A Comparison of Association Statistics between Pooled and Individual Genotypes

Jo Knight^a Scott F. Saccone^b Zhehao Zhang^b Dennis G. Ballinger^c
John P. Rice^b

^aSocial Genetic & Developmental Psychiatry MRC Centre, Institute of Psychiatry, Kings College London, London, UK; ^bDepartment of Psychiatry, Washington University School of Medicine, St Louis, Mo.,
^cPerlegen Sciences Inc, Mountain View, Calif., USA

Key Words

Pooling • Power • Association

Abstract

Background: Markers for individual genotyping can be selected using quantitative genotyping of pooled DNA. This strategy saves time and money. **Methods:** To determine the efficacy of this approach, we investigated the bivariate distribution of association test statistics from pooled and individual genotypes. We used a sample of approximately 1,000 samples with individual and pooled genotyping on 40,000 SNPs. **Results and Conclusions:** We found that the distribution of the joint test statistics can be modelled as a mixture of two bivariate normal distributions. One distribution has a correlation of zero, and is probably due to SNPs whose pooled genotyping was unsuccessful. The other distribution has a correlation of approximately 0.65 in our data. This latter distribution is probably accounted for by SNPs whose pooled genotyping accurately predicts the underlying allele frequency. Approximately 87% of the data belongs to this distribution. We also derived a method to investigate the effect of both the correlation and selection cut-off on the relative power of pooling studies. We demonstrate that pooled genotyping has good power to detect SNPs that are truly associated with disease-causing variants for SNPs showing good correlation between pooled and individual genotyping. Therefore, this approach is a cost effective tool for association studies.

Copyright © 2009 S. Karger AG, Basel

Introduction

DNA pooling has been used as a cost effective method for screening markers for more than 20 years [1]. Proof of principle studies has been undertaken and has demonstrated the accuracy of the estimation of allele frequencies in pooled DNA on platforms that allow researchers to simultaneously analyse hundreds of thousands of markers [2–6]. The relative efficiency of different platforms has also been compared [7]. Furthermore, promising results have been gained from two-stage case control studies of a number of different diseases which have been undertaken using pooled DNA to screen for interesting markers and individual genotyping to confirm results [8–17]. It is known that measurement error from pooled studies is higher than measurement error in individual studies, and that differential hybridisation can make allele frequency estimation more difficult, but continuing efforts are being made to improve on the accuracy of the data [6, 18, 19].

There have also been a number of papers that have attempted to establish the efficiency of pooled DNA analysis. The study of Barratt et al. [20] was one of the earliest: they used estimates of the different variances to approximate the effective sample size. Although important work is continuing into the nature of the variation underlying these studies, these parameters are platform dependent and will continue to change as platforms are modified. More recently, Prentice and Qi [21] have developed meth-

ods to consider the efficiency of pooling based on the ability to identify SNPs with a certain odds ratio. Pearson et al. [22] looked at the efficiency of pooling using simulation to deal with the difficulties of accounting for linkage disequilibrium (LD).

In this paper we investigate the relationship between the association test statistics derived from the quantitative genotypes from pools and individual genotypes. We fit a mixture of bivariate normal distributions to the joint statistics and determine the correlation and relative proportions of these distributions. We go on to investigate the relative power of pooling studies to correctly identify markers that show association in individual genotyping studies. In addition, we investigate the effect of the cut-off used for marker selection, the correlation between the association statistics from the two types of study, and the p value of the marker in the individual genotyping study.

Methods

Approximately 2.4 million SNPs were genotyped in pools by Perlegen Sciences using custom Affymetrix chips. The SNPs were divided into 49 chips with about 50,000 SNPs on each chip. The sample of 482 cases and 466 controls was split into 8 case and 8 control pools containing about 60 people each. Details of pool construction, pooled allele frequency estimation, association tests of the pooled data, SNP selection, and genotyping methodology for the individual genotypes, are outlined in the paper by Bierut et al. [14]; hence only a summary of this information is provided here.

Pool Construction

DNA concentration of each individual was normalised and equimolar amounts were included in the pools.

Allele Frequency Calculation and Significance Testing in Pooled Sample

Allele frequencies were determined from the ratio of the intensity of the reference strand to the sum of intensities from both the reference and alternate strand. Trimmed means from perfect-match (PM) and mis-match (MM) intensities were both utilised. The three quality control measures used were designed to ensure that (i) the PM has a higher intensity signal than the MM, (ii) the signal to noise ratio was acceptably low, and (iii) no features were saturated.

Corrected t tests were used to evaluate association. This method involved correcting the standard error by a chip-specific additive constant to ensure there was not biased selection of SNPs with small standard errors. Empirical p values were calculated using the rank of the corrected t test value divided by the number of successful markers on the chip.

SNP Selection for Individual Genotyping

Only those SNPs in which at least two pools of both the cases and controls had allele frequency estimates that met the quality

control standard required were considered for selection. The main selection criteria was a p value below 0.05; however, some SNPs were selected for population stratification and quality control. 41,402 SNPs were individually genotyped: 39,213 of these were selected from the pools and 2,189 were selected for QC and population stratification. After QC we were left with 29,314.

Individual Genotyping and Significance Tests

Genotyping was performed with a clustering technique which was a combination of K-means and multiple linear regression. There was a stringent quality control procedure for the individually genotyped data. The call rate had to be above 80%, HWE p values from both cases and controls were considered, and extra checks relating to the X and Y chromosomes were undertaken. Furthermore, a metric was derived from 15 inputs relating to SNP quality and genotype, which correlated with the probability of having a discordant call (this was calculated using a non-Perlegen HapMap project genotypes). Monomorphic SNPs were also excluded from the analysis. Allele-wise analysis was carried out to derive p values for association.

Investigation of Test Statistics

We examined the relationship between the association test statistics rather than the allele frequencies. While some investigators prefer to study the allele frequency, we believe the association test statistics is more relevant for determining the efficacy of using pooled versus individual genotyping for an association study. That is, if a SNP would have been significant under individual genotyping, would it meet the pooling threshold for individual genotyping and thus be detected in a pooling design? This is the focus of our analyses.

For each p value resulting from the tests carried out on both the quantitative genotypes derived from the pools and the individual genotypes, we determined the value of the standard normal deviant for a two-sided test (X_{pool} and X_{igt}). We ensured that the direction of deviation was comparable for the pooled and individual genotyping.

After selecting only those SNPs that had been genotyped using both pooled DNA and individual DNA, we fitted a mixture of bivariate normal distributions to the distribution of the joint X statistics. One of the bivariate normal distributions – defined as α – was fixed to have a correlation of zero because it would comprise of data from ‘bad’ SNPs, i.e. SNPs with allele frequencies that can not be accurately estimated from pools.

For the proportion $(1 - \alpha)$ of ‘good’ SNPs we assumed that the joint distribution ($X_{\text{pool}}, X_{\text{igt}} | X_{\text{pool}} \in S$), where S is the set of SNPs chosen for individual genotyping. The density for this distribution is $K\phi(r)$, where the constant K does not depend on the correlation r and corresponds to the selected area under the marginal density of X_{pool} .

However, the statistics are only available from SNPs selected for individual genotyping ($X_{\text{pool}}, X_{\text{igt}} | X_{\text{pool}} \in S$), where S is the set of SNPs chosen for individual genotyping. The density for this distribution is $K\phi(r)$, where the constant K does not depend on the correlation r and corresponds to the selected area under the marginal density of X_{pool} .

The final likelihood is given by:

$$(1 - \alpha) K\varphi(r) + \alpha K\varphi(r = 0).$$

The parameters r and α may then be estimated by maximizing the likelihood over the observations.

Investigation of the Relative Power of Pooling

Established characteristics of both standard and bivariate normal distributions can be used to determine the ability to identify markers that would demonstrate association in an individual genotyping study from markers genotyped in pools. Essentially, we calculated the probability of a marker being selected for the individual genotyping study given its expected standard deviate in that study (X_{igt}), as well as the cut-off used for marker selection (t) from the pooled study and the correlation between the distributions of the test statistics from the pooled and individual genotyping results (r).

$$P(X_{pool} > t) = \varphi\left(\frac{t - rX_{igt}}{\sqrt{1 - r^2}}\right)$$

We derived the expected standard deviate of t given both X_{igt} and the correlation of the distributions. We then determined the probability of X_{pool} being above this value.

Results

In the study by Bierut et al. [14] the correlation of the allele frequency estimates had been calculated as 0.87 for the cases and 0.84 for the controls. The distribution of X_{pool} was approximately normal, as expected. It had a mean of 0.019 and a standard deviation of 1.000.

We demonstrated that the bivariate distribution of the joint test statistics from the case-control tests of pooled and individual data can be modelled as a mixture of normal distributions with proportions α and $1 - \alpha$ ($n = 29,314$). In our data the proportion α with correlation zero equalled 0.133 and the correlation of the $1 - \alpha$ proportion of the distribution was estimated to be 0.65 (± 0.0022). We propose that some markers are not typed successfully in pools and it is the results from these that comprise the 13% of the data that has no correlation. Such markers pass QC of the pooled data but do not accurately reflect the true allele frequency. This could be due a number of problems such as extreme differential hybridisation or non-specificity of the probe. Differences allele frequency estimates could be used to ascertain the identity of 'bad' SNPs, however such information is platform specific with limited relevance to the wider community hence is not presented here.

Figure 1a shows a simulated mixture of bivariate normal distributions at correlation 0 and 0.65 with propor-

tions 0.13 and 0.87. This demonstrates the pattern we would expect to observe when we plot our data. Figure 1b is the plot of our data and although the data points from the two distributions are not distinguishable, a similar pattern is seen.

We also analysed data on which a stricter criterion for selection was used – i.e. markers in which at least 7 case and control pools passed QC, α decreased to 0.077, and the correlation of the other distribution increased to 0.67 (± 0.0021) ($n = 24,879$). This suggests more stringent use of quality control statistics could increase the overall accuracy of test statistics from the analysis of pooled genotypes. From figure 2 we can see that 7 or more pools pass QC in the vast majority of markers and that over 90% of markers pass QC in all pools.

We also looked at the correlation in the 286 stratification SNPs in which 7 or more of both the case and control pools had passed QC. The correlation was 0.72 (± 0.023). (There was no evidence of stratification in these samples.)

When looking at the proportion of markers we expect to successfully select for individual genotyping, the patterns we would expect to find are present:

- The higher the correlation between the test statistics the greater the power.
- The lower the selection threshold the greater the power (however, this will, of course, also lead to an increase of overall numbers of markers selected).
- Markers with lower p values in the individual genotyping experiments are more likely to be selected.

These statements are all illustrated in figure 3 where the proportion of markers selected is plotted against minus log 10 of the expected p value of a causative SNP. Different data series represent three different correlations and two different cut-offs. For simplicity the proportion of the SNPs with an uncorrelated bivariate distribution are not included in the figure but would have an impact on overall selection.

Although these statements are self evident, our method is a useful tool for researchers to derive actual estimates of relative power for pooling studies.

Discussion

We demonstrate that the bivariate distribution of the joint test statistics from the case-control tests of pooled and individual genotypes can be modelled as a mixture of normal distributions. We also demonstrate that for 'good' SNPs pooling has essentially the same power to

Fig. 1. a A simulated data set from a mixture of bivariate normal distributions. The squares represent data points from 13% of the data which have a correlation of zero and the crosses represent data points from the other 87% of the data which have a correlation of 0.65. **b** The association test statistics from the pooled DNA analysis and the individual DNA analysis. x = test statistics derived from pooled DNA; z = test statistics derived from individually typed DNA.

Fig. 2. The percentage of markers plotted against the number of pools passing QC. All 16 pools passed QC in 74.6% of the markers but this is not shown on the graph so that the pattern from the rest of the results is easier to see. We only have data on those SNPs that have at least two pools passing QC in both the case and control samples.

Fig. 3. The proportion of markers selected is plotted against minus log 10 of the expected p value of a causative SNP. The different data series represent three different correlations and two different cut-offs. For simplicity the proportion of the SNPs with an uncorrelated bivariate distribution are not included in the figure.

detect SNPs that are truly associated with disease-causing variants and is therefore a cost effective tool for association studies. If, for example, we would require a p value of 10^{-8} to deem a SNP significant genome-wide in a GWAS study, such a SNP would meet the 0.05 cut-off for individual genotyping in the pooled analysis over 99% of the time. Accordingly, for a given power P (determined by effect size, sample size and analysis used), the power would be 0.99 P if the same sample size and SNP density is used. 'Bad' SNPs are unlikely to be identified using pooling approaches but with high SNP density signals can be picked up by markers in high LD.

The importance of good quality control for pooling studies is also evident as it increases the correlation of the test statistic. Furthermore, we provide a statistical method that will allow researchers to evaluate a suitable cut-off for selection, taking into account information specific to their laboratory set-up and the aims of their study. Alongside the work of others who have investigated different aspects of pooling methodologies, our research demonstrates the usefulness of this technique.

Our sample was particularly useful for the purposes of the study. We had a large amount of good quality data. More than 40,000 markers had been genotyped in pools and individually; of these, about 30,000 met the high QC criteria we demanded. Because 295 SNPs were analysed before the pooling stage to assess population stratification, we could also be confident of the quality of each individual sample. Although we would have been able to use the individual genotyping data to estimate differential hybridisation (k) for each SNP we did not. Our aim was to investigate the power of pooling in circumstances where k would be unknown for example on most custom built chips used to follow up linkage regions.

Despite the proven utility of pooling, it does provide less information, hence less power, than individual genotyping. As a consequence, the usual motivation to pool is the reduction in cost of the overall experiment, although accessibility of certain equipment can also play a role. There are massive set-up costs for the technologies that genotype hundreds of thousands of markers at one time; many groups do not have such equipment and would need to outsource samples for this kind of analysis. Although extra work is required to ensure the exact consistency of the concentration of all DNA in a pooling study, it can be undertaken on machinery that is cheaper and more commonly available. This may make it more feasible for a laboratory to pool DNA and outsource the pools than outsource all their samples individually.

Cost has been evaluated in a number of studies but due to the continual changes in both the available technology and its price, cost is a constantly changing variable. Furthermore, researchers have to consider whether any given saving outweighs the disadvantages of pooling and this will depend on many factors.

In determining the best approach one also has to consider what the goals and possibilities of the screening procedure are, bearing in mind the reduced relative power of a pooling study in relation to an individual genotyping study. The number of markers being investigated, along with any knowledge of their expected effect size, should impact the decision making process. If a researcher has a small sample size but would like to undertake a genome wide association study (GWAS), and if pooling is the only way they can afford to do it, they would need to consider the implication on power very carefully. Placed alongside the necessity of correction for multiple testing, pooling may reduce power to a level where any money spent on a GWAS is a waste, making it more prudent to use individual genotyping to undertake an extensive candidate gene study. If a GWAS is being performed on a large sample like those being gathered by the various biobanks, pooled DNA analysis would still have considerable power. The number of markers needing to be followed up could alter the method and subsequent cost of the individual genotyping study. For example, the current costs of Illumina technology mean the cost of a custom chip with 4000 markers is the same as the cost of the 300K chip, hence pooling is only cost effective when significantly fewer markers are followed up. If pooling is to be undertaken, there is a lot of information available on how to do it most efficiently, including a review by Sham et al. [23] and a recent paper by MacGregor [24]. The latter looks at issues of variability, suggesting that although multiple pools are useful, there are diminishing returns when one gets to a certain point.

The main disadvantages of pooling are the increase in measurement error and a decrease in the ability to undertake secondary analysis or control for multiple covariates. In a GWAS – where hundreds of thousands of markers are typed – the increase in measurement error would not be as much of a problem as in the context of a tag SNP study. The former would result in considerable redundancy of information from markers in complete or very high LD. This redundancy is also useful when one considers the need for strict quality control. Furthermore, new methods are still developing to help keep the measurement error to a minimum in all studies.

If pooling studies are designed carefully, some level of secondary analysis can be undertaken. For example, one

can construct pools based on covariates and incorporate this information into the analysis using meta-regression techniques [25]. This could be important because of the documented effect of a certain covariate or because the DNA has come from different sources (buccal vs. blood). However, this approach is limited by the number of pools it is feasible to create, as well as by the nature of the covariate. Categorical variables are easy to work with but quantitative ones are not, and something like admixture, which may require assessment by individual genotype data, would be impossible to use without extensive preliminary analysis. Genotype and haplotype analysis can only be undertaken in very small pools, reducing the cost-effectiveness and perhaps the usefulness of the method [26]. Depending on the suspected genetic architecture of the trait, such secondary analysis may not be considered important.

In summary, we suggest that pooling studies are a useful screening technique in some situations, and we provide an approach for researchers to investigate their power in this context.

References

- 1 Arnheim N, Strange C, Erlich H: Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proc Natl Acad Sci USA* 1985;82:6970–6974.
- 2 Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC: Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* 2006;7:27.
- 3 Butcher LM, Meaburn E, Liu L, Fernandes C, Hill L, Al-Chalabi A, Plomin R, Schalkwyk L, Craig IW: Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav Genet* 2004;34:549–555.
- 4 Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I, Schalkwyk LC: Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. *BMC Genomics* 2005;6:52.
- 5 Meaburn E, Butcher LM, Schalkwyk LC, Plomin R: Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. *Nucleic Acids Res* 2006;34:e27.
- 6 Hinds DA, Seymour AB, Durham LK, Banerjee P, Ballinger DG, Milos PM, Cox DR, Thompson JE, Frazer KA: Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. *Hum Genomics* 2004;1:421–434.
- 7 Macgregor S, Zhao ZZ, Henders A, Nicholas MG, Montgomery GW, Visscher PM: Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. *Nucleic Acids Res* 2008;36:e35.
- 8 Sawcer S, Compston A: The genetic analysis of multiple sclerosis in Europeans: concepts and design. *J Neuroimmunol* 2003;143:13–16.
- 9 Butcher LM, Meaburn E, Knight J, Sham PC, Schalkwyk LC, Craig IW, Plomin R: SNPs, microarrays, and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6,000 children. *Hum Mol Genet* 2005.
- 10 Reneland RH, Mah S, Kammerer S, Hoyal CR, Marnellos G, Wilson SG, Sambrook PN, Spector TD, Nelson MR, Braun A: Association between a variation in the phosphodiesterase 4D gene and bone mineral density. *BMC Med Genet* 2005;6:9.
- 11 Liu QR, Drgon T, Johnson C, Walther D, Hess J, Uhl GR: Addiction molecular genetics: 639,401 SNP whole genome association identifies many 'cell adhesion' genes. *Am J Med Genet B Neuropsychiatr Genet* 2006;141:918–925.
- 12 Liu QR, Drgon T, Walther D, Johnson C, Poleskaya O, Hess J, Uhl GR: Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci USA* 2005;102:11864–11869.
- 13 Johnson C, Drgon T, Liu QR, Walther D, Edenberg H, Rice J, Foroud T, Uhl GR: Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am J Med Genet B Neuropsychiatr Genet* 2006;141:844–853.
- 14 Bierut LJ, Madden PA, Breslau N, Johnson EO, Hatsukami D, Pomerleau OF, Swan GE, Rutter J, Bertelsen S, Fox L, Fugman D, Goate AM, Hinrichs AL, Konvicka K, Martin NG, Montgomery GW, Saccone NL, Saccone SF, Wang JC, Chase GA, Rice JP, Ballinger DG: Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum Mol Genet* 2007;16:24–35.
- 15 Meaburn EL, Harlaar N, Craig IW, Schalkwyk LC, Plomin R: Quantitative trait locus association scan of early reading disability and ability using pooled DNA and 100K SNP microarrays in a sample of 5760 children. *Mol Psychiatry* 2008;13:729–740.

Funding

JK was supported by an MRC Bioinformatics Advanced Training Fellowship, Grant No. G0501329.

The NICSNP project is a collaborative research group and part of the NIDA Genetics Consortium. Subject collection was supported by NIH grants CA89392 (PI-L Bierut) from the National Cancer Institute and DA012854 (PI-P Madden) from the National Institute on Drug Abuse. Genotyping work at Perlegen Sciences was performed under NIDA Contract HHSN271200477471C. Phenotypic and genotypic data are stored in the NIDA Center for Genetic Studies (NCGS) at <http://nidagenetics.org/under NIDA Contract HHSN271200477471C> (Pls J Tischfield and J Rice).

JP Rice, DG Ballinger and SF Saccone are listed as inventors on a patent (US 20070258898) held by Perlegen Sciences, Inc., covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction.

Acknowledgements

We would like to thank Laura Bierut, Nick Martin and Stuart MacGregor who assisted in the preparation of this manuscript.

- 16 Shifman S, Bhomra A, Smiley S, Wray NR, James MR, Martin NG, Hettema JM, An SS, Neale MC, van den Oord EJ, Kendler KS, Chen X, Boomsma DI, Middeldorp CM, Hottenga JJ, Slagboom PE, Flint J: A whole genome association study of neuroticism using DNA pooling. *Mol Psychiatry* 2008;13: 302–312.
- 17 Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze TG, Cichon S, Rietschel M, Nothen MM, Georgi A, Schumacher J, Schwarz M, Abou Jamra R, Hofels S, Propping P, Satagopan J, Detera-Wadleigh SD, Hardy J, McMahon FJ: A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 2008;13:197–207.
- 18 Simpson CL, Knight J, Butcher LM, Hansen VK, Meaburn E, Schalkwyk LC, Craig IW, Powell JF, Sham PC, Al-Chalabi A: A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. *Nucleic Acids Res* 2005; 33:e25.
- 19 Craig DW, Huentelman MJ, Hu-Lince D, Zismann VL, Kruer MC, Lee AM, Puffenberger EG, Pearson JM, Stephan DA: Identification of disease causing loci using an array-based genotyping approach on pooled DNA. *BMC Genomics* 2005;6:138.
- 20 Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG: Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Ann Hum Genet* 2002; 66:393–405.
- 21 Prentice RL, Qi L: Aspects of the design and analysis of high-dimensional SNP studies for disease risk estimation. *Biostatistics* 2006;7:339–354.
- 22 Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL, Webster JA, Beach T, Sando SB, Aasly JO, Heun R, Jensen F, Kolsch H, Tsohaki M, Daniilidou M, Reiman EM, Papassotiropoulos A, Hutton ML, Stephan DA, Craig DW: Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am J Hum Genet* 2007;80:126–139.
- 23 Sham P, Bader JS, Craig I, O'Donovan M, Owen M: DNA Pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002;3: 862–871.
- 24 Macgregor S: Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur J Hum Genet* 2007;15:501–504.
- 25 Knight J, Sham P: Design and analysis of association studies using pooled DNA from large twin samples. *Behav Genet* 2006;36: 665–677.
- 26 Zeng D, Lin DY: Estimating haplotype-disease associations with pooled genotype data. *Genet Epidemiol* 2004;28:70–82.