Supplemental Materials of some Implementation details

S1 Implementation details

This section includes implementation details with handling arithmetic problems.

1. Apply expression $\Gamma(x)\Gamma(y) = \Gamma(x+y)B(x,y)$ (Wikipedia, 2003) to Eqn. (**??**), we have the following:

$$L(q_i|D) \propto \frac{B(a + 100 \times \pi_i, r + 100 \times (1-\pi)}{B(100 \times \pi_i, 100 \times (1-\pi_i))}.$$

Take the log likelihood expression is obtained:

$$l(q_i\big|D) = \log(B(a + 100 \times \pi_i, r + 100 \times (1-\pi)) - \log(B(100 \times \pi_i, 100 \times (1-\pi_i)))).$$

2. During reference panel building stage, we use the PLAF as the prior probability in Eqn. (**??**), and use P_0 and P_1 to denote $P(g_s = 0)$ and $P(g_s = 1)$ respectively. Let l_0 and l_1 denote the log likelihood of $g_s = 0$, $g_s = 1$ given data. Let $L = \max(L_0, L_1)$ and $l = \max(l_0, l_1)$

We normalize the posterior probabilities as:

$$\begin{cases} P(g_s = 0|D) & = \frac{P(g_s=0|D)}{P(g_s=0|D)+P(g_s=1|D)} \\ P(g_s = 1|D) & = \frac{P(g_s=0|D)}{P(g_s=0|D)+P(g_s=1|D)} \end{cases}$$

where

$$\begin{aligned} P(g_s = 0|D) &= \frac{P(g_s = 0|D)}{P(g_s = 0|D) + P(g_s = 1|D)} \\ &= \frac{P(g_s = 0) \cdot L_0}{P(g_s = 0) \cdot L_0 + P(g_s = 1) \cdot L_1} = \frac{(P(g_s = 0) \cdot L_0)/L}{(P(g_s = 0) \cdot L_0 + P(g_s = 1) \cdot L_1)/L} \\ &= \frac{P(g_s = 0) \cdot L_0/L}{P(g_s = 0) \cdot L_0/L + P(g_s = 1) \cdot L_1/L} \end{aligned}$$

Similarly, we have

$$P(g_s = 1|D) = \frac{P(g_s = 1) \cdot L_1/L}{P(g_s = 0) \cdot L_0/L + P(g_s = 1) \cdot L_1/L},$$

where we substitue L_0/L and L_1/L as $\exp(l_0 - l)$ and $\exp(l_1 - l)$ respectively.

We normalize the log likelihood with its maximum at every site, in order to avoid truncation errors occured during probability summations. This approach is also applied to equations (**??**), (**??**) and

(??).

References

Wikipedia (2003). Relationship between gamma function and beta function. [Online; accessed 2016-02-01].

S2 Method validation on lab controlled strains

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, such as 3D7, Dd2, HB3 and 7G8, and mixed with different ratios of mixed infection (Table S2.2), and submitted to the MalariaGEN (MalariaGEN, 2008) pipeline for Illumina sequencing.

sample	3D7 (F)	Dd2 (C)	HB3 (C)	7G8 (C)
PG0389C	90	10	0	0
PG0390C	80	20	0	0
PG0391C	67	33	0	0
PG0392C	33	67	0	0
PG0393C	20	80	0	0
PG0394C	10	90	0	0
PG0395C	0	33.3	33.3	33.3
PG0396C	0	25	25	50
PG0397C	0	14.3	14.3	71.4
PG0398C	0	0	100	0
PG0399C	0	0	99	1
PG0400C	0	0	95	5
PG0401C	0	0	90	10
PG0402C	0	0	85	15
PG0403C	0	0	80	20
PG0404C	0	0	75	25
PG0405C	0	0	70	30
PG0406C	0	0	60	40
PG0407C	0	0	50	50
PG0408C	0	0	40	60
PG0409C	0	0	30	70
PG0410C	0	0	25	75
PG0411C	0	0	20	80
PG0412C	0	0	15	85
PG0413C	0	0	5	95
PG0414C	0	0	1	99
PG0415C	0	0	0	100

Table S2.1: caption

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.

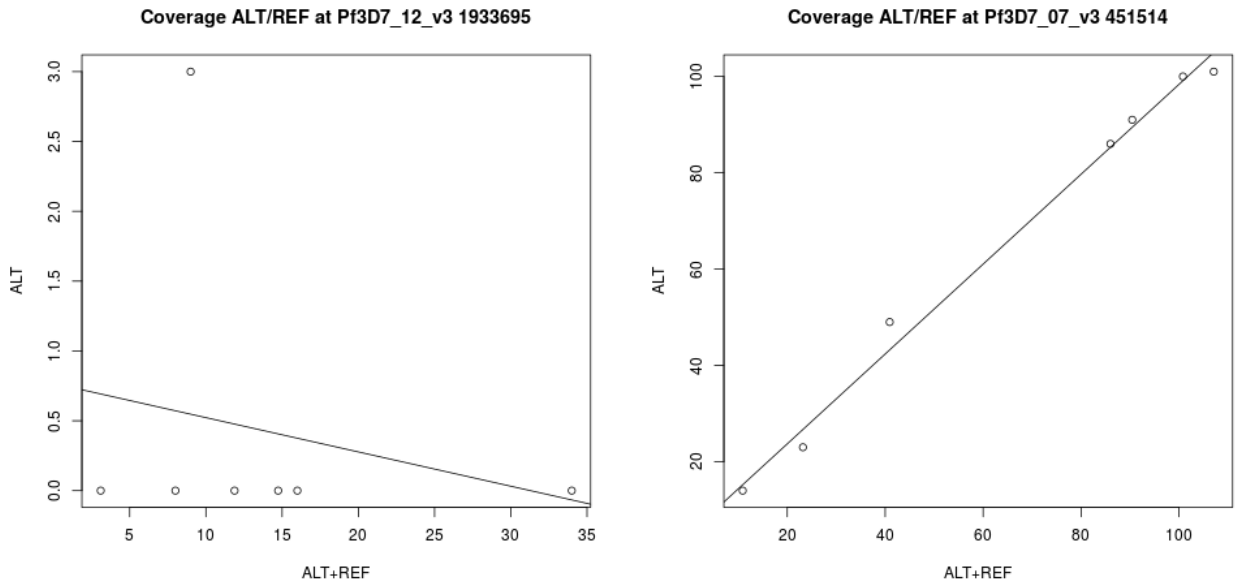
S2.1 Use inference method to reconstruct the reference strains

1. Mixtures of strains 3D7 and Dd2 Since 3D7 is reference strain, we can assume that strain Dd2 is the only source of 'ALT' reads in samples `PG0389-C`, `PG0390-C`, `PG0391-C`, `PG0392-C`, `PG0393-C` and `PG0394-C`. Assume markers are independent from each other, let y be the read count for 'ALT' allele and x be the weighted coverage, of which the weight are the proportions that are used during the mixing (see Table S2.2), we use the following regression model to infer the Dd2 variant calling,

$$y = \beta_0 + \beta_{Dd2} x,$$

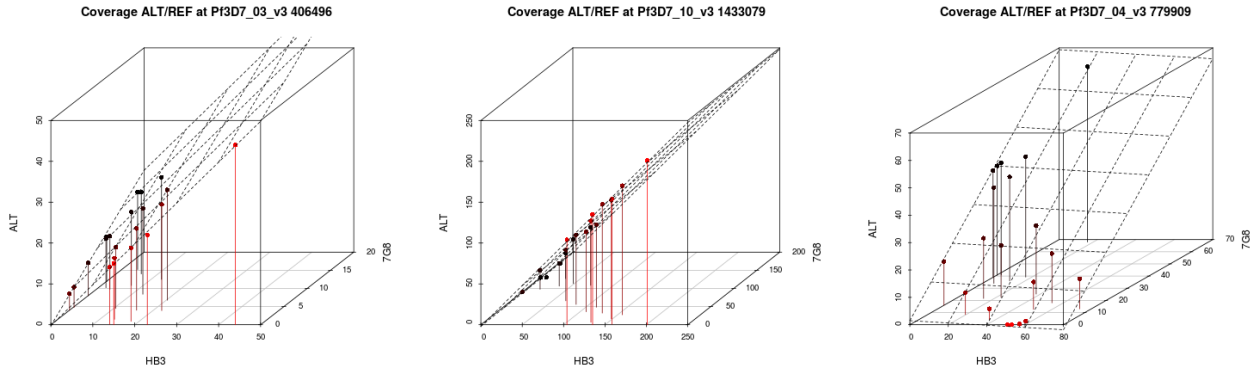
from which significant coefficent β_{Dd2} implies a Dd2 variant (Fig. S2.1b).

2. Mixtures of strains HB3 and 7G8. Similarly, for sample from `PG0398-C` to `PG0415-C`, we let variables x_1, x_2 be the weighted coverages, of which the weights are the mixing proportions for strains HB3 and 7G8 respectively. We use regression model $y = \beta_0 + \beta_{Hb3} x_1 + \beta_{7G8} x_2$ to investigate the relationships between the total allele count and weighted coverage of HB3 and 7G8. Hb3 variant is inferred as coefficients β_{Hb3} is significant (Fig. S2.2a and S2.2b), so is 7G8 (Fig. S2.2a and S2.2c).



(a) (b)

Figure S2.1: XXXXXXXXXXXXXX



(a) (b) (c)

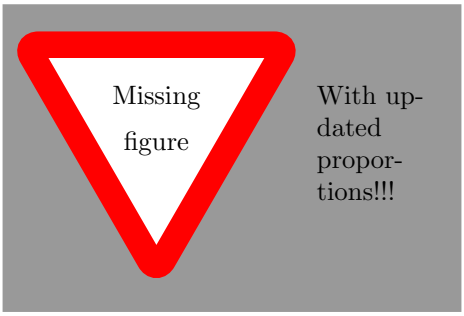
Figure S2.2: XXXXXXXXXXXXXX

S2.2 Validation performance

S2.2.1 Assessing quality of the proportion inference

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Table S2.2: Inferred proportions from Jason's samples

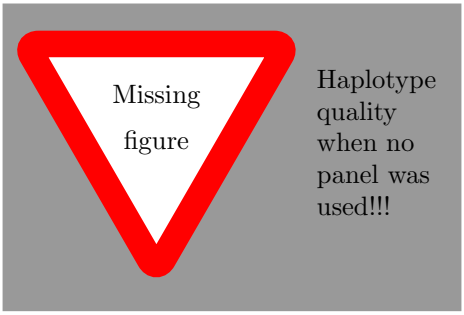


S2.2.2 Assesing haplotype qualities when given different panels

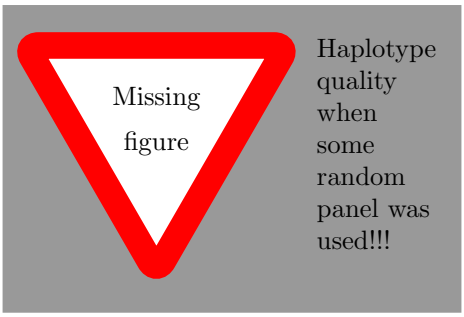
References

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

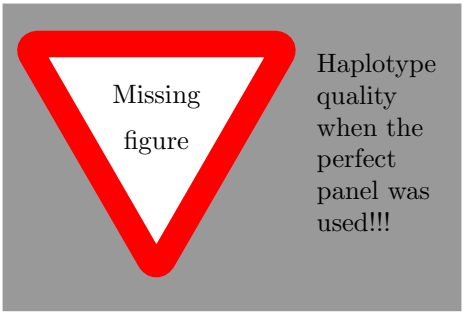
Miles, A., Z. Iqbal, P. Vauterin, R. Pearson, S. Campino, M. Theron, K. Gould, D. Mead, E. Drury, J. O'Brien, V. Ruano Rubio, B. MacInnis, J. Mwangi, U. Samarakoon, L. Ranford-Cartwright, M. Ferdig,



Haplotype quality when no panel was used!!!



Haplotype quality when some random panel was used!!!



Haplotype quality when the perfect panel was used!!!

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*.

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