Tests for difference in population structure between two samples with application to HapMap genotype data

Kai Wang

Department of Biostatistics, University of Iowa, Iowa City, IA 52242

Received _________________________; accepted _________________________

Contact information: Kai Wang, PhD, Department of Biostatistics, C227 GH, College of Public Health, University of Iowa, Iowa City, IA 52242. E-mail: `kai-wang@uiowa.edu`, phone: (319) 384-5175, fax: (319) 384-5018.

ABSTRACT

Population structure is a phenomenon caused by population stratification or cryptic relatedness. It exists even in populations that appear to be homogeneous. Difference in population structure between cases and controls can inflate false positive rates in genetic association studies. Many statistical methods have been proposed to eliminate this side effect. However, statistical tests for detecting the existence of such difference are lacking. Statistical tests are proposed to fill this gap. Performance of these tests are evaluated through extensive simulation studies. These tests are applied to the HapMap genotype data on the Yoruba people of Ibadan, Nigeria (YRI), U.S. residents with northern and western European ancestry (CEU) by the Centre d'Etude du Polymorphisme Humain, Han Chinese from Beijing, China (CHB), and Japanese from Tokyo, Japan (JPT). Difference in population structure is found to be the largest between YRI and CHB+JPT, followed by CEU and YRI, and CEU and CHB+JPT while no difference between CHB and JPT is detected.

Subject headings: Population stratification, F parameter, case-control design, genetic association

Introduction

It has been long recognized that population structure is a potential confounder in genetic association analysis. It has gained more and more attention with the increased popularity of large scale genetic association studies. Population structure is caused by population stratification or cryptic relatedness. It exists even in populations that seem to be homogeneous, for instance, European American[1] and Han Chinese[2,3]. See Astle and Balding[4] for a recent review on this topic. Many methods have been proposed for association studies to eliminate the confounding effect of population structure[5–13].

Despite the serious consequences population structure could have on genetic association studies, there is a lack of formal statistical tests to help researchers to decide whether or not to use methods that take population structure into account. Indeed, there is no need to sacrifice power when it is unnecessary. Sometimes the existence of population structure may be obvious, for instance, if the study subjects are of different ethnic background, but sometimes it may be not. A common approach for identifying population structure is to use multivariate analysis methods such as cluster analysis, multidimensional scaling, or principle component analysis. Subjective judgements are often made based on scatter plots of proxy variables such as principal components[14].

Existence of population structure does not necessarily cause inflated false positive rates in case-control association studies. It does only when there is a difference in the pattern of population structure between the cases and the controls. Suppose that there are J sub-populations. The two alleles at a biallelic marker are denoted by a and A, respectively. The frequency of genotypes aa, aA, and AA are denoted by $P_0^{(j)}$, $P_1^{(j)}$, and $P_2^{(j)}$, respectively, in sub-population j. Let a_j denote the proportion of sub-population j in cases and b_j in controls. If this marker is not associated with the case-control phenotype status, frequencies of these genotypes would be the same in cases as in controls. The frequency of allele A

would be $p_1 = \sum_{j=1}^J a_j(P_2^{(j)} + P_1^{(j)}/2)$ for cases and $p_2 = \sum_{j=1}^J b_j(P_2^{(j)} + P_1^{(j)}/2)$ for controls. So it is still possible that $p_1 = p_2$ even though there exists population structure in either sample. The condition $p_1 = p_2$ is the null hypothesis underlying the Armitage test for trend that is popular for association studies. Hence the false positive rate is not necessarily inflated when population structure is present.

Population structure not only affects the allele frequency but also causes departure from the Hardy-Weinberg equilibrium (HWE) even if each sub-population is in HWE[15]. The extent of departure can be measured by a parameter F, which can be interpreted as the proportionate reduction in heterozygosity relative to a population in HWE[15]. It is also a measure of the variation of the allele frequency among sub-populations. In the context of the previous paragraph, since the frequency of A allele is the same in cases as in controls in each sub-population, the departure from HWE in cases would be different than in controls. As evidenced by the work to be presented, it is possible to use the difference of the parameter F between cases and controls as a surrogate of the difference in population structure.

In what follows, we describe the setup of the new methods. We introduce two single-marker test statistics and their multi-marker generalizations. Simulation studies will be used to assess the type I error rate and the power of these statistics. Finally, we report an application of these statistics to the HapMap genotype data.

The work in this report is motivated by the issue of population structure in genetic association studies. The methods are presented through cases and controls. However, these methods are applicable to any two samples of interest, as illustrated by their application to the HapMap genotype data.

Method

Consider a biallelic marker such as a single nucleotide polymorphism (SNP). Denote the two alleles by A and a, respectively. The frequencies of genotypes aa, aA, and AA are denoted by P_{10}, P_{11}, and P_{12}, respectively, in cases and by P_{20}, P_{21}, and P_{22}, respectively, in controls. Suppose that there are n_{10}, n_{11}, and n_{12} individuals of these genotypes, respectively, in cases and n_{20}, n_{21}, and n_{22} individuals of these genotypes, respectively, in controls. Let $n_{1+} = n_{10} + n_{11} + n_{12}$ be the total number of individuals in cases and $n_{2+} = n_{20} + n_{21} + n_{22}$ in controls. The total number of individuals involved in the study is denoted by n_{++} $(= n_{1+} + n_{2+})$.

Let p_1 be the frequency of allele A in cases and p_2 in controls. Let F_1 and F_2 be parameter F for cases and controls, respectively. Besides its interpretations given in the Introduction section, this F parameter can be interpreted as the probability that a pair of alleles in a population are identical by descent or the correlation coefficient between the indicators of a pair of alleles when the mating is random[16]. The former interpretation necessarily requires F to be in the interval $[0, 1]$ while the latter does not. The genotype frequencies can be written in terms of p_i and F_i as follows:

$$\begin{aligned} P_{i2}(p_i, F_i) &= F_i p_i + (1 - F_i)p_i^2, \quad i = 1, 2, \\ P_{i1}(p_i, F_i) &= 2(1 - F_i)p_i(1 - p_i), \quad i = 1, 2, \\ P_{i0}(p_i, F_i) &= F_i(1 - p_i) + (1 - F_i)(1 - p_i)^2, \quad i = 1, 2. \end{aligned} \tag{1}$$

Obviously, p_i and F_i provide an alternative parameterization to P_{i2}, P_{i1}, and P_{i0}.

The vector of genotype counts (n_{10}, n_{11}, n_{12}) in cases follows a trinomial distribution with parameters n_{1+} and (P_{10}, P_{11}, P_{12}). The vector of genotype counts (n_{20}, n_{21}, n_{22}) in controls follows a similar trinomial distribution as well. The log-likelihood function for the data (cases and controls) is

$$l(p_1, p_2, F_1, F_2) = l_1(p_1, F_1) + l_2(p_2, F_2),$$

where

$$l_i(p_i, F_i) = \sum_{j=1}^{3} n_{ij} \log[P_{ij}(p_i, F_i)].$$

We are interested in testing whether $F_1 = F_2$ holds. For this purpose, we consider two sets of hypotheses. One set assumes $p_1 = p_2$ while the other one does not. Hence the first set of hypotheses are

$$H_0' : F_1 = F_2, p_1 = p_2, \text{ versus } H_1' : F_1 \neq F_2, p_1 = p_2$$

and the second set of hypotheses are

$$H_0'' : F_1 = F_2, p_1, p_2, \text{ versus } H_1'' : F_1 \neq F_2, p_1, p_2.$$

It is straightforward to compute the likelihood ratio statistic for either set of hypotheses. However, no explicit formulae are available. For the ease of computation (especially in the case of multiple markers to be discussed later), we would consider the score statistic for the first set of hypotheses and the Wald statistic for the second set.

Let $\psi = n_{1+}/n_{++}$ be the proportion of case individuals out of total individuals. Define $T = n_{12}/n_{+2} + n_{10}/n_{+0} - 2n_{11}/n_{+1}$. In the appendix, it is shown that the score statistic for testing H_0' against H_1' is

$$S = \frac{T^2}{\psi(1-\psi) \cdot (1/n_{+2} + 4/n_{+1} + 1/n_{+0})}.$$

Statistic S approximately follows a chi-square distribution with 1 degree of freedom under the null H_0'. This can be directly verified as follows. Fixing the counts n_{+2}, n_{+1}, and n_{+0}, each of n_{12}, n_{11}, and n_{10} follows independently a binomial distribution with a common probability of "success" ψ under the null H_0'. Hence the variance of T equals

$$\begin{aligned} & Var(n_{12}/n_{+2}) + 4Var(n_{11}/n_{+1}) + Var(n_{10}/n_{+0}) \\ &= \psi(1-\psi)/n_{+2} + 4\psi(1-\psi)/n_{+1} + \psi(1-\psi)/n_{+0} \\ &= \psi(1-\psi)(1/n_{+2} + 4/n_{+1} + 1/n_{+0}). \end{aligned}$$

There is a numerical problem in the computation of S statistic when any one of n_{+0}, n_{+1}, or n_{+2} is 0 as they appear in the denominators of the fractions. When such situation occurs, the value of statistic S is set to missing value.

How is T related to allele frequency and genotype frequencies? Let $\hat{P}_{ij} = n_{ij}/n_{i+}, i = 1, 2, j = 0, 1, 2$, be the observed genotype frequencies in cases ($i = 1$) and in controls ($i = 2$) and $\hat{P}_j = n_{+j}/n_{++} = \psi\hat{P}_{1j} + (1-\psi)\hat{P}_{2j}, j = 0, 1, 2$, be the observed genotype frequencies in the pool of cases and controls. Then

$$\begin{aligned} T &= \left(\frac{n_{12}}{n_{12}+n_{22}} - \frac{n_{11}}{n_{11}+n_{21}}\right) + \left(\frac{n_{10}}{n_{10}+n_{20}} - \frac{n_{11}}{n_{11}+n_{21}}\right) \\ &= \frac{n_{12}n_{21} - n_{11}n_{22}}{(n_{12}+n_{22})(n_{11}+n_{21})} + \frac{n_{10}n_{21} - n_{11}n_{20}}{(n_{10}+n_{20})(n_{11}+n_{21})} \\ &= \frac{\psi(1-\psi)}{\hat{P}_1}\left[\frac{\hat{P}_{12}\hat{P}_{21} - \hat{P}_{11}\hat{P}_{22}}{\hat{P}_2} + \frac{\hat{P}_{10}\hat{P}_{21} - \hat{P}_{11}\hat{P}_{20}}{\hat{P}_0}\right]. \end{aligned}$$

Let $p = P_2 + P_1/2$ be the pooled frequency of allele A for cases and controls and $q = 1 - p$. It is straightforward to verify that, under H_1', T converges to

$$\frac{\psi(1-\psi)}{1-F}\left(\frac{p}{P_2} + \frac{q}{P_0}\right)(F_1 - F_2), \tag{2}$$

as $n_{1+} \to \infty$, $n_{2+} \to \infty$, and $n_{1+}/(n_{1+} + n_{2+}) = \psi$. In other words, $T' = T/\alpha$ is a consistent estimator of $F_1 - F_2$ under hypothesis H_1', where

$$\alpha = \frac{\psi(1-\psi)}{1-\hat{F}}\left(\frac{\hat{p}}{\hat{P}_2} + \frac{\hat{q}}{\hat{P}_0}\right).$$

Here $\hat{p} = \hat{P}_2 + \hat{P}_1/2$, $\hat{q} = 1-\hat{p}$, $\hat{P}_2 = n_{+2}/n_{++}$, $\hat{P}_0 = n_{+0}/n_{++}$, and $\hat{F} = 1-(1-\hat{P}_0-\hat{P}_2)/(2\hat{p}\hat{q})$ are the respective estimates of the corresponding parameters. The asymptotic variance of $n_{++}^{1/2}T/\alpha$ can be estimated by $\psi(1-\psi)\beta/\alpha^2$, where $\beta = 1/\hat{P}_2 + 4/\hat{P}_1 + 1/\hat{P}_0$ with the additional notation $\hat{P}_1 = n_{+1}/n_{++}$.

To introduce the multiple-marker version of the S statistic for a collection of SNPs in linkage equilibrium with each other, we consider the weighted sum of the T' statistic with weights proportional to the inverse of their variances. It is well known that such a linear combination has the smallest variance when each T' has the same expectation. Let subscript k index the kth SNP. A multi-marker version of the S statistic is defined as

$$\begin{aligned}\bar{S} &= \frac{(\sum_k [Var(T'_k)]^{-1} T'_k)^2}{\sum_k [Var(T'_k)]^{-1}} \\ &= \frac{n_{++}}{\psi(1-\psi)} \cdot \frac{\left(\sum_k \alpha_k^2 T'_k/\beta_k\right)^2}{\sum_k \alpha_k^2/\beta_k}.\end{aligned}$$

Asymptotically, $\bar{S}$ follows a chi-square distribution with 1 degree of freedom. Markers at which the statistic S takes the missing value are excluded.

Now we derive the Wald statistic for testing H_0'' against H_1''. Since $P_{i1} = (1-F_i)\cdot 2p_i(1-p_i)$, F_i can be written in terms of P_{i0} and P_{i2} as

$$\begin{aligned}F_i(P_{i0}, P_{i2}) &= 1 - \frac{P_{i1}}{2p_i(1-p_i)} \\ &= 1 - \frac{1 - P_{i0} - P_{i2}}{2[1/2 + (P_{i2} - P_{i0})/2][1/2 - (P_{i2} - P_{i0})/2]} \\ &= 1 - \frac{1 - P_{i0} - P_{i2}}{1/2 - (P_{i2} - P_{i0})^2/2}.\end{aligned}$$

Define ζ_i and ξ_i as

$$\zeta_i = \frac{\partial F_i(P_{i0}, P_{i2})}{\partial P_{i0}} = \frac{1}{2p_i q_i} + \frac{P_{i1}}{(2p_i q_i)^2}(P_{i2} - P_{i0})$$

and

$$\xi_i = \frac{\partial F_i(P_{i0}, P_{i2})}{\partial P_{i2}} = \frac{1}{2p_i q_i} - \frac{P_{i1}}{(2p_i q_i)^2}(P_{i2} - P_{i0}).$$

Applying the Delta method, the variance of $\hat{F}_i = F_i(\hat{P}_{i0}, \hat{P}_{i2})$ is

$$V_i \equiv n_{i+}^{-1}(\hat{\zeta}_i^2 \hat{P}_{i0}(1-\hat{P}_{i0}) - 2\hat{\zeta}_i\hat{\xi}_i\hat{P}_{i0}\hat{P}_{i2} + \hat{\xi}_i^2\hat{P}_{i2}(1-\hat{P}_{i2}))$$

where $\hat{\zeta}_i$ and $\hat{\xi}_i$ are the values of ζ_i and ξ_i evaluated at $\hat{p}_i$, $\hat{q}_i$, $\hat{P}_{i0}$, $\hat{P}_{i1}$, and $\hat{P}_{i2}$. Let $\hat{F}_i$ denote the value of $F_i(P_{i0}, P_{i2})$ evaluated at $\hat{P}_{i0}$ and $\hat{P}_{i2}$ and $X = \hat{F}_1 - \hat{F}_2$. A test statistic W for testing H_0'' against H_1'' is

$$W = \frac{\hat{X}^2}{V_1 + V_2}.$$

Asymptotically, W follows a chi-square distribution with 1 degree of freedom. It can be shown that this is the Wald statistic, for example, using the equation (4.5.4) of Amemiya[17]. Similar to the derivation of $\bar{S}$, a multiple-SNP version of W is

$$\bar{W} = \frac{[\sum_k (V_{1k} + V_{2k})^{-1} \hat{X}_k]^2}{\sum_k (V_{1k} + V_{2k})^{-1}}.$$

If the SNPs are in linkage equilibrium with each other, $\bar{W}$ also follows an asymptotically chi-square distribution with 1 degree of freedom.

Numerically, the value of $\hat{F}_1$ and V_1 are highly sensitive to rare count of the heterozygous genotype in either sample. For instance, when $n_{11} = 0$, there is $\hat{\zeta}_1 = \hat{\xi}_1 = 1/2\hat{p}_1\hat{q}_1$. Because of $\hat{P}_{10} + \hat{P}_{12} = 1$, the estimate of F_1 is $\hat{F}_1 = 1$ with $V_1 = 0$ regardless of the counts of the two homozygous genotypes. However, if $n_{11} \neq 0$, for instance, $(n_{10}, n_{11}, n_{12}) = (100, 1, 1)$, direct computation shows that $\hat{F}_1 = 0.6617$ and $V_1 = 0.1010$. They are sensitive to rare homozygous genotypes, too. For instance, when $(n_{10}, n_{11}, n_{12}) = (100, 1, 3)$, direct computation shows that $\hat{F}_1 = 0.8522$ and $V_1 = 0.0212$ which are quite different from their values for $(n_{10}, n_{11}, n_{12}) = (100, 1, 1)$. In summary, statistic W is very sensitive to rare genotypes in either sample, making it less appealing than statistic S. This property of statistic W may explain the tendency found in the simulation study to be reported that W tends to be larger than a random variable that follows a chi-square distribution with 1 degree of freedom. Statistic W takes missing value whenever one sample has no observed heterozygous genotype or there are no observed homozygous genotypes in either sample.

Simulation studies were carried out to assess the performance of the proposed methods. It is assumed that the minor allele frequency is the same in cases as in controls, but the F

parameters are allowed to be different. For specified allele frequency p and F coefficient, the genotype frequencies are computed using relationship presented in (1). Ten thousand simulation replicates are used to assess the type I error rate and power of the proposed statistics.

Table 1 presents the type I error rates in the case of single marker. Simulated rejection rates for S statistic are very close to the respective nominal levels. But W statistic tends to be inflated, maybe due to its sensitivity to rare genotypes discussed previously. Table 2 presents the rejection rates when F_1 and F_2 are not equal. This rejection rate (power) is not high for the situations considered.

[Table 1 about here.]

[Table 2 about here.]

To investigate the performance of the multi-SNP versions of statistics S and W, 50 markers were generated independently. The minor allele frequencies of these markers are taken to be the same, which are further assumed to be the same in cases as in controls. The 50 markers in cases (or controls) share a common parameter F. Table 3 presents the type I error rate. Again, statistic $\bar{W}$ tends to be liberal but statistic $\bar{S}$ remains valid. The power of $\bar{S}$ and $\bar{W}$ in the same settings as in table 3 is very high (> 0.999, data not shown).

[Table 3 about here.]

Application to HapMap genotype data

SNP genotype data were obtained from the official HapMap website (`http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/latest_ncbi_build35/rs_strand/non-redundant/`).

These genotypes are on 90 U.S. residents with northern and western European ancestry from 30 CEPH family trios (CEU), 90 individuals from 30 family trios from the Yoruba people of Ibadan, Nigeria (YRI), 45 unrelated Han Chinese from Beijing China (CHB), and 45 unrelated Japanese from Tokyo, Japan (JPT). To avoid known relative relationship, only the parents of the trios are kept for further analysis. So the sample size is 60 for CEU, 60 for YRT, 45 for CHB, and 45 for JPT. The number of SNPs genotyped for these samples are not exactly the same. There are close to 3.8 million Single-nucleotide polymorphisms (SNPs) for the 22 autosomes. Due to its hyper-sensivitity to rare genotype counts, results for statistic W are not reported since the sample size of this HapMap genotype data is not large, nor do results for its multi-marker version $\bar{W}$.

The following comparisons are made: CEU versus CHB+JPT, CEU versus YRI, YRI versus CHB+JPT, and CHB versus JPT. Table 4 presents the averages of statistic S for each chromosome along with its variance. These averages surprisingly do not vary much across chromosomes, so do the variances for each chromosome. It is very interesting to see that the averages of S for the JPT versus CHB comparison are very close to 1 (the mean of the chi-square disribution with 1 degree of freedom) on each chromosome and the variances on each chromosome is even less than 2, the variance of the chi-square distribution with 1 degree of freedom. This observation suggests that there is no difference detected between JPT population and CHB population. CHB+JPT may represent eastern Asian population. Results in this table also show that the difference between CHB+JPT and YRI is the largest, followed by the difference between CEU and YRI, and then by the difference between CRU and CHB+JPT. To view the distribution of statistic S on each chromosome for each comparison, Q-Q plots are provided (figures 1 – 4).

[Table 4 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

Statistic $\bar{S}$, the multiple-marker version of S, is also computed. To make sure all markers involved are mutually independent, 22 markers that all three genotypes are present in the pool of two samples being compared (so statistic S does not return a missing value) are randomly selected from the 22 autosomes, one from each chromosome. Statistic $\bar{S}$ is then computed using these 22 markers. Being from different chromosomes, these 22 markers are guaranteed to be mutually independent. This process is repeated 10,000 times for each comparison. The QQ plots for all four comparisons are presented in figure 5. The pattern of population structure difference remains the same as revealed by the single marker statistic S for all the comparisons. Surprisingly, the QQ plot for CHB versus JPT conforms better to the 45-degree line. Since the sample size is not large, there are many rare genotypes that cause the statistic W (also $\bar{W}$) to be large. For instance, for the comparison CHB versus JPT, the mean of W on chromosome 1 is 6.30 with an variance of 30011.48. In comparison, the mean of statistic S on chromosome 1 is 1.00 and its variance is 1.83. As stated before, results from statistic W and $\bar{W}$ are thus not presented.

[Figure 5 about here.]

Discussion

Statistical tests are proposed to test whether there is a difference in population structure between two samples. Results from such tests would be of many important uses.

Fro instance, it would be useful for genetic association studies. If a difference exists, there would be a need to use statistical methods that account for population stratification. For the ease of exposition, these methods are presented in terms of a case sample and a control sample. However, they can be used to any two samples. Application to the HapMap genotype data reveals successfully the difference in population structure between various populations.

Difference in population structure between two samples is reflected in their different genotype distributions. It can be shown that $T = n_{10}/n_{+0} + n_{12}/n_{+2} - 2n_{11}/n_{+1}$ is asymptotically independent of the allele frequency difference in two samples under the assumption that the population genotypes are the same in two samples. The allele frequency difference between cases and controls is often used to construct test of association in genetic association studies. For instance, the popular Armitage test for trend normalizes the square of this difference with a variance estimate. This observation suggests that information used in the statistic S presented here may be different from that used for association studies.

Two types of statistics are proposed. One assumes equal allele frequency between two samples and the other does not need such an assumption. In the context of an association study, it may be natural to carry out these tests at “null” markers, i.e., markers that are known to be not associated with case-control status. The statistic W allows the allele frequency in two samples to be different. It is highly sensitive to rare genotype counts making chi-square approximation to its distribution invalid. Studying its distribution in such situation could be an interesting research problem.

The proposed statistics have been implemented in R. The code is available from the author upon request.

REFERENCES

1. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nature Genetics 37, 868–872.

2. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., He, Y., Yang, Y., Wang, Y., An, Y., Fu, W., Wang, J., Tan, J., Qian, J., Chen, X., Zhang, X., Sun, Y., Zhang, X., Wu, B., and Jin, L. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet 85(6), 762–774.

3. Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X., Zhang, X., and Liu, J. (2009). Genetic structure of the Han Chinese population revealed by genome-wide snp variation. Am J Hum Genet 85(6), 775 – 785.

4. Astle, W. and Balding D.J. (2010). Population structure and cryptic relatedness in genetic association studies. Statistical Science, in press.

5. Horvath, S., and Laird, N.M. (1998). A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63, 1886–1897.

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

7. Bacanu, S., Devlin, B., and Roeder, K. (2000). The power of genomic control. Am J Hum Genet 66, 1933–1944.

8. Pritchard, J.K. and Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratication in association studies. Am J Hum Genet 65, 220–228.

9. Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P. (2000). Association mapping in structured populations. Am J Hum Genet 67, 170–181.

10. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38, 904–909.

11. Kimmel, G., Jordan, M.I., Halperin, E., Shamir, R., and Karp, R.M. (2007). A randomization test for controlling population stratification in whole-genome association studies. Am J Hum Genet 81, 895–905.

12. Zhu, X., Li, S., Cooper, R.S., and Elston, R.C. (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 82, 352–365.

13. Wang, K. (2009). Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 33, 637-645.

14. Foulkes, A.S. (2009). Applied Statistical Genetics with R for Population-Based Association Studies. Springer-Verlag.

15. Halliburton, R. (2004). Introduction to Population Genetics. Prentice-Hall (Pearson Education).

16. Weir, B.S. and Hill, W.G. (2002). Estimating F-statistics. Annu Rev Genet 36, 721–50.

17. Amemiya, T. (1985). Advanced Econometrics. Harvard Press.

This manuscript was prepared with the AAS LaTeX macros v5.2.

Appendix Derivation of the score statistic

We found that it is easier to work on genotype frequencies than parameter F and allele frequency p. In hypotheses H_0' and H_1', the frequency of allele A is held constant, which holds if and only if $P_{12} - P_{10} = P_{22} - P_{20}$. So the log-likelihood function can be re-parameterized through P_{10}, P_{20} and δ: $l(P_{10}, P_{20}, \delta) = \sum_{i=1,2}[n_{i2}\log(\delta + P_{i0}) + n_{i1}\log(1 - \delta - 2P_{i0}) + n_{i0}\log P_{i0}]$. The hypotheses H_0' and H_1' can be re-formulated in the following equivalent forms: $H_0' : P_{10} = P_{20}, \delta$ and $H_1' : P_{10} \neq P_{20}, \delta$.

The vector of first-order derivatives of the log-likelihood function is

$$\begin{pmatrix} \partial l/\partial P_{10} \\ \partial l/\partial P_{20} \\ \partial l/\partial \delta \end{pmatrix} = \begin{pmatrix} n_{12}/(\delta + P_{10}) - 2n_{11}/(1 - \delta - 2P_{10}) + n_{10}/P_{10} \\ n_{22}/(\delta + P_{20}) - 2n_{21}/(1 - \delta - 2P_{20}) + n_{20}/P_{20} \\ \sum_{i=1,2}[n_{i2}/(\delta + P_{i0}) - n_{i1}/(1 - \delta - 2P_{i0})] \end{pmatrix}.$$

The second order derivatives are

$$\begin{aligned}
\frac{\partial^2 l}{\partial P_{10}^2} &= -\frac{n_{12}}{(\delta + P_{10})^2} - \frac{4n_{11}}{(1 - \delta - 2P_{10})^2} - \frac{n_{10}}{P_{10}^2}, \\
\frac{\partial^2 l}{\partial P_{10}\partial P_{20}} &= 0, \\
\frac{\partial^2 l}{\partial P_{10}\partial \delta} &= -\frac{n_{12}}{(\delta + P_{10})^2} - \frac{2n_{11}}{(1 - \delta - 2P_{10})^2}, \\
\frac{\partial^2 l}{\partial P_{20}^2} &= -\frac{n_{22}}{(\delta + P_{20})^2} - \frac{4n_{21}}{(1 - \delta - 2P_{20})^2} - \frac{n_{20}}{P_{20}^2}, \\
\frac{\partial^2 l}{\partial P_{20}\partial \delta} &= -\frac{n_{22}}{(\delta + P_{20})^2} - \frac{2n_{21}}{(1 - \delta - 2P_{20})^2}, \\
\frac{\partial^2 l}{\partial \delta^2} &= \sum_{i=1,2}\left[-\frac{n_{i2}}{(\delta + P_{i0})^2} - \frac{n_{i1}}{(1 - \delta - 2P_{i0})^2}\right].
\end{aligned}$$

So the expectation of the negative of the matrix of the second-order derivatives is

$$\begin{pmatrix} n_{1+}\lambda & 0 & n_{1+}\mu \\ 0 & n_{2+}\lambda & n_{2+}\mu \\ n_{1+}\mu & n_{2+}\mu & n_{1+}\nu_1 + n_{2+}\nu_2 \end{pmatrix},$$

where

$$\begin{aligned}
\lambda &= \frac{1}{\delta + P_{10}} + \frac{4}{1 - \delta - 2P_{10}} + \frac{1}{P_{10}}, \\
\mu &= \frac{1}{\delta + P_{10}} + \frac{2}{1 - \delta - 2P_{10}}, \\
\nu_i &= \frac{1}{\delta + P_{i0}} + \frac{1}{1 - \delta - 2P_{i0}}.
\end{aligned}$$

Under H_0, it is easy to see the maximum likelihood estimates of P_{10} and P_{20} are given by $\hat{P}_{10} = \hat{P}_{20} = n_{+0}/n_{++}$ and the maximum likelihood estimate of δ satisfy $n_{+1}/n_{++} = 1 - \hat{\delta} - 2\hat{P}_{10}$. Define $\hat{P}_{11} = \hat{P}_{21} = n_{+1}/n_{++}$ and $\hat{P}_{12} = \hat{P}_{22} = n_{+2}/n_{++}$. The vector of first-order derivatives becomes $Tn_{++}(1, -1, 0)^t$ where $T = n_{12}/n_{+2} - 2n_{11}/n_{+1} + n_{10}/n_{+0}$. The expectation of the negative of the matrix of second-order derivatives becomes $n_{++}^2 A$ where

$$A = \begin{pmatrix} \frac{bn_{1+}}{n_{++}} & 0 & \frac{n_{1+}}{n_{++}}(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}) \\ 0 & \frac{bn_{2+}}{n_{++}} & \frac{n_{2+}}{n_{++}}(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}) \\ \frac{n_{1+}}{n_{++}}(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}) & \frac{n_{2+}}{n_{++}}(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}) & \frac{1}{n_{+1}} + \frac{1}{n_{+2}} \end{pmatrix},$$

in which $b = 1/n_{+2} + 4/n_{+1} + 1/n_{+0}$. Let

$$\begin{aligned}
A^{11} &= \frac{bn_{2+}}{n_{++}}\left(\frac{1}{n_{+2}} + \frac{1}{n_{+1}}\right) - \left[\frac{n_{2+}}{n_{++}}\left(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}\right)\right]^2, \\
A^{12} &= \frac{n_{1+}n_{2+}}{n_{++}^2}\left(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}\right)^2, \\
A^{21} &= A^{12}, \\
A^{22} &= \frac{bn_{1+}}{n_{++}}\left(\frac{1}{n_{+2}} + \frac{1}{n_{+1}}\right) - \left[\frac{n_{1+}}{n_{++}}\left(\frac{1}{n_{+2}} + \frac{2}{n_{+1}}\right)\right]^2.
\end{aligned}$$

According to equation (4.5.5) of Amemiya[17], the score statistic is

$$
\begin{aligned}
S &= T^2(1,-1,0)A^{-1}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \\
&= \frac{T^2}{|A|}(1,-1)\begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\
&= \frac{T^2}{|A|}\left[b\left(\frac{1}{n_{+2}}+\frac{1}{n_{+1}}\right)-\left(\frac{2}{n_{+1}}+\frac{1}{n_{+2}}\right)^2\right] \\
&= \frac{T^2 n_{++}}{|A|\cdot n_{+0}n_{+1}n_{+2}}.
\end{aligned}
$$

The determinant of A equals

$$
\begin{aligned}
|A| &= \frac{b^2 n_{1+}n_{2+}}{n_{++}^2}\left(\frac{1}{n_{+2}}+\frac{1}{n_{+1}}\right)-\frac{bn_{1+}^2n_{2+}}{n_{++}^3}\left(\frac{1}{n_{+2}}+\frac{2}{n_{+1}}\right)^2-\frac{bn_{1+}n_{2+}^2}{n_{++}^3}\left(\frac{1}{n_{+2}}+\frac{2}{n_{+1}}\right)^2 \\
&= \frac{bn_{1+}n_{2+}}{n_{++}^2}\left[b\left(\frac{1}{n_{+2}}+\frac{1}{n_{+1}}\right)-\left(\frac{1}{n_{+2}}+\frac{2}{n_{+1}}\right)^2\right] \\
&= \frac{bn_{1+}n_{2+}}{n_{++}^2}\cdot\frac{n_{++}}{n_{+0}n_{+1}n_{+2}}.
\end{aligned}
$$

Hence the score statistic S is

$$
\begin{aligned}
S &= \frac{n_{++}^2}{n_{1+}n_{2+}}\cdot\frac{T^2}{b} \\
&= \frac{n_{++}^2}{n_{1+}n_{2+}}\cdot\frac{(n_{12}/n_{+2}-2n_{11}/n_{+1}+n_{10}/n_{+0})^2}{1/n_{+2}+4/n_{+1}+1/n_{+0}}.
\end{aligned}
$$

Fig. 1.— Chromosome-wise Q-Q plot of statistic S against the 1-df chi-square distribution for 22 autosomes using HapMap genotype data: CHB versus JPT

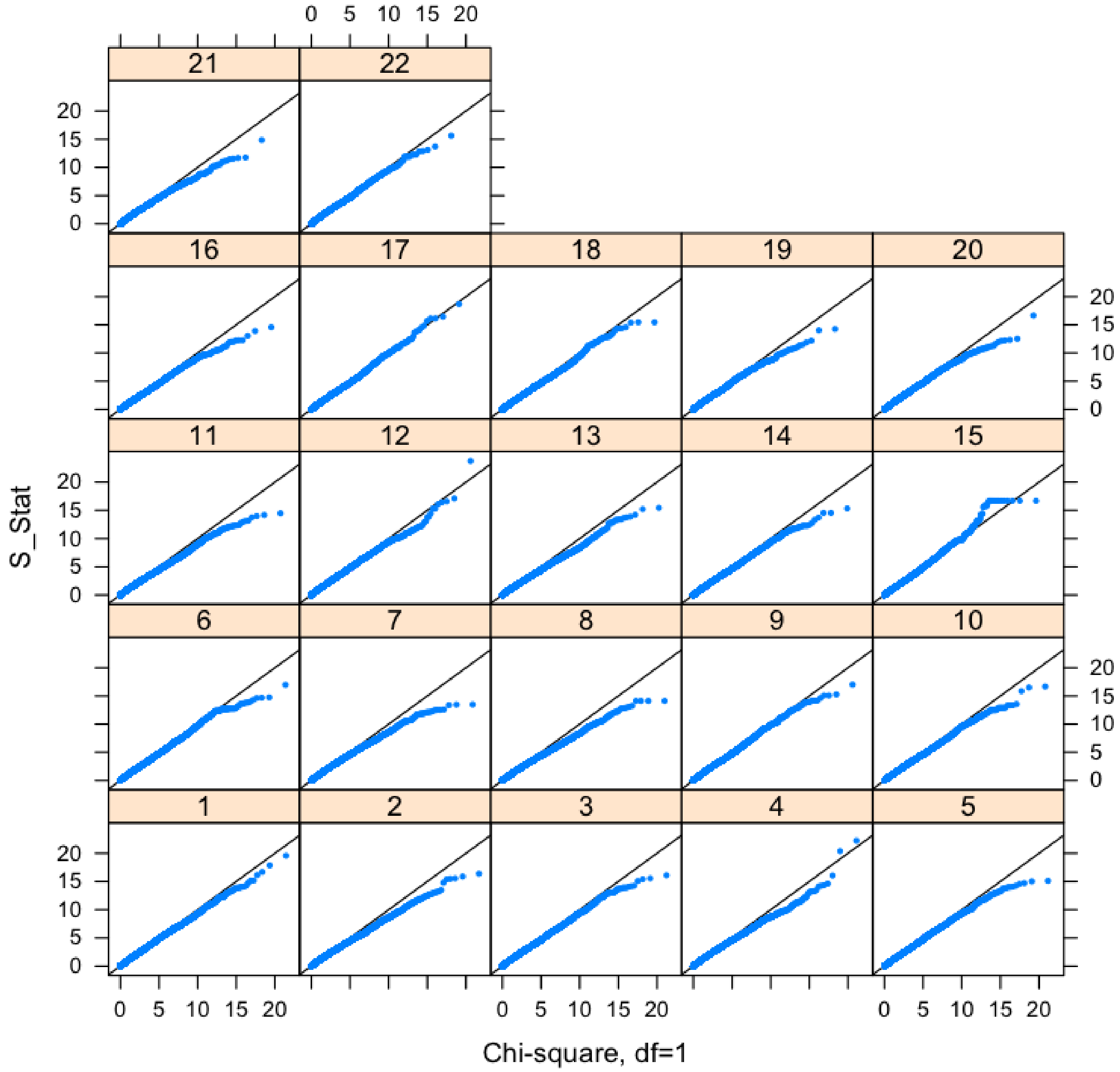

Fig. 2.— Chromosome-wise Q-Q plot of statistic S against the 1-df chi-square distribution for 22 autosomes using HapMap genotype data: CEU versus CHB+JPT

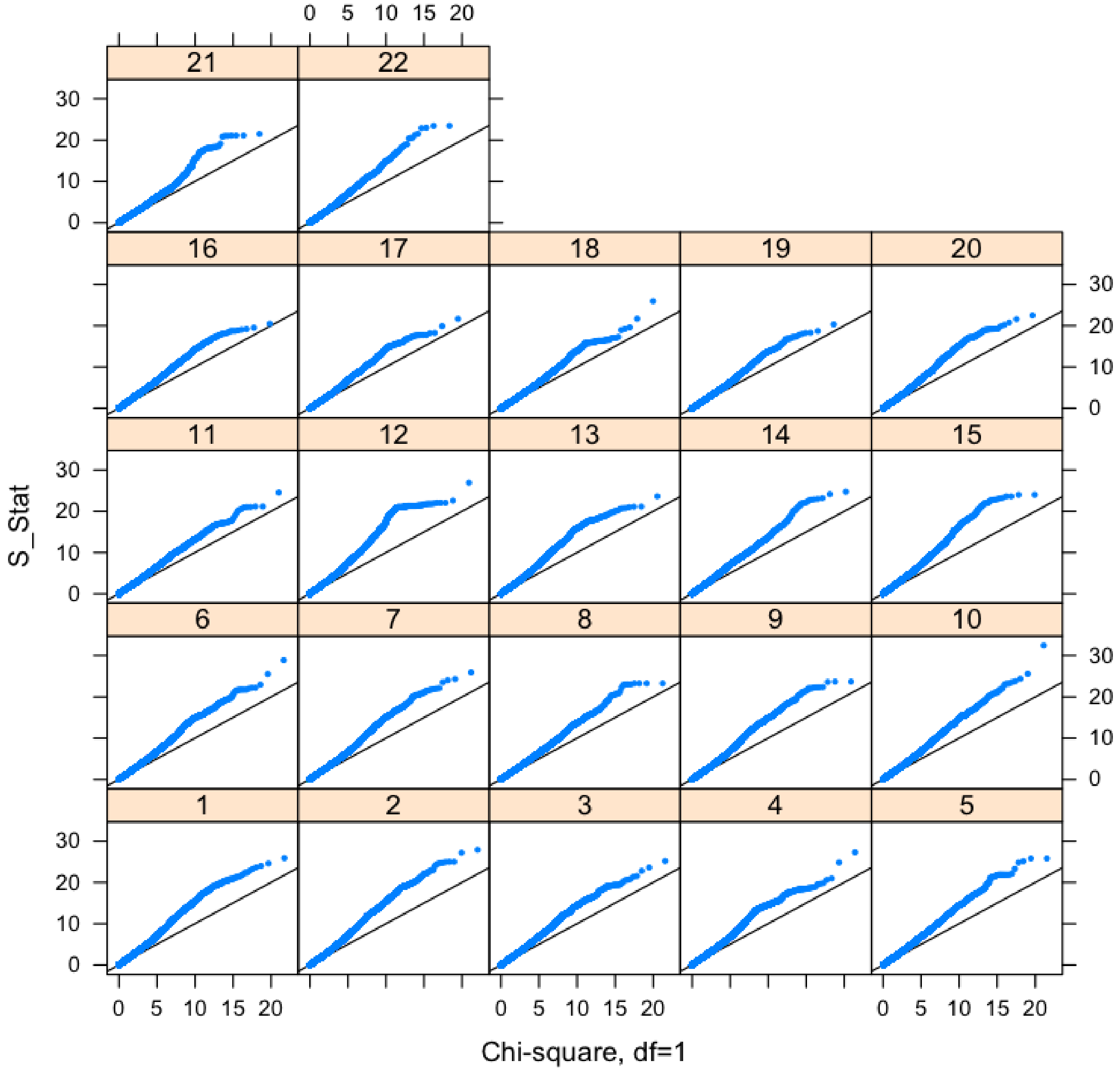

Fig. 3.— Chromosome-wise Q-Q plot of statistic S against the 1-df chi-square distribution for 22 autosomes using HapMap genotype data: CEU versus YRI

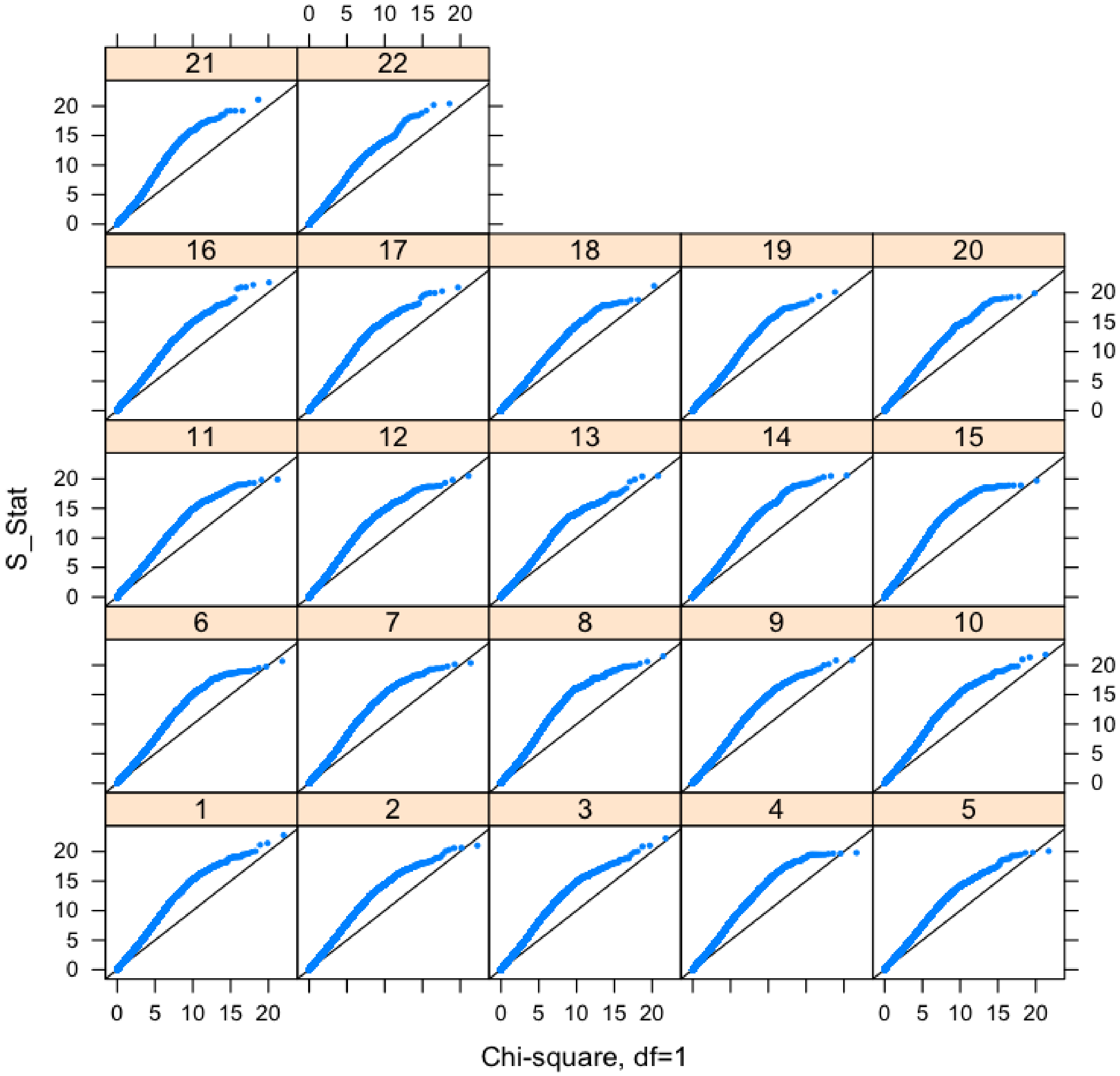

Fig. 4.— Chromosome-wise Q-Q plot of statistic S against the 1-df chi-square distribution for 22 autosomes using HapMap genotype data: YRI versus CHB+JPT

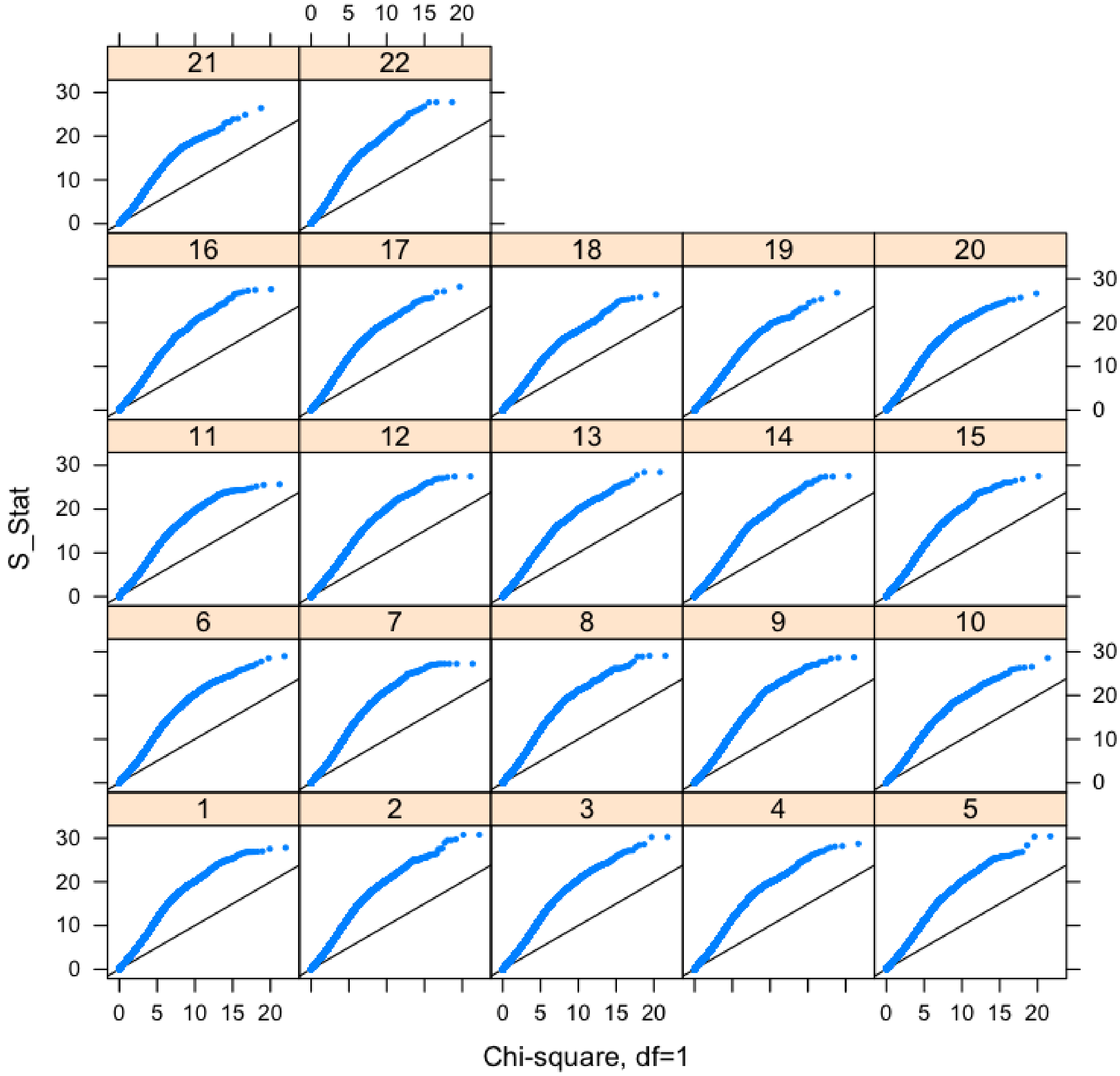

Fig. 5.— Q-Q plot of the multi-marker statistic $\bar{S}$ against the 1-df chi-square distribution. Statistic $\bar{S}$ is computed 10,000 times. Each time it is computed using 22 randomly selected markers, one from each autosome.

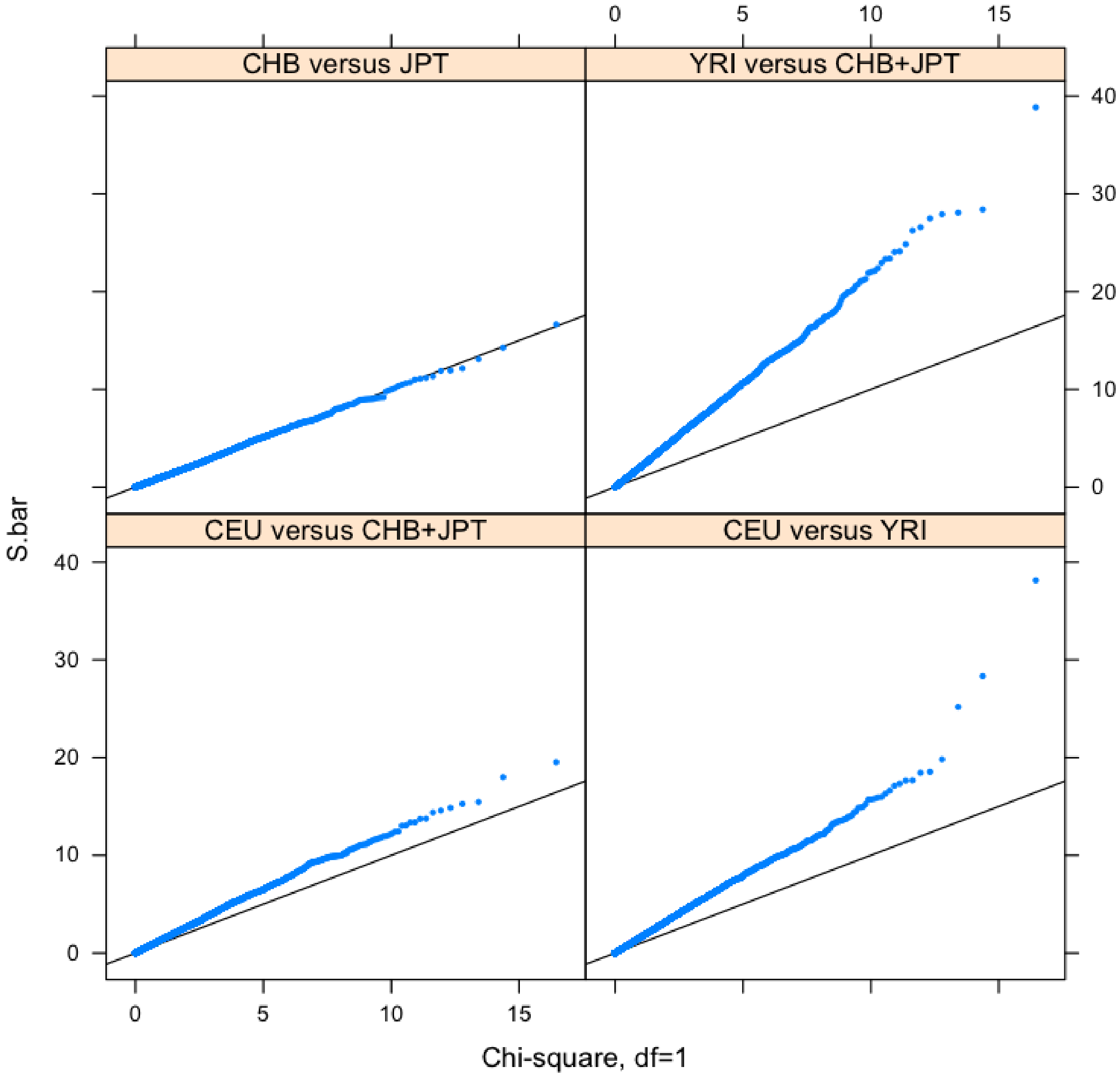

Table 1: Simulated type I error rate for statistics S and W. Computed from 10,000 simulation replicates.

			Statistic S			Statistic W		
			Significance Level			Significance Level		
n	p	$F_1(=F_2)$	0.01	0.005	0.001	0.01	0.005	0.001
300	0.1	0.1	0.0108	0.0043	0.0009	0.0164	0.0099	0.0035
		0.2	0.0103	0.0051	0.0007	0.0148	0.0086	0.0032
		0.3	0.0102	0.0047	0.0009	0.0156	0.0105	0.0033
	0.2	0.1	0.0098	0.0040	0.0007	0.0139	0.0083	0.0015
		0.2	0.0094	0.0047	0.0008	0.0123	0.0058	0.0013
		0.3	0.0103	0.0052	0.0010	0.0114	0.0064	0.0018
500	0.1	0.1	0.0081	0.0034	0.0010	0.0130	0.0077	0.0018
		0.2	0.0101	0.0044	0.0009	0.0104	0.0060	0.0016
		0.3	0.0098	0.0045	0.0008	0.0122	0.0060	0.0013
	0.2	0.1	0.0088	0.0043	0.0009	0.0120	0.0062	0.0013
		0.2	0.0115	0.0061	0.0015	0.0107	0.0063	0.0014
		0.3	0.0094	0.0051	0.0011	0.0125	0.0067	0.0019

Table 2: Simulated power for statistics S and W computed from 10,000 simulation replicates.

			Statistic S			Statistic W		
			Significance Level			Significance Level		
n	p	$F_1(=F_2)$	0.01	0.005	0.001	0.01	0.005	0.001
300	0.1	(0.1, 0.2)	0.0430	0.0262	0.0085	0.0654	0.0447	0.0181
		(0.1, 0.3)	0.2040	0.1403	0.0535	0.2440	0.1839	0.0904
		(0.2, 0.3)	0.0395	0.0225	0.0056	0.0539	0.0348	0.0143
	0.2	(0.1, 0.2)	0.0699	0.0448	0.0142	0.0750	0.0471	0.0171
		(0.1, 0.3)	0.3458	0.2650	0.1313	0.3632	0.2849	0.1535
		(0.2, 0.3)	0.0683	0.0428	0.0114	0.0713	0.0443	0.0151
500	0.1	(0.1, 0.2)	0.0848	0.0500	0.0144	0.0919	0.0608	0.0228
		(0.1, 0.3)	0.3883	0.2976	0.1414	0.4072	0.3266	0.1849
		(0.2, 0.3)	0.0649	0.0413	0.0134	0.0753	0.0488	0.0178
	0.2	(0.1, 0.2)	0.1223	0.0826	0.0291	0.1300	0.0911	0.0360
		(0.1, 0.3)	0.6024	0.5097	0.3201	0.6068	0.5157	0.3359
		(0.2, 0.3)	0.1190	0.0792	0.0302	0.1240	0.0837	0.0326

Table 3: Simulated type I error rate with 50 independent markers computed from 10,000 simulation replicates.

			Statistic S			Statistic W		
			Significance Level			Significance Level		
n	p	$F_1(=F_2)$	0.01	0.005	0.001	0.01	0.005	0.001
300	0.1	0.2	0.0089	0.0042	0.0006	0.0154	0.0091	0.0017
		0.3	0.0091	0.0048	0.0015	0.0156	0.0082	0.0018
	0.2	0.1	0.0088	0.0041	0.0008	0.0119	0.0064	0.0010
		0.2	0.0107	0.0056	0.0013	0.0122	0.0066	0.0015
		0.3	0.0106	0.0061	0.0009	0.0131	0.0066	0.0015
500	0.1	0.1	0.0102	0.0053	0.0014	0.0139	0.0077	0.0022
		0.2	0.0097	0.0057	0.0004	0.0128	0.0065	0.0014
		0.3	0.0095	0.0046	0.0010	0.0164	0.0096	0.0015
	0.2	0.1	0.0102	0.0052	0.0012	0.0094	0.0056	0.0010
		0.2	0.0102	0.0047	0.0011	0.0100	0.0050	0.0006
		0.3	0.0112	0.0057	0.0008	0.0125	0.0063	0.0011

Table 4: Mean and variance of statistic S on each chromosome for the population comparisons using Hapmap genotype data

	Statistic S (mean and variance)			
Chromosome	CEU vs. CHB+JPT	CEU vs. YRI	YRI vs. CHB+JPT	JPT vs. CHB
1	1.37 (4.10)	1.56 (4.72)	2.16 (9.50)	1.00 (1.83)
2	1.36 (4.14)	1.55 (4.60)	2.21 (10.08)	0.97 (1.59)
3	1.35 (3.75)	1.58 (4.77)	2.15 (9.63)	0.99 (1.77)
4	1.34 (3.80)	1.55 (4.71)	2.15 (9.43)	0.98 (1.67)
5	1.34 (3.72)	1.48 (4.19)	2.07 (8.95)	0.99 (1.78)
6	1.34 (3.71)	1.49 (4.37)	2.11 (9.19)	1.00 (1.83)
7	1.33 (3.71)	1.55 (4.80)	2.13 (9.99)	0.98 (1.66)
8	1.27 (3.34)	1.59 (5.18)	2.23 (10.50)	0.95 (1.54)
9	1.33 (3.90)	1.49 (4.50)	2.09 (9.62)	1.03 (1.88)
10	1.35 (3.80)	1.55 (4.74)	2.11 (9.12)	0.99 (1.74)
11	1.28 (3.34)	1.52 (4.40)	2.06 (8.82)	0.96 (1.62)
12	1.42 (4.55)	1.59 (4.97)	2.11 (8.92)	1.01 (1.83)
13	1.38 (4.18)	1.46 (4.13)	2.12 (9.25)	0.95 (1.57)
14	1.37 (3.76)	1.54 (4.70)	2.07 (8.84)	0.99 (1.76)
15	1.40 (4.22)	1.72 (5.81)	2.17 (9.62)	1.00 (1.95)
16	1.35 (3.63)	1.59 (4.74)	2.16 (9.51)	0.96 (1.68)
17	1.27 (3.50)	1.68 (5.38)	2.26 (10.40)	1.01 (1.87)
18	1.23 (3.19)	1.51 (4.23)	2.06 (8.35)	0.98 (1.77)
19	1.24 (3.22)	1.50 (4.40)	2.04 (8.58)	0.99 (1.76)
20	1.35 (3.93)	1.53 (4.57)	2.22 (10.10)	1.00 (1.78)
21	1.23 (3.19)	1.56 (5.17)	2.07 (9.16)	0.99 (1.67)
22	1.34 (3.88)	1.53 (4.68)	2.35 (11.32)	1.03 (1.87)