Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Thompson et al., J Biomet Biostat 2013, 4:5
DOI: 10.4172/2155-6180.1000e126

Editorial Open Access

Using Ancestral Information to Inform Analyses of Complex Data Sets

Katherine L Thompson[1], Richard Charnigo[1,2]* and Catherine R Linnen[3]

[1]Department of Statistics, University of Kentucky, USA
[2]Department of Biostatistics, University of Kentucky, USA
[3]Department of Biology, University of Kentucky, USA

Abstract

Over the last decade, improvements in sequencing technologies coupled with active development of association mapping methods have made it possible to link genotypes and quantitative traits in humans. Despite substantial progress in the ability to generate and analyze large data sets, however, genotype-phenotype associations are often difficult to find, even in studies with large numbers of individuals and genetic markers. This is due, in part, to the fact that effects of individual loci can be small and/or dependent on genetic variation at other loci or the environment. Tree-based mapping, which uses the evolutionary relatedness of sampled individuals to gain information during association mapping, has the potential to significantly improve our ability to detect loci impacting human traits. However, current tree-based methods are too computationally intensive and inflexible to be of practical use. Here, we compare tree-based methods with more classical approaches for association mapping and discuss how the limitations of these newer methods might be addressed. Ultimately, these advances have the potential to advance our understanding of the molecular mechanisms underlying complex diseases.

Introduction

A central goal in the biological and biomedical sciences is to identify the genetic basis of morphological, physiological, behavioral, and disease traits. Over the last decade, improvements in deoxyribonucleic acid (DNA) sequencing technologies coupled with active development of genome-wide association (GWA) methods have made it possible to link genetic variation and quantitative traits in a wide range of organisms, including humans. However, despite substantial progress in our ability to generate and analyze large data sets, important statistical and bioinformatic challenges remain [1,2]. For example, while GWA studies have identified a large number of loci contributing to human disease, these loci rarely map to individual genes, let alone individual mutations [3-5]. Moreover, identified loci typically account for only a fraction of the total heritable variation in quantitative traits. To date, multiple overlapping explanations have been proposed to account for this "missing heritability" [6,7]. These explanations, some of which are described in more detail below, implicate several strategies for improving on current GWA methodology, including: increased sampling (of genetic regions and individuals), better measurements of traits and environmental variables, and improvements of existing statistical methodology. Here, we focus on the potential for using a novel statistical framework—tree-based association mapping—for improving our ability to map complex traits (i.e., those due to multiple genes and that are influenced by environment, genotype-by-environment, and genotype-by-genotype effects).

One of the leading explanations for missing heritability in human GWA studies is that many common diseases (e.g., cancer, diabetes, and heart disease) likely stem from the combined action of a large number of rare variants with individually small impacts on disease susceptibility. For example, despite the hundreds of GWA studies that have been performed to date, large-effect variants (e.g., APOE4 in Alzheimer's disease and CFH in age-related macular degeneration) remain the exception rather than the rule [3]. Using currently available mapping methods, small effect loci will be extremely difficult to detect without massive sample sizes.

Another potential explanation for missing heritability is that many genetic variants could be largely dependent on the environmental and genetic contexts in which they occur. For example, variation at the monoamine oxidase A (MAOA) gene is associated with violent behavior in humans, but only if the individual was abused as a child [8]. Also, gene-gene interactions (epistasis) are well-documented in controlled laboratory crosses in model organisms such as fruit flies and mice. While identification of epistasis remains elusive in humans [9,10], likely due to limited statistical power to detect gene-gene interactions in GWA studies of genetically diverse human populations, it is suspected to be widespread [3,11,12]. Disease susceptibility variants can also depend on sex [13] or on the parent from which the allele was inherited [14]. In short, the impact of a given genetic variant on a disease trait is often highly context-dependent. Such variants may be very difficult to detect when traits are measured in multiple genetic backgrounds and/or multiple environments, as is often the case in GWA studies. Ignoring such information during analyses may reduce the power to identify associated genetic loci when multiple factors (genetic or otherwise) influence a quantitative trait. Developing methods that can adapt to the different contexts of GWA study data sets may increase the power to detect associated loci using quantitative trait mapping.

Many of the current limitations of association mapping methods ultimately stem from limitations on the power to detect and localize causal variants (either because they have small effect sizes, are context-dependent, or both). While increasing sample size is one way to approach this problem, another strategy is to develop more powerful statistical methods that can take greater advantage of the information contained within the data. In particular, in contrast to most commonly used methods for association mapping, tree-based methods use the

***Corresponding author:** Richard Charnigo, Department of Biostatistics, University of Kentucky, Room 203, Multidisciplinary Science Building, Lexington, Kentucky 40536, USA, Tel: 859-218-2072; E-mail: RJCharn2@aol.com

Received November 01, 2013; **Accepted** November 02, 2013; **Published** November 05, 2013

Citation: Thompson KL, Charnigo R, Linnen CR (2013) Using Ancestral Information to Inform Analyses of Complex Data Sets. J Biomet Biostat 4: e126. doi:10.4172/2155-6180.1000e126

evolutionary relatedness of sampled individuals to gain information during analysis. Thus, there is potential for these methods to show increased power in detecting small effects and/or context-dependent loci. However, most currently available tree-based methods [15-19] are computationally inefficient and/or cannot take into account external covariates that may influence variation in quantitative traits. Extending tree-based methods to the wider variety of contexts provided by GWA study data sets may improve the power in association mapping compared to existing non-tree based approaches. Before describing tree-based methods in more detail, we discuss the rationale for genetic mapping and the advantages and limitations of existing association mapping methods.

Background

The conceptual basis of quantitative trait mapping (QTM), in which statistical correlations are sought between quantitative traits and polymorphic DNA variants ("markers"), stems back to the early part of the 20th century [20]. However, it has only been relatively recently that widespread availability of variable markers (e.g., single nucleotide polymorphisms (SNPs), insertion/deletion mutations (indels), or simple sequence repeats (microsatellites)) has made QTM feasible in humans [3]. In some cases, QTM is performed when the relationships among sampled individuals are known (linkage mapping), while in other cases, the relationships among individuals are unknown (association mapping). The present editorial focuses on the techniques developed for association mapping, although the reader is referred to the existing literature for information about the analysis of data with known familial relationships among sampled individuals [21-24]. Thus far, methodology proposed for association mapping either uses information in the evolutionary history present among genes and excludes other available information (including but not limited to experimental information and external covariate information) or uses external covariates to inform the analysis through the direct application of classical techniques that ignore the evolutionary relationships present within a particular SNP.

Classical statistical techniques applied in association mapping include the t-test, Analyses of Variance (ANOVA), and generalized linear model approaches that can be applied either marginally at each SNP or jointly on small neighboring sets of SNPs. Using generalized linear models allows straightforward adjustment for covariates during analyses [24-28]. More generally, classical statistical approaches are simple and readily-available, so that they are computationally efficient to implement on large GWA study data sets, making them popular approaches to association mapping. However, the precise localization of associated SNPs is not readily addressed by current techniques that are flexible enough to allow for covariates [25,29]. Additionally, these approaches assume independence among sampled individuals at each SNP, while the evolutionary relationships among these sampled individuals could be a potential source of covariation. By failing to consider shared evolutionary history among sampled individuals, classical statistical techniques could lose power to identify causal locations compared to methods that utilize this information.

Information about the evolutionary history for sampled observations can be represented by a bifurcating phylogenetic tree, as in Figure 1. The tips of the tree represent the sampled individuals at the present time, and the leftmost point on the tree represents the most recent common ancestor of the genetic variant under study. The lengths of the branches represent time, so that, if two observations share a branch, they share that part of their evolutionary history. Observations evolve independently after a split in their evolutionary history (represented by a split in a branch on the phylogenetic tree when viewed from left to right). If two observations (such as those denoted by red squares in Figure 1 share a large part of their evolutionary history, they are expected to have greater similarity in their trait(s) than two observations that share only a small portion of their evolutionary history (such as those denoted by blue circles). In fact, the technique in Thompson and Kubatko [17] suggests that, at a causal SNP, the covariance between two sampled observations could be approximated by the length of shared evolutionary history for that particular SNP. Phylogenetic methods provide an avenue to use the evolutionary relatedness among sampled individuals in the analysis of GWA study data, which may also be beneficial to association mapping [30].

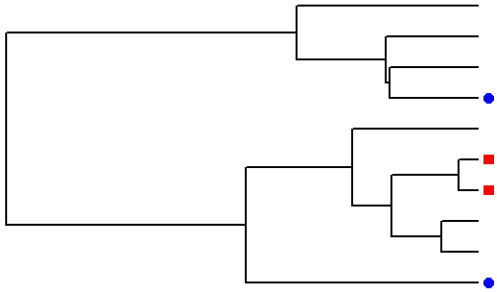

Figure 1: Example of a phylogeny at a particular SNP. In the phylogenetic tree, time moves from past (left) to present (right) across the tree, and the tips of the tree represent observations from the present time. The amount of shared evolutionary history among the two observations with red squares is large, so that a large covariance is expected among their trait values. In contrast, the two observations denoted by blue circles share a smaller portion of their evolutionary history, so that little covariance in their trait values is expected from shared evolutionary history.

Tree-based methods use estimated phylogenetic trees to gain information about the evolutionary history of a set of randomly sampled outbred individuals, and these methods show increased power compared to classical statistical techniques that ignore this information. Previous tree-based methods include those in Zöllner and Pritchard [19], which are not computationally feasible for large data sets, and Besenbacher et al., Pan et al., Zhang et al. [15,16,18], which consider all possible groups of observations compatible with the estimated phylogenies during association analysis. Because these methods use estimated phylogenies for each sampled SNP, they are especially computationally intensive. The method in Thompson and Kubatko [17] limits the required number of computations at the expense of considering only groups of observations defined by the earliest evolutionary events (edges) along an estimated phylogeny. In addition, current tree-based methods for data from randomly sampled individuals are limited by their inability to incorporate covariate information or any other existing information during association mapping. By extending these methods and remaining cognizant of the computational difficulties often associated with them, phylogenetic tools may provide an avenue for researchers to use external information during association mapping and achieve superior power over classical statistical techniques.

Future Directions

While the immediate goal of QTM is to identify loci that are statistically associated with complex trait variation, the ultimate goal is to use this information to uncover the biological underpinnings of quantitative trait variation and human disease. To these ends, finding

genomic regions harboring causal variants is not enough—it is only through finding the causal mutations themselves that we can dissect the molecular mechanisms that connect changes at the DNA level to traits expressed at the organism level. Moreover, many longstanding evolutionary questions are best informed by the identification of mutations rather than genomic regions [31], such as: Does adaptation proceed via a few large mutational steps or many small ones? Do individual mutations tend to impact few or many traits? How often do populations adapting to similar conditions utilize the same mutational solutions? Unfortunately, few GWA studies have achieved gene-level resolution, and even fewer have achieved mutation-level resolution. Detection and localization are especially challenging when individual effects are small and/or context-dependent. By extracting more information from the data, tree-based methods have the potential to significantly improve our ability to find causal mutations. In particular, the performance of association mapping methods may be improved by estimating covariance structures using ancestral information within genes, which can be done using phylogenetic techniques. If successful, these methods may help recover some of the "missing heritability" that has plagued GWA studies of complex diseases to date.

However, two significant challenges in the development of tree-based association methods remain. First, existing methods are too computationally intensive to be of practical use for large GWA studies. The methods in Besenbacher et al. [15] and Thompson and Kubatko [17] propose the use of broad-scale evolutionary relationships to address this limitation. Second, while context-dependence is pervasive in quantitative traits, current tree-based methods are not flexible enough to take into account environmental or gender-specific covariates. Importantly, context-dependent effects are more than just nuisance parameters in association mapping—gene-environment and gene-gene interactions may provide essential clues to the molecular pathways underlying complex traits. Thus, methods that show an improved ability to detect and quantify epistatic effects and genotype-by-environment interactions would represent a significant advance in GWA methodology. Together, these improvements have the potential to yield novel insights into the genetics of complex diseases that may better inform disease prediction and treatment strategies.

Acknowledgments

This material is based, in part, upon work supported by the National Science Foundation under Grant No. DEB-1257739. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-791.
2. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445-455.
3. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881-888.
4. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Nature 456: 728-731.
5. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.
6. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450.
7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
8. Caspi A, McClay J, Moffitt TE, Mill J, Martin J, et al. (2002) Role of genotype in the cycle of violence in maltreated children. Science 297: 851-854.
9. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, et al. (2001) Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science 293: 2256-2259.
10. Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, et al. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434: 857-863.
11. Flint J, Mackay TF (2009) Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res 19: 723-733.
12. Mackay TF, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10: 565-577.
13. Ober C, Loisel DA, Gilad Y (2008) Sex-specific genetic architecture of human disease. Nat Rev Genet 9: 911-922.
14. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. Nature 462: 868-874.
15. Besenbacher S, Mailund T, Schierup MH (2009) Local phylogeny mapping of quantitative traits: higher accuracy and better ranking than single-marker association in genomewide scans. Genetics 181: 747-753.
16. Pan F, McMillan L, Pardo-Manuel De Villena F, Threadgill D, Wang W (2009) TreeQA: quantitative genome wide association mapping using local perfect phylogeny trees. Pac Symp Biocomput.
17. Thompson KL, Kubatko LS (2013) Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. BMC Bioinformatics 14: 200.
18. Zhang Z, Zhang X, Wang W (2012) HTreeQA: Using Semi-Perfect Phylogeny Trees in Quantitative Trait Loci Study on Genotype Data. G3 (Bethesda) 2: 175-189.
19. Zöllner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169: 1071-1092.
20. Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology 14: 43-59.
21. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.
22. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709-1723.
23. Lynch M, Walsh B (1998) Genetics and Analysis of Quantitative Traits, chapter 26. Sinauer Associates, Inc.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
25. González JR, Armengol L, Solé X, Guinó E, Mercader JM, et al. (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23: 644-645.
26. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425-434.
27. Sinnwell JP, Schaid DJ (2009) haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R package version 1.4.4.
28. Solé X, Guinó E, Valls J, Iniesta R, Moreno V (2006) SNPStats: a web tool for the analysis of association studies. Bioinformatics 22: 1928-1929.
29. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5: 1780-1815.
30. Roses AD (2012) Post-GWAS: Phylogenetic analysis in the hunt for complex disease-associated loci. Journal of Pharmacogenomics and Pharmacoproteomics 3: 3.
31. Linnen CR, Poh YP, Peterson BK, Barrett RD, Larson JG, et al. (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312-1316.