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

???Genetics and population analysis

???Genome analysis

DEploid: Untangling multiplicity 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, 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: Multiplicity 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 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 in the field.

Results: The software package *DEploid* learns haplotype structure from a reference panel of clonal isolates, and deconvolutes sequences of mixed samples. It reports the number of strains, the mixing proportions and the haplotypes present in an isolate. allowing researchers to study malaria infection history 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.

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 is still 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 increase disease development (de Roode et al., 2005), higher 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 very challenging since genetic differences between the genetic types of this haploid organism will render as heterozygous loci. Mixed calls also confounds methods that exploit haplotype data to detect, among other phenomena, the occurrence of natural selection or recent demographic events. 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 solving the mixed proportion from allele frequencies, yet the haplotypes within a mixed isolate remain unclear.

As part of the Pf3k project (Pf3k, 2016), an effort to map the genetic diversity of *P. falciparum* at global scale, we have developed the 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 and infection history at fine scale.

2 Methods

Overall, we use Markov chain Monte Carlo (MCMC) methods to learn the number of parasite strains and the proportions of allele frequencies, and use sampling method to infer the haplotype of each strain. The goal is firstly construct a high quality reference panel from the clonal samples, and then deconvolute the mixed samples with the reference panel using Li and Stephens (2003)'s hidden Markov model.

2.1 Notations

Let's first introduce some of the notation (see Table 1). Suppose that our data D are the allele counts of sample j at a given site i, denoted as $r_{j,i}$ and $a_{j,i}$ for reference and alternative alleles respectively. The allele frequencies within sample (WSAF) $p_{j,i}$ and at the 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 is subjected to the same sample, we drop the subscript j from now on. Let $\mathbf{w} = [w_1, \ldots, w_k]$ and $\mathbf{h}_i = [h_{1,i}, \ldots, h_{k,i}]$ denote the proportions and haplotypes of k parasite strain at site i. We inherit O'Brien et al. (2015)'s expression for the expected WSAF q_i as:

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

.

i	Marker index
j	Sample index
r	Read count for reference allele
a	Read count for alternative allele
f	Population level allele frequency (PLAF)
k	Number of strains within sample
$\mathbf{w}$	Proportion of strains
$\mathbf{h}_i$	haplotypes of k parasite strain at site i
p	Observed within sample allele frequency (WSAF)
q	Unadjusted expected WSAF
π	Adjusted expected WSAF
Ξ	Reference panel

Table 1. Notation summary

2.1.1 Likelihood of data given the expected WSAF

We adjust the allele frequency q_i by taking into account of read errors 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, we adjust the WSAF take into account of read error as follows:

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

We take into account over-dispersion in read counts, modelling the count distribution as a Beta-binomial distribution. Specifically, the read counts of 'ALT' are identically and independently distributed (i.i.d.) with probability π_i (adjusted), i.e. $a_i \sim Binom(\pi_i, a_i + r_i)$, and $\pi_i \sim Beta(\alpha, \beta)$, where $\pi_i = \alpha/(\alpha + \beta)$. From experience, we set $\alpha = 100 \cdot q_i$ and $\beta = 100 \cdot (1 - q_i)$. Hence we can formalise the likelihood of the data using:

$$L(q_i|D) = P(D|q_i) \propto \frac{\Gamma(a_i + 100 \cdot \pi_i)\Gamma(r_i + 100 \cdot (1 - \pi_i))}{\Gamma(100 \cdot \pi_i)\Gamma(100 \cdot (1 - \pi_i))}, \tag{3}$$

of which expected WSAF q_i is adjusted through Eqn.(2).

2.2 Technical details

Overall, our method generates MCMC samples for the proportions $\mathbf{w}$ and the haplotypes $\mathbf{h}$ for given number of strains. In particular, we assume there are more strains than we actually need, start the MCMC chain with a fixed k. As the values of proportion drops, "zero-out" the "noisy" strains. As for the MCMC updates, we use a Metropolis-Hastings algorithm to sample proportions $\mathbf{w}$ given $\mathbf{h}$ (section 2.2.1); and use Gib sampler to update $\mathbf{h}$ of given $\mathbf{w}$, which are further divided into cases when building the reference panel (section 2.3) and deconvolute mixed samples (section 2.4). At last, we take the best fit (section2.5) of the MCMC sample as a point estimate to infer the haplotypes and proportion.

2.2.1 MCMC update for proportions

We use a sparse update on $\mathbf{w}$. We introduce a multivariate normal variable titre $\mathbf{x} = [x_1, \ldots, x_k]$, where each x is i.i.d. normally distributed from $N(0, 3)$, with the density function $d(x)$. We sample xs, then transform to $w = e^x$. We then normalise vector $\mathbf{w}$ by the sum, to obtain a new sample $\mathbf{w}$.

2.3 Infer the reference strains

In practice, we assume clonal sample haplotypes capture the diversities of haplotype structures given all samples. We use them as the reference strains for start, and deconvolute the rest mixed samples from them. We start with a set of clonal sample candidate, and run the algorithm to confirm they are in fact clonal. We use Gib sampler to update $\mathbf{h}$ of given $\mathbf{w}$, randomly select one strain at the time, or a pair of strains to update in order to improve the mixing of the MCMC process.

2.3.1 Update a single strain at one time

Choose haplotype strain s uniformly at random from these K strains, consider both cases of 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 firstly remove it from the current WSAF, i.e. subtract $w_s \cdot h_s$ from Eqn. (1), which gives

$$q_{i,-s} = \sum_{k \neq s} w_k \cdot h_k = \text{Eqn. (1)} - w_s \cdot h_s \tag{4}$$

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

$$q_{i,g_s=0} = \text{Eqn. (4)} \tag{5}$$

$$q_{i,g_s=1} = \text{Eqn. (4)} + w_s \times 1 \tag{6}$$

Substitute equations (5) and (6) into Eqn. (3) to compute associated likelihood $L(q_{i,g_s}|D)$, which is expressed as $L(g_s|D)$ in short, for the rest of the paper.

Assume Independence between site i and $i+1$, we use PLAF at position i as prior, we then can compute the posterior probability as follows:

$$P(g_s|D) \propto L(g_s|D) \times P(g_s). \tag{7}$$

2.3.2 Update two haplotypes at one time

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. Similar to Eqn. (4), we first remove their states from the WSAF:

$$\begin{aligned} q_{i,-s_1,-s_2} &= \sum_{k \neq s_1,s_2} w_k \cdot h_k \\ &= \text{Eqn. (1)} - w_{s_1} \cdot h_{s_1} - w_{s_2} \cdot h_{s_2} \end{aligned} \tag{8}$$

Consider all four possible combination of genotypes, we have

$$q_{i,g_{s_1}=0,g_{s_2}=0} = \text{Eqn. (8)} \tag{9}$$

$$q_{i,g_{s_1}=0,g_{s_2}=0} = \text{Eqn. (8)} + \cdot w_{s_1} \times 1 \tag{10}$$

$$q_{i,g_{s_1}=0,g_{s_2}=1} = \text{Eqn. (8)} + \cdot w_{s_2} \times 1 \tag{11}$$

$$q_{i,g_{s_1}=0,g_{s_2}=1} = \text{Eqn. (8)} + \cdot w_{s_1} \times 1 + w_{s_2} \times 1 \tag{12}$$

Substitute expressions. (9) to (12), into Eqn. (3), we then obtain their associated likelihood $L(q_{i,g_{s_1},g_{s_2}}|D)$, which is denoted as $L(g_{s_1}, g_{s_2}|D)$ in the rest of the paper.

Thus, similar to Eqn. (7), we obtain the following posterior probability, and sample the state (genotype) of strains s_1 and s_2 simultaneously at site i:

$$P(g_{s_1}, g_{s_2}|D) \propto L(g_{s_1}, g_{s_2}|D) \times P(g_{s_1}, g_{s_2}), \tag{13}$$

where $P(g_{s_1}, g_{s_2}) = P(g_{s_1}) \cdot P(g_{s_2})$, assume independence between s_1 and s_2.

2.4 Deconvolute the mixed isolates

We use Li and Stephens (2003)'s hidden Markov model frame work as a starting point. The following modifications are 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 equal probability.
- simplifying the mutation model with a fixed miss copying operation.

2.4.1 Update single haplotype with LS

Recombination map model We assume a uniform recombination map, genetic distances between loci i and $i+1$ are computed by $G_i = D_i/morgan$ where D_i denotes the physical distance between loci i and $i+1$ in nucleotide, $morgan$ is the average morgan distance, which we use 1500000. Suppose the recombination rate ψ_i is given by $\psi_i = N_e G_i$, with $N_e = 10$ being the effective population size. Note that **we scale the probabilities with the number of haplotypes in the reference panel**. Let Ξ denote the set of the strains in the reference panel. For position $i > 1$, let ρ'_i denote the probability of **no** recombination from site $i-1$ to i, we have $\rho'_i = \exp(-\psi_i)$. Thus, the probability of recombining from any strain in the panel is $\dfrac{1-\rho'_i}{|\Xi|}$, where $|\Xi|$ is the size of the panel.

A crucial difference between our method and Li and Stephens (2003)'s model is that mixed samples can have more than two strains, with unknown proportions. We randomly choose the strains to update, then apply LS's algorithm to sample the path using Gibbs sampler given the proportion $\mathbf{p}$ (see example in Fig. 1), rather than 50/50 in the cases of diploid samples.

In addition to updating the haplotypes from the panel, we take into account of miss copying (see example shown in Fig. 1), which allow the actual genotype differ from the path. Therefore, this method favours mutation rather than recombination for one-off events. 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 from data.

Specific steps are as following:

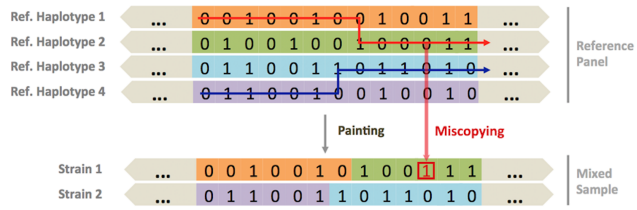

Fig. 1. Illustration of Li and Stephens (2003)'s algorithm. Strain 1 haplotype is made up from reference haplotype segments of 1 and 2; and strain 2 haplotype is made up from reference haplotype segments of 3 and 4. With miss copying, we allow strain states differ from the path: At the third last position 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 g_s to denote the genotype of the copied path and the updated strain respectively. We have:

$$\begin{aligned} L(g_p = *|D) &= L(g_s = *|D) \times P(g_p = g_s) + \\ &\quad L(g_s = 1 - *|D) \times P(g_p \neq g_s) \end{aligned}$$

where $* \in \{0, 1\}$, and $1 - *$ indicates the event that g_s takes value that differs from g_p. Let μ denote the probability of miss copying, we have

$$\begin{cases} P(g_p = g_s) &= 1 - \mu, \\ P(g_p \neq g_s) &= \mu. \end{cases}$$

2. Compute the probability of path at each position using forward algorithm. Therefore, we have the posterior probability of path (reference strain) p at position i as:

$$\begin{aligned} P_i(g_p|D) \;\propto\; & L(g_p|D) \times \\ & \left(\rho'_i \cdot P_{i-1}(g_p|D) + \tfrac{1-\rho'_i}{|\Xi|} \cdot \sum_{x \in \Xi} P_{i-1}(g_x|D) \right). \end{aligned} \tag{14}$$

In the HMM frame work, $L(g_p|D)$ is the emission probability of observing data D given the hidden state of the path, ρ'_i and $\frac{1-\rho'_i}{|\Xi|}$ are the transition probabilities from position $i-1$ to i, of which reflect the recombination event in our context.

3. Sample the path backwards once the forward posterior probabilities are calculated. We start at the end position, first propose if a recombination events have had happened with the probabilities proportional to

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

if it recombines, sample path backwards according $P_i(g_p|D)$.

4. Ultimately, given the state of the path at each site, we now want to sample the genotype according to the posterior probabilities:

$$P(g_s = *|D) = \begin{cases} P(g_p = *|D) \cdot (1 - \mu), & g_s = g_p; \\ P(g_p = 1 - *|D) \cdot \mu, & g_s \neq g_p. \end{cases} \tag{15}$$

2.4.2 Update pair of haplotypes with LD

Similarly to the previous section, we need to

1. Compute the emission probabilities

$$L\left(\begin{smallmatrix} g_{p_1}=*, \\ g_{p_2}=\# \end{smallmatrix}\middle|D\right) = L\left(\begin{smallmatrix} g_{s_1}=*, \\ g_{s_2}=\# \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}=g_{s_1}, \\ g_{p_2}=g_{s_2} \end{smallmatrix}\right) + L\left(\begin{smallmatrix} g_{s_1}=*, \\ g_{s_2}=1-\# \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}=g_{s_1}, \\ g_{p_2}\neq g_{s_2} \end{smallmatrix}\right) + L\left(\begin{smallmatrix} g_{s_1}=1-*, \\ g_{s_2}=\# \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}\neq g_{s_1}, \\ g_{p_2}=g_{s_2} \end{smallmatrix}\right) + L\left(\begin{smallmatrix} g_{s_1}=1-*, \\ g_{s_2}=1-\# \end{smallmatrix}\middle|D\right) \times P\left(\begin{smallmatrix} g_{p_1}\neq g_{s_1}, \\ g_{p_2}\neq g_{s_2} \end{smallmatrix}\right)$$

where

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

2. Compute the probability of path at each position using forward algorithm. Similar to Equation (14), for 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.

$$\begin{aligned} P_i(g_{p_1}, g_{p_2}|D) \propto & L(g_{p_1}, g_{p_2}|D)\times \\ \Bigg[& \sum_{x\in\Xi} P_{i-1}(g_{p_1}, g_x|D)\cdot\rho'_i\cdot\frac{1-\rho'_i}{|\Xi|} + \\ & P_{i-1}(g_{p_1}, g_{p_2}|D)\cdot\rho'_i\cdot\rho'_i + \\ & \sum_{y\in\Xi} P_{i-1}(g_y, g_{p_2}|D)\cdot\rho'_i\cdot\frac{1-\rho'_i}{|\Xi|} + \\ & \sum_{x,y\in\Xi\cdot\Xi} P_{i-1}(g_x, g_y|D)\cdot\frac{1-\rho'_i}{|\Xi|}\cdot\frac{1-\rho'_i}{|\Xi|} \Bigg] \end{aligned} \tag{16}$$

3. Sample the path up to position i, i.e. backwards, start from the end of the panel, and sampling backwards to the $i-1$ position, first sample if a recombination events had happened given the probabilities of

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

If both strains recombine, sample the path, according to $P_i(g_{p_1}, g_{p_2}|D)$. If one of the strains recombine, sample the path according to the marginal probability of $P_i(g_p|D)$.

4. Ultimately, we consider add miss copies similar to the previous section, and sample the strain state given the path state with probabilities:

$$P\left(\begin{smallmatrix} g_{s_1}=*, \\ g_{s_2}=\# \end{smallmatrix}\middle|D\right) = \begin{cases} P\left(\begin{smallmatrix} g_{p_1}=*, \\ g_{p_2}=\# \end{smallmatrix}\middle|D\right)\cdot(1-\mu)\cdot(1-\mu), \\ P\left(\begin{smallmatrix} (g_{p_1}=*, \\ g_{p_2}=1-\# \end{smallmatrix}\middle|D\right)\cdot(1-\mu)\cdot\mu, \\ P\left(\begin{smallmatrix} (g_{p_1}=1-*, \\ g_{p_2}=\# \end{smallmatrix}\middle|D\right)\cdot\mu\cdot(1-\mu), \\ P\left(\begin{smallmatrix} (g_{p_1}=1-*, \\ g_{p_2}=1-\# \end{smallmatrix}\middle|D\right)\cdot\mu\cdot\mu, \end{cases}$$

consider all cases of if the path the same as the strain.

2.5 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 then use the lowest deviance information criterion (DIC) to select the best fit model. 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.

3 Validation and Performance

A set of *in vitro* mixtures of parasites were created by Wendler (2015) to simulate mixed infection, which is an ideal validation data set in our use. In this data set, DNA was extracted from four laboratory parasite lines: 3D7, Dd2, HB3 and 7G8, and mixed with different ratios of mixed infection (see Table 2 in brackets), and submitted to the MalariaGEN pipeline (MalariaGEN, 2008) for Illumina sequencing.

Note that this data set only contains two clonal-ish samples. Due to the limited size, they are not ideal for constructing a reference panel. Moreover, the *P. falciparum* genetic crosses project (Miles et al., 2015) finds that due to sequencing error or applying different variant calling methods, genotype calls vary at the same position given the same strain of *P. falciparum*. Thus we apply inference methods to mutiple samples that contains the same parasite strains, and infer the genotypes of a reference strain.

Infer haplotypes for Dd2 strain. Since 3D7 is the reference strain, we can assume that strain Dd2 is the only source of 'ALT' reads in samples from PG0389-C to PG0394-C. Assume 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 regression model $y = \beta_0 + \beta_1 x$ to infer the Dd2 genotype: 1 if β_1 is significant; 0 otherwise.

Infer haplotypes for HB3 and 7G8. Similarly, for sample from 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; use 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; 0 otherwise; similarly for 7G8.

3.1 Accuracy

3.1.1 Proportions and number of strains

We apply our program to 27 lab-mixed *in vitro* samples to validate our methods and program. As described in section 2.2.1, we start our method with the assumption of at most three strains present in the mixtures; and discard the strains less than 1%. 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%.

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 over-fits 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 have subtle difference with the same parasite line, but overall vastly different from the last strain. The subtle variation is caused by few heterozygous sites with high coverage resulting high leverage in our model (supplemental material Figure S2.3(a)). The source of the noisy markers are possibly from sequencing or variant calling process, which are not recalabriated by our program.

We experimented deconvoluting the 27 lab-mixed samples with the following different 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.
- 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 take into account of both switch errors and genotype discordance, which reflect to the recombination and miss copying events in the method section. Intuitively, one may suggest

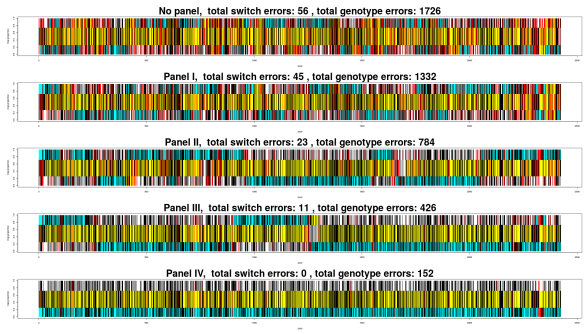

Fig. 2. Haplotypes comparison of sample PG0396-C chromosome 14 deconvolution without any reference strain (top) versus with using reference panels I to IV (from the second to the bottom). Black bars indicate alternative alleles; red bars mark wrongly inferred positions. The yellow, cyan and white background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

to use the Li and Stephens model to compute the vertabi path or posterior probabilities to assess the how different our inferred haplotypes differ from the truth. However, we find that such methods overestimate switch errors at short segments of sequence due to variable reference panel quality.

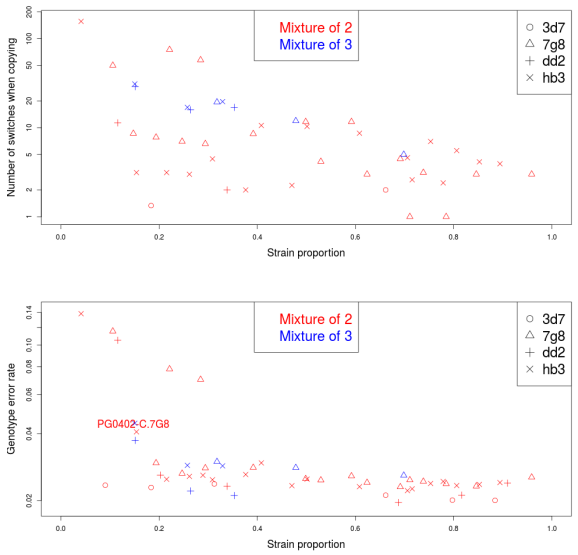

Fig. 3. Error rates vs. strain proportions. We use reference panel V to deconvolute all 27 samples. Each marker represent a deconvoluted haplotype with 18570 sites.

Moreover, we want to assess the switch errors by taking into account of switches from the other haplotypes. 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 vertabi path or posterior probabilities fail to align the switches of all strains. Therefore, we have taken a heuristic approach by dividing the inferred haplotypes into segments with length of 50, then mapped onto reference strains of panel V according to the indices. The switch errors are obtained by counting the changes of a strain mapped to reference strains; the genotype errors are the discordance between the strain and the mapped reference segments.

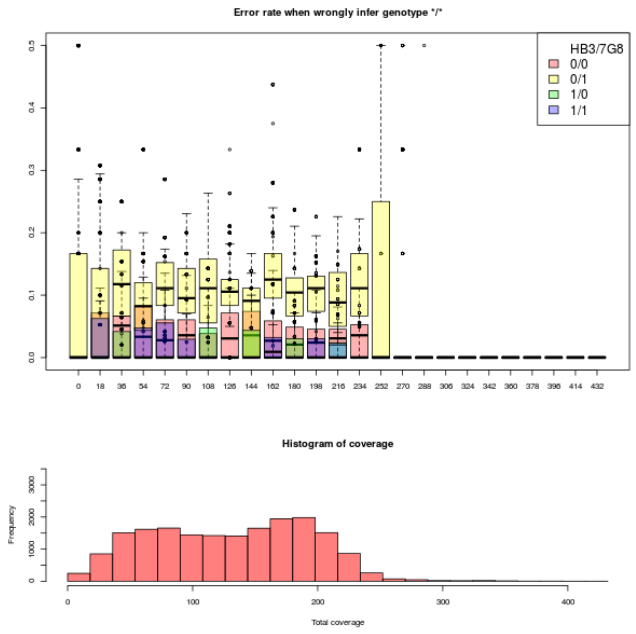

Fig. 4. Error rate of a particular genotype inference at different converages. Sample PG0402-C deconvolution with panel V. Compare with the error rate of strain PG0402-C.7G8 in Figure 3, we find most of errors in infering heterozygous sites, where alternative allele is coming from the minor strain.

Our assessment on haplotype inference can be concluded as the following:

- The proportion inference do not seem to be affected by the presence of reference panel nor the quality of a panel (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 of with proportions over 20% accurately, but struggle with minor strains due to insufficient data, in particularly at sites when the minor strain should be alternative allele, and the dominate strain should be reference allele (see Figure 3).

We extended our experiments further to test how sensitive the inference result is to the sequence coverage. Data was simulated by sampling read counts according to $Binom(n, p)$ models, where n was the number alternative and reference alleles at each site. Three different probabilities p: 0.2, 0.5, 0.8 were used for creating the scenarios of lower, median and high coverage data. In general, we observed that higher coverage data was more informative about the genotype, which accompanied with reduced error rates, with exceptions when using a perfect reference panel: The haplotype inference heavily relied on the reference panel for low coverage data, which again addresses the importance of using appropriate reference panels (see supplement materials).

3.2 Run-time

The complexity of our program is $\mathcal{O}(n^2 m)$ (see Fig 5), where n and m are the number of reference strains and sites respectively. In practice, we divide Pf3k samples into several geographical region and perform deconvolution, with ten vastly diversed 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.

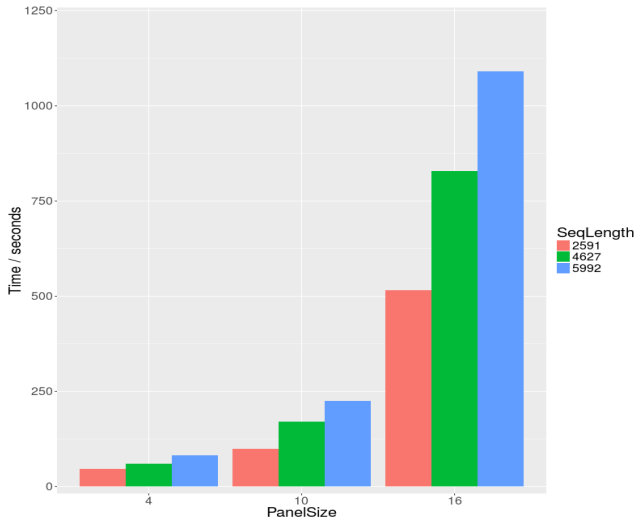

Fig. 5. As a demonstration, we deconvolute chromosome 14, chromosomes 13 and 14, chromosomes 12 to 14 of sample PG0412-C with reference panels I, V and VI. The runtime is almost linear respect to the number of sites; and shows quadratic trend against the number of reference strains.

4 Discussion

The program DEploid and its analysis pipeline is originally developed for *P. falciparum* studies. With some specific minor parameter changes, DEploid can be used for deconvolute *P. vivax* sequence data (Pearson et al., 2016). 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 Ebola virus genomes.

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 naming the program 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., R. Culleton, A. Bell, and A. Read (2004). Competitive release of drug resistance following drug treatment of mixed *Plasmodium Chabaudi* infections. *Malaria Journal 3*(33), 1–6.

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 of the United States of America 102*(21), 7624–7628.

Galinsky, K., Valim, C., Salmier, A., de Thoisy, B., Legrand, E., Faust, A., Baniecki, M. L., Ndiaye, D., Daniels, R. F., Hartl, D. L., Sabeti, P. C., Wirth, D. F., Volkman, S. K., Neafsey, Daniel E.(2015). COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. *Malaria Journal 14*(4), 1–9.

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., P. Donnelly, and J. Marchini (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet 5*(6), 1–15.

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.

Miles, A., Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O'Brien, V. Ruano Rubio, B. MacInnis, J. Mwangi, U. Samarakoon, L. Ranford-Cartwright, M. Ferdig, K. Hayton, X. Su, T. Wellems, J. Rayner, G. McVean, and D. Kwiatkowski (2015). Genome variation and meiotic recombination in plasmodium falciparum: insights from deep sequencing of genetic crosses. *bioRxiv*.

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, June). 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., Iqbal Z, Wendler J, Amenga-Etego L (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

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*.