Ordered Subset Analysis in Genetic Linkage Mapping of Complex Traits

Elizabeth R. Hauser,[1*] Richard M. Watanabe,[2] William L. Duren,[3] Meredyth P. Bass[1] Carl D. Langefeld[4] and Michael Boehnke[3]

[1]*Section of Medical Genetics, Department of Medicine, Center for Human Genetics, Duke University Medical Center, Durham, North Carolina*
[2]*Division of Biostatistics, Department of Preventive Medicine, USC Keck School of Medicine, Los Angeles, California*
[3]*Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan*
[4]*Section on Biostatistics, Department of Public Health Sciences, Wake Forest University of Medicine, Winston-Salem, North Carolina*

Etiologic heterogeneity is a fundamental feature of complex disease etiology; genetic linkage analysis methods to map genes for complex traits that acknowledge the presence of genetic heterogeneity are likely to have greater power to identify subtle changes in complex biologic systems. We investigate the use of trait-related covariates to examine evidence for linkage in the presence of heterogeneity. Ordered-subset analysis (OSA) identifies subsets of families defined by the level of a trait-related covariate that provide maximal evidence for linkage, without requiring a priori specification of the subset. We propose that examining evidence for linkage in the subset directly may result in a more etiologically homogeneous sample. In turn, the reduced impact of heterogeneity will result in increased overall evidence for linkage to a specific region and a more distinct lod score peak. In addition, identification of a subset defined by a specific trait-related covariate showing increased evidence for linkage may help refine the list of candidate genes in a given region and suggest a useful sample in which to begin searching for trait-associated polymorphisms. This method provides a means to begin to bridge the gap between initial identification of linkage and identification of the disease predisposing variant(s) within a region when mapping genes for complex diseases. We illustrate this method by analyzing data on breast cancer age of onset and chromosome 17q [Hall et al., 1990, Science 250:1684–1689]. We evaluate OSA using simulation studies under a variety of genetic models. © 2004 Wiley-Liss, Inc.

Key words: linkage analysis; genetic heterogeneity; covariate; complex traits

Grant sponsor: NIH; Grant numbers: MH59528, DK09525, HG00040, HG00376; Grant sponsor: ADA.

*Correspondence to: Elizabeth R. Hauser, Ph.D., Section of Medical Genetics, Department of Medicine, Center for Human Genetics, Duke University Medical Center, Box 3445, Durham, NC 27710.
E-mail: Elizabeth.Hauser@duke.edu

Received for publication 27 August 2002; revision accepted 8 September 2003
Published online 8 April 2004 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/gepi.20000

INTRODUCTION

Complex diseases, such as type 2 diabetes, cancer, hypertension, and cardiovascular disease, are characterized by a multifactorial etiology. Genes and environment likely interact in a complex fashion to cause disease. Presumably, development of a complex disease is a result of perturbations in biological pathways. Genetic polymorphisms may affect various components of these pathways resulting in differential effects on quantitative measurements related to the complex disease. Detecting the effects of genes and their genetic polymorphisms on a complex trait may require taking intermediate traits or covariates into account so that the susceptibility gene is detectable in a subset of the families. This model suggests that genetic heterogeneity will be a fundamental feature of complex disease etiology. Furthermore, genetic linkage analysis methods to map genes for complex traits should acknowledge the presence of genetic heterogeneity to maximize power to identify subtle changes in complex biologic systems.

The large genome screens for the genetic analysis of complex traits provide tremendous amounts of information. These screens typically produce lod scores or estimated IBD sharing for hundreds of families typed for hundreds of genetic markers across the genome. This vast amount of information has not often translated directly into impressive lod scores, presumably because of etiologic complexity. The presence of multiple independent and/or interacting disease genes and environmental factors causes significant problems for genetic linkage analysis. Etiologic heterogeneity reduces power to detect linkage, but explicitly taking heterogeneity into account can help minimize this information loss [Arnett et al., 1997; Lunetta and Rogus, 1998; Greenwood and Bull, 1999; Leal and Ott, 2000]. Even when it is possible to detect linkage to a region, etiologic heterogeneity can bias the estimate of gene location and result in broader confidence intervals [Ott, 1983; Falk, 1997; Chiano and Yates, 1994].

Several linkage analysis methods have been proposed to recognize and correct for heterogeneity. The most popular is the admixture test and its extensions, which allow for two or more disease-causing loci [Ott, 1983; Bhat et al., 1999]. This test uses evidence for or against linkage at a marker to suggest the presence of heterogeneity and may not be particularly sensitive when the sample is composed of small families, such as those usually used for late-onset diseases [Falk, 1997].

Another method to assess genetic heterogeneity is the pre-divided sample test [Morton, 1956]. When *a priori* evidence for genetic heterogeneity exists, subsets of families may be defined prior to analysis and examined for differences in the evidence for linkage. This idea was applied to a genetic linkage study of 26 Finnish families from the Bothnia region in which no evidence for linkage was obtained in the overall sample but suggestive evidence for linkage was obtained in families in the lowest quartile of 30-minute insulin [Mahtani et al., 1996]. While it may be straightforward to select trait-related covariates based on biological or clinical considerations, the value of the covariate to choose as a cutoff in defining a genetically homogeneous subset often is not obvious.

One way to avoid having to define arbitrary *a priori* cutoffs is to use a function of a covariate to rank the families and then to add in the families one or a few at a time based on the covariate. This idea was exploited in the initial report of the linkage of breast and ovarian cancer to the gene subsequently identified as BRCA1 [Hall et al., 1990; Merette et al., 1992; Miki et al., 1994; Futreal et al., 1994]. Hall et al. [1990] ranked the families by mean age of onset of breast cancer and plotted the cumulative sum of the lod scores. The results indicate that the evidence for linkage was concentrated in the families with early age of onset. Recognition of this key feature was important in subsequent studies to narrow the region of linkage and to identify BRCA1. Had age of onset been ignored, linkage to chromosome 17 might have been excluded as the overall lod score was substantially negative due to the heterogeneous mixture of families in the sample. We expand upon the idea of using a trait-related covariate to examine evidence for linkage in the presence of heterogeneity. Our ordered subset method identifies subsets of families defined by level of a trait-related covariate that provide maximal evidence for linkage, without requiring *a priori* specification of the subset. The significance of the subset and its evidence for linkage is evaluated using a permutation procedure to estimate empirical *P* values.

We propose that examining evidence for linkage in the subset directly can result in a more etiologically homogeneous sample. In turn, the reduced impact of heterogeneity may result in increased overall evidence for linkage to a specific region and the potential to produce a more distinct lod score peak with more precise gene localization. The identified ordered subset may be a useful sample in which to begin searching for trait-associated polymorphisms or mutations through sequencing or other fine-mapping techniques. Identification of a trait-related covariate using this ordered subset analysis (OSA) also may help refine the list of candidate genes in a given region based on biological pathways, which connect the covariate and the trait. In this report, we apply OSA to linkage results of breast cancer to D17S74 as published by Hall et al. [1990]. We present the results of a simulation study evaluating type I error rates and power for a variety of genetic models consistent with a complex genetic trait. We find that OSA preserves the appropriate type I error rates. Furthermore, compared to analyses that do not use the covariate information but do allow for heterogeneity among families, OSA has substantially greater power to detect ordered subsets of families defined by a covariate.

METHODS

Ordered subset analysis (OSA) may be used at any time during a linkage analysis. The requirements are additive linkage statistics and covariates for each family. Typically, use of this method will begin when the genome screen linkage analysis has been completed. We assume in what follows that the results of the linkage analysis are lod scores for all families, although other additive test statistics could also be used (see Discussion).

The input for OSA is the lod scores $Z_i(d,\gamma)$ for family i, at position d where d may range across a region, chromosome, or the genome, and γ represents the genetic model. We will refer to the maximum of the sum of the lod scores over all families as the overall baseline lod score, $Z(d,\gamma)$. The genetic model may take a variety of forms from completely unspecified, to a vector of IBD sharing probabilities or recurrence risk ratios (λ) or a full Mendelian model with a genetic model parameter set including penetrances and allele frequencies. Depending on the analysis method used for generating the family-specific lod scores, the genetic model may include reduced penetrance, liability classes, or phenocopies.

The covariates may be one or more continuous or ordinal variables thought to be related to the trait of interest and are often a function of trait values for members in the family. Examples of covariates include mean, median, or minimum covariate values for affected individuals in the family. We use the term "covariate" very generally to include disease-related quantitative traits as well as IBD sharing or linkage statistics at another locus.

FINDING THE MAXIMUM ORDERED SUBSET STATISTIC

We begin by ordering the families based on some covariate value. Let $Z_{(j)}(d, \gamma)$ be the lod score matrix for ordered family j. For example, if we are ordering families for low to high mean age of onset, we would start with the family with the lowest mean age of onset, with lod score matrix $Z_{(1)}(d, \gamma)$. We find the maximum lod score in that family and the estimates of $\widehat{d}^1$ and $\widehat{\gamma}^1$ at which the maximum occurs. Next we use element-wise addition to add the matrix for the family with the next lowest mean age of onset, $Z_{(2)}(d, \gamma)$, to the matrix for the first family. In general, we create the jth partial sum by adding each element of the lod score matrix for each family up to and including ordered family j:

$$Z^j(d,\gamma) = \sum_{k=1}^{j} Z_{(k)}(d,\gamma)$$

After the addition of each family j, we note the maximum summed lod score, $Z^j(\widehat{d}^j,\widehat{\gamma}^j)$ and estimates of the disease location $(\widehat{d}^j)$ and disease model parameters $(\widehat{\gamma}^j)$. After all of the N families have been added in turn, we have maxima for each partial sum of the lod scores $(Z^1(\widehat{d}^1,\widehat{\gamma}^1),\ldots,(Z^N(\widehat{d}^N,\widehat{\gamma}^N))$, ordered by the family-specific covariate value. Note that the values of d and γ can and often are maximized at different values depending on the subset of families. We define the maximum ordered subset lod score

$$Z^*(\widehat{d}^*,\widehat{\gamma}^*) = \max_{\mathrm{j}}\left[Z^j(\widehat{d}^j,\widehat{\gamma}^j)\right]$$

and the Δ lod score

$$\Delta = Z^*(\widehat{d}^*,\widehat{\gamma}^*) - Z(\widehat{d}^*,\widehat{\gamma}^*)$$

to be the difference between the maximum ordered subset lod score and the lod score summed over all families at the same position and model. Since a genetically more homogeneous subset may be identified either through high values or low values of a covariate, we generally perform the summation in both ascending and descending order unless there is strong prior evidence to consider only one order or the other.

EVALUATING LINKAGE EVIDENCE IN SUBSETS OF FAMILIES

Given the maximization over subsets procedure we have employed, the maximum ordered subset lod score will always be at least as large as the overall lod score. Therefore, the distribution of the maximum ordered subset lod score cannot be the same as the distribution of the lod score without this maximization. The observation of a large lod score in a subset of the families must be evaluated in the context of the evidence for linkage in the entire sample and the fact that we have selected a subset of the data.

To assess significance of linkage evidence for a subset conditional on linkage evidence in the entire sample, we employ a permutation strategy to test the null hypothesis that there is no relationship between the trait-related covariate and the evidence for linkage. In the permutation test procedure, we assess the significance of the increase in the lod score in the identified ordered subset compared to the baseline lod score in all families. We sample from the distribution of all N! permutations by randomly ordering the N families and calculating the maximum ordered subset lod score for each permutation p, $Z_p^*(d_p^*,\gamma_\pi^*)$, in exactly the same way as for the original data. We estimate the P value for the ordered subset lod score as the proportion of permutations giving maximum ordered subset lod scores as large or larger than that observed. The permutation test can be used to estimate *P* values when maximizing the ordered subset lod score at a single point, for a chromosomal region or whole chromosome, or for the entire genome.

An assumption of permutation tests is that the observations to be permuted are exchangeable. When all families are of the same size and structure, this assumption is met. When families are of different sizes or structures, the *P* values are approximate.

COMPUTATIONAL EFFICIENCY

While this analysis and particularly the requirements of the permutation test may appear computationally burdensome, this burden can be minimized. If the number of family-position-model-specific lod scores $Z_i(d,\gamma)$ is not too large, these lod scores can be calculated once and stored. OSA then may be performed for many different functions of covariates and for as many permutations as desired by simply reordering the families and carrying out the summation and maximization steps. We routinely compute permutation tests for analyses with over 500 families, 400 positions, and 25 disease models in this way.

To devote more computation time to assessing significance of the most interesting results, we have implemented an inverse-sampling method for empirical *P* value estimation. We set the maximum number of permutations desired as a parameter for the empirical *P* value calculation. We continue sampling until the estimated variance of the empirical *P* value $\widehat{p}$ is sufficiently small so that $\frac{1}{2}\widehat{p}$ is less than the 95% lower bound for $\widehat{p}$ or until the maximum number of permutations is reached. This procedure uses the Poisson approximation to the Binomial for small *P*. The variance of $\widehat{p}$ is evaluated after every 10th permutation replicate. If $\widehat{p}$ exceeds this ''boredom threshold,'' the permutation routine stops before the maximum number of permutations is reached. We also follow the recommendation of North et al. [2002] that the number of permutations giving a larger statistic than the observed statistic be >10. Thus, we can set a high number of maximum permutations but only reach that maximum when trying to estimate small *P* values, increasing computational efficiency.

APPLICATION TO BREAST CANCER DATA

We applied OSA to the data published by Hall et al. [1990] in 23 families ascertained for multiple cases of breast cancer. We used the data presented in their table I that includes, for each family, the mean age of onset of breast cancer in the affected family members and lod scores at five positions relative to marker D17S74. We used their figures 1 and 2 to define minimum age of onset, maximum age of onset, and range of age of onset for each family. Hall et al. [1990] considered only one genetic model, so that we stored only five lod scores (one for each position) for each family.

SIMULATION STUDIES

We performed a simulation study to evaluate the false-positive rate of OSA in affected sib pair (ASP) linkage analysis using the recurrence risk ratio to sibs λ to describe the genetic effect [Risch 1990a–c; Hauser and Boehnke 1998]. As seen in power studies of ASP linkage analysis, λ is a critical factor in determining the power of the linkage analysis [Hauser et al., 1996]. We used the simulation package SIMLA that we developed for simulation of complex genetic traits incorporating both linkage and association [Bass et al., 2004]. We simulated ASPs with genotyped parents under disease models with locus-specific λ values of 1.2 and 1.4. We first simulated data sets to represent two null hypotheses: (1) no linkage in any of the ASPs, λ=1, and 2) linkage ($1.5<\lambda<5.0$) in 20–50% of the ASPs but no difference in the covariate distribution between the linked and unlinked subsets (OSA null hypothesis of no covariate effect). We simulated a covariate value for each individual from a normal distribution. To evaluate the power of OSA to identify a linked subset, we simulated a subset in which the covariate values were chosen from a mixture of normal distributions with means different in the linked and unlinked ASPs. Table I shows the parametric models used to perform the simulations to examine power. These models were chosen to yield population prevalences of $\sim$5% as well as to cover a range of models for complex diseases. Two values of λ are given for each model, the overall λ for the combined sets of families and the λ in the linked subset. These values allow comparisons between different genetic models resulting in the same λ values. We simulated genotype data for a 90-cM map of 10 markers evenly spaced at 10-cM intervals. In simulations with $\lambda > 1$, we placed the disease locus at 45 cM. We simulated 5,000 replicate data sets of 400 ASPs for each condition to examine type I error rates under the two null hypotheses and 500 replicate data sets for the power studies.

We performed two linkage analyses of the simulated data sets to generate family-specific lod scores for input into OSA. First, we carried out multipoint ASP based on the IBD sharing method of Risch [1990a–c] as implemented in Siblink [Hauser and Boehnke, 1998]. This method provides OSA input files of lod scores calculated at a grid of map positions and λ values assuming an additive genetic model. Second we performed two-point parametric linkage analysis with Vitesse [O'Connell and Weeks, 1995] using the parametric model from the linked subset to calculate the family-specific lod scores at a marker 5 cM from the true disease location. We used these family-specific lod-scores as input to the HOMOG package [Ott, 1983, 1999], which implements the admixture test. HOMOG performs nested likelihood ratio tests for linkage with or without genetic heterogeneity and provides a χ^2 test statistic for the hypotheses of linkage with homogeneity among the families ($\theta<0.5$, $\alpha=1$) vs. linkage with heterogeneity ($\theta<0.5$, $\alpha<1$). To assess whether the incorporation of a trait-related covariate would provide information related to heterogeneity among the families above and beyond differences in the maximum lod scores, we compared the power of OSA with the power of the test for heterogeneity provided by the admixture test.

RESULTS

BREAST CANCER EXAMPLE

Table II shows the results of OSA applied to the linkage results for breast cancer and chromosome 17q [Hall et al., 1990] that localized BRCA1 and ultimately led to the identification of BRCA1 [Miki et al., 1994]. OSA replicates the results presented in the paper with a maximum subset lod score of 5.98 and Δ lod of 11.46 occurring in the 7 families with youngest mean age of onset. Hall et al. [1990] did not estimate a *P* value for their result. The approximate empirical *P* value for this subset from OSA is 0.0009, suggesting that the result would be highly unlikely if there were no relationship between age of onset and evidence for linkage. Lowest maximum age of onset and range of age of onset in the family also gave significant improvements in linkage evidence with lod scores of 5.10 and 4.66 (Δ lods of 10.58 and 11.14, *P* values of .02 and .01, respectively).

TABLE I. Genetic models used for the simulation studies[a]

						Penetrances		
Model number	λ	λ in linked subset	Genetic model	Proportion of families in the linked subset	Disease allele frequency P(A)	P(D∣AA)	P(D∣Aa)	P(D∣aa)
1	1.2	5.0	Dominant	0.20	0.02	0.70	0.70	0.02
2	1.2	4.0	Recessive	0.22	0.18	0.80	0.02	0.02
3	1.2	2.5	Dominant	0.28	0.06	0.32	0.32	0.02
4	1.2	2.5	Recessive	0.28	0.28	0.40	0.02	0.02
5	1.2	2.5	Additive	0.28	0.06	0.58	0.29	0.02
6	1.2	1.5	Additive	0.50	0.02	0.58	0.29	0.04
7	1.4	2.3	Additive	0.50	0.06	0.50	0.25	0.02

[a] λ is the locus-specific recurrence risk ratio in the entire sample. Penetrances are expressed as the probability of disease (D) given the genotype. The disease prevalence was constrained to be ~0.05.

TABLE II. Results of ordered subset analysis applied to breast cancer linkage data for 23 families and marker D17S74 [Hall et al. 1990] for four age-related covariates

		Subste				
Age of onset covariate	Order	$\hat{\theta}$	Maximum lod	Δ lod at $\hat{\theta}$	Empirical *P*-value	Families in the subset
Mean	Ascending	0.001	5.98	11.46	0.0009	1–7
	Descending	0.20	2.35	0.00	0.89	1–23
Minimum	Ascending	0.10	3.03	1.49	0.14	1–5,7–10,13,15,16
	Descending	0.20	2.35	0.00	0.82	1–23
Maximum	Ascending	0.001	5.10	10.58	0.02	1–4,7,8
	Descending	0.20	2.35	0.00	0.66	1–23
Range	Ascending	0.001	4.66	11.14	0.01	1,3–7
	Descending	0.20	2.35	0.00	0.91	1–23

TABLE III. Type I error rates: simulation results for nonparametric multipoint linkage analysis with 400 ASPs using ordered subset analysis under no linkage (λ=l), and no covariate effect with 20–50% of the families linked (λ>1)[a]

Model number	λ/Linked λ model	Mean overall lod score	Mean maximum subset lod score	Mean Δ lod	Mean proportion of families in subset	$P(\widehat{p}<.05)$	$P(\widehat{p}<.01)$
—	1.0	0.28	1.34	1.06	0.25	0.052	0.014
1	1.2/5.0 D	2.99	4.78	1.79	0.58	0.057	0.010
2	1.2/4.0 R	4.97	6.37	1.40	0.73	0.050	0.009
3	1.2/2.5 D	2.79	4.57	1.77	0.57	0.047	0.009
4	1.2/2.5R	4.19	5.72	1.53	0.68	0.049	0.010
5	1.2/2.5 A	2.69	4.47	1.78	0.55	0.045	0.008
6	1.2/1.5 A	2.58	4.38	1.80	0.54	0.050	0.008
7	1.4/2.3 A	6.19	7.46	1.27	0.78	0.043	0.010

[a] For the simulations with λ>1, the trait-related covariate was distributed as a normal random variable. Models are described in Table I; 5,000 replicates were simulated per model.

TABLE IV. Power: simulation results for nonparametric multipoint linkage analysis with 400 ASPs using ordered subset analysis with a proportion of the families linked (λ>1)[a]

Model number	λ/linked λ model	Mean overall lod score	Mean maximum subset lod score	Mean Δ lod	Mean proportion of families in subset	$P(\widehat{p}<.05)$	$P(\widehat{p}<.01)$
1	1.2/5.0 D	2.65	13.16	10.51	0.22	0.964	0.910
2	1.2/4.0R	4.50	19.08	14.57	0.22	0.998	0.988
3	1.2/2.5 D	2.79	9.84	7.06	0.31	0.804	0.638
4	1.2/2.5 R	3.93	13.70	9.77	0.29	0.942	0.876
5	1.2/2.5 A	2.45	9.38	6.93	0.29	0.784	0.630
6	1.2/1.5 A	2.59	6.34	3.74	0.46	0.334	0.156
7	1.4/2.3 A	6.10	13.46	7.36	0.51	0.894	0.754

[a] The trait-related covariate was distributed as a mixture of normals with means two standard deviations apart in the linked and unlinked subsets. Models are described in Table I, 500 replicates were simulated per model.

The families included in these subsets are very similar to those included for the mean age subset, reflecting the high correlation in these family-specific functions of age of onset. The minimum age of onset gave a nonsignificant result with a maximum subset lod score of 3.03, Δ lod of 1.49, and *P* value=0.14. No subsets for any of the age-related covariates showed an increase in the lod score when the 23 families were ordered from oldest to youngest or largest to smallest maximum, minimum, or range so that the results for these analyses are the same as the results for the complete set of families.

SIMULATION STUDY

Table III shows the results of a simulation study using multipoint affected sib pair linkage analysis to assess the overall behavior of OSA under no linkage (λ=1) and linkage (λ>1) but no covariate effect, i.e., the possible null hypotheses for OSA. As expected, the mean maximum subset lod score is larger than the overall lod score in every case. The Δ lod measures the difference between the maximum subset lod score and the overall baseline lod score at that map position and for that model. For the no linkage and no covariate effect simulations we performed, the mean Δ lod scores were >1 lod unit. The type I error rates estimated as the proportion of replicates with empirical *P* values <0.05 and <0.01 are close to the nominal levels for both the no linkage and no covariate effect cases. This observation suggests that, as expected, the permutation approach adequately adjusts for the increase in the lod score when maximizing over subsets.

Table IV presents the power of OSA using the multipoint nonparametric affected sibling pair lod scores generated by Siblink for the genetic models listed in Table I. In these models, 20–50% of the families are linked; the covariate distribution is a mixture of normals with means two standard deviations apart in linked and unlinked families. The power of OSA is estimated by the proportion of replicates with empirical *P* values $\widehat{p}$ less than 0.05 or 0.01. Not surprisingly, the power increases as the λ in the linked families increases. We applied a variety of different genetic models, holding the λ values constant. As discussed by Olson [1995], the underlying genetic model can

TABLE V. Simulation results for two-point parametric linkage analysis with 400 ASPs using ordered subset analysis and the Admixture Test implemented in HOMOG with a subset of the families linked ($\lambda>1$)[a]

Model number	λ/linked λ model	Mean max HLOD	Mean chi-square for heterogeneity	$P(\widehat{p}<.05)$	Mean overall lod score	Mean maximum subset lod score	$P(\widehat{p}<.05)$
1	1.2/5.0 D	5.92	3.68	0.50	−13.63	14.21	1.00
2	1.2/4.0 R	2.78	3.39	0.52	−32.08	10.89	0.98
3	1.2/2.5 D	2.37	1.42	0.20	−6.08	6.13	0.94
4	1.2/2.5 R	2.00	1.57	0.18	−10.98	6.35	0.94
5	1.2/2.5 A	2.04	1.40	0.20	−6.59	5.55	0.89
6	1.2/1.5 A	2.12	0.89	0.08	−0.05	4.12	0.72
7	1.4/2.3 A	4.28	1.58	0.24	−0.28	8.18	0.96

[a] The trait-related covariate was distributed as a mixture of normals with means two standard deviations apart in the linked and unlinked subsets. The generating model was used for the linkage analysis. Models are described in Table I, 500 replicates were simulated per model.

have a significant impact on the power to detect linkage using nonparametric affected sibling pair analysis. The OSA results reflect this difference as seen by comparing the maximum lod scores, maximum subset lod scores, and the OSA power for genetic models with overall $\lambda=1.2$ or with linked subset $\lambda=2.5$. The recessive models provide more evidence for linkage across all measures. However, even in the dominant and additive models, when the subset is large (proportion of linked families 50%) or when the linked subset λ is large ($\lambda\geq2.5$), OSA had greater than 78% power at significance level 0.05 to detect the increased evidence for linkage in the subset using the trait-related covariate.

Table V compares the power of OSA to the power of the admixture test as implemented in HOMOG for detecting heterogeneity among the families. The parametric model from the linked subset was used as the analysis model to provide the most information possible for detecting linkage in these small families. The power of the test of heterogeneity in the admixture model as implemented in HOMOG is considerably lower for all genetic models tested for these small families. OSA, using the same family-specific parametric two-point lod scores, provides substantially greater power than that observed with the admixture test to detect heterogeneity using a trait-related covariate. As in the multipoint nonparametric case, the power varies substantially across models.

DISCUSSION

We propose ordered subset analysis (OSA) as a means to address the etiologic and genetic heterogeneity in the analysis of complex genetic traits. The goal of OSA is to identify genetically more homogeneous subsets and to refine disease gene location estimates. This method is a fast and simple way to identify potentially interesting subsets of families and to help prioritize the list of candidate genes in a region. It can provide direction for follow-up and confirmatory analyses.

This work grew out of our attempts to perform stratified linkage analysis on pre-specified subsets based on disease-related covariates. When there is solid *a priori* evidence for defining strata, procedures such as the predivided sample test [Morton, 1955] should be used. However, when the strata are less obvious it is difficult to choose and defend cutpoints and even more difficult to interpret the results. The advantage of OSA is that it does not require pre-specified cutpoints but allows choices based on the data and then performs a permutation test to evaluate the evidence for linkage in the context of the results for the entire sample.

OSA is quite flexible; d and γ can be held fixed to highlight a specific location, d, or model, γ. In addition, $(Z(\widehat{d}^1,\widehat{\gamma}^1)\dots Z^N(\widehat{d}^N,\widehat{\gamma}^N))$ can be plotted against the covariate values for the families to get a graphic representation of the change in the subset lod scores as families are added [Hall et al., 1990]. OSA can use a variety of covariates, including evidence for linkage at other loci as suggested by Cox et al. [1999], and provides empirical P values for these conditional analyses. OSA is not limited to a single method of linkage analysis but may be applied wherever grids of additive linkage statistics by family can be obtained [Watanabe et al., 1999]. For example, $Z_i(d,\gamma)$ may be any additive statistic including lod scores, χ^2 values, or squared normal Z-scores.

The identification of the BRCA1 gene provided an important example of the challenges inherent in mapping complex human disease and of the successes that could be obtained by careful examination of the data. The small empirical *P* value obtained for the families with low mean age of onset clearly rejected the null hypothesis that these families are a random sample from families with breast cancer. In reality, no *P* value was necessary to identify the obvious clustering of linkage evidence in the seven families with the earliest mean age of onset of breast cancer [Hall et al., 1990]. However, applying this paradigm to studies of common diseases has been somewhat more difficult because of the greater degree of genetic heterogeneity and the smaller genetic effects of single genes accompanied by the much larger number of families required to detect them. Thus, we sought a flexible test to identify homogeneous subsets for further study.

OSA IN A GENOME SCREEN

An appealing feature of OSA in the genome scan context is the ability to re-estimate the disease gene location in the subset. When OSA is used on genome screen data, the entire multipoint lod score curves for each chromosome are used to define the maximum subset lod score. Our OSA software provides plot-ready files so that these multipoint curves may be compared in the subset and in the overall group. By re-estimating the gene location in the maximum ordered subset, it may be possible to narrow a large region of linkage to something more manageable, with a narrower one-lod down support interval [Ghosh et al., 2000; Shao et al., 2003]. It may also happen that the maximum likelihood disease gene location in the subset may be quite different from the maximum likelihood disease gene location in the entire sample. In some cases, it may be desirable to compare the OSA maximum subset lod score at a particular point on a map, say at the estimated disease gene location of the maximum in the entire sample, to understand the contribution of the covariate to evidence for linkage at that point. Our OSA software allows for analysis of a particular point and reports empirical *P* values for the test at that point.

INTERPRETATION OF THE MAXIMUM SUBSETS LOD SCORE

The maximum ordered subset lod score must be at least as large as the maximum lod score in the original sample. Thus, the value of the maximum subset lod score is sensitive to the linkage evidence in the total sample. If there is no evidence for linkage in the total sample, the empirical *P* value may be quite significant for a subset showing only moderate evidence for linkage. If there is substantial and perhaps diffuse evidence for linkage, the *P* values for the maximum subset lod score may not be significant. While the *P* values may be disappointing in themselves, the resulting subset may be useful in suggesting a group for further study or for refining the location in a diffuse area of linkage. Thus, it is important to consider the significance of a given result in the context of the overall evidence for linkage. Cox et al. [1999] propose examining the difference in the overall and conditional lod scores (Δ lod) to evaluate the effects of epistasis or genetic heterogeneity. As we have shown, that idea may also be applied in OSA.

COMPARISON TO OTHER METHODS

We compared the results of OSA to the results of the admixture test for heterogeneity as implemented in HOMOG [Ott, 1983, 1999]. These results demonstrate that when a covariate related to evidence for linkage is identified, power to detect linkage in subsets can be enhanced using OSA, even when the overall genetic effect is low or when the families are small. It is to be expected that the power to detect the heterogeneity and to identify a subset using the admixture test implemented in HOMOG, which uses only the evidence for linkage in each family, would be low in the case of ASPs. When the families are larger and more variability across the family-specific lod scores is possible, as in the breast cancer example, the admixture test will have increased power to detect heterogeneity. However, for the family samples collected to study common complex genetic traits, OSA provides an additional means of identifying a subset of families with increased evidence for linkage by accumulating that evidence across families with similar covariate levels.

There have been a number of other recent methods proposed to identify and control for genetic heterogeneity. The conditional logistic regression method of Olson [1999; Olson et al., 2001] allows for modeling the relationship between evidence for linkage and dependence on covariates. Goddard et al. [2001] showed how this conditional logistic approach can be applied to a genome screen and how the lod score, as a function of a likelihood ratio test on proposed models, can be increased as covariates are included in the model. Langefeld and colleagues [Langefeld et al., 2001; Davis et al., 2001] propose nonparametric linkage (NPL) regression to allow conditional or simultaneous tests of multiple genetic loci, phenotypic or environmental covariates and their interactions. Bull et al. [2002] propose modeling ASP IBD sharing probabilities using covariates, including examination of experiment-specific covariates such as plate assignment. Devlin et al. [2002] apply mixture modeling methods, similar to the original heterogeneity models proposed by Smith [1963]. These modeling approaches can be powerful for examining well-defined models for specific genes or genetic locations including covariates. OSA is a simpler approach that can be easily applied to a wide variety of covariates. As a screening tool, OSA does not require a particular location or genetic model to be chosen and, thus, provides greater flexibility in detecting heterogeneity when heterogeneity might not be otherwise suspected on the basis of overall lod scores.

SUITABLE COVARIATES

For many complex traits, there are multiple choices of trait-related covariates to use in OSA. This method is well suited to any quantitative covariate, either continuous or ordered. The covariate may be adjusted by variables such as age and gender if they are known to influence the value of the covariate. Analyses of data from the FUSION study of the genetics of type 2 diabetes used the mean of the covariate value in all affected sibs to order the families [Ghosh et al., 2000]. Other summary statistics such as the sibship median, minimum, or maximum could be used. Complex functions of the data could be used, including results of data reduction techniques, such as principal components or cluster analysis [Merette et al., 1999; Shao et al., 2003] or residuals from fitting a linear model [Tores et al., 1999]. OSA may be applied to an ordinal categorical covariate such as number of affected family members or the proportion of family members with the specific characteristic, such as presence or absence of a specific allele, or evidence for linkage at another locus [see also Cox et al., 1999].

MULTIPLE COMPARISONS

We control for the inflation in the false positive rate induced by examining multiple family subsets for a given covariate by generating empirical *P* values using a permutation test, which appears to give the proper type 1 error rate in limited simulations. However, we have not controlled for ordered subset comparisons over multiple trait-related covariates or multiple regions of the genome as conditioning loci. It is likely that correlation between some trait-related covariates will be operating. This suggests that a Bonferroni-type correction of the *P* value will be very conservative. Due to the exploratory nature of the analysis, we do not feel compelled to apply a correction for multiple comparisons. We strive for consistency of results across the various analyses within our study as well as comparisons of results to analyses performed by other groups. In the absence of a correction for multiple comparisons, we stress the hypothesis-generating nature of these results and the need for follow-up.

POWER VS. SAMPLE SIZE

In analyses that subset the total data set, there is an implicit tradeoff between the power to detect a given effect and the decreasing sample size in the subset. Leal and Ott [2000] show that if the relative risk in the subset is close to one, then the marginal increase in the genetic effect in the strata does not offset the decrease in power due to the reduced sample size. In this case, subsetting the data will not help and only a large increase in sample size will provide sufficient power to detect small genetic effects. In what we present here, we presume that the increase in the apparent genetic effect afforded by increasing homogeneity in the subset will offset the loss of power induced by the reduced sample size [Kovac et al., 1999]. Another factor that impacts power for OSA is the correlation between the evidence for linkage and the levels of the covariate. Our simulations used a large (two standard deviation) difference in the means for the linked and unlinked ASPs. Additional simulations with a one standard deviation difference in the means for linked and unlinked ASPs suggest that, while there is a decrease in the power, it is not substantial (e.g., 0.804 for 2SD vs. 0.778 for 1SD in the 1.2/2.5D model in Table IV). We are continuing to explore sensitivity of OSA to the covariate values in additional simulations. So far, our power studies suggest that OSA has reasonable power to detect a subset with increased evidence for linkage when using a trait-related continuous covariate. As a result OSA appears to be a useful tool in the set of statistical methods for the analysis of complex genetic traits.

SUMMARY

We have developed the ordered subset method to use trait-related covariates to identify a more genetically homogeneous sample. We generate an empirical *p* value to test the hypothesis that the families providing the maximum subset evidence for linkage are a random sample from all possible ordered subsets. We believe that such analyses will help identify homogenous sets of families that will, in turn, provide better estimates of a disease gene location. These families along with the trait-related covariates may help prioritize candidate genes in regions of linkage in these families. We have developed software to perform these analyses, which is available from the authors via the internet.

ACKNOWLEDGMENTS

We thank our FUSION colleagues for data and encouragement. We thank Chun Li, Heather Stringham, Julie Douglas, Mike Epstein, Margaret Pericak Vance, Eden Martin, and Bill Scott for helpful discussions. E.R.H. is supported by NIH grant MH59528. R.M.W. was previously supported by an individual NSRA from the NIH (DK09525) and a Career Development Award from the ADA and is currently supported by an ADA Research Award. W.L.D. and C.D.L. were supported by NIH training grant HG00040. M.B. is supported by NIH grant HG00376.

ELECTRONIC DATABASE INFORMATION

OSA, SIBLINK and SIMLA software: http://wwwchg.mc.duke.edu/software/index.html

REFERENCES

Arnett DK, Pankow JS, Atwood LD, Sellers TA. 1997. Impact of adjustments for multiple phenotypes on the power to detect linkage. Genet Epidemiol 14:749–754.

Bass MP, Martin ER, Hauser ER. 2004. Pedigree generation for analysis of genetic linkage and association. Proceedings of the Pacific Symposium on Biocomputing 2004:93–103.

Bhat A, Heath SC, Ott J. 1999. Heterogeneity for multiple disease loci in linkage analysis. Hum Hered 49:229–231.

Bull SB, Greenwood CM, Mirea L, Morgan K. 2002. Regression models for allele sharing: analysis of accumulating data in affected sib pair studies. Stat Med 21:431–444.

Chiano MN, Yates JR. 1994. Bootstrapping in human genetic linkage. Ann Hum Genet 58:129–143.

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A. 1999. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat Genet 21:213–215.

Davis CC, Brown WM, Lange EM, Rich SS, Langefeld CD. 2001. Nonparametric linkage regression. II: Identification of influential pedigrees in test for linkage. Genet Epidemiol 21(Suppl):S123–S129.

Devlin B, Jones BL, Bacanu SA, Roeder K. 2002. Mixture models for linkage analysis of affected sibling pairs and covariates. Genet Epidemiol 22:52–65.

Falk C. 1997. Effect of genetic heterogeneity and assortative mating on linkage analysis: A simulation study. Am J Hum Genet 61:1169–1178.

Futreal PA, Liu Q, Shattuck-Eidens D, Cochran C, Harshman K, Tavtigian S, Bennett LM, Haugen-Strano A, Swensen J, Miki Y, Eddington K, McClure M, Frye C, Weaver-Feldhaus J, Ding W, Gholami Z, Soderkvist P, Terry L, Jhanwar S, Berchuck A, Iglehart JD, Marks J, Ballinger DG, Barrett JC, Skolnick MH, Kamb A, Wiseman R. 1994. BRCA1 mutations in primary breast and ovarian carcinomas. Science 266:120–126.

Ghosh S, Watanabe RM, Valle T, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, Mohlke K, Silander K, Kohtamaki K, Chines P, Balow J, Birznieks G, Chang J, Eldridge W, Erdos MR, Karanjawala ZE, Knapp JI, Kuelko K, Martin C, Morales-Mena A, Musick A, Musick T, Pfahl C, Porter R, Rayman JB, Rha D, Segal L, Shapiro S, Sharaf R, Shurtleff B, So A, Tannenbaum J, Te C, Tovar J, Unni A, Whiten R, Witt A, Blaschak-Harven J, Douglas JA, Duren WL, Epstien MP, Fingerlin TE, Kaleta HS, Lange EM, Li C, McEachin RC, Stringham HM, Trager E, White PP, Erikkson J, Toivanen L, Vidren G, Nylund SJ, Tuomilehto-Wolf E, Ross EH, Demirchyan E, Hagopian WA, Buchanan TA, Tuomilehto J, Bergman RN, Collins FS and Boehnke M. 2000. The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetic (FUSION) study: I. An autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67:1174–1185.

Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM. 2001. Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. Am J Hum Genet 68:1197–1206.

Greenwood CM, Bull SB. 1999. Analysis of affected sib pairs, with covariates–with and without constraints. Am J Hum Genet 64:871–875.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–1689.

Hauser ER, Boehnke M. 1998. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. Biometrics 54:1238–1246.

Hauser ER, Boehnke M, Guo SW, Risch N. 1996. Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. Genet Epidemiol 13:117–137.

Kovac I, Rouillard E, Merette C, Palmour R. 1999. Exploring the impact of extended phenotype in stratified samples. Genet Epidemiol 17:S211–S216.

Langefeld CD, Davis CC, Brown WM. 2001. Nonparametric linkage regression. I: Combined Caucasian CSGA and German genome scans for asthma. Genet Epidemiol 21(Suppl): S136–S141.

Leal SM Ott J. 2000. Effects of stratification in the analysis of affected sib-pair data: Benefits and costs. Am J Hum Genet 66:567–575.

Lunetta KL Rogus JJ. 1998. Strategy for mapping minor histocompatibility genes involved in graft-versus-host disease:

a novel application for the discordant sib pair methodology. Genet Epidemiol 15:595–607.

Mahtani MM, Widen E, Lehto M, Thomas J, McCarthy M, Brayer J, Bryant B, Chan G, Daly M, Forsblom C, Kanninen T, Kirby A, Kruglyak L, Munnelly K, Parkkonen M, Reeve-Daly MP, Weaver A, Brettin T, Duyk G, Lander ES, Groop LC. 1996. Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. Nat Genet 14:90–94.

Merette C, King MC, Ott J. 1992. Heterogeneity analysis of breast cancer families by using age at onset as a covariate. Am J Hum Genet 50:515–519.

Merette C, Cayer M, Rouillard E, Roy AK, Guibord P, Kovac I, Ghazzali N, Szatmari P, Roy MA, Maziade M, Palmour R. 1999. Evidence of linkage in subtypes of alcoholism. Genet Epidemiol 17:S253–S258.

Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W, Bell R, Rosenthal J, and 33 others. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266:66–71.

Morton NE. 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318.

Morton NE. 1956. The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet 8:80–96.

North BV, Curtis D, Sham PV. 2002. A note on the calculation of empirical p values from Monte Carlo procedures. Am J Hum Genet 71:439–441.

O'Connell JR, Weeks DE. 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance [see comments]. Nat Genet 11:402–408.

Olson JM. 1995. Multipoint linkage analysis using sibpairs: An interval mapping approach for dichotomous outcomes. Am J Hum Genet 56:788–798.

Olson JM. 1999. A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 65: 1760–1769.

Olson JM, Goddard KA, Dudek DM. 2001. The amyloid precursor protein locus and very-late-onset Alzheimer disease. Am J Hum Genet 69:895–899.

Ott J. 1983. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–320.

Ott J. 1999. Analysis of human genetic linkage, 3rd ed. Baltimore, MD: The Johns Hopkins University Press

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.

Risch N. 1990c. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253.

Shao Y, Cuccaro ML, Hauser ER, Raiford KL, Menold MM, Wolpert CM, Ravan SA, Elston L, Decena K, Donnelly SL, Abramson RK, Wright HH, DeLong GR, Gilbert JR, Pericak-Vance MA. 2003. Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. Am J Hum Genet 72:539–548.

Smith CAB. 1963. Testing for heterogeneity of recombination fractions values in human genetics. Ann Hum Genet 27: 175–182.

Tores F, Uhry Z, Detroyes B, Demenais F, Martinez M. 1999. Sib-pair linkage analysis of alcohol dependence taking into account covariates and age-of-onset variability: Evaluation of the residual approach. Genet Epidemiol 17: S349–S354.

Watanabe RM, Ghosh S, Birznieks G, Duren WL, Mitchell BD. 1999. Application of an ordered subset analysis approach to the genetics of alcoholism. Genet Epidemiol 17:S385–S390.