Proc. Assoc. Advmt. Anim. Breed. Genet. 18:243-246

POPULATION STRATIFICATION, NOT GENOTYPE ERROR, CAUSES SOME SNPS TO DEPART FROM HARDY-WEINBERG EQUILIBRIUM

Y.D. Zhang and B. Tier

Cooperative Research Centre for Beef Genetic Technologies
Animal Genetics and Breeding Unit*, University of New England, Armidale NSW 2351

SUMMARY

Large scale whole genome scans generate massive amounts of genotype data. It is essential to check genotype integrity and identify genotype errors prior to association analysis. Departure from Hardy-Weinberg Equilibrium has been adopted as one of the main methods to identify genotype errors. However population stratification also causes departure from Hardy-Weinberg Equilibrium, which is a disadvantage of this approach. This study used 2 sets of SNP genotypes to show that after basic editing using Call Rate and minor allele frequency, up to 13% of SNPs departed from Hardy-Weinberg Equilibrium (HWD) and about one third of these HWD SNPs could be falsely identified as genotype errors, were attributable to population subdivision (eg herd of origin, cohort) for one dataset and corresponding numbers for the second dataset are 21% and 16%, respectively. This approach can avoid improper culling of a considerable proportion of SNPs.

INTRODUCTION

Genotype errors exist in mass generated SNP genotype data. Genotype error may have significantly deleterious effects on genetic tests, such as affecting LD measures and linkage distance in family-design. Such errors may also lead to a high false positive rate, in particular in case-control design (Mitchell *et al.* 2003); for example in the case-control design, difference in allele frequencies at a locus between case and control groups could be interpreted as linkage between this locus and the causal gene.

There are many sources of genotype errors, eg assay failure. Prior to association analysis, it is necessary to identify genotype errors and control them at a certain level. Though some study argued that various combinations of quality control did not reduce much the proportion of false association (Chan *et al.* 2009). Methods proposed to detect genotype errors can be classified into 4 groups: testing Mendelian Inheritance using family based or/and trio data; checking closely linked loci for double recombination events; checking experimental error using duplicates (not useful for systematic genotyping errors); checking errors in population-based or controls of case-control design using Hardy-Weinberg Equilibrium (HWE) test. Deviation from Hardy-Weinberg Equilibrium (HWD) has been widely used for detecting genotype errors (eg Hosking *et al.* 2004). However, besides genotype error, there are a number of other factors causing HWD, such as small population variation and population structure (inbreeding, assortative mating, stratification/admixture). This is particularly relevant to genotypes from livestock, because livestock populations have been subject to decades of selective breeding and breed formation (inbreeding, assortative mating, selection). As illustrated by Hartl and Clark (2007) using Wright's example, that estimated frequencies of the recessive allele for blue flower colour in a population of *Linanthus parryae* in approximately 900 square miles of the Mohave desert exhibited in remarkable geographical subdivision. The average allele frequencies in the East, West and Central regions were 0.515, 0.013 and 0.189, respectively, and the corresponding heterozygosity were 0.50, 0.03 and 0.31. Geographic isolation in the East, West and Central regions, implying

*AGBU is a joint venture of NSW Department of Primary Industries and University of New England.

population subdivision, causes a reduction in heterozygous genotypes, relative to that expected under random mating.

This demonstrates that low levels of heterozygosity and population sub-division can lead to marked deviation from HWE. This implies that some genotype errors are falsely identified by HWE test, because departure from HWE may be actually due to population subdivision. The aim of this study is to describe a method for identifying HWD caused by population structure using SNP genotypes, and demonstrate the method in beef data..

MATERIAL AND METHODS

Reduction in heterozygosity is calculated as the difference between the expected heterozygosity under random mating and that observed in the whole population or subpopulation. The fixation index or Wright's F-statistic is defined as the reduction in heterozygosity expected with random mating at any one level of a population hierarchy relative to another, more inclusive level of the hierarchy (Hartl and Clark 2007). In this assessment, the hierarchical F-statistics is expressed as $F_{ST} = (H_T - H_S) / H_T$, where H_T is heterozygosity of the total population and H_S is the average heterozygosity of subpopulations, for instance the sire groups. Wright (1978) suggested the following interpretations for F_{ST}:

- The range of 0 to 0.05 indicates little genetic differentiation.
- The range of 0.05 to 0.15 indicates moderate genetic differentiation.
- The range of 0.15 to 0.25 indicates great genetic differentiation.
- The value of F above 0.25 indicates very great genetic differentiation.

Modified χ^2 test. In general, Hardy-Weinberg Equilibrium tests were performed using a χ^2 test. This study employed a modified χ^2 test to carry out HWE test. The expected frequencies for three genotypes were adjusted using F_{ST} as the weight factor:

- Aa: $2pq - 2pq\ F_{ST}$
- AA: $p^2 + pq\ F_{ST}$
- aa: $q^2 + pq\ F_{ST}$

Data. SNP genotypes in this assessment were mainly derived from two beef cattle whole genome scan projects (designated as P1 and P2), generated using Affymetrix 10K platform. The SNP genotype data from P1 were the main data which were generated for 579 heifers on 9065 SNPs. In P1, animals can be further classified into subclasses by herd of origin, cohort and sire group (half-sib family). In P2, 9421 SNPs were genotyped for 191 animals derived from 7 breeds. The P2 SNP genotypes were used in this assessment for comparison purposes. Three parameters were available to assess the integrity of SNP data: SNP Call Rate (CR, an indication of genotype completeness, at a scale of 0 to 100%), minor allele frequency (MAF) and deviation from Hardy-Weinberg Equilibrium (HWE). The SNP genotypes were initially edited against MAF and CR. The empirical culling thresholds for CR and MAF were suggested as >93% and >0.05. In 9065 SNPs of the P1 animals, 1043 SNPs showed significant departure from HWE ($p<0.05$) and 1740 out of 9241 SNPs in P2 dataset. In this process, SNPs that showed departure from HWE were used in this subdivision test. The subdivision tests were applied to sire groups, herd of origin and cohort for the P1 dataset and to breed for the P2 dataset.

RESULTS AND DISCUSSIONS

Individual SNP Call Rate is the quality control indicator for an experimental assay. For the P1 dataset (Table 1), after applying CR>93%, there were 8716 SNPs remaining, and 831 of 8716 SNPs departed from HWE. When SNPs were culled with CR>93% and MAF>0.05, 5678 SNPs

remained and 751 (13.5%) of them departed from HWE ($p<0.05$). The majority of the HWD SNPs (91%) showed high reduction of heterozygosity (0.15).

Two examples illustrated in Table 2. Although having high Call Rate (99.8 and 96.0) and moderate MAF (0.14 and 0.33), SNPs A and B showed significant departure from HWE and high reduction in heterozygosity. As indicated by their F-statistics, their HWD were clearly due to subdivision caused by sire. Both SNPs showed reduction in heterozygosity (0.34 and 0.47).

Table 1. Distribution of SNPs in P1 dataset after culling against Call Rate (CR), Minor Allele Frequency (MAF) and Hardy-Weinberg Equilibrium test (χ^2 $p<0.05$). The hierarchical F-statistics were assessed against sire group, cohort and herd of origin. The modified Chi Square test showed that the departure from HWE of about 30% of SNPs was due to subdivision of sire group or 10% due to cohort or herd of origin (χ^2_F $p>0.05$)

CR	MAF	Total	HWD χ^2 ($p<0.05$)	Modified χ^2_F ($p>0.05$)
				Sire Group
0.0	0.0	9065	1043	355
93	0.0	8716	831	300
0.0	0.05	5908	930	296
93	0.05	5678	751	262
				Cohort
0.0	0.0	9065	1043	119
93	0.0	8716	831	106
0.0	0.05	5908	930	85
93	0.05	5678	751	79
				Herd of Origin
0.0	0.0	9065	1043	111
93	0.0	8716	831	100
0.0	0.05	5908	930	80
93	0.05	5678	751	76

Table 2. Examples of departure from Hardy-Weinberg Equilibration due to subdivision, illustrated using 2 SNPs with high Call Rate and moderate MAF

	Genotype 0	Genotype 1	Genotype 2	χ^2	P
SNP A					
Observed	37	97	461		
Expected	12.3	146.4	436.3	23.3	0.0001
Expected, F_{ST}(sire)=0.34	37.3	96.3	461.3	0.004	0.998
SNP B					
Observed	323	133	177		
Expected	264.8	249.5	58.8	60.53	0.000001
Expected, F_{ST}(sire)=0.34	311.2	156.7	105.2	2.78	0.24

The hierarchical F-statistics values for each SNP were assessed against sire groups, cohort or herd of origin (as shown in Table 1). Using the hierarchical F-statistics derived against sire groups in the modified χ^2 test on the 831 HWD SNPs, 531 showed significant departure from HWE ($p<0.05$), *ie.* 300 SNPs were not eliminated due to HWD ($p<0.05$). The modified Chi Square test (χ^2_F) showed that about one-third of SNPs that departed from HWE were attributable to subdivision caused by sire groups. When using the F-statistics derived from cohort or herd of

origin for P1 SNPs, an additional 106 or 100 SNPs remained, respectively. Collectively, 357 of 831 SNPs were retained. As a result, after applying HWE test using the hierarchical F-statistics 374 of 8716 SNPs were culled due to departure from HWE ($p<0.05$). This result suggested that about 10% of SNPs departed from HWE due to subdivision by cohort or herd of origin.

On examination of the P2 SNPs, the F-statistics were estimated within breed. After culling on CR and MAF 7461 SNPs remained (Table 3), 1536 of the remained SNPs (21%) departed from HWE ($p<0.05$). Similarly, the majority of these SNPs showed reduction in heterozygosity. After applied with the modified χ^2 test, 1233 of 1536 SNPs (80%) were in Hardy Weinberg Equilibrium (χ^2_F $p>0.05$).

Table 3. Distribution of SNP genotypes for the P2 animals after culling against Call Rate (CR), Minor Allele Frequency (MAF) and Hardy-Weinberg Equilibrium test (χ^2 $p<0.05$). The hierarchical F-statistics was assessed against breed. The modified χ^2 (χ^2_F) showed about 80% of SNPs departed from HWE were due to subdivision of breed (χ^2_F $p>0.05$)

CR	MAF	Total	HWD χ^2 ($p<0.05$)	F-statistics χ^2_F ($p> 0.05$)
0.0	0.0	9241	1740	1353
93	0.0	8634	1562	1242
0.0	0.05	7944	1709	1344
93	0.05	7461	1536	1233

CONCLUSIONS

The HWE test can be used to detect genotype errors. However, in populations with some sub-structure, steps should be taken to identify possible sources underlying the HWD other than genotype errors. Possible sources of subdivision could be natural grouping, management process etc. We have demonstrated that application of the modified χ^2 test using the hierarchical F-statistics can identify some SNPs with HWD due to subpopulation. In this assessment for the P1 dataset, subdivision was assessed against sire group, cohort and herd of origin. Sire group is the main source causing subdivision. Collectively, about one-third of SNPs showing HWD can be corrected by accounting for sub classification and should be attempted when data are analysed for trait-genotype associations. In the P2 dataset, when genotypes from 7 breeds were pooled as population data, significant subdivision due to breed was apparent. When HWE test is applied to this dataset, about 16% of SNPs (1233 out of 7461) could be wrongly culled because of their departure from HWE.

ACKNOWLEDGMENTS

The work is part of the Cooperative Research Centre for Beef Genetic Technologies and data was made available from the Cooperative Research Centre for Beef Genetic Technologies.

REFERENCES

Chan, E. K. F., Hawken, R. and Reverter, A. (2009) *Animal Genetics* **40**:149.

Hartl, D. L. and Clark, A. G. (2007) "Principles of Population Genetics" 5th ed. Sinauer Associates, Sunderland, MA, USA.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I. and Xu, C. F. (2004) *Eur. J. Hum. Genet.* **12**:395.

Mitchell, A. A., Cutler, D. J. and Chakravarti, A. (2003) *Am. J. Hum. Genet.* **72**(3):598.

Wright, S. (1978) "Evolution and the Genetics of Populations", Vol. 4: Variability within and among Natural Populations. University of Chicago Press, Chicago.