Discussion

Our method provides a powerful tool for detecting genetic changes in multiple genomes associated with a common phenotype. We show improvements in the false discovery rate over a typical genome wide association study (GWAS) while accommodating data with relatively large amounts of inter-genome variability. We accomplish this with a very sensitive and tunable phylogenetic screen followed by more rigorous tests of alternative mechanisms. By systematically addressing alternative mechanisms of convergence related to chance, rather than merely tuning a false discovery rate parameter, we provide strong evidence for the association between positive results and the phenotype of interest while maintaining high power to detect small but important changes.

Methods currently in use include simple pattern matching (Foote et al. 2015) or associations across large number of very similar genomes (GWAS). With unlimited computing power finding genes which are molecularly convergent would be a trivial task. However, even if one could find all such homologies one would still need to determine if these were likely to be associated with the phenotype of interest. Pattern matching is computationally intensive and the results are difficult to interpret. GWAS break down when there is a large amount of heterogeneity among the genomes under consideration and are therefore ill suited for comparisons across species. Moreover, correctly targeting the false discovery rate in GWAS is a vexing problem (Korte and Farlow 2013; Andersson et al. 2009). Our method scales well with large data sets and uses a novel approach to reduce rate of false discovery while maintaining high sensitivity.

Our method is limited to detecting changes in only the protein coding portion of the genome. This is due to the difficulty in assessing the likelihood of DNA changes which are not associated with a change in amino acid. This limits our ability to detect important changes in potential regulatory pathways and micro-RNA (Ward and Kellis 2012). Future work on this method could incorporate entropy from biochemical processes, linkage disequilibrium, and copy number changes to assess the likelihood of single base-pair changes in non-coding portions of the genome. This would undoubtedly increase the utility of our analysis but would likely also increase the false discovery rate, a key advantage of our current method.

False positives in genome wide association studies can potentially waste both time and resources through the extended pursuit of noise. We believe our method can significantly reduce the number of false positive results thereby increasing the speed of valuable scientific discovery. However, as with all analyses which rely on association, false discoveries are possible. Therefore, we advise the use of rigorous experimental methodology to confirm a causal link for any genotype-phenotype association detected with our method.

Literature Cited

Andersson U, McKean-Cowdin R, Hjalmars U, and B Malmer. 2009. Genetic variants in association studies -- review of strengths and weaknesses in study design and current knowledge of impact on cancer risk. Acta Oncol. 48(7); 948-954

Korte A, and A Farlow. 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9; 9:29

Ward LD, and M Kellis. 2012. Interpreting noncoding genetic variation in complex traits and human disease. Nature Biotechnology 30; 1095-1106