Supplemental Materials of the Deconvolution Method

S1 Methods

S1.1 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 strais with variable proportion rather than two sequences with equal probability.
- simplifying the mutation model with a fixed miss copying operation.

S1.1.1 Update single haplotype with LD

Recombination map model The first case refers to staying on the same path and the second to a recombination event (i.e switch). Let ψ_i is given by $\psi_i = N_e G_i$, with N_e being the effective population size and G_i the genetic distance between loci i and $i+1$. We assume a uniform recombination map, genetic distances 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, $N_e = 10$.

Whereas recombination probabilities for a segment are computed by the following function. Note that **we scale the probabilities with the number of haplotypes in the reference panel**. Let RP denote the set of the strains in the reference panel. For position $i > 1$, let ρ_i' denote the probability of **no** recombinations 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 $\dfrac{1-\rho_i'}{|RP|}$, where $|RP|$ 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 **p** (see example in Fig. S1.1.

In addiontion to updating the haplotypes from the panel, we take into account of miss copying (see example shown in Fig. S1.1), which allow the actual genotype differ from the path, in order to improve the likelihood of data.

1. Consider the likelihood as the emission probabilities at site i. Let's use g_p and g_s to denote the genotype

Reference haplotype 1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1
Reference haplotype 2	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1
Reference haplotype 3	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1
Reference haplotype 4	0	1	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0
Reference haplotype 5	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	1

Strain haplotype 1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	1
Strain haplotype 2	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1
Strain haplotype 3	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1
Strain haplotype 4	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1

Figure S1.1: Illustration of Li and Stephens (2003)'s algorithm. Strain haplotype 1 is made up from reference haplotype segments of 1, 2, and 4; and strain haplotype 3 is made up from reference haplotype segments of 1, 2, 3 and 4. With miss copying, we allow strain states differ from the path: At the end position of strain 3, the path is copied from reference haplotype 4, with the state of "0".

of the copied path and the updated strain respectively. We have:

$$L(g_p = *|D) = P(g_p = g_s) \cdot L(g_s = *\big|D) + P(g_p \neq g_s) \cdot L(g_s = 1 - *\big|D) \tag{S1.1}$$

where $g_p = * \in \{0, 1\}$, 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 probabilility of path (reference strain) p at position i as:

$$P_i(g_p\big|D) \propto \left(\rho_i' \cdot P_{i-1}(g_p\big|D) + \frac{1 - \rho_i'}{|RP|} \cdot \sum_{x \in R} P_{i-1}(g_x\big|D)\right) \cdot L(g_p\big|D). \tag{S1.2}$$

In the HMM frame work, $L(g_p\big|D)$ is the emmission probability of oberving data D given the hidden state of the path, ρ_i' and $\frac{1-\rho_i'}{|RP|}$ are the transition probabilities from position $i-1$ to i, of which reflect the recombination event in our context.

3. Sample the path up to position i, i.e. backwards, start from the end of the sequence. At the end position, sample path according $f_{u,end}$. for the $i-1$ position, first sample if a recombination events

had happened with the probabilities proportional to

$$\begin{cases} \rho'_i \cdot f_{u,i-1} & \text{no recombined,} \\ (1-\rho'_i) \cdot \sum_{x \in R} f_{x,i-1} & \text{recombined.} \end{cases}$$

If it was recombined, sample the path u, according to $f_{u,i-1}$.

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

$$P(g_s = *\big|D) = \begin{cases} P(g_p = *\big|D) \cdot (1-\mu), & g_s = g_p; \\ P(g_p = 1-*\big|D) \cdot \mu, & g_s \neq g_p. \end{cases} \tag{S1.3}$$

S1.1.2 Update pair of haplotypes with LD

Similarly to the previous section, we need to

1. Compute the emission probabilities

$$\begin{aligned} L(g_{p_1} = *, g_{p_2} = \#\big|D) = & P(g_{p_1} = g_{s_1}, g_{p_2} = g_{s_2}) \cdot L(g_{s_1} = *, g_{s_2} = \#\big|D) + \\ & P(g_{p_1} = g_{s_1}, g_{p_2} \neq g_{s_2}) \cdot L(g_{s_1} = *, g_{s_2} = 1-\#\big|D) + \\ & P(g_{p_1} \neq g_{s_1}, g_{p_2} = g_{s_2}) \cdot L(g_{s_1} = 1-*, g_{s_2} = \#\big|D) + \\ & P(g_{p_1} \neq g_{s_1}, g_{p_2} \neq g_{s_2}) \cdot L(g_{s_1} = 1-*, g_{s_2} = 1-\#\big|D) \end{aligned} \tag{S1.4}$$

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 (S1.2), for all possible pair of the copying strain, we take into account of the possiblility of one strain recombines and the other does not with the probability of $\rho'_i \cdot \frac{1-\rho'_i}{|RP|}$; both recombines, with the probability of $\rho'_i \cdot \rho'_i$; neither recombines, with the probability of $\frac{1-\rho'_i}{|RP|} \cdot \frac{1-\rho'_i}{|RP|}$,

assuming that recombination events of two copying strains are independent from each other.

$$
\begin{aligned}
P_i(g_{p_1}, g_{p_2}|D) \propto [&P_{i-1}(g_{p_1}, g_{p_2}|D) \cdot \rho'_i \cdot \rho'_i + \\
&\sum_{x \in R} P_{i-1}(g_{p_1}, g_x|D) \cdot \rho'_i \cdot \frac{1-\rho'_i}{|RP|} + \\
&\sum_{y \in R} P_{i-1}(g_y, g_{p_2}|D) \cdot \rho'_i \cdot \frac{1-\rho'_i}{|RP|} + \\
&\sum_{x,y \in R \cdot R} P_{i-1}(g_x, g_y|D) \cdot \frac{1-\rho'_i}{|RP|} \cdot \frac{1-\rho'_i}{|RP|}] \cdot L(g_{p_1}, g_{p_2}|D)
\end{aligned} \tag{S1.5}
$$

3. Sample the path up to position i, i.e. backwards, start from the end of the panel. At the end position, sample path according $P(p_1 = u, p_2 = v) = f_{u,v,end}$. for the $i-1$ position, first sample if a recombination events had happened given the probabilities of

$$
\begin{cases}
f_{u,v,i-1} \cdot \rho'_i \cdot \rho'_i, & \text{no recombined,} \quad \text{(S1.6)} \\
\sum_{* \in RP} f_{u,*,i-1} \cdot \frac{1-\rho'_i}{|RP|} \cdot \rho'_i, & \text{u recombined ,} \quad \text{(S1.7)} \\
\sum_{* \in RP} f_{*,v,i-1} \cdot \frac{1-\rho'_i}{|RP|} \cdot \rho'_i, & \text{v recombined ,} \quad \text{(S1.8)} \\
\sum_{*,* \in RP \cdot RP} f_{*,*,i-1} \cdot \frac{1-\rho'_i}{|RP|} \cdot \frac{1-\rho'_i}{|RP|}), & \text{both recombined.} \quad \text{(S1.9)}
\end{cases}
$$

If it both recombined, sample the path, according $P(p_1 = u, p_2 = v) = f(u, v, i-1)$. If one of them recombined, sample the path according to the marginal probability of $P(p_1 = u) = f(u, i-1)$.

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

$$
P(g_{s_1} = *, g_{s_2} = \#|D) = \begin{cases}
P(g_{p_1} = *, g_{p_2} = \#|D) \cdot (1-\mu) \cdot (1-\mu), & g_{s_1} = g_{p_1} \text{ and } g_{s_2} = g_{p_2}; \\
P(g_{p_1} = *, g_{p_2} = 1-\#|D) \cdot (1-\mu) \cdot \mu, & g_{s_1} = g_{p_1} \text{ and } g_{s_2} \neq g_{p_2}; \\
P(g_{p_1} = 1-*, g_{p_2} = \#|D) \cdot \mu \cdot (1-\mu), & g_{s_1} \neq g_{p_1} \text{ and } g_{s_2} = g_{p_2}; \\
P(g_{p_1} = 1-*, g_{p_2} = 1-\#|D) \cdot \mu \cdot \mu, & g_{s_1} \neq g_{p_1} \text{ and } g_{s_2} \neq g_{p_2}.
\end{cases} \tag{S1.10}
$$

References

OBrien, 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 Comput Biol 12*(6), 1–20.

Li, N. and M. Stephens (2003, December). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics 165*(4), 2213–2233.

Supplemental Materials of some Implementation details

S2 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 (S1.1), (**??**) and

(S1.4).

References

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

Supplemental Materials of the Deconvolution Method Validation

S3 Method validation on lab controlled strains

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.

S3.1 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 **??**), 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. S3.1b).

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. S3.2a and S3.2b), so is 7G8 (Fig. S3.2a and S3.2c).

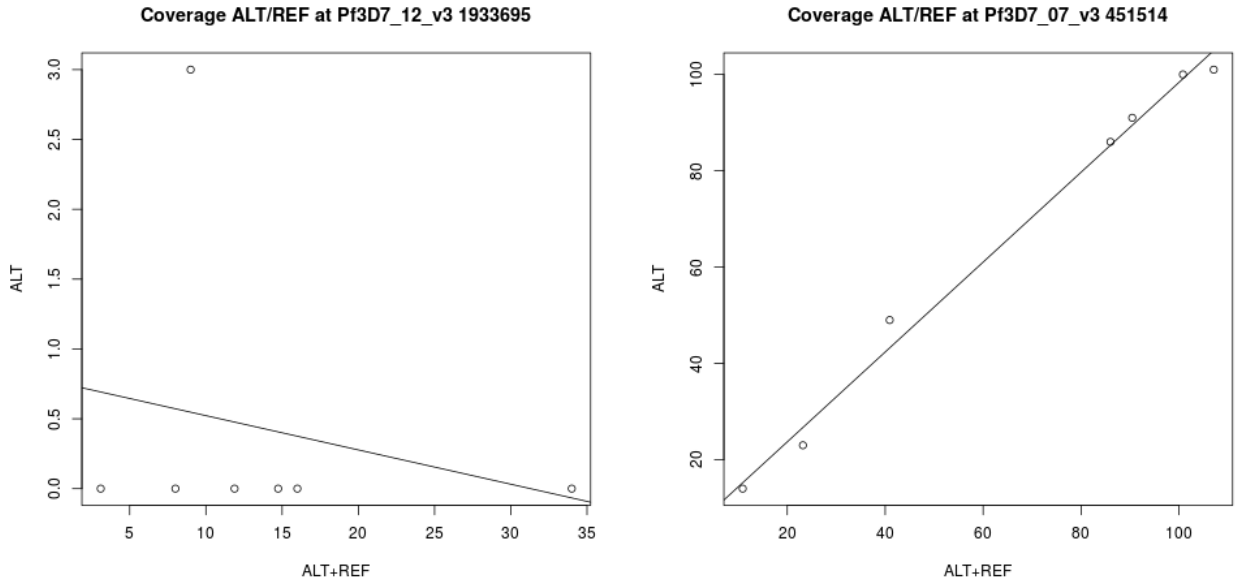
S3.2 Validation performance

S3.2.1 Assessing quality of the proportion inference

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

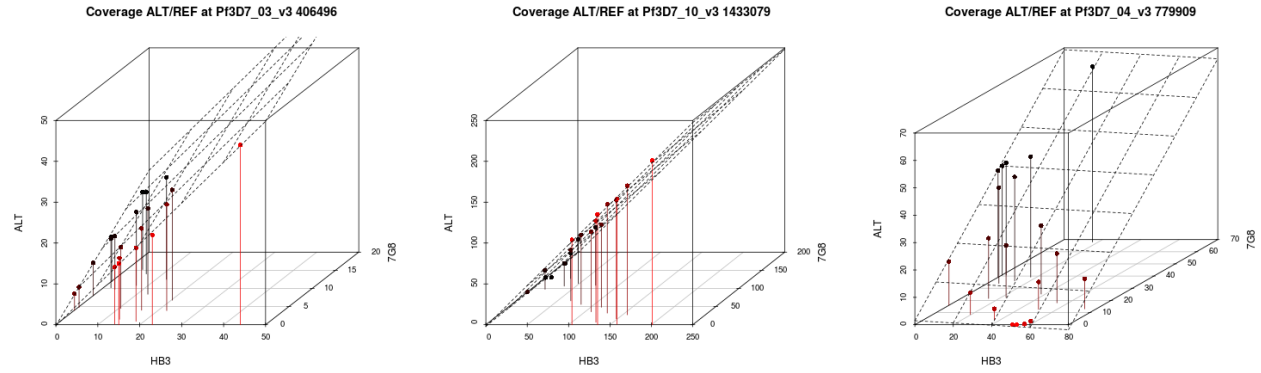
References

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,



(a) (b)

Figure S3.1: XXXXXXXXXXXXXX



(a) (b) (c)

Figure S3.2: XXXXXXXXXXXXXX

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*.

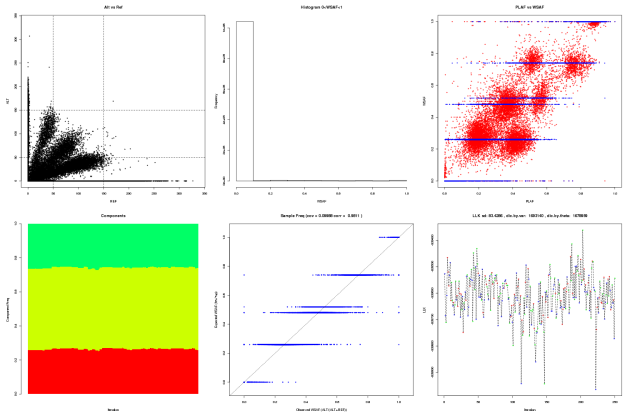
Wendler, J. (2015). *Accessing complex genomic variation in Plasmodium falciparum natural infection*. Ph. D. thesis, University of Oxford.

Supplemental Materials of DEploid

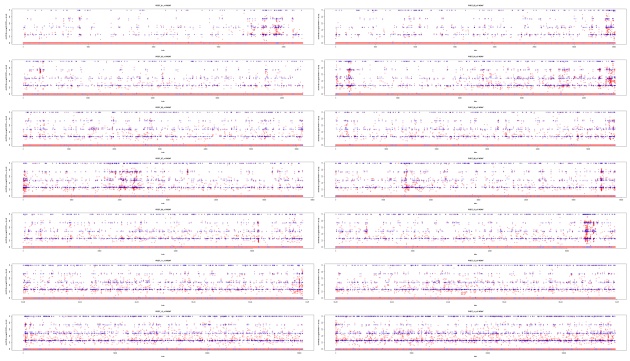
S4 DEploid

Our program *DEploid* s freely available at `https://github.com/mcveanlab/DEploid` under the conditions of the GPLv3 license. A detailed document can be found at `http://deploid.readthedocs.io/en/latest/`.

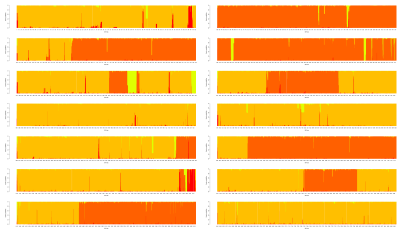
- (a)
    1. alt vs ref
    2. wsaf hist
    3. wsaf vs plaf
    4. proportion
    5. wsaf obs vs est
    6. llk
- (b)
- (c) (d) and (e)



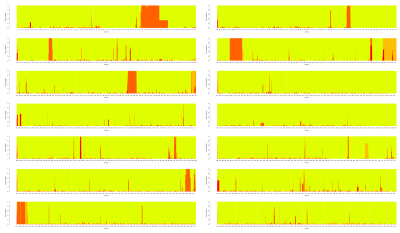
(a)



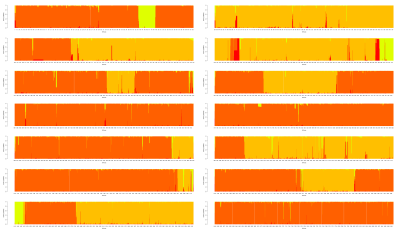
(b)



(c)

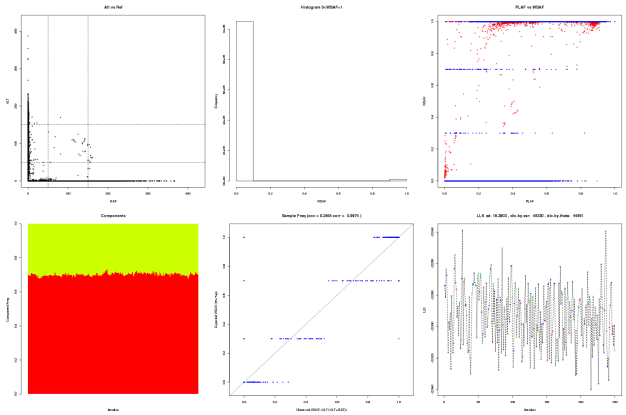


(d)

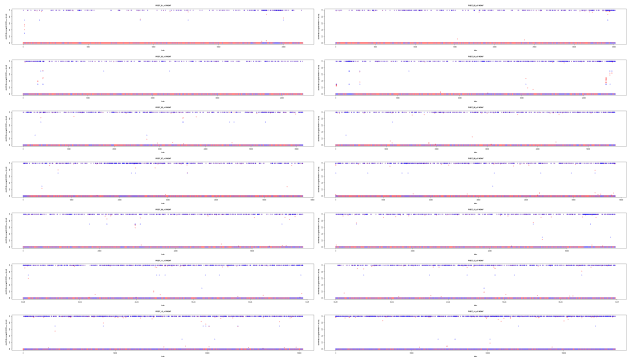


(e)

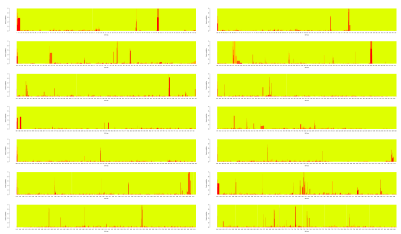
Figure S4.1



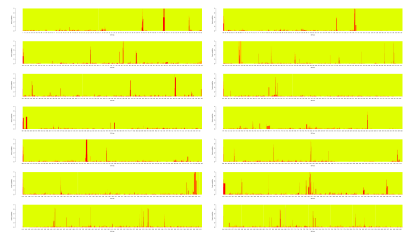
(a)



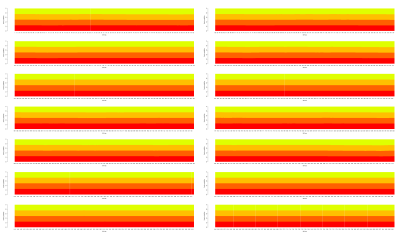
(b)



(c)



(d)



(e)

Figure S4.2