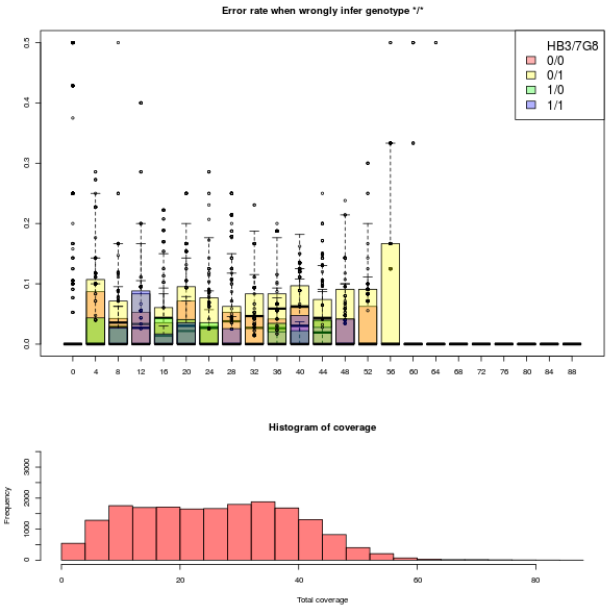
Supplemental Materials of using data with different coverages

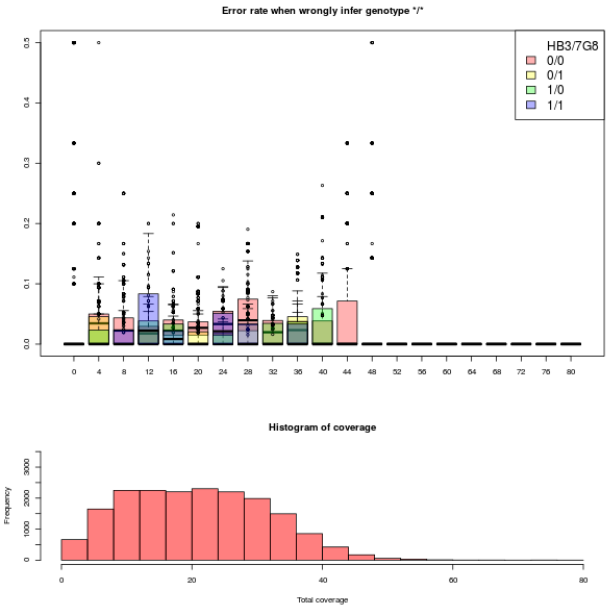
S1 Method validation on lab controlled strains

We experimented 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. We summerise our findings as follows:

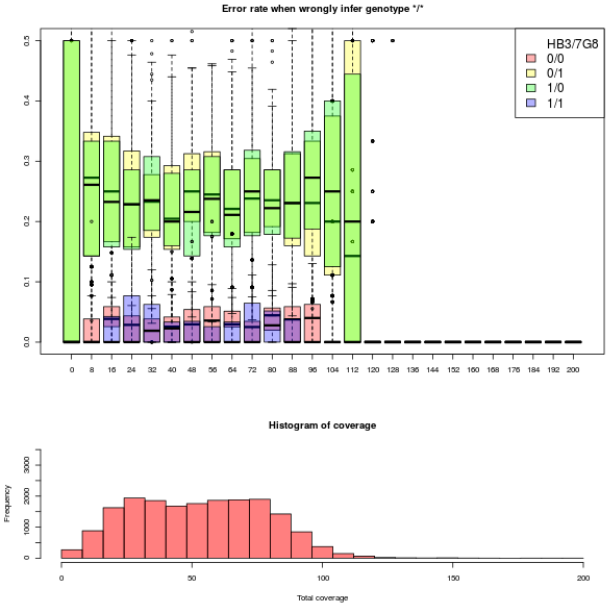
- Figure S1.1(a) suggests that for the haplotype inference, the minor strain in particular, heavily relies on the reference panel for low coverage data.
- Figure S1.1(b) suggests that sequence coverage have little, almost none affect on haplotype inference of balanced mixtures, when a perfect panel is provided.
- Figure S1.1(c) and (d) show the reduced error rate in higher coverage data.
- All figures show that deconvoluting heterozygous sites is challenging.
- All figures show that even at homozygous sites, there is still error for the inference.



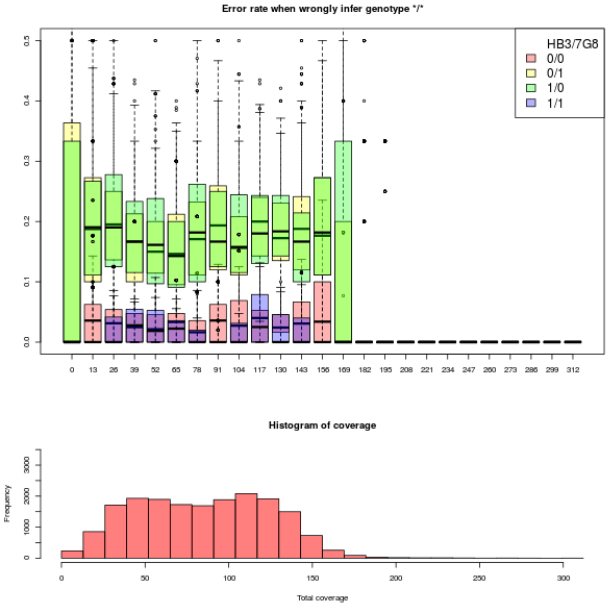
(a) PG402-C low coverage data decovolution with panel V.



(b) PG406-C low coverage data decovolution with panel V.



(c) PG406-C median coverage data decovolution with panel I.



(d) PG406-C high coverage data decovolution with panel I.

Figure S1.1: Error rate of a particular genotype inference at different converages

Supplemental Materials of DEploid

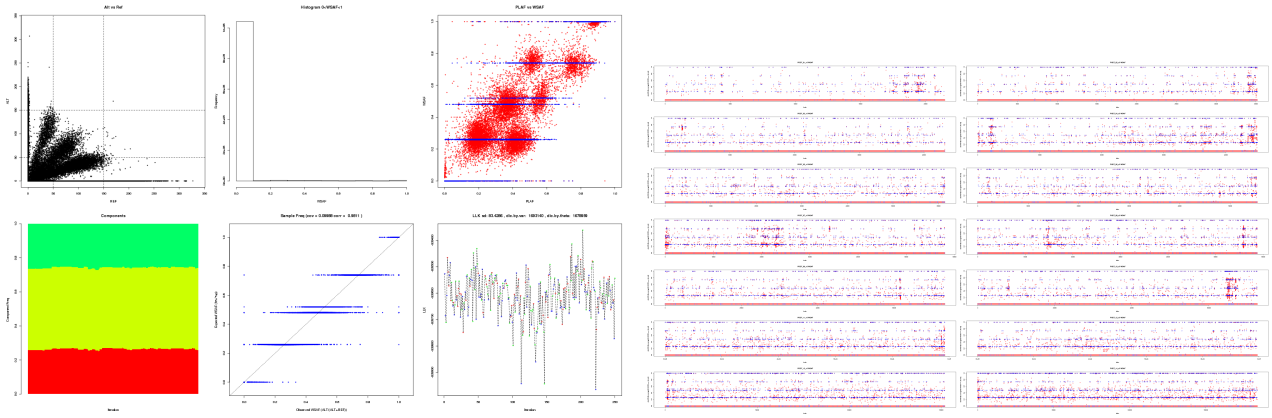
S2 DEploid

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/`.

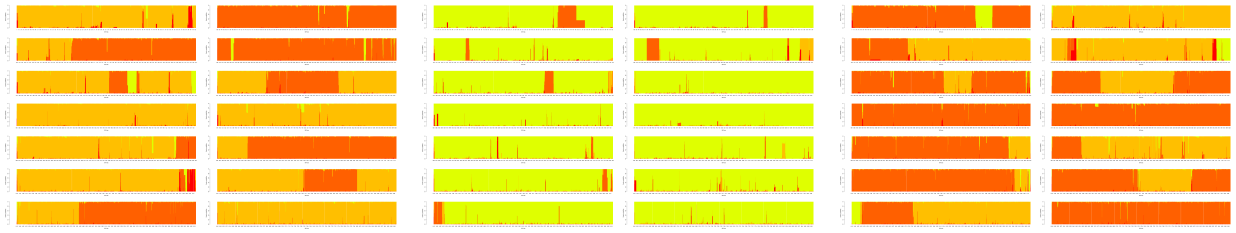
Here we show examples of deconvolution with reference panel V: 3D7, HB3, 7G8 and Dd2 strains. All figures are generated from plotting utilities with DEploid.

- (a) Panel figures of interpreting DEploid output. Figures from the top to the bottom, the left to the right show:
    1. Alternative read count vs reference red count, which is used for exploring the data.
    2. Histogram of the allele frequencies within sample.
    3. Allele frequencies at the population level vs allele frequencies at within sample. The PLAF is calculated from the total read counts. Red dots show observed WSAF, blue does show the expected WSAF inferred from our model.
    4. MCMC samples of the proportions.
    5. Expected WSAF vs observed WSAF.
    6. Log likelihood of the MCMC chain.
- (b) Expected WSAF (blue) and observed WSAF (red) at every site.
- (c) (d) and (e) Posterior probabilities (Li and Stephen's model) of deconvoluted strain with strains in panel V.

Figure S2.2 (a) suggests that our model overfits this clonal sample as a mixture of two strains. Figure S2.2 (c) and (d) suggest the two strains are in fact the same strain with subtle differences.

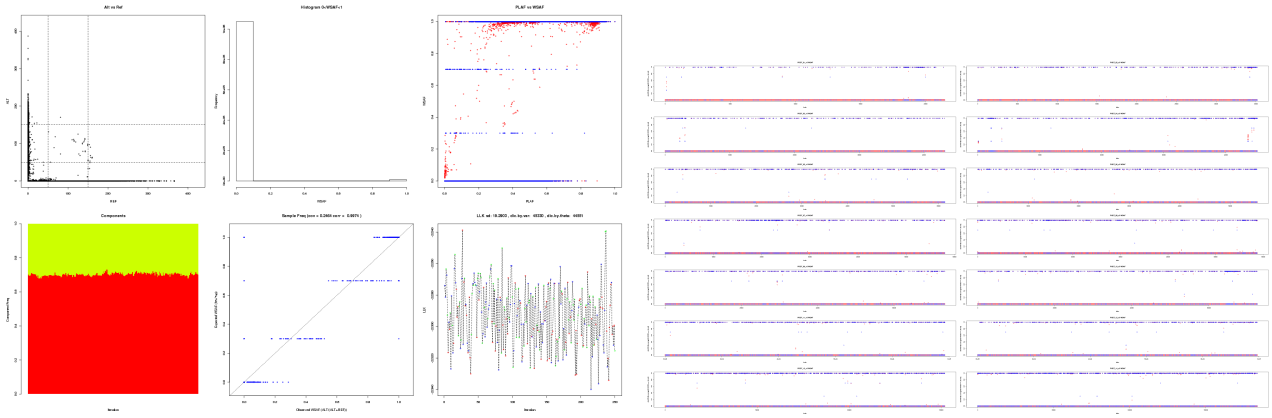


(a) (b)

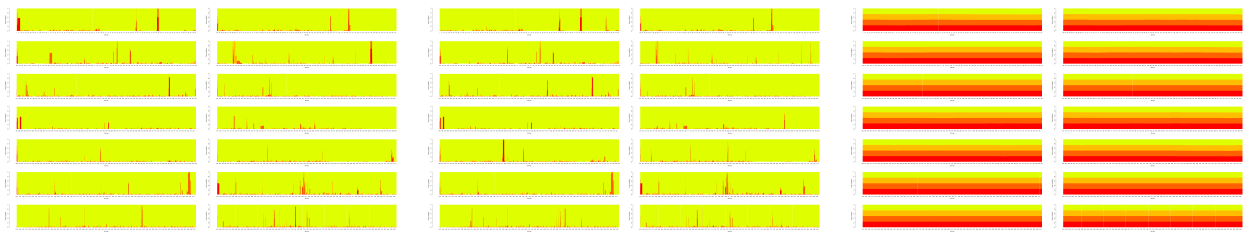


(c) (d) (e)

Figure S2.1: Sample PG0396-C deconvolution with reference panel V.



(a) (b)



(c) (d) (e)

Figure S2.2: Sample PG0415-C deconvolution with reference panel V.