Genetic Epidemiology 29(Supplement 1): S120–S124 (2005)

Genotyping Errors, Pedigree Errors, and Missing Data

Anthony L. Hinrichs[1]* and Brian K. Suarez[1,2]

[1]*Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri*
[2]*Department of Genetics, Washington University School of Medicine, St. Louis, Missouri*

Our group studied the effects of genotyping errors, pedigree errors, and missing data on a wide range of techniques, with a focus on the role of single-nucleotide polymorphisms (SNPs). Half of our group used simulated data, and half of our group used data from the Collaborative Study on the Genetics of Alcoholism (COGA). The simulated data had no missing genotypes and no genotyping errors, so our group, as a whole, removed data and introduced artificial errors to study the robustness of various techniques. Our teams showed that genotyping errors are less detectable and may have a greater impact on SNPs than on microsatellites, but recently developed methods that account for genotyping errors help reduce false positives, and the assumptions of these methods appear to be supported by observations from repeated genotyping. The ability to detect linkage disequilibrium (LD) was also substantially reduced by missing data; this in turn could affect tagging SNPs chosen to generate haplotypes. In the COGA sample, genotyping measurements were repeated in three ways. First, full-genome screens were performed on three sets of markers: 328 microsatellites, 11,560 SNPs from the Affymetrix GeneChip Mapping 10 K Array marker set, and 4,720 SNPs from the Illumina Linkage III panel. Second, the entire Affymetrix marker set was typed on the same 184 individuals by two different laboratories. Finally, the Affymetrix and Illumina marker panels had 94 SNPs in common. Our teams showed that both SNPs and microsatellites can be readily used to identify pedigree errors, and that SNPs have fewer genotyping errors and a low inconsistency rate. However, a fairly high rate of no-calls, especially for the Affymetrix platform, suggests that the inconsistency rate may be higher than observed. *Genet. Epidemiol.* 29(Suppl. 1):S120–S124, 2005. © 2005 Wiley-Liss, Inc.

Key words: missing genotypes; error detection; SNPs; type I error; power

*Correspondence to: Anthony L. Hinrichs, Department of Psychiatry, Washington University School of Medicine, Campus Box 8134, 660 S. Euclid, St. Louis, MO 63110. E-mail: tony@silver.wustl.edu
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20120

INTRODUCTION

Single-nucleotide polymorphisms (SNPs) have typically been used for association studies, but have more recently been considered for linkage studies. Genetic Analysis Workshop (GAW) 14 presented participants with the opportunity to analyze real and simulated data sets in order to shed light on this topic. The real data set, taken from the Collaborative Study of the Genetics of Alcoholism (COGA), provided extended pedigrees with full-genome scans performed with three markers sets: microsatellites, the Illumina Linkage III panel, and the Affymetrix 10 K Mapping Array. The SNP genotyping was facilitated and partially duplicated by the Center for Inherited Disease Research (CIDR). The simulated data set provided several different populations genotyped for both microsatellite markers and SNPs. As a whole, GAW14 investigated the extension to SNPs of many techniques traditionally applied to microsatellites. The diallelic nature of SNPs and the inherent linkage disequilibrium (LD) found in dense marker maps complicated many of these techniques. Our group studied the effects of genotyping errors, pedigree errors, and missing data on a wide range of techniques, with a focus on the role of SNPs. The diallelic nature of SNPs can be especially difficult in this area because of the reduced information from each marker considered individually. For example, given a single highly polymorphic microsatellite in a large sibship even without genotyped parents, one has a chance of identifying genotyping errors and (since there may be four different alleles present in the sibship) directly determining identity-by-descent (IBD) status. For a single SNP, on the other hand, for any sibship without genotyped parents, no genotyping errors can be detected because two heterozygous parents would

be completely compatible with any observed genotypes. Because only two alleles occur, IBD status is estimated with less accuracy. This increased susceptibility to undetectable genotyping error and reduced information content make SNPs an interesting challenge.

METHODS

SIMULATED DATA

Half of our group used the simulated data [Barral et al., 2005; McCaskie et al., 2005; Thompson et al., 2005]. Because the simulated data had no missing values and no simulated genotyping errors, the authors modified the data by removing genotypes and randomly changing genotypes to simulate errors. They then examined the impact of the missing data and genotyping errors, and methods to detect or correct for the errors, for the transmission disequilibrium test (TDT), LD detection, and linkage analyses.

Barral et al. [2005] modified the GAW14 simulated data (SNP genotypes only) to introduce errors under the Sobel-Papp-Lange model (SPL) [Sobel et al., 2002]. Under this model, for each of the three possible genotypes of an SNP, one specifies the probabilities of observing each of the coded genotypes. Values used are presented in Table I. Note the decreased probability of observing a 2/2 genotype, given a true 1/1 genotype (and vice versa), in an attempt to simulate more realistic genotyping errors than a simple uniform probability. Barral et al. [2005] also removed 10% of the parental genotypes independent of all other variables, i.e., genotypes were missing completely at random (MCAR) [Little and Rubin, 1987]. After introducing these errors and removing data, Barral et al. [2005] compared the traditional TDT to a version of the TDT which allows for errors (TDTae) [Gordon et al., 2001, 2004]. The TDTae performs a likelihood-based TDT in which chance of genotyping error is incorporated into the likelihood calculation. The method assumes that genotyping errors are random and independent, and that the genotypes have not been cleaned to remove Mendelian inconsistencies.

TABLE I. Sobel-Papp-Lange error model penetrance values [Barral et al., 2005]

	Observed genotype		
True genotype	1/1	1/2	2/2
1/1	0.989	0.01	0.001
1/2	0.01	0.98	0.01
2/2	0.001	0.01	0.989

Similarly, McCaskie et al. [2005] deleted genotype values (MCAR) from the Aipotu data set at rates of 1%, 5%, and 10%. They applied their new LD plotting program, JLIN [Carter et al., 2004], to study the effect of missing data on computation of the disequilibrium coefficients D′. They also examined the role of missing data in haplotype association analyses, using their SIMHAP software [McCaskie et al., 2004]; this software uses an expectation-maximization algorithm to impute diplotypes (i.e., a pair of haplotypes: the haplotype equivalent of genotype), and then simulates multiple data sets to determine the empirical distribution of parameter estimates. Using the simulated data set with missing data, analyses were performed with tagged SNPs that had been shown to be associated with affection status. Coefficients (with significance values), as well as means and 95% confidence intervals, were extracted from 10,000 simulations.

Thompson et al. [2005] assessed the ability to detect genotyping errors in sibships without genotyped parents. They used the three simulated nuclear family populations (Aipotu, Danacaa, and Karangar) and removed all parental genotyping data. Random genotyping was simulated at error rates of 0.14% and 2.8%. For their analyses, genotypes were chosen randomly at the specified error rate, and one of the two alleles was selected at random: in the case of SNPs, the chosen allele was replaced with the other allele; in the case of microsatellites, the chosen allele was replaced with an adjacently sized one (either one more or one less repeat). Although different from the SPL error model, this attempts to mimic laboratory conditions, and disallows transitions from one homozygote to the other. As previously mentioned, no genotyping errors are detectable with a single SNP in this circumstance, because all possibilities are consistent with Mendelian inheritance. However, microsatellites were screened using each marker individually and testing for Mendelian inheritance. Although double recombinants (a recombination occurring immediately before and after a marker) are possible, due to their scarcity, a search for double recombinants is also often used to identify genotyping errors. The authors used GENIBD, part of the SAGE package [Statistical Solutions, 2004], to compute the change in sharing probabilities both before and after each

locus in turn. If the change in probabilities was high (above a predetermined threshold) on both sides of the marker for two or more sib pairs that included the same individual, then the marker was deemed to be the site of a double recombinant and a potential genotyping error. By this definition, it would be impossible to detect genotyping errors for the first or last marker on a chromosome or for a sibship containing only a single sib pair. False-positive and false-negative rates of genotype error detection were computed for the microsatellite and SNP markers. The Shannon information content (SIC) was computed to evaluate error rates as a function of the SIC.

Thompson et al. [2005] also examined power and type I error rates for microsatellite and SNP markers. They computed power to detect a signal exceeding a predefined threshold within 20 cM of the true location of a disease gene. Type I errors were considered to be signals above a given threshold at least 40 cM from a true location of a disease gene, and more than 20 cM from any other signal (to allow for multiple false positives on a chromosome). They also examined the power and type I error rate under reduced genotypic penetrances, to simulate mistyping.

COGA DATA

Half of our group used the real COGA data. Suarez et al. [2005] and Tintle et al. [2005] examined replicated genotyping for the SNP genotypes. Wang et al. [2005] compared pedigree errors detected by SNPs and by microsatellites.

Suarez et al. [2005] identified 94 SNPs typed on all individuals, using both the Illumina and Affymetrix platforms. The authors studied variations in no-call rates and patterns of discrepancies (where both platforms reported a genotype, but the reported genotypes differed).

Tintle et al. [2005] examined a set of individuals genotyped with the same platform but by two different laboratories. In particular, CIDR and Affymetrix genotyped the same set of 11,560 SNPs on 184 individuals using the Affymetrix GeneChip Mapping 10 K Array. The authors distinguished between inconsistency (where two genotypes for a particular SNP and subject exist and are different) and nonreplication (two genotypes for a particular SNP and subject exist and are different, or one of the two genotypes is missing). Of the 11,560 SNPs, 440 SNPs were dropped from the analysis because they were not included in the final map information. Five of 184 subjects were dropped: two with the same CIDR identification numbers, and three genotyped only on a subset of the SNPs.

Wang et al. [2005] examined the role of SNPs in the identification of pedigree errors. They first selected 239 unrelated founders of white, non-Hispanic pedigrees to test the neutrality of the autosomal markers, and 160 unrelated female founders of white, non-Hispanic pedigrees to test the neutrality of the X-chromosome markers. They identified outlier loci-markers with deviation from Hardy-Weinberg equilibrium under either balancing selection (with lower than expected homozygosity values) or directional selection (with higher than expected homozygosity values). They divided the data set into 21 pedigrees with the majority of individuals of self-reported black ethnicity (both Hispanic and non-Hispanic), and 122 pedigrees with the majority of individuals of self-reported white ethnicity (again, Hispanic and non-Hispanic). They ran PREST [McPeek and Sun, 2000] on these subsets separately for all three marker sets, first including and then excluding the outlier loci.

RESULTS

SIMULATED DATA

Barral et al. [2005] demonstrated a dramatic inflation of the false-positive rate for traditional TDT in the presence of undetected genotyping errors and missing parental genotypes. They further demonstrated increased precision using TDTae compared to traditional TDT, i.e., the distance between a disease gene position and markers reaching a given significance level is much smaller with TDTae.

McCaskie et al. [2005] found that missing data might increase the amount of strong LD observed in the data. The pattern of LD across a chromosome could thus become more fractionated, causing the partitioning of haplotype blocks into smaller blocks, affecting tag SNP selection and haplotype formation. For haplotypic analyses, they observed a trend toward increased 95% confidence intervals for both odds ratios and *P*-values as the amount of missingness increased, i.e., average values of odds ratios and *P*-values became less precise. However, there were no obvious differences in the odds ratios and *P*-values themselves for the haplotypes.

Thompson et al. [2005] found that, as expected, none of the genotyping errors in SNPs were

detected using a single-marker test of Mendelian inheritance, and that 35.6% of the genotyping errors in microsatellites were detected using a single-marker test. The use of large changes in estimated IBD (EIBD) to identify genotyping errors performed poorly, with high false-positive rates at low thresholds, and low true-positive rates at higher thresholds. Under the most rigorous condition (delta = 0.9, i.e., a change of EIBD by at least 0.9 before and after the marker in question), 53.1% of genotyping errors were detected for the microsatellites, whereas only 2.4% of genotyping errors were detected for the SNPs. The authors found that even after adjusting for different levels of SIC within a pedigree, genotyping errors were still easier to detect with microsatellites than SNPs.

Thompson et al. [2005] also showed that for all levels of genotyping error, SNPs were more powerful for linkage than microsatellites (spaced at an average distance of 10.5 cM), but also had a higher false-positive rate in the presence of genotyping errors compared to microsatellites. They suggested that different thresholds may be useful for declaring significant evidence of linkage for the two different types of markers. Because SNP genotyping error rates may be lower than microsatellites, the authors compared a 0.14% SNP error rate to a 2.8% microsatellite error rate. In this case, the SNPs were still slightly more powerful for linkage, but had a false-positive rate similar to the microsatellites. Finally, the authors suggested that allowing for genotyping errors in linkage analyses will result in a greater increase in power for microsatellites than for SNPs.

COGA DATA

The real data were shown to have errors by multiple techniques in multiple ways. However, given the lack of an ''infallible'' data source, the nature and localization of the errors were somewhat speculative.

After adjusting for differences in allelic designation between the Illumina and Affymetrix platforms, Suarez et al. [2005] observed no cases in which a homozygote observed under one platform was scored as the opposite homozygote with the other platform. They also found that although the concordance rate was high (99.85%), there was a substantially higher no-call rate with the Affymetrix platform that appeared to be genotype- and SNP-specific. The SNP with the largest number of differences (rs958883) was near an *Xba*I site that also contained an SNP (rs17150546). Because the Affymetrix technology uses the *Xba*I restriction enzyme to produce fragments of 250–1,000 nucleotides, it may be that individuals with the polymorphism in the *Xba*I site will be missing the allele typed at the rs958883 SNP. Moreover, because all discrepant individuals with the Affymetrix platform typed as ''22'' homozygotes, it is reasonable to hypothesize that the SNP that obliterates the *Xba*I site is in strong LD with the ''1'' allele at rs958883.

Tintle et al. [2005] found a low inconsistency rate (0.2%), but a high nonreplication rate (9.5%). As with Suarez et al. [2005], their evidence supports the hypothesis that errors incorrectly reporting a homozygote as the opposite homozygote are rare. They found this quite encouraging, because recently developed methods rely on this assumption. Analysis of missing data suggests that some may have dependencies, while other missing data appear to be independent. Dependent missing data suggest that particular individuals or SNPs may be difficult to classify. Independent missing data suggest the presence of no-call regions, which were recently shown to have no value in association tests [Kang et al., 2004].

Wang et al. [2005] found that outlier SNPs have little impact on the identification of pedigree errors. Furthermore, there were several clear pedigree errors present in the COGA data set. Although some of these pedigree errors were detected in all three marker sets, some errors were only detected in a single set. This suggests possible sample swaps occurring only in one of the three genotyping laboratories.

DISCUSSION

There are three themes throughout the papers in this group. First, our group found that genotyping errors are less detectable and may have a greater impact for SNPs than for microsatellites. In the presence of genotyping error, one team found that single- and multiple-locus tests for Mendelian inheritance, when parents were missing, detected substantially fewer errors for SNPs than for microsatellites (2.4% vs. 53.1%). Another team found that the power to detect a true linkage signal was greater for SNP (75%) than microsatellite (67%) marker maps, although there were also slightly more false-positive signals using SNP marker maps (five compared with three). The ability to detect LD was also substantially reduced

by missing data; this could in turn affect tagging SNPs chosen to generate haplotypes. Second, our group found that recently developed methods to account for genotyping errors helped reduce false positives, and the assumptions of these methods appear to be supported by observations from repeated genotyping. In a comparison of TDT and TDTae, TDT showed substantially increased type I error (rates of 28.8%, 14.8%, 5.4%, and 1.7% at the 5%, 1%, 0.1%, and 0.01% significance levels, respectively), while TDTae maintained the correct false-positive rate. TDTae also showed an increased ability to localize disease loci. Third, although SNPs appear to have fewer genotyping errors than microsatellites and can be readily used for detecting pedigree errors, the Affymetrix platform appears to suffer from a high no-call rate that is concentrated in a subset of SNPs that are difficult to genotype. A search for pedigree structure errors using three marker sets separately identified 15 errors in 143 pedigrees. However, some of these errors appeared only with one set of markers, indicating a potential sample mix-up in one genotyping laboratory. A comparison of genotyping performed by both CIDR and Affymetrix showed that while the inconsistency rate (two different genotypes for the same subject) was low (0.2%), the nonreplication rate (two different genotypes for the same subject, or one identified genotype and one missing genotype) was substantial (9.5%). This also suggests that the actual inconsistency rate is higher than reported and may have a significant impact on power. Analysis of 94 SNPs common to both platforms showed significant agreement when both platforms made a call (99.85%); however, the no-call rate for the Affymetrix platform was approximately 8.6 times higher than for the Illumina platform. When genotypes were inconsistent, the number of inferred recombinants for Affymetrix genotypes was substantially higher compared to Illumina genotypes. Finally, for at least two SNPs, familial clustering of inconsistency may have been due to the presence of a second segregating SNP that obliterated an *Xba*I site (the restriction enzyme used in the Affymetrix platform), resulting in a fragment too long (>1,000 bp) to be amplified. Taken together, these themes suggest that although SNPs may eventually replace the role of microsatellites for many applications, they also present new problems for missing data and genotyping errors.

REFERENCES

Barral S, Haynes C, Levenstien MA, Gordon D. 2005. Precision and type I error rate in the presence of genotype errors and missing parental data: a comparison between the original transmission disequlibrium test (TDT) and TDTae statistics. BMC Genet [Suppl] 6:150.

Carter KW, McCaskie PA, Palmer LJ. 2004. JLIN: a Java based linkage disequilibrium plotter [http://www.genepi.com.au/projects/jlin].

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

Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J. 2004. 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 12:752–761.

Kang SJ, Gordon D, Brown AM, Ott J, Finch SJ. 2004. Tradeoff between no-call reduction in genotyping error rate and loss of sample size for genetic case/control association studies. Pacif Symp Biocomput 2004:116–127.

Little RJA, Rubin DB. 1987. Statistical analysis with missing data. New York: John Wiley.

McCaskie PA, Carter KW, McCaskie SR, Palmer LJ. 2004. SimHap: a simulation approach to haplotype analysis for population data [http://www.genepi.com.au/projects/simhap].

McCaskie PA, Carter KW, McCaskie SR, Palmer LJ. 2005. The effect of missing data on linkage disequilibrium mapping and haplotype association analysis in the GAW14 simulated datasets. BMC Genet [Suppl] 6:151.

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

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

Statistical Solutions. 2004. S.A.G.E. Statistical analysis for genetic epidemiology. Cork, Ireland: Statistical Solutions.

Suarez BK, Taylor C, Bertelsen S, Bierut LJ, Dunn G, Jin CH, Kauwe JSK, Paterson AD, Hinrichs AL. 2005. An analysis of identical single-nucleotide polymorphisms genotyped by two different platforms. BMC Genet [Suppl] 6:152.

Thompson CL, Baechle D, Lu Q, Mathew G, Song Y, Iyengar SK, Gray-McGuire C, Goddard KAB. 2005. Effects of genotyping error in model-free linkage analysis using microsatellite or single-nucleotide polymorphism marker maps. BMC Genet [Suppl] 6:153.

Tintle NL, Ahn K, Mendell NR, Gordon D, Finch SJ. 2005. Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and Center for Inherited Diseases Research. BMC Genet [Suppl] 6:154.

Wang K-S, Liu M, Paterson AD. 2005. Evaluating outlier loci and their effect on the identification of pedigree errors. BMC Genet [Suppl] 6:155.