Online haplotype inference and Compression of personal genomes (progress report)

2017/02/13

Yosuke Tanigawa (ytanigaw@stanford.edu)
M. Rivas Lab (rotation student) | Biomedical Informatics Ph.D. Program

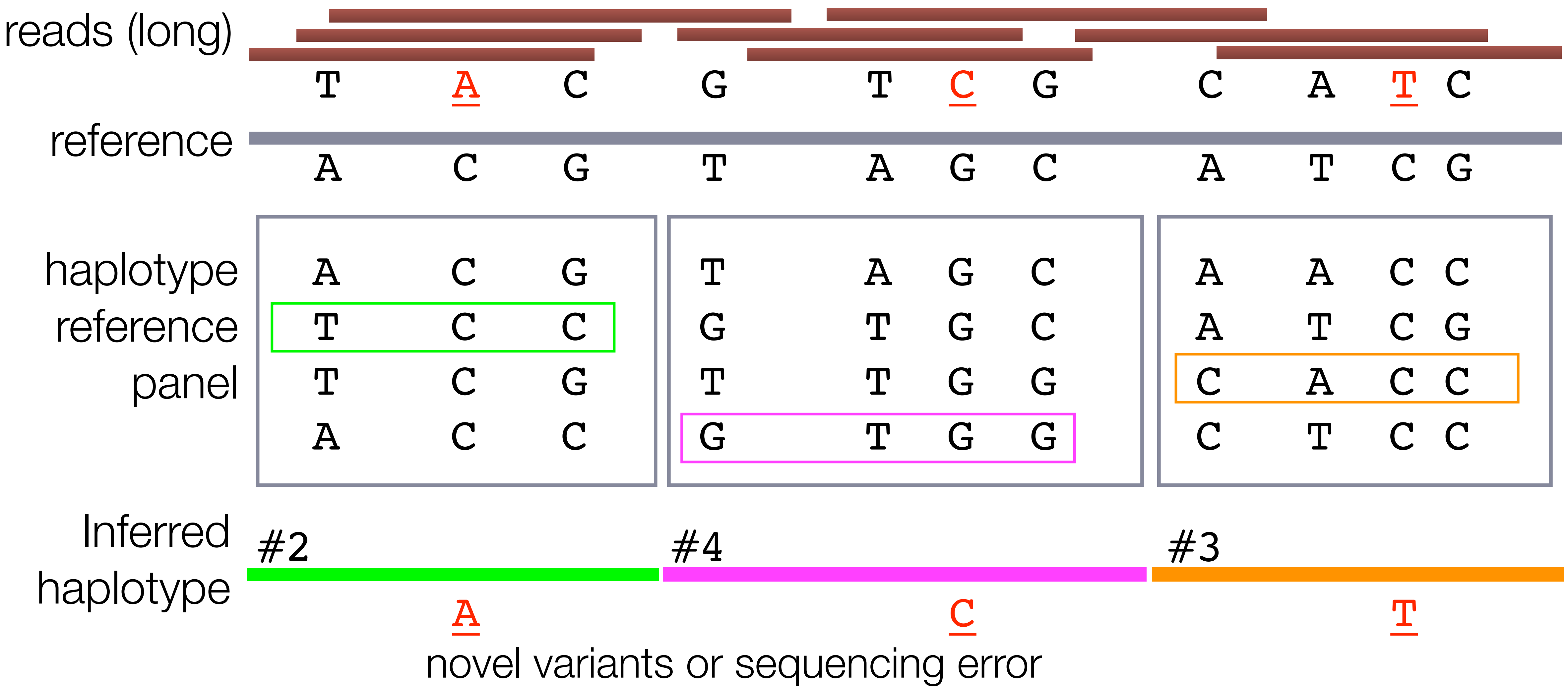
Background
Long read sequencer & Haplotype reference panel

- Long read sequencers (Oxford Nanopore, PacBio)
  - Read length >= 8 kb
  - Review paper of Nanopore

- Haplotype reference
  - UK BioBank (152,729 imputed haplotypes)
  - Haplotype reference consortium (not available yet?)
- Compressed data representation of haplotypes
  - PLINK2

[Research Question]
Compressed representation of personal genomes

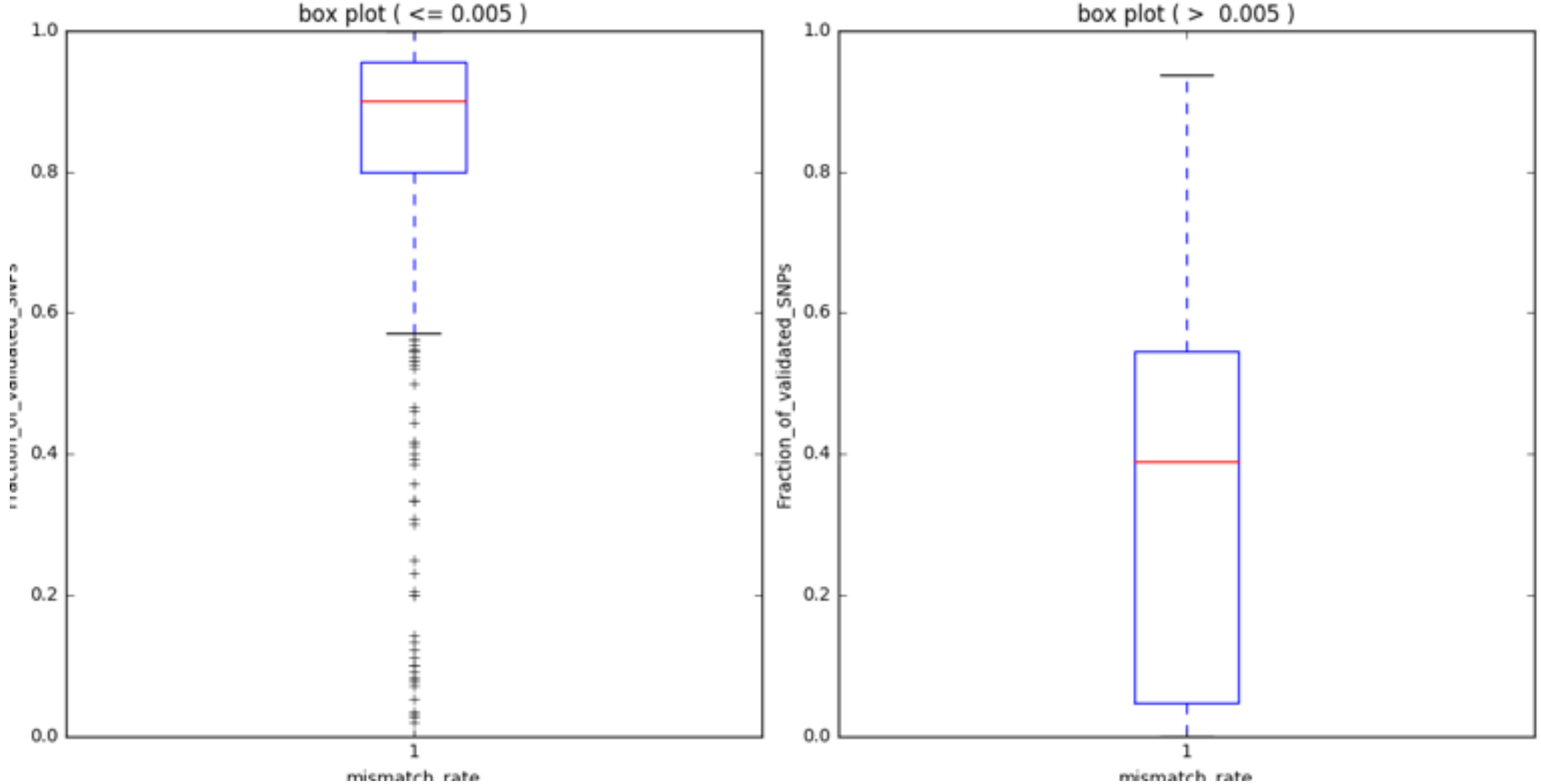
- How can we compress personal human genome ?



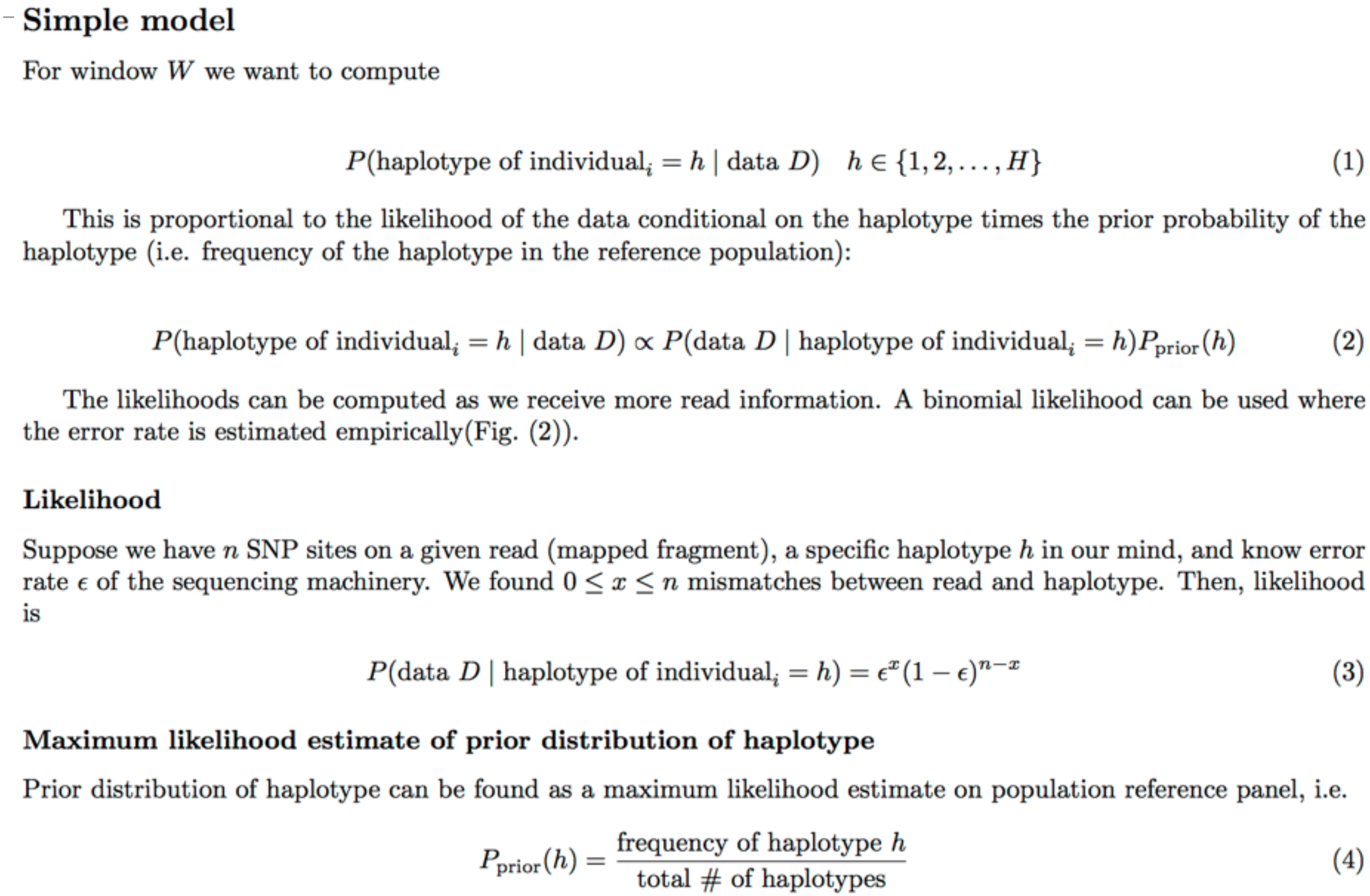
- Personal genome = Indices for haplotypes + novel variants

Estimate of error rate with NA12878

- Mismatch rate on fragement: <= 0.005, right > 0.005
y-axis: 1 - error rate (#SNPs called by pipeline / truth)



Model: haplotype inference



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

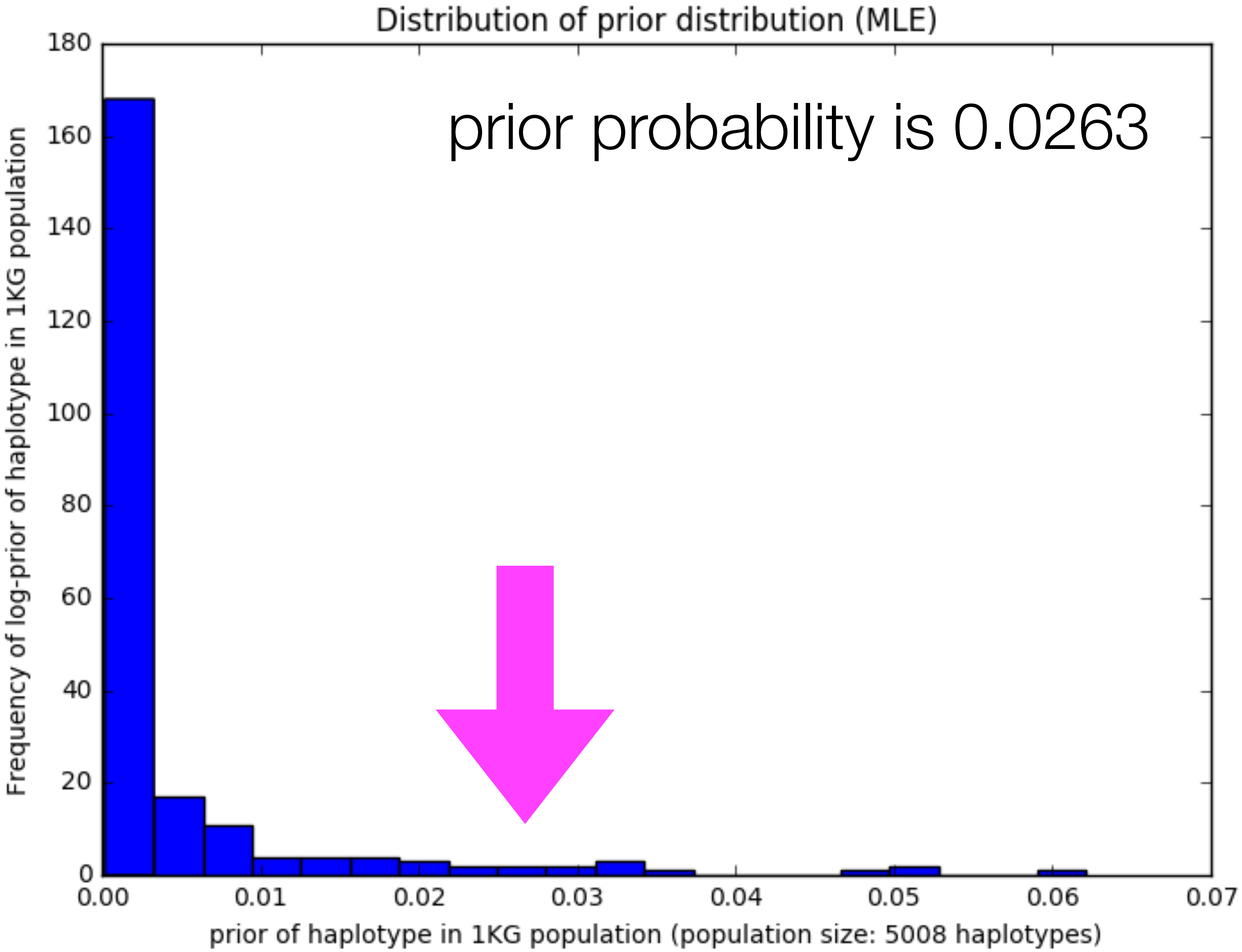
Prior distribution of haplotype can be found as a maximum likelihood estimate on population reference panel, i.e.

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

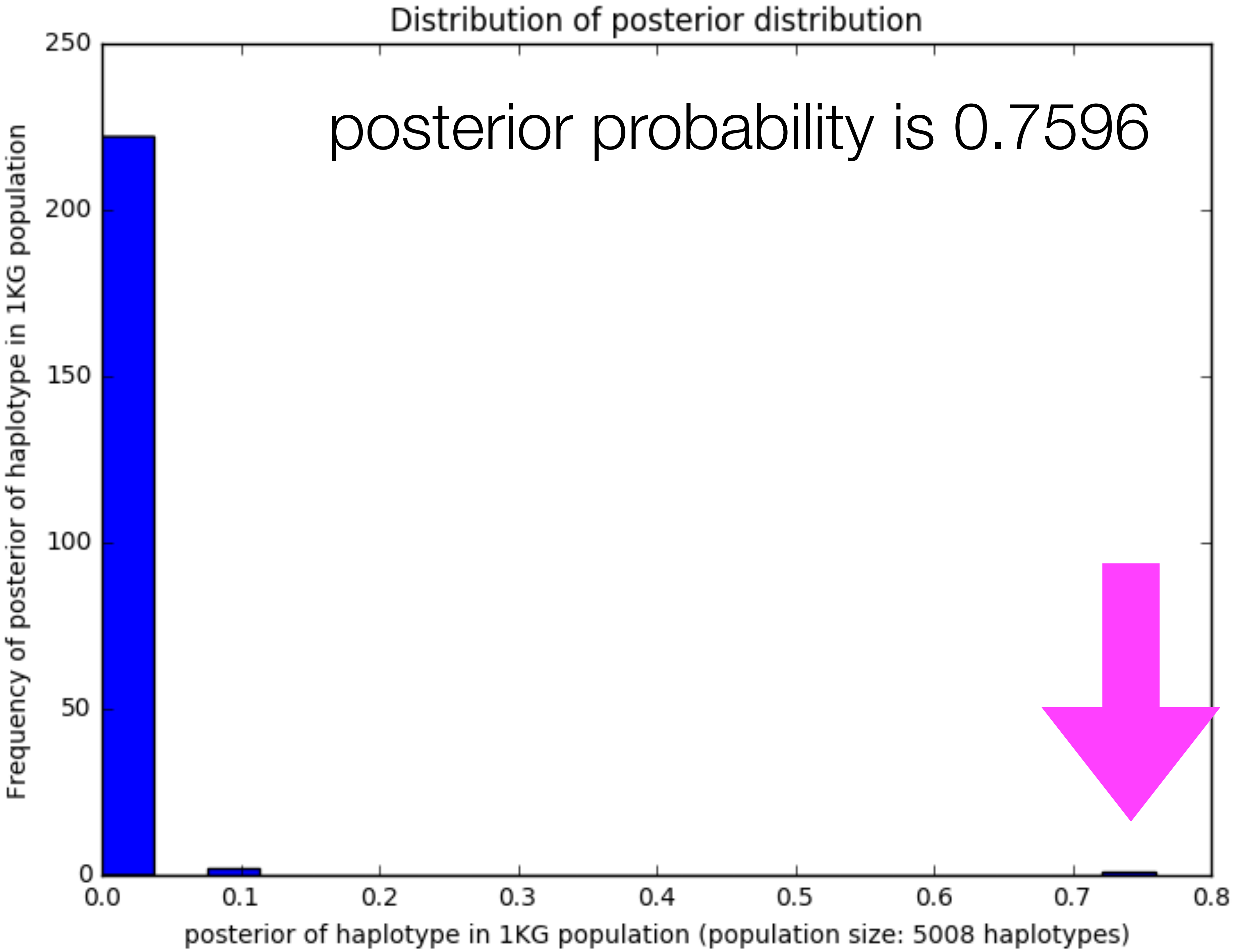
Preliminary results on chr. 20 with 1000 genome

- Prior: Maximum likelihood estimate (1KG reference)
- Dataset Nanopore consortium
- Example mapped fragment on chromosome 20
    - Length = 30 kb, 15 (validated) SNPs
- Jupyter notebook is on GitHub

Preliminary results on chr. 20 with 1000 genome



Preliminary results on chr. 20 with 1000 genome



posterior probability is 0.7596

Estimate of compression level

- ~250 haplotypes in 30kb region
  8bit (1byte) for 30kb region would be sufficient

- Genome = 2 * 3 * 10^9 = 2 * 10^5 * (30 kb)
  2 * 10^5 * 1byte = 200 k byte
  for 2 sets of haplotype

- Current representation
  vcf file of NA12878: 40M byte (incl. rare variant)

Replacing hashing with encryption ?

- Haplotype is a bit string
  000100100101011: haplotype for 15 SNPs
  2^{15} possible haplotype,
  but not all of them are present in population

- Haplotype —> index (relatively small integer)
  we will use **hash** function for this mapping

- Haplotype —> codeword (shorter bit string)
  compression & **encryption** of personal genome ?

What’s the next?

- How long does it takes to infer haplotype ?
  - we can estimate it
  - look at time stamp on the raw data

- How much coverage do we need ?

- Output format
  - want to perform inference on compressed genome

- Rare variants

- Applications (HLA, forensic, etc.)