Published in final edited form as:

Commun Stat Theory Methods. 2009 January 1; 38(16-17): 2843–2855. doi:10.1080/03610920902947188.

Assessment of Population Structure and Its Effects on Genome-Wide Association Studies

Hongyan Xu and Varghese George

Department of Biostatistics, Medical College of Georgia, Augusta, Georgia, USA

Abstract

Large-scale genome-wide association studies are promising for unraveling the genetic basis of complex diseases. However, population structure is a potential problem, the effects of which on genetic association studies are controversial. Quantification of the effects of population structure on large scale genetic association studies is needed for valid analysis of data and correct interpretation of results. In this study, we performed extensive coalescent-based simulation study with varying levels of population structure to investigate the effects of population structure on large-scale genetic association studies. The effects of population structure are measured by the multiplicative changes of the probability of type I error, which is then correlated with the levels of population structure. It is found that at each nominal level of association tests, there is a positive relationship between the level of population structure and its effects, which could be summarized well with a regression function. It is also found that at a specific level of population structure, its effect on association study increases drastically as the significance level of the test decreases. The type I error is inflated by an amount approximately equal to Wright's F_{ST} , a measure that is used to quantify the magnitude of population structure. Therefore, in genome-wide association studies, the effects of population structure cannot be safely ignored, and must be accounted for with proper methods. This study provides quantitative guidelines to account for the effects of population structure on genome-wide association studies in admixed populations.

1. Introduction

Complex human diseases such as hypertension, diabetes and cancer pose great public health burden on human society. These diseases are “complex” because the etiology involves both environmental factors and multiple genes, with each contributing small to moderate effects. Identification of the underlying genetic factors would be the initial major step toward understanding the molecular mechanisms of these diseases. There are two primary statistical approaches to mapping genes of complex diseases, linkage and association. Genetic linkage analysis is the traditional approach that tests for correlation of transmission patterns of genetic markers and disease traits in families, while association studies tests for the correlation of a specific genetic variant (allele) with disease trait in populations. Generally speaking, association studies have more power and provide greater resolution than linkage analysis (Risch & Merikangas, 1996). Association studies also have simpler study designs, and samples are relatively easier to ascertain than for linkage analysis. Consequently, genetic association studies are becoming more and more popular in recent years. With the development of genomic technologies, large-scale genome-wide associations are becoming increasingly feasible and hold great promises for unraveling the genetic basis of complex diseases. This is made possible by the recent advances in human genome research, in particular, the completion of the human

genome sequence (International Human Genome Sequencing Consortium, 2004; Lander, et al., 2001; Venter, et al., 2001), the availability of millions of genetic markers, especially single nucleotide polymorphisms (SNPs) in public databases, the initiation of the International HapMap Project (The International HapMap Consortium, 2003), and improvements in genotyping technologies such as 500K SNP chip. However, the size of the resulting data sets raises significant issues of analysis and interpretation.

One of the potential problems in association studies is the presence of population structure, which raises the potential for confounding that leads to spurious significant results. For example, if the sample comes from a population having several subpopulations with different allele frequencies, and if the proportions of cases and controls sampled from each subpopulation are not matched, differences in allele frequencies between cases and controls will result in mimicking a statistical signal of association and leading to false positive results. This concern has influenced the design and interpretation of the association studies (Spence, Greenberg, Hodge, & Vieland, 2003). However, the magnitudes of the effects of population structure in real large-scale association studies are still controversial. While some investigators believe that it is not a major concern because the levels of population structure in the general human population are small (Ardlie, Lunetta, & Seielstad, 2002), others argue strongly against ignoring the effect of population structure especially for large-scale association studies (Marchini, Cardon, Phillips, & Donnelly, 2004). Several statistical methods have been proposed to correct for population structure in genetic association studies (Devlin & Roeder, 1999; Price, et al., 2006; Pritchard, Stephens, Rosenberg, & Donnelly, 2000). These approaches are quite different and have different theoretical basis. It is unclear when one should use a specific method as opposed to a different choice. Therefore, a systematic assessment is needed to quantify the levels of population structure and its effects on genetic association studies.

The first step to quantify the effects of population structure is to choose an appropriate measure of population structure. A commonly used measure is Wright's F_{ST} (Wright, 1943). In a large population, consider a genetic locus with 2 alleles, A_1 and A_2 . There are 3 possible genotypes at this locus in the population - A_1A_1 , A_1A_2 , A_2A_2 . Among these 3 genotypes, A_1A_1 and A_2A_2 are called homozygotes and A_1A_2 are called heterozygotes. The frequency of heterozygotes in the population is called heterozygosity, denoted H , and it is a measure of the genetic variation at this particular locus. Under random mating and other regular assumptions, the genotype distribution will reach Hardy-Weinberg equilibrium (HWE). That is, the frequencies for A_1A_1 , A_1A_2 , and A_2A_2 will be p^2 , $2pq$, q^2 , respectively, where p is the frequency of allele A_1 and $q = 1 - p$ is the frequency of allele A_2 . When a large population splits into several small sub-populations, these sub-populations evolve somewhat independently and the sub-populations will diverge genetically. In this case, it is useful to differentiate genetic variations between subpopulations and those within sub-populations. F_{ST} is a measure of population structure that quantifies the proportion of genetic variations between sub-populations with respect to the total genetic variation. In the presence of population structure, the observed heterozygosity within sub-populations will deviate from the heterozygosity under HWE, with fewer heterozygotes than what is expected under HWE. We could measure the genetic variation within subpopulation using the average heterozygosity across sub-populations, denoted H_S . Similarly, the overall genetic variation in the total population could be measured using the expected heterozygosity under HWE, denoted H_T . Then F_{ST} is given by $F_{ST} = (H_T - H_S)/H_T$.

In section 2, we review methods of estimating F_{ST} and procedures for testing association in the presence of population structure. We then present the simulation model used to generate samples with population structure, and the results from the simulation study. In section 3, we present a new estimator of population structure using all the marker information simultaneously, and investigate its relationship with F_{ST} .

2. Methods

2.1 Estimation of F_{ST}

F_{ST} is estimated using the unbiased estimator, $\hat{F}_{ST}$, described by (Weir & Cockerham, 1984). Specifically, suppose samples are drawn from P subpopulations and there are two alleles, A and a , at any given SNP. Let the frequency of allele A in the i^{th} population be p_i , the average allele frequency across populations be $\bar{p}$, and the sample size from the i^{th} population be n_i . Then the observed mean square error of allele frequency within a population, denoted MSI , is given by

$$MSI = \frac{1}{\sum_{i=1}^P n_i - 1} \sum_{i=1}^P n_i p_i (1 - p_i).$$

The observed mean square error of allele frequency between populations, denoted MSP , is given by

$$MSP = \frac{1}{P - 1} \sum_{i=1}^P n_i (p_i - \bar{p})^2.$$

F_{ST} is then estimated as,

$$\hat{F}_{ST} = \frac{MSP - MSI}{MSI + (n_c - 1)MSI},$$

where n_c is the average sample size across subpopulations to correct for the different sample sizes from each population, which is given by

$$n_c = \frac{1}{P - 1} \left(\sum_{i=1}^P n_i - \frac{\sum_{i=1}^P n_i^2}{\sum_{i=1}^P n_i} \right).$$

2.2 Measuring association at an SNP locus in the presence of population structure

In order to quantify the effects of population structure on tests for association, we simulated samples under the null hypothesis of no association where we randomly assigned case/control status while keeping population structure in the total sample. For each SNP locus, 1,000 random assignments of case/control status were performed. The association for each assignment was measured using Armitage's trend test under an additive genetic model (Armitage, 1955), which has been shown to be robust against deviations of genotype frequencies from Hardy-Weinberg proportions (Sasieni, 1997). Suppose the two alleles at an SNP locus are denoted as A and a , then this test statistic is given by

$$Y^2 = \frac{N[N(k_1 + 2k_2) - k(n_1 + 2n_2)]^2}{k(N - k)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]},$$

where N is the total sample size, k is the number of cases, n_1 and n_2 are the number of individuals with genotypes Aa and AA , respectively, in the total sample, and, k_1 and k_2 are the number of cases with genotypes Aa and AA , respectively. Under the null hypothesis of no association, and assuming no population structure, the test statistic Y^2 follows a χ^2 distribution with 1 degree of freedom. Y^2 was computed for each of the 1000 samples with randomly assigned case/control status at each SNP. The empirical distribution of Y^2 from samples with population structure was compared with the χ^2_1 distribution to study the effects of population structure under the null hypothesis.

2.3 Simulation models for the assessment of population structure

In order to systematically examine the effect of population structure on genetic association studies, we took a computer simulation approach to generate samples with various levels of population structure. All the simulations were performed under the null hypothesis of no association between the SNPs and the disease status. The basic study design is a case-control design. The samples were simulated using the standard coalescent-based algorithm assuming a finite-island model (Wright, 1943), which is commonly used to model subdivided populations. Specifically, a homogeneous ancestor population was split into k isolated sub-populations, T generations ago from current sampling time. The sub-populations are assumed to evolve independently after they were separated from the ancestor population. Varying levels of population structure could be created in each simulation, as it is a realization of a stochastic evolution process leading to the simulated samples. Coalescent theory is the recent development in theoretical population genetics concerning the stochastic evolution process (Kingman, 1982a, 1982b, 2000). It has reshaped the field of population genetics and is efficient in simulating samples under various demographic models, because it focuses on the history of samples rather than the statistical properties of the entire population (Hudson, 1991). It differs from the traditional approach in that it looks backward in time from current sampling time. Looking backward, we see that samples join together or coalesce successively at common ancestors and, thus, the history of the sample could be described as a tree. The coalescent-based algorithm first generates a binary tree from the sample to their common ancestor with the branch length following an exponential distribution with mean $4N/k(k-1)$, where N is the effective population size and k is the number of samples at a particular time. Given the branch length t , the algorithm then simulates the number of mutations along each branch from a Poisson distribution with mean μt , where μ is the mutation rate per sequence per generation. In the finite island model, the samples from each subpopulation were allowed to coalesce in their respective subpopulations until the separation time T , and the remaining samples were pooled together to continue the coalescence process in the ancestor population until one single common ancestor for all the sample sequences was reached (Hudson, 1991).

We simulated samples of 1000 individuals from 2 subpopulations. Among them, 900 are from population 1 and 100 are from population 2. Genotypes at SNP markers were simulated from 100 consecutive genomic regions among which recombination events were allowed. The SNP genotypes were derived from the simulated coalescent tree assuming infinite site model, where each mutation creates a new SNP site. Therefore, the number of SNPs in each simulation is a random variable with value equal to the number of mutation in the simulation given a fixed mutation rate. For each simulated sample, we randomly selected 500 individuals from population 1 as cases and the remaining 400 individuals from population 1 and 100 individuals from population 2 were treated as controls. We are mimicking a sampling scenario where cases were from one population, while the controls were mismatched with a small fraction from the other population.

2.4 Simulation Results

We performed 1000 simulations. From each replicate, we generated SNP genotypes from 100 consecutive genomic regions. The number of markers varies for each replicate, and ranges from 433 to 1078 with a mean of 719. Figure 1 gives the allele frequency distribution of the simulated markers in the two sub-populations. The number of alleles in each bin decreases as the allele frequency increases, which is similar to the allele frequency distribution for SNPs reported from the HapMap project (Frazer, et al., 2007). The frequency in each bin is also similar. For each simulated sample, we also generated 1000 samples with randomly assigned case/control status as outlined in the Methods section. Therefore, for each simulated sample with a specific level of population structure, the number of association tests is in the order of 10^6 , which ensures that we have enough tests to generate the empirical distribution of the Y^2 statistic and examine the effects of population structure even at very low significance levels. Samples were simulated using the coalescent algorithm assuming finite island model as detailed in section 2.3. Each simulated sample is a realization of a stochastic evolutionary process. The levels of population structure in our simulated samples, measured by F_{ST} , range from 0.022 to 0.487, with an estimated mean value of 0.15 and standard deviation of 0.08. Figure 2 shows the histogram of the distribution of F_{ST} in the samples.

The effects of population structure on genetic association studies were assessed by comparing the empirical distribution of Y^2 test statistic under the null hypothesis to the expected χ^2_1 distribution. Given a specific significance level α , the number of tests in which the null hypothesis is rejected can be counted among all the tests for a specific level of population structure. The empirical type I error probability, α_E , is estimated by the ratio of the number of tests rejected under the null hypothesis over the total number of tests. The empirical type I error probability is then compared to the nominal α level. The ratio of the empirical type I error probability over nominal type I error probability is defined as multiplicative change of type I error, denoted M_α . Since the simulations were performed under the null hypothesis, allele frequencies in cases and controls differ not by true association with the genetic marker but because of the presence of population structure. If population structure has little effect on the association tests, the empirical type I error probability and the nominal level will agree, in general, and the multiplicative change will be close to 1. On the other hand, if population structure has significant effect on the association test, the empirical type I error probability tends to disagree with the nominal level, and the multiplicative change will deviate from the null value of 1. Therefore, the magnitude of the multiplicative change of type I error could be used as a measure to quantify the effects of population structure on association studies.

Figure 3 shows the multiplicative changes of type I error at various nominal significance levels. The multiplicative changes are shown in logarithm scale with base 10. Two results are obvious from the figure. First, the type I error probability is inflated because of population structure, and the level of inflation can be very high. For example, empirical type I error probability is inflated by over 1000-fold for all simulated samples with varying F_{ST} when the tests were performed at 10^{-5} level. Thus, in this case, ignoring population structure will result in highly spurious significance and false positives. As expected, the magnitude of the inflation increases with the level of population structure as reflected by the increase in multiplicative changes. The other interesting result is that the magnitude of inflation of type I error probability depends on the significance level of the test. For example, when $F_{ST} = 0.065$, which is a likely value for major human populations (Nei, 1975; Rosenberg, et al., 2002), type I error probability was inflated on average by 8, 15, 29, and 54 times when the significance level of the tests were 10^{-2} , 10^{-3} , 10^{-4} , and 10^{-5} , respectively. In all cases, type I error probability was inflated further when the significance levels were lower.

To further characterize the effects of population structure on genetic association studies and provide quantitative guidelines to gauge the magnitude of the inflation of type I error probabilities, we fit a linear regression function for the multiplicative changes for each significance level of the tests, with the measure of population structure as the predictor variable. i.e., the regression function is, $M_\alpha = \alpha_0 + \beta F_{ST} + \varepsilon$, where α_0 and β are the regression coefficients.

Table 1 gives the slope, intercept and R^2 values of the regression function at different significance levels. The standard errors of the slope and the intercept are given in parentheses. From the R^2 values, it can be seen that the regression line fits the data well. With the regression line, we could make quantitative predictions of the actual type I error probability for a given level of population structure at a nominal significance level.

As evident from table 1, there is approximately a ten-fold increase in slope as the nominal α decreases by 10^{-1} . Also, the slope is approximately equal to the reciprocal of the nominal α . The theoretical value for the intercept is 1 (the value of the multiplicative change under no population structure). Noting that the outcome variable is the ratio of the actual (empirical) type I error to the nominal significance level, this implies that the actual type I error is approximately the sum of the nominal significance level and F_{ST} , which is a very interesting and, potentially, a highly useful result. However, further confirmatory investigation is needed to support this conclusion.

3. A New measure of population structure

Given the genotype information at several markers, F_{ST} is usually estimated at each marker separately. In general, these estimates are different for distinct markers. For a particular set of subpopulations in one population, the general practice is to take the average of the estimates from each marker and use it as an overall measure of population structure. With the development of advanced genotyping technology, we now have access to genetic information on genome-wide markers, especially on SNPs which are the polymorphic sites at single nucleotides in the population. It is therefore feasible to measure population structure utilizing the information at genome-wide markers simultaneously through a genomic approach. Recently, Nicholson, et. al. proposed a method for estimating population differentiation and isolation from SNP data (Nicholson, et al., 2002), which is subpopulation-specific and common for all loci. For the j^{th} subpopulation, let c_j denote this measure of population structure. Each c_j is comparable to the subpopulation-specific version of the F_{ST} , and it measures the divergence of the j^{th} subpopulation from the common ancestor. However, it would be more desirable if one single measure could summarize the level of population structure across all subpopulations. We propose a new measure, denoted C , by combining all j measures of c_j for each subpopulation. For each sample in the simulated data set, c_j is estimated for each subpopulation using a Bayesian approach as implemented in the software package Popgen (Nicholson, et al., 2002; Pritchard, Stephens, & Donnelly, 2000). Uniform priors are used for each c_j , and Markov Chain Monte Carlo (MCMC) is employed to sample from the posterior distribution. The Markov Chain was run for 20,000 iterations and the first 10,000 iterations were discarded as burn-in. We estimate c_j from the mean of the posterior distribution. When summarizing the level of population structure across subpopulations, it is desirable to have a single statistic, instead of one for each subpopulation. Therefore, we take the weighted mean of the all values of c_j as the new measure of the overall population structure. Specifically, suppose we have P subpopulations, and let c_j be the measure of population structure for the j^{th} subpopulation, as defined above. The new overall measure of population structure is defined as

$$C = \sum_{j=1}^P w_j c_j$$

where w_j is the weight for the j^{th} subpopulation. There are many possible weighting schemes that can be employed. Here, we use sample size from each subpopulation as the weight. That

is, $w_j = \frac{n_j}{\sum_{i=1}^P n_i}$, where n_j is the number of individuals from subpopulation j .

The new measure C is specific for SNPs and takes account of the ascertainment bias in the process of SNP discovery. Since SNP discovery is generally conducted in small samples, SNPs with high minor allele frequencies are more likely to be discovered than SNPs with low minor allele frequencies, thus creating potential ascertainment bias. It has been shown that the ascertainment bias could affect the estimation of population parameters in genetic analysis (Wakeley, Nielsen, Liu-Cordero, & Ardlie, 2001). This ascertainment bias has been explicitly accounted for in the method for estimating the individual c_j for each sub-population, and hence in the estimation of the overall C . Another advantage, as pointed out by Nicholson, et. al. (2002) is that for point estimation, the estimates of c_j s and hence that of C can substantially outperform F_{ST} and other similar methods. Also, this approach appropriately handles differing sample sizes across loci and/or subpopulations.

Table 2 gives the correlation coefficient and R^2 between the estimates of F_{ST} and C for varying sample sizes, from replicates of simulated samples. The number in the sample size column is the number of individuals simulated from each sub-population. Clearly, the correlation between estimates of F_{ST} and the new measure C increases as the sample size increases though the correlation is high at all the sample sizes we have tried. Even when the sample size is as small as 30 individuals from each sub-population, we still have $R^2 = 0.89$. Given this high correlation, similar results for the magnitude of the inflated type I error, as those obtained in Section 2, should hold using C instead of F_{ST} .

4. Discussion

Presence of population structure could potentially be a confounding factor in genetic association studies in admixed populations, and can lead to spurious significance and false positives. However, whether it could pose serious problem to the studies in human populations is controversial. Some studies argued that population structure is not a major issue (Ardlie, et al., 2002; Wacholder, Rothman, & Caporaso, 2002), while others conclude that population structure cannot be safely ignored, especially in large-scale studies (Freedman, et al., 2004; Marchini, et al., 2004). The general solution is to match samples in cases and controls by ancestry. However, this may be difficult for association studies in admixed populations such as African American population, because the degree of admixture varies among individuals, and population structure can be considered as a continuous variable in this case. Moreover, using genome-wide markers, genetic structures have been discovered among those previously regarded as “homogenous” populations (Campbell, et al., 2005; Salmela, et al., 2008). Therefore, matching by ancestry or geographic region may not be sufficient.

In this study, we took a computer simulation approach to assess the effect of population structure on genetic association studies. Samples with varying levels of population structure were generated using a coalescent-based algorithm. Effects of population structure on association studies were assessed quantitatively in terms of changes in actual type I error probability as a function of population structure and the nominal significance level. Results

shown in Figure 2 indicate that population structure can cause inflation of type I error probability. The degree of inflation increases with the level of population structure and it can be very serious in some cases. The results in Figure 2 also indicate that as the nominal significance level decreases, the problem posed by a given amount of population structure becomes more serious. This could have important implications for large scale genome-wide association studies in which millions of markers are usually used. Owing to the large number of markers tested in such studies, correcting for multiple testing would mean that any "significant" result must have a much lower p-value in order for the association results to be considered "significant" at genomic level. Usually, genome-wide significance level is set a very low range of 10^{-4} to 10^{-8} . Our results indicate that in this range, the problem posed by population structure becomes very serious indeed with type I error probability inflated over 1000 times. As a consequence, the effects of population structure cannot be safely ignored in genome-wide association studies, and appropriate steps must be taken to correct for the effects of population structure.

Through systematic simulations of samples with levels of population structure in a wide range and using a regression analysis approach, we describe the multiplicative changes of type I error probability as a function of population structure and nominal significance level of association tests. This result could be used as a quantitative guideline to assess the effects of population structure on genetic association studies. For a given level of population structure and significance level α , we could use the regression line to find the expected change in type I error probability. The regression function includes those with very low significance levels that are commonly used in genome-wide studies. Therefore, they could provide guidelines for genome-wide association studies.

A strong empirical finding of our study is that the actual type I error can be approximated as the sum of the nominal type I error and F_{ST} . In other words, the type I error is inflated by a magnitude approximately equal to F_{ST} . Thus, we give a simple method to assess the actual type I error in the presence of population structure. It is noteworthy that the minimum value of type I error that can be used in an association study must be larger than F_{ST} , imposing type I error restrictions on association studies of admixed populations. This is especially a critical issue in genome-wide association studies wherein extremely small α -level is typically preferred, for reasons described above.

The proposed estimator, C , is highly correlated with $\hat{F}_{ST}$, and it has some added advantages. First, it is specific for SNPs and takes account of the ascertainment bias in the process of SNP discovery. Second, since the estimator of C is derived within the Bayesian framework, it is very flexible in modeling and can incorporate prior information on the parameters. In our simulation studies, we used non-informative prior for the c_j parameter. However, any prior knowledge regarding the parameters can be incorporated in the estimation, which can lead to more precise estimates than the moment-based estimates of F_{ST} (Nicholson, et al., 2002).

References

- Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 2002;71(2):304–311. [PubMed: 12096349]
- Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955;11:375–386.
- Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, et al. Demonstrating stratification in a European American population. *Nat Genet* 2005;37(8):868–872. [PubMed: 16041375]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55(4):997–1004. [PubMed: 11315092]

- Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449(7164):851–861. [PubMed: 17943122]
- Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004;36(4):388–393. [PubMed: 15052270]
- Hudson, RR. Gene genealogies and the coalescent process. In: Futuyma, D.; Antonovics, J., editors. *Oxford surveys in evolutionary biology*. Vol. Vol. 7. Oxford: Oxford University Press; 1991. p. 1–44.
- International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431(7011):931–945. [PubMed: 15496913]
- Kingman JFC. The coalescent. *Stochastic Processes and their Applications* 1982a;13:235–248.
- Kingman JFC. On the genealogy of large populations. *Journal of Applied Probability A* 1982b;19:27–43.
- Kingman JFC. Origins of the coalescent. 1974–1982. *Genetics* 2000;156(4):1461–1463. [PubMed: 11102348]
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409(6822):860–921. [PubMed: 11237011]
- Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;36(5):512–517. [PubMed: 15052271]
- Nei, M. *Molecular Population Genetics and Evolution*. New York: North-Holland – American Elsevier; 1975.
- Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Ser. B Stat. Methodol* 2002;64(4):695–715.
- 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(8):904–909. [PubMed: 16862161]
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155(2):945–959. [PubMed: 10835412]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000;67(1):170–181. [PubMed: 10827107]
- Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273(5281):1516–1517. [PubMed: 8801636]
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. *Science* 2002;298(5602):2381–2385. [PubMed: 12493913]
- Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, et al. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* 2008;3(10):e3519. [PubMed: 18949038]
- Sasieni PD. From genotypes to genes: doubling the sample size. *Biometrics* 1997;53(4):1253–1261. [PubMed: 9423247]
- Spence MA, Greenberg DA, Hodge SE, Vieland VJ. The emperor's new methods. *Am J Hum Genet* 2003;72(5):1084–1087. [PubMed: 12704569]
- The International HapMap Consortium. The International HapMap Project. *Nature* 2003;426(6968):789–796. [PubMed: 14685227]
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* 2001;291(5507):1304–1351. [PubMed: 11181995]
- Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 2002;11(6):513–520. [PubMed: 12050091]
- Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. The discovery of single-nucleotide polymorphisms--and inferences about human demographic history. *Am J Hum Genet* 2001;69(6):1332–1347. [PubMed: 11704929]
- Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* 1984;38:1358–1370.

Wright S. Isolation by Distance. *Genetics* 1943;28(2):114–138. [PubMed: 17247074]

Figure 1.
Allele frequency distribution in the two sub-populations from the simulated samples.

Figure 2.
Histogram of the distribution of estimates of F_{ST} from the simulated samples.

Figure 3. Effects of population structure on genetic association studies as measured by multiplicative changes of type I error probability versus F_{ST} at several significance levels.

Table 1

Estimates of regression coefficients of multiplicative changes of type I error probability on F_{ST} at various significance levels.

α	Slope (SE)	Intercept (SE)	R^2
10^{-2}	114.96 (3.44)	4.10 (2.63)	0.919
10^{-3}	1120.7 (27.27)	-8.27 (5.03)	0.945
10^{-4}	10807.97 (305.35)	-368.83 (56.36)	0.927
10^{-5}	104095 (3459.9)	-5533 (638.6)	0.902

Table 2Correlation between estimates of F_{ST} and C from simulations with varying sample sizes

Sample size	Correlation coefficient	R^2
30	0.944	0.89
60	0.959	0.92
90	0.973	0.95