Online haplotype inference and Compression of personal genomes

Yosuke Tanigawa (Biomedical Informatics Ph.D. program, ytanigaw@stanford.edu)
Manuel A. Rivas (Department of Biomedical Data Science, mrivas@stanford.edu)

February 12, 2017

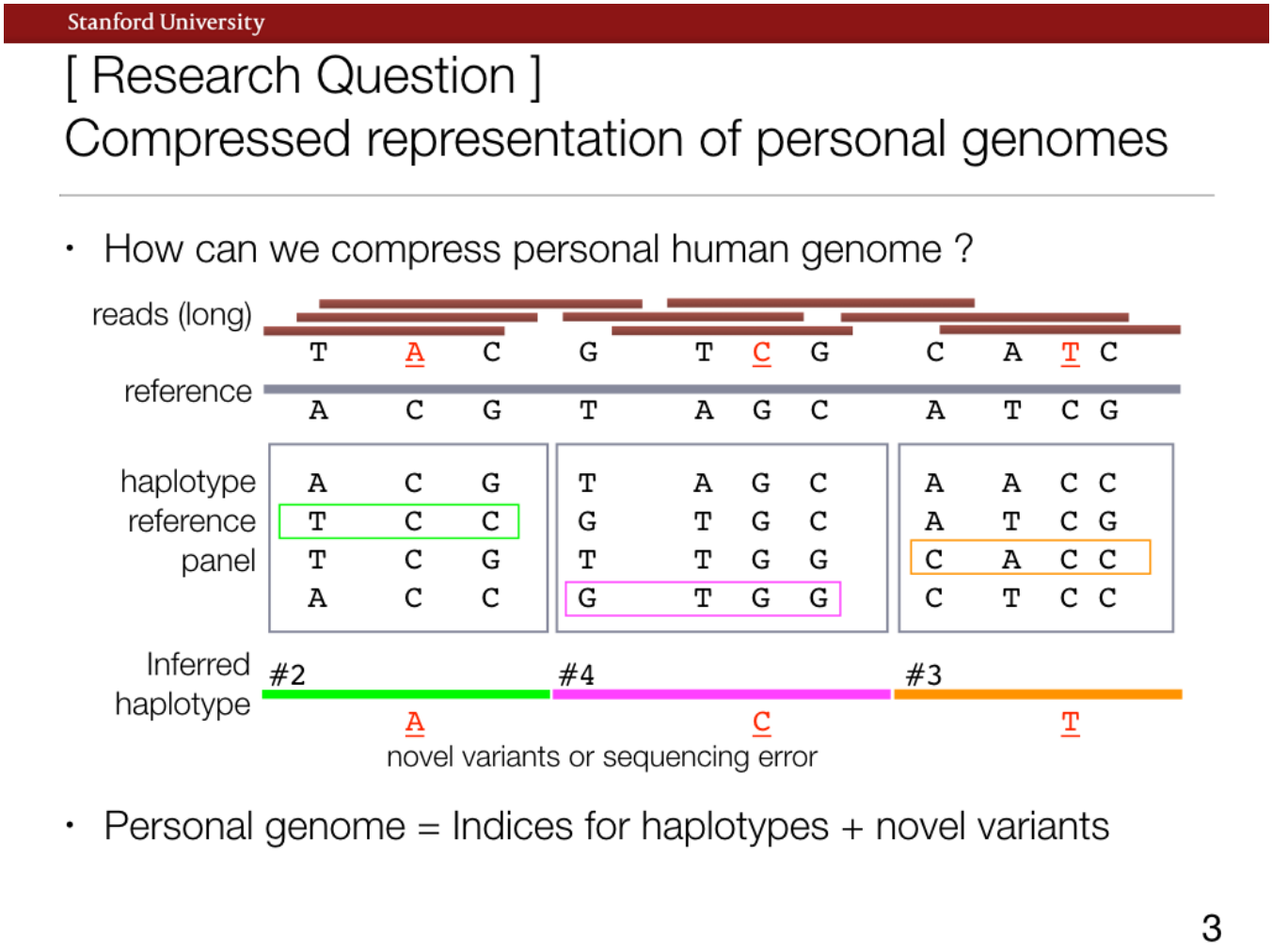

Figure 1: Frequency of mismatches in ONT data sets

Motivation

We are interested in the problem of using a reference genome, along with reference haplotypes (say from the UK Biobank), to compress online calling of variants in the genome. Fig. (1) is a schematic diagram.

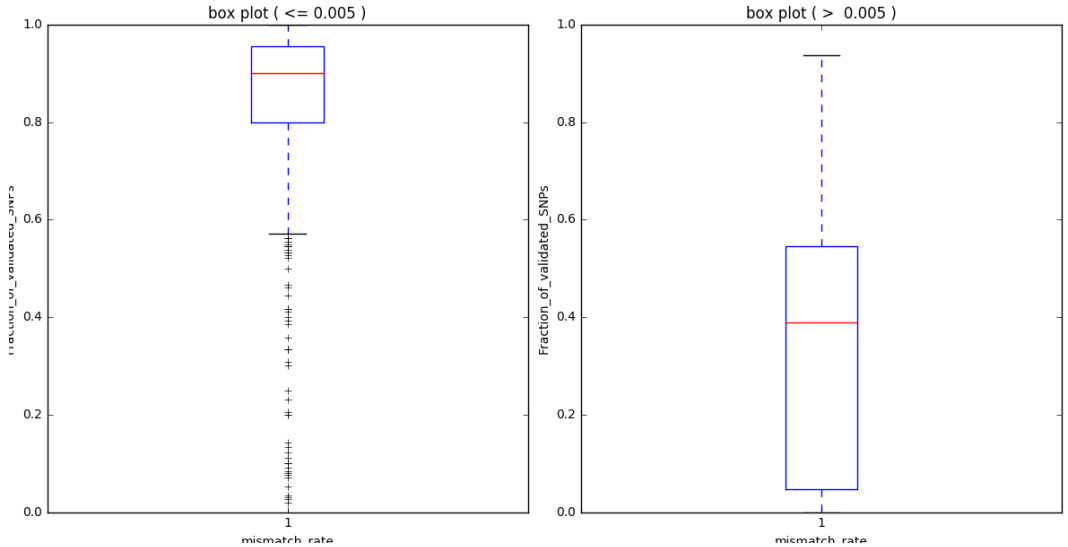

Figure 2: Error rate estimate on NA12878 sample. (left) $1 -$ error rate of mapped fragments with mismatch rate ≤ 0.005. (right) $1 -$ error rate of mapped fragments with mismatch rate > 0.005. We have implemented a pipeline that outputs informative mismatch sites on selected mapped fragments.

Simple model

For window W we want to compute

$$P(\text{haplotype of individual}_i = h \mid \text{data } D) \quad h \in \{1, 2, \ldots, H\} \tag{1}$$

This is proportional to the likelihood of the data conditional on the haplotype times the prior probability of the haplotype (i.e. frequency of the haplotype in the reference population):

$$P(\text{haplotype of individual}_i = h \mid \text{data } D) \propto P(\text{data } D \mid \text{haplotype of individual}_i = h) P_{\text{prior}}(h) \tag{2}$$

The likelihoods can be computed as we receive more read information. A binomial likelihood can be used where the error rate is estimated empirically(Fig. (2)).

Likelihood

Suppose we have n SNP sites on a given read (mapped fragment), a specific haplotype h in our mind, and know error rate ϵ of the sequencing machinery. We found $0 \leq x \leq n$ mismatches between read and haplotype. Then, likelihood is

$$P(\text{data } D \mid \text{haplotype of individual}_i = h) = \epsilon^x (1 - \epsilon)^{n-x} \tag{3}$$

Maximum likelihood estimate of prior distribution of haplotype

Prior distribution of haplotype can be found as a maximum likelihood estimate on population reference panel, i.e.

$$P_{\text{prior}}(h) = \frac{\text{frequency of haplotype } h}{\text{total \# of haplotypes}} \tag{4}$$

Preliminary results

We applied this simple model to a single mapped fragment to validate our approach. 1000 genome phase 3 data was used as a reference panel.

Fig. (3) shows the prior probability of haplotypes and Fig. (4) shows the posterior probability.

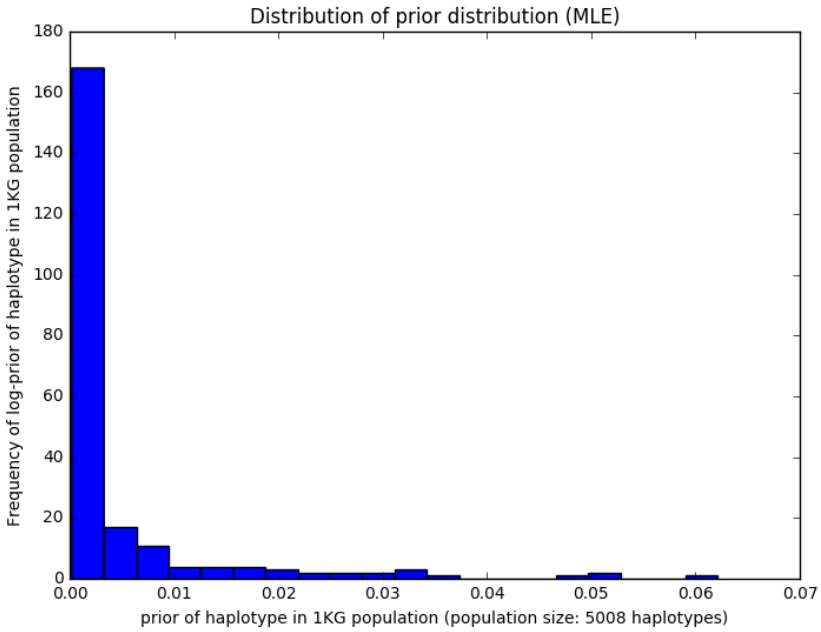

Figure 3: Prior distribution

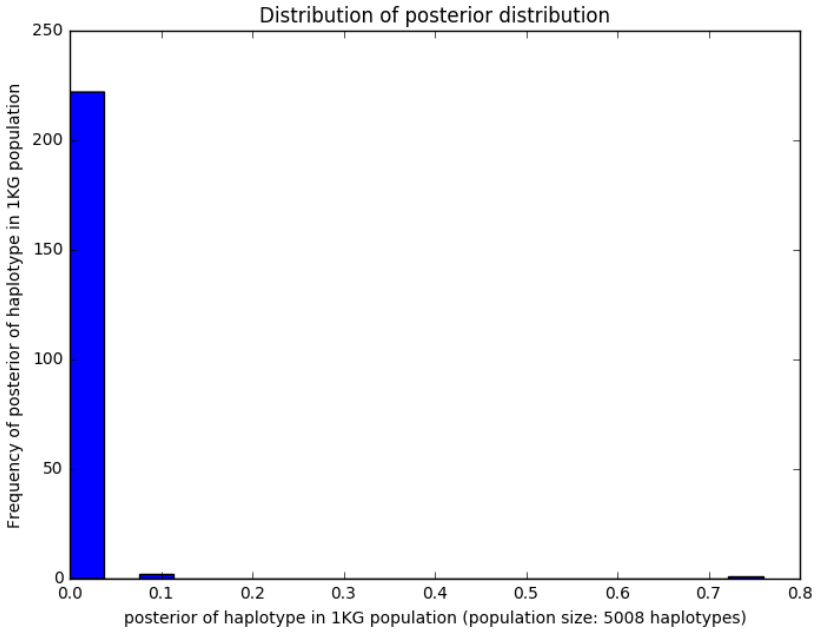

Figure 4: Prior distribution