Minireview Genomics & Informatics Vol. 3(1) 1-7, March 2005

DNA Pooling as a Tool for Case-Control Association Studies of Complex Traits

Chul Ahn[1]*, Terri M. King[1], Kyusang Lee[2] and Seung-Ho Kang[3]

[1]Department of Medicine, University of Texas Medical School, Houston, TX 77030, USA, [2]Samsung Advanced Institute of Technology, Seoul 135-710, Korea, [3]Department of Statistics, Ewha Womans University, Seoul 120-750, Korea

Abstract

Case-control studies are widely used for disease gene mapping using individual genotyping data. However, analyses of large samples are often impractical due to the expense of individual genotyping. The use of DNA pooling can significantly reduce the number of genotyping reactions required; hence reducing the cost of large-scale case-control association studies. Here, we discuss the design and analysis of DNA pooling genetic association studies.

Genetic Study

Two types of statistical methods have been widely used to attempt to identify genetic determinants of diseases. One is a linkage analysis and the other is an association study. Linkage analysis methods attempt to find out the rough location of the disease gene relative to another DNA sequence called a genetic marker, which has its position on the chromosome already known. Linkage analysis has been extremely useful in the identification of genes responsible for diseases with simple Mendelian inheritance such as cystic fibrosis (Rommens *et al.*, 1989). Since complex diseases are likely to be influenced by the factors such as genetic heterogeneity, phenocopies, incomplete penetrance, genotype-by-environment interactions, and multilocus effects, the application of linkage analysis to complex diseases has been much less successful.

Complex diseases are likely to be influenced by multiple genes of small effect (Elston, 1995). Association studies provide the most powerful approach to identify such genes underlying complex traits (Risch and Merikangas, 1996). Genetic association studies investigate whether there is a relationship between a "genetic marker" and the frequency or severity of a particular trait. Risch and Merikangas (1996) have shown that association studies can be a very powerful approach to find genetic determinants of a complex disorder. Genetic association studies have been applied in a variety of complex human diseases such as cancer, alcoholism, pulmonary disease, heart disease, diabetes and Alzheimer disease (Silverman and Palmer, 2000; Risch and Merikangas, 1996; Chakravarti, 1999).

Association Studies

Different experimental designs can be used to conduct genetic association studies. Cohort studies prospectively assemble a group of individuals, and then follow them to determine the frequency of developing disease. Cohort studies investigate the frequency of the DNA variant in the entire cohort for the estimation of risk ratio and predictive values. Cohort studies are labor intensive and costly. However, phenotype may be more clearly defined for genetic analysis through the use of longitudinal data due to the ability to observe the natural history of the disease and potentially significant comorbidities. This is an important issue in complex human disease since research into such diseases can be complicated by phenotype heterogeneity. Cohort studies are potentially less prone to ascertainment bias than are case-control studies, although geographic and population factors must be examined explicitly.

Case-control designs for genetic association studies are not different in conception from the case-control designs that have been well developed for use in epidemiological methods (Romero *et al.*, 2002). Case-control studies examine the frequency of a DNA variant in individuals affected by a disease (cases) and in those not affected by the disorder (controls). Genetic association studies have been mostly performed in a case-control setting with unrelated affected subjects compared with unrelated unaffected subjects. Significant differences in allele frequencies or genotype frequencies between cases and controls are taken as evidence for involvement of an allele or genotype in disease susceptibility.

Case-control studies are potentially susceptible to bias if case and controls are not in fact comparable. One

*Corresponding author: E-mail Chul.W.Ahn@uth.tmc.edu, Tel +1-713-500-6701, Fax +1-713-500-6722
Accepted 28 February 2005

source of this bias arises due to population stratification. Population stratification occurs when the differences in gene frequency are wholly or partially attributable to inherent underlying population structures between cases and controls rather than to association between disease and gene. Case-control studies are also susceptible to a spurious association by a false-positive result due to chance. The problem is particularly acute for studies involving a large number of markers, such as are used in regional or genome-wide association studies. This association will be less likely if a very stringent threshold for significance is met. Risch and Merikangas (1996) suggest a Bonferroni correction with a p-value of $5x10^{-8}$ to demonstrate a significant association if one million SNPs (single nucleotide polymorphisms) were typed across the genome. While Bonferroni corrections (Bonferroni, 1936) are often used when multiple associations are measured in a study, there are some features of a genetic association study which make it unattractive. Bonferroni corrections are estimated by dividing the type I error rate (α) by the number of tests performed; after this procedure, a test is significant only if the p-value is less than this adjusted type I error rate. However, with the number of tests done in a typical genetic association study, it is clear that the Bonferroni correction can significantly lower power for detecting true association. The correction also does not account for the fact that often the test statistics of markers located in near proximity are not independent test and are, in fact highly correlated. For instance, investigating 2 SNPs that are 50 kb apart are likely highly correlated due to linkage disequilibrium. To address this issue, Kaplan *et al.* (1997) and McIntyre *et al.* (2000) have proposed a Monte Carlo approach for evaluating associations within the transmission/disequilibrium test (TDT) construct for a single locus and multilocus tests, respectively.

The case-control study remains a popular approach in genetic epidemiology since case-control studies are economically and statistically efficient. Even though ascertainment of disease, selection of controls, and measurement of exposure present substantial difficulties in most case-control studies, a large body of epidemiologic theory provides guidance to meet these challenges. Bias arising from population stratification should be mitigated by proper design and analysis of case-control studies and by new statistical methods such as genomic control. This method, proposed by Devlin and Roeder (1999), does not require information about the genealogy of the population and corrects for population heterogeneity, poor choice of controls, and cryptic relatedness of cases.

The validity of genetic association studies depends on the selection of appropriate controls since misclassification of controls leads to a false-negative result. Controls may be pre-symptomatic or asymptomatic instead of disease-free for disease such as stroke which tends to have an advanced age of onset. Small cerebral infarctions are visible only in CT scans. So, the subjects classified as controls may have asymptomatic disease. Therefore, as with standard epidemiologic studies, phenotypic characterizations of controls and cases are crucial. Care and attention in the design, such as having older controls (i.e., outside of the risk profile) can mitigate this bias. Unlike risk factors in traditional case-control studies, genes do not have the potential for risk factor recall bias.

Population Substructure in Case-Control Studies

Population stratification is a form of confounding (Wacholder *et al.*, 2000) that may cause spurious associations in a case-control study when allelic frequencies vary across subpopulations within the study sample.

In association studies, the standard methods for dealing with potential confounding effects are matching and statistical adjustment. In association studies using DNA pools from many individuals, significant causal disease associations may not be distinguished from associations due to the differences in confounding factors between cases and controls. Thus, matching by confounding factors prior to DNA pooling is essential. Population stratification may be solved by using matched case-control designs which match controls to cases on potential confounding factors such as age, sex, and ethnicity, etc. That is, DNA pools needs to be matched to have similar socio-demographic composition to minimize the risk of spurious associations due to confounding. Then, allele-frequency estimates in the matched DNA pools will give a more reliable indication of causal disease association. Statistical methods are proposed to adequately control for population stratification (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999; Pritchard *et al.*, 2000).

When a disorder has known risk factors, it might be desirable to construct multiple pools so that pools differ in the level of exposure to the risk factors. Use of multiple pools incorporating risk factors might increase the power to detect an allelic association with disease. However, it can be argued that the use of multiple pools can be avoided by using the risk factors as covariates in the statistical analysis at the second stage since DNA pooling is used as a screen to be followed by individual genotyping.

Family-based case-control designs can be used to solve the stratification problem. That is, stratification problem can be solved using parents as controls or using

unaffected sibs as controls. However, family-based case-control designs are more expensive than the case-control designs using controls that are unrelated to the cases. Family-based designs will have less power compared to a well-designed study involving unrelated controls.

Issues in Genetic Case-Control study

Genetic case-control study still poses great challenges. One of the difficulties is to obtain the large number of genotypes needed. Additionally, sample sizes of hundreds or even thousands of subjects may be needed to achieve statistical power to detect loci with modest effect size.

Genetic association studies using a set of SNPs that covers the human genome densely would be very expensive, and beyond the reach of most laboratories even though the cost of large-scale single-nucleotide polymorphisms (SNP) determination dropped dramatically in recent years. As a result, the development of innovative study designs that reduce the cost is warranted.

Haplotype-tagging SNPs and DNA pooling have potential to reduce the cost barriers of the genetic association study. Botstein and Risch (2003) provide an excellent review for the comparison of genome-wide haplotype map-based versus sequenced-based strategies. The use of haplotype-tagging SNPs is a map-based approach while DNA pooling is a sequenced-based approach. A comprehensive map-based approach may require genotyping 10-fold more SNPs than a sequence-based approach (500,000-1,000,000 for map- based versus 50,000-100,000 for sequenced-based). Bostein and Risch (2003) suggest a sequence-based approach for the initial stage of a major program aimed at genome-wide association studies. Bansal *et al.* (2002) demonstrate the potential of a sequence-based DNA pooling techniques and their associated technologies as an initial screen in the search for genetic association.

DNA Pooling

The use of DNA pooling technique has been proposed as a means of reducing the number of genotyping reactions required, hence reducing the cost of large-scale case-control association studies. DNA pooling is a powerful and efficient tool for high throughput association analysis. DNA pooling allows measurement of allele frequencies in groups (or pools) of individuals, thereby reducing the number of PCR reactions and genotyping assays dramatically. Therefore, the use of DNA pooling can significantly reduce the number of genotyping reactions required; hence reducing the cost of large-scale case-control association studies, and offering an approach to this economic impasse. Furthermore, DNA pooling can be an extremely effective method for conserving precious DNA.

The most powerful methods for detecting the association between a marker and a phenotype require individual genotyping. Experimental savings can be achieved by testing allele frequency differences between DNA pools chosen by phenotypic value (Darvasi and Soller, 1994; Barcellos *et al*, 1997). In DNA pooling, the equal amounts of DNA from each subject are used to create a pool to estimate the allele frequency. That is, DNA of individuals is mixed together to generate a pool before estimating allele frequencies. In general, one pool is created out of all cases and a second out of all controls in a case-control study. Allele frequencies are then estimated in each pool and compared directly between two pools. Various methods have been used for allele frequency estimation, including mass spectrometry, denaturing high-performance liquid chromatography (HPLC), and photo-lithography. Whatever the method of estimating allele frequency is used, allele frequencies are estimated in each DNA pool by quantifying the relative amounts of DNA products representing each allele.

In DNA pooling, allele frequency is measured in a large number of SNP markers in each of the pools as an efficient screen to enrich for SNPs with significant allele frequency differences (Bansal *et al.*, 2002). The SNPs with the large allele-frequency differences in the pooled data are then selected for individual genotyping. The pooled genotyping step reduces the number of SNPs that must be individually genotyped to confirm allele-frequency differences between cases and controls. The larger the sample, the greater the saving, so that the design with minimal genotyping would involve comparing just two pools, each containing DNA from numerous individuals. These two pools could be constituted from cases and controls for a disease trait, or from individuals with trait values at the two extremes of a quantitative trait.

For the quantitative traits such as blood pressure and cholesterol levels, Bader *et al.* (2001) provide power estimates for two pooled DNA designs which classify the individuals as affected or unaffected, analogous to a case-control design. The optimal design for the quantitative phenotypes is to pool the top and bottom 27% of individuals. This optimal design requires a sample size only 1.24 times larger than that required for individual genotyping when we ignore the experimental measurement error. When a measurement error is included, the pooled

DNA association test serves well as a pre-screen to identify candidate markers which then proceed to individual genotyping. This DNA pooling strategy can still provide a substantial savings over individual genotyping.

There are disparate recommendations on the pool sizes in DNA pooling. While Barratt *et al.* (2002) suggest the pool sizes of 50 or fewer, Mohlke *et al.* (2002) and Le Hellard *et al.* (2002) suggest larger pool size of up to 500 cases and 500 controls. Feng *et al.* (2004) thinks the larger pool sizes are preferable since the quality of frequency estimates does not seem to degrade with larger pool size. Effect of pool sizes on DNA pooling needs to be investigated.

There are some potential error sources in the stages of DNA quantification and formation of pools (Barratt *et al.*, 2002). One of the potential error sources is that alleles with different sequences may not be amplified equally in the competitive reaction. Many high-throughput genotyping platforms measure the amount of each allele from competitive amplification reaction, and determine the genotype based on the abundance of alleles. A standard procedure is to genotype some heterozygotes individually using competitive reaction and then to estimate the average relative speed of amplification reaction from the experimentally. If each allele of an SNP site is equally represented in the assay, the ratio (k) of the amplifications of competing alleles would be one. The frequency of allele A in a pool can be estimated more accurately as $A/(A+k \cdot B)$, where A and B are the measured abundance of alleles corresponding to two polymorphic alleles at the SNP locus. Another potential error sources are due to unequal amounts of DNA per individual and due to experimental errors.

The foundation of successful DNA pooling association test is a precise and accurate estimation of allele frequency. Tang *et al.* (2004) show that the SNP allele frequency estimates from pooled analysis are comparable to those from individual genotyping. The coefficient of determination (R-square) of the frequency estimates between DNA pooling and individual genotyping is 0.975 using individual heterozygous samples. The quantitative abilities of various platforms used for estimating allele frequencies are confirmed by multiple studies (Sham *et al.*, 2002; Le Hellard *et al.*, 2002; Shifman *et al.*, 2002; Moskvina *et al.*, in press). Mohlke *et al.* (2002) provide an example of comparison of SNP allele frequency estimates from pooled analysis and from individual genotyping.

Due to the dramatic reduction in the large-scale SNP determination, the cost of SNP determination is now approximately 1 cent per SNP (Feng *et al.*, 2004). Epidemiologic genetic association studies would be still prohibitively expensive even though the genotyping cost has recently dropped dramatically. Feng *et al.* (2004) describe that genotyping of 2 million SNPs for 500 cases and 500 matched controls would cost about $20 million. When a single case and a single control pools are used for pooled analysis, the minor allele frequency analysis for 2 million SNPs can be obtained at a genotyping cost of about $40,000. If quadruplicate frequency estimation is obtained from the single case and control pools, genotype cost will be approximately $160,000. DNA pooling technique has the potential to reduce the cost barrier of the large-scale genetic association study.

It is expected that the differences in allele frequencies between cases and controls may be only small in the analysis of complex disease. Therefore, the success of DNA pooling crucially depends on reproducibility and high accuracy in the estimation of allele frequencies of cases and controls. Replicates are needed for each reaction in order to achieve high accuracy.

Statistical Analysis

The most effective use of DNA pooling might be a two-stage design in which markers that show evidence of association from the pooled association are followed up by individual genotyping. Allele frequencies are separately estimated from cases and controls at the first stage, and then individual genotyping will be done for the selected markers with the large allele-frequency differences at the second stage. Thus, pooling can be used as an efficient and sensitive method of screening numerous markers to identify a subset of markers for more detailed studies.

A common observation with the use of DNA pools is that two alleles at a polymorphic SNP locus are not amplified in equal amounts in heterozygous individuals depending on the design of assay. In addition, there are pool-specific errors so that there is variation in the estimates of allele frequencies from different pools that are from the same individuals. As a result of these additional sources of variation, the outcome of an experiment is an estimated count of alleles rather than the usual outcome in terms of observed counts.

In the first stage, the simple method to analyze SNP-based case-control association studies using DNA pooling is to multiply the estimated allele frequencies by the number of chromosomes in the pool and perform a chi-square test. This simple chi-square test has been used in several published studies (Shifman *et al.*, 2002; Williams *et al.*, 2002; Norton *et al.*, 2003, 2004). Visscher and Le Hellard (2003) and Le Hellard *et al.* (2002) show that simply substituting estimated count for observed

count can lead to an inflation of type I error rates due to unequal amplification in DNA pooling. Adjustment was made to account for unequal amplification by various researchers (Mohlke *et al.* 2002; Yang *et al.*, 2005; Visscher and Le Hellard, 2003). Visscher and Le Hellard (2003) modify the standard chi-square test by incorporating the variation inflation factor to control the type I error rate in the presence of experimental variation.

As an alternative approach for the analysis of pooled data, Visscher and Le Hellard (2003) suggest an over-dispersed model, which is widely used for the analysis of clustered binary data in statistical literatures (Ahn *et al.*, 2003; Jung *et al.*, 2001; Kang *et al.*, 2003). The parameters from the over-dispersed model can be estimated from a nested design of population samples and replicated pools within samples. Kang *et al.* (2004) investigate the allelic chi-square test used in genetic association studies in terms of empirical type I error rates and empirical powers.

If the large allele-frequency differences are observed in the pools of cases and controls, genotyping is performed individually. One of the important issues in DNA pooling study is the selection of markers to be genotyped individually. Development of statistical methods is needed to determine methods for selecting markers to proceed to the second stage. Konig and Ziegler (2004) propose decision-theoretic models to determine individual genotyping based on the results from pooled DNA. Bonferroni correction, false discovery rate (FDR) or related concepts can be also used for the selection of markers to move to the second stage. In the second stage, genotyping is performed individually and markers are analyzed conventionally. A more comprehensive control of confounding factors can be done at the second stage. The second stage can enable the study of gene-gene interactions and gene-environment interactions.

The importance of correcting for multiple comparisons in genomic screens is well known (Lander and Kruglyak, 1995). Benjamini and Hochberg (1995) introduce the false discovery rate (FDR) for multiple testing situations. Sabatti *et al.* (2003) show that the simple step-down procedure introduced by Benjamini and Hochberg (1995) controls the FDR for the dependent tests on which association genome screens are based through simulation.

Even though complex diseases are generally caused by multiple genetic variations, most available association methods are based on the assumption that a single genetic variation is primarily responsible for the disease under study. Only a few approaches consider interactions of multiple genes and environmental factors in identifying susceptibility loci for complex disease (Hoh *et al.*, 2001; Ritchie *et al.*, 2001; Nelson *et al.*, 2001; Kim *et al.*, 2003; Hao *et al.*, 2004). Hoh *et al.* (2001) develop a novel test procedure, called a set association approach, to identify genetic variation responsible for complex diseases when multiple genes are involved. This approach is appropriate for many study designs, such as case-control, trio and extended families. The method uses a score statistic that is weighted by the allele contribution to a Hardy-Weinberg equilibrium measurement. All alleles are jointly estimated and the minimal p-value identifies the combination of alleles, across genes that appear to act in concert to alter the risk of disease. They applied the set association approach to a real restenosis data set and could identify several SNPs of interest that were in linkage disequilibrium with susceptibility locus for restenosis, and the re-blockage of the coronary after treatment. Zee *et al.* (2002) use this approach to define a panel of contributory genes in instant restenosis. Hoh *et al.* (2001) did not evaluate the empirical the type I errors and empirical powers of the set association approach. Hao *et al.* (2004) systematically evaluate the performances of multiple SNP association test in terms of power and accuracy in capturing the real disease SNPs. Hao *et al.* (2004) show that the inclusion of Hardy-Weinberg Disequilibrium (HWD) reduces the power through simulation. They demonstrate that the test procedure could capture the SNPs associated with disease fairly successfully.

Romero *et al.* (2002) propose guidelines for the evaluation of reports of genetic association studies including selection of SNPs, study design, assay characteristics, sample size determination, multiple tests and statistical analysis. Proposed guidelines for the reporting of genetic association studies facilitate the peer review process, publication, and availability of the data for future studies and systematic reviews.

Conclusion

DNA pooling technique is ideal for screening a large number of markers for associations although positive results will require confirmation through individual genotyping. Considerable savings can be achieved concerning DNA, cost and labor through the use of DNA pooling. All markers identified in the initial discovery process are subjected to a follow-up program. Additional work is necessary to eliminate false-positive associations that are likely due to sampling errors and potential population substructures. The most powerful method is the application of observations in one or more independent samples. That is, replicate studies using pooled DNA from two or more studies are expected to

play an efficient role in eliminating false positive findings, and for the efficient identification of meaningful association.

Considerable cost reduction can be achieved through the use of DNA pooling, whereby DNA samples from multiple individuals are pooled before genotyping. It is suggested that DNA pooling should be considered especially in the initial stages of a major program aimed at genome-wide association studies since large-scale association studies can be accelerated with the use of DNA pooling.

References

Ahn, C., Jung, S., and Kang, S. (2003). An evaluation of weighted chi-square statistics for site specific data. *Drug Information Journal* 37, 91-99.

Bader, J., Bansal, A., and Sham, P. (2001). Efficient SNP-based test of association for quantitative phenotypes using pooled DNA. *Genescreen* 143-150.

Bansal, A., Boom, D., Krammerer, S., Honisch, C., Adam, G., Cantor, C., Kleyn, P., and Braun, A. (2002). Association testing by DNA pooling: An effective initial screen. *Proc. Natl. Acad. Sci. USA* 99,16871-16874.

Barcellos, L.F., Klitz, W., Field, L.L., Tobias, R., Bowcock, A.M., Wilson, R., Nelson, M.P., Nagatomi, J., and Thomson, G. (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. *American Journal of Human Genetics* 61,734-747.

Barratt, B.J., Payne, F., Rance, H.E., Nutland, S., Todd, J.A., and Clayton, D.G. (2002). Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. *Annals of Human Genetics* 66,393-405.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of Royal Statistical Society B* 57, 289-300.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilit? *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8,3-62.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* 33, Suppl. S228-S237.

Chakravarti, A. (1999). Population genetics-making sense out of sequence. *Nature Genetics* 21, Suppl.1, S56-S60.

Darvasi, A. and Soller, M. (1994). Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics* 138, 1365-1373.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics* 55, 997-1004.

Elston, R. (1995). The genetic dissection of multifactorial traits. *Clin. Exp. Allergy* 2, 103-106.

Feng, Z., Prentice, R., and Srivastava, S. (2004). Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. *Pharmacogenomics* 5, 709-719.

Hao, K., Laid, N., Wang, X., and Xu, X. (2004). Power estimation of multiple SNP association test of case-control study and application. *Genetic Epidemiology* 26, 22-30.

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research* 11, 2115-2119.

Jung, S., Ahn, C., and Donner, A. (2001). Evaluation of an adjusted chi-square statistic applied to observational studies involving a clustered binary data. *Statistics in Medicine* 20, 2149-2161.

Kang, S., Ahn, C., and Jung, S. (2003). Sample size calculations in cluster randomized studies with varying cluster size: a binary case. *Drug Information Journal* 37, 109-114.

Kang, S., Shin, D., Oh, M., and Ahn, C. (2004). An investigation on the allelic chi-square test used in genetic association studies. *Biometrical Journal* 46, 699-706.

Kaplan, N.L., Martin, E.R., and Weir, B.S. (1997). Power studies for the transmission/disequilibrium tests with multiple alleles. *American Journal of Human Genetics* 60, 691-702.

Kim, S., Zhang, K., and Sun, F. (2003). Detecting susceptibility genes in case-control studies using set association. *BMC Genetics* 4, Suppl. 1, S9.

Konig, I. and Ziegler, A. (2004). Analysis of SNPs in pooled DNA: a decision theoretic model. *Genetic Epidemiology* 26, 31-43.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* 11, 241-247.

Le Hellard, S., Ballereau, S., Visscher, P., Torrance, H., Pinson, J., Morris, S., Thomson, M., Semple, C., Muir, W., Blackwood, D., Porteous, D., and Evans, K. (2002). SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Research* 30, e74.

Moskvina, V., Norton, N., Williams, N., Holmans, P., Owen, M., and O'Donovan, M. Streamlined analysis of pooled genotype data in SNP-based association studies. *Genetic Epidemiology* (in press).

McIntyre, L.M., Martin, E.R., Simonsen, K.L., and Kaplan, N.L. (2000). Circumventing multiple testing: a multilocus

Monte Carlo approach. *Genetic Epidemiology* 19, 18-29.

Mohlke, K.L., Erdos, M.R., Scott, L.J., Fingerlin, T.E., Jackson, A.U., Silander, K., Hollstein, P., Boehnke, M., and Collins, F.S. (2002). High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. *Proc. Natl. Acad. Sci. USA* 99, 16928-16933.

Nelson, M.R., Kardia, S.L., Ferell, R.E., and Sing, C.F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research* 11, 458-470.

Norton, N., Williams, H.J., Williams, N.M., Spurlock, G., Zammit, S., Jones, G., Jones, S., Owen, R., O'Donovan, M.C., and Owen, M.J. (2003). Mutation screening of the Homer gene family and association analysis in schizophrenia. *American Journal of Medical Genetics* 120B, 18-21.

Norton, N., Williams, N., O'Donovan, M., and Owen, M. (2004). DNA pooling as a tool for large-scale association studies in complex traits. *Annals of Medicine* 36, 146-152.

Pritchard, J., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics* 67, 170-181.

Pritchard, J. and Rosenberg, N. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* 65, 220-228.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* 273, 1516-1517.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Pari, F.F., and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* 69, 138-147.

Romero, R., Kuivaniemi, H., Tromp, G., and Olson, J. (2002). The design, execution, and interpretation of genetic association studies to decipher complex disease. *Am. J. Obstet. Gynecol.* 187, 1299-1312.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., *et al.* (1989). Identification of cystic fibrosis gene: chromosome walking and jumping. *Science* 245, 1059-1065.

Sabatti, C., Service, S., and Freimer, N. (2003). False discovery rate in linkage and associations genome screens for complex disorders. *Genetics* 164, 829-833.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA pooling: a tool for large scale association studies. *Nature Reviews Genetics* 3, 862-871

Shifman, S., Bronstein, M., Sternfeld, M., Pisante-Shalom, A., Lev-Lehamn, E., Weizman, A., Reznik, I., Spivak, B., Grisaru, N., Karp, L., Schiffer, R., Kotler, M., Strous, R.D., Swartz-Vanetik, M., Knobler, H.Y., Shinar, E., Beckmann, J.S., Yakir, B., Risch, N., Zak, N.B., and Darvasi, A. (2002). A highly significant association between a COMT haplotype and schizophrenia. *American Journal of Human Genetics* 71, 1296-1302.

Silverman, E. and Palmer, L. (2000). Case-control association studies for the genetics of complex respiratory diseases. *Am. J. Respir. Cell. Mol. Biol.* 22, 645-648.

Tang, K., Oeth, P., Krammerer, S., Denissenko, M., Ekblom, J., Jurinke, C., Boom, D., Braun, A., and Cantor, C. (2004). Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry. *Journal of Proteome Research* 3, 218-227.

Visscher, P.M. and Le Hellard, S. (2003). Simple method to analyze SNP-based association studies using DNA pools. *Genetic Epidemiology* 24, 291-296.

Wacholder, S., Rothman, N., and Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *Journal of the National Cancer Institute* 92, 1151-1158.

Williams, N.M., Spurlock, G., Norton, N., Williams, H.J., Hamshere, M.L., Krawczak, M., Kirov, G., Nikolov, I., Georgieva, L., Jones, S., Cardno, A.G., O'Donovan, M.C., and Owen, M.J. (2002). Mutation screening and LD mapping in the VCFS deleted region of chromosome 22q11 in schizophrenia using a novel DNA pooling approach. *Molecular Psychiatry* 7, 1092-1100.

Yang, H.C., Pan, C.C., Lu, R., and Fann, C. (2005). New adjustment factors and sample size calculation in a DNA-pooling experiment with preferential amplification. *Genetics* 169, 399-410.

Zee, R.Y., Hoh, J., Cheng, S., Reynolds, R., Grow, M.A., Silbergleit, A., Walker, K., Steiner, L., Zangenberg, G., Fernandez-Ortiz, A., Mayaca, C., Pinto, E., Fernandez-Cruz, A., Ott, J., and Lindpainter, K. (2002). Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. *Pharmacogenomics Journal* 2, 197-201.