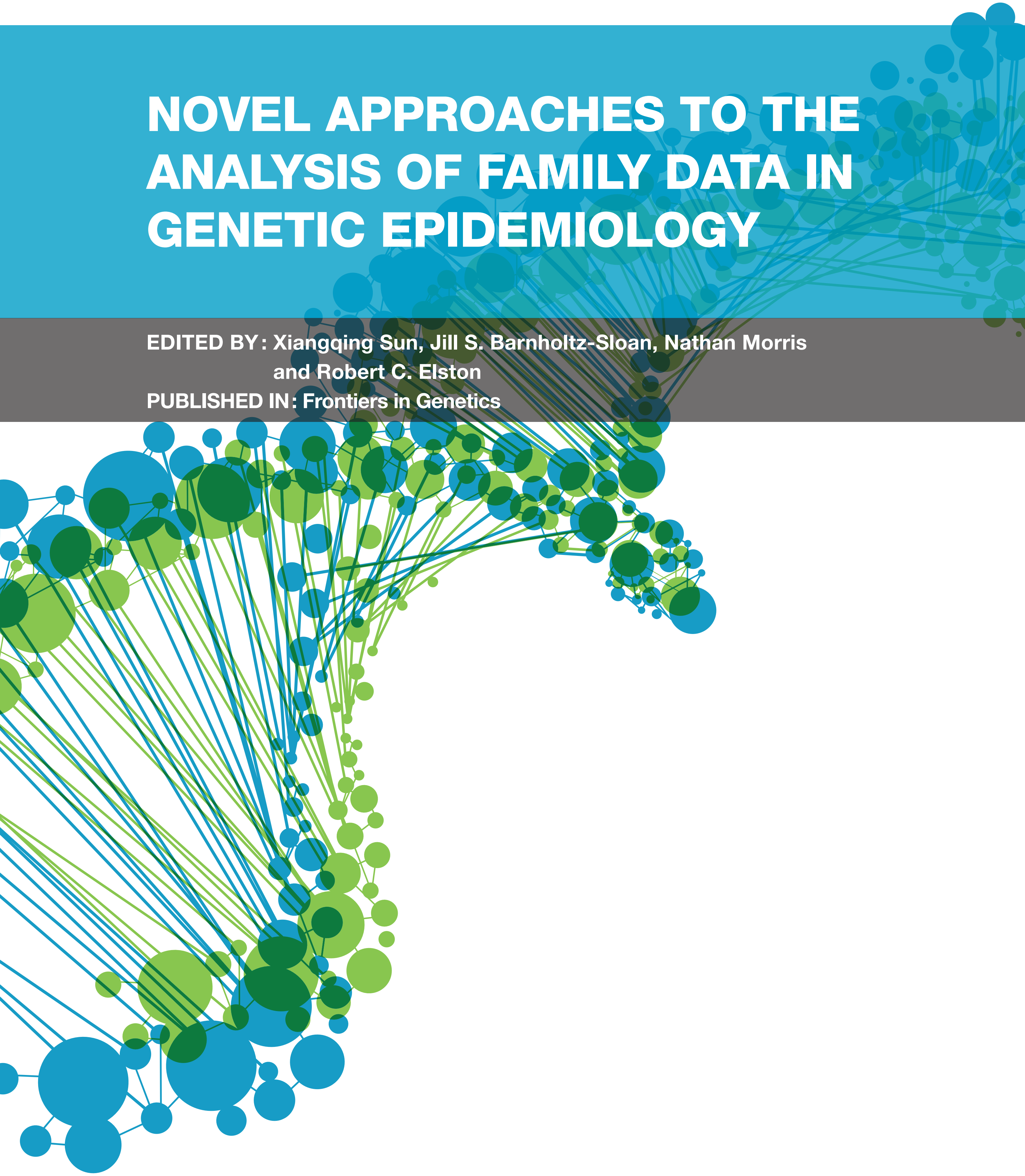
NOVEL APPROACHES TO THE ANALYSIS OF FAMILY DATA IN GENETIC EPIDEMIOLOGY

EDITED BY: Xiangqing Sun, Jill S. Barnholtz-Sloan, Nathan Morris and Robert C. Elston

PUBLISHED IN: Frontiers in Genetics



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-932-7
DOI 10.3389/978-2-88919-932-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

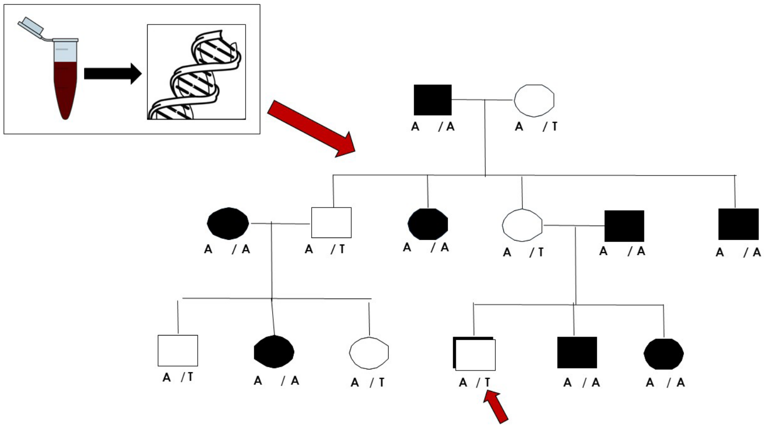
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

NOVEL APPROACHES TO THE ANALYSIS OF FAMILY DATA IN GENETIC EPIDEMIOLOGY

Topic Editors:
Xiangqing Sun, Case Western Reserve University School of Medicine, USA
Jill S. Barnholtz-Sloan, Case Western Reserve University School of Medicine, USA
Nathan Morris, Case Western Reserve University School of Medicine, USA
Robert C. Elston, Case Western Reserve University School of Medicine, USA



Example pedigree for linkage analysis where the A/A genotype confers risk for affection status of a complex disease.

Figure by Jill S. Barnholtz-Sloan

Genome-wide association studies (GWAS) for complex disorders with large case-control populations have been performed on hundreds of traits in more than 1200 published studies (http://www.genome.gov/gwastudies/) but the variants detected by GWAS account for little of the heritability of these traits, leading to an increasing interest in using family based designs. While GWAS studies are designed to find common variants with low to moderate attributable risks, family based studies are expected to find rare variants with high attributable risk. Because family-based designs can better control both genetic and environmental background, this study design is robust to heterogeneity and population stratification. Moreover, in family-based analysis, the background genetic variation can be modeled to control the residual variance which could increase the power to identify disease associated rare variants. Analysis of families can also help us gain knowledge about disease transmission and inheritance patterns.

Although a family-based design has the advantage of being robust to false positives, novel and powerful methods to analyze families in genetic epidemiology continue to be needed, especially for the interaction between genetic and environmental factors associated with disease. Moreover, with the rapid development of sequencing technology, advances in approaches to the design and analysis of sequencing data in families are also greatly needed.

The 11 articles in this book all introduce new methodology and, by using family data, substantial new findings are presented in the areas of infectious diseases, diabetes, eye traits, autism spectrum.

Citation: Sun, X., Barnholtz-Sloan, J. S., Morris, N., Elston, R. C., eds. (2016). Novel Approaches to the Analysis of Family Data in Genetic Epidemiology. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88919-932-7

Table of Contents

frontiers in
GENETICS

EDITORIAL
published: 06 February 2015
doi: 10.3389/fgene.2015.00027

Novel approaches to the analysis of family data in genetic epidemiology

Nathan Morris*[1,2], *Robert C. Elston*[1,3], *Jill S. Barnholtz-Sloan*[1,3] and *Xiangqing Sun*[1]

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University, OH, USA
[2] Center for Clinical Investigation, Case Western Reserve University, OH, USA
[3] Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, OH, USA
**Correspondence: x.sun@case.edu*

Edited and reviewed by:
Anthony Gean Comuzzie, Texas Biomedical Research Institute, USA

Keywords: genome-wide association, family studies, study designs, genetic factors, environmental factors

THE IMPORTANCE OF FAMILY DATA

The study of Genetic Epidemiology has historically focused on the inheritance of genetic factors and phenotypes within families. In fact, much of genetics involves the study of patterns of familial resemblance and identifying the factors that explain the observed patterns. However, in recent years the most common study design for investigating the genetic determinants of diseases has become that of genome wide association studies (GWAS) utilizing samples of *unrelated* individuals. The popularity of this approach has been driven primarily by a flood of ever improving technologies. Unfortunately, while GWAS using unrelated individuals have revealed a great many interesting disease associated variants, these variants are typically of small effect and cannot explain the observed patterns of heritability for many traits. In contrast there are numerous examples of highly penetrant rare segregating alleles that have been discovered using family based approaches. Furthermore, family based approaches have other advantages: the ability to overcome confounding factors such as population stratification, and the numerous studies that have collected large amounts of family data and which should continue to be leveraged. Unfortunately, family based approaches to genetics have an added layer of complexity at all stages from design to analysis.

This editorial introduces the Frontiers in Genetics Research Topic and Ebook: "Novel approaches to the analysis of family data in genetic epidemiology." The papers in this issue reveal that, even with easy access to high-throughput genotyping tools such as SNP arrays and next generation sequencing, family based study designs still play an important role in untangling the complex web of environmental and genetic factors that lead to disease.

FAMILY BASED STUDY DESIGNS

A number of articles in this issue shed light on unique study designs and approaches to analyzing family data. Stein et al. (2013) describe a household contact study design which involves collecting data on households that may include both related and unrelated individuals. They argue that this research study design may be a powerful approach for jointly studying genetic and environmental exposures. Similarly, Estus et al. (2013) describe an approach to combining family based and population based data by utilizing a combined association test. Wang et al. (2013) describe an approach of using only the independent probands from a family based study of autism to investigate genetic factors that account for IQ differences in autism patients. Nelson et al. (2013) describe a unique population based registry in Utah that contains pedigree information for all residents of the state and dates back many decades. Using this information they show that certain subsets of prostate cancer, such as early onset, high BMI, and lethal prostate cancer, cluster in families more strongly than other forms of prostate cancer. They further suggest that future studies should focus on families that display a clear clustering of a more carefully defined cancer phenotype to reduce the signal to noise ratio. Uemoto et al. (2013) discuss the power of regional heritability mapping with a mixed model approach applicable to both related and unrelated persons. This approach leverages the fact that even distantly related individuals share small regions of the genome that are inherited from a common ancestor.

ANALYSIS OF FAMILY DATA

The analysis of family data is generally more complex than the analysis of unrelated samples, and, thus, specialized statistical methods and software are often needed. Huang et al. (2013) propose a novel method of linkage analysis using sequence data on large pedigrees. This method, which uniquely combines MCMC based approximations with non-stochastic approaches, can be used to map disease genes using linkage and/or association evidence. Song and Elston (2013a) investigate the distributional properties of a commonly used linkage analysis statistic. These authors also describe a new web based software package which, among other things, plots pedigrees, calculates genetic similarity coefficients and performs visualization of the relatedness among family members (Song and Elston, 2013b). Similarly, Lutz et al. (2013) describe a method of using data from family based studies to test for a direct genetic effect, an extension of a method previously used for analysis of unrelated individuals. Additionally, Lutz et al. (2014) describe an approach to look at secondary phenotypes in case-control genetic association studies that circumvents the computational issues of a former approach.

CONCLUSION

Although GWAS with unrelated samples have become one of the most common study designs currently used in human genetics, utilizing a family based design has many advantages. If a variant can be observed to co-segregate with a phenotype within a family, the evidence for its association with the disease is greatly strengthened. Family data provide excellent opportunities to find highly penetrant rare variants, and thus discover important biology informing us about disease. The articles in this issue illustrate how family based genetic designs remain a foundational part of human genetics.

REFERENCES

Estus, J. L., Family Investigation of Nephropathy and Diabetes Research Group, and Fardo, D. W. (2013). Combining genetic association study designs: a GWAS case study. *Front. Genet.* 4:186. doi: 10.3389/fgene.2013.00186

Huang, Y., Thomas, A., and Vieland, V. J. (2013). Employing MCMC under the PPL framework to analyze sequence data in large pedigrees. *Front. Genet.* 4:59. doi: 10.3389/fgene.2013.00059

Lutz, S. M., Hokanson, J. E., and Lange, C. (2014). An alternative hypothesis testing strategy for secondary phenotype data in case-control genetic association studies. *Front. Genet.* 5:188. doi: 10.3389/fgene.2014.00188

Lutz, S. M., Vansteelandt, S., and Lange, C. (2013). Testing for direct genetic effects using a screening step in family-based association studies. *Front. Genet.* 4:243. doi: 10.3389/fgene.2013.00243

Nelson, Q., Agarwal, N., Stephenson, R., and Cannon-Albright, L. A. (2013). A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer. *Front. Genet.* 4:152. doi: 10.3389/fgene.2013.00152

Song, Y. E., and Elston, R. C. (2013a). The null distribution of likelihood-ratio statistics in the conditional-logistic linkage model. *Front. Genet.* 4:244. doi: 10.3389/fgene.2013.00244

Song, Y. E., and Elston, R. C. (2013b). PedWiz: a web-based tool for pedigree informatics. *Front. Genet.* 4:189. doi: 10.3389/fgene.2013.00189

Stein, C. M., Hall, N. B., Malone, L. L., and Mupere, E. (2013). The household contact study design for genetic epidemiological studies of infectious diseases. *Front. Genet.* 4:61. doi: 10.3389/fgene.2013.00061

Uemoto, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Wilson, J. F., et al. (2013). The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. *Front. Genet.* 4:232. doi: 10.3389/fgene.2013.00232

Wang, H. Z., Qin, H. D., Guo, W., Samuels, J., and Shugart, Y. Y. (2013). New insights into the genetic mechanism of IQ in autism spectrum disorders. *Front. Genet.* 4:195. doi: 10.3389/fgene.2013.00195

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 December 2014; accepted: 19 January 2015; published online: 06 February 2015.

Citation: Morris N, Elston RC, Barnholtz-Sloan JS and Sun X (2015) Novel approaches to the analysis of family data in genetic epidemiology. Front. Genet. ***6****:27. doi: 10.3389/fgene.2015.00027*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 19 April 2013
doi: 10.3389/fgene.2013.00059

Employing MCMC under the PPL framework to analyze sequence data in large pedigrees

Yungui Huang*[1], **Alun Thomas**[2] **and Veronica J. Vieland**[1,3]*

[1] *Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA*
[2] *Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA*
[3] *Departments of Pediatrics and Statistics, Ohio State University, Columbus, OH, USA*

Edited by:
Xiangqing Sun, Case Western Reserve University, USA

Reviewed by:
Jeffrey O'Connell, University of Maryland School of Medicine, USA
Jianzhong Ma, University of Texas MD Anderson Cancer Center, USA
Rob Igo, Case Western Reserve University, USA

****Correspondence:***
Yungui Huang, Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, WB5131, 575 Children's Crossroad, Columbus, OH 43215, USA.
e-mail: yungui.huang@nationwidechildrens.org

The increased feasibility of whole-genome (or whole-exome) sequencing has led to renewed interest in using family data to find disease mutations. For clinical phenotypes that lend themselves to study in large families, this approach can be particularly effective, because it may be possible to obtain strong evidence of a causal mutation segregating in a single pedigree even under conditions of extreme locus and/or allelic heterogeneity at the population level. In this paper, we extend our capacity to carry out positional mapping in large pedigrees, using a combination of linkage analysis and within-pedigree linkage trait-variant disequilibrium analysis to fine map down to the level of individual sequence variants. To do this, we develop a novel hybrid approach to the linkage portion, combining the non-stochastic approach to integration over the trait model implemented in the software package Kelvin, with Markov chain Monte Carlo-based approximation of the marker likelihood using blocked Gibbs sampling as implemented in the McSample program in the JPSGCS package. We illustrate both the positional mapping template, as well as the efficacy of the hybrid algorithm, in application to a single large pedigree with phenotypes simulated under a two-locus trait model.

Keywords: linkage analysis, linkage disequilibrium, MCMC, genome-wide association, PPL, PPLD, epistasis, whole-genome sequence

INTRODUCTION

The increased feasibility of whole-genome (or whole-exome) sequencing has led to renewed interest in using family data to find disease mutations. For clinical phenotypes that lend themselves to study in large families, this approach can be particularly effective, because it may be possible to obtain strong evidence of a causal mutation segregating in a single pedigree even under conditions of extreme locus and/or allelic heterogeneity at the population level.

The template for this type of "single large pedigree" design is straightforward. Linkage analysis can be used to narrow the region of interest to a relatively small locus. From there, linkage disequilibrium (LD, or association) analysis can be used for fine-mapping within the linked locus. This step can be based on all sequence variants within the region (whether measured directly in all individuals or partially imputed from selected individuals with sequence and single nucleotide polymorphism (SNP)-chip data in remaining family members). That is, rather than relying solely on bioinformatic filtering approaches to reduce the set of all observed sequence variants down to a manageable number, the set of candidate sequence variants is obtained by (i) restricting the region of interest based on co-segregation with the phenotype, and then within that region, further restricting the set of interesting variants to specific individual mutations co-segregating with the phenotype. Of course, in the presence of appreciable LD among mutations, further filtering and follow-up experiments may be needed to resolve which among a set of correlated mutations is the functional one.

One challenge to this approach is that linkage analysis of large pedigrees is itself not trivial. As is well-known, the Elston–Stewart (ES) algorithm (Elston and Stewart, 1971) can handle relatively large pedigrees, but only a small number of markers at a time. This was less of an issue in the era of microsatellite marker maps, but renders ES relatively ineffective when conducting multipoint analyses using SNPs, because relying on a small number of SNPs per calculation leaves substantial gaps in map informativeness. On the other hand, the Lander–Green (LG) algorithm (Lander and Green, 1987), which can make simultaneous use of large numbers of SNPs, is constrained to smaller pedigrees. Pedigrees with more than around 25 individuals can exceed the limits of the LG algorithm, but these are precisely the pedigrees that can show strong evidence on their own. Trimming or breaking up pedigrees to circumvent LG limitations can lead to substantial loss of information and potentially to misleading results. This is also true of the practice of selecting a small number of affected individuals to use for identity-by-state (IBS) sharing of rare sequence variants, rather than utilizing identity-by-descent (IBD) methods to track variants through the full pedigree structure.

One widely used approach to circumventing the computational complexity of large pedigree calculations is to use statistical methods that avoid calculation of the full pedigree likelihood, such as variance-components (as implemented, e.g., in Almasy and Blangero, 1998). Another familiar alternative is to use Markov chain Monte Carlo (MCMC). This supports the use of the full likelihood, but the difficulties of optimizing performance of samplers tends to limit flexibility in handling the trait model. In particular, we have developed a suite of linkage methods with a very flexible underlying framework for handling the trait model (Vieland et al., 2011) by integrating trait parameters out of the likelihood, one advantage of which is the ease with which new trait models or additional trait parameters can be added to the calculation. MCMC would require separate development and tuning of samplers for each variation of the model, and success in developing well-behaved samplers for all variations is far from guaranteed. For this reason, we have been reluctant to turn to MCMC in the past.

Here we take a novel hybrid approach, combining MCMC to handle the marker data, while retaining the non-stochastic approach to trait–model integration implemented in Kelvin (Vieland et al., 2011). Specifically, we use the graphical-model-based MCMC approach of (Thomas et al., 2000) for the marker data combined with the adaptive numerical integration algorithm described in detail in Seok et al. (2009) for the trait data. This allows us to exploit the power of MCMC in the context of the posterior probability of linkage (PPL) framework (Vieland et al., 2011). We illustrate the application of this new approach by applying it to a single large family.

MATERIALS AND METHODS

In this section, we (i) present background on Kelvin, the software package in which the PPL framework is implemented, and (ii) on McSample, which implements the underlying MCMC techniques used here. We restrict attention to background directly relevant to this paper (see Vieland et al., 2011 for details on the PPL framework and Thomas et al., 2000 for details on the MCMC methodology). We then (iii) describe the software engineering used to implement the new hybrid method, and (iv) describe the application of the new method to a single large pedigree.

KELVIN

The PPL framework, as implemented in the software package Kelvin (Vieland et al., 2011), can be used to calculate two primary statistics, both illustrated here: the PPL and the PPLD (posterior probability of linkage disequilibrium, or trait–marker association). The PPL framework is designed to accumulate evidence both for linkage and/or LD and also against linkage and/or LD. All statistics in the framework are on the probability scale, and they are interpreted essentially as the probability of a trait gene being linked (and/or associated) to the given location (or marker). The PPL assumes a prior probability of linkage of 2%, based on empirical calculations (Elston and Lange, 1975), while the PPLD assumes a prior probability of trait–marker LD of 0.04% based on reasoning in Huang and Vieland (2010). This is one caveat to interpretation of the statistics as simple probabilities, since values below the prior indicate evidence against linkage (or LD), while values above the prior indicate evidence in favor. Note too that the small prior probabilities constitute a form of "penalization" of the likelihood; moreover, as posterior probabilities rather than *p*-values, statistics in the PPL framework do not require correction for multiple testing (see, e.g., Edwards, 1992; Vieland and Hodge, 1998 for further discussion of this issue).

One distinguishing feature of this framework is how it handles the trait parameter space. An underlying likelihood in a vector of trait parameters is used. The base models are a dichotomous trait (DT) model parameterized in terms of a disease allele frequency, three genotypic penetrances, and the admixture parameter α of Smith (1963) to allow for intra-data set heterogeneity; and a quantitative trait (QT) model parameterized in terms of a disease allele frequency, three genotypic means and variances corresponding to normally distributed data at the genotypic level, and α. The QT model has been shown to be highly robust to non-normality at the population level and it is inherently ascertainment corrected, so that no transformations of QTs are necessary prior to analysis (Bartlett and Vieland, 2006). Models assuming χ^2 distributions at the genotypic level are also available to handle QTs with floor effects. The basic QT model can also be extended to cover left- or right-censoring, using a QT threshold (QTT) model (Bartlett and Vieland, 2006; Hou et al., 2012).

Whatever specific model is used, Kelvin handles the unknown parameters of the model by integrating over them for a kind of model-averaging. [Independent uniform priors are assumed for each (bounded) parameter, with an ordering constraint imposed on the penetrances (DT) or genotypic means (QT); see Vieland et al., 2011 for details.]. Kelvin also uses Bayesian sequential updating to accumulate evidence across data sets, integrating over the trait parameter space separately for each constituent data set. This is an explicit allowance for inter-data set heterogeneity with respect to trait parameters, and it also means that the number of parameters being integrated over does not go up with the number of data sets analyzed (see below). A related technique is Kelvin's use of liability classes (LCs): individuals are assigned to an LC, and the integration over the penetrances or means is done separately for each LC. This is an explicit allowance for dependence of the penetrances (or means) on a classification variable. While current computational restrictions preclude the use of more than three or four LCs at a time, one very important use of this model is incorporation of gene–gene interaction by classifying individual based on their status at a known gene or SNP; we illustrate this approach below.

Due to the nature of the underlying trait models, which are formulated based on genetic considerations without regard to computational convenience, analytic solutions to the resulting multi-dimensional integrals are not possible. Instead, Kelvin carries out the integration over the trait parameters using a modified version of DCUHRE (Berntsen et al., 1997; Seok et al., 2009), a sub-region adaptive or dynamic method, tailored to the specific features of our application. While non-stochastic in nature, the method tunes the amount of resampling of the parameter space to the shape of that space (peakedness) on a position-by-position basis for each data set, resulting in a highly efficient approach to obtaining accurate estimates of the integral. The algorithm is theoretically guaranteed to be accurate for up to 13–15 dimensions, a limit that we generally do not exceed (see above); and because the method is non-stochastic, we do not need to worry about burn-in, convergence or other issues that can complicate Monte Carlo-based approaches.

Kelvin source code is available for download at http://kelvin.mathmed.org/ and Kelvin documentation is accessible on the same site. Help with access, installation, and use can be requested by emailing kelvin@nationwidechildrens.org.

McSAMPLE

McSample is a program for sampling the inheritance states in a pedigree of relatives from the conditional distribution given the structure of the pedigree, observed genotypes and/or phenotypes for individuals in the pedigree, and a model for the founder haplotypes. It is written in Java and is part of the Java Programs for Statistical Genetics and Computational Statistics (JPSGCS) package available from Alun Thomas (http://balance.med.utah.edu/wiki/index.php/Download). The sampling is done using blocked Gibbs updates of two types: ones involving all the inheritance states associated with a locus, and ones involving inheritance states associated with sets of individuals as described by Thomas et al. (2000). Founder haplotype models can be derived under the assumption of linkage equilibrium from the allele frequencies in a sample. It is also possible to estimate models under LD using the FitGMLD program that is also available in JPSGCS, as described by Thomas (2010) and Abel and Thomas (2011). In the case that LD is allowed, only locus block Gibbs updates can be made which typically leads to poorer mixing of the MCMC sampler. The input to McSample must be provided in the format used by the LINKAGE programs (Ott, 1976) with extensions when there is LD. Missing data are allowed in the input. In McSample output, the inheritances are specified by labeling each founder allele uniquely and listing the alleles inherited by each person in the pedigree. There are no missing data in the output. A different output file is created for each iteration. These output files can then be used as input, e.g., to standard lod score calculating programs, with the results averaged over iterations. Note that a standard application would consist of averaging over MCMC-based marker likelihoods for a single, fixed trait model.

SOFTWARE ENGINEERING

The only difficulty in combining MCMC to handle the marker data with Kelvin's non-stochastic algorithm for the trait parameter space is one of order of operations. On the MCMC side, calculations are done on a per-pedigree basis for an entire chromosome at a time, and likelihoods are averaged across iterations. For the trait model, however, the adaptive algorithm works by averaging the likelihood ratio (LR, not likelihood; see Vieland et al., 2011 for details) across pedigrees, one calculating position at a time as we walk down each chromosome. Thus there are two iterative processes that need to be decoupled and properly tracked: first, repeated MCMC marker-sample generation for each pedigree across the chromosome; second, repeated (adaptive) trait-space sampling across pedigrees at each position on each chromosome, conditional upon the marker data obtained from the MCMC runs and the trait data. In order to minimize confusion in the exposition that follows, we use "iteration" to describe each individual marker configuration as generated by the MCMC routine in obtaining the marker likelihood, and "trait vector" to describe each individual vector of values for the trait parameters generated by Kelvin to calculate the trait likelihood conditional on the marker information.

To address the required bookkeeping issues while maintaining modular code with minimal changes to existing logic, we adapted Kelvin by simply inserting a set of McSample runs at the beginning of the calculation. At this step, multiple MCMC iterations are generated for each pedigree conditional on the marker data only. Each iteration creates a set of pedigree files with fully informative, phased marker genotypes for each pedigree and each chromosome. We create a single pedigree file incorporating all iterations for each pedigree, with the pedigree label modified to reflect both the pedigree and the iteration. To calculate the LR for a pedigree, we first calculate the LR for each iteration as if it represented a unique pedigree. For each trait vector we average these LRs across iterations for each pedigree at each calculation position along the chromosome, returning a set of LRs by pedigree by position for each trait vector. These LRs are multiplied across pedigrees to obtain the LR by position across pedigrees for each vector, and averaged over all trait vectors. The average LR per position is then evaluated, on the basis of which additional trait vectors may be added in an iterative process until the adaptive trait–model integration algorithm terminates.

The marker likelihood calculation itself is done using the ES algorithm, based on the two markers flanking each calculation position in turn. Because each individual MCMC iteration is fully phased and fully informative, using two markers is equivalent to using all markers with computational complexity no longer a function of the total number of markers. (Indeed a single marker could be used, but because of Kelvin's built-in algorithm for walking down each chromosome in multipoint analysis, three-point calculations were simpler to implement.) Trait calculations per position are also done based on the ES algorithm regardless of pedigree complexity (Wang et al., 2007). Thus the overall complexity of the MCMC-PPL analysis is linear to the product of the number of iterations, the number of pedigrees, the number of individuals and the number of trait vectors, the last of which differs across calculating positions.

In order to decouple the adaptive trait–model integration process from the likelihood calculations, we use the software engineering trick of employing a client–server architecture together with a database to facilitate the operations (see **Figure 1**). The client is the driver for the generation of trait vectors, deciding which trait vectors are needed for the likelihood evaluation at each position, as described in detail in Seok et al. (2009). The client requests likelihoods for the trait vectors from the server using the database as an intermediary. If requested trait vectors are not available in the database, the client adds the required entries to the database for each pedigree for the given calculation position. Once the likelihoods are available for all pedigrees, the client uses them to calculate integrals for the current set of trait vectors and to decide whether additional trait vectors are needed, in which case the process is repeated until the client determines that no additional sampling of the trait vector space is needed.

On the server side, once initiated the server searches the database for trait vector entries flagged as new. It performs the needed likelihood calculations, stores the results in the database, and marks the entry for that trait vector, pedigree, and position as complete/available. Here the server is not a physical node, but rather a likelihood-calculation process. Typically our analyses

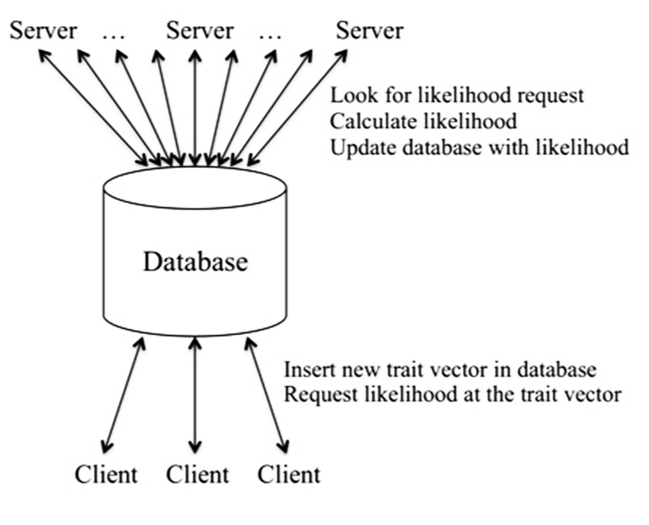

Server … Server … Server
Look for likelihood request
Calculate likelihood
Update database with likelihood
Database
Insert new trait vector in database
Request likelihood at the trait vector
Client Client Client

FIGURE 1 | Client–server architecture in Kelvin.

involve a small number of client processes and many likelihood servers. (Thus this is the reverse of the typical client–server model with a small number of servers and many clients. Nonetheless, our likelihood client plays the usual client role, by sending many requests to the likelihood servers.) The integration process is fast and efficient, requiring very little in terms of computing resources, and for this reason only a few client processes are required. By contrast, the likelihood calculations are highly computationally intensive. Thus the more servers, the faster the overall speed of the analysis. Here the database serves not only as a bookkeeping device, but also as the single server interface to a large pool of server processes.

The client–server architecture supports considerable flexibility in overall Kelvin functionality. It allows us to dynamically add and delete servers as needed. It also allows us to dedicate each server to one pedigree, with the amount of memory and number of cores tailored to the complexity of the pedigree, for efficient use of a distributed computing resource. The client is also by design indifferent as to how the underlying marker likelihood is calculated, i.e., the mechanism used to request and retrieve likelihoods is the same regardless of what approach was used to generate the likelihood. This allows us in principle to mix and match approaches to the marker data, e.g., using the LG algorithm for pedigrees small enough for LG to handle while simultaneously employing MCMC for larger pedigrees, all within the same data set.

APPLICATION TO SIMULATED DATA

To illustrate the use of this new hybrid MCMC–Kelvin approach, we selected a single large pedigree from an ongoing study of real human data. The pedigree has 48 individuals spanning four generations (see **Figure 2**); all but 10 individuals were genotyped. We used actual genotypes for 664,278 SNPs (after comprehensive cleaning) from the Illumina Human OmniExpress 12 V1.0. However, we simulated a new phenotype (for all but the 10 individuals missing genotypes) by selecting two SNPs (rs6851302@178.68cM on chromosome 4, which we call locus 1, and rs1145787@102.65cM on chromosome 6, which we call locus 2), with population frequencies (based on additional data not used here) matching our generating model as specified below; these SNPs were selected additionally for entering the pedigree through the top-most founders and segregating to the next generation at least four times to ensure they would be at least moderately informative in this pedigree.

Phenotypes for each individual were generated assuming an underlying two-locus (2L) disease model based on genotypes at this pair of SNPs. The generating model stipulated disease gene frequency of 1% (locus 1) and 20% (locus 2), and a fully penetrant dominant–dominant (DD) model. This model was selected from a set of 2L models considered in Vieland et al. (1992), which suggested that locus 1 would be moderately easy to map given sufficient meiotic information, while locus 2 might be very difficult to map; the model also represents a major gene effect with a modifier, something we might be interested in studying individual pedigrees. However, the purpose here is not to undertake a comprehensive study of power under different models, but simply to illustrate our approach in application to a single, albeit possibly atypical, pedigree.

Our overall approach to analyzing the pedigree is as follows:

1. We thinned the marker map following standard procedures to eliminate marker–marker LD, after filtering out markers with minor allele frequencies lower than 25%, and applied the new hybrid MCMC–Kelvin method to perform genome-wide linkage analysis. For purposes of this analysis the locus 1 and 2 SNPs were omitted from the marker set analyzed. We based the analysis on 2,000 MCMC iterations combined from 10 independent sampling processes (with different seeds), each with a 1,000-sample burn-in and 200 iterations/sampling run. (See below for rationale.) Linkage calculations were made every 2 cM under Kelvin's standard single-locus (SL) DT model.
2. We applied the PPLD to fine map under the (primary) linkage peak obtained in the first step, now utilizing all of the available SNPs (including those trimmed out during the first step and the locus 1 and 2 SNPs). While we did not have whole-genome sequence available for this pedigree, if such data were available, then this step would be applied to each variant in turn under the peak(s).
3. We repeated step 1, this time conditioning on genotypes at the most highly associated SNP from step 2, under a 2L model. Specifically, we assigned each individual to a LC based on the individual's SNP genotype. Kelvin then integrates over the trait parameters separately within LC as described above, which allows for dependence of penetrances on LC. We rescanned the genome under this model in order to look for possible modifier loci interacting with the gene under the primary linkage peak. We also carried out conditional 2L-PPLD analyses to see if we could fine map under a secondary linkage peak down to the level of the individual modifier SNP (or sequence variant, if we had sequence available).

In addition to these analyses, we also used the simulated pedigree to assess variability of the MCMC portion of the calculations. First, we repeated the entire MCMC process as described above five times, and examined variability of the results across these five runs. Second, we ran a single, much longer sampling process

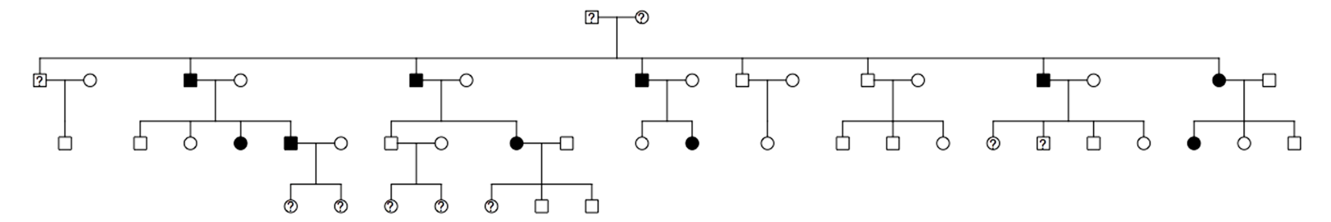


FIGURE 2 | Structure of the analyzed pedigree (filled, affected; empty, unaffected; ?, unknown phenotype and genotype).

(20,000 iterations) for which convergence was almost certainly achieved, then compared our results as described in step 1 with the final 5,000 iterations from the tail (post-convergence) end of this run. Finally, we considered variability across individual runs of 200 iterations with a 1,000-sample burn-in, that is, the length of runs that were averaged over in step 1 above.

RESULTS

In this section we (i) show results of the analysis of the single large pedigree. We then (ii) consider the accuracy of the MCMC component of the analysis.

DATA ANALYTIC RESULTS

Figure 3A shows the initial linkage scan. A peak on chromosome 4 clearly stands out above background noise, and we considered this to be our primary linkage finding. The PPL is elevated across a broad region of the chromosome (**Figure 3B**). However, the strongest evidence of linkage spans a relatively short region at approximately 175–181 cM.

For purposes of fine-mapping, we considered any positions on this chromosome with PPL $\geq$ 10%. The resulting (non-contiguous) region contained 9,433 SNPs from the full original marker set. Forty-nine percent of the analyzed SNPs within the linked regions gave evidence against LD (PPLD $<$ 0.0004), while only six SNPs (0.064%) showed PPLD $\geq$ 5% (**Table 1**). Two SNPs (rs6851302 and rs654089) clearly stand out from the rest, with PPLD = 0.43 in both cases. These two are in complete LD with one another ($R^2 = 1$) and in fact they share the same genotypes across this pedigree; rs6851302 and rs11934037 also show some LD (R^2=0.28). Note that even had we restricted fine-mapping to just the best supported region (175–181 cM), we would have successfully found this LD peak. Also for reference purposes, had we selected all 15,531 SNPs from all regions across the entire genome with PPL $\geq$ 10%, only one additional SNP would have given PPLD $\geq$ 5% (rs9916791, at 21.73 cM on chromosome 17, PPLD = 0.05).

Table 1 | Chromosome 4 SNPs with PPLD ≥ 5%.

Chromosome	SNP	cM	BP	PPLD
4	rs1800792	157.60	155753857	0.07
4	rs11100000	158.54	156542439	0.1
4	rs1460128	158.54	156544989	0.09
4	rs11934037	178.57	176255309	0.06
4	rs6851302	178.68	176328488	0.43
4	rs654089	178.71	176347501	0.43

We then conditioned on rs6851302 in order to rescan the genome for evidence of modifier loci. (Clearly choosing to use rs11934037 instead would yield identical results.) **Figure 4A** shows the 2L genome scan and **Figure 4B** shows the difference between the 2L and SL-PPLs across the genome (a measure of how much the data "prefer" the 2L model over the SL model). There are no

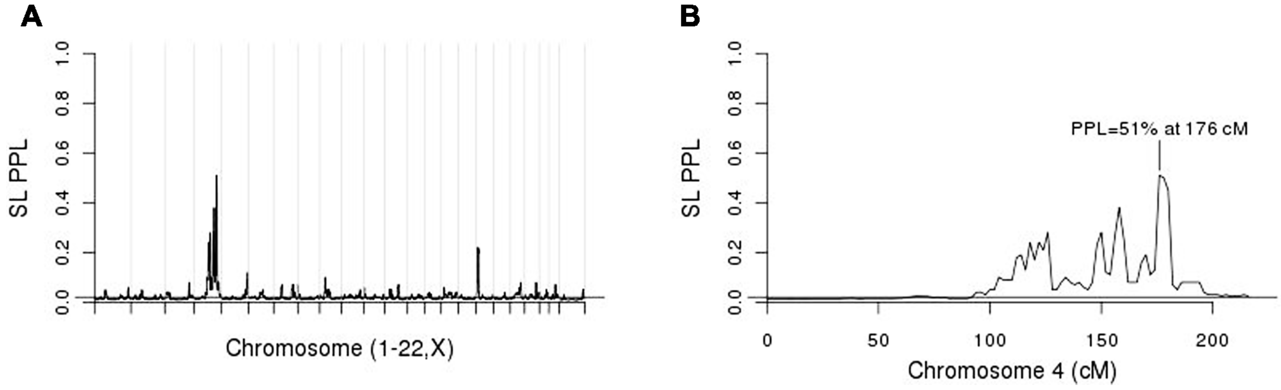


FIGURE 3 | (A) Single-locus (SL) genome scan; (B) chromosome 4 alone.

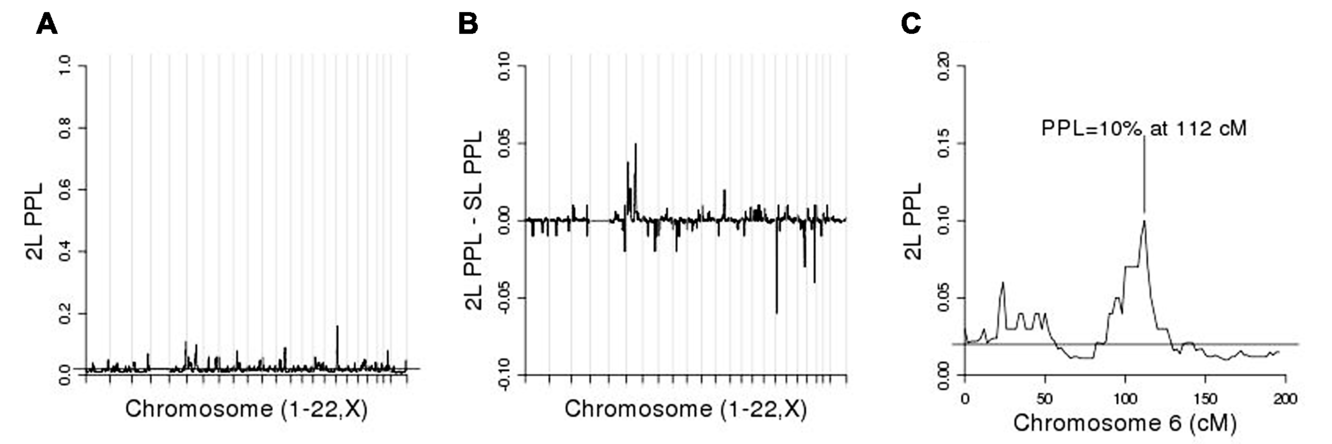


FIGURE 4 | (A) Two-locus (2L) genome scan; **(B)** 2L-PPL – SL-PPL across the genome; note that the scale of the *y*-axis is [−0.1, 0.1]. **(C)** Chromosome 6 alone.

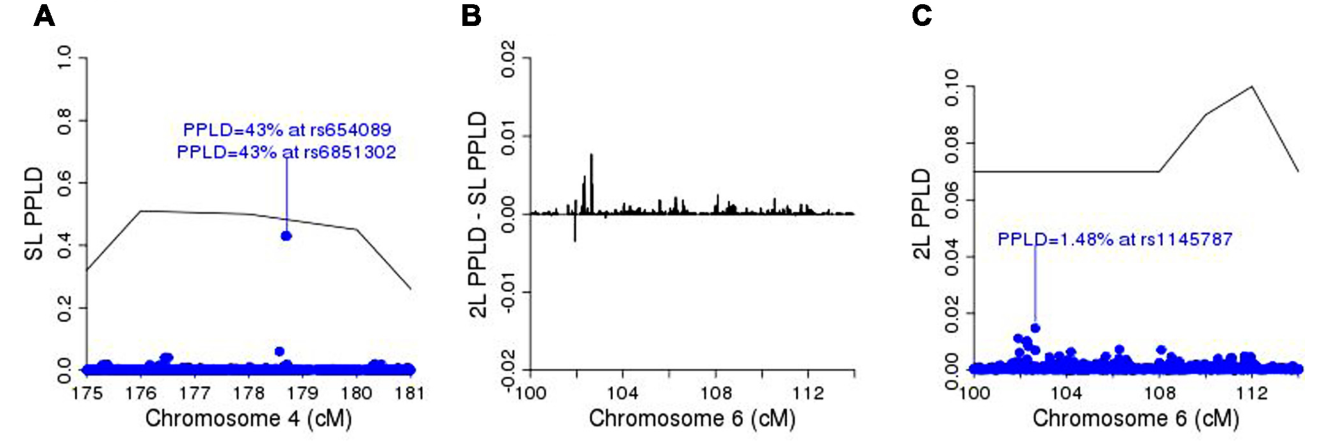


FIGURE 5 | (A) SL-PPLD under linkage peak region on chromosome 4 with solid line depicting PPL; **(B)** 2L-PPLD – SL-PPLD across linked region on chromosome 6; **(C)** 2L-PPLD under the 2L linkage peak on chromosome 6.

large 2L peaks (**Figure 4A**). However, using the difference between the 2L and SL-PPLs as an indication of how much the data "prefer" the 2L model over the SL model (**Figure 4B**), the largest positive difference occurs on chromosome 6 at 112 cM (SL-PPL = 5%; 2L-PPL = 10%). The doubling of the PPL under the epistasis model suggested a possible modifier gene location. We determined the width of the linkage peak by visual inspection as covering approximately 100–114 cM (see **Figure 4C**), and ran conditional 2L-PPLD analyses on all 3,120 SNPs in this region.

Figure 5A shows the SL-PPLD under the linkage region on chromosome 4, and **Figure 5B** shows the 2L-PPLD – SL-PPLD across the linkage region on chromosome 6; again a single region is elevated in the 2L analysis, with the highest positive change in the PPLD occurring at rs1145787 (SL-PPLD = 0.71%, 2L-PPLD = 1.48%; see **Figure 5C**). While these numbers are very small, they are still considerably higher than the prior probability of LD, and viewed in terms of 2L–SL differences, rs1145787 is clearly salient.

In summary, SL linkage analysis in this single pedigree enabled us to narrow the primary genomic region of interest to 6 cM on chromosome 4, while fine-mapping based on LD within this region detected the true causal variant (locus 1) within this region along with one other variant in complete LD with the causal one. The modifier locus was not salient in the initial linkage scan, however, 2L analysis conditioning on genotype at locus 1 led to discovery of the true modifier variant. While both the PPL and the PPLD at this locus were relatively small, they were easily detected based on the amount of increase of the 2L signals relative to the original SL signals.

Kelvin can also be used to estimate the trait model using maximum likelihood estimators (m.l.e.'s) following the theory developed in Clerget-Darpoux et al. (1986), Elston (1989), Greenberg (1989), and Vieland and Hodge (1998). While our numerical integration routine is not optimized for maximization and therefore returns approximate rather than exact m.l.e.'s, it is interesting to note the models obtained from these analyses (**Table 2**). The

Table 2 | Approximate maximum likelihood trait parameter estimates.

Analysis	Locus	Disease allele frequency	Penetrances
SL-PPL	1	0.011	0.75, 0.56, 0.006
SL-PPLD	1	0.022	0.50, 0.49, 0.01
2L-PPL	2	0.125	0.99, 0.97, 0.011
2L-PPLD	2	0.25	0.99, 0.98, 0.011

Penetrances are given for: (SL), D_1D_1, D_1d_1, and d_1d_1 genotypes, respectively, where "D_1" indicates the putative disease allele at locus 1; (2L) D_2D_2, D_2d_2, and d_2d_2 genotypes, respectively at locus 2, among those individuals who carry D_1D_1 or D_1d_1.

disease allele frequency is estimated quite accurately by both PPL and PPLD analyses at locus 1; while at locus 2, the 2L-PPLD in particular returns an estimate reasonably close to the generating model. (Particularly at locus 2 where the PPL and PPLD themselves are quite low, the standard error of these estimates is likely to be substantial. Kelvin itself has no direct way to calculate these, but see Nouanesengsy et al., 2009 for further discussion.) More interesting, however, are the penetrance estimates. While there is no exact analog of the random reduced penetrance parameter of the SL model for a 2L generating model, using the approach described in Vieland et al. (1993), we obtain a SL penetrance vector "corresponding" to the generating model of (0.62, 0.62, 0) for the putative disease genotypes, respectively. This vector is approximated very closely by both the PPL and PPLD m.l.e.'s at locus 1. At the modifier locus, considering only individuals coded in the "dominant" LC based on locus 1, the estimated penetrance vectors indicate a virtually fully penetrant 2L dominant–dominant epistatic model. Thus overall, we were able not only to map both loci to the level of the individual variant, but also to determine the correct generating model with great accuracy, all in a single, highly informative pedigree.

MCMC ACCURACY

As seen in **Figure 6A**, repeating the entire MCMC sampling process five times produced very similar, albeit not identical, PPL profiles across chromosome 4. The marker log likelihood for chromosome 4 from the single long MCMC run still showed some upward convergence up to about 14,000 iterations, at which point it remained essentially flat. Comparing the final (post-convergence) 5,000 iterations with the original results (**Figure 6B**) again supported the accuracy of the original analysis in terms of the PPLs themselves. Again, however, the results are not identical. **Figure 6C** shows PPLs based on each of the component shorter sampling runs (as averaged over to obtain the original results) considered independently. There is considerable variation, particularly at positions further away from the true casual SNP. This strongly suggests, not surprisingly, that shorter runs of this length are not individually sufficient.

However, averaging across this set of shorter runs did enable us to achieve accurate results. Compared to a single, extremely long run, this is also a highly cost-effective approach insofar as it enables us to distribute the MCMC iterations to run concurrently on separate processors. On our hardware, the pooled-iteration process (using 10 servers with 2.5 GHz CPUs and 8 G memory) required 4 h, 40 min to complete chromosome 4, while the single long run (using one server) required 3 days, 5.5 h. Additional simulation studies are needed to further compare averaging across shorter sampling runs with use of single long sampling runs, especially across different pedigree structures with different patterns of missing data.

DISCUSSION

We have illustrated an approach to gene discovery based on a single, highly informative family. This approach involves narrowing the genomic region(s) of interest using linkage analysis, followed by fine-mapping based on targeted LD (association) analysis in the same family. We have additionally illustrated how not just primary but even modifier genes can in principle be detected within a single pedigree.

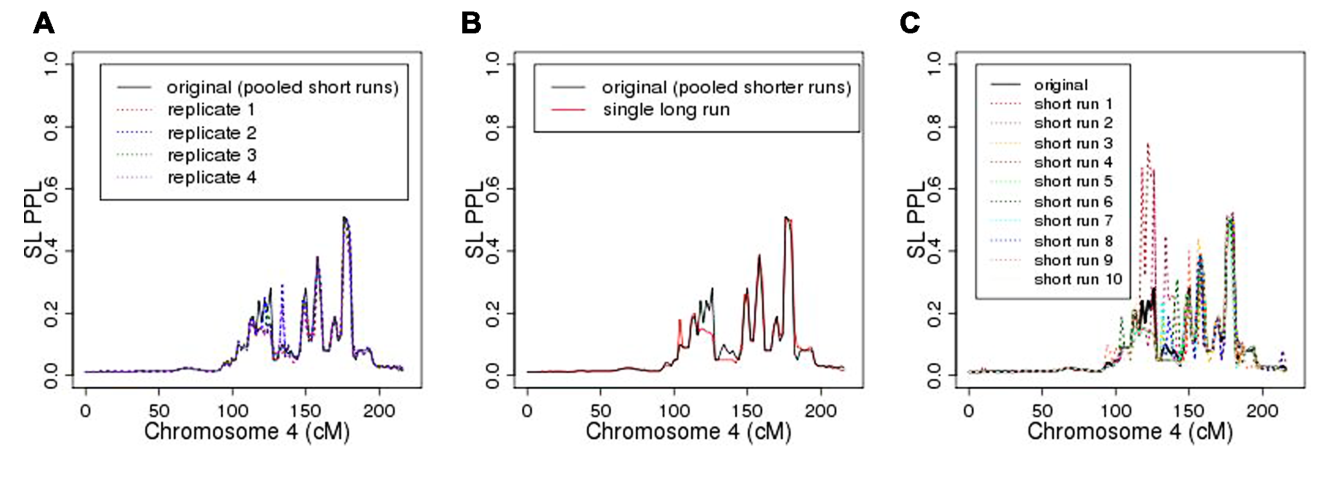


FIGURE 6 | (A) Five replicates of chromosome 4 analysis; **(B)** original analysis compared to single long sampling run; **(C)** original (averaged) analysis compared to individual component short runs.

Of course, we "cheated" by including the two causal SNPs in our association panel. In general, we might expect to have data from a standard SNP chip available on most family members for purposes of linkage and association mapping, together with sequence on a subset of individuals. In this case, the association analyses could be conducted on every observed sequence variant in the regions of interest, ensuring that the true mutation would be included (assuming that the relevant disease-causing element is a SNP). Of course to do this, the sequence variants would need to be measured in many family members, but at least in principle this could be done in part through imputation of sequence using sequencing in a subset of individuals combined with SNP data on the remaining individuals.

We chose our 2L generating model to be moderately mappable at the primary locus but with a modifier locus that was much harder to find. Of course in practice, realistic models may present more difficult challenges at all component loci, and this illustration is in no way to be construed as an estimate of any kind of power to find the genes. However, one salient feature of this approach is that it is not dependent on bioinformatic "filtering" approaches to prioritizing sequence variants as likely candidates. Instead, following the now classical reverse genetic paradigm, we rely entirely on positional mapping even at the variant-selection stage. Again, in practice this is likely to still leave a number of variants as candidates, since highly correlated variants under a peak may still be difficult to resolve statistically. Nevertheless, the number of such variants likely to be left on the list of candidates is greatly reduced by focusing on the linked and associated regions.

As noted above, the PPL framework is designed to measure strength of evidence, and not to test hypothesis or serve as a decision making algorithm. Thus at no point in the discussion did we consider "significance levels" or decide whether the evidence was "strong enough" to declare success. Rather, we relied on the accuracy of the framework overall as an evidence measurement technique, and simply followed up on the strongest evidence wherever that occurred. In this particular case, doing so led us to find both genes and both causal SNPs, without any "false positive" results. In practice, of course, difficult decisions would need to be made before, e.g., expending substantial resources following up functionally on the locus 2 SNP, given the very low PPLD. Nevertheless, had we set very stringent significance criteria from the outset and refused to follow-up on the strongest evidence regardless of the absolute numbers involved, we would have missed the modifier locus entirely. We note too that in consortium settings, Kelvin's use of Bayesian sequential updating to accumulate evidence across data sets provides an alternative to traditional meta-analysis. Access to primary data, and not just summary measures such as *p*-values, is required for this. However, Kelvin outputs posterior marginal distributions, which can be used to sequentially update results across data sets without the need to actually pool the data themselves across sites.

The study design utilized here presented us with one salient computational challenge: how to compute the (parametric) likelihood for so large a pedigree. For this purpose we engineered a hybrid version of Kelvin using MCMC for the marker data and a non-stochastic method for integration over the trait parameters. This method proved to be quite accurate and computationally feasible, at least for data of this type. Of course the method can also be applied to sets of large pedigrees, and as noted, combined with ES- or LG-based analyses of smaller pedigrees or pedigrees with sparser marker maps for greater computational efficiency when analyzing data sets with variable family sizes.

Further studies in additional pedigree structures are needed to make specific recommendations regarding burn-in lengths and number of iterations needed to maximize the chances of accurate results for the MCMC portion of the calculation. In this regard, our new method is no better and no worse than McSample itself. However, we have some reason to think that the PPL and PPLD themselves may be relatively robust to some level of sampling variability in the underlying marker likelihood, possibly in part because integration over the trait model protects against modest amounts of imprecision at the marker level. This remains a subject for further investigation.

In this particular application, however, 2,000 samples derived from pooling the results of 10 independent sampling processes, each with 200 iterations following a 1,000-sample burn-in, appears to have been highly accurate. Still, this approach remains out of reach for genome-wide analysis on a typical desktop machine, requiring instead a distributed cluster environment to make real-time completion of results feasible. As high performance computing environments become more common for purposes of whole-genome sequence analysis and other "-omics" applications, we hope that this will become less of an impediment to analyses of the sort proposed here. Given the costs of data collection, we would argue that the additional computational demands are worth while if the methods are effective. The most definitive demonstration that they were effective in the current application is in the final results: successful mapping of two interacting disease loci down to the level of the individual causal variants.

ACKNOWLEDGMENTS

This work was funded in part by National Institute of Mental Health grant U24 MH068457 (under a subcontract to Veronica J. Vieland; J. Tischfield, PI) and by NIH R01GM081417 (Alun Thomas). We would like to thank Dr. Irma Järvelä for permission to use her data and Jaana Oikkonen for curation of the marker data used here.

REFERENCES

Abel, H. J., and Thomas, A. (2011). Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation. *Stat. Appl. Genet. Mol. Biol.* 10, 1–15.

Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* 62, 1198–1211.

Bartlett, C. W., and Vieland, V. J. (2006). Accumulating quantitative trait linkage evidence across multiple datasets using the posterior probability of linkage. *Genet. Epidemiol.* 31, 91–102.

Berntsen, J., Espelid, T. O., and Genz, A. (1997). An adaptive multidimensional integration routine for a vector of integrals. *ACM Trans. Math. Softw.* 17, 452–456.

Clerget-Darpoux, F., Bonaiti-Pellie, C., and Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* 42, 393–399.

Edwards, A. (1992). *Likelihood*. Baltimore: Johns Hopkins University Press.

Elston, R. C. (1989). Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. *Am. J. Med. Genet.* 34, 487–488.

Elston, R. C., and Lange, K. (1975). An approximation for the prior probability of autosomal linkage. *Cytogenet. Cell Genet.* 14, 290–292.

Elston, R. C., and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21, 523–542.

Greenberg, D. A. (1989). Inferring mode of inheritance by comparison of lod scores. *Am. J. Med. Genet.* 34, 480–486.

Hou, L., Wang, K., and Bartlett, C. W. (2012). Evaluation of a Bayesian model integration-based method for censored data. *Hum. Hered.* 74, 1–11.

Huang, Y., and Vieland, V. J. (2010). Association statistics under the PPL framework. *Genet. Epidemiol.* 34, 835–845.

Lander, E. S., and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.* 84, 2363–2367.

Nouanesengsy, B., Seok, S.-C., Shen, H.-W., and Vieland, V. J. (2009). "Using projection and 2D plots to visually explore multidimensional genetic likelihood spaces" in *IEEE Symposium on Visual Analytics Science and Technology IEE VAST 2009*, 171–178.

Ott, J. (1976). A computer program for linkage analysis of general human pedigrees. *Am. J. Hum. Genet.* 28, 528–529.

Seok, S., Evans, M., and Vieland, V. J. (2009). Fast and accurate calculation of a computationally intensive statistic for mapping disease genes. *J. Comput. Biol.* 16, 659–676.

Smith, C. A. B. (1963). Testing for heterogeneity of recombination fraction values in human genetics. *Ann. Hum. Genet.* 27, 175–182.

Thomas, A. (2010). Assessment of SNP streak statistics using gene drop simulation with linkage disequilibrium. *Genet. Epidemiol.* 34, 119–124.

Thomas, A., Gutin, A., Abkevich, V., and Bansal, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Stat. Comput.* 10, 259–269.

Vieland, V. J., Greenberg, D. A., and Hodge, S. E. (1993). Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. *Hum. Hered.* 43, 329–336.

Vieland, V. J., and Hodge, S. E. (1998). Review of statistical evidence: a likelihood paradigm. *Am. J. Hum. Genet.* 63, 283–289.

Vieland, V. J., Hodge, S. E., and Greenberg, D. A. (1992). Adequacy of single-locus approximations for linkage analyses of oligogenic traits. *Genet. Epidemiol.* 9, 45–59.

Vieland, V. J., Huang, Y., Seok, S. C., Burian, J., Catalyurek, U., O'Connell, J., et al. (2011). KELVIN: a software package for rigorous measurement of statistical evidence in human genetics. *Hum. Hered.* 72, 276–288.

Wang, H., Segre, A., Huang, Y., O'Connell, J., and Vieland, V. (2007). "Rapid computation of large numbers of LOD scores in linkage analysis through polynomial expression of genetic likelihoods," in *Proceedings of IEEE Workshop on High-Throughput Data Analysis for Proteomics and Genomics IEEE 2007*, 197–204.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2013; accepted: 03 April 2013; published online: 19 April 2013.

Citation: Huang Y, Thomas A and Vieland VJ (2013) Employing MCMC under the PPL framework to analyze sequence data in large pedigrees. Front. Genet. ***4****:59. doi: 10.3389/fgene.2013.00059*

This article was submitted to Frontiers in Applied Genetic Epidemiology, a specialty of Frontiers in Genetics.

frontiers in
GENETICS

PERSPECTIVE ARTICLE
published: 30 April 2013
doi: 10.3389/fgene.2013.00061

The household contact study design for genetic epidemiological studies of infectious diseases

Catherine M. Stein[1,2]*, Noémi B. Hall[1], LaShaunda L. Malone[2] and Ezekiel Mupere[1,2,3]

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA
[2] Uganda – Case Western Reserve University Research Collaboration, Kampala, Uganda
[3] Mulago Hospital, Makerere University School of Medicine, Kampala, Uganda

Edited by:
Karen T. Cuenco, University of PIttsburgh, USA

Reviewed by:
Peter A. Kanetsky, Perelman School of Medicine at the University of Pennsylvania, USA
Karen T. Cuenco, University of PIttsburgh, USA
Hui-Qi Qu, The University of Texas School of Public Health, USA

***Correspondence:**
Catherine M. Stein, Department of Epidemiology and Biostatistics, Case Western Reserve University, 2103 Cornell Road, Wolstein Research Building 1316, Cleveland, OH 44106, USA.
e-mail: catherine.stein@case.edu

Most genetic epidemiological study designs fall into one of two categories: family based and population-based (case–control). However, recent advances in statistical genetics call for study designs that combine these two approaches. We describe the household contact study design as we have applied it in our several years of study of the epidemiology of tuberculosis. Though we highlight its applicability for genetic epidemiological studies of infectious diseases, there are many facets of this design that are appealing for modern genetic studies, including the simultaneous enrollment of related and unrelated individuals, closely and distantly related individuals, collection of extensive epidemiologic and phenotypic data, and evaluation of effects of shared environment and gene by environment interaction. These study design characteristics are particularly appealing for current sequencing studies.

Keywords: extended pedigrees, genetic association, whole genome studies, cohort study, case-contact

INTRODUCTION

The advantages of family studies for genetic epidemiology have long been established (Stein and Elston, 2009). Early methods in genetic epidemiology utilized twins, sibling pairs, and other relative pairs to establish the relative recurrence risk of a disease. Segregation analysis and traditional linkage analysis can only be conducted using pedigree data. Concerns of population stratification are easily accounted for. In addition to these analytical issues, family studies have the advantage of investment of relatives; if someone in the family has a particular disease, family members are more likely to participate in research in order to somehow help their relative and others affected with the disease. Today, with the advent of whole exome and whole genome sequencing technologies, there are additional advantages of family studies, which we shall review below.

These advantages of family studies are further amplified for genetic epidemiological studies of infectious diseases. It was once believed that tuberculosis (TB) was a familial disease because it occurred within families. Once the disease was determined to be caused by a mycobacteria, the ideas surrounding the familial component recessed to the background. Now decades after the causal pathogen, *Mycobacterium tuberculosis* (Mtb), has been identified, many studies have shown that human genetic factors influence risk for development of TB infection and disease (Moller and Hoal, 2010; Stein, 2011). Development of TB infection and disease is essentially a phenotype resulting from a gene by environment interaction, so a well-constructed genetic epidemiological study must account for host genetics, shared environment, and gene x environment interaction. In this paper, we provide an overview of our household contact (HHC) study of TB and its advantages for genetic epidemiological studies, particularly in light of study designs best suited to identify rare genetic variants.

OVERVIEW OF THE HOUSEHOLD CONTACT STUDY DESIGN

In its natural history, TB is a two-stage process of infection followed by disease (Comstock, 1982). The household provides a natural setting to study TB because the genetic epidemiology of the two stages of infection and disease can be characterized. In our previous studies (Guwatudde et al., 2003), we defined a household as a group of people living within one residence and share meals together with a head of family who makes decisions for the household. Extensive epidemiological data are collected on individual risk factors, such as proximity and frequency of contact with the index case as well as other factors that may increase susceptibility, characteristics of the home that may increase the risk of transmission, as well as clinical data. Blood samples are obtained at baseline and longitudinally for genetic and immunologic studies.

In our HHC study, the first TB patient is identified in the household and referred to as the index case. Thereafter individuals who reside in the same household with the index case for a certain period prior to the diagnosis of the index case are identified and screened for TB as HHCs. Each HHC is also evaluated clinically for latent Mtb infection with the tuberculin skin test (or interferon-γ response assay in the future). Individuals who are tuberculin skin test negative have repeated skin tests several times over the 2-year study follow-up. Thus, the HHC evaluation is efficient in identification of individuals with different phenotypes or stages of TB infection in a household including: (1) exposed and uninfected, (2) exposed and infected without disease, (3) recent infection, and (4) active TB. These different household phenotypes or categories can provide the basis to compare genetic factors associated with TB infection and disease. As all of these stages of infection and disease are diagnosed, both the index case and his/her contacts receive appropriate clinical care and treatment, which is an immediate benefit to all study participants.

The design of the HHC study is ideal for evaluating genetic susceptibility to TB (Stein et al., 2003, 2005, 2007, 2008). The family structure and the ability to identify sibling pairs can form the basis for linkage analysis studies. Evaluation for new candidate genes for TB can be done through conduct of association studies such as case–control, family based, and/or case–parent studies. Heritability to TB can be determined using standard quantitative genetic approaches which can be based on host immune responses as intermediate phenotypes (Stein et al., 2005; Tao et al., 2013). Studies of HHCs have demonstrated that young children are at greater risk for developing TB and the clustering of cases within families does give hint at a familial susceptibility (Brailey, 1940; Puffer et al., 1952).

In sum, the essence of an HHC design is the recruitment of an entire household through an index case/proband, and collection of extensive clinical and epidemiological data. All age ranges and relative pair types are enrolled, and the entire spectrum of disease is captured. There is flexibility for collection of biological samples and a longitudinal component to observe changes in phenotypes and biomarkers.

ADVANTAGES OF THE HHC DESIGN FOR CURRENT GENETIC EPIDEMIOLOGICAL STUDIES

RECRUITMENT AND PHENOTYPE COLLECTION

As summarized above, the household is ascertained through an index case with TB (aka proband). Thus, as long as each individual in the household provides informed consent (or assent in the case of children), an entire family is enrolled in the study. Sometimes, there is another individual with TB in the household at the time of enrollment (co-prevalent case). In some households, another individual develops TB later on during the course of study follow-up (incident case). In this respect, no additional recruitment efforts are needed to identify additional affected individuals. The longitudinal component of the HHC design is valuable, especially for TB, where individuals have a 5–10% lifetime risk of developing active disease after exposure. In our studies, we have observed incident cases develop 2 years after initial enrollment of the household. If related individuals are desired for analytical and study design reasons (see "Analytical Considerations" below), the HHC design allows for easier enrollment of relatives, particularly in settings where literacy is low and roads are impassable (Bennett et al., 2002). Since both HIV co-infected and uninfected individuals may live within the same household, both will be enrolled in the study; this enables the examination of gene by HIV interaction effects (Stein et al., 2007). Finally, the ideal setting for a case-contact study is where the balance of household vs. community spread of disease is in favor of the household (Hill and Ota, 2010).

Both pediatric and adult TB cases may be diagnosed because the HHC design does not restrict enrollment by age. Studies suggest that the genetic influences on pediatric vs. adult TB differ (Malik et al., 2005; Alcais et al., 2010) and the HHC study design is an efficient method for ascertaining both types of cases. By contrast, studies that focus solely on recruitment of pediatric TB cases are challenging – school-based studies are limited because children living in poverty may not have access to education, and hospital- and clinic-based studies may also miss out on enrolling children because many babies are born at home in developing countries and families in poverty who are most at risk for developing TB may not have access to medical care. Door-to-door case finding strategies would require a great number of resources in order to identify a sufficient number of pediatric cases.

The HHC design also enables the enrollment of appropriate "controls." For a proper case–control study, controls must be similar in every way to the cases except that they do not have the disease of interest. For infectious diseases like TB, this is especially true, and in order for an individual to have the opportunity to become a case, he/she must have been exposed to an infectious TB case. This is particularly important for TB, because clinical status of the controls determines whether observed genetic associations are with susceptibility to latent infection or progression to active disease (Stein, 2011). By virtue of the HHC design, all the household members have been exposed to the index case. The selection of appropriate controls in community-based studies of TB is problematic (Hill and Ota, 2010).

Finally, studies of large pedigrees often include extensive and highly detailed phenotype information (Wijsman, 2012). This is extraordinarily useful for infectious diseases such as TB for a number of reasons. As the natural history of Mtb infection and disease follows a two-stage process, the longitudinal HHC design captures all of these stages, and progression from one stage to another. Furthermore, the HHC design can also include collection of extensive immunological data. The HHC design therefore is flexible enough to analyze immunological correlates of the natural history of TB (Whalen et al., 2006; Mahan et al., 2012), and also genetic influences on the immune response to Mtb (Stein et al., 2007, 2008). Omics technologies, such as gene expression and proteomic arrays, can also be incorporated into a study that has an established blood draw protocol and rigorous clinical classification. Finally, as we describe later, data are also collected on important epidemiological factors, which can be incorporated as covariates as well as in gene by environment interaction models.

ANALYTICAL CONSIDERATIONS

One unique aspect of HHC studies is that households may contain all sorts of relationship types – nuclear families, extended relatives, and unrelated individuals. Half-siblings are common in African settings where polygamy is practiced (Bennett et al., 2002). Similarly, adoption by extended relatives is common when children are orphaned, which may be particularly relevant in areas with a heavy AIDS burden.

A few studies have developed strategies for jointly analyzing family based and case–control/population-based data (Chen and

Lin, 2008; Gray-McGuire et al., 2009; Lasky-Su et al., 2010; Zheng et al., 2010; Mirea et al., 2012). Though they differ in how they combine data from these two different study designs – some analyze them all together, and some combine *p*-values or test statistics – there are some common themes. First, joint analysis of data from these two different study designs results in increased power due to increased sample size, enabling the detection of smaller effect sizes. Second, family based data have the advantage of controlling for population substructure, which alleviates this common concern of population-based studies.

There have been many recent reports detailing the usefulness of extended pedigrees for the analysis of sequence data and detection of rare variants. Cirulli and Goldstein (2010) explain how the analysis of distantly related, co-affected individuals is an economical design, because there will be fewer genetic variants in common, thereby reducing the search space for rare variants. Stringent filtering could use identity-by-descent sharing to capitalize on this biological phenomenon (Akula et al., 2011). Large pedigrees also have increased power to detect linkage, even in the presence of linkage heterogeneity among families, and are enriched for variants of interest (Wijsman, 2012). Linkage analysis with pedigree data can be used as a filtering strategy of chromosomal regions, and can guide the selection of subjects to sequence (Wijsman, 2012). In addition, linkage analysis may be conducted to examine co-segregation between the trait and variant(s) of interest (Clerget-Darpoux and Elston, 2007; Ziegler and Sun, 2012). Consanguineous marriages are common in West Africa, which increases the power to detect rare recessive alleles (Bennett et al., 2002). To summarize, all of the relationship types that are useful for the identification of rare variants are easily obtainable in the HHC design.

IMPACT OF ENVIRONMENT

A well-designed HHC study includes vast epidemiologic data about environmental risk factors for transmission of disease within homes. For TB, these include factors related to ventilation and crowding within the home, poverty, clinical characteristics of the index case that make him/her more infectious, and proximity to the index case that increase degree of contact (Stein et al., 2005; Mandalakas et al., 2012). Risk of infection by Mtb is determined by a number of epidemiological risk factors (Guwattude et al., 2003; Lienhardt et al., 2003; Mandalakas et al., 2012), and many variables associated with high risk of TB transmission are automatically present in the HHC design. Analysis of foster relationships as seen in adoptions may be useful for the estimation of effects due to shared environment (Bennett et al., 2002), and many such relationships occur in HHC studies in the developing world.

Genetic substrains of Mtb may differ in their transmissibility. All of these factors relate to the risk of an individual to acquire infection, and develop disease, and thus are important in epidemiological characterization of affected individuals. Furthermore, recent studies have also suggested that substrains of Mtb have synergistic effects with host genes, thus resulting in gene x environment interaction effects related to TB risk (Caws et al., 2008). Case-only designs can be nested within HHC studies to examine these gene x environment effects (Bennett et al., 2002). Because exposure to the index case is generally highest, and in turn exposure to that individual's strain of Mtb, the HHC design provides a natural setting to test both transmissibility, gene x environment interaction, and role of shared environment.

Nutrition and nutritional status are also important factors in TB-related outcomes (Jaganath and Mupere, 2012; Mupere et al., 2012a). We have shown that nutritional status of a patient may be an indicator on how the food basket is shared in the household and the subsequent macro- and micronutrient intake (Mupere et al., 2012b). Because of the shared environmental and genetic components of diet and obesity (or in the case of TB, malnutrition), the HHC design provides a robust setting to test the role of nutritional status on infectious disease outcomes.

EXAMPLES FROM OUR STUDIES

Our genetic association studies have taken the approach by Gray-McGuire et al. (2009). We identified the first reported association between TNFR1 gene and TB and also a gene by HIV interaction for this same gene (Stein et al., 2007). Our genome-wide linkage scan (Stein et al., 2008) and subsequent fine mapping studies (Baker et al., 2011) replicated previously a novel set of genes on chromosome 20, CTSZ, and MC3R. We have also identified novel chromosomal regions linked to a unique resistance phenotype (Stein et al., 2008); we are uniquely able to clinically and epidemiologically characterize this phenotype because of our solid study design. Our future plans will incorporate structural equation modeling (SEM to multivariately analyze the influences of host genetics, immunology, and environment on clinical outcome; this shall be done using a SEM approach that jointly models familial relationship and covariance among variables (Morris et al., 2011).

CONCLUSION

Certainly HHC designs may be expensive to implement, because they include repeated clinical visits, longitudinal data collection, and travel to the homes. However, the wealth of data collected through HHC studies is invaluable for genetic epidemiological studies, as described here. HHC study designs offer unique advantages for genetic epidemiological studies, including the presence of related and unrelated individuals, and the ability to quantify environmental factors that are important for both shared environmental influences on the phenotype as well as gene x environment interaction. Though our focus has been primarily on studies of TB, this study design has advantages for the study of infectious diseases in general (Hill and Ota, 2010).

ACKNOWLEDGMENTS

We would like to thank Drs. Christopher Whalen and W. Henry Boom for their key roles in the design and implementation of the Household Contact Studies in Kampala, Uganda.

Funding for this work was provided by the Tuberculosis Research Unit (grant N01-AI95383 and HHSN266200700022C/N01-AI70022 from the NIAID), and National Heart Lung and Blood Institutes (NHLBI) grants R01HL096811, R01HL-10566113, and T32HL007567.

REFERENCES

Akula, N., Detera-Wadleigh, S., Shugart, Y., Nalls, M., Steele, J., and McMahon, F. J. (2011). Identity-by-descent filtering as a tool for the identification of disease alleles in exome sequence data from distant relatives. *BMC Proc.* 5(Suppl. 9):S76. doi: 10.1186/1753-6561-5-S9-S76

Alcais, A., Quintana-Murci, L., Thaler, D. S., Schurr, E., Abel, L., and Casanova, J. L. (2010). Life-threatening infectious iseases of childhood: single-gene inborn errors of immunity? *Ann. N. Y. Acad. Sci.* 1214, 18–33.

Baker, A. R., Zalwango, S., Malone, L. L., Igo, R. P. Jr, Qiu, F., Nsereko, M., et al. (2011). Genetic Susceptibility to Tuberculosis Associated with CTSZ Haplotype in Ugandan Household Contact Study. *Hum. Immunol.* 72, 426–430.

Bennett, S., Lienhardt, C., Bah-Sow, O., Gustafson, P., Manneh, K., Del Prete, G., et al. (2002). Investigation of environmental and host-related risk factors for tuberculosis in Africa. II. Investigation of host genetic factors. *Am. J. Epidemiol.* 155, 11, 1074–1079.

Brailey, M. (1940). Mortality in the children of tuberculous households. *Am. J. Public Health Nations Health* 30, 816–823.

Caws, M., Thwaites, G., Dunstan, S., Hawn, T. R., Lan, N. T., Thuong, N. T., et al. (2008). The influence of host and bacterial genotype on the development of disseminated disease with *Mycobacterium tuberculosis*. *PLoS Pathog.* 4:e1000034. doi: 10.1371/journal.ppat.1000034

Chen, Y. H., and Lin, H. W. (2008). Simple association analysis combining data from trios/sibships and unrelated controls. *Genet. Epidemiol.* 32, 520–527.

Cirulli, E. T., and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415–425.

Clerget-Darpoux, F., and Elston, R. C. (2007). Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum. Hered.* 64, 91–96.

Comstock, G. (1982). Epidemiology of tuberculosis. *Am. Rev. Respir. Dis.* 125, 8–15.

Gray-McGuire, C., Bochud, M., Goodloe, R., and Elston, R. C. (2009). Genetic association tests: a method for the joint analysis of family and case–control data. *Hum. Genomics* 4, 2–20.

Guwattude, D., Nakakeeto, M., Jones-Lopez, E. C., Maganda, A., Chiunda, A., Mugerwa, R. D., et al. (2003). Tuberculosis in household contacts of infectious cases in Kampala, Uganda. *Am. J. Epidemiol.* 158, 887–898.

Guwatudde, D., Zalwango, S., Kamya, M. R., Debanne, S. M., Diaz, M. I., Okwera, A., et al. (2003). Burden of tuberculosis in Kampala, Uganda. *Bull. World Health Organ.* 81, 799–805.

Hill, P. C., and Ota, M. O. (2010). Tuberculosis case-contact research in endemic tropical settings: design, conduct, and relevance to other infectious diseases. *Lancet Infect. Dis.* 10, 723–732.

Jaganath, D., and Mupere, E. (2012). Childhood tuberculosis and malnutrition. *J. Infect. Dis.* 206, 1809–1815.

Lasky-Su, J., Won, S., Mick, E., Anney, R. J., Franke, B., Neale, B., et al. (2010). On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls. *Am. J. Hum. Genet.* 86, 573–580.

Lienhardt, C., Fielding, K., Sillah, J., Tunkara, A., Donkor, S., Manneh, K., et al. (2003). Risk factors for tuberculosis infection in sub-Saharan Africa: a contact study in The Gambia. *Am. J. Respir. Crit. Care Med.* 168, 448–455.

Mahan, C. S., Zalwango, S., Thiel, B. A., Malone, L. L., Chervenak, K. A., Baseke, J., et al. (2012). Innate and adaptive immune responses during acute M. tuberculosis infection in adult household contacts in Kampala, Uganda. *Am. J. Trop. Med. Hyg.* 86, 690–697.

Malik, S., Abel, L., Tooker, H., Poon, A., Simkin, L., Girard, M., et al. (2005). Alleles of the *NRAMP1* gene are risk factors for pediatric tuberculosis disease. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12183–12188.

Mandalakas, A. M., Kirchner, H. L., Lombard, C., Walzl, G., Grewal, H. M., Gie, R. P., et al. (2012). Well-quantified tuberculosis exposure is a reliable surrogate measure of tuberculosis infection. *Int. J. Tuberc. Lung Dis.* 16, 1033–1039.

Mirea, L., Infante-Rivard, C., Sun, L., and Bull, S. B. (2012). Strategies for genetic association analyses combining unrelated case–control individuals and family trios. *Am. J. Epidemiol.* 176, 70–79.

Moller, M., and Hoal, E. G. (2010). Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis (Edinb.)* 90, 71–83.

Morris, N. J., Elston, R. C., and Stein, C. M. (2011). A framework for structural equation models in general pedigrees. *Hum. Hered.* 70, 278–286.

Mupere, E., Malone, L., Zalwango, S., Chiunda, A., Okwera, A., Parraga, I., et al. (2012a). Lean tissue mass wasting is associated with increased risk of mortality among women with pulmonary tuberculosis in urban Uganda. *Ann. Epidemiol.* 22, 466–473.

Mupere, E., Parraga, I. M., Tisch, D. J., Mayanja, H. K., and Whalen, C. C. (2012b). Low nutrient intake among adult women and patients with severe tuberculosis disease in Uganda: a cross-sectional study. *BMC Public Health* 12:1050. doi: 10.1186/1471-2458-12-1050.:1050-12

Puffer, R. R., Zeidberg, L. D., Dillon, A., Gass, R. S., and Hutcheson, R. H. (1952). Tuberculosis attack and death rates of household associates; the influence of age, sex, race, and relationship. *Am. Rev. Tuberc.* 65, 111–127.

Stein, C. M. (2011). Genetic epidemiology of tuberculosis susceptibility: impact of study design. *PLoS Pathog.* 7:e1001189. doi: 10.1371/journal.ppat.1001189

Stein, C. M., and Elston, R. C. (2009). Finding genes underlying human disease. *Clin. Genet.* 75, 101–106.

Stein, C. M., Nshuti, L., Chiunda, A. B., Boom, W. H., Elston, R. C., Mugerwa, R. D., et al. (2005). Evidence for a major gene influence on tumor necrosis factor-alpha expression in tuberculosis: path and segregation analysis. *Hum. Hered.* 60, 109–118.

Stein, C. M., Zalwango, S., Chiunda, A. B., Millard, C., Leontiev, D. V., Horvath, A. L., et al. (2007). Linkage and association analysis of candidate genes for TB and TNFalpha cytokine expression: evidence for association with IFNGR1, IL-10, and TNF receptor 1 genes. *Hum. Genet.* 121, 663–673.

Stein, C. M., Zalwango, S., Malone, L. L., Won, S., Mayanja-Kizza, H., Mugerwa, R. D., et al. (2008). Genome scan of *M. tuberculosis* infection and disease in Ugandans. *PLoS ONE* 3:e4094. doi: 10.1371/journal.pone.0004094

Stein, C., Guwatudde, D., Nakakeeto, M., Peters, P., Elston, R. C., Tiwari, H. K., et al. (2003). Heritability analysis of cytokines as intermediate phenotypes of tuberculosis. *J. Infect. Dis.* 187, 1679–1685.

Tao, L., et al. (2013). Genetic and shared environmental influences on interferon-gamma production in response to *Mycobacterium tuberculosis* antigens in a Ugandan population. *Am. J. Trop. Med. Hyg.* (in press).

Whalen, C. C., Chiunda, A., Zalwango, S., Nshuti, L., Jones-Lopez, E., Okwera, A., et al. (2006). Immune correlates of acute *Mycobacterium tuberculosis* infection in household contacts in Kampala, Uganda. *Am. J. Trop. Med. Hyg.* 75, 55–61.

Wijsman, E. M. (2012). The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet.* 131, 1555–1563.

Zheng, Y., Heagerty, P. J., Hsu, L., and Newcomb, P. A. (2010). On combining family-based and population-based case–control data in association studies. *Biometrics* 66, 1024–1033.

Ziegler, A., and Sun, Y. V. (2012). Study designs and methods post genome-wide association studies. *Hum. Genet.* 131, 1525–1531.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 January 2013; accepted: 05 April 2013; published online: 30 April 2013.

Citation: Stein CM, Hall NB, Malone LL and Mupere E (2013) The household contact study design for genetic epidemiological studies of infectious diseases. Front. Genet. ***4****:61. doi: 10.3389/fgene.2013.00061*

This article was submitted to Frontiers in Applied Genetic Epidemiology, a specialty of Frontiers in Genetics.

frontiers in GENETICS

ORIGINAL RESEARCH ARTICLE
published: 20 August 2013
doi: 10.3389/fgene.2013.00152

A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer

Quentin Nelson[1], **Neeraj Agarwal**[1,2], **Robert Stephenson**[2,3,4] **and Lisa A. Cannon-Albright**[1,2,3]*

[1] Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA
[2] Huntsman Cancer Institute, Salt Lake City, UT, USA
[3] George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT, USA
[4] Surgery, University of Utah Health Sciences Center, Salt Lake City, UT, USA

Edited by:
Robert C. Elston, Case Western Reserve University, USA

Reviewed by:
Lara E. Sucheston, Roswell Park Cancer Institute, USA
Tao Wang, Albert Einstein College of Medicine, USA

***Correspondence:**
Lisa A. Cannon-Albright, Internal Medicine, University of Utah School of Medicine, 30 N 1900 E, Salt Lake City, UT 84132, USA
e-mail: lisa.albright@utah.edu

Background: Prostate cancer is a common and often deadly cancer. Decades of study have yet to identify genes that explain much familial prostate cancer. Traditional linkage analysis of pedigrees has yielded results that are rarely validated. We hypothesize that there are rare segregating variants responsible for high-risk prostate cancer pedigrees, but recognize that within-pedigree heterogeneity is responsible for significant noise that overwhelms signal. Here we introduce a method to identify homogeneous subsets of prostate cancer, based on cancer characteristics, which show the best evidence for an inherited contribution.

Methods: We have modified an existing method, the Genealogical Index of Familiality (GIF) used to show evidence for significant familial clustering. The modification allows a test for excess familial clustering of a subset of prostate cancer cases when compared to all prostate cancer cases.

Results: Consideration of the familial clustering of eight clinical subsets of prostate cancer cases compared to the expected familial clustering of all prostate cancer cases identified three subsets of prostate cancer cases with evidence for familial clustering significantly in excess of expected. These subsets include prostate cancer cases diagnosed before age 50 years, prostate cancer cases with body mass index (BMI) greater than or equal to 30, and prostate cancer cases for whom prostate cancer contributed to death.

Conclusions: This analysis identified several subsets of prostate cancer cases that cluster significantly more than expected when compared to all prostate cancer familial clustering. A focus on high-risk prostate cancer cases or pedigrees with these characteristics will reduce noise and could allow identification of the rare predisposition genes or variants responsible.

Keywords: familiality, prostate cancer, lethal, UPDB

INTRODUCTION

Prostate cancer is the most commonly diagnosed cancer in men and is the second leading cause of cancer deaths among men (ACS, 2013). While there is significant evidence of a genetic contribution (Cannon et al., 1982; Carter et al., 1993; Stanford and Ostrander, 2001; Langeberg et al., 2007), decades of investigation into the genetic causes of familial prostate cancer has yet to clearly identify genes or variants which explain much more than a small number of pedigrees with an excess of prostate cancer. Traditional linkage analysis of thousands of high-risk prostate cancer pedigrees has elucidated little in the identification of predisposition genes responsible for prostate cancer pedigrees. This may reflect the heterogeneous nature of prostate cancer, and this could confound identification of informative homogeneous pedigrees segregating rare predisposition variants.

We hypothesize that there exist rare prostate cancer predisposition variants that are responsible for our observation of high risk prostate cancer pedigrees including homogeneous prostate cancer cases (defined by clinical characteristics). We present a methodology to compare subsets of prostate cancer cases and identify those that show more familial clustering than expected for all prostate cancer cases.

Using a population-based resource in Utah that combines genealogy and cancer data, we identified 3 subsets of prostate cancer cases that cluster in pedigrees more than expected: prostate cancer which is diagnosed before age 50 years, lethal prostate cancer (leading to metastasis and death from prostate cancer), and prostate cancer in men with BMI ≥ 30. We propose that analysis of the high-risk prostate cancer cases or pedigrees with an excess of prostate cancer cases with these characteristics could lead to identification of the rare predisposition variants responsible.

DATA AND METHODS

The Utah Population Data Base (UPDB) integrates three key electronic datasets: a Genealogy of the Utah pioneers constructed in the 1970s and kept current (Skolnick, 1980), death certificates for Utah, and a statewide cancer registry. The original Utah genealogy had approximately 1.6 million individual records for

186,000 three-generation families. Since the genealogy was created in the 1970s, state vital records have been used to create genealogy triplets (mother, father, and child) to extend the genealogy to present day. The UPDB has become a person-oriented database with information on 7 million Utahns, some 2.5 million of whom have at least three generations of genealogy. The Utah Cancer Registry (UCR) was created in 1966 to collect data on all cancer diagnosed in Utah. It became a SEER (Surveillance, Epidemiology, and End-Results) Registry of the National Cancer Institute in 1973. The UCR individual records are linked to the Utah genealogy annually; approximately 2/3 of UCR cases link to a record in the UPDB. Cause of death from Utah state death certificates from 1904 to present have been coded to ICD Revisions 6–10, and record linked to the UPDB. Utah Drivers License records from 1970 have been linked to the UPDB and include height and weight measurements for calculation of body mass index (BMI). The combination of genealogy, death certificates, drivers license data, and cancer registry data facilitates the identification of all Utah prostate cancer cases and the genetic relationships between them.

To perform the genetic analyses presented here we restrict ourselves to those individuals in the UPDB with ancestral genealogy data. We identified all individuals in the UPDB who were born before 1972 (when the original Utah genealogy was constructed) and whose parents, four grandparents, and six (of eight total) great grandparents are present in the UPDB genealogy data. This identifies 1.2 million individuals with ancestral genealogy data who are used for all analyses.

We have extended a well-published analysis method, the Genealogical Index of Familiality (GIF), to enable comparison of the relatedness of a subset of prostate cancer cases to the relatedness of *all* prostate cancer cases. Those subsets with evidence for significantly more relatedness than all prostate cancer cases are hypothesized to represent homogeneous genetic subsets that will be most informative for gene identification studies.

GENEALOGICAL INDEX OF FAMILIALITY (GIF) METHOD

For decades the GIF statistic has been used to quantify familial clustering of cancer and other phenotypes in the UPDB. This well-established statistical method has yielded strong evidence of heritability for several cancer phenotypes (Cannon et al., 1982; Cannon-Albright et al., 1994; Larson et al., 2006; Albright et al., 2012). The GIF was developed to test the hypothesis of excess relatedness of individuals with a common phenotype. Excess relatedness is measured by comparing the average relatedness between all pairs of cases of interest to the expected relatedness of matched controls from the Utah population. Since record linkage of any subset of UPDB records may indicate better or different quality data, for individuals with a death certificate, we select controls from all UPDB individuals who have a Utah death certificate. Since the UCR is statewide, we select controls for cancer cases from the entire UPDB resource.

The relatedness of a pair of individuals in a set is measured using the Malécot coefficient of kinship. The Malécot coefficient of kinship mathematically expresses Mendelian inheritance pattern probabilities that randomly selected homologous chromosomes are identical due to inheritance from a common ancestor. For example, the Malécot coefficient for siblings is 1/4, avunculars is 1/8, and first cousins 1/16. The GIF analysis tests excess relatedness by comparing all pairwise relationships within a set of cases to the expected relatedness measured in all pairwise relationships in 1000 sets of matched controls randomly selected from the UPDB. Controls were matched on characteristics that might be associated with record linking and disease rates, including five-year birth year cohort, sex, and birth state (Utah or not).

The overall GIF analysis tests for significant excess relatedness (over what is expected in the UPDB population) among a group of individuals. It can be performed on all prostate cancer cases, and on subsets of cases based on cancer characteristics. It cannot, however, determine which, if any, of these subsets exhibits the best evidence for a genetic predisposition, and which therefore might be the best set of high-risk pedigrees in which to search for genes.

NEW SUBSETGif TEST

Here we consider a modified GIF test and test the relatedness of multiple subsets of prostate cancer cases to identify those which exhibit excess relatedness above the observed relatedness among *all Utah prostate cancer cases*. This modified GIF test is referred to as the SubsetGif. Evidence for significant excess relatedness for a subset of prostate cancer cases above the expected for *all prostate cancer cases* could indicate the presence of a common genetic cause shared by the homogeneous subset. The identification and subsequent study of pedigrees including cases of such a homogeneous subset might facilitate the identification of rare predisposition genes.

CONTRIBUTION TO THE GIF BY GENETIC DISTANCE

It is possible to view the distribution of the contribution to the GIF statistic by the pairwise genetic distance of the different relationships observed in cases (and controls). The genetic distance represents the number of paths between a pair of individuals. Genetic distance 1 represents parent/offspring pairs, genetic distance 2 represents siblings or grandparent/grandchild, genetic distance 3 represents avunculars, and so forth.

RESULTS

In the UPDB resource, 18,291 prostate cancer cases were identified who also had ancestral genealogical records. The available prostate cancer subsets and their corresponding sample sizes are outlined in **Table 1**.

Table 1 | Subsets of prostate cancer and sample size.

Set of prostate cancer cases	*n*
All prostate cancers	18,291
Age at diagnosis <50 years	213
Metastatic disease at diagnosis	912
With at least 1 primary cancer of other site	2922
Gleason score >7 at diagnosis	4784
Short survival (0–9 months)	1180
Long survival (240 + months)	806
High BMI (≥30)	2459
Prostate cancer cause of death (lethal prostate cancer)	3982

ANALYSIS OF EXCESS RELATEDNESS

Previous studies have strongly supported evidence for a genetic contribution to predisposition to prostate cancer in the Utah population, as well as other populations (Cannon et al., 1982; Cannon-Albright et al., 1994, 2005). When all prostate cancer cases with genealogy data in the UPDB are analyzed there is evidence of excess relatedness (represented by both close and distant genetic relationships) over expected relatedness in matched Utah population controls. **Table 2** shows the traditional GIF test for excess relatedness compared to matched Utah population controls for all prostate cancer cases, and for each subset. The mean relatedness for cases and controls is shown. All prostate cancer cases and subsets, except prostate cases who survived less than 10 months after diagnosis, show strong evidence for excess clustering compared to Utah population controls. These results suggest a genetic contribution to prostate cancer predisposition, and suggest that study of almost all subsets of prostate cancer could be fruitful, but the results do not allow identification of which, if any, of the subsets are significantly more related than expected when compared to all prostate cancer cases, and thus show the best evidence for a genetic contribution.

In order to consider the hypothesis that a subset of prostate cancer cases represents a more homogeneous subset of highly related cases, we propose use of the SubsetGif analysis. The average pairwise relatedness of each subset of cases is compared to the average pairwise relatedness of 1000 sets of matched "controls"; these controls are selected from the set of 18,291 Utah prostate cancer cases. The results for this SubsetGif test are shown in **Table 3**. The average pairwise relatedness of the cases does not change for any subset (as expected), but the mean control GIF statistic is higher than in **Table 2** for each subset because the "controls" here are randomly selected prostate cancer cases, who are more closely related than random members of the Utah population.

Table 3 results show that the average pairwise relatedness of three different subsets of prostate cancer cases is significantly higher than expected among prostate cancer cases, supporting the hypothesis that these subsets of cases cluster more than all prostate cancer cases and represent sets on which to focus for predisposition gene identification. The three subsets include prostate cancer cases diagnosed before age 50 years, prostate cancer cases with BMI ≥ 30, and prostate cancer cases whose cause of death is prostate cancer (lethal prostate cancer).

It is difficult to determine whether these three subsets represent independent groups of interest or whether there is overlap between the groups because not all cases have BMI and death certificate data. There were 222 prostate cancer cases with BMI ≥ 30 among the 3982 cases with prostate cancer as a cause of death (6% total and 17% of the 1300 lethal cases with BMI data), and 58 prostate cancer cases with BMI ≥ 30 of the 213 cases who were

Table 2 | GIF analysis of prostate cancer relatedness compared to expected relatedness in the UPDB population.

Group	*n*	Case GIF	Mean control GIF	Empirical significance
All prostate cancers	18,291	5.54	4.74	<0.001
Age at diagnosis <50 years	213	11.72	4.54	<0.001
Metastatic disease at diagnosis	912	5.94	4.89	<0.001
With at least 1 primary cancer of other site	2922	5.58	4.74	<0.001
Gleason score >7 at diagnosis	4784	5.41	4.69	<0.001
Short survival (0–9 months)	1180	5.19	4.92	0.138
Long survival (240 + months)	806	5.64	4.75	0.005
BMI ≥ 30	2459	5.81	4.71	<0.001
Prostate cancer cause of death* (lethal)	3982	5.98	4.93	<0.001

Because the subset of lethal prostate cancer cases differs from all prostate cancer cases with respect to the identification of a linked death certificate record, and because the fact of record linking may suggest different data quality, we performed the GIF analysis for the subset of cases with prostate cancer contributing to death in **Tables 2, 3 using only the 10,421 prostate cancer cases with a linked Utah death certificate as controls; this is the standard for analysis of sets of individuals selected from Utah death certificate data (Cannon-Albright, 2008).*

Table 3 | Subset prostate cancer relatedness compared to expected prostate cancer case relatedness in the UPDB.

Prostate cancer subsets	*n*	Case GIF	Mean control GIF	Empirical significance
Age at diagnosis <50 years	213	11.72	7.51	0.024
Metastatic disease at diagnosis	912	5.94	5.95	0.506
With at least 1 primary cancer of other site	2922	5.58	5.51	0.303
Gleason Score >7 at diagnosis	4784	5.41	5.39	0.417
Short Survival (0–9 months)	1180	5.19	6.08	1.000
Long Survival (240 + months)	806	5.64	5.56	0.400
BMI ≥ 30	2459	5.81	5.27	<0.001
Prostate cancer cause of death (lethal)	3982	5.98	5.76	0.030

Controls randomly selected from 18,291 prostate cancer cases.

diagnosed before age 50 years (27%). Overall, 11,536 prostate cancer cases had BMI data, and 21.3% were BMI ≥ 30. There were 26 prostate cancer cases diagnosed before age 50 years (0.7%) among the 3982 lethal prostate cancer cases, and overall the 213 prostate cancer cases diagnosed before age 50 years represented 1% of all cases.

In order to determine the overall distribution of excess relatedness we can view the contribution to the GIF statistic by the pairwise genetic distance for cases and for controls. **Figure 1** shows the GIF distribution for all 18,291 prostate cancer cases compared to the distribution for the 1000 sets of matched Utah population controls. The comparison shows that the relatedness for prostate cancer cases exceeds that expected in the Utah population, as observed in random matched Utah controls, for genetic distances up to 7 (e.g., second cousins once removed).

Figures 2–4 show the contribution to the GIF statistic for the three subsets of cases, with matched controls randomly selected from all Utah prostate cancer cases. **Figure 2** shows this distribution for prostate cancer cases with BMI ≥ 30; as seen in **Table 3** there is significant excess relatedness for prostate cases with BMI ≥ 30. This excess extends to a genetic distance of 5, equivalent to first cousins once removed, for example. **Figure 3** shows this distribution for prostate cancer cases diagnosed before age 50 years, which is also observed to show significant excess relatedness. The excess relatedness is irregular, but is clearly observed for genetic distance = 2 (siblings primarily), and distance = 8 (third cousins, for example). **Figure 4** shows the GIF distribution for lethal prostate cancer cases, also observed to show significant excess clustering when compared to all deceased prostate cancer cases. The excess extends to genetic distance = 4, equivalent to first cousins, for example.

Figures 5–7 show examples Utah high-risk prostate cancer pedigrees for each of the subset characteristics identified.

DISCUSSION

Analysis of a population-based Utah resource linking cancer characteristics data with genealogy data has previously shown evidence for a genetic contribution to prostate cancer predisposition (Cannon et al., 1982; Cannon-Albright et al., 1994, 2005; Albright et al., 2012; Teerlink et al., 2012). Here we have extended a well-published analysis method which tests for excess relatedness in a set of individuals to allow the identification of subsets of prostate cancer cases who show the strongest evidence for excess familial clustering. The subsets identified might be argued to represent the most informative sets of cases or pedigrees to be studied for rare predisposition gene identification.

Some of the subsets of prostate cancer cases that show significant evidence of clustering in excess of expected for prostate cancer were expected, some represent new subsets of interest for genetic studies. The subset of men diagnosed with prostate cancer before age 50 years is not surprising; there is much literature suggesting a strong genetic contribution to cancer of most sites that is diagnosed early (Goldgar et al., 1994; Brandt et al., 2008) and much analysis of this subset of prostate cancer cases and pedigrees has been performed (Gronberg et al., 1999; Xu et al., 2005). However, the other two groups of prostate cancer

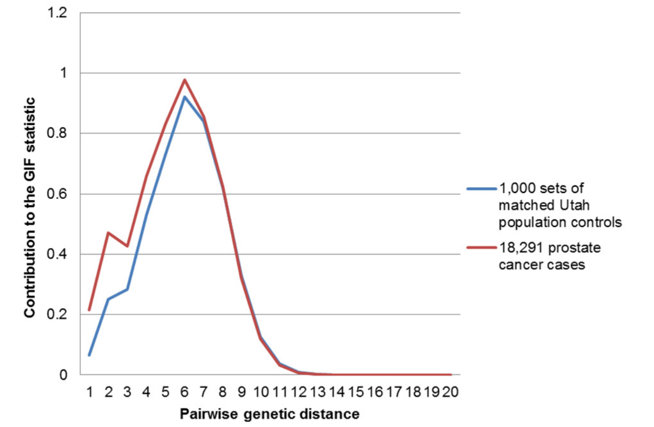


FIGURE 1 | Contribution to the GIF statistic by pairwise genetic distance for cases and controls for all prostate cancers vs. population.

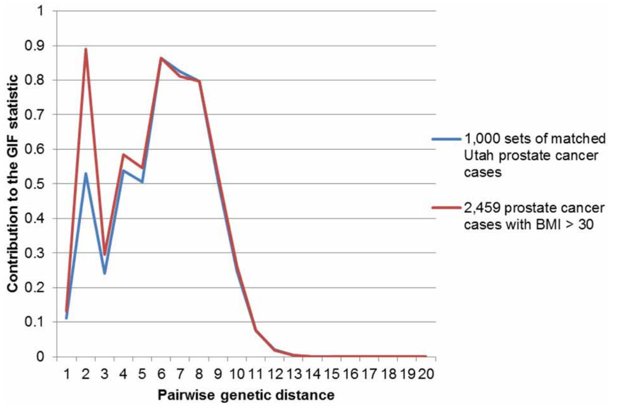


FIGURE 2 | Contribution to the GIF statistic by pairwise genetic distance for cases and controls for prostate cancer cases with a BMI of 30 or greater.

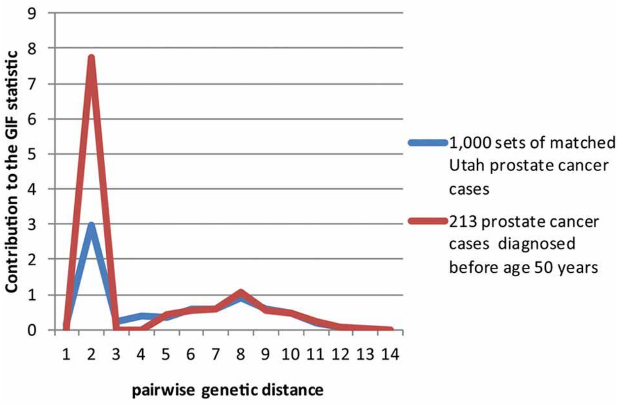


FIGURE 3 | Contribution to the GIF statistic by pairwise genetic distance for cases and controls for prostate cancer cases diagnosed before age 50.

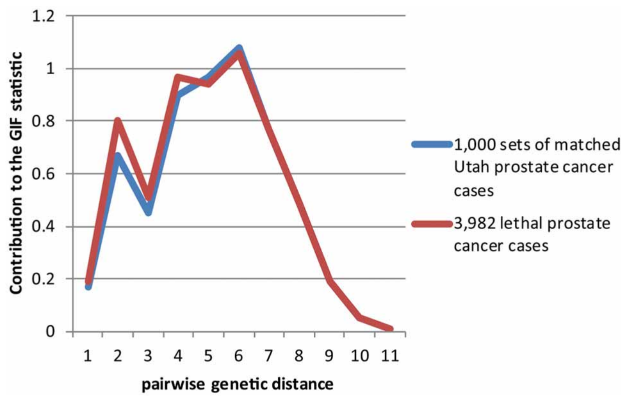


FIGURE 4 | Contribution to the GIF statistic by pairwise genetic distance for cases and controls for prostate cancer cases that have prostate cancer as a cause of death.

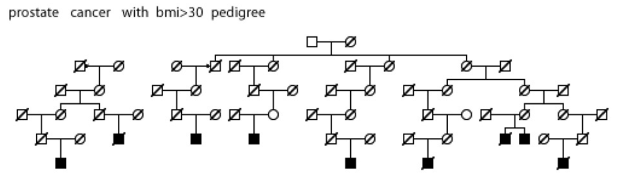


FIGURE 5 | High risk Utah prostate cancer pedigree (56 prostate cancer cases observed among descendants of the pedigree founder, 36 expected, $p = 0.001$); cases with BMI ≥ 30 are shown.

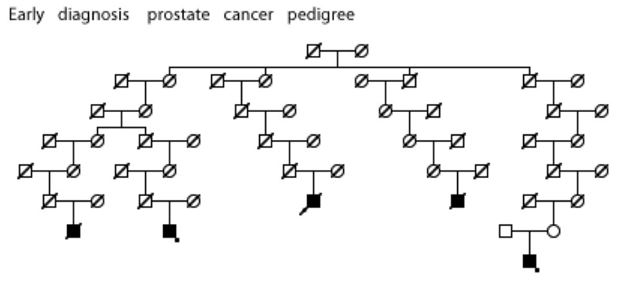


FIGURE 6 | High risk Utah prostate cancer pedigree (173 prostate cancers observed among descendants of the pedigree founder, 131 expected, $p = 0.0003$); cases diagnosed before age 50 years are shown. The two cases with an asterisk were also observed to have BMI ≥ 30 (data not available for all cases).

cases identified, high BMI (≥30) and lethal prostate cancer cases, have not been suggested previously as associated with a strong genetic contribution for prostate cancer. There was some overlap of prostate cancer cases between these sets; further investigation of specific high-risk pedigrees will determine whether they are independent.

Although epidemiologic studies have shown that systemic metabolic disorders including obesity might increase risk for prostate cancer, BMI in the context of high risk prostate cancer pedigrees does not appear to have been studied. Since there is evidence for familial clustering of high BMI or obesity (independent of cancer status), it is possible that these results are due, at least in part, to a shared predisposition to obesity. Nevertheless, these results suggest this is an informative set of pedigrees to be studied for prostate cancer risk.

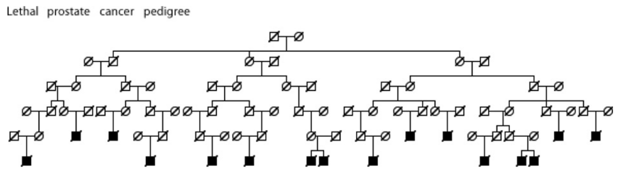


FIGURE 7 | High risk Utah prostate cancer pedigree (76 prostate cancer cases observed among descendants of the pedigree founder, 51.5 expected, $p = 0.0008$); cases known to have died from prostate cancer are shown.

The familiality of *aggressive* prostate cancer has been noted, and subsets of aggressive prostate cancer cases have been studied, without any gene identifications (Paiss et al., 2003; Lange et al., 2006; Schaid et al., 2006; Christensen et al., 2007). Little progress has been made in understanding why 30% of all patients with localized prostate cancer eventually develop recurrent, and subsequently fatal, prostate cancer. Rather than subset aggressive prostate cancers, we specifically targeted the pathogenesis of lethal prostate cancer. This subtle definition difference focuses on the subtype of prostate cancer which is associated with the worst prognosis i.e., which kills, but our definition ignores age at onset and pathology grading data for the individual, both of which are more commonly used to classify prostate cancer cases for aggressive status, but which can be poor markers for survival. This subset of lethal prostate cancer cases, among all others, is the most clinically significant and that which could yield the most translational opportunities were genes to be identified.

The Utah population has proven valuable to the study of many common cancers, and to the isolation of multiple cancer predisposition genes. The University of Utah group has been studying high-risk cancer pedigrees since 1972, and has built a resource of thousands of extended high-risk pedigrees that includes over 35,000 DNA samples. The study of extended pedigrees allowed our research group to isolate *BRCA1* (Miki et al., 1994), to localize and isolate *BRCA2* (Wooster et al., 1994; Tavtigian et al., 1996), to localize and isolate *p16* (Cannon-Albright et al., 1992, 1994; Kamb et al., 1994), and to localize and isolate *HPC2/ELAC2* (Tavtigian et al., 2001). These findings of excess relatedness in the UPDB for three subsets of prostate cancer cases represent multiple Utah high-risk prostate cancer pedigrees for each of the subsets. Analysis of these high risk pedigrees will lead to identification of the predisposition genes responsible, which might otherwise not be identifiable in studies of all high-risk prostate cancer pedigrees combined.

We have identified significant evidence for three characteristics of prostate cancer that independently coaggregate in both close and distant relatives. We have identified multiple high-risk prostate cancer pedigrees that independently include multiple prostate cancer cases with the characteristics of interest.

Figures 5–7 show an example Utah high-risk prostate cancer pedigree for each of the three characteristics identified. We propose that linkage analysis or shared genomic segment (Thomas et al., 2008) analysis can identify chromosomal regions shared in the related cases and that sequence analysis of predisposition carriers in the targeted regions located will lead to identification of the responsible predisposition genes. Rather than studying all high-risk prostate cancer pedigrees, we instead will focus on those that exhibit multiple cases with those characteristics most likely to have a genetic contribution. These studies will examine fewer pedigrees than a typical prostate cancer pedigree study, but will focus on the homogeneous subsets most likely to represent rare segregating predisposition genes or variants.

These findings should be generalizable to the U.S.A. population. Utah was originally settled by ~10,000 Mormons of British, Scandinavian, and German origin. They, and the more than 50,000 migrants from the same areas who arrived in the next generations, have typical Northern European gene frequencies (McLellan et al., 1984) and low to normal levels of inbreeding compared to the U.S. (Jorde, 1989). These characteristics make this population appropriate for inferences in populations of Northern European descent. The predisposition genes identified in Utah are represented similarly in other studies in terms of frequency, penetrance, and interactions with risk factors and modifier genes. Utah cancer rates are lower than U.S. rates, most likely due to lower rates of smoking and alcohol use.

Recent advances in mapping the genome, combined with the unique resources of Utah, provide a rare opportunity for a successful search for predisposition genes or variants for prostate cancer and the definition of their role at a population level. Recent evidence has shown the advisability and efficiency of rare predisposition gene identification by study of extended pedigrees (Ewing et al., 2012; Roberts et al., 2012). Here we identify characteristics of prostate cancer that can be used to more specifically focus gene identification efforts on appropriate pedigrees. The eventual identification of predisposition genes for prostate cancer, accompanied by a greater understanding of how these genes contribute to morbidity and mortality, will lead to the development of diagnostic tests and more personalized treatments for prostate cancer.

REFERENCES

Albright, F., Teerlink, C., Werner, T. L., and Cannon-Albright, L. A. (2012). Significant evidence for a heritable contribution to cancer predisposition: a review of cancer familiality by site. *BMC Cancer* 12:138. doi: 10.1186/1471–2407-12-138

American Cancer Society. (2013). *Cancer Facts and Figures 2013*. Atlanta, GA: American Cancer Society.

Brandt, A., Bermejo, J. L., Sundquist, J., and Hemminki, K. (2008). Age of onse in familial cancer. *Ann. Oncol.* 19, 2084–2088. doi: 10.1093/annonc/mdn527

Cannon, L., Bishop, D. T., Skolnick, M. H., Hunt, S., Lyon, J. L., and Smart, C. (1982). Genetic epidemiology of prostate cancer in the Utah Mormon genealogy. *Cancer Surv.* 1, 48–69.

Cannon-Albright, L. A. (2008). Utah family-based analysis: past, present and future. *Hum. Hered.* 65, 209–220. doi: 10.1159/000112368

Cannon-Albright, L. A., Goldgar, D. E., Meyer, L. J., Lewis, C. M., Anderson, D. E., Fountain, et al. (1992). Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* 258, 1148–1152. doi: 10.1126/science.1439824

Cannon-Albright, L. A., Goldgar, D. E., Neuhausen, S., Gruis, N. A., Anderson, D. E., Lewis, C. M., et al. (1994). Localization of the 9p melanoma susceptibility locus (MLM) to a 2-cM region between D9S736 and D9S171. *Genomics* 23, 265–268. doi: 10.1006/geno.1994.1491

Cannon-Albright, L. A., Schwab, A., Camp, N. J., Farnham, J. S., and Thomas, A. (2005). Population-based risk assessment for other cancers in relatives of hereditary prostate cancer (HPC) Cases. *Prostate* 64, 347–355. doi: 10.1002/pros.20248

Carter, B. S., Bova, G. S., Beaty, T. H., Steinberg, G. D., Childs, B., Isaacs, W. B., et al. (1993). Hereditary prostate cancer: epidemiologic and clinical features. *J. Urol.* 150, 797–802.

Christensen, G. G., Camp, N. J., Farnham, J. M., and Cannon-Albright, L. A. (2007). Genome-wide linkage analysis for aggressive prostate cancer in Utah high-risk pedigrees. *Prostate* 67, 605–613. doi: 10.1002/pros.20554

Ewing, C. M., Ray, A. M., Lange, E. M., Zuhlke, K. A., Robbins, C. M., Tembe, W. D., et al. (2012). Germline mutations in HOXB13 and prostate-cancer risk. *N. Engl. J. Med.* 366, 141–149. doi: 10.1056/NEJMoa1110000

Goldgar, D. E., Easont, D. F., Cannon-Albright, L. A., and Skolnick, M. H. (1994). Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J. Natl. Cancer Inst.* 86, 1600–1608. doi: 10.1093/jnci/86.21.1600

Gronberg, H., Smith, J., Emanuelsson, M., Jonsson, B. A., Bergh, A., Carpten, J., et al. (1999). In Swedish families with hereditary prostate ccancer,linkage to the HPC1 locus on chromosome 1q24-25 is restricted to families with early-onset prostate cancer. *Am. J. Hum. Genet.* 65, 134–140. doi: 10.1086/302447

Jorde, L. B. (1989). Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann. Hum. Genet.* 53, 339–355.

Kamb, A., Shattuck-Eidens, D., Eeles, R., Liu, Q., Gris, N. A., Ding, W., et al. (1994). Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat. Genet.* 8, 23–26. doi: 10.1038/ng0994-22

Lange, E. M., Ho, L. A., Beebe-Dimmer, J. L., Wang, Y., Gillanders, E. M., Trent, J. M., et al. (2006). Genome-wide linkage scan for prostate cancer susceptibility genes in men with aggressive disease: significant evidence for linkage at chromosome 15q12. *Hum. Genet.* 119, 400–407.

Langeberg, W. J., Isaacs, W. B., and Stanford, J. L. (2007). Genetic etiology of hereditary prostate cancer. *Front. Biosci.* 12, 4101–4110.

Larson, A. A., Leachman, S. A., Eliason, M. J., and Cannon-Albright, L. A. (2006). Population-based assessment of non-melanoma cancer risk in relatives of cutaneous melanoma probands. *J. Invest. Dermatol.* 127, 183–188. doi: 10.1038/sj.jid.5700507

McLellan, T., Jorde, L. B., and Skolnick, M. H. (1984) Genetic distances between the Utah Mormons and related populations. *Am. J. Hum. Genet.* 36, 836–857.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA 1. *Science* 266, 66–71. doi: 10.1126/science.7545954

Paiss, T., Worner, S., Kurtz, F., Haeussler, J., Hautmann, R. E., Gschwend, J. E., et al. (2003). Linakge of aggressive prostate cancer to chromosome 7q31-33 in German prostate cancer families. *Eur. J. Hum. Genet.*11, 17–22. doi: 10.1038/sj.ejhg.5200898

Roberts, N. J., Jiao, Y., Yu, J., Kopelovich, L., Petersen, G. M., Bondy, M. L., et al. (2012). ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discov.* 2, 41–46. doi: 10.1158/2159-8290.CD-11-0194

Schaid, D. J., McDonnell, S. K., Zarfas, K. E., Cunningham, J. M., Hebbring, S., Thibodeau, S. N., et al. (2006). Pooled genome linkage scan of aggressive prostate cancer: results from the international consortium for prostate cancer genetics. *Hum. Genet.* 120, 471–485. doi: 10.1007/s00439-006-0219-9

Skolnick, M. H. (1980). "The Utah geneological data base: a resource for genetic epidemiology," in *Banbury Report 4: Cancer Incidence in Defined Populations*, eds J. Cairns, J. H. Lyon, and M. H. Skolnick (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory), 285–297.

Stanford, J. L., and Ostrander, E. A. (2001). Familial prostate cancer. *Epidemiol. Rev.* 23, 19–23, doi: 10.1093/oxfordjournals.epirev.a000789

Tavtigian, S. V., Simard, J., Rommens, J., Couch, F., Shattuck-Eidens, D., Neuhausen, S., et al. (1996). The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat. Genet.* 12, 333–337. doi: 10.1038/ng0396-333

Tavtigian, S. V., Simard, J., Teng, D. H., Abtin, V., Baumgard, M., Beck, A., et al. (2001). A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat. Genet.* 27, 172–180. doi: 10.1038/84808

Teerlink, C. C., Albright, F. S., Lins, L., and Cannon-Albright, L. A. (2012). A comprehensive survey of cancer risks in extended families. *Genet. Med.* 14, 107–114. doi: 10.1038/gim.2011.2

Thomas, A., Camp, N. J., Farnham, J. M., Allen-Brady, K., and Cannon-Albright, L. A. (2008). Shared genomic segment analysis. mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann. Hum. Genet* 72(Pt 2), 279–287.

Wooster, R., Neuhausen, S., Mangion, J., Quirk, Y., Ford, D., Collins, N., et al. (1994). Localization of a breast cancer susceptibility gene (BRCA2) to chromosome 13q12-13. *Science* 265, 2088–2090. doi: 10.1126/science.8091231

Xu, J., Dimitrov, L., Chang, B. L., Adams, T. S., Turner, A. R., Meyers, D. A., et al. (2005). A Combined genomewide linkage scan of 1,233 families for prostate cancer-susceptibility genes conducted by the International Consortium for porstate cancer genetics. *Am. J. Hum. Genet* 77, 219–229.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 April 2013; accepted: 22 July 2013; published online: 20 August 2013.

Citation: Nelson Q, Agarwal N, Stephenson R and Cannon-Albright LA (2013) A population-based analysis of clustering identifies a strong genetic contribution to lethal prostate cancer. Front. Genet. ***4****:152. doi: 10.3389/fgene.2013.00152*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 25 September 2013
doi: 10.3389/fgene.2013.00189

PedWiz: a web-based tool for pedigree informatics

Yeunjoo E. Song and Robert C. Elston*

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Edited by:
Karen T. Cuenco, University of Pittsburgh, USA

Reviewed by:
Lijun Ma, Wake Forest University Health Sciences, USA
Thomas A. Sellers, H. Lee Moffitt Cancer Center & Research Institute, USA

Correspondence:
Robert C. Elston, Case Western Reserve University, Department of Epidemiology and Biostatistics, Wolstein Research Building, 2013 Cornell Road, Cleveland, OH, 44106, USA
e-mail: robert.elston@cwru.edu

A novel web-based tool PedWiz that pipelines the informatics process for pedigree data is introduced. PedWiz is designed to assist researchers in the analysis of pedigree data. It provides a convenient tool for pedigree informatics: descriptive statistics, relative pairs, genetic similarity coefficients, the variance-covariance matrix for three estimated coefficients of allele identical-by-descent sharing as well as mean allele sharing, a plot of the pedigree structures, and a visualization of the identity coefficients. With a renewed interest in linkage and other family based methods, PedWiz will be a valuable tool for the analysis of family data.

Keywords: pedigree, informatics, genetic similarity, identity-by-descent, relative pairs, family data

INTRODUCTION

When a researcher has collected or is provided with a set of nuclear family or extended pedigree data for genetic analysis, the first thing that needs to be done is to find out what information is available on the family or families before proceeding in the analysis of phenotype and/or genotype data to study the characteristics of a certain disease or trait, i.e., pedigree informatics. This can include descriptive statistics, visualization of family data, the degree of genetic relatedness among members of a family, and so on.

Descriptive statistics summarize and provide basic information on the family data, as done in the PEDINFO program in S.A.G.E. (2012). The visualization of family data is a fundamental task for both family studies and genetic counseling. There are many computer programs available that provide the graphical representation of pedigree data, including the R packages *kinship* (Zhao, 2006) and *pedantics* (Morrissey, 2010). The concept of genetic relatedness is essential in modern genetic analysis, and the applications of kinship and condensed identity coefficients are everywhere in analyses that have a genetic component. In human genetics, they are used in genotype prediction, calculation of genetic risk ratios for binary disease status, calculations of correlations between relatives, and robust linkage analysis. Robust linkage analysis, a powerful approach to map disease genes, is based on comparing the genetic marker profiles, i.e., allele identical-by-descent (IBD) sharing, of pairs of relatives. There are many software programs that calculate kinship and inbreeding coefficients, but not many for the nine condensed coefficients of IBD sharing.

A brief survey of available R packages with their relevant components of pedigree informatics is shown in **Table 1**. As can be seen, there is no program that provides all the different genetic similarity measurements together with the variance-covariance matrix of the estimated coefficients of IBD. Abney (2009)'s graphical algorithm for the computation of the generalized kinship coefficients is implemented in *idcoefs2* (written in C++, and implemented as the R package *identity*), and this is the only currently available program that outputs the nine condensed coefficients of IBD. The R package *ibdreg* by Schaid et al. (2007) has two functions, *sim.ibd.var* and *exact.ibd.var*, to calculate the variance-covariance of mean allele sharing, but not the variance-covariance of the individual coefficients of IBD. An essential part of score tests is the choice of the denominator variance, and some of these tests for genetic linkage require the variance-covariance of allele IBD sharing statistics under the null, i.e., of the coefficients of IBD. It would be useful to make available the variance-covariance matrix of these coefficients for a pedigree independent of the choice of test statistics, so that it can be used for different choices of test statistics. Currently, no such tools are available.

PedWiz (**Ped**igree Informatics **Wiz**ard) is designed to fulfill this need as a web-based tool for pedigree informatics, to assist researchers in the analysis of pedigree data. It provides a convenient "one-stop-shop" for pedigree informatics. It provides all the genetic similarity coefficients mentioned above, including the nine condensed coefficients of IBD and the variance-covariance matrix of the one-locus three marginal coefficients of allele IBD sharing, as well as other pedigree descriptive statistics. Additionally, it provides a plot of the pedigree structure and a visualization of the identity coefficients, something that no other program provides. PedWiz is an automated pipeline for extracting pedigree informatics before conducting specialized analyses of phenotype and/or genotype data.

MATERIAL AND METHODS

IMPLEMENTATION

The web interface of PedWiz is implemented using a combination of XHTML (eXtensible HyperText Markup Language), CSS (Cascading Style Sheets), and PHP (Hypertext Preprocessor) on an Apache web server. The interactivity is provided by JavaScript and Ajax technologies. Custom Python modules handle the overall

Table 1 | R packages available for pedigree informatics.

Name	Plot	Stat	*F*	Φ	Δ	VC(2Φ)	VC(Δ)	Simulation
adegenet			√					
gap	√			√				
geneland			√					
ibdreg						√		
identity					√			
kinship	√			√				
pedantics	√	√	√					√
pedigree			√					
pedigreemm			√					
GeneticsPed			√					

Plot, pedigree plot; stat, descriptive statistics; F, inbreeding coefficient; Φ, kinship coefficient; Δ, 9 condensed IBD coefficients; VC(2Φ), variance-covariance matrix of mean allele sharing; VC(Δ), variance-covariance matrix of 3 IBD coefficients.

flow of the pipeline by calling pre-existing programs written in C++ or R.

USER INPUT

PedWiz accepts a plain ASCII text file format for pedigree input. Since PedWiz extracts the information contained in a pedigree structure, it requires a pedigree file to have five essential columns: pedigree ID, individual ID, the two parents' IDs and sex. These five columns do not need be in any specific order, nor need they be consecutive. If a pedigree file contains other columns, they are ignored. The pedigree file is required to be in either tab-delimited or comma-delimited format. It may optionally contain a header line specifying the names of the columns. The user inputs configuration information and the location of the pedigree file through a user-friendly interface, and then submits it to start the analysis pipeline.

ANALYSIS TOOLS

Once the user submits a pedigree file and configuration information, the informatics process starts by running the first tool. Currently, the PedWiz process consists of six main tools (**Figure 1**). The complete process utilizes many internal Python scripts (which are not detailed here) to create junctions between the programs (format compatibility) and to create the necessary R scripts.

The descriptive statistics tool

This tool is used to calculate the descriptive statistics for each pedigree contained in the user-submitted pedigree file. PedWiz utilizes the existing C++ program PEDINFO of the S.A.G.E. package (v6.3 with *each_pedigree* = *true* option). PEDINFO provides many useful descriptive statistics on pedigree data including means, standard deviations; family, sibship and pedigree sizes; and counts of each type of relative pair. The results are parsed and reported to the user by PedWiz as a table on the website. From here, the user selects a pedigree to proceed with other tools.

The pedigree plot tool

This tool is used to visualize a pedigree. PedWiz utilizes the R package *kinship* to generate the plot (Zhao, 2005). As in a typical pedigree diagram, males and females are shown as squares and circles, respectively. The resulting pedigree plot is reported to the user as a pdf file on the website.

The relative pairs tool

This tool is used to report all relative pairs existing in a pedigree. PedWiz uses an internal C++ program that finds all existing relative pairs by traversing the pedigree structure recursively as done in the FCOR program in S.A.G.E. (2012). The results are reported to the user on the website as a text file containing the relative pair matrix and the list of relative pairs for each relative type.

The genetic similarity tool

This tool is used to provide the various genetic similarity coefficients. PedWiz uses an internal C++ program to perform this task. The results include two matrices; one is the matrix of kinship/inbreeding coefficients (inbreeding coefficients on the diagonal and kinship coefficients off the diagonal), and the other is the matrix of nine condensed coefficients of IBD. The coefficients of relationship, which are twice the kinship coefficients, can be easily calculated from the kinship/inbreeding coefficients. The resulting matrices are reported to the user on the website as a text file.

The visualization of genetic similarity tool

This tool is used to visualize the two matrices generated by the genetic similarity tool. PedWiz uses a custom R script to represent a matrix graphically as a heat map. The resulting heat maps are reported to the user as a pdf file on the website.

The variance-covariance of genetic similarity tool

This tool is used to find the variance-covariance matrix of the coefficients reported by the genetic similarity tool. PedWiz uses an internal C++ program to perform this task. The variance-covariance matrix of kinship coefficients is estimated by an exact method given by Chen and Abecasis (2006). The variance-covariance matrix of IBD coefficients is estimated by a simulation method, given a pedigree structure (MacCluer et al., 1986), based on 500 simulation replicates. The simulation method approximates the distribution of IBD states by gene dropping, so it can be used regardless of pedigree size and structure. The results are reported to the user on the website as a text file.

RESULTS

We developed a novel web-based tool that pipelines the informatics process for pedigree data. PedWiz may be accessed at http://darwin.cwru.edu/∼song/pedwiz. Here we present an application example using pedigree data from the Madeline 2.0 website (Trager et al., 2007). These pedigree data contain a consanguineous marriage between cousins. The user inputs configuration information and the location of the pedigree file through the interface on the website as shown in **Figure 2**.

After configuration information and the location of the pedigree file have been submitted by the user, PedWiz produces a table with the descriptive statistics for each pedigree on the website as shown in **Figure 3**. All results are accessed through a set of buttons under the descriptive statistics table for each pedigree. The user uses a radio button to select a pedigree for an analysis pipeline.

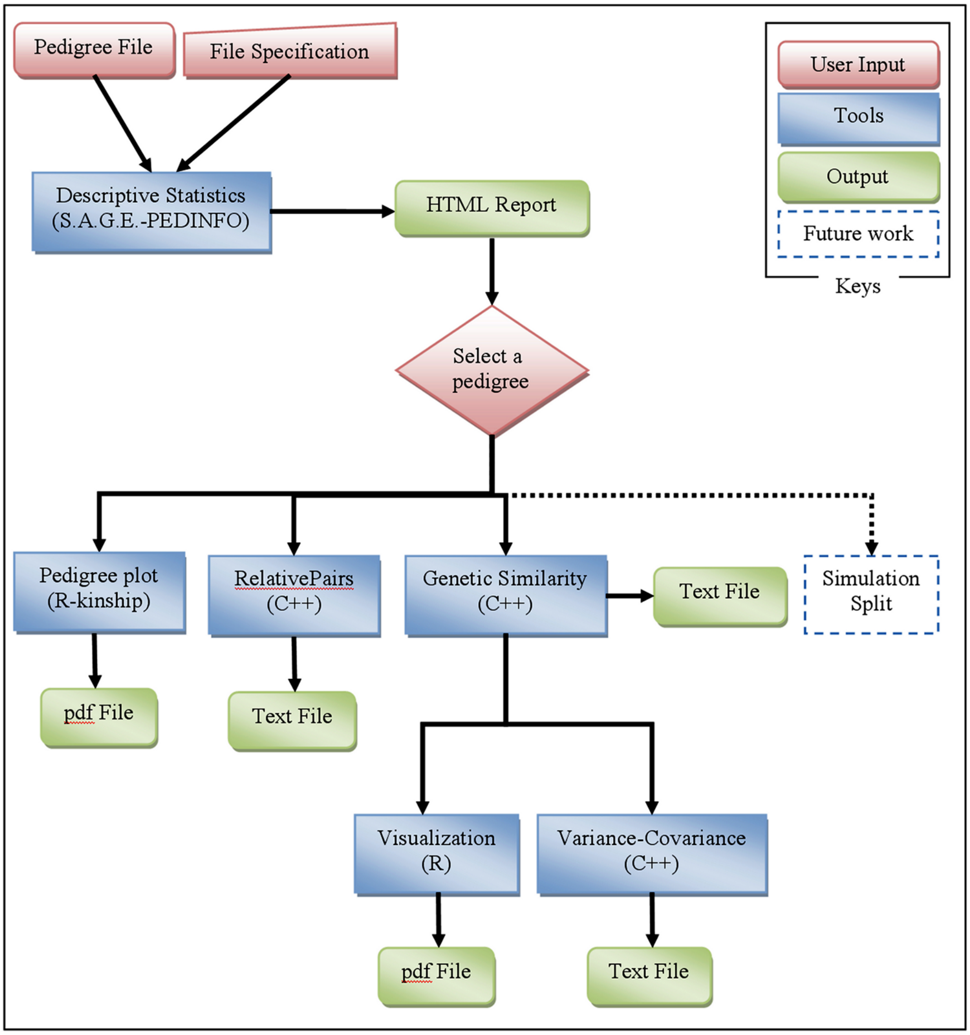


FIGURE 1 | PedWiz overview. This figure illustrates the analysis pipeline implemented in PedWiz. It consists of six tools to mine the information in a pedigree structure: descriptive statistics, pedigree plot, relative pairs, genetic similarity coefficients, visualization of genetic similarity coefficients, and the variance-covariance matrix of coefficients of IBD. The tools denoted by dotted lines are anticipated future extensions.

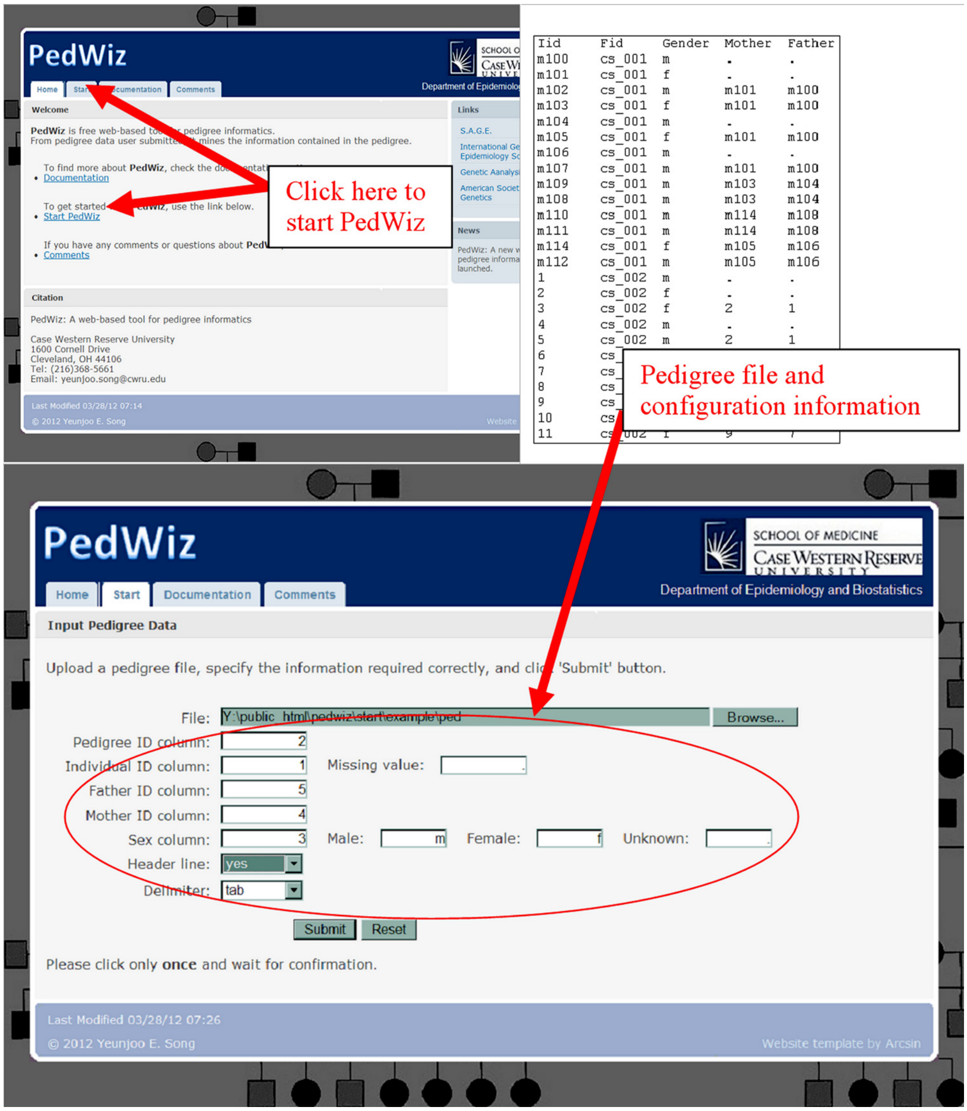
This selection information is reflected under the table (shown in the green eclipse). The resulting output from each tool for the example pedigree is shown also.

DISCUSSION

We developed a novel web-based tool PedWiz that pipelines the informatics process for pedigree data. PedWiz is designed to assist researchers in the analysis of pedigree data. It provides a convenient tool for pedigree informatics: descriptive statistics, relative pairs, genetic similarity coefficients, the variance-covariance matrix of three coefficients of allele IBD sharing, as well as mean allele sharing, a plot of the pedigree structure, and visualization of identity coefficients. PedWiz is an automated pipeline for extracting pedigree informatics before conducting specialized analysis of phenotype and/or genotype data.

Emerging availability of whole genome sequence data has led to a renewed interest in linkage and other family based methods (Ott et al., 2011). Many researchers have been emphasizing the importance and advantages of family studies all along (Clerget-Darpoux and Elston, 2007; Stein and Elston, 2009), especially to interpret next generation sequence data (Bailey-Wilson and Wilson, 2011; Wijsman, 2012). Family study designs provide not only the enrichment of genetic loci containing rare variants, but also methods to control for genetic heterogeneity and population stratification. PedWiz is a valuable tool for initial analysis of those family data.

Additionally, the results from each tool in Pedwiz will be useful for later analysis of phenotype and/or genotype data. As stated before, an essential part of score tests is the choice of the denominator variance, and some of these tests for genetic linkage require the variance-covariance of the coefficients of IBD. No software tools are currently available to provide this information independent of the choice of test statistics. The variance-covariance of the genetic similarity tool of PedWiz provides this need, so that it


PedWiz
Home
Start
Documentation
Comments
SCHOOL OF MEDICINE
CASE WESTERN RESERVE UNIVERSITY
Department of Epidemiology and Biostatistics
Links
S.A.G.E.
International Genetic Epidemiology Society
Genetic Aanalysis
American Society of Human Genetics
News
Browse...
Submit
Reset

Click here to start PedWiz

Welcome

PedWiz is free web-based tool for pedigree informatics.
From pedigree data user submitted, it mines the information contained in the pedigree.

To find more about **PedWiz**, check the documentation.
- Documentation

To get started with PedWiz, use the link below.
- Start PedWiz

If you have any comments or questions about PedWiz,
- Comments

PedWiz: A new web-based tool for pedigree informatics launched.

Citation

PedWiz: A web-based tool for pedigree informatics

Case Western Reserve University
1600 Cornell Drive
Cleveland, OH 44106
Tel: (216)368-5661
Email: yeunjoo.song@cwru.edu

Last Modified 03/28/12 07:14
© 2012 Yeunjoo E. Song

Iid	Fid	Gender	Mother	Father
m100	cs_001	m	.	.
m101	cs_001	f	.	.
m102	cs_001	m	m101	m100
m103	cs_001	f	m101	m100
m104	cs_001	m	.	.
m105	cs_001	f	m101	m100
m106	cs_001	m	.	.
m107	cs_001	m	m101	m100
m109	cs_001	m	m103	m104
m108	cs_001	m	m103	m104
m110	cs_001	m	m114	m108
m111	cs_001	m	m114	m108
m114	cs_001	f	m105	m106
m112	cs_001	m	m105	m106
1	cs_002	m	.	.
2	cs_002	f	.	.
3	cs_002	f	2	1
4	cs_002	m	.	.
5	cs_002	m	2	1
6	cs			
7	cs			
8	cs			
9	cs			
10	cs			
11	cs_002	f	9	7

Pedigree file and configuration information

Input Pedigree Data

Upload a pedigree file, specify the information required correctly, and click 'Submit' button.

File: Y:\public_html\pedwiz\start\example\ped
Pedigree ID column: 2
Individual ID column: 1 Missing value: .
Father ID column: 5
Mother ID column: 4
Sex column: 3 Male: m Female: f Unknown: .
Header line: yes
Delimiter: tab

Please click only **once** and wait for confirmation.

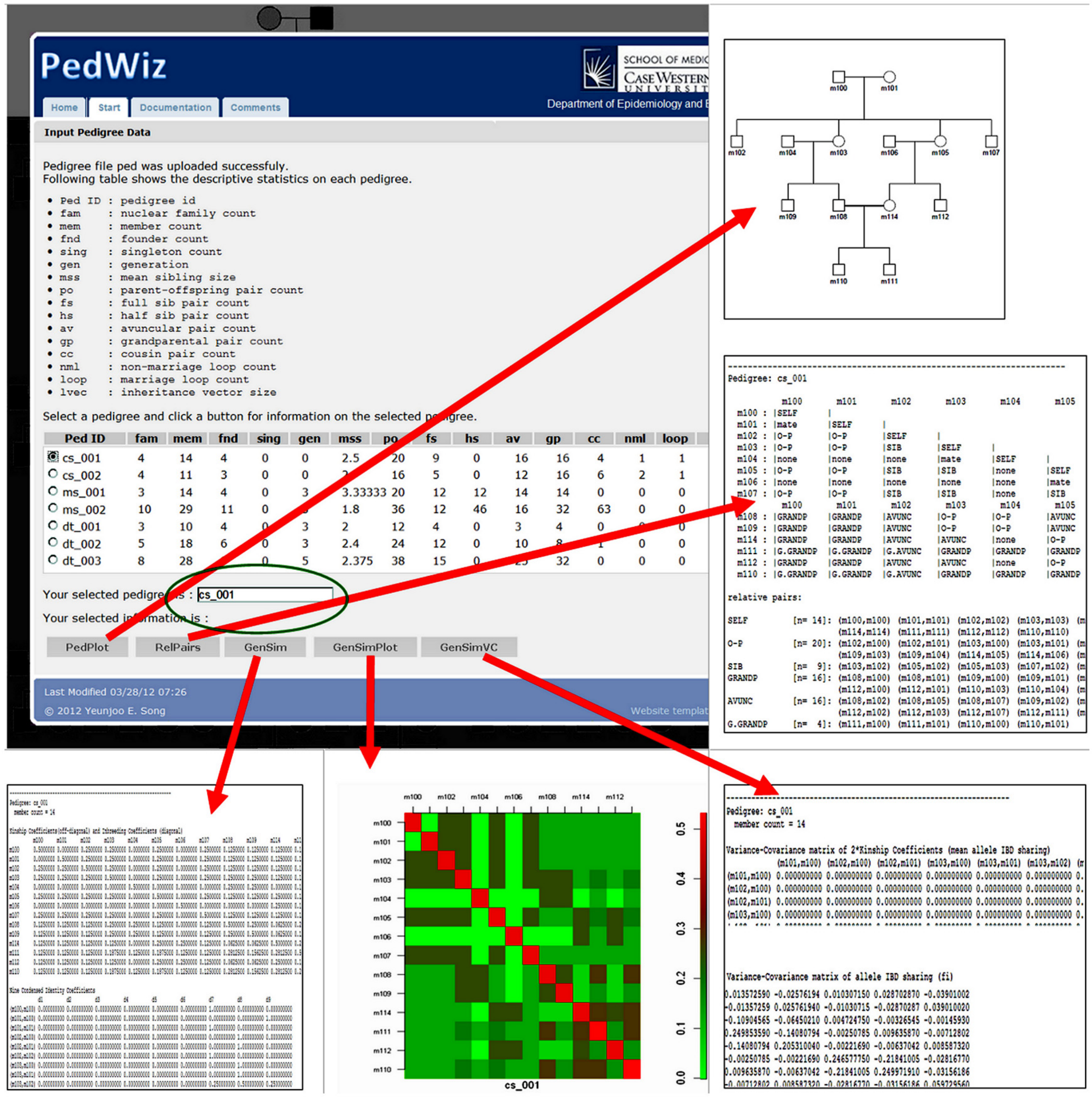
Last Modified 03/28/12 07:26
© 2012 Yeunjoo E. Song
Website template by Arcsin

FIGURE 2 | Starting PedWiz. This figure illustrates the user interface to start PedWiz.

can be used for different choices of test statistics. The information from the genetic similarity tool of PedWiz can be used for weighting pedigrees of different sizes. Another potential use of this tool is for selecting families with the most information in terms of genetic relatedness that would best suit a phenotype/genotype analysis of choice. Selecting families with multiple affected subjects, or families with extreme values, is known to provide improved ability to measure, and detect, the effects of rare variants (Ionita-Laza and Ottman, 2011; Wijsman, 2012). The strategy of selecting "large linked families" for initial screening has long been a successful strategy (Bowden et al., 2010). To be successful with this approach, selecting families with a real linkage signal in specific regions is essential. This new tool will be useful for selecting such families when used together with phenotype/genotype information.

With a modular design, each analysis tool within PedWiz is independent of the others, so it is very easy to extend and add more tools. Planned additions in the near future are simulation and pedigree split tools, shown in **Figure 1** with dotted lines. PedWiz is currently specialized to deal with the information contained within pedigree structures only. Therefore, it is very fast and safe with regard to data transfer over the web. However, it is always possible to add more pipeline modules that could process the information from phenotype and/or genotype data. Good candidates for this addition would be simulation conditional on given phenotype and/or genotype data, and imputation. Another extension that could be added on is the inclusion of a backend database to save data and results for reuse.

The genetic similarity tool of PedWiz is specifically designed to provide the information on within-pedigree relatedness. As a reviewr pointed out, a tool that addresses between-pedigree relatedness, especially for pedigrees from a relatively isolated population like the Hutterites, would be a useful addition to



PedWiz

Home | Start | Documentation | Comments

SCHOOL OF MEDIC
CASE WESTERN
UNIVERSIT
Department of Epidemiology and B

Input Pedigree Data

Pedigree file ped was uploaded successfully.
Following table shows the descriptive statistics on each pedigree.

- Ped ID : pedigree id
- fam : nuclear family count
- mem : member count
- fnd : founder count
- sing : singleton count
- gen : generation
- mss : mean sibling size
- po : parent-offspring pair count
- fs : full sib pair count
- hs : half sib pair count
- av : avuncular pair count
- gp : grandparental pair count
- cc : cousin pair count
- nml : non-marriage loop count
- loop : marriage loop count
- lvec : inheritance vector size

Select a pedigree and click a button for information on the selected pedigree.

Ped ID	fam	mem	fnd	sing	gen	mss	po	fs	hs	av	gp	cc	nml	loop
● cs_001	4	14	4	0	0	2.5	20	9	0	16	16	4	1	1
○ cs_002	4	11	3	0	0	2	16	5	0	12	16	6	2	1
○ ms_001	3	14	4	0	3	3.33333	20	12	12	14	14	0	0	0
○ ms_002	10	29	11	0	3	1.8	36	12	46	16	32	63	0	0
○ dt_001	3	10	4	0	3	2	12	4	0	3	4	0	0	0
○ dt_002	5	18	6	0	3	2.4	24	12	0	10	8	1	0	0
○ dt_003	8	28	[illegible]	0	5	2.375	38	15	0	23	32	0	0	0

Your selected pedigree is : cs_001

Your selected information is :

Last Modified 03/28/12 07:26

© 2012 Yeunjoo E. Song

Website templat

```
--------------------------------------------------------------------
Pedigree: cs_001

          m100      m101      m102      m103      m104      m105
  m100 : |SELF     |
  m101 : |mate     |SELF     |
  m102 : |O-P      |O-P      |SELF     |
  m103 : |O-P      |O-P      |SIB      |SELF     |
  m104 : |none     |none     |none     |mate     |SELF     |
  m105 : |O-P      |O-P      |SIB      |SIB      |none     |SELF
  m106 : |none     |none     |none     |none     |none     |mate
  m107 : |O-P      |O-P      |SIB      |SIB      |none     |SIB
          m100      m101      m102      m103      m104      m105
  m108 : |GRANDP   |GRANDP   |AVUNC    |O-P      |O-P      |AVUNC
  m109 : |GRANDP   |GRANDP   |AVUNC    |O-P      |O-P      |AVUNC
  m114 : |GRANDP   |GRANDP   |AVUNC    |AVUNC    |none     |O-P
  m111 : |G.GRANDP |G.GRANDP |G.AVUNC  |GRANDP   |GRANDP   |GRANDP
  m112 : |GRANDP   |GRANDP   |AVUNC    |AVUNC    |none     |O-P
  m110 : |G.GRANDP |G.GRANDP |G.AVUNC  |GRANDP   |GRANDP   |GRANDP

relative pairs:

SELF        [n= 14]: (m100,m100) (m101,m101) (m102,m102) (m103,m103) (m
                     (m114,m114) (m111,m111) (m112,m112) (m110,m110)
O-P         [n= 20]: (m102,m100) (m102,m101) (m103,m100) (m103,m101) (m
                     (m109,m103) (m109,m104) (m114,m105) (m114,m106) (m
SIB         [n=  9]: (m103,m102) (m105,m102) (m105,m103) (m107,m102) (m
GRANDP      [n= 16]: (m108,m100) (m108,m101) (m109,m100) (m109,m101) (m
                     (m112,m100) (m112,m101) (m110,m103) (m110,m104) (m
AVUNC       [n= 16]: (m108,m102) (m108,m105) (m108,m107) (m109,m102) (m
                     (m112,m102) (m112,m103) (m112,m107) (m112,m111) (m
G.GRANDP    [n=  4]: (m111,m100) (m111,m101) (m110,m100) (m110,m101)
```

```
-------------------------------------------------------
Pedigree: cs_001
  member count = 14

Kinship Coefficients(off-diagonal) and Inbreeding Coefficients (diagonal)
       m100      m101      m102      m103      m104      m105      m106      m107      m108      m109      m114
m100   0.5000000 0.0000000 0.2500000 0.2500000 0.0000000 0.2500000 0.0000000 0.2500000 0.1250000 0.1250000 0.1250000
m101   0.0000000 0.5000000 0.2500000 0.2500000 0.0000000 0.2500000 0.0000000 0.2500000 0.1250000 0.1250000 0.1250000
m102   0.2500000 0.2500000 0.5000000 0.2500000 0.0000000 0.2500000 0.0000000 0.2500000 0.1250000 0.1250000 0.1250000
m103   0.2500000 0.2500000 0.2500000 0.5000000 0.0000000 0.2500000 0.0000000 0.2500000 0.2500000 0.2500000 0.1250000
m104   0.0000000 0.0000000 0.0000000 0.0000000 0.5000000 0.0000000 0.0000000 0.0000000 0.2500000 0.2500000 0.0000000
m105   0.2500000 0.2500000 0.2500000 0.2500000 0.0000000 0.5000000 0.0000000 0.2500000 0.1250000 0.1250000 0.2500000
m106   0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.5000000 0.0000000 0.0000000 0.0000000 0.2500000
m107   0.2500000 0.2500000 0.2500000 0.2500000 0.0000000 0.2500000 0.0000000 0.5000000 0.1250000 0.1250000 0.1250000
m108   0.1250000 0.1250000 0.1250000 0.2500000 0.2500000 0.1250000 0.0000000 0.1250000 0.5000000 0.2500000 0.0625000
m109   0.1250000 0.1250000 0.1250000 0.2500000 0.2500000 0.1250000 0.0000000 0.1250000 0.2500000 0.5000000 0.0625000
m114   0.1250000 0.1250000 0.1250000 0.1250000 0.0000000 0.2500000 0.2500000 0.1250000 0.0625000 0.0625000 0.5000000
m111   0.1250000 0.1250000 0.1250000 0.1875000 0.1250000 0.1875000 0.1250000 0.1250000 0.2812500 0.1562500 0.2812500
m112   0.1250000 0.1250000 0.1250000 0.1250000 0.0000000 0.2500000 0.2500000 0.1250000 0.0625000 0.0625000 0.2500000
m110   0.1250000 0.1250000 0.1250000 0.1875000 0.1250000 0.1875000 0.1250000 0.1250000 0.2812500 0.1562500 0.2812500

Nine Condensed Identity Coefficients
            d1         d2         d3         d4         d5         d6         d7         d8         d9
(m100,m100) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000 0.00000000
(m101,m100) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000
(m101,m101) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000 0.00000000
(m102,m100) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000
(m102,m101) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000
(m102,m102) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000 0.00000000
(m103,m100) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000
(m103,m101) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 1.00000000 0.00000000
(m103,m102) 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.25000000 0.50000000 0.25000000
```

```
-------------------------------------------------------------
Pedigree: cs_001
  member count = 14

Variance-Covariance matrix of 2*Kinship Coefficients (mean allele IBD sharing)
            (m101,m100) (m102,m100) (m102,m101) (m103,m100) (m103,m101) (m103,m102) (m
(m101,m100) 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.
(m102,m100) 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.
(m102,m101) 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.
(m103,m100) 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.


Variance-Covariance matrix of allele IBD sharing (fi)
0.013572590 -0.02576194 0.010307150 0.028702870 -0.03901002
-0.01357259 0.025761940 -0.01030715 -0.02870287 0.039010020
-0.10904565 -0.06450210 0.004724750 -0.00326545 -0.00145930
0.249853590 -0.14080794 -0.00250785 0.009635870 -0.00712802
-0.14080794 0.205310040 -0.00221690 -0.00637042 0.008587320
-0.00250785 -0.00221690 0.246577750 -0.21841005 -0.02816770
0.009635870 -0.00637042 -0.21841005 0.249971910 -0.03156186
-0.00712802 0.008587320 -0.02816770 -0.03156186 0.059729560
```

FIGURE 3 | Different types of outputs from PedWiz. All results are accessed through a set of buttons under the descriptive statistics table for each pedigree.

PedWiz. Cryptic relatedness among unrelated individuals can be estimated by incorporating a number of dense markers across different chromosomes (Weir et al., 2006; Bink et al., 2008; Astle and Balding, 2009; Sillanpää, 2011). There are many software tools available to estimate the genome-average relatedness, for example, SPAGeDi (Hardy and Vekemans, 2002), PLINK (Purcell et al., 2007), FEST (Skare et al., 2009), CoCoa (Maenhout et al., 2009), CrypticIBDcheck (Nembot-Simo et al., 2013). Adding this to PedWiz would require an extension to process information from phenotype and/or genotype data, as mentioned above.

ACKNOWLEDGMENTS

Some of the results of this paper were obtained by using the program package S.A.G.E., which was supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources.

REFERENCES

Abney, M. (2009). A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. *Bioinformatics* 25, 1561–1563. doi: 10.1093/bioinformatics/btp185

Astle, W., and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* 24, 451–471. doi: 10.1214/09-STS307

Bailey-Wilson, J. E., and Wilson, A. F. (2011). Linkage analysis in the next generation sequencing era. *Hum. Hered.* 72, 228–236. doi: 10.1159/000334381

Bink, M. C., Anderson, A. D., van de Weg, W. E., and Thompson, E. A. (2008). Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theor. Appl. Genet.* 117, 843–855. doi: 10.1007/s00122-008-0824-1

Bowden, D. W., An, S. S., Palmer, N. D., Brown, W. M., Norris, J. M., Haffner, S. M., et al. (2010). Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. *Hum. Mol. Genet.* 19, 4112–4120. doi: 10.1093/hmg/ddq327

Chen, W. M., and Abecasis, G. R. (2006). Estimating the power of variance component linkage analysis in large pedigrees. *Genet. Epidemiol.* 30, 471–484. doi: 10.1002/gepi.20160

Clerget-Darpoux, F., and Elston, R. C. (2007). Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum. Hered.* 64, 91–96. doi: 10.1159/000101960

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Ionita-Laza, I., and Ottman, R. (2011). Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* 189, 1061–1068. doi: 10.1534/genetics.111.131813

MacCluer, J. W., Vandeburg, J. L., Read, B., and Ryder, O. A. (1986). Pedigree analysis by computer simulation. *Zoo Biol.* 5, 149–160. doi: 10.1002/zoo.1430050209

Maenhout, S., De Baets, B., and Haesaert, G. (2009). CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. *Bioinformatics* 25, 2753–2754. doi: 10.1093/bioinformatics/btp499

Morrissey, N. E. (2010). Pedantics: an R package for pedigree-based genetic simulation and pedigree manipulation, characterization and viewing. *Mol. Ecol. Resour.* 10, 711–719. doi: 10.1111/j.1755-0998.2009.02817.x

Nembot-Simo, A., Graham, J., and McNeney, B. (2013). CrypticIBDcheck: an R package for checking cryptic relatedness in nominally unrelated individuals. *Source Code Biol. Med.* 8, 5 doi: 10.1186/1751-0473-8-5

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* 12, 465–474. doi: 10.1038/nrg2989

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

S.A.G.E. (2012). *Statistical Analysis for Genetic Epidemiology, Release 6.3.* Available at: http://darwin.cwru.edu/sage/.

Schaid, D. J., Sinnwell, J. P., and Thibodeau, S. N. (2007). Testing genetic linkage with relative pairs and covariates by quasi-likelihood score statistics. *Hum. Hered.* 64, 220–233. doi: 10.1159/000103751

Sillanpää, M. J. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analysis. *Heredity* 106, 511–519. doi: 10.1038/hdy.2010.91

Skare, O., Sheehan, N., and Egeland, T. (2009). Identification of distant family relationships. *Bioinformatics* 25, 2376–2382. doi: 10.1093/bioinformatics/btp418

Stein, C. M., and Elston, R. C. (2009). Finding genes underlying human disease. *Clin. Genet.* 75, 101–106. doi: 10.1111/j.1399-0004.2008.01083.x

Trager, E. H., Khanna, R., Marrs, A., Siden, L., Branham, K. E. H., Swaoop, A., et al. (2007). Madeline 2.0 PDE: a new program for local and web-based pedigree drawing. *Bioinformatics* 23, 1854–1856. doi: 10.1093/bioinformatics/btm242

Weir, B., Anderson, A., and Hepfer, A. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960

Wijsman, E. M. (2012). The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet.* 131, 1555–1563. doi: 10.1007/s00439-012-1190-2

Zhao, J. H. (2005). Mixed-effects Coz models of alcohol dependence in extended pedigrees. *BMC Genet.* 6:S127. doi: 10.1186/1471-2156-6-S1-S127

Zhao, J. H. (2006). Pedigree-drawing with R and graphviz. *Bioinformatics* 22, 1013–1014. doi: 10.1093/bioinformatics/btl058

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 Jul 2013; accepted: 04 September 2013; published online: 25 September 2013.

Citation: Song YE and Elston RC (2013) PedWiz: a web-based tool for pedigree informatics. Front. Genet. ***4****:189. doi: 10.3389/fgene.2013.00189*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 27 September 2013
doi: 10.3389/fgene.2013.00186

Combining genetic association study designs: a GWAS case study

Janice L. Estus*[1], *Family Investigation of Nephropathy and Diabetes Research Group*[2] *and David W. Fardo*[1]

[1] Department of Biostatistics, University of Kentucky, Lexington, KY, USA
[2] Genetic Analysis and Data Coordinating Center, Family Investigation of Nephropathy and Diabetes, Case Western Reserve University, Cleveland, OH, USA

Edited by:
Robert C. Elston, Case Western Reserve University, USA

Reviewed by:
Andrew Skol, University of Chicago, USA
Claire Infante-Rivard, McGill University, Canada

****Correspondence:***
David W. Fardo, Department of Biostatistics, University of Kentucky, 205E Multidisciplinary Science Building, 725 Rose St, Lexington, KY 40536, USA
e-mail: david.fardo@uky.edu

Genome-wide association studies (GWAS) explore the relationship between genome variability and disease susceptibility with either population- or family-based data. Here, we have evaluated the utility of combining population- and family-based statistical association tests and have proposed a method for reducing the burden of multiple testing. Unrelated singleton and parent-offspring trio cases and controls from the Genetics of Kidneys in Diabetes (GoKinD) study were analyzed for genetic association with diabetic nephropathy (DN) in type 1 diabetics (T1D). The Cochran-Armitage test for trend and the family-based association test were employed using either unrelated cases and controls or trios, respectively. In addition to combining single nucleotide polymorphism (SNP) *p*-values across these tests via Fisher's method, we employed a novel screening approach to rank SNPs based on conditional power for more efficient testing. Using either the population-based or family-based subset alone predictably limited resolution to detect DN SNPs. For 384,197 SNPs passing quality control (QC), none achieved strict genome-wide significance (1.4×10^{-7}) using 1171 singletons (577/594 cases/controls) or 1738 pooled singletons and offspring probands (841/897). Similarly, none of the 352,004 SNPs passing QC in 567 family trios (264/303 case/control proband trios) reached genome-wide significance. Testing the top 10 SNPs ranked using aggregated conditional power resulted in two SNPs reaching genome-wide significance, rs11645147 on chromosome 16 ($p = 1.74 \times 10^{-4} < 0.05/10 = 0.005$) and rs7866522 on chromosome 9 ($p = 0.0033$). Efficient usage of mixed designs incorporating both unrelated and family-based data may help to uncover associations otherwise difficult to detect in the presence of massive multiple testing corrections. Capitalizing on the strengths of both types while using screening approaches may be useful especially in light of large-scale, next-generation sequencing and rare variant studies.

Keywords: genome-wide association, combined study design, family-based association analysis, case-control study, diabetic nephropathies

INTRODUCTION

The successes and failures of genome-wide association studies (GWAS) have made for both interesting scientific dialog and the development of innovative statistical methodologies. While debate continues around reasons for the so-called missing heritability of GWAS, the sheer number of replicable genetic associations discovered using this approach is unarguable (Hindorff et al., 2013). Next-generation sequencing has taken the baton (or at least begun its own race) to continue the search for genetic association with complex disease outcomes. Many unique analytical issues have arisen with sequencing data, but two paramount themes of concern, in particular, persist regardless of the assay technology—quality control (QC) and study design. Here, we examine the latter in the context of the Genetics of Kidneys in Diabetes (GoKinD) study, a GWAS comprising one subset of unrelated subjects and another of mother-father-proband trios.

The relative merits of a genetic association study being designed around either families or unrelated subjects, most often cases and controls, has been addressed previously (Fardo et al., 2012). Briefly, case-control studies are generally considered easier to implement, less costly and more powerful than studies incorporating related subjects. Family-based studies on the other hand are robust to the discovery of spurious association due to unresolved population substructure and also provide more textured information such as improved haplotype resolution, Mendelian error checking and the ability to test for imprinting effects. This obviously oversimplifies the comparison of two very broad classes of designs—in this work we are concerned with implications of combining the two rather than simply choosing one or the other.

Many genetic association studies spawn from existing cohorts that either had previously employed linkage analysis with pedigree recruitment (Clerget-Darpoux and Elston, 2007) or had initially not explored genetic risk factors. Studies in these scenarios can then quite naturally comprise both unrelated subjects and families. Because this is not uncommon, there are many statistical methodologies that have been developed to combine information from unrelated cases and controls with family pedigrees, and several of these have been compared via simulation (Fardo et al., 2011). Our focus here is not to again compare distinct methodologies across a simulation study but rather to compare simple, easily implementable approaches in handling the unrelated and family subsets from the GoKinD study.

The GoKinD study aimed to identify genes associated with diabetic nephropathy (DN) in type 1 diabetics (T1D). T1D patient probands were screened to identify cases with kidney disease and controls with normal renal status despite long-term diabetes. When possible, both parents of the proband were enrolled to form family trios and DNA was collected for all T1D patients and participating parents (Mueller et al., 2006). In the original GWAS, Pezzolesi et al. (2009) combined all GoKinD cases and controls (unrelated singletons and trio offspring) to test for genetic association with DN. No single nucleotide polymorphism (SNP) reached genome-wide significance but several loci were "suggestive" ($p < 1 \times 10^{-5}$). This strategy of combining the offspring from trios, or, more generally, a randomly selected non-founder from a pedigree, with the unrelated cases and controls has been common practice (Infante-Rivard et al., 2009).

Here, we propose a simple, intuitive, and straightforwardly implementable strategy to combine association metrics from unrelateds and families while providing a working solution to the multiple testing problem when these types of data are available. The main goal of this work, however, is two-fold: to thoroughly examine the differences between first-pass approaches and those using all available information; and to make the case for using and developing methods for aggregation while suggesting a direction for this methodological research. Due to the nature of the study designs employed, the GoKinD study is an ideal dataset to present these comparisons. In what follows, we further describe the motivating GoKinD dataset and QC procedures employed, we outline the various methodological approaches explored including our initial suggestion of a combined screening and testing method, and finally we thoroughly compare results from the GoKinD study.

METHODS

THE GoKinD STUDY

Subjects

Detailed information regarding these data can be found in Mueller et al. (2006). Briefly, the GoKinD study comprises 1869 T1D patients with and without kidney impairment who were recruited through the George Washington University Biostatistical Center (GWU) and the Joslin Diabetes Center, section of Genetics and Epidemiology (JDC). Patients were 18–59 years old at the time of enrollment, received a T1D diagnosis before age 31 and had diabetes duration of more than 10 years in cases and more than 15 years in controls. DN cases were defined as either persistent proteinuria or end stage renal disease requiring dialysis or renal transplant. Controls were defined as having normal renal function and normal urine albumin. Of the 1285 unrelated singletons (664/621 DN cases/controls) and 584 mother-father-offspring trios (268/316 DN case/control offspring) recruited and genotyped, 1270 unrelated singletons (651/619 DN cases/controls) and 571 mother-father-offspring trios (266/305 DN case/control offspring) were released for analysis through dbGaP (Mailman et al., 2007; Pluzhnikov et al., 2010).

Quality control

We replicated the extensive and well-documented QC procedures conducted in the original GoKinD GWAS which employed the Affymetrix 5.0 500K SNP array (Pezzolesi et al., 2009). To maintain consistency, we repeated the entire QC pipeline with and without the addition of trio offspring cases and controls using the 469,094 SNPs provided by dbGaP. The former mirrors the original study that incorporated family data which allowed for additional Mendelian error QC filtering and the latter comprises the QC for the population-based subset within the proposed methodology and typical of case-control GWAS studies. Over 35,000 SNPs were removed due to the detection of 3 or more trios exhibiting a Mendelian error (Supplemental Table 2). Principal component analysis (PCA) was applied to both population-based subsets to minimize spurious associations due to population substructure by removing potential ethnic outliers (Price et al., 2006) (**Supplemental Figure 2**). More details on QC can be found in the Supplementary Materials.

STATISTICAL ANALYSIS

We first compared the approach of separating subjects into subsets of unrelated population-based cases and controls (singletons) and family-based subjects (trios), against adding the trio offspring into a pooled unrelated subset, to analyze using common case-control statistics as in the initial analysis of Pezzolesi et al. (2009). We then implemented a two-step approach combining statistical tests across unrelated and family-based study designs (**Figure 1**).

Population-based association

Genetic association using the subset of unrelateds was examined using the Cochran-Armitage test for trend assuming an additive genetic mode of inheritance. The trend test was adjusted for sex and stratified by center using a Cochran-Mantel-Haenszel test as in the original GoKinD GWAS. These tests were conducted with and without the addition of offspring cases and controls in order to replicate the original findings and to use within the proposed framework, respectively (**Figure 1**; Singletons Only vs. Singletons and Trios). All analyses were conducted using the freely-available softwares PLINK (Purcell et al., 2007) and R (R Development Core Team, 2010).

Incorporation of trio parents

Along with adding resolution for QC, the addition of parents makes possible traditional family-based association testing (FBAT). FBATs were calculated using the FBAT package (Laird et al., 2000) assuming a DN prevalence within T1D of 30% (Krolewski et al., 1996; Steinke, 2009). Using true prevalence as an offset in the FBAT numerator is known to maximize power for the test in population samples (Whittaker and Lewis, 1998; Lunetta et al., 2000; Lange and Laird, 2002). Because ascertainment was not conditioned on DN status in GoKinD, this estimate should perform optimally.

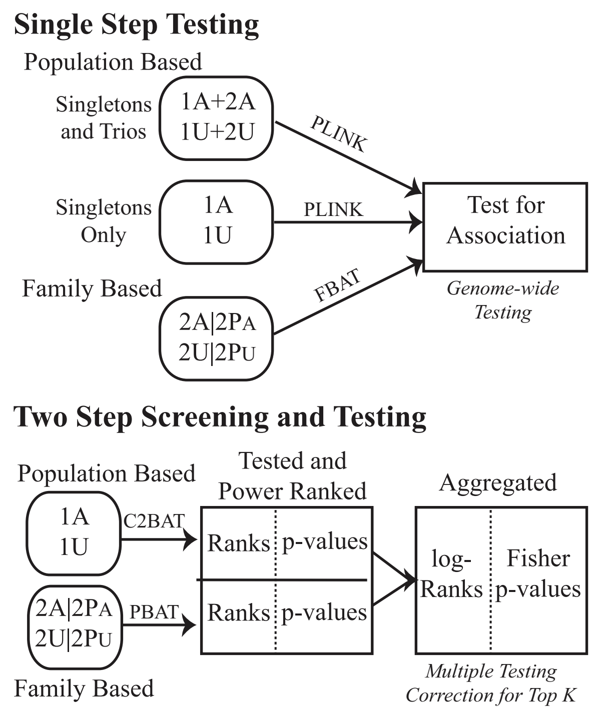


FIGURE 1 | Testing schematic for the GoKinD collection. Subjects with type 1 diabetes with (affected) and without (unaffected) diabetic nephropathy were studied for genetic association. Population- and family-based subsets were either tested in a typical straight-forward single-step strategy or in a two-step combination strategy with conditional screening for power, association testing and subsequent combination of the two. To distinguish between the data subsets used, 1′ s indicate unrelateds and 2′ s are from the family samples. A denotes affected and U unaffected. The analytic methods used are indicated above the corresponding arrows.

Fisher's combined probability test

We adopted a simple procedure to combine test statistics across study designs. Fisher's method (Fisher, 1925), often used in meta analyses, is a commonly used approach to aggregate independent p-values. Here, our testing is done in two separate subsets, family trios and case-control singletons, which maintains the independence necessary to implement this test. There are other methodologies to combine *p*-values, and all of our work could be adapted straightforwardly to accommodate alternative choices.

Dealing with multiple comparisons

For the trio subset, offspring genotypes are treated as missing and then imputed assuming Mendelian transmissions from parental genotypes in order to provide information for screening that is completely independent of the actual family-based association test. That is, offspring genotypes are not used in the screening step so that they may be used in a completely independent testing step. SNPs with favorable configurations (i.e., enough allelic variation and informative families) will be ranked highly by virtue of providing more likelihood of finding an association that is present. More formally, the Van Steen algorithm (Van Steen et al., 2005) decomposes the joint data likelihood into two independent pieces [i.e., $P(Y,X,S) = P(X|Y,S)P(Y,S)$, where Y is the offspring phenotype, X is the offspring genotype score (e.g., the count of minor alleles) and S are the sufficient statistics for offspring genotypes which are equal to the parental genotypes when available]. SNPs are screened based on information from $P(Y,S)$, either from obtaining significance rankings from regression of Y on $E(X|S)$ or from analytically calculating the conditional power for a SNP-phenotype pair; we employed the latter approach throughout. Note that $E(X|S)$ is simply the expected offspring genotype score given the parental genotypes. These analyses were conducted using the freely-available PBAT software (Lange et al., 2004). The SNP rankings produced in this step use information that is completely independent of the offspring genotypes so that FBAT test statistics are orthogonal and do not require adjustment from the screening step. Thus, the top 10 SNPs, for example, can be tested with only a multiple testing adjustment for the 10 tests conducted. Extensions to the top K approach have been developed and could easily be employed (Ionita-Laza et al., 2007). The screening step is susceptible to effects of population stratification, but the testing step remains robust to spurious association.

C2BAT as proposed by McQueen et al. (http://rss.acs.unt.edu/Rdoc/library/pbatR/html/c2bat.html) and described by Sharma et al. (2012) was developed as the case-control analog to the Van Steen screening approach. Information from each SNP is split in order to screen for highly powered SNPs and then independently test for association. Similar to conditioning on the sufficient statistics for offspring genotypes, the random variables in the family-based testing framework, the margins of the affection-by-genotype contingency table are the appropriate sufficient statistics for the corresponding cell counts, which are the random variables in the population-based framework. Briefly, the C2BAT algorithm splits subjects from the contingency table into a non-informative table for screening and a testing table. These splits can be done to preferentially over-select minor homozygotes for the testing step. We employed the default selection of 75%, 50%, 25% minor homozygotes, heterozygotes, and major homozygotes to the testing table, respectively. The margins of the resulting testing table are used to randomly impute (under the null) cell counts, which are then combined with the non-informative table to rank SNPs. The testing table is then used to perform an orthogonal test for association for the highest ranking SNPs. We used the C2BAT version implemented in the pbatR package (Hoffmann and Lange, 2013).

To combine the rankings between the trio and case-control subsets, we averaged log-transformed rankings to come up with an aggregate ranking. The top 10 SNPs were then assessed for statistical significance at a lower multiple testing penalty (i.e., $0.05/10 = 0.005$).

Note that our selection approach results in rankings equivalent to those from multiplying the rankings from both subsets. Importantly, this method is subjectively chosen and can likely be optimized in future research.

Methodological comparisons

Our primary methodology to combine information across study designs employs p-value aggregation, so we compare our approach to METAL (Willer et al., 2010), an efficient meta-analysis program that incorporates the disparate association information via sample size weights and effect directionality into an aggregate statistic. Briefly, each of the trio and singleton test p-values is converted into a Z-score and then weighted by the square root of the subset sample size to comprise a meta-analytic Z-score. In addition to meta-analytic methodologies, we also compared to an approach the aggregates data across the subsamples rather than the p-values (Zhang et al., 2009). The method proposed by Zhang et al. was chosen due to its implementation in the Genetic Association analysis Platform (GAP) and its superior performance in a previous comparison between other similar data aggregation methodologies (Fardo et al., 2011). The proposed score test comprises components from each subsample separately and is explained in detail in Zhang et al. (2009). Similar to the FBAT approach, we employ a phenotypic offset equal to the estimated prevalence of DN in T1D.

RESULTS

In the population-based analyses, no SNP achieved Bonferroni adjusted genome-wide significance for association with DN in T1D ($0.05/384{,}197 = 1.3 \times 10^{-7}$). Areas of suggestive association noted in the pooled population-based analysis (**Figure 2**) are diminished in the singletons alone analysis (**Figure 3**). In the singleton alone subset, only four SNPs exceeded a suggestive p-value of 1×10^{-5}.

In the family-based analysis, no SNP achieved genome-wide significance using an FBAT statistic (**Figure 4**). Suggestive areas of associations in chromosome 11p in the *CARS* gene (cysteinyl-tRNA synthetase) were similar to results from Pezzolesi et al. (2009). New areas of interest in chromosome 6p within the major histocompatibility complexes (MHC) class II and III and in chromosome 7p are noted (Supplemental Table 3). The 13q chromosomal peak reported by Pezzolesi et al. (2009) was not observed.

No SNPs achieved significance using Fisher's combined probability method without the benefit of Van Steen-type screening approaches (**Figure 5**). The SNPs of suggestive significance in the population-based singleton only and pooled singleton and trio proband analysis were diminished by the addition of family-based information, suggesting potential population structure correction. Compared to the family-based subset, associations remained similar in other regions.

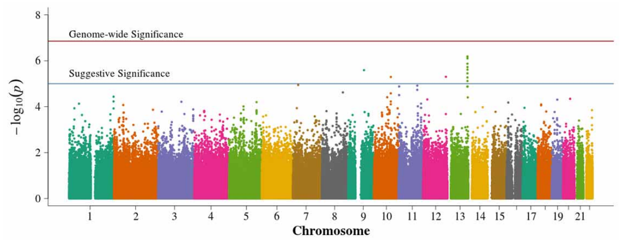


FIGURE 2 | Manhattan plot for population-based study with pooled singletons and trio probands. Summary of genome-wide association scan results in the GoKinD population-based singletons and trios combined. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy among subjects with type 1 diabetes using the Cochran-Armitage test for trend for an additive genetic model adjusted for sex and stratified by center ascertainment using Cochran-Mantel-Haenszel method. The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/357, 887 = 1.4 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

There were no genome-wide significant SNPs from either METAL or GAP, although six and four SNPs reach the suggestive significance level for METAL (**Figure 6**) and GAP (**Figure 7**), respectively. Three of these variants were not identified using other approaches. GAP analysis supports the chromosome 6p finding from the FBAT. This region harbors multiple genome-wide significant SNPs when employing either FBAT or GAP without the optimal phenotypic offset (not shown) and may actually be testing for T1D associations rather than those from DN within a TID population since, without the offset, the analysis reduces to a traditional, affecteds-only TDT.

Selection of the top 10 ranked SNPs from screening approaches combined across the unrelated and trio subsets and testing with

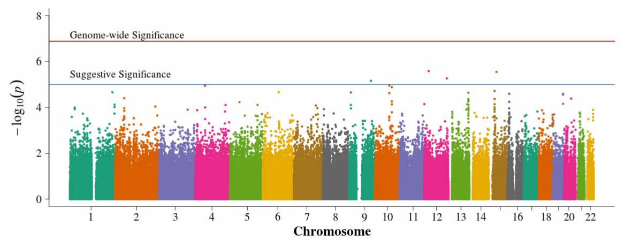


FIGURE 3 | Manhattan plot for population-based study with case and control singletons only. Summary of genome-wide association scan results in the GoKinD cases and controls, singletons only. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy among subjects with type 1 diabetes using the Cochran-Armitage test for trend for an additive genetic model adjusted for sex and stratified by center ascertainment using Cochran-Mantel-Haenszel method. The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/384, 094 = 1.3 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

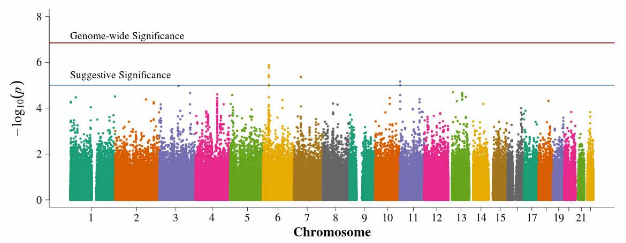


FIGURE 4 | Manhattan plot for family-based study. Summary of genome-wide association scan results in the GoKinD cases and controls family-based trios and duo parent/offspring pairs. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy among subjects with type 1 diabetes using the generalized FBAT method with an offset of 0.3 (the prevalence of diabetic nephropathy in type 1 diabetics). The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/351{,}951 = 1.4 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

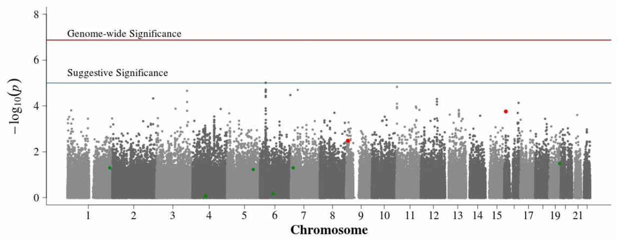


FIGURE 5 | Manhattan plot for Fisher's combined probability of population- and family-based studies. Summary of genome-wide association scan results in the GoKinD collection of combined probability of the population-based and family-based *P*-values. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy among subjects with type 1 diabetes by combining each study *P*-values using Fisher's combined probability method. Power ranking was obtained using conditional mean model for family-based data and data partitioning for population-based cases and controls. Rankings were obtained for each subset and then log transformed and summed. The top ten ranked SNPs were tested; the two SNPs significant at $0.05/10 = 0.005$ are indicated in red, while the other eight are in green. The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/374,042 = 1.3 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

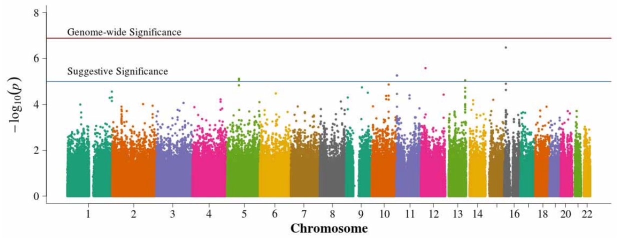


FIGURE 6 | Manhattan plot for METAL. Summary of genome-wide association scan results in the GoKinD collection of the meta-analyzed population-based and family-based P-values. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy among subjects with type 1 diabetes by combining each study *P*-value using the METAL sample size method. The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/385,830 = 1.3 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

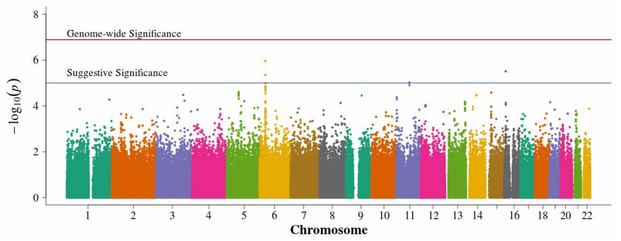


FIGURE 7 | Manhattan plot for GAP. Summary of genome-wide association scan results in the GoKinD collection of combined population-based and family-based data. The $-\log_{10}P$-values were calculated for SNP association with diabetic nephropathy using the method of Zhang et al. (2009). The red horizontal line corresponds to genome-wide significance (P-value $= 0.05/386,822 = 1.3 \times 10^{-7}$). The blue horizontal line corresponds to suggestive significance (P-value $= 1 \times 10^{-5}$).

Fisher's test resulted in two SNPs achieving genome-wide significance ($p = 0.05/10 = 0.005$; **Table 1**, Supplemental Table 3, **Figure 5**). SNP rs7866522 on chromosome 9p (p-value $= 0.0033$) is contained in the protein tyrosine phosphatase, receptor, D gene (PTPRD). Members of the protein tyrosine phosphatase family are known to be signaling molecules which regulate processes such as cell growth, differentiation, mitotic cycle, and oncogenic transformation (Wheeler et al., 2007). This region has been in identified in type 2 diabetic risk genome-wide studies (Tsai et al., 2010; Below et al., 2011; Chang et al., 2012) potentially related to glucose homeostasis and insulin sensitivity (Ren et al., 1998; Chagnon et al., 2006). SNP rs11645147 on chromosome 16p (p-value $= 0.00017$) is located in proximity to the glutamate receptor, ionotropic, N-methyl D-aspartate 2A gene (GRIN2A).

We sought to replicate the two genome-wide significant SNPs using the Family Investigation of Nephropathy and Diabetes (FIND) study (Knowler et al., 2005; Iyengar et al., 2007; Igo et al., 2011). The FIND study recruited diabetic subjects with and without nephropathy. Most FIND participants with GWAS have type 2 diabetes (between 90 and 95%), and the majority of nephropathy controls used in this sub-study are relatives of index cases. To be consistent with the GoKinD population, we examined only European American subjects. Rs11645147 conferred a p-value of 0.004 assuming a dominant mode of inheritance; rs7866522 failed to reach significance. While FIND shares the nephrotic phenotype with GoKinD, it includes primarily type 2 diabetics as opposed to type 1, making comparisons inexact. In addition, the dominant mode of inheritance was the only one for which rs11545147 garnered nominal significance, although it still reached significance after adjusting for testing multiple modes of inheritance.

DISCUSSION

The primary finding of this study is that analysis of GoKinD collection by any of a strict population-based design, a family-based design or the combined approach without any screening, did not detect genome-wide significant SNPs. Simply combining family-based association results with those from population-based data actually suppressed areas of suggestive genome-wide significance compared to the original GoKinD GWAS, possibly by correcting for previously unrecognized population substructure; however, the definitive reason for this is unknown. Conversely, the incorporation of family-based information also uncovered new areas of possible interest. Two SNPs reached significance in our combined data analysis by the novel two-step approach using Van Steen screening with the family-based trios and C2BAT data partitioning in the unrelated case-control data, which ranks markers by conditional power and then selects the top 10 overall ranked markers.

Suggestive findings using only population-based association tests with all unrelated cases and controls, when pooled with trio probands as in the Pezzolesi study, were not replicated by either the family-based or combined analyses. This finding could suggest the presence of unresolved population structure despite using PCA to select a homogenous population, and that earlier

Table 1 | Top ten SNPs by population- and family-based screening aggregation.

Power rank	SNP	CHR	BP	Minor Allele Freq.	Fisher's *p*-value	METAL *p*-value	GAP *p*-value	FBAT *p*-value	Singletons only *p*-value	Pooled cases/ controls *p*-value
1	rs11645147	16	9802457	0.377	1.74E-04	3.28E-07	3.09E-06	1.16E-03	5.71E-05	7.93E-04
2	rs7866522	9	8812704	0.284	3.28E-03	3.02E-03	5.61E-03	3.95E-01	1.10E-04	1.26E-01
3	rs847986	7	12454877	0.089	4.98E-02	6.81E-03	5.53E-03	2.71E-01	1.14E-03	3.52E-02
4	rs980519	4	72180823	0.065	8.42E-01	1.27E-01	5.90E-02	8.63E-01	1.23E-01	2.47E-03
5	rs11673097	19	57119434	0.293	3.25E-02	3.80E-04	1.50E-03	8.38E-01	2.77E-05	2.85E-03
6	rs4707991	6	73493822	0.332	6.78E-01	1.00E-01	1.28E-01	8.81E-01	7.15E-02	1.13E-03
7	rs17689531	4	72022775	0.120	8.81E-01	9.33E-01	9.15E-01	9.23E-01	9.59E-01	9.26E-02
8	rs1901712	4	72147303	0.065	9.25E-01	9.37E-02	3.70E-02	6.38E-01	1.08E-01	2.01E-03
9	rs17470789	5	144584265	0.125	5.88E-02	2.72E-01	1.78E-01	1.86E-02	8.41E-01	5.92E-01
10	rs8179278	1	234379913	0.091	5.01E-02	5.19E-05	5.25E-05	8.67E-01	2.18E-05	2.66E-03

Aggregated SNP rankings based on conditional power for family-based and population-based studies were calculated. The top ten ranked SNPs were selected per convention to minimize the need for multiple comparison correction (0.05/10 = 0.005). Fisher's combined p-values for diabetic nephropathy trait association were obtained using the corresponding association method used for power analysis from FBAT and C2BAT. METAL and Genetic Association analysis Platform (GAP) aggregated p-values were also obtained but are subjected to genome-wide multiple comparison correction (0.05 /384,197 = 1.3 × 10^{-7}). Genome-wide family-based association testing (FBAT) included trio probands and parents. Cochran-Armitage test for trend for an additive genetic model adjusted for sex and stratified by center ascertainment using Cochran-Mantel-Haenszel method were obtained using unrelated cases and controls only (singletons only) and by combining cases and controls from unrelated subjects and trio probands (pooled cases and controls).

suggestive SNPs were likely false positive associations. It also could be a result of a decrease in power from using family-based tests. This balance of increased robustness against problems of population stratification and a decrease in power are common factors when considering family-based tests.

Compared to analyzing either of the unrelated case-control or trio datasets alone, the additional sample size via the combined Fisher's method increases study power, and this may explain the new areas of suggestive significance. The lack of findings of genome-wide significant SNPs may reflect that there are truly no associations between DN and genotyped markers among the GoKinD dataset or that the study reflects the difficulties encountered with the multiple testing problem inherent to GWAS.

Applying screening methods due to Van Steen et al. (2005) and McQueen et al. (http://rss.acs.unt.edu/Rdoc/library/pbatR/html/c2bat.html), statistically independent assessments of each SNP's power to detect an association allows for more efficient genome-wide testing. Here we aggregated the independently obtained marker rankings using parental information in the family-based data and data partitioning in the population-based data. By limiting testing to the conventionally-used top 10 highest ranked markers (Herbert et al., 2006), two SNPs reached genome-wide significance. While this result is appealing, without extensive simulation to establish operating characteristics of the suggested approach in other settings, caution must be taken to not over interpret. It does suggest, at the least, that future methodological work in this regard is warranted. We plan to investigate the performance of this approach in other scenarios and examine implications of varying the choice of the number of SNPs to carry to the testing stage as well as the function for rank aggregation.

With the growing availability of GWAS and now sequencing data, association studies have increasingly reported positive results. Multiple-hypothesis testing, low power, study design variability, phenotypic definition, and population structure continue to pose investigational difficulties (Laird and Lange, 2006). Family-based and population-based case control designs each have unique strengths and weaknesses, but when used in a complementary fashion as proposed, they may overcome these challenges.

ACKNOWLEDGMENTS

The authors thank Drs. Jim Hartsfield and Dick Kryscio for helpful comments and useful discussion as well as two anonymous reviewers whose critiques improved the manuscript. A full list of Family Investigation of Nephropathy and Diabetes Research Group contributors can be found in the online appendix of Schelling et al. (2008).This work was funded in part by grants from the National Institute of General Medical Sciences (1K25AG043546-01A1, 8P20GM103436-12, 8P20GM103538-09). Funding support for the GAIN Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes (GoKinD study participants) study was provided by the Juvenile Diabetes Research Foundation (JDRF), and the Centers for Disease Control (CDC) (PL 105-33, 106-554, and 107-360 administered by the National Institute of Diabetes and Digestive and Kidney Diseases, NIDDK) and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). The dataset(s) used for the analyses described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000018.v1.p1 and phs000088.v1.p1. Samples and associated phenotype data for the Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes study were provided by James H. Warram, MD, Joslin Diabetes Center. This manuscript was not prepared in collaboration with Investigators of the GoKinD study and does not necessarily reflect the opinions or views of the GoKinD study or the NIDDK.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Applied_Genetic_Epidemiology/10.3389/fgene.2013.00186/abstract

Figure S1 | Q-Q Plots for association within controls and cases. When controls and cases from each center of ascertainment are combined by affection status, an over dispersion of the Cochran-Armitage test statistic for trend is noted. The deviation from expected, confirmed by an elevated genomic control inflation factor ($\lambda_{GC} > 1.05$), suggests underlying confounding and stratification by center ascertainment between the Joslin Diabetes Center and the George Washington University Biostatistical Center.

Figure S2 | Mendelian errors per GoKinD family trio. In the GoKinD family-based study, 551 trios (two parents and one offspring) were assessed for Mendelian errors of transmission. Number of Mendel SNP errors per family was $\log_{10}$ transformed. A single outlier family was determined by a greater than 5% Mendel error rate (>20,000 errors, indicated by the red arrow). Mendelian errors reflect poor quality SNP genotype calling, poor DNA sample quality or inconsistent familial relationship.

Figure S3 | Projection of principal components of population-based GoKinD subjects onto HapMap populations. A pruned set of SNPs (85,051) from the population-based cases and controls were projected onto a similar set of SNPs from the original three International HapMap populations [GoKinD subjects in blue, **(A)**]. Using *Z*-scores based on median absolute deviation, a homogenous population was selected for association analysis [selected GoKinD population shown in blue and outliers in red, **(B)**]. HapMap populations: CEU (Eastern and Western European) samples are shown in green, YRI (Yoruba in Ibadan, Nigeria) are in black, and JPT + CHB (Japanese in Tokyo, Japan and Hans Chinese in Beijing) are shown in violet.

REFERENCES

Below, J. E., Gamazon, E. R., Morrison, J. V., Konkashbaev, A., Pluzhnikov, A., McKeigue, P. M., et al. (2011). Genome-wide association and meta-analysis in populations from Starr County, Texas, and Mexico City identify type 2 diabetes susceptibility loci and enrichment for expression quantitative trait loci in top signals. *Diabetologia* 54, 2047–2055. doi: 10.1007/s00125-011-2188-3

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Chagnon, M. J., Elchebly, M., Uetani, N., Dombrowski, L., Cheng, A., Mooney, R. A., et al. (2006). Altered glucose homeostasis in mice lacking the receptor protein tyrosine phosphatase sigma. *Can. J. Physiol. Pharmacol.* 84, 755–763. doi: 10.1139/y06-020

Chang, Y. C., Chiu, Y. F., Liu, P. H., Shih, K. C., Lin, M. W., Sheu, W. H., et al. (2012). Replication of genome-wide association signals of type 2 diabetes in Han Chinese in a prospective cohort. *Clin. Endocrinol. (Oxf.)* 76, 365–372. doi: 10.1111/j.1365-2265.2011.04175.x

Clerget-Darpoux, F., and Elston, R. C. (2007). Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. *Hum. Hered.* 64, 91–96. doi: 10.1159/000101960

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997–1004. doi: 10.1111/j.0006-341X.1999.00997.x

Fardo, D. W., Charnigo, R., and Epstein, M. P. (2012). Families or unrelated: the evolving debate in genetic association studies. *J. Biomet. Biostat.* 3:e108. doi: 10.4172/2155-6180.1000e108

Fardo, D. W., Druen, A. R., Liu, J., Mirea, L., Infante-Rivard, C., and Breheny, P. (2011). Exploration and comparison of methods for combining population- and family-based genetic association using the Genetic Analysis Workshop 17 mini-exome. *BMC Proc.* 5(Suppl. 9):S28. doi: 10.1186/1753-6561-5-S9-S28

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Edinburgh, London: Oliver and Boyd.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., et al. (2003). The international HapMap project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., et al. (2006). A common genetic variant is associated with adult and childhood obesity. *Science* 312, 279–283. doi: 10.1126/science.1124779

Hindorff, L. A., MacArthur, J. (European Bioinformatics Institute)., Morales, J. (European Bioinformatics Institute)., Junkins, H. A., Hall, P. N., Klemm, A. K., et al. (2013). *A Catalog of Published Genome-Wide Association Studies.* Available online at: www.genome.gov/gwastudies. Accessed August 1, 2013.

Hoffmann, T., and Lange, C. (2013). *pbatR: P2BAT. R package* Version 2.2-9. Available online at: http://CRAN.R-project.org/package=pbatR

Igo, R. P. Jr., Iyengar, S. K., Nicholas, S. B., Goddard, K. A., Langefeld, C. D., Hanson, R. L., et al. (2011). Genomewide linkage scan for diabetic renal failure and albuminuria: the FIND study. *Am. J. Nephrol.* 33, 381–389. doi: 10.1159/000326763

Infante-Rivard, C., Mirea, L., and Bull, S. B. (2009). Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am. J. Epidemiol.* 170, 657–664. doi: 10.1093/aje/kwp180

Ionita-Laza, I., McQueen, M. B., Laird, N. M., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am. J. Hum. Genet.* 81, 607–614. doi: 10.1086/519748

Iyengar, S. K., Abboud, H. E., Goddard, K. A., Saad, M. F., Adler, S. G., Arar, N. H., et al. (2007). Genome-wide scans for diabetic nephropathy and albuminuria in multiethnic populations: the family investigation of nephropathy and diabetes (FIND). *Diabetes* 56, 1577–1585. doi: 10.2337/db06-1154

Knowler, W. C., Coresh, J., Elston, R. C., Freedman, B. I., Iyengar, S. K., Kimmel, P. L., et al. (2005). The Family Investigation of Nephropathy and Diabetes (FIND): design and methods. *J. Diabetes Complications* 19, 1–9. doi: 10.1016/j.jdiacomp.2003.12.007

Krolewski, M., Eggers, P. W., and Warram, J. H. (1996). Magnitude of end-stage renal disease in IDDM: a 35 year follow-up study. *Kidney Int.* 50, 2041–2046. doi: 10.1038/ki.1996.527

Laird, N. M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* 19(Suppl. 1), S36–S42. doi: 10.1002/1098-2272(2000)19:1+<::AID-GEP16>3.0.CO:2-M

Laird, N. M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394. doi: 10.1038/nrg1839

Lange, C., Demeo, D., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2004). PBAT: tools for family-based association studies. *Am. J. Hum. Genet.* 74, 367–369. doi: 10.1086/381563

Lange, C., and Laird, N. M. (2002). Power calculations for a general class of family-based association tests: dichotomous traits. *Am. J. Hum. Genet.* 71, 575–584. doi: 10.1086/342406

Lunetta, K. L., Faraone, S. V., Biederman, J., and Laird, N. M. (2000). Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.* 66, 605–614. doi: 10.1086/302782

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi: 10.1038/ng1007-1181

Mueller, P. W., Rogus, J. J., Cleary, P. A., Zhao, Y., Smiles, A. M., Steffes, M. W., et al. (2006). Genetics of Kidneys in Diabetes (GoKinD) study: a genetics collection available for identifying genetic susceptibility factors for diabetic nephropathy in type 1 diabetes. *J. Am. Soc. Nephrol.* 17, 1782–1790. doi: 10.1681/ASN.2005080822

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2:e190. doi: 10.1371/journal.pgen.0020190

Pezzolesi, M. G., Poznik, G. D., Mychaleckyj, J. C., Paterson, A. D., Barati, M. T., Klein, J. B., et al. (2009). Genome-wide association scan for diabetic nephropathy susceptibility genes in type 1 diabetes. *Diabetes* 58, 1403–1410. doi: 10.2337/db08-1514

Pluzhnikov, A., Below, J. E., Konkashbaev, A., Tikhomirov, A., Kistner-Griffin, E., Roe, C. A., et al. (2010). Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am. J. Hum. Genet.* 87, 123–128. doi: 10.1016/j.ajhg.2010.06.005

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. 2.25.2 ed. Vienna: R. Foundation for Statistical Computing.

Ren, J. M., Li, P. M., Zhang, W. R., Sweet, L. J., Cline, G., Shulman, G. I., et al. (1998). Transgenic mice deficient in the LAR protein-tyrosine phosphatase exhibit profound defects in glucose homeostasis. *Diabetes* 47, 493–497. doi: 10.2337/diabetes.47.3.493

Schelling, J. R., Abboud, H. E., Nicholas, S. B., Pahl, M. V., Sedor, J. R., Adler, S. G., et al. (2008). Genome-wide scan for estimated glomerular filtration rate in multi-ethnic diabetic populations: the Family Investigation of Nephropathy and Diabetes (FIND). *Diabetes* 57, 235–243. doi: 10.2337/db07-0313

Sharma, S., Poon, A., Himes, B. E., Lasky-Su, J., Sordillo, J. E., Belanger, K., et al. (2012). Association of variants in innate immune genes with asthma and eczema. *Pediatr. Allergy Immunol.* 23, 315–323. doi: 10.1111/j.1399-3038.2011.01243.x

Steinke, J. M. (2009). The natural progression of kidney injury in young type 1 diabetic patients. *Curr. Diab. Rep.* 9, 473–479. doi: 10.1007/s11892-009-0077-7

Tsai, F. J., Yang, C. F., Chen, C. C., Chuang, L. M., Lu, C. H., Chang, C. T., et al. (2010). A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. *PLoS Genet.* 6:e1000847. doi: 10.1371/journal.pgen.1000847

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., Demeo, D. L., et al. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37, 683–691. doi: 10.1038/ng1582

Weale, M. E. (2009). *EIGENSOFT Notes*. Available online at: http://sites.google.com/site/mikeweale/software/eigensoftplus. (Accessed 2010).

Weale, M. E. (2010). "EIGENSOFTplus," in *Bio informatics, Statistical Genetics, and Epigenetics*. 10th Edn.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 35, D5–D12. doi: 10.1093/nar/gkl1031

Whittaker, J. C., and Lewis, C. M. (1998). The effect of family structure on linkage tests using allelic association. *Am. J. Hum. Genet.* 63, 889–897. doi: 10.1086/302008

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Zhang, L., Pei, Y. F., Li, J., Papasian, C. J., and Deng, H. W. (2009). Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PLoS ONE* 4:e6502. doi: 10.1371/journal.pone.0006502

Ziegler, A., Konig, I. R., and Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biom. J.* 50, 8–28. doi: 10.1002/bimj.200710398

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 March 2013; paper pending published: 28 April 2013; accepted: 03 September 2013; published online: 27 September 2013.

*Citation: Estus JL, FIND Research Group and Fardo DW (2013) Combining genetic association study designs: a GWAS case study. Front. Genet. **4**:186. doi: 10.3389/fgene.2013.00186*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

APPENDIX

RECRUITMENT

Of the 1879 T1D subjects initially recruited, 10 failed genotyping on an Affymetrix 5.0 (500 K) SNP array conducted by the GAIN genotyping laboratory at the Broad Institute (Cambridge, MA) and the Central Biochemistry Laboratory at the University of Minnesota. Of the remaining 1869 subjects, 21 were excluded from the data release due to sample duplication detected by identifying cryptic relatedness and eight were removed due to assay plate failure via genotype calling interference (https://www.niddkrepository.org/GOKIND) (Pluzhnikov et al., 2010). Of the 1840 remaining, none were detected as cryptically related using identity-by-descent proportion estimation ($\widehat{\pi} < 0.1621$).

GoKinD samples were recruited under two separate ascertainment protocols at JDC and GWU. Using Q-Q plots, over dispersion of the Cochran-Armitage test statistic for trend for JDC vs. GWU, among controls and cases, separately were demonstrated (**Supplemental Figure 1**). To test if the observed overall inflation factor (λ_{GC}) (Devlin and Roeder, 1999) for cases ($\lambda_{GC} = 1.097$) and controls ($\lambda_{GC} = 1.115$) were truly significant for stratification, centers were permutated by affection status. For 1000 permutations in cases and controls, no λ_{GC} were more extreme (p-value $< 10^{-3}$); hence further association testing was stratified by center.

SINGLETONS

Of the 1270 population-based singletons remaining, four were removed for sex mismatch and one for high individual genotype missingness (>0.10), which left 1265 (650 cases and 615 controls) (Supplemental Table 1).

TRIOS

Of the 571 family-based trios, 551 included genotyping for both parents (full trios) while 20 included only a single founder. Three subjects and their parents were excluded for sex mismatch. Families were evaluated for Mendelian error rates to assess validity of relatedness and the degree of genotyping error: one was excluded with a Mendelian error rate greater than 5% (**Supplemental Figure 2**). This subject was added to the singletons but was excluded due to high individual genotype missingness, which confirms the original Mendelian error finding. A total of 567 parent(s)/offspring were included (264 case and 303 control offspring) (Supplemental Table 1).

SNP QUALITY CONTROL

For autosomal chromosomes, both population- and family-based SNPs were filtered for an overall minor allele frequency (MAF) <0.01, Hardy-Weinberg equilibrium probability $= 1 \times 10^{-5}$, duplicate SNPs, and sequential missingness by MAF; 95% overall minimum completeness, 97% for MAF between 5–10%, and 99% for infrequent SNPs with MAF between 1 and 5% (Burton et al., 2007; Ziegler et al., 2008; Pezzolesi et al., 2009). For population-based C2BAT power analysis, an overall MAF <0.05 screening was used per computational software restriction. In addition, family-based SNPs were filtered for a Mendelian error rate of 3 per SNP based on a subset of full trios excluding families with > 10,000 errors per family. A final 384,197 and 338,970 singleton SNPs (PLINK and C2BAT analysis, respectively) and 352,004 trio SNPs were analyzed (Supplemental Table 2).

POPULATION STRUCTURE

To select a homogenous population in the singleton cases and controls, PCA was performed by projection of a pruned subset of SNPs (85,051) onto the three original HapMap populations [Utah residents with ancestry from northern and western Europe (CEU), Yoruba in Ibadan, Nigeria (YRI), Japanese in Tokyo, Japan and Hans Chinese in Beijing, China (JPT_CHB), http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml#hapmap, Phase 2, release 23] (Gibbs et al., 2003) using EIGENSOFT (Patterson et al., 2006; Price et al., 2006) and EIGENSOFTplus (Weale, 2009, 2010) software. Singleton genotypes were pruned with PLINK's (Purcell et al., 2007) in-depth pairwise option (500 SNP sliding window, 5 SNP step), with additional removal of long range linkage disequilibrium areas (**Supplemental Figure 3A**). Outliers were determined using visual assessment and calculated Z-scores based on median absolute deviation, i.e., median (|X—median (X) |). Ninety-four subjects were excluded at a Z-score of 9.1 for a final singleton sample of 1171 (576 cases, 597 controls) (**Supplemental Figure 3B**).

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 18 October 2013
doi: 10.3389/fgene.2013.00195

New insights into the genetic mechanism of IQ in autism spectrum disorders

Harold Z. Wang*[1], *Hai-De Qin*[1], *Wei Guo*[1], *Jack Samuels*[2] *and Yin Yao Shugart*[1]

[1] Unit on Statistical Genomics, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA
[2] Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

Edited by:
Xiangqing Sun, Case Western Reserve University, USA

Reviewed by:
Jing Hua Zhao, University of Cambridge, UK
Tao Feng, Case Western Reserve University, USA

****Correspondence:***
Yin Yao Shugart, Unit on Statistical Genomics, Intramural Research Program, National Institute of Mental Health, National Institutes of Health, Building 35, Room 3A 1000, 35 Convent Drive, Bethesda, MD 20892, USA
e-mail: kay1yao@mail.nih.gov

Autism spectrum disorders (ASD) comprise a number of underlying sub-types with various symptoms and presumably different genetic causes. One important difference between these sub-phenotypes is IQ. Some forms of ASD such as Asperger's have relatively intact intelligence while the majority does not. In this study, we explored the role of genetic factors that might account for this difference. Using a case–control study based on IQ status in 1657 ASD probands, we analyzed both common and rare variants provided by the Autism Genome Project (AGP) consortium via dbGaP (database of Genotypes and Phenotypes). We identified a set of genes, among them HLA-DRB1 and KIAA0319L, which are strongly associated with IQ within a population of ASD patients.

Keywords: GWAS, functional variants, rare variants, common variants, autism, cognitive development

INTRODUCTION

Autism gained recognition in the 1940s as a mental disorder characterized by social deficits, communication difficulties, and other abnormalities. Since then, scientists have increasingly recognized that autism is not one but a family of conditions that share certain clinical characteristics. Currently, classical autism, Asperger's syndrome, Rett's syndrome, childhood disintegrative disorder, and pervasive developmental disorder not otherwise specified (PDD-NOS) are grouped together as autism spectrum disorders (ASD). However, the recent revision of in the Diagnostic and Statistical Manual of Mental Disorders version 5 replaced this categorization with a continuous scale of severity (Halfon and Kuo, 2013).

There is considerable evidence for the role of inheritance in the etiology of autism and related disorders. Studies have consistently reported that the prevalence of autism in siblings of autistic children is approximately 15–30 times greater than the rate in the general population (Szatmari, 1999). More recently, identified genetic variants include inherited mutations, *de novo* mutations, single point mutations, and copy number variants (CNVs). In particular, researchers reported hundreds of ASD risk factors, ranging from *de novo* to inherited, CNVs to single point mutations (Anney et al., 2012).

Some variants found to be associated with ASD were discovered only when researchers restricted the study subjects to a specific population group. The distinction by IQ may be particularly relevant in ASD research, helping to separate Asperger's syndrome, an ASD sub-type which spares language development, from autism, which does not. For example, in a recent study, Anney et al. (2012) identified a variant, rs1718101, which was strongly associated with ASD only in Europeans with high-IQ. In the current study, we hypothesized that the genetic etiology of ASD may be different based on IQ status. To test this hypothesis, we compared genotypic frequencies in high-IQ ASD probands with those of the low-IQ probands. We analyzed both common and rare variant. Specifically, we used the sequence kernel association test (SKAT) developed by Wu et al. (2011) to analyze the rare variants with minor allele frequency (MAF) less than 0.05.

MATERIALS AND METHODS

DATA DESCRIPTION

The study was conducted using a genome-wide association study (GWAS) data set of ASD families evaluated by the Autism Genome Project (AGP) consortium [provided by dbGaP (database of Genotypes and Phenotypes); Anney et al., 2012]. The AGP consortium represented more than 50 centers in North America and Europe. The centers collected clinical information from 2705 ASD families for the combined stage 1 and 2 study. Autism Diagnostic Interview-Revised (ADI-R) (2) and Autism Diagnostic Observation Schedule (ADOS) (3) were used for research diagnostic evaluation. Individuals were classified into "strict" or "spectrum" (i.e., includes strict) disorders, based on ADI-R and ADOS classification. Individuals with known karyotypic abnormalities, fragile X mutations, or other genetic disorders were excluded. Genotyping was performed by using the Illumina Human 1M-single Infinium BeadChip array (Anney et al., 2012). This resulted in 2665 ASD families (7880 individuals). We checked for Mendelian errors using PedCheck, and found none (O'Connell and Weeks, 1998). We further checked for per-individual genotyping missing rate, and removed those with more than 50%, leaving 7769 individuals within 2604 pedigrees. Because our research aim was to investigate the role of genetic variants associated with IQ difference in IQ in ASD patients, we focused on the probands and excluded their parents from this study.

ANALYTICAL METHODS

High-IQ probands in the AGP data set were defined by the AGP committee as those with IQ greater than 80, while low-IQ probands were defined as those with IQ of between 25 and 70. Using this definition, out of 2095 probands with non-missing IQ statues included in the data, 1034 were classified as high-IQ, 623 as low-IQ, and 438 as normal-IQ. Probands with missing IQ statuses were not included in the analyses. In this paper, we compared the 1034 high-IQ probands to the 623 low-IQ probands for a total of 1657 individuals. Of these 1657 individuals, 918 high-IQ individuals and 511 low-IQ individuals for a total of 1429 were Caucasian. This required us to account for population stratification in this study.

Our approach differed for common and rare variants. We used MAF of 0.05 as the threshold to differentiate between the two types of variants. For common variants, we used PLINK's (v1.07) built in function to account for population stratification. We first calculated the pair wise identity by state (IBS) matrix, and then performed a multidimensional scaling (MDS) analysis using two dimensions. We then used the two-dimensional MDS statistics along with sex as covariates to perform a logistic regression for each individual common single nucleotide polymorphism (SNP).

The analysis of rare variants is more complicated since, given the low numbers of informative individuals, association results for single rare variants tend to be unreliable. For this study, we used the SKAT (Wu et al., 2011). As with many other methods designed for rare variant analysis, SKAT analyzes multiple variants together as a unit. This remedied the lack of power for single rare variants by combining the effects of multiple variants. However, unlike the burden tests such as collapsing methods, which aggregate variants into a single variable before performing statistical regression, SKAT combines individual variant-test statistics after analyzing each variant independently. This is advantageous compared to collapsing methods when large numbers of variants affect the phenotype to increase or decrease the risk, and also when a large fraction of variants is non-causal. We used a gene-based method in our approach to rare variants, in which rare variants outside of known genes were not included in our analysis and the rest analyzed collectively via SKAT on a gene-by-gene basis. Dealing with population stratification via MDS analysis was not satisfactory for rare variants; thus, we included only Caucasian probands in this analysis.

RESULTS

POPULATION STRATIFICATION

Of 1657 probands, 1429 are of Caucasian descent. The MDS plot obtained during the common variant analysis process is shown in **Figure 1**. Population stratification is significant for the sample. The Caucasian probands were relatively close genetically, while non-Caucasian individuals showed wide genetic differences among themselves. Specifically, non-Caucasians seemed to group themselves into two clusters. These could be different non-Caucasian ethnicities, but data were not available for proper identification. We presented a QQ-plot with the p-value of our adjusted analysis (**Figure 2**).

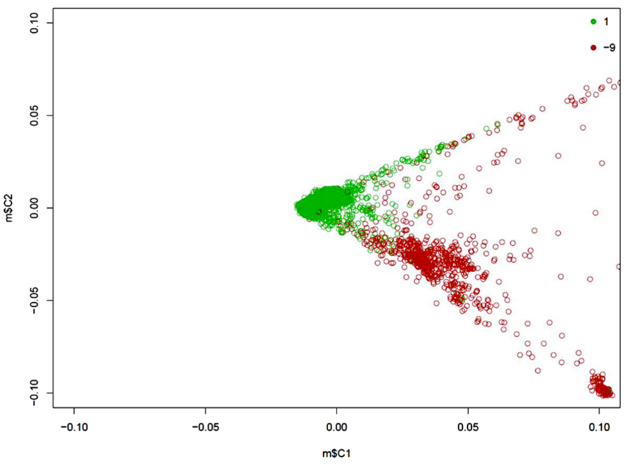


FIGURE 1 | Two-dimensional MDS plot of the AGP population. The green circles are Caucasian individuals; the red circles are those of other ethnicities.

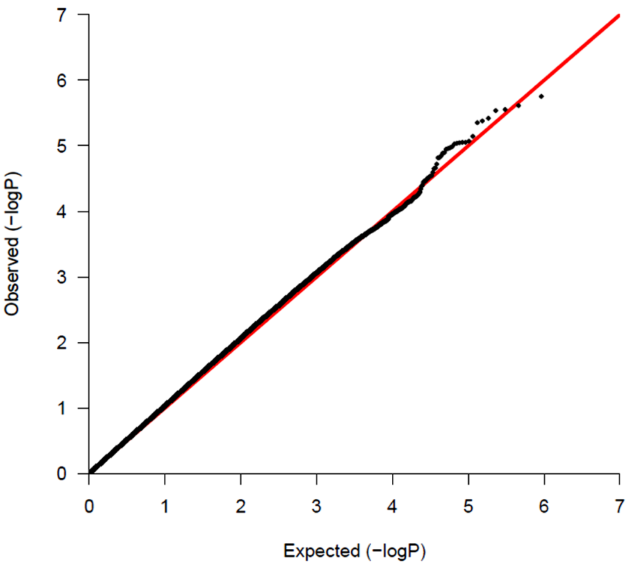


FIGURE 2 | QQ-plot of the *p*-values of common variant analysis.

COMMON VARIANTS

We analyzed a total of 878,930 SNPs. Fifteen SNPs had associations with p-value lower than 10^{-5}, and 82 with p-values lower than 10^{-4} (data not shown). Forty-eight of the variants found in the high-IQ vs. low-IQ comparison have odds-ratio of less than 1, indicating an association with low-IQ, while the remainders are associated with high-IQ. We probed into the biological relevance of all SNPs with p-values lower than 10^{-4} in the NCBI SNP database, by analyzing genes that contain or are situated close to the SNP. Seventeen SNPs out of 192 in the high-IQ vs. low-IQ analysis fell

Table 1 | Common variant analysis results of high-IQ vs. low-IQ.

CHR	SNP	BP	Risk allele	TEST	Sample size	OR	STAT	*p*-Value	Gene
6	rs9268880	32539336	A	ADD	1657	0.7089	−4.443	8.85×10^{-6}	HLA-DRB1
6	rs6903608	32536263	G	ADD	1656	0.7121	−4.391	1.13×10^{-5}	HLA-DRB1
6	rs6923504	32536164	C	ADD	1656	0.7135	−4.365	1.27×10^{-5}	HLA-DRB1
4	rs17012830	88670120	A	ADD	1650	0.6437	−4.242	2.22×10^{-5}	SPARCL1
6	rs4715377	13622276	G	ADD	1656	1.809	4.176	2.97×10^{-5}	GFOD1
2	rs10190416	36405331	G	ADD	1652	1.41	4.141	3.46×10^{-5}	CRIM1
18	rs238129	3478151	A	ADD	1657	1.359	4.131	3.62×10^{-5}	DLGAP1
17	rs12453363	45576919	A	ADD	1655	0.6792	−4.076	4.58×10^{-5}	PPP1R9B
13	rs12872448	98393248	A	ADD	1653	0.6773	−4.041	5.31×10^{-5}	DOCK9
7	rs805803	122842791	A	ADD	1653	1.449	4.032	5.52×10^{-5}	IQUB
2	rs968796	3872434	G	ADD	1657	1.35	3.976	7.02×10^{-5}	DCDC2C
10	rs10884381	108676055	G	ADD	1655	0.7411	−3.95	7.80×10^{-5}	SORCS1
3	rs9289026	116838866	G	ADD	644	0.5493	−3.917	8.98×10^{-5}	GAP43
8	rs1469039	140720961	A	ADD	1653	1.527	3.913	9.10×10^{-5}	KCNK9
10	rs10786981	108671720	A	ADD	1657	0.7437	−3.907	9.33×10^{-5}	SORCS1
17	rs8066520	24400717	A	ADD	1657	1.498	3.898	9.72×10^{-5}	DCC

within or near genes that have a significant role in the nervous system and neurodevelopment. The details are listed in **Table 1**.

RARE VARIANTS

We used the hg19 database as the standard for gene annotation. Excluding genes that do not have rare variants, we analyzed 8060 genes for high-IQ vs. low-IQ comparisons. The top 15 ranked genes are presented in **Table 2**. Genes that are functionally relevant to the nervous system and neurodevelopment are discussed below.

DISCUSSION

The AGP dataset consists of ASD probands and their parents sequenced using a GWAS platform. Its purpose is to explore the role of common variants in ASD by using a transmission disequilibrium test (TDT) approach. In this study, we focused on the probands themselves and excluded their parents. We speculated that by using a case-comparison design, we could potentially identify the specific variants that differentiate high- vs. low-functioning ASD individuals.

A total of 15 SNPs met the *p*-value threshold of 10^{-5} while 82 genes met the less stringent significance threshold of 10^{-4}. We then examined the properties of genes that contain or are close to these SNPs using the NCBI database. We were particularly interested in genes known to be related to neurological disorders and neurodevelopment. These genes, as well as their related biological functions are summarized in **Table 3**.

The most interesting finding is that three of the SNPs are included within the human leukocyte antigen (HLA) region on chromosome 6, very close to the gene HLA-DRB1, which was implicated in a paper by Torres et al. (2012) to be protective against ASD. All three of the SNPs (rs9268880, $p = 8.85 \times 10^{-6}$; rs6903608, $p = 1.13 \times 10^{-5}$; rs6923504, $p = 1.27 \times 10^{-5}$) near HLA-DRB1 are associated with lower IQ.

Table 2 | Rare variant results of high-IQ vs. low-IQ.

Gene	*p*-Value	*N.* marker test
LTA4H	0.000132	1
STEAP2	0.000201	2
ALK	0.000268	29
ZMYM4	0.000303	5
LINC00550	0.000316	1
FKTN	0.000402	2
KIAA0319L	0.000536	4
TFAP2E	0.000639	1
NRD1	0.000659	7
SEMA6A	0.000662	9
ACAD11	0.000769	1
UBA5	0.000769	1
SLC16A4	0.000782	2
RAB3B	0.000991	1

N. marker test is the number of markers to test for an association after excluding non-polymorphic or high missing rates markers.

Among the remaining genes, there are three general categories. The first category includes genes related to neurodevelopment. One of these is the gene DCDC2C, a member of the doublecortin gene family, which has been implicated in neuronal migration, neurogenesis, and retina development through regulation of cytoskeletal structure and microtubule-based transport. Mutations in genes of this family have been implicated in epilepsy

Table 3 | Summary of known biologically relevant genes found in common variant analysis.

Gene	Effect
HLA-DRB1	ASD
SPARCL1	Astroglial cells
gfod1	ADHD
crim1	CNS development
dlgap1	Schizophrenia
ppp1r9b	Dendritic spines
DOCK9	Bipolar
IQUB	Intelligence
dcdc2c	Neurogenesis
sorcs1	Memory
GAP43	Neurogenesis
DCC	Axon guidance

and developmental dyslexia, among other disorders (Dijkmans et al., 2010). Another gene of this class is GAP43, named growth associated protein 43 because it is expressed at high levels in neuronal growth cones during development and axonal regeneration, and considered a crucial component of regenerative response in the nervous system (Skene et al., 1986; Aigner et al., 1995). The third of these genes is DCC, which encodes a netrin 1 receptor that acts as a cue for axon growth and guidance (Forcet et al., 2002). The fourth gene, SPARCL1, has been implicated in multiple cellular processes during brain development. Specifically, SPARCL1 is prominently expressed in radial glia, where it terminate radial glial guided neuronal migration, and is further expressed in the proliferative ventricular zone (VZ) of the embryonic cortex (Weimer et al., 2008). Another gene, CRIM1 has also been implicated in central nervous system (CNS) development, possibly via growth factor binding (Kolle et al., 2000).

The second category contains genes that are related to neural function. PPP1R9B belongs to this category. This gene encodes spinophilin, which is a regulatory subunit of protein phosphatase-1 catalytic subunit (PP1) and is highly enriched in dendritic spines. Allen et al. (1997) suggested that spinophilin may serve as a neuronal targeting subunit for PP1 and might be responsive to neuronal inputs.

The third category contains genes linked to neurological conditions via bioinformatic methods, but has not yet been verified via biological experiments. These include GFOD1, which is associated with attention deficit hyperactivity disorder (ADHD), DLGAP1 which is associated with schizophrenia, DOCK9 associated with bipolar disorder, and SORCS1 which is associated with memory (Detera-Wadleigh et al., 2007; Lasky-Su et al., 2008; Reitz et al., 2011; Li et al., 2013). Interestingly, the SNP rs805803 is in close proximity (75 kb) to rs7791660, which was shown to be associated with mathematical ability (Docherty et al., 2010).

Considering rare variants, three genes are noteworthy. The first is ALK, which is an oncogene whose mutation also disrupts CNS development (de Pontual et al., 2011). The second is KIAA0319L located on chromosome 1, which has been identified as a candidate for dyslexia. This gene is expressed in the brain and, based on its structural similarities to the gene KIAA0319, has been suggested to play a role in neuronal migration (Couto et al., 2008). The third gene SEMA6A is expressed in developing neural tissue and is required for proper development of the thalamocortical projection (Leighton et al., 2001).

CONCLUSION

In this study, we used a case–control approach to investigate the association of genetic variants with IQ in the ASD population. We analyzed common variants and rare variants separately and in different ways, using a standard case–control association test implemented in PLINK for common variants, and the SKAT for rare variants. Considering their previously reported biological roles, we were able to identify several genes that are plausible candidates for involvement in brain development in ASD patients. To our knowledge, this is among the first studies that addresses this issue.

These genes are biologically relevant to CNS and neurodevelopment based on published literature, the most prominent examples being the genes KIAA0319L and HLA-DRB1. These genes warrant further investigation of their properties, both in regard to their connection with intelligence and relationship to ASD.

We acknowledge that the findings reported are preliminary, and it is possible that at least some of the associated genes are false positives. Thus, further molecular validations are warranted.

ACKNOWLEDGMENTS

Harold Z. Wang, Hai-De Qin, Wei Guo, and Yin Yao Shugart are all supported by the Intramural Research Programs of the National Institute of Mental Health (grant number MH002930-03). The datasets used for the analyses described in this manuscript were obtained from the database of Genotypes and Phenotypes (phs000267.v3.p2). We would like to thank the people of the Autism Genome Project for their generosity in providing the data to dbGaP. The views expressed in this presentation do not necessarily represent the views of the NIMH, NIH, HHS, or the United States Government.

REFERENCES

Aigner, L., Arber, S., Kapfhammer, J. P., Laux, T., Schneider, C., Botteri, F., et al. (1995). Overexpression of the neural growth-associated protein GAP-43 induces nerve sprouting in the adult nervous system of transgenic mice. *Cell* 83, 269–278. doi: 10.1016/0092-8674(95)90168-X

Allen, P. B., Ouimet, C. C., and Greengard, P. (1997). Spinophilin, a novel protein phosphatase 1 binding protein localized to dendritic spines. *Proc. Natl. Acad. Sci. U.S.A.* 94, 9956–9961. doi: 10.1073/pnas.94.18.9956

Anney, R. Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., et al. (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum. Mol. Genet.* 21, 4781–4792. doi: 10.1093/hmg/dds301

Couto, J. M., Gomez, L., Wigg, K., Cate-Crter, T., Archibald, J., Anderson, B., et al. (2008). The KIAA 0319-like (KIAA0319L) gene on chromosome 1p34 as a candidate for reading disabilities. *J. Neurogenet.* 22, 295–313. doi: 10.1080/01677060802354328

de Pontual, L., Kettaneh, D., Gordon, C. T., Oufadem, M., Boddaert, N., Lees, M., et al. (2011). Germline gain-of-function mutations of ALK disrupt central nervous system development.

Hum. Mutat. 32, 272–276. doi: 10.1002/humu.21442

Detera-Wadleigh, S. D., Liu, C. Y., Maheshwari, M., Cardona, I., Corona, W., Akula, N., et al. (2007). Sequence variation in DOCK9 and heterogeneity in bipolar disorder. *Psychiatr. Genet.* 17, 274–286. doi: 10.1097/YPG.0b013e328133f352

Dijkmans, T. F., van Hooijdonk, L. W., Fitzsimons, C. P., and Vreuqdenhil, E. (2010). The doublecortin gene family and disorders of neuronal structure. *Cent. Nerv. Syst. Agents Med. Chem.* 10, 32–46. doi: 10.2174/187152410790780118

Docherty, S. J., Davis, O. S., Kovas, Y., Meaburn, E. L., Dale, P. S., Petrill, S. A., et al. (2010). A genome-wide association study identifies multiple loci associated with mathematics ability and disability. *Genes Brain Behav.* 9, 234–247. doi: 10.1111/j.1601-183X.2009.00553.x

Forcet, C., Stein, E., Pays, L., Corset, V., Llambi, F., Tessier-Lavigne, M., et al. (2002). Netrin-1-mediated axon outgrowth requires deleted in colorectal cancer-dependent MAPK activation. *Nature* 417, 443–447. doi: 10.1038/nature748

Halfon, N., and Kuo, A. A. (2013). What DSM-5 could mean to children with autism and their families. *JAMA Pediatr.* 167, 608–613. doi: 10.1001/jamapediatrics.2013.2188

Kolle, G., Georgas, K., Holmes, G. P., Little, M. H., and Yamada, T. (2000). CRIM1, a novel gene encoding a cysteine-rich repeat protein, is developmentally regulated and implicated in vertebrate CNS development and organogenesis. *Mech. Dev.* 90, 181–193. doi: 10.1016/S0925-4773(99)00248-8

Lasky-Su, J., Neale, B. M., Franke, B., Anney, R. J., Zhou, K., Maller, J. B., et al. (2008). Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 147B, 1345–1354. doi: 10.1002/ajmg.b.30867

Leighton, P. A., Mitchell, K. J., Goodrich, L. V., Lu, X., Pinson, K., Shcerz, P., et al. (2001). Defining brain wiring patterns and mechanisms through gene trapping in mice. *Nature* 410, 174–179. doi: 10.1038/35065539

Li, J. M., Lu, C. L., Cheng, M. C., Luu, S. U., Hsu, S. H., and Chen, C. H. (2013). Genetic analysis of the DLGAP1 gene as a candidate gene for schizophrenia. *Psychiatry Res.* 205, 13–17. doi: 10.1016/j.psychres.2012.08.014

O'Connell, J. R., and Weeks, D. E. (1998). PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.* 63, 259–266. doi: 10.1086/301904

Reitz, C., Lee, J. H., Rogers, R. S., and Mayeurx, R. (2011). Impact of genetic variation in SORCS1 on memory retention. *PLoS ONE* 6:e24588. doi: 10.1371/journal.pone.0024588

Skene, J. H., Jacobson, R. D., Snipes, G. J., McGuire, C. B., Norden, J. J., and Freeman, J. A. (1986). A protein induced during nerve growth (GAP-43) is a major component of growth-cone membranes. *Science* 233, 783–786. doi: 10.1126/science.3738509

Szatmari, P. (1999). Heterogeneity and the genetics of autism. *J. Psychiatry Neurosci.* 24, 159–165.

Torres, A. R., Westover, J. B., and Rosenspire, A. J. (2012). HLA immune function genes in autism. *Autism Res. Treat.* 2012, 959073. doi: 10.1155/2012/959073

Weimer, J. M., Stanco, A., Cheng, J. G., Vargo, A. C., Voora, S., and Anton, E. S. (2008). A BAC transgenic mouse model to analyze the function of astroglial SPARCL1 (SC1) in the central nervous system. *Glia* 56, 935–941. doi: 10.1002/glia.20666

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 April 2013; accepted: 13 September 2013; published online: 18 October 2013.

*Citation: Wang HZ, Qin H-D, Guo W, Samuels J and Shugart YY (2013) New insights into the genetic mechanism of IQ in autism spectrum disorders. Front. Genet. **4**:195. doi: 10.3389/fgene.2013.00195*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 19 November 2013
doi: 10.3389/fgene.2013.00232

The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits

Yoshinobu Uemoto[1,2], Ricardo Pong-Wong[1], Pau Navarro[3], Veronique Vitart[3], Caroline Hayward[3], James F. Wilson[4], Igor Rudan[4], Harry Campbell[4], Nicholas D. Hastie[3], Alan F. Wright[3] and Chris S. Haley[1,3]*

[1] Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK
[2] National Livestock Breeding Center, Fukushima, Japan
[3] MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[4] Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Edited by:
Nathan Morris, Case Western Reserve University, USA

Reviewed by:
Xuefeng Wang, Harvard University, USA
Tao Feng, Case Western Reserve University, USA

***Correspondence:**
Chris S. Haley, MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK
e-mail: chris.haley@igmm.ed.ac.uk

Genome-wide association studies (GWAS) have provided valuable insights into the genetic basis of complex traits. However, they have explained relatively little trait heritability. Recently, we proposed a new analytical approach called regional heritability mapping (RHM) that captures more of the missing genetic variation. This method is applicable both to related and unrelated populations. Here, we demonstrate the power of RHM in comparison with single-SNP GWAS and gene-based association approaches under a wide range of scenarios with variable numbers of quantitative trait loci (QTL) with common and rare causal variants in a narrow genomic region. Simulations based on real genotype data were performed to assess power to capture QTL variance, and we demonstrate that RHM has greater power to detect rare variants and/or multiple alleles in a region than other approaches. In addition, we show that RHM can capture more accurately the QTL variance, when it is caused by multiple independent effects and/or rare variants. We applied RHM to analyze three biometrical eye traits for which single-SNP GWAS have been published or performed to evaluate the effectiveness of this method in real data analysis and detected some additional loci which were not detected by other GWAS methods. RHM has the potential to explain some of missing heritability by capturing variance caused by QTL with low MAF and multiple independent QTL in a region, not captured by other GWAS methods. RHM analyses can be implemented using the software REACTA (http://www.epcc.ed.ac.uk/projects-portfolio/reacta).

Keywords: common and rare variants, GWAS, regional heritability mapping, multiple independent effects, missing heritability

INTRODUCTION

Genome-wide association studies (GWAS) have provided valuable insights into the genetic basis of complex traits. However, the reported SNPs associated with a trait typically explain only a small proportion of genetic variance. For example, the heritability of human height is about 80% (Visscher et al., 2008), but the SNPs significantly associated with height explain only 10% of the phenotypic variance (Lango Allen et al., 2010). This has been called the "missing heritability" problem (Maher, 2008). Recently, Yang et al. (2010) showed that 45% of the phenotypic variance for human height is accounted for by common SNPs, and the difference between 10 and 45% was due to many SNPs with small effects that fail to reach significance in GWAS. Yang et al. (2010) suggested that the remaining variance evaded capture due to imperfect linkage disequilibrium (LD) between the genotyped SNPs and causal variants. Causal variants may have lower minor allele frequency (MAF) than genotyped SNPs if they are subject to purifying natural selection. In this case the variation explained by the genotyped SNPs will be lower than that due to causal variants because of low LD. A pressing need is analytical approaches adapted to capturing genetic variation due to causal variants with low MAF.

Recent studies have shown that multiple independent loci with different allele frequencies and effects are often located on the same gene region or narrow segment region. For example, seven independent alleles at 8q24 region affect prostate cancer (Haiman et al., 2007), three at the IRF5 gene affect systemic lupus erythematosus (Graham et al., 2007), and two at the IL23R gene affect Crohn's disease (Duerr et al., 2006). Such loci may escape detection by single SNP analyses if the individual allele effects are not large enough to be detected even though the cumulative effect of the whole locus on trait variance is quite large. An alternative method to analyze GWAS data is to consider an association between a trait and a composite P-value generated by all markers within a segment of the genome or a gene region, as opposed to individual SNPs. For a gene region, this method is called gene-based association (Neale and Sham, 2004), and can potentially increase the power to identify a causal gene that harbors several functional alleles. A new variance component approach called regional heritability mapping (RHM) that screens the genome by analyzing small regions has been suggested to capture more of the missing genetic variation (Nagamine et al., 2012). In RHM, a mixed model framework based on restricted maximum likelihood (REML) is used, and two variance components, one contributed by the whole genome and a second one by a specific genomic region, are fitted in the model to estimate genomic and regional heritabilities, respectively. RHM facilitates the capture of genetic variation that is associated with each segment of the genome by combining the effects of both common and rare variants in a region. By analyzing real data, Nagamine et al. showed that the results of RHM are correlated with results from GWAS but capture more of the missing genetic variation and identify additional quantitative trait loci (QTL).

The objective of this study is to investigate the effectiveness of RHM to capture QTL variance that is potentially not detected by single-SNP GWAS and gene-based association analyses. Such variance may be due to low MAF alleles and multiple independent QTL with small effects located on a narrow genomic region. We investigated the power to detect significant regions and accuracy of estimating regional heritability using simulation based on real genotype data from a human population. We used imputation to generate a dense map of SNPs from which to randomly select subsets at different frequencies to represent causative variants (QTL) in our simulations, using only the genotyped SNPs in the analyses. We studied the impact of different window sizes in RHM on its power and accuracy, and compared them to those of other methods that include single-SNP GWAS and a range of gene-based association approaches under several different scenarios. In addition, we also applied a RHM to analyze three eye traits to evaluate the effectiveness of this method in real data analysis.

MATERIALS AND METHODS

POPULATION AND SNP ARRAY INFORMATION USED IN THE SIMULATION STUDY

Samples were available from two Croatian cohorts recruited from two Dalmatian islands, Vis and Korcula, and both cohorts were approved by the Ethical Committee of the Medical School, University of Zagreb and the Multi-Centre Research Ethics Committee for Scotland. All participants gave written informed consent. The cohorts are usually referred to as CROATIA-Vis and CROATIA-Korcula, but will be referred to as Vis and Korcula in the remaining of the manuscript.

DNA samples were genotyped using the Illumina Human Hap300 (282,415 autosomal SNPs) for Vis and Illumina CNV370 (302,507 autosomal SNPs) for Korcula. Our quality control protocol excluded SNPs with MAF <0.0005, call rate <0.98 or Hardy–Weinberg equilibrium (HWE) P-value $< 1.0 \times 10^{-6}$. The exclusion criterion for individuals was call rate <0.97. A total of 269,706 SNPs on autosomal chromosomes were common to Vis and Korcula samples and were used in this study. In total, 953 individuals passed all quality control thresholds from Vis and 898 from Korcula, and the total of 1851 individuals were then used in the simulation study.

SNP IMPUTATION FOR SIMULATION ANALYSIS

SNPs were imputed to provide a dense map from which to select simulated causative variants (QTL). SNP imputation was performed using the IMPUTE2 program (Howie et al., 2009), incorporating 1000 Genomes Phase I (interim) data as reference panel for the Vis and Korcula genotypes, respectively. This imputation yielded posterior probabilities for genotypes at ~35 million SNPs, and an estimate of imputation quality [IMPUTE2-info score ranging from 0 (low confidence) to 1 (high confidence)]. Imputed SNPs were assigned to one of two groups depending on their IMPUTE2-info score (high_info group: $0.7 \leq$ IMPUTE2-info score ≤ 1.0 and low_info group: $0.0 \leq$ IMPUTE2-info score ≤ 0.5) in both populations. IMPUTE2 gives posterior probabilities for all three genotypes at each locus for each individual. Individual genotypes at each imputed SNP locus were randomly assigned according to the posterior probabilities for the three genotypes from IMPUTE2. These imputed SNPs were then assessed by the exclusion criteria of very rare MAF <0.0005 and HWE P-value $< 1.0 \times 10^{-6}$. The total number of selected SNPs were 3,793,540 SNPs in the low_info group and 6,704,137 SNPs in the high_info group. Comparison of imputed SNPs in the high_info group with the same SNPs genotyped on a commercial exome array indicates that the LD structure of the real population is well-represented by the imputed SNPs (see **Figure S1** in Supplementary Material).

GENERATING PHENOTYPES UNDER THE NULL HYPOTHESIS

We simulated phenotypes under the null hypothesis based on the observed genotype data of 1851 individuals at 269,706 SNPs. The phenotypes under the null hypothesis were simulated using a polygenic model in which all SNPs were assumed to have a very small effect on the phenotype. The polygenic model was $y_i = \sum_j^n x_{ij}b_j + e_i$, where x_{ij} is the genotype for j-th causal variant of the i-th individual (coded as 1, 2, or 3), b_j is the allele effect of the j-th causal variant generated from $N(0, 1)$, and e_i is the residual effect generated from $N(0, \sigma_g^2(1/h^2 - 1))$. σ_g^2 is the total genetic variance of $\sum_j^n x_{ij}b_j$ and h^2 is the setting value of genome heritability. Three setting values of genome heritability ($h^2 = 0.20$, 0.40, and 0.80) were used for generating phenotypes. These generated phenotypes were under the null hypothesis of no phenotype-window correlation (i.e., there was no significant effect for RHM), and were then used for simulation analysis (see the details in **Figure S2** in Supplementary Material).

SIMULATION DESIGN AND ANALYSES

The genotyped and imputed SNPs were assigned to genomic regions on the basis of their location in 1000 Genomes Phase I (interim) information. Once the polygenic background was simulated based on the genotyped SNPs as described above, we added to it regional effects, based on the simulated genotypes at imputed SNPs. For imputed SNPs, two MAF categories were defined as low MAF (MAF < 0.10) and high MAF (MAF ≥ 0.10), respectively. We carried out simulations in the high_info and low_info imputed SNPs categories. The parameters considered in the simulation are summarized in **Table 1**, and shown in detail below.

Table 1 | Settings criteria in the simulation study.

Condition	Criteria	
	Low_info group	**High_info group**
Minor allele frequency	Low MAF (0.0005 < MAF < 0.10)	Low MAF (0.0005 < MAF < 0.10), high MAF (0.10 ≤ MAF ≤ 0.50)
Number of QTL	1, 5, 10	1, 5, 10
QTL heritability	0.05	0.025, 0.050
Genome heritability	0.20, 0.40, 0.80	0.20, 0.40, 0.80

The division of the genome into regions was based on numbers of genotyped markers. A window containing 100 adjacent genotyped SNPs was named as win100. A total of 2686 win100 without overlap covered the autosomes and from these 300 win100s were randomly picked for the simulation analysis. Each win100 was divided equally into 10 10-SNP-windows (named as win10). One win10 was randomly selected from the six centermost win10s of a win100, and assumed as a gene region (i.e., we simulated causal variants in the chosen win10). This gene region contained at least 10 imputed SNPs with high MAF or low MAF. Of these 1, 5, or 10 with either high MAF or low MAF were randomly selected and assumed as QTL with joint heritabilities of either 0.025 or 0.05, which are based on the proportion of total genetic variance generated under the null hypothesis. The effect of these selected SNPs was generated, and then added to the phenotypic value generated under null hypothesis and an error value to generate a new phenotypic value with genome heritability (0.20, 0.40, and 0.80). Each selected SNP had an equal effect (and a randomly selected effect direction) that contributed to the total ("joint") QTL variance. Each win100 was divided equally into 2, 5, and 10 windows, and each window with 50 SNPs (named as win50), 20 SNPs (named as win20) and 10 SNPs, respectively, was then used to calculate the power to detect QTL and estimate regional heritability in order to assess the optimum analysis window size for these simulated data. Average window length across all autosomal chromosomes was 1030.2 kbp for win100, 515.1 kbp for win50, 206.0 kbp for win20, and 103.0 kbp for win10.

A total of 18 RHM analyses were performed per 100-SNP window (1 win100, 2 win50, 5 win20, and 10 win10 analyses), and a *P*-value of win100 and the minimum *P*-values results of win50, win20, and win10 were selected in each window size. To determine the threshold value of win100, win50, win20, and win10, a Bonferroni correction was applied by using 2686 windows, 5372 windows, 13,430 windows, and 26,860 windows, respectively. The power to achieve 5% genome-wide significance was calculated as the proportion of replicates with a significant window for each window size, genomic heritability, number of QTL, IMPUTE2-info score levels, MAF, and QTL heritability. The regional heritability and minimum *P*-value were also computed in all replicates for win100, win50, win20, and win10, and the average value of estimated regional heritability in all simulation replicates was calculated for each window size.

We wanted to compare the power and estimated regional heritability of RHM and a range of single SNP or gene-based association methods. We used two single-SNP GWAS analyses based on the Genome-wide rapid association using mixed model and regression (GRAMMAR) method (Aulchenko et al., 2007) and the genome-wide efficient mixed-model association (GEMMA) method (Zhou and Stephens, 2012). GRAMMAR is a two-step method that first estimates the residuals from mixed model without a SNP effect and then treats these residuals as corrected phenotypes for GWAS by simple linear regression. GEMMA is an exact mixed model approach that tests for association efficiently by using the mixed model with a SNP effect at one step. The whole genomic relationship matrix used in RHM was also used to perform the GRAMMAR and GEMMA analyses. The minimum *P*-values of GWAS were recorded in each win100 replicate. The *P*-value of thresholds for genome-wide significance came from the Bonferroni correction accounting for 268,600 SNPs, and the power to achieve 5% genome-wide significance was calculated as the proportion of replicates with a significant association. The heritability at the most significant SNP was calculated assuming Hardy–Weinberg proportions for the SNP genotypes; SNP heritability at the SNP with the minimum *P*-value, h^2_{SNP}, was calculated as $h^2_{\mathrm{SNP}} = 2p(1 - p)b^2/\sigma^2$, where p was the SNP MAF, b was the SNP effect (regression coefficient estimated from the analysis), and σ^2 was the residual variance for GRAMMAR and the phenotypic variance for GEMMA (Falconer and Mackay, 1996). An average value of SNP heritability across simulation replicates was calculated for the GRAMMAR and GEMMA analyses.

To investigate the power of RHM and other GWAS methods that consider several variants in a (gene) region simultaneously, we analyzed the data using three recently reported gene-based association tests. These GWAS methods implement gene-based association approaches which consider an association between a trait and all markers within a gene rather than each marker individually, and generate one new *P*-value as a representative value of the gene. These methods can account for the number of independent effects within a gene. Three gene-based association approaches were as follows:

A versatile gene-based test for genome-wide association studies (VEGAS): VEGAS proposed by Liu et al. (2010) sums the SNP-based chi-square test statistics from all the SNPs within a gene and then corrects the sum for LD to generate a gene-based test statistic. VEGAS requires the pairwise LD correlation matrix of the SNPs from HapMap genotype information calculated by the PLINK software (Purcell et al., 2007). In this study, a custom set of individual genotypes was used to estimate an LD correlation matrix by using genotype data from our population, instead of HapMap genotype information, because the selected region is not a gene locus. The VEGAS test was performed by using the *P*-values obtained from GEMMA analysis.

Sequence kernel association test (SKAT): As a kernel machine based test, SKAT proposed by Wu et al. (2011) aggregates genetic information across the region using a kernel function and uses a computationally efficient variance component test to test for association. This method has an advantage if the causal mutation is rare. SKAT's power is greater than that of several burden tests such as the cohort allelic sum test (Morgenthaler and Thilly, 2007). In this study, the GRAMMAR method was used obtain a phenotype adjusted for population stratification that was then used in SKAT analysis. We used the default beta (1, 25) weight in this study.

Canonical Correlation Analysis (CCA): Tang and Ferreira (2012) explored the gene-based association test using canonical correlation to test multiple SNPs for association with a single or multiple phenotypes measured in unrelated individuals. CCA removes any multicollinearity between SNPs by accounting for pairwise (LD) correlations and variance inflation factor, calculates canonical correlations between selected SNPs and phenotypes, and tests the significance of all canonical correlations. Tang and Ferreira (2012) showed that the power of this method was greater than that of GWiS (Huang et al., 2011) and single-SNP GWAS. We used the GRAMMAR-adjusted phenotypes as input in the CCA analysis.

In our simulated 300 "gene regions," for the high_info group, RHM with win10, single-SNP GWAS by GEMMA and three gene-based association approaches were performed, and the power to achieve 5% genome-wide significance was calculated. For single-SNP GWAS, only GEMMA was performed in this analysis, because the power to detect QTL using GEMMA was greater than that obtained using GRAMMAR in all simulations (see Results), and the minimum P-value was calculated in a gene region. For the low_info group, there was no significant result for any methods in all simulations (see Results), and therefore results of these analyses are not presented.

REGIONAL HERITABILITY MAPPING

We performed RHM based on two-step variance component method described by Nagamine et al. (2012) using ASReml software (Gilmour et al., 2006). The mixed model is as follows;

$$\mathrm{y} = 1_{\mathrm{n}}\mu + \mathrm{Xu} + \mathrm{Zw} + \mathrm{e} \qquad (1)$$

where **y** is the vector of phenotypic values and **X** and **Z** are the design matrices for random effects. 1_n is a vector of 1s and μ is the mean. $u \sim N(0, G\sigma_u^2)$ is the whole genomic additive genetic effect, $w \sim N(0, Q\sigma_w^2)$ is the regional genomic additive genetic effect and $e \sim N(0, I\sigma_e^2)$ is the residual effect. Matrices **G**, **Q**, and **I** are a whole genomic relationship matrix, a regional genomic relationship matrix using SNPs within the short region of genome, and an identity matrix, respectively. Elements of matrices **G** and **Q** are based on genomic kinship and inbreeding coefficient between individual i and j using identity by state (IBS), and element f_{ij} of both **G** and **Q** is defined as follows,

$$f_{ij} = \frac{2}{n}\sum_{k=1}^{n}\frac{(x_{ik}-p_k)(x_{jk}-p_k)}{p_k(1-p_k)}, \quad (i \neq j)$$

$$f_{ij} = 1 + \frac{1}{n}\sum_{k=1}^{n}\frac{Obs(\#hom)_{ik} - E(\#hom)_k}{1 - E(\#hom)_k}, \quad (i = j)$$

where x_{ik} (x_{jk}) is the genotype of the i-th (j-th) person at the k-th SNP (coded as 0, 0.5, and 1 for AA, AB, and BB, respectively). Here n represents the total genomic SNPs for matrix **G** or the number of SNPs in the region for matrix **Q**. The frequency p_k is for the B allele at the k-th SNP, and n is the number of SNPs. $Obs(\#hom)_{ik}$ and $E(\#hom)_k$ are the observed and expected number of homozygous genotypes in the i-th person at the k-th SNP. Regional heritability h^2_{RH} and genome heritability h^2_{GH} are calculated as follows,

$$h^2_{RH} = \frac{\sigma_w^2}{\sigma_u^2 + \sigma_w^2 + \sigma_e^2}$$

$$h^2_{GH} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_w^2 + \sigma_e^2}$$

where σ_u^2, σ_w^2, σ_e^2 are whole genome additive genetic variance, regional genomic additive variance, and residual variance, respectively.

TEST STATISTICS AND THEIR DISTRIBUTION

To test for the presence of QTL effect against the null hypothesis (no regional variance) at a test region (window), the likelihood ratio test statistics (LRT) $= -2\ \ln(L_0\text{–}L_1)$ was calculated, where L_0 and L_1 represent the likelihood values under the hypothesis of no presence (H_0) and presence (H_1) of regional variance, respectively. The L_1 was calculated by using the model (1), and the L_0 was calculated by using the following mixed model (2) that does not include regional genomic additive genetic effect from the model (1).

$$\mathrm{y} = 1_n\mu + \mathrm{Xu} + \mathrm{e} \qquad (2)$$

Statistical theory states that the LRT follows a χ^2 distribution with the degrees of freedom equal to the number of random parameters being tested (Wilks, 1938). However, for testing a single variance component in a REML context, the asymptotic distribution of the LRT under the null hypothesis follows a mixture of χ^2 distributions with different degrees of freedom (e.g., Visscher, 2006). Hence for the RHM method, the LRT follows a 50:50 mixture distribution, where one mixture component is a peak at 0 and the other component is a χ^2_1 distribution (Nagamine et al., 2012). In this study, phenotypes under the null hypothesis were generated, and LRTs for each non-overlapping win100 were calculated to obtain an empirical distribution of $-\log_{10}$(P-value) under the null hypothesis and compared with the theoretical distribution. The results show that the 50:50 mixture distribution is more appropriate (see **Figure S2** in Supplementary Material).

ANALYSES OF REAL POPULATION DATA ON THREE BIOMETRICAL EYE TRAITS

To illustrate the applicability of RHM in the real population data, we considered three eye traits measured in four populations [three Croatian (CROATIA-Vis, CROATIA-Korcula, CROATIA-Split) and one from Orkney (ORCADES), including axial length (AL), central corneal thickness (CCT), and spherical equivalent refraction (SER)]. These are quantitative endophenotypes related to common eye disorders; AL and SER are related to incidence of myopia and hyperopia and CCT is related to the incidence of corneal disorders and probably glaucoma. All cohorts have contributed to large single-SNP GWAS meta-analyses efforts studying these phenotypes (Lu et al., 2013; Verhoeven et al., 2013). All the Croatian cohorts (that will be referred from here as Vis, Korcula, and Split) received ethical approval from the Ethics

Committee of the Medical School, University of Split and the NHS Lothian (South East Scotland Research Ethics Committee). The ORCADES cohort, referred to as Orkney from now on received ethical approval from the NHS Orkney Research Ethics Committee and North of Scotland Research Ethics Committee. All studies followed the tenets of the Declaration of Helsinki and all participants gave written informed consent. A total of 2245 individuals for AL, 2261 individuals for CCT, and 2251 individuals for SER were measured, and descriptive statistics for these three traits were shown in **Table S1** in Supplementary Material. The Vis cohort genotyping was performed using the Illumina HAP300v1 SNP array, the Korcula and Split cohorts were genotyped using the Illumina HAP370CNV SNP array, and the Orkney cohort used the Illumina HumanHap300 beadchip. A total of 3210 individuals in four populations were genotyped (the number of individuals in each population is shown in **Table S1** in Supplementary Material). A total of 344,065 SNPs with overlap among four populations were assessed by the same protocol as above, and 272,315 SNPs on autosomal chromosomes passed the quality control (the number of SNPs in each chromosome is shown in **Table S2** in Supplementary Material). We performed single-SNP GEMMA analysis and RHM across the whole genome to detect any significant regions. To account for non-genetic effects in these two analyses, population, and sex were included as fixed effects, and age (and height in AL) was used as a covariate in these analyses. The significance threshold value for single-SNP GEMMA was determined by Bonferroni correction with 272,315 SNPs. For RHM, we applied a two-step approach to reduce computation. At first, RHM with win100 was performed across all autosomes. The window was shifted every 50 SNPs to overlap a region, and a total of 5412 windows were tested across chromosomes. In the second step, the top 100 win100s with higher LRT were selected from all 5412 windows, and then each win100 was divided equally into 10 win10s and 5 win20s, and RHM with win10 and win20 was performed. To evaluate the power of other GWAS methods, the windows with *P*-value $< 1.0 \times 10^{-5}$ in RHM analyses were then analyzed by three gene-based association approaches (VEGAS, CCA, and SKAT), and the window was assumed as a "gene region" in these methods. The methodologies of these gene-based association approaches were the same as above. To determine the significance threshold value of RHM and the three gene-based association approaches with win20 and win10, the Bonferroni correction was applied by using 27,060 and 54,120 windows, respectively.

RESULTS

IMPUTED SNPs

After removing markers with the exclusion criteria we have described, a total of 6,704,137 SNPs in the high_info group and 3,793,540 SNPs in the low_info group were available. **Table 2** shows the summary of imputed SNP number within a win10 region for low_info and high_info groups. In the low_info group, almost all SNPs had low MAF, and therefore only SNPs with low MAF were used in the simulation. In the high_info group, 45% of SNPs had low MAF and 55% of SNPs had high MAF. The density distributions of MAF for imputed SNPs within the high_info group and for genotyped SNPs are plotted in **Figure 1**. The MAF distribution shows a very low ratio of genotyped to imputed SNPs at low MAF, pointing to the difficulty of capturing genetic variance if imputed SNPs at low MAF are assumed to be QTL.

Table 2 | Total number of imputed SNPs and summary of SNP number in a window containing 10 genotyped SNPs (win10) for two different IMPUTE2-info scores in the simulation study.

IMPUTE2-info score	Total number of SNPs		Number of SNPs in win10		
			Total	Low MAF	High MAF
Low_info group	3,793,540	Mean	141	140	1
		Max	3749	3241	508
		Min	0	0	0
High_info group	6,704,137	Mean	250	112	138
		Max	5126	1965	3161
		Min	0	0	0

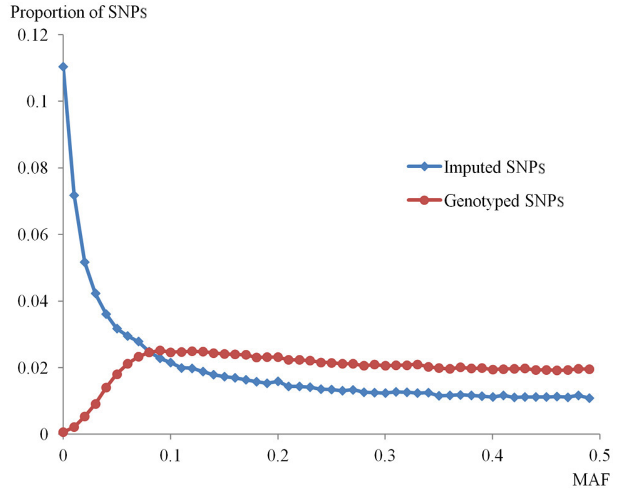


FIGURE 1 | Distribution of minor allele frequencies (MAFs) for genotyped and imputed SNPs. The distributions of MAF for imputed SNPs within the high_info group and for genotyped SNPs in this population are shown. The x-axis indicates the MAF of both groups of SNPs. The y-axis represents the proportion of SNPs in each MAF category.

THE POWER OF RHM AND SINGLE-SNP GWAS IN THE 100-SNP WINDOW

In the low_info group, there was no significant replicate in all simulations, indicating that the power was low for both methods. For the high_info group, the power to detect QTL for the phenotype with genome heritability 0.4 is shown in **Figure 2**. For RHM, as the number of QTL increased, the power to detect QTL was almost constant in all simulated scenarios, except when the QTL had low MAF and 0.05 QTL heritability. For RHM, using smaller window sizes (win10 and win20) yielded greater power than using

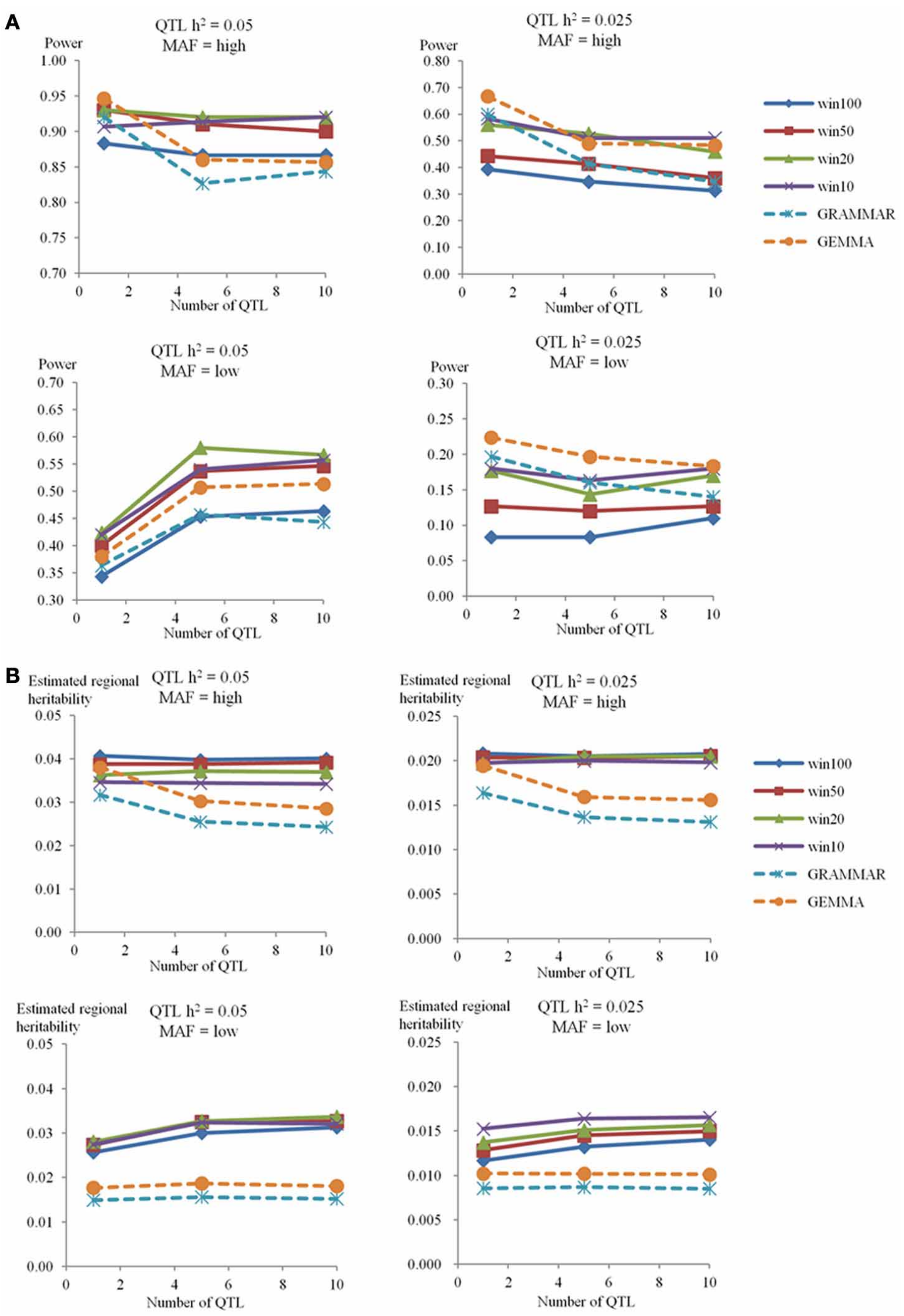


FIGURE 2 | The power to achieve 5% genome-wide significance and estimated regional heritability in the 100-SNP window. The powers **(A)** and estimated regional heritabilities **(B)** in the simulation study for genome heritability 0.4 were calculated by RHM with four different window sizes (100 SNPs as win100, 50 SNPs as win50, 20 SNPs as win20, and 10 SNPs as win10), and two single-SNP GWAS methods (GRAMMAR and GEMMA) in the different situations. The number of QTL is on the x-axis, and the power to detect QTL **(A)** or the estimated regional heritability **(B)** are on the y-axis. Each graph shows the different situations for genome heritability 0.4 (QTL heritability is 0.05 or 0.025, and MAF is high or low).

larger window sizes (win50 and win100), and the difference in power among window sizes was almost the same for different numbers of simulated QTL. There was no significant difference in power among simulations with different genome heritability (see **Figure S3** in Supplementary Material). For single-SNP GWAS, as the number of QTL increased, the power to detect QTL decreased, except for QTL with low MAF and 0.05 QTL heritability, where it increased as was also the case for RHM. Changes in genome heritability, had no large impact in power for the GEMMA analyses, but the power of GRAMMAR analyses decreased as the genome heritability increased (see **Figure S3** in Supplementary Material). The difference in power between

RHM and single-SNP GWAS varied with QTL MAF. For high MAF QTL, the power of single-SNP GWAS was greater than that of RHM when the number of QTL was one. But as the number of QTL increased, the power of RHM was greater than that of single-SNP GWAS. For low MAF, the power of RHM was higher than that of single-SNP GWAS for 0.05 QTL heritability, but lower for 0.025 QTL heritability for all numbers of QTL.

THE ESTIMATED REGIONAL HERITABILITY OF RHM AND SINGLE-SNP GWAS IN THE 100-SNP WINDOW

In the high_info group for a genome heritability of 0.40, the estimated regional and genome heritabilities are shown in **Figure 2** and **Table S3** in Supplementary Material, respectively. For all methods the mean heritability captured was less than that actually simulated but RHM generally captured a substantially greater proportion than GEMMA, the best of the single SNP methods, although RHM and GEMMA captured a similar proportion of simulated heritability when there was a single high MAF QTL. For RHM, as the number of QTL increased, the estimated regional heritability remained almost constant (averaging about 80% of the amount simulated) for QTL with high MAF, but it increased slightly with the number of QTL at low MAF (averaging around 60% of the amount simulated). Overall, there were no large differences in the amount of heritability captured by RHM using different window sizes and no overall trend in the size of window capturing most heritability. There was also no big difference for estimated regional heritability with different genome heritabilities (see **Figure S4** in Supplementary Material). For single-SNP GWAS, as the number of QTL increased, the estimated regional heritability decreased for high MAF QTL, but was almost constant for low MAF QTL. On average GEMMA estimates of the QTL heritability were almost 80% of that simulated for a single high MAF QTL, but the estimates dropped to around 60% of the simulated values for 5 or 10 high MAF QTL and were only about 40% of the simulated values for 1, 5, and 10 simulated low MAF QTL. Varying the genome heritability produced no big difference in the QTL heritability captured by GEMMA. As the simulated genome heritability increased, the regional heritability estimated by GRAMMAR decreased (see **Figure S4** in Supplementary Material). **Table S3** in Supplementary Material also showed the genome heritability estimated by model (2). The genome heritability estimated for high MAF was close to the simulated value, but genome heritability was underestimated for low MAF QTL.

THE POWER OF RHM AND OTHER METHODS IN THE GENE REGION

For the genome heritability of 0.40, **Figure 3** shows the results of power for RHM with win10, single-SNP GWAS (GEMMA), and three gene-based association approaches (VEGAS, SKAT, and CCA) in a gene region. The power of RHM was higher than that of all other methods for most simulation conditions, with the exception of the single QTL with 0.025 heritability, for which GEMMA had slightly higher power than RHM. As the number of QTL increased, the power to detect QTL generally remained almost constant or slightly reduced in all methods, but it increased slightly for all methods with low MAF and 0.05 QTL heritability. As for the other methods, CCA was the most powerful for QTL with high heritability, and GEMMA was the most powerful for QTL with low heritability. The power of VEGAS and SKAT was the lowest for QTL with low MAF and high MAF, respectively. The magnitude of the genome heritability had no great impact of on the power of RHM or GEMMA based methods (single-SNP GWAS and VEGAS), but the power of methods using GRAMMAR-adjusted phenotype (SKAT and CCA) decreased as the genome heritability increased (see **Figure S5** in Supplementary Material).

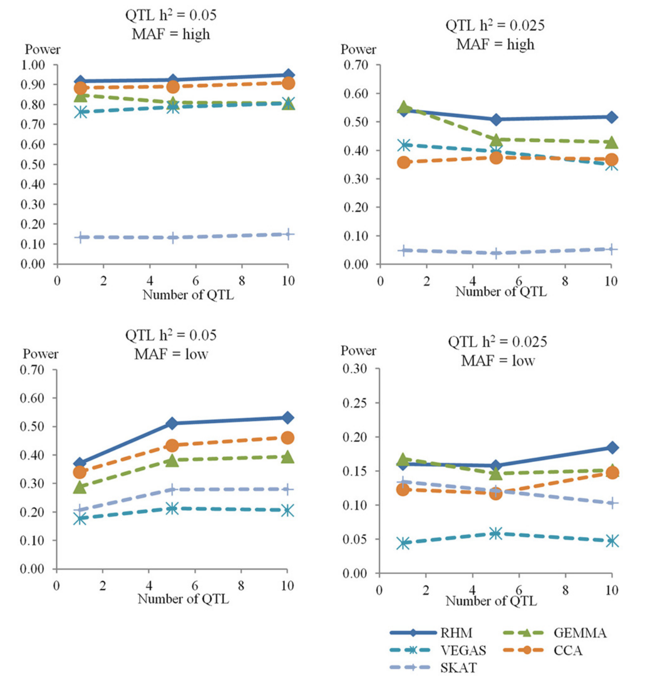


FIGURE 3 | The power to achieve 5% genome-wide significance in the gene region. The powers in the simulation study for genome heritability 0.4 were calculated by RHM with window size 10 (RHM), single-SNP GWAS (GEMMA), and three gene-based association approaches (VEGAS, CCA, and SKAT) in the different situations. The number of QTL is on the x-axis, and the power to detect QTL is on the y-axis. Each graph shows the different situations for genome heritability 0.4 (QTL heritability is 0.05 or 0.025, and MAF is high or low).

The Venn diagrams for comparisons of the significantly associated regions identified by three different methods (RHM, GEMMA, and gene-based association approach) are shown in **Figure 4**. As the number of QTL increased, the probability that QTL were detected only by RHM increased. For GEMMA and gene-based association approaches, as the number of QTL increased, the power to detect QTL by each method increased for low MAF but decreased or stayed constant for high MAF. In addition, RHM identifies some additional loci, even where GEMMA has higher power than RHM as is the case for the single QTL with 0.025 heritability. By using RHM and GEMMA, more than 90% of the QTL which were detected in all methods can be captured in all simulations.

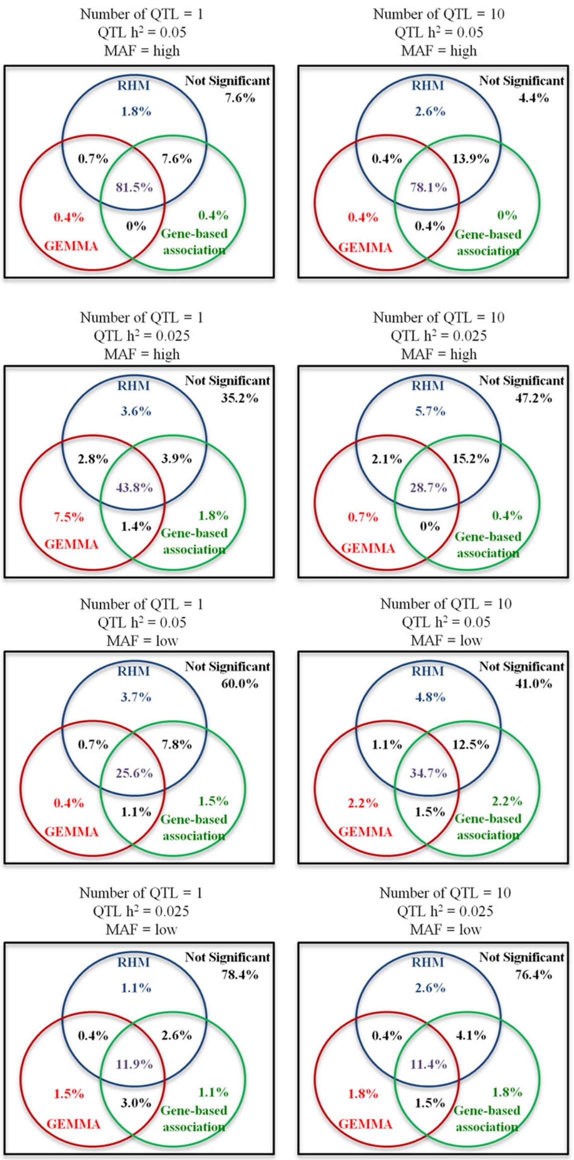


FIGURE 4 | Venn diagrams for comparisons of the significantly associated gene regions identified by three different methods. The percentages in circles are the proportions of the significantly associated gene regions identified by three different methods: blue circles for RHM, read circles for single-SNP GWAS (GEMMA), and green circles for gene-based association approaches including VEGAS, CCA, and SKAT (Gene-based association). Percentages in purple represent the significant replicates shared by all three methods, percentages in black represent the significant replicates shared only by two methods, and percentages in other colors are the significant replicates identified only by the corresponding method. The percentages in the squares are the proportion of not-significantly associated replicates. Each Venn diagram shows the different situations for genome heritability 0.4 (Number of QTL is 1 or 10, QTL heritability is 0.05 or 0.025, and MAF is high or low).

ANALYSES OF REAL POPULATION DATA ON THREE BIOMETRICAL EYE TRAITS

Quantile-quantile plots for the GEMMA results shown in **Figure S6** in Supplementary Material demonstrate that population stratification was successfully accounted by this method. Genome-wide plots of *P*-values for AL, CCT, and SER by GEMMA are shown in **Figure S7**. For GEMMA, two significant SNPs were detected for CCT, the significant SNPs being rs1536482 (P-value $= 1.0 \times 10^{-7}$) on chromosome 9 and rs12447690 (P-value $= 3.3 \times 10^{-11}$) on chromosome 16. These hits represent the *RXRA-COL5A1* and *ZNF469* loci as reported by Vitart et al. (2010) and both replicated in multiple studies (Lu et al., 2013). For RHM, the top 100 win100s with higher LRT were selected for further analysis using win10 and win20. The results from these latter analyses that gave P-values $< 1.0 \times 10^{-5}$ are given in **Figure 5** and **Table 3**. For AL, there was no significant region, but a novel region with a P-value $< 1.0 \times 10^{-5}$ was detected on chromosome 10 by RHM with win10. For CCT, there was a significant region on chromosome 16 that included the significant SNP detected by GEMMA and with the *ZNF469* gene located near this region (**Figure S8** in Supplementary Material). For SER, there were two significant novel regions [unreported in the largest single-SNP GWAS meta-analyses published by Verhoeven et al. (2013); Kiefer et al. (2013)] detected by RHM with win20 on chromosome 2 and with win10 on chromosome 10, this latter was the same region as detected for AL, a trait phenotypically correlated to SER. On chromosome 2, the two genes (*CREG2* and *RNF149* loci) and four genes (CREG2, *RNF149*, *SNORD89*, and *C2orf29* loci) were located within win10 with the lowest *P*-value and the significant win20, respectively, and there was no coding gene in the significant region of chromosome 10 (**Figure S8** in Supplementary Material). On chromosome 9, the *RXRA-COL5A1* locus detected by GEMMA was not significant by RHM. To evaluate the power of three gene-based association approaches, these windows were analyzed, and the results were shown in **Table 3**. The significant region was detected by VEGAS on chromosome 16, but there were no other significant regions detected by VEGAS or other methods (SKAT and CCA).

DISCUSSION

Nagamine et al. (2012) introduced a new variance component-based mapping methodology, referred to as regional genomic relationship mapping or RHM, to localize some of the genetic variation that cannot be detected by single-SNP GWAS analyses. Here, we study in depth the implementation and power of RHM in a range of circumstances. In particular, we describe the power to detect regions harboring different numbers of QTL with different MAFs (common and rare) explaining different proportions of the trait variance and the accuracy for estimating regional heritability. We also compare these results to those obtained using a range of single-SNP GWAS and gene-based association approaches. In addition, we applied RHM to the analysis of eye traits to show the effectiveness of this method.

Our simulation was based on real genotype data from a human population in an attempt to accurately account for LD found in real populations between marker SNP and QTL. Using imputed SNPs as the simulated QTL allowed us to generate a number of QTL in a region at both high and low MAF whilst retaining the genotyped SNPs as the markers for analysis. As might be expected, in our analyses of QTL based on poorly imputed SNPs

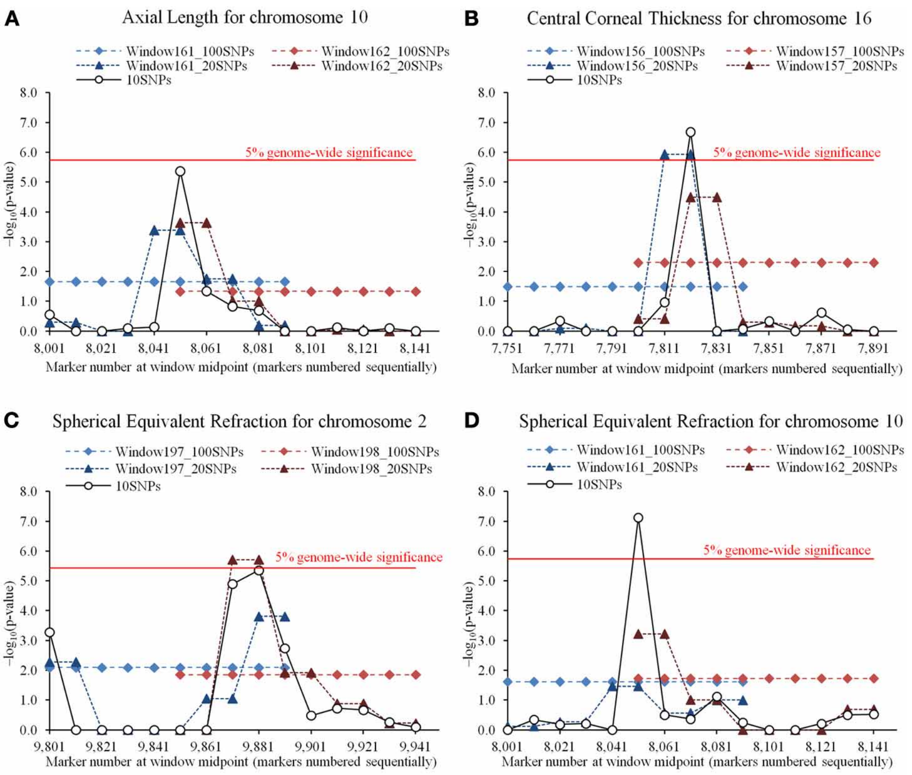


FIGURE 5 | Comparisons of regional heritability mapping (RHM) among different window sizes on a significant region for three eye traits. Comparisons shown on the results of higher $-\log_{10}(P\text{-value})$ (>5.0) for axial length (AL), central corneal thickness (CCT), and spherical equivalent refraction (SER). **(A)** The results of AL on chromosome 10. The plot shows the $-\log_{10}(P\text{-value})$ of 100-SNP-window number 161 for RHM with win100 (Window161_100SNPs) and win20 (Window161_20 SNPs), 100-SNP-window number 162 for RHM with win100 (Window162_100SNPs) and win20 (Window162_20SNPs), and 100-SNP-window numbers 161 and 162 for RHM with win10 (10SNPs). The red horizontal line is drawn at the 5% genome-wide significance for RHM with win10. **(B)** The result of CCT on chromosome 16. The plot shows the $-\log_{10}(P\text{-value})$ of 100-SNP-window number 156 for RHM with win100 (Window156_100SNPs) and win20 (Window156_20 SNPs), 100-SNP-window number 157 for RHM with win100 (Window157_100SNPs) and win20 (Window157_20SNPs), and 100-SNP-window numbers 156 and 157 for RHM with win10 (10SNPs). The red horizontal line is drawn at the 5% genome-wide significance for RHM with win10. **(C)** The result of SER on chromosome 2. The plot shows the $-\log_{10}(P\text{-value})$ of 100-SNP-window number 197 for RHM with win100 (Window197_100SNPs) and win20 (Window197_20 SNPs), 100-SNP-window number 198 for RHM with win100 (Window198_100SNPs) and win20 (Window198_20SNPs), and 100-SNP-window numbers 197 and 198 for RHM with win10 (10SNPs). The red horizontal line is drawn at the 5% genome-wide significance for RHM with win20. **(D)** The result of SER on chromosome 10. The plot shows the $-\log_{10}(P\text{-value})$ of 100-SNP-window number 161 for RHM with win100 (Window161_100SNPs) and win20 (Window161_20 SNPs), 100-SNP-window number 162 for RHM with win100 (Window162_100SNPs) and win20 (Window162_20SNPs), and 100-SNP-window numbers 161 and 162 for RHM with win10 (10SNPs). The red horizontal line is drawn at the 5% genome-wide significance for RHM with win10.

(information score <0.5) no method was able to detect the simulated QTL. With QTL simulated based on well-imputed SNPs (information score >0.7) all methods we used had some power and often they were quite similar. Nonetheless, overall RHM was similar or greater in power to detect QTL than single SNP GWAS and had greater power than other gene-based methods. In particular, RHM had greater power to detect low MAF QTL and/or multiple independent QTL effects acting in a region than any of the methods of single-SNP GWAS and gene-based association approaches we tested, especially when RHM was performed using smaller analysis window sizes. RHM also captured a larger proportion of the QTL variance caused by multiple independent QTL and/or low MAF QTL. Importantly, for QTL with low MAF, RHM was capable of capturing more of the QTL variance than single-SNP GWAS for all magnitudes of QTL heritability.

GEMMA had slightly higher power than RHM when we simulated a single QTL with 0.025 QTL heritability. However, even in this case RHM found additional loci not detected by GEMMA. RHM also had greater power to detect QTL than GEMMA when several QTL in a region contribute trait variation and all have low MAF.

The effect of QTL MAF was evaluated by simulating QTL in the low MAF (MAF < 0.1) and high MAF (MAF ≥ 0.10) groups. As the number of QTL per window increased, the power to detect QTL also increased when the QTL had low MAF and 0.05 QTL heritability (**Figure 2** and Supplementary **Figure S3**). When a single low MAF QTL is randomly selected, it is likely to be very rare (as very rare SNPs are more common within the low MAF group than moderately rare ones, see **Figure 1**) and hence not well-captured by genotyped SNPs. When multiple (5 or 10) low MAF

Table 3 | Summary of the significant regions for Regional heritability mapping (RHM) and other methods in Axial Length, Central Corneal Thickness, and Spherical Equivalent Refraction.

Trait[1]	Chromosome	Window size								RHM				Other methods			
			SNP number		SNP name and position		MAF					Heritability		*P*-value[4]			
			Start	End	Start	End	Min	Mean	Max	LRT[2]	*P*-value[3]	Regional	Genome	GEMMA[5]	VEGAS	CCA	SKAT
AL	10	10	8051	8060	rs1877591 79,784,171	rs11002552 79,870,936	0.08	0.14	0.28	21.2	4.2×10^{-6}	0.049	0.438	4.7×10^{-3}	6.9×10^{-3}	1.4×10^{-2}	6.2×10^{-2}
CCT	9	10	13,711	13,720	rs1536482 134,666,473	rs4304399 134,746,881	0.05	0.22	0.49	8.1	4.4×10^{-3}	0.013	0.818	1.0×10^{-7}**	1.1×10^{-3}	1.8×10^{-4}	9.8×10^{-1}
	16	20	7811	7830	rs7403882 86,639,452	rs8044502 87,003,009	0.08	0.27	0.52	23.6	1.2×10^{-6}**	0.026	0.799	3.3×10^{-11}**	1.0×10^{-6}**	4.2×10^{-5}	1.8×10^{-1}
	16	10	7821	7830	rs12597413 86,782,559	rs8044502 87,003,009	0.16	0.34	0.52	27.0	2.1×10^{-7}**	0.026	0.795	3.3×10^{-11}**	0**	3.0×10^{-6}	7.5×10^{-1}
SER	2	20	9871	9890	rs4851411 101,260,596	rs3923053 101,444,879	0.10	0.25	0.53	22.6	2.0×10^{-6}**	0.150	0.349	6.9×10^{-2}	6.0×10^{-1}	9.2×10^{-1}	7.0×10^{-1}
	2	10	9881	9890	rs7568067 101,385,329	rs3923053 101,444,879	0.10	0.21	0.48	21.1	4.4×10^{-6}	0.112	0.369	6.9×10^{-2}	5.1×10^{-1}	5.5×10^{-1}	8.1×10^{-1}
	10	10	8051	8060	rs1877591 79,784,171	rs11002552 79,870,936	0.08	0.14	0.28	28.9	7.5×10^{-8}**	0.092	0.362	2.9×10^{-2}	7.2×10^{-3}	2.0×10^{-1}	1.4×10^{-1}

[1] *AL, Axial Length; CCT, Central Corneal Thickness; SER, Spherical Equivalent Refraction.*

[2] *Likelihood ratio test statistics.*

[3] *For guidance, the 5% genome-wide significance for RHM with window size 10 (P-value = 1.8×10^{-6}) and window size 20 (P-value = 3.7×10^{-6}).*

[4] *For guidance, the 5% genome-wide significance for GEMMA (P-value = 1.8×10^{-7}), 5% genome-wide significance for gene-based association approaches with window size 10 (P-value = 9.2×10^{-7}) and window size 20 (P-value = 1.8×10^{-6}).*

[5] *The minimum P-value in the window was selected.*

[6] ***5% genome-wide significance level.*

QTL are selected, one or more of the less rare ones within the low MAF group may well be chosen. These will contribute much of the variance and are likely to be better captured by genotyped SNPs leading to increased power when there were more QTL per window.

The power to detect QTL by RHM was greater than that of the three gene-based association approaches studied. We found that these gene-based association methods are strongly affected by the QTL MAF. The power of VEGAS and SKAT was greatly decreased for low MAF or high MAF QTL, respectively. SKAT was developed as a rare-variant association test (Wu et al., 2011), and uses a weighting scheme that upweights the contribution of rare variants and downweights the contribution of common variants in its default setting. Therefore, this default setting would be less powerful when variants have high MAF. VEGAS corrects the test statistics by LD between genotyped SNPs (Liu et al., 2010), and this correction might lose the power in the condition with low MAF because of incomplete LD between genotyped SNPs and QTL. In addition, these methods are also affected by the genome heritability. In this simulation, GRAMMAR-adjusted phenotypes are used to correct the effect of population stratification in SKAT and CCA, because these methods are not designed within a mixed model framework and cannot readily account for family relatedness among samples. The power for high genome heritability is lower than that for low genome heritability in these methods. But RHM was also more powerful than all gene-based association approaches at low genome heritability. For comparisons of the significant regions identified by RHM, GEMMA, and gene-based association approaches, more than 90% of the QTL can be captured by only RHM and GEMMA. Therefore, we suggest that RHM should be used as the complementary method which detects a different set of QTL when the power to detect QTL is not complete.

RHM has the potential to capture some of the "missing heritability." Yang et al. (2010) estimated that common SNP variation explained more than half of the expected heritability of human height, and suggested that missing heritability is due to imperfect LD between genotyped SNPs and causal variants. Yang et al. (2010) also simulated a quantitative trait by randomly sampling causal variants from the SNPs with MAF ≤ 0.10, and showed that estimated genome heritability was underestimated in comparison with the true genome heritability. In this study, the genome heritability estimated by model (2) for low MAF QTL was also underestimated in comparison with that for high MAF QTL. However, RHM captured more QTL variance with low MAF QTL than single-SNP GWAS and hence may capture heritability missed by single SNP GWAS.

Many explanations for the missing heritability have been suggested: a large number of common variants with small effect, a moderate number of rare variants with large effect, and some of combination of genotypic, environmental and epigenetic interactions (Manolio et al., 2009; Gibson, 2012). In this study, we show that RHM has the potential to explain some of the missing heritability through identification of trait-associated low MAF QTL by using common SNPs. However, some genetic variance could not be captured as some of the QTL variance is not in LD with individual common SNPs. An alternative method to capture QTL variance using common SNPs would be haplotype-based association, and some of the unknown low MAF QTL might be recovered by re-constructing haplotypes using common SNPs. However, some rare variants will be unique to particular populations and it will be difficult to detect QTL which are in linkage equilibrium with common SNPs. In this case we suggest that using exome sequencing or exome genotyping arrays combined with RHM on these types of data has the potential to capture even more of the missing variance.

In the study of real population data, some significant regions were detected by single-SNP GWAS, RHM or gene-based association approaches, corresponding to known loci but, additionally, two loci were newly identified, only by RHM, for SER. For the first one on chromosome 2, the *P*-value of win20 (SNP number 9871–9890) was lower than that of win10 (SNP number 9881–9890), and the win20 had high regional heritability 0.150 and contains four loci genes not previously implicated in refractive error control. There, multiple independent QTL of low MAF might be located on this narrow segment region. For the putative second novel SER locus, on chromosome 10, the regional signal was also suggestive for the phenotypically correlated trait AL making it unlikely to be a false positive finding. The closest genome-wide significant hit reported in the large GWAS meta-analyses of similar traits (SER or myopia) is a megabase away [Myopia GWAS SNP rs6480859 reported by Kiefer et al. (2013)] and although it is unlikely that the two findings reflect the same causal signal, they may highlight the same gene. Further analyses using other populations will be needed to validate these findings but this may be difficult if the variants are rare and their contribution to the trait variance large enough to be detectable in specific populations only. Functional analysis of the regions highlighted may also help confirming involvement of these regions. In this study, the significant *RXRA-COL5A1* CCT lead SNP on chromosome 9 detected by single-SNP GWAS was not detected by RHM. These mirrored the same trend as our simulation study, and also suggest that RHM should be an important complementary method to single-SNP GWAS, where multiple variants of low effect size and a range of MAFs may be segregating.

Nagamine et al. (2012) introduced RHM approach, and we present the effectiveness and implementation of RHM by assuming QTL in a narrow segment region, evaluating the impact of window size, and comparing with other single-SNP GWAS and gene-based association approaches under many different conditions. In addition, we detected some additional loci which were not detected by single-SNP GWAS and gene-based association approaches in real population data. We suggest in this study that RHM using common SNPs has the potential to explain some of missing heritability by capturing QTL variance with low MAF and localizing multiple independent QTL in a segment region. In conclusion, the results reported in this study support that RHM is more powerful to detect QTL and capture QTL variance than other single-SNP GWAS and gene-based association approaches under most conditions in populations structured similarly to those we studied, which include both related and unrelated individuals.

AUTHOR NOTE

For software capable of implementing Regional Heritability Mapping (RHM) analyses in populations of related and/or unrelated individuals, see REACTA: Regional Heritability Advanced Complex Trait Analysis at http://www.epcc.ed.ac.uk/projects-portfolio/reacta.

ACKNOWLEDGMENTS

This work was supported by funding from the Medical Research Council (UK) and the Biotechnology and Biological Sciences Research Council (UK). Yoshinobu Uemoto was supported by the Japanese National Livestock Breeding Centre. CROATIA-Korcula: The CROATIA-Korcula study was funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947) and Republic of Croatia Ministry of Science, Education and Sports research grants to Igor Rudan (108-1080315-0302). We would like to acknowledge the invaluable contributions of the recruitment team in Korcula, the administrative teams in Croatia and Edinburgh and the people of Korcula. The SNP genotyping for the CROATIA-Korcula cohort was performed in Helmholtz Zentrum München, Neuherberg, Germany. CROATIA-Split: The CROATIA-Split study is funded by grants from the Medical Research Council (UK), European Commission Framework 6 project EUROSPAN (Contract No. LSHG-CT-2006-018947) and Republic of Croatia Ministry of Science, Education and Sports research grants to Igor Rudan (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools and the Croatian Institute for Public Health. The SNP genotyping for the CROATIA-Split cohort was performed by AROS Applied Biotechnology, Aarhus, Denmark. CROATIA-Vis: The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants to Igor Rudan (108-1080315-0302). We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health. The SNP genotyping for the CROATIA-Vis cohort was performed in the core genotyping laboratory of the Wellcome Trust Clinical Research Facility at the Western General Hospital, Edinburgh, Scotland. Orkney Complex Disease Study (ORCADES): ORCADES was supported by the Chief Scientist Office of the Scottish Government, the Royal Society, the MRC Human Genetics Unit, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh. We would like to acknowledge the invaluable contributions of Lorraine Anderson and the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2013.00232/abstract

Figure S1 | Average r^2-value plotted against inter-marker distance and Correlation plot. To evaluate the quality of SNP imputation in this study, the difference of linkage disequilibrium (LD) between exome SNP and imputed SNP was investigated to evaluate whether the relationship between imputed SNPs is based on actual LD in this population or linkage equilibrium (LE). A total of 820 DNA samples from 898 Korcula samples were genotyped using the Illumina HumanExome-12v1 SNP array, that genotypes in excess of 250,000 exonic variants. These exome SNPs were then assessed by the exclusion criteria of minor allele frequency (MAF) <0.0005, call rate <0.98 and Hardy–Weinberg Equilibrium (HWE) $< 1.0 \times 10^{-6}$; SNPs included in Illumina CNV370 array were also excluded. A total of 7283 SNPs which were included in the low_info group and 25,313 SNPs in the high_info group were extracted from the exome array data. We estimated r^2, a measure of LD, for all segregating pairs of SNPs less than 10 Mbp apart in each of these groups using the PLINK software (Purcell et al., 2007). Average r^2-values for a given inter-marker distance, with markers distances grouped in 250 bp bins, were calculated in each autosome and plotted for each group. For the high_info group, r^2-values for low MAF (MAF <0.10) and high MAF (MAF $\geq$ 0.10) SNPs were also calculated and plotted. The imputed SNPs for individuals with corresponding exome SNP data were extracted from the imputed SNP data, and r^2-values were calculated as above. The correlation of r^2-values obtained from genotyped SNPs (i.e., exome array data) and imputed SNPs was also estimated. The average r^2-value was plotted against inter-SNP distance for exome SNPs and imputed SNPs in **(A)**, and the correlation of r^2-value between exome SNPs and imputed SNPs was also plotted in **(B)**. In the high_info group, 12,636 SNPs with low MAF and 12,677 SNPs with high MAF were also used to calculate r^2-values separately and then plotted. In the low_info group, there was no relationship between r^2-value and marker distance, and no correlation of r^2-value between exome SNPs and imputed SNPs. This result indicates that a high proportion of these exome and imputed SNPs in low_info group are estimated to be in linkage equilibrium (LE). On the other hand, in the high_info group, the r^2-value in shorter inter-marker distances was higher than that in greater inter-marker distances, and there was high correlation of r^2-values between exome SNPs and imputed SNPs. For the high_info group, the results within each MAF showed the same trend as the result for all SNPs and are not shown. The distribution and correlation of r^2-values for high MAF was tighter than that for low MAF. In addition, the slopes of correlation in the high_info group were about 1.0 in all results, and the magnitude of LD between exome SNP and imputed SNPs was almost the same. This result indicates that the LD structure of the real population is still preserved in these imputed SNPs. **(A)** Average r^2-value plotted against inter-marker distance for exome SNPs and imputed SNPs. The inter-marker distance grouped in 250 bp bins is on the x-axis, and average r^2-value is on the y-axis. Each figure shows the results of SNPs with low MAF in the low_info group, and all SNPs, SNPs with low MAF, and SNPs with high MAF the in high_info group. **(B)** Correlation plot between average r^2-values of exome SNPs and imputed SNPs. The average r^2-value of imputed SNPs is on the x-axis, and the average r^2-value of exome SNPs is on the y-axis. Each figure shows the results of SNPs with low MAF in the low_info group, and all SNPs, SNPs with low MAF, and SNPs with high MAF in the high_info group.

Figure S2 | Quantile-quantile plot of the *P*-values for Regional Heritability Mapping (RHM) with 100-SNP-windows. For each value of simulated genome heritability (0.20, 0.40, and 0.8), the phenotype was generated, and then genome heritability was estimated by using model (2). The estimated genome heritabilities were 0.20 ± 0.07, 0.40 ± 0.07, and 0.80 ± 0.06 for the simulated values of 0.20, 0.40, and 0.80, respectively. We then performed RHM analyses using 100-SNP window (win100) using these phenotypes, to obtain empirically a distribution of test statistics under the null hypothesis. For each genome heritability, a quantile-quantile plot of the *P*-values of the RHM analyses with win100 assuming that they follow either a 50:50 mixture distribution of a χ^2_1 and a pick at 0 or a χ^2_1 distribution are shown. Results of analysis of generated phenotypes with genome heritability = 0.2, 0.4, and 0.8 are presented. The red circles represent the $-\log_{10}$(*P*-value) value assumed as following the χ^2_1 distribution, and the blue triangles represent the $-\log_{10}$(*P*-value) value assumed as following the 50:50 mixture (one component mixture is a peak at 0 and the other is a χ^2_1) distribution. The black line represents where the dots are expected to fall under the null hypothesis of no association. The plots show that the 50:50 mixture is more appropriate, and also reflect the fact that our simulations generated appropriate phenotypes under the null hypothesis of no phenotype-window correlation.

Figure S3 | The power to achieve 5% genome-wide significance for 100-SNP windows: the case of genome heritability 0.2 and 0.8. The powers in the simulation study for genome heritability 0.2 and 0.8 were calculated by regional heritability mapping (RHM) with four different window sizes (100 SNPs as win100, 50 SNPs as win50, 20 SNPs as win20, and 10 SNPs as win10), and two single-SNP GWAS methods (GRAMMAR and GEMMA) in the different situations. The number of QTL is on the x-axis, and the power to detect QTL is on the y-axis. The results for genome heritability 0.2 are shown in **(A)** and 0.8 in **(B)**. The parameters considered in this simulation are QTL heritability (0.05 or 0.025) and MAF (low or high) in each genome heritability.

Figure S4 | The estimated regional heritability for 100-SNP windows: the case of genome heritability 0.2 and 0.8. The regional heritabilities in the simulation study for genome heritability 0.2 and 0.8 were estimated by regional heritability mapping (RHM) with four different window sizes (100 SNPs as win100, 50 SNPs as win50, 20 SNPs as win20, and 10 SNPs as win10), and two single-SNP GWAS methods (GRAMMAR and GEMMA) in the different situations. The number of QTL is on the x-axis, and the estimated regional heritability is on the y-axis. The results for genome heritability 0.2 are shown in **(A)** and 0.8 in **(B)**. The parameters considered in this simulation are QTL heritability (0.05 or 0.025) and MAF (low or high) in each genome heritability.

Figure S5 | The power to achieve 5% genome-wide significance in the gene region: the case of genome heritability 0.2 and 0.8. The powers in the simulation study for genome heritability 0.2 and 0.8 were calculated by regional heritability mapping with window size 10 (RHM), single-SNP GWAS (GEMMA), and three gene-based association approaches (VEGAS, CCA, and SKAT) in the different situations. The number of QTL is on the x-axis, and the power to detect QTL is on the y-axis. The results for genome heritability 0.2 are shown in **(A)** and 0.8 in **(B)**. The parameters considered in this simulation are QTL heritability (0.05 or 0.025) and MAF (low or high) in each genome heritability.

Figure S6 | Quantile–quantile plots for genome-wide association scan for three eye traits. Quantile-quantile plots of 272,315 SNPs in the genome-wide association scan were shown for Axial Length, Central Corneal Thickness, and Spherical Equivalent Refraction by single-SNP GEMMA analysis. The red circles represent the observed statistics, and the black line represents where the dots are expected to fall under the null hypothesis of no association. The plots show that this method successfully accounts for population stratification.

Figure S7 | Genome-wide plots of $-\log_{10}$ (*P*-values) for an association with three eye traits. Manhattan plots for Axial Length, Central Corneal Thickness, and Spherical Equivalent Refraction analyses by single-SNP GEMMA are shown. The genomic position is represented along the x-axis (chromosome number is indicated at the bottom of the plot).
The $-\log_{10}$(*P*-value) is on the y-axis. The red dotted horizontal line is drawn at the 5% genome-wide significance. The significant threshold of genome-wide significance at 5% by Bonferroni correction was *P*-value = 1.8×10^{-7}. For Central Corneal Thickness, there were two significant SNPs which were reported by Lu et al. (2013).

Figure S8 | Regional association plots for three eye traits near the significant region by regional heritability mapping (RHM) with window size 10 (win10). The results of regional association signals [higher $-\log_{10}$(*P*-value) > 5.0] are shown for Axial Length, Central Corneal Thickness, and Spherical Equivalent Refraction by RHM with win10. Plots were generated using LocusZoom (Pruim et al., 2010), and the color of each dot represents the SNP's linkage disequilibrium r^2 in the HapMap Phase II CEU with the labeled SNP (1st SNP within win10 with the lowest *P*-value) plotted as a purple diamond. The blue bars show the recombination rate based on HapMap phase II CEU population, and the bottom panels illustrate the locations of known genes.

Table S1 | Descriptive statistics for Axial Length, Central Corneal Thickness, and Spherical Equivalent Refraction.

Table S2 | SNP number and the average distance between SNPs by chromosome.

Table S3 | Estimated genome heritability for Regional heritability mapping (RHM) in the high_info group of the simulation study.

REFERENCES

Aulchenko, Y. S., de Koning, D. J., and Haley, C. S. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177, 577–585. doi: 10.1534/genetics.107.075614

Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., et al. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314, 1461–1463. doi: 10.1126/science.1135245

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* England: Longman.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Gilmour, A. R., Gogel, B. J., Cullis, B. R., and Thompson, R. (2006). *ASReml User Guide Release 2.0.* Orange, NSW: Agriculture.

Graham, R. R., Kyogoku, C., Sigurdsson, S., Vlasova, I. A., Davies, L. R., Baechler, E. C., et al. (2007). Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6758–6763. doi: 10.1073/pnas.0701266104

Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A., et al. (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* 39, 638–644. doi: 10.1038/ng2015

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5:e1000529. doi: 10.1371/journal.pgen.1000529

Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E. (2011). Gene-based tests of association. *PLoS Genet.* 7:e1002177. doi: 10.1371/journal.pgen.1002177

Kiefer, A. K., Tung, J. Y., Do, C. B., Hinds, D. A., Mountain, J. L., Francke, U., et al. (2013). Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet.* 9:e1003299. doi: 10.1371/journal.pgen.1003299

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838. doi: 10.1038/nature09410

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A versatile gene-basedtest for genome-wide association studies. *Am. J. Hum. Genet.* 87, 139–145. doi: 10.1016/j.ajhg.2010.06.009

Lu, Y., Vitart, V., Burdon, K. P., Khor, C. C., Bykhovskaya, Y., Mirshahi, A., et al. (2013). Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat. Genet.* 45, 155–163. doi: 10.1038/ng.2506

Maher, B. (2008). Personal genomes: the case of missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Morgenthaler, S., and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* 615, 139–145. doi: 10.1016/j.mrfmmm.2006.09.003

Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., et al. (2012). Localizing loci underlying complex trait variation using regional genomic relationship mapping. *PLoS ONE* 7:e46501. doi: 10.1371/journal.pone.0046501

Neale, B. M., and Sham, P. C. (2004). The future of association studies: Gene-based analysis and replication. *Am. J. Hum. Genet.* 75, 353–362. doi: 10.1086/423901

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., et al. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337. doi: 10.1093/bioinformatics/btq419

Tang, C. S., and Ferreira, M. A. R. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics* 28, 845–850. doi: 10.1093/bioinformatics/bts051

Verhoeven, V. J., Hysi, P. G., Wojciechowski, R., Fan, Q., Guggenheim, J. A., Höhn, R., et al. (2013). Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat. Genet.* 45, 314–318. doi: 10.1038/ng.2554

Visscher, P. M. (2006). A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Res. Hum. Genet.* 9, 490–495. doi: 10.1375/twin.9.4.490

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era – concepts and misconceptions. *Nat. Rev. Genet.* 9, 255–266. doi: 10.1038/nrg2322

Vitart, V., Bencic, G., Hayward, C., Herman, J. S., Huffman, J., Campbell, S., et al. (2010). New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. *Hum. Mol. Genet.* 19, 4304–4311. doi: 10.1093/hmg/ddq349

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* 9, 60–62. doi: 10.1214/aoms/1177732360

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 July 2013; accepted: 21 October 2013; published online: 19 November 2013.

Citation: Uemoto Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Wilson JF, Rudan I, Campbell H, Hastie ND, Wright AF and Haley CS (2013) The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Front. Genet. ***4****:232. doi: 10.3389/fgene.2013.00232*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 19 November 2013
doi: 10.3389/fgene.2013.00244

The null distribution of likelihood-ratio statistics in the conditional-logistic linkage model

Yeunjoo E. Song and Robert C. Elston *

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Edited by:
Brahim Aissani, University of Alabama at Birmingham, USA

Reviewed by:
Kui Zhang, University of Alabama at Birmingham, USA
Marie-Hélène Roy-Gagnon, University of Ottawa, Canada

***Correspondence:**
Robert C. Elston, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2013 Cornell Road, Cleveland, OH 44106, USA
e-mail: robert.elston@cwru.edu

Olson's conditional-logistic model retains the nice property of the LOD score formulation and has advantages over other methods that make it an appropriate choice for complex trait linkage mapping. However, the asymptotic distribution of the conditional-logistic likelihood-ratio (CL-LR) statistic with genetic constraints on the model parameters is unknown for some analysis models, even in the case of samples comprising only independent sib pairs. We derive approximations to the asymptotic null distributions of the CL-LR statistics and compare them with the empirical null distributions by simulation using independent affected sib pairs. Generally, the empirical null distributions of the CL-LR statistics match well the known or approximated asymptotic distributions for all analysis models considered except for the covariate model with a minimum-adjusted binary covariate. This work will provide useful guidelines for linkage analysis of real data sets for the genetic analysis of complex traits, thereby contributing to the identification of genes for disease traits.

Keywords: linkage analysis, affected sib pairs, identity-by-descent, conditional-logistic model, genetic constraints, null distribution, likelihood-ratio statistics

INTRODUCTION

In the study of human data by genetic linkage analysis, the traditional LOD score method, also called a "parametric" or "model-based" method because it requires information about an assumed genetic model, is efficient for single-gene Mendelian traits but is much less well suited for the analysis of traits with complex non-Mendelian modes of inheritance. In the absence of a well-defined disease inheritance model, alternative robust "non-parametric," "weakly-parametric" or "model-free" linkage methods, which do not require the specification of a disease model, have been used for deciphering the genetic basis of complex traits.

One such approach that has been extremely useful in the analysis of human genetic diseases is the affected sib pair (ASP) study design, as in tests based on the mean proportion of identity-by-descent (IBD) sharing (Blackwelder and Elston, 1985) or tests based on the likelihood-ratio (LR) defined by Risch (1990a,b) that uses the same one-parameter model to analyze ASPs or any other affected unilineal relative pairs by producing a LOD score. Holmans (1993) extended Risch's maximum LOD score method into a two-parameter model for ASPs, but with the genetic constraints required for single locus Mendelian inheritance; here we call this the Risch and Holmans (RH) model. Olson (1999) proposed a general conditional-logistic (CL) model that combines several extensions and modifications (Cordell et al., 1995; Rogus and Krolewski, 1996; Greenwood and Bull, 1997, 1999; Olson, 1997; Lunetta and Rogus, 1998) into a unified framework: the likelihood is conditioned on sampling affected relative pairs (ARPs) and the parameterization is done in terms of the logarithm of allele sharing specific relative risks, instead of allele sharing probabilities as in the RH model. The CL model not only retains the "nice" property of the LOD score formulation of the RH model, i.e., it is additive over independent sets of data, but it also has advantages over the RH model. It is valid for any type of ARPs with the same allele sharing specific parameters. In contrast, the RH model is parameterized in terms of relative-type specific IBD probabilities, so it can accommodate only one ARP type at a time. The other advantage of this CL model is that it can allow for incorporation of covariate effects by re-parameterizing the model in terms of the logarithms of genetic relative risk parameters. A modification of this original two-parameter CL model into a one-parameter model was proposed by Goddard et al. (2001). Linkage analysis using the CL model has been proven to be an effective tool for evaluating genetic linkage (Goddard et al., 2001; Arcos-Burgos et al., 2004; Reck et al., 2005; Doan et al., 2006; Rybicki et al., 2007; Stein et al., 2007; Zandi et al., 2007; Song et al., 2011).

One limitation of the general two-parameter CL model is the unknown asymptotic distribution of certain cases when single-locus genetic constraints are imposed on the model parameters, even in the case of analyzing only independent ASPs. Because of the genetic constraints (Holmans, 1993), the distribution of the CL-LR (i.e., 2ln(10) $*$ LOD score) statistics for linkage are mixtures of χ^2 distributions that are difficult to specify. The use of simulation methods to obtain p-values has been recommended to ensure accuracy of the inference in complex situations (Olson, 1999). Although gene-dropping techniques can be used for this purpose, the ideal method to infer the statistical significance of a test statistic is to compare it with its permutation distribution. When analyzing affected pairs alone, however, permuting the allele sharing of relative pairs does not lead to a useful permutation distribution. As an alternative, Sinha et al. (2006) developed regression prediction models that provide more accurate p-values under the CL model framework. However, their results are limited to the cases they evaluated, so it is not a general solution for the unknown distribution of the CL-LR statistic.

Here, we first derive approximations to the asymptotic distributions of the CL-LR statistics when using the constrained two-parameter analysis model for independent ASPs. The derivation is done under the null hypothesis of no linkage and assuming complete marker information, by following Self and Liang (1987), as done for the RH model (Holmans, 1993; Whittemore and Tu, 1998; Feng et al., 2006). Next, we study the empirical null distributions of the CL-LR statistics by simulation, again for independent ASPs, examining several analysis models with different constraints on the model parameters when using the LODPAL program in the S.A.G.E. package (2012). Then, we compare these distributions to the derived asymptotic distributions - either known or approximated in the previous step.

MATERIALS AND METHODS

CONDITIONAL-LOGISTIC MODEL

We first briefly describe the original two-parameter CL model from Olson (1999). The unconditional (prior) probability that a pair of type r relatives shares i alleles IBD is denoted as f_{ri}, and the estimated probability that the pair shares i alleles IBD conditional on the available marker data I_m is denoted as $\hat{f}_{ri}$. Then the likelihoods under the null hypothesis (H_0) of no linkage and under the alternative (H_1) can be written as

$$H_0 : L(\lambda_1 = 1, \lambda_2 = 1) = P(I_m|r)$$

and

$$H_1 : L(\lambda_1, \lambda_2) = P(I_m|r)\,\frac{\sum\limits_{i=0,1,2} \lambda_i \hat{f}_{ri}}{\sum\limits_{i=0,1,2} \lambda_i f_{ri}},$$

where λ_i is the relative risk to an individual who shares i alleles IBD (i =0, 1, 2) with an affected relative: equating with the notation used in the RH model,$\lambda_0 = \lambda_u (= 1)$ is the relative risk for unrelated individuals, $\lambda_1 = \lambda_o$ is the offspring relative risk, and $\lambda_2 = \lambda_m$ is the MZ-twin relative risk. The CL model is parameterized in terms of the logarithms of relative risk, so $\lambda_i = e^{\beta_i}$. Under the null hypothesis of no linkage, the parameters $(\beta_1, \beta_2) = (0, 0)$ correspond to Risch's allele sharing probability parameters $(z_1, z_2) = (½, ¼)$, where z_1 and z_2 are the respective probabilities an ASP shares 1 and 2 alleles IBD at a locus. The LR contribution for an ARP of type r is LR $= \frac{\sum_{i=012} \lambda_i \hat{f}_{ri}}{\sum_{i=012} \lambda_i f_{ri}}$, and for a sample of independent ARPs the LOD score is obtained by summing the base-10 logarithms of the pair-specific LRs. For the test of linkage, this LOD score is maximized over a possible range of the parameter space that depends on the constraints imposed, as discussed in the following section. For details of the derivation of the LR and the equivalence of the LR whether the parameterization is in terms of allele sharing probabilities or allele sharing relative risks, we direct the reader to Olson (1999).

When the parameters β_1 and β_2 are completely free without any constraints, the parameter space is the whole 2-dimensional plane with two coordinate axes defined by the two parameters. The values of the two parameters under the null hypothesis fall into interior points of this parameter space, and so the CL-LR statistic under the null hypothesis of no linkage is distributed as χ_2^2 asymptotically. We refer to this model as the *unconstrained two-parameter model.*

When the (pure single-locus etiology) genetic constraints (Holmans, 1993) are imposed, the parameter β_1 and β_2 are constrained to be $\beta_1 \geq 0$ and $\beta_2 \geq \log_e\left(2e^{\beta_1} - 1\right)$, or equivalently, $\lambda_1 \geq 1$ and $\lambda_2 \geq 2\lambda_1 - 1$, to reflect the possible allele sharing probabilities for ASPs. In this case, the values of the parameters under the null hypothesis are on the edge of the parameter space, so that the LR statistic is asymptotically distributed as the mixture $\left(\frac{1}{2} - c\right)\chi_0^2 + \frac{1}{2}\chi_1^2 + c\chi_2^2$ with the mixing proportion c representing the probability that the allele sharing estimates fall inside a triangle that is part of the two-dimensional plane. We refer to this model as the *constrained two-parameter model.*

MIXING PROPORTION c

The mixing proportion c is a function of the expected information matrix. For the RH model with allele sharing parameters, it has been derived to be $c \approx 0.098$ when there is complete marker information (Holmans, 1993; Whittemore and Tu, 1998; Feng et al., 2006), regardless of the choice of any two free parameters, i.e., (z_0, z_1), (z_0, z_2), or (z_1, z_2). However, for the CL model with the parameters in terms of the logarithms of relative risk, this value is unknown. We apply the method of Self and Liang (1987), as for the RH model, to derive the mixing proportion c for the LR statistic in the CL genetic constrained, two-parameter model.

As shown in **Figure 1**, let (β_1,β_2) represent a point in the 2-dimensional plane with two coordinate axes that are defined by the parameters β_1 and β_2, constrained to be $\beta_1 \geq 0, \beta_2 \geq \log_e\left(2e^{\beta_1} - 1\right)$ (gray area). We first define the three vertices of possible triangles in the (β_1,β_2) plane. Let $N = (0, 0)$ be the null point, A denote an additive inheritance point, and D a dominant inheritance point. The point A will be on the line $\beta_2 = \log_e\left(2e^{\beta_1} - 1\right)$. We define D =(0, β_2) as a point on the β_2 axis where the value of β_2 is the same as the point A, as in **Figure 1**. Let I be the Fisher information matrix of the likelihood function $L\left(data|\hat{\beta}_1, \hat{\beta}_2\right)$ evaluated at the null values. Assuming complete information, the variance-covariance matrix of the parameters is the inverse of I, i.e., $I^{-1} = \begin{pmatrix} 6 & 4 \\ 4 & 8 \end{pmatrix}$. Let $\mathrm{P}\Lambda\mathrm{P}^T$ be the spectral decomposition of I^{-1}, and Y_N, Y_A, and Y_D be the orthogonally transformed vertices of N, A and D such that $Y = \Lambda^{1/2}\mathrm{P}^T\left(\hat{\beta} - N\right)$. Let y_N, y_A, and y_D be the rotated vertices of Y_N, Y_A and Y_D such that Y_A lies on the β_1 axis and the ray defined by two points Y_N and Y_D becomes the hypotenuse in the upper right quadrant of the plane. Now, the three rotated vertices y_N, y_A, and y_D define the triangle area in the orthogonal space, and the angle θ formed by the two rays $\overrightarrow{y_N y_A}$ and $\overrightarrow{y_N y_D}$ represents the mixing proportion c. Letting the end point of the hypotenuse be (x, y), $\theta = arctan\left(\frac{y}{x}\right)$ and $c = \frac{\theta}{2\pi}$.

If a model with no dominance genetic variance is to fit, then $\beta_2 = \log_e\left(2e^{\beta_1} - 1\right)$, as shown by a solid red line in **Figure 1**. Owing to the fact that this line is not straight, the angle θ differs

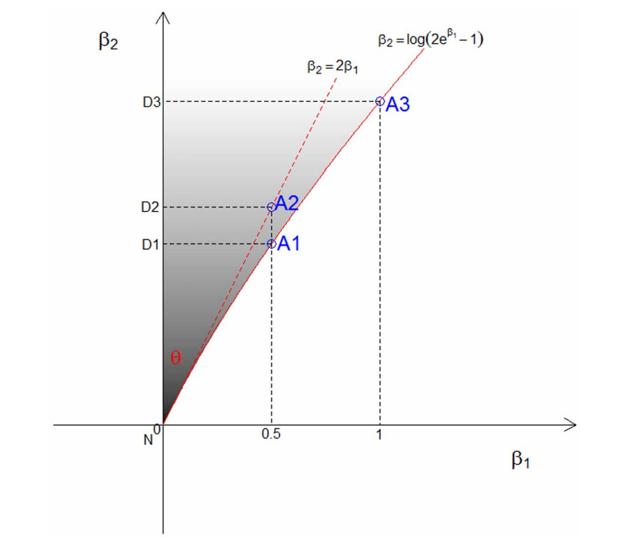


FIGURE 1 | The three points (A1, A2, and A3) used to approximate the relation between β_1 and β_2 and the upper bound of β_1 under genetic constraints in the CL model. The corresponding dominant points are denoted (D1, D2, and D3), and the shaded area is the possible triangle area in the CL model.

according to the choice of the point A on the line. The point A depends on both the assumption we make about the relation between β_1 and β_2, and the upper value of β_1 that is chosen. We consider 3 different points for A, denoted A1, A2, and A3, as shown in **Figure 1**. First, under the A1 assumption, we take the exact relation between β_1 and β_2, i.e., $\beta_2 = \log_e\left(2e^{\beta_1} - 1\right)$, and approximate the angle θ under the assumption that β_1 represents the allele sharing probability z_1, which has maximum value ½. Second, with the A2 assumption, we approximate a straight line about the null value using a Taylor series expansion, i.e., $\beta_2 = 2\beta_1$ (dotted red line in **Figure 1**). In this case, the upper bound of β_1 is irrelevant. This is equivalent to using the triangle obtained from the constraints on λ, i.e., $\lambda_2 = 2\lambda_1 - 1$. Third, with the A3 assumption, we take the exact relation between β_1 and β_2 and approximate the angle θ under the assumption that β_1 can go up to 1. This is equivalent to assuming the maximum offspring relative risk $\lambda_1 = \lambda_0 \approx 2.718$. We derive the resulting mixing proportions for these 3 cases and expand them for more values in the results section.

ONE-PARAMETER MODEL

Goddard et al. (2001) proposed to modify the two-parameter model into a one-parameter model on the basis of the minmax model developed by Whittemore and Tu (1998). In this one-parameter model, the constraint $\lambda_2 = (\pi + 1)\lambda_1 - \pi$ was imposed, where π is a parameter associated with the mode of inheritance and is fixed to be 2.634, i.e., $\beta_2 = \log_e\left(3.634e^{\beta_1} - 2.634\right)$ (Olson, 2002). This constraint assumes a genetic model approximately halfway between a recessive and a dominant mode of inheritance, which has been shown to be usually more powerful for most genetic models.

For this one-parameter model, the CL-LR statistic is known to be asymptotically distributed as a χ^2_1 when β_1 is free without any constraints, because its null value is an interior point of the parameter line. Even though Whittemore and Tu's minmax constraint is already imposed to make it a one-parameter model, we refer to this model as the *unconstrained one-parameter model* because β_1 is completely free without any genetic constraints. When the parameter space for β_1 is constrained by $\beta_1 \geq 0$ (equivalently $\lambda_1 \geq 1$) to reflect non-negative allele sharing probabilities, the CL-LR statistic is asymptotically distributed as a 50:50 mixture of a point mass at 0 and χ^2_1. We refer to this as the *constrained one-parameter model*.

COVARIATES

If there are K covariates in the model, assuming a log-linear (i.e., multiplicative) effect of the covariate on genetic relative risk, which is a common, natural, and flexible way to model relative risk in general epidemiology (Olson, 1999), the relative risk is $\lambda_i = \exp\left(\beta_i + \sum_{j=1}^{K} \delta_{ij}x_j\right)$, where the δ_{ij} are the two parameters associated with the covariate x_j, with $\beta_0 = \delta_{0j} = 0$. Therefore, each covariate added requires two additional parameters for the two-parameter model but only one additional parameter for the one-parameter model.

When there are no constraints imposed on the covariate parameters, with the addition of K covariates the CL-LR statistic is asymptotically distributed as $\chi^2_{2(k+1)}$ in the unconstrained two-parameter model. For the triangle-constrained two-parameter model, with the addition of K covariates the distribution of the CL-LR statistic is a mixture of a point mass at 0 and several χ^2s with up to $2(K + 1)$ df, asymptotically. However, no covariates are allowed in the two-parameter model in the LODPAL program in the S.A.G.E. package (2012), owing to the practical difficulty of maximizing the likelihood of models with two additional parameters for each covariate. Therefore, in this study we did not consider the two-parameter models with covariates.

For the one-parameter model, addition of covariates requires one additional parameter for each covariate. With the addition of K covariates, without any additional constraints imposed on covariate parameters the CL-LR statistic is asymptotically distributed as χ^2_{k+1} in the unconstrained one-parameter model. Addition of K covariates in the constrained one-parameter model, again without any additional constraints imposed on the covariate parameters, gives a CL-LR statistic with a distribution that is asymptotically a 50:50 mixture of a χ^2 with K df and a χ^2 with $K + 1$ df, (Goddard et al., 2001). In this study, we only included the constrained one-parameter model with covariate(s), and this is referred to as the *covariate model*.

Depending on additional constraints on the covariates, we define two covariate models. By including a "mean-centered" covariate $(x - \bar{x})$, no constraints on the δ_{1j} are required (Olson, 1999), so the CL-LR statistic is asymptotically distributed as a 50:50 mixture of two χ^2s depending on the number of such covariates, as stated previously. This is reasonable for many covariates, in particular continuous covariates such as age. We refer to this as the *unconstrained covariate model*.

However, for some covariates, such as indicator variables that represent different populations or a binary factor, the offset from the minimum value of the covariate, i.e., "minimum-adjusted," $[x_a = x - \min(x)]$ is included in the model, so that the smallest value of the covariate equals zero. For such covariates, the constraint $\min_{x_{aj}>0} \sum_j x_{aj}\delta_{1j} \geq -\beta_1$ is applied; it is not then feasible to derive the asymptotic distribution of the CL-LR statistic under the null hypothesis theoretically, since it depends on the distribution of the covariate values in the given data. We refer to this as the *constrained covariate model*.

SIMULATIONS

To examine the precision of the expected asymptotic distributions in the previous section, we used simulation to determine the empirical null distributions of the CL-LR statistics. We considered 6 different analysis models described in the previous section. We considered the covariate model with just one covariate. For the unconstrained covariate model, we included one with a mean-centered continuous covariate. For the constrained covariate, we included one with a minimum-adjusted binary covariate.

We first simulated 100,000 replicates of 500 nuclear families having two parents and two affected siblings, i.e., 500 independent ASPs. For each case, one fully informative unlinked marker was simulated by assigning a unique allele to each founder, and then the alleles were randomly segregated to all offspring. For covariate models, under the null hypothesis of no linkage and no covariate effect, the covariate was simulated such that it was correlated with affection status but not with genotype. A random continuous value from a normal distribution with mean 0 and variance 1 was first assigned to each individual, regardless of affection status. Then a continuous covariate was simulated by adding a pre-fixed covariate effect to this value. A binary covariate was generated by dichotomizing this continuous covariate such that its population prevalence was 0.2. Given the covariate values for each member of the pair, the pair-level covariate for a pair was created by summing the two individual-level covariates. The continuous pair-wise covariate values for the unconstrained covariate model are mean-centered, and the binary pair-wise values for the constrained covariate model are minimum-constrained when they are included in the analysis.

To check the performance of the expected asymptotic null distribution for each analysis model under different sample sizes, we also simulated 100,000 replicates of 30, 50, and 100 families, as above. Additionally, the precision of the approximated asymptotic null distributions of the CL-LR statistics for the constrained two-parameter model was compared with the empirical null distributions under different marker information levels. We simulated 100,000 replicates of 100 independent ASPs for markers with 2, 4, 8, and 20 equally frequent alleles. These numbers correspond to PIC values of 0.38, 0.70, 0.86, and 0.95, respectively. We checked two cases, when both parents are typed and when neither is typed.

The empirical p-value corresponding to the LOD score was determined by assigning $p = (r+1)/(100,000+1)$ to the rth of the ranked LOD scores from 100,000 replicates. The asymptotic p-value corresponding to the same LOD score was calculated using the known or approximated asymptotic distribution, as described above.

RESULTS

ASYMPTOTIC NULL DISTRIBUTIONS UNDER TRIANGLE CONSTRAINTS

The resulting triangles under assumption A1 are graphically illustrated in **Figure 2**, showing the steps to derive the mixing proportion for a given value of A. In this figure, the possible triangle space for ASPs on the original (β_1, β_2) plane is in black, formed by the three vertices (N, A, D) $= \{[0, 0], [½, \log_e(2e^{1/2} - 1)], [0, \log_e(2e^{1/2}$ - $1)]\}$. Then, we have

$$Y_N = \begin{pmatrix} 11.12 & 0 \\ 0 & 2.88 \end{pmatrix}^{1/2} \begin{pmatrix} 0.615 & -0.788 \\ 0.788 & 0.615 \end{pmatrix}^T \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$Y_A = \begin{pmatrix} 11.12 & 0 \\ 0 & 2.88 \end{pmatrix}^{1/2} \begin{pmatrix} 0.615 & -0.788 \\ 0.788 & 0.615 \end{pmatrix}^T \begin{pmatrix} 0.5 \\ \log_e\left(2e^{0.5} - 1\right) \end{pmatrix}$$
$$= \begin{pmatrix} 3.213 \\ 0.199 \end{pmatrix},$$

$$Y_D = \begin{pmatrix} 11.12 & 0 \\ 0 & 2.88 \end{pmatrix}^{1/2} \begin{pmatrix} 0.615 & -0.788 \\ 0.788 & 0.615 \end{pmatrix}^T \begin{pmatrix} 0 \\ \log_e\left(2e^{0.5} - 1\right) \end{pmatrix}$$
$$= \begin{pmatrix} 2.187 \\ 0.868 \end{pmatrix};$$

and then $y_N = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $y_A = \begin{pmatrix} 3.219 \\ 0 \end{pmatrix}$, and $y_D = \begin{pmatrix} 2.236 \\ 0.731 \end{pmatrix}$.

The corresponding orthogonally transformed triangle (Y_N, Y_A, Y_D) is in blue, and the green dashed triangle (y_N, y_A, y_D) is the same orthogonally transformed triangle after rotation such that Y_A lies on the β_1 axis and the ray defined by Y_N and Y_D becomes the hypotenuse in the upper right quadrant of the plane. Then the angle θ formed by the two rays $\overrightarrow{y_N y_A}$ and $\overrightarrow{y_N y_D}$ in the green triangle is *arctan* $\left(\frac{0.731}{2.236}\right) \approx 0.316$, and the corresponding mixing proportion c_1 is $\frac{\theta}{2\pi} \approx 0.050$. By following the same steps, we find the mixing proportions to be $c_2 \approx 0.044$ and $c_3 \approx 0.054$, respectively, under the A2 and A3 assumptions.

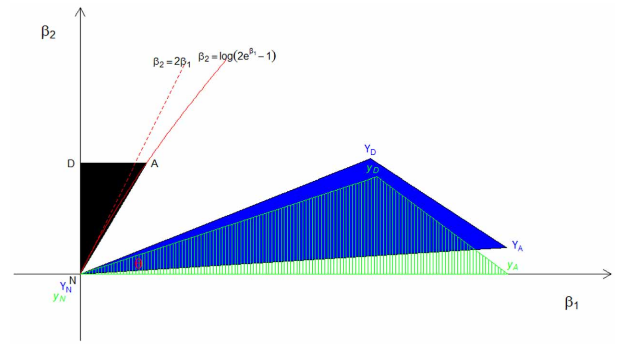


FIGURE 2 | The distribution of constrained CL-LR statistics under the A1 approximation. The black area (*N*, *A*, and *D*) is the original possible triangle space for ASPs, the blue area (Y_N, Y_A, and Y_D) is the orthogonally transformed triangle, and the green dashed triangle (y_N, y_A, and y_D) is the space after rotation. The angle θ formed by the two rays $y_N y_A$ and $y_N y_D$ represents the mixing probability *c*.

The value of c_2 obtained from the A2 assumption provides the minimum bound for c and, from the A1 and A3 assumptions, we can see that the mixing proportion value c becomes larger as we take a larger upper value for β_1. **Figure 3** shows how the value of c depends on the value of the parameter β_1. It can be seen that the maximum value converges to around 0.070, which is smaller than the value for the RH model. The critical LOD score values corresponding to the test sizes 0.05, 0.01, 0.001, 0.0001 [the classical "LOD score 3" criterion given by Morton (1955)], 0.000049 [significant evidence for linkage given by Lander and Kruglyak (1995)] and 0.00001 are given in **Table 1** for the different mixing proportion values. Given the same size of test, the critical LOD scores for the CL model are smaller than those for the RH model. Therefore, the null hypothesis is more likely to be rejected using the CL-LR test, and the CL-LR statistic is more powerful.

EMPIRICAL NULL DISTRIBUTIONS

Two-parameter model

In **Figure 4**, we show plots of –$\log_{10}$(empirical p-value) against –$\log_{10}$(asymptotic p-value) corresponding to the observed CL-LR statistics with a sample size 500 for two two-parameter models. For the unconstrained model, the empirical p-values well matched the asymptotic p-values from the expected chi-square distribution with 2 df. For the constrained model, the mixture distribution from the A1 assumption was also close to the empirical distribution. Since the mixing proportions from the three approximations are so close to each other, the empirical distributions matched the asymptotic distributions well for all three different mixing proportions (results not shown).

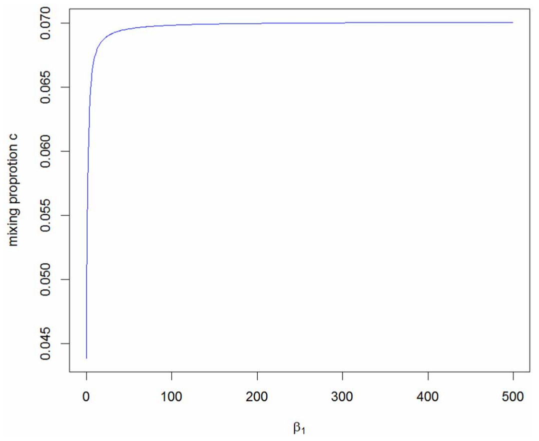


FIGURE 3 | The range of the mixing proportion values according to the different beta1 values for the distribution of the CL-LR statistics from the constrained two-parameter model.

Table 1 | Critical LOD scores obtained from the constrained two-parameter models for different mixing proportion values; *CL* − c_{min} and *CL* − c_{max} are the minimum and maximum *c* values for the CL model, *A1-c* is the value from the A1 approximation, and *RH-c* is the mixing proportion for the RH model.

Mixing proportion	Size of test					
	0.05	0.01	0.001	0.0001	0.000049	0.00001
CL-c_{min}	0.662	1.276	2.202	3.154	3.452	4.118
A1-c	0.672	1.289	2.219	3.172	3.470	4.138
CL-c_{max}	0.702	1.328	2.265	3.225	3.524	4.195
RH-c	0.742	1.377	2.324	3.290	3.591	4.265

For each sample size simulated, the specific LOD score values corresponding to the empirical p-values 0.05, 0.01, 0.001, and 0.0001 for these two models are given in **Figure 5**, compared with the theoretical values (shown as a red line for each p-value). These values are the critical values for the type I error rates equal to the given empirical p-values. Overall, for all sample sizes, the critical LOD scores from the empirical distributions were similar and very close to the values from the asymptotic distributions, well up to about –$\log_{10}$(p-value) = 3. When the type I error rate is 0.0001, the critical LOD scores varied depending on the sample size.

The empirical null distributions under different marker information levels for the constrained two-parameter model are shown in **Figure 6** (A for parents typed, B for parents not typed). For the two types of parental information, the specific LOD score values corresponding to the empirical p-values 0.05, 0.01, 0.001, and 0.0001 are again compared with the theoretical values from the A1 assumption (shown as a red line for each p-value). Again, it can be seen that the approximated asymptotic null distribution well matched the empirical distribution for the different levels of marker information, both in terms of the number of alleles and the amount of parental information.

One-parameter model

Here again, we found that the distribution of LOD scores follows the theoretical distribution well (results not shown). For both one-parameter models, the empirical p-values well matched the asymptotic p-values from the expected chi-square distributions. For the unconstrained case, the CL-LR statistic was distributed as a χ^2_1, as expected. The empirical distribution of the CL-LR statistics for the constrained model followed closely a 50:50 mixture of a point mass at 0 and a χ^2_1, which again agrees with the asymptotic distribution. For all sample sizes, the critical LOD scores from the empirical distributions were again similar and very close to the values from the asymptotic distributions well, up to about -$\log_{10}$ (p-value) = 3, and they varied depending on the sample size when the type I error rate is 0.0001, as for the two-parameter model.

Covariate model

In **Figure 7**, we show the distributions of empirical p-values under the null hypothesis of no linkage for the unconstrained covariate model. The empirical p-values for the covariate model with one unconstrained continuous covariate matched well the asymptotic p-values from a 50:50 mixture of a χ^2_1 and a χ^2_2 distribution when the sample size was 500, as expected. However, unlike other analysis models, the distribution of LOD scores did not follow the theoretical distribution for the smaller sample sizes. We found the empirical null distribution departed more from the asymptotic null distribution the smaller the sample size, as expected. For example, the critical LOD scores were over 10.0 for sample sizes 30, 50, and 100, compared to 3.77 from the asymptotic distribution for the test size 0.0001.

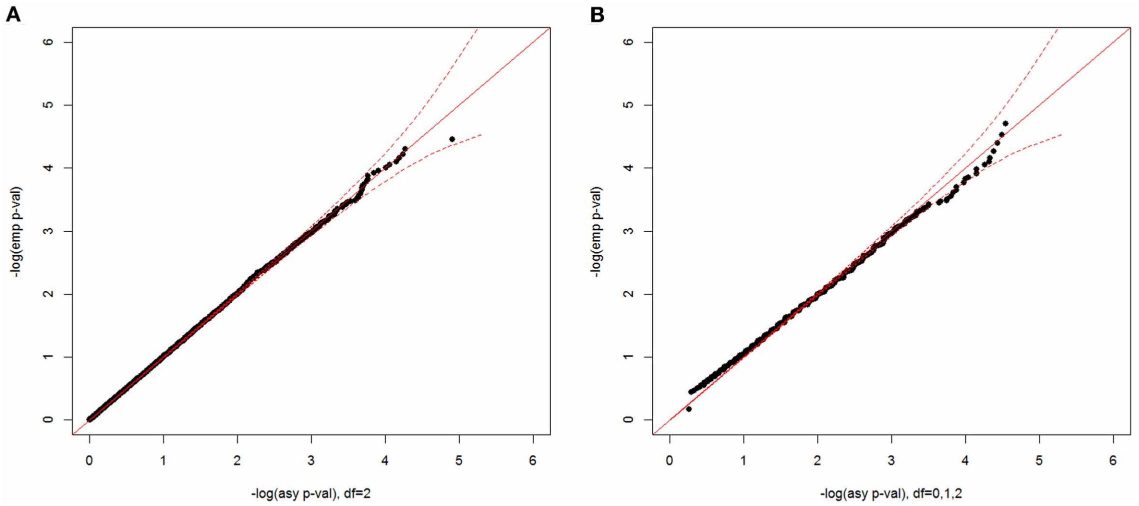


FIGURE 4 | Null distributions of the CL-LR statistics for the two-parameter models, using 500 independent ASPs and a fully informative marker. The empirical *p*-values for the observed LR statistics (y-axis) are plotted against the asymptotic *p*-values from known chi-square distribution (x-axis) for the unconstrained model **(A)** and for the constrained model **(B)** Note that the asymptotic distribution for the constrained model is under the A1 assumption, and a 95% confidence interval is shown by the dotted red line.

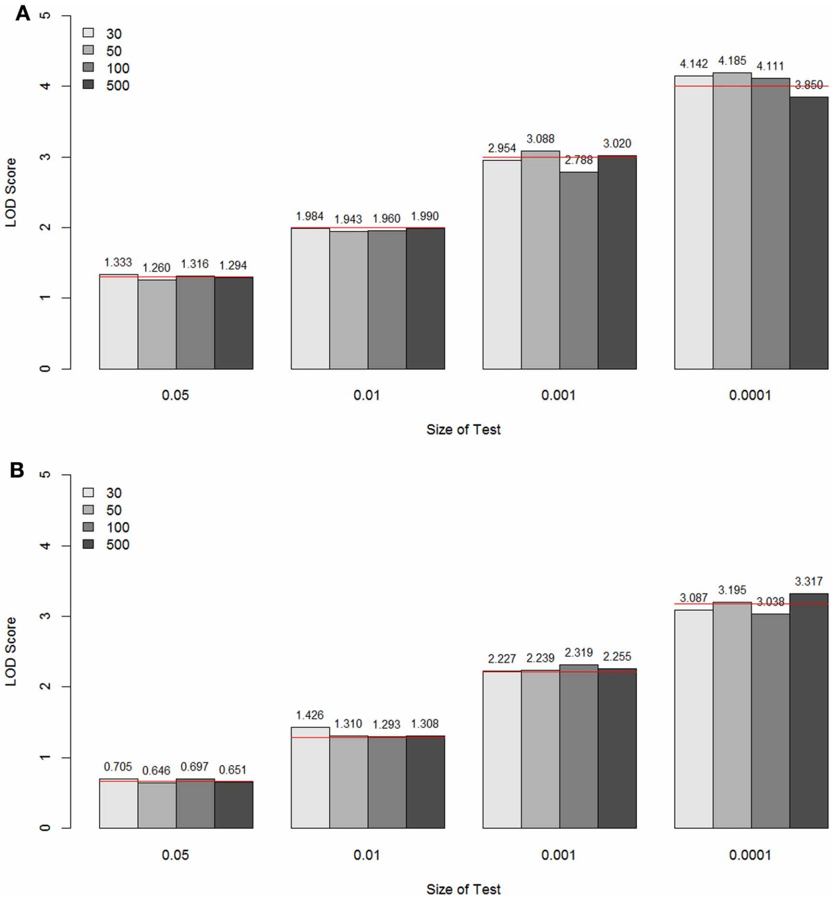


FIGURE 5 | The LOD score values corresponding to the empirical *p*-values 0.05, 0.01, 0.001, and 0.0001 for the unconstrained two-parameter model (A) and the constrained two-parameter model (B), by sample size and size of the test. These values are the critical values for the type I error rates equal to the given empirical p-values. The theoretical values are shown as a red line for each *p*-value.

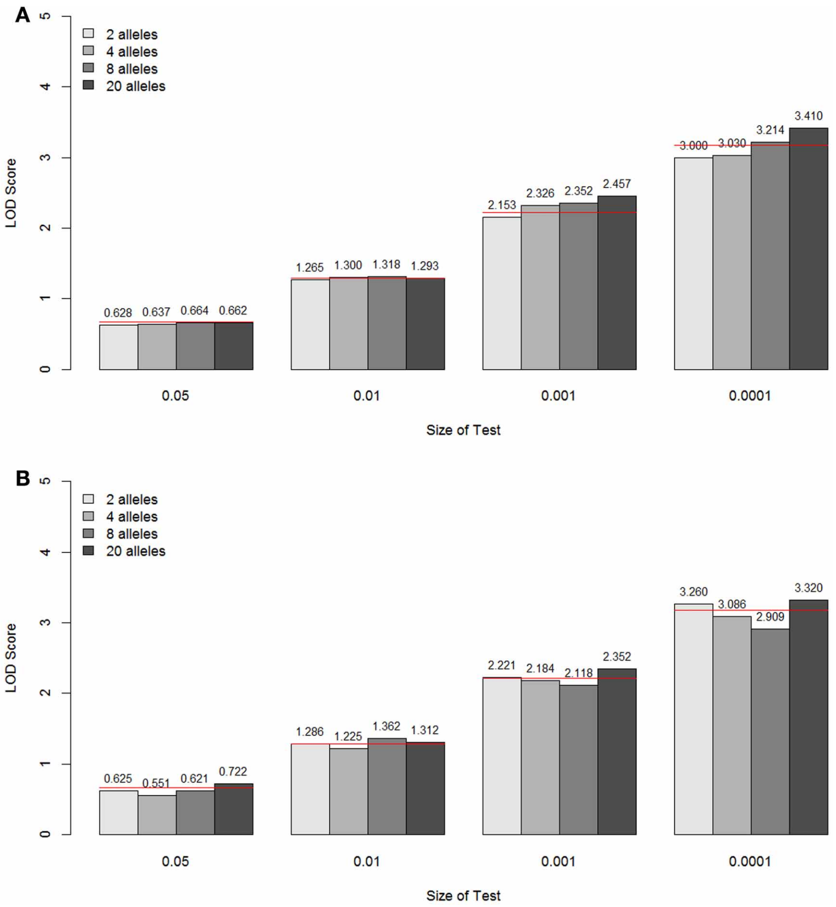


FIGURE 6 | LOD score values corresponding to the empirical *p*-values 0.05, 0.01, 0.001, and 0.0001 under different marker information levels for the constrained two-parameter model, when the parents are typed (A) and not typed (B). These values are the critical values for the type I error rates equal to the given empirical p-values. The theoretical values are shown as a red line for each *p*-value.

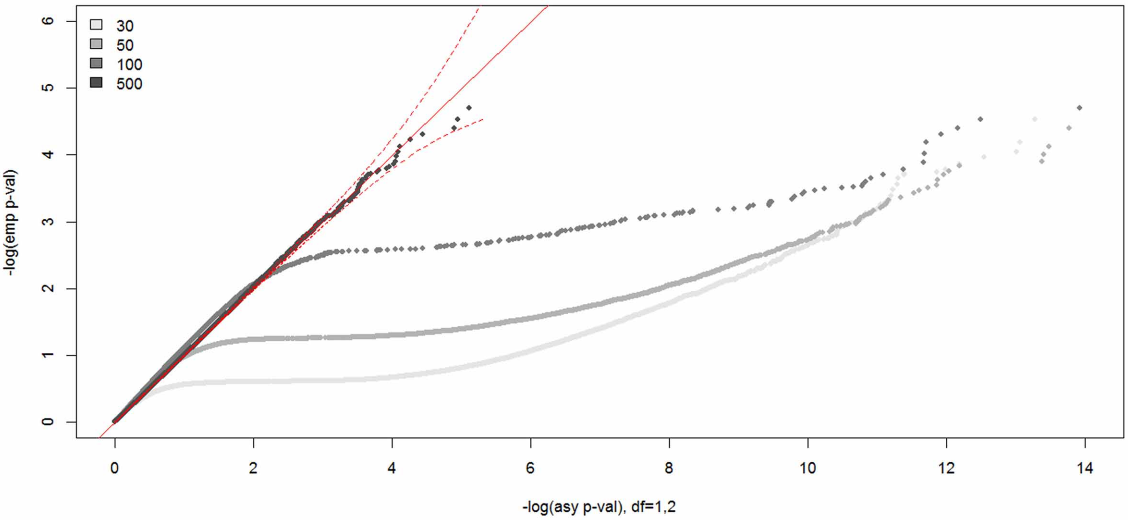


FIGURE 7 | Null distributions of the CL-LR statistics for the unconstrained covariate models, using 30, 50, 100, and 500 independent ASPs and a fully informative marker. The empirical *p*-values for the observed LR statistics (*y*-axis) are plotted against the asymptotic *p*-values from the known chi-square distribution (*x*-axis) for the unconstrained covariate model. The dotted red line is the 95% confidence interval.

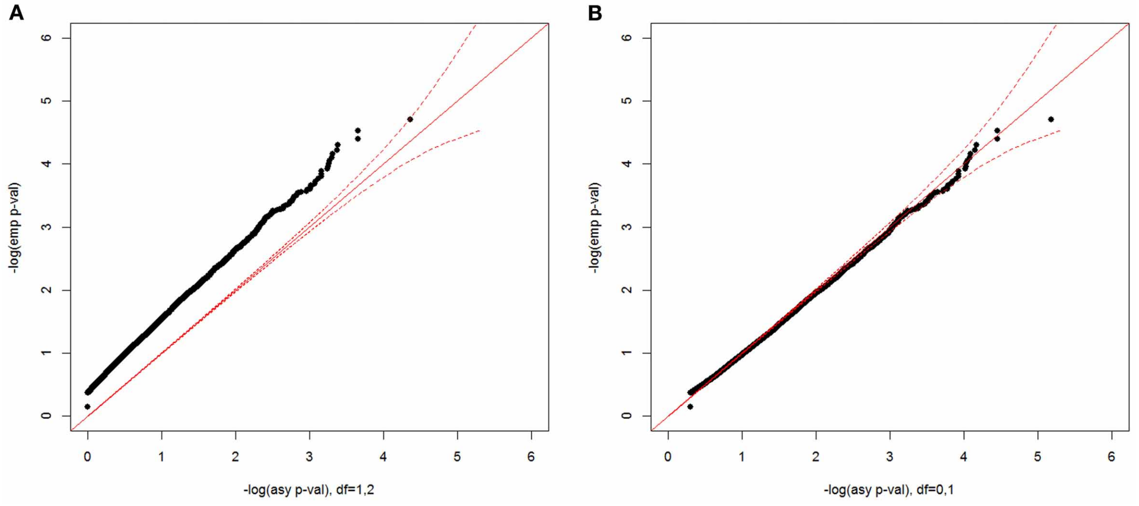


FIGURE 8 | Null distributions of the CL-LR statistics for the constrained covariate model, using 500 independent ASPs and a fully informative marker. The empirical *p*-values for the observed LR statistics (y-axis) are plotted against the asymptotic *p*-values from a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution **(A)**, and from a 50:50 mixture of a point mass at 0 and a χ_1^2 **(B)** The dotted red line is the 95% confidence interval.

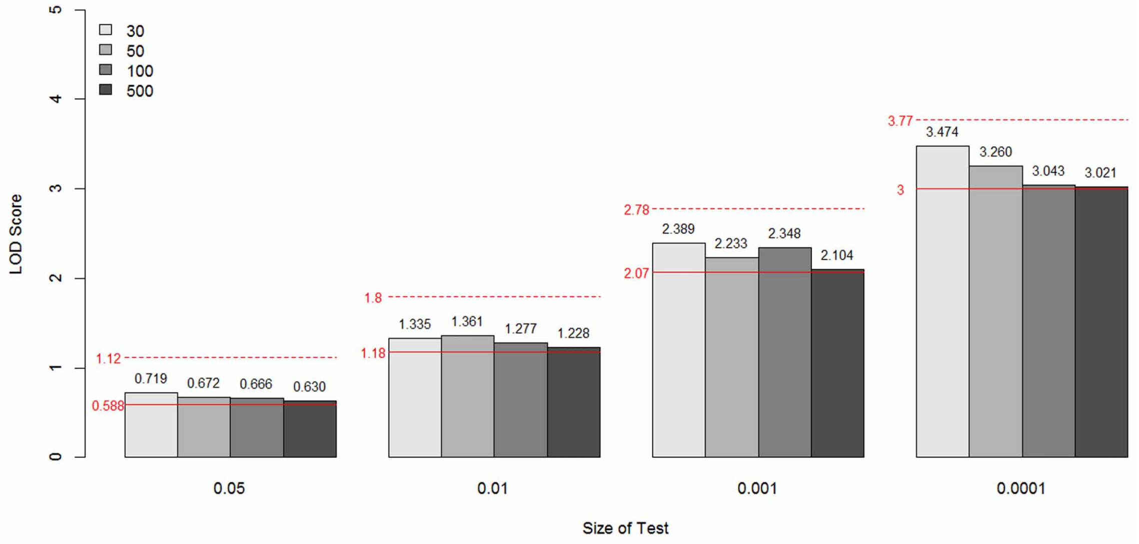


FIGURE 9 | LOD scores corresponding to the empirical *p*-values 0.05, 0.01, 0.001, and 0.0001 for the constrained covariate model by sample size and size of test. These values are the critical values for the type I error rates equal to the given empirical p-values. The theoretical values are shown as a red line for each *p*-value. The dotted lines are from a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution and the solid lines are from a 50:50 mixture of a point mass at 0 and a χ_1^2.

For the constrained covariate model with a minimum-adjusted binary covariate, we show the empirical null distribution compared with two asymptotic distributions in **Figure 8**, one with a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution (A) and the other with a 50:50 mixture of a point mass at 0 and χ_1^2 distribution (B). The asymptotic *p*-values from a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution were too conservative, while the asymptotic *p*-values from a point mass at 0 and χ_1^2 distribution well matched the empirical *p*-values. In the simulated data for this model, the possible pair-wise covariate values are 0, 1, or 2, since we included the sum of two individual binary covariate values. Since $\beta_1 \geq 0$ and $\min_{x_{aj} > 0} \sum_k x_{aj}\delta_k \geq -\beta_1$, $\delta_1 \geq 0$ when $\beta_1 = 0$. When $\beta_1 > 0$, the minimum value of δ_1 is $\frac{-\beta_1}{2}$. Therefore, the two-parameter space is constrained to be 1/3 of the whole plane, instead of 1/2 of the plane, which causes the asymptotic *p*-values from a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution to be too conservative. In practice, the distribution will depend on the distribution of the covariate values in the data.

In **Figure 9**, the specific LOD score values corresponding to the empirical *p*-values 0.05, 0.01, 0.001, and 0.0001 are given for each sample size simulated. These values are again the critical values for the type I error rates equal to the given empirical p-values, compared with theoretical values (shown as a red line for each *p*-value). The dotted lines are from a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution, and the solid lines are from a 50:50 mixture of a point mass at 0 and a χ_1^2.

DISCUSSION

In the RH model, the mixing probability *c* (which represents the probability that the allele sharing estimates fall inside the possible triangle) is the same for any two allele-sharing parameters. However, this is not so in the CL model owing to the non-straight line relation between the two parameters β_1 and β_2, the logarithms of relative risks. In this paper, we developed three approximations to the asymptotic distributions of the CL-LR statistics for the constrained two-parameter model, under the null hypothesis of no linkage, for independent ASPs. We derived the mixing probability *c* assuming complete information, as was done for the RH model with Risch's allele sharing parameters, following the method given by Self and Liang (1987). From these three approximations, we also investigated the relation between the parameter values for β_1 and *c*. We found the range of the *c* values to be (0.0439–0.070), which is lower than the value obtained for the RH model. This results in critical LOD score values lower by 5–11% (0.702–0.662 vs. 0.742) for a test size 0.05, and by 3–5% (2.265–2.202 vs. 2.324) for a test size 0.001, compared to the RH model. Therefore, the test using the CL-LR statistic will be more powerful, though perhaps not significantly so. In practice, the estimate of β_1 can be used to decide on an appropriate value for *c* to obtain a reasonably accurate test of linkage for a particular set of data.

By simulation, the performance of the approximate asymptotic distribution was checked for various sample sizes both when there is perfect information and under different marker information levels. This was done for two different parental information cases (typed and not typed) for a fixed sample size of 100 independent ASPs. Generally, for all sample sizes and the different levels of information content investigated, we found the empirical null distribution of the CL-LR statistic from the constrained two-parameter model matches well the approximated asymptotic distribution. This result shows the applicability of the approximated asymptotic distribution to real data analysis for any marker.

For the unconstrained two-parameter model, the unconstrained one-parameter model, and the constrained one-parameter model, we also found that the known asymptotic distributions matched the empirical distributions well. Therefore, for these models, the test of linkage using the CL-LR statistic can be performed using the known asymptotic null distribution to find the *p*-value. The unconstrained models may not be biologically plausible, but could be useful for the purpose of comparison, or when the data include ASPs with a different direction of genetic effect caused by other factors, as investigated by Dizier et al. (2000).

Unlike for the other models, a large sample size was needed for the asymptotic distribution to hold well for the unconstrained covariate model, i.e., the constrained one-parameter model with an unconstrained covariate. Sinha et al. (2006) also reported this vast discrepancy between the asymptotic *p*-values and the empirical *p*-values for this model. Their result was based on average sample sizes of 20, 40, 80, 120, and 320 affected pairs. To determine the sample size necessary for the asymptotic *p*-values to be applicable, we additionally simulated 200 and 300 ASPs. This showed that with 200 ASPs the empirical distribution matched well the asymptotic distribution (results not shown). Therefore, in practice, for this model we recommend the use of simulation methods or the Sinha et al. method when the sample size is less than 200, to ensure accurate *p*-values.

Though the results are not shown, from additional simulations with two and three covariates and 500 ASPs, except in the tail, the distributions of CL-LR statistics for the unconstrained covariate model with two covariates also closely matched a 50:50 mixture of a $\chi_2^2$and a χ_3^2, and that for three covariates a 50:50 mixture of a χ_3^2 and a χ_4^2, as expected from the asymptotic distributions. These results confirm that the empirical distribution of the CL-LR statistic for comparing nested unconstrained covariate models that differ by *J* covariates has a χ^2 distribution with *J* df, as expected from the asymptotic distribution. Therefore, in large samples it is valid to test the significance of the contribution of a covariate using the asymptotic distribution.

It was interesting to find in our simulated data that the empirical null distribution for the constrained covariate model, i.e., constrained one-parameter model with a constrained covariate, was closer to a 50:50 mixture of a point mass at 0 and χ_1^2 distribution than to a 50:50 mixture of a χ_1^2 and a χ_2^2 distribution. This is due to the functional dependency of δ_1 on the maximum covariate value in the data when $\beta_1 > 0$. This dependency effectively reduces the degrees of freedom and hence changes the distribution. To show how the range of the covariate values in the data changes the null values of the parameters, and therefore the distribution of the CL-LR statistics, we additionally simulated datasets with pair-wise covariate values (0 or 1), (0, 1, 2, or 3), (0, 1, 2, 3, or 4), and a random number in the range (0, 8). In **Figure 10**, we show a plot of the estimates of

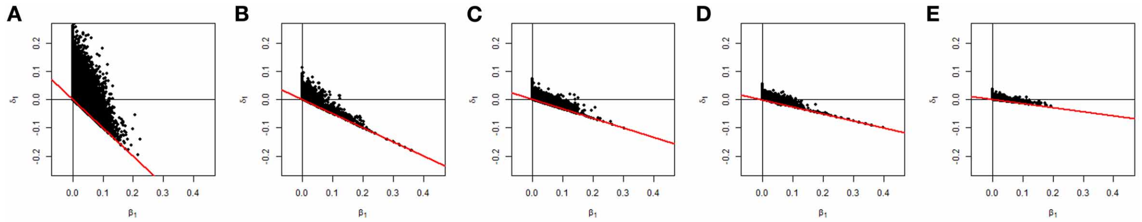


FIGURE 10 | Distributions of the estimates of β_1 and δ_1 under the constrained covariate model for different covariate distributions. The covariate values were (0 or 1) **(A)**, (0, 1 or 2) **(B)**, (0, 1, 2, or 3) **(C)**, (0, 1, 2, 3, or 4) **(D)** and a random number in the range (0, 8) **(E)**

the parameters β_1 and δ_1, including the result from the (0, 1, or 2) case in the previous simulation. We can see that the space for two parameters becomes smaller as the maximum value of the minimum-adjusted covariate increases. For the (0 or 1) case, it seems the CL-LR statistics will be closely distributed as the mixture $c_0\chi_0^2 + c_1\chi_1^2 + c_2\chi_2^2$. In other cases, a 50:50 mixture of a point mass at 0 and χ_1^2 distribution closely matched the empirical distribution. Therefore, in practice, the distribution will depend on the distribution of the covariate values in the dataset, so careful examination of the distributions of the covariates in the dataset is needed before including them in any analysis.

We did not include any power analysis in this study because our purpose was to find an approximation to the theoretically unknown null distributions and to compare them with the empirical null distribution, to provide guidelines for testing linkage when using the CL-LR statistics in various analysis models. To our knowledge, there has not been any study of the null distribution of LOD scores for the CL model, neither theoretical nor empirical. The results from this study should provide useful guidelines for the linkage analysis of real datasets since our results are based on both a perfect scenario as well as on non-perfect cases. Our results for various sample sizes will also provide guidelines for cases with missing data, since these will in general correspond to a reduced sample size. We assumed no errors in the relationship between pairs. When the information content in the marker and/or pedigree structure in real data are reduced due to errors in the data, we would generally expect the power to be lower for given type I error; but the test of linkage based on our results will still be valid, as long as the analysis is done on independent pairs.

REFERENCES

Arcos-Burgos, M., Castellanos, F. X., Pineda, D., Lopera, F., Palacio, J. D., Palacio, L. G., et al. (2004). Attention-deficit/hyperactivity disorder in a population isolate: linkage to loci at 4q13.2, 5q33.3, 11q22, and 17p11. *Am. J. Hum. Genet.* 75, 998–1014. doi: 10.1086/426154

Blackwelder, W. C., and Elston, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet. Epidemiol.* 2, 85–97. doi: 10.1002/gepi.1370020109

Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y., and Farrall, M. (1995). Two-locus maximum LOD score analysis of a multifactorial trait: joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes. *Am. J. Hum. Genet.* 57, 920–934.

Dizier, M. H., Quesneville, H., Prum, B., Selinger-Leneman, H., and Clerget-Darpoux, F. (2000). The triangle test statistic (TTS): a test of genetic homogeneity using departure from the triangle constraints in IBD distribution among affected sib-pairs. *Ann. Hum. Genet.* 64, 433–442. doi: 10.1046/j.1469-1809.2000.6450433.x

Doan, B. Q., Sorant, A. J., Frangakis, C. E., Bailey-Wilson, J. E., and Shugart, Y. Y. (2006). Covariate-based linkage analysis: application of a propensity score as the single covariate consistently improves power to detect linkage. *Eur. J. Hum. Genet.* 14, 1018–1026. doi: 10.1038/sj.ejhg.5201650

Feng, Z. Z., Chen, J., and Thompson, M. E. (2006). Asymptotic properties of the likelihood ratio statistics with the possible triangle constraint in affected-sib-pair analysis. *Can. J. Stat.* 35, 351–364. doi: 10.1002/cjs.5550350302

Goddard, K. A., Witte, J. S., Suarez, B. K., Catalona, W. J., and Olson, J. M. (2001). Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am. J. Hum. Genet.* 68, 1197–1206. doi: 10.1086/320103

Greenwood, C. M. T., and Bull, S. B. (1997). Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. *Genet. Epidemiol.* 14, 635–640. doi: 10.1002/(SICI)1098-2272(1997)14:6<635::AID-GEPI14>3.0.CO;2-R

Greenwood, C. M. T., and Bull, S. B. (1999). Analysis of affected sib pairs, with covariates—with and without constraints. *Am. J. Hum. Genet.* 64, 871–885. doi: 10.1086/302288

Holmans, P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* 52, 362–374.

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247. doi: 10.1038/ng1195-241

Lunetta, K. L., and Rogus, J. J. (1998). Strategy for mapping minor histocompatibility genes involved in graft-versus-host disease: a novel application of discordant sib pair methodology. *Genet. Epidemiol.* 15, 595–607. doi: 10.1002/(SICI)1098-2272(1998)15:6<595::AID-GEPI4>3.0.CO;2-4

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7, 277–318.

Olson, J. (1997). Likelihood-based models for genetic linkage analysis using affected sibpairs. *Hum. Hered.* 47, 110–120. doi: 10.1159/000154398

Olson, J. (1999). A general conditional-logistic model for affected-relative pair linkage studies. *Am. J. Hum. Genet.* 65, 1760–1769. doi: 10.1086/302662

Olson, J. (2002). Letter to the editor - rejoinder. *Genet. Epidemiol.* 23, 456–457.

Reck, B. H., Mukhopadhyay, N., Tsai, H. J., and Weeks, D. E. (2005). Analysis of alcohol dependence phenotype in the COGA families using covariates to detect linkage. *BMC Genet.* 6(Suppl. 1):S143. doi: 10.1186/1471-2156-6-S1-S143

Risch, N. (1990a). Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46, 222–228.

Risch, N. (1990b). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* 46, 229–241.

Rogus, J. J., and Krolewski, A. S. (1996). Using discordant sib pairs to map loci for qualitative traits with high sibling recurrence risk. *Am. J. Hum. Genet.* 59, 1376–1381.

Rybicki, B. A., Sinha, R., Iyengar, S. K., Gray-McGuire, C., Elston, R. C., Iannuzzi, M. C., et al. (2007). Genetic linkage analysis of sarcoidosis phenotypes: the sarcoidosis genetic analysis (SAGA) study. *Genes Immun.* 8, 379–386. doi: 10.1038/sj.gene.6364396

S.A.G.E. package (2012). *Statistical Analysis for Genetic Epidemiology.* Available online at: http://darwin.cwru.edu/sage/

Self, S. G., and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* 82, 605–610. doi: 10.1080/01621459.1987.10478472

Sinha, M., Song, Y., Elston, R. C., Olson, J. M., and Goddard, K. A. B. (2006). Prediction of empirical p values from asymptotic p values for conditional logistic affected relative pair linkage analysis. *Hum. Hered.* 61, 45–54. doi: 10.1159/000092552

Song, Y., Namkung, J., Shields, R., Baechle, D., Song, S., and Elston, R. C. (2011). A method to detect single-nucleotide polymorphisms accounting for a linkage signal using covariate-based affected relative pair linkage analysis. *BMC Proc.* 5:S84. doi: 10.1186/1753-6561-5-S9-S84

Stein, C. M., Zalwango, S., Chiumda, A. B., Millard, C., Leontiev, D. V., Horvath, A. L., et al. (2007). Linkage and association analysis of candidate genes for TB and TNFα cytokine expression: evidence for assocation with IFNGR1, IL-10, and TNF receptor 1 genes. *Hum Genet.* 121, 663–673. doi: 10.1007/s00439-007-0357-8

Whittemore, A. S., and Tu, I. P. (1998). Simple, robust linkage tests for affected sibs. *Am. J. Hum. Genet.* 62, 1228–1242. doi: 10.1086/301820

Zandi, P. P., Badner, J. A., Steele, J., Willour, V. L., Miao, K., MacKinnon, D. F., et al. (2007). Genome-wide linkage scan of 98 bipolar pedigrees and analysis of clinical covariates. *Mol. Psychiatry* 12, 630–639. doi: 10.1038/sj.mp.4002027

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 July 2013; accepted: 26 October 2013; published online: 19 November 2013.

Citation: Song YE and Elston RC (2013) The null distribution of likelihood-ratio statistics in the conditional-logistic linkage model. Front. Genet. ***4****:244. doi: 10.3389/fgene.2013.00244*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

frontiers in
GENETICS

METHODS ARTICLE
published: 21 November 2013
doi: 10.3389/fgene.2013.00243

Testing for direct genetic effects using a screening step in family-based association studies

Sharon M. Lutz[1,2]*, **Stijn Vansteelandt**[3] **and Christoph Lange**[2]

[1] Department of Biostatistics, University of Colorado, Aurora, CO, USA
[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Edited by:
Xiangqing Sun, Case Western Reserve University, USA

Reviewed by:
Nathan Morris, Case Western Reserve University, USA
Sungho Won, Chung-Ang University, South Korea

***Correspondence:**
Sharon M. Lutz, Department of Biostatistics, University of Colorado, 13001 E. 17th Place, B119 Building 500, Room W3128, Anschutz Medical Campus, Aurora, CO 80045, USA
e-mail: sharon.lutz@ucdenver.edu

In genome wide association studies (GWAS), family-based studies tend to have less power to detect genetic associations than population-based studies, such as case-control studies. This can be an issue when testing if genes in a family-based GWAS have a direct effect on the phenotype of interest over and above their possible indirect effect through a secondary phenotype. When multiple SNPs are tested for a direct effect in the family-based study, a screening step can be used to minimize the burden of multiple comparisons in the causal analysis. We propose a 2-stage screening step that can be incorporated into the family-based association test (FBAT) approach similar to the conditional mean model approach in the Van Steen-algorithm (Van Steen et al., 2005). Simulations demonstrate that the type 1 error is preserved and this method is advantageous when multiple markers are tested. This method is illustrated by an application to the Framingham Heart Study.

Keywords: family-based association analysis, causal inference, genetic pathway, mediation, pleiotropy

INTRODUCTION

Some of the recently published genome-wide association studies identified the same genetic locus as a disease susceptibility locus for different complex diseases (Amos et al., 2008; Thorgeirsson et al., 2008). One possible mechanism is that the marker locus is pleiotropic and has genetic effects on several, different phenotypes. Determining whether the marker acts directly on each of these phenotypes or only indirectly via one or more intermediate phenotypes is an important step in understanding the biological significance of the genetic associations. In order to understand and characterize the underlying genetic effect, methods have been proposed to disentangle these potential direct and indirect genetic effects (Vansteelandt et al., 2009; Vansteelandt, 2010; Berzuini et al., 2012; Vansteelandt and Lange, 2012; VanderWeele et al., 2012). All currently available methods focus on the direct and indirect genetic effects relative to one (group of) secondary phenotypes. Because the magnitude of the indirect effect depends on how strongly these secondary phenotypes affect the primary phenotype, these methods consider adjustment for confounding of the relationship between these phenotypes by measured extraneous factors. Some of these methods quantify both the direct and indirect genetic effects, but assume that none of these extraneous confounding factors is influenced by the considered marker (VanderWeele et al., 2012). Some of these methods allow for some of the extraneous confounding factors to be influenced by the considered marker, but they merely quantify direct genetic effects (Vansteelandt et al., 2009; Vansteelandt, 2010; Berzuini et al., 2012).

Regardless of the considered framework, all available methods only test one gene at a time and need to be corrected for multiple comparisons. This concern over multiple comparisons becomes an issue in family-based genome wide association studies (GWAS). When there is a region with a strong association with both the endo-phenotype and phenotype, identifying SNPs in the region that are still associated with the phenotype of interest after accounting for the association with the endo-phenotype requires testing for a direct causal effect for every SNP in the region. In order to increase power to detect this direct genetic effect, we propose a 2-stage testing strategy to minimize the burden of multiple comparisons in the causal analysis (Van Steen et al., 2005; Murphy et al., 2008; Won et al., 2009). The application of a screening step when testing for direct genetic effects is an important advantage in this scenario where the multiple-comparison problem is a major hurdle. The power of our approach is assessed by simulation studies. We show that the type-1 error is preserved and the method is shown to be advantageous when multiple SNPs are tested for a direct effect on the phenotype of interest.

METHODS

Suppose that in the family-based study, n trios (offspring and both parents) have been genotyped at a specific marker locus. Assuming there is no bias due to ascertainment conditions, the variable X_i denotes the coded genotype of the offspring and S_i denotes the parental genotypes for individual i. If genotypic data is unavailable for the parents but genotypic information is available on the subject's siblings, the variable S_i denotes the sufficient statistic by Rabinowitz and Laird (2000) For offspring i, Y_i denotes the target phenotype in the association study and K_i denotes the secondary phenotype in the study.

Suppose that an association has been observed between the secondary phenotype of interest, K_i, and the marker locus. Given this association, the goal is to test for an association between the target phenotype Y_i and the marker locus that cannot be explained by a possible indirect effect mediated by K_i. To achieve this goal, data is needed on all risk factors of the secondary phenotype K_i that are also associated with the primary phenotype (Cole and Hernan, 2002). Let L_i denote this collection of measured confounding variables. Because L may be high-dimensional, we do not assume that it is only related with Y by means of a causal effect, but allow for their association to be itself confounded by potentially unmeasured factors U. This is shown in the causal diagram of **Figure 1**, where the presence of U additionally captures the potential for confounding of the genetic association as a result of population admixture (Vansteelandt and Lange, 2012). Throughout, in contrast to other mediation analysis techniques (namely those based on so-called natural direct and indirect effects), we will allow for the possibility that some of these confounding variables are themselves affected by the studied marker, as illustrated via the edge from X to L in the causal diagram (VanderWeele et al., 2012).

Consider model

$$E[Y_i|X_i, K_i, L_i] = \gamma_0 + \gamma_1 K_i + \gamma_2 X_i + \gamma_3 L_i \tag{1}$$

where γ_j for $j = 0, 1, \ldots, 3$ denote the mean parameters and can be estimated by ordinary least squares. Note that γ_1 represents the true effect of K_i on Y_i and not a spurious association because, by assumption, the above model includes all relevant risk factors of K_i. In order to construct an adjustment principle that tests for a direct genetic effect of the marker locus X on the target phenotype Y, the effect of the secondary phenotype K has to be estimated. Vansteelandt et al. use an estimate for γ_1 based on model (1) to adjust the phenotype Y_i to $Y_i - \gamma_1 K_i$. A family-based association test (FBAT) on this adjusted phenotype is then a test for the direct genetic effect in the family-based setting (provided that the distribution of the test statistic acknowledges the uncertainty in the estimated effect γ_1) (Vansteelandt et al., 2009).

To reduce the number of multiple comparisons, we adapt the conditional mean model approach in the VanSteen-algorithm (Van Steen et al., 2005) to model (1). By replacing the observed marker score in model (1) by the expected marker score conditional upon the parental genotypes or sufficient statistic, the genetic effects of locus X_i can be assessed without having to adjust the α-level of any subsequently computed FBATs (Lange et al., 2003a,b; Van Steen et al., 2005). Similar to the idea of the conditional mean model approach, model (1) can be rewritten by substituting X_i with its expected value $E(X_i|S_i)$,

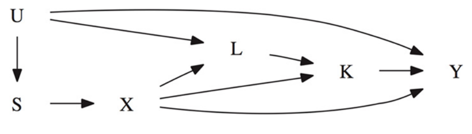


FIGURE 1 | Causal diagram illustrating the confounding of the target phenotype *Y* and the marker locus *X*. *S* denotes the parental genotype or Rabinowitz and Laird's sufficient statistic. *K* denotes the secondary phenotype of interest. *L* allows for confounding between *K* and *Y*. *U* represents a collection of unmeasured factors that allow for confounding due to population stratification or confounding between the two phenotypes *K* and *Y*. Note that causal diagrams assume that all variables that jointly affect any two variables are included. The absence of an arrow between any two variable denotes that there is no direct causal effect. For instance, *U* has no direct causal effect on *X*.

$$E[Y_i|K_i, L_i, S_i] = \beta_0 + \gamma_1 K_i + \beta_2 L_i + \beta_3 E(X_i|S_i)\,, \tag{2}$$

As shown in the proof given in the appendix, the parameter γ_1 is the same in both model (1) and model (2) when the null hypothesis holds that there is no direct effect and, moreover, there is no confounding due to population substructure. For testing the null hypothesis of no direct genetic effect, model (2) can thus be used to estimate the parameter γ_1 in a screening step without biasing the significance level since X_i is not included in this model, provided there is no confounding due to population substructure.

For the screening step, each subject contributes

$$T_i^* = \{E(X_i|S_i)\}\,\tilde{Y}_i^* \tag{3}$$

where $\tilde{Y}_i^* = Y_i - \bar{y} - \hat{\gamma}_1^*(K_i - \bar{k})$ and $\hat{\gamma}_1^*$ is the ordinary least squares estimate for γ_1 in model (2), which does not involve the genetic marker X. $\tilde{Y}_i^*$ is not adjusted for the covariates L_i since including factors such as L_i in the phenotypic adjustment would introduce bias if the common risk factor L_i is associated with the DSL X_i (Vansteelandt et al., 2009). The parameters $\bar{y}$ and $\bar{k}$ are the observed phenotypic means of Y and K in the sample, respectively. Then the test statistic for the screening step is

$$\frac{\left(\sum_{i=1}^{n} T_i^*\right)^2}{\sum_{i=1}^{n} \operatorname{var}(\tilde{T}_i^*)} \tag{4}$$

where

$$\tilde{T}_i^* = T_i^* - E\left[\{E(X_i|S_i)\}\,(K_i - \bar{k})\right] \frac{\left(K_i - \mu_k^{*(i)}\right)}{\sigma_k^{*2}}\,\epsilon_i^* \tag{5}$$

where $\operatorname{var}(\tilde{T}_i^*)$ is calculated based on the sample variance of $\tilde{T}^*$ and ϵ_i^* denotes the residual from model (2). $\mu_k^{*(i)} = E(K|L_i, E(X_i|S_i))$ is the predicted value for K under a linear regression model for K with the covariates L_i and $E(X_i|S_i)$, and σ_k^{*2} denotes the residual variance in that model. The variance correction given in Equation (5) is needed to account for estimating γ_1 in the proposed phenotype adjustment $\tilde{Y}_i^*$ (Vansteelandt et al., 2009).

For step 1, the test statistic given in Equation (4) can be used for the screening step to pick the SNPs with the highest power since X is not used in this test statistic. For step 2, this smaller subset of SNPs are used to test the null hypothesis of no direct effect using the test statistic based on Equation (1) proposed by Vansteelandt et al. (2009)

$$T_i = \{X_i - E(X_i|S_i)\}\tilde{Y}_i \tag{6}$$

where $\tilde{Y}_i = Y_i - \bar{y} - \hat{\gamma}_1(K_i - \bar{k})$ and $\hat{\gamma}_1$ is the ordinary least square estimate for γ_1 in model (1), which does involve the genetic marker X. Using this association test with the adjusted phenotype $\tilde{Y}_i$ as the target phenotype provides a robust and valid test for the null hypothesis that there is no direct effect between the target phenotype Y_i and the DSL; i.e., the association between the target phenotype Y_i and the DSL is solely a result of the association between the secondary phenotype K_i and the DSL. Adjusting for estimating γ_1 based on model (1), the test statistic is distributed chi-square with one degree of freedom under the null hypothesis of no direct effect of X on Y and has the following form

$$\frac{\left(\sum_{i=1}^{n} T_i\right)^2}{\sum_{i=1}^{n} \mathrm{var}(\tilde{T}_i)} \tag{7}$$

where

$$\tilde{T}_i = T_i - E[\{X_i - E(X_i|S_i)\}\, K_i]\, \frac{\left(K_i - \mu_k^{(i)}\right)}{\sigma_k^2}\epsilon_i \tag{8}$$

where $\mathrm{var}(\widetilde{T_i})$ is calculated based on the sample variance of $\tilde{T}$ and ϵ_i denotes the residual from model (1). $\mu_k^{(i)} = E(K|L_i, X_i, E(X_i|S_i))$ is the predicted value for K under a linear regression model for K with the covariates L_i, X_i, and $E(X_i|S_i)$, and σ_k^2 denotes the residual variance in that model. The variance correction given in Equation (8) is needed to account for estimating γ_1 in the proposed phenotype adjustment $\tilde{Y}_i$ (Vansteelandt et al., 2009). Note that Equation (3) is similar to Equation (6), but Equation (6) contains the genetic marker X_i. Similarly, Equation (5) is similar to Equation (8), but Equation (8) contains the genetic marker X_i.

Note that under the alternative hypothesis, the association between K and Y is different in models (1) and (2), even in the absence of population admixture. Model (1) represents the causal effect of K on Y under the alternative hypothesis, but model (2) does not represent the causal effect of K on Y because there is a remaining spurious association between X and Y along the path $K \leftarrow X \rightarrow Y$ in **Figure 1**. Under the null hypothesis, this path does not exist. As a result, the proposed approach is valid for testing in the absence of population stratification, but may have less power when either the $X \rightarrow K$ or the $X \rightarrow Y$ link is strong. This scenario is explored further in the simulation section of this paper.

Because the test statistic for the screening step given in Equation (4) is susceptible to population stratification, we examined this scenario in the simulation section as well. Principal component analysis (PCA) can be used in the screening step to correct for population stratification.

SIMULATIONS

Using simulation studies, we asses the type-1 error rate, the power, and robustness of this new approach which uses a trait that estimates γ_1 based on model (2) in the screening step and compare it to the approach proposed by Vansteelandt et al. (2009) which uses a trait that estimates γ_1 based on model (1). Similar to Vansteelandt et al. (2009), both methods are evaluated under various conditions. All simulations use a sample size of 1000 trios and are based on 5000 replications. The simulations are run for allele frequencies 5, 10, 15, 20, 25, 30, 35, 40, and 45%.

To reflect a realistic setting, the data is simulated to reflect covariances found in the Framingham Heart Study (Herbert et al., 2006). The phenotype of interest Y is simulated such that it resembles FEV1. The secondary phenotype K resembles weight and the set of common confounding variables resemble height and age. As seen in **Figure 2**, the first scenario assumes there is a direct genetic effect of the marker on the intermediate phenotype K and on the common covariate L. Each genetic effect has a locus specific heritability of 1%. The intermediate phenotype K explains 1% of the phenotypic variation in Y, creating an association between the SNP and Y. The second scenario is similar to the first scenario except that there is no genetic effect on the confounder L. The genetic association with the intermediate phenotype K is still present. The third scenario is similar to the first scenario except there is no association between K and Y. The fourth scenario is similar to the second scenario except that there is no genetic effect on the intermediate phenotype K.

As seen in **Table 1**, the type-1 error rate is similar whether model (1) or model (2) is used to estimate γ_1. For lower allele frequencies, under scenario 1 and 3, the type-1 error rate is 1–2% higher than expected. For higher allele frequencies under all four scenarios, the type-1 error rate is 0.5% lower than expected. In general, the type-1 error rate is close to 0.05 regardless of how γ_1 is estimated. As seen in **Table 2**, the power is similar whether model (1) or model (2) is used to estimate γ_1 assuming no population admixture. For lower allele frequencies, the method by

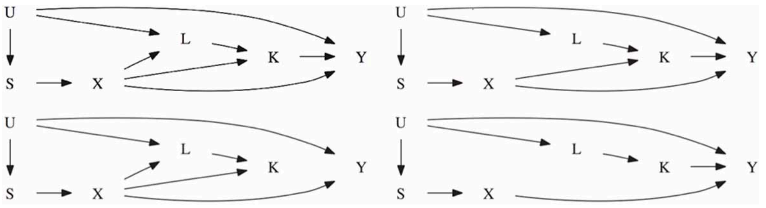


FIGURE 2 | The top left figure represents scenario 1. The top right figure represents scenario 2 which is the same as scenario 1 except that *X* does not cause *L*. The bottom left figure represents scenario 3 which is the same as scenario 1 except that *K* does not cause *Y*. The bottom right figure represents scenario 4 which is the same as scenario 2 except that *X* does not cause *K*.

Table 1 | This table displays the type-1 error rate for the test statistics using Model 1 [the Vansteelandt et al. test statistic (Vansteelandt et al., 2009)] or Model 2 (the screening test statistic) to estimate γ_1 for different allele frequencies.

Allele frequency (%)	Type-1 error rate when 1 SNP is tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.071	0.059	0.049	0.047	0.045	0.047	0.049	0.051	0.050
Scenario 1: Model 2	0.069	0.058	0.048	0.046	0.046	0.046	0.049	0.050	0.051
Scenario 2: Model 1	0.044	0.045	0.045	0.045	0.045	0.045	0.045	0.043	0.045
Scenario 2: Model 2	0.045	0.044	0.045	0.045	0.045	0.043	0.045	0.043	0.045
Scenario 3: Model 1	0.058	0.048	0.043	0.045	0.045	0.046	0.044	0.047	0.044
Scenario 3: Model 2	0.052	0.050	0.044	0.046	0.044	0.046	0.045	0.047	0.046
Scenario 4: Model 1	0.044	0.045	0.045	0.043	0.046	0.044	0.045	0.045	0.042
Scenario 4: Model 2	0.044	0.044	0.045	0.043	0.046	0.044	0.046	0.045	0.042

Table 2 | This table displays the power for the test statistics using Model 1 [the Vansteelandt et al. test statistic (Vansteelandt et al., 2009)] or Model 2 (the screening test statistic) to estimate γ_1 for different allele frequencies.

Allele frequency (%)	Power when 1 SNP is tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.264	0.363	0.448	0.504	0.576	0.629	0.669	0.692	0.706
Scenario 1: Model 2	0.241	0.361	0.444	0.508	0.581	0.633	0.671	0.696	0.710
Scenario 2: Model 1	0.180	0.302	0.406	0.492	0.564	0.610	0.649	0.667	0.686
Scenario 2: Model 2	0.180	0.302	0.408	0.491	0.563	0.610	0.646	0.666	0.685
Scenario 3: Model 1	0.265	0.365	0.449	0.504	0.581	0.632	0.669	0.696	0.712
Scenario 3: Model 2	0.246	0.361	0.451	0.510	0.586	0.634	0.671	0.699	0.716
Scenario 4: Model 1	0.175	0.304	0.408	0.499	0.558	0.607	0.647	0.671	0.681
Scenario 4: Model 2	0.174	0.303	0.407	0.498	0.557	0.605	0.648	0.672	0.682

Vansteelandt et al. (2009) has higher power and for higher allele frequencies the proposed method has higher power. However, this difference in power is negligible; the power never differs by more than 2%.

The advantage of our approach becomes clear when testing multiple SNPs. **Table 4** shows how the power to detect the causal SNP for our approach compares to Vansteelandt et al. (2009) when one SNP has a direct effect on the phenotype as simulated above in **Table 2** and 49 other SNPs are not associated with the phenotype of interest. **Table 1** shows the type-1 error rate in this scenario where the one SNP has an indirect effect on the phenotype as simulated above in **Table 1** and 49 other SNPs are not associated with the phenotype of interest or any of the other phenotypes. **Table 6** shows how the power to detect the causal SNP for our approach compares to Vansteelandt et al. (2009) when one SNP has a direct effect on the phenotype as simulated above in **Table 2** and 99 other SNPs are not associated with the phenotype of interest. **Table 5** shows the type-1 error rate in this scenario where the one SNP has an indirect effect on the phenotype as simulated above in **Table 1** and 99 other SNPs are not associated with the phenotype of interest or any of the other phenotypes.

Our approach allows for a screening step similar to the Van Steen algorithm (Van Steen et al., 2005) where the top 3 SNPs out of 50 and the top 5 SNPs out of 100 with the highest test statistic given by Equation (4) are chosen. We chose 3 SNPs out of 50 and 5 SNPs out of 100 since this is roughly 5% of the SNPs. After the top 3 or 5 SNPs are chosen based on the screening step, the test statistic described in Equation (7) is used to obtain a p-value which is compared to $\alpha/3$ and $\alpha/5$, respectively. We compare our approach with the screening step to the approach by Vansteelandt et al. (2009) with a Sidak correction. Since our approach allows for a screening step, we are better able to detect the SNP that has a direct causal effect on the target phenotype as seen in **Tables 4**, **6**.

Note that the power in **Tables 4**, **6** is lower than that in **Table 2** which is expected since multiple SNPs are tested. For more common allele frequencies, the power of using the proposed method with a screening step is more than double that of the Vansteelandt algorithm with a Sidak correction while the type-1 error rates are similar as seen in **Tables 3**, **5**. Therefore, if multiple SNPs are tested, the proposed approach has better power to detect the SNP that has a direct effect on the phenotype of interest.

Since the proposed approach is valid for testing, but may have less power when either the $X \rightarrow K$ or the $X \rightarrow Y$ link is strong, we looked at the effect of increasing the association between X and K when K influences Y ($X \rightarrow K$) and X and Y ($X \rightarrow Y$). We increased the correlation between X and K from 0.025 to 0.05 and then 0.075. We also increased the correlation between X and Y from 0.05 to 0.10 and then 0.15. The power of both statistics remained very close. At most, the power of the Vansteelandt et al. statistic (Vansteelandt et al., 2009) was 0.9% better than our approach.

Since the test statistic for the screening step given in Equation (4) is susceptible to population stratification, we examined a few scenarios where population stratification was present. We simulated half of the subjects to have allele frequency of 5, 5, 20, and 40% and the other half of the subjects to have allele frequency of 10, 45, 25, and 45%, respectively. Similar to **Tables 3**, **4**, one SNP has a direct effect on the phenotype of interest and 49 other SNPs are not associated with the phenotype of interest in

Table 3 | This table displays the significance rate when one SNP does not have a direct effect on the phenotype Y but acts as seen in Figure 2 without the arrow from X to Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Type-1 error rate when 50 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.0018	0.0008	0.0008	0.0006	0.0004	0.0006	0.0008	0.0006	0.0010
Scenario 1: Model 2	0.0014	0.0006	0.0002	0.0006	0.0012	0.0012	0.0006	0.0012	0.0006
Scenario 2: Model 1	0.0014	0.0006	0.0008	0.0012	0.0004	0.0008	0.0004	0.0008	0.0002
Scenario 2: Model 2	0.0004	0.0010	0.0012	0.0016	0.0012	0.0006	0.0010	0.0004	0.0006
Scenario 3: Model 1	0.0018	0.0006	0.0008	0.0014	0.0006	0.0010	0.0008	0.0008	0.0002
Scenario 3: Model 2	0.0014	0.0006	0.0008	0.0016	0.0012	0.0010	0.0012	0.0004	0.0006
Scenario 4: Model 1	0.0014	0.0006	0.0008	0.0012	0.0004	0.0008	0.0004	0.0008	0.0002
Scenario 4: Model 2	0.0008	0.0010	0.0013	0.0016	0.0012	0.0006	0.0010	0.0004	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (8) are tested.

Table 4 | This table displays the power when one SNP has a direct effect on the phenotype Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Power when 50 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.031	0.039	0.073	0.075	0.120	0.150	0.176	0.191	0.191
Scenario 1: Model 2	0.038	0.073	0.133	0.188	0.255	0.321	0.356	0.368	0.431
Scenario 2: Model 1	0.013	0.030	0.040	0.074	0.110	0.112	0.158	0.162	0.172
Scenario 2: Model 2	0.015	0.056	0.117	0.18	0.236	0.292	0.344	0.356	0.378
Scenario 3: Model 1	0.031	0.039	0.074	0.083	0.121	0.130	0.185	0.191	0.201
Scenario 3: Model 2	0.038	0.073	0.136	0.194	0.257	0.312	0.368	0.370	0.445
Scenario 4: Model 1	0.012	0.030	0.063	0.076	0.110	0.113	0.159	0.176	0.177
Scenario 4: Model 2	0.015	0.057	0.107	0.181	0.235	0.290	0.344	0.376	0.416

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (8) are tested.

Table 5 | This table displays the significance rate when one SNP does not have a direct effect on the phenotype Y but acts as seen in Figure 2 without the arrow from X to Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Type-1 error rate when 100 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.0010	0.0006	0.0004	0.0006	0.0007	0.0006	0.0002	0.0004	0.0005
Scenario 1: Model 2	0.0008	0.0004	0.0004	0.0006	0.0006	0.0006	0.0008	0.0002	0.0006
Scenario 2: Model 1	0.0006	0.0000	0.0008	0.0000	0.0000	0.0004	0.0002	0.0006	0.0002
Scenario 2: Model 2	0.0004	0.0004	0.0008	0.0002	0.0004	0.0006	0.0010	0.0004	0.0008
Scenario 3: Model 1	0.0010	0.0010	0.0002	0.0004	0.0000	0.0004	0.0002	0.0008	0.0000
Scenario 3: Model 2	0.0008	0.0004	0.0002	0.0002	0.0002	0.0006	0.0002	0.0002	0.0004
Scenario 4: Model 1	0.0006	0.0003	0.0004	0.0006	0.0007	0.0006	0.0002	0.0004	0.0005
Scenario 4: Model 2	0.0002	0.0004	0.0004	0.0006	0.0006	0.0006	0.0008	0.0002	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (8) are tested.

Table 6 | This table displays the power when one SNP has a direct effect on the phenotype Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency (%)	Power when 100 SNPs are tested								
	5	10	15	20	25	30	35	40	45
Scenario 1: Model 1	0.014	0.028	0.049	0.048	0.084	0.109	0.111	0.147	0.142
Scenario 1: Model 2	0.021	0.056	0.099	0.136	0.196	0.262	0.277	0.332	0.351
Scenario 2: Model 1	0.004	0.018	0.040	0.055	0.076	0.099	0.098	0.116	0.123
Scenario 2: Model 2	0.014	0.042	0.088	0.145	0.178	0.246	0.249	0.284	0.332
Scenario 3: Model 1	0.018	0.028	0.038	0.049	0.087	0.094	0.112	0.128	0.139
Scenario 3: Model 2	0.023	0.057	0.099	0.137	0.198	0.229	0.283	0.315	0.368
Scenario 4: Model 1	0.006	0.018	0.040	0.041	0.076	0.086	0.098	0.116	0.123
Scenario 4: Model 2	0.011	0.042	0.088	0.126	0.178	0.209	0.249	0.284	0.332

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (8) are tested.

Table 7 | This table displays the significance level when one SNP has an indirect effect on the phenotype Y as seen in Figure 2 without the arrow from X to Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency	Type-1 error rate when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.0012	0.0014	0.0011	0.0013
Scenario 1: Model 2	0.0006	0.0006	0.0004	0.0005
Scenario 2: Model 1	0.0010	0.0006	0.0004	0.0006
Scenario 2: Model 2	0.0012	0.0013	0.0018	0.0020
Scenario 3: Model 1	0.0009	0.0002	0.0004	0.0011
Scenario 3: Model 2	0.0008	0.0012	0.0016	0.0008
Scenario 4: Model 1	0.0006	0.0014	0.0008	0.0009
Scenario 4: Model 2	0.0009	0.0006	0.0006	0.0012

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 in a screening step where the three SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Table 8 | This table displays the power when one SNP has a direct effect on the phenotype Y and 49 SNPs are not associated with the phenotype Y.

Allele frequency	Power when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.025	0.070	0.111	0.171
Scenario 1: Model 2	0.064	0.199	0.248	0.394
Scenario 2: Model 1	0.016	0.070	0.103	0.163
Scenario 2: Model 2	0.040	0.205	0.227	0.366
Scenario 3: Model 1	0.025	0.070	0.113	0.172
Scenario 3: Model 2	0.064	0.202	0.249	0.396
Scenario 4: Model 1	0.016	0.064	0.103	0.163
Scenario 4: Model 2	0.040	0.186	0.227	0.366

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the three SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Tables 7, **8**. Similar to **Tables 5**, **6**, one SNP has a direct effect on the phenotype of interest and 99 other SNPs are not associated with the phenotype of interest in **Tables 9**, **10**. As seen in **Tables 7**, **9**, the type-1 error rates are similar for both methods. As seen in **Tables 8**, **10**, even though there is some population stratification present, the proposed method with a screening step still performs better than the Vansteelandt algorithm, especially when the allele frequencies are more common.

DATA ANALYSIS: AN APPLICATION TO THE FRAMINGHAM STUDY

We evaluated the practical relevance of the proposed adjustment principle by an application to the Framingham Heart Study with 1400 probands (Herbert et al., 2006). For the target phenotype, we selected the lung-function measurement FEV1. For the secondary phenotype K, we selected height. Gender, and age represent L, the collection of common risk factors between FEV1 and height. For rs2415815 a SNP associated with both height and FEV1, the test statistic equals 0.044 with corresponding p-value equal 0.83. As a result, we fail to reject the null hypothesis and conclude that there is no evidence that the SNP acts directly on FEV1 other than via body height.

DISCUSSION

Our proposed FBAT assesses the direct genetic effect of a marker locus on the phenotype of interest, other than through another correlated phenotype. The adjustment is based on the conditional mean model approach and can be incorporated into the FBAT-approach in a straightforward fashion. The power of the approach is assessed by simulation studies and shown to be similar to the Vansteelandt et al. method when only one SNP is being tested and superior when multiple SNPs are being tested (Vansteelandt et al., 2009). Unlike the Vansteelandt et al. method, this method

Table 9 | This table displays the significance level when one SNP has an indirect effect on the phenotype Y as seen in Figure 2 without the arrow from X to Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency	Type-1 error rate when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.0011	0.0005	0.0007	0.0003
Scenario 1: Model 2	0.0009	0.0006	0.0008	0.0003
Scenario 2: Model 1	0.0004	0.0015	0.0009	0.0005
Scenario 2: Model 2	0.0003	0.0011	0.0012	0.0005
Scenario 3: Model 1	0.0004	0.0010	0.0008	0.0004
Scenario 3: Model 2	0.0006	0.0009	0.0010	0.0006
Scenario 4: Model 1	0.0008	0.0013	0.0007	0.0004
Scenario 4: Model 2	0.0010	0.0008	0.0011	0.0006

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

Table 10 | This table displays the power when one SNP has a direct effect on the phenotype Y and 99 SNPs are not associated with the phenotype Y.

Allele frequency	Power when the following population stratification is present			
	5 and 10%	5 and 45%	20 and 25%	40 and 45%
Scenario 1: Model 1	0.022	0.050	0.073	0.157
Scenario 1: Model 2	0.044	0.141	0.170	0.324
Scenario 2: Model 1	0.014	0.046	0.071	0.148
Scenario 2: Model 2	0.036	0.137	0.161	0.298
Scenario 3: Model 1	0.022	0.050	0.076	0.159
Scenario 3: Model 2	0.045	0.143	0.174	0.326
Scenario 4: Model 1	0.014	0.046	0.071	0.148
Scenario 4: Model 2	0.036	0.137	0.161	0.298

Model 1 is used to estimate γ_1 with a Sidak correction and Model 2 to is used estimate γ_1 with a screening step where the five SNPs with the largest test statistic given by Equation (4) are tested. Population stratification is present such that half of the subjects have one of the allele frequencies listed and the other half of the subjects have the other allele frequency listed.

uses a screening step and has the unique advantage in situations in which a large number of SNPs are tested for a direct effect on the phenotype of interest. Since the number of tests will be much smaller than the total number of SNPs, this will lead to substantial reduction in the adjustment for multiple-comparisons and will result in improved overall statistical power. In this process, the screening step works under the assumption of no population admixture, but the final analysis of the selected SNPs is robust against it.

While we considered several causal scenarios, if the causal relationships assumed in the DAGs are not true this could cause problems for the proposed method. For example, a causal arrow $K \leftarrow Y$ or $L \rightarrow Y$ could introduce spurious association for this method. Therefore, one needs to makes sure that the assumptions of the DAG are met before using the proposed approach. While the simulations considered 50 and 100 SNPs, a realistic application could involve thousands of GWAS SNPs. This leads to extreme multiple test corrections and may lead to very different behavior than the behavior observed in the simulation studies (Morris and Elston, 2011). Furthermore, if phenotypes of the founders are known, the proposed method could perform poorly compared to population-based approaches.

For the screening step in the Simulations section, we chose 3 out of 50 and 5 out of 100 SNPs since this is roughly 5% of the tested SNPs. Another number of SNPs could be chosen for the screening step. Although, if the majority of SNPs are chosen in the screening step (i.e., 40 out of 50 SNPs), this increases the number of multiple comparisons and can decrease power. If too few SNPs are chosen in the screening step (i.e., 1 out of 50 SNPs), this decreases the number of multiple comparisons, but one may fail to detect the causal SNP since too few SNPs were chosen. Care needs to be given to the number of SNPs chosen in the screening step (Van Steen et al., 2005). One cannot simply choose different numbers of SNPs for the screening step until significant results are found since this will inflate the type-1 error rate (Van Steen et al., 2005).

ACKNOWLEDGMENTS

This work was funded by NIH/NHLBI U01 HL089856 Edwin K. Silverman, PI. COPDGene is supported by NHLBI Grant Nos U01HL089897 and U01Hl089856. This work was also funded by FWO grant G.0111.12.

REFERENCES

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. *Nat. Genet.* 40, 616–622. doi: 10.1038/ng.109

Berzuini, C., Vansteelandt, S., Foco, L., Pastorino, R., and Bernardinelli, L. (2012). Direct genetic effects and their estimation from matched case-control data. *Genet. Epidemiol.* 36, 652–662. doi: 10.1002/gepi.21660

Cole, S. R., and Hernan, M. A. (2002). Fallibility in estimating direct effects. *Int. J. Epidemiol.* 31, 163–165. doi: 10.1093/ije/31.1.163

Herbert, A., Gerry, N. P., McQueen, M. B., Heid, I. M., Pfeufer, A., Illig, T., et al. (2006). A common genetic variant is associated with adult and childhood obesity. *Science* 312, 279–283. doi: 10.1126/science.1124779

Lange, C., DeMeo, D., Silverman, E. K., Weiss, S. T., and Laird, N. M. (2003a). Using the noninformative families in family-based association tests: a powerful new testing strategy. *Am. J. Hum. Genet.* 73, 801–811. doi: 10.1086/378591

Lange, C., Lyon, H., DeMeo, D., Raby, B., Silverman, E. K., and Weiss, S. T. (2003b). A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. *Hum. Hered.* 56, 10–17. doi: 10.1159/000073728

Morris, N., and Elston, R. (2011). A note on comparing the power of test statistics at low significance levels. *Am. Stat.* 65, 164. doi: 10.1198/tast.2011.10117

Murphy, A., Weiss, S. T., and Lange, C. (2008). Screening and replication using the same data set: testing strategies for family-based studies in which all probands are affected. *PLoS Genet.* 4:e1000197. doi: 10.1371/journal.pgen.1000197

Rabinowitz, D., and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* 50, 211–223. doi: 10.1159/000022918

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 452, 638–642. doi: 10.1038/nature06846

VanderWeele, T. J., Asomaning, K., Tchetgen, E. J., Han, Y., Spitz, M. R., Shete, S., et al. (2012). Genetic variants on 15q25.1, smoking and lung cancer: an assessment of mediation and interaction. *Am. J. Epidemiol.* 175, 1013–1020. doi: 10.1093/aje/kwr467

Vansteelandt, S. (2010). Estimating direct effects in cohort and case-control studies. *Epidemiology* 21, 278. doi: 10.1097/EDE.0b013e3181cd72aa

Vansteelandt, S., Goetgeluk, S., Lutz, S., Waldman, I., Lyon, H., Schadt, E. E., et al. (2009). On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. *Genet. Epidemiol.* 33, 394–405. doi: 10.1002/gepi.20393

Vansteelandt, S., and Lange, C. (2012). Causation and causal inference for genetic effects. *Hum. Genet.* 131,1665–1676. doi: 10.1007/s00439-012-1208-9

Van Steen, K., McQueen, M. B., Herbert, A., Raby, B., Lyon, H., Demeo, D. L., et al. (2005). Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37, 683–691. doi: 10.1038/ng1582

Won, S., Bertram, L., Becker, D., Tanzi, R. E., and Lange, C. (2009). Maximizing the power of genome-wide association studies: a novel class ofpowerful family-based association tests. *Stat. Biosci.* 1, 125–143. doi: 10.1007/s12561-009-9016-z

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 June 2013; accepted: 25 October 2013; published online: 21 November 2013.

Citation: Lutz SM, Vansteelandt S and Lange C (2013) Testing for direct genetic effects using a screening step in family-based association studies. Front. Genet. ***4****:243. doi: 10.3389/fgene.2013.00243*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

APPENDIX

The following proof shows that the test statistics in the first and second screening steps are uncorrelated under the null hypothesis. As discussed in the introduction and methods sections, $\tilde{Y} = Y - \bar{y} - \gamma_1 K - \bar{k}$ is the adjusted phenotype for the effect that the phenotype K has on the target phenotype Y. For ease of notation, we will use $\tilde{Y} = Y - \gamma_1(K)$ for this proof. Suppose that the null hypothesis is true that X has no effect on Y other than through K. Let

$$E(Y|X, K, U) = E(Y|K, U) = \Phi\{w(U) + \gamma_1 K\} \tag{9}$$

where Φ equals the identity link or exponential link and $w(U)$ is an arbitrary function. Without loss of generality, for the following proof, let Φ equal the identity link. This model does not involve X because we are working under the null hypothesis of no direct effect. Furthermore, the parameter γ_1 in this model is the same as in model

$$E(Y|X, K, L, S) = w^*(X, L, S) + \gamma_1 K \tag{10}$$

for some function $w^*(X, L, S)$ of (X, L, S), which can be seen by inferring this model from model (9) upon noting that $Y \perp\!\!\!\perp (L, S)|X, K, U$ and $U \perp\!\!\!\perp K|L, X, S$. Using model (9) and model (10) and noting that $Y \perp\!\!\!\perp S|X, K, U$ and $X \perp\!\!\!\perp U|S$, then

$$\begin{aligned} E[\tilde{Y}(X - E[X|S])] &= E[(Y - \gamma_1 K)(X - E[X|S])] \\ &= E[w(U)(E[X|S, U] - E[X|S])] = 0 \end{aligned} \tag{11}$$

As a result of Equation (11) and model (9), the $\text{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S])$ can be written as follows

$$\begin{aligned} &\text{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) \\ &\quad = E[(Y - \gamma_1 K)^2 E(X|S)(X - E[X|S])] \\ &\quad = E[(Y - E(Y|X, K, U) + w(U))^2 E(X|S)(X - E[X|S])] \\ &\quad = \text{Part}_1 + \text{Part}_2 + \text{Part}_3 \end{aligned}$$

where

$$\begin{aligned} \text{Part}_1 &= E[\big(w(U)^2\big)\, E[X|S](X - E[X|S])] \\ \text{Part}_2 &= E[(2(w(U)(Y - E[Y|X, K, U]))\, E[X|S](X - E[X|S])] \\ \text{Part}_3 &= E[\big((Y - E[Y|X, K, U])^2\big)\, E[X|S](X - E[X|S])] \end{aligned} \tag{12}$$

We will show that the $\text{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) = 0$ by showing that each part of the above equation equals zero.

$$\begin{aligned} \text{Part}_1 &= E[w(U)^2 E[X|S](X - E[X|S])] \\ &= E[w(U)^2 E[X|S](E[X|S, U] - E[X|S])] = 0 \end{aligned} \tag{13}$$

because $X \perp\!\!\!\perp U|S$.

$$\text{Part}_2 = E[2w(U)(Y - E[Y|X, K, U])E[X|S](X - E[X|S])] = 0 \tag{14}$$

because $Y \perp\!\!\!\perp S|X, K, U$.

$$\begin{aligned} \text{Part}_3 &= E[(Y - E[Y|X, K, U])^2 E[X|S](X - E[X|S])] \\ &= E[E[X|S](X - E[X|S])\, Var[Y|K, U]] \end{aligned} \tag{15}$$

because $Y \perp\!\!\!\perp S|X, K, U$ and $Y \perp\!\!\!\perp X|K, U$.

Assuming that $\text{Var}(Y|K, U)$ is constant, as we do throughout, it is immediate that the term Part_3 is zero. As a result, this shows that $\text{Cov}(\tilde{Y}(X - E[X|S]), \tilde{Y}E[X|S]) = 0$.

frontiers in
GENETICS

METHODS ARTICLE
published: 01 July 2014
doi: 10.3389/fgene.2014.00188

An alternative hypothesis testing strategy for secondary phenotype data in case-control genetic association studies

Sharon M. Lutz[1]*, John E. Hokanson[2] and Christoph Lange[3,4,5,6]

[1] Department of Biostatistics, University of Colorado, Aurora, CO, USA
[2] Department of Epidemiology, University of Colorado, Aurora, CO, USA
[3] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[4] Channing Laboratory, Harvard Medical School, Boston, MA, USA
[5] Institute for Genomic Mathematics, University of Bonn, Bonn, Germany
[6] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

Edited by:
Nathan Morris, Case Western Reserve University, USA

Reviewed by:
Qing Lu, Michigan State University, USA
Geetha Chittoor, University of North Carolina at Chapel Hill, USA

Correspondence:
Sharon M. Lutz, Department of Biostatistics, University of Colorado, 13001 E. 17th Place, B119 Bldg. 500, W3128, Aurora, CO 80045, USA
e-mail: sharon.lutz@ucdenver.edu

Motivated by the challenges associated with accounting for the ascertainment when analyzing secondary phenotypes that are correlated with case-control status, Lin and Zeng have proposed a method that properly reflects the case-control sampling (Lin and Zeng, 2009). The Lin and Zeng method has the advantage of accurately estimating effect sizes for secondary phenotypes that are normally distributed or dichotomous. This method can be computationally intensive in practice under the null hypothesis when the likelihood surface that needs to be maximized can be relatively flat. We propose an extension of the Lin and Zeng method for hypothesis testing that uses proportional odds logistic regression to circumvent these computational issues. Through simulation studies, we compare the power and type-1 error rate of our method to standard approaches and Lin and Zeng's approach.

Keywords: secondary phenotype, case-control study, ascertainment, genetic association, proportional odds logistic regression

INTRODUCTION

For the analysis of secondary phenotype data collected in a case-control study, Lin and Zeng have proposed a method that properly reflects the case-control sampling (Lin and Zeng, 2009). This work is motivated by the challenges associated with accounting for the ascertainment when analyzing secondary phenotypes that are correlated with case-control status. Several methods have been proposed that accurately estimate the odds ratio of genetic variants for binary secondary phenotypes associated with case-control status, but most of these methods do not readily accommodate continuous secondary phenotypes (Greenland, 2003; Kraft, 2007; Richardson et al., 2007; Monsees et al., 2009; Li et al., 2010; Wang and Shete, 2011a,b; He et al., 2012; Li and Gail, 2012). While two of these methods use an inverse probability weighted (IPW) regression approach that can accommodate continuous secondary phenotypes, these methods focus on correcting for the bias in the estimator due to the ascertainment conditions and involve a known disease rate (Richardson et al., 2007; Monsees et al., 2009). Since this paper focuses on hypothesis testing versus estimation of disease-association parameters with an equal number of cases and controls, we do not present these methods here.

Alternatively, the Lin and Zeng method has the advantage of accurately estimating effect sizes for secondary phenotypes that are normally distributed or dichotomous (Lin and Zeng, 2009). Under the null hypothesis when the likelihood surface that needs to be maximized can be relatively flat, this method can be computationally intensive in practice. To circumvent these computational issues, we propose an extension of the Lin and Zeng method for hypothesis testing that uses proportional odds logistic regression. Since the approach by Lin and Zeng has the advantage that effect sizes can also be estimated, we recommend the following work-flow for the analysis of continuous secondary phenotypes.

1. Test all SNPs with our approach using proportional odds logistic regression since the vast majority of SNPs will be under the null hypothesis.
2. For the significant SNPs, apply Lin and Zeng's method to obtain parameter estimates and confidence intervals.

This proposed approach circumvents the computational issues encountered in the Lin and Zeng approach under the null hypothesis, but utilizes the Lin and Zeng's method to accurately estimate effect sizes for significant SNPs found in Step 1. Through simulation studies, we compare the power and type-1 error rate of our method to standard approaches and Lin and Zeng's approach.

METHODS

When the secondary phenotype is normally distributed, Lin and Zeng propose an adjusted score test that incorporates genetic associations with affection status into the test statistic and models the likelihood function as follows (Lin and Zeng, 2009):

$$\prod_{i=1}^{n} P(Y_i, X_i|D_i) = \prod_{i=1}^{n} \left\{ \frac{P(D_i = 1|X_i, Y_i)P(Y_i|X_i)P(X_i)}{P(D_i = 1)} \right\}^{D_i} \left\{ \frac{P(D_i = 0|X_i, Y_i)P(Y_i|X_i)P(X_i)}{P(D_i = 0)} \right\}^{1-D_i} \quad (1)$$

where D denotes the case-control status (1 = case and 0 = control), Y denotes the secondary phenotype, n denotes the total number of subjects, and X denotes the genotype of interest.

Lin and Zeng calculate $P(D_i = 1) = \sum_y \sum_x P(D_i = 1|x, y)P(y|x)P(x)$. The probability $P(D|X, Y)$ is defined as a logistic regression model. They model $P(Y|X)$ as a logistic regression for dichotomous Y or a linear regression for normally distributed Y. They maximize the likelihood with respect to $P(X)$ by the Newton Raphson algorithm. In this framework, likelihood based statistics (i.e., Wald, score, and likelihood-ratio statistics) can be used to make inference.

The Lin and Zeng approach requires the secondary phenotype to be normally distributed and the method can be problematic under the null hypothesis since the likelihood surface that needs to be maximized can be relatively flat. Since Lin and Zeng's method estimates the parameters in the model by maximizing the likelihood given in Equation (1), the approach is numerically exhaustive when testing a large number of SNPs where a majority of the SNPs are under the null hypothesis. This is a result of the maximization of the likelihood function being difficult under the null hypothesis, since the surface can be flat due to the ascertainment condition.

If the primary goal of the secondary phenotype analysis is hypothesis testing as opposed to estimation of disease-association parameters, an alternative approach is to use the following likelihood composition, which ultimately does not require maximizing a relatively flat likelihood surface. Therefore, for the association testing of secondary phenotypes in case-control studies, we propose using a simpler break down of the likelihood that requires few assumptions.

$$\prod_{i=1}^{n} P(Y_i, X_i|D_i) = \prod_{i=1}^{n} P(X_i|Y_i, D_i)P(Y_i|D_i) \tag{2}$$

Under the null hypothesis, X is independent of Y given D and any confounders. The likelihood ratio test becomes

$$LRT = -2ln\left(\frac{\prod_{i=1}^{n} P(X_i|D_i)P(Y_i|D_i)}{\prod_{i=1}^{n} P(X_i|Y_i, D_i)P(Y_i|D_i)}\right)$$

$$= -2ln\left(\frac{\prod_{i=1}^{n} P(X_i|D_i)}{\prod_{i=1}^{n} P(X_i|Y_iD_i)}\right) \sim \chi^2_{1df} \tag{3}$$

As a result, one only needs to model $P(X|D)$ and $P(X|Y, D)$. For an additive genetic model, i.e., $X = 0, 1, 2$, corresponding to allele counts, instead of modeling the likelihood function, one can use a cumulative logistic regression model with proportional odds proportional for $P(X|D)$ and the $P(X|Y, D)$ such that

$$\begin{gathered} logit[P(X \le j|Y, D)] = \alpha_{1j} + \delta_{1Y}Y + \delta_{1D}D \\ logit[P(X \le j|D)] = \alpha_{0j} + \delta_{0D}D \end{gathered} \tag{4}$$

for $j = 0, 1$. To control for any known confounders, these covariates can be added to Equation (4). This model assumes the same effect for different cumulative logits (Agresti, 2002). If assumptions are not met then we recommend a link function for which the response curve is non-symmetric or adding a dispersion parameter. For imputed dosages, j becomes the number of dosage levels minus one, meaning the levels of X in the cumulative logistic regression are increased to the number of dosage levels minus one.

SIMULATIONS

To assess the performance of this approach and compare it to Lin and Zeng's method, we conducted simulation studies following Lin and Zeng's manuscript with a MAF of 0.3, an additive mode of inheritance, and $\alpha = 0.01$ level of significance (Lin and Zeng, 2009). We also compared both of these methods to the standard case-only method, control only method and combined case and control method where both cases and controls are included in the analysis. For the model of the secondary quantitative trait Y and the disease D,

$$Y|X \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right) \tag{5}$$

$$P(D = 1|X, Y) = \frac{exp\,(\gamma_0 + \gamma_1 X + \gamma_2 Y)}{1 + exp\,(\gamma_0 + \gamma_1 X + \gamma_2 Y)} \tag{6}$$

where $\beta_0 = \sigma^2 = 1$, $\beta_1 = 0$ under the null hypothesis and $\beta_1 = -0.12$ under the alternative hypothesis. We let $\gamma_2 = log(2)$, γ_1 varies from 0 to $log(1.5)$, and γ_0 was chosen such that the disease rate is 1% or 5%. For each combination of simulation parameters, we generated 1000 data sets with 500 cases and 500 controls.

Figure 1 shows the type 1 error rates and power for a disease rate of 1% and 5%. Our method, using the proportional odds logistic regression, maintains the type 1 error rate and has slightly higher power as compared to Lin and Zeng's method and superior power compared to the other methods. While the proposed method and Lin and Zeng's method have similar power, the proposed method is computationally more feasible under the null hypothesis than Lin and Zeng's method since it does not involve maximizing a relatively flat likelihood surface. The computing time for the proposed approach is under 1 s per SNP where as the software associated with the Lin and Zeng approach needs to be run multiple times if there are issues with convergence which can take 5 min to an hour per SNP. When running a GWAS with about 500,000 SNPs, this difference in computing time per SNP can be substantial. To examine this concept further, the plot on the left in **Figure 2** shows the log Likelihood specified by Lin and Zeng for varying values of β_0 and β_1 with all other parameters fixed at their true values and for data generated under the null hypothesis with $\gamma_1 = log(1.5)$ and the disease rate equal 5%. The plot on the right is the log Likelihood specified by Lin and Zeng for varying values of γ_1 and γ_2 with all other parameters fixed at their true values, and for data generated under the null hypothesis with $\gamma_1 = log(1.5)$ and the disease rate equals 5%. The red dots on the

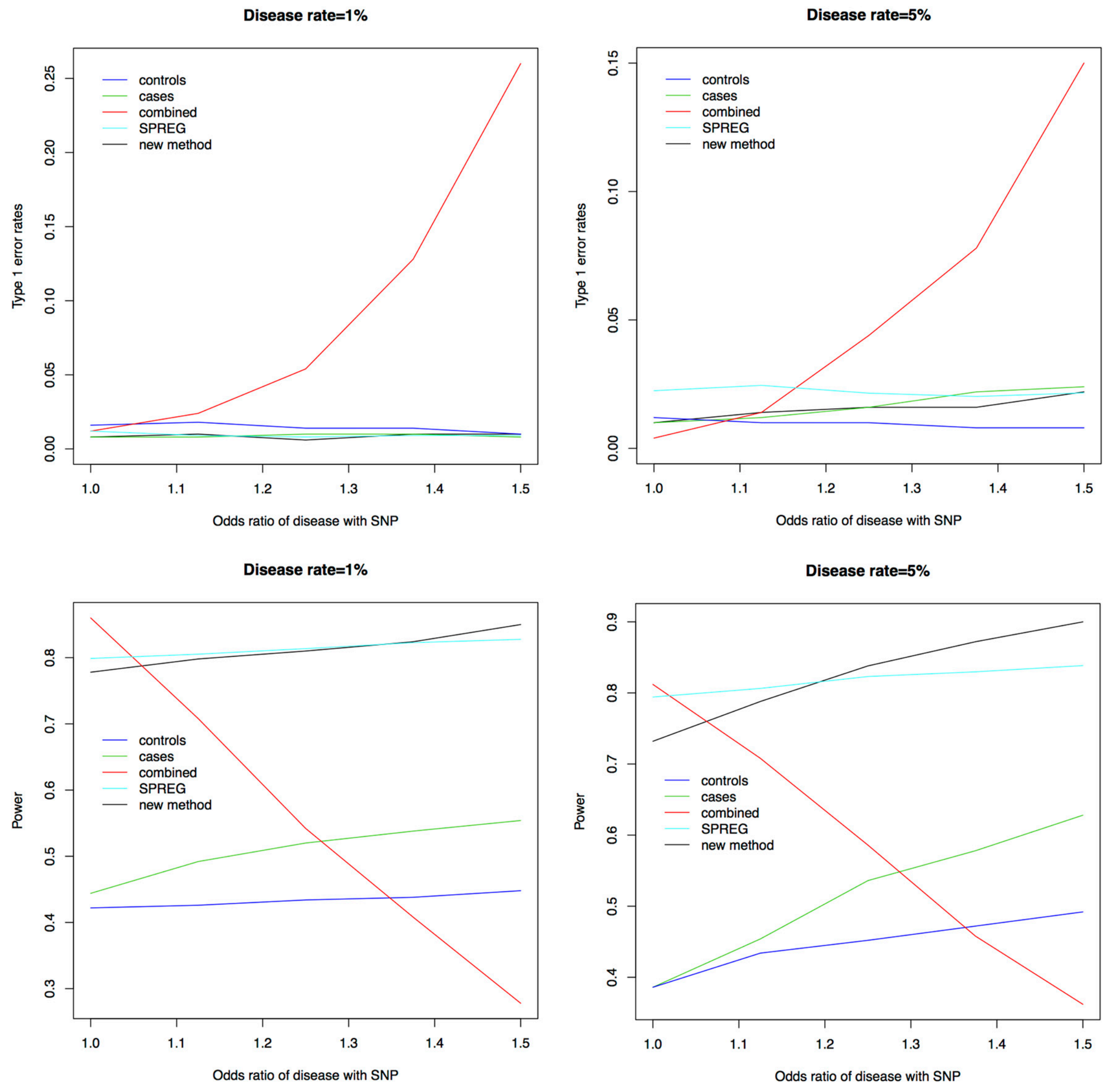


FIGURE 1 | Type 1 error rates and power for a disease rate of 1% and 5%. As seen in the plots above the new method using proportional odds logistic regression maintains the type 1 error rate. The new method has similar power compared to Lin and Zeng's method called SPREG and superior power compared to the other methods.

plots represent the true maximum. The surface for β_0 and β_1 has a clear maximum whereas the surface for γ_1 and γ_0 is relatively flat, demonstrating the difficulty in maximizing the likelihood surface defined by Lin and Zeng under the null hypothesis.

DISCUSSION

While the power of the proposed method is comparable to the method of Lin and Zeng, the proposed approach does not have the issue of maximizing a flat likelihood surface under the hull hypothesis that can be computationally intensive. Since the proposed approach is limited in it's ability to accurately estimate effect sizes while the approach by Lin and Zeng has the advantage that effect sizes can be accurately estimated, we recommend the following work-flow for the analysis of secondary phenotypes.

1. Test all SNPs with the proposed approach using proportional odds logistic regression since the vast majority of SNPs will be under the null hypothesis.
2. For the significant SNPs, apply Lin and Zeng's method to obtain parameter estimates and confidence intervals.

By using our approach to test all the SNPs in the GWAS, the hypothesis testing can be done quickly and efficiently since our approach does not suffer from this issue of maximizing a flat likelihood surface under the null hypothesis. By obtaining parameter estimates for only the significant SNPs with Lin and Zeng's method, one can make sure that the likelihood is properly maximized which is too computational exhaustive to apply to the entire GWAS.

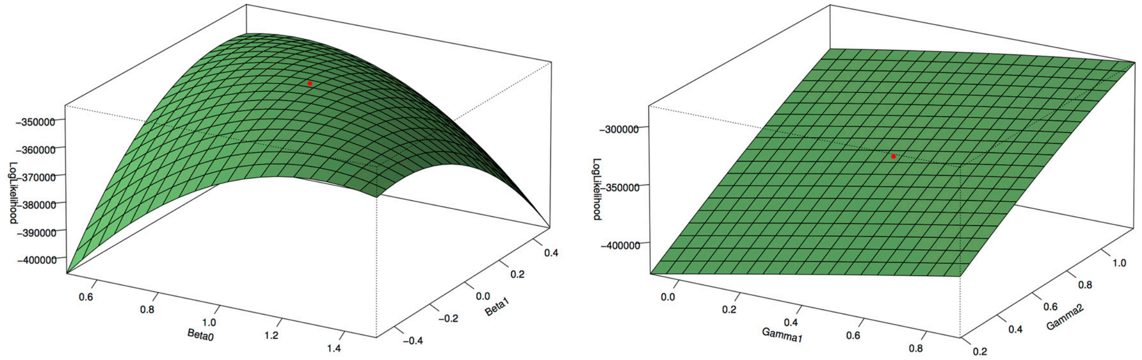


FIGURE 2 | Log Likelihood surface specified by Lin and Zeng. The plot on the left is the log Likelihood specified by Lin and Zeng for varying values of β_0 and β_1 with all other parameters fixed at their true values and for data generated under the null hypothesis with $\gamma_1 = log(1.5)$ and the disease rate equal 5%. The plot on the right is the log Likelihood specified by Lin and Zeng for varying values of γ_1 and γ_2 with all other parameters fixed at their true values and for data generated under the null hypothesis with $\gamma_1 = log(1.5)$ and the disease rate equal 5%. The red dots on the plots represent the true maximum. The surface for β_0 and β_1 has a clear maximum whereas the surface for γ_1 and γ_0 is relatively flat, demonstrating the difficulty in maximizing the likelihood surface defined by Lin and Zeng under the null hypothesis.

There are potential limitations associated with this strategy of combining two methodological approaches to reduce the computational burden while still being able to estimate the parameters of interest. While the two approaches have comparable power, a relatively small number of SNPs that are significant from the new approach may not be significant in the Lin and Zeng's method and vice versa. Also both approaches may have issues if the case control status is extremely correlated with the secondary phenotype. In this case, the secondary phenotype is not providing new information compared to the case-control status and these methods for testing secondary phenotypes in case-control genetic association studies are not applicable.

ACKNOWLEDGMENTS

This work was funded by NIH/NHLBI U01 HL089856 Edwin K. Silverman, PI. COPDGene is supported by NHLBI Grant Nos U01HL089897 and U01Hl089856.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: Wiley Series in Probability and Statistic. doi: 10.1002/0471249688

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 14, 300–306. doi: 10.1097/00001648-200305000-00009

He, J., Li, H., Edmondson, A. C., Rader, D. J., and Li, M. (2012). A Gaussian copula approach for the analysis of secondary phenotypes in case control genetic association studies. *Biostatistics* 3, 497–508. doi: 10.1093/biostatistics/kxr025

Kraft, P. (2007). Letter to the editor: analyses of genome-wide association scans for additional outcomes. *Epidemiology* 18, 838. doi: 10.1097/EDE.0b013e318154c7e2

Li, H., and Gail, M. H. (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Hum. Hered.* 73, 159–173. doi: 10.1159/000338943

Li, H., Gail, M. H., Berndt, S., and Chatterjee, N. (2010). Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet. Epidemiol.* 34, 427–433. doi: 10.1002/gepi.20495

Lin, D. Y., and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *J. Genet. Epidemiol.* 33, 256–265. doi: 10.1002/gepi.20377

Monsees, G. M., Tamimi, R. M., and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genet. Epidemiol.* 33, 717–728. doi: 10.1002/gepi.20424

Richardson, D. B., Rzehak, P., Klenk, J., and Weiland, S. K. (2007). Analyses of case control data for additional outcomes. *Epidemiology* 18, 441–445. doi: 10.1097/EDE.0b013e318060d25c

Wang, J., and Shete, S. (2011a). Power and type I error results for a bias-correction approach recently shown to provide accurate odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet. Epidemiol.* 35, 739–743. doi: 10.1002/gepi.20568

Wang, J., and Shete, S. (2011b). Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet. Epidemiol.* 35, 190–200. doi: 10.1002/gepi.20611

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 04 June 2014; published online: 01 July 2014.

Citation: Lutz SM, Hokanson JE and Lange C (2014) An alternative hypothesis testing strategy for secondary phenotype data in case-control genetic association studies. Front. Genet. ***5****:188. doi: 10.3389/fgene.2014.00188*

This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics.

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on