With the availability of genotyping data of very large samples, there is an increasing need for tools that can efficiently identify genetic relationships among all individuals in the sample. Modern biobanks cover genotypes up to 0.1%–1% of an entire large population. At this scale, genetic relatedness among samples is ubiquitous. However, current methods are not efficient for uncovering genetic relatedness at such a scale. We developed a new method, Random Projection for IBD Detection (RaPID), for detecting Identical-by-Descent (IBD) segments, a fundamental concept in genetics in large panels. RaPID detects all IBD segments over a certain length in time linear to the sample size. We take advantage of an efficient population genotype index, Positional BWT (PBWT) developed by Richard Durbin. PBWT achieves linear time query of perfectly identical subsequences among all samples. However, the original PBWT is not tolerant to genotyping errors which often interrupt long IBD segments into short fragments. The key idea of RaPID is that the problem of approximate high-resolution matching over a long range can be mapped to the problem of exact matching of low-resolution subsampled sequences with high probability. PBWT provides an appropriate data structure for bi-allelic data. With the increasing sample sizes, more multi-allelic sites are expected to be observed. Hence, there is a necessity to handle multi-allelic genotype data. We also introduce a multi-allelic version of the original Positional Burrows-Wheeler Transform (mPBWT).

Some regions of genome may be shared by multiple individuals rather than only a pair. Clusters of identical haplotypes could reveal information about the history of intermarriage, isolation of a population and could also be medically important. We propose an efficient method to find clusters of identical segments among individuals in a large panel, called cPBWT, using the PBWT data structure. The time complexity of finding all clusters of identical matches is linear to the sample size. The human genome harbors several runs of homozygous sites (ROHs) where identical haplotypes are inherited from each parent. We applied cPBWT on UKBiobank and searched for clusters of ROH region that are shared among multiple. We discovered strong associations between ROH regions and some non-cancerous diseases, specifically autoimmune disorders.

The increasingly large cohorts of whole genome genotype data present an opportunity for searching genetically related people within a large cohort to an individual. At the same time, doing so efficiently presents a challenge. The PBWT algorithm offers constant time matching between one haplotype and an arbitrarily large panel at each position, but only for the set maximal matches. We used the PBWT data structure to develop a method to search for all matches of a given query in a panel. The matches larger than a given length correspond to the all shared IBD segments of certain lengths between the query.

Major: Computer Science

Educational Career:
Bachelor's of Computer Science, BS, 2010, Technical University of Darmstadt
Master's of Computer Science, MS, 2012, Technical University of Darmstadt

Committee in Charge:
Shaojie Zhang, Chair, Computer Science
Charles E. Hughes, Computer Science
Shibu Yooseph, Computer Science
Degui Zhi, University of Texas Health Science Center at Houston

Approved for distribution by Shaojie Zhang, Committee Chair, on October 11, 2018.
The public is welcome to attend.