Compressed representation of personal genomes and its application for real-time haplotype inference

Yosuke Tanigawa[1], Manuel Rivas[2]
[1]Biomedical Informatics, Stanford University, [2]Department of Biomedical Data Science, Stanford University

Abstract

Recent advancements in sequencing and genotyping technologies enabled characterization of personal genomes for millions of individuals. However, the current de facto analysis pipeline is not scalable in terms of both data storage cost and computation time. Here, we present a novel compressed representation of personal genomes with an aid of large-scale population reference. We demonstrate that our approach can (1) reduce the storage space by orders of magnitude and (2) accelerate statistical inference. As an application, we demonstrate a method to perform real-time inference of haplotypes from long-read sequencing data. We hope our approach facilitates biomedical findings from personal genomes of large-scale populations.

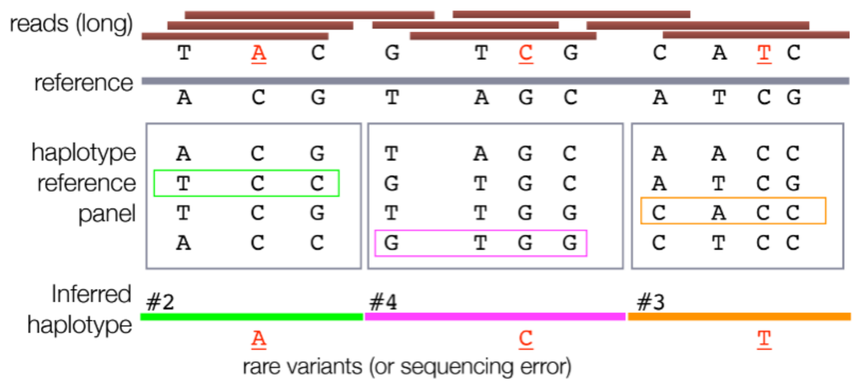
Research Questions

- Can we infer haplotypes of personal genomes **on the fly** with portable sequencers?
- What would be the **compact & efficient** representation of personal genomes?
- How can people **share personal genomes** with researchers while maintaining participants's **privacy**?

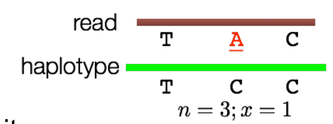
Oxford Nanopore Technologies SmidgION

[Method] Compact representation of personal genome

- Personal genome = haplotype indices + personal rare variants

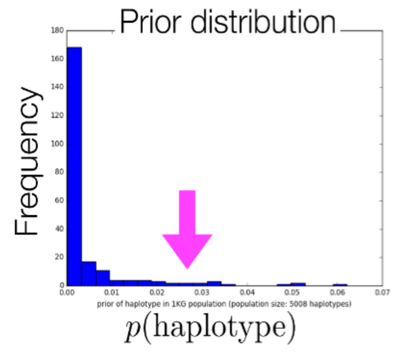


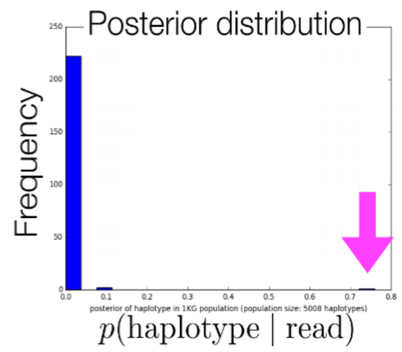
- Population reference captures most of the common variants
- Long-reads contain multiple polymorphic sites
- $p(\text{haplotype} \mid \text{read}) \propto p(\text{read} \mid \text{haplotype}; \epsilon) p(\text{haplotype})$
- Prior $p(\text{haplotype})$: Maximum Likelihood Estimate
- Likelihood $p(\text{read} \mid \text{haplotype}; \epsilon)$:
  Binomial model $p(\text{read} \mid \text{haplotype}; \epsilon) = \epsilon^x (1-\epsilon)^{n-x}$
  - n : # of polymorphic sites
  - x : # of mismatches
  - ϵ : error rate
- We can estimate error rate ϵ
  - $\hat{\epsilon}$: mismatch rate on non-polymorphic sites



[Preliminary Results] Rapid haplotype inference

- Our Bayesian approach can infer haplotype with a few reads





- Data: Nanopore consortium NA12878
- Population reference: 1000 genome project phase 3
- Example read: chr20; 30kb long regions; 15 SNPs
- Prior probability of the correct haplotype = 0.0263
- Posterior probability of the correct haplotype = 0.7596
- Only one read is sufficient to find the correct haplotype

Future prospects

- Explicitly model linkage-disequilibrium (LD) structure
- Build an end-to-end pipeline
- Validate on external dataset (e.g. UK Biobank population)