European Journal of Human Genetics (2006) 14, 469–477
© 2006 Nature Publishing Group All rights reserved 1018-4813/06 $30.00
www.nature.com/ejhg

ARTICLE

The impact of data quality on the identification of complex disease genes: experience from the Family Blood Pressure Program

Yen-Pei Christy Chang[1], James Dae-Ok Kim[1], Karen Schwander[2], Dabeeru C Rao[2], Mike B Miller[3], Alan B Weder[4], Richard S Cooper[5], Nicholas J Schork[6], Michael A Province[2], Alanna C Morrison[7], Sharon LR Kardia[8], Thomas Quertermous[9] and Aravinda Chakravarti*[,1]

[1]*McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA;* [2]*Division of Biostatistics, Washington University School of Medicine, St. Louis, MO, USA;* [3]*Division of Epidemiology, University of Minnesota, Minneapolis, MN, USA;* [4]*Division of Hypertension, University of Michigan School of Medicine, Ann Arbor, MI, USA;* [5]*Department of Preventive Medicine and Epidemiology, Loyola Stritch School of Medicine, Maywood, IL, USA;* [6]*Department of Psychiatry, University of California, San Diego, CA, USA;* [7]*Division of Human Genetics, University of Texas, Houston, TX, USA;* [8]*Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA;* [9]*Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA*

The application of genome-wide linkage scans to uncover susceptibility loci for complex diseases offers great promise for the risk assessment, treatment, and understanding of these diseases. However, for most published studies, linkage signals are typically modest and vary considerably from one study to another. The multicenter Family Blood Pressure Program has analyzed genome-wide linkage scans of over 12 000 individuals. Based on this experience, we developed a protocol for large linkage studies that reduces two sources of data error: pedigree structure and marker genotyping errors. We then used the linkage signals, before and after data cleaning, to illustrate the impact of missing and erroneous data. A comprehensive error-checking protocol is an important part of complex disease linkage studies and enhances gene mapping. The lack of significant and reproducible linkage findings across studies is, in part, due to data quality.

European Journal of Human Genetics (2006) **14,** 469–477. doi:10.1038/sj.ejhg.5201582; published online 22 February 2006

Keywords: genome-wide linkage studies; data quality; family relationship

Introduction

Whole genome linkage scans have been enormously successful for the identification of monogenic and some oligogenic disorders. Applying the same approach to locate major genes for complex human diseases has, unfortunately, been much less rewarding. A review of 101 studies of complex human diseases revealed that few consistent and significant linkage results have been obtained, and even fewer novel susceptibility genes have been identified.[1] The failure to detect linkage or to reproduce previous linkage findings has been attributed to inadequate sample size, phenotypic misclassification, genotyping errors, and the standard hallmarks of complex diseases such as low effect sizes of variant alleles (ie, low locus-specific heritability), phenocopies, and genetic heterogeneity. Data quality is at least as important as the study design and the methods of analysis, particularly since this is under the

*Correspondence: Professor A Chakravarti, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 North Broadway, BRB 580, Baltimore, MD 21205, USA. Tel: +1 410 502 7525; Fax: +1 410 502 7544; E-mail: aravinda@jhmi.edu

Received 23 March 2005; revised 19 October 2005; accepted 25 November 2005; published online 22 February 2006

investigator's control. While there is extensive literature on how data quality impacts linkage study results, few recommendations exist on the logical sequence in how to detect, correct, and remove erroneous data in preparation for analysis. In this report, we present our protocol for cleaning genotyping and pedigree errors based on our experience with the large multicenter NHLBI Family Blood Pressure Program (FBPP).[2] Using specific examples from one of the FBPP network, GenNet, we also discuss the impact of erroneous data on linkage results.

When mapping monogenic diseases, the location of the disease-causing locus is often defined by a few individuals with recombinations near the disease locus. The phenotypes and genotypes at relevant markers of these individuals are confirmed and repeated before one starts searching for candidate genes. Finding genes underlying complex diseases necessitate a much larger sample size to distinguish a linkage signal from random noise. Moreover, late age of onset and mortality in these diseases make collecting large pedigrees a real challenge if not an impossibility. As the evidence of linkage and the locations of the susceptibility loci are based on a large number of small families, it is both costly and impractical to repeat the phenotyping and genotyping protocols. A comprehensive error checking protocol to ensure data accuracy is therefore an important part of complex disease linkage studies.

Generally speaking, four types of errors are common to all family-based genetic studies: errors in phenotypes, pedigree structure, marker information, and marker genotypes. All four can potentially reduce linkage signal and provide misleading information on the linkage region. Since the characteristics of phenotype data vary from one study to another, addressing phenotypic data quality issues usually requires customized protocols for each study. Marker-related errors, such as marker order and recombination distance, have been discussed elsewhere and are not considered here because the human genome sequence and maps of increasing resolution have reduced this category of errors.[3,4] Pedigree structure errors, that is, discrepancies between reported relationships and actual biological relationships, are mostly due to erroneous information reported by study participants. Sample switching and data entry errors can also give the appearance of pedigree structure errors. Genotyping errors are associated both with nonrandom errors at specific markers and a low rate of random errors at all markers. Given a true pedigree structure, genotyping error can be identified by inconsistencies in Mendelian transmission of alleles but only in informative families. Conversely, when a sufficiently large number of markers are genotyped, relationships among members of the same family can be inferred based on allele sharing. This reciprocal error-checking process is especially effective in large pedigrees examined for a large number of highly polymorphic markers. For complex diseases, the pedigree size is usually small and hence has less power to detect genotyping errors. The large number of markers genotyped for a typical 10 cM linkage scan (~400), however, is more than sufficient to infer the pair-wise relationships among all members within and between families.

Materials and methods

The family blood pressure program

The Family Blood Pressure Program (FBPP)[2] was established by the NHLBI in 1995 to identify genes underlying blood pressure (BP) regulation and hypertension. It consists of four collaborating multicenter networks: GenNet, GENOA, HyperGEN and SAPPHIRe. The four Networks pursued different study designs and protocols with multiple ethnic groups including Caucasian Americans, African Americans, Hispanic Americans, and Asians. Over 120 common phenotypic measurements were collected in each Network so that phenotypic and genotypic data can be pooled across Networks for maximum statistical power. During the first 5 years (1995–2000), the project has enrolled a total of 13 592 individuals.

Genotyping

Genome-wide scans using short tandem repeat (STR) markers were carried out by the NHLBI Mammalian Genotyping Service, using a mixture of di-, tri, and tetranucleotide repeat markers. To date, a total of 12 041 individuals have been genotyped. As a reslut of the enormous size of the FBPP, genotyping was performed in several batches spread over several years, using slightly variable marker sets. Some STR markers were intrinsically more difficult to genotype accurately and were phased out in later sets of markers. Thus, the number of markers genotyped varied somewhat over the batches. For example, in GenNet, there were from 365 to 396 markers per batch. A total of 352 markers were genotyped in all seven batches and 54 markers were genotyped in one or a few batches only. An average of 310 samples was genotyped in each batch and heterozygosities ranged from 0.76–0.79. All analyses were carried out using the Marshfield genetic map.[5]

Data cleaning protocol

Our data cleaning protocol consists of three steps. Step 1: identification and removal of error-prone markers. When STR genotyping of a dataset is performed in multiple batches over several years, identical alleles are sometimes called differently because different flanking primers, allele sizing software, or allele binning methods have been used.[6] Examples of this phenomenon, that we term 'allele shifting' or AS, are demonstrated in Figure 1. For marker A, between the first and the last 2 batches, the size of all alleles have been altered by 38 nucleotides by using different primers to amplify the sequence containing the

STR. For Marker B, only those alleles sized smaller than 154 base pairs showed evidence of AS, due to the use of a different binning method or size standard. Step 2: identification and correction of errors in family structure. Once the genotypes affected by AS were deleted, family structure errors due to misreporting or sample switching are corrected based on allele sharing statistics of pairwise relationships within and between pedigrees. Step 3: identification and deletion of random genotype errors. Finally, using confirmed or corrected family structure, genotyping errors that are Mendelian-inconsistent can be identified and removed. Details of our data cleaning protocol are presented in online Supplementary Information.

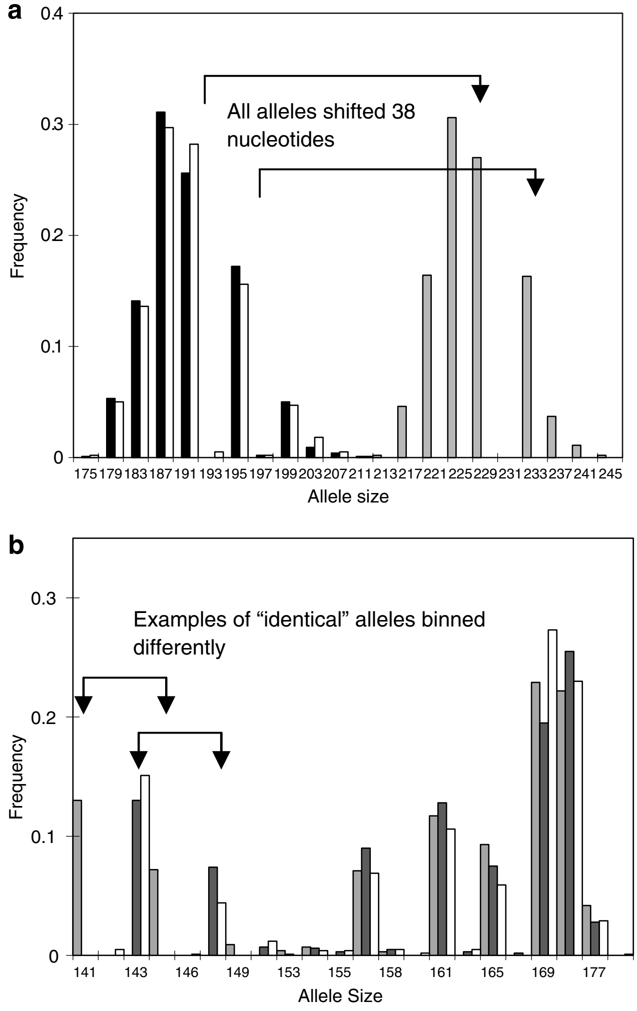

Figure 1 Examples of allele shifting (AS) markers. Gray = Batch 1, Black = Batch 2, White = Batch 3. Marker A: GATA3H06, and Marker B = GATA71E08.

Genome-wide linkage analysis

Variance components linkage analysis (GENEHUNTER,[7] version 2.1) was applied to the GenNet data using three blood pressure (BP)-related phenotypes: systolic blood pressure (SBP), diastolic blood pressure (DBP), and pulse pressure (PP) and body mass index (BMI). Details of the GenNet study population and linkage results are published elsewhere.[8] When comparing our linkage results to 28 other published genome-wide linkage studies of hypertension-related traits, we used stringent criterion in deciding what constituted a replication: we compared only linkage results with maximum lod scores (LODs) greater than 2 and considered a linkage region validated when the markers closest to the LOD were identical in both studies. If different marker panels were used, replication was accepted only if the physical locations (based on NCBI Build 34) of the markers closest to the LOD in both studies are within the one-lod interval defined in each paper.

Results

Summary of the QC results for the entire FBPP

At the end of the data cleaning process, data loss for each Network could be classified into four categories (Table 1) based on how many genotypes there were at each stage of data cleaning relative to the total number of expected genotypes based on the numbers of samples and markers genotyped. Overall, data loss ranges from 7.1 to 9.3% among the four Networks. In addition, a total of 59 full sibling pairs were recovered in the entire FBPP based on the inter-pedigree relationships found by the GRR.

Five markers were found to have clear evidence of AS in all four networks (Supplementary Tables 1 and 2), and FBPP investigators opted to delete all genotypes of these five

Table 1 Source of missing data from each individual network's cleaning process

	GenNet (%)	*GENOA (%)*	*HyperGEN (%)*	*SAPPHIRe (%)*
Genotyping failure	3.5	3.9	4.3	3.2
Inconsistent allele calling	2.4	1.7	1.4	0.8
Family structure	2.8	0.03	1.2	1.1
Genotyping error	0.6	0.5	0.7	4.0
Total	9.3	7.1	7.6	9.1

markers in the entire networks. In doing so, we essentially created gaps in the linkage map and increased the intermarker distances from an average of 8.1 cM (no markers deleted) to 16.2 cM in five chromosomal regions. In addition, other markers were found to be problematic in some, but not all, networks. Two procedures were adopted. First, only those genotypes that contained inconsistently called alleles in the specific batch or ethnic group were deleted. This procedure made the data consistent within each network across all batches of genotype data. Second, all genotypes in all AS markers were deleted so that data from all Networks can pooled for further analysis.

Data cleaning and linkage study: the GenNet experience

From 1995 to 2000, GenNet has analyzed data from a total of 1921 subjects from 603 families that are either Caucasian Americans individuals from Tecumseh, MI, or African Americans from Maywood, IL. The subjects were recruited in nuclear families, which have an average of 3.2 subjects per family. During the data cleaning process, genotypes of 10 markers were deleted because of AS. In addition, 10 families were excluded from linkage analysis because a clear genetic relationship could not be established.

How much of the genome is covered by a 10 cM linkage scan?

If all markers were successfully and correctly genotyped, and were fully informative, then the average intermarker distance in our study would be 9.3 cM and only 0.5% of the genome would be flanked by markers <20 cM apart. As a reslut of <100% marker heterozygosity, genotyping failure, and data deletion in the cleaning step, the distance between informative markers is, as expected, much greater. Postcleaning, the actual distance between two informative (heterozygous) markers is 12.5 cM with >13.5% of the genome flanked by informative markers >20 cM apart (Figure 2). Thus, in a typical genome-wide linkage scan, a significant portion of the genome remains unexamined in various samples. When such segments are concentrated in a particular genomic region, then the power of detecting linkage in this region is compromised.

Does data cleaning improve the likelihood of finding true linkage?

Postcleaning, 11 chromosomal regions had maximum lod scores (LODs) >2 with BP-related phenotypes (1q25, 2q33, 10p14, 10q21, 14q13, 14q32, 15q12, and 17q21), and BMI (3p25, 3q26, and 3q28). Overall, there are remarkable similarities between our linkage results after data cleaning and other independent linkage scans of similar traits. For example, six out of the eight BP-related linkage regions were confirmed by other independent genome-scan linkage studies and four regions are also validated by BP-related QTLs identified in mouse and rat models of hypertension (manuscript submitted). As linkage evidences in these replicated regions are more likely to be true positives, we use them to illustrate the impact of hidden (Mendelian-consistent) errors on linkage results.

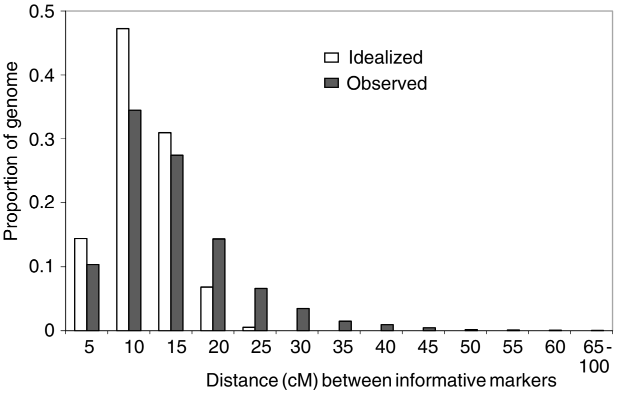

Figure 2 Idealized and observed distance between two informative markers when markers are placed approximately 10 cM apart. The average observed distance was calculated using 487 unrelated founders in the GenNet study and the Marshfield sex-averaged map5.

Do allele shifting markers affect linkage evidence?

Linkage signals including and excluding AS markers were compared to see how erroneous information from such markers effect linkage results. Chromosomes 1, 14 and 17 all demonstrated linkage (LOD $\geq$2) to BP-related phenotypes in our data and at least one other study[9–12] and have at least one AS marker. The location of the LOD and AS markers, the distance between them, and the linkage analysis results including and excluding the AS markers, are summarized in Table 2.

For chromosome 1 and 17, including AS markers caused no or minor reduction in LODs and no change in their locations (Table 2). On chromosome 14, there are two regions with evidence of linkage to PP: LOD = 2.5 at 41 cM and LOD = 3.0 at 106 cM, based on Tecumseh and Maywood samples combined. Marker Mfd190 (D14S53, 86.4 cM) had allele-shifting problem in both samples. Deleting this marker created a gap of 10.8 cM on chromosome 14. Inclusion of Mfd190 caused a global change of lod scores reducing the peak at 41 cM from 2.5 to 1.3 and the second peak at 106 cM from 3.1 to 2.3 (Figure 3). In contrast, excluding data from non-AS markers (chromosomes 1, 14 and 17: N = 55) led to a localized change without effecting lod scores elsewhere on the chromosome (data not shown). Therefore some, but not all, AS markers can quantitatively alter linkage outcome. The extent of AS, distance to the linkage region, and the percentage of the sample affected, all contribute to the effect an AS marker can have on the lod scores for the entire chromosome.

Table 2 The effect of AS markers on chromosome 1, 14 and 17

				Maximum Lod Score (LOD)		
Chr.	*Location of LOD (cM)*	*Distance between AS markers and LOD (cM)*	*Amount of data with evidence of AS*	*Excluding AS marker*	*Including AS marker*	*Excluding any other marker*[a]
1	182	51	All	3.3	3.2	3.0–3.4
14	41	45	All	2.5	1.3	2.2–2.9
	106	20	All	3.0	2.3	2.9–3.1
17	67	10	Partial	2.2	2.0	1.9–2.6

Maximum lod scores (LOD) decreased if AS markers were included in the analysis but the location of LOD did not change.
[a]The range of LODs when one non-AS marker is removed from analysis, excluding the marker at the location of LOD. Partial = When the linkage evidence came from Tecumseh and Maywood samples combined but evidence AS was detected in one sample set only.

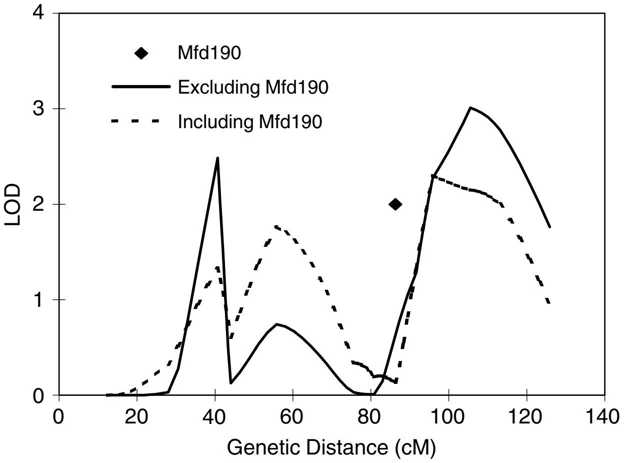

Figure 3 Chromosome 14 lod scores (Tecumseh and Loyola samples combined, phenotype = PP) including and excluding marker Mfd190. Data derived from excluding Mfd190 are in black.

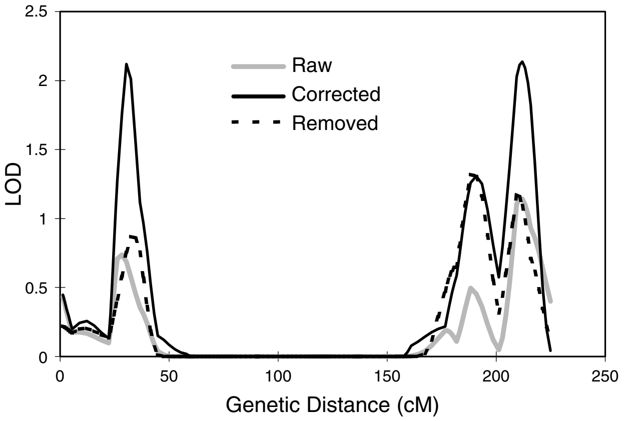

Figure 4 The impact of family structure errors on linkage (BMI, chromosome 3). The lod score plot is based on the GenNet samples from Maywood that have the most family structure errors identified. Raw = lod score using the original family structure; Corrected = family structure based on allele sharing of >300 autosomal microsatellite markers; removed = Using the original family structure but removing all genotypes involved in nonMendelian transmission of alleles.

Do family structure errors significantly reduce linkage signal?

Most of the family structure corrections were performed for Maywood samples, which, postcleaning, had evidence of linkage to BMI on chromosome 3. There were three replicated linkage peaks on chromosome 3: LOD = 2.1 at 30.4 cM; LOD = 1.3 at 191 cM (LOD = 2.1 at the same position when Maywood and Tecumseh samples were combined); and LOD = 2.2 at 212 cM.[13,14] Using the raw family structure file reduced the lod scores of all three regions (no LOD >1.5, Figure 4). Since pedigrees with family structure errors typically have non-Mendelian segregation of alleles at many markers, these pedigrees are deleted by most linkage analysis programs before calculating lod scores. However, by removing all genotypes involved in non-Mendelian allele transmission, the incorrect family structures still reduced lod scores of all three linkage peaks.

Is removal of Mendelian inconsistencies an adequate substitution for comprehensive data cleaning?

Two examples of linkage analysis (on BMI and PP phenotypes) before and after data cleaning are shown in Figure 5. For the precleaned data, the GENEHUNTER linkage program automatically deleted entire families at each marker where non-Mendelian transmission of alleles was detected. Examples of falsely elevated lod scores from incorrect estimation of IBS due to undetected errors are found on chromosome 8 and 9 (phenotype = BMI, Figure 5a). Importantly, the linkage signal on chromosomes 2 and 14, regions where linkage to BP-related phenotypes have been reported,[11,15] became more significant once detectable errors are either removed or corrected (Figure 5b).

Discussion

We reviewed 74 published genome-wide linkage scans of complex traits (hypertension, diabetes, obesity-related phenotypes and psychiatric disorders) published between 1994 and 2004 to assess to what extent data quality is routinely addressed. A significant proportion of these

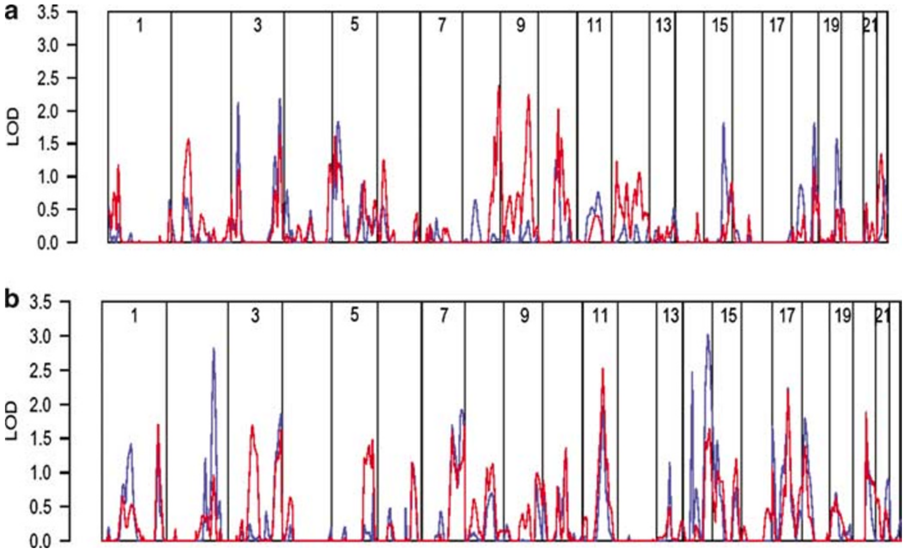

Figure 5 Genomewide pre- and postdata cleaning linkage results. (**a**) Maywood samples, body mass index (BMI). (**b**) Tecumseh and Maywood samples combined, PP. Precleaning lod scores are shown in red and postcleaning lod scores in blue.

publications did not mention data quality (39%). When it was mentioned, 40% relied solely to the occurrence of Mendelian inconsistency or double recombination between closely linked markers to identify genotyping errors and 60% did not validate the biological relationships among the samples analyzed. Importantly, not a single publication discussed post cleaning data completeness and to what extend it impacted on their linkage results. Data cleaning should be a critical part of linkage analysis, however, it is clear from the literature that this important issue is not routinely addressed.

The actual effect of data cleaning is difficult to assess because most linkage analysis programs do not run in the presence of inheritance errors. By using only genotypes that were consistent with Mendelian inheritance, we demonstrated that both AS markers and erroneous family structure can significantly reduce lod scores and alter linkage outcome. Often, in an attempt to confirm a linkage finding, more samples are genotyped, only to find that the LOD score from the combined data is less significant than the original. While there is no doubt that a significant LOD score from the original study is in part due to sampling variation, lower lod scores due to inconsistencies in allele calling should be considered before a linkage result is abandoned.

The problem of relationship error on genetic inference has been extensively studied and algorithms to determine the most likely relationships within a small family, such as parent–child, full siblings, and half siblings, are well established.[16–19] Recently, a pedigree error detection method that extends the IBS-based test and the likelihood calculation to general outbred relative pairs, including avuncular and first-cousin relationships, has been developed[20,21] and is ideal for studies that utilize large extended pedigrees.

It is well known that the accuracy of phenotypes also dramatically affects the power and the outcome of both linkage and association studies.[22,23] Ideally, estimated phenotype error rate can be derived from repeat sampling and should be modeled into power calculations and sample size estimations. For quantitative phenotypes, multiple measurements also improve power of detecting linkage by reducing phenotypic variance. In our experience, the average of multiple BP measurements gave more significant linkage results in confirmed linkage regions than individual measurements. Similarly, linkage analysis using BP from a more accurate and reproducible method gave a more significant linkage result than a less reliable method (data not shown).

Detection of genotyping errors in genetic studies depends greatly on study design (large pedigree, sib-pair and case/control) and marker heterozygosity. In all cases, even modest genotyping errors (1–2%) can significantly reduce the significance of a linkage or association result.[24–28] When family members are available, errors most likely to contribute to false recombination can be identified and removed to restore most of the lost linkage information. Theoretically, only 51–77% of the error can be detected for multiallelic markers such as STR, and these rates are even lower for biallelic markers such as SNPs.[29] Most error-detecting algorithms identify only errors that lead to blatant inheritance inconsistencies. By using a multipoint method, Mendelian-consistent genotyping errors most likely to affect linkage analysis can also be detected.[27] Four such methods (SIBMED, MENDEL, Merlin and SimWalk) were compared and reviewed recently.[30] Regardless of methods, the error-detection threshold should be decided based on marker density and family size specific to individual studies, since the default thresholds used by all four programs have relatively low power. Multipoint methods are effective only when the marker density is relative high (1–3 cM) and therefore not applicable to genome-scan data such as ours.

Potential sources of STR genotyping error have been described elsewhere.[31,32] High throughput genotyping centers report a current genotyping error rate of <1%. The Mammalian Genotyping Service (MGS, http://research.marshfieldclinics.org/genetics/) reports a rate of 0.7% while the Center for Inherited Disease Research (CIDR, http://www.cidr.jhmi.edu) reports an error rate of 0.14%. As a result of changes in primer design and evolving technologies in genotyping instruments and allele sizing software, inconsistent allele calling can occur even when genotyping is carried out in the same center over time. Inconsistent allele calling is even more serious if genotyping data generated by multiple laboratories are combined. When comparing genome scan data obtained from CIDR and MGS, most alleles are concordantly called but 18.4% of the markers show some degree of allele calling discrepancy.[33] In our case, we elected to delete all the genotypes of AS markers. Otherwise, genotypes at problematic loci can be recovered only if there are sufficient reference alleles from samples genotyped in all batches or laboratories. While consistent allele calling is important for linkage studies, it is absolutely crucial for association studies.

Both STR and SNP genotyping failure and error can cluster in a small number of markers (from lack of sufficient sequence uniqueness) and samples (due to poor DNA quality or contamination). Higher than expected missing genotypes, Mendelian errors, and deviation from Hardy–Weinberg equilibrium should alert investigators to consider deleting all genotypes associated with such samples or markers. As genotyping technology moves from STR to SNPs, and study designs from pedigree-based linkage to population-based association studies, error detection becomes an even greater challenge. While the probability of detecting genotype errors of biallelic markers in unrelated samples is low, this problem is partially remedied by the fact that large scale SNP genotyping has a significantly lower error rate than STR genotyping.[34] One area of active research focus on incorporate phenotype and genotype errors in genetic analysis.[35–40] The extent and type of phenotype and genotype errors can be estimated by repeating phenotyping and genotyping in at least a subset of samples.[41] This costly procedure is best done in the beginning of a study so the expected error rate can be modeled into power calculation to estimated. Packages such as Mega2 allow users to introduce genotyping errors to simulated data to calculate the effects of such error to both linkage and association studies.[42] PAWE-3D calculates power for association studies using multiple error parameters.[28,43] Others calculate posterior probability of genotyping error in studies based on sibling pair data (SIBMED[27]), nuclear families (MENDEL, version 5), and large pedigrees (SimWalk3)[44] so that genotypes with high probability of error can be removed or corrected by additional genotyping before analysis. Ideally, both Mendelian-consistent and -inconsistent error should be handled automatically (no data deletion/correction required) and the likelihood calculation of routine genetic analysis (such haplotype construction, test of linkage or association) will incorporate the likelihood of the reported genotypes *versus* other possible genotypes.

To achieve adequate power, genetic studies of complex traits often involve thousands of samples collected over time by multiple groups of investigators. The samples are usually processed, genotyped, and analyzed by different groups, potentially introducing heterogeneity and more error. Improvements in three areas are necessary for genetic studies of this scale and complexity. First, by simply increasing marker density in a linkage scan, the inevitable missing and erroneous genotypes can be identified with greater ease and will have less effect on the overall linkage signal. Second, improved genotyping technology can no doubt reduce genotyping errors and between-batch variation. Reliable and high-throughput SNP genotyping might replace STR markers as the choice of genome scans in the near future. For studies in progress, having all the genotyping done in one laboratory in the shortest time possible can also avoid some of the batch-to-batch variability. Finally, we strongly recommend investing in study design and infrastructure. Laboratory automation and centralized data management that minimize human errors in sample handling, data entry, as well as subsequent data transfer and storage, are of paramount importance in the ultimate success of finding genes underlying complex diseases. It is our hope that the experience discussed here will be instructive to investigators of similar studies and encourage some to take another look at their own data.

Acknowledgements

The following investigators are associated with the Family Blood Pressure Program. GenNet Network*: Alan B Weder (Network Director), Lillian Gleiberman, Anne E Kwitek, Aravinda Chakravarti, Richard S Cooper, Carolina Delgado, Howard J Jacob, and Nicholas J Schork.* GENOA Network*: Eric Boerwinkle (Network Director), Andy Brown, Christy Thielmier, Robert E Ferrell, Craig Hanis, Sharon Kardia, and Stephen Turner.* HyperGEN Network*: Steven C Hunt (Network Director), Janet Hood, Donna Arnett, John H Eckfeldt, R Curtis Ellison, Chi Gu, Gerardo Heiss, Paul Hopkins, Jean-Marc Lalouel, Mark Leppert, Albert Oberman, Michael A Province, DC Rao, Treva Rice, and Robert Weiss* – SAPPHIRe Network*: David Curb*

(Network Director), David Cox, Timothy Donlon, Victor Dzau, John Grove, Kamal Masaki, Richard Myers, Richard Olshen, Richard Pratt, Tom Quertermous, Neil Risch and Beatriz Rodriguez. National Heart, Lung, and Blood Institute: Dina Paltoo and Cashell E Jaquish. Website: www.sph.uth.tme.edu/hgc/fbpp.

References

1 Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M: Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 2001; **69**: 936–950.

2 Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* 2002; **39**: 3–9.

3 DeWan AT, Parrado AR, Matise TC, Leal SM: The map problem: a comparison of genetic and sequence-based physical maps. *Am J Hum Genet* 2002; **70**: 101–107.

4 Daw EW, Thompson EA, Wijsman EM: Bias in multipoint linkage analysis arising from map misspecification. *Genet Epidemiol* 2000; **19**: 366–380.

5 Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 1998; **63**: 861–869.

6 Ghosh S, Karanjawala ZE, Hauser ER *et al.*: Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group. *Genome Res* 1997; **7**: 165–178.

7 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 1996; **58**: 1347–1363.

8 Thiel BA, Chakravarti A, Cooper RS *et al.*: A genome-wide linkage analysis investigating the determinants of blood pressure in whites and African Americans. *Am J Hypertens* 2003; **16**: 151–153.

9 Hunt SC, Ellison RC, Atwood LD, Pankow JS, Province MA, Leppert MF: Genome scans for blood pressure and hypertension: the National Heart, Lung, and Blood Institute Family Heart Study. *Hypertension* 2002; **40**: 1–6.

10 Perola M, Kainulainen K, Pajukanta P *et al.*: Genome-wide scan of predisposing loci for increased diastolic blood pressure in Finnish siblings. *J Hypertens* 2000; **18**: 1579–1585.

11 Von Wowern F, Bengtsson K, Lindgren CM *et al.*: A genome wide scan for early onset primary hypertension in Scandinavians. *Hum Mol Genet* 2003; **12**: 2077–2081.

12 Levy D, DeStefano AL, Larson MG *et al.*: Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the framingham heart study. *Hypertension* 2000; **36**: 477–483.

13 Hsueh WC, Mitchell BD, Schneider JL *et al.*: Genome-wide scan of obesity in the Old Order Amish. *J Clin Endocrinol Metab* 2001; **86**: 1199–1205.

14 Wu X, Cooper RS, Borecki I *et al.*: A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program. *Am J Hum Genet* 2002; **70**: 1247–1256.

15 Hsueh WC, Mitchell BD, Schneider JL *et al.*: QTL influencing blood pressure maps to the region of PPH1 on chromosome 2q31–34 in Old Order Amish. *Circulation* 2000; **101**: 2810–2816.

16 Goring HH, Ott J: Relationship estimation in affected sib pair analysis of late-onset diseases. *Eur J Hum Genet* 1997; **5**: 69–77.

17 Boehnke M, Cox NJ: Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet* 1997; **61**: 423–429.

18 Ehm M, Wagner M: A test statistic to detect errors in sib-pair relationships. *Am J Hum Genet* 1998; **62**: 181–188.

19 O'Connell JR, Weeks DE: PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet* 1998; **63**: 259–266.

20 McPeek MS, Sun L: Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet* 2000; **66**: 1076–1094.

21 Sun L, Wilder K, McPeek MS: Enhanced pedigree error detection. *Hum Hered* 2002; **54**: 99–110.

22 Martinez M, Khlat M, Leboyer M, Clerget-Darpoux F: Performance of linkage analysis under misclassification error when the genetic model is unknown. *Genet Epidemiol* 1989; **6**: 253–258.

23 Zheng G, Tian X: The impact of diagnostic error on testing genetic association in case-control studies. *Stat Med* 2005; **24**: 869–882.

24 Buetow KH: Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am J Hum Genet* 1991; **49**: 985–994.

25 Abecasis GR, Cherny SS, Cardon LR: The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet* 2001; **9**: 130–134.

26 Goldstein DR, Zhao H, Speed TP: The effects of genotyping errors and interference on estimation of genetic distance. *Hum Hered* 1997; **47**: 86–100.

27 Douglas JA, Boehnke M, Lange K: A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet* 2000; **66**: 1287–1297.

28 Gordon D, Finch SJ, Nothnagel M, Ott J: Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered* 2002; **54**: 22–33.

29 Douglas JA, Skol AD, Boehnke M: Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet* 2002; **70**: 487–495.

30 Mukhopadhyay N, Buxbaum SG, Weeks DE: Comparative study of multipoint methods for genotype error detection. *Hum Hered* 2004; **58**: 175–189.

31 Ewen KR, Bahlo M, Treloar SA *et al.*: Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet* 2000; **67**: 727–736.

32 Weber JL, Broman KW: Genotyping for human whole-genome scans: past, present, and future. *Adv Genet* 2001; **42**: 77–96.

33 Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB: A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Res* 2002; **12**: 430–435.

34 Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P: International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.

35 Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D: Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genet* 2005; **6**: 18.

36 Goring HH, Terwilliger JD: Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet* 2000; **66**: 1107–1118.

37 Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 2001; **69**: 371–380.

38 Bernardinelli L, Berzuini C, Seaman S, Holmans P: Bayesian trio models for association in the presence of genotyping errors. *Genet Epidemiol* 2004; **26**: 70–80.

39 Morris RW, Kaplan NL: Testing for association with a case-parents design in the presence of genotyping errors. *Genet Epidemiol* 2004; **26**: 142–154.

40 Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J: A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur J Hum Genet* 2004; **12**: 752–761.

41 Gordon D, Yang Y, Haynes C *et al.*: Increasing power for tests of genetic association in the presence of phenotype and/or

genotype error by use of double-sampling. *Stat Appl Genet Mol Biol* 2004; **3** (1), article 26.

42 Mukhopadhyay N, Almasy L, Schroeder M, Mulvihill WP, Weeks DE: Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics* 2005; **21**: 2556–2557.

43 Gordon D, Haynes C, Blumenfeld J, Finch SJ: PAWE-3D: visualizing power for association with error in case/control genetic studies of complex traits. *Bioinformatics* 2005.

44 Sobel E, Papp JC, Lange K: Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet* 2002; **70**: 496–508.

Supplementary information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)