European Journal of Human Genetics (2004) 12, 395–399
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $25.00
www.nature.com/ejhg

ARTICLE

Detection of genotyping errors by Hardy–Weinberg equilibrium testing

Louise Hosking*,[1], Sheena Lumsden[1], Karen Lewis[1], Astrid Yeo[2], Linda McCarthy[1], Aruna Bansal[2], John Riley[1], Ian Purvis[1] and Chun-Fang Xu[1]

[1]*GlaxoSmithKline Medicines Research Centre, Gunnels Wood Rd, Stevenage, Hertfordshire SG1 2NY, UK;*
[2]*GlaxoSmithKline Medicines Research Centre, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, UK*

Genotyping data sets may contain errors that, in some instances, lead to false conclusions. Deviation from Hardy–Weinberg equilibrium (HWE) in random samples may be indicative of problematic assays. This study has analysed 107 000 genotypes generated by TaqMan, RFLP, sequencing or mass spectrometric methods from 443 single-nucleotide polymorphisms (SNPs). These SNPs are distributed both within genes and in intergenic regions. Genotype distributions for 36 out of 313 assays (11.5%) whose minor allele frequencies were >0.05 deviated from HWE ($P<0.05$). Some of the possible reasons for this deviation were explored: assays for five SNPs proved nonspecific, and genotyping errors were identified in 21 SNPs. For the remaining 10 SNPs, no reasons for deviation from HWE were identified. We demonstrate the successful identification of a proportion of nonspecific assays, and assays harbouring genotyping error. Consequently, our current high-throughput genotyping system incorporates tests for both assay specificity and deviation from HWE, to minimise the genotype error rate and therefore improve data quality.

European Journal of Human Genetics (2004) **12,** 395–399. doi:10.1038/sj.ejhg.5201164
Published online 11 February 2004

Keywords: Hardy–Weinberg equilibrium; genotyping; error

Introduction

As a consequence of the generation of large numbers of genotypes in both family- and population-based genetic studies, much work is currently focusing on the identification and possible integration of error within such data. Error can be introduced into genetic data sets from a variety of sources, which include inconsistencies within family pedigrees,[1,2] sample mishandling,[3] and errors introduced by the genotyping process itself.[4]

Inclusion of incorrect data in genetic analysis can lead to the generation of false conclusions[5,6] and a reduction of power to fine map trait loci,[7,8] and is a recognised problem in the statistical analysis of genetic data sets.[9–13] Previous simulation studies have considered the impact of genotyping errors in data sets generated from pedigrees;[5,10] even genotyping error rates as low as 1–2% can affect both linkage and sib-pair studies.[5] Within family studies, incorrect genotypes may inflate map distance between markers and also reduce the power to detect linkage,[7,14–16] and contribute to an inflated false positive rate among transmission disequilibrium test (TDT)-derived associations.[17]

Within family studies, a proportion of genotyping errors can be detected by incorporating checks for consistency with Mendelian inheritance.[2,9,18] Checking for Mendelian inconsistency will not, however, exclude all genotyping errors, and cannot be applied to population-based studies. The presence of errors within genotyping data sets generated from unrelated individuals has considerable impact on subsequent data analysis, as no checks for Mendelian consistency are possible within such data sets.[19] It has also been demonstrated in a simulation study that

*Correspondence: Dr L Hosking, GlaxoSmithKline Medicines Research Centre, Gunnels Wood Rd, Stevenage, Hertfordshire SG1 2NY, UK.
Tel: +44 1438 764364; Fax: +44 1438 768097;
E-mail: louise.k.hosking@gsk.com

Received 28 August 2003; revised 27 November 2003; accepted 9 December 2003

genotyping error rates as low as 3% can adversely affect linkage disequilibrium (LD) measures.[20] This could limit attempts to identify complex disease genes, because it has been demonstrated that genotyping errors always decrease the power of certain statistical tests for linkage and/or association. For example, the χ^2 test of independence applied to case:control data always loses power in the presence of genotyping errors.[8,21,22]

Statistical tools that are able to take error into account have been developed. The majority of these models are applicable to linkage studies[23–26] or TDTs.[17,27] Genotyping error within data sets generated from unrelated individuals is also currently being addressed within certain statistical models.[19,28]

This study used the Hardy–Weinberg equilibrium (HWE) test to identify genotyping error within population-based data sets. In large enough randomly mating populations, not subject to genetic and population parameters affecting allele frequencies, the genotypes for an individual marker should distribute according to the principle of HWE.[29] Technical reasons, such as assay nonspecificity and genotyping errors, can also impact on the distribution of genotypes for any one marker. These technical reasons for distribution deviation were explored in this study. As a result, an improved genotyping process was implemented to reduce the occurrence of genotype errors.

Subjects and methods

DNA samples

A total of 2750 Caucasian samples have been employed in this analysis. A total of 2008 of these samples have already been reported elsewhere.[30–32] In addition, 588 samples were collated from North European Caucasians within GlaxoSmithKline with consent for nonidentified genotyping. In all, 92 Caucasian DNA samples were collected from North America with informed consent for nonidentified genotyping, and 62 Caucasian DNA samples were purchased from Coriell cell repositories (Camden, New Jersey, USA) (Table 1).

Genotyping

A total of 443 single-nucleotide polymorphisms (SNPs) were typed using different sets of the DNA samples and different methodologies (Table 1) generating 107 068 genotypes.

Table 1 Number of SNPs and samples used for each of the four technologies

Genotyping methodology	*Number of SNPs*	*Number of samples*
Taqman	96	162–590
Maldi-tof	38	1018
RFLP	10	590
Sequencing	299	62–92

Determination of deviation of genotype distribution from HWE

Minor allele frequencies were recorded for all of the 443 SNPs (Table 2). A subset of 313 SNPs, whose minor allele frequencies were >5%, was analysed for deviation of genotype distribution from HWE, using the χ^2 statistic with one degree of freedom. SNPs whose genotype distribution deviated from HWE ($P<0.05$) were identified (Table 3), and are also referred to as HWD SNPs.

Assay specificity

Nucleotide homology searches were performed on the primer and probe sequences defining reaction specificity for each of the SNPs deviating from HWE ($P<0.05$), using the NRNUC and NRHTG (Human Genome from EMBL, Sanger Centre and Washington University) databases in March 2002.

Results

In all, 107068 genotypes were generated from 443 SNPs, with minor allele frequencies ranging between 0.002 and 0.49 (Table 2). Of the 443 SNPs, 81% (353/443) were distributed throughout the genome. Of the remaining 90 SNPs, 38 mapped to a 400 kb, region on chromosome 22,[32] 24 to a region on chromosome 19,[30] and 28 to a region on chromosome 3.[31] A subset of 313 SNPs, whose minor allele frequencies were >5%, was selected for the estimation of deviation from HWE. A total of 36 SNPs (11.5%), with minor allele frequencies ranging between 0.06 and 0.49, were found to deviate from HWE ($P<0.05$) (Table 3). Of the 36 SNPs demonstrating HWD, 20 displayed deviation from HWE at the $P<0.01$ level (Table 3). Controlling the false discovery rate,[33] 16 of them would be considered significant at the 5% level.

Possible explanations for SNPs that showed deviation from HWE were explored. An SNP assay was classified as 'nonspecific' if a primer and/or probe set showed 100% homology with multiple regions in the genome. Five of the 36 SNPs (14%) were found to have 'nonspecific' assays.

Table 2 Distribution of minor allele frequencies of 443 SNPs

Minor allele frequency	*Number of SNPs*	*%*
0.002–0.01	23	5
0.01–0.05	107	24
0.05–0.10	67	15
0.10–0.20	74	17
0.20–0.30	70	16
0.30–0.40	47	11
0.40–0.50	55	12

Table 3 SNPs (36) exhibiting HWD ($P<0.05$)

SNP ID	*Minor allele frequency*	*Number of samples*	*Genotyping technology*
1[a,b]	0.18	1018	Maldi-tof
2[a,b]	0.33	1018	Maldi-tof
3[a,b]	0.26	1018	Maldi-tof
4[a,b]	0.24	1018	Maldi-tof
5[a,b]	0.27	590	RFLP
6[a,b]	0.10	590	Taqman
7[a,b]	0.21	400	Taqman
8	0.10	602	Taqman
9	0.34	162	Taqman
10	0.41	162	Taqman
11[a,b]	0.45	380	Taqman
12	0.42	368	Taqman
13[a]	0.46	380	Taqman
14[a]	0.12	62	Sequencing
15[a]	0.08	62	Sequencing
16[a,b]	0.36	62	Sequencing
17[a,b]	0.11	62	Sequencing
18	0.17	62	Sequencing
19	0.21	62	Sequencing
20	0.23	62	Sequencing
21[a,b]	0.29	62	Sequencing
22	0.27	62	Sequencing
23	0.27	62	Sequencing
24	0.27	62	Sequencing
25[a,b]	0.21	62	Sequencing
26[a,b]	0.06	94	Sequencing
27[a,b]	0.09	94	Sequencing
28	0.26	62	Sequencing
29	0.49	62	Sequencing
30	0.47	62	Sequencing
31	0.10	62	Sequencing
32[a]	0.13	62	Sequencing
33[a,b]	0.13	62	Sequencing
34	0.28	62	Sequencing
35[a,b]	0.09	62	Sequencing
36	0.38	62	Sequencing

[a] $P<0.01$. [b] $P<0.05$ controlling the false discovery rate.

Genotyping errors were identified in 21 assays, accounting for 58% of the SNPs showing deviation from HWE. For the remaining 10 SNPs (28%), no reasons for the observed HWD were identified. When analysing assays that deviated from HWE at the $P<0.01$ level (Table 4), the percentage of SNPs associated with genotyping errors was slightly lower, and the proportion associated with nonspecificity slightly higher, than all HWD ($P<0.05$) assays.

For the 21 assays where deviation of genotype distribution from HWE ($P<0.05$) was due to genotyping error (Table 4), the data sets were stratified according to each genotyping technology used (Table 5), and according to the level of deviation from HWE ($0.01<P<0.05$, or $P<0.01$). Sources of error appeared to be dependent, at least in part, on the methodology used to type the SNP. One type of error seen in SNPs analysed by directly sequencing PCR products was the inability to distinguish accurately genotypes if the background signal was too high on the sequence trace. An error frequently associated with data generated using Taqman methodology was the inaccurate calling of individual genotypes if those individual genotypes fell between the three main genotype clusters.

Table 4 Identifiable reasons for deviation from HWE in 36 SNPs

Deviation from HWE	*Reason for deviation*	*Number of SNPs*	*%*
$P<0.05$	Genotyping error	21	58
	Nonspecific	5	14
	Unknown	10	28
	Total	36	
$P<0.01$	Genotyping error	9	45
	Nonspecific	4	20
	Unknown	7	35
	Total	20	

In general, the proportion of assays associated with genotyping error is slightly lower (2.9%) in the $P<0.01$ assays than in the $0.01< P<0.05$ assays (3.8%), although following this stratification (Table 5), the numbers of assays studied are low. A greater proportion of RFLP assays appear to harbour genotyping error, but the number of assays studied (10) was low.

Discussion

Generation and analysis of large SNP genotyping data sets for the investigation of human complex disorders is currently the subject of much discussion, and focus for activity.[34] Large genotyping data sets will inevitably contain some error, which has long been recognised as a problem in the accurate statistical analysis of genetic data.[1,9] As genotyping errors are known specifically to affect certain genetic measurements such as LD, upon which association studies depend,[20] and also to affect family-based studies of linkage and association,[17] the identification of error is critical to accurate analysis and subsequent interpretation of the data. Current interest also surrounds the development of statistical methods that are able to take error into account in genetic data analysis.[17,19,23–26] This large, empirical study reports the measurement of genotype distribution deviation from HWE as a method to identify and reduce genotyping errors generated as a result of the genotyping process itself in population-based studies.

The study measured deviation from HWE in 313 SNPs and revealed 36 HWD ($P<0.05$) assays, which is ~2.3 × more than expected by chance. When considering these data, it is important to remember that the sensitivity of measurement of deviation from HWE will also depend on the minor allele frequencies of the SNPs typed (0.06–0.49), and the number of samples analysed (62–1018). Further investigation of the 36 HWD ($P<0.05$) assays revealed that 58% of them harboured genotyping error.

Table 5 Stratification of 21 HWD ($P<0.05$) assays due to genotyping error by methodology

Deviation from HWE	*Methodology*	*Number of genotype data sets generated by each methodology*	*Number of data sets associated with genotyping error*	*%*
$0.01<P<0.05$	Taqman	87	4	4.6
	Sequencing	184	8	4.3
	Maldi-tof	32	0	0
	RFLP	10	0	0
	Total	313	12	3.8
$P<0.01$	Taqman	87	2	2.3
	Sequencing	184	4	2.2
	Maldi-tof	32	2	6.2
	RFLP	10	1	10
	Total	313	9	2.9

In this study, when the primers defining assay specificity were designed, high-level repeats in the genome sequence, including Alus and LINE, were masked. However, low-level repeat sequences are more difficult to monitor. In order to address this, primers and/or probes defining the reaction specificity for each of the 36 assays in HWD ($P<0.05$) were retrospectively analysed, by searching against NRNUC and NRHTG databases. In order to identify possible sequence homologies, no sequence filters were used at this stage. Assays developed for five SNPs appeared to be nonspecific. Two of these were SNPs that mapped to a 390 kb region on chromosome 22 flanking *CYP2D6*.[32] Two pseudogenes, *CYP2D7* and *CYP2D8*, lie adjacent to *CYP2D6* and the primers defining assay specificity for the two nonspecific SNPs were found to map to either or both of the pseudogenes. Pseudogenes are clearly abundant,[35] and therefore experimental design must take these sequences into consideration. The other three nonspecific SNP assays demonstrated 100% homology to more than one chromosomal region.

After analysis of the 36 HWD ($P<0.05$) SNPs, and the identification of those assays associated with genotyping error or nonspecificity, 10 SNPs remained. It is possible that the deviation from HWE observed in these SNPs is occurring by chance. However, as a large proportion of low-level duplications have not been represented following the assembly of the draft human genome sequence,[36] it is conceivable that within this group of HWD ($P<0.01$) SNPs there are some assays that may be nonspecific, but all the sequences that their probes and primers are homologous to are not captured in the human genome sequence assembly.

The data reported here were generated during 1998–2000 using various different genotyping technologies. As a result of this retrospective analysis, a high-throughput standardised semi-automated genotyping process has been developed. This incorporates automatic 'electronic PCR' of primers before running the assay. All SNP assays are tested for deviation from HWE in 94 Caucasian DNA samples prior to running the SNP assay across the DNA sample set of interest. Genotypes are scored by highly trained scientists, and data accuracy is not compromised for individual assay genotype success rates. Following this process, analysis of genotypes generated from 94 unrelated Caucasians for 1434 SNPs revealed only 10 HWD ($P<0.01$) SNPs (unpublished data). This is slightly less than the number expected to occur by chance (~14), suggesting improved data quality.

In conclusion, this study demonstrates the successful identification of a proportion of nonspecific assays and assays harbouring genotyping errors, by using a simple test for HWE. The genotyping process was subsequently modified in order to generate data of improved quality.

Acknowledgements

We are grateful to all the scientists within Discovery Genetics Europe, GlaxoSmithKline for sharing their data with us. In particular, we thank Nicola White, Jennifer O'Sullivan, and Christine Caine. We also thank scientists within Population Genetics, GlaxoSmithKline, particularly Peter Boyd.

References

1 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.

2 Gordon D, Heath SC, Ott J: True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum Hered* 1999; **49**: 65–70.

3 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.

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

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

6 Terwilliger JD, Weeks DE, Ott J: Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross over-estimations of the recombination fraction. *Am J Hum Genet* 1990; **47**: A201.

7 Gordon D, Matise TC, Heath SC, Ott J: Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genet Epidemiol* 1999; **17** (Suppl 1): S587–S592.

8 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.

9 Ehm MG, Kimmel M, Cottingham RW: Error detection for genetic data, using likelihood methods. *Am J Hum Genet* 1996; **58**: 225–234.

10 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.

11 Ott J: Linkage analysis with misclassification at one locus. *Clin Genet* 1977; **12**: 119–124.

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

13 Cherny SS, Abecasis GR, Cookson WO, Sham P, Cardon L: The effect of genotype and pedigree error on linkage analysis: analysis of three asthma genome scans. *Genet Epidemiol* 2001; **25**: 36–47.

14 O'Connell JR: Genotyping and error checking. in Elston R, Olson J, Palmer L (eds): *Biostatistical Genetics and Genetic Epidemiology.* New York: Wiley; 2002, pp 348–352.

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

16 Shields DC, Collins A, Buetow KH, Morton NE: Error filtration, interference and the human linkage map. *Proc Nat Acad Sci USA* 1991; **88**: 6501–6505.

17 Mitchell AA, Cutler DJ, Chakravarti A: Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* 2003; **72**: 598–610.

18 Gordon D, Leal SM, Heath SC, Ott J: An analytical solution to sinlge nucleotide polymorphism error-detection rates in nuclear families: implications for study design. *Pac Symp Biocomput* 2000; 663–674.

19 Gordon D, Ott J: Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac Symp Biocomput* 2001; 18–29.

20 Akey JM, Zhang K, Xiong M, Doris P, Jin L: The effect that genotyping errors have on the robustness of common linkage disequilibrium measures. *Am J Hum Genet* 2001; **68**: 1447–1456.

21 Mote VL, Anderson RL: An investigation of the effect of misclassification on the properties of chi-square tests in the analysis of categorical data. *Biometrika* 1965; **52**: 95–109.

22 Gordon D, Levenstein MA, Finch SJ, Ott J: Errors and linkage disequilibrium interact multiplicatively when computing sample sizes for genetic case–control association studies. *Pac Symp Biocomput* 2003; 490–501.

23 Goring HH, Terwilliger JD: Linkage analysis in the presence of errors I: complex valued recombination fractions and complex phenotypes. *Am J Hum Genet* 2000; **66**: 1095–1106.

24 Goring HH, Terwilliger JD: Linkage analysis in the presence of errors II: marker-locus genotyping errors modelled with hypercomplex recombination fractions. *Am J Hum Genet* 2000; **66**: 1107–1108.

25 Goring HH, Terwilliger JD: Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *Am J Hum Genet* 2000; **66**: 1298–1309.

26 Goring HH, Terwilliger JD: Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 2000; **66**: 1310–1327.

27 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.

28 Rice KM, Holmans P: Allowing for genotyping error in analysis of unmatched cases and controls. *Ann Hum Genet* 2003; **67**: 165–174.

29 Hardy GH: Mendelian proportions in a mixed population. *Science* 1908; **28**: 49–50.

30 McCarthy LC, Hosford DA, Riley JH *et al*: Single nucleotide polymorphism alleles in the insulin receptor gene are associated with typical migraine. *Genomics* 2001; **78**: 135–149.

31 Hewett D, Samuelsson L, Polding J *et al*: Identification of a psoriasis susceptibility gene by linkage disequilibrium mapping with a localised single nucleotide polymorphism map. *Genomics* 2002; **79**: 305–314.

32 Hosking LK, Boyd PR, Xu C-F *et al*: Linkage disequilibrium mapping identifies a 390 kb region associated with CYP2D6 poor drug metabolising activity. *Pharmacogenomics J* 2002; **2**: 165–175.

33 Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* 1995; **57**: 289–300.

34 Ranade K, Chang M-S, Ting C-T *et al*: High-throughput genotyping with single nucleotide polymorphisms. *Genet Res* 2001; **11**: 1262–1268.

35 Harrison PM, Hegyi H, Balasubramanian S *et al*: Molecular fossils in the human genome identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genet Res* 2002; **12**: 272–280.

36 Bailey J, Gu Z, Clark R *et al*: Recent segmental duplications in the human genome. *Science* 2002; **297**: 1003–1007.