$$E\{H\} = 1 - \sum_{i=1}^{\infty} \Pr\{\mathbf{T} = i\} \times (1-u)^{2i}$$

Let T_C be the mean time to coalescence measured in generations. If u is sufficiently small (say $< 10^{-5}$) we can safely ignore factors of u^2 and above to make the approximation

$$E\{H\} \approx 1 - \sum_{i=1}^{\infty} \Pr\{\mathbf{T} = i\} + 2u \sum_{i=1}^{\infty} i \Pr\{\mathbf{T} = i\}$$

$$E\{H\} \approx 2uT_C$$

2.1 Nucleotide diversity of the global parasite population

The MalariaGEN Pf6 and Pf7 datasets [18,19] both contain estimates of π based on coding regions of the core *P. falciparum* genome. The results differ between the two datasets as can be seen by comparing the upper and lower panels of figure 28. In African samples, the median value of π is $\sim 2.5 \times 10^{-4}$ in Pf6 and $\sim 5 \times 10^{-4}$ in Pf7. This difference could be partly because the Pf6 estimate is restricted to biallelic SNPs whereas the Pf7 estimate includes multiallelic SNPs. Another factor is that Pf7 has approximately twice as many high-quality SNPs and short indels as Pf6 (6 versus 3 million), probably due to a combination of increased sample size and minor alterations to the variant calling algorithms.

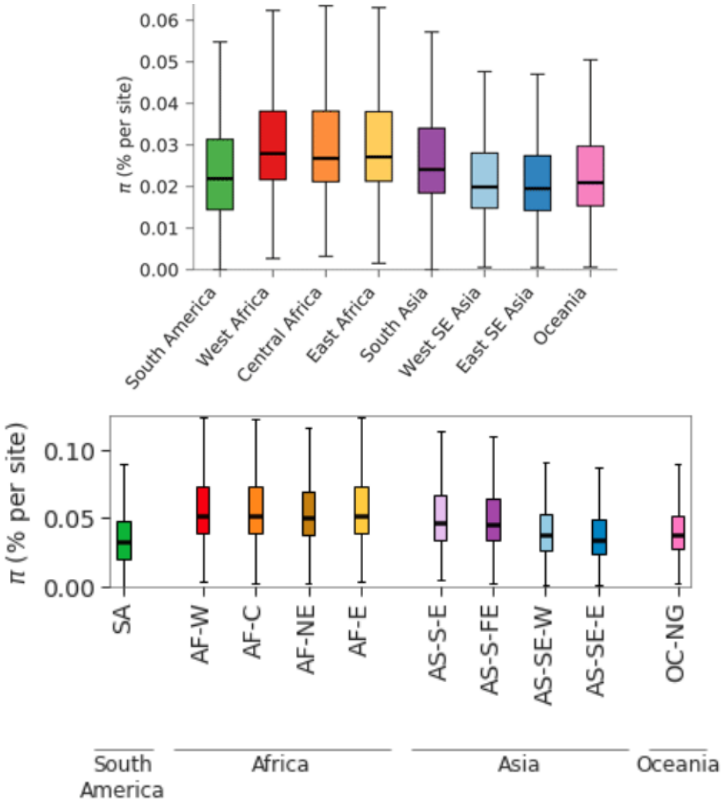

Figure 28: **Estimates of nucleotide diversity in coding regions of the *P. falciparum* genome.** Upper panel is copied from the MalariaGEN Pf6 dataset (reference [18] fig. 2) and lower panel is from the MalariaGEN Pf7 dataset (reference [19] supp. fig. 6). Thick lines represent median values, boxes show the interquartile range, and whiskers represent the bulk of the distribution, discounting outliers. Possible reasons for differences between the Pf6 and Pf7 results are discussed in the text.

The reason for considering only coding regions is that non-coding regions are error-prone for variant calling due to their very high AT content with many short tandem repeats. Our algorithms are attuned to minimise variant calling errors but both false positive and false negative results are possible. Note also that some *P. falciparum* genes are under purifying selection (which tends to reduce π) while others are under diversifying selection (which tends to increase π).

In this paper we use $\pi \approx 4 \times 10^{-4}$ as a first approximation for the global parasite population, but it will be clear that there are multiple potential sources of error and this could be either an overestimate or an underestimate.

3 Genetic variation at a haplotype locus

3.1 Shared haplotype segments

Shared haplotype segments are segments of the genome where two parasites have identical DNA sequences. Unrelated parasites often have some shared haplotype segments that extend over a few kilobases, but if we observe a substantial number of shared haplotype segments that are hundreds of kilobases long, this suggests that the parasites share a recent common ancestor.

We would like to evaluate the expected proportion of the genome that is occupied by shared haplotype segments. Imagine a chromosome of length c centimorgans that is divided into non-overlapping blocks of one centimorgan size as illustrated in figure 29. We are interested in shared haplotype segments of > 2 centimorgans and so we will make the simplifying assumption that haplotype loci are constructed of an integral number of these one centimorgan blocks. More precisely, we specify that a haplotype locus that occupies i blocks of the imaginary chromosome

has a real length in the interval $(i-1, i]$ centimorgans.

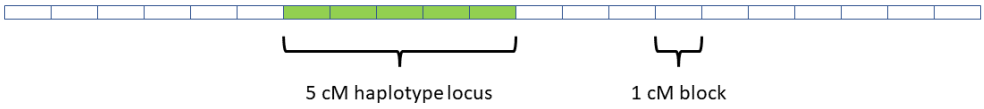

Figure 29: **Simplified architecture of a haplotype locus.** We imagine that a chromosome is divided into non-overlapping blocks of 1 centimorgan size, and that each haplotype locus is constructed of an integral number of these blocks. Thus a haplotype locus in the interval $(4, 5]$ centimorgans is considered to be 5 blocks.

Take a haplotype locus of i blocks and let g_i be the probability that this is a shared haplotype segment with respect to two alleles randomly sampled from the population. We can evaluate g_i from equation 4 since:

$$g_i = E\{G_{i-1}\} - E\{G_i\}$$

Let ω_i be the expected number of shared haplotype segments of i blocks. The number of i-block segments that can be fitted into a chromosome is c/i, so we will make the very crude approximation that

$$\omega_i \approx g_i \times \frac{c}{i}$$

To quantify identity by descent (IBD) we are interested in shared haplotype segments of above a certain size. The total length of shared haplotype segments of $> k$ centimorgans is given by the sum

$$\sum_{i=k+1}^{c} i\omega_i \approx \sum_{i=k+1}^{c} \frac{ig_i c}{i} \approx \sum_{i=k+1}^{c} g_i c$$

Thus the proportion of the chromosome occupied by shared haplotype segments of $> k$ centimorgans is very approximately given by

$$\sum_{i=k+1}^{c} i\omega_i \div c = \sum_{i=k+1}^{c} g_i = E\{G_k\} \tag{14}$$

We can extrapolate this result to the whole genome. This tells us that the proportion of the genome occupied by shared haplotype segments of > 2 centimorgans can be crudely approximated by γ, the mean haplotype homozygosity of a 2 centimorgan locus (figure 12).

4 Migration in a hierarchical population structure

Imagine that we sample two alleles from a local subpopulation of parasites that is embedded within a much larger metapopulation, and follow the two lineages back in time until they coalesce.

If we sample two alleles from the subpopulation and go back in time, the two lineages have six possible states: (i) in different hosts in the subpopulation; (ii) in the same host in the subpopulation; (iii) in different hosts in the metapopulation; (iv) in the same host in the metapopulation; (v) one lineage in the subpopulation and the other in the metapopulation; (vi) coalesced. We could work out the transition probabilities between these six states, but if the metapopulation is much larger than the subpopulation then we can make some simplifying assumptions that will help to clarify the population dynamics as well as speeding up our Markov chain simulations.

Let m be the probability that a host within the local subpopulation acquired their infection from the metapopulation, and let the number of such hosts per generation be $N_m = mN_h$. These

migrant hosts could be either immigrants from the metapopulation or local residents who have been travelling outside the local area. Let m' be the probability that a host within the metapopulation acquired their infection from the local parasite subpopulation, and let N'_m and N'_h be corresponding terms for the metapopulation.

Coalescence must occur either in the subpopulation or the metapopulation. If the metapopulation is much larger than the local subpopulation ($N'_h \gg N_h$) and if absolute rates of migration between the two populations are approximately symmetrical ($m'N'_h \approx mN_h$) then it must be the case that $m' \ll m$. Thus if one lineage moves from the subpopulation into the metapopulation (as we proceed back in time) then it is very unlikely that it will move back into the subpopulation before the other lineage joins it in the metapopulation. In other words, as soon as one lineage has entered the metapopulation, then it becomes highly probable that coalescence will eventually occur in the metapopulation.

Thus we can consider the metapopulation as an absorbing state from the perspective of the subpopulation, because once a lineage has entered the metapopulation we might as well treat both lineages as being in the metapopulation, as that is where they must coalesce. This allows us to organise our simulation into two compartments:

1. When we model the behaviour of two lineages within the subpopulation, there are four possible states:
   (a) in different hosts in the subpopulation
   (b) in the same host in the subpopulation
   (c) coalesced in the subpopulation (absorbing state)
   (d) entered the metapopulation (absorbing state)
2. If a lineage goes into the metapopulation, we treat both lineages as being in the metapopulation and consider three possible states:
   (a) in different hosts in the metapopulation
   (b) in the same host in the metapopulation
   (c) coalesced within the metapopulation (absorbing state)

Framed in this way, the metapopulation is equivalent to the simple population whose transition probability matrix is given by table 1. The transition probabilities of the subpopulation can be worked out as follows.

If two lineages are separated within the subpopulation and we go back a generation, these are the possible outcomes:

1. They are in the same host. $\Pr = 1/N_h$
   (a) They coalesce. $\Pr = 1/Q$
   (b) They are cotransmitted. $\Pr = 1 - 1/Q$
      i. They remain in the subpopulation. $\Pr = 1 - m$
      ii. They enter the metapopulation. $\Pr = m$
2. They are not in the same host. $\Pr = 1 - 1/N_h$
   (a) Both remain in the subpopulation. $\Pr = (1-m)^2$
   (b) One or both enter metapopulation. $\Pr = 2m - m^2$

From this we obtain

$$\Pr\{\text{SEPARATED} \to \text{COALESCED}\} = \frac{1}{N_h Q}$$

$$\Pr\{\text{SEPARATED} \to \text{COTRANSMITTED}\} = \frac{(Q-1)(1-m)}{N_h Q}$$

$$\Pr\{\text{SEPARATED} \to \text{SEPARATED}\} = \frac{(N_h-1)(1-m)^2}{N_h}$$

$$\Pr\{\text{SEPARATED} \to \text{METAPOPULATION}\} = \frac{(Q-1)m}{N_h Q} + \frac{(N_h-1)(2m-m^2)}{N_h}$$

$$= \frac{m(Q-1)+Q(N_h-1)(2m-m^2)}{N_h Q}$$

If two lineages are cotransmitted within the subpopulation and we go back a generation, these are the possible outcomes:

1. If the current host is multiply infected. Pr $= \chi$
   (a) They remain cotransmitted. Pr $= (Q-1)/(2Q-1)$
      i. They coalesce. Pr $= 1/Q$
      ii. They do not coalesce. (Pr $= 1-1/Q$)
         A. They remain in the subpopulation. Pr $= 1-m$
         B. They enter the metapopulation. Pr $= m$
   (b) They become separated. Pr $= Q/(2Q-1)$
      i. They both remain in the subpopulation. Pr $= (1-m)^2$
      ii. One or both enters the metapopulation. Pr $= 2m-m^2$
2. If the current host is not multiply infected. Pr $= 1-\chi$
   (a) They coalesce. Pr $= 1/Q$
   (b) They do not coalesce, i.e. they remain cotransmitted. Pr $= 1-1/Q$
      i. They remain within the subpopulation. Pr $= 1-m$
      ii. They enter the metapopulation. Pr $= m$

From this we obtain

$$\Pr\{\text{COTRANSMITTED} \to \text{COALESCED}\} = \frac{2Q-Q\chi-1}{Q(2Q-1)}$$

$$\Pr\{\text{COTRANSMITTED} \to \text{COTRANSMITTED}\} = \frac{(Q-1)(1-m)(2Q-\chi Q-1)}{Q(2Q-1)}$$

$$\Pr\{\text{COTRANSMITTED} \to \text{SEPARATED}\} = \frac{\chi Q(1-m)^2}{2Q-1}$$

$$\Pr\{\text{COTRANSMITTED} \to \text{METAPOPULATION}\} = \frac{mQ(2Q-3+\chi+\chi Q-m\chi Q)+m}{Q(2Q-1)}$$

This gives us a matrix of transition probabilities as shown in table 5

5 Analysis of within-host variation

5.1 Within-host heterozygosity in a population with $\chi = 0$

First we consider a population in which there is no superinfection, i.e. $\chi = 0$. We assume that evolution is neutral, i.e. there is no effect of natural selection.

Imagine that host A transmits parasites to host B. Let G_A be the probability that two alleles sampled from host A are homozygous, and H_A the probability that they are heterozygous, where $H_A = 1 - G_A$. Likewise G_B and H_B for host B. The relationship between H_A and H_B will be determined by the combination of genetic drift and mutation in just the same way as the Wright-Fisher model.

State	Separated	Cotransmitted	Coalesced	Metapopulation
Separated	$\frac{(N_h-1)(1-m)^2}{N_h}$	$\frac{(Q-1)(1-m)}{N_hQ}$	$\frac{1}{N_hQ}$	$\frac{m(Q-1)+Q(N_h-1)(2m-m^2)}{N_hQ}$
Cotransmitted	$\frac{Q\chi(1-m)^2}{2Q-1}$	$\frac{(Q-1)(1-m)(2Q-Q\chi-1)}{Q(2Q-1)}$	$\frac{2Q-Q\chi-1}{Q(2Q-1)}$	$\frac{mQ(2Q-3+\chi+\chi Q-m\chi Q)+m}{Q(2Q-1)}$
Coalesced	0	0	1	0
Metapopulation	0	0	0	1

Table 5: **Transition probabilities for the four possible states of two lineages within a subpopulation**. At any point in time, two lineages must be (1) separated within the subpopulation or (2) cotransmitted within the subpopulation or (3) coalesced within the subpopulation or (4) entered the metapopulation. As discussed in the text, if the metapopulation is much larger than the subpopulation then we can treat it as an absorbing state. Row i column j of the table gives the probability that lineages in state i will transition to state j if we go back a single generation.

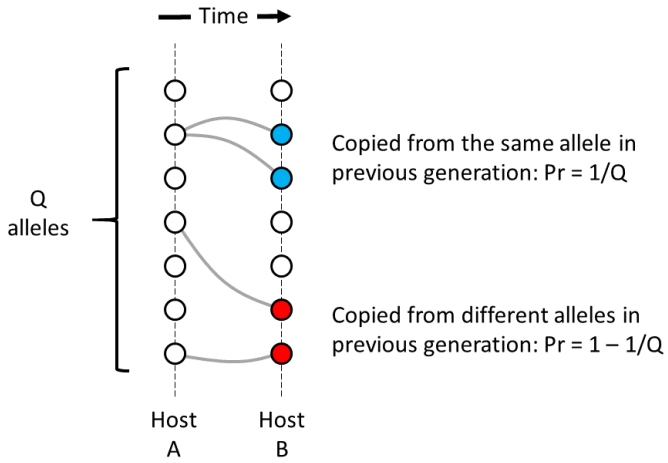

Figure 30: **Copying of alleles from one host to the next along a transmission chain with no superinfection.** Alleles are copied with replacement from Q alleles in the previous generation, exactly analogous to the Wright-Fisher model.

Genetic drift. If we sample two alleles from host B (Figure 30) there are two ways in which they can be homozygous:

1. they are copied from the same allele in host A (Pr $= 1/Q$);
2. they are copied from different alleles in host A (Pr $= 1-(1/Q)$) but these are homozygous (Pr $= G_A$).

Since these are mutually exclusive possibilities

$$G_B = \frac{1}{Q} + G_A(1-\frac{1}{Q}) \tag{15}$$

and after substitution and rearrangement

$$H_B = H_A(1-\frac{1}{Q})$$

Mutation. Let u be the probability that an allele is altered by mutation during one generation of transmission. We assume infinite alleles, i.e. if two alleles are initially homozygous then a mutation in one or both alleles will make them heterozygous. We can account for mutation by incorporating into equation 15 the probability of $(1-u)^2$ that neither of the two alleles is affected by mutation:

$$G_B = \Big(\frac{1}{Q} + G_A(1-\frac{1}{Q})\Big)(1-u)^2$$

Since u is generally very small we can usually ignore factors of u^2, so after substitution and rearrangement we obtain

$$H_B \approx H_A\Big(1-\frac{1}{Q}-2u+\frac{2u}{Q}\Big)+2u \tag{16}$$

Hosts that are x generations apart on the same transmission chain. Now imagine two hosts that are x generations apart on the same transmission chain. Let the within-host heterozygosity of the first host (i.e. the one that exists earlier in time) be H and that of the second host be H'. If we apply equation 16 over multiple generations we obtain

$$H' \approx H\alpha^x + 2u\sum_{i=0}^{x-1}\alpha^i \tag{17}$$

where

$$\alpha = (1-\frac{1}{Q})(1-2u) \tag{18}$$

The first part of equation 17 describes geometric decay of the initial heterozygosity due to genetic drift, while the second part describes a new source of heterozygosity that gradually builds up due to the accumulation of mutations, attenuated by drift.

Equilibrium state of heterozygosity in a non-crossing transmission chain. If a transmission chain continues for many generations without crossing with another transmission chain, its within-host heterozygosity H_W equilibrates when the effect of genetic drift is equal and opposite to the effect of mutation. We can evaluate this equilibrium value theoretically by letting $x \to \infty$ in equation 17.

$$H_W \approx H\alpha^\infty + 2u\sum_{i=0}^{\infty}\alpha^i$$

$$= \frac{2u}{1-1+1/Q+2u-2u/Q}$$

$$H_W \approx \frac{2uQ}{1+2uQ-2u} \tag{19}$$

If we replace u with μ, the genome-wide single nucleotide substitution rate we can obtain π_W, the expected level of within-host nucleotide diversity in a population where there has been no superinfection for many generations:

$$\pi_W \approx \frac{2\mu Q}{1+2\mu Q-2\mu}$$

and by rearrangement

$$Q \approx \frac{\pi_W(1-2\mu)}{2\mu(1-\pi_W)} \approx \frac{\pi_W}{2\mu}$$

5.2 Estimating within-host nucleotide diversity π_W from genome sequencing data.

To examine the frequency distribution of within-host nucleotide diversity π_W in the MalariaGEN Pf6 dataset, we begin by selecting high-quality samples (n = 5,970) and high-quality biallelic coding SNPs with vqslod > 3 (n = 502,221). We then calculate

- within-host heterozygosity for each SNP in each sample
- mean within-host heterozygosity for each SNP across all samples ($\widehat{H}_W$)
- mean within-host heterozygosity for each sample across all SNPs ($\widehat{H}_{sample}$)

As in section 2.1, we restrict our analysis to coding regions because non-coding regions are error-prone due to their very high AT content with many short tandem repeats. The MalariaGEN Pf6 dataset has quality control processes that endeavour to minimise variant calling errors but both false positive and false negative results are possible. Also, some *P. falciparum* genes are under purifying selection (which tends to reduce π) while others are under diversifying selection (which tends to increase π).

A particularly important source of error for this analysis is overestimation of heterozygosity due to genome sequence alignment artefacts. This affects some SNPs much more severely than others: we call these hyperhets and they are discussed in some detail in the supplementary material to reference [32]. To reduce the number of hyperhet artefacts, we filter out SNPs with $\widehat{H}_W \geq 0.02$, i.e. a mean minor allele frequency of > 0.01. However this is a crude method which may lead to us overestimate heterozygosity (if we have failed to exclude all artefactual heterozygote calls) or underestimate it (if we have over-corrected by filtering out valid heterozygote calls).

With these caveats, we are left with 494,829 coding SNPs to analyse. There are 12,028,350 coding positions in the *P. falciparum* genome [53]. If we make the simplifying assumption that there is no variation in the (12,028,350 - 494,829) coding positions that lie outside our set of 494,829 coding SNPs, then the within-host nucleotide diversity of coding positions is given by:

$$\pi_W = \widehat{H}_{sample} \times \frac{494829}{12028350}$$

A histogram of the number of samples with different values of π_W is shown in figure 15.

TBD. Show histograms for different filter cutoffs for hyperhet SNPs and for different regions e.g. West Africa vs Southeast Asia.

5.3 The effect of two transmission chains crossing, i.e. an episode of superinfection.

Imagine an episode of superinfection in which a host acquires infection from two sources, host A and host B. Let the Q alleles acquired from host A have heterozygosity H'_A, and let the alleles acquired from host B have heterozygosity H'_B.

Note the way that we have framed the problem. H'_A is not exactly the same as the heterozygosity of parasites within host A, as it allows for genetic drift and mutation that have occurred in the process of transmission from host A to the superinfected host. Here we are imagining that we have already taken account of equation 16 and this is factored into H'_A and H'_B.

We are left with the question of what is the overall level of within-host heterozygosity when we combine Q alleles from A with Q alleles from B?

We approach this by randomly sampling a pair of alleles from the superinfected host and asking if they are homozygous. We have already sampled with replacement from the previous generation, and now we are sampling without replacement from the $2Q$ alleles acquired by the superinfected host.

There are three ways in which two alleles from the superinfected host could be homozygous:

1. they are both acquired from host A ($\Pr = (Q-1)/(2(2Q-1))$ and they are homozygous ($\Pr = 1 - H'_A$)

2. they are both acquired from host B (Pr $= (Q-1)/(2(2Q-1))$ and they are homozygous (Pr $= 1 - H'_B$)

3. they are acquired from different hosts (Pr $= Q/(2Q-1)$ and they are homozygous (Pr $= 1 - H_S$)

By substitution and rearrangement this gives us the within-host heterozygosity H_W of the superinfected host:

$$H_W = \frac{(Q-1)(H'_A + H'_B) + 2QH_S}{2(2Q-1)} + \frac{QH_S}{2Q-1} \tag{20}$$

Thus superinfection acts to boost within-host heterozygosity because H_S will generally be much greater than either H'_A or H'_B.

5.4 The relationship between H_W and H_S in a population with $\chi \geq 0$

We now consider the more general case of a population in which superinfection may or may not occur, i.e. $\chi \geq 0$.

Imagine that we are following a transmission chain that crosses with other transmission chains with a probability of χ per generation. Let $\mathbf{X}$ be a random variable representing the number of generations that separate two crossing events on this transmission chain:

$$Pr\{\mathbf{X} = i\} = \chi(1-\chi)^{i-1} \tag{21}$$

.

Each crossing event causes within-host heterozygosity to rise abruptly to a peak, and then genetic drift causes it to decline gradually to a trough before it is boosted by another crossing event. These peaks and troughs will vary in magnitude according to the number of generations that separate crossing events and other factors. Let $\mathbf{H}$ and $\mathbf{H'}$ be random variables representing the peaks and troughs, respectively, of within-host heterozygosity along our transmission chain. We can think of $\mathbf{H'}$ and $\mathbf{H}$ as the states of our transmission chain immediately before and after a crossing event has occurred in a superinfected host, analogous to H'_A and H_W in equation 20.

Select any crossing event and follow the transmission chain to the next crossing event which occurs $\mathbf{X}$ generations later. Genetic drift causes heterozygosity to decline from $\mathbf{H}$ immediately after the first crossing event to $\mathbf{H'}$ immediately before the next crossing event.

$$\mathbf{H'} \approx \mathbf{H}\alpha^{\mathbf{X}} \tag{22}$$

This is essentially a truncated version of equations 17 and 18 that ignores the accumulation of new mutations. The approximation is justifiable in these circumstances, because mutation will generally have a much smaller effect than drift if there is a significant level of superinfection.

At each crossing event, our transmission chain crosses with another transmission chain which is assumed to be independent but to have the same probability distributions for $\mathbf{H}$ and $\mathbf{H'}$. We use equation 20 to estimate ΔH, the increase in within-host heterozygosity of our transmission chain that occurs as a result of a crossing event:

$$\Delta H = \frac{Q(H_S - \mathbf{H'})}{2Q-1} \tag{23}$$

For the system to be in equilibrium, the expected value of ΔH must equal the difference between the expected values of $\mathbf{H}$ and $\mathbf{H'}$, i.e. $\Delta H = E\{\mathbf{H}\} - E\{\mathbf{H'}\}$. By combining this with equations 22 and 23, we obtain

$$E\{\mathbf{H}\} - E\{\mathbf{H}\alpha^{\mathbf{X}}\} \approx \frac{QH_S}{2Q-1} - \frac{Q.E\{\mathbf{H}\alpha^{\mathbf{X}}\}}{2Q-1} \tag{24}$$

This equation contains two products of non-independent random variables, but here we make a substantial approximation by supposing that $\mathbf{H}$ and $\mathbf{X}$ are independent, allowing us to rearrange the equation:

$$E\{\mathbf{H}\} \approx E\Big\{\frac{QH_S}{2Q-(Q-1)\alpha^{\mathbf{X}}-1}\Big\}$$

From equation 21 we know that

$$E\{f(\mathbf{X})\} = \sum_{i=1}^{\infty} \chi(1-\chi)^{i-1} f(\mathbf{X}=i)$$

hence

$$E\{\mathbf{H}\} \approx H_S \sum_{i=1}^{\infty} \frac{Q\chi(1-\chi)^{i-1}}{2Q-(Q-1)\alpha^i-1}$$

Let H_W be the heterozygosity within a host that is sampled from some random point on our transmission chain, and let $\mathbf{S}$ be a random variable representing the number of generations between the time of sampling and the most recent crossing event.

$$Pr\{\mathbf{S}=i\} = \chi(1-\chi)^i$$

Note that the probability distribution of $\mathbf{S}$ is different from that of $\mathbf{X}$ in equation 21 because it is possible that $\mathbf{S}=0$, i.e. that we sample a host that is superinfected. The value of H_W will depend on how much genetic drift and mutation have occurred since the most recent crossing event. It is convenient to introduce mutation into the picture at this point by returning to equation 17 and using α as defined in equation 18

$$H_W = E\{\mathbf{H}\}\alpha^{\mathbf{S}} + 2u\sum_{i=0}^{\mathbf{S}-1}\alpha^i$$

In the special case of $\mathbf{S}=0$, i.e. if we sample a host that is superinfected, then $\alpha^{\mathbf{S}}=1$ and the right-hand term becomes an empty sum (from $i=0$ to -1). This gives the desired result of $H_W = E\{\mathbf{H}\}$.

In the special case of $\chi=0$, i.e. if transmission chains never cross, then there is a probability of 1 that $\mathbf{S}=\infty$. Thus $\alpha^{\mathbf{S}}=0$ (since $\alpha<1$) and the right hand term must equal $2u\sum_{i=0}^{\infty}\alpha^i$

By summing over the probability distribution of $\mathbf{S}$ we can obtain $\widehat{H}_W$, the mean value of within-host heterozygosity across our transmission chain:

$$\widehat{H}_W = E\{\mathbf{H}\}\sum_{j=0}^{\infty}\chi(1-\chi)^j\alpha^j + 2u\sum_{j=0}^{\infty}\sum_{k=0}^{j-1}\chi(1-\chi)^j\alpha^i \tag{25}$$

This gives us the interesting result

$$\widehat{H}_W \approx \kappa H_S + \lambda \tag{26}$$

where

$$\kappa = \sum_{i=1}^{\infty}\frac{Q\chi(1-\chi)^{i-1}}{2Q-(Q-1)\alpha^i-1} \times \sum_{j=0}^{\infty}\chi(1-\chi)^j\alpha^j$$

$$\lambda = 2u\sum_{j=0}^{\infty}\sum_{k=0}^{j-1}\chi(1-\chi)^j\alpha^k$$

For completeness let us compare equation 26 with our previous analysis of within-host heterozygosity in a population without superinfection. If $\chi=0$ then $\kappa=0$ and in evaluating λ we must treat this as a special case, in which the most recent crossing event happened an infinite number of generations ago, so that:

$$\widehat{H}_W = \lambda = 2u\sum_{k=0}^{\infty}\alpha^k = \frac{2uQ}{1+2uQ-2u}$$

This agrees with the expected value for within-host heterozygosity in a population without superinfection, that we previously derived in equation 19.

5.5 Comparing methods for analysing the relationship of F_{WS} to χ and Q.

We would like to know whether equation 26 gives the same result as Markov chain simulation in describing the relationship of F_{WS} to χ and Q. Figure 31 compares the two methods and confirms that they give extremely similar results when the effective number of hosts is large. The results deviate when the effective number of hosts is small and this is expected as the simplifying assumptions used to derive equation 13 become unreliable in these circumstances.

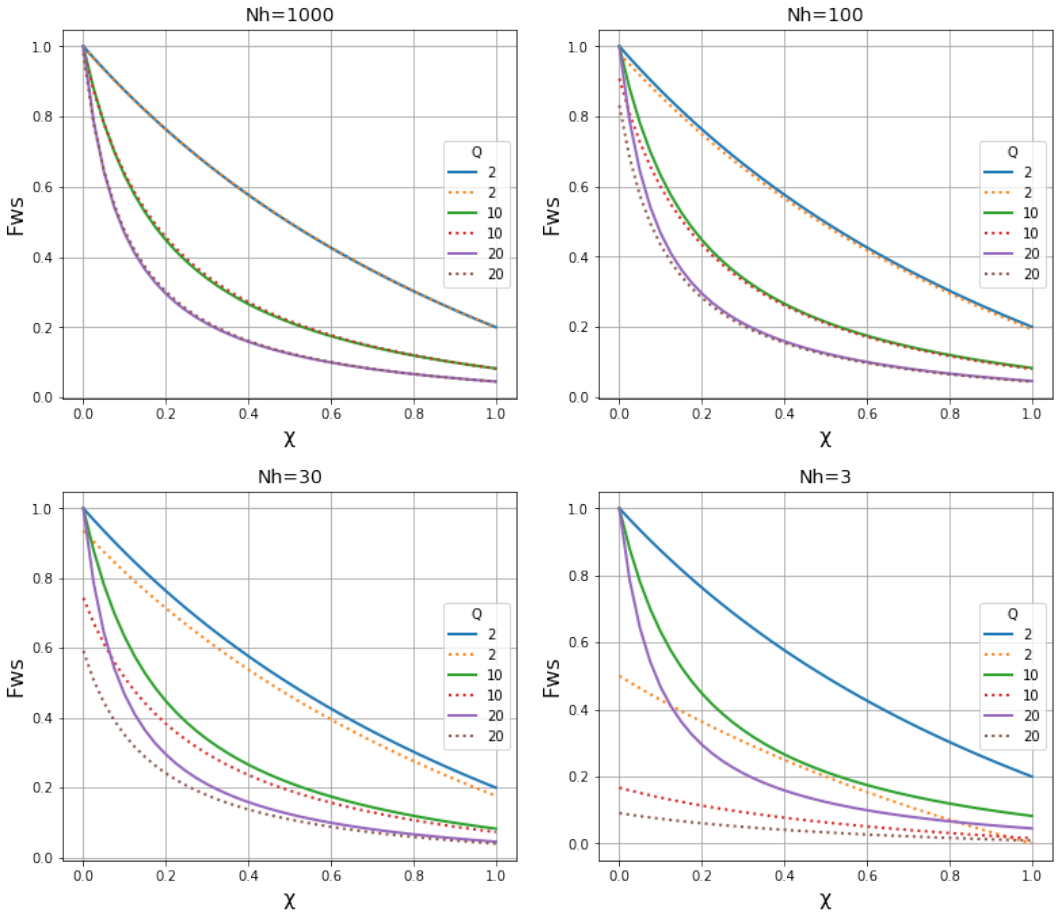

Figure 31: **The inbreeding coefficient F_{WS} is inversely related to χ.** Colours represent different values of Q. Solid lines show the results obtained from equation 26 and dotted lines show the results obtained by Markov chain simulation of coalescence times. When N_h is above 100 (top panels) the two methods give very similar results. When $N_h = 30$ (bottom left panel) equation 26 tends to overestimate F_{WS} at low values of χ as compared with the results obtained by Markov chain simulation. When $N_h = 3$ (bottom right panel) these differences are magnified and equation 26 is very unreliable. View code

5.6 Confounding of F_{WS} by local population structure.

Estimates of F_{WS} might be unreliable if there is a high degree of local population structure within the geographical area that we are sampling from, e.g. if we are sampling from a region with extremely mountainous or densely forested terrain, such that people and parasites rarely move between different villages. If we let H_R be the heterozygosity of the region that we are sampling from, and if H_S is the heterozygosity of a single village, and $\widehat{H}_W$ the mean of within-host heterozygosity in a village, then

$$F_{WS} = \frac{F_{WR} - F_{SR}}{1 - F_{SR}}$$

where $F_{WR} = 1 - \widehat{H}_W / H_R$ and $F_{SR} = 1 - H_S / H_R$.

If F_{SR} is small, then $F_{WS} \approx F_{WR}$ so it does not matter if we aggregate samples between villages. However if local subpopulations are highly differentiated from each other, i.e. if F_{SR} is large, then it is essential to use samples from a specific village when estimating F_{WS} because aggregating samples from different villages could cause a substantial overestimate.

This might explain the surprising observation that F_{WS} appears to be close to 1 in regions of Papua New Guinea where *P. falciparum* infection is at a very high prevalence [18, 32]. This could be due to confounding by local population structure, since these are mountainous jungle regions where the human population is divided into many small isolated communities, which might cause local parasite subpopulations to become highly differentiated from each other.