The combined effect of SNP-marker and phenotype attributes in genome-wide association studies

Eva KF Chan, Rachel Hawken, and Antonio Reverter

Cooperative Research Centre for Beef Genetic Technologies

CSIRO Livestock Industries, Queensland Bioscience Precinct

306 Carmody Rd., St. Lucia, QLD 4067, Australia

Corresponding author:

Antonio Reverter

Fax: +61 7 3214 2900

Ph: +61 7 3214 2392

Email: Tony.Reverter-Gomez@csiro.au

Summary

The last decade has seen rapid improvements in high-throughput SNP genotyping technologies that have consequently made genome-wide association studies (GWAS) possible. With tens to hundreds of thousands of SNP markers being tested simultaneously in GWAS, it is imperative to appropriately pre-process, or filter out, those SNPs that may lead to false associations. This paper explores the relationships between various SNP genotype and phenotype attributes and their effects on false associations. We show that *i*) uniformly-distributed ordinal data as well as binary data are more easily influenced, though not necessarily negatively, by differences in various SNP attributes compared to normally-distributed data; *ii*) filtering SNPs on minor allele frequency (MAF) and extent of Hardy-Weinberg equilibrium (HWE) deviation, has little effect on the overall false positive rate; *iii*) in some cases, filtering on MAF only serves to exclude SNPs from the analysis without reduction to the overall proportion of false associations; and iv) HWE, MAF, and heterozygosity are all dependent on minor genotype frequency, a newly proposed measure for genotype integrity.

Keywords: Genome-wide association studies, SNP, minor allele frequency, Hardy-Weinberg equilibrium, minor genotype frequency, quantitative traits, trait-distribution.

Introduction

Genome-wide association studies (GWAS) using SNP markers have become increasingly popular for dissecting the genetics of complex traits (reviewed in Hirschhorn *et al.* 2002 and McCarthy *et al.* 2008). Therefore, it is invaluable to recognise and understand how confounding factors embedded within genotypic and/or phenotypic data may lead to spurious associations. This is particularly important in GWAS because associations are tested at tens to hundreds of thousands of SNP markers, inflating the rate of false associations (type I error).

A filtering process, defined by a set of rules, is generally applied to remove markers from an analysis. The deduction of these rules may be arbitrary (e.g. Sladek *et al*. 2007; Easton *et al*. 2007) or empirical (Wellcome Trust Case Control Consortium 2007), and are typically based on various measures or attributes calculated to reflect the markers' integrity and usefulness. These attributes may include genotyping call-rate, missing data, monomorphism, loss of heterozygosity (LOH), observed heterozygosity (H_{obs}), minor allele frequency (MAF), and extent of Hardy-Weinberg equilibrium (HWE) deviations. In this paper we also propose minor genotype frequency (MGF) as a filtering criterion and explore its value as a quality control measure.

Call-rate and missing data can be used as an indicator of genotyping error and they remain the most commonly used measures of genotyping integrity (Moorhead *et al.* 2006; Di *et al.* 2005; Shen *et al.* 2005, Sladek *et al.* 2007; Easton *et al.* 2007; Shifman *et al.* 2008). Monomorphic SNPs are uninformative in genetic association studies as there is no genotypic difference. LOH (H_{obs}=0) SNPs may impact on statistical power

due to loss of information. SNPs with excessively high H_{obs} may reflect contamination and poor genotyping integrity (Teo *et al.* 2007). SNPs with low MAF have a frequency imbalance between the two allelic groups, which may in fact reflect functional importance (Cargill *et al.* 1999). SNPs deviating from HWE may confound trait-allele association as they are thought to reflect genotyping error (Clayton *et al.* 2005; Salanti *et al.* 2005); although the contrary has also been argued (Cox and Kraft 2006). Together, these warrant the need to understand the cost and benefits of filtering SNPs based on these properties.

To date, little research has been conducted using genome-wide SNP genotyping in cattle (e.g. Barendse *et al.* 2007; Khatkar *et al.* 2007; Hayes *et al.* 2007; Hayes and Goddard 2008), and only one group (Barendse *et al.* 2007) has reported a GWAS using cattle. Further, the majority of GWAS have adopted a case-control design whereby the traits of interest are binary (McCarthy *et al.* 2008). Appreciating many complex traits are continuous or ordinal, and recognising the growing attention on these traits (e.g. Weedon *et al.* 2008; Scuteri *et al.* 2007), we also focus on the effects of trait properties on GWAS. We first introduce and report on the SNP attributes of an empirical data, then we proceed to examine the combined effects of various genotype and phenotype properties on false associations in GWAS.

Methods and Materials

Samples and SNP genotype data

565 Brahman cows were genotyped at 9,075 SNPs using the MegAllele™ Genotyping Bovine 10k SNP Panel (Hardenbol *et al.* 2005). Genotyping calls were

made, as part of Affymetrix's genotyping service, using TrueCall™ Analyzer (ParAllele BioScience; Moorhead *et al.* 2006).

Partial or full parentage for 486 animals is known. They were sired by 55 bulls averaging 10 (±7.6 SD) progenies/bull (max 47 progenies/bull) and 478 dams averaging one (±0.2 SD) progeny/dam (max three progenies/dam). Kinship coefficients were estimated using pedigree information of 9,082 animals spanning up to seven generations and the *parente* program of the PEDIG package (Boichard 2002).

Simulated phenotype data

Five trait types were simulated according to the following distributions reflecting the majority of real data structures:

- *i)* continuous data with normal distribution, Normal(μ=0, σ^2=1);
- *ii)* ordered categorical data with normal distribution, Binomial(n=10, p=0.5);
- *iii)* ordered categorical data with discrete distribution, Binomial(n=10, p=X), where X~Uniform(a=0, b=1);
- *iv)* ordered categorical data with uniform distribution, Uniform(a=0, b=1);
- *v)* binary data with binomial distribution, Binomial(n=1, p=0.5).

For each trait-type, 1,000 simulations were generated under the null hypothesis of no association where in each simulation 565 random deviates were generated from the corresponding distribution.

Test for Hardy-Weinberg Equilibrium (HWE)

Deviation from HWE was assessed using the χ^2 goodness-of-fit test and Fisher's Exact test on the null hypothesis that $p^2+2pq+q^2=1$, where p and q are the two allelic

frequencies (Emigh 1980). P-values for the two tests were obtained from the χ^2 (1 d.f.) and hypergeometric distributions, respectively, as per the *pchisq*() and *fisher.test*() functions in R/stats (R Core Development Team 2007).

Genome-wide association test

Association between each trait at each polymorphic SNP was assessed using linear regression, where the simulated trait values across the 565 individuals were regressed onto the numeric code of each SNP genotype (i.e. 0, 1, or 2 copies of the alleles); this tested the null hypothesis of the additive allelic effect on the trait. Regression analyses were performed using *lm*() and P-values obtained from the F-distribution using *pf*() in R/stats. Significant associations were defined at point-wise $P<0.001$ to ensure an average of one significant (and spurious) association per SNP across the 1,000 replicates.

Test for uniform distribution of P-values

To test whether association is independent of SNP attributes, we compared, using the Kolmogorov-Smirnov (KS) test, the observed distribution of the 8,623 P-values (one from each polymorphic SNP) against the null distribution: a uniform distribution in the [0, 1] interval. P-values were obtained using the *ks.test*() function in R/stats. The median P-values from the 1,000 KS tests are 0.14±0.30SD for continuous normal traits, 0.12±0.24SD for categorical normal traits, 0.12±0.27SD for categorical discrete traits; 0.12±0.27SD for categorical uniform traits, and 0.02±0.14SD for binary traits.

Correlation tests

To ascertain the relationship between a SNP attribute and the number of FPs, Spearman's correlation coefficients (ρ) were calculated. Significant correlation was only asserted if $|\rho| \geq 0.1$ at $P<0.05$ (two-sided test against the null that $\rho=0$). As per the *cor.test*() function in R/stats, P-values were computed using the AS 89 algorithm.

The null hypothesis of the numbers of SNPs across eight FP bins (FP=0, 1, 2, 3, 4, 5, 6-10, >10) are the same between the "good" and "bad" SNP sets was tested for each trait type using Pearson's χ^2 test with P-values obtained from 10,000 permutations using *chisq.test*() in R/stats.

Two tests were used for comparing the distributions of the same SNP attribute between FP-free (FP=0) and FP-prone (FP≥4) SNPs: *i*) Pearson's χ^2 test for LOH; and *ii*) Mann-Whitney test for all other (non-binary) SNP attributes. P-values for the χ^2 test were determined from 10,000 simulations using *chisq.test*() and those for the Mann-Whitney tests were approximated from a Gaussian distribution using *wilcox.test*() in R/stats.

Results

SNP attributes

Each SNP has a median call-rate of 99.8% (85% – 100%), a median of one (range: 0 – 90) missing genotype, and 5% of SNPs are monomorphic. Excluding monomorphic SNPs, $H_{obs}=0.21\pm0.17$, of which 0.4% (33/8623) have LOH ($H_{obs}=0$).

In this paper we introduce and examine the effects of MGF on GWAS. The necessity to include MGF in addition to MAF is justified because SNPs with low MGF do not

always imply low MAF (Fig. 1). An extreme example is LOH; of the 33 LOH SNPs, two have MAF>0.4 suggesting equal selection pressure on the two homozygotes. Furthermore, the inclusion of MGF in addition to test of HWE is because SNPs with low MGF do not necessarily deviate from HWE, as in the case when the minor genotype is one of the homozygotes (Fig. 1 insert). Of the 638 SNPs with 0<MGF<0.002 (averaging only one individual harbouring the minor genotype), 507 (79.5%) are in HWE.

MAF is 0.10±0.14 SD across all SNPs and MGF is 0.05±0.07 SD, with the former figure increasing to 0.16±0.14 SD following the exclusion of monomorphic markers, but remains unchanged for MGF. Depending on the test statistic and associated criteria, between 13.6% (Fisher's Exact test at $P<0.0001$ for autosomal SNPs with MAF≥0.05 as in Khatkar *et al.* (2007)) and 23.6% (Pearson's χ^2 test at $P<0.05$ for autosomal SNPs with at least five expected samples per genotypic group as in Barendse *et al.* (2007)) SNPs deviate from HWE. Our notably left skewed MAF distribution (relative to that reported in Barendse *et al.* (2007)) and large numbers of HWE deviations are attributed to the elevated shared ancestry within our samples: average kinship coefficient is 0.020±0.024 SD. In this paper, we use this to our advantage to explore the effect of HWE deviation on the extent of type I errors.

Effects of SNP and phenotypic attributes on GWAS

We examined the effects of SNP attributes on type I errors in GWAS in consideration of five types of phenotypic traits. As we are interested in the extent of false associations, we chose to simulate these traits under the null hypothesis of no

association: traits were purely simulated under the specified distribution independent of the animals and their genotypes; i.e. no genetic structure was simulated.

Extent of False Associations

Under our null hypothesis, two observations are expected: *i*) P-value distributions should be uniform for each GWAS (i.e. each simulated trait); and *ii*) an average of one false positive (FP) should be observed for each SNP. Here, FP is the number of 1,000 simulated traits passing the significance threshold of $P<0.001$, thus each SNP is expected to falsely associate with one of the 1,000 simulated traits by chance alone (FP=1).

The first expectation is satisfied by four trait-types; only simulated binary traits have P-values that are significantly non-uniform (median P=0.02 for tests of uniformity), signifying an increased sensitivity of binary traits to various SNP attributes. The second expectation is satisfied by all but categorical-discrete traits (Figure 2 top panel): instead of the majority of SNPs having FP=1, only 10% SNPs complied, while >78% show no significant association (FP=0).

What SNP properties affect FP?

To identify SNP attributes that may influence false associations, we assessed the level of correlations between FP and each SNP attributes. Here, significant correlation is only asserted if $|\rho|\geq 0.1$ and corresponding $P<0.01$. Results show only significant correlations for categorical-uniform and binary traits (Table 1).

The extent of false associations is not affected by call-rate, missing data, or LOH. It is however significantly affected by H_{obs} for categorical-uniform ($\rho \approx 0.2$) and binary ($\rho \approx 0.3$) traits. Due to the relationships between MAF, MGF, and Ho (MAF=x+½Ho, where $0 \leq x \leq 1$; MAF≥MGF×1.5), FPs are also significantly influenced by MAF and MGF with $0.16 \leq \rho \leq 0.28$ for categorical-uniform and binary traits.

Can filtering of SNPs reduce the extent of FPs?

Significant correlations between FP and various SNP attributes suggest FP should decrease if problematic, or "bad", SNPs are eliminated prior to association. Here we assess this by comparing the extents of FPs from "good" and "bad" SNPs. As our objective is to investigate the impact of various SNP attributes on false associations, our null hypothesis here was that the extent of false positives are equal between the set of "good" and "bad" SNPs. In GWAS, SNPs are commonly excluded based on several criteria that generally reflect their informaitveness and level of variation. These criteria are variable in the literatures and for the purpose of this study, "good" SNPs are defined as those passing the following set of criteria derived from recent literatures:

- Call-rate ≥ 95% (e.g. Sladek *et al.* 2007; Easton *et al.* 2007; Shifman *et al.* 2008);
- MAF ≥ 0.01 (e.g. Sladek *et al.* 2007); and
- HWE P ≥ 0.001 (e.g. Sladek *et al.* 2007; Shifman *et al.* 2008; Cupples *et al.* 2007).

These criteria classified 25% of polymorphic SNPs as "bad" and 75% as "good".

The extent of false associations between "good" and "bad" SNPs is not significantly different ($P>0.05$; Fig. 2) for continuous-normal traits. Conversely, and paradoxically, the proportion of "good" SNPs with FP=0 is lower compared to "bad" SNPs (Fig. 2 bottom two panels) for the remaining four trait-types, suggesting "bad" SNPs are *less* vulnerable to spurious associations. This phenomenon extends to FP>0: there is significant difference in the proportion "good" and "bad" SNPs across the eight FP bins ($P<0.01$) for all but continuous-normal traits. In particular, >59% of "bad" SNPs have FP=0 for categorical-uniform traits and <40% of the "good" SNPs have FP=0 for binary traits.

Retropectively, these results are unsurprising as "bad" SNPs are more uninformative than "good" SNPs and so should be more likely to incur false negatives. Yet, as our interest is in false positives, these results raise the question of which of the SNP attributes, if any, can "protect" against FP. To address this we compared various attributes of SNPs that are FP-free (FP=0) and FP-prone (FP≥4) and found (Table 2): *i*) FP-prone SNPs have significantly higher frequencies (but not the absence) of heterozygotes compared to FP-free SNPs in non-normally distributed traits, *ii*) FP-prone SNPs have significantly higher MAF and MGF for all but continuous-normal traits, and *iii*) many more FP-free SNPs have MGF=0 (35%-58%) compared to FP-prone SNPs (10%-19%). These observations suggest low Ho, MAF, or MGF can limit false associations, particularly for ordinal and binary traits.

FP-prone SNPs deviate from HWE more often than FP-free SNPs, but this difference is abolished when we exclude 49% SNPs with at least one of the three genotypes represented by less than five (or <1%) individuals. We infer from this that deviation

from HWE alone does not affect false associations, rather FP is dependent on low MGF-induced HWE deviation. Again, continuous-normal traits appear unaffected by this.

Trade-off between reduction in false positives and loss of SNPs

Finally, we explored the trade-off between the amount of FP we can reduce and the number of useful SNPs we can retain. In particular, we examined the effects of all three-way combinations of MAF, MGF, and HWE, threshold at:

- MGF≥0, MGF>0, MGF>0.005, MGF>0.01, MGF>0.05, MGF>0.1;
- MAF>0, MAF>0.005, MAF>0.01, MAF>0.05, MAF>0.1; and
- HWE: P≥0, P>10^{-6}, P>10^{-5}, P>10^{-4}, P>10^{-3}, P>10^{-2}, P>0.05.

For most traits the rate of FP reduction is proportional to the rate of SNP loss (Fig. 3): i.e. removing *x*% of the SNPs removes ~*x*% of FP. This is particularly true for continuous-normal traits reaffirming the loss (and gain) of FP is random and thus proportional to the number of SNPs excluded from analysis.

However, for binary, categorical-discrete, and categorical-uniform traits, some (combination of) SNP filtration criteria result in more rapid SNP loss than FP loss. Specifically, increase in MAF stringency only serves to increase the number of excluded SNPs but does not reduce the extent of false associations (Fig.3: shift of data-points above line of negative unity with increasing MAF stringency). And finally, we show that the reduction in SNPs (and FPs) is more rapid from no filtration on MGF (circle) to MGF≤0.05 (upside-down triangle) compare to no filtration on HWE deviation (smallest circle) to deviation at P≤0.05 (largest circle).

Discussion

Association studies are based on the fundamental assumption that the genetic variants underlying a phenotypic trait will co-segregate with the trait of interest in a given population. The statistical analyses are thus aimed at identifying the markers whose genotypes correlate best with the trait values across a population of individuals. Clearly, factors affecting the characteristics of either or both the phenotypic or genotypic data can severely affect the power and accuracy of detection.

In this paper, we showed that some, but not all, of the examined SNP-attributes can influence spurious associations, and the effect is not always negative and certainly not applicable to all trait-types. In particular, none of the SNP attributes appear to have major effects on normally-distributed traits, be it continuous or ordered-categorical (Table 1). Only when we compare attributes of FP-prone and FP-free SNPs do we notice the effects of several SNP attributes on false associations of the latter trait-type (Table 2).

One such attribute is MGF. The influence of zero or near zero MGF is not limited to categorical-normal traits and its effect is, surprisingly, not negative in regards to type I error. Repeatedly we showed that SNPs with low MGF tend to have fewer false associations across all trait-types. This ironically is a consequence of reduced statistical power in association tests which would prevent, or reduce, true as well as false associations. Thus, although we have shown that SNPs with zero or near zero MGF tend to protect against false associations, we suspect it would conversely inflate false negatives (type II error).

Additionally, and in some cases as a consequence of, low to zero MGF, low MAF, low H_{obs} and deviation from HWE can also protect against false associations; this is especially true for categorical-uniform and binary traits. Again, this is because SNPs with these attributes are susceptible to false negatives. In the case of deviation from HWE, and possibly for low H_{obs} and MAF, its effect is only manifested when the corresponding SNP also has near zero MGF. In fact, we failed to establish any connection between deviation from HWE and false associations with any trait-type for SNPs with MGF<0.009 (corresponding to fewer than five individuals per genotype). This finding is of particular importance in GWAS because deviation from HWE is a widely used SNP quality control measure.

While HWE deviation-induced FP for binary traits have been noted previously (Schaid and Jacobsen 1999), we have further demonstrated that the effect extends to categorical-uniform traits and that the effect is likely restricted to low MGF-induced HWE deviation. Moreover, while LOH (H_{obs}=0) markers with sufficiently low MAF to escape detection from HWE deviation have been shown to cause false associations in transmission-disequilibrium tests (Hirschhorn and Daly 2005), here we demonstrated the effect of near zero H_{obs} is only a subclass of the larger problem of near zero MGF in GWAS. For this reason, we strongly advise HWE deviation be used with caution or in conjunction with MGF as an inclusion/exclusion measure for genetic association studies.

To allow for easy comparison of the effects of genotype attributes on different trait-types, we have chosen to use a linear regression model for test of association for all

trait-types. This is generally acceptable for quantitative traits are which are either normal or can be transformed to normality (e.g. Scuteri *et al.* 2007). However, this is not applicable to truly non-normal data; and for this reason, such data type can, retrospectively, be more prone to type I errors. We have shown this to be particularly true for binary and uniformly-distributed ordinal traits due to the relative increased probability of sampling from the tails of these distributions. For binary traits, alternate association test methods such as logistic regression (e.g. WTCCC 2007) and Cochran-Armitage test (e.g. Fellay *et al*. 2007) are well-developed and commonly adopted. Conversely, there is little research into more appropriate methods for analysing ordinal and non-normally distributed traits. With the increasing popularity of GWAS, perhaps it is time for the community to divert more attention to this area.

Finally, two technical points are of note here. First, although we recognise the genotype data used in this study are from one cattle population with its inherent family structure, the relationship between SNP and phenotypic attributes and their effects on spurious genetic associations are population independent and thus should be applicable to other (non-cattle) populations. For example, although this population demonstrated a relatively low MAF across all SNPs (32% polymorphic SNPs with MAF<0.05), the only difference compare to a population with a higher average MAF is the extent of FP but the nature of the effect of low MAF and the fact that the effect would be more prominent for categorical-uniform and binary traits are indisputable. Clearly, in order to make inference on statistical power and type II error, one would have to model family structure into the phenotype data and then account for it in the association test (e.g. Marchini *at al.* 2004).

Secondly, several studies have claimed that genotyping error can confound association studies due to distortion of allele frequencies (e.g. Gomes *et al.* 1999, Hosking *et al.* 2004, and Salanti *et al*. 2005). Although we did not find any effect of genotyping call-rate and genotyping failure (missing data) on GWAS, we acknowledge these are not true measures of genotyping accuracy. These measures are highly dependent on the genotyping platform, corresponding genotype-calling algorithm, and their inherent limitations (Hardenbol *et al.* 2005). Thus, it is unclear whether a more accurate measure of genotyping call-rate that is more reflective of genotyping error would reveal significant impact on GWAS; again, further study is needed.

In conclusion, we emphasise that whether a SNP is FP-free or FP-prone is highly dependent on H_{obs}, MAF, and MGF, as well as the characteristic and distribution of the trait in which the SNP is to be tested against. Furthermore, a SNP that is FP-free does not necessarily imply it will be more efficient in a test of association because the FP-free nature may simply be a reflection of the SNP's inherent lack of statistical power for such a purpose.

Acknowledgements

We thank Peter Visscher and CRC for Beef Genetic Technologies Underpinning Science Committee for their inputs into this manuscript. We also like to thank two anonymous reviewers and Dr. De Koning for their insightful inputs. Genotyping of the 565 Brahman samples was funded through the CRC for Beef Genetic Technologies.

References

Barendse W., Reverter A., Bunch R.J., Harrison B.E., *et al.* (2007) A validated whole-genome association study of efficient food conversion in cattle. *Genetics* **176**(3): 1893 – 1905.

Boichard D. (2002) Pedig: a Fortran package for pedigree analysis suited to large populations. *7th World Congress on Genetics Applied to Livestock Production, Montpellier*, 19-23 August 2002, paper 28-13.

Cargill M., Altshuler D., Ireland J., Sklar P., *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* **22**(3), 231 – 238.

Clayton D.G., Walker N.M., Smyth D.J., Pask R., *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37**: 1243 – 1246.

Cox D.G. and Kraft P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Human Heredity* **61**(1): 10 – 14.

Cupples L.A., Arruda H.T., Benjamin E.J., D'Agostino R.B. *et al.* (2007) The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Medical Genetics* **8 Suppl 1**:S1.

Di X., Matsuzaki H., Webster T.A., Hubbell E., *et al.* (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics* **21**(9): 1958 – 1963.

Easton D.F., Pooley K.A., Dunning A.M., Pharoah P.D., *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**: 1087-1093.

Emigh T.H. (1980) A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* **36**: 627 – 642.

Gomes I., Collins A., Lonjou C., Thomas N.S., *et al.* (1999) Hardy-Weinberg quality control. *Annals of Human Genetics* **63**: 535 – 538.

Hardenbol P., Yu F., Belmont J., Mackenzie J., *et al.* (2005) Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Research* **15**(2): 269 – 275.

Hayes B.J., Chamberlain A.J., McPartlan H., Macleod I., *et al.* (2007) Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genetical Research* **89**(4): 215 – 220.

Hayes B.J. and Goddard M.E. (2008) Technical note: prediction of breeding values using marker derived relationship matrices. *Journal of Animal Science* **AOP** doi: 10.2527/jas.2007-0733.

Hirschhorn J.N., Lohmueller K., Byrne E., Hirschhorn K. (2002) A comprehensive review of genetic association studies. *Genetics in Medicine* **4**(2): 45 – 61.

Hirschhorn J.N. and Daly M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics* **6**: 95 – 108.

Hosking L., Lumsden S., Lewis K., Yeo A., *et al.* (2004) Detection of genotyping errors by Hardy–Weinberg equilibrium testing. *European Journal of Human Genetics* **12**, 395 – 399.

Khatkar M.S., Zenger K.R., Hobbs M., Hawken R.J., *et al.* (2007) A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in holstein-friesian cattle. *Genetics* **176**(2): 763 – 772.

Marchini J., Cardon L.R., Phillips M.S., & Donnelly P. (2004) The effects of human population structure on large genetic association studies. *Nature Genetics* **36**: 512 – 517.

McCarthy M.I., Abecasis G.R., Cardon L.R., Goldstein D.B., *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics* **9**(5): 356 – 369.

Moorhead M., Hardenbol P., Siddiqui F., Falkowski M., *et al.* (2006) Optimal genotype determination in highly multiplexed SNP data. *European Journal of Human Genetics* **14**: 207 – 215.

R Development Core Team (2007) R version 2.5.1: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Salanti G., Amountza G., Ntzani E.E., and Ioannidis J.P. (2005) Hardy-Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *European Journal of Human Genetics* **13**(7), 840 – 848.

Schaid D.J. and Jacobsen S.J. (1999) Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg Proportions. *American Journal of Epidemiology* **149**: 706 – 711.

Scuteri A., Sanna S., Chen W.M., Uda M., *et al.* (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genettics* **3**(7): e115.

Shen R., Fan J.B., Campbell D., Chang W., *et al.* (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research* **573**: 70 – 82.

Shifman S., Johannesson M., Bronstein M., Chen S.X., *et al.* (2008) Genome-Wide Association Identifies a Common Variant in the Reelin Gene That Increases the Risk of Schizophrenia Only in Women. *PLoS Genetics* **4**(2): e28.

Sladek R., Rocheleau G., Rung J., Dina C., *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**: 881 – 885.

Teo Y.Y., Inouye M., Small K.S., Gwilliam R., *et al.* (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**(20):2741-2746.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661 – 678.

Weedon M.N., Lango H., Lindgren C.M., Wallace C., *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics* **40**(5): 575 – 583.

Legends

Figure Legends

Figure 1 Relationship between minor genotype frequency (MGF) and minor allele frequency (MAF) for 9,075 SNPs from 565 individuals.

Figure 2 Proportions of SNPs with the corresponding number of false associations for the five trait-types. Shown are the proportions of all SNPs (top), "good" SNPs (middle), and "bad" SNPs (bottom).

Figure 3 Rates of reduction in the proportion of false associations to the proportion of excluded SNPs at various combinations of MAF, MGF, and HWE deviation thresholds.

Table Legends

Table 1 Spearman's ρ correlation between number of false associations (FP) and various SNP attributes.

Table 2 The significance of testing the null hypothesis of no difference between FP-free (FP=0) and FP-prone (FP≥4) SNPs.

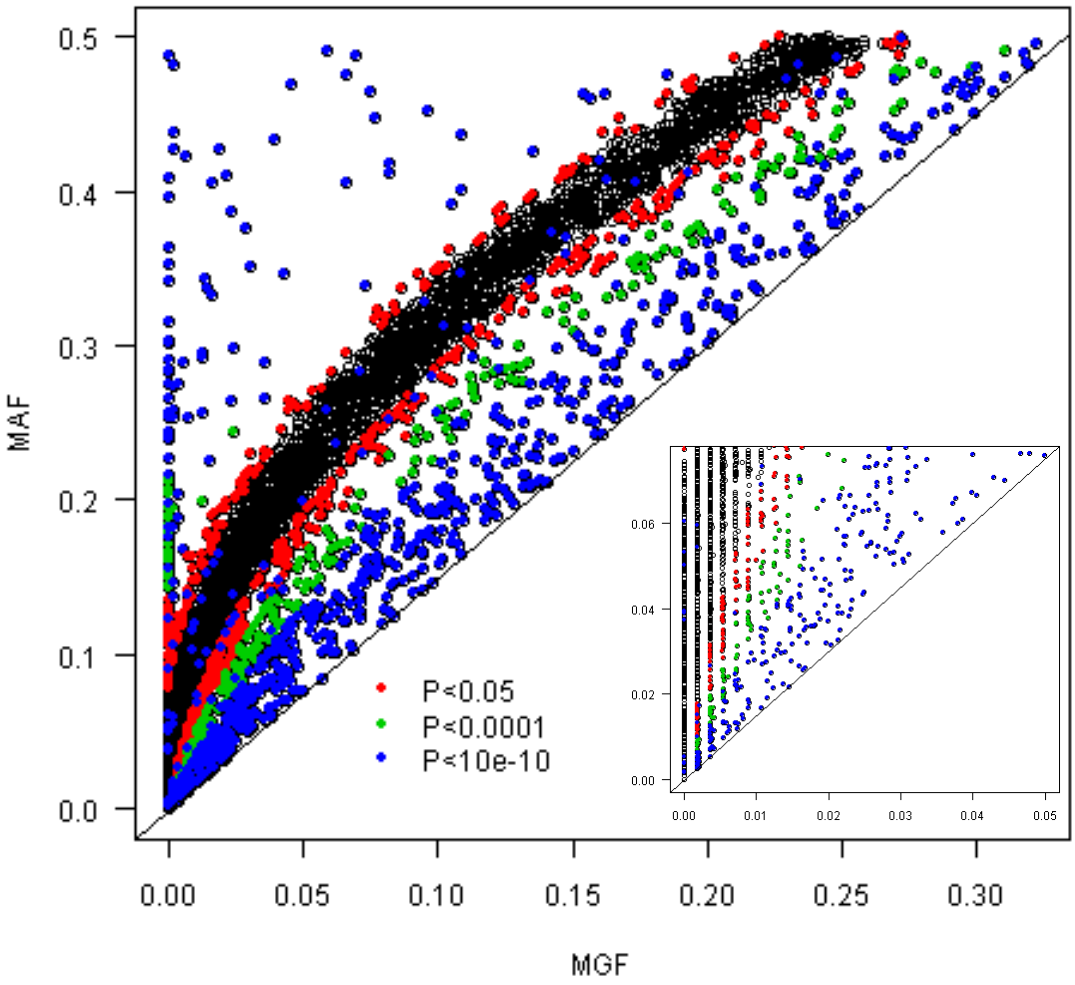

Figure 1 Relationship between minor genotype frequency (MGF) and minor allele frequency (MAF) for 9,075 SNPs from 565 individuals. Shaded in colour are the SNPs deviating from HWE at various P-values using Pearson's χ^2 test. The black line satisfies MAF = 1.5×MGF, the lower-bound for MAF. Insert is a zoomed-in of the bottom left corner.

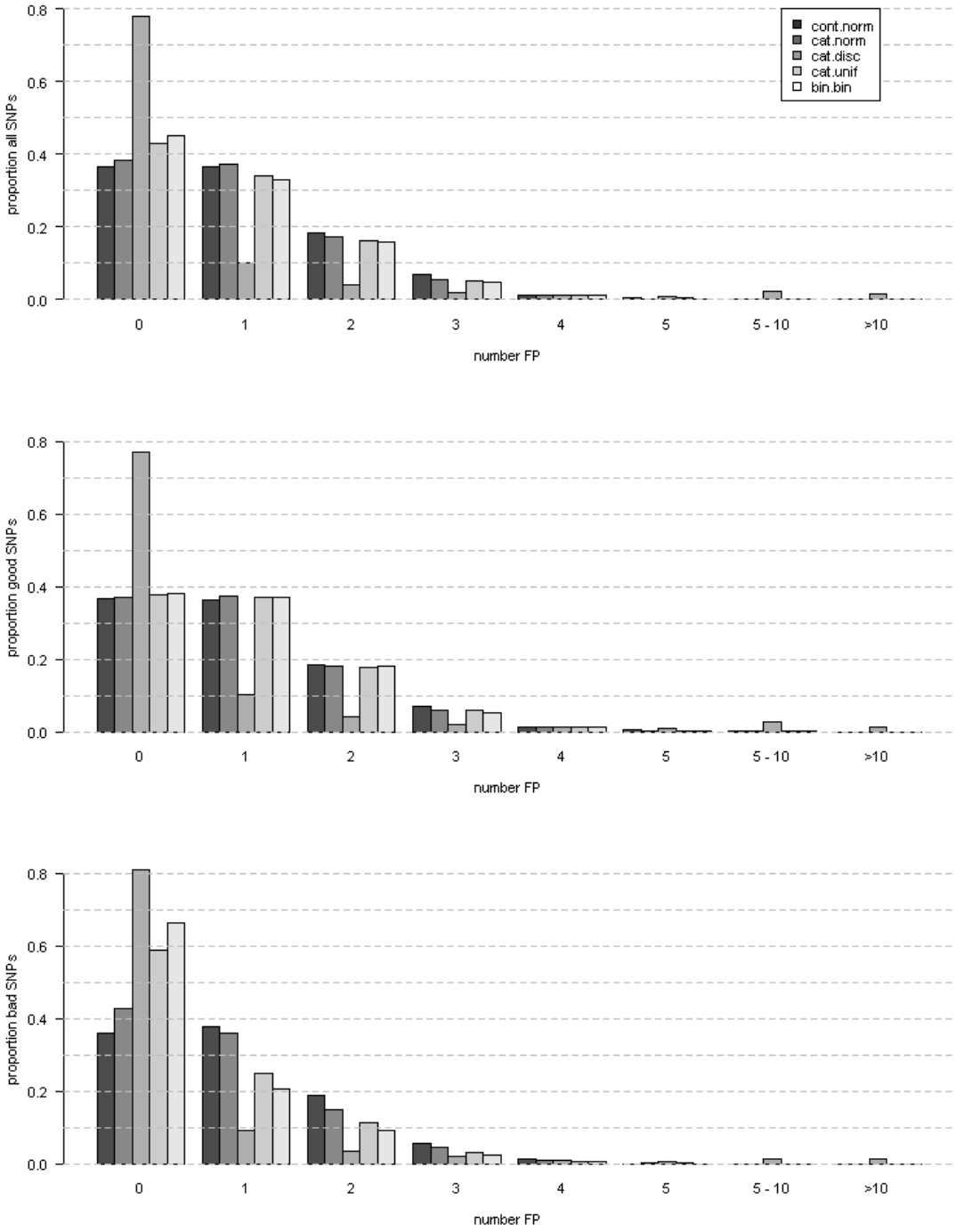

Figure 2 Proportions of SNPs with the corresponding number of false associations for the five trait-types. Shown are the proportions of all SNPs (top), "good" SNPs (middle), and "bad" SNPs (bottom). The five types of quantitative traits are: normally-distributed continuous data (cont norm), normally-distributed ordered-categorical data (cat norm), discretely-distributed ordered-categorical data (cat disc), uniformly-distributed ordered-categorical data (cat unif), and binomially-distributed binary data (bin.bin).

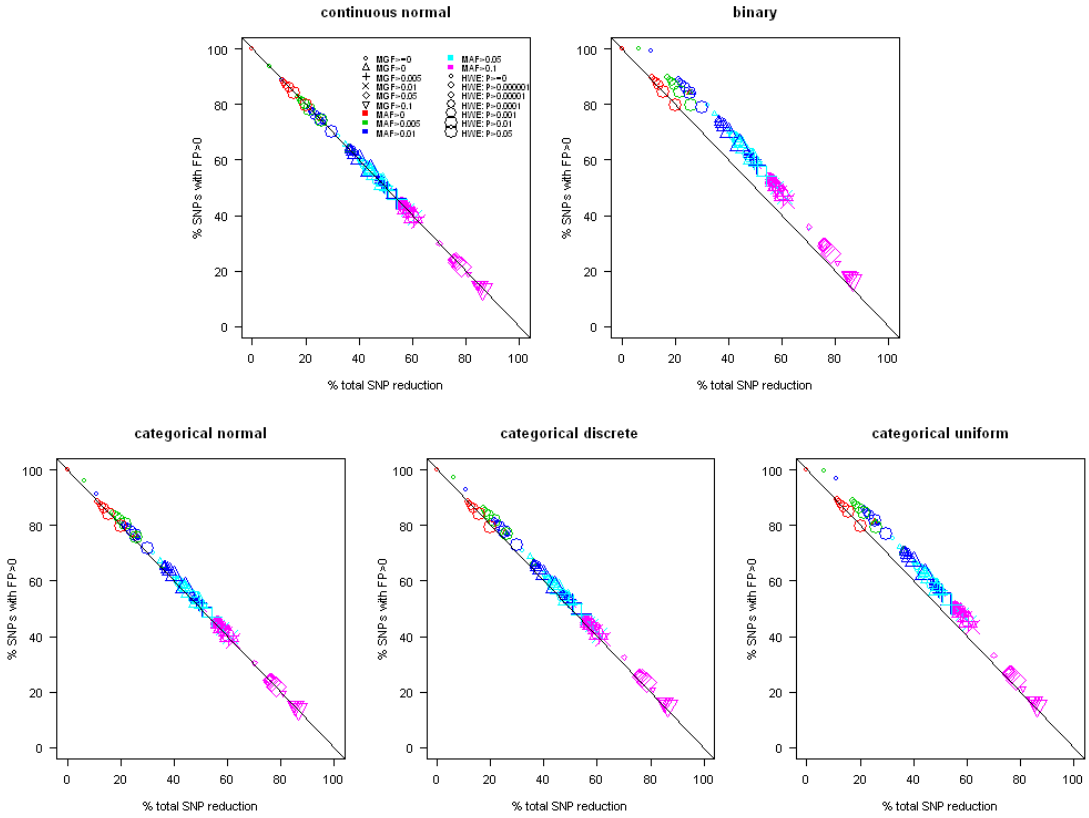

Figure 3 Rates of reduction in the proportion of false associations to the proportion of excluded SNPs at various combinations of MAF, MGF, and HWE deviation thresholds. Filtration on MGF is indicated by different plotting symbols (circle: no filtration on MGF, triangle: MGF>0, plus: MGF>0.005, cross: MGF>0.01, diamond: MGF>0.05, inverse triangle: MGF>0.01), filtration on MAF is indicated by different colours (red: polymorphic (MAF>0), green: MAF>0.005, blue: MAF>0.01, cyan: MAF>0.05, magenta: MAF>0.01), and filtration on HWE deviation is indicated by different plotting sizes (smallest: no filtration on HWE deviation, to largest: P>0.05). Black line indicates line of unity.

Table 1 Spearman's ρ correlation between number of false associations (FP) and various SNP attributes. Only correlations where either $|\rho| \geq 0.1$ or the corresponding $P < 0.05$ are shown, otherwise '–' is indicated, and only when both criteria are satisfied is significance asserted (**bold**). For test of HWE, the χ^2 test was used for all SNPs, and Fisher's Exact test was used only on SNPs with $n \geq 5$.

SNP attributes	Continuous normal	Categorical normal	Categorical discrete	Categorical uniform	Binary
Call-rate	–	–	–	–	–
Missing values	–	–	–	–	–
LOH	–	–	$\|\rho\|<0.1$ ($P=0.009$)	–	–
Ho	–	$\|\rho\|<0.1$ ($P<10^{-7}$)	$\|\rho\|<0.1$ ($P<10^{-4}$)	**$\rho=0.20$ ($P<10^{-16}$)**	**$\rho=0.29$ ($P<10^{-16}$)**
MAF	–	$\|\rho\|<0.1$ ($P<10^{-7}$)	$\|\rho\|<0.1$ ($P<10^{-4}$)	**$\rho=0.20$ ($P<10^{-16}$)**	**$\rho=0.28$ ($P<10^{-16}$)**
MGF	–	$\|\rho\|<0.1$ ($P<10^{-6}$)	$\|\rho\|<0.1$ ($P<10^{-3}$)	**$\rho=0.16$ ($P<10^{-16}$)**	**$\rho=0.23$ ($P<10^{-16}$)**
HWE: χ^2 statistic	–	$\|\rho\|<0.1$ ($P<10^{-4}$)	$\|\rho\|<0.1$ ($P=0.017$)	**$\rho=0.11$ ($P<10^{-16}$)**	**$\rho=0.12$ ($P<10^{-16}$)**
HWE: Fisher's odds ratio	–	–	–	–	$\|\rho\|<0.1$ ($P<10^{-4}$)

Table 2 The significance of testing the null hypothesis of no difference between FP-free (FP=0) and FP-prone (FP≥4) SNPs. Only significant differences (P<0.05) are shown, otherwise, '–' is indicated. "higher" & "lower" in parentheses indicate if the distributions are right or left shifted, respectively, in FP-prone compare to FP-free SNPs. For test of HWE, the χ^2 test was used for all SNPs, and Fisher's Exact test was used only on SNPs with n≥5.

FP=0 vs. FP≥4	**Continuous normal**	**Categorical normal**	**Categorical discrete**	**Categorical uniform**	**Binary**
Call-rate	–	–	–	–	–
Number of missing	–	–	–	–	–
LOH	–	–	–	–	–
Ho	–	–	$<10^{-3}$ (higher)	$<10^{-9}$ (higher)	$<10^{-16}$ (higher)
MAF	–	0.023 (higher)	$<10^{-3}$ (higher)	$<10^{-10}$ (higher)	$<10^{-16}$ (higher)
MGF	–	–	0.001 (higher)	$<10^{-8}$ (higher)	$<10^{-13}$ (higher)
MGF=0	–	0.001 (lower)	$<10^{-3}$ (lower)	$<10^{-4}$ (lower)	$<10^{-4}$ (lower)
HWE: χ^2 test	–	0.007 (higher)	0.008 (higher)	0.017 (higher)	0.009 (higher)
HWE: Fisher's Exact test	–	–	–	–	–