Genetic Epidemiology 33:299–307 (2009)

Selection of the Most Informative Individuals From Families With Multiple Siblings for Association Studies

Chunyu Liu,[1]* Qiong Yang,[2] L. Adrienne Cupples,[2] James B. Meigs[3,4] and Josée Dupuis[2]

[1]*Genetics and Genomics, Biogen Idec, Cambridge, Massachusetts*
[2]*Department of Biostatistics, Boston University, Boston, Massachusetts*
[3]*General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts*
[4]*Harvard Medical School, Boston, Massachusetts*

Association analyses may follow an initial linkage analysis for mapping and identifying genes underlying complex quantitative traits and may be conducted on unrelated subsets of individuals where only one member of a family is included. We evaluate two methods to select one sibling per sibship when multiple siblings are available: (1) one sibling with the most extreme trait value; and (2) one sibling using a combination score statistic based on extreme trait values and identity-by-descent sharing information. We compare the type I error and power. Furthermore, we compare these selection strategies with a strategy that randomly selects one sibling per sibship and with an approach that includes all siblings, using both simulation study and an application to fasting blood glucose in the Framingham Heart Study. When genetic effect is homogeneous, we find that using the combination score can increase power by 30–40% compared to a random selection strategy, and loses only 8–13% of power compared to the full sibship analysis, across all additive models considered, but offers at least 50% genotyping cost saving. In the presence of genetic heterogeneity, the score offers a 50% increase in power over a random selection strategy, but there is substantial loss compared to the full sibship analysis. In application to fasting blood sample, two SNPs are found in common for the selection strategies and the full sample among the 10 highest ranked single nucleotide polymorphisms. The EV strategy tends to agree with the IBD-EV strategy and the analysis of the full sample. *Genet. Epidemiol.* 33:299–307, 2009.

Key words: linkage analysis; association study; linkage disequilibrium; identity-by-descent (IBD)

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: NIH NCRR; Contract grant number: 1S10RR163736-01A1; Contract grant sponsor: National Heart, Lung and Blood Institute's Framingham Heart Study; Contract grant number: N01-HC-25195; Contract grant sponsor: American Diabetes Association.
*Correspondence to: Chunyu Liu, Department of Drug Discovery, Biogen Idec, 12 Cambridge Center, Cambridge, MA.
E-mail: chunyu.liu@biogenidec.com
Received 17 May 2008; Revised 17 May 2008; Accepted 13 September 2008
Published online 21 November 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20380

INTRODUCTION

Mapping and identifying complex disease genes are difficult because there is no simple correspondence between genotype and phenotype [Palmer and Cardon, 2005]. Some complex traits are measured on a continuous scale, such as weight and height. Some complex disease phenotypes are defined by threshold models applied to continuous traits. For example, obesity is often defined as high body mass index (BMI). Therefore, it is important to understand the genetics of underlying continuous traits in order to understand the traits themselves or their associated disease phenotypes. Because most complex quantitative traits may be the result of multiple loci, each contributing modestly to the total heritability, more efficient study designs need to be considered to increase power to detect disease signals for both linkage analysis and association studies.

Sibling pairs are often used for linkage analysis. Previous studies have shown that sibling pairs ascertained on their phenotypes are more powerful than randomly selected sibling pairs from the population for quantitative trait linkage analysis [Carey and Williamson, 1991; Eaves and Meyer, 1994; Risch and Zhang, 1995]. The effect of ascertainment on power has also been explored in population-based quantitative association studies. Truncated selection (TS) picks a group of individuals with high trait values, which serve as "cases," and a random group, which serve as "controls," and a typical case-control analysis is performed [Slatkin, 1999]. Chen et al. [2005] extend the TS approach to a two-sided selection approach (t-TS), in which the controls are selected with low trait values rather than a random sample. This extension is shown to further improve power. Chen et al. [2005] also suggest a more feasible approach of extreme random selection. Wang and Elston [2006b] examine various selection strategies analytically and with simulations. They find that the t-TS method is most powerful for a trait locus segregating common alleles with similar effects.

Association studies often follow an initial linkage analysis, targeting these regions identified by linkage analysis for fine mapping purposes. In order to address the question of how to design an efficient association study given information provided by the previous linkage analysis, Fingerlin et al. [2004] propose using the identity-by-descent (IBD) sharing information to select one case per sibship from sibships with multiple affected siblings. In the context of quantitative traits, Wang and Elston

[2006a] derive a quantitative linkage score (QLS) based on Haseman-Elston regression [Elston et al., 2000; Haseman and Elston, 1972] and make use of this score to select a subsample from the linkage study sample. They show that subjects selected using the QLS tend to be more homogenous than a random sample, and therefore greatly improve the power of the association study. In addition, they propose a test to determine if an associated variant in a linkage region can explain, in part, the detected linkage signal.

An association study using all available siblings in a family should be more powerful than a study using a single selected sibling per family, simply because the effective sample size is larger. However, genotyping all siblings (AS) is costly and budgets remain a major limiting factor for most studies. The aim of our study is to select the most informative sibling per sibship with multiple siblings for an association study following a quantitative linkage analysis. In the "Methods" section, we propose two methods to select one sibling per sibship with multiple siblings: (1) one sibling with the most extreme trait value; and (2) one sibling using a combination score statistic based on trait values and IBD-sharing information. The proposed score statistic, $S(j)$, belongs to the framework of Haseman and Elston (H-E) revisited [Elston et al., 2000] and is an extension of the QLS statistic introduced by Wang and Elston [Wang and Elston, 2006a]. The QLS statistic is a sibship-specific score and is used to select sibships with evidence of linkage in the context of population heterogeneity, resulting in more homogeneous samples. The $S(j)$ statistic is a sibling-specific score and is used to select an individual with an extreme trait value and increased allele sharing with other siblings in a sibship with similar trait values.

Furthermore, we describe a simulation study to compare these selection strategies with a strategy that randomly selects one sibling per sibship and the strategy that includes AS. The simulation study is carried out in both homogeneous and heterogeneous sample conditions. In the "Results" section, we describe our simulation results along with an application to fasting plasma glucose (FPG) in the Framingham Heart Study (FHS). We put our selection strategies in the context of other ascertainment approaches and discuss their advantages and limitations in the final section.

METHODS

We first define a score statistic to be used as one of the strategies in selecting siblings from families with multiple siblings. Definitions of six possible selection strategies follow, with a description of a simulation study to compare these six selection approaches. We evaluate these selection strategies in homogeneous samples and in heterogeneous samples consisting of subpopulations with different quantitative trait locus (QTL) effects. Finally, we describe the FHS sample to illustrate our proposed methods.

SCORE STATISTIC

We define a combination score statistic for the jth sibling as

$$S(j) = \sum_{i \neq j} \frac{(x_i - \mu)(x_j - \mu)[\hat{\pi}_{(i,j)} - \frac{1}{2}]}{N_{sib} - 1},$$

where the sum is taken over by only those sibling pairs that include sibling j. Here, $\hat{\pi}_{(i,j)}$ is the estimated multipoint IBD sharing for siblings i and j in a sibship at the locus with maximum evidence for linkage or with maximum LOD score (MLS); and x_i and x_j are phenotype values for sib i and j, respectively. The population mean for the phenotype is μ, and N_{sib} is the number of siblings in the sibship. The use of the N_{sib} in the denomination in $S(j)$ does not alter ranking of the sibs within a sibship; however, it is used to standardize $S(j)$ across sibships. Because the siblings have equal $S(j)$ scores for sibpairs, we consider shipships containing more than two siblings in our study. The siblings selected using the score statistic $S(j)$ tend to fall in two categories: (1) a sibling with large absolute trait value and high IBD sharing with other siblings with similar trait values; (2) a sibling with large absolute trait value and low IBD sharing with other siblings with dissimilar trait values.

SELECTION STRATEGIES

We consider six selection strategies to select one sibling from a sibship with multiple siblings. For the "all random" or AR strategy, one sibling is randomly selected from each sibship yielding a sample of unrelated subjects. In the second approach one sibling with the highest absolute phenotype value $|x_i|$ (i.e., the extreme value or the EV strategy) is selected. The third approach consists of selecting one sibling with the highest score statistic $S(j)$ using the population mean (IBD-EV). The fourth strategy is an alternative to the IBD-EV strategy, replacing the population mean with the sibship-specific mean in $S(j)$ $S(j)_2 = \sum_{i \neq j} \frac{(x_i - m)(x_j - m)[\hat{\pi}_{(i,j)} - \frac{1}{2}]}{n_{sib} - 1}$ or the IBD-EV2 strategy. When a sample consists of some large sibships or when the study population is heterogeneous (i.e., the QTL explains different genetic variance of the subpopulations), the sibship-specific mean might be more appropriate as an offset in the $S(j)$ score and thus provide better power. The fifth strategy is based on one sibling with the highest absolute phenotype value $|x_i|$, restricted to linked families (ML-EV). The sixth selection is based on one sibling with the highest score statistic $S(j)$, restricted to linked families (ML-IBD-EV). In order to evaluate cost-effectiveness for these selection strategies, we also consider a reference group that consists of AS.

COMPARISON OF SELECTION STRATEGIES

We compare the power and type I error for the selection strategies in tests for association between a single nucleotide polymorphism (SNP) in linkage disequilibrium (LD) with a QTL influencing the phenotype of interest, assuming that genotypes at the QTL are not observed. For each sample of unrelated individuals selected from different strategies, the difference between the phenotype means of the genotype groups is assessed using an F statistic derived from an analysis of variance (ANOVA) approach. Because some selection strategies result in a sample with non-normally distributed traits, statistical

significance is further evaluated with a permutation test [Fisher, 1935; Good, 2005]. In a permutation test, the phenotypic values are randomly shuffled among the unrelated individuals and the test statistic is computed using the shuffled data. Permuted phenotypes should not be associated with their non-permutated genotypic counterparts and by repeating the permutation multiple times (20,000 in our case). One obtains an estimate of distribution of the statistic under the null hypothesis.

Generalized estimating equations (GEE) methods are used for the analysis of all siblings in order to account for correlation among siblings within the same sibship.

SIMULATIONS

The goal of this research is to utilize linkage information and quantitative trait values to select the most informative individuals for follow up association studies. Therefore, we consider a QTL that might consist of a cluster of genetic loci with moderate to high heritability, which could be detected by linkage analysis. A set of microsatellite markers, 1 or 5 cM apart, are simulated on both sides of the QTL, assuming Hardy-Weinberg and linkage equilibrium. Parental chromosomes are generated according to population allele frequencies and are passed down to offspring after introducing crossovers between markers according to the Haldane mapping function with a rate of one crossover per Morgan [Haldane, 1919]. The phenotype values for parents and offspring are generated according to a variance component model [Amos, 1994].

In order to evaluate type I error and power of different selection strategies, we generate six SNPs in the region of interest (i.e., under the linkage peak). One SNP marker is in LD (measured by r^2) with the QTL and referred to as "LD SNP." By varying the r^2, the LD SNP can represent one of a cluster of genetic loci that contribute small to modest genetic effect toward the QTL. Five other SNPs, with MAFs $p = 0.05$, 0.1, 0.2, 0.35 and 0.5, are also generated in the same region. These five SNPs are not in LD with the QTL, and thus, represent the null and can be used to estimate type I error.

The power of associations is estimated at 0.1% significance level ($\alpha = 0.001$) with 10,000 simulation replicates, because our methods are likely to be applied to multiple SNPs under a linkage peak, and therefore some correction for multiple testing is appropriate. We choose three nominal type I error rates ($\alpha = 0.001$, 0.01, and 0.05) for all genetic models and generate 100,000 simulation replicates for evaluation of type I error.

A sample of 300 unascertained sibships are generated for both homogeneous and heterogeneous samples. When the sample is homogeneous, we consider a QTL with total heritability of 30% for most genetic models. Without loss of generality, we set $\sigma_T^2 = 1$ and calculate the additive and dominant effects, a and d, on the basis of the assumed QTL heritability (H) and minor allele frequency, p. Note that the additive effect corresponds to half the average phenotype difference between the two homozygous genotype groups, and d is defined by the genetic model: $d = 0$, a and $-a$ for an additive, dominant, and recessive model, respectively. We consider a number of scenarios with different combinations of QTL allele frequencies (p), r^2, sibship sizes, and mode of inheritance. In more detail, we compare power for selection strategies with (1) a range of $p = 0.1$, 0.2, and 0.4 for additive and dominant models and $p = 0.3$, 0.4, and 0.7 for recessive models with $s = 3$ and a fixed $r^2 = 0.1$; (2) varying s ($s = 3$, 4, and 5) with $p = 0.4$ and a fixed $r^2 = 0.1$; and (3) varying effect size of SNP (LD = 0.05–0.20, incremented by 0.05) with $s = 3$ and $p = 0.4$. These scenarios are summarized in Table I.

Because a study population is likely to be heterogeneous in many situations, we further evaluate the selection strategies in 300 sibships consisting of subpopulations. The simulation of heterogeneous samples is similar to what has been described in Wang and Elston [2006a] with minor modifications. The first sample structure consists of two subpopulations. A QTL with 40% total heritability is present in subpopulation 1 only. There is no QTL effect in subpopulation 2. We vary the proportion of the subpopulation 1 (q_1) from 0.5 (this sample is denoted as Het1), 0.7 (Het2), and 0.9 (Het3), resulting in samples with increasing homogeneity. Another sample structure consists of four equally sampled subpopulations with a QTL effect explaining 0, 10, 20, and 40% of the total variance in subpopulations 1–4, respectively, this sample is denoted as Het4. For these four heterogeneity models, we compare power and type I error for the selection strategies. The heterogeneity scenarios are summarized in Table III.

All simulations are carried out using the R (version 2.6.1) language (http://www.r-project.org). Merlin software [Abecasis et al., 2002] is utilized to estimate multipoint IBD sharing among individuals and to carry out variance component linkage analysis to identify the MLS and sibships with evidence for linkage. We define a sibship as showing evidence for linkage if the sibship-specific LOD ≥ 0 at the location of the MLS.

TABLE I. Characteristics of the simulated genetic models: homogeneous populations

Model	H%	p	a	MLS (N_{link})	Model R^2 (%)
Additive					
[A1]	30	0.1	1.3	2.7 (169)	3.4
[A2]	30	0.2	1.0	2.7 (170)	3.3
[A3]	30	0.4	0.8	2.7 (170)	3.2*
[A4]	20	0.4	0.7	1.4 (161)	3.3
Dominant					
[D1]	30	0.1	0.7	2.7 (170)	3.2
[D2]	30	0.2	0.6	2.6 (172)	3.0
[D3]	30	0.4	0.6	2.4 (173)	2.6*
[D4]	20	0.4	0.5	1.3 (170)	2.6
Recessive					
[R1]	30	0.3	1.1	1.7 (169)	1.7
[R2]	30	0.4	0.8	2.2 (169)	2.2*
[R3]	30	0.7	5.5	2.5 (186)	2.9
[R4]	20	0.4	0.6	1.2 (169)	2.1

H, heritability; p, minor allele frequency for the QTL; a, the additive effect; MLS, the maximum LOD score; N_{link}, average number of linked families; R^2, the proportion of phenotypic variability explained by the LD SNP; $s = 3$ for all models in this table. The $r^2 = 0.1$ between the LD SNP and the QTL for all scenarios described in this table. The asterisk (*) denotes that additional genetic models are evaluated with the same H, p, and a under various LD ($r^2 = 0.05$, 0.1, 0.15, and 0.20) (Table II).

TABLE II. Phenotype variability that can be explained by LD-SNP markers (%)

$r^2_{(\text{LD SNP-QTL locus})}$	0.05	0.10*	0.15	0.20	1
R^2 [A3]	1.9	3.2	4.8	6.3	30
R^2 [D3]	1.5	2.6	3.9	5.1	30
R^2 [R2]	1.2	2.2	3.1	5.7	30

The asterisk (*) denotes that scenario is also described in Table I; $r_{(\text{SNP}-\text{D locus})}$ represents linkage disequilibrium between an LD SNP and the QTL locus.

TABLE III. Summary statistics of the simulated genetic models: heterogeneous populations

Model	p	a	MLS (N_{link})	Model R^2 (%)
Additive				
[Het1]	0.1	1.5	1.5 (161)	0.02
[Het2]	0.1	1.5	2.4 (165)	0.03
[Het3]	0.1	1.5	3.9 (170)	0.05
[Het4]	0.1	0, 0.7, 1.1 and 1.5	1.4 (160)	0.02
[Het1]	0.2	1.1	1.5 (162)	0.02
[Het2]	0.2	1.1	2.4 (167)	0.03
[Het3]	0.2	1.1	3.8 (174)	0.05
[Het4]	0.2	0, 0.6, 0.8, and 1.1	1.3 (163)	0.02
[Het1]	0.4	0.9	1.4 (166)	0.02
[Het2]	0.4	0.9	2.4 (171)	0.03
[Het3]	0.4	0.9	3.8 (175)	0.05
[Het4]	0.4	0, 0.5, 0.6 and 0.9	1.2 (164)	0.02

Het1, Het2, and Het3 consist of two subpopulations. The QTL effect with total heritability of 40% only exists in subpopulation 1. There is no QTL effect in subpopulation 2. We vary proportion of the subpopulation 1 or q_1 at 0.5 (this resulting sample is denoted as Het1), 0.7 (Het2), and 0.9 (Het3) resulting in samples with increase in homogeneity. The Het4 consists of four equally distributed subpopulations with the QTL effect explaining 0, 10, 20, and 40% of the total variance. p, minor allele frequency for the QTL; a, the additive effect; MLS, the maximum LOD score; N_{link}, average number of linked families; R^2, the proportion of phenotypic variability explained by the LD SNP; $s = 3$ for all models in this table. The $r^2 = 0.15$ between the LD SNP and the QTL for all scenarios described in this table.

APPLICATION TO THE FHS DATA

Detailed information for the study sample and genotyping methods are described elsewhere [Cupples et al., 2007]. In brief, a subset of 1345 FHS participants were selected from the largest 330 extended pedigrees and genotyping was conducted using the Affymetrix 100K GeneChip. Original genome-wide linkage analyses were performed using 1,341 subjects from the 310 full pedigrees with a subset of SNPs selected to minimize LD (all pair-wise $D' < 0.5$) combined with 613 microsatellite markers [Cupples et al., 2007]. A region of linkage was identified for FPG measured at the 5th exam LOD > 3.0 at about 72–77 cM on chromosome 10. We chose this region for comparing the selection strategies for association analysis.

Because the selection strategies considered application to nuclear families, we break down the 330 pedigrees into 441 sibships with two or more siblings with phenotypes and genotypes available. The 441 sibships include 119 sibling pairs, 161 sibships of size 3, 98 sibships of size 4, and 63 sibships of size 5 or more. These sibships are used for selection of siblings for association studies. A set of 467 SNPs with pair-wise maximum LD (measured as D') < 0.5 are selected on chromosome 10 to evaluate the linkage evidence in this sub-sample. The average distance between any two SNPs is about 289.5 kb. Multipoint IBD sharing is calculated using Merlin [Abecasis et al., 2002]. In order to control for confounding from known risk factors and to increase the ability to detect genetic signals, standardized residuals were created using multiple linear regression models [Cupples et al., 2007] adjusting for sex, age, age squared, and BMI. Furthermore, rank normalized residuals are created for linkage analysis because departure from the normality assumption may lead to spurious evidence for linkage [Allison et al., 1999].

Three selection strategies are applied to the FHS data to select one sibling per sibship, yielding a sample of 441 subjects or about 33% of 1,345 siblings for association study: (i) the AR strategy; (ii) the EV strategy; (iii) the strategy based on the combination score statistic (IBD-EV). Association analyses are performed for the SNPs that are within the corresponding 1.0-LOD support interval from the linkage analysis in this study. For each SNP, we compare association test results using these three strategies to that of GEE using the original 330 pedigrees structure. The distribution of the standardized residuals of FPG is skewed to the right (data not shown) and thus, permutation test are performed for the three selection strategies.

RESULTS

We first present summary statistics for the genetic models, followed by power and type I error comparisons for the different selection strategies in our simulations. Then we describe the results obtained with FPG in the FHS.

SIMULATION RESULTS

GENETIC MODELS

Table I presents summary statistics of the simulated genetic models with sibship size $s = 3$ in homogeneous samples. The 10,000 simulation replicates are used to estimate the MLS, the number of linked families and the model R-square (R^2). Across additive, dominant, and recessive models, the average MLS is under 3.0 and the average number of linked sibships is less than 60% (56–58%) for $H = 30\%$ and 300 sibships of size $s = 3$.

The average R^2, corresponding to the estimate of genetic heritability explained by the LD SNP, is displayed in Table II for selected additive, dominant, and recessive genetic models for a range of LD (r^2). The QTL ($r^2 = 1$) explains about 30% phenotypic variability and equals model heritability H. In general, R^2 is largest for additive models and smallest for recessive models when other parameters are held constant; moreover, R^2 increases when LD increases for all genetic models.

Table III displays the average model R^2 and the average MLS over 10,000 simulation replicates of the simulated genetic models for heterogeneous samples.

TYPE I ERROR

Table IV displays the type I error for an additive model with $H = 30\%$, $p = 0.2$, and $s = 3$ when the study population is homogeneous. Across all models, the type I error is not affected by the unassociated SNPs with MAFs in the range of 0.05–0.5 for all selection strategies at all α levels. The sample with all siblings (GEE analysis), in general, exhibits slightly inflated type I error rate at each α level and this inflation of the type I error is similar for all sibship sizes. We further evaluate the type I error rates for the EV, IBD-EV, ML-EV, and ML-IBD-EV strategies using the permutation test at $\alpha = 0.001$, 0.01, and 0.05 based on a total 100,000 replicates. The permutation type I error rates are similar to the ANOVA type I error (data not shown). The type I error for other genetic models in both homogeneous and heterogeneous samples is similar to that presented in Table IV.

POWER

The power to detect the LD SNP—quantitative trait association is estimated by simulation in a sample of 300 unascertained sibships for all selections strategies and the approach using AS. The AR strategy randomly selects one sib per sibship, resulting in a group of subjects that represents the underlying population and is expected to have lower power. The AS sample uses all siblings and is expected to be more powerful than all selection strategies. Thus, the AR strategy and the AS sample define the lower and upper power range and are used as reference points for power comparisons for other selection strategies. For both homogeneous and heterogeneous sample conditions, we compare power for additive models and point out where dominant and recessive model results differ from the additive model result.

TABLE IV. Type I error (%) of the five selection strategies and the approach that uses all siblings for five SNP markers not associated with the quantitative trait

	α	AR	EV	IBD-EV	ML-EV	ML-IBD-EV	AS (GEE)
SNP1	0.10%	0.12	0.10	0.09	0.06	0.06	0.23
SNP2		0.13	0.07	0.11	0.10	0.10	0.12
SNP3		0.11	0.09	0.10	0.11	0.11	0.15
SNP4		0.11	0.09	0.10	0.11	0.11	0.13
SNP5		0.11	0.10	0.10	0.10	0.10	0.11
SNP1	1.00%	0.96	0.97	0.98	0.91	0.91	1.45
SNP2		1.03	1.00	0.94	0.99	0.99	1.21
SNP3		0.96	0.90	0.93	0.96	0.96	1.20
SNP4		0.97	0.96	1.02	1.05	1.06	1.16
SNP5		0.99	0.99	1.02	1.01	1.01	1.10
SNP1	5.00%	5.02	4.91	4.95	4.89	4.89	6.03
SNP2		5.13	5.00	4.94	4.95	4.95	5.39
SNP3		4.87	4.88	4.83	4.88	4.87	5.51
SNP4		4.89	5.09	5.04	5.03	5.03	5.35
SNP5		5.05	5.00	4.95	5.02	5.02	5.26

Minor Allele Frequencies (MAFs) for SNP1–SNP5 are 0.05, 0.1, 0.2, 0.35, and 0.5; respectively; AR, randomly selects a sibling/per sibship; EV, one sibling with the highest absolute phenotype value $|x_i|$ is selected; IBD-EV, one sibling with the highest score statistic $S(j)$ is selected; ML-EV, the EV strategy is applied using only linked sibships; ML-EBD-EV, the IBD-EV strategy is applied using only linked sibships; AS, all siblings are included using model generalized estimation equation (GEE).

Homogeneous samples. Figures 1 and 2 display power comparisons for homogeneous samples. Figure 1 displays power for the six selection strategies and the AS reference for additive models with $H = 30\%$, $r^2 = 0.1$, and $s = 3$ and a range of p. Figure 2 presents power for additive models with $H = 30\%$, $r^2 = 0.1$, and $p = 0.4$ for sibship sizes $s = 3$, 4, and 5. The power of the AR strategy remains virtually unchanged regardless of sibship size and QTL minor allele frequency. For additive models, all selection strategies except the AR strategy exhibit increased power when p increases. For recessive models, all strategies exhibit increased power as p increases (Supplementary Fig. 1). However, dominant models exhibit a different

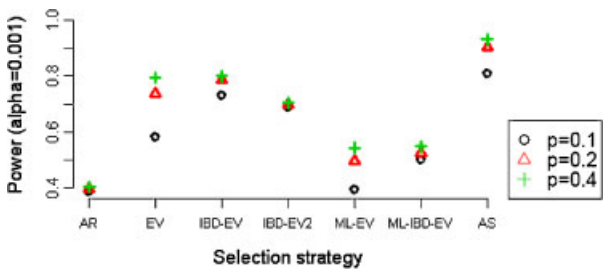

Fig. 1. Additive models with $H = 30\%$, $s = 3$, $r^2 = 0.1$ for $p = 0.1$, 0.2, and 0.4: Power comparisons for six strategies and the full sample in homogeneous samples. AR, randomly selects a sibling/per sibship; EV, one sibling with the highest absolute phenotype value $|x_i|$ is selected; IBD-EV, one sibling with the highest score statistic $S(j)$ is selected; ML-EV, the EV strategy is applied using linked sibships only; ML-EBD-EV, the IBD-EV strategy is applied using linked sibships only; AS, all siblings are included using Generalized Estimation Equation (GEE) to account for sibling correlation.

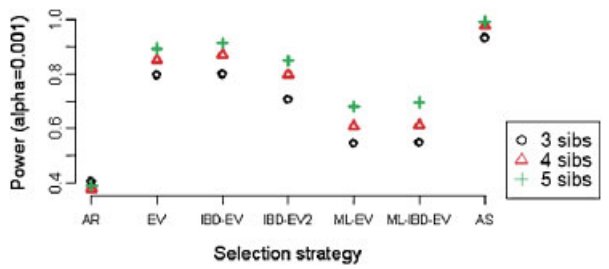

Fig. 2. Additive models with $H = 30\%$, $p = 0.4$, $r^2 = 0.1$: Power comparison for six strategies and the full sample with sibship sizes $s = 3$, 4, and 5 in homogeneous samples. AR, randomly selects a sibling/per sibship; EV, one sibling with the highest absolute phenotype value $|x_i|$ is selected; IBD-EV, one sibling with the highest score statistic $S(j)$ is selected; ML-EV, the EV strategy is applied using linked sibships only; ML-EBD-EV, the IBD-EV strategy is applied using linked sibships only; AS, all siblings are included using Generalized Estimation Equation (GEE) to account for sibling correlation.

trend in that all selection strategies have a decrease in power for larger QTL allele frequencies (Supplementary Fig. 2).

The IBD-EV strategy is the most powerful method of all selection strategies across all additive models considered. It exhibits 34% higher power than the AR strategy with $p = 0.1$, and about 40% higher power than the AR strategy for $p = 0.2$ and 0.4 (Fig. 1). Compared to using all individuals, the IBD-EV strategy shows 8, 12, and 13% loss of power with $p = 0.1$, 0.2, and 0.4, respectively (Fig. 1). In general, the IBD-EV2 strategy has about 4 to 10% lower power than the IBD-EV strategy for all genetic models we evaluated for homogeneous samples except for a recessive model with $p = 0.25$ and $s = 3$ (Supplementary Fig. 3). When $s = 3$, the IBD-EV2 strategy has 4, 9, and 10% lower power than the IBD-EV strategy for $p = 0.1$, 0.2, and 04, respectively (Fig. 1). For additive models with $s = 3$, the EV strategy has 14% lower power than the IBD-EV strategy with $p = 0.1$ and 5% lower power with $p = 0.2$; when $p = 0.4$, the two strategies have almost the same power (Fig. 1). However, the EV strategy is slightly more powerful than the IBD-EV strategy for both recessive and dominant models, and the power difference between these two strategies become smaller for larger p (Supplementary Figs. 1 and 2). When sibship size increases, the IBD-EV strategy tends to be more powerful than the EV strategy, and this is true for all additive, dominant, and recessive genetic models (Fig. 2 and Supplementary Figs. 3 and 4). In Fig. 2, the IBD-EV strategy exhibits 2% higher power than the EV strategy for both $s = 4$ and 5.

In general, the sample size for the ''linked'' strategies is about 60% of that for the AR/EV/IBD-EV strategies. However, the two ''linked'' strategies are more powerful than the AR strategy for all genetic models investigated. The ML-IBD-EV strategy gains about 11, 13, and 14% of power for $p = 0.1$, 0.2, and 0.4, respectively, over the AR strategy, despite about 40% of decrease in sample size (Fig. 1). There is a slight gain in power for the ML-EV strategy over the AR strategy when $p = 0.1$, the gains being 10 and 14% when $p = 0.2$, and 0.4, respectively (Fig. 1). Despite the power gains over the AR strategy, the power to detect the marker-phenotype association is 20–50% lower for the two ''linked'' strategies compared to the power using all siblings. This observation holds for all genetic models investigated, and is due to the great reduction in sample size by restricting the analysis to linked sibships.

We vary the effect size of the SNP marker by varying the LD between the LD SNP and the QTL. The results indicate that the power of all strategies increases in parallel with r^2 increases from 0.05 to 0.10, 0.15, and 0.20, for additive models with $H = 30\%$, $p = 0.4$, and $s = 3$. The corresponding model R^2 is 1.92, 3.24, 4.77, and 6.26%, respectively. The biggest power increases occurs when R^2 changes from 1.92 to 3.24% (or r^2 changes from 0.05 to 0.10). For R^2 of 5% or more, the power for all strategies, but the AR strategy, approach 1 at $\alpha = 0.001$.

Because the trait values for individuals selected from the EV, IBD-EV, ML-EV and ML-IBD-EV strategies are no longer normally distributed, we perform a permutation test on selected genetic models. We simulated a total 20,000 replicated and evaluate power for these four strategies at $\alpha = 0.001$. The power estimated from the permutation test for the strategies are similar to the power estimated from ANOVA models.

HETEROGENEOUS SAMPLES

Figure 3 displays power comparisons for the selection strategies for the four heterogeneity models with $s = 3$, $r^2 = 0.15$, and QTL allele frequency $p = 0.2$. We use this example to compare power in heterogeneous samples and point out differences with other genetic models. For the four heterogeneous samples we consider, the power increases as the homogeneity in the sample increases [i.e., Power (Het3) > power (Het2) > power (Het4) > power (Het1)]. The IBD-EV selection strategy remains most powerful among all the selection strategies for additive genetic models and is more powerful than the IBD-EV2 strategy in most instances (Fig. 3 and Supplementary Figs. 5–8). For the model presented in Figure 3, the IBD-EV strategy retains about 33, 70, 96, and 48% of power achieved using all individuals (AS reference) for Het1, Het2, Het3, and Het4 samples, respectively. The EV strategy is equivalent to or slightly better than the IBD-EV strategy for the additive model at $p = 0.4$, dominant models at $p = 0.1$, 0.2, and 0.4 and recessive models at $p = 0.4$ and 0.7 (Supplementary Figs. 5, 7, and 8). The linked strategies have lower power compared to EV and IBD-EV strategies. Compared to the AR strategy, the ML-IBD-EV strategy is 40, 31, 36, and 52% more powerful for Het1, Het2, Het3, and Het4 samples, respectively (Fig. 3).

Despite the power difference among the selection strategies, the power loss for the best selection strategy (IBD-EV in Fig. 3) is large compared to the use of all siblings (the AS reference) when the study sample is more heterogeneous: 67 and 52% power loss for Het1 and Het4 sample conditions compared to AS for the example in Figure 3.

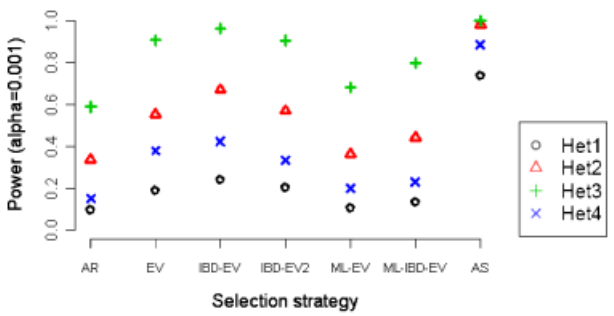

Fig. 3. Additive model with $p = 0.2$, $s = 3$, and $r^2 = 0.15$: Power comparison for six selection strategies and the full sample in heterogeneous samples. There are two subpopulations (with proportion q_1 and q_2, $q_1 + q_2 = 1$) in Het1, Het2, and Het3. The QTL effect with $H = 40\%$ only exists in subpopulation 1. We vary q_1 from 0.5 (this sample is denoted as Het1), 0.7 (Het2), and 0.9 (Het3). The Het4 sample consists of four equally sampled subpopulations with a QTL effect explaining 0, 10, 20, and 40% of the total variance in subpopulations 1–4, respectively. AR, randomly selects a sibling/per sibship; EV, one sibling with the highest absolute phenotype value $|x_i|$ is selected; IBD-EV/or IBD-EV2, one sibling with the highest score statistic $S(j)$, using the population mean/or sibship-specific mean, is selected; ML-EV, the EV strategy is applied using only linked sibships; ML-EBD-EV, the IBD-EV strategy is applied using only linked sibships; AS, all siblings are included using Generalized Estimation Equation (GEE) to account for sibling correlation.

APPLICATION TO THE FHS DATA

We perform variance component linkage analysis using 441 sibships and selected 467 SNPs on chromosome 10. The resulting LOD scores from the linkage analysis using the 441 sibships are generally lower than using the 330 extended pedigrees with the same set of SNPs (data not shown). However, the linkage peaks fall at the same region (70–80 cM on chromosome 10) for both analyses. To follow up on this linkage peak, 580 SNPs within the corresponding 1.0-LOD support interval are further tested for association with FPG; among them, 55 are included in the 467 SNP set that is used in linkage analysis.

Table V displays the 10 highest ranked SNPs from each of the IBD-EV, EV strategies and analysis of the full sample using GEE model. Table V(a) displays *P*-values ordered by IBD-EV *P*-values, (b) by EV *P*-values, and (c) by GEE *P*-values. Table V(b) contains 11 SNPs because the last two SNPs have the same *P*-values for the EV strategy. Two SNPs, rs876705 and rs780654, are in common for the two selection strategies and the full sample. The IBD-EV and EV strategies yield more consistent results and six SNPs are in common for these two selection strategies. Five SNPs identified from the analysis of the full sample are also in the EV strategy analysis. The IBD-EV and analysis of the full sample only have two SNPs in common (rs876705 and rs780654). In contrast, the AR strategy yields no *P*-values below 0.05. Seven SNPs from the IBD-EV strategy and six SNPs from the EV strategy fall within or very close to genes, while only three SNPs from analysis of the full sample are within genes. We include the gene names in the footnote of Table V. Among them, the hexokinase domain containing 1 (*HKDC1*, ∼5 kb away from rs5030938); and hexokinase 1 (*HK1*; ∼55 kb away

TABLE V. The 10 smallest *P*-values from EV, IBD-EV strategies and full sample (GEE) within 1.0 - LOD support interval for linkage peak on chromosome 10 with fasting plasma glucose at exam 5

SNP ID	IBD-EV	IBD-EV Permutation	EV	EV Permutation	AR	GEE	Known Gene
(a) Ordered by IBD-EV *P* value							
rs4074715	0.0005	0.0011	0.0002	0.0002	0.2215	0.0228	*CTNNA3* (*intron*)
rs10509276	0.0022	0.0026	0.0008	0.0013	0.3289	0.0859	*CTNNA3* (*intron*)
rs5030938	0.003	0.0023	0.0237	0.0236	0.3181	0.1451	*HKDC1*|*HK1* (*flanking*)
rs2043089	0.0036	0.0036	0.0047	0.0042	0.1622	0.0288	
rs4082516	0.0053	0.0061	0.0118	0.0119	0.3594	0.0955	*COL13A1* (*intron*)
rs876705	0.0067	0.0065	0.0059	0.0057	0.1405	0.0113	
rs10509378	0.0085	0.0094	0.0681	0.0664	0.3365	0.2838	*KCNMA1* (*intron*)
rs10509379	0.0112	0.0106	0.056	0.0526	0.3987	0.553	*KCNMA1 (intron)*
rs3851252	0.0113	0.0106	0.0057	0.0061	0.3053	0.1041	
rs780654	0.0126	0.0137	0.0037	0.0043	0.281	0.0065	*SLC29A3* (*intron*)
(b) Ordered by EV *P* value							
rs4074715	0.0005	0.0011	0.0002	0.0002	0.2215	0.0228	*CTNNA3* (*intron*)
rs10509276	0.0022	0.0026	0.0008	0.0013	0.3289	0.0859	*CTNNA3* (*intron*)
rs4074716	0.0191	0.0173	0.0015	0.0016	0.3236	0.1131	*CTNNA3* (*intron*)
rs4622198	0.014	0.0137	0.0016	0.0017	0.3138	0.1234	*KCNMA1* (*intron*)
rs1516510	0.019	0.0199	0.0026	0.0026	0.3378	0.158	*KCNMA1* (*intron*)
rs780654	0.0126	0.0137	0.0037	0.0043	0.281	0.0065	*SLC29A3* (*intron*)
rs2043089	0.0036	0.0036	0.0047	0.0042	0.1622	0.0288	
rs3851252	0.0113	0.0106	0.0057	0.0061	0.3053	0.1041	
rs876705	0.0067	0.0065	0.0059	0.0057	0.1405	0.0113	
rs2140391	0.0589	0.0565	0.0099	0.008	0.2739	0.0093	
rs1880065	0.0589	0.0565	0.01	0.008	0.2744	0.0097	
(c) Ordered by GEE *P* value							
rs780654	0.0126	0.0137	0.0037	0.0043	0.281	0.0065	*SLC29A3* (*intron*)
rs2140391	0.0589	0.0565	0.0099	0.008	0.2739	0.0093	
rs1354038	0.068	0.0656	0.1175	0.1183	0.2639	0.0096	
rs1880065	0.0589	0.0565	0.01	0.008	0.2744	0.0097	
rs2140390	0.0589	0.0565	0.01	0.0082	0.2746	0.0097	
rs7895188	0.0499	0.0477	0.0368	0.0349	0.2132	0.0105	*DLG5* (*intron*)
rs876705	0.0067	0.0065	0.0059	0.0057	0.1405	0.0113	
rs1500737	0.0867	0.0803	0.1438	0.1426	0.2753	0.0119	
rs4399260	0.0274	0.0256	0.0367	0.0396	0.1755	0.012	
rs2120989	0.1735	0.1706	0.3323	0.3287	0.4128	0.012	*CCDC6* (*intron*)

Genotype annotation sources are described in Cupples et al. [2007]; CTNNA3, α-T-catenin; *COL13A1* type XIII collagen α 1 (*COL13A1*); *KCNMA1*, potassium large conductance calcium-activated channel, subfamily M, α member 1; HKDC1, hexokinase domain containing 1; *HK*1, hexokinase 1; *SLC29A3*, solute carrier family 29 (nucleoside transporters), member 3; DLG5, DLG5 discs, large homology 5; CCDC6, coiled-coil domain containing 6; AR, all random or randomly select a sibling/per sibship; EV, one sibling with the highest absolute phenotype value $|x_i|$ is selected; IBD-EV, one sibling with the highest score statistic $S(j)$ is selected.

from rs5030938) are involved in glucose metabolism. In addition, DLG5 discs, large homology 5 (DLG5) (contains intron SNP rs7895188) is associated with inflammatory bowel diseases and Crohn's disease.

DISCUSSION

In this article, we extend the selection of the most informative individuals from families with multiple siblings from a binary trait [Fingerlin et al., 2004] to a quantitative trait to propose the use of the most extreme trait value and a score statistic, $S(j)$ that incorporates an individual's trait value and his/her IBD sharing information with other siblings in the selection of the most informative individuals. The $S(j)$ score is an extension of the QLS statistic [Wang and Elston, 2006a] with modifications and is used in a different manner in the selection process. The $S(j)$ score contains a modified QLS statistic, replacing the sibship-specific mean with the sample mean. In addition, the $S(j)$ score for an individual is a summation of the sibpair scores over all pairs that includes sibling j, while the QLS statistic is a summation of the scores for all the pairs in a sibship. As a result, the QLS statistic is sibship-specific and selects linked sibships from a heterogeneous sample, and $S(j)$ is sibling-specific and selects the most informative individuals from sibships. Despite the difference, the individuals selected based on these two score statistics tend to carry genetic variants influencing the quantitative trait and be more homogeneous and thus have greater power to detect an association compared to the random selection strategy.

We evaluated different selection strategies in both homogeneous and heterogeneous study samples. When the sample was or was close to homogeneous, we found that the IBD-EV strategy saves most in genotyping cost when selecting one ''best'' sibling from sibships with multiple siblings for additive genetic models included in our simulations. The IBD-EV strategy only lost 8–13% of power compared to the full sample that uses all siblings, across all additive models considered, but offered at least 50% genotyping cost saving (Figs. 1 and 3). Compared to the alternative IBD-EV2 strategy, which incorporates the sibship-specific mean as the offset in the $S(j)$, the IBD-EV strategy, using the sample mean as the offset in $S(j)$, retains more power, for most genetic models in both homogeneous and heterogeneous conditions and different sibship sizes. The GEE model that was used to accommodate correlated family structure posed somewhat increased type I error (Table IV), as observed by Cupples et al. [2007]. If we take this fact into consideration, the power loss of the IBD-EV strategy is even less.

One limitation of this score statistic is that the IBD sharing information does not accurately reflect the expected trait differences when trait values follow either dominant or recessive models. We illustrate this point with a dominant model, where the mean phenotype of individuals with the AB and BB genotypes are the same but different from the mean phenotype of individuals with the AA genotypes. For a nuclear family with three siblings having genotypes of AA, AB, and BB, the trait difference for sib-pair 1–2 will be bigger, on average, than that trait difference for sib-pair 2–3 for a dominant model despite the IBD sharing probability being the same for sib-pair 1–2 and sib-pair 2–3. A similar situation occurs for recessive models. For additive models, the IBD-based statistic reflects the expected difference in phenotype more accurately. Thus, the IBD-EV strategy is the most powerful strategy among all selection strategies in the context of additive models.

Across all additive models, the EV strategy loses little power compared to the IBD-EV strategy except for additive models with $p = 0.1$. For dominant and recessive models, the EV strategy is, in general, slightly more powerful than or equivalent to the IBD-EV strategy. The EV strategy does not require IBD sharing information and thus can be performed for sibling selection without prior linkage scans, and result in a sample that can be used for testing association across the whole genome, not just under a particular linkage peak.

We performed permutation tests to validate our findings for the EV, IBD-EV, ML-EV, and ML-IBD-EV strategies with a range of selected genetic models, because the trait values based on these strategies are no longer normally distributed. Permutation test results agree with the results from the ANOVA F-test in this study, mainly because the trait values based on the EV and IBD-EV strategies are symmetric. In practice, the phenotype studied may not be normally distributed because of either ascertainment, outlier observations, or a skewed distribution. For non-normal data, trait values based on the EV and IBD-EV strategies may not be symmetric, and the trait values based on the AR strategy may be skewed in an ascertained sample unlike the traits simulated in our study. Even though the ANOVA method is robust to non-normal data, we recommend performing the permutation test to obtain empirical P-values for association studies for all strategies. The GEE model is also susceptible to non-normal data and might has larger elevated type I error [Cupples et al., 2007], however; the permutation test cannot be directly used in the GEE model because of correlation structure in the full sample.

We applied three selection strategies to an FHS continuous trait—FPG at exam 5. Only two SNPs are in common in the EV and IBD-EV strategies and the analysis of the full sample. There may be two reasons for the inconsistent results. First, there is a trade-off when we try to select the most informative individuals from sibships. As stated previously, 441 siblings only represent about 30% of sample formed by all available siblings. On the other hand, the selected 441 siblings may represent a more homogeneous population than the full sibships. Therefore, the EV and IBD-EV strategy may detect signals that would go undetected by analyzing the full sample. Despite the inconsistency among all three approaches, six SNPs from the EV strategies are in common with the IBD-EV strategies and five SNPs yielded from the analysis of the full sample are also in the EV strategy analysis, indicating that the EV strategy might be the best approach in practice, given the results in this application and those from simulations.

In association studies using family data, if trait values and linkage information are available, selection of the most informative individuals from each family using the score statistic or only the trait value is a worthwhile effort. This selection offers at least 50% genotyping cost saving, sacrificing little power loss (at most 13% in this simulation) if a study population is or close to homogenous. If the study population is heterogeneous, selection might represent a more homogeneous population.

ACKNOWLEDGMENTS

This research was conducted using the Boston University Linux Cluster for Genetic Analysis (LinGA) funded by the NIH NCRR (National Center for Research Resources) Shared Instrumentation grant (1S10RR163736-01A1). Part of this work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195), and by a Career Development Award from the American Diabetes Association (to J.B.M.).

REFERENCE

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet 65:531–544.

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543.

Carey G, Williamson J. 1991. Linkage analysis of quantitative traits: increased power by using selected samples. Am J Hum Genet 49:786–796.

Chen Z, Zheng G, Ghosh K, Li Z. 2005. Linkage disequilibrium mapping of quantitative-trait Loci by selective genotyping. Am J Hum Genet 77:661–669.

Cupples LA, Arruda HT, Benjamin EJ, D'Agostino Sr RB, Demissie S, Destefano AL, Dupuis J, Falls KM, Fox CS, Gottlieb DJ et al. 2007. The Framingham Heart Study 100 K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8:S1.

Eaves L, Meyer J. 1994. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. Behav Genet 24:443–455.

Elston RC, Buxbaum S, Jacobs KB, Olson JM. 2000. Haseman and Elston revisited. Genet Epidemiol 19:1–17.

Fingerlin TE, Boehnke M, Abecasis GR. 2004. Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. Am J Hum Genet 74:432–443.

Fisher RA, editor. 1935. The Design of Experiment. New York: Hafner.

Good PI, editor. 2005. Permutation, Parametric and Bootstrap of Hypotheses, 3rd edition. Springer.

Haldane JBS. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet 8:299–309.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19.

Palmer LJ, Cardon LR. 2005. Shaking the tree: mapping complex disease genes with linkage disequilibrium. Lancet 366: 1223–1234.

Risch N, Zhang H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584–1589.

Slatkin M. 1999. Disequilibrium mapping of a quantitative-trait locus in an expanding population. Am J Hum Genet 64:1764–1772.

Wang T, Elston RC. 2006a. A quantitative linkage score for an association study following a linkage analysis. BMC Genet 7:5.

Wang T, Elston RC. 2006b. Sample selection to perform association study for quantitative-trait loci. BMC Genet Nov. 16–17; St. Petersburg Beach. FL.