Published in final edited form as:

Ann Hum Genet. 2012 May ; 76(3): 237–245. doi:10.1111/j.1469-1809.2012.00708.x.

Effect of Population Stratification On False Positive Rates Of Population-based Association Analyses Of Quantitative Traits

Tanushree Haldar and Saurabh Ghosh

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Summary

It is now well established that population stratification can result in spurious association findings in genetic case-control studies. However, very few studies have addressed similar issues for mapping quantitative traits. Since quantitative phenotypes are often precursors of clinical end-point traits and carry more information on within-genotype trait variability, it has been argued that studying these quantitative traits may be a more powerful strategy to map genes than the binary clinical end-points. Thus, it is of interest to evaluate the adverse effects of population stratification on the analyses of quantitative traits. The popular statistical tests of association for quantitative traits using population level data are ANOVA, linear regression with an additive allelic effect and Kruskal-Wallis. We have theoretically studied the marginal effects of genetic heterogeneity and phenotypic heterogeneity as well as their joint effects on the false positive rate of the three tests mentioned above. We have carried out extensive simulations under different genetic models and probability distributions of quantitative traits to assess the rate of false positives in the presence of population stratification. We find that the rate of false positives increases at a very fast rate with simultaneous increase in differences in the standardized phenotypic means and marker allele frequencies in the subpopulations.

Keywords

allelic additivity; ANOVA; genetic heterogeneity; Kruskal-Wallis test; phenotypic heterogeneity

Introduction

Case-control studies are the most common design used for genetic association mapping of complex disorders. They are more popular than family-based association studies such as the Transmission Disequilibrium Test [TDT] (Spielman et al. 1993) not only because of the comparative ease in data collection, but also because it has been shown that they are statistically more powerful to detect association (Haldar and Ghosh 2011, McGinnis et al. 2002). However, one of the major limitations of the case-control design is its susceptibility to population stratification. Association analyses based on a combined sample of genetically heterogeneous subpopulations (i.e., having different allele frequencies at a locus) can result in inflated false positive rates. This is not only of theoretical importance, but also is a major problem in practice. Empirical studies have shown that allele frequencies at many loci vary substantially across populations (Perez et al. 1997). The problem of population stratification is of special relevance in the Indian context as multiple studies have revealed the genetic diversity of Indian subpopulations (Basu et al. 2003, Thangaraj et al. 2005, Indian Genome Variation Initiative 2008). While the problem has been addressed and solutions explored in

the context of binary traits (Pritchard et al. 2000, Devlin et al. 2001, Price et al. 2006), the issue is of current research interest in the context of quantitative traits. Clinical end-point traits are often governed by quantitative precursors which carry more information on inter-individual trait variability compared to the end-point traits. Thus, it has been argued that analyzing these quantitative precursors may be statistically a more prudent strategy to detect association. In this article, we study the possible effects of population stratification on association analyses of quantitative traits.

A relook at Stratification Effects in the Context of Binary Traits

Consider an autosomal biallelic disease locus with alleles D_1 and D_2 . Suppose there are k subpopulations and let us consider a biallelic marker locus with alleles M_1 and M_2 which is in perfect linkage equilibrium with the disease locus in each of these subpopulations. Let m_i and d_i respectively be the frequency of M_1 and D_1 in the i^{th} subpopulation, $i = 1, 2, \dots, k$. Then the difference in the frequency of the allele M_1 between cases and controls can be expressed as

$$\Delta = P(M|cases) - P(M|controls) = \frac{\sum_{i=1}^k \alpha_i d_i m_i}{\sum_{i=1}^k \alpha_i d_i} - \frac{\sum_{i=1}^k \alpha_i (1 - d_i) m_i}{\sum_{i=1}^k \alpha_i (1 - d_i)} \quad (1)$$

where, α_i is the proportion of individuals from the i^{th} subpopulation, $i = 1, 2, \dots, k$. The details are provided in the Appendix. The null hypothesis of a case-control association test, that there is no linkage disequilibrium between alleles at the disease locus and the marker locus is equivalent to the hypothesis that there is no difference in the allele frequencies at the marker locus between cases and controls (i.e., $\Delta = 0$). However, it is clear from the Equation (1) that even if there is no allelic association between the two loci in any of the subpopulations, Δ will be zero if and only if at least one of the following two conditions is satisfied.

- The marker allele frequency is same in all the subpopulations, i.e., there is no genetic heterogeneity across the subpopulations.
- The disease prevalence is same in all the subpopulations, i.e., there is no phenotypic heterogeneity across the subpopulations.

Thus, the case control design will suffer from the problem of inflated false positive rates only in the presence of both genetic heterogeneity and differential disease prevalence among the subpopulations.

We note here that from logistic considerations, the numbers of cases and controls in a case-control design are usually fixed a priori and are not random variables in the analyses. Unless the proportional representation of each sub-population is equal in cases and controls, the condition on prevalence is not satisfied in practice and genetic heterogeneity is sufficient to inflate the false positive rate. Thus, the focus of the problem of population stratification has traditionally been only on genetic heterogeneity.

Stratification Effects in the Context of Quantitative Traits

In a quantitative trait design where individuals are sampled randomly from the population, the marker genotypes as well as the quantitative trait values are random variables. Hence it is of interest to explore theoretically whether either genetic heterogeneity or phenotypic heterogeneity (differences in distributions of quantitative traits across subpopulations) is sufficient to yield an inflated rate of false positives in existing association methodologies of quantitative traits or are both necessary?

Among the most popular methods used to test for association between a marker locus and a quantitative trait locus (QTL) are the Analysis of Variance (ANOVA), the additive allelic effect regression model and the Kruskal Wallis test. It has been established that these statistical tests are equivalent to testing for linkage disequilibrium between the two loci (Ghosh and De 2007). ANOVA and the linear regression analysis are based on model assumptions such as normality of the QT values and homoskedasticity across genotypic groups at a locus. Violations of these assumptions may lead to incorrect inferences on association, especially, high false positive rates (Ghosh and De 2007). On the other hand, the Kruskal Wallis test is model-free and hence is more robust compared to ANOVA and the regression analysis. However, if the underlying assumptions are satisfied, ANOVA and the linear regression analysis are expected to be more powerful than the Kruskal-Wallis test.

We shall study the effects of population stratification on these three tests. We consider a quantitative trait Y controlled by an autosomal biallelic locus with alleles A_1 and A_2 . Suppose there are k subpopulations $S_1, S_2, \dots, S_k$. We also consider a biallelic marker locus with alleles M_1 and M_2 such that there is no linkage disequilibrium between the marker locus and the QTL in any of these subpopulations.

Effect on ANOVA

The ANOVA tests for the equality of means of the quantitative trait across the three genotypic groups at a marker locus. In the above formulation, let us assume that μ_i is the mean of Y given any marker genotype in the i^{th} subpopulation and the proportion of individuals from the i^{th} subpopulation is α_i , $i = 1, 2, \dots, k$. Then, the mean of Y given a marker genotype g in the combined population is given by:

$$\frac{\sum_{i=1}^k \mu_i P(g|S_i) \alpha_i}{\sum_{i=1}^k P(g|S_i) \alpha_i} \quad (2)$$

The details are provided in the Appendix. Thus, even if there is no linkage disequilibrium between the marker locus and the QTL, the conditional mean of Y given any marker genotype will be the same if and only if at least one of the following two conditions is satisfied.

- The marker genotype frequency and hence the allele frequency of M_1 is same in all the subpopulations, i.e., there is no genetic heterogeneity at the marker locus across the subpopulations.
- The mean value of the quantitative trait is same in all the subpopulations, i.e., there is no phenotypic heterogeneity in the quantitative trait across the subpopulations.

Hence, ANOVA will produce an inflated rate of false positives only if both the above conditions are violated.

Additive Allelic Effect Regression Model

An alternative to ANOVA for detecting association between a marker locus and a quantitative trait is a linear regression model with an additive allelic effect. The advantage of the regression framework is its flexibility to incorporate possible confounding covariates as explanatory variables. However, the assumption of an additive allelic effect constrains the model to be more restrictive compared to ANOVA. Suppose that the regression model is $E(Y|x) = \beta_0 + \beta_1 x$ where, $x = 2, 1$ or 0 according as the marker genotype (g) of an individual is M_1M_1 , M_1M_2 or M_2M_2 . The test for association is equivalent to the test for $\beta_1 = 0$ versus $\beta_1 \neq 0$. We have $\beta_1 = Cov(Y, x) / Var(x)$.

Now, $Cov(Y, x) = E\{Cov(Y, x|S)\} + Cov\{E(Y|S), E(x|S)\}$. Since there is no association between the marker locus and the quantitative trait in any of the subpopulations, $Cov(Y, x|S) = 0 \forall S$. Thus,

$$\begin{aligned} Cov(Y, x) &= Cov\{E(Y|S), E(x|S)\} \\ &= \sum_{i=1}^n 2m_i \mu_i \alpha_i - \left\{ \sum_{i=1}^n 2m_i \alpha_i \right\} \left\{ \sum_{i=1}^n \mu_i \alpha_i \right\} \\ &= 2 \left\{ \sum_{i=1}^n m_i \mu_i \alpha_i - \left(\sum_{i=1}^n m_i \alpha_i \right) \left(\sum_{i=1}^n \mu_i \alpha_i \right) \right\} \end{aligned} \quad (3)$$

Thus, even if there is no linkage disequilibrium between the marker locus and the QTL, β_1 will be zero if and only if at least one of the following two conditions is satisfied.

- The marker allele frequency of M_1 is same in all the subpopulations, i.e., there is no genetic heterogeneity at the marker locus across the subpopulations.
- The mean value of the quantitative trait is same in all the subpopulations, i.e., there is no phenotypic heterogeneity in the quantitative trait across the subpopulations.

Kruskal-Wallis Test

The Kruskal-Wallis test attempts to differentiate the distribution of the quantitative trait across the different genotypic groups at the marker locus. Under the same formulation of the subpopulations as mentioned above, let us assume F_i be the distribution of the quantitative trait given any marker genotype in the i^{th} subpopulation and the proportion of individuals from the i^{th} subpopulation is α_i , $i = 1, 2, \dots, k$. Then, the distribution of Y given a marker genotype g in the combined population can be expressed as:

$$\frac{\sum_{i=1}^k F_i P(g|S_i) \alpha_i}{\sum_{i=1}^k P(g|S_i) \alpha_i} \quad (4)$$

The details are provided in the Appendix. Thus, in the same lines as ANOVA, it is possible that there is no linkage disequilibrium between the two loci, but the conditional distribution of Y given any marker genotype will be the same if and only if at least one of the two following conditions is satisfied:

- The marker genotype frequency and hence the allele frequency of M_1 is same in all the subpopulations, i.e., there is no genetic heterogeneity at the marker locus across the subpopulations.
- The distribution of the quantitative trait is same across all the subpopulations, i.e., there is no phenotypic heterogeneity in the quantitative trait across the subpopulations.

Thus, the Kruskal Wallis test will yield an inflated rate of false positives if both the above conditions are violated.

It is interesting to observe in this context that if the quantitative trait values have different distributions but the same mean across the subpopulations, the Kruskal-Wallis test may suffer from inflated false positive rates, but ANOVA will not.

A closer look at Stratification Effects for two subpopulations

While we have derived conditions for inflated false positive rates, we wish to restrict our attention to the case of two heterogeneous subpopulations for a better theoretical understanding of effects of stratification on ANOVA.

Under no genetic association in any of these two subpopulations, the ANOVA statistic has the property that:

$$\frac{E(MSG)}{E(MSE)} = 1 + (\mu_1 - \mu_2)^2 \frac{\sum_{i=1}^3 n_i (u_i - \bar{u})^2}{2\sigma^2}$$

where, MSG and MSE are the mean between genotype sum of squares and the mean within genotype sum of squares, respectively; u_i is the proportion of subpopulation 1 and n_i is the frequency in the i^{th} genotype group, $\bar{u}$ is the mean of u_i s, $i = 1, 2, 3$. The details are provided in the Appendix.

Under the null hypothesis of ANOVA, $\frac{E(MSG)}{E(MSE)}$ is equal to 1. Under the alternative, this quantity is greater than 1. We note that the expression given above is equal to 1 if and only if $\mu_1 = \mu_2$ or u_i s are all equal. Thus, deviation of the quantity from 1 will result in an inflated false positive rate. The extent of inflation depends on:

- the difference between the QT means in the two subpopulations ($|\mu_1 - \mu_2|$)
- the variance of the QT within each genotype (σ^2)
- the variation in the proportional representation of each subpopulation across different marker genotypes in the combined population

Simulations

In order to assess the adverse effects of genetic and phenotypic heterogeneity on association results based on ANOVA, the proposed linear regression model and the Kruskal Wallis test, we carried out extensive simulations over a wide spectrum of QTL models. In all our simulations, we generated data on 100 individuals. We assumed that our sample is a mixture of two subpopulations in equal proportions. Since the set-up under consideration is one where there is no association between the QTL and the marker locus, the quantitative trait values are independent of the genotypes at the marker locus and hence, the distribution of the QT values conditioned on the marker genotypes is equivalent to the marginal distribution of the QT values. Thus, it is not necessary to generate genotypes at the QTL. We generate the genotype of an individual at a biallelic marker locus and the corresponding quantitative trait value independently depending on the subpopulation to which the individual belongs. If m_1 and m_2 are, respectively, the frequencies of the marker allele M_1 in subpopulations 1 and 2, we have considered three levels of allelic heterogeneity : ($m_1 = 0.1, m_2 = 0.9$); ($m_1 = 0.3, m_2 = 0.6$) and ($m_1 = 0.2, m_2 = 0.4$). To induce phenotypic heterogeneity between the two subpopulations, we have simulated the QT values from five different probability models: (i) two different normal distributions; (ii) two chi-squares distributions with 4 degrees of freedom, but with different location shifts such that they have different means; (iii) two chi-squares distributions with 1 degree of freedom, but with different location shifts; (iv) a normal distribution in subpopulation 1 and a chi-squares distribution with 4 degrees of freedom in subpopulation 2; and (v) a normal distribution in subpopulation 1 and a chi-squares distribution with 1 degree of freedom in subpopulation 2. The details of the five phenotypic distributions are summarized in Table 1. The different models represent possible shapes of the probability distributions of the quantitative trait. For example, model (i) pertains to a quantitative trait having a symmetric distribution, models (ii) and (iii) pertain to a quantitative trait having skewed distributions (such as fasting glucose levels among obese individuals) with the degree of skewness more in model (iii) while models (iv) and (v) pertain to a quantitative trait having a symmetric distribution in one subpopulation and a skewed distribution in the other.

Variables with skewed probability densities are often subjected to logarithmic transformations with the aim of reducing inherent variability and inducing normality. Hence, we have also evaluated the performance of ANOVA with logarithmic transformed quantitative trait values. The false positive rates have been evaluated empirically as a

function of $d = \frac{|\mu_1 - \mu_2|}{\sigma}$, the standardized absolute difference in the mean QT values in the two subpopulations and have been presented in Figures 1–5 corresponding to the five different phenotypic distributions described above. Each figure comprises a trio of graphs corresponding to the three levels of allelic heterogeneity. In each figure, the continuous curve represents the results of ANOVA, the dashed curve represents the results of the linear regression with an additive allelic effect, the darker dotted curve represents the results of the Kruskal-Wallis test while the lighter dotted curve represents the results of ANOVA based on logarithmic transformed QT values. The empirical false positive rates are computed based on 10000 replications.

Results

We first note that for two genetically homogeneous subpopulations, we found that all the three methods maintain the appropriate rate of false positives, irrespective of the phenotypic distribution and heterogeneity measured in terms of d (data not presented for brevity). However, it is clear from Figures 1–5 that the false positive rate increases at a very fast rate with the increase in the standardized absolute difference between means of the quantitative trait (d) between two genetically heterogeneous subpopulations. Even when the QT values in both the subpopulations are distributed as normal and marker allele frequencies 0.2 and 0.4 in the two subpopulations (the set-up with the least adverse effect on false positive rates), a value of $d=4$ can yield a false positive rate as high as 0.7 for ANOVA or the regression model and 0.6 for the Kruskal-Wallis test. In each figure, we find that the rate of increase with d in the false positive rate is maximum in the first graph and minimum in the third. This is expected as the first graph corresponds to the most genetically heterogeneous set-up while the third graph corresponds to the least genetically heterogeneous set-up. On the other hand, the rate of increase in the false positive rate is much higher for the Kruskal Wallis test compared to ANOVA with increase in the extent of genetic heterogeneity, particularly when the quantitative trait has a skewed distribution. Thus, the effect of population stratification is more adverse when the distribution is chi-squares compared to when it is normal (Figure 1 compared to the remaining Figures) and the most adverse effect is observed when the underlying distribution is chi-squares with 1 d.f. (Figure 3 compared to Figure 2 and Figure 5 compared to Figure 4). While ANOVA has a lower false positive rate than the Kruskal Wallis test when the genetic heterogeneity is high (the first graph in each Figure), it has a higher rate when the heterogeneity is low (the third graph in each Figure). The additive allelic effect regression model has a much higher false positive rate compared to both ANOVA and the Kruskal Wallis test unless the genetic heterogeneity between the subpopulations is very high, in which case, it has a lower rate compared to the Kruskal Wallis test. This phenomenon was observed for all the phenotypic models considered except when the quantitative trait values were distributed as normal in both the subpopulations. We found that ANOVA based on logarithmic transformation of QT values exhibited similar false positive rates compared to that based on the original QT values, except when the QT values was distributed as chi-squares. As expected, the logarithmic transformation reduces the skewness in the chi-squares distribution. However, the transformation lowers the false positive rate when the genetic heterogeneity is low, but increases it when the heterogeneity is high (the third graph compared to the first graph in Figures 2 and 3).

Discussion

In this article, we have explored the conditions that affect the false positive rates of population-based association analyses of quantitative traits. We found that neither genotypic heterogeneity (varying allele frequencies at a locus) nor phenotypic heterogeneity (different QT distributions) across constituent sub-populations in the sample can alone result in inflated rates of false positives. A combination of both the above conditions is necessary to yield inflated false positive rates. These conditions are similar to those for case-control studies on binary traits where both allelic heterogeneity as well as differential disease prevalences are necessary across subpopulations to result in spurious association findings. However, there is a basic difference in the sampling designs of case-control studies and quantitative trait association analyses. Case-control studies are retrospective in nature: the sample sizes of the phenotypic classes are fixed apriori and the genotypic/allelic distributions are studied conditioned on the phenotypic classes. Thus, the condition that each of the subpopulations has the same disease prevalence is not usually satisfied by the study design itself, and hence, the problem of population stratification in the context of case-control studies has been emphasized only with respect to the condition on allelic heterogeneity. On the other hand, ANOVA, the proposed linear regression approach and the Kruskal-Wallis test are based on a prospective study design: the phenotypic distribution is modeled conditioned on the genotype class. In this set-up, both the genotypes as well as quantitative trait values are random in nature. Thus, the conditions on both genetic as well as phenotypic heterogeneities are relevant in practice for understanding the problem of population stratification in the context of quantitative traits.

The Kruskal-Wallis test is non-parametric in nature and hence expected to be more robust with respect to violations in distributional assumptions compared to ANOVA or the linear regression model, which are based on certain model assumptions. However, in the presence of population stratification, we find that while the Kruskal Wallis test is less severely affected by population stratification compared to ANOVA or the linear regression model for low levels of genetic heterogeneity, it is more adversely affected for high levels of heterogeneity especially when the distribution of the QT is skewed. This can be explained by the fact that while the Kruskal-Wallis test tries to differentiate between distributions of underlying groups, ANOVA tries to differentiate between the means of the different groups. Thus, under population stratification, it is possible that the Kruskal-Wallis test rejects the null since the QT distributions are substantially different across genotypes, but ANOVA may not reject the null if the QT means across the genotypes are close to each other. On the other hand, the linear regression model assumes an additive allelic effect and possible violations to this assumption in the presence of genetic as well as phenotypic heterogeneities leads to higher inflation of false positive rates compared to not only ANOVA, but also the Kruskal-Wallis test. Logarithmic transformation of QT values having skewed distributions is able to decrease false positive rates only if the extent of genetic heterogeneity is low. In summary, ANOVA, the additive allelic effect linear regression and the Kruskal-Wallis test all perform poorly in the presence of genetic and phenotypic heterogeneities, but contrary to traditional belief, ANOVA may be a more prudent strategy for analyzing quantitative traits if there is suspicion of possible population stratification in the selected sample.

The focus of the current study was to assess the adverse effect of genetic and phenotypic heterogeneities on the false positive rates yielded by the three most popular approaches of association analyses of quantitative traits. However, an interesting issue in this context is that it is statistically possible for an association finding in each of two genetically heterogeneous subpopulations to yield a false negative in the combined population. This is true for both binary as well as quantitative traits. However, we wish to emphasize that simulations under the assumption of linkage disequilibrium will not directly provide good

estimates of false negative rates as the size of the test is, in general, not maintained in the presence of population stratification. Thus, the evaluation of false negative rates can be performed only after statistical corrections such as Structure (Pritchard et al. 2000), Genomic Controls (Devlin et al. 2001) and Eigenstrat (Price et al. 2006) are implemented to control for the inflated false positive rates in the presence of population stratification. While data on genotypes at single SNPs suffice the evaluation of false positive rates, genome-wide SNP data will be required for the statistical corrections to control for the inflated false positive rates and hence, evaluate the false negative rates. Moreover, the estimates of the false negative rates may depend on the choice of the statistical correction implemented. We are currently carrying out extensive simulations and implementing the different statistical corrections to address the issue of false negatives.

Acknowledgments

This work was supported by the Fogarty International Center, National Institutes of Health through grant R01 TW006604 to Saurabh Ghosh and by the Council of Scientific and Industrial Research (CSIR) fellowship 09/093(0111)/2008-EMR-I to Tanushree Haldar.

References

1. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP. Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* 2003; 13:2277–2290. [PubMed: 14525929]
2. Devlin B, Roder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol.* 2001; 60:155–166. [PubMed: 11855950]
3. Ghosh S, De G. Association analysis of population-based quantitative trait data: An assessment of ANOVA. *Hum Hered.* 2007; 64:82–88. [PubMed: 17483600]
4. Haldar T, Ghosh S. Power comparison between population-based case-control studies and family-based transmission-disequilibrium tests: An empirical study. *Indian J Hum Genet.* 2011; 17:S27–31. [PubMed: 21747584]
5. Indian Genome Variation Consortium. Genetic landscape of the people of India: a canvas for disease gene exploration. *J Genet.* 2008; 87:3–20. [PubMed: 18560169]
6. McGinnis R, Shifman S, Darvasi A. Power and Efficiency of the TDT and Case-Control Design for Association Scans. *Behav Genet.* 2002; 32(2):135–144. [PubMed: 12036111]
7. Perez-Lezaun A, Calafell F, Mateu E, Comas D, Bosch E, Bertranpetit J. Allele frequencies for 20 microsatellites in a worldwide population survey. *Hum Hered.* 1997; 47:189–196. [PubMed: 9239505]
8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association analysis. *Nat Genet.* 2006; 38:904–909. [PubMed: 16862161]
9. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet.* 2000; 67:170–181. [PubMed: 10827107]
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993; 52:506–516. [PubMed: 8447318]
11. Thangaraj K, Sridhar V, Kivisild T, Reddy AG, Chaubey G, Singh VK, Kaur S, Agarawal P, Rai A, Gupta J, Mallick CB, Kumar N, Velavan TP, Suganthan R, Udaykumar D, Kumar R, Mishra R, Khan A, Annapurna C, Singh L. Different population histories of the Mundari- and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/insertion polymorphism in Indian populations. *Hum Genet.* 2005; 116:506–517.

Appendix

Expressions for the marker allele frequencies in cases and controls in presence of population stratification:

$$P(M|case) = \frac{P(M \cap case)}{P(case)} = \frac{\sum_{i=1}^k \sum_{G \in \mathcal{G}} P(M \cap G|S_i) f_G P(S_i)}{\sum_{i=1}^k \sum_{G \in \mathcal{G}} P(M \cap G|S_i) (1 - f_G) P(S_i)} = \frac{\sum_{i=1}^k \alpha_i m_i d_i}{\sum_{i=1}^k \alpha_i d_i}$$

$$P(M|control) = \frac{P(M \cap control)}{P(control)} = \frac{\sum_{i=1}^k \sum_{G \in \mathcal{G}} P(M \cap G|S_i) (1 - f_G) P(S_i)}{\sum_{i=1}^k \alpha_i (1 - d_i)} = \frac{\sum_{i=1}^k \alpha_i m_i (1 - d_i)}{\sum_{i=1}^k \alpha_i (1 - d_i)}$$

where f_G is the penetrance of the disease for disease genotype G and $\mathcal{G} = \{D_1 D_1, D_1 D_2, D_2 D_2\}$

ANOVA and the Kruskal-Wallis test in the presence of population stratification:

If F_i is the distribution function of Y in the i^{th} subpopulation, $i = 1, 2, \dots, k$, irrespective of the QTL genotype, then the distribution function F of Y for marker genotype g in the combined population can be computed as:

$$F(y|g) = \sum_{i=1}^k F(y|S_i) P(S_i|g) = \sum_{i=1}^k F(y|S_i) \frac{P(S_i) P(g|S_i)}{\sum_{i=1}^k P(S_i) P(g|S_i)} = \frac{\sum_{i=1}^k F_i(y) P(g|S_i) \alpha_i}{\sum_{i=1}^k P(g|S_i) \alpha_i}$$

Thus, the distribution function of Y in the combined population is a convex combination of the distribution functions (F_i s) of Y for the same marker genotype in the subpopulations as provided in Equation (3).

Since the marker allele M_1 is in perfect linkage equilibrium with the QTL in each subpopulation, the distribution functions (F_i s) are independent of g . Hence,

$$E(Y|g) = \int y dF(y|g) = \frac{\sum_{i=1}^k P(g|S_i) \alpha_i \int y dF_i(y|g)}{\sum_{i=1}^k P(g|S_i) \alpha_i} = \frac{\sum_{i=1}^k P(g|S_i) \alpha_i \mu_i}{\sum_{i=1}^k P(g|S_i) \alpha_i}$$

Thus the mean of the quantitative trait (Y) for a given marker genotype g in the combined population can be expressed in terms of the means of the quantitative trait for the same genotype in the subpopulations as provided in Equation (2).

ANOVA in the case of two-subpopulations

The mean of Y given a marker genotype g in the combined population is given by:

$$E(Y|g) = \frac{\mu_1 P(g|S_1) \alpha_1 + \mu_2 P(g|S_2) \alpha_2}{P(g|S_1) \alpha_1 + P(g|S_2) \alpha_2}$$

For the genotype g_i at the marker locus,

Define $s_i = P(g_i|S_1) \alpha_1$, $t_i = P(g_i|S_2) \alpha_2$ and $u_i = \frac{s_i}{s_i + t_i}$, $i = 1, 2, 3$.

Then, $E(Y|g_i) = \frac{\mu_1 s_i + \mu_2 t_i}{s_i + t_i} = (\mu_1 - \mu_2)u_i + \mu_2 = w_i$ (say) for $i = 1, 2, 3$.

$\bar{w} = (\mu_1 - \mu_2) \bar{u} + \mu_2$, where $\bar{u} = \frac{\sum_{i=1}^3 n_i u_i}{n}$

The test statistic for ANOVA is:

$$\frac{MSH}{MSE} = \frac{\sigma^2 + \frac{\sum_{i=1}^3 n_i (w_i - \bar{w})^2}{2}}{\sigma^2} = 1 + \frac{\sum_{i=1}^3 n_i (w_i - \bar{w})^2}{2\sigma^2} = 1 + (\mu_1 - \mu_2)^2 \frac{\sum_{i=1}^3 n_i (u_i - \bar{u})^2}{2\sigma^2}$$

Thus, the ratio of MSH to MSE differs from the null value of 1 by a quantity

$$(\mu_1 - \mu_2)^2 \frac{\sum_{i=1}^3 n_i (u_i - \bar{u})^2}{2\sigma^2}.$$

Figure 1.

False positive rates when the QT is distributed as normal in both the subpopulations, but with different means.

Figure 2. False positive rates when the QT is distributed as chi-squares with 4 d.f. in both the subpopulations, but with different location shifts.

Figure 3. False positive rates when the QT is distributed as chi-squares with 1 d.f. in both the subpopulations, but with different location shifts.

Figure 4. False positive rates when the QT is distributed as normal in sub-population 1 and chi-squares with 4 d.f. in subpopulation 2.

Figure 5. False positive rates when the QT is distributed as normal in sub-population 1 and chi-squares with 1 d.f. in subpopulation 2.

Table 1

Simulation Parameters

Model	Distribution of QT		Marker Allele Frequency	
	Population 1	Population 2	Population 1	Population 2
1	$N(9, 1)$	$N(9 + d, 1)$	0.1	0.9
			0.3	0.6
			0.2	0.4
			0.1	0.9
2	χ_4^2	$\chi_4^2 + d$	0.1	0.9
			0.3	0.6
			0.2	0.4
			0.1	0.9
3	χ_1^2	$\chi_1^2 + d$	0.1	0.9
			0.3	0.6
			0.2	0.4
			0.1	0.9
4	$N(9 + d, 1)$	$\chi_4^2 + 9$	0.1	0.9
			0.3	0.6
			0.2	0.4
			0.1	0.9
5	$N(9 + d, 1)$	$\chi_1^2 + 9$	0.1	0.9
			0.3	0.6
			0.2	0.4
			0.1	0.9