



Published in final edited form as:

Trends Parasitol. 2023 January ; 39(1): 17–25. doi:10.1016/j.pt.2022.11.002.

Measurably recombining malaria parasites

Flavia Camponovo^{1,†}, Caroline O Buckee¹, Aimee R Taylor^{2,†,*}

¹Harvard T.H. Chan School of Public Health, Boston, USA.

²Institut Pasteur, Paris, France.

Abstract

Genomic epidemiology has guided research and policy for various viral pathogens, and there has been a parallel effort towards using genomic epidemiology to combat diseases that are caused by eukaryotic pathogens, such as the malaria parasite. However, the central concept of viral genomic epidemiology, namely that of measurably mutating pathogens, does not apply easily to sexually recombining parasites. Here we introduce the related but different concept of measurably recombining malaria parasites to promote convergence around a unifying theoretical framework for malaria genomic epidemiology. Akin to viral phylodynamics, we anticipate that an inferential framework developed around recombination will help guide practical research, and thus realise the full public health potential of genomic epidemiology for malaria parasites and other sexually recombining pathogens.

Keywords

Plasmodium; Molecular Epidemiology; Recombination; Pedigree

A new genomic epidemiological concept

The public health value of malaria genomic epidemiology (see Glossary) has been demonstrated in several recent studies. A few examples include studies identifying local versus imported transmission in Bangladesh [1] and Southern Africa [2], tracking the rise and spread of drug resistance in the Greater Mekong Region [3], quantifying transmission changes in Senegal [4], or informing on the feasibility of malaria elimination in Sri Lanka [5]. The public health value of malaria genomic epidemiology is recognized beyond the research community, particularly in light of the recent COVID19 pandemic [6], and in

This work is licensed under a Creative Commons Attribution 4.0 International License, which allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

*Correspondence: ataylor@pasteur.fr (A.R. Taylor).

†Contributed equally

Declaration of Interests

The authors declare no competing interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the context of malaria elimination, where it is used to identify transmission hotspots and imported cases, for example [7, 8]. In order to capitalize on advances in data generation (e.g. [9]), and efforts by country-level stakeholders to build capacity and integrate genomic epidemiology into policy and practice [7], methodological advances are needed to make best use of parasite genetic data [10].

Genomic epidemiology relies on the concept of measurably evolving pathogens [11, 12]. A population can be said to evolve measurably if differences among DNA sequences, sampled at different points in time, are statistically significant [11]. If a pathogen population is measurably evolving on epidemiologically relevant timescales, genomic data sampled from infections can be used to measure and map different aspects of disease transmission [12]. For example, epidemiological timescales may be on an individual host level, between serial infections or symptom onset, or on a host population level, between groups of infected individuals separated in space or time.

The conventional definition of a measurably evolving pathogen assumes genetic differences are generated by mutation [11, 12]. Pathogen genomic epidemiology as a field developed around fast-mutating RNA viruses because these viruses mutate so rapidly that differences among them can be detected with limited genomic data, typical of the pre-genomic era [12]. Whole genome sequencing has since enabled genomic epidemiology of some more slowly-mutating pathogens [12]; see <https://nextstrain.org/pathogens>. In general, malaria parasites are not counted among them, partly because the coherent inferential framework that applies to fast-mutating RNA viruses (phylodynamics) does not apply readily to malaria parasites since they sexually recombine.

In this article we compare the genetic consequences of recombination versus mutation in the context of malaria genomic epidemiology, examine methodological gaps, and propose an approach towards an unifying inferential framework, something akin to phylodynamics in viral genomic epidemiology. Although we focus on *Plasmodium*, the concepts apply to a broader range of sexually recombining pathogens.

Malaria parasites mutate and sexually recombine

Both mutation and recombination generate genetic variation [12, 11]: mutation creates differences, while recombination creates new combinations of those differences. Mutational differences, δ , can be modelled simply as linear function of time t , the rate of mutation per locus per time μ , and the number of loci l : $\delta = \mu lt$ [12]. Sexual recombination also depends on some fixed parameters (crossover rate, number of loci, chromosomes and meioses); however, it is only “effective” when genetically distinct individuals recombine (effective recombination). Therefore, to model recombinational differences, one must also consider the external processes that bring individuals together (mating system), the amount of pre-existing variation among those individuals (population diversity) and how this variation is distributed (population structure). Malaria parasites are eukaryotes and mutate at a typical eukaryotic rate, which is slow compared to other pathogens. They recombine sexually every life-cycle, but the effectiveness of recombination can range from one, when completely unrelated parasites recombine, to zero, when clones recombine, a plausible event even in

diverse populations. Moreover, the effectiveness of recombination depends on the processes that unite genetically distinct parasites: coinfection and/or superinfection with genetically distinct parasites. In this section, we discuss how and when malaria parasites mutate measurably and recombine effectively on an epidemiologically relevant timescale. We also discuss the known and unknown aspects of the processes that shape effective recombination. We focus on *Plasmodium falciparum* and *Plasmodium vivax*, the two malaria parasite species most frequently responsible for human malaria [13].

Mutation

Malaria parasites are single-celled and, throughout the human stage of their lifecycle, haploid. Compared with viral pathogens, they have larger genomes but slower mutation rates: *P. falciparum* has a 23 megabase nuclear genome [14] and a SNP mutation rate on the order of 10^{-10} mutations per base pair per asexual generation (48 hours) [15].

Although this process generates many mutations - given the vast amount of parasites within a single malaria infection - the majority of those mutations occur singularly and are purged [9]. A well-defined core genome is often used for *P. falciparum* genomic epidemiology [16]. Among a population of infecting parasites, it accrues an estimated 0.84 ± 1.8 non-purged mutations per month [17]. That value increases to 2.92 ± 2.3 non-purged mutations per month (comparable to measurably mutating viruses) when advanced technologies are used to extend the accessible region of the genome [17]. Thus, with a generation interval of around 3 months for *P. falciparum* [18], it is theoretically possible to differentiate malaria parasites along a transmission chain using mutation. However, those mutations are only identifiable when parasites from different infections do not recombine, e.g. in near-elimination settings where transmission is extremely low and clonal propagation is extensive [17].

Recombination

While malaria parasites might accrue a small number of non-purged mutations over the course of one lifecycle, 50% of the genome is expected to differ if recombination with an unrelated parasite occurs. This means that recombination has greater potential to generate measurable variation on epidemiologically relevant timescales. This potential has been demonstrated by various studies (e.g. [19, 20, 21]), using either identity-by-descent (IBD) as a measure of recombinational relatedness or identity-by-state (IBS), a correlate of IBD [22].

Recombination is obligate in the malaria parasite lifecycle. Human blood stage parasites differentiate into gametocytes that are imbibed by the mosquito, where they differentiate promptly into gametes, and pair to sexually recombine approximately three hours after ingestion [23], each pair resulting in an oocyst, with usually <5 oocysts per mosquito [24]. The speed of fertilization impedes recombination between parasites from different blood meals and thus different people, unless a mosquito feeds on different people in very quick succession - a phenomena that likely does not contribute significantly to malaria epidemiology. *P. falciparum* gametes are estimated to crossover with probability 7.4×10^{-7} per base pair [16]. This implies, on average, 0.01 crossovers per 13500 base pairs and approximately one crossover per chromosome, of which *P. falciparum* has 14. This means that after recombination between unrelated parasites, we expect offspring to be 50% related

to their parents with, on average, one contiguous IBD segment per chromosome. Even without crossovers, sexual reproduction can generate variation because offspring inherit a random combination of their parental chromosomes.

Effective recombination

Although recombination is obligate, it is not always effective. Malaria parasites can self, i.e. genetically identical parasites can recombine, in which case recombination is ineffective. Selfing is inevitable when a mosquito feeds on a single monoclonal infection. Otherwise, selfing, inbreeding and/or outcrossing can occur, where inbreeding refers to partially effective recombination between related parasites, outbreeding refers to fully effective recombination between unrelated parasites, and the occurrence of one or more events depends on the number of parasite pairs that recombine. The extent of effectiveness depends principally on the composition of multiclonal human infections on which mosquitoes feeds (Figure 1A).

A multiclonal human infection can be generated in two, non-mutually exclusive ways: by a single mosquito bite transmitting genetically distinct parasites (cotransmission) and/or by several mosquito bites (superinfection), e.g. Figure 1A. Parasites from different mosquitoes cannot belong to the same brood. Within the mosquito, parasites can belong to the same brood, which can contain clones, strangers, and siblings [25, 26]. Inter-brood relatedness of parasites depends on the diversity and structure of the parasite population, and intra-brood relatedness depends on the relative occurrence of clones, strangers, and siblings within the brood and on the relatedness of the parental gametes. This means that the level of effective recombination between parasite genotypes depends on the relative frequency of both brood and non-brood mating between parasites, which, in turn, depends on cotransmission and superinfection between hosts (Figure 1A). That is to say, malaria parasites are not panmictic and the generation of diversity is linked to transmission intensity in a non-trivial way.

Both cotransmission and superinfection are expected to increase with transmission intensity, thereby increasing the overall prevalence of multiclonal infections (as has been observed inversely [27]) and the frequency of occurrence of effective recombination (Figure 1B). It is more difficult to predict how the effectiveness of recombination will be impacted by transmission intensity: given pre-existing variation, more infectious bites lead to a higher rate of superinfection and thus more opportunities for non-brood mating, which leads to outcrossing if the population is diverse and largely unstructured. However, more outcrossing leads to more mosquito-to-human cotransmission of outbred offspring that can consequently brood mate, which almost certainly leads to some inbreeding. As such, although superinfection and outcrossing both lead to effective recombination (Figure 1B), superinfection likely amplifies its effectiveness while mosquito-to-human cotransmission likely attenuates it (not shown).

Observations from field studies testify to the complexity of this system, using descriptive statistics of parasite genetic data as proxy indicators. Generally, high estimates of diversity and average multiplicity of infection (MOI) suggest high transmission, while evidence of prevalent clonal clusters and monoclonal infections suggest low transmission [28]. However, this relationship is sometimes unclear [29], especially in the presence of gene

flow [30, 5, 31, 32, 33, 34, 35]. Moreover, interpretation is hampered by extensive spatial heterogeneity [36], which is accentuated as transmission declines [37, 38], but does exist in high transmission [39, 40], though it is harder to detect [41]. Relapses add additional complexity for *P. vivax*, where MOIs can reflect present or past inoculations and thus are generally higher than those of *P. falciparum* [42].

For either species, what these processes collectively mean for the effectiveness of recombination is unclear: in low transmission settings, evidence of high *P. vivax* population diversity and average MOI has been observed together with significant linkage disequilibrium (*LD*, indicative of low effective recombination) [5, 37]; while in similarly low transmission settings, evidence of low *P. falciparum* population diversity and prevalent monoclonal infections has been observed together with low *LD* [43]. In high transmission settings, evidence of *P. falciparum* inbreeding persists [44, 45], consistent with the expected effect of brood-mating, and in both low and high transmission settings, *P. falciparum* multiclonal infections contain highly related parasites [46, 26, 47].

To summarise, effective recombination has greater potential than mutation to generate variation that is measurable on an epidemiological scale, but, unlike mutation, its effectiveness is inextricably linked to the epidemiological context in a complicated way (Figure 1B). Although some models of the mosquito stage of this highly complex system exist [48, 25, 26], its entirety is not understood well enough to translate into a functional form (see Outstanding Questions).

Malaria genomic epidemiology at present

For the practical application of pathogen genomic epidemiology, data should be used to infer parameters of epidemiological interest under a cohesive statistical model that links the processes that generate the genetic data to epidemiological ones (e.g. in viral genomic epidemiology, phylogenetic models are linked to coalescent or birth-deaths models in a framework called phylodynamics [49, 50, 51, 52]). Under a statistical model, interpretation is straightforward (phenomena of interest can be expressed explicitly as parameters and their dependence on hypothesised predictors evaluated [53]), as is prospective study design (e.g. using posterior predictive simulation or by maximizing the Fisher information of parameters of interest, as in [54]). Various steps build up to this model (Figure 2). Typically, malaria genomic epidemiological projects culminate in hypotheses generated by descriptive analyses (step three of Figure 2) because a cohesive inferential framework is lacking.

Descriptive analyses in malaria genomic epidemiology are related to those across malaria genomics more generally (see Box 1). They generate valuable hypotheses but they are also liable to generate some spurious associations. Moreover, descriptive analyses cannot provide conclusive answers to the questions malaria genomic epidemiology ultimately seeks to answer [55]. For example, a clustering analysis might reveal population structure that suggests gene flow to a region is restricted [56] and thus that the region is a suitable candidate for targeted intervention, but without a model under which this hypothesis can be falsified, one cannot reject competing processes, such as drug selection.

Because of recombination, phylodynamic frameworks cannot be applied directly to malaria (phylodynamic methods that accommodate recombination treat it as noise and not signal [12]) and an equivalent framework for malaria is lacking. However, efforts to develop simulation-based models are ongoing (e.g. the R package SIMPLEGEN, <https://mrc-ide.github.io/SIMPLEGEN/>). Agent-based models linked to genomic processes have been used to estimate R_0 and changes in transmission intensity [57, 4], to investigate the relationship between different descriptive statistics of parasite genetic data and transmission intensity [58, 55], to study the effect of heterogeneity on the spatial distribution of multiclonal infections and on the stability of transmission [38], and to study the effect of selective pressures on evolution under different transmission settings [59]. These models are used to simulate data under arbitrarily complex scenarios whose parameters are known. They are thus very versatile, but at a cost: in general, they are too complex for full statistical inference. However, they can be calibrated by comparing model predictions to real data, and then used to design prospective studies.

The future of malaria genomic epidemiology

The ultimate unifying inferential framework for malaria genomic epidemiology would centre around ancestral recombination graphs (ARGs), in the same way phylodynamics centers around phylogenetic trees. In the case of malaria, the mating system would link host-level parameters of epidemiological interest to the parasite genomes that feature in the ARG.

An ARG is a graph that links DNA sequences by both mutation and effective recombination. It can also be viewed as a sequence of phylogenetic trees, one tree for each locus along the genome, where trees from one loci to the next are transformed if effective recombination events occur between the loci [60]. It is a summary of all the coalescence and effective recombination events in the genealogical history of a set of nucleotide sequences and thus very powerful [60, 61]. Moreover, if inferred under a framework that includes an epidemiological model whose parameters can be expressed as a function of parasite ancestry, an inferred ARG leads to estimates of epidemiological interest (Figure 3). ARG-based malaria genomic epidemiology does not exist yet in large part because ARG inference is expensive both computationally and operationally [60, 62, 63], but human population genetic studies are advancing ARG inferential methods [61]. ARG-based methods require adjustment for malaria populations, specifically to account for dynamic rates of selfing, inbreeding and outbreeding. In particular, an advanced ARG-based model is needed since the effectiveness of recombination is an emergent property of the parasite mating system, and a fixed estimate or average rate (as for selfing in [64]) would not represent underlying transmission, which is ultimately the target of inference.

The locations of effective recombination events in a malaria parasite ARG and thus the effectiveness of recombination are governed by the parasite mating system: branches are sampled uniformly at random among branches allocated to different broods when non-brood mating occurs, branches are sampled among those allocated to the same brood when brood mating occurs; otherwise, when recombination is not effective, branches are propagated from one generation to the next. How to model the malaria parasite mating

system is a difficult open question but models from population ecology provide some inspiration. The use of IBD, though relatively new to malaria genomic epidemiology, is not new to population ecological studies of eukaryotes, where sexual recombination is the primary source of genomic variation. For example, Close Kin Mark Recapture, a method recently developed to estimate time-series of adult population size and survival of fish or other species [54], such as mosquitoes [65], defines priors for kinship probabilities using demographic models with parameters such as the adult population size, birth rate, and individual survival probability, while accounting for possible covariates including date and location of capture. Theoretically, this framework could be adapted to malaria, where a transmission model would replace the demographic model with epidemiological covariates (e.g. case-specific characteristics) that modify the probability of kinship among malaria parasites.

Concluding Remarks

Malaria genomic epidemiology is an exciting field of research which has proven useful for informing malaria surveillance and in which interest is growing among malaria control programs and policy makers. However, it is largely dominated by descriptive genomic data analysis that are disconnected from routine epidemiological analyses and are frequently retrospective. Although insightful, these types of analyses lack a clear common framework and are limited to speculative interpretation of the underlying transmission dynamics. We introduce the concept of measurably recombining malaria parasites in the hope that it will encourage development around a unifying inferential framework under which models can be developed and thus used for hypothesis-driven analyses and statistically robust prospective study design. This is by no means an easy task (see Outstanding Questions) but ongoing progress towards it will advance malaria genomic epidemiology thereby helping to promote its full potential for public health.

Acknowledgements

F.C. and C.O.B. are supported by a Maximizing Investigators' Research Award for Early-Stage Investigators (R35 GM-124715). A.R.T is supported by the Vivax Serology Partnership (VISPA) funded by the Bill & Melinda Gates Foundation.

Glossary

ancestral recombination graphs

Graphs that link DNA sequences by both mutation and effective recombination.

brood

Used herein to refer to a collection of parasites produced when one or more oocysts hatch collectively in a mosquito. Note that because of the speed of fertilisation, parasites ingested from different humans in a superinfected mosquito likely do not have an opportunity to mate and likely have staggered hatchings. As such parasites from a superinfected mosquito can either belong to the same or different broods.

brood and non-brood mating

Used herein to refer to mating between parasites from the same and different broods, respectively. Brood mating is comparable to non-random mating in population genetics more generally, since mates are not sampled uniformly from the population at large. We avoid the term non-random mating, however, because it could be misconstrued: when parasites brood mate, they are sampled randomly, but from the brood.

cotransmission

The transmission of genetically distinct parasites from mosquito to human or vice versa upon a single mosquito bite.

effective recombination

Used herein to refer to recombination between genetically distinct individuals. The effectiveness of recombination can range from low (inbreeding - recombination between genetically distinct but related individuals) to high (outcrossing - recombination between genetically distinct and unrelated individuals). When genetically identical individuals recombine (selfing), recombination is ineffective.

generation interval

The time between infection onset in consecutive human hosts in the transmission chain.

genomic epidemiology

Using genomics to study disease determinants in epidemiology. Used herein to refer to the use of pathogen genomic data to track pathogen populations in space and time for public health purposes, as opposed to the study of human genetic determinants of noninfectious diseases.

identity-by-descent

IBD; two alleles are identical-by-descent (IBD) if they are both copies of an ancestral allele; a chromosomal segment is IBD if it is descended intact (unbroken by recombination) from a common ancestor.

identity-by-state

IBS; two alleles are identical-by-state (IBS) if they are biochemically alike, e.g. both adenine, regardless of whether they are identical-by-descent or identical-by-chance.

multiplicity of infection

MOI; also referred to as complexity of infection, COI; the number of genetically distinct parasite genotypes within an infection, where genotype is used here to refer to a specific example of the malaria parasite genome.

phylodynamics

Models linking phylogenetic and epidemiological processes in order to estimate epidemiologically relevant parameters.

phylogenetic models

Also referred to as phylogenomics. The study of phylogeny, which is the evolutionary history between groups of organisms according to a taxon rank (such as genus, species, or strains). Molecular phylogeny uses nucleotide sequences to reconstruct phylogenetic trees.

superinfection

The transmission of parasites to an already infected human or mosquito, via infectious bites from multiple mosquitoes to one human (or via a single mosquito biting multiple infectious people, respectively).

References

- [1]. Chang H-H, Wesolowski A, Sinha I, Jacob CG, Mahmud A, Uddin D, Zaman SI, Hossain MA, Faiz MA, Ghose A, Sayeed AA, Rahman MR, Islam A, Karim MJ, Rezwana MK, Shamsuzzaman AKM, Jhora ST, Aktaruzzaman MM, Drury E, Gonçalves S, Kekre M, Dhorda M, Vongprommek R, Miotto O, Engø-Monsen K, Kwiatkowski D, Maude RJ, and Buckee C.(2019). Mapping imported malaria in Bangladesh using parasite genetic and human mobility data. *eLife*, 8.
- [2]. Tessema S, Wesolowski A, Chen A, Murphy M, Wilhelm J, Mupiri A-R, Ruktanonchai NW, Alegana VA, Tatem AJ, Tambo M, Didier B, Cohen JM, Bennett A, Sturrock HJ, Gosling R, Hsiang MS, Smith DL, Mumbengegwi DR, Smith JL, and Greenhouse B.(2019). Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *Elife*, 8.
- [3]. Imwong M, Suwannasin K, Kunasol C, Sutawong K, Mayxay M, Rekol H, Smithuis FM, Hlaing TM, Tun KM, van der Pluijm RW, Tripura R, Miotto O, Menard D, Dhorda M, Day NPJ, White NJ, and Dondorp AM (2017). The spread of artemisinin-resistant *Plasmodium falciparum* in the greater mekong subregion: a molecular epidemiology observational study. *Lancet Infect Dis*, 17(5):491–497. [PubMed: 28161569]
- [4]. Lee A, Ndiaye YD, Badiane A, Deme A, Daniels RF, Schaffner SF, Fall FB, Ndiop M, Gueye AB, Diallo I, Battle KE, Wenger EA, Bever CA, Sene D, MacInnis B, Wirth DF, Ndiaye D, Hartl DL, Volkman SK, and Proctor JL (2021). Modeling the levels, trends, and connectivity of malaria transmission using genomic data from a health facility in thiès, senegal. *medRxiv*.
- [5]. Gunawardena S, Ferreira M, Kapilananda G, Wirth D, and Karunaweera N.(2014). The Sri Lankan paradox: High genetic diversity in *Plasmodium vivax* populations despite decreasing levels of malaria transmission. *Parasitology*, 141:1–11. [PubMed: 24401336]
- [6]. Inzaule SC, Tessema SK, Kebede Y, Ouma AEO, and Nkengasong JN (2021). Genomic-informed pathogen surveillance in Africa: opportunities and challenges. *The Lancet Infectious Diseases*, 21(9):e281–e289. [PubMed: 33587898]
- [7]. Noviyanti R, Miotto O, Barry A, Marfurt J, Siegel S, Thuy-Nhien N, Quang HH, Anggraeni ND, Laihah F, Liu Y, et al. (2020). Implementing parasite genotyping into national surveillance frameworks: feedback from control programmes and researchers in the asia–pacific region. *MalariaJournal*.
- [8]. Dalmat R, Naughton B, Kwan-Gett TS, Slyker J, and Stuckey EM (2019). Use cases for genetic epidemiology in malaria elimination. *Malaria Journal*, 18(1):163. [PubMed: 31064369]
- [9]. Dia A.and Cheeseman IH (2021). Single-cell genome sequencing of protozoan parasites. *Trends in Parasitology*, 37(9):803–814. [PubMed: 34172399]
- [10]. Neafsey DE, Taylor AR, and MacInnis BL (2021). Advances and opportunities in malaria population genomics. *Nature Reviews Genetics*, 22(8):502–517.
- [11]. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, and Rodrigo AG (2003). Measurably evolving populations. *Trends in ecology & evolution*, 18(9):481–488.
- [12]. Biek R, Pybus OG, Lloyd-Smith JO, and Didelot X.(2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, 30(6):306–313. [PubMed: 25887947]
- [13]. WHO et al. (2021). World malaria report 2021.
- [14]. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, 419(6906):498–511. [PubMed: 12368864]
- [15]. Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, and Kwiatkowski D.(2017). Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic acids research*, 45(4):1889–1901. [PubMed: 27994033]

- [16]. Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, et al. (2016). Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome research*, 26(9):1288–1299. [PubMed: 27531718]
- [17]. Redmond SN, MacInnis BM, Bopp S, Bei AK, Ndiaye D, Hartl DL, Wirth DF, Volkman SK, and Neafsey DE (2018). De novo mutations resolve disease transmission pathways in clonal malaria. *Molecular biology and evolution*, 35(7):1678–1689. [PubMed: 29722884]
- [18]. Huber JH, Johnston GL, Greenhouse B, Smith DL, and Perkins TA (2016). Quantitative, model-based estimates of variability in the generation and serial intervals of *Plasmodium falciparum* malaria. *Malaria Journal*, 15(1).
- [19]. Omedo I, Mogeni P, Bousema T, Rockett K, Amambua-Ngwa A, Oyier I, Stevenson JC, Baidjoe AY, De Villiers EP, Fegan G, et al. (2017). Micro-epidemiological structuring of *Plasmodium falciparum* parasite populations in regions with varying transmission intensities in Africa. *Wellcome open research*, 2.
- [20]. Taylor AR, Schaffner SF, Cerqueira GC, Nkhoma SC, Anderson TJC, Sriprawatt K, Phyo AP, Nosten F, Neafsey DE, and Buckee CO (2017). Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. *PLOS Genetics*, 13(10):e1007065.
- [21]. Briggs J, Kuchta A, Murphy M, Tessema S, Arinaitwe E, Rek J, Chen A, Nankabirwa JI, Drakeley C, Smith D, et al. (2021). Within-household clustering of genetically related *Plasmodium falciparum* infections in a moderate transmission area of Uganda. *Malaria journal*, 20(1):1–8. [PubMed: 33386070]
- [22]. Taylor AR, Jacob PE, Neafsey DE, and Buckee CO (2019). Estimating relatedness between malaria parasites. *Genetics*, 212(4):1337–1351. [PubMed: 31209105]
- [23]. Baton LA and Ranford-Cartwright LC (2005). Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito. *Trends in parasitology*, 21(12):573–580. [PubMed: 16236552]
- [24]. Rosenberg R. (2008). Malaria: some considerations regarding parasite productivity. *Trends in parasitology*, 24(11):487–491. [PubMed: 18805735]
- [25]. Wong W, Wenger EA, Hartl DL, and Wirth DF (2018). Modeling the genetic relatedness of *Plasmodium falciparum* parasites following meiotic recombination and cotransmission. *PLOS Computational Biology*, 14(1):e1005923.
- [26]. Zhu SJ, Hendry JA, Almagro-Garcia J, Pearson RD, Amato R, Miles A, Weiss DJ, Lucas TC, Nguyen M, Gething PW, et al. (2019). The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria. *Elife*, 8:e40845.
- [27]. Nkhoma SC, Nair S, Al-Saai S, Ashley E, McGready R, Phyo AP, Nosten F, and Anderson TJ (2013). Population genetic correlates of declining transmission in a human pathogen. *Molecular ecology*, 22(2):273–285. [PubMed: 23121253]
- [28]. Volkman SK, Neafsey DE, Schaffner SF, Park DJ, and Wirth DF (2012). Harnessing genomics and genome biology to understand malaria biology. *Nature Reviews Genetics*, 13(5):315–328.
- [29]. Razakandrainibe FG, Durand P, Koella JC, Meeus TD, Rousset F, Ayala FJ, and Renaud F. (2005). “Clonal” population structure of the malaria agent *Plasmodium falciparum* in high-infection regions. *Proceedings of the National Academy of Sciences*, 102(48):17388–17393.
- [30]. Roh ME, Tessema SK, Murphy M, Nhlabathi N, Mkhonta N, Vilakati S, Ntshalintshali N, Saini M, Maphalala G, Chen A, Wilhelm J, Prach L, Gosling R, Kunene S, Hsiang MS, and Greenhouse B. (2019). High genetic diversity of *Plasmodium falciparum* in the low-transmission setting of the kingdom of Eswatini. *The Journal of infectious diseases*, 220(8):1346–1354. [PubMed: 31190073]
- [31]. Arnott A, Barry AE, Senn N, Barnadas C, Reeder JC, Mueller I, and Siba P. (2013). High genetic diversity of *Plasmodium vivax* on the north coast of Papua New Guinea. *The American journal of tropical medicine and hygiene*, 89(1):188–194. [PubMed: 23690553]
- [32]. Ferreira MU, Karunaweera ND, da Silva-Nunes M, da Silva NS, Wirth DF, and Hartl DL (2007). Population structure and transmission dynamics of *Plasmodium vivax* in rural Amazonia. *The Journal of infectious diseases*, 195(8):1218–1226. [PubMed: 17357061]

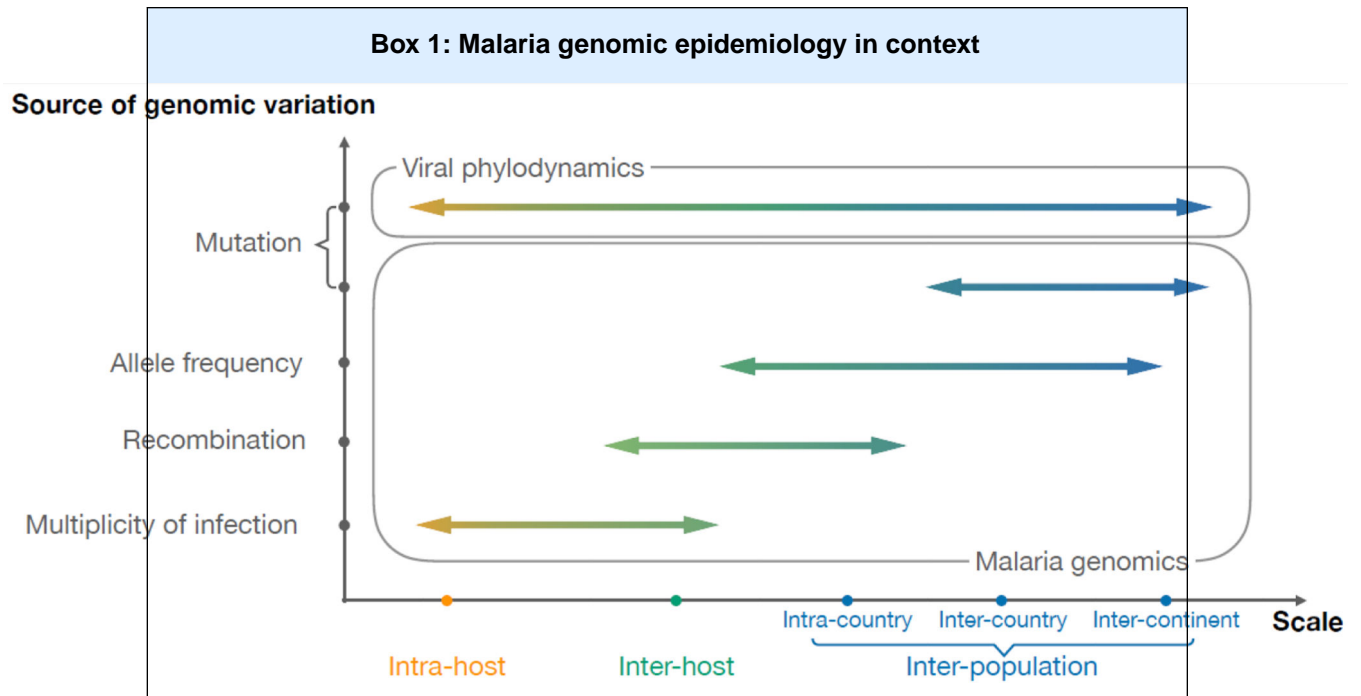
- [33]. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, Ouedraogo J-B, Michon P, Mueller I, Siba P, Nzila A, Borrmann S, Kiara SM, Marsh K, Jiang H, Su X-Z, Amaratunga C, Fairhurst R, Socheat D, Nosten F, Imwong M, White NJ, Sanders M, Anastasi E, Alcock D, Drury E, Oyola S, Quail MA, Turner DJ, Ruano-Rubio V, Jyothi D, Amenga-Etego L, Hubbart C, Jeffreys A, Rowlands K, Sutherland C, Roper C, Mangano V, Modiano D, Tan JC, Ferdig MT, Amambua-Ngwa A, Conway DJ, Takala-Harrison S, Plowe CV, Rayner JC, Rockett KA, Clark TG, Newbold CI, Berriman M, MacInnis B, and Kwiatkowski DP (2012). Analysis of plasmodium falciparum diversity in natural infections by deep sequencing. *Nature*, 487(7407):375–379. [PubMed: 22722859]
- [34]. Imwong M, Nair S, Pukrittayakamee S, Sudimack D, Williams JT, Mayxay M, Newton PN, Kim JR, Nandy A, Osorio L, Carlton JM, White NJ, Day NP, and Anderson TJ (2007). Contrasting genetic structure in *Plasmodium vivax* populations from asia and south america. *International journal for parasitology*, 37(8–9):1013–1022. [PubMed: 17442318]
- [35]. Chenet SM, Schneider KA, Villegas L, and Escalante AA (2012). Local population structure of plasmodium: impact on malaria control and elimination. *Malaria journal*, 11(1).
- [36]. Bejon P, Williams TN, Nyundo C, Hay SI, Benz D, Gething PW, Otiende M, Peshu J, Bashraheil M, Greenhouse B, et al. (2014). A micro-epidemiological analysis of febrile malaria in coastal kenya showing hotspots within hotspots. *elife*, 3.
- [37]. Waltmann A, Koepfli C, Tessier N, Karl S, Fola A, Darcy AW, Wini L, Harrison GA, Barnadas C, Jennison C, et al. (2018). Increasingly inbred and fragmented populations of plasmodium vivax associated with the eastward decline in malaria transmission across the southwest pacific. *PLoS neglected tropical diseases*, 12(1):e0006146.
- [38]. Karl S, White MT, Milne GJ, Gurarie D, Hay SI, Barry AE, Felger I, and Mueller I. (2016). Spatial effects on the multiplicity of plasmodium falciparum infections. *PLoS One*, 11(10):e0164054.
- [39]. Cattani J, Tulloch J, Vrbova H, Jolley D, Gibson F, Moir J, Heywood P, Alpers M, Stevenson A, and Clancy R. (1986). The epidemiology of malaria in a population surrounding madang, papua new guinea. *The American journal of tropical medicine and hygiene*, 35(1):3–15. [PubMed: 3511748]
- [40]. Forsyth KP, Anders RF, Cattani JA, and Alpers MP (1989). Small area variation in prevalence of an s-antigen serotype of plasmodium falciparum in villages of madang, papua new guinea. *The American journal of tropical medicine and hygiene*, 40(4):344–350. [PubMed: 2653060]
- [41]. Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, Ghani A, Drakeley C, and Gosling R. (2012). Hitting hotspots: spatial targeting of malaria for control and elimination. *PLoS medicine*, 9(1):e1001165.
- [42]. Fola AA, Harrison GLA, Hazairin MH, Barnadas C, Hetzel MW, Iga J, Siba PM, Mueller I, and Barry AE (2017). Higher complexity of infection and genetic diversity of plasmodium vivax than plasmodium falciparum across all malaria transmission zones of papua new guinea. *Am. J. Trop. Med. Hyg.*, pages 16–0716.
- [43]. Echeverry DF, Nair S, Osorio L, Menon S, Murillo C, and Anderson TJ (2013). Long term persistence of clonal malaria parasite plasmodium falciparum lineages in the colombian pacific region. *BMC genetics*, 14(1):1–13. [PubMed: 23280002]
- [44]. Paul R, Packer M, Walmsley M, Lagog M, Ranford-Cartwright L, Paru R, and Day K. (1995). Mating patterns in malaria parasite populations of papua new guinea. *Science*, 269(5231):1709–1711. [PubMed: 7569897]
- [45]. Babiker HA, Ranford-Cartwright LC, and Walliker D. (1999). 3. genetic structure and dynamics of plasmodium falciparum infections in the kilombero region of tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 93:11–14. [PubMed: 10450420]
- [46]. Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, and Anderson TJ (2012). Close kinship within multiple-genotype malaria parasite infections. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738):2589–2598.
- [47]. Nkhoma SC, Trevino SG, Gorena KM, Nair S, Khoswe S, Jett C, Garcia R, Daniel B, Dia A, Terlouw DJ, et al. (2020). Co-transmission of related malaria parasite lineages shapes within-host parasite diversity. *Cell host & microbe*, 27(1):93–103. [PubMed: 31901523]

- [48]. Childs LM and Prosper OF (2017). Simulating within-vector generation of the malaria parasite diversity. *PloS one*, 12(5):e0177941.
- [49]. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, and Frost SDW (2009). Phylogenetics of Infectious Disease Epidemics. *Genetics*, 183(4):1421–1430. [PubMed: 19797047]
- [50]. Rodrigo AG, Shpaer EG, Delwart EL, Iversen AKN, Gallo MV, Brojatsch J, Hirsch MS, Walker BD, and Mullins JI (1999). Coalescent estimates of HIV-1 generation time in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2187–2191. [PubMed: 10051616]
- [51]. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, and Harvey PH (2001). The Epidemic Behavior of the Hepatitis C Virus. *Science*, 292(5525):2323–2325. Publisher: American Association for the Advancement of Science. [PubMed: 11423661]
- [52]. Stadler T.(2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404. [PubMed: 20851708]
- [53]. Borchers DL and Efford MG (2008). Spatially explicit maximum likelihood methods for capture–recapture studies. *Biometrics*, 64(2):377–385. [PubMed: 17970815]
- [54]. Bravington MV, Skaug HJ, Anderson EC, et al. (2016). Close-kin mark-recapture. *Statistical Science*, 31(2):259–274.
- [55]. Watson OJ, Okell LC, Hellewell J, Slater HC, Unwin HJT, Omedo I, Bejon P, Snow RW, Noor AM, Rockett K, Hubbart C, Nankabirwa JI, Greenhouse B, Chang H-H, Ghani AC, and Verity R.(2021). Evaluating the Performance of Malaria Genetics for Inferring Changes in Transmission Intensity Using Transmission Modeling. *Molecular Biology and Evolution*, 38(1):274–289. [PubMed: 32898225]
- [56]. Schultz L, Wapling J, Mueller I, Ntsuke PO, Senn N, Nale J, Kiniboro B, Buckee CO, Tavul L, Siba PM, Reeder JC, and Barry AE (2010). Multilocus haplotypes reveal variable levels of diversity and population structure of *Plasmodium falciparum* in Papua New Guinea, a region of intense perennial transmission. *Malaria Journal*, 9(1).
- [57]. Daniels RF, Schaffner SF, Wenger EA, Proctor JL, Chang H-H, Wong W, Baro N, Ndiaye D, Fall FB, Ndiop M, Ba M, Milner DA, Taylor TE, Neafsey DE, Volkman SK, Eckhoff PA, Hartl DL, and Wirth DF (2015). Modeling malaria genomics reveals transmission decline and rebound in Senegal. *Proceedings of the National Academy of Sciences*, 112(22):7067–7072.
- [58]. Hendry JA, Kwiatkowski D, and McVean G.(2021). Elucidating relationships between *P. falciparum* prevalence and measures of genetic diversity with a combined genetic-epidemiological model of malaria. *PLOS Computational Biology*, 17(8):e1009287.
- [59]. Cárdenas P, Corredor V, and Santos-Vega M.(2022). Genomic epidemiological models describe pathogen evolution across fitness valleys. *Science Advances*, 8(28):eabo0173.
- [60]. Arenas M.(2013). The importance and application of the ancestral recombination graph. *Frontiers in Genetics*, 4.
- [61]. Harris K.(2019). From a database of genomes to a forest of evolutionary trees. *Nature Genetics*, 51(9):1306–1307. [PubMed: 31477932]
- [62]. McGill JR, Walkup EA, and Kuhner MK (2013). Graphml specializations to codify ancestral recombinant graphs. *Frontiers in Genetics*, 4:146. [PubMed: 23967010]
- [63]. Rasmussen MD, Hubisz MJ, Gronau I, and Siepel A.(2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342.
- [64]. Nordborg M.(2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics*, 154(2):923–929. [PubMed: 10655241]
- [65]. Jasper M, Schmidt TL, Ahmad NW, Sinkins SP, and Hoffmann AA (2019). A genomic approach to inferring kinship reveals limited intergenerational dispersal in the yellow fever mosquito. *Molecular Ecology Resources*, 19(5):1254–1264. [PubMed: 31125998]
- [66]. Müller NF, Wüthrich D, Goldman N, Sailer N, Saalfrank C, Brunner M, Augustin N, Seth-Smith HM, Hollenstein Y, Syedbasha M, Lang D, Neher RA, Dubuis O, Naegele M, Buser A, Nickel CH, Ritz N, Zeller A, Lang BM, Hadfield J, Bedford T, Battegay M, Schneider-Sliwa R, Egli A, and Stadler T.(2020). Characterising the epidemic spread of influenza A/H3N2 within a city through phylogenetics. *PLOS Pathogens*, 16(11):e1008984. Publisher: Public Library of Science.

- [67]. Alizon S, Lion S, Murall CL, and Abbate JL (2014). Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence*, 5(8):825–827. Publisher: Taylor & Francis _eprint: 10.4161/21505594.2014.976514. [PubMed: 25495064]
- [68]. Attwood SW, Hill SC, Aanensen DM, Connor TR, and Pybus OG (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics*.
- [69]. Rougeron V, Boundenga L, Arnathau C, Durand P, Renaud F, and Prugnolle F.(2021). A population genetic perspective on the origin, spread and adaptation of the human malaria agents *Plasmodium falciparum* and *Plasmodium vivax*. *FEMS Microbiology Reviews*, 46(1).
- [70]. Brown TS, Arogbokun O, Buckee CO, and Chang H-H (2021). Distinguishing gene flow between malaria parasite populations. *PLoS genetics*, 17(12):e1009335.
- [71]. Taylor AR, Echeverry DF, Anderson TJC, Neafsey DE, and Buckee CO (2020). Identity-by-descent with uncertainty characterises connectivity of *Plasmodium falciparum* populations on the Colombian-Pacific coast. *PLOS Genetics*, 16(11):e1009101.
- [72]. Henden L, Lee S, Mueller I, Barry A, and Bahlo M.(2018). Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLoS genetics*, 14(5):e1007279.
- [73]. Watson JA, Taylor AR, Ashley EA, Dondorp A, Buckee CO, White NJ, and Holmes CC (2020). A cautionary note on the use of unsupervised machine learning algorithms to characterise malaria parasite population structure from genetic distance matrices. *PLoS genetics*, 16(10):e1009037.

Highlights

- The recent pandemic has further highlighted the public health potential of infectious disease genomic epidemiology.
- For viruses, epidemiological parameters can be estimated under powerful phylodynamic models using both epidemiological and genomic data jointly. An equivalent framework for malaria parasites is lacking because they recombine.
- Recombination between malaria parasites can generate epidemiologically relevant variation, but recombination is sometimes ineffective, depending dynamically on transmission. This makes it hard to model. It also means it could link epidemiological and genomic processes if they were modelled jointly.
- Given the potential of recombination, efforts to build a unifying inferential framework around the malaria parasite ancestral recombination graph (ARG) are merited. ARG-based genomic epidemiology could someday be an equivalent of phylodynamics.

**Figure I:**

Sources of genomic variation used in pathogen genomic analyses on different scales.

Pathogen genomic analyses use different sources of variation that cover different spatio-temporal scales, not all of which are suited to epidemiology. Figure I provides an overview, to which there are exceptions, e.g. [17].

Mutation is the main source of genomic variation used in viral phylodynamic studies, which range from intra-host to inter-continental scales. Examples of phylodynamic studies of RNA-viruses across different scales include a retrospective study of city-scale spatial spread of Influenza A/H3N2 [66]; the investigation of the 2014 Ebola outbreak in Sierra Leone [67]; and tracking the SARS-CoV-2 pandemic (reviewed in [68]).

Mutation is also the main source of genomic variation used in malaria phylogenetic studies, which often explore species origin on a large spatial scale [69]. These studies often use genomic data where the assumption of no recombination holds (e.g. genomic data from different parasite species, or DNA data on non-recombining mitochondria).

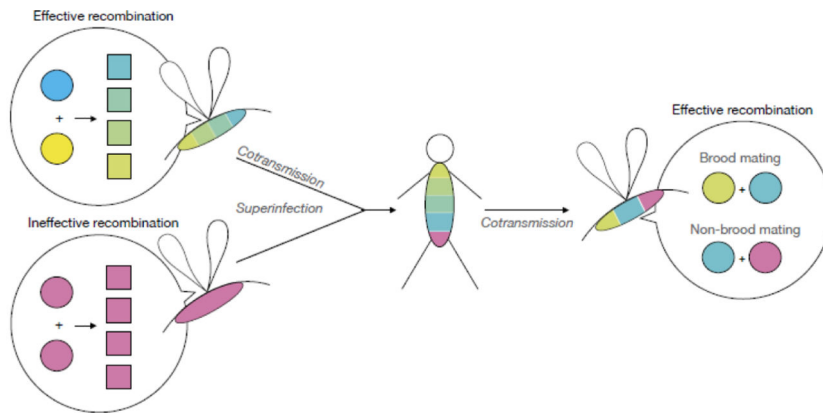
Analyses of allele frequencies within and across populations feature in many population genetic analyses of malaria parasites. Allele frequencies change with selection (e.g. from drug pressure), gene-flow (e.g. due to the mobility of infected hosts) or genetic drift (particularly in small populations), typically at a rate slower than that which is epidemiologically relevant on an individual level [22, 70]. That said, in small populations with limited effective recombination, allele frequencies may vary at a rate equivalent to recombination (e.g. [71]). In any case, descriptive statistics of allele frequency variation (population-level) provide less resolution than those of recombination-based metrics (individual-level).

Recombination-based metrics (e.g. summaries of IBD along the genome and relatedness, which averages over the genome) are popular when studying signals of selection, malaria parasite population connectivity and population structure (e.g. [20, 72, 2, 1]). To characterise population structure, IBD- and IBS-based similarity matrices are often input into clustering algorithms, which include tree-like algorithms but are not strictly phylogenetic models [73]. Unlike allele frequencies, which may or may not change over a short period of time, parasites always recombine between different human hosts. Therefore, recombination-based metrics have the potential to vary between individuals, providing recombination is effective.

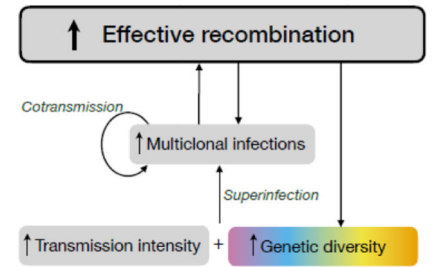
Multiplicities of infection vary between infected hosts. They do not require any change on the parasite level. Otherwise stated, variation on the parasite level might be fixed, but partitioned differently among hosts. As such, among all genetic metrics, estimates of the multiclonal infection prevalence and MOIs respond most quickly to changes in transmission [27, 55, 58]. They can also change between initial and recurrent infections within individuals.

Outstanding questions

1. How does transmission impact the effectiveness of malaria parasite recombination? And how do these processes translate into a model of the malaria parasite mating system? More specifically,
 - a. How does transmission translate into the frequency of occurrence of cotransmission and superinfection between hosts, and thus brood and non-brood mating between parasites.
 - b. For a given transmission setting at equilibrium, how does brood and non-brood mating between parasites translate into outcrossing, inbreeding, and selfing between parasite genotypes.
 - c. For a given transmission setting at equilibrium, what is a representative distribution of stranger, sibling, and clonal parasites within a brood.
 - d. For a given transmission setting at equilibrium, what are the representative distributions of relatedness and IBD between stranger, sibling, and clonal parasites within a brood, and between stranger parasites from different broods.
 - e. How do the underlying processes translate into measurable correlates of transmission (e.g. the prevalence of human infection and the entomological inoculation rate) and measurable correlates of effective recombination (e.g. measures of genetic diversity, linkage disequilibrium, and IBD).
2. Can statistical inference be improved for simulation models in malaria genomic epidemiology.
3. Can a computationally efficient joint inference framework be developed for ARG-based genomic epidemiology?
4. Could an ARG-based epidemiological framework be adapted to the malaria parasite mating system? Specifically, dynamic rates of selfing, inbreeding and outbreeding that link to epidemiological processes.



A) Multiclonally infected humans lead to effective recombination



B) Modelling effective recombination

Figure 1: Effective recombination.

(A) Human-to-mosquito cotransmission leads to effective recombination but its effectiveness depends on the type of mating and thus the processes that generated the multiclonal human infection (cotransmission and superinfection). Different colours represent genetically distinct parasites within infections (host fill) and during recombination (parental gametes, circles; offspring, squares). Brood mating following mosquito-to-human cotransmission is more likely to have low effectiveness due to probable inbreeding, whereas non-brood mating following superinfection is likely to have high effectiveness due to probable outbreeding if the population is largely outbred. (B) The frequency of occurrence of effective recombination is expected to increase with higher transmission intensity. Given genomic diversity in the population, greater transmission results in more superinfections, leading to more multiclonal infections in humans and mosquitoes, which in turn increase the opportunity for cotransmission. Multiclonal infections allow for effective recombination, which in turn gives rise to more multiclonal infections and greater genomic diversity. The effectiveness of recombination (not shown) depends on the routes via which multiclally infected mosquitoes are generated: more routes via superinfection will lead to more non-brood mating, while more routes via mosquito-to-human cotransmission will lead to more brood mating (panel A).

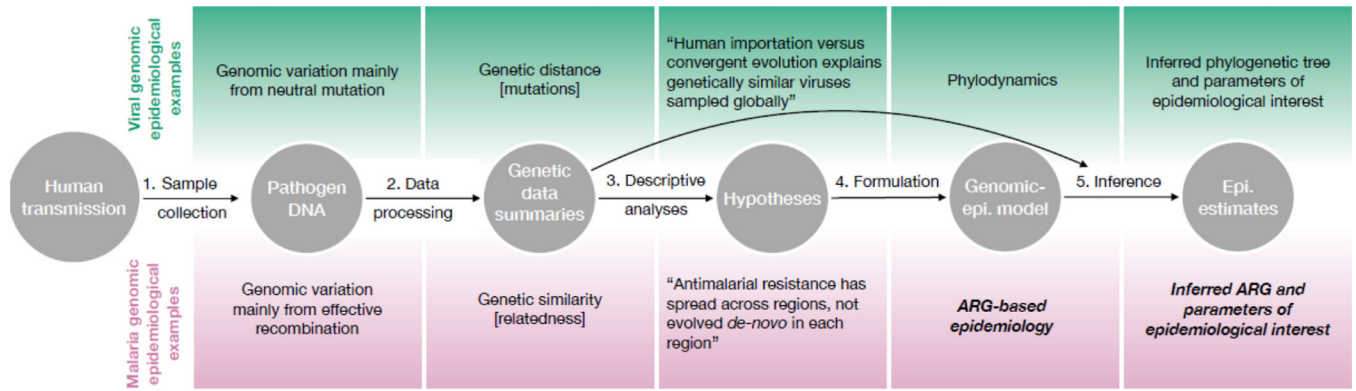


Figure 2: Possible series of genomic steps in pathogen genomic epidemiological studies. First, genomic data are collected. To be useful, those data must contain variation that has accumulated on an epidemiological relevant scale, e.g. due to mutation or due to recombination, processes that data summaries generated in step two often reflect. Step three involves a suite of descriptive analyses; for example, the computation of descriptive statistics of pathogen diversity and differentiation, population assignment and clustering analyses. A model that connects the genomic processes to epidemiological ones is formulated in the fourth step, typically using mathematics to articulate hypotheses concretely. This process incites clarification and thus is valuable in and of itself. An arrow connects data to the fifth step, because data are used to infer the parameters of the model. This step also links to epidemiological data; however, these are not shown. Above and below the steps, examples are provided for viral and malaria genomic epidemiology, respectively. Those that do not currently exist are highlighted in italic and bold.

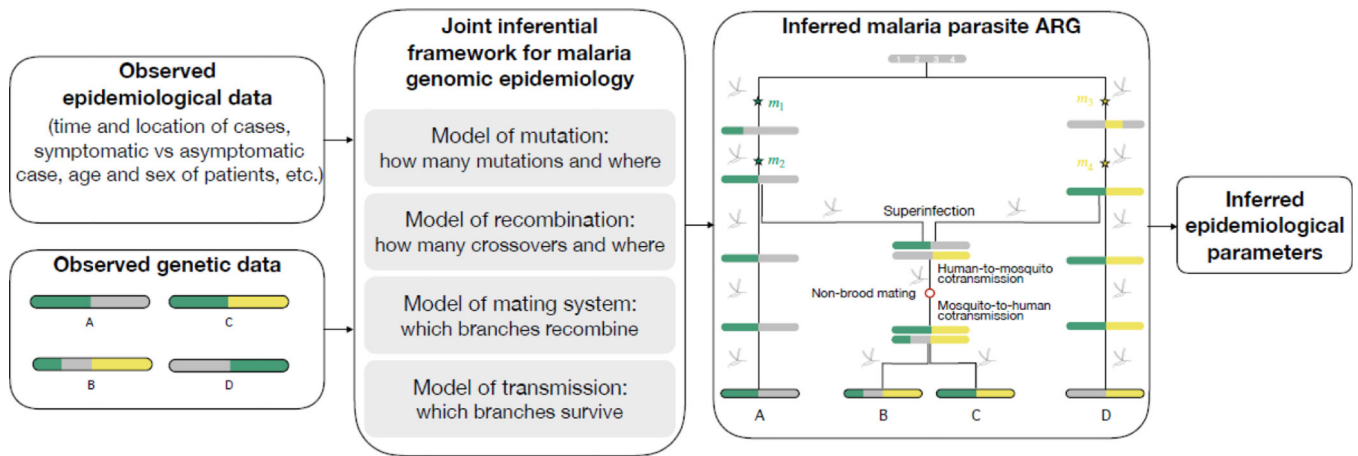


Figure 3: Overview of ARG-based malaria genomic epidemiology.

The goal of genomic epidemiology is to use observed genomic and epidemiological data (left) to infer epidemiologically relevant parameters (right). One way to do this would be to formulate a model around the malaria parasite ARG. In this example, four genetically distinct DNA segments (A, B, C, D) each of length equal to four loci (indexed by i) are linked back to a recent common ancestor (grey) via four per-locus mutations (m_i) and one effective recombination event (red circle) over five generations. The observed sequences are outlined in black whereas unobserved inferred sequences deeper within the ARG are not. This example only depicts sequences sampled from the human host and thus some sequences present in the mosquito are omitted; specifically, two of the four haploid meiotic products that were produced after effective recombination.