Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

OXFORD

???Genome analysis

DEploid: Untangling complexity of infection in *Plasmodium falciparum*.

Sha Joe Zhu [1,*], Jacob Almagro-Garcia [1] and Gil McVean [1,2,*]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
[2] Big data institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7BN, UK

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Complexity of infection in the malarial parasite *Plasmodium falciparum* affects key phenotypic traits, including drug resistance and risk of severe disease. Advances in protocols and sequencing technology have made it possible to obtain high-coverage genome-wide sequence data from blood samples taken in the field. However, analysing and interpreting such data is challenging because of the high rate of multiple infections present.
Results: The software package *DEploid* deconvolutes sequences of mixed samples by learning haplotype structure from a reference panel of clonal isolates. It reports the number of strains, their relative proportions and their haplotypes present in an isolate, allowing researchers to study complexity of infection in malaria with an unprecedented level of detail.
Availability and implementation: The open source implementation *DEploid* is freely available at https://github.com/mcveanlab/dEploid under the conditions of the GPLv3 license. An R version is available at https://github.com/mcveanlab/DEploid-r.
Contact: joe.zhu@well.ox.ac.uk or mcvean@well.ox.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Malaria remains one of the top global health problems. Transmitted by mosquitoes of the genus *Anopheles*, the majority of malaria related deaths are caused by the *Plasmodium falciparum* parasite (WHO, 2016). Patients are often infected with more than one parasite strain, due to bites from multiple mosquitoes, mosquitoes carrying multiple genetic types or a combination of both. Multiplicity of infection can lead to competitions among co-existing strains and may influence disease development (de Roode et al., 2005), transmission rates (Arnot, 1998) and even the spread of drug resistance (de Roode et al., 2004).

The presence of multiple strains of *P. falciparum* makes fine scale analysis of genetic variation challenging, since genetic differences between the genetic types of this haploid organism will appear as heterozygous loci. Mixed calls also confound methods that exploit haplotype data to detect, among other phenomena, the occurrence of natural selection or recent demographic events (Harris and Nielsen, 2013; Lawson et al., 2012; Mathieson and McVean, 2014; Sabetil et al., 2002). In light of these difficulties, researchers usually focus on clonal infections or resort to heuristics methods for resolving heterozygous genotypes. The former approach discards valuable information regarding genetic diversity and inbreeding whereas the latter tends to create chimeric haplotypes that are not suitable for analysis, unless mixed calls are very sparse.

Phasing or deconvoluting the strains of a mixed infection is a harder problem than phasing diploid organisms because the levels of mixture within isolates (i.e. the abundance of each genetic type) vary greatly and are unknown. Existing tools for phasing diploid organisms, such as Beagle (Browning and Browning, 2007) and IMPUTE2 (Howie et al., 2009), are not designed to cope with this. Galinsky et al. (2015) and O'Brien et al. (2015) have attempted to address the mixed infection problem by inferring the mixed proportions from allele frequencies within samples, since they do not provide haplotypes and their models have limited use.

As part of the Pf3k project (Pf3k, 2016), an effort to map the genetic diversity of *P. falciparum* at global scale, we have developed DEploid, a software package for deconvoluting mixed infections. The program provides estimates for the number of different genetic types present in the isolate, the proportion or abundance of each strain and their sequences (i.e. haplotypes). To our knowledge, DEploid is the first package able

to deconvolute strain haplotypes and provides a unique opportunity for researchers to study inbreeding patterns and complexity of infection, leaving open the possiblity to investigate infection history at fine scale.

2 Methods

2.1 Notations

Let's first introduce some notation (see Table 1). Suppose that our data D are the allele read counts of sample j at a given site i, denoted as $r_{j,i}$ and $a_{j,i}$ for reference (REF) and alternative (ALT) alleles respectively. The allele frequencies within sample (WSAF) $p_{j,i}$ and at population level (PLAF) f_i can be calculated by $\frac{a_{j,i}}{a_{j,i}+r_{j,i}}$ and $\frac{\sum_j a_{j,i}}{\sum_j a_{j,i}+\sum_j r_{j,i}}$. Since all data in this section refers to the same sample, we drop the subscript j from now on.

i	Marker index
j	Sample index
r	Read count for reference allele
a	Read count for alternative allele
f	Population level allele frequency (PLAF)
n	Number of strains within sample
l	Sequence length
$\mathbf{w}$	Proportions of strains
$\mathbf{h}_i$	allelic states of n parasite strains at site i
$h_{k,i}, h_k$	allelic states of k parasite strains k at site i
p	Observed within sample allele frequency (WSAF)
q	Unadjusted expected WSAF
π	Adjusted expected WSAF
Ξ	Reference panel
g_p	allelic states of reference strain p
e	Probability of read error

Table 1. Table summarising the notation we use in this article.

2.2 Model

We descibe the mixed infection problem by considering the number of strains n, the relative abundance of each strain $\mathbf{w}$ and their allelic states $\mathbf{h}$. Similar to O'Brien et al. (2015), we use a Bayesian approach and define the posterior probabilities of n, $\mathbf{w}$ and $\mathbf{h}$ for some reference panel Ξ and the read error rate e given the data as:

$$P(n, \mathbf{w}, \mathbf{h}, \Xi, e|D) \propto L(n, \mathbf{w}, \mathbf{h}, \Xi, e|D) \times P(n, \mathbf{w}, \mathbf{h}, \Xi, e).$$

We assume that the proportion of parasite strains $\mathbf{w}$ given n strains, the chosen reference panel Ξ and sequencing error rate are independent from each other, leading to:

$$\begin{aligned} P(n, \mathbf{w}, \mathbf{h}, \Xi, e|D) \propto & L(n, \mathbf{w}, \mathbf{h}, \Xi, e|D) \times P(\mathbf{h}|\mathbf{w}, \Xi, e) \times \\ & P(\mathbf{w}|n) \times P(n) \times P(\Xi) \times P(e). \end{aligned} \tag{1}$$

Since we are not making inference on the referenece panel nor the error rate, the prior probabilities are fixed. For convience, we fix the number of strains n and simplify Eqn. (2) as:

$$P(\mathbf{w}, \mathbf{h}, \Xi, e|D) \propto L(\mathbf{w}, \mathbf{h}, \Xi, e|D) \times P(\mathbf{h}|\mathbf{w}, \Xi, e) \times P(\mathbf{w}|n). \tag{2}$$

2.2.1 Likelihood of the data given the WSAF

Let $\mathbf{w} = [w_1, \ldots, w_n]$ and $\mathbf{h}_i = [h_{1,i}, \ldots, h_{n,i}]$ denote the proportions and alleic state of n parasite strains at site i. We use O'Brien et al. (2015)'s expression for the expected WSAF q_i as:

$$q_i = (\mathbf{w} \cdot \mathbf{h}_i) = \sum_{k=1}^{n} w_k \cdot h_{k,i}. \tag{3}$$

Thus, the likelihood of $L(\mathbf{w}, \mathbf{h}, \Xi, e|D)$ in Eqn. (2) becomes $L(q_i, \Xi, e|D)$.

We adjust the allele frequency q_i by taking into account the read error rate e. This implies that the expected allele frequency of 'REF' read as 'ALT' is $(1 - q_i)e$, and the expected allele frequency of 'ALT' read as 'REF' is $q_i e$. Thus, the adjusted WSAF becomes:

$$\pi_i = q_i + (1 - q_i)e - q_i e = q_i + (1 - 2q_i)e. \tag{4}$$

We model over-dispersion in read counts relative to the Binomial using a Beta-binomial distribution. Specifically, the read counts of 'ALT' are identically and independently distributed (i.i.d.) with probability v_i, i.e. $a_i \sim Binom(a_i + r_i, v_i)$, and $v_i \sim Beta(\alpha, \beta)$, where $v_i = \alpha/(\alpha + \beta)$. Let $\alpha = c \cdot \pi_i$, $\beta = c \cdot (1 - \pi_i)$ and π_i is calculated with π_i given by Eqn. (4), we can formalise the likelihood of the data using:

$$L(q_i, \Xi, e|D) \propto \frac{\Gamma(a_i + c \cdot \pi_i)\Gamma(r_i + c \cdot (1 - \pi_i))}{\Gamma(c \cdot \pi_i)\Gamma(c \cdot (1 - \pi_i))}. \tag{5}$$

2.2.2 Priors

- We use a multinormal prior for the proportions $\mathbf{w}|n$: A multivariate normal variable titre $\mathbf{x}|n = [x_1, \ldots, x_n]$ is proposed, where x_i is i.i.d. normally distributed from $N(\eta, \sigma^2)$. We then compute w_k as

  $$w_k = \frac{\exp(x_k)}{\sum_{k=1}^{n} \exp(x_k)}.$$

  Since x_is are i.i.d. normally distributed. The prior $P(\mathbf{w}|n) = \prod_{k=1}^{n} \Phi(x_i)$, where $\Phi(x_i)$ is the distribution function of x_i.
- For the prior on $\mathbf{h}|\omega, \Xi$, we use Li and Stephens (2003)'s hidden Markov model framework as a starting point, with the following modifications made:

  * likelihood of data given the expected WSAF rather than the "product of approximate conditionals" (PAC).
  * multiple strains with variable proportion rather than two sequences with strict diploidity.
  * simplifying the mutation model with a fixed miss copying operation.

  Let ξ_k denote the path in the reference panel Ξ that h_k is copying from; and μ denote the probability of miss copying:

  $$\begin{cases} P(\xi_k = h_k) & = 1 - \mu, \\ P(\xi_k \neq h_k) & = \mu, \end{cases}$$

  We can therefore express the prior on each h_k as

  $$\begin{aligned} P(h_k|\mathbf{w}, \Xi, e) = & P(\xi_k = a|\mathbf{w}, \Xi, e)(1 - \mu) + \\ & P(\xi_k = 1 - a|\mathbf{w}, \Xi, e)\mu, \end{aligned} \tag{6}$$

  where $a \in \{0, 1\}$, and $1 - a$ indicates the event that allelic state h_k differs from ξ_k. We can therefore transform the prior on $\mathbf{h}_k$, and express it as a prior on ξ_k, and the prior of the path is simply the transmission probabilities between two loci of a given recombination model. For position $i > 1$, let ρ_i' denote the probability of **no** recombination from site $i - 1$ to i, we have $\rho_i' = \exp(-\psi_i)$, where ψ is the recombnation rate. Thus, the probability of recombining from any strain in the panel is $\frac{1 - \rho_i'}{|\Xi|}$, where $|\Xi|$ is the size of the panel.

2.3 Inference

Overall, our method generates Markov chain Monte Carlo (MCMC) samples for the proportions $\mathbf{w}$ and the haplotypes $\mathbf{h}$ given a fixed number of strains. As for the MCMC updates on the prior, we use a Metropolis-Hastings algorithm to sample proportions ($\mathbf{w}$) given $\mathbf{h}$; and use a Gibbs sampler to update $\mathbf{h}$ for a given $\mathbf{w}$, which are further divided into cases of updating only one haplotype and updating a pair of haplotypes.

2.3.1 Metropolis-Hastings update for proportions

We use a sparse update on $\mathbf{w}|n$, by updating $\mathbf{x}|n$. More specifically, we propose new x_i's from $x_i' = x_i + \delta x$, where $\delta x \sim N(0, \sigma^2/s)$, and s is a scaling factor. The new proposed proportion is therefore $\frac{\exp(x_k')}{\sum_{k=1}^{n}\exp(x_k')}$. Since the proposal distribution is symmetrical, we have hasting ratio of 1. A new update is accepted with probability:

$$\min\left(1, \frac{P(\mathbf{w}'|n)}{P(\mathbf{w}|n)}\frac{L(\mathbf{w}', \mathbf{h}, \Xi, e|D)}{L(\mathbf{w}, \mathbf{h}, \Xi, e|D)}\right).$$

2.3.2 Gibbs update for single haplotype

We choose haplotype strain s uniformly at random from K strains, considering both cases for updating the state of strain s at position i to 0 and 1, we compute the WSAF and its associated likelihood as follows: Regardless the state of strain s at position i, we first remove it from the current WSAF, i.e. subtract $w_s \cdot h_s$ from Eqn. (3), which gives

$$q_{i,-s} = \sum_{k\neq s} w_k \cdot h_k = \text{Eqn. (3)} - w_s \cdot h_s. \tag{7}$$

Therefore, updating the allelic state of strain s to 0 and 1, the expected WSAF becomes

$$q_{i,h_s=0} = \text{Eqn. (7)} \tag{8}$$

$$q_{i,h_s=1} = \text{Eqn. (7)} + w_s \times 1 \tag{9}$$

We substitute equations (8) and (9) into Eqn. (5) to compute the associated likelihood $L(q_{i,h_s}, \Xi, e|D)$, which is expressed as $L(h_s|D)$ in the rest of the paper. Therefore, Eqn. (2) becomes

$$P(h_s|D) \propto L(h_s|D) \times P(h_s|\mathbf{w}, \Xi, e). \tag{10}$$

Instead of computing $P(h_s|D)$ directly, we consider the posterior probabilities of the path p where h_s is copying from:

$$P(g_p|D) \propto L(g_p|D) \times P(g_p|\mathbf{w}, \Xi, e). \tag{11}$$

Similar to the prior on $\mathbf{h}|\Xi, e$ (see Eqn.(6)), we can compute the posterior probabilities of h_k by $P(g_p|D)$ via:

$$P(h_s = a\big|D) = \begin{cases} P(g_p = a\big|D) \cdot (1-\mu), & h_s = g_p; \\ P(g_p = 1-a\big|D) \cdot \mu, & h_s \neq g_p. \end{cases} \tag{12}$$

where $a \in \{0, 1\}$, and $1 - a$ indicates the event that allelic state h_k differs from g_p. Note that g_p is the same as ξ_k, but indexed differently: g_p is indexed by the reference strain index p from the panel.

In addition to updating the haplotypes from the panel, we take into account misscopying (see example shown in Fig. 1), which allows the actual genotype to differ from the path. Therefore, this method favours mutation rather than recombination for a single-site difference from a reference strain. Our model benefits from combining information from both the reference haplotypes as well as the data. For *de novo* mutations which are not found reference panel, our method will infer mutations based on read count in the data. Our method proceeds according to the following steps:

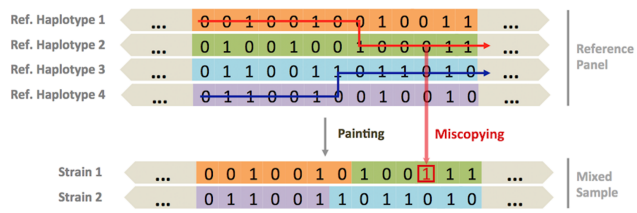

Fig. 1. Illustration of our adaptation of Li and Stephens (2003)'s algorithm. Strain 1 haplotype is made up from reference haplotype segments of 1 and 2; and strain 2 haplotype is made up from reference haplotype segments of 3 and 4. With miss copying, we allow strain states differ from the path: At the third last position of strain 1, the path is copied from reference haplotype 2, with the state of "0".

1. Consider the likelihood as the emission probabilities at site i. Let's use g_p and h_s to denote the allelic state of the copied path and the updated strain respectively.

$$\begin{aligned} L(g_p = a|D) \;=\; & L(h_s = a\big|D) \times P(g_p = h_s) + \\ & L(h_s = 1-a\big|D) \times P(g_p \neq h_s) \end{aligned} \tag{13}$$

   where $a \in \{0, 1\}$, and $1 - a$ indicates the event that h_s takes value that differs from g_p.
2. For the prior on $g_p|\mathbf{w}, \Xi, e$, it is simply the Li and Stephens (2003) HMM transition probabilities from position $i-1$ to i: the probabilites ρ_i' and $\frac{1-\rho_i'}{|\Xi|}$ reflect no recombination and recombination events respectively. Therefore, we have

$$P(g_p|\mathbf{w}, \Xi, e) = \rho_i' \cdot P_{i-1}(g_p\big|D) + \frac{1-\rho_i'}{|\Xi|} \cdot \sum_{x\in\Xi} P_{i-1}(g_x\big|D). \tag{14}$$

   Combine equations (13) and (14), the posterior probability of path (reference strain) p at position i is then obtained (Eqn. (11)).
3. Given the posterior probabilities of path (reference strain) p at position i, we sample the path backwards to generate a MCMC sample for the path that strain s copied from. We start at the end position, and sample/propose recombination events with probabilities proportional to

$$\begin{cases} \rho_i' \cdot P_i(g_p\big|D) & \text{no recombination,} \\ (1-\rho_i') \cdot \sum_{x\in\Xi} P_{i-1}(g_x\big|D) & \text{recombination,} \end{cases}$$

   if a recombination event is chosen, we sample the path backwards according $P_i(g_p\big|D)$.
4. Finally, we generate MCMC samples of the haplotype using the probabilities expressed in Eqn.(12).

2.3.3 Gibbs update for a pair of haplotypes

In order to improve the MCMC mixing, we consider pairs of haplotypes and update both strain simultaneously. Suppose random sampling two strains to update, namely, s_1 and s_2. Similarly to Eqn. (7), we first remove their states from the WSAF:

$$\begin{aligned} q_{i,-s_1,-s_2} \;&=\; \sum_{k\neq s_1,s_2} w_k \cdot h_k \\ &=\; \text{Eqn. (3)} - w_{s_1} \cdot h_{s_1} - w_{s_2} \cdot h_{s_2} \end{aligned} \tag{15}$$

Considering all four possible combination of genotypes, we have

$$q_{i,h_{s_1}=0,h_{s_2}=0} = \text{Eqn. (15)} \tag{16}$$

$$q_{i,h_{s_1}=0,h_{s_2}=0} = \text{Eqn. (15)} + \cdot w_{s_1} \times 1 \tag{17}$$

$$q_{i,h_{s_1}=0,h_{s_2}=1} = \text{Eqn. (15)} + \cdot w_{s_2} \times 1 \tag{18}$$

$$q_{i,h_{s_1}=0,h_{s_2}=1} = \text{Eqn. (15)} + \cdot w_{s_1} \times 1 + w_{s_2} \times 1 \tag{19}$$

Substitute expressions. (16) to (19), into Eqn. (5), we then obtain their associated likelihood $L(q_{i,h_{s_1},h_{s_2}}, \Xi, e|D)$, which is denoted as $L(h_{s_1}, h_{s_2}|D)$ in the rest of the paper. Thus, similar to Eqn. (10), we obtain the following posterior probability, and sample the state (genotype) of strains s_1 and s_2 simultaneously at site i:

$$P(h_{s_1}, h_{s_2}|D) \propto L(h_{s_1}, h_{s_2}|D) \times P(h_{s_1}, h_{s_2}|\mathbf{w}, \Xi, e). \tag{20}$$

As in the previous section, instead of working directly with posterior probabilities $P(h_{s_1}, h_{s_2}|D)$, we consider the posterior probabilites of the path p_1 and p_2 where s_1 and s_2 are copying from respectively:

$$P(g_{p_1}, g_{p_2}|D) \propto L(h_{p_1}, h_{p_2}|D) \times P(g_{p_1}, g_{p_2}|\mathbf{w}, \Xi, e). \tag{21}$$

Then we can compute the posterior probabilities of the pair h_{s_1} and h_{s_2} from

$$P\left(\begin{smallmatrix} h_{s_1}=a, \\ h_{s_2}=b \end{smallmatrix}\middle|D\right) = \begin{cases} P\left(\begin{smallmatrix} g_{p_1}=a, \\ g_{p_2}=b \end{smallmatrix}\middle|D\right)\cdot(1-\mu)\cdot(1-\mu), \\ P\left(\begin{smallmatrix} (g_{p_1}=a, \\ g_{p_2}=1-b \end{smallmatrix}\middle|D\right)\cdot(1-\mu)\cdot\mu, \\ P\left(\begin{smallmatrix} (g_{p_1}=1-a, \\ g_{p_2}=b \end{smallmatrix}\middle|D\right)\cdot\mu\cdot(1-\mu), \\ P\left(\begin{smallmatrix} (g_{p_1}=1-a, \\ g_{p_2}=1-b \end{smallmatrix}\middle|D\right)\cdot\mu\cdot\mu, \end{cases} \tag{22}$$

where $a, b \in \{0, 1\}$, and $1-a$ indicates the event that allelic state h_{s_1} differs from g_{p_1}, and $1-b$ indicates the event that allelic state h_{s_2} differs from g_{p_2}.

specifically, we carry out the following steps:

1. Compute the likelihood of the allelic state pairs of the path given data:

$$\begin{aligned} L\left(\begin{smallmatrix} g_{p_1}=a, \\ g_{p_2}=b \end{smallmatrix}\middle|D\right) =& L\left(\begin{smallmatrix} h_{s_1}=a, \\ h_{s_2}=b \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}=h_{s_1}, \\ g_{p_2}=h_{s_2} \end{smallmatrix}\right) + \\ & L\left(\begin{smallmatrix} h_{s_1}=a, \\ h_{s_2}=1-b \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}=h_{s_1}, \\ g_{p_2}\neq h_{s_2} \end{smallmatrix}\right) + \\ & L\left(\begin{smallmatrix} h_{s_1}=1-a, \\ h_{s_2}=b \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}\neq h_{s_1}, \\ g_{p_2}=h_{s_2} \end{smallmatrix}\right) + \\ & L\left(\begin{smallmatrix} h_{s_1}=1-a, \\ h_{s_2}=1-b \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}\neq h_{s_1}, \\ g_{p_2}\neq h_{s_2} \end{smallmatrix}\right) \end{aligned} \tag{23}$$

where

$$\begin{aligned} P(g_{p_1}=h_{s_1}, g_{p_2}=h_{s_2}) &= (1-\mu)\cdot(1-\mu), \\ P(g_{p_1}\neq h_{s_1}, g_{p_2}=h_{s_2}) &= \mu\cdot(1-\mu), \\ P(g_{p_1}=h_{s_1}, g_{p_2}\neq h_{s_2}) &= \mu\cdot(1-\mu), \\ P(g_{p_1}\neq h_{s_1}, g_{p_2}\neq h_{s_2}) &= \mu\cdot\mu, \end{aligned}$$

and $a, b \in \{0, 1\}$, and $1-a$ indicates the event that allelic state h_{s_1} differs from g_{p_1}, and $1-b$ indicates the event that allelic state h_{s_2} differs from g_{p_2}.

2. For the prior of the allelic state pairs of the path, we consider all possible pair of the copying strain:

   - we take into account of the possibility of one strain recombines and the other does not with the probability of $\rho'_i \cdot \frac{1-\rho'_i}{|\Xi|}$;
   - both recombines, with the probability of $\rho'_i \cdot \rho'_i$;
   - neither recombines, with the probability of $\frac{1-\rho'_i}{|\Xi|} \cdot \frac{1-\rho'_i}{|\Xi|}$, assuming that recombination events of two copying strains are independent from each other.

   The prior for the path takes the form

$$\begin{aligned} P(g_{p_1}, g_{p_2}|\mathbf{w}, \Xi, e) =& P_{i-1}(g_{p_1}, g_{p_2}|D)\cdot\rho'_i\cdot\rho'_i + \\ & \sum_{x\in\Xi} P_{i-1}(g_{p_1}, g_x|D)\cdot\rho'_i\cdot\frac{1-\rho'_i}{|\Xi|} + \\ & \sum_{y\in\Xi} P_{i-1}(g_y, g_{p_2}|D)\cdot\rho'_i\cdot\frac{1-\rho'_i}{|\Xi|} + \\ & \sum_{x,y\in\Xi\cdot\Xi} P_{i-1}(g_x, g_y|D)\cdot\frac{1-\rho'_i}{|\Xi|}\cdot\frac{1-\rho'_i}{|\Xi|}. \end{aligned} \tag{24}$$

   Combining equations (23) and (24), we obtain Eqn. (21).
3. Given the posterior probabilities of path pair p_1, p_2 at position i, we sample the path backwards to generate a MCMC sample for the path pair that strain s_1 and s_2 copied from. We start at the end position, first propose if a recombination events have happened with probabilities proportional to

$$\begin{cases} \sum_{x\in\Xi} P_{i-1}(g_{p_1}, g_x|D)\cdot\frac{1-\rho'_i}{|\Xi|}\cdot\rho'_i, & p_1 \text{ recombines}, \\ P_{i-1}(g_{p_1}, g_{p_2}|D)\cdot\rho'_i\cdot\rho'_i, & \text{no recombination}, \\ \sum_{y\in\Xi} P_{i-1}(g_y, g_{p_2}|D)\cdot\frac{1-\rho'_i}{|\Xi|}\cdot\rho'_i, & p_2 \text{ recombines}, \\ \sum_{x,y\in\Xi\cdot\Xi} P_{i-1}(g_x, g_y|D)\cdot\frac{1-\rho'_i}{|\Xi|}\cdot\frac{1-\rho'_i}{|\Xi|}, & \text{both recombine}. \end{cases}$$

   If both strains recombine, we sample the path, according to $P_i(g_{p_1}, g_{p_2}|D)$. If one of the strains recombine, we sample the path according to the marginal probability of $P_i(g_p|D)$.
4. Finally, we generate MCMC samples of the haplotype pairs using the probabilities provided in Eqn.(22).

2.4 Implementation details

- **Fixed number of strains**. We assume there are more strains than actually present, starting the MCMC chain with a fixed n. As the proportion values drop, we "zero-out" "noisy" strains. For example, suppose that $n = 5$ and the inferred proportions $\mathbf{w} = [0.848, 6.36e-05, 6.81e-05, 0.152, 3.31e-05]$, we drop the strains with proportions less than 0.01, and keep the first and the forth strains.
- **Parameters used**. In practice, we set the parameters $c = 100$ (Eqn. (5)); $\eta = 0$, $\sigma^2 = 3$ and $s = 40$ (sections 2.2.2 and 2.3.1).
- **Update with linkage disequilibrium information**. We assume a uniform recombination map, genetic distances between loci i and $i+1$ are computed by $G_i = D_i/d_m$ where D_i denotes the physical distance between loci i and $i+1$ in nucleotides, d_m denotes the average genetic distance in morgans, (as average genetic distance for *P. Falciparum* we use 15,000 base pairs per centimorgan, in line with tthe estimates provided by Miles et al. (2016)). Suppose the recombination rate ψ_i is given by $\psi_i = N_e G_i$, with $N_e = 10$ being the effective

population size. In reality, Ne is just a scaling parameter (not the effective population size). Note that **we scale the probabilities with the number of haplotypes in the reference panel**.

- **Update without LD** When assume there is free recombination between two adjacent loci, the LD structure hardly play any role in the inference. In which case, we can assume independence between site i and $i+1$. In particular, we use the PLAF as the prior of $P(h_s|\mathbf{w}, \Xi, e)$ in Eqn. (10). Furthermore, we assume independence between s_1 and s_2 in Eqn. (20), we compute $P(h_{s_1}, h_{s_2}|\mathbf{w}, \Xi, e)$ as a product of $P(h_{s_1}|\mathbf{w}, \Xi, e)$ and $P(h_{s_2}|\mathbf{w}, \Xi, e)$, where $P(h_{s_*}|\mathbf{w}, \Xi, e) =$ PLAF.
- **Model selection** Since the final iteration of the MCMC is taken as a point estimate to infer the haplotypes and proportion, the deconvolution process is repeated with different random seeds. We use the lowest deviance information criterion (DIC) to select the best fit model (best chain for the same inferred n). The DIC is calculated from the generated MCMC simulation, and penalised by the average deviance. More specifically, we compute the deviance by $D_{\mathbf{w},\mathbf{h}} = -2\log(L(\mathbf{w}, \mathbf{h}|D))$ and $DIC = 2\bar{D} - D_{\mathbf{w},\mathbf{h}}$, where $\bar{D}$ is the average deviance of the MCMC chain.
- **Pf3k Deconvolution**. When deconvoluting the Pf3k(Pf3k, 2016) project data, we assume that clonal sample haplotypes capture the diversity of the haplotype structures of the local population. In particular, we deconvolute clonal samples without LD, and then use the deconvoluted clonal haplotypes as reference strains to deconvolute mixed samples.

3 Validation and Performance

As a validation set we used a set of *in vitro* mixtures created by Wendler (2015) to simulate mixed infections, which is an ideal validation data set for our purposes. In this data, DNA was extracted from four laboratory parasite lines: 3D7, Dd2, HB3 and 7G8, and mixed by different proportions (see Table 2 in brackets), and submitted to the MalariaGEN pipeline (MalariaGEN, 2008) for Illumina sequencing and genotyping (Manske et al., 2012).

Note that this data set only contains two clonal samples. Due to its limited size, they are not ideal for constructing a reference panel. Moreover, the *P. falciparum* genetic crosses project (Miles et al., 2016) found that due to sequencing error or applying different variant calling methods, genotype calls vary at the same locus for the same strain of *P. falciparum*. We therefore apply inference methods to mutiple samples that contains the same parasite strains, to infer the genotypes of a reference strain.

Inferring haplotypes for Dd2 strain. Since 3D7 is the reference strain, we assume that strain Dd2 is the only source of 'ALT' reads in samples PG0389-C to PG0394-C. Assuming markers are independent from each other, let y be the read count for 'ALT' allele and x be the total read count weighted by the Dd2 mixing proportion (see Table 2 in brackets), we use a regression model ($y = \beta_0 + \beta_1 x$) to infer the Dd2 genotype: 1 if β_1 is significant with p-values below 0.001; 0 otherwise.

Inferring haplotypes for HB3 and 7G8. Similarly, for samples PG0398-C to PG0415-C, we let variables x_1, x_2 be the coverages weighted by the mixing proportions of HB3 and 7G8 respectively; we use a regression model ($y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$) to infer the genotypes of HB3 and 7G8: HB3 is 1 if β_2 is significant with p-values below 0.001; 0 otherwise; similarly for 7G8.

sample	3D7	Dd2	HB3	7G8
PG0389-C	88.5 (90)	11.5 (10)	0	0
PG0390-C	79.8 (80)	20.2 (20)	0	0
PG0391-C	66.1 (67)	33.9 (33)	0	0
PG0392-C	31.2 (33)	68.8 (67	0	0
PG0393-C	18.4 (20)	81.6 (80)	0	0
PG0394-C	9.1 (10)	90.1 (90)	0	0
PG0395-C	0	33.6 (33.3)	35 (33.3)	31.3 (33.3)
PG0396-C	0	25.9 (25)	26.1 (25)	48 (50)
PG0397-C	0	14.7 (14.3)	15.3 (14.3)	69.9 (71.4)
PG0398-C	0	0	45.1+54.9 (100)	0
PG0399-C	0	0	56.7+40.9 (99)	2.4 (1)
PG0400-C	0	0	39.5+57.5 (95)	3 (5)
PG0401-C	0	0	33.3+56.7 (90)	10 (10)
PG0402-C	0	0	85.2 (85)	14.8 (15)
PG0403-C	0	0	80.1 (80)	19.3 (20)
PG0404-C	0	0	75.4 (75)	24.6 (25)
PG0405-C	0	0	70.6 (70)	29.4 (30)
PG0406-C	0	0	61 (60)	39 (40)
PG0407-C	0	0	50.5 (50)	49.5 (50)
PG0408-C	0	0	40.1 (40)	59.2 (60)
PG0409-C	0	0	30.1 (30)	69.1 (70)
PG0410-C	0	0	25.9 (25)	73,4 (75)
PG0411-C	0	0	21.4 (20)	78.5 (80)
PG0412-C	0	0	15.2 (15)	84.8 (85)
PG0413-C	0	0	3.8 (5)	96.2 (95)
PG0414-C	0	0	0 (1)	29.9+70.1 (99)
PG0415-C	0	0	0	30.0+70.0 (100)

Table 2. Inferred percentages (true in brackets) of the mixed samples. Our model overfits clonal sample PG0398-C as a mixture of two strains each with the abundance proportions of 45.1% and 54.9%. We then find that both inferred haplotypes are referring to the same strain HB3. Hence we use the "+" sign to denote the overfitting events, and both haplotypes are inferring the same strain.

3.1 Accuracy

3.1.1 Proportions and number of strains

To validate our method we applied our program DEploid to 27 lab-mixed *in vitro* samples. As described in section 2.3.1, we start by assuming at most three strains present in the mixtures; and discard strains with less than 1% relative abundance. Our method successfully recovers the proportions with haplotypes of the input (see Table 2). The deviation between our proportion estimates and the truth is at most 2%.

Our model overfits the noisy lab-mixed sample with additional strains. Note that in Table 2, we infer six of the HB3 and 7G8 mixtures as mixing of three, two of which haplotypes are in fact inferring the same parasite line, but are seperated because of a few heterozygous sites with high coverage resulting in high leverage in our model (supplemental material Figure S2.3(a)). The noisy markers are not recalabriated by our program.

We experimented deconvoluting the 27 lab-mixed samples with the following reference panels:

- panel I: five Asian and five African clonal strains from the Pf3k(Pf3k, 2016) data base: PD0498-C, PD0500-C, PD0660-C, PH0047-Cx, PH0064-C, PT0002-CW, PT0007-CW, PT0008-CW, PT0014-CW, PT0018-CW.
- panel II: panel I with the addition of HB3;
- panel III: panel II with the addition of 7G8;
- panel IV: panel III with the addition of Dd2;
- panel V: 3D7, HB3, 7G8 and Dd2 strains.

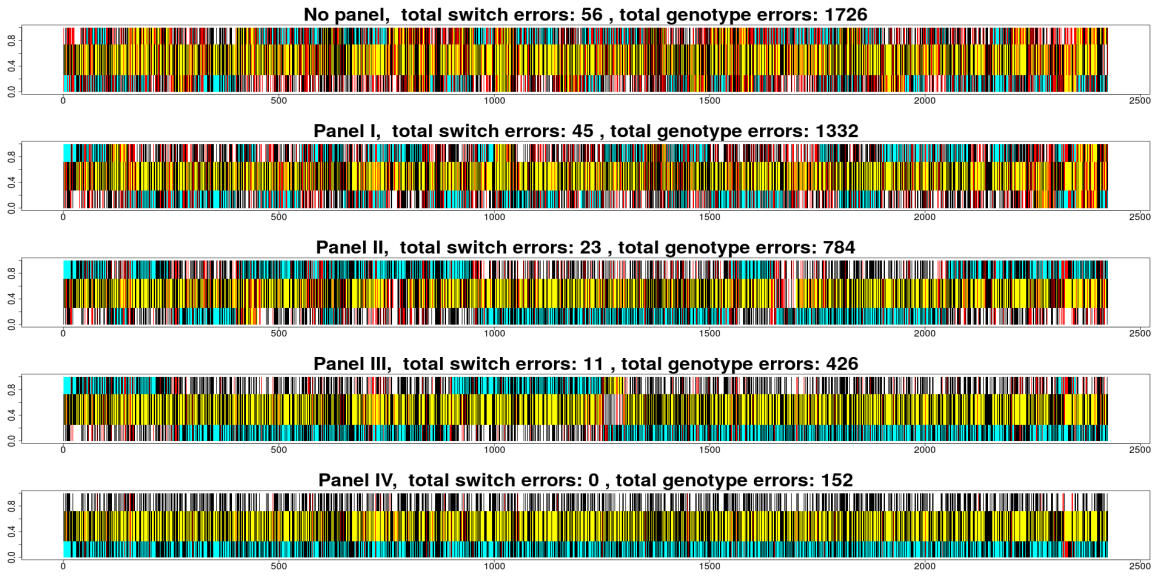

Fig. 2. Haplotypes comparison of sample PG0396-C chromosome 14 deconvolution without any reference strain (top) versus with using reference panels I to IV (from the second to the bottom). Black bars indicate alternative alleles; red bars mark wrongly inferred positions. The yellow, cyan and white background label the haplotype segments from strains 7G8, HB3 and Dd2 respectively. The switch errors are obtained by counting the changes of a strain segment mapped to reference strains; the genotype errors are the discordance between the strain and the mapped reference segments.

- panel VI: panel I with the addition of six (three each) clonal strains from Asia and Africa: PH0193-C, PH0283-C, PH0305, PT0060-C, PT0146-C and PT0158-C.

In all cases we estimated the number and proportion of strains accurately, for example Figure 2 y-axises show the proportions of strains Dd2/7G8/HB3 as approximately $\frac{1}{4}/\frac{1}{2}/\frac{1}{4}$.

3.1.2 Haplotypes

Our accuracy assessment for inferred haplotypes takes into account switch errors and genotype discordance, which reflects recombination and misscopying events in the method Section. Intuitively, one may think of using the Li and Stephens model to compute the viterbi path or posterior probabilities to assess how our inferred haplotypes differ from the truth. However, we find that such methods overestimate switch errors at short segments due to variable reference panel quality.

Note that in Figure 2, the top and bottom strains have similar proportions, which are difficult to phase without a perfect reference panel. One may flip parts of the two strains to resolve the errors. Unfortunately, our Li and Stephens implementation focuses on a single strain (two strains at the most), the viterbi path or posterior probabilities fail to align the switches of all strains. Instead, we have taken a heuristic approach: we first divide the inferred haplotypes into segments of length 50, and map onto the truth. The switch errors are obtained by counting the changes of a strain segment mapped to reference strains; the genotype errors are the discordance between the strain and the mapped reference segments.

From our assessment of haplotype inference, we conclude:

- The inference of relative proportions do not seem to be affected by the use of a reference panel or its quality (Figure 2).
- The accuracy of the haplotype is dependent on having an appropriate reference panel (Figure 2).
- The strain proportion affects the haplotype inference (see Fig 3). Our method infer strains with proportions over 20% accurately, but struggles with minor strains due to insufficient data, in particularly at sites when the minor strain should be alternative allele, and the dominant strain should be the reference allele (see Figure 3).

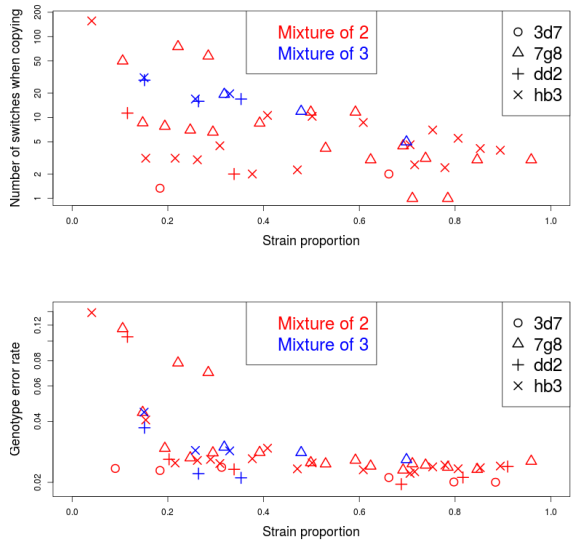

Fig. 3. Error rates vs. strain proportions. We use reference panel V to deconvolute all 27 samples. Each marker represent a deconvoluted haplotype with 18,570 sites.

3.2 Run-time

The complexity of our program is $\mathcal{O}(lm^2)$ (see Fig 4), where l and m are the number of reference strains and sites respectively. In practice, we divide Pf3k samples into several geographical regions and perform deconvolution, with ten most different local clonal strains as reference panel. The run time of deconvoluting a field sample range between 1 and

6 hours, depending on the number variants in a sample: For example, it takes $5\frac{1}{2}$ hours to process sample QG0182-C over 372,884 sites.

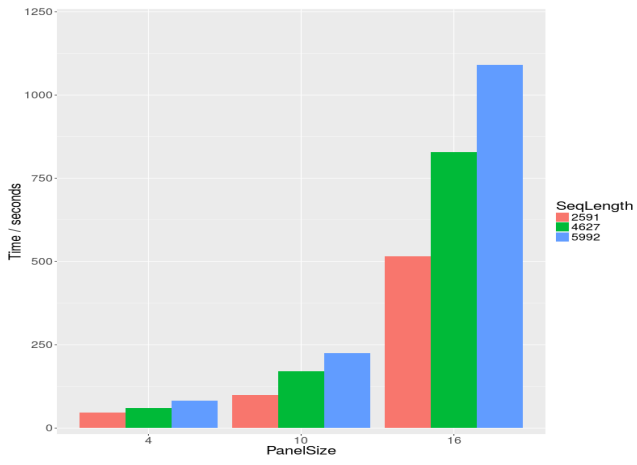

Fig. 4. As a demonstration, we deconvolute chromosome 14, chromosomes 13 and 14, chromosomes 12 to 14 of sample PG0412-C with reference panels I, V and VI. The runtime is almost linear respect to the number of sites; and shows quadratic trend against the number of reference strains.

4 Discussion

The program DEploid and its analysis pipeline has been originally developed for *P. falciparum* studies. Nonetheless, with some minor parameter changes, DEploid can be used for deconvolute *Plasmodium vivax* samples (Pearson et al., 2016). Since the framework is suitable for deconvoluting mixed genomes with unknown proportions. It can thus be extended to a wider range of applications, such as deconvoluting cancer tumour cell genomes or the genomes of other organisms like the Ebola virus.

Acknowledgements

We thank valuable insights and suggestions from Roberto Amato, John O'Brien, Richard Pearson, and Jason Wendler for providing the data of artificial samples. We thank Zam Iqbal for suggesting the name DEploid.

Funding

This project is funded by the Wellcome Trust grant [100956/Z/13/Z].

Conflict of Interest: none declared.

References

Anita, D. (1998). Unstable malaria in Sudan: the influence of the dry season: clone multiplicity of *Plasmodium falciparum* infections in individuals exposed to variable levels of disease transmission. *Transactions of The Royal Society of Tropical Medicine and Hygiene 92*(6), 580–585.

Browning, S. R. and B. L. Browning (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localised haplotype clustering. *The American Journal of Human Genetics 81*(5), 1084–1097.

de Roode, J. *et al*. (2004) Competitive release of drug resistance following drug treatment of mixed *Plasmodium Chabaudi* infections. *Malaria Journal 3*(33), 1–6.

de Roode, J. C. *et al*. (2005) Virulence and competitive ability in genetically diverse malaria infections. *Proceedings of the National Academy of Sciences of the United States of America 102*(21), 7624–7628.

Galinsky, K. *et al*. (2015) COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria Journal 14*(4), 1–9.

Harris K., Nielsen R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genet 9*(6). doi: 10.1371/journal.pgen.1003521

Hastings, I. and U. D'Alessandro (2000). Modelling a predictable disaster: the rise and spread of drug-resistant malaria. *Parasitology Today 16*(8), 340–347.

Howie, B. N. *et al*. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet 5*(6), 1–15.

Lawson D. J. *et al*. (2012) Inference of Population Structure using Dense Haplotype Data. *PLoS Genet 8*(1). doi: 10.1371/journal.pgen.1002453

Li, N. and M. Stephens (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics 165*(4), 2213–2233.

MalariaGEN (2008) A global network for investigating the genomic epidemiology of malaria. *Nature 456*(7223), 732 – 737.

Manske, M. *et al*. (2012) Analysis of plasmodium falciparum diversity in natural infections by deep sequencing. *Nature 487*(7407), 375–379.

Mathieson I. and McVean G. (2014). Demography and the Age of Rare Variants. *PLoS Genet 10*(8), doi: 10.1371/journal.pgen.1004528

Miles, A. *et al*. (2015) Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res 26*, 1288–1299.

Pearson, R. D. *et al*. (2016) Genomic analysis of local variation and recent evolution in Plasmodium vivax. *Nat Genet 48*, 959–964.

The Pf3k Project: pilot data release 5 (2016) www.malariagen.net/data/pf3k-5 [accessed 1 June 2016]

O'Brien D,J. *et al*. (2016) Inferring Strain Mixture within Clinical *Plasmodium falciparum* Isolates from Genomic Sequence Data. *PLoS Comput Biol 12*(6): e1004824. doi: 10.1371/journal.pcbi.1004824

Sabeti1. P. C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature 419*, 832–837. doi:10.1038/nature01140

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

WHO (2016) World Malaria Report 2015. *World Health Organization*.