A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data

Douglas W. Bjelland[1], Uday Lingala[1], Piyush S. Patel[1], Matt Jones[2], and Matthew C. Keller[1,2,*]

[1]Institute for Behavioral Genetics, University of Colorado at Boulder, Boulder, CO, 80303

[2]Department of Psychology & Neuroscience, University of Colorado at Boulder, Boulder, CO, 80301

*Corresponding author: Matthew Keller, matthew.c.keller@gmail.com

Abstract

Identical by descent (IBD) segments are used to understand a number of fundamental issues in genetics. IBD segments are typically detected using long stretches of identical alleles between haplotypes in whole-genome SNP data. Phase or SNP call errors in genomic data can degrade accuracy of IBD detection and lead to false positive calls, false negative calls, and under- or overextension of true IBD segments. Furthermore, the number of comparisons increases quadratically with sample size, requiring high computational efficiency. We developed a new IBD segment detection program, FISHR (Find IBD Shared Haplotypes Rapidly), in an attempt to accurately detect IBD segments and to better estimate their endpoints using an algorithm that is fast enough to be deployed on the very large whole-genome SNP datasets. We compared the performance of FISHR to three leading IBD segment detection programs: GERMLINE, refinedIBD, and HaploScore. Using simulated and real genomic sequence data, we show that FISHR is slightly more accurate than all programs at detecting long (greater than 3 cM) IBD segments but slightly less accurate than refinedIBD at detecting short (about 1 cM) IBD segments. Moreover, FISHR outperforms all programs in determining the true endpoints of IBD segments, which is important for several reasons. FISHR takes two to four times longer than GERMLINE to run, whereas both GERMLINE and FISHR were orders of magnitude faster than refinedIBD and HaploScore. Overall, FISHR provides accurate IBD detection in unrelated individuals and is computationally efficient enough to be utilized on large SNP datasets greater than 20,000 individuals.