Abstract: With the wide availability of whole-genome genotype data, there is an increasing need for conducting genetic genealogical searches efficiently. Computationally, this task amounts to identifying shared DNA segments between a query individual and a very large panel containing millions of haplotypes. We have developed two algorithms to find all L-long matches, matches longer than or equal to a given length L, between a query and a panel. In the first algorithm, PBWT-Query, we introduce “virtual insertion” of the query into the PBWT matrix of the panel, and then scanning up and down for the PBWT match blocks with length greater than L. In our second algorithm, L-PBWT-Query, we further speed up PBWT-Query by introducing additional data structures that allow us to avoid iterating through blocks of incomplete matches. The efficiency of PBWT-Query and L-PBWT-Query is demonstrated using the simulated data and the UK Biobank data. Our results show that our proposed algorithms can detect related individuals for a given query efficiently in very large cohorts which enables a fast on-line query search.

Biography: Shaojie Zhang received his Ph.D. degree in Computer Science from the University of California, San Diego in 2007. He is currently an Associate Professor of the Department of Computer Science at the University of Central Florida. His research predominantly focuses on algorithms related to bioinformatics and computational biology, and in particular, on computational methods for non-coding RNA discovery, genomics and genetics data analysis, and high-throughput sequencing data analysis. Zhang also works on algorithms related to hardware gate-level netlist analysis.