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

2017/02/27

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

Background

- Most of the variants are shared in the population
- Long-read sequencing technologies
  - Oxford Nanopore MinION sequencer (N50: 12.5 kb)

- Biobank projects -> Haplotype reference panel data
  - UK BioBank (152,729 imputed haplotype)
- Compressed representation of haplotype in a population
  - PLINK2

Research questions

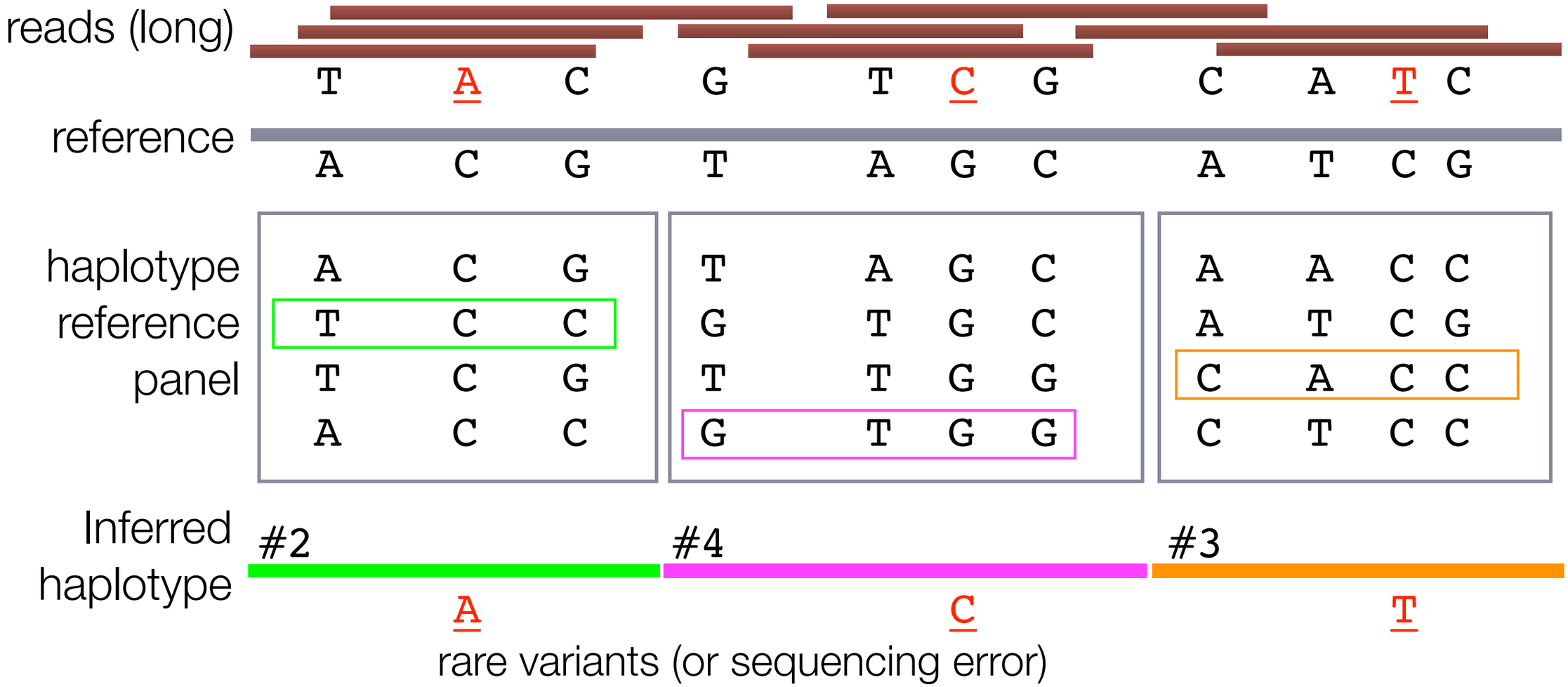
1. Can we infer haplotypes of personal genomes **on the fly** with portable sequencers?

2. What would be the **compact** & **efficient** representation of personal genomes?

3. How people can **share personal genomes** with researchers while maintaining their **privacy**?

[Method]
Compressed representation of personal genomes

- How can we compress personal human genome ?



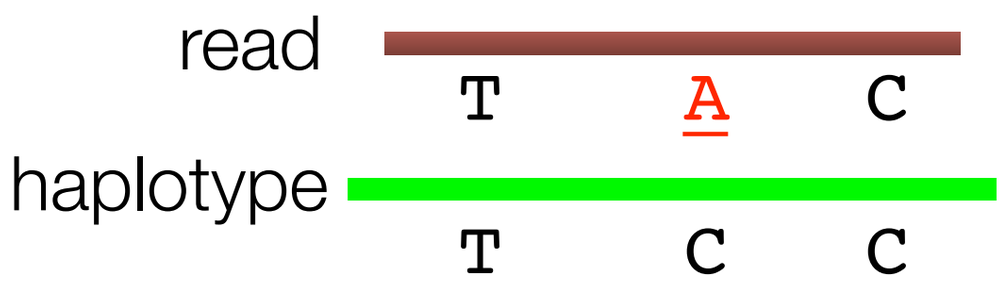
- Personal genome = Indices for haplotype + rare variants

[Method]

$$p(\text{haplotype} \mid \text{read}) \propto p(\text{read} \mid \text{haplotype}; \epsilon) p(\text{haplotype})$$

- Prior $p(\text{haplotype})$
  - Maximum likelihood estimate on population

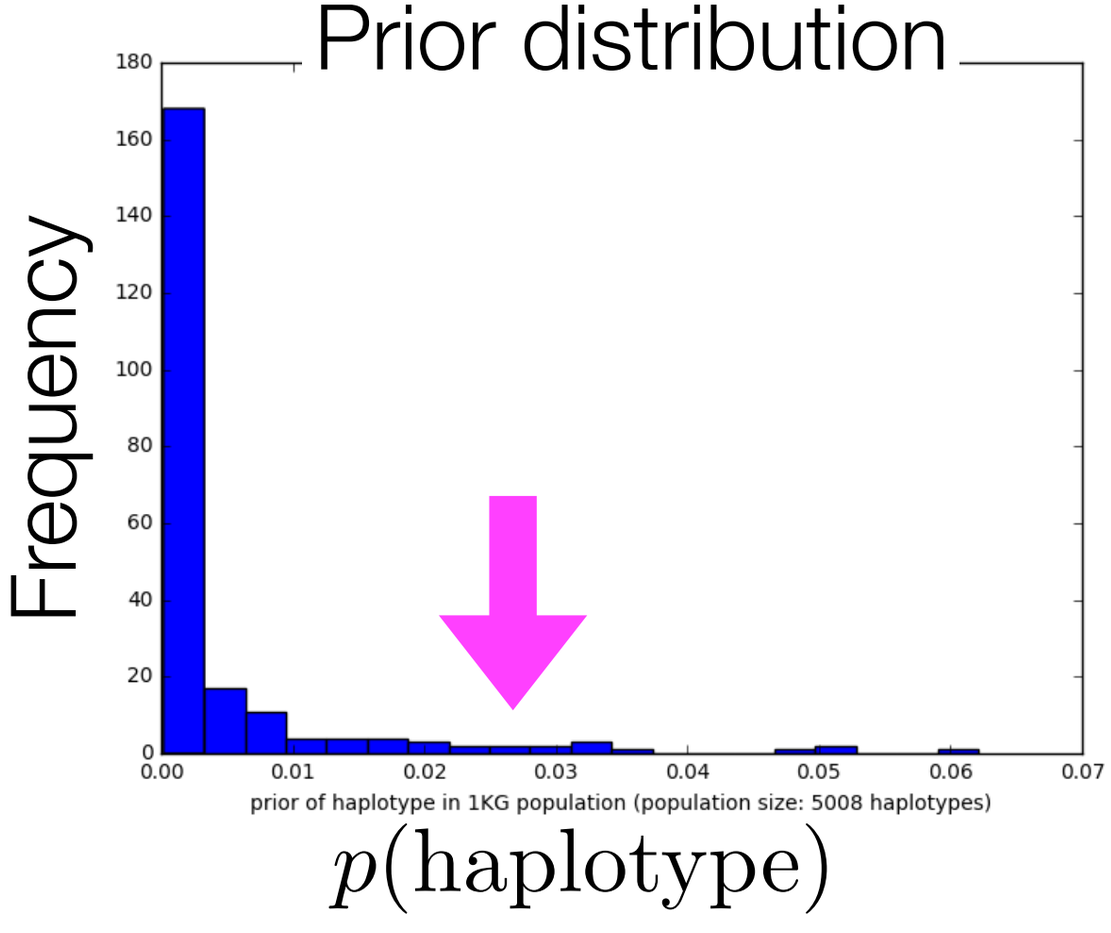
- Likelihood $p(\text{read} \mid \text{haplotype}; \epsilon)$
  - Binomial model $p(\text{read} \mid \text{haplotype}; \epsilon) = \epsilon^x (1-\epsilon)^{n-x}$
  - n : # polymorphic sites
  - x : # mismatches
  - ϵ : error rate

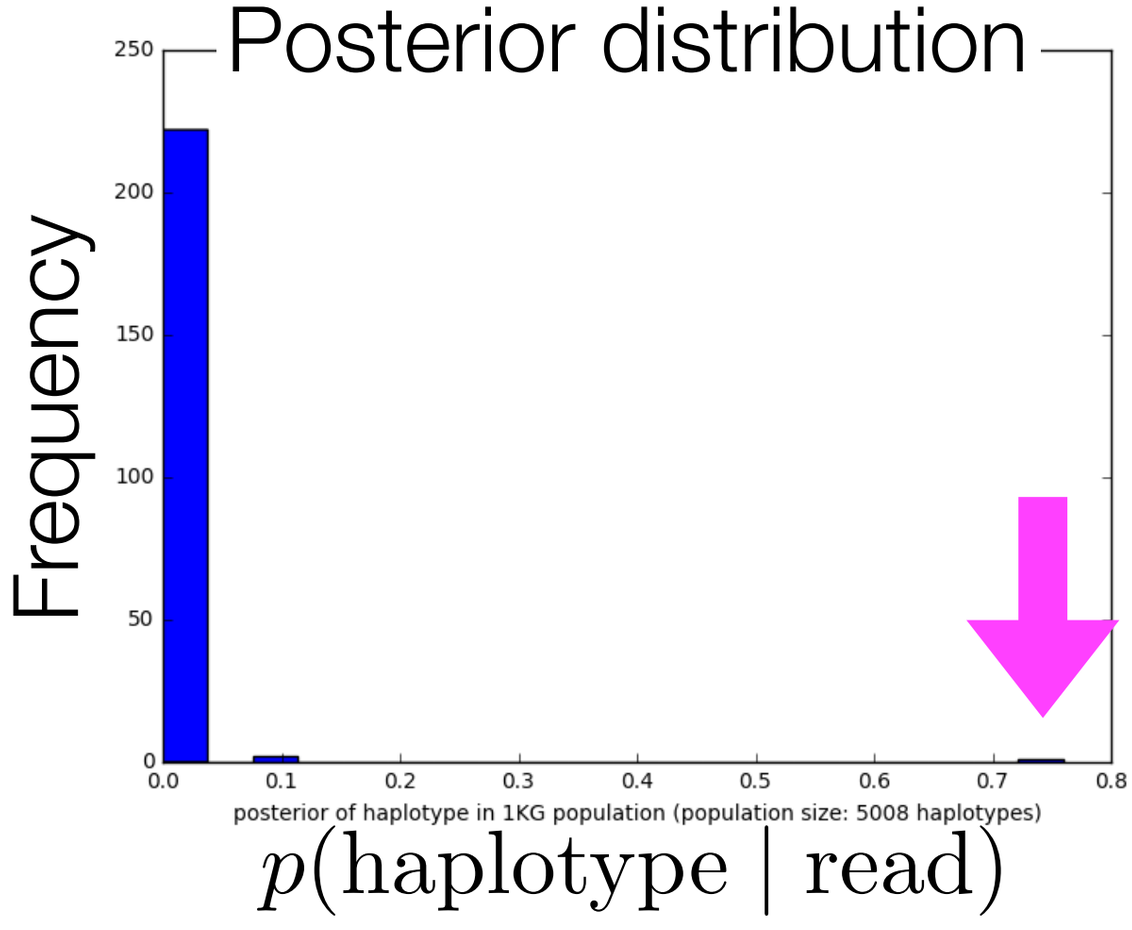


- We can estimate error rate ϵ
  - $\widehat{\epsilon}$ mismatch rate on non-polymorphic site

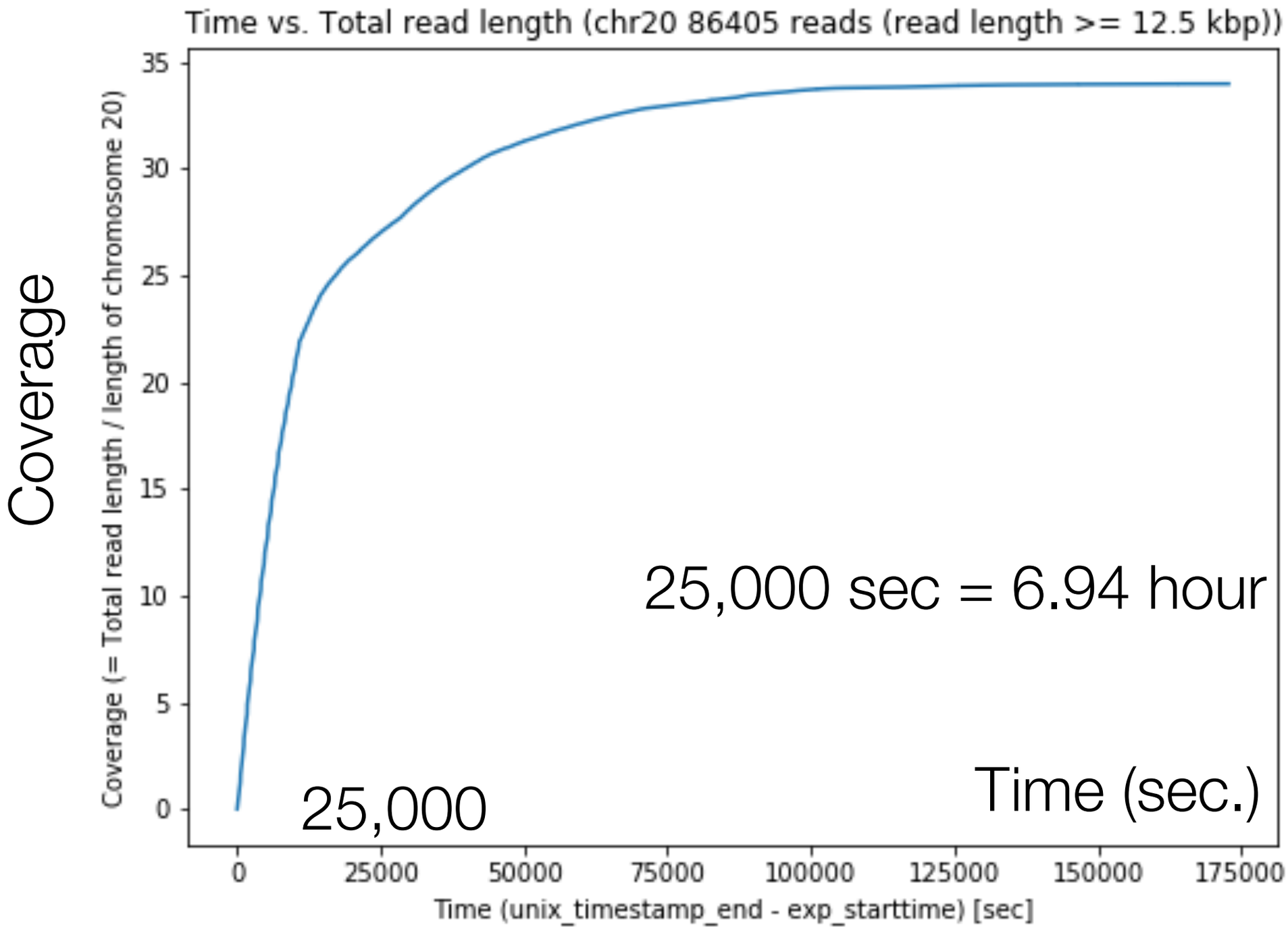
[Preliminary results]
read: NA12878; population: 1000 Genome (phase 3)

- Example read: chr20; 30kb long; 15 SNPs
- prior = 0.0263 —> posterior = 0.7596
- Only one read is sufficient to find the right haplotype





How long does it take to infer haplotype?



Nanopore consortium rel.3 has ~60x coverage