Genetic Epidemiology 21(Suppl 1): S341–S346 (2001)

Association Analysis in a Variance Components Framework

Gonçalo R. Abecasis, Lon R. Cardon, William O.C. Cookson, Pak C. Sham, and Stacey S. Cherny

Wellcome Trust Centre for Human Genetics (G.R.A., L.R.C., W.O.C.C., S.S.C.), University of Oxford, Oxford; Social, Genetic and Developmental Psychiatry Research Center and Department of Psychiatry (P.C.S.), Institute of Psychiatry, London, United Kingdom

Association analyses conducted in a variance components framework can include information from all available individuals but remain unbiased in the presence of familiality or linkage. Models that include both linkage and association parameters provide different estimates of the effect of a single locus and can be used to distinguish causal polymorphisms from other types of variation. We examine some of these models and their properties in a blind analysis of the simulated Genetic Analysis Workshop12 data sets. © 2001Wiley-Liss, Inc.

Key words: association, extended pedigree, QTL, TDT

INTRODUCTION

Allele-sharing analysis can be used to establish linkage between a chromosomal region and a phenotype, but is usually too coarse for gene-identification in complex disease. Linkage disequilibrium and association analyses are theoretically more suited to fine-mapping and are currently being used to refine promising linkage findings.

We have developed a linkage disequilibrium model based on allelic transmission that includes both linkage and association parameters [Abecasis et al., 2000a,c; Fulker et al., 1999]. When a single mutation is present, the combined information from the linkage and association parameters can be used to predict the location (expressed as a disequilibrium coefficient) and allele frequency of the etiological mutation [Cardon and Abecasis, 2000]. For example, if the locus being analyzed is the trait locus or is in complete disequilibrium with it, modelling linkage and association simultaneously results in no detectable linkage component, since the allele means then account for all of the

Address reprint requests to Gonçalo Abecasis or Stacey Cherny, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.

variance. With decreasing levels of disequilibrium, the variance explained by linkage is expected to increase. This simple relationship is readily quantified, and allows estimation of both the extent of disequilibrium between the putative trait locus and the marker locus, and of the allele frequencies of the unobserved trait locus.

Here, we explore the properties of these tests and location estimates in the complex situations encompassed within the Genetic Analysis Workshop (GAW) 12 simulated data sets.

METHODS

Data. We analyzed simulated replicates 1 through 50 in the outbred population. We focus on dissecting the association between Q1 and candidate gene 6 on chromosome 19.

IBD calculations. Single-point identity-by-descent (IBD) calculations were carried out using all available genotypes for all microsatellites. To compensate for low marker heterozygosity and missing parental genotypes, IBD matrices calculated at the nearest microsatellite marker were used as surrogates for IBD matrices at each candidate gene. Allele-sharing probabilities were calculated using the Lander and Green [1987] algorithm as implemented in the program MERLIN [Abecasis et al., 2000b].

Linkage analysis. Linkage was tested using standard variance components methods based on the biometrical model. Briefly, the expected phenotype for each individual i in family j was modeled as a function of the population mean μ and sex-specific mean differences β_{sex}, that is $\hat{y}_{ij} = \mu + \beta_{sex} i_{sex}$ (i_{sex} is an indicator variable for sex). Variance components were used to model deviations from this expectation for each pair of individuals j and k in family i as:

$$\Omega_{ijk} = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\ \pi_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2 & \text{if } j \neq k \end{cases},$$

where σ_a^2 is the additive genetic variance of the QTL, σ_g^2 is the variance attributable to polygenes and σ_e^2 is the residual environmental variance. π_{ijk} and φ_{ijk} are the proportion of alleles shared IBD and the kinship coefficient between individuals j and k in family i.

Association analysis. In regions where linkage was identified, candidate gene polymorphisms were tested for association and linkage disequilibrium using the between-within model of association [Abecasis et al., 2000a; Fulker et al., 1999]. Association was partitioned into two orthogonal components, β_b and β_w, measuring population strata (on the basis of allele frequency differences between population strata defined on the basis of ancestral genotypes) and linkage disequilibrium (on the basis of allelic transmission through the pedigree), which can be calculated efficiently in extended pedigrees [Abecasis et al., 2000c]. Modelling association as a fixed effect gives $\hat{y}_{ij} = \mu + \beta_{sex} i_{sex} + \beta_b b_{ij} + \beta_w w_{ij}$.

Model fitting. Significance of parameter estimates was evaluated using likelihood ratio chi-squared tests. For each parameter of interest (e.g., β_w when testing for linkage disequilibrium), the likelihood of the data was maximized under the null (e.g., $\beta_w = 0$) and alternative hypothesis (e.g., no constraints on β_w) using QTDT [Abecasis et al., 2000a]. In all cases, the following likelihood of the data, expressed in terms of the observed phenotypes $\mathbf{y}_i$ and expected phenotypes and the random effects in Ω_i was used:

$$L = \prod_i (2\pi)^{-n_i/2} \left|\hat{\Omega}_i\right|^{-1/2} e^{-1/2[(y_i-\hat{y}_i)'\hat{\Omega}_i^{-1}(y_i-\hat{y}_i)]}.$$

RESULTS

We focus on dissecting the association between Q1 and candidate gene 6 on chromosome 19, where a single single-nucleotide polymorphism (SNP) accounts for 24% of the phenotypic variance. This exercise illustrates some of the challenges in fine-mapping and candidate gene studies using SNPs. In a single-point analysis using all 105 chromosome 19 microsatellite markers, the average peak lod score for each replicate was 2.67 (varying between 0.22 for replicate 30 and 6.23 for replicate 40). While 68% percent of replicates exhibited suggestive evidence for linkage (lod > 2.0), only 34% percent of replicates exhibited conclusive evidence for linkage (lod > 3.0) and 6% exhibited virtually no evidence for linkage (peak lod < 1.0).

Table I presents a summary of association results for the most common SNP polymorphisms. On average, the strongest association observed between Q1 and candidate gene 6 corresponded to a lod score of about 18.57 (varying between 10.35 for replicate 2 and 28.97 for replicate 24). In addition to the etiological variant (SNP 5782), another six markers show significant association at genome-wide significance levels (that is, $p < 5 \times 10^{-8}$, corresponding to 1,000,000 independent tests) on all samples, while a further three displayed significant association at this level in over 40% of replicates. The peak of association only corresponds to the functional mutation in 64% of the samples, but the functional mutation is within 1 lod unit of the strongest observed association in all samples. All strongly associated polymorphisms (5007, 6805, 7073, 7332, 8067, and 8226) are in nearly complete disequilibrium with 5782 (pair-wise $\Delta^2 \geq 0.97$), while the polymorphisms displaying intermediate levels of association (993, 1748, 1987, and 4411) are in moderate disequilibrium with 5782 (pair-wise $\Delta^2 \geq 0.43$). While the table summarizes the effect of more common alleles, note that patterns of association are more haphazard when rare alleles are considered. For example, polymorphism 5663 (with a frequency of just less than 5%) shows an association lod score of less than 1.0 in all replicates, but is within 200 bp of the etiological mutation.

Although fixed effects (such as association) and random effects (such as linkage) are measured on different scales, it is important to decide whether an observed association can account for previously observed linkage for a particular region. In large samples, linkage effects disappear when association with a functional polymorphism is modeled using the linkage and association model of Fulker et al. [1999]. Table II presents a summary of evidence for linkage in each replicate after accounting for association. Evidence for linkage was measured at the nearest microsatellite marker (D19G032) while allowing for association to each individual SNP and 21 replicates exhibiting lod scores < 1.0 before modelling association were excluded from this analysis. Again the functional variant (5782) and six neighboring SNPs are indistinguishable and account for over 80% of the linkage in 93% of all samples. Using the derivations of Cardon and Abecasis [2000], this indicates that disequilibrium between these markers and the trait mutation is $\Delta^2 \geq 0.90$. The small proportion of linkage that is unexplained by the disease mutation in each sample is likely attributable to random fluctuations expected in samples of this size.

TABLE I. Summary of Association lod Scores[a]

SNP Position	Average lod score		Proportion of significant hits			Peak of association	
			Genome	Region	Point	Exact	Near
993	7.33	(1.96 -16.13)	46%	84%	100%	-	-
1748	7.35	(1.96 -16.13)	44%	86%	100%	-	-
1987	7.36	(2.50 -16.13)	46%	86%	100%	-	-
2275	0.39	(0.00 - 2.77)	-	-	12%	-	-
3477	0.39	(0.00 - 2.77)	-	-	12%	-	-
4411	7.41	(2.41 -15.80)	42%	88%	100%	-	-
4848	1.37	(0.03 - 4.10)	-	-	46%	-	-
5007	18.35	(10.35 -28.32)	100%	100%	100%	54%	98%
5782	**18.47**	**(10.35 -28.78)**	**100%**	**100%**	**100%**	**64%**	**100%**
6805	18.43	(10.35 -28.78)	100%	100%	100%	58%	100%
7073	18.39	(10.35 -28.78)	100%	100%	100%	58%	98%
7332	18.43	(10.35 -28.78)	100%	100%	100%	58%	100%
8067	18.41	(10.35 -28.78)	100%	100%	100%	58%	98%
8226	18.06	(9.71 -28.97)	100%	100%	100%	38%	82%
9616	0.64	(0.00 - 2.69)	-	-	10%	-	-
9954	0.27	(0.00 - 2.17)	-	-	2%	-	-
10054	0.61	(0.00 - 2.74)	-	-	8%	-	-
10955	0.61	(0.00 - 2.61)	-	-	8%	-	-
11146	0.22	(0.00 - 1.00)	-	-	-	-	-
11782	0.71	(0.00 - 3.96)	-	-	18%	-	-
11981	0.42	(0.00 - 1.26)	-	-	-	-	-
12408	0.47	(0.00 - 2.37)	-	-	10%	-	-
12716	0.22	(0.00 - 1.86)	-	-	4%	-	-
13869	0.72	(0.00 - 4.49)	-	2%	18%	-	-
14425	0.28	(0.00 - 1.04)	-	-	-	-	-
14544	0.31	(0.00 - 1.03)	-	-	-	-	-
15021	3.54	(0.28 - 7.63)	2%	36%	90%	-	-

[a]Association results between phenotype Q1 and each SNP in candidate gene 6 using the between-within model. The table indicates the average, minimum and maximum association lod scores for all replicates and the proportion of tests reaching genome-wide (1,000,000 tests), candidate region (1,000 tests) and point-wise (one test) significance at the 0.05 level. The last two columns indicate the proportion of times the marker displayed the highest lod score in the region or was within 1 lod unit of the highest lod score in the region.

CONCLUSIONS

We show that the between-within model can distinguish linkage disequilibrium and association in extended pedigrees. In these replicates, which correspond to unselected random samples, the effect of the 5782 mutation on Q1 could be established at genome-wide significance levels in all the replicates. However, the simulated effect is relatively large and in practice selected samples are likely to provide a more powerful setting. It is important to note that while 5782 was within one lod unit of the strongest association in all replicates, other polymorphisms exhibited stronger association in 36% of replicates. Distinguishing functional polymorphisms from other strongly associated alleles, through comparison of competing statistical models, may require prohibitively large sample sizes.

Association mapping is much more powerful than linkage mapping when the number of disease alleles is modest. However, association mapping can be successful even for multi-allelic systems of modest complexity. For example, sequential analysis

TABLE II. Summary of Linkage lod Scores After Modeling Association[a]

		Proportion of lod score		
SNP position	Proportion of explained lod score	= 1.00	> 0.95	> 0.80
993	0.62 (0.07 - 1.00)	7%	7%	24%
1748	0.62 (0.07 - 1.00)	7%	7%	24%
1987	0.63 (0.07 - 1.00)	7%	7%	24%
2275	0.11 (0.00 - 0.39)	0%	0%	0%
3477	0.11 (0.00 - 0.39)	0%	0%	0%
4411	0.63 (0.11 - 1.00)	7%	7%	21%
4848	0.20 (0.00 - 0.88)	0%	0%	3%
5007	0.94 (0.58 - 1.00)	41%	69%	93%
5782	**0.94 (0.58 - 1.00)**	**41%**	**66%**	**93%**
6805	0.94 (0.58 - 1.00)	41%	66%	93%
7073	0.94 (0.58 - 1.00)	41%	66%	93%
7332	0.94 (0.58 - 1.00)	41%	66%	93%
8067	0.94 (0.58 - 1.00)	41%	66%	93%
8226	0.93 (0.58 - 1.00)	41%	66%	90%
9616	0.15 (0.00 - 0.72)	0%	0%	0%
9954	0.12 (0.00 - 0.37)	0%	0%	0%
10054	0.15 (0.00 - 0.70)	0%	0%	0%
10955	0.16 (0.00 - 0.70)	0%	0%	0%
11146	0.12 (0.00 - 0.43)	0%	0%	0%
11782	0.17 (0.00 - 0.68)	0%	0%	0%
11981	0.15 (0.00 - 0.45)	0%	0%	0%
12408	0.14 (0.00 - 0.49)	0%	0%	0%
12716	0.13 (0.00 - 0.42)	0%	0%	0%
13869	0.17 (0.00 - 0.68)	0%	0%	0%
14425	0.10 (0.00 - 0.37)	0%	0%	0%
14544	0.11 (0.00 - 0.43)	0%	0%	0%
15021	0.29 (0.00 - 0.71)	0%	0%	0%

[a]Tests for linkage between phenotype Q1 and marker D19G032, after discounting association to each SNP in candidate gene 6 using the between-within model. The table indicates the average, minimum, and maximum proportion of the original lod score explained. The last three columns indicate the proportion of times that the SNP mutation accounted for 1.00, > 0.95, or > 0.80 of the linkage evidence. Only the 29 replicates exhibiting linkage between D19G032 and Q1 (lod > 1.0) were examined.

with the between-within model identifies seven SNPs in candidate gene 2 that account for > 90% of the linkage between Q5 and chromosome 1 (data not shown).

The between-within association model of Fulker et al. [1999] incorporates a test for population stratification and we find no significant evidence for population substructure effects leading to spurious association. If it is not necessary to guard against population stratification, a more powerful association test can be constructed by setting the between and within components of association to be equal. This alternative test almost doubles the lod scores for association, but does not improve separation of the 5782 from strongly associated alleles.

ACKNOWLEDGMENTS

Supported by the Wellcome Trust and grant EY-12562 from the National Institutes of Health.

REFERENCES

Abecasis GR, Cardon LR, Cookson WOC. 2000a. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279-92.

Abecasis GR, Cherny SS, Cookson WOC, et al. 2000b. MERLIN – Multipoint engine for rapid likelihood inference. Am J Hum Genet 67(Suppl. 2):326.

Abecasis GR, Cookson WOC, Cardon LR. 2000c. Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8:545-51.

Cardon LR, Abecasis GR. 2000. Some properties of variance-components model for fine-mapping of quantitative trait loci. Behav Genet 30:235-43.

Fulker DW, Cherny SS, Sham PC, et al. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259-67.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363-7.