



Published in final edited form as:

*Genet Epidemiol.* 2009 ; 33(Suppl 1): S88–S92. doi:10.1002/gepi.20478.

## Population Stratification and Patterns of Linkage Disequilibrium

Anthony L. Hinrichs<sup>1</sup>, Emma K. Larkin<sup>2,3</sup>, and Brian K. Suarez<sup>1,4</sup>

<sup>1</sup>Department of Psychiatry, Washington University School of Medicine, St. Louis, MO

<sup>2</sup>Department of Epidemiology and Biostatistics, Division of Genetic and Molecular Epidemiology, Case Western Reserve University School of Medicine, Cleveland, OH

<sup>3</sup>Center for Clinical Investigation, Case Western Reserve University School of Medicine, Cleveland, OH

<sup>4</sup>Department of Genetics, Washington University School of Medicine, St. Louis, MO

### Abstract

Although the importance of selecting cases and controls from the same population has been recognized for decades, the recent advent of genome-wide association studies has heightened awareness of this issue. Because these studies typically deal with large samples, small differences in allele frequencies between cases and controls can easily reach statistical significance. When, unbeknownst to a researcher, cases and controls have different substructures, the number of false-positive findings is inflated. There have been three recent developments of purely statistical approaches to assessing the ancestral comparability of case and control samples: genomic control, structured association, and multivariate reduction analyses. The widespread use of high-throughput technology has allowed the quick and accurate genotyping of the large number of markers required by these methods.

Group 13 dealt with four population stratification issues: single-nucleotide polymorphism marker selection, association testing, non-standard methods, and linkage disequilibrium calculations in stratified or mixed ethnicity samples. We demonstrated that there are continuous axes of ethnic variation in both datasets of Genetic Analysis Workshop 16. Furthermore, ignoring this structure created *p*-value inflation for a variety of phenotypes. Principal-components analysis (or multidimensional scaling) can control inflation as covariates in a logistic regression. One can weight for local ancestry estimation and allow the use of related individuals. Problems arise in the presence of extremely high association or unusually strong linkage disequilibrium (e.g., in chromosomal inversions). Our group also reported a method for performing an association test controlling for substructure when genome-wide markers are not available to explicitly compute stratification.

### Keywords

genetic association; genome-wide association study; principal components; multidimensional scaling; ethnic substructure

### Introduction

We are reminded of Justice Potter Stewart's famous opinion from the 1964 Supreme Court case *Jacobellis v. Ohio* dealing with pornography. He wrote: "I shall not today attempt

further to define the kinds of material I understand to be embraced... But I know it when I see it..." So it is with the main issue addressed by most of the contributions in Group 13 (Population Stratification and Patterns of Linkage Disequilibrium), to wit: What is the definition of a human breeding population and what are its boundaries? Unfortunately, unlike Justice Stewart, most of us do not know what a "breeding population" is – even when we "see" it.

No natural population, human or otherwise, is "homogeneous" in the sense of being uniformly homozygous. The evolutionary forces of mutation, drift, migration, and selection have shaped all human groups and produced unique patterns of variation. What is clear is that human groups more or less differ from one another depending on the length of time since they shared common ancestry. Today, most markers used in linkage studies or case-control genome-wide association studies (GWAS) appear to be selectively neutral. This is true for most variable-number tandem repeats, microsatellites, and single-nucleotide polymorphisms (SNPs). Furthermore, human groups differ in their allele frequencies for these markers (e.g., Goddard et al. [2000]).

Genetic association tests identify differences in allele frequency between cases and controls. True positives occur when the marker in question is related to disease status. False positives occur when the apparent difference in allele frequency is due to measurement error or when an actual difference in allele frequency is unrelated to disease status. Thorough data cleaning can help eliminate or reduce many false positives due to measurement error. Genotyping errors and missing genotypes that are not erroneous or missing completely at random (such as plate effects or missingness due to degraded samples) can create false positives when cases and controls are located on different plates or cases and controls have different DNA extraction methods. Cryptic relatedness or cryptic duplicates skew allele frequency estimates because observations are not independent. When cases and controls are drawn from different randomly mating breeding populations, allele frequencies truly are different, but these differences may not be related to disease status. In practice, we distinguish between two types of population stratification. The first type, well studied by the field, is that of genetically, and usually geographically, distinct populations. An example would be controls of European ancestry and cases of Native American ancestry. An individual's self-reported ancestry may be sufficient to control for this stratification. The second type, only recently studied, is that of continuous genetic variation. An example would be when the sample is of European ancestry, but if cases tend to be from Northern Europe while controls tend to be from Southern Europe, false-positive rates will be elevated due to true differences in allele frequency. Multiple continuous axes are possible (Eastern and Western European, for example), and in practice a complex mix of discrete and continuous stratification may be observed.

Although the importance of selecting cases and controls from the same breeding population has been recognized for decades [Suarez and Hampe, 1994], the recent advent of large scale GWAS has heightened awareness of this issue because relatively small differences in allele frequencies (such as found in subtle continuous stratification) can reach statistical significance when large samples are studied. This is particularly true for control samples that have been collected with the intention that they will be used repeatedly in a variety of different GWAS. (A recent example is the publicly available control sample collected by a marketing research company (Knowledge Networks, Menlo Park, CA), for the NIH (<http://www.nimhgenetics.org>.) The strategy of using and reusing an all-purpose, "one-size-fits all" control sample is attractive from a budgetary perspective. However, differences in ascertainment can create subtle differences in ancestry, thereby creating allele frequency differences that are unrelated to the disease of interest. Indeed, one of the datasets from the Genetic Analysis Workshop – the North American Rheumatoid Arthritis Consortium

(NARAC) data – may fall into this category. Thus, while the case sample was gathered from rheumatology clinics across North America, all of the controls were selected from participants of the New York Cancer Project [Plenge et al., 2007].

There are many experimental designs that attempt to minimize ancestry differences between cases and controls. One can, for instance, match cases and controls on the country of origin of each individual's four grandparents. Another popular design is to restrict controls to spouses. In practice, however, experience has shown that these designs are often difficult to implement and, moreover, neither guarantees that cases and controls come from the same breeding population. The widespread use of array technology has allowed the quick and accurate genotyping of tens of thousands to millions of markers per individual. These technological advances, in turn, have allowed the development of purely statistical approaches to assessing the ancestral comparability of case and control samples. Three of these approaches are in common use: 1) genomic control analyses, 2) structured analyses, and 3) multivariate reduction analyses.

The method of genomic control [Bacanu et al., 2000; Devlin and Roeder, 1999; Devlin et al., 2001; Reich and Goldstein, 2001; Zheng et al., 2005; Zheng et al., 2006] surveys markers with a low prior probability of association with disease (“null markers”). These are preferably a large number of unlinked loci across the genome. The observed median value of the chi-squared statistic for the null markers divided by the expected median value of the chi-squared statistic (approximately 0.456 for 1 df tests) is the “inflation factor,” lambda. If lambda is less than or equal to 1, no adjustment is necessary. When lambda is greater than one, all subsequent chi-squared statistics on a set of candidate markers are divided by lambda. In a case of many markers with no particular prior hypotheses (such as a GWAS), the set of null markers and the set of candidate markers is taken to be the same. This may be a poor assumption in the case of some polygenic phenotypes such as height and weight that have established high heritabilities. The genomic control method has the disadvantage of assuming that stratification creates uniform inflation across the genome, potentially biasing association tests conservatively in some regions and freely in other regions. Furthermore, the inclusion of true positives in the null set can overestimate lambda; in the rheumatoid arthritis (RA) dataset, this occurs if one includes the major histocompatibility complex (MHC) markers on chromosome 6p in the null set.

The method of structured association [Falush et al., 2003; Falush et al., 2007; Pritchard and Rosenberg, 1999; Pritchard et al., 2000a; Pritchard et al., 2000b; Satten et al., 2001] uses a Markov-chain Monte Carlo process to determine allele frequency distribution for each marker for each of  $K$  clusters. Individuals are assigned probability of membership in each cluster. This method applies to microsatellites as well as SNPs and produces easily interpreted results. However, the number of clusters must be determined heuristically, and the method is computationally intractable for more than a few hundred markers. Therefore, to use this method to discern subtle variation in structure, such as northern to southern European, it is necessary to use *a priori* knowledge to select markers that are highly informative for this variation [Price et al., 2008].

The third method, and the one predominantly used by Group 13, is multivariate data reduction [Patterson et al., 2006; Zhang et al., 2003; Zhu et al., 2002]. This method is primarily applied to large-scale SNP data. Although there are a number of possible approaches, the most frequently used is an application of traditional principal-component analysis (PCA) implemented in the EIGENSOFT package [Price et al., 2006]. In PCA, for  $N$  individuals and  $M$  markers, where  $M \gg N$ , we first create an  $M \times N$  standardized genotype matrix,  $G$ . In particular, for each marker we recode the three genotypes (1/1, 1/2, 2/2) to the values  $(d-a, d, d+a)$  to have an additive dose effect, where  $a$  and  $d$  are chosen so that the

marker has mean zero and standard deviation one. We filter out markers with very rare minor alleles. This approximates the normalization process from standard PCA. We then compute  $G^T G$ , the  $N \times N$  covariance matrix and perform an eigenvalue decomposition. This produces a new coordinate system such that the greatest genotypic variance is encoded by the first component, the second greatest variance by the second component, and so on. One may then either look for discrete groups (ethnically discrete populations) or use the components themselves as covariates (for continuous population distributions).

When using the components themselves as covariates, various methods are used to determine the number of components to retain. Feng et al. [2009] and Kang et al. [2009] retained the first ten principal components, as suggested by Price et al. [2006]. Hinrichs et al. [2009] applied a graphical scree analysis. Peloso et al. [2009] retained only the PCs that were correlated with the disease status. Wang et al. [2009] used the Tracy-Widom statistic. They found that the number of significant PCs reported by the statistic was inflated if using all SNPs but was appropriate after removing linkage disequilibrium (LD). The cost of choosing too few components is an increase in false positives; the cost of choosing too many components is a reduction of power due to potentially over-fitting the model. However, the more conservative approach, and one favored by all contributions to the group, is to include all plausible components despite the risk of over-fitting.

One important recurring issue: can genomic regions of high LD dominate the PCA components? This was a reasonable expectation for a case-control study of an autoimmune disease like RA with its known MHC involvement. Indeed, among the cases in the NARAC dataset, a minuscule 2.3% were found to have low-risk genotypes, while 56.4% of the controls had low risk genotypes. The MHC, coincidentally, is a region of extensive LD. Accordingly, the ascertainment procedure itself could result in the appearance of regional substructure.

There are four separate questions dealt with by Group 13: first, can one use a subset of SNPs and how does the selection of a subset affect the outcome; second, how can we control for case-control association in a stratified sample; third, are there advantages to modifying the method of dimensionality reduction; and finally, because LD estimates vary by population and are fundamental to most of these methods, how can one best compute LD? Because all teams used real data, often a quantile-quantile (Q-Q) plot and a measurement of the inflation factor (for some phenotype) were generated in order to assess the effectiveness of the methods. Alternatively, some teams present a comparison to a stratification “gold standard” (usually PCA).

## Methods and Results

For the question of subsets of SNPs, Kang et al. [2009] considered local versus global ancestry in unrelated individuals from the Framingham Heart Study Offspring Cohort. They suspected that local genome regions harboring functional polymorphisms may be subject to subtle forms of population stratification. An important example of this is the lactase (*LCT*) gene, which reveals a Northern-Southern European cline and shows a spurious association to height. To test this hypothesis, they examined the inflation factor (for height as a phenotype) when adjusting for ancestry with principal components (PCs) derived from SNPs across the genome (defined as global ancestry) versus SNPs from distinct 20-Mb regions (defined as local ancestry). Inspection of the Q-Q plot and the inflation factors reveals that including local PCs as covariates still results in moderate inflation, whereas controlling for either global PCs alone or global and local PCs combined results in minimal inflation. In the specific case of *LCT*, global PCs, local PCs, and combined global and local PCs all performed well at controlling for spurious association. Local and global PCs provided

different information about ancestry; however, it is unclear which method of adjustment is optimal for the polygenic trait, height.

Peloso et al. [2009] also considered PCA using a subset of SNPs in the RA data. In particular, they considered various levels of thinning to remove LD and examined effects of the inclusion of the MHC region and the region of inversion on chromosome 8. The first two PCs (PC1 and PC2) were highly correlated under pairwise comparison for all subsets. For additional PCs, the inclusion or exclusion of the MHC region made the most significant impact on consistency. When included in the PCA, SNPs from the MHC region and the chromosome 8 inversion provided extremely high weights to the components. When these regions are excluded, only the small region on chromosome 2 containing *LCT* shows unusually high weights to the components. Interestingly, the impact of LD seems minimal with genome-wide data, except for the inversion, which preserves LD across 4 Mb.

Peloso et al. [2009] and colleagues also addressed our second question: how can we control for case-control association in a stratified sample? They found that logistic regression using PC1-PC5 as linear covariates produced the smallest inflation factors. Including both linear covariates and discrete clusters did not improve the inflation. Because this population is all of self-identified European ancestry, this finding may not generalize to a more ethnically diverse sample in which ancestral populations are geographically separated.

Feng et al. [2009] proposed a novel segregation model to perform association analyses in pedigree data. They performed PCA on the parents (or if none had data, then a randomly chosen pedigree member was used) in the Framingham Heart Study data set and then used factor loadings to compute PCs for all individuals. A multivariate logistic model was then used, controlling for 10 PCs, to fit an additive genotype and an index variable indicating whether an individual was chosen for PCA. Using hypertension as phenotype, they examined the Q-Q plots and found that their method successfully controls inflation while simultaneously incorporating pedigree data. This method can also be extended to larger pedigrees. However, they found that the model would frequently fail to estimate the variance matrix of all covariates and familial correlations due to insufficient data.

Zhang et al. [2009b] also proposed a novel association test that does not involve directly computing stratification and can be performed on a single genotyped marker. In particular, in the presence of subpopulations, a deviation from Hardy-Weinberg equilibrium will be observed. The value  $F_{ST}$  is the proportion of the total heterozygosity in the population due to differences in allele frequencies among each subpopulation. Letting  $F_1$  and  $F_2$  denote  $F_{ST}$  in cases and controls, respectively, they construct a likelihood function where  $F_1$  and  $F_2$  are nuisance parameters. This has the general form of a 2-df Pearson chi-square test, but does not have a closed-form solution. Analysis of the RA data shows that when the estimates of  $F_1$  and  $F_2$  are similar for a given marker, the resulting  $p$ -values are also similar. However, discordant values of  $F_1$  and  $F_2$  (when the case and control populations have different substructure at a particular marker) reveal dramatic inflation of  $p$ -values for the standard Pearson test while the proposed method appears to control the inflation.

The third question for our group was to examine different methods of dimensionality reduction. Wang et al. [2009] compared the population structures from PCA and multidimensional scaling (MDS) and evaluated the performance of the two approaches in the RA dataset. First, the team performed PCA using all SNPs and examined the PC loadings on individual SNPs to determine whether these components were dominated by relatively few chromosome regions with extended LD. The team found that regions of high LD did dominate the most significant components when all of the SNPs were used. Indeed, 4,413 of 9,980 SNPs that deviated from their expected quantiles with a distance greater than



1 were in the MHC. A second round of PCA removed SNPs with high loadings and pruned the remaining SNPs based on LD. This round of PCA identified nine outlier individuals whose PC loadings were more than six standard deviations away from the mean.

MDS was performed using PLINK [Purcell et al., 2007]. The MDS analyses started with a list of pruned SNPs based on LD. A pairwise identity-by-state distance analysis was conducted to identify (and remove) individuals deemed to be population outliers. In the MDS analysis a small number of subjects ( $N=7$ ) had to be excluded as outliers. Interestingly, five of these seven outliers were also among the nine outliers excluded by PCA.

Both PCA and MDS gave similar results for the NARAC data. The correlation coefficients between the first four components/dimensions exceeded 0.88 before dropping off rather precipitously. Both methods detected strong (and similar) population stratification, with genomic inflation parameter estimated at 1.447. Logistic regression with the significant PCs as covariates for PCA or the leading dimensions for MDS successfully corrected the inflation factor to 1.037 and 1.045, respectively.

Zhang et al. [2009a] and Hinrichs et al. [2009] both used Laplacian matrices in an attempt to improve or extend the standard PCA. In the present context, the Laplacian matrix is a specially formatted matrix,  $L$ , which is included in the computation of the covariance: instead of  $G^T G$  as previously presented,  $G^T L G$  is computed and then eigenvalue decomposition is performed. This is essentially a weighted PCA. The first team used the RA data (consisting only of unrelated individuals), and defined the weights based on the genetic correlation between pairs of individuals. More distantly correlated individuals play less of a role in the final stratification results than more closely correlated individuals. The final stratification results are then based on local rather than global comparisons. In the case of the RA data, the results show very clearly two stratification axes rather than the diffuse cloud observed without this method.

The second Laplacian team used the related individuals in the Framingham Heart Study data set. In this case, the Laplacian was weighted to allow for related individuals based on the kinship matrix. In particular, the best linear unbiased estimate [McPeck et al., 2004] used for allele frequency estimates in related individuals was adapted to compute pairwise correlation. The results showed consistency with other methods of computation but allowed for use of more genotyped individuals.

He and Willcox [2009] tackled the fourth and final question: how can one best compute LD? Starting with 332 genotyped trios (two parents and an offspring) in the Framingham Heart Study data, they examined LD estimates using different numbers of trios and singletons. Interestingly, they found that 30 to 40 trios produced LD values very close to the values produced by all trios, whereas estimates with unrelated individuals were inaccurate with few or many unrelated individuals. The ability to phase haplotypes with related individuals is likely key to this observation. This is especially important because of the four commonly used HapMap samples; the Japanese and Chinese samples contain only unrelated individuals.

## Discussion

Our group dealt with four issues: SNP selection, association testing, non-standard methods, and LD calculations in stratified or mixed ethnicity samples. We collectively demonstrated that there are continuous axes of ethnic variation in both the RA and the Framingham Heart Study datasets, which are populations of European descent. We further showed that ignoring this creates  $p$ -value inflation for a variety of phenotypes. This can be corrected by standard application of PCA or MDS. One can use local ancestry (as defined above) to assess

population stratification. Related individuals can also be utilized to assess population stratification without creating bias from large pedigrees. On the whole, the method is robust to SNP selection except for areas of extremely high association (such as the MHC region in the RA data) or unusually strong LD in chromosomal inversions. For association tests in these data sets, one can simply include PCs or clusters as covariates in a standard logistic regression. Our group also reported a method for performing an association test controlling for substructure when genome-wide markers are not available to explicitly compute stratification. One limitation of the data set analyzed by our group was the lack of geographically diverse ethnic groups. Our data sets were self-reported Caucasian and the stratification results throughout are consistent with subtle variation in a European ancestry population. Therefore, some of the linear corrections may not be suitable in a more diverse population. Finally, we note that although the LD estimates from the HapMap populations are widely used, there may be systematic problems due to lack of related individuals in all but two of the original four samples (and in six of the eleven Phase III HapMap samples).

## Acknowledgments

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. This work was supported in part by the NIH grant AA015572 from the National Institute on Alcohol Abuse and Alcoholism. The authors are grateful for the many contributions of the Group 13 participants, especially Drs. Kathryn Lunetta, Franz Quehenberger, and Dabao Zhang.

## References

- Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet.* 2000; 66:1933–44. [PubMed: 10801388]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol.* 2001; 60:155–66. [PubMed: 11855950]
- Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics.* 2003; 164:1567–87. [PubMed: 12930761]
- Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol Ecol Notes.* 2007; 7:574–8. [PubMed: 18784791]
- Feng Q, Abraham J, Feng T, Song Y, Elston RC, Zhu X. A method to correct for population structure using a segregation model. *BMC Proc.* 2009; 3(Suppl 7):S104. [PubMed: 20017968]
- Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele frequency distribution for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet.* 2000; 66:216–34. [PubMed: 10631153]
- He Q, Willcox BJ. Linkage disequilibrium of single-nucleotide polymorphism data: How sampling methods affect estimates of linkage disequilibrium. *BMC Proc.* 2009; 3(Suppl 7):S105. [PubMed: 20017969]
- Hinrichs AL, Culverhouse R, Jin CH, Suarez BK. Detecting population stratification using related individuals. *BMC Proc.* 2009; 3(Suppl 7):S106. [PubMed: 20017970]
- Kang SJ, Larkin EK, Song Y, Barnholtz-Sloan J, Baechle D, Feng T, Zhu X. Assessing the impact of global versus local ancestry in association studies. *BMC Proc.* 2009; 3(Suppl 7):S107. [PubMed: 20017971]
- McPeck MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics.* 2004; 60:359–67. [PubMed: 15180661]
- Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genetics.* 2006; 2:e190. [PubMed: 17194218]
- Peloso GM, Timofeev N, Lunetta KL. Principal-component-based population structure adjustment in the North American Rheumatoid Arthritis Consortium data: Impact of single-nucleotide polymorphism set and analysis method. *BMC Proc.* 2009; 3(Suppl 7):S108. [PubMed: 20017972]

- Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *New Engl J Med.* 2007; 357:1119–209.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–9. [PubMed: 16862161]
- Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P, Seligsohn U, Waliszewska A, Schirmer C, Ardlie K, Ramos A, Nemes J, Arbeitman L, Goldstein DB, Reich D, Hirschhorn JN. Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* 2008; 4:e236. [PubMed: 18208327]
- Prichard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet.* 1999; 65:220–8. [PubMed: 10364535]
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000a; 155:945–59. [PubMed: 10835412]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet.* 2000b; 67:170–81. [PubMed: 10827107]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]
- Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol.* 2001; 20:4–16. [PubMed: 11119293]
- Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet.* 2001; 68:466–77. [PubMed: 11170894]
- Suarez BK, Hampe CL. Linkage and association. *Am J Hum Genet.* 1994; 54:554–9. [PubMed: 8116626]
- Wang D, Sun Y, Stang P, Berlin JA, Wilcox MA, Li Q. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: Principal-component analysis versus multidimensional scaling. *BMC Proc.* 2009; 3(Suppl 7):S109. [PubMed: 20017973]
- Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol.* 2003; 24:44–56. [PubMed: 12508255]
- Zhang J, Weng C, Niyogi P. Graphic analysis of population structure on genome-wide rheumatoid arthritis data. *BMC Proc.* 2009a; 3(Suppl 7):S110. [PubMed: 20017975]
- Zhang Y, Xiao X, Wang K. Accommodating population stratification in case-control association analysis: A new test and its application to genome-wide study on rheumatoid arthritis. *BMC Proc.* 2009b; 3(Suppl 7):S111. [PubMed: 20017976]
- Zheng G, Freidlin B, Li Z, Gastwirth JL. Genomic control for association studies under various genetic models. *Biometrics.* 2005; 61:186–92. [PubMed: 15737092]
- Zheng G, Freidlin B, Gastwirth JL. Robust genomic control for association studies. *Am J Hum Genet.* 2006; 78:350–6. [PubMed: 16400614]
- Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. *Genet Epidemiol.* 2002; 23:181–96. [PubMed: 12214310]