Genetic Epidemiology 21(Suppl 1): S117–S122 (2001)

The Effect of Genotype and Pedigree Error on Linkage Analysis: Analysis of Three Asthma Genome Scans

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

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

The effects of genotype and relationship errors on linkage results are evaluated in three of the Genetic Analysis Workshop 12 asthma genome scans. A number of errors are detected in the samples. While the evidence for linkage is not striking in any data set with or without error, in some cases the difference in test statistic could support different conclusions. The results provide empirical evidence for the predicted effects of genotype and relationship error and highlight the need for rigorous detection and elimination of data error in complex trait studies.
© 2001 Wiley-Liss, Inc.

Key words: asthma, genome scan, genotype error, linkage, pedigree error

INTRODUCTION

The lack of success in complex disease linkage genome scans and their lack of replication has been attributed to genetic heterogeneity, lack of power (sometimes due to inappropriate analytical frameworks), over-simplistic models of oligogenic inheritance, and the failure to allow for multi-locus models. The lack of consistency may also be due, at least in part, to genotype error, which has recently been shown to substantially reduce power to detect linkage [Douglas et al., 2000]. This is most pronounced in the case of affected sib-pair studies, in which genotype error reduces sharing of alleles identical by descent (IBD) and, therefore, decreases the test statistic for linkage [Abecasis et al., 2001]. Similarly, the effect of genotype error on the power to detect linkage with quantitative traits depends on the specific method of sample ascertainment.

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

Additionally, errors in specification of familial relationships can lead to incorrect conclusions regarding evidence for linkage [reviewed in Epstein et al., 2000], which further detracts from reproducibility of linkage outcomes. Several methods have been developed that make use of genome-wide marker data for the detection of such errors [McPeek and Sun, 2000].

For the present report, we analyzed the Genetic Analysis Workshop (GAW) 12 asthma genome scan data sets using affected relative pair methods (for qualitative traits) and a distribution-free statistic for analysis of quantitative traits. We first analyzed the data as they were made available for GAW12. We then examined the data for relationship and genotype errors, corrected these errors, and re-analyzed all data sets.

METHODS

Data

The Collaborative Study on the Genetics of Asthma (CSGA), Busselton, and German asthma data sets were analyzed. While the Busselton and German data sets consist mainly of two-generation families, the CSGA data set includes several three-generation pedigrees. The phenotypes used comprised asthma affection status (discrete) and log-IgE (quantitative). The samples were analyzed separately and combined before and after error correction. Log-IgE was standardized within each sample.

IBD Calculations

All three data sets employed sparse (10 cM) microsatellite markers. Multipoint IBD calculations were carried out for all chromosomes using the Marshfield marker map. Allele-sharing probabilities were calculated using the Lander and Green [1987] algorithm as implemented in the program MERLIN [Abecasis et al., 2000]. Allele frequencies were estimated separately within each data set, using all available genotypes. Population-specific allele frequencies were used when analyzing the combined data set.

Analysis Methods

For assessment of asthma affection, the NPL-ALL linkage statistic [Whittemore and Halpern, 1994] was used. For quantitative trait analyses, we used the distribution-free method implemented in MERLIN [Abecasis et al., 2000]. This method allows for direct comparison of the different data sets that were collected using different ascertainment criteria, as it is designed to avoid assumptions of trait or residual normality that are implicit in most other quantitative trait linkage tests. In can be evaluated in arbitrary pedigrees and is robust to the method of sample ascertainment. The value of the statistic, Z_{QTL}, increases when relatives with similar phenotypes share alleles IBD and decreases otherwise. For phenotype vector y_i in the i^{th} family, with population mean, μ, the test is constructed as:

$$S_{allele} = \sum_{\text{all carriers of allele}} (y_i - \mu)$$

$$S_{QTL} = S_{allele}^2$$

$$Z_{QTL} = \frac{E(S_{QTL} \mid \text{phenotypes, genotypes}) - E(S_{QTL} \mid \text{phenotypes}) \cdot}{\sqrt{\text{Var}(S_{QTL} \mid \text{phenotypes})}}$$

Pedigree Error Detection

The genetic relationship between any two individuals defines an expected pattern of allele sharing between them. The details of this pattern can be complex and will depend on the exact type of relationship, marker characteristics, population history, and inbreeding. Errors in relationship specification were detected using a simple graphics-based method that detects family member pairs where patterns of their identity-by-state (IBS) sharing across the genome are inconsistent with their coded relationship. The method involves plotting the mean IBS across the genome against its standard deviation, for each relationship pairing, and locating outliers from this bivariate distribution.

Genotype Error Detection

Genotype errors were detected using the likelihood-based procedures in MERLIN [Abecasis et al., 2000]. MERLIN uses a novel algorithm to identify genotypes that imply unlikely recombinants in general pedigrees. After dropping each genotype, the pedigree likelihood is re-evaluated both conditional on the current map and assuming unlinked markers. Markers that produce relatively large changes in the first likelihood are flagged as probable errors. In dense maps, all genotypes that introduce obligate recombinants can be identified.

RESULTS

An illustration of the graphical output used to detect errors of relationship in these data is shown in Figure 1. For each coded pair of individuals in the CSGA sample, the IBS standard deviation is plotted against the IBS mean, both of which were calculated using all available marker genotypes. For clarity, we only plot three types of relationship, as outliers in other relationships could not be identified with a high level of certainty. The figure shows clear mean/standard deviation relationships for each relative pair; half-sibs form the left-most cluster, parent-offspring pairs form the tight, linear distribution in the lower right quadrant, and full-siblings cluster in the right, top-most area. These two-dimensional relationships clearly demarcate the empirical range for each relative pair and thereby highlight outliers, which almost certainly reflect pedigree errors. In the data illustrated, there are obvious erroneous full-sibling pairs (open squares) that are closely aligned with the half-sib IBS distribution (these misclassifications would reduce allele sharing statistics) or monozygotic (MZ) twin IBS distribution (these misclassifications would inflate allele sharing statistics). There are also “half-siblings” (open triangles) with genome-wide IBS resembling that of full-siblings.

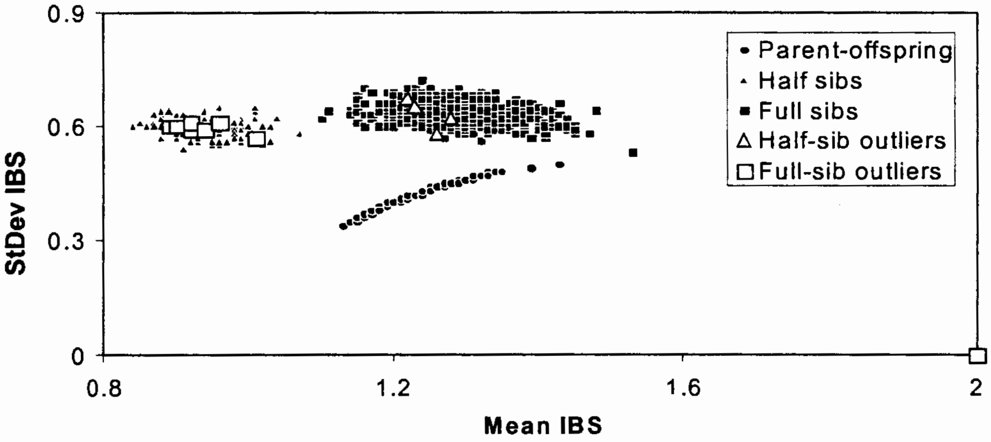

Fig. 1. CSGA relationship error detection.

Family relationship errors identified using the graphical procedure are summarized in Table I, in which the 'coded,' or expected, relationship is compared with the 'true' relationship defined on the basis of the genomic distribution of IBS mean and variance. Three types of relationship errors are apparent in the samples: half-sibs who are actually full-siblings, full-siblings who are actually MZ twins, and full-siblings that are actually half-siblings. The first two of these errors are expected to increase type I error, while the latter decreases linkage power. Interestingly, the former errors inflate, rather than deflate, allele-sharing statistics in affected-pair analyses. In total, there are 14 familial errors in the CSGA sample and 2 errors in the Busselton data. No relationship errors were detected in the German collection.

Genotype error detection procedures identified 501 likely errors in the genome-scan panels. Of these errors, 236, 170, and 95 were attributable to the Busselton, CSGA and German collections, respectively. Overall, these errors reflect respective error rates of at least 0.25%, 0.04%, and 0.07%. Since there are a substantial number of missing genotypes and the maps are sparse, many undetected errors may be present. For example, the CSGA data set has the smallest proportion of detected errors but also the greatest proportion of missing genotypes (20%).

TABLE I. Relationship Errors in Asthma Samples

Study	Coded	Predicted	Number
CSGA	Half sib	Full sib	4
	Full sib	Half sib	8
	Full sib	MZ twin	2
Busselton	Full sib	MZ twin	2

No relationship errors were detected in the German sample.

Table II presents the highest lod scores obtained in the genome scans. An arbitrary cutoff of lod ≥ 1.80 in the corrected sample was empirically selected so that several regions could be compared before and after error removal.

The results show that genotype and relationship errors can have a substantial impact on linkage results. The Busselton sample, which offers the most dramatic outcomes, illustrates a gain of as much as 59% in lod score (chromosome 15; lod 1.62 vs. 2.57) as a result of removing data points with likely errors. Conversely, the same collection shows four loci for which the lod score is reduced slightly as a result of error removal. These discrepancies may be due to the competing effects of lod deflation due to genotype error versus lod inflation due to the mislabeling of MZ twins as full-siblings. Naturally, in these small samples the observed changes will include random fluctuations.

The lack of any detectable relationship errors in the German data, coupled with the low genotype error rate, lends robustness to the linkage results. Interestingly, the CSGA data have the fewest lod scores that meet even our relaxed lod-score criterion of 1.80 (one trait at one marker in the entire genome scan). While this may simply reflect a lack of linkage power in this data set, it is intriguing that the CSGA also has the most relationship errors (Table I). We find no striking results in the combined data set before or even after removing error.

DISCUSSION

Pedigree errors and incorrectly assigned genotypes are known to have dramatic consequences in linkage studies. There are a number of relationship errors in some of the GAW12 asthma samples and the genotyping error rates, which appear low by many

TABLE II. Lod Scores ≥ 1.8 in Each Genome Scan

Study	Trait	Chromosome/position	Lod before correction	Lod after correction
CSGA	Q_Asthma[a]	Chr 11 / 69.9	1.90	1.85
Busselton	Q_Asthma	Chr 1 / 224.1	1.99	1.83
	ln(IgE)	Chr 4 / 117.1	1.57	1.88
	Q_Asthma	Chr 6 / 148.3	2.51	2.22
	Q_Asthma	Chr 7 / 97.4	2.99	2.77
	Q_Asthma	Chr 10 / 19.0	2.05	2.00
	ln(IgE)	Chr 15 / 22.0	1.62	2.57
Germany	ln(IgE)	Chr 2 / 229.2	2.96	2.97
	ln(IgE)	Chr 12 / 119.8	2.84	2.71
	ln(IgE)	Chr 15 / 86.0	1.76	1.80
Combined	ln(IgE)	Chr 2 / 96.4	1.80	1.82
	Asthma	Chr 6 / 49.3	1.83	1.85
	ln(IgE)	Chr 6 / 52.9	1.88	1.93
	Q_Asthma	Chr 11 / 67.5	2.08	1.97
	Q_Asthma	Chr 12 / 75.1	2.04	2.00

[a] The phenotype Q_Asthma was coded as 1/-1 for affected/unaffected and analyzed using the distribution-free quantitative trait test.

standards (all < 1%), are still of sufficient magnitude to create substantial losses in linkage power. Assessment of the effects of such error on the GAW12 asthma data sets revealed some large changes in linkage results. While none of the data sets showed striking evidence for linkage with or without error, the scale of the test statistic was such that different conclusions about the presence or absence of linkage could have been reached. The outcomes emphasize the necessity of rigorous data examination in linkage studies of multifactorial traits.

ACKNOWLEDGMENTS

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

REFERENCES

Abecasis GR, Cherny SS, Cardon LR. 2001. The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet 9:130-4.

Abecasis GR, Cherny SS, Cookson WOC, et al. 2000. MERLIN: Multipoint engine for rapid likelihood inference. Am J Hum Genet 67:327.

Douglas JA, Boehnke M, Lange K. 2000. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet 66:1287-97.

Epstein MP, Duren WL, Boehnke M. 2000. Improved inference of relationship for pairs of individuals. Am J Hum Genet 67:1219-31.

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

McPeek MS, Sun L. 2000. Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66:1076-94.

Weiss KM, Terwilliger JD. 2000. How many diseases does it take to map a gene with SNPs? Nat Genet 26:151-7.

Whittemore AS, Halpern J. 1994. A class of tests for linkage using affected pedigree members. Biometrics 50:118-27.