bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# The origins and relatedness structure of mixed infections vary with local prevalence of *P. falciparum* malaria

Sha Joe Zhu[1†], Jason A. Hendry[1†], Jacob Almagro-Garcia[1,2,3,4], Richard D. Pearson[2,3,4], Roberto Amato[2,3,4], Alistair Miles[1,2,3,4], Daniel J. Weiss[1], Tim C.D. Lucas[1], Michele Nguyen[1], Peter W. Gething[1], Dominic Kwiatkowski[1,2,3,4], Gil McVean[1,3*], and for the Pf3k Project[5]

[1]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford

[2]Wellcome Centre for Human Genetics, University of Oxford

[3]Medical Research Council Centre for Genomics and Global Health, University of Oxford

[4]Wellcome Sanger Institute

[5]See Supplementary Note

**Abstract**

Individuals infected with the *Plasmodium falciparum* malaria parasite can carry multiple strains with varying levels of relatedness. Yet, how parameters of local epidemiology and the biology of transmission affect the rate and relatedness of such mixed infections remains unclear. Here, we develop an enhanced method for strain deconvolution from genome sequencing data, which estimates the number of strains, their proportions, identity-by-descent (IBD) profiles and individual haplotypes. We validate the method through experimental and *in silico* simulations and apply it to the Pf3k data set, consisting of 2,344 field samples from 13 countries. We find that the rate of mixed infection varies from 18% to 63% across countries and that 51% of all mixed infections involve more than two strains. By modelling the structure of IBD resulting from different infection mechanisms we estimate that 55% of dual infections contain sibling strains likely to have been co-transmitted from a single mosquito, and find evidence of mixed infections propagated over successive infection cycles. By combining genetic data with epidemiological estimates of prevalence from the Malaria Atlas Project, we find that, at the country level, prevalence

[†]These authors contributed equally to this work

[*]For correspondence: gil.mcvean@bdi.ox.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlates with both the rate of mixed infection (Pearson $r = 0.65$, $P = 3.7 \times 10^{-6}$) and the level of IBD ($r = -0.51$, $P = 6.0 \times 10^{-4}$). Genomics is becoming a standard tool in pathogen surveillance. In this work, we conclude that monitoring fine-scale patterns of mixed infections and within-sample relatedness will be highly informative for assessing the impact of interventions and to inform malaria control programs.

***Key words:*** Malaria, genome, epidemiology, relatedness

# 1 Introduction

Individuals infected with malaria-causing parasites of the genus *Plasmodium* often carry multiple, distinct strains of the same species (Bell et al., 2006). Such mixed infections, also known as complex infections, are likely indicative of intense local exposure rates, being common in regions of Africa with high rates of prevalence (Howes et al., 2016). However, they have also been documented for *P. vivax* and other malaria-causing parasites (Ivo Mueller, 2007; Collins, 2012), even in regions of much lower prevalence (Howes et al., 2016; Steenkeste et al., 2010). Mixed infections have been associated with increased disease severity (de Roode et al., 2005) and also facilitate the generation of genomic diversity within the parasite, enabling co-transmission to the mosquito vector where sexual recombination occurs (Mzilahowa et al., 2007). Mixed infections are transient (Bruce and Day, 2002; Zimmerman et al., 2004), but little is known about the distribution of their duration. Whether the clearance of one or more strains results purely from host immunity (Borrmann and Matuschewski, 2011) or can be influenced by interactions between the distinct strains (Enosse et al., 2006; Bushman et al., 2016), are also open questions.

Although mixed infections can be studied from genetic barcodes (Galinsky et al., 2015) or single nucleotide polymorphisms (SNPs) (O'Brien et al., 2016), genome sequencing provides a more powerful approach for detecting mixed infections (Chang et al., 2017). Genetic differences between co-existing strains manifest as polymorphic loci in the DNA sequence of the isolate. The higher resolution of sequencing data allows the use of statistical methods for estimating the number of distinct strains, their relative proportions, and genome sequences (Zhu et al., 2018). Although genomic approaches cannot identify individuals infected multiple times by identical strains, and are affected by sequencing errors and problems of incomplete or erroneous reference assemblies, they provide a rich characterisation of within host diversity (Manske et al., 2012; Auburn et al., 2012; Pearson et al., 2016).

Previous research has highlighted that co-existing strains can be highly related (Nair et al., 2014; Trevino et al., 2017). For example, in *P. vivax*, 58% of mixed infections show long stretches of within host homozygosity (Pearson et al., 2016). In addition, Nkhoma et al. (2012) reported an average of 78.7% *P. falciparum* allele sharing in Malawi and 87.6% sharing in Thailand. Relatedness can arise through different mechanisms. First, a mosquito vector may acquire distinct strains from biting a single multiply-infected individual, in

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which case sexual reproduction and onward transmission can result in $F_2$ inbred progeny. A similar process may occur from biting multiple infected individuals. If these subsequently undergo sexual reproduction in the mosquito midgut, then transmission may result in an individual being infected with multiple, sibling strains with lower levels of inbreeding than in the previous case. Alternatively, relatedness can occur through independent infection events but in a population where genetic diversity is low, such as during the early stages of an outbreak or following severe population bottlenecks; for instance, those provoked by an intervention (Mouzin et al., 2010; Wong et al., 2017; Daniels et al., 2015).

The rate and relatedness structure of mixed infections are therefore highly relevant for understanding regional epidemiology. However, progress towards utilising this source of information is limited by three problems. Firstly, while strain deconvolution within mixed infections has received substantial attention (Galinsky et al., 2015; O'Brien et al., 2016; Chang et al., 2017; Zhu et al., 2018), currently, no methods perform joint deconvolution of strains and estimation of relatedness. Because existing deconvolution methods assume equal relatedness along the genome, differences in relatedness that occur, for example, through infection by sibling strains can lead to errors in the estimation of the number, proportions and sequences of individual strains (Figure 1). Recently, progress has been made in the case of dual-infections with balanced proportions (Henden et al., 2018), but a general solution is lacking. The second problem is that little is known about how the rate and relatedness structure of mixed infections relates to underlying epidemiological parameters. Informally, mixed infections will occur when prevalence is high; an observation exploited by Cerqueira et al. (2017) when estimating changes in transmission over time. However, the quantitative nature of this relationship, the key parameters that influence mixed infection rates and how patterns of relatedness relate to infection dynamics are largely unexplored.

Here, we develop, test and apply an enhanced method for strain deconvolution, which builds on our previously-published `DEploid` software. The method separates estimation of strain number, proportions, and relatedness (specifically the identity-by-descent, or IBD, profile along the genome) from the problem of inferring genome sequences. This strategy provides substantial improvements in accuracy under complex settings or when dealing with low coverage data. We apply the approach to 2,344 field isolates of *P. falciparum* collected from 13 countries over a range of years (2001-2014) and available through the Pf3k Project (see Supplementary Note), and characterise the rate and relatedness patterns of mixed infections. In addition, we develop a statistical framework for characterising the processes underlying mixed infections, estimating that more than half of mixed infections arise from the transmission of siblings, as well as demonstrating the propagation of mixed infections through cycles of host-vector transmission. Finally, we investigate the relationships between statistics of mixed infection and epidemiological estimates of pathogen prevalence (MAP, 2017), showing that country-level rates of mixed infection are highly correlated with estimates of malaria parasite prevalence.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 2 Strain deconvolution in the presence of relatedness

Existing methods for deconvolution of mixed infections typically assume that the different genetic strains present in mixed infections are unrelated. This assumption allows for efficient computation of priors for allele frequencies within samples, either through assuming independence of loci (O'Brien et al., 2016) or as sequences generated as imperfect mosaics of some (predefined) reference panel (Zhu et al., 2018). However, when strains are related to each other, and particularly when patterns of IBD vary along the genome (for example through being siblings, or sibs for short), the constraints imposed on within-sample allele frequencies through IBD can cause problems for deconvolution methods, which can try to fit complex strain combinations (with relatedness) as simpler configurations (without relatedness). Below we outline the approach we take to integrating IBD into `DEploid` Further details are provided in the Supplementary Materials.

## 2.1 Decoding genomic relatedness among strains

A common approach to detecting IBD between two genomes is to employ a hidden Markov Model that transitions into and out of IBD states (Chang et al., 2015; Gusev et al., 2009, 2011). We have generalised this approach to the case of $k$ haploid *Plasmodium* genomes (strains). In this setting, there are $2^k$ possible genotype configurations, as each of the $k$ strains can be either reference, i.e. same as the reference genome used during assembly, or alternative at a given locus (we assume all variation is bi-allelic). If each of the $k$ strains constitutes a unique proportion of the infection, each genotype configuration will produce a distinct alternative within sample allele frequency (WSAF; Figure 1A), which defines the expected fraction of total sequencing reads that are alternative at a given locus in the sequenced infection.

The effect of IBD among these $k$ strains is to limit the number of distinct genotype configurations possible, in a way that depends on the pattern of IBD sharing. Consider that, for any given locus, the $k$ strains in the infection are assigned to $j \leq k$ possible reference haplotypes. IBD exists when two or more strains are assigned to the same haplotype. In this scenario, the total number of possible patterns of IBD for a given $k$ is equal to $\sum_{j=1}^{k} S(k,j)$ where $S(k,j)$ is the number of ways $k$ objects can be split into $j$ subsets (a Stirling number of the second kind (Graham et al., 1988)). Thus, for two strains, there are two possible IBD states (IBD or non-IBD), for three strains there are five states (all IBD, none IBD and the three pairwise IBD configurations), for four strains there are fifteen states (see Supplementary Materials), and so on. We limit analysis to a maximum of four strains for computational efficiency and because higher levels of mixed infection are rarely observed. Finally, for a given IBD state, only $2^j$ rather than $2^k$ genotype configurations are possible, thereby restricting the set of possible WSAF values.

Moving along the genome, recombination can result in changes in IBD state, hence changing WSAF values at those loci (Figure 1B). To infer IBD states we use a hidden Markov model, which assumes linkage equilibrium between variants for computational efficiency, with a Gamma-Poisson emission model for read

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

counts (see Supplementary Materials). Population-level allele frequencies are estimated from isolates obtained from a similar geographic region. Given the structure of the hidden Markov model, we can compute the likelihood of the strain proportions by integrating over all possible IBD sharing patterns, yielding a Bayesian estimate for the number and proportions of strains (see Methods). We then use posterior decoding to infer the relatedness structure across the genome (Figure 1B). To quantify relatedness, we compute the mean IBD between pairs of strains, and statistics of IBD tract length (mean, median and N50, the length-weighted median IBD tract length, Figure 1C).

In contrast to our previous work, `DEploidIBD` infers strain structure in two steps. In the first we estimate the number and proportions of strains using Markov Chain Monte-Carlo (MCMC), allowing for IBD as described above. In the second, we infer the individual genomes of the strains, using the MCMC methodology of Zhu et al. (2018), which can account for linkage disequilibrium (LD) between variants, but without updating strain proportions. The choice of reference samples for deconvolution is described in Zhu et al. (2018) and in the Supplementary Materials. During this step we do not use the inferred IBD constraints *per se*, though the inferred haplotypes will typically copy from the same (or identical) members of the reference panel within the IBD tract.

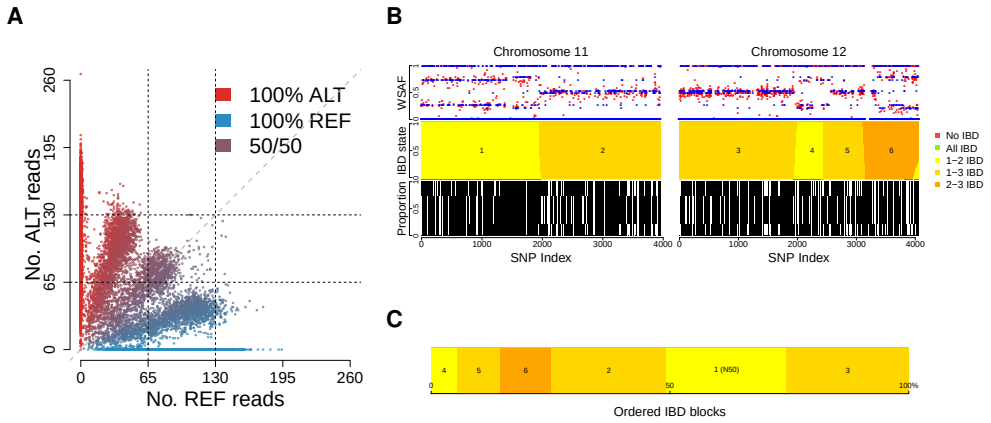


Figure 1: Deconvolution of a complex field sample PD0577-C from Thailand. (A) Scatter-plot showing the number of reads supporting the reference (REF: x-axis) and alternative (ALT: y-axis) alleles. The multiple clusters indicate the presence of multiple strains, but cannot distinguish the exact number or proportions. (B) The profile of within-sample allele frequency along chromosomes 11 and 12 (red points) suggests a changing profile of IBD with three distinct strains, estimated to be with proportions of 22%, 52% and 26% respectively (other chromosomes omitted for clarity, see Figure 1-Supplement 1); blue points indicate expected allele frequencies within the isolate. However, the strains are inferred to be siblings of each other: green segments indicate where all three strains are IBD; yellow, orange and dark orange segments indicate the regions where one pair of strains are IBD but the others are not. In no region are all three strains inferred to be distinct. (C) Statistics of IBD tract length, in particular illustrating the N50 segment length. A graphical description of the modules and workflows for `DEploidIBD` is given in Figure 1-Supplement 2.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 3 Results

## 3.1 Method validation

We validated `DEploidIBD` through both experimental mixtures using lab strains and *in silico* mixtures using clonal field samples. First, to test consistency with `DEploid` (Zhu et al., 2018), we re-analysed the 27 experimental mixtures from (Wendler, 2015). This data set includes 27 samples of various mixtures of four laboratory parasite lines (3D7, Dd2, HB3 and 7G8; Figure 2–Supplement 3). Allowing for mixtures of up to four strains and using optimal reference panels, we found comparable performance with the single-step `DEploid` method, with the exception of three strains of equal proportions where LD information is necessary to achieve accurate deconvolution (Figure 2–Supplement 3).

To test the accuracy of `DEploidIBD` in a more realistic setting, we created *in silico* mixtures of two strains from 212 clonal samples of Asian origin (proportions ranging from 10/90% to 45/55%) using Chromosome 14 data (8,070 sites). A further 20 randomly chosen samples were used as the reference panel. In order to compare the accuracy of the two methods at different levels of relatedness, we set 25%, 50% and 75% of the second haplotype to be the same as the first haplotype to mimic scenarios of low, medium and high relatedness. This operation sets a lower limit to the relatedness between two strains, as background relatedness may also exist. To simulate data, we used empirical read depths and drew read counts for the two alleles from binomial proportions. We inferred strain proportions (summarised by the effective number of strains: $K_e = 1/\sum w_i^2$), and haplotypes. Both `DEploid` and `DEploidIBD` correctly estimate strain proportions with low relatedness (Figure 2A). However, for moderate and high relatedness mixtures, `DEploid` fails to recover the correct proportion, when the minor strain proportion is below 30%.

`DEploidIBD` is a substantial improvement on `DEploid`. In addition to estimating proportions and number of strains, `DEploidIBD` also estimates identity-by-descent (IBD) profiles. However, due to background relatedness `DEploidIBD` typically over-estimates IBD fraction by a few percentage points (Figure 2B). Rates of genotype error are similar for the two approaches in settings of low relatedness (error rate of 0.4% per site for 25/75 mixtures and 1.0% for 45/55 mixtures). However, for the 25/75% mixtures with high relatedness, genotype error for the non-IBD approach increases to 0.6%, while error in the IBD approach remains at 0.4% (Figure 2C). Switch errors in haplotype estimation are comparable between the two methods and decrease with increased relatedness due to the higher homozygosity (Figure 2D). In summary, joint inference of IBD profiles and strain haplotypes is expected to improve estimates of strain proportions (and hence haplotypes), particularly in regions with high rates of IBD. Moreover, direct estimates of IBD within mixed infections can be used as an additional feature to characterise isolates.

We repeated the *in silico* experiment with mixtures of two strains from 197 clonal African samples, with mixing proportions of 10/90%, 25/75% and 45/55%, using 92,780 sites from Chromosome 14. `DEploidIBD`

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

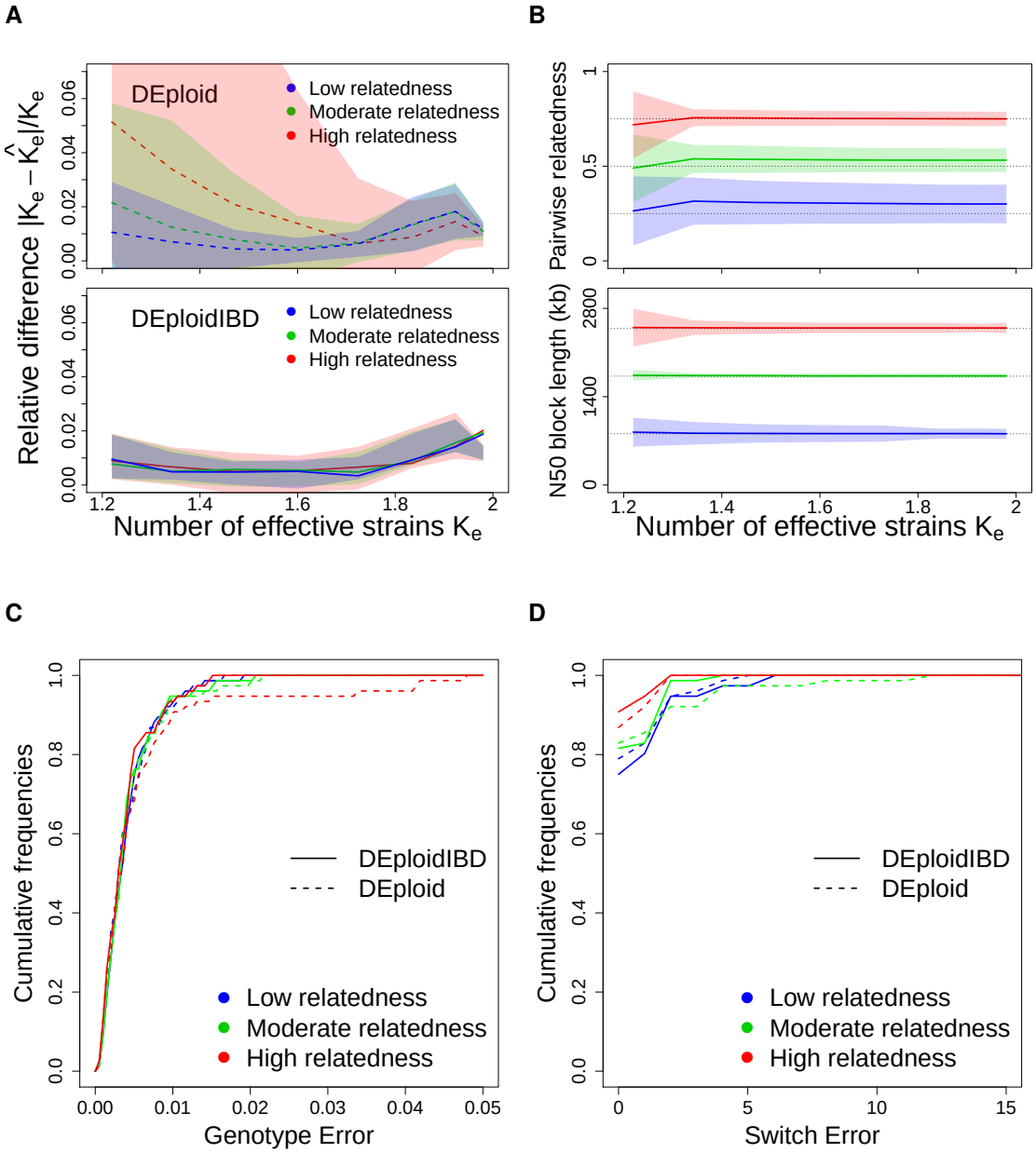


Figure 2: Comparison of `DEploidIBD` and `DEploid` on 76 *in silico* mixtures of two strains from Asia for 8,070 sites on Chromosome 14. (A) Relative differences of inferred effective number of strains using `DEploid` and `DEploidIBD`. The relative difference is calculated as the difference between inferred and expected effective number of strains divided by the expected value. (B) Inferred pairwise relatedness and N50 IBD tract length using `DEploidIBD`. Dotted lines indicate parameters used in the simulation. (C) Cumulative distribution of the average per site genotype error across simulated mixtures with three levels of IBD (25%, 50% and 75%) for a mixture proportion of 25/75%. We performed 100 simulations, but excluded eight cases where simulated haplotypes were over 99% identical and another 16 cases where average coverage was below 20. (D) Cumulative distribution of haplotype switch errors with three levels of IBD (25%, 50% and 75%) for a mixture proportion of 25/75%.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimates the correct proportions at all relatedness levels (Figure 2–Supplement 1), although with a greater relative difference in effective $K$ compare to Asia ($\sim 2\%$ vs. $\sim 1\%$). `DEploidIBD` also recovers the correct level of relatedness and IBD tract length (note that in Africa background relatedness is typically low). The per site genotype error rate remains below 1%. The number of haplotype switch errors is higher than in Asia, but by a factor much less than the 11-fold increase in the number of SNPs.

Finally, we extended benchmarking to *in silico* mixtures of three Asian strains (Figure 2–supplement 2). We set one strain to have the highest proportion (the dominant strain) and constructed the two minor strains to be IBD with the dominant strain over distinct halves of the chromosome, such that at any point there are only two distinct haplotypes present. We find that `DEploidIBD` outperforms (lower relative difference) `DEploid` in all cases and typically provides accurate estimates of proportions (Figure 2–Supplement 2) with the exception of two cases. For the case of (0.10, 0.40, 0.50), the minor strain creates very weak allele frequency imbalance, leading `DEploidIBD` to infer the number of strains as two (with proportions $\sim 45/55\%$) in 90/100 cases. For the case of (0.30, 0.30, 0.40), the problem is fundamentally unidentifiable and `DEploidIBD` fits the data as a mixture of two strains. In these cases, `DEploidIBD` also underestimates the pairwise relatedness and N50 tract lengths.

## 3.2 Geographical variation in mixed infection rates and relatedness

To investigate how the rate and relatedness structure of mixed infections varies among geographical regions with different epidemiological characteristics, we applied `DEploidIBD` to 2,344 field samples of *P. falciparum* released by the Pf3k project (Pf3k Consortium, 2016). These samples were collected under a wide range of studies with heterogeneous designs, with the majority of samples being taken from symptomatic individuals seeking clinical treatment. A summary of the data sources is presented in Table 1 and full details regarding study designs can be found at `https://www.malariagen.net/projects/pf3k#sampling-locations`. Details of data processing are given in the Methods. For deconvolution, samples were grouped into geographical regions by genetic similarity; four in Africa, and three in Asia. (Table 1). Reference panels were constructed from the clonal samples found at each region. Since previous research has uncovered severe population structure in Cambodia (Miotto et al., 2013), we stratified samples into West and North Cambodia when performing analysis at the country level. Diagnostic plots for the deconvolution of all samples can be found at `https://github.com/mcveanlab/mixedIBD-Supplement` and inferred haplotypes can be accessed at URL. We identified 787 samples where low sequencing coverage or the presence of low-frequency strains resulted in unusual haplotypes (see Supplementary Material). Estimates of strain number, proportions and IBD states from these samples are used in subsequent analyses, but not the haplotypes. We also confirmed that reported results are not affected by the exclusion of all metrics from samples with haplotypes with low

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

confidence.

We find substantial variation in the rate and relatedness structure of mixed infections across continents and countries. Within Africa, rates of mixed infection vary from 18% in Senegal to 63% in Malawi (Figure 3A). In Southeast Asian samples, mixed infection rates are in general lower, though also vary considerably; from 21% in Thailand to 54% in Bangladesh. Where data for a location is available over multiple years, we find no evidence for significant fluctuation over time (though we note that these studies are typically not well powered to see temporal variation and collection dates are very heterogeneous). We observe that between 5.1% (Senegal) and 40% (Malawi) of individuals have infections carrying more than two strains.

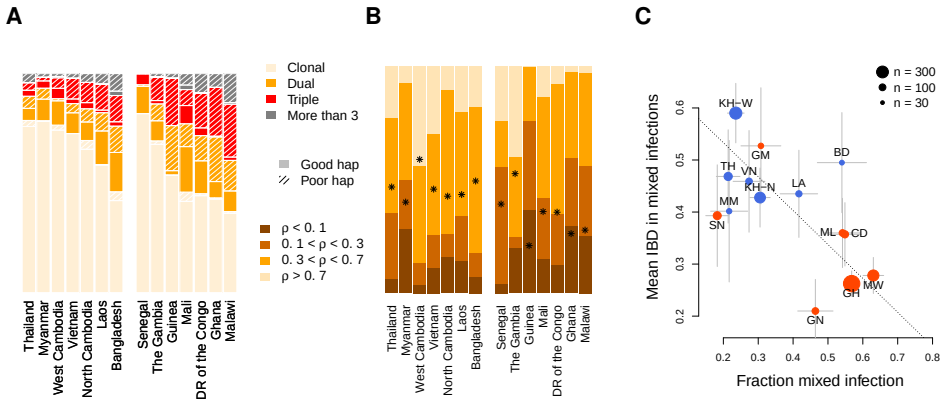


Figure 3: Characterisation of mixed infections across 2,344 field samples of *Plasmodium falciparum*. (A) The fraction of samples, by population, inferred to carry one (clonal), two, three, or more than three strains. Populations are ordered by rate of mixed infections within each continent. We use shaded regions to indicate the distribution of 787 samples that have low-confidence deconvolved haplotypes. (B) The distribution of IBD within mixed infections (including dual, triple and quad infections), broken down into unrelated (where the fraction of the genome inferred to be IBD, $\rho$, is $< 0.1$), low IBD ($0.1 \geq \rho < 0.3$), sib-level ($0.3 \geq \rho < 0.7$) and high ($\rho \geq 0.7$). Stars indicate the average IBD scaled between 0 and 1 from bottom to the top. Populations follow the same order as in Panel A. (C) The relationship between the rate of mixed infection and level of IBD. Populations are coloured by continent, with size reflecting sample size and error bars showing $\pm 1$ s.e.m.. The dotted line shows the slope of the regression from a linear model. Abbreviations: SN-Senegal, GM-The Gambia, NG-Nigeria, GN-Guinea, CD-The Democratic Republic of Congo, ML-Mali, GH-Ghana, MW-Malawi, MM-Myanmar, TH-Thailand, VN-Vietnam, KH-Cambodia, LA-Laos, BD-Bangladesh.

Relatedness between samples and populations also varies substantially. In dual infections, the average fraction of the genome inferred to be IBD ranges from 21% in Guinea to 59% in West Cambodia (Figure 3B). Asian populations show, on average, a higher level of relatedness within dual infections (48%) compared to African populations (29%). Levels of IBD in samples with three or more strains are comparable to those seen in dual infections (average IBD being 50% in Asia and 29% in Africa) and significantly correlated at the country level, with weighted correlation of 0.76 (P = 0.0017, weighted by the number of mixed samples). Overall, 53% of all mixed infections involve strains with over 30% of the genome being IBD.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Summary of Pf3k samples.

| Country | Year | Location | $PfPR$ | $ss$ | $\bar{D}$ (s.e.) | $\bar{\rho}$ | $\bar{K}$ | Reference |
|---|---|---|---|---|---|---|---|---|
| Gambia | 2008 | Brikam | 0.06 | 65 | 129 ( 9.4 ) | 0.53 | 1.3 | Amambua-Ngwa et al. (2012) |
| Ghana | 2009 | Navrongo | 0.79 | 121 | 86 ( 5.7 ) | 0.25 | 1.6 | Duffy et al. (2015); Kamau et al. (2015); MalariaGEN *Plasmodium falciparum* Community Project (2016) |
| | 2010 | Navrongo | 0.79 | 171 | 127 ( 10.3 ) | 0.26 | 1.5 | |
| | 2011 | Navrongo | 0.72 | 97 | 76 ( 5.3 ) | 0.24 | 1.5 | |
| | | Kintampo | 0.58 | 6 | 89 ( 13.5 ) | 0.16 | 1.5 | |
| | 2012 | Navrongo | 0.52 | 47 | 111 ( 3.8 ) | 0.31 | 1.6 | |
| | | Kintampo | 0.41 | 40 | 157 ( 8.1 ) | 0.25 | 1.6 | |
| | 2013 | Navrongo | 0.31 | 88 | 119 ( 4 ) | 0.29 | 1.6 | |
| | | Kintampo | 0.29 | 4 | 172 ( 38.4 ) | 0.53 | 1.1 | |
| Malawi | 2011 | Chikwawa | 0.19 | 230 | 101 ( 3 ) | 0.28 | 1.7 | Ocholla et al. (2014) |
| | | Zomba | 0.34 | 35 | 89 ( 9.1 ) | 0.28 | 1.6 | |
| Mali | 2007 | Bandiagara | 0.43 | 9 | 95 ( 25.2 ) | 0.39 | 1.8 | MalariaGEN *Plasmodium falciparum* Community Project (2016) |
| | | Faladje | 0.37 | 36 | 75 ( 10.1 ) | 0.34 | 1.3 | |
| | | Kolle | 0.21 | 51 | 82 ( 10.5 ) | 0.36 | 1.6 | |
| Guinea | 2011 | Nzerekore | 0.49 | 97 | 77 ( 4.6 ) | 0.21 | 1.4 | |
| Congo DR | 2013 | Kinshasa | 0.24 | 113 | 49 ( 3.2 ) | 0.36 | 1.5 | |
| Senegal | 2004 | Thies | 0.09 | 2 | 130 ( 68.2 ) | 0.03 | 1.4 | Wong et al. (2017) |
| | 2009 | Thies | 0.04 | 43 | 175 ( 14.9 ) | 0.47 | 1.1 | |
| | 2010 | Thies | 0.04 | 24 | 159 ( 9.7 ) | 0.36 | 1.3 | |
| | 2011 | Thies | 0.03 | 32 | 97 ( 6 ) | 0.4 | 1.1 | |
| West Cambodia | 2009 | Pursat | 0.0071 | 19 | 75 ( 8.8 ) | 0.39 | 1.3 | Amato et al. (2017); MalariaGEN *Plasmodium falciparum* Community Project (2016) |
| | 2010 | Pursat | 0.0071 | 105 | 95 ( 6.8 ) | 0.65 | 1.2 | |
| | 2011 | Pailin | 0.0025 | 49 | 54 ( 4.1 ) | 0.43 | 1.1 | |
| | | Pursat | 0.0096 | 103 | 49 ( 3.1 ) | 0.63 | 1.2 | |
| | 2012 | Pailin | 0.00096 | 31 | 46 ( 5.6 ) | 0.43 | 1.0 | |
| | | Pursat | 0.0079 | 7 | 37 ( 19.1 ) | 0.58 | 1.4 | |
| North Cambodia | 2010 | Ratanakiri | 0.0039 | 50 | 71 ( 6.1 ) | 0.44 | 1.3 | |
| | 2011 | Preah Vihear | 0.02 | 73 | 51 ( 5.3 ) | 0.36 | 1.2 | |
| | | Ratanakiri | 0.0032 | 81 | 45 ( 4.3 ) | 0.48 | 1.4 | |
| | 2012 | Preah Vihear | 0.0075 | 30 | 43 ( 6.7 ) | 0.38 | 1.0 | |
| | | Ratanakiri | 0.0016 | 15 | 44 ( 8.9 ) | 0.32 | 1.3 | |
| Thailand | 2011 | Mae Sot | 0.00011 | 35 | 66 ( 7.5 ) | 0.35 | 1.2 | Miotto et al. (2013); MalariaGEN *Plasmodium falciparum* Community Project (2016) |
| | | Sisakhet | 1e-04 | 5 | 112 ( 25.4 ) | 0.17 | 1.3 | |
| | 2012 | Mae Sot | 5.7e-05 | 69 | 83 ( 4.9 ) | 0.59 | 1.3 | |
| | | Ranong | 0.00018 | 11 | 82 ( 12.4 ) | 0.34 | 1.2 | |
| | | Sisakhet | 0 | 13 | 89 ( 13 ) | 0.37 | 1.1 | |
| | 2013 | Sisakhet | 0 | 3 | 62 ( 8.8 ) | 0.09 | 1.2 | |
| Bangladesh | 2012 | Ramu | 0.0021 | 50 | 53 ( 4.2 ) | 0.49 | 1.5 | |
| Viet Nam | 2011 | Bu Gia Map | 0.0073 | 43 | 67 ( 5 ) | 0.44 | 1.3 | |
| | | Phuoc Long | 0.0053 | 27 | 68 ( 7.2 ) | 0.38 | 1.2 | |
| | 2012 | Bu Gia Map | 0.0072 | 19 | 115 ( 8 ) | 0.67 | 1.1 | |
| | | Phuoc Long | 0.0048 | 5 | 107 ( 6.3 ) | 0.82 | 1.2 | |
| Myanmar | 2011 | Bago Division | 0.0076 | 12 | 59 ( 7.1 ) | 0.26 | 1.2 | |
| | 2012 | Bago Division | 0.0084 | 47 | 62 ( 5.2 ) | 0.46 | 1.2 | |
| Laos | 2011 | Attapeu | 0.0094 | 59 | 71 ( 4.2 ) | 0.37 | 1.4 | |
| | 2012 | Attapeu | 0.02 | 25 | 77 ( 7.2 ) | 0.69 | 1.3 | |

Table 2: Summary of Pf3k samples in data release 5.1, where $D$ denotes mean read depth and $ss$ is sample size. Genotyping, including both indel and SNP variants, was performed using a pipeline based on GATK best practices, see Methods. Data available from `ftp://ngs.sanger.ac.uk/production/pf3k/release_5/5.1`. $PfPR$ is the inferred parasite prevalence rate in a $5 \times 5$ km resolution grid from the MAP project, centred at the Pf3k sample collection sites; Relatedness $\rho$ and effective number of strains $K_e$ are summary metrics from `DEploidIBD` output.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We next considered the relationship between mixed infection rate and the level of IBD. We find that populations with higher rates of mixed infection tend to have lower levels of IBD within mixed infections (linear model P = 0.06 after accounting for a continental level difference and weighted by sample size). However, the continental level effect is driven by Senegal, which has an unusual combination of low mixed infections and also low IBD. Excluding Senegal, we find a consistent pattern across populations (Figure 3C), with a strong negative correlation between mixed infection rate and the level of IBD (Pearson $r = -0.84$, P $= 3 \times 10^{-4}$). Previous work has demonstrated how a recent and dramatic decline in *P.* falciparum prevalence within Senegal has left an impact on patterns of genetic variation (Daniels et al., 2015), which may explain its unusual profile.

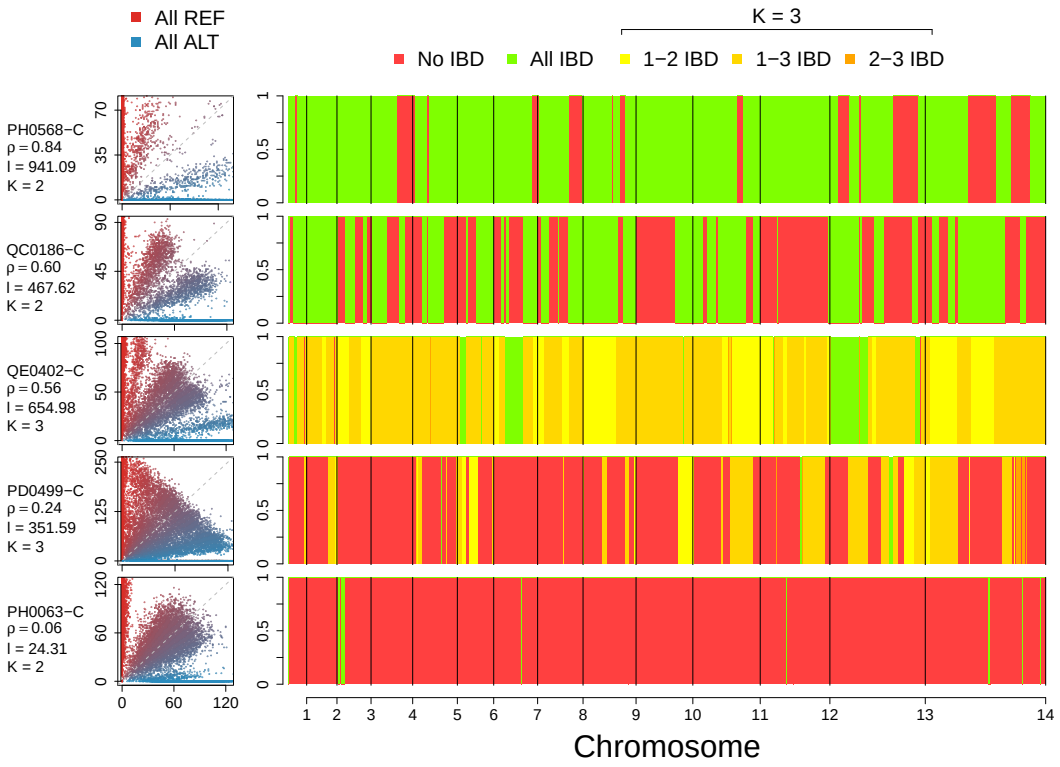


Figure 4: Example IBD profiles in mixed infections. Plots showing the ALT versus REF plots (left hand side) and inferred IBD profiles along the genome for five strains of differing composition. From top to bottom: A dual infection of highly related strains ($\rho = 0.84$); a dual infection of two sibling strains ($\rho = 0.6$); a triple infection of three sibling strains (note the absence of stretches without IBD); a triple infection of two related strains and one unrelated strain; and a triple infection of three unrelated strains. The numbers below the sample IDs indicate the average pairwise IBD, $r$, and the mean length of IBD segments, $l$, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 3.3 Inferring the origin of IBD in mixed infections

The high levels of IBD observed in many mixed infections suggest the presence of sibling strains (Figure 4). To quantify the expected IBD patterns between siblings, we developed a meiosis simulator for *P. falciparum* (`pf-meiosis`), incorporating relevant features of malaria biology that can impact the way IBD is produced in a mosquito and detected in a human host. Most importantly, a single infected mosquito can undergo multiple meioses in parallel, one occurring for each oocyst that forms on the mosquito midgut (Ghosh et al., 2000). In a mosquito infected with two distinct strains, each oocyst can either self (the maternal and paternal strain are the same) or outbreed (the maternal and paternal strains are different). We model a $K = n$ mixed infection as a sample of $n$ strains (without replacement, as drawing identical strains yields $K = n - 1$) from the pool of strains created by all oocysts. Studies of wild-caught *Anopheles Gambiae* suggest that the distribution of oocysts is roughly geometric, with the majority of infected mosquitoes carrying only one oocyst (Beier et al., 1991; Collins et al., 1984). Surprisingly, in such a case, a $K = 2$ infection will have an expected IBD of $1/3$ (see Supplementary Materials). Conditioning on at least one progeny originating from an outbred oocyst (such that a detectable recombination event has occurred), the expected IBD asymptotically approaches $1/2$ as the total number of oocysts grows.

Using this simulation framework, we sought to classify observed mixed infections based on their patterns of IBD. We used two summary statistics to perform the classification: mean IBD segment length and IBD fraction. We built empirical distributions for these two statistics for each country in Pf3k, by simulating meiosis between pairs of clonal samples from that country. In this way, we control for variation in genetic diversity (as background IBD between clonal samples) in each country. Starting from a pair of clonal samples ($M = 0$, where $M$ indicates the number of meioses that have occurred), we simulated three successive rounds of meiosis ($M = 1, 2, 3$), representing the creation and serial transmission of a mixed infection (Figure 5A). Each round of meiosis increases the amount of observed IBD. For example, in Ghana, the mean IBD fraction for $M = 0$ was 0.002, for $M = 1$ was 0.41, for $M = 2$ was 0.66, and for $M = 3$ was 0.80 (Figure 5B). West Cambodia, which has lower genetic diversity, had a mean IBD fraction of 0.08 for $M = 0$ and consequently, the mean IBD fractions for higher values of $M$ were slightly increased, to 0.46, 0.68, 0.81 for $M = 1, 2$ and 3, respectively (Figure 5B).

From these simulated distributions, we used Naive Bayes to classify $k = 2$ mixed infections in Pf3k (Figure 5C). Of the 404 $K = 2$ samples containing only high-quality haplotypes (see Supplementary Materials), 288 (71%) had IBD statistics that fell within the range observed across all simulated $M$. Of these, more than half (221, 55%) were classified as siblings ($M > 0$, with ¿ 99% posterior probability). Moreover, we observe geographical differences in the rate at which sibling and unrelated mixed infections occur. Notably, in Asia a greater fraction of all mixed infections contained siblings (65% vs. 51% in Africa), driven by a higher frequency of $M = 2$ and $M = 3$ mixed infections (Figure 5D).

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

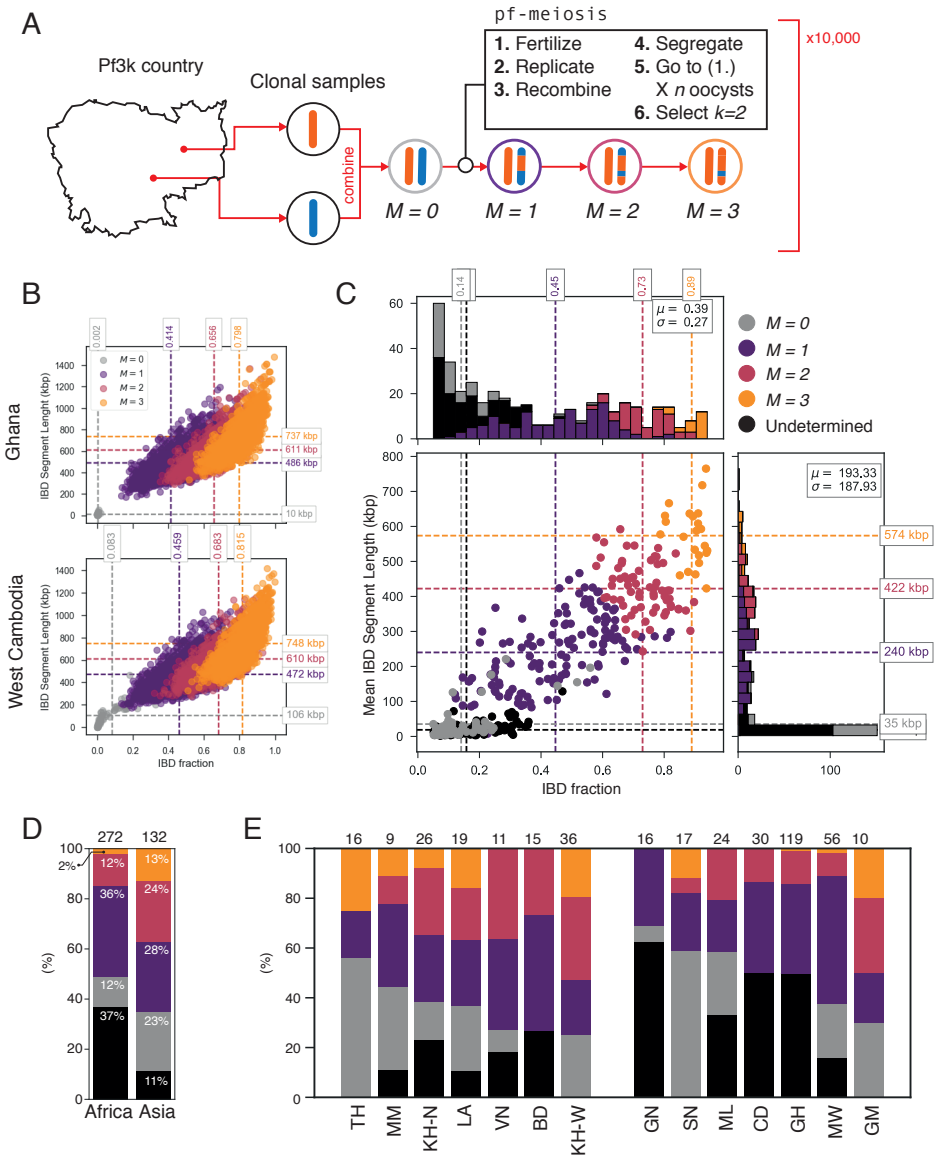


Figure 5: Identifying sibling strains within mixed infections. (A) Schematic showing how IBD fraction and IBD segment length distributions are created for $k = 2$ mixed infections using `pf-meiosis`. Two clonal samples from a given country are combined to create an unrelated ($M = 0$, where $M$ is number of meioses that have occurred) mixed infection. The $M = 0$ infection is then passed through 3 rounds of pf-meioses to generate $M = 1, 2, 3$ classes, representing serial transmission of the mixed infection ($M = 1$ are siblings). (B) Simulated IBD distributions for $M = 0, 1, 2, 3$ for Ghana (top) and West Cambodia (bottom). A total of 10,000 mixed infections are simulated for each class, from 500 random pairs of clonal samples. (C) Classification results for 404 $K = 2$ mixed infections from 13 countries. Undetermined indicates mixed infections with IBD statistics that were never observed in simulation. (D) Breakdown of class percentage by continent. Total number of samples is given above bars. Colours as in panel C ($M = 0$, grey; $M = 1$, purple; $M = 2$, pink; $M = 3$, orange; Undeteremined, black). (E) Same as (D), but by country. Abbreviations as in Figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 3.4 Characteristics of mixed infections correlate with local parasite prevalence

To assess how characteristics of mixed infections relate to local infection intensity, we obtained estimates of *P. falciparum* prevalence ($PfPR_{2-10}$) from the Malaria Atlas Project (MAP, 2017, see Table 1). The country level prevalence estimates range from 0.01% in Thailand to 55% in Ghana, with African countries having up to two orders of magnitude greater values than Asian ones (mean of 36% in Africa and 0.6% in Asia). However, seasonal and geographic fluctuations in prevalence mean that, conditional on sampling an individual with malaria, local prevalence may be much higher than the longer-term (and more geographically widespread) average. We summarise mixed infection rates by the average effective number of strains, which reflects both the number and proportion of strains present. This metric both avoids the problem of having to estimate a threshold for determining the presence of a very low proportion strain and is sensitive to the presence of triply (and more) infected samples.

We find that the effective number of strains is a significant predictor of $PfPR_{2-10}$ in African populations ($r = 0.48, P = 0.04$), but is uncorrelated within Asian populations. Similarly, within-sample IBD and background IBD are both negatively correlated with $PfPR_{2-10}$ only in Africa ($r = -0.67, P = 0.0017$ and $r = -0.53, P = 0.02$, respectively). The rate of sibling infection ($M = 1$) is not correlated with the parasite prevalence ($r = -0.06$, $P = 0.70$). However, the super-sibling infection rate ($M = 2, 3$) does exhibit a marginally significant correlation with $PfPR_{2-10}$ ($r = -0.29$, $P = 0.06$), albeit only at the continental scale. Interestingly, all statistics relating to IBD are positively correlated with $PfPR_{2-10}$ in Asian populations (though not significantly so), in contrast to the negative (and significant) associations seen within African populations.

## 4 Discussion

It has long been appreciated that mixed infections are an integral part of malaria biology, determining the number, proportions, and haplotypes of the strains that comprise them has proven a formidable challenge. Previously we developed an algorithm, `DEploid`, for deconvolving mixed infections (Zhu et al., 2018). However, we subsequently noticed the presence of mixed infections with highly related strains in which the algorithm performed poorly, particularly with low-frequency minor strains. Mixed infections containing highly related strains represent an epidemiological scenario of particular interest, because they are likely to have been produced from a single mosquito bite, itself multiply infected, and in which meiosis has occurred to generate sibling strains. Thus, we developed an enhanced method, `DEploidIBD`, capable not only of deconvolving highly related mixed infections, but also providing a profile of IBD segments between all pairs of strains present in the infection. We note that technical difficulties remain, including analysing data with multiple infecting species, coping with low-coverage data, and selecting appropriate reference panels from

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

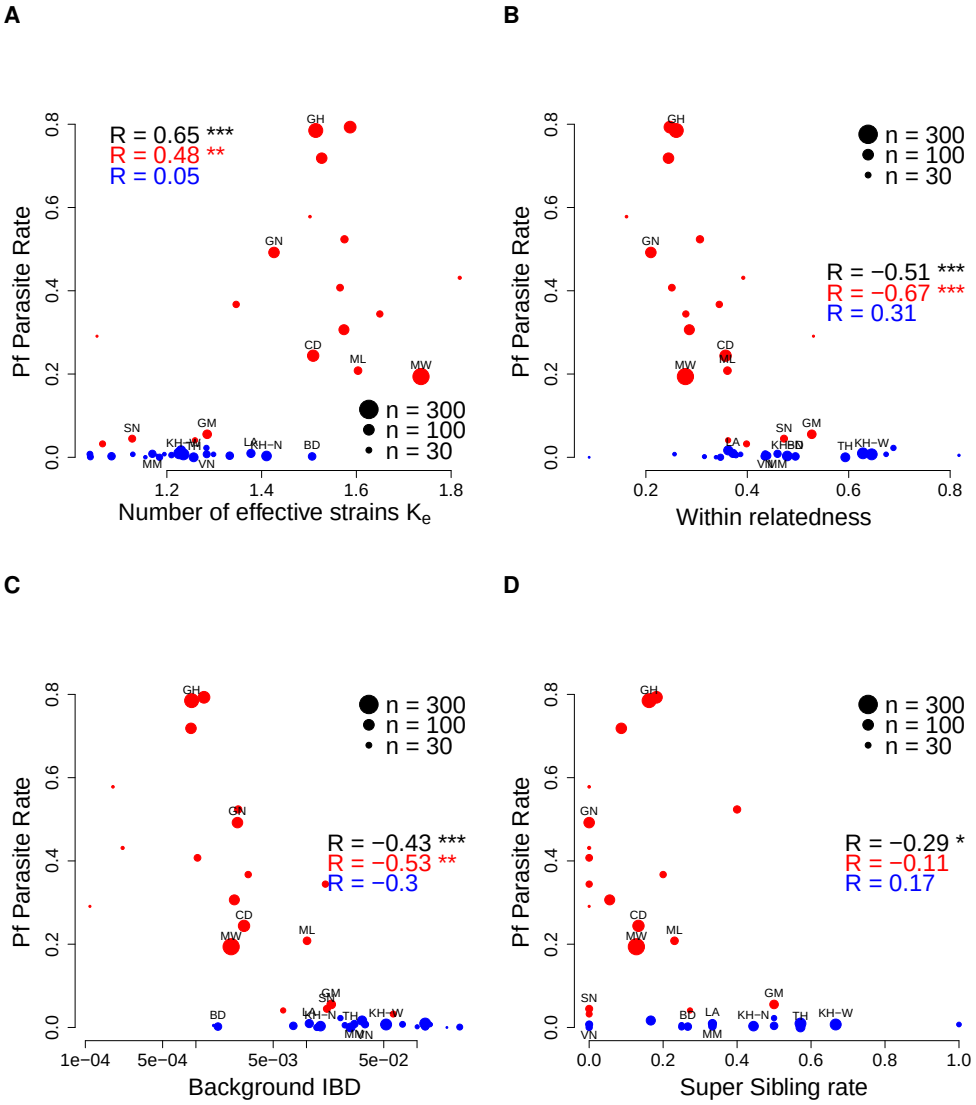


Figure 6: The relationship between *P. falciparum* prevalence and characteristics of mixed infection including (A) the average effective number of strains, given by $K_e = (\sum w_i^2)^{-1}$, where $w_i$ is the proportion of the $i$th strain; (B) within sample relatedness (average IBD fraction) in mixed samples; (C) background relatedness between clonal samples; and (D) super sibling rate ($M > 1$) in dual infections. Each point relates to a row in Table 1 from different sampling locations and years. Point size is proportional to Pf3k sample size. We label points with the country of the sampling site contributing the largest number of samples. Black R is the R for the combined dataset.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the growing reference resources.

The application of `DEploidIBD` to the 2,344 samples in the Pf3k project has revealed the extent and structure of relatedness among malaria infections and how these characteristics vary between geographic locations. We found that 1,026 (44%) of all samples in Pf3k were mixed, being comprised of 480 $K = 2$ infections, 372 $K = 3$ and 127 $K = 4$ infections. Across the entire data set, the total number of genomes extracted from mixed infections is nearly double the number extracted from clonal infections (2,584 genomes from $K > 1$ vs. 1,365 from $K = 1$). We also found considerable variation, between countries and continents in the characteristics of mixed infections, suggesting that they are sensitive to local epidemiology. For example, in West Africa, Senegal (which has undergone a recent and effective malaria control campaign) has a rate of mixed infections less than half that of neighbouring Guinea and Mali. Previous work has highlighted the utility of mixed infection rate in discerning changes in regional prevalence, and we re-enforce that finding here, observing a significant correlation between the effective number of strains and parasite prevalence across Pf3k collection sites. Similarly, using `DEploidIBD` we also observe significant geographical variation in the relatedness profiles of strains within mixed infections. Interestingly, this variation is structured such that regions with high rates of mixed infection tend to contain strains that are less related, resulting in a significant negative correlation between mixed infection rate and mean relatedness within those infections.

The ability to identify the extent and genomic structure of IBD enables inference of the mechanisms by which mixed infections can arise. A mixed infection of $K$ strains can be produced by either $K$ independent infectious bites or by $j < K$ infectious bites. In the first case, parasites are delivered by separate vectors and no meiosis occurs between the distinct strains, thus any IBD observed in the mixed infection must have pre-existed as background IBD between the individual strains. In the second case, meiosis may occur between strains, resulting in long tracts of IBD. The exact amount of IBD produced by meiosis is a random variable, dependent on outcomes of meiotic processes, such as the number of recombination events, the distance between them, and the segregation of chromosomes. Importantly, the mean IBD produced during meiosis in *P. falciparum* also depends on the number and type (selfed vs outbred) of oocysts in the infectious mosquito. Consequently, the amount of IBD expected in a single-bite mixed infection produced from two unrelated parasites strains will always be slightly less than 1/2, and as low as 1/3.

To quantify the distribution of IBD statistics expected through different mechanisms of mixed infection, we developed a Monte Carlo simulation tool, `pf-meiosis`, which we used to infer the recent transmission history of individuals with dual ($K = 2$) infections. We considered mixed infection chains, in which $M$ successive rounds of meiosis, transmission to host, and uptake by vector can result in sibling strain infections with very high levels of IBD. Overall, we found that 56% of all mixed infections are from sibling strains and, particularly within Asian population samples, evidence for long mixed infection chains ($M > 1$). This observation is not a product of lower genetic diversity in Asia, as differences in background IBD between

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

countries have been controlled for in the simulations. Rather, it reflects true differences in transmission epidemiology between continents. These findings have three important consequences. First, it suggests that successful establishment of multiple strains through a single infection event is major source of mixed infection. Second, it implies that the bottlenecks imposed at transmission (to host and vector) are relatively weak. Finally, it indicates that the source of mixed infections reflects aspects of local epidemiology.

We note that a non-trivial fraction (29%) of all mixed infections had patterns of IBD inconsistent with the simulations (typically higher IBD than background but lower than among siblings). We suggest two explanations. Firstly, our estimate of background IBD, generated by combining pairs of random clonal samples from a given country into an artificial $M = 0$ mixed infection, will underestimate true background IBD if there is very strong local population structure. Second, we only simulated simple mixed infection transmission chains, at the exclusion of more complex transmission histories, such as involving strains related at the level of cousins. The extent to which such complex histories can be inferred with certainty remains to be explored.

Finally, our results show that the rate and relatedness structure of mixed infections correlate with estimated levels of parasite prevalence, at least within Africa, where prevalence is typically high (Smith et al., 1993). In Asia, which has much lower overall prevalence, as well as greater temporal (and possibly spatial) fluctuations, we do not observe such correlations. However, it may well be that other genomic features that we don't contemplate in this work could provide much higher resolution, in space and time, for capturing changes in prevalence than traditional methods. Testing this hypothesis will lead to a much greater understanding of how genomic data can potentially be used to inform global efforts to control and eradicate malaria.

# 5 Methods and Materials

The data analysed within this paper were collected and made openly available to researchers by member of the Pf3k Consortium. Information about studies within the data set can be found at `https://www.malariagen.net/projects/pf3k#sampling-locations`. Detailed information about data processing can be found at `https://www.malariagen.net/data/pf3k-5`. Briefly, field isolates were sequenced to an average read depth of 86 (range 12.6 – 192.5). After removing human-derived reads and mapping to the 3D7 reference genome, variants were called using GATK best practice and approximately one million variant sites were genotyped in each isolate. After filtering samples for low coverage and cross-species contamination, 2,344 samples remained. The Supplementary Material provides details on the filters used and data availability. For deconvolution, samples were grouped into geographical regions by genetic similarity; four in Africa, and three in Asia. (Table 1). Reference panels were constructed from the clonal samples found at each region. Since previous research has uncovered severe population structure in Cambodia (Miotto et al., 2013), we

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stratified samples into West and North Cambodia when performing analysis at the country level.

## 6 Acknowledgements

This study was supported by the Wellcome Trust (206194, 090770, 204911, 100956/Z/13/Z to GM), the Medical Research Council (G0600718), and the UK Department for International Development (M006212). This study used data from the MalariaGEN Pf3k Project. Genome sequencing was done by the Wellcome Sanger Institute (WSI), and sample collections were coordinated by the MalariaGEN Resource Centre. The samples from Senegal were was supported by funding from the Bill and Melinda Gates Foundation to Dyann Wirth, and sequenced by the Broad Institute. We thank the staff of the WSI Sample Logistics, Sequencing, and Informatics facilities for their contribution; all patients and collaborators contributing samples and data to the Pf3k project.

## 7 Data availability

Metadata on samples is available from `ftp://ngs.sanger.ac.uk/production/pf3k/release_5/pf3k_release_5_metadata_20170804.txt.gz`. Sequence data (aligned to *P*lasmodium falciparum strain 3D7 v3.1 reference genome sequences, for details see `ftp://ftp.sanger.ac.uk/pub/project/pathogens/gff3/2015-08/Pfalciparum.genome.fasta.gz`) is available from `ftp://ngs.sanger.ac.uk/production/pf3k/release_5/5.1/`. Diagnostic plots for the deconvolution of all samples can be found at `https://github.com/mcveanlab/mixedIBD-Supplement` and deconvolved haplotypes can be accessed at XXX. Code implementing the algorithms described in this paper, `DEploidIBD`, is available at `https://github.com/mcveanlab/DEploid`.

## 8 Disclosure Declaration

None declared.

## References

Amambua-Ngwa, A., K. K. A. Tetteh, M. Manske, N. Gomez-Escobar, L. B. Stewart, M. E. Deerhake, I. H. Cheeseman, C. I. Newbold, A. A. Holder, E. Knuepfer, O. Janha, M. Jallow, S. Campino, B. MacInnis, D. P. Kwiatkowski, and D. J. Conway (2012, 11). Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLOS Genetics 8*(11), 1–14.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Amato, R., P. Lim, O. Miotto, C. Amaratunga, D. Dek, R. D. Pearson, J. Almagro-Garcia, A. T. Neal, S. Sreng, S. Suon, E. Drury, D. Jyothi, J. Stalker, D. P. Kwiatkowski, and R. M. Fairhurst (2017, 2018/03/25). Genetic markers associated with dihydroartemisinin-piperaquine failure in *plasmodium falciparum* malaria in cambodia: a genotype-phenotype association study. *The Lancet Infectious Diseases 17*(2), 164–173.

Auburn, S., S. Campino, O. Miotto, A. A. Djimde, I. Zongo, M. Manske, G. Maslen, V. Mangano, D. Alcock, B. MacInnis, et al. (2012). Characterization of within-host *plasmodium falciparum* diversity using next-generation sequence data. *PloS one 7*(2), e32891.

Beier, J. C., F. K. Onyango, M. Ramadhan, J. K. Koros, C. M. Asiago, R. A. Wirtz, D. K. Koech, and C. R. Roberts (1991). Quantitation of malaria sporozoites in the salivary glands of wild afrotropical anopheles. *Med Vet Entomol 5*(1), 63–70.

Bell, A. S., J. C. De. Roode, D. Sim, and A. F. Read (2006). Within-host competition in genetically diverse malaria infection: parasite virulence and competitive success. *Evolution 60*(7), 1358–1371.

Borrmann, S. and K. Matuschewski (2011). Protective immunity against malaria by 'natural immunization': a question of dose, parasite diversity, or both? *Current Opinion in Immunology 23*(4), 500 – 508.

Bruce, M. C. and K. P. Day (2002). Cross-species regulation of malaria parasitaemia in the human host. *Current Opinion in Microbiology 5*(4), 431 – 437.

Bushman, M., L. Morton, N. Duah, N. Quashie, B. Abuaku, K. A. Koram, P. R. Dimbu, M. Plucinski, J. Gutman, P. Lyaruu, S. P. Kachur, J. C. de Roode, and V. Udhayakumar (2016). Within-host competition and drug resistance in the human malaria parasite *plasmodium falciparum. Proceedings of the Royal Society of London B: Biological Sciences 283*(1826).

Cerqueira, G. C., I. H. Cheeseman, S. F. Schaffner, S. Nair, M. McDew-White, A. P. Phyo, E. A. Ashley, A. Melnikov, P. Rogov, B. W. Birren, F. Nosten, T. J. C. Anderson, and D. E. Neafsey (2017, Apr). Longitudinal genomic surveillance of *plasmodium falciparum* malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. *Genome Biology 18*(1), 78.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience 4*(1), 1–16.

Chang, H.-H., C. J. Worby, A. Yeka, J. Nankabirwa, M. R. Kamya, S. G. Staedke, G. Dorsey, M. Murphy, D. E. Neafsey, A. E. Jeffreys, C. Hubbart, K. A. Rockett, R. Amato, D. P. Kwiatkowski, C. O. Buckee, and B. Greenhouse (2017, 01). 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), 1–18.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Collins, F. H., F. Zavala, P. M. Graves, A. H. Cochrane, R. W. Gwadz, J. Akoh, and R. S. Nussenzweig (1984). First field trial of an immunoradiometric assay for the detection of malaria sporozoites in mosquitoes. *Am J Trop Med Hyg 33*(4), 538–43.

Collins, W. E. (2012). Plasmodium knowlesi: A malaria parasite of monkeys and humans. *Annual Review of Entomology 57*(1), 107–121.

Daniels, R. F., S. F. Schaffner, E. A. Wenger, J. L. Proctor, H.-H. Chang, W. Wong, N. Baro, D. Ndiaye, F. B. Fall, M. Ndiop, M. Ba, D. A. Milner, T. E. Taylor, D. E. Neafsey, S. K. Volkman, P. A. Eckhoff, D. L. Hartl, and D. F. Wirth (2015, Jun). Modeling malaria genomics reveals transmission decline and rebound in senegal. *Proceedings of the National Academy of Sciences of the United States of America 112*(22), 7067–7072.

de Roode, J. C., R. Pansini, S. J. Cheesman, M. E. H. Helinski, S. Huijben, A. R. Wargo, A. S. Bell, B. H. K. Chan, D. Walliker, and A. F. Read (2005). Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences 102*(21), 7624–7628.

Duffy, C. W., S. A. Assefa, J. Abugri, N. Amoako, S. Owusu-Agyei, T. Anyorigiya, B. MacInnis, D. P. Kwiatkowski, D. J. Conway, and G. A. Awandare (2015, Jul). Comparison of genomic signatures of selection on *plasmodium falciparum* between different regions of a country with high malaria endemicity. *BMC Genomics 16*(1), 527.

Enosse, S., C. Dobaño, D. Quelhas, J. J. Aponte, M. Lievens, A. Leach, J. Sacarlal, B. Greenwood, J. Milman, F. Dubovsky, J. Cohen, R. Thompson, W. R. Ballou, P. L. Alonso, D. J. Conway, and C. J. Sutherland (2006, 05). Rts,s/as02a malaria vaccine does not induce parasite csp t cell epitope selection and reduces multiplicity of infection. *PLOS Clinical Trials 1*(1), 1–10.

Galinsky, K., C. Valim, A. Salmier, B. de Thoisy, L. Musset, E. Legrand, A. Faust, M. L. Baniecki, D. Ndiaye, R. F. Daniels, D. L. Hartl, P. C. Sabeti, D. F. Wirth, S. K. Volkman, and D. E. Neafsey (2015, Jan). Coil: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria Journal 14*(1), 4.

Ghosh, A., M. J. Edwards, and M. Jacobs-Lorena (2000). The journey of the malaria parasite in the mosquito: hopes for the new century. *Parasitol Today 16*(5), 196–201.

Graham, R. L., D. E. Knuth, and O. Patashnik (1988). *Concrete Mathematics*. Reading MA: Addison-Wesley.

Gusev, A., E. E. Kenny, J. K. Lowe, J. Salit, R. Saxena, S. Kathiresan, D. M. Altshuler, J. Friedman, J. L. Breslow, and I. Pe'er (2011, Jun). Dash: A method for identical-by-descent haplotype mapping uncovers association with recent variation. *The American Journal of Human Genetics 88*(6), 706–717.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gusev, A., J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, and I. Pe'er (2009, Feb). Whole population, genome-wide mapping of hidden relatedness. *Genome Research 19*(2), 318–326.

Henden, L., S. Lee, I. Mueller, A. Barry, and M. Bahlo (2018, 05). Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. *PLOS Genetics 14*(5), 1–31.

Howes, R. E., K. E. Battle, K. N. Mendis, D. L. Smith, R. E. Cibulskis, J. K. Baird, and S. I. Hay (2016). Global epidemiology of *plasmodium vivax*. *The American Journal of Tropical Medicine and Hygiene 95*, 15–34.

Ivo Mueller, Peter A. Zimmerman, J. C. R. (2007). *Plasmodium malariae* and *plasmodium ovale* – the "bashful" malaria parasites. *Trends in Parasitology 23*, 278–283.

Kamau, E., S. Campino, L. Amenga-Etego, E. Drury, D. Ishengoma, K. Johnson, D. Mumba, M. Kekre, W. Yavo, D. Mead, M. Bouyou-Akotet, T. Apinjoh, L. Golassa, M. Randrianarivelojosia, B. Andagalu, O. Maiga-Ascofare, A. Amambua-Ngwa, P. Tindana, A. Ghansah, B. MacInnis, D. Kwiatkowski, and A. A. Djimde (2015). K13-propeller polymorphisms in *plasmodium falciparum* parasites from sub-saharan africa. *The Journal of Infectious Diseases 211*(8), 1352–1355.

MalariaGEN *Plasmodium falciparum* Community Project (2016, mar). Genomic epidemiology of artemisinin resistant malaria. *eLife 5*, e08714.

Manske, M., O. Miotto, S. Campino, S. Auburn, J. Almagro-Garcia, G. Maslen, J. O'Brien, A. Djimde, O. Doumbo, I. Zongo, J.-B. Ouedraogo, P. Michon, I. Mueller, P. Siba, A. Nzila, S. Borrmann, S. M. Kiara, K. Marsh, H. Jiang, X.-Z. Su, C. Amaratunga, R. Fairhurst, D. Socheat, F. Nosten, M. Imwong, N. J. White, M. Sanders, E. Anastasi, D. Alcock, E. Drury, S. Oyola, M. A. Quail, D. J. Turner, V. Ruano-Rubio, D. Jyothi, L. Amenga-Etego, C. Hubbart, A. Jeffreys, K. Rowlands, C. Sutherland, C. Roper, V. Mangano, D. Modiano, J. C. Tan, M. T. Ferdig, A. Amambua-Ngwa, D. J. Conway, S. Takala-Harrison, C. V. Plowe, J. C. Rayner, K. A. Rockett, T. G. Clark, C. I. Newbold, M. Berriman, B. MacInnis, and D. P. Kwiatkowski (2012). Analysis of *plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature 487*, 375–379.

MAP (2017). The malaria atlas project: developing global maps of malaria risk. Project Date: December 8, 2017.

Miotto, O., J. Almagro-Garcia, M. Manske, B. MacInnis, S. Campino, K. A. Rockett, C. Amaratunga, P. Lim, S. Suon, S. Sreng, J. M. Anderson, S. Duong, C. Nguon, C. M. Chuor, D. Saunders, Y. Se, C. Lon, M. M. Fukuda, L. Amenga-Etego, A. V. Hodgson, V. Asoala, M. Imwong, S. Takala-Harrison, F. Nosten,

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

X.-z. Su, P. Ringwald, F. Ariey, C. Dolecek, T. T. Hien, M. F. Boni, C. Q. Thai, A. Amambua-Ngwa, D. J. Conway, A. A. Djimdé, O. K. Doumbo, I. Zongo, J.-B. Ouedraogo, D. Alcock, E. Drury, S. Auburn, O. Koch, M. Sanders, C. Hubbart, G. Maslen, V. Ruano-Rubio, D. Jyothi, A. Miles, J. O'Brien, C. Gamble, S. O. Oyola, J. C. Rayner, C. I. Newbold, M. Berriman, C. C. Spencer, G. McVean, N. P. Day, N. J. White, D. Bethell, A. M. Dondorp, C. V. Plowe, R. M. Fairhurst, and D. P. Kwiatkowski (2013, Jun). Multiple populations of artemisinin-resistant *plasmodium falciparum* in cambodia. *Nature Genetics 45*(6), 10.1038/ng.2624.

Mouzin, E., P. M. Thior, M. B. Diouf, and B. Sambou (2010). Focus on senegal roll back malaria: Progress and impact series. *Geneva: World Health Organization*.

Mzilahowa, T., P. J. McCall, and I. M. Hastings (2007, 07). "sexual" population structure and genetics of the malaria agent *p. falciparum*. *PLOS ONE 2*(7), 1–8.

Nair, S., S. C. Nkhoma, D. Serre, P. A. Zimmerman, K. Gorena, B. J. Daniel, F. Nosten, T. J. Anderson, and I. H. Cheeseman (2014, Jun). Single-cell genomics for dissection of complex malaria infections. *Genome Research 24*(6), 1028–1038.

Nkhoma, S. C., S. Nair, I. H. Cheeseman, C. Rohr-Allegrini, S. Singlam, F. Nosten, and T. J. C. Anderson (2012). Close kinship within multiple-genotype malaria parasite infections. *Proceedings of the Royal Society of London B: Biological Sciences 279*(1738), 2589–2598.

O'Brien, J. D., Z. Iqbal, J. Wendler, and L. Amenga-Etego (2016, 06). Inferring strain mixture within clinical *plasmodium falciparum* isolates from genomic sequence data. *PLOS Computational Biology 12*(6), 1–20.

Ocholla, H., M. D. Preston, M. Mipando, A. T. R. Jensen, S. Campino, B. MacInnis, D. Alcock, A. Terlouw, I. Zongo, J.-B. Oudraogo, A. A. Djimde, S. Assefa, O. K. Doumbo, S. Borrmann, A. Nzila, K. Marsh, R. M. Fairhurst, F. Nosten, T. J. C. Anderson, D. P. Kwiatkowski, A. Craig, T. G. Clark, and J. Montgomery (2014). Whole-genome scans provide evidence of adaptive evolution in malawian *plasmodium falciparum* isolates. *The Journal of Infectious Diseases 210*(12), 1991–2000.

Pearson, R. D., R. Amato, S. Auburn, O. Miotto, J. Almagro-Garcia, C. Amaratunga, S. Suon, S. Mao, R. Noviyanti, H. Trimarsanto, J. Marfurt, N. M. Anstey, T. William, M. F. Boni, C. Dolecek, H. T. Tran, N. J. White, P. Michon, P. Siba, L. Tavul, G. Harrison, A. Barry, I. Mueller, M. U. Ferreira, N. Karunaweera, M. Randrianarivelojosia, Q. Gao, C. Hubbart, L. Hart, B. Jeffery, E. Drury, D. Mead, M. Kekre, S. Campino, M. Manske, V. J. Cornelius, B. MacInnis, K. A. Rockett, A. Miles, J. C. Rayner, R. M. Fairhurst, F. Nosten, R. N. Price, and D. P. Kwiatkowski (2016). Genomic analysis of local variation and recent evolution in *plasmodium vivax*. *Nature Genetics 48*, 959–964.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pf3k Consortium (2016). The pf3k project: pilot data release 5. accessed 17 July 2018.

Smith, T., J. Charlwood, J. Kihonda, S. Mwankusye, P. Billingsley, J. Meuwissen, E. Lyimo, W. Takken, T. Teuscher, and M. Tanner (1993). Absence of seasonal variation in malaria parasitaemia in an area of intense seasonal transmission. *Acta Tropica 54*(1), 55 – 72.

Steenkeste, N., W. O. Rogers, L. Okell, I. Jeanne, S. Incardona, L. Duval, S. Chy, S. Hewitt, M. Chou, D. Socheat, F.-X. Babin, F. Ariey, and C. Rogier (2010, Apr). Sub-microscopic malaria cases and mixed malaria infection in a remote area of high malaria endemicity in rattanakiri province, cambodia: implication for malaria elimination. *Malaria Journal 9*(1), 108.

Trevino, S. G., S. C. Nkhoma, S. Nair, B. J. Daniel, K. Moncada, S. Khoswe, R. L. Banda, F. Nosten, and I. H. Cheeseman (2017). High-resolution single-cell sequencing of malaria parasites. *Genome Biology and Evolution 9*(12), 3373–3383.

Wendler, J. (2015). *Accessing complex genomic variation in Plasmodium falciparum natural infection*. Ph. D. thesis, University of Oxford.

Wong, W., A. D. Griggs, R. F. Daniels, S. F. Schaffner, D. Ndiaye, A. K. Bei, A. B. Deme, B. MacInnis, S. K. Volkman, D. L. Hartl, D. E. Neafsey, and D. F. Wirth (2017, Jan). Genetic relatedness analysis reveals the cotransmission of genetically related *plasmodium falciparum* parasites in thiès, senegal. *Genome Medicine 9*, 5.

Zhu, S. J., J. Almagro-Garcia, and G. McVean (2018). Deconvolution of multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics 34*, 9–15.

Zimmerman, P. A., R. K. Mehlotra, L. J. Kasehagen, and J. W. Kazura (2004, 2018/02/05). Why do we need to know more about mixed *plasmodium* species infections in humans? *Trends in Parasitology 20*(9), 440–447.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

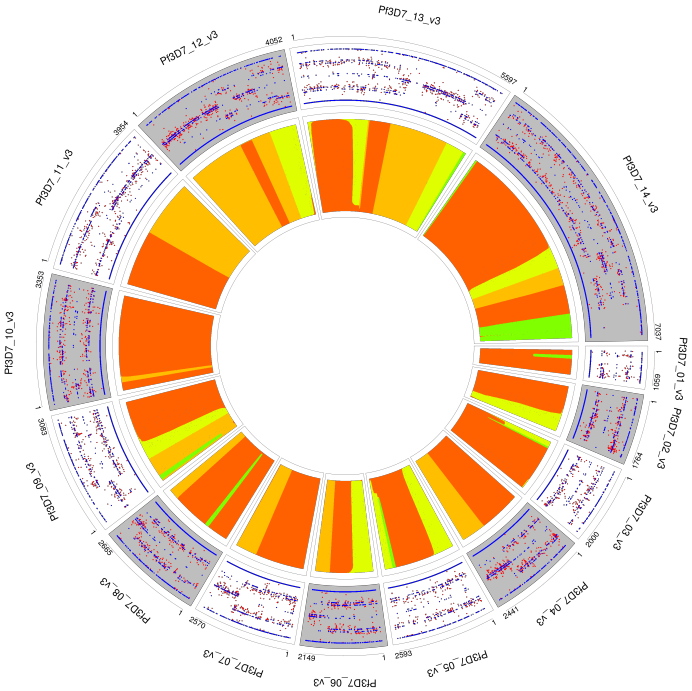


**Figure 1–Figure supplement 1.** Whole genome deconvolution of field sample PD0577-C. The outer ring shows the expected within-sample allele frequency (WSAF) (blue) and observed WSAF (red) across the genome. Red and blue points indicate observed and expected allele frequencies within the isolate. The inner ring indicates the IBD states among the three strains: green segments indicate where all three strains are IBD; yellow, orange and dark orange segments indicate the regions where one pair of strains are IBD but the others are not. In no region are all three strains inferred to be distinct, suggesting that all three strains are siblings.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

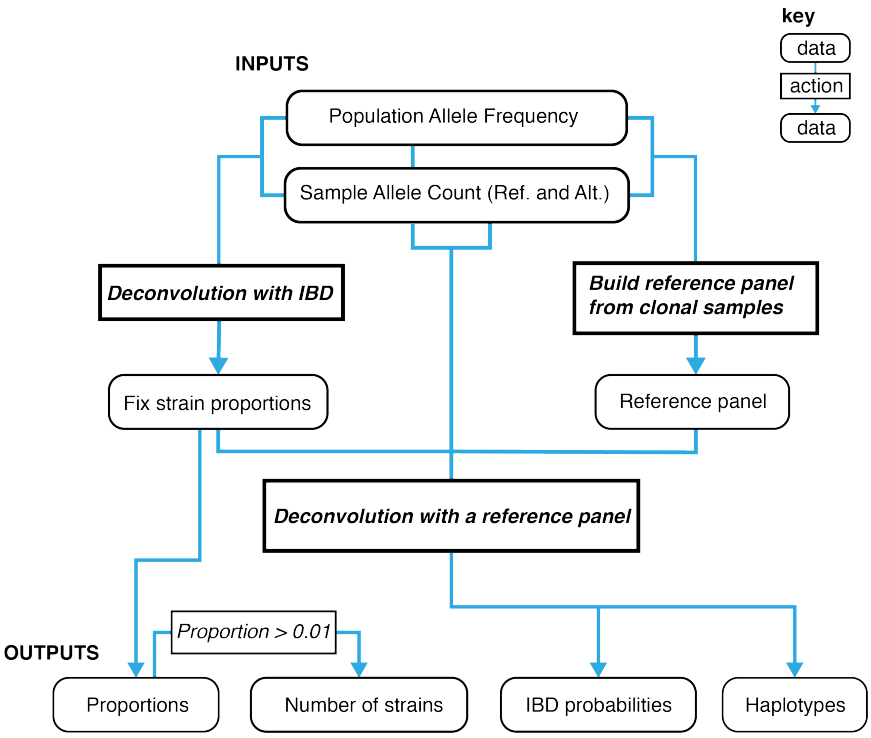


**Figure 1–Figure supplement 2.** A graphical overview of the data types and work flows for `DEploidIBD`. The boxes at the bottom represent final outputs of the pipeline. The rectangular boxes indicate when `DEploidIBD` is executed, with inputs highlighted by blue arrows. The process has three key steps: Step 1. A reference panel for the set of samples is constructed from high confidence clonal haplotypes, either identified from within a study or from an external resource, such as Pf3k. Step 2: `DEploidIBD`, using population level allele frequencies, is used to infer the number of strains, strain proportions and IBD profile within each sample. Step 3: `DEploidIBD` is re-run on each sample to infer haplotypes, but with the proportions estimated in Step 2 fixed and this time using the haplotype (LD-aware) method previously implemented in `DEploid`.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

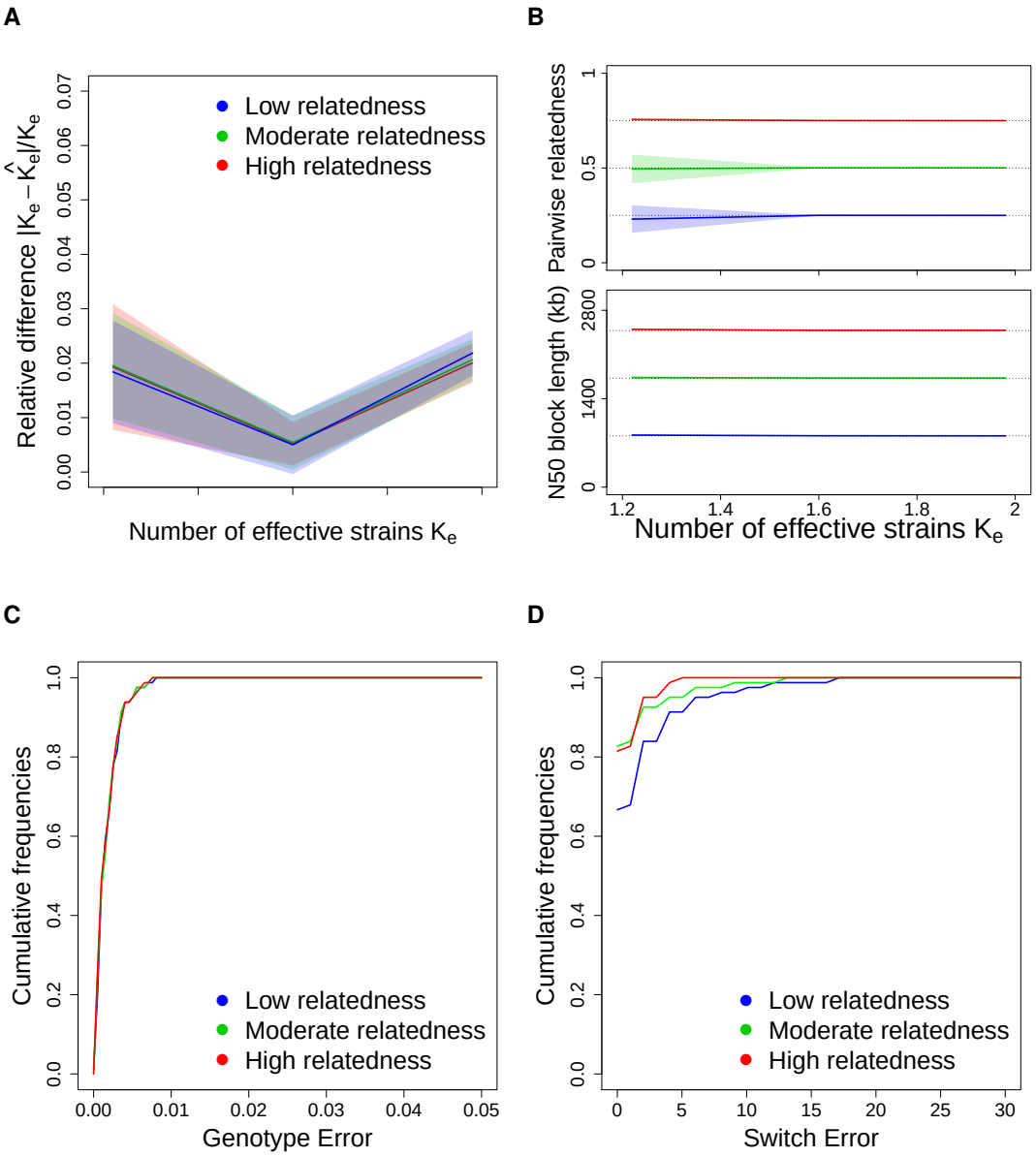


**Figure 2–Figure supplement 1.** Performance of `DEploidIBD` on 81 *in silico* mixtures of two strains from Africa using 92,780 sites from Chromosome 14. (A) Relative differences of inferred effective number of strains. (B) Inferred pairwise relatedness and N50 IBD tract length. Dotted lines mark parameters used in the simulation. We performed 100 simulations, but excluded eight cases that samples with very low coverage (below 13) and another 11 cases that sample coverage is between 13 and 29, and the reference panel contains at least one low confidence haplotype. (C) Cumulative distribution of average per site genotyping error for three levels of IBD (25%, 50% and 75%). (D) Cumulative distribution of haplotype switch error for the same three levels of IBD.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

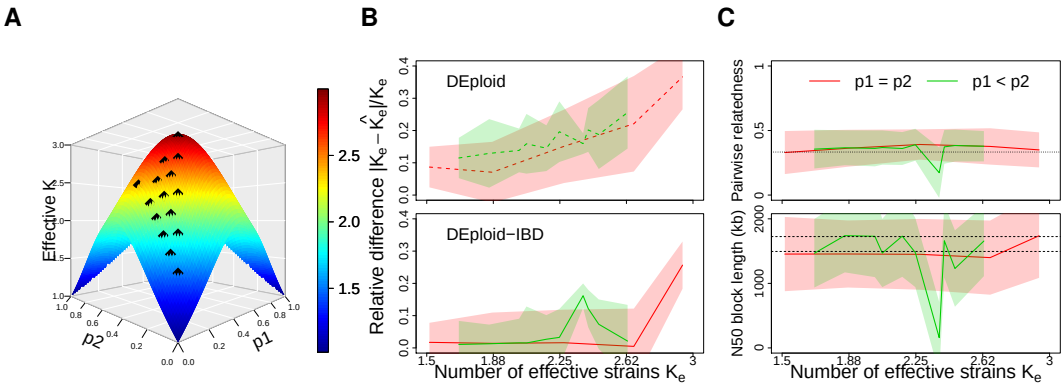


**Figure 2–Figure supplement 2.** Performance of `DEploidIBD` and `DEploid` on *in silico* mixtures of three strains from Asia. $p_1$ and $p_2$ denote the proportions of the minor strains 1 and 2 respectively. The third strain, with proportion of $1 - p_1 - p_2$, always has the largest proportion. We consider two scenarios: five cases with $p_1 = p_2$ in the range $0.1 \leq p_1 \leq 0.35$) and 10 cases with $p_1 < p_2$: (0.10, 0.15, 0.75), (0.10, 0.20, 0.70), (0.10, 0.25, 0.65), (0.15, 0.20, 0.65), (0.10, 0.30, 0.60), (0.15, 0.25, 0.60), (0.10, 0.40, 0.50), (0.15, 0.30, 0.55), (0.20, 0.25, 0.55) and (0.20, 0.30, 0.50). (A) Illustration of the mixture proportion profile over the $K_e$ surface. Let red and green color represent scenarios that $p_1 = p_2$ and $p_1 < p_2$. (B) Relative difference of inferred effective number of strains using `DEploid` and `DEploidIBD`. (C) Inferred pairwise relatedness and N50 IBD tract length. Dashed lengths indicate the tract lengths of the first and the second 50% of the SNPs.

bioRxiv preprint doi: https://doi.org/10.1101/387266; this version posted August 9, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

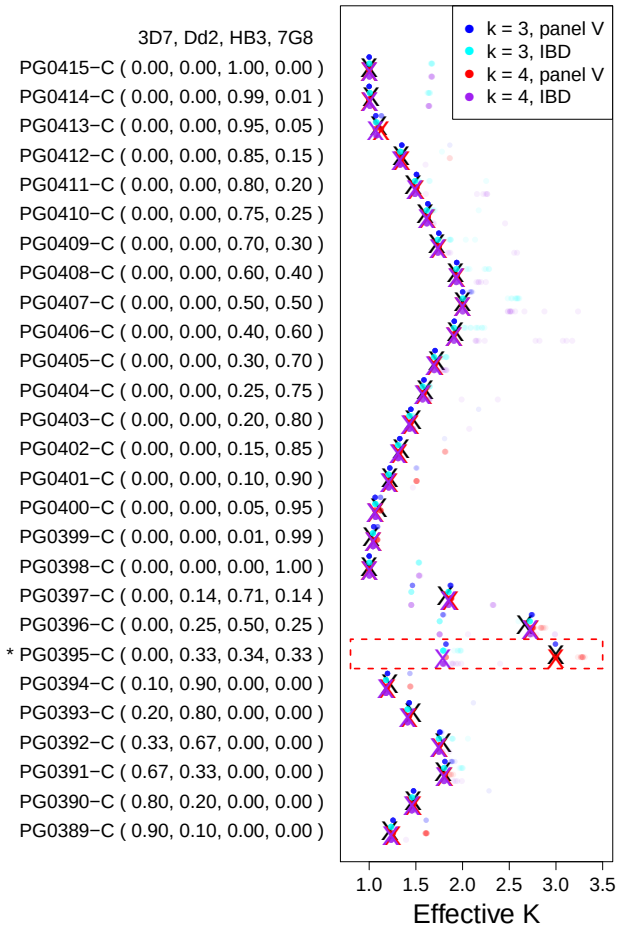


**Figure 2–Figure supplement 3.** Validation of `DEploidIBD` using 27 experimental lab mixtures. A reference panel of the laboratory strains (3D7, Dd2, HB3 and 7G8; Panel V) was used to deconvolve samples with `DEploid`. Each experiment is performed with and without IBD inference and with two values of the maximum number of strains (3 and 4). Black crosses indicate the true effective number of strains. Coloured crosses indicate median values obtained from 30 replicates using the algorithm/strain number indicated in the legend. The coloured dots show the inferred effective number of strains across replicates with intensity proportional to fraction. Note one sample (indicated by an asterisk and dotted red line) where balanced proportions of three strains results in the LD-free approach fitting the data as a mixture of two strains with proportions of $1/3$ and $2/3$.