Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 5

Detecting Genotyping Error Using Measures of Degree of Hardy-Weinberg Disequilibrium

John Attia[*] Ammarin Thakkinstian[†] Patrick McElduff[‡]
Elizabeth Milne[**] Somer Dawson[††] Rodney J. Scott[‡‡]
Nicholas de Klerk[§] Bruce Armstrong[¶] John Thompson[||]

[*] University of Newcastle, john.attia@newcastle.edu.au
[†] Mahidol University, raatk@mahidol.ac.th
[‡] University of Newcastle, patrick.mcelduff@newcastle.edu.au
[**] University of Western Australia, lizm@ichr.uwa.edu.au
[††] University of Western Australia, somerd@ichr.uwa.edu.au
[‡‡] University of Newcastle, rodney.scott@newcastle.edu.au
[§] University of Western Australia, nickdk@ichr.uwa.edu.au
[¶] University of Sydney, brucea@health.usyd.edu.au
[||] University of Leicester, trj@leicester.ac.uk

Detecting Genotyping Error Using Measures of Degree of Hardy-Weinberg Disequilibrium[*]

John Attia, Ammarin Thakkinstian, Patrick McElduff, Elizabeth Milne, Somer Dawson, Rodney J. Scott, Nicholas de Klerk, Bruce Armstrong, and John Thompson

Abstract

Tests for Hardy-Weinberg equilibrium (HWE) have been used to detect genotyping error, but those tests have low power unle the sample size is very large. We aeed the performance of measures of departure from HWE as an alternative way of screening for genotyping error. Three measures of the degree of disequilibrium (α, ,D, and F) were tested for their ability to detect genotyping error of 5% or more using simulations and a real dataset of 184 children with leukemia genotyped at 28 single nucleotide polymorphisms. The simulations indicate that all three disequilibrium coefficients can usefully detect genotyping error as judged by the area under the Receiver Operator Characteristic (ROC) curve. Their discriminative ability increases as the error rate increases, and is greater if the genotyping error is in the direction of the minor allele. Optimal thresholds for detecting genotyping error vary for different allele frequencies and patterns of genotyping error but allele frequency-specific thresholds can be nominated. Applying these thresholds would have picked up about 90% of genotyping errors in our actual dataset. Measures of departure from HWE may be useful for detecting genotyping error, but this needs to be confirmed in other real datasets.

KEYWORDS: association study, Hardy-Weinberg equilibrium, genotyping error, disequilibrium coefficient

[*]This research was supported by the National Health and Medical Research Council of Australia project grant number 254539. Bruce Armstrong's research is supported by a University of Sydney Medical Foundation Program Grant.

Introduction

An increasing number of epidemiological studies are incorporating large-scale genotyping of polymorphisms, including genome-wide association studies (GWAs), to explore gene-disease associations and gene and environment interactions. Given that no genotyping method is 100% accurate and that genotype mistakes can lead to increased random error and bias in gene-disease associations (Gordon, et al. 2001; Gordon, et al. 1999b; Govindarajulu, et al. 2006), methods have been developed to detect and, where possible, deal with genotyping error. These methods have followed five main avenues:

- Genotyping in duplicate to confirm results (Gordon, et al. 2004; Rice and Holmans 2003; Tintle et al. 2007), although at substantial added expense;
- Dropping ambiguous or difficult to call genotypes, which leads to reduced power (Kang, et al. 2004);
- Developing analytic methods that correct for some degree of genotyping error (Hao and Wang 2004);
- In family-based designs, checking for Mendelian inconsistencies, although this appears to have low sensitivity, detecting only 25-30% of errors (Gordon, et al. 1999a); and
- Using tests of Hardy-Weinberg equilibrium (HWE) to prompt re-checking of genotype information (Tiret and Cambien 1995; Xu, et al. 2002). This latter method is based on the assumption that in a large, randomly mating population, genotype frequencies should comply with HWE proportions. Deviation from these proportions can be caused by many factors, one of which is genotyping error.

The last approach has received much attention, but both actual data (Hosking, et al. 2004) and simulations(Cox and Kraft 2006; Leal 2005; Zou and Donner 2006) have indicated that tests of HWE have poor power to detect genotyping error at common allele frequencies. Compounding the problem of low power at small sample sizes is the separate problem of excessive power at very large sample sizes. The consequence of viewing the detection of genotyping errors as a problem of statistical testing is that, like all statistical tests, there will be too many false negatives at low sample sizes and the detection of negligible and practically unimportant deviations at very large sample sizes (Wigginton et al. 2005; Rohlfs and Weir 2008). For example, if genotyping of 102 people at a diallelic locus leads to genotype frequencies of 48, 41 and 13 (minor allele frequency of 33%), a test of HWE indicates no evidence of deviation (Chi-square P = 0.370, Exact test P = 0.373). However, increasing the sample size 10-fold while keeping the genotype proportions the same, i.e. 480, 410, and 130, produces a different conclusion based on HWE testing (Chi-square P = 0.005, Exact test P = 0.006). This dependence of tests for HWE on sample size is a major drawback when using HWE to detect genotyping error, and highlights the well described pitfall of relying on a p-value of testing to judge "significance" (Rothman 2002). Some authors have sought to get around this problem by routinely adjusting the Chi-square test for genetic association to account for deviations from HWE rather than attempting to detect whether or not such deviations are present (Weir 1996; Zou and Donner 2006).

Instead of focusing simply on whether proportions are or are not consistent with HWE, we propose to look at measures that quantify the **degree** of deviation from HWE for detecting genotyping error and to view the problem as one of screening rather than testing for genotyping error. Three commonly used measures have already been defined in the literature - the "inbreeding coefficient" (F) (Weir 1996), the "disequilibrium parameter" (D) (Hernandez and Weir 1989), and the "alpha parameter" (α) (Lindley 1998) - but to date their use to detect genotyping error has not been explored. They all have the favorable property that they vary according to genotype proportions and not sample size. For instance, in the example above, the values of α, F, and D remain at 0.198, 0.089, and 0.020 despite the 10-fold increase in sample size.

Most previous simulation studies of genotyping error have assumed non-differential error (Gordon, et al. 2004; Zou and Donner 2006) or error that is differential between cases and controls (Moskvina et al. 2006; Ahn et al. 2009). In reality however, genotyping error may also be differential in hybridization or amplification, in that heterozygotes are more likely to be miscalled as homozygotes than vice versa (Milne, et al. 2006) and this pattern is incorporated in our simulations. We compared the disequilibrium measures (α, F, and D) to see which would best be able to detect differential hybridization or amplification genotyping error and what the optimal threshold for detection might be. These thresholds were then used for data from our study of childhood leukemia (Milne, et al. 2006), under the assumption that whole–genome amplified samples from buccal swabs have more genotyping errors than samples taken from whole blood in the same person.

Methods

AUS-ALL study and laboratory methods

The Australian Study of Causes of Acute Lymphoblastic Leukaemia in Children (AUS-ALL) is a population-based case-control study that began in 2003. Case families were identified through the ten pediatric oncology centers in Australia, and control families through national random-digit dialing. Case children and their parents provided blood samples during routine hospital visits, and buccal samples were also collected from case children. The latter were collected to provide a comparison of genotyping results from blood and buccal DNA. The study was approved by the Human Research Ethics Committees of the ten participating hospitals and parents also completed DNA consent forms.

For this study we included genotype data from 184 case children who provided usable blood and buccal samples. Blood samples were collected in EDTA blood collection tubes and buccal samples were collected using FTA® Micro cards in accordance with the manufacturer's instructions (Whatman International Ltd., Maidstone, United Kingdom). All samples were sent to the laboratory within 24 hours of collection. Genomic DNA was isolated from blood samples using the Wizard Genomic DNA Purification Kit (Promega, Sydney Australia) and DNA was extracted from FTA® cards as recommended by the manufacturer (Whatman International Ltd.). Samples extracted from FTA® cards were whole genome amplified using GenomiPhi DNA amplification Kit (Amersham, GE Healthcare) to increase the quantity of DNA for, and for ease of, genotyping.

Because of their specific expertise, three separate laboratories genotyped the bloods for a number of polymorphisms in genes coding for folate metabolizing enzymes, DNA repair enzymes, and xenobiotic metabolizing enzymes. We conducted between- and within-laboratory quality control checks. In relation to the former, the MTHFR C667T polymorphism was analyzed at each laboratory; this SNP was selected because the primary hypothesis for AUS-ALL relates to folate intake and metabolism. Genotyping was performed using TaqMan® SNP Genotyping Assays (Applied Biosystems) and polymorphism specific probes and primers were used according to standard laboratory protocols. All laboratory personnel were blinded to the DNA source and case or control status of each sample. We report here only the results of 28 diallelic SNPs (including MTHFR C677T at the three laboratories) where the blood genotyping proportions show no evidence of deviation from HWE (Chi square $P > 0.05$).

Simulations and analysis

Simulations were run assuming a diallelic locus. Letting A and a represent the major (common) and minor alleles respectively, the allele frequencies are p_A and p_a respectively, where $p_A+p_a=1$. The genotyping error rate was represented by ε and three types of genotyping error were simulated:

- Scenario 1: Some of the heterozygotes are miscalled as being homozygous for the common allele, i.e. there is loss of heterozygosity in the direction of the major allele. The resulting genotype frequencies, P, are:

$$P_{AA} = p_A^2 + 2p_A p_a \varepsilon$$
$$P_{Aa} = 2p_A p_a(1-\varepsilon)$$
$$P_{aa} = p_a^2$$

- Scenario 2: Some of the heterozygotes are miscalled as being homozygous for the minor allele, i.e. there is loss of heterozygosity in the direction of the minor allele. The genotype frequencies here are represented by:

$$P_{AA} = p_A^2$$
$$P_{Aa} = 2p_A p_a(1-\varepsilon)$$
$$P_{aa} = p_a^2 + 2p_A p_a \varepsilon$$

- Scenario 3: Some of the heterozygotes are miscalled in such a way that they are equally likely to be homogygous for the major or minor alleles. The genotype frequencies here are represented by:

$$P_{AA} = p_A^2 + 2p_A p_a(\frac{\varepsilon}{2})$$
$$P_{Aa} = 2p_A p_a(1-\varepsilon)$$
$$P_{aa} = p_a^2 + 2p_A p_a(\frac{\varepsilon}{2})$$

Because of the balanced errors in this scenario, the allele frequencies do not change, unlike scenarios 1 and 2.

In each scenario the proportions reduce to those predicted under HWE (p_A^2, $2p_Ap_a$, p_a^2) when the error rate, ε, is zero.

In each simulation, a set of SNPs was randomly created; half were given a very low genotyping error rate (below 5%) and the other half were given a higher error rate (5% or more). The measures of departure from HWE were calculated and used as the basis of a screen to try to differentiate between the SNPs with high and low rates of genotyping error.

In the simulations the following parameters were varied:

- The higher genotyping error rate: the percentage of heterozygotes that are miscalled was varied (5%, 10%, 20%, 30% and 50%). Although the highest genotyping error rates are unlikely, these values were chosen to indicate trends.
- Minor allele frequency was varied from 5, 10, 20, 30 and 50%.

Calculation of disequilibrium measures

The three measures of disequilibrium were calculated as follows; assuming that AA, Aa, and aa are the three genotype groups, and P_{xx} and p_x refer to the observed genotype proportions and allele frequencies respectively, as above:

a) *Inbreeding coefficient (F)*(Weir 1996)

$$F = P_{aa}/p_a + P_{AA}/p_A - 1$$

where the bounds on F are functions of the allele frequencies:

$$\max[-p_a/p_A, - p_A/p_a] \leq F \leq 1$$

b) Disequilibrium parameter (D)(Hernandez and Weir 1989)

$$D = P_{AA} - p_A^2$$

where the bounds on D are functions of the allele frequencies:

$$\max[-p_a^2, -p_A^2] \leq D \leq p_a p_A$$

c) *Alpha* (α) (Lindley 1998)

$$\alpha = ½\log (4P_{aa}P_{AA}/P_{Aa}^2)$$

where the bounds on alpha are:

$$-\infty < \alpha < \infty$$

Estimates were obtained by plugging the sample proportions into these formulae. However, to avoid problems with zero cells we added 0.5 to the numbers in each category before calculating the alpha statistic.

These measures are related to each other. For example, the relationship between F and D is:

$$F=D/p_A(1-p_A)$$

and they are also related to the test statistic of the Chi-square test of HWE:

$$X^{2=} n\hat{}D^2/p_A^2 (1-p_A)^2$$

$$X^2=n\hat{}F^2$$

All scenarios and combinations of values for these measures were simulated with study sample sizes of 500 and were repeated 100 times before their results were averaged. We assessed these measures as one might a diagnostic or screening test, measuring their ability to detect the SNPs with the higher rates of genotyping error using a receiver-operator characteristic (ROC) curve. An ROC curve plots the sensitivity on the y-axis vs (1-specificity) on the x axis as functions of a varying threshold. The area under the ROC curve (AUC) is an estimate of the diagnostic accuracy across the spectrum of possible thresholds. An AUC that is much greater than 0.5 (chance agreement) indicates a potentially useful measure for identifying SNPs with error; a perfect diagnostic test would have an AUC of 1 (Soreide 2009; Zou et al 2007). A threshold was then chosen as the value of alpha, D or F that maximized Youden's index (sensitivity + specificity – 1), and this was also expressed as a likelihood ratio (Sackett, et al. 1991). Simulations were run using STATA version 10 (StataCorp 2007, Texas, USA).

The derived thresholds were then applied to data from our case-control study of childhood leukemia (Milne, et al. 2006) in which we had the opportunity to genotype a series of children diagnosed with leukemia for a number of SNPs; genotyping was performed in two ways:

- directly from blood samples (used as the reference standard) and
- whole-genome amplified (WGA) DNA derived from buccal cells.

Samples where genotyping failed were excluded. Most of the discordant results involved a heterozygous blood result and a homozygous buccal result. In this report, we describe the error rate for the SNPs assuming that where the results were discordant, the genotype from the blood sample is correct and the genotype from buccal WGA DNA is incorrect. Support for this interpretation comes from the fact that such problems with WGA fidelity have been reported before (Pinard, et al. 2006). The α, F, and D measures for each SNP were estimated. The results of applying the threshold derived from simulations to our real data is summarized as a 2x2 table, treating the threshold as a "diagnostic" test for detecting genotyping error in the buccal swabs compared with the reference standard of the blood DNA result.

Results

Simulations

The results of simulations for the first two scenarios, i.e. genotyping error for heterozygotes in the direction of the major allele and then minor allele, are shown in Table 1. From this table, a number of trends may be seen:

- Genotyping error in the direction of the minor allele (Scenario 2) is more easily detected than in the direction of the major allele (Scenario 1). Even at low allele frequencies and low error rates, there is reasonable detection ability with AUCs in the range of 0.65 to 0.8 (for allele frequencies up to 20% and error rates up to 10%)
- The three disequilibrium measures perform very similarly with results for alpha and F being almost identical; D appears to perform slightly worse

for detecting error in the direction of the major allele but slightly better for detecting error in the direction of the minor allele.

- As the allele frequencies get higher, the AUCs get better at detecting error in the direction of the major allele but get worse for detecting error in the direction of the minor allele; as expected, the AUC values converge for the two scenarios as the allele frequency reaches 50%.

Results for scenario 3, i.e. heterozygote error in the direction of both homozygotes equally, were partway between scenarios 1 and 2 (data not shown). Of note is that the balanced error in this scenario leads to an F value equal to the genotype error rate.

Table 1. Area Under the ROC Curve (AUC) for detecting genotyping error in heterozygotes, in the direction of the major allele (Scenario 1, Aa to AA) , and in the direction of the minor allele (Scenario 2, Aa to aa).

SNP	Area under the ROC Curve (AUC)										
	Error rate (%)	Allele frequency of a (%)									
		5		10		20		30		50	
		Scenario									
		1	2	1	2	1	2	1	2	1	2
α	5	0.53	0.64	0.52	0.64	0.53	0.62	0.54	0.61	0.56	0.56
	10	0.57	0.75	0.54	0.76	0.57	0.76	0.59	0.74	0.69	0.68
	20	0.65	0.88	0.58	0.92	0.66	0.93	0.74	0.93	0.88	0.88
	30	0.73	0.94	0.63	0.98	0.77	0.98	0.87	0.99	0.97	0.97
	50	0.87	0.99	0.82	0.99	0.93	1.00	0.99	1.00	1.00	1.00
D	5	0.52	0.65	0.51	0.67	0.51	0.65	0.53	0.63	0.56	0.56
	10	0.55	0.76	0.52	0.80	0.54	0.80	0.57	0.76	0.68	0.68
	20	0.61	0.89	0.53	0.94	0.60	0.96	0.69	0.95	0.87	0.88
	30	0.67	0.95	0.54	0.98	0.69	1.00	0.82	0.99	0.97	0.97
	50	0.76	0.99	0.62	0.99	0.82	1.00	0.94	1.00	1.00	1.00
F	5	0.53	0.65	0.51	0.66	0.52	0.64	0.53	0.62	0.56	0.56
	10	0.57	0.76	0.54	0.79	0.56	0.79	0.58	0.75	0.69	0.68
	20	0.63	0.89	0.57	0.94	0.65	0.95	0.72	0.94	0.88	0.88
	30	0.70	0.95	0.61	0.98	0.75	0.99	0.85	0.99	0.97	0.97
	50	0.82	0.99	0.76	0.99	0.91	1.00	0.98	1.00	1.00	1.00

Given that the estimates of the 3 disequilibrium measures may be indicators of genotyping error (as judged by the AUC), the next question is what threshold to use. A hypothesis test at the 5% level sets the specificity of the screen to 95% and in doing so penalizes the sensitivity. A better overall performance may be possible and so Table 2 shows the optimal cutoffs of α, F, and D as judged by maximizing Youden's index. That table also shows the value of the positive likelihood ratio by allele frequency and size of the higher rate of genotyping error. At lower allele frequencies and lower error rates, the thresholds for all three measures of disequilibrium are variable, and depend on the direction of the error; as the allele frequency increases beyond 20%, the estimate of the threshold is more stable.

Table 2. Optimal threshold values (and likelihood ratios) for the three measures of Hardy-Weinberg disequilibrium under different scenarios; scenario 1 is genotyping error in heterozygotes, in the direction of the major allele (Aa to AA), and scenario 2 in the direction of the minor allele (Aa to aa).

SNP		Optimal Threshold values for HWD measures									
	Error rate (%)	Allele frequency of a (%)									
		5		10		20		30		50	
		Scenario									
		1	2	1	2	1	2	1	2	1	2
α	5	0.45 (1.1)	0.75 (1.6)	0.59 (1.2)	0.59 1.9	0.27 (1.2	0.29 (1.5)	0.26 (1.2)	0.28 (1.7)	0.19 (1.2)	0.19 (1.3)
	10	0.55 (1.3)	0.79 (2.3)	0.68 (1.6	0.63 (3.2)	0.29 (1.3)	0.34 (2.4)	0.21 (1.4)	0.28 (2.4)	0.24 (2.1)	0.24 (2.2)
D	5	0.003 (1.1)	0.006 (1.6)	0.014 (1.3)	0.009 1.7	0.02 (1.1	0.02 (1.9)	0.03 (1.2)	0.03 (1.7)	0.03 (1.3)	0.02 (1.3)
	10	0.003 (1.2)	0.01 (2.6)	0.01 (1.2)	0.01 (3.2)	0.02 (1.4)	0.02 (3.3)	0.02 (1.3)	0.03 (2.8)	0.03 (2.1)	0.03 (2.2)
F	5	0.06 (1.0)	0.13 (1.6)	0.14 (1.3)	0.11 (1.7)	0.08 (1.1)	0.12 (1.8)	0.11 (1.2)	0.13 (1.8)	0.12 (1.3)	0.10 (1.3)
	10	0.08 (1.1)	0.17 (2.6)	0.11 (1.4)	0.13 (2.7)	0.11 (1.4)	0.14 (3.2)	0.09 (1.4)	0.13 (2.7)	0.12 (2.1)	0.12 (2.2)

Leukemia data

Table 3 gives the blood and FTA/WGA results for each of the 28 SNPs for all 184 children. As seen in the last column, 10 of the 28 SNPs had an error rate of 5% or more, with the highest error rate reaching almost 13%. There were 207 individual genotyping reactions that were incorrect; in almost 90% of these, the buccal sample was incorrectly identified as a homozygote.

Using D as an example, Table 4 gives the 2x2 table obtained using the lowest and highest threshold values according to allele frequency as taken from table 2; the reference standard is the blood genotyping. Using the lowest D threshold (0.003) would have identified 19 out of 28 SNPs as potentially problematic, of which 9 would have been truly erroneous; one erroneous SNP would not have been detected. Using the highest threshold D (0.006) yielded similar results.

Table 3. Details of SNPs and genotype frequencies in the entire dataset. Wt= homozygous wildtype; het=heterozygous; mut=homozygous mutant

SNP	Blood				Buccal mucosa				Type of error n(blood/ buccal)
	wt	het	mut	% minor allele	wt	het	mut	% minor allele	
ATM D1835N	131	52	1	15	131	52	1	15	-
ATM V2424G	184	0	0	0	184	0	0	0	-
CBS 844 ins 68	151	33	0	9	156	25	1	7	8 (WT/Ins/WT/WT) 1 (WT/Ins/Ins/Ins) 2 (WT/WT/WT/Ins)
CBS T2199C	63	81	35	42	66	72	39	42	8 (TC/TT) 4 (TC/CC) 2 (TT/CC) 1 (CC/TC) 1 (CC/TT)
CYP1A1 Ile-Val	166	14	1	4	166	15	1	5	-
CYP3A4*1B	167	16	0	4	159	15	4	6	-
CYPE1D ra1	147	34	3	11	145	30	5	11	3 (TA/TT) 1 (TT/TA) 2 (TA/AA)
ERCC1 N118N	77	82	25	36	81	74	28	36	4 (GA/AA) 3 (GA/GG)
ERCC2 D312N	73	82	29	38	87	57	39	37	14 (GA /GG) 10 (GA/AA)
ERCC2 K751Q	77	87	20	35	88	71	25	33	11 (AC/AA) 6 (AC/CC) 1 (CC/AC)
GST pi EX5	87	79	18	31	92	71	19	30	6 (AG/AA) 2 (AG/GG) 1 (AA/AG)
GST pi EX6	155	28	0	8	154	22	7	10	6 (CT/TT) 1 (CC/TT)
MS A2756G	121	62	1	17	122	52	8	19	1 (AG / AA) 9 (AG/GG)
MS C5049A	60	97	27	41	62	92	28	41	2 (CA /CC) 1 (CA/AA)
MSR A66G	42	101	41		44	81	56	53	5 (AG/AA) 15 (AG/GG)

				50					2 (AA/AG)
MTHFR A1298C	80	86	17	33	84	68	30	35	13 (AC / CC) 4 (AC/AA)
MTHFR C677T (lab 1)	89	79	16	30	90	75	16	30	3 (CT/CC) 2 (CC/CT)
MTHFR C677T (lab 2)	86	81	17	31	90	74	18	30	6 (CT/CC) 1 (TT/CC)
MTHFR C677T (lab 3)	88	80	16	30	89	76	18	31	3 (CT/TT) 1 (CT/CC)
NAT2 M1	63	75	46	45	62	76	46	46	1 (CC/CT) 2 (CT/CC)
NAT2 M2	98	72	14	27	99	70	15	27	1 (GA / AA) 1 (GA/GG)
NAT2 M3	172	10	1	3	173	9	2	4	1 (GA/GG) 1 (GA/AA)
NAT2*5B C481T	63	75	46	45	61	78	45	46	3 (CT/TT) 4 (TT/CT) 2 (CC/CT)
NAT2*5B T341C	59	76	48	47	58	73	51	48	3 (TC/CC) 1 (CC/TC) 1 (TT/TC)
NQ01 C609T	126	48	10	18	126	47	10	18	3 (CT/CC)
XRCC3 T241M	68	86	26	38	73	76	27	37	5 (GA/GG) 3 (GA/AA) 1 (AA/GA)
XRCC5	150	33	1	10	151	31	1	9	1 (AG / AA)
hMSH3 A1036T	99	75	10	26	102	65	14	26	4 (AG / GG) 3 (AG / AA)

It is also possible that other users may wish to choose other thresholds that optimize specificity or sensitivity for detecting genotyping error >5%. For example, if we apply a general cutoff of 0.025 for D to the actual dataset, rather than a more specific cutoff according to allele frequency, the resulting AUC is 0.69 with a sensitivity of 60% and specificity of 77.8%, and +LR of 2.7. Alternatively, some may choose to detect even lower levels of genotyping error, e.g. 3% or less. For example, applying a cutoff for D of 0.025 to detect a more conservative error rate of $\geq$ 3%, the AUC, sensitivity, specificity, and +LR were 0.72, 52.9%, 90.9%, and 5.8 respectively.

Table 4. Predicted error for 28 SNPs using buccal swab DNA and WGA.

D Low threshold	Actual genotyping error		D High threshold	Actual genotyping error	
	Yes	No		Yes	No
≥ 0.003	9	10	≥ 0.006	9	11
< 0.003	1	8	< 0.006	1	7
Sensitivity	90.0		Sensitivity	90.0	
Specificity	44.4		Specificity	38.9	
LR+	1.6		LR+	1.5	
AUC	0.67		AUC	0.64	

It is also interesting to note that these disequilibrium measures appear to complement other approaches to detecting genotyping error. Table 3 lists differences in minor allele frequency between blood (reference) and buccal swabs, and percentage of missing data in buccal swabs compared with blood. In our dataset, neither of the latter 2 measures appears to correlate with SNPs that have high discordance.

Discussion

In this study, we explored the ability of three existing measures of the degree of deviation from HWE (D, F and α) to detect differential genotyping error. Estimates of these measures of disequilibrium have the desirable property that, unlike the p-value, they are dependent on genotype proportions only and independent of sample size. Our simulations indicate that all three measures have similar ability to detect differential genotyping error, as evidenced by AUCs that are above 0.5, and perform better if the direction of the error is from the heterozygous to the homozygous minor allele. The AUCs also increase in magnitude as the genotyping error increases, but do not necessarily increase as the allele frequency increases; this depends on the direction of the error. Even for allele frequencies under 20% and error rates under 10%, AUCs range from 0.65 to 0.8 for detecting error in the direction of the minor allele indicating potential for use as a discriminative test. There is little to choose between the measures and it appears that any one of the three might reasonably be used.

As with any diagnostic test, one must define a threshold at which it can be said the outcome is positive. In the real data set we defined a positive outcome as 5% or more genotyping error. This value was based on a previous review that indicated genotyping error rates in genetic association studies ranging from 0.5% up to 15% per SNP (Pompanon, et al. 2005); our choice is in the middle of this range and corresponds roughly to the threshold where error rates can start to significantly impact power for detection of association (Lincoln and Lander 1992). The thresholds for screening that we chose were based on maximizing both sensitivity and specificity simultaneously. Depending on the needs of the user, one may choose different thresholds, for example the traditional statistical approach is to set specificity at 95% (i.e. one minus the significance level of the statistical test) and let sensitivity (i.e. power) be determined by the sample size.

Maximising sensitivity means detecting more cases of genotyping error at the cost of re-genotyping many samples with no error, whereas maximizing specificity means minimizing the re-genotyping at the cost of missing more samples with error. One may also choose to modify thresholds to detect lower levels of genotyping error, e.g. 3%. It was clear that no “one size fits all” threshold could be applied across all allele frequencies, and we propose that if these measures are developed then thresholds based on allele frequency should be explored.

The relationships between χ^2 and F (i.e. χ^2=n$\hat{}$F^2) and χ^2 and D (i.e. χ^2= n$\hat{}$D^2/p$_A{}^2$ (1-p$_A$)^2) are monotonic for a given allele frequency, which indicates that these three measures have the same ability to detect genotyping error. So it would be possible to calculate the threshold corresponding to a particular choice of sensitivity and specificity by using the appropriate central and non-central chi-squared distributions instead of simulations.

The blood samples used in this study were from children diagnosed with acute lymphoblastic leukaemia. To allow for the selected nature of the study subjects, we only included SNPs where the genotype proportions from the blood samples showed no evidence of deviation from HWE using the chi-square test. As discussed above, however, this approach only detects genotyping results with extreme deviation from HWE. This methodological study was based on the premise that, where there was genotype discordance between blood and buccal samples, the results from the blood samples were correct. We believe this to be a justifiable assumption, given that the within-lab error rate from our quality control tests was less than 1% (unpublished observations). We considered genotyping error per SNP rather than per allele or per PCR since it reflects a combination of reliability of laboratory and experimental procedures, and can be compared with other markers (Pompanon, et al. 2005).

Similar patterns were found to hold for all three of the patterns of genotyping error that we tested, all of which had loss of heterozygosity, i.e. error away from the heterozygotes. This corresponds to the most common form of genotyping error, in that secondary structure of the DNA can make one allele more difficult to amplify than another; indeed most errors in our leukemia dataset were in the direction of loss of heterozygosity. Errors in the direction of homozygote going to heterozygote, i.e. gain of heterozygosity, are due to contamination, and it is possible that contamination may influence these 3 measures differently. It is also important to note that previous simulations have assumed non-differential genotyping error and indicate greatly reduced ability of HWE to detect this kind of error (Zou, 2006 #27). Non-differential genotyping error can affect statistical inference in genetic association studies by decreasing power; differential genotyping error however can lead to bias, which is more serious and hence the focus of our simulations. Although some methods have been developed to incorporate genotyping error into data analysis (Gordon and Ott 2001; Hao and Wang 2004), these methods involve assumptions and it may be more prudent to detect actual error and correct it.

In summary, these measures of disequilibrium show favourable characteristics as potential tools for detecting differential genotyping error and merit further work. Their discriminative ability is reasonable, i.e. AUCs in the range of 0.65 to 0.8, at low allele frequencies and low error rates, and is independent of sample size. Hence, in the context of large-scale genotyping, they may have a place among the many other quality control checks, e.g. checking data for missingness at random, checking minor allele frequencies against population values, etc. Further work needs to be done to confirm these results and extend them to other kinds of genotyping error, and to define other thresholds that may increase sensitivity or specificity as desired. We plan to explore the use of measures of disequilibrium in the context of genome-wide association analyses in follow-up studies. Finally we note again that genotyping error is only one of the possible reasons for departure from HWE and for that reason SNPs that fail the screen will require further study before genotyping error can be declared.

References

Ahn K, Gordon D, Finch SJ (2009) Increase of rejection rate in case-control studies with the differential genotyping error rates. Stat Appl Genet Mol Biol 8 (1):Article25.

Cox DG, Kraft P. (2006). Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered 61:10-4.

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 69:371-80.

Gordon D, Heath SC, Ott J. (1999a). True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. Hum Hered 49:65-70.

Gordon D, Matice T, Heath S, Ott J. (1999b). Power loss for multiallelic transmission/disequilibrium test when errors introduced:GAW11 simulated data. Genet Epidemiol 17:587-92.

Gordon D, Ott J. (2001). Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput:18-29.

Gordon D, Yang Y, Haynes C, Finch SJ, Mendell NR, Brown AM, Haroutunian V. 2004. Increasing power for tests of genetic association in the presence of phenotype and/or genotype error by use of double sampling. Statistical Applications in Genetics and Molecular Biology.

Govindarajulu US, Spiegelman D, Miller KL, Kraft P. (2006). Quantifying bias due to allele misclassification in case-control studies of haplotypes. Genet Epidemiol 30:590-601.

Hao K, Wang X. (2004). Incorporating individual error rate into association test of unmatched case-control design. Hum Hered 58:154-63.

Hernandez JL, Weir BS. (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics 45:53-70.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF. (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet 12:395-9.

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. Pac Symp Biocomput:116-27.

Leal SM. (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol 29:204-14.

Lincoln SE, Lander ES. (1992). Systematic detection of errors in genetic linkage data. Genomics 14:604-10.

Lindley D. (1998). Statistical inference concerning Hardy-Weinberg equilibrium. In: Bernardo JM BJ, Dawid AP, and Smith AFM, editor. Bayesian Statistics. 3 ed. Oxford: Oxford University Press. p 307-26.

Milne E, van Bockxmeer FM, Robertson L, Brisbane JM, Ashton LJ, Scott RJ, Armstrong BK. (2006). Buccal DNA collection: comparison of buccal swabs with FTA cards. Cancer Epidemiol Biomarkers Prev 15:816-9.

Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC (2006) Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered 61:55-64.

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7:216.

Pompanon F, Bonin A, Bellemain E, Taberlet P. (2005). Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847-59.

Rice KM, Holmans P. (2003). Allowing for genotyping error in analysis of unmatched case-control studies. Ann Hum Genet 67:165-74.

Rohlfs RV, Weir BS (2008) Distributions of Hardy-Weinberg equilibrium test statistics. Genetics 180:1609-16.

Rothman K. (2002). Random error and the role of statistics. Epidemiology: An introduction. Oxford: Oxford University press.

Sackett DL, Haynes RB, Guyatt GH, Tugwell P. (1991). Clinical Epidemiology A Basic Science for Clinical Medicine. Toronto: Little, Brown and Company

Soreide K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J Clin Pathol. 62:1-5.

Tintle NL, Gordon D, McMahon FJ, Finch SJ. (2007) Using duplicate genotyped data in genetic analyses: testing association and estimating error rates. Stat Appl Genet Mol Biol 6 (1); article 4.

Tiret L, Cambien F. (1995). Departure from Hardy-Weinberg equilibrium should be systematically tested in studies of association between genetic markers and disease. Circulation 92:3364-5.

Weir BS. (1996). Genetic data analysis II: Methods for discrete population genetic data. Sunderland, MA, USA: Sinauer Associates.

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887-93.

Xu J, Turner A, Little J, Bleecker ER, Meyers DA. (2002). Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Hum Genet 111:573-4.

Zou KH, O'Malley AJ, Mauri L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115:654-7.

Zou GY, Donner A. (2006). The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann Hum Genet 70:923-33.