Statistical Approaches to Combine Genetic Association Data

Sharon M Lutz1,*, Tasha Fingerlin1, and David W Fardo2
1Department of Biostatistics, University of Colorado, 13001 E. 17th St, Aurora CO 80045, USA
2Department of Biostatistics, University of Kentucky, 725 Rose St, Lexington KY 40536, USA

Abstract

In an attempt to discover and unravel genetic predisposition to complex traits, new statistical methods have emerged that utilize multiple sources of data. This appeal to data aggregation is seen on various levels: across genetic variants, across genomic/biological/environmental measures and across different studies, often with fundamentally differing designs. While combining data can increase power to detect genetic variants associated with disease phenotypes, care must be taken in the design, analysis, and interpretation of such studies. Here, we explore methodologies employed to combine sources of genetic data and discuss the prospects for novel advances in the fields of statistical genetics and genetic epidemiology.

Background

Common genetic variants identified through genome-wide association studies often explain only a small proportion of a disease trait’s variability [1]. This so-called problem of missing heritability has been the impetus for exploring additional ways to test for genetic associations with complex disease [2]. For example, currently identified single nucleotide polymorphisms (SNPs) that are well-replicated and highly significant (often p<5e-8) for the association with height only explain about 5% of the phenotypic variance [3]. Recently, Yang et al. showed that simultaneously considering 294,831 SNPs genotyped on 3,925 unrelated individuals explained 45% of the phenotypic variance for height [4]. These SNPs account for substantially less than the estimated heritability of ~80%. While the remaining heritability may be captured in part by genomic features not captured in the sequence (e.g. heritable epigenetic features [5]), it is also likely that some of the missing heritability has not previously been detected because the individual effects of SNPs are small and are not well identified in single-SNP association models. As such, traditional and new strategies for combining various sources of data are being widely used in genetic studies of complex traits. Below we describe some of the most prominent approaches and discuss potential problems in implementation or interpretation.

Aggregating Over Genetic Variants

As opposed to the conventional single-SNP association testing, resolution can be gained by incorporating multiple variants simultaneously. The variants included may be from the same genomic region, often defined by a gene, or from multiple genomic locations, depending on the hypotheses of interest. In addition, the number of variants included can range from two
to hundreds of thousands or millions. In each case described below, the strategy is to exploit the relationships among genetic variation and the trait of interest to increase power, either by model-based data reduction or prudently combining the raw variant information.

**Simultaneously Testing Multiple SNPs**

In order to increase power by better modeling the true underlying biological process, one option is to test for an association with the phenotype of interest and multiple variants along a chromosome at the same time. Haplotype analysis captures the sequence of the underlying chromosome and has been widely used, especially in candidate gene association studies, but usually requires estimation of haplotype phase (the assignment of alleles to the maternal and paternal chromosomes) [6]. Other methods use traditional modeling techniques to examine multiple variants at a time, for instance in a logistic regression model using either frequentist or Bayesian approaches. Because there are often many variants of interest, strategies for reducing the number of SNPs to be included in a given model are attractive. One such approach is Principal Component Analysis (PCA), which uses an orthogonal transformation to produce a subset of linearly uncorrelated variables [7]. These variables can then be used in a multivariable regression model rather than each of the individual SNPs, reducing the number of statistical tests. A drawback to this approach is the lack of standard approaches to make inference on the relative importance of individual variants should an association with a particular PC be observed. Another approach to the problem of many variants compared to the total number of individuals included in a study is the use of penalized regression, which penalizes the size of the regression coefficients, and is most helpful and appropriate when multicollinearity is an issue (i.e., the covariates are highly correlated such as SNPs in LD) [8].

**Rare Variant Aggregation**

Due to the low frequency of many variants, single variant-based analysis and many existing tests developed for common variants are not valid or are extremely underpowered for most studies. Burden tests assess the cumulative effects of multiple variants in a genomic region by collapsing, or otherwise summarizing, the rare variants within a region [9]. A disadvantage of many burden tests is the loss of power if all the included rare variants do not influence the phenotype in the same direction (e.g. all lower blood cholesterol) and with the same magnitude of effect. In addition, most variants have little or no effect and collapsing all variants together can introduce noise. Methods based on variance components such as the Sequence Kernel Association Test (SKAT) [10] specifically address both of these issues, but a recent comparison of rare variant association methods found that no single method gave consistently acceptable power across the range of scenarios examined [9].

**Gene-Gene Interactions**

Another approach to jointly considering multiple variants is to explicitly allow for epistasis, the interaction of effects from one or more variants. The most common approach is to use traditional regression-based models [11]. Since the space of possible epistasis models can be extraordinarily large even for a relatively small number of variants, methods have been developed to reduce the number of interactions considered. A few of these machine learning methods include tree-based methods like random forests, focused interaction testing framework (FITF), model-based multifactor dimensionality reduction (MB-MDR), and Bayesian epistasis association mapping (BEAM) [11]. Each of these strategies requires careful consideration to avoid over-interpretation, but with an emphasis on replication, can be widely applicable.
Aggregating Across Genomic/Biological/Environmental Features

While the above methods focus on combining genetic sequence variation information, there is growing interest in the incorporation of measures of other genomic, transcriptomic, proteomic, and metabolomic variation as well as relevant environmental factors in order to gain a better understanding of the disease of interest. A common and easily applied approach is to examine the association between sequence variation and gene expression, an approach that has been labeled “genetical genomics” [12]. Pathway analysis tries to gain insight into the underlying biology of a trait by using differentially expressed genes and proteins to identify common biological pathways and their inter-relationships [13]. The broader field of exploration in the area of combining data across experiments and organisms is often referred to as systems biology [14]. Methods to determine gene-environment interactions investigate how environmental factors modify the relationship between genetic variant and disease risk [15]. While these approaches have great promise for better aggregating and summarizing the volumes of data that are now available, important challenges remain for making inference about the relationships between the identified patterns and pathways and traits.

Aggregating Across Studies

Either raw data or summary results from multiple genetic association studies can be combined in a mega- or meta-analytic framework, respectively. When study designs differ fundamentally, e.g. considering family-based and case-control studies, association metrics are often not commensurate, thus requiring more care than as in standard approaches [16,17]. All approaches described above can be used within the study and then aggregated across, but meta-analyses of genetic association studies may have biased results if one does not properly adjust for population stratification [18]. Won et al. [19] propose a method for the meta-analysis of genome-wide association studies that combines family-based and unrelated samples and properly adjusts for population stratification. Even using this approach, still there may be issues with interpretation and heterogeneity that arise from utilizing different trait measurements and study designs, different ethnic groups, different environmental exposures, and different genotyping chips [20].

Discussion

The last two decades have seen a dramatic increase in the development and use of statistical methodology for genetic association studies. Because the focus of most studies has been on common, complex diseases, we expect most individual genetic variants to have relatively little marginal effect due to low frequency and/or modest increase in risk. The realities of limited resources and ever-increasing collaboration have resulted in an emphasis on development of methods that combine data within and across studies. Many of the other approaches described in this paper require similar rigorous attention to assumptions regarding sample ascertainment, definition of phenotypes, and the potential for genetic heterogeneity that can lead to both false positive and false negative associations. One of the most important areas with a need for continued methods development is in the area of inference related to complex systems and their relationship to traits of interest.

Acknowledgments

This work was supported in part by grants from the NHLBI [Grants R01HL089897 and R01HL089856 (SML) and 1R21HL113543 (TEF)], the NIA [5P30AG028383-07 (DWF)] and the NIGMS [5P30AG028383-07 (DWF)].
References

1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]

2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

3. Visscher PM. Sizing up human height variation. Nat Genet. 2008; 40:489–490. [PubMed: 18443579]

4. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]

5. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010; 465:721–727. [PubMed: 20535201]

6. Zhao H, Pfeiffer R, Gail MH. Haplotype analysis in population genetics and association studies. Pharmacogenomics. 2003; 4:171–178. [PubMed: 12605551]

7. Horne BD, Camp NJ. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genet Epidemiol. 2004; 26:11–21. [PubMed: 14691953]

8. Ayers KL, Cordell HJ. SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. Genet Epidemiol. 2010; 34:879–891. [PubMed: 21104890]

9. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. Plos Genet. 2012; 8:e1002496. [PubMed: 22319458]

10. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]

11. Van Steen K. Travelling the world of gene–gene interactions. Brief Bioinform. 2012; 13:1–19. [PubMed: 21441561]

12. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet. 2005; 14:R163–R169. [PubMed: 16244315]

13. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012; 8:e1002375. [PubMed: 22383865]

14. Kitano H. Systems Biology: A Brief Overview. Science. 2002; 295:1662–1664. [PubMed: 11872829]

15. Ashard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, et al. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet. 2012; 131:1591–1613. [PubMed: 22760307]

16. Mirea L, Infante-Rivard C, Sun L, Bull SB. Strategies for Genetic Association Analyses Combining Unrelated Case-Control Individuals and Family Trios. Am J Epidemiol. 2012; 176:70–79. [PubMed: 22573432]

17. Fardo DW, Druen AR, Liu J, Mirea L, Infante-Rivard C, et al. Exploration and comparison of methods for combining population- and family-based genetic association using the Genetic Analysis Workshop 17 mini-exome. BMC Proceedings. 2011; 5:S28. [PubMed: 22373349]

18. Kavvoura FK, Ioannidis JP. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Hum Genet. 2008; 123:1–14. [PubMed: 18026754]

19. Won S, Lu Q, Bertram L, Tanzi RE, Lange C. On the meta-analysis of genome-wide association studies: a robust and efficient approach to combine population and family-based studies. Hum Hered. 2011; 73:35–46. [PubMed: 22261799]

20. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 2010; 34:60–66. [PubMed: 19847795]