Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations

Martin et al

ONLINE SUPPLEMENTAL MATERIAL

Supplemental Methods and Results

S1. METHODS FOR SIMULATIONS

We performed simulations to test our theoretical conclusions and further analyze more complex population demographic scenarios. We employed simuPOP (Peng & Kimmel, 2005), a forward-time population genetics simulation program. It provides a Python interface for developing more advanced simulations, including evolutionary processes such as migration. An overview of the simulation design is shown in Supplemental Figure 8.

S1.1 Initial Reference Populations

We downloaded genotype data of CEU and YRI individuals from HapMap3 (International HapMap Consortium et al., 2007). Due to memory capacity and time constraints, we selected 20,000 SNPs from the first ~41Mb of chromosome 2 to represent our simulated "chromosome" (resulting in ~2kb intermarker distance between adjacent pairs). To mimic the founder effect in evolving admixed populations, we randomly selected 10 individuals among unrelated CEU and YRI individuals to represent founders of our simulated European and African ancestral populations.

S1.2 Generating Admixed Subpopulations

Starting with the CEU and YRI founder populations, we expanded two ancestral populations for 100 generations (assuming linear growth with random mating) until each of them reached a population size of 20,000. We assumed a mutation rate of 10^{-8} per nucleotide, recombination intensity 10^{-8} (rate=intensity $\times$ physical distance), and random mating within each population. As a measure of diversity between simulated CEU and YRI populations, we found that Fst was 0.15 in the founder generation, then increased to 0.21 in the first generation, leveling off and

ending at 0.249 in the final (100^{th}) generation. To simulate admixture, we mixed the ancestral populations by setting different migration scenarios. For the first admixed subpopulation, we set the migration rate to $q_0 = 0.1$ from CEU to YRI and $q_1 = 0.1$ from YRI to CEU for subpopulation 0 and 1, respectively. Admixture occurred for the first 5 generations, followed by 20 generations of random mating within the admixed subpopulations. Finally, we randomly chose 24,000 individuals from each of the admixed subpopulations to form admixed subpopulation 0 and 1.

S1.3 Generating Quantitative Trait

In order to assign quantitative trait values, we chose a SNP (rs11897611) near the center of the simulated chromosome to represent the QTL. This SNP was chosen because of its different allele frequencies in the ancestral populations (p_0=0.932 for YRI and p_1 =0.125 for CEU). We simulated quantitative trait values for all individuals depending on QTL genotype based on the mixture distributions described in the main text (equations 9 and 13). For the single-admixed population model, we assessed two genetic models: one with the trait mean to $a = 0$ to simulate the null hypothesis of no genetic effect and another with $a = 1$ to simulate a genetic effect. We used a trait variance $\sigma^2 = 2$ for both models. For the stratified population model, we added a constant c=0, 1 or 2 to the trait mean of the second admixed population.

S.1.4 Statistical Analyses in Simulated Data

The regression models described in the main text were fit in samples from the simulated populations. We used the known local ancestry for any locus; that is by tracking the ancestral population of origin for each locus in simulations. Global ancestry was computed as the average of local ancestries across a haplotype, and then averaged over haplotypes within an individual.

For the examples in a single admixed population, we analyzed 500 individuals chosen randomly from subpopulation 0 for each realization. For the examples in a stratified admixed population, each realization used 250 individuals from both subpopulation 0 and 1 were randomly chosen and combined for joint analysis. We averaged the statistics over 1000 realizations. We used in-house scripts written in Python and R to perform linear regression analysis. The genotype variable (1, 0 or -1) was extracted for the individuals using simuPOP's provided functions.

S2. RESULTS FROM SIMULATIONS

We conducted simulations to examine type I error and power (i.e., rejection rate) of tests based on adjusted and unadjusted regression models in the single admixed population model and the stratified admixed population model discussed above. All estimates are based on tests in 1000 replicate random samples of 500 individuals each. Each test uses a significance level of 0.05. We considered tests at the QTL and 10 other test markers across the simulated chromosome to illustrate our theoretical results (Supplemental Table 1).

S2.1 Single admixed population

We first examine results in the simulated single admixed population. Supplemental Table 1 shows the values of allele frequencies and LD in the ancestral and admixed populations. The SNPs were selected to show a range of LD values and ancestral allele frequency differences. Supplemental Figure 9 shows estimates of the rejection rates for tests of the genotype term from regression analyses using the unadjusted model (UN), the local-ancestry adjusted model (LA), the global-ancestry adjusted model (GA), and a model that adjusts for both global and

local ancestry (GLA). As predicted by theory, when $a = 0$ (Supp. Figure 9A) the QTL and all test markers have rejection rates near the nominal level of 0.05.

When $a = 1$ (Supp. Figure 9B), rejection rates vary dependent on LD. Our theoretical derivations showed that when $W \approx 0$, the regression parameter for the models adjusted by local ancestry, but not necessarily the unadjusted or global adjusted models, should be close to 0 and hence the rejection rate should be close to the nominal level. We further showed that when the LD values in the admixed population (D^*) are close to 0, we expect the UN model to have close to the nominal rejection rate. SNP 1 provides an example for which both W and D^* are close to 0, and indeed the rejection rates are close to the nominal level for the unadjusted model (UN) and the models adjusted for local ancestry (LA and GLA). SNPs 2, 3, 5 and 10 have $W \approx 0$ but D^* is moderate or large due to the differences in ancestral allele frequencies. As predicted by theory, the UN and GA models have higher rejection rates than the LA and GLA adjusted models, which are close to the nominal level since this represents the null hypothesis for the local-ancestry adjusted model. For SNP 9 $D^* \approx 0$ giving UN and GA tests that are close to the nominal level but W is slightly larger (in absolute value) than D^*, resulting in LA and GLA tests that are somewhat higher than the nominal level.

For SNPs 4, 6, 7, and 8 both $|W|$ and $|D^*|$ are greater than 0 and rejection rates can be interpreted as power. For SNPs 6 and 7 the UN and GA models outperform the LA and GLA models. These are SNPs for which W is larger than 0 but $D^* > W$. For SNPs 4 and 8, the LA and GLA models have higher rejection rates than the GA and UN models. These are SNPs with W is large relative to D^*.

In general, we found that the GA model performs similarly to the UN model or between the UN and LA models. This is consistent with our theory that showed that the regression parameter for the GA model should be close to the UN model if there has been sufficient recombination and close to the LA model if there has been little recombination. Our simulations used 20 generations of random mating after initial admixture and simulated a single chromosome. We observed on average 8 recombination events in any random chromosome from the final generation, which is what we expected after 20 generations of random mating with 20,000 variants with adjacent intermarker distance ~2kb. This is large enough to give similar parameter estimates for the GA model similar to the UN model but not so large as to make the estimates equivalent. We did not derive the regression parameters for the model adjusted for both local and global ancestry (GLA), but for the examples considered here, the GLA model always performed similarly to the LA model.

For the QTL itself, all models have rejection rates (power) of 1 for the sample size of 500, with the given QTL model. To better distinguish relative power at the QTL, we also simulated samples of size 100. We found relative power for UN, GA, LA and GLA models of 0.883, 0.843, 0.57 and 0.569, respectively. These results are consistent with our ARP calculations showing that at the QTL power should be ordered UN>GA>LA.

S2.2 Stratified admixed populations

We next examine results in simulated stratified admixed populations using the same QTL and set of test markers as above. Supplemental Table 2 summarizes the values for the LD parameters for the two simulated admixed subpopulations and the stratified population as a whole. As a side note, subpopulation 0 is the same as the admixed population used in the single

admixed population examples above. In addition to the regression models considered above, we also include a model adjusted for subpopulation membership (MA). Supplemental Figure 10 shows rejection rates for the test of the genotype term for the different models when $a = 0$ for $c = 0$ and $c = 2$. For $c = 0$, there is no trait difference between the two strata, but there is random mating only within strata making this different from the single admixed population model. As shown by our theoretical calculations in this case when there is no genetic effect ($a = 0$), we expect all regression models (unadjusted and adjusted) to give rejection rates close to the nominal rate, and our simulations bear this out (Supp. Figure 10A). When $c \neq 0$, such that there is a difference in trait mean between strata but still no genetic effect ($a = 0$), we showed that only the regression models adjusting by local ancestry and population membership should be generally valid tests of no genetic effect (since they do not depend on c). As expected, Supplemental Figure 10B shows that these models (LA, GLA and MA) give consistently correct rejection rates. Results from the UN model show inflated rejection rates at markers with large differences in ancestral allele frequencies (SNPs 1, 2, 3, 6, 7 and 10). Interestingly, the GA model, has correct rejection rates, suggesting that the specification bias discussed in the methods section is small for these examples.

When $a > 0$ (Supplemental Figure 11), rejection rates depend on LD (as well as other parameters). We showed theoretically that when $W^* \approx 0$ (or $r_{gl} = 0$), the regression parameter for the models adjusted by local ancestry should be close to 0. SNPs 1, 2, 3, 5 and 10 provide examples for which $W^* \approx 0$ and show LA and GLA models with nominal rejection rates. We also showed that when there was no LD in both subpopulations, the regression parameter for the model adjusted for subpopulation membership should be close to 0. SNP 1 provides

such an example and shows the nominal rejection rate for the MA model. SNP 8 is particularly interesting because it has moderate LD in both of the subpopulations, and in fact showed power in the single admixed population examples above, but because the LD in the subpopulations is in the opposite directions, the test in the MA model loses power in the stratified population compared to the single admixed population (Figure 9B).

Only SNPs 4 and 9 show higher rejection rates for the local-ancestry adjusted models (LA and GLA) than the other models. These are SNPs that have relatively high values of LD in the ancestral populations compared to LD in the admixed populations or stratified population at large. In general, the GA models have rejection rates similar to the MA models or between the LA and MA models.

Rejection rates for the UN are greater than the nominal rate for all markers except SNPs 4, 8 and 9. These exceptions all have relatively small values of LD in the stratified population (D^{**}) and small differences between marker allele frequencies in the ancestral populations (small Δ_L); consequently, as theory showed, the regression parameter for the genotype term will be close to 0. Interestingly SNP 1 also has a small value of D^{**} but has a severely inflated rejection rate, particularly for $c = 2$ (Figure 11B). This is because the marker allele frequency difference in the ancestral populations is large which leads to a non-zero regression parameter regardless of the value of D^{**}.

REFERENCES

International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., . . . Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861. doi:10.1038/nature06258

Peng, B., & Kimmel, M. (2005). simuPOP: A forward-time population genetics simulation environment. *Bioinformatics (Oxford, England), 21*(18), 3686-3687. doi:bti584 [pii]

Supplemental Table 1. Properties of the QTL and 10 test markers: distance, allele frequencies (p_0, p_1) and LD measures between the QTL and test markers (D_0 and D_1 are computed in the simulated ancestral populations; W is computed as $W = 0.7D_0 + 0.3D_1$; D^* is computed in the simulated single admixed population).

SNP	Distance from QTL (bp)	p_0 (CEU)	p_1 (YRI)	D_0 (CEU)	D_1 (YRI)	W	D*
1	20451825	0.149	0.693	-0.0005	-0.0003	-0.0004	0.0041
2	9638088	0.195	0.885	-0.0004	0.0006	-0.0001	0.0228
3	5754699	0.342	0.934	0.0012	-0.0011	0.0005	0.0395
4	89846	0.277	0.155	0.0882	0.0062	0.0636	0.0456
5	59661	0.273	0.395	0.0012	0.0024	0.0015	0.0186
6	19418	0.063	0.817	-0.0079	0.0554	0.0111	0.1391
QTL	*0*	*0.125*	*0.932*	*1.0000*	*1.0000*	*1.0000*	*1.0000*
7	21771	0.092	0.754	0.0796	0.0505	0.0709	0.1838
8	63686	0.404	0.363	0.0687	-0.0410	0.0358	0.0375
9	303562	0.043	0.073	-0.0053	-0.0298	-0.0127	-0.0077
10	1088150	0.017	0.66	-0.0001	-0.0015	-0.0005	0.0857

Supplemental Table 2. LD between the QTL and 10 test markers in the two admixed subpopulations and stratified population as a whole: W_0 and W_1 are computed as $W_0 = 0.7D_0 + 0.3D_1$ and $W_1 = 0.3D_0 + 0.7D_1$ (where D_0 and D_1 are computed in the simulated ancestral populations); W^* is computed as $W^* = 0.5W_0 + 0.5W_1$; D_0^* and D_1^* are computed in the simulated admixed subpopulations; D^{**} is computed in the simulated stratified population.

SNP	W_0	W_1	W^*	D^*_0	D^*_1	D^{**}
1	-0.0004	-0.0004	-0.0004	0.0041	0.0020	0.0226
2	-0.0001	0.0003	0.0001	0.0228	0.0226	0.0464
3	0.0005	-0.0004	0.0000	0.0395	0.0351	0.0567
4	0.0636	0.0308	0.0472	0.0456	0.0071	0.0214
5	0.0015	0.0020	0.0018	0.0186	0.0210	0.0237
6	0.0111	0.0365	0.0238	0.1391	0.1607	0.1761
QTL	*1.0000*	1.0000	1.0000	*1.0000*	*1.0000*	*1.0000*
7	0.0709	0.0592	0.0650	0.1838	0.1642	0.1960
8	0.0358	-0.0081	0.0139	0.0375	-0.0187	0.0083
9	-0.0127	-0.0225	-0.0176	-0.0077	-0.0190	-0.0120
10	-0.0005	-0.0011	-0.0008	0.0857	0.0877	0.1111

SUPPLEMENTAL FIGURE LEGEND

Supplemental Figure 1. Plots of φ and φ_l as a function of s, the expected number of recombination events along a chromosome since initial admixture. φ_l is shown for l at the end of the end ($f = 0$) and middle ($f = 1/2$) of the chromosome maps. We considered K =1 or 22 chromosomes and varied the recombination probability between chromosomes: r_u: **A.** $K = 1$; **B.** $K = 22, r_u = 1$; **C.** $K = 22, r_u = 0.5$; and **D.** $K = 22, r_u = 0$.

Supplemental Figure 2. Plots of ratios of φ, $\varphi_l$and φ_g as a function of s, the expected number of recombination events along a chromosome since initial admixture. The function $\frac{\varphi_l^2}{\varphi}$ is shown for l at the end of the end ($f = 0$) and middle ($f = \frac{1}{2}$) of the chromosome maps. The function $\varphi_g \varphi_l / \varphi$ is shown for one locus at the end of the chromosome ($f = 0$) and the other locus in the middle ($f = \frac{1}{2}$) of the chromosome. We considered K =1 or 22 chromosomes and varied the recombination probability between chromosomes: r_u: **A.** $K = 1$; **B.** $K = 22, r_u = 1$; **C.** $K = 22, r_u = 0.5$; and **D.** $K = 22, r_u = 0$.

Supplemental Figure 3. ARPs for adjusted (local ancestry) model relative to unadjusted model, $\mathrm{ARP}(\hat{\beta}_G^*, \hat{\beta}_G)$, for the single admixed population with a range of values for p_1 and q, for A)$p_0 = 0.5$ and B) $p_0 = 0.9$. The test is conducted at a QTL with $\sigma^2 = 1$.

Supplemental Figure 4. ARPs for global-ancestry adjusted model relative to the unadjusted model, ARP($\hat{\beta}_G', \hat{\beta}_G$), for the single admixed population with a range of values for p_1 and q, for

$p_0 = 0.5$ with A)$r_u = 0.5$ and $s = 2$ and B)$r_u = 1$ and $s = 10$. The test is conducted at a QTL with $\sigma^2 = 1$.

Supplemental Figure 5. ARPs for local-ancestry adjusted model (LA) relative to the model adjusted for population membership (MA), $\mathrm{ARP}(\hat{\beta}_G^*, \hat{\beta}_G'')$, for a stratified admixed population with $Q = 0.5$, $q_0 = 0.65$, and a range of values for p_1 and q_1. Panels show A)$p_0 = 0.5$, $c = 1$; B) $p_0 = 0.9$, $c = 1$, C)$p_0 = 0.5$, $c = 2$; D) $p_0 = 0.9$, $c = 2$. The test is conducted at a QTL with $\sigma^2 = 1$.

Supplemental Figure 6. Violin plots for absolute value of estimates of disequilibrium coefficients (D and D^*) from seven admixed datasets and three non-admixed ("ancestral") datasets. Results are binned by intermarker distance, with the final bin being pairs of unlinked markers.

Supplemental Figure 7. Violin plots for absolute value of estimates of disequilibrium coefficients three admixed datasets. D^* is the estimate in the admixed dataset and W is the average of disequilibrium coefficients in the ancestral populations, weighted by estimates of mixing proportions: Honduran (0.62 European, 0.09 African, 0.29 Native American), Puerto Rican (0.80 European, 0.13 African, 0.07 Native American) and Mexican (0.50 European, 0.05 African, 0.45 Native American). Results are binned by intermarker distance, with the final bin being pairs of unlinked markers.

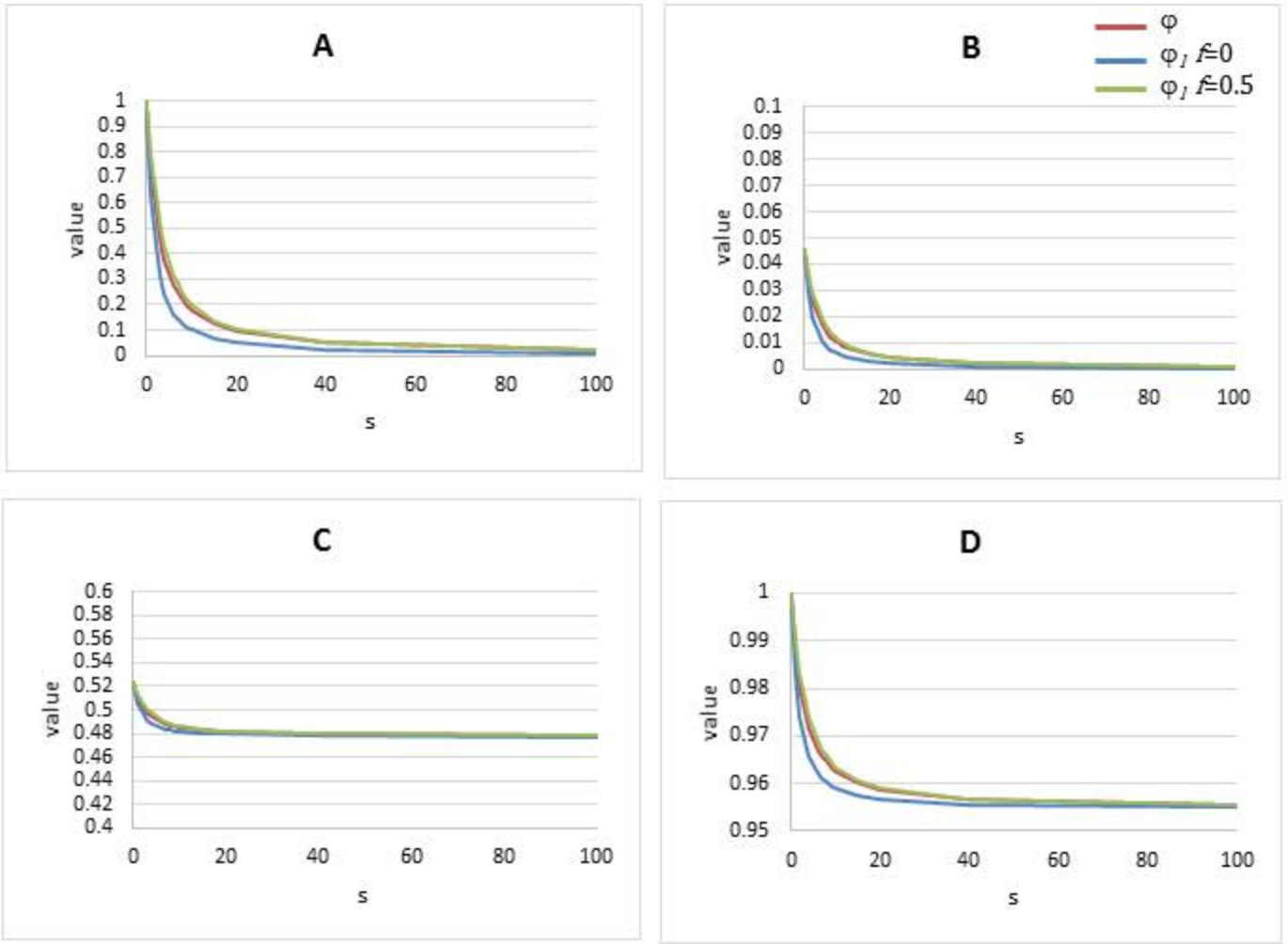
Supplemental Figure 8. Diagram of population models. Two ancestral populations migrate to form initial admixed populations (Admixed Pop i has mixing proportions q_i from Ancestral Pop 0 and $1 - q_i$ from Ancestral Pop 1). Followed by generations of random mating within each admixed subpopulation. For the single admixed population model only Admixed Pop 0 is considered. The stratified admixed population model considers the two admixed subpopulations as a whole with proportions Q for Admixed Pop 0 and $1 - Q$ for Admixed Pop 1.

Supplemental Figure 9. Rejection rates from regression analyses using the unadjusted model (UN), the local-ancestry adjusted model (LA), the global-ancestry adjusted model (GA), and a model adjusted by both global and local ancestry (GLA) from simulations of a single admixed population with a sample size of N=500. A. significance level of 0.05 (dashed, red line) was used for each test: A. No genetic effect ($a = 0$), B. Genetic effect ($a = 1$).

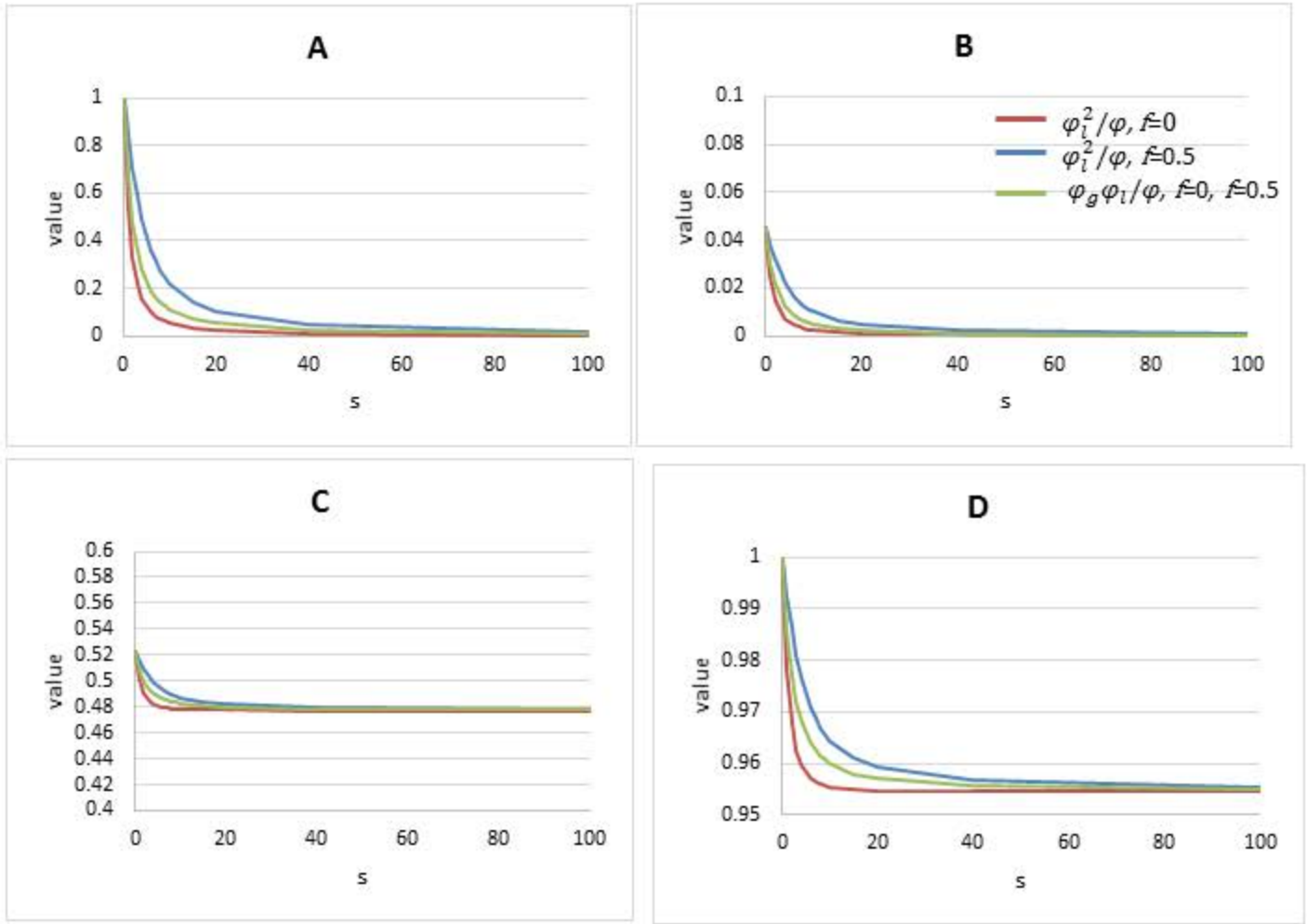
Supplemental Figure 10. Rejection rates from regression analyses using the unadjusted model (UN), the local-ancestry adjusted model (LA), the global-ancestry adjusted model (GA), a model adjusted by both global and local ancestry (GLA), and a model adjusted by subpopulation membership (MA) from simulations of stratified admixed populations with a sample size of N=500 and no genetic effect ($a = 0$): A. significance level of 0.05 (dashed, red line) was used for each test: A. No trait mean difference between strata ($c = 0$), B. Trait mean difference between strata ($c = 2$).

Supplemental Figure 11. Rejection rates from regression analyses using the unadjusted model (UN), the local-ancestry adjusted model (LA), the global-ancestry adjusted model (GA), a model adjusted by both global and local ancestry (GLA), and a model adjusted by subpopulation membership (MA) from simulations of stratified admixed populations with a sample size of N=500 and genetic effect ($a = 1$): A. significance level of 0.05 (dashed, red line) was used for each test: A. No trait mean difference between strata ($c = 0$), B. Trait mean difference between strata ($c = 2$).

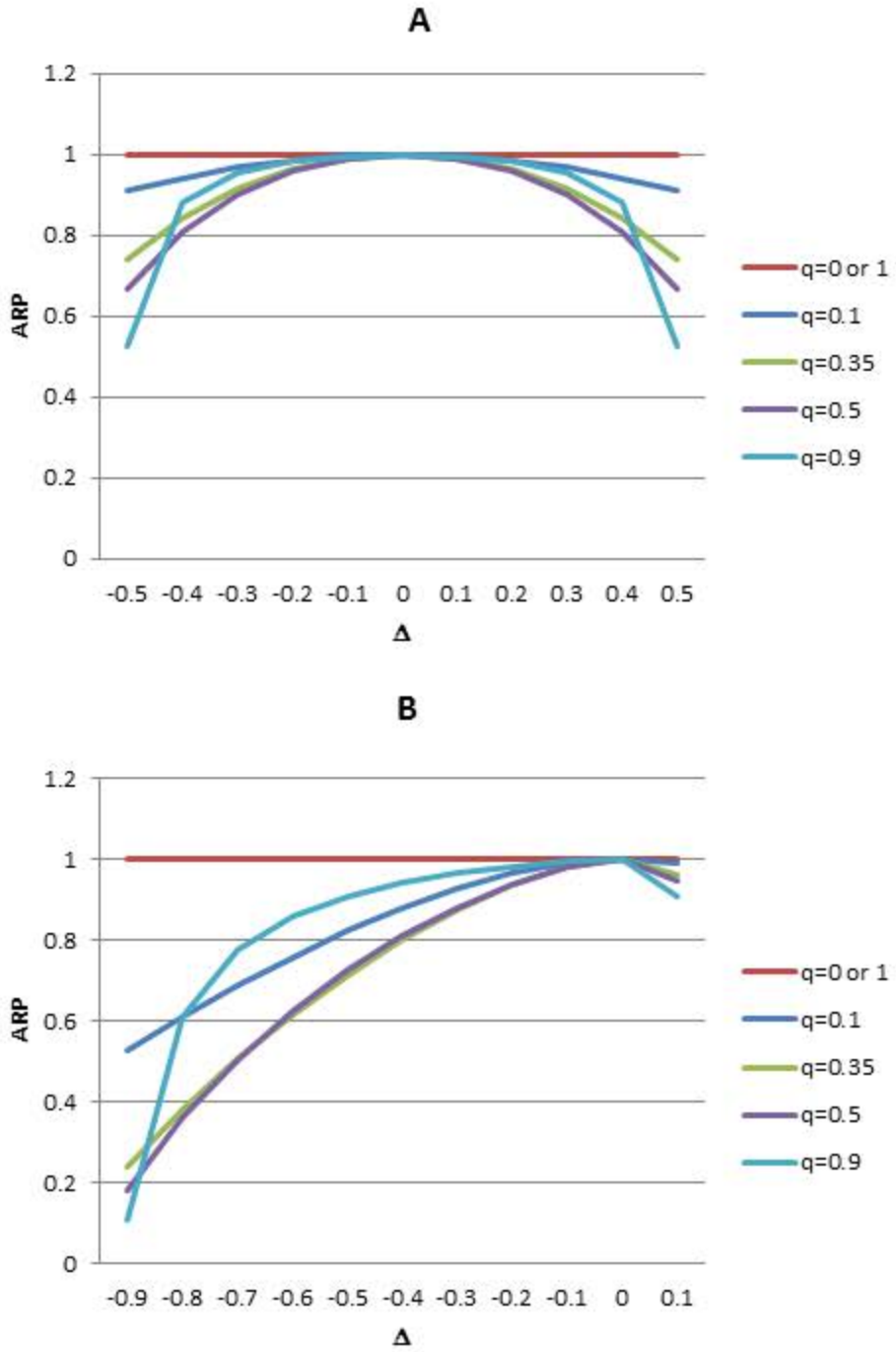
Supplemental Figure 1



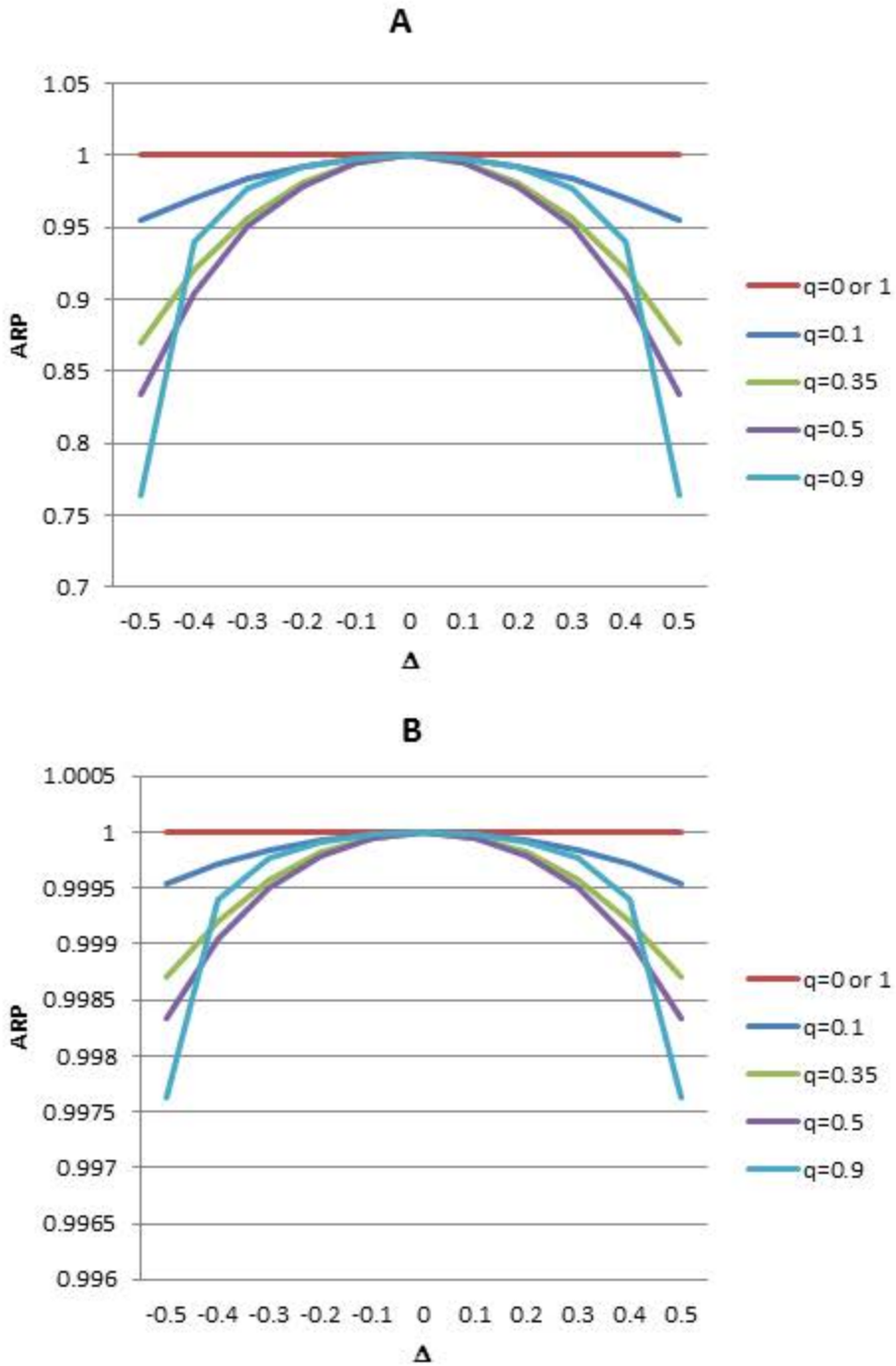
Supplemental Figure 2



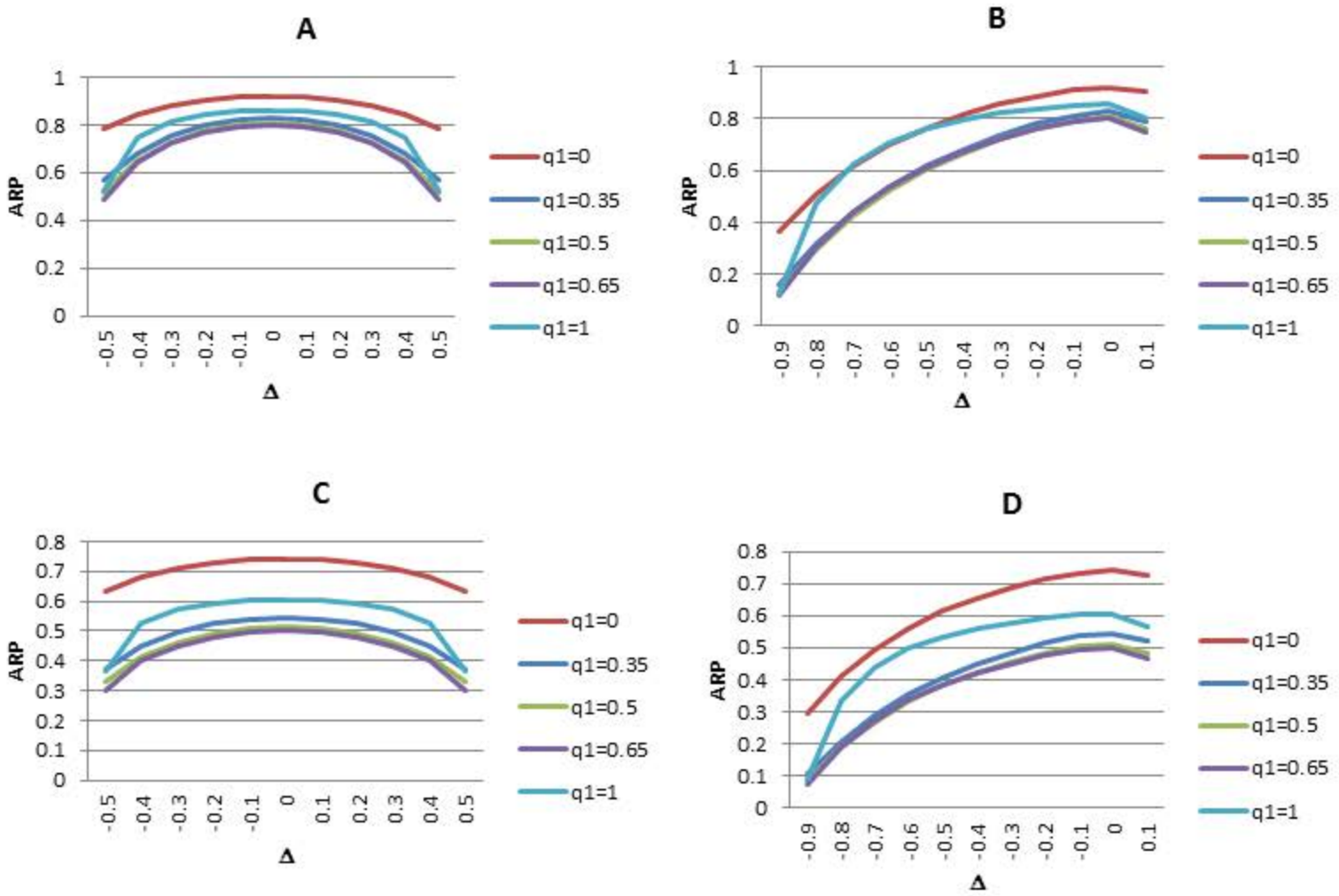
Supplemental Figure 3

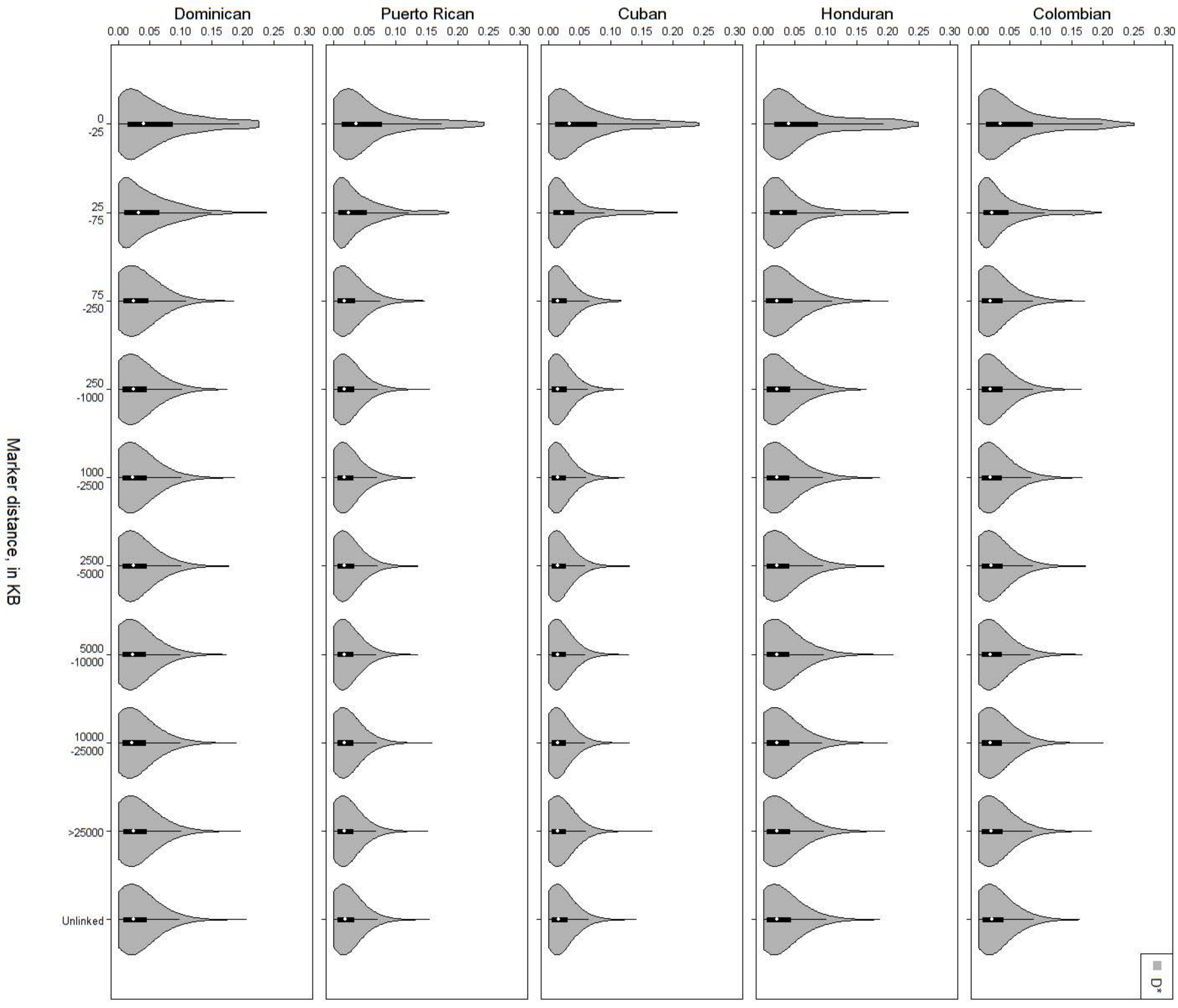


Supplemental Figure 4



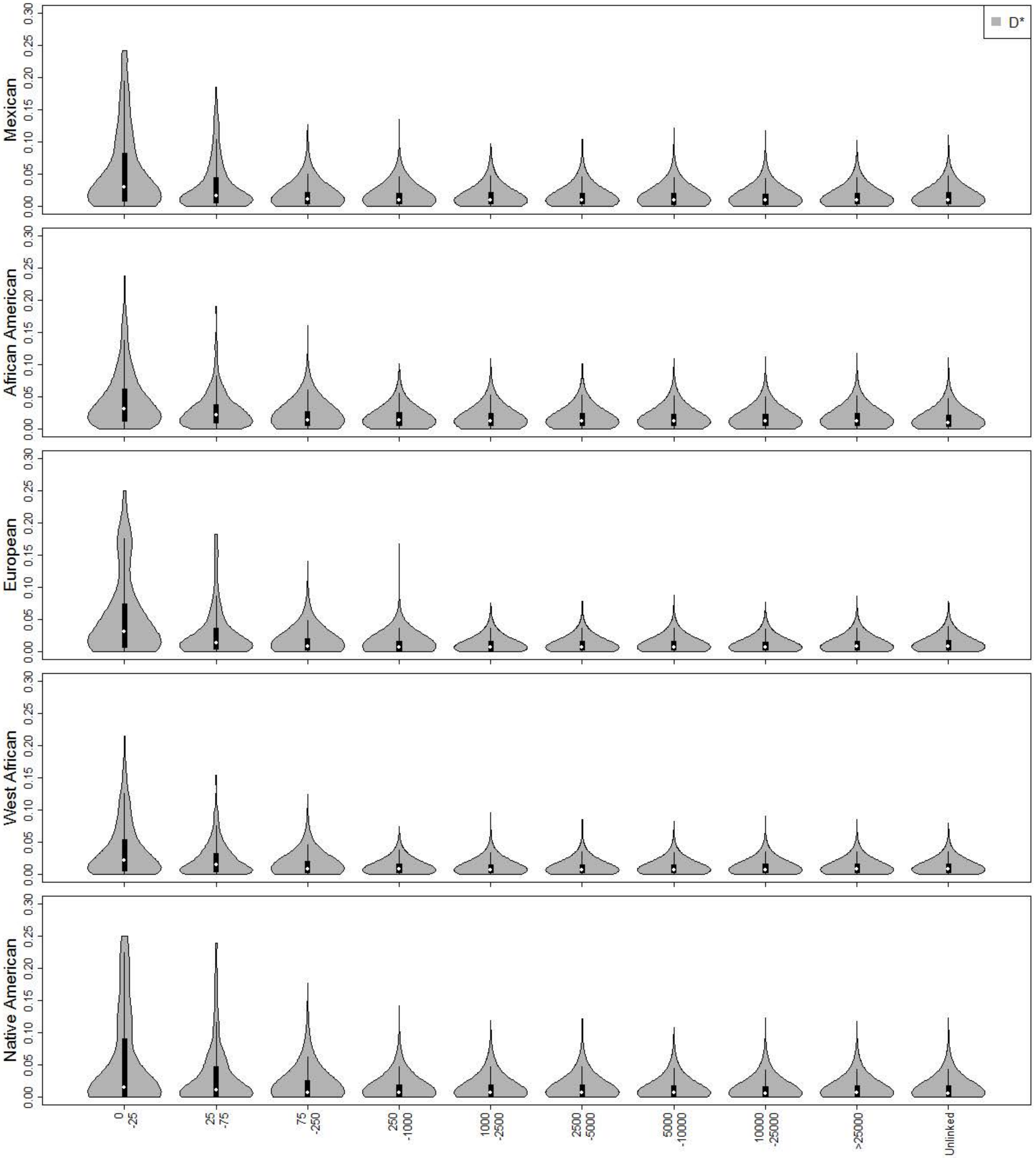
Supplemental Figure 5



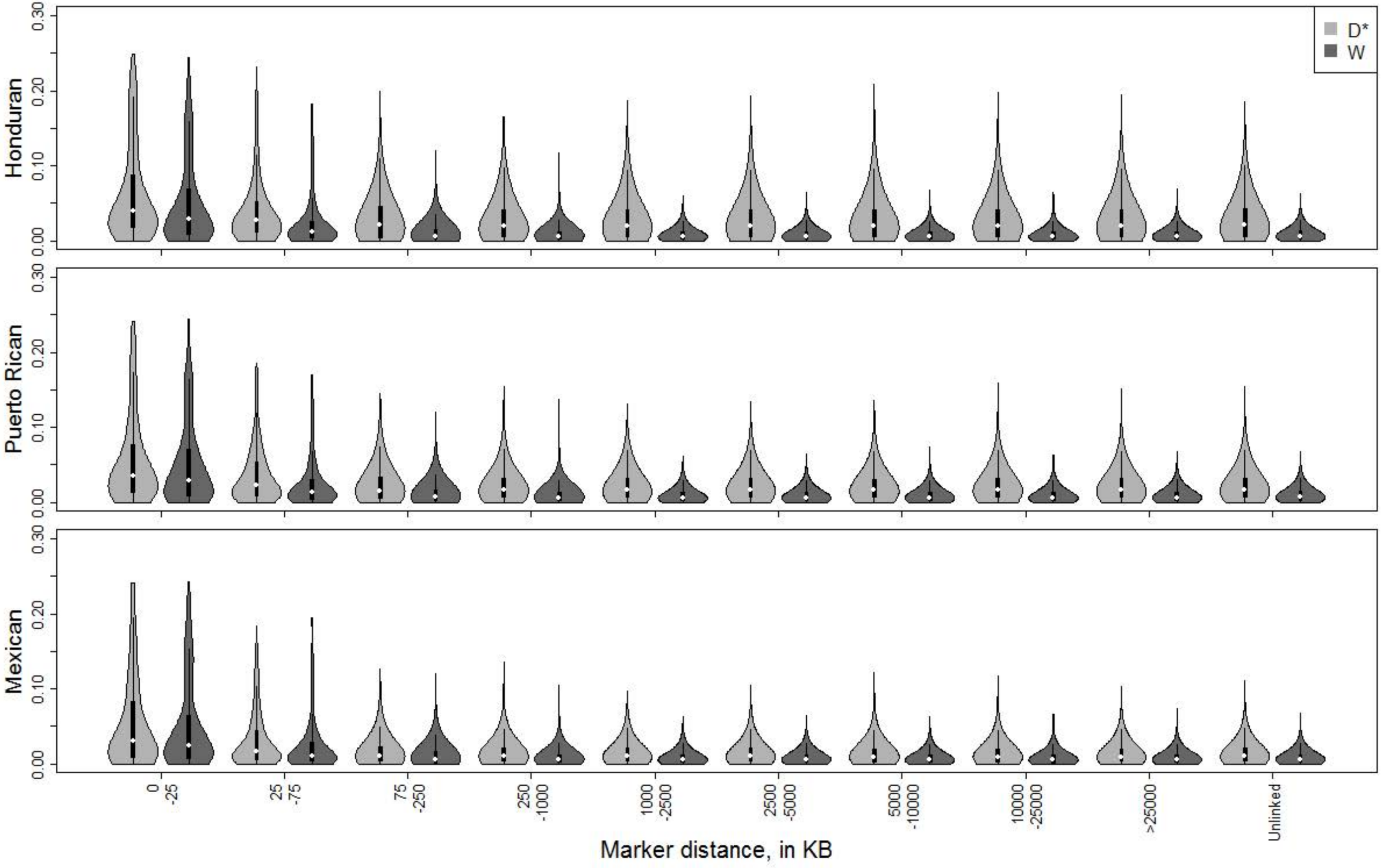


Supplemental Figure 6- page 1

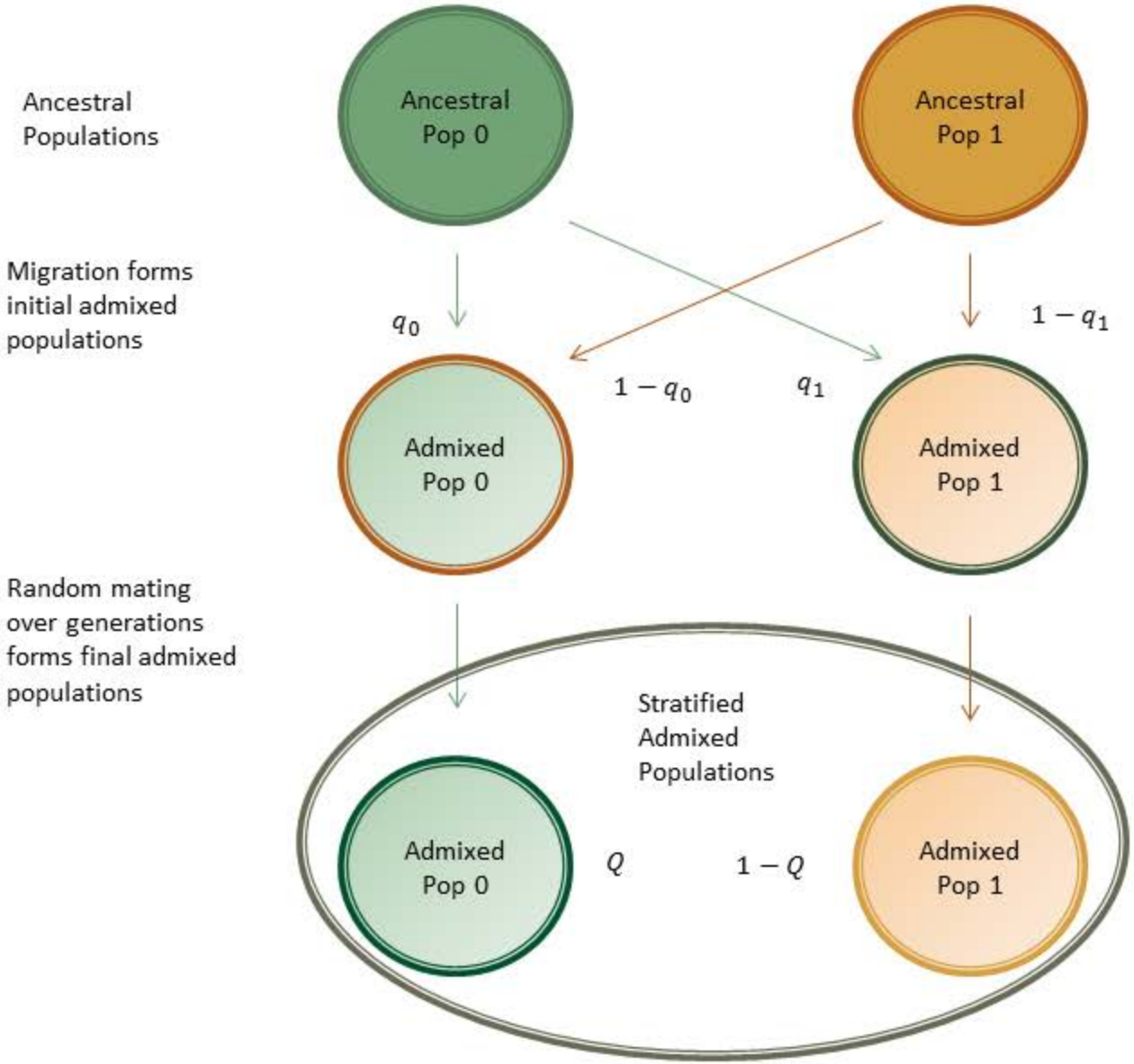
Supplemental Figure 6-page 2



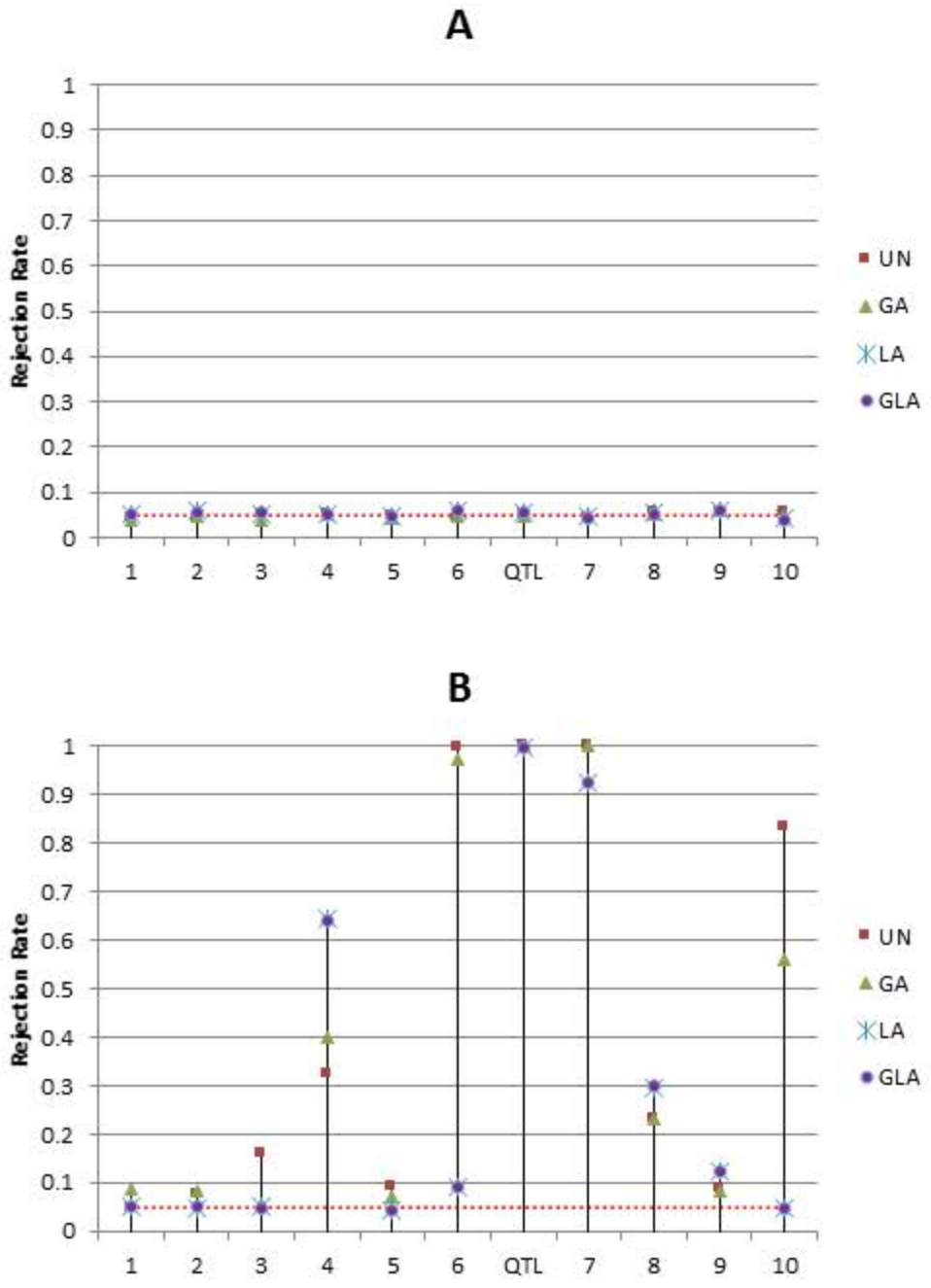
Supplemental Figure 7



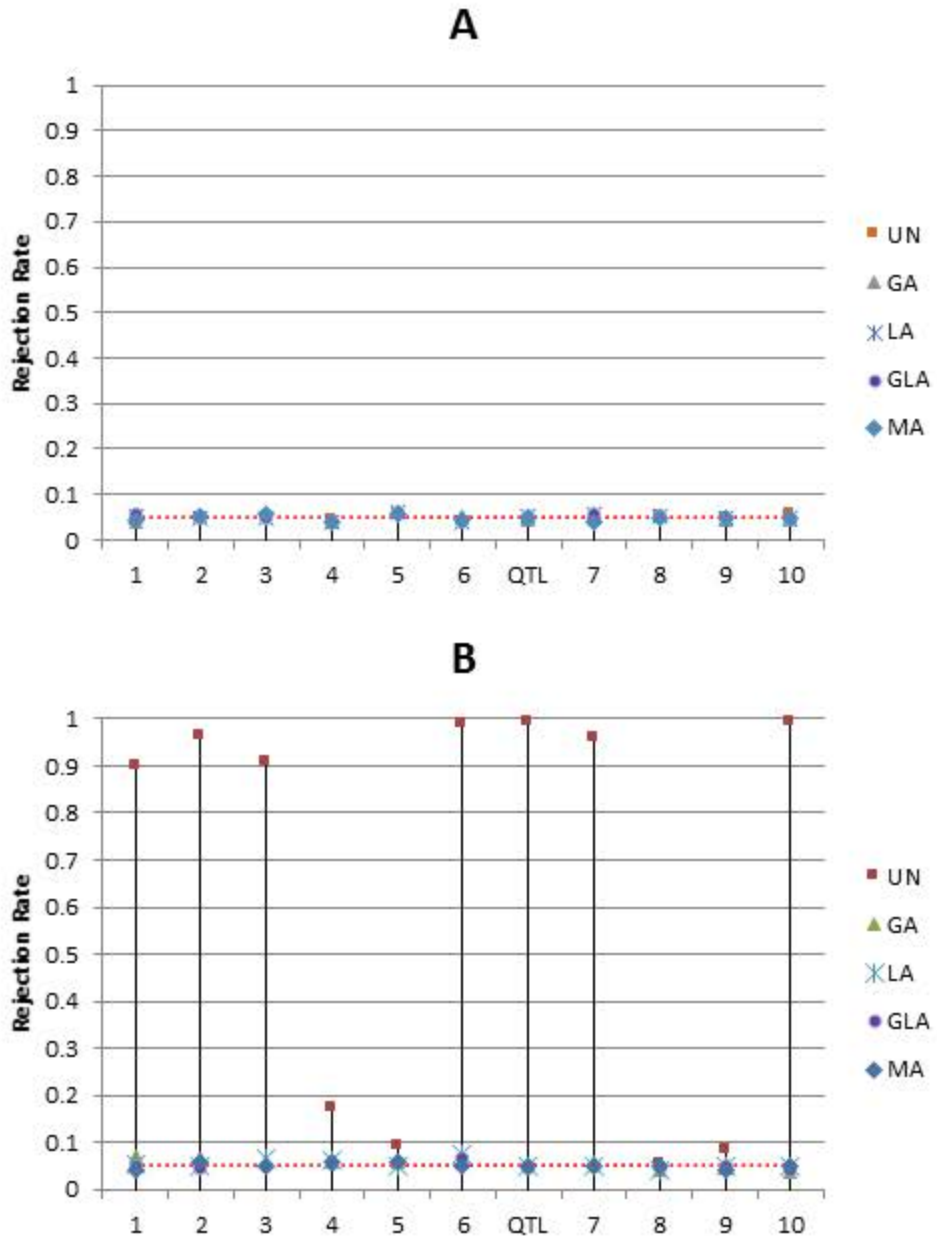
Supplemental Figure 8



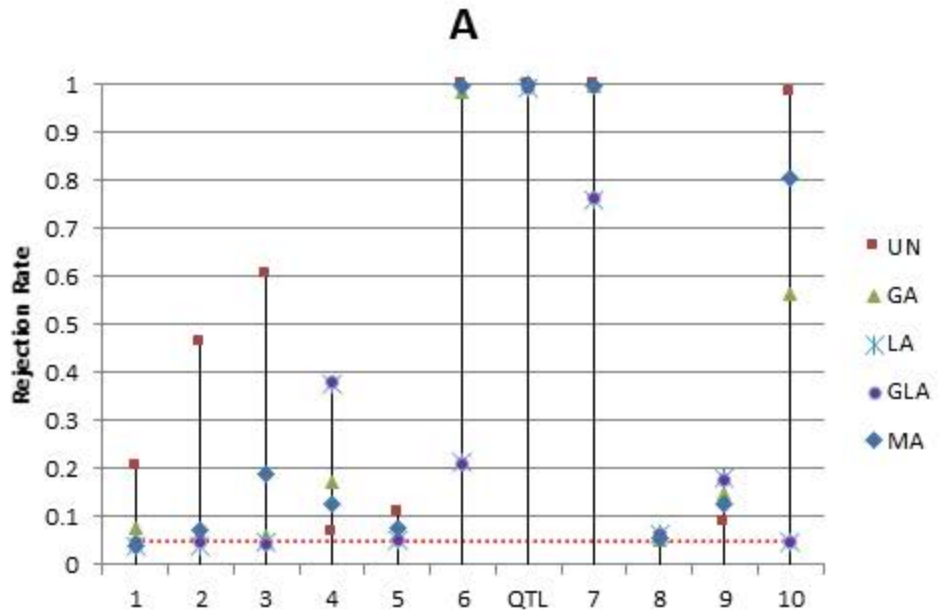
Supplemental Figure 9

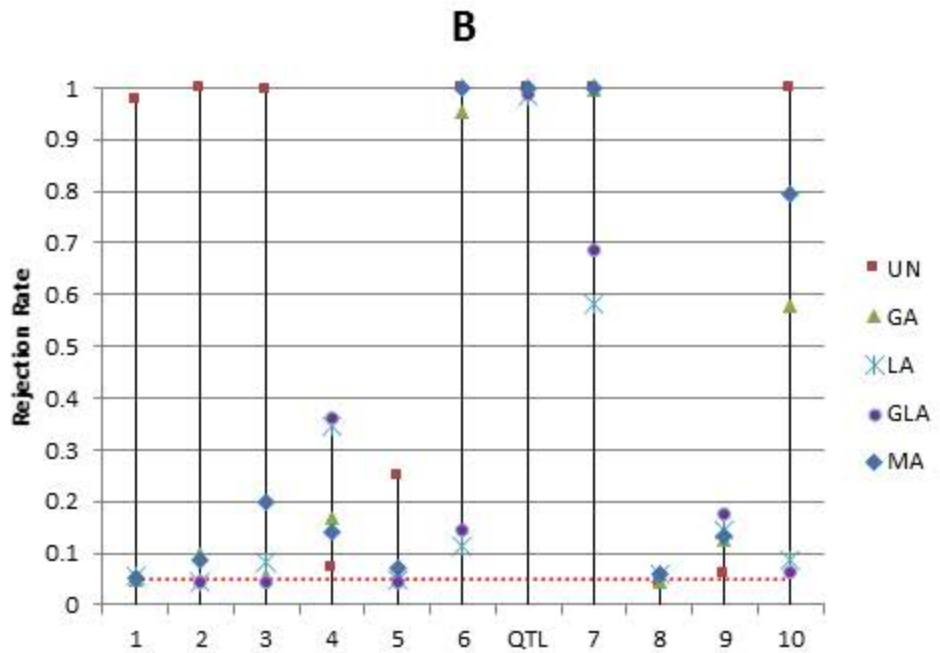


Supplemental Figure 10



Supplemental Figure 11





Appendices *(Online Supplemental Material)*

Appendix A.1. Variances, covariances, correlations and regression parameters for testing at the QTL in a single admixed population model.

A.1.1 Definitions

Consider a single admixed population generated by migration between two ancestral populations. We assume that following the initial migration, there has been random mating within the admixed population and no further migration (see main text for details). Let

$q =$ probability a locus on a randomly sampled haplotype is from ancestral population 1.

Suppose there is a QTL with alleles T_1 and T_2, and let

p_i=the frequency of allele T_1 from ancestral population i, for $i = 0,1$.

We define the following functions of parameters:

$\Delta = p_1 - p_0$ is the ancestral allele frequency difference

$p^* = qp_1 + (1-q)p_0$ is the frequency T_1 of in the admixed population,

and the following random variables for any individual:

For QTL genotype $G = \begin{cases} 1 & \text{if } T_1T_1 \\ 0 & \text{if } T_1T_2 \\ -1 & \text{if } T_2T_2 \end{cases}$;

and for the quantitative trait $Y \sim \begin{cases} N(a,\sigma^2) & \text{if } T_1T_1 \\ N(0,\sigma^2) & \text{if } T_1T_2 \\ N(-a,\sigma^2) & \text{if } T_2T_2 \end{cases}$.

For the jth locus, the average ancestry over the two haplotypes, $A_{.j} = (A_{1j}+A_{2j})/2$, where for the hth haplotype,

A_{hj}= 1 if the jth locus comes from ancestral population 1

0 if the jth locus comes from ancestral population 0 .

As described in the main text, we consider both local measures of ancestry, $A = A_{.j}$, and global measures of ancestry averaged over M loci in the genome,

$$\bar{A} = \frac{\sum_{j=1}^{M} A_{.j}}{M} .$$

We can then derive the means, (μ), variances (V) and covariances (C) for the above random variables as follows.

A.1.2 Means and Variances

We let $G = X_1 + X_2 - 1$, where X_1 and X_2 are indicator variables for the presence of the T_1 allele on haplotype 1 and 2, respectively. Under the assumption of HWE X_1 and X_2 are independent Bernoulli random variables with probability p^* gives:

$$\mu_G = 2p^* - 1 \qquad \text{and} \qquad V_G = 2p^*(1-p^*).$$

The phenotypic variable,Y, follows a mixture distribution dependent on G, where $Y = aG + N(0,\sigma^2)$. This gives:

$$\mu_Y = a\mu_G \qquad \text{and} \qquad V_Y = a^2V_G + \sigma^2$$

$$= a(2p^* - 1) \qquad = 2a^2p^*(1-p^*) + \sigma^2.$$

Local ancestry at any locus is the average of two independent Bernoulli random variables (one for each haplotype) with probabilities q, which gives:

$$\mu_A = q \qquad \text{and} \qquad V_A = q(1-q)/2.$$

The distribution of global ancestry depends on the probability of recombination events along the chromosomes. Let r_{jk} be the probability that there has been a recombination event between the jth and kth locus on the same chromosome at some point since the initial admixture. The variable r_{jk} is a function of the recombination fraction between the loci and the number of generations of random mating since initial admixture. For unlinked loci on different chromosomes, let this recombination probability be denoted r_U. This is the probability refers to the exchange of non-gametic alleles (and for example would be ½ in a single generation). We can write global ancestry as an average over the two independent haplotypes:

$$\bar{A} = 1/2\left(\frac{\sum_{j=1}^{M} A_{1j}}{M} + \frac{\sum_{j=1}^{M} A_{2j}}{M}\right).$$

Since the measures on the two haplotypes are iid, we can focus on the distribution of ancestry along a single haplotype, e.g., A_{1j}. We can easily see that for each locus,

$$\mu_{A1} = q \qquad \text{and} \qquad V_{A1} = q(1-q).$$

For two loci j and k on the same chromosome, we can derive

$$Cov\left(A_{1j}, A_{1k}\right) = q(1-q)(1-r_{jk}).$$

For two loci j and j' on different chromosomes, we can derive

$$Cov\left(A_{1j}, A_{1j'}\right) = q(1-q)(1-r_U).$$

Suppose that there are K chromosomes and for simplicity, we assume the same number of loci (m) have been observed for each chromosome, so that the total number of loci is $M = Km$. Then

$$Var\left(\sum_{j=1}^{M} A_{1j}\right) = \sum_{j=1}^{M} Var(A_{1j})$$

$$+K\sum_{j=1}^{m}\sum_{k\neq j}^{m} Cov(A_{1j}, A_{1k}) + K(K-1)\sum_{j=1}^{m}\sum_{j\prime=1}^{m} Cov(A_{1j}, A_{1j\prime}),$$

where the first term sums all variances, the second term is the sum of all pairwise covariances along a single chromosomes and the third term is the sum of pairwise covariances between pairs of loci on different chromosomes. Plugging in the derivations above, we have

$$Var\left(\sum_{j=1}^{M} A_{1j}\right) = M^2 q(1-q)[1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq k}^{m} r_{jk}].$$

Letting $\varphi = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq j}^{m} r_{jk}$, it follows that

$$\mu_{\bar{A}} = q \qquad \text{and} \qquad V_{\bar{A}} = \frac{q(1-q)\varphi}{2}.$$

A.1.3 Covariances

Next, we derive the covariances between the random variables. The covariance between the trait and QTL genotype is

$$C_{YG} = aV_G$$
$$= 2ap^*(1-p^*).$$

The covariance between the QTL genotype and the local ancestry at the QTL can be derived easily by recognizing that we can write G and A in terms of the two haplotypes: $G = (X_1 - 1/2) + (X_2 - 1/2)$ and $A = A_1/2 + A_2/2$. It follows that

$$C_{GA} = 2Cov\left(\left(X_1 - \frac{1}{2}\right), \frac{A_1}{2}\right)$$

$$= Cov(X_1, A_1)$$

$$= q(1-q)\Delta$$

and

$$C_{YA} = aC_{GA}$$

$$= aq(1-q)\Delta\,.$$

For covariances with global ancestry, we assume, without loss of generality, that the QTL is at the gth locus on the first chromosome. Then

$$C_{G\bar{A}} = \frac{1}{M}\sum_{j=1}^{M} Cov(G, A_{.j})$$

$$= \frac{1}{M}\left[C_{GA} + \sum_{j\neq g}^{M} Cov(G, A_{.j})\right].$$

The second term sums over loci on the same chromosome as the QTL and different chromosomes from the QTL. For locus j on the same chromosome with the QTL,

$$Cov\left(G, A_{.j}\right) = q(1-q)\Delta\left(1 - r_{gj}\right).$$

For locus j on a different chromosome from the QTL,

$$Cov\left(G, A_{.j}\right) = q(1-q)\Delta(1 - r_U).$$

This gives

$$C_{G\bar{A}} = q(1-q)\Delta\varphi_g.$$

where $\varphi_g = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{M}\sum_{j\neq g}^{m} r_{gj}$ (the final term is summed over all markers on the same chromosome with the QTL).

The covariance between the trait and global ancestry is a function of the covariance between the QTL genotype and global ancestry:

$$C_{Y\bar{A}} = aC_{G\bar{A}}$$
$$= aq(1-q)\Delta\varphi_g.$$

A.1.4 Correlations

With variances and covariances derived, we are now in a position to derive relevant correlations to assess confounding and validity of regression tests. First, we consider local ancestry. As noted in the main text, we need to examine two correlations to assess confounding: ρ^2_{GA} and $\rho^2_{YA,G}$. It is straightforward to show

$$\rho^2_{GA} = \frac{q(1-q)\Delta^2}{p^*(1-p^*)}.$$

It is convenient to define a new parameter

$$\gamma = qp_1(1-p_1) + (1-q)p_0(1-p_0)\,,$$

which allows us to write

$$\rho^2_{GA} = \frac{q(1-q)\Delta^2}{q(1-q)\Delta^2 + \gamma}.$$

The partial correlation is

$$\rho^2_{YA,G} = \frac{(\rho_{YA} - \rho_{YG}\rho_{GA})^2}{(1-\rho^2_{YG})(1-\rho^2_{GA})}$$

$$= \frac{(C_{YA}V_G - C_{YG}C_{GA})^2}{(V_Y V_G - C^2_{YG})(V_A V_G - C^2_{GA})}$$

$$= 0\ .$$

Noticing that $C_{YA} = aC_{GA}$ and $C_{YG} = aV_G$ shows immediately that this partial correlation is 0.

For global ancestry, the correlations are the following:

$$\rho^2_{G\bar{A}} = \left(\frac{q(1-q)\Delta^2}{q(1-q)\Delta^2 + \gamma}\right)\left(\frac{\varphi^2_g}{\varphi}\right)$$

and

$$\rho^2_{Y\bar{A},G} = 0\ .$$

The partial correlation is 0 because of the same relationships shown above for local ancestry.

A.1.5 Regression Parameters

Given the calculations in the previous sections, it is trivial to show that for the unadjusted model: $E(Y|G) = \beta_0 + \beta_G\ G$, we have

$$\beta_G = \frac{C_{YG}}{V_G} = a\ .$$

For the model adjusted by local ancestry: $E(Y|G, A_l) = \beta^*_0 + \beta^*_G G + \beta^*_A A_l$, the unstandardized multiple regression coefficient for genotype is

$$\beta_G^* = \sqrt{\frac{V_Y}{V_G}}\left(\frac{\rho_{YG} - \rho_{YA}\rho_{GA}}{1 - \rho_{GA}^2}\right)$$

$$= \frac{C_{YG}V_A - C_{YA}C_{GA}}{(V_G V_A - C_{GA}^2)}$$

$$= a\,,$$

which can be seen easily by noting that that $C_{YA} = aC_{GA}$ and $C_{YG} = aV_G$.

Similarly for the model adjusted by global ancestry: $E(Y|G, \bar{A}) = \beta_0' + \beta_G' G + \beta_{\bar{A}}' \bar{A}$, noting $C_{Y\bar{A}} = aC_{G\bar{A}}$ and $C_{YG} = aV_G$ shows

$$\beta_G' = a\,.$$

These calculations demonstrate validity of all three models for a test of the null hypothesis of no genetic effect ($a = 0$).

Appendix A.2. Variances, covariances, correlations and regression parameters for testing a marker locus in a single admixed population model.

A.2.1 Definitions

Consider a single admixed population generated by migration between two ancestral populations. Let q, p_i andp^* be defined as in Appendix A.1 as the mixing proportion, the QTL ancestral allele frequency and QTL allele frequency in the admixed population, respectively. However, suppose that instead of testing the QTL itself, we test a marker locus. Let the marker locus have two alleles, L_1, L_2, with

p_{Li} =the frequency of allele L_1 from ancestral population i, for $i = 0,1$.

We do not require that the alleles at the marker and QTL be independent, and define the linkage disequilibrium (LD) coefficient in ancestral population i:

$$D_i = P_{TLi} - p_{Li}p_i,$$

where P_{TLi} is the haplotype frequency of the haplotype carrying T_1 and L_1 from ancestral population i.

We define the following functions of parameters:

$$\Delta_L = p_{L1} - p_{L0}$$

$$p_L^* = qp_{L1} + (1-q)p_{L0}$$

$$W = qD_1 + (1-q)D_0.$$

We assume the same distributions for trait (Y) and QTL genotype (G) defined in Appendix A.1.1, and introduce a genotypic random variable based on the marker locus:

$$L = \begin{cases} 1 & \text{if } L_1L_1 \\ 0 & \text{if } L_1L_2 \\ -1 & \text{if } L_2L_2 \end{cases}.$$

We measure local and global ancestry as in Appendix A1, but local ancestry ($A_{.l}$) is measured at marker locus l which is being tested.

A.2.2 Mean and Variance The mean and variance of L are analogous to those of G:

$$\mu_L = 2p_L^* - 1 \qquad \text{and} \qquad V_L = 2p_L^*(1 - p_L^*).$$

The means and variances of the other random variables given in A.1.2 still hold.

A.2.3 Covariances

To derive the covariance between the QTL and the marker genotype random variables, we write G and L as a function of Bernoulli random variables,

$$G = X_1 + X_2 - 1 \text{ and } L = X_{L1} + X_{L2} - 1,$$

where X_1 and X_{L1} and Bernoulli random variables with probabilities p^* and p_L^*, respectively. The covariance between X_1 and X_{L1} can be written as follows and represents the disequilibrium coefficient in the admixed population, which we denote D^*, where

$$D^* = Cov(X_1, X_{L1}) = \left(1 - r_{gl}\right)(W + q(1-q)\Delta\Delta_L).$$

It follows that

$$C_{GL} = 2D^*.$$

Recognizing that $C_{YL} = aC_{GL}$ gives

$$C_{YL} = 2aD^*.$$

The covariance between the marker genotype and local ancestry at the marker is the same as for the QTL, replacing QTL parameters with marker parameters:

$$C_{LA_l} = q(1-q)\Delta_L \, .$$

The same is true for global ancestry:

$$C_{L\bar{A}} = q(1-q)\Delta_L \varphi_l \, .$$

$\varphi_l = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{M}\sum_{j \neq l}^{m} r_{lj}$ (the final term is summed over all markers on the same chromosome with the marker being tested).

Finally, we can show that the covariance between the QTL genotype and the local ancestry at the marker is

$$C_{GA_l} = q(1-q)\Delta\left(1 - r_{gl}\right) .$$

It then follows that

$$C_{YA_l} = aq(1-q)\Delta\left(1 - r_{gl}\right) .$$

A.2.4 Correlations

We now derive the correlations necessary to assess confounding and validity. For local ancestry, we have the following two to assess confounding:

$$\rho^2_{LA} = \frac{q(1-q){\Delta_L}^2}{p^*_L(1-p^*_L)}$$

$$= \frac{q(1-q){\Delta_L}^2}{q(1-q){\Delta_L}^2 + \gamma_L} \, ,$$

where $\gamma_L = qp_{L1}(1-p_{L1}) + (1-q)p_{L0}(1-p_{L0})$.

The partial correlation is more complicated. After some algebra, we can show

$$\rho^2_{YA,L}$$

$$= \frac{\Psi^2}{\Psi^2 + \left(q(1-q){\Delta_L}^2 + \gamma_L\right)\left(2a^2\left(\gamma\gamma_L - (1-r_{gl})^2W^2 + (1-\left(1-r_{gl}\right)^2)q(1-q)\Delta^2\gamma_L\right) + \gamma_L\sigma^2\right)}$$

where $\Psi = \sqrt{2q(1-q)}a(1-r_{gl})(\Delta\gamma_L - \Delta_L W)$.

For global ancestry, we can show

$$\rho^2_{L\bar{A}} = \left(\frac{q(1-q){\Delta_L}^2}{q(1-q){\Delta_L}^2 + \gamma_L}\right)\left(\frac{\varphi_l^2}{\varphi}\right)$$

$$= \rho^2_{LA}\left(\frac{\varphi_l^2}{\varphi}\right).$$

The partial correlation is complex. The numerator of $\rho^2_{Y\bar{A},L}$ becomes $(\Psi + \psi)^2$, where

$$\psi = 2aq(1-q)\left[D^*\Delta_L(1-\varphi_l) - p_L^*(1-p_L^*)\Delta\left(1-\varphi_g\right)\right].$$

The denominator is more complex and is not necessary to evaluate confounding, since we need only evaluate whether the numerator is 0, so we do not derive the denominator here.

A.2.5 Regression Parameters

Regression parameters for adjusted and unadjusted models can be derived as a function of the correlation coefficients. First consider the unadjusted model: $E(Y|L) = \beta_{L0} + \beta_L\ L$. We can show that

$$\beta_L = \frac{C_{YL}}{V_L}$$

$$= \frac{aD^*}{p_L^*(1-p_L^*)}$$

$$= \frac{aD^*}{q(1-q)\Delta_L^2 + \gamma_L}.$$

Next, we consider the model adjusted for local ancestry: $E(Y|L, A_l) = \beta^*_{L0} + \beta^*_L L + \beta^*_{LA} A_l$. We can show

$$\beta^*_L = \frac{C_{YL}V_A - C_{YA}C_{LA}}{(V_L V_A - C^2_{LA})}$$

$$= \frac{a(1 - r_{gl})W}{\gamma_L} .$$

Finally, we consider the model adjusted for global ancestry: $E(Y|L, \bar{A}) = \beta'_{L0} + \beta'_L L + \beta'_{L\bar{A}} \bar{A}$. The genotype coefficient for this adjusted model is

$$\beta'_L = \frac{C_{YL}V_{\bar{A}} - C_{Y\bar{A}}C_{L\bar{A}}}{\left(V_L V_{\bar{A}} - C^2_{L\bar{A}}\right)}$$

$$= \frac{a\left[D^*\varphi - q(1-q)\Delta\Delta_L\varphi_g\varphi_l\right]}{\left(q(1-q){\Delta_L}^2 + \gamma_L\right)\varphi - q(1-q)\Delta^2_L\varphi^2_l}$$

$$= \frac{a\left[D^* - q(1-q)\Delta\Delta_L\dfrac{\varphi_g\varphi_l}{\varphi}\right]}{q(1-q){\Delta_L}^2\left(1 - \dfrac{\varphi^2_l}{\varphi}\right) + \gamma_L}.$$

Appendix B.1. Variances, covariances, correlations and regression parameters for testing at the QTL in a stratified admixed population model (2 strata).

B.1.1 Definitions

Consider a stratified population that is composed of two admixed subpopulations (subpopulation 0 and 1), where each admixed subpopulation is derived from two ancestral populations (ancestry 0 and 1). There is random mating within admixed subpopulations but not between (see main text for details). We define the following probabilities:

Q =the probability that a random individual from the whole population is from subpopulation 1.

q_s = probability a locus is from ancestral population 1, for subpopulation $s = 0,1$.

Consider the QTL defined in Appendix A with alleles T_1, T_2 and with

p_i=the frequency of allele T_1 from ancestral population i, for $i = 0,1$. It follows that the allele frequency in subpopulation s is

$$p_s^* = q_s p_1 + (1 - q_s) p_0;$$

and the allele frequency in the stratified population as a whole is

$$p' = Q p_1^* + (1 - Q) p_0^*$$

$$= \alpha p_1 + (1 - \alpha) p_0 ,$$

where $\alpha = Q q_1 + (1 - Q) q_0$is the probability that a haplotype (at any specific position) sampled from the stratified population is from ancestral population 1.

We define the following functions of parameters:

$\Delta = p_1 - p_0$

$\gamma' = \alpha p_1(1-p_1) + (1-\alpha)p_0(1-p_0)$

$\delta = Q(1-Q)(q_1-q_0)^2$

$\omega = \alpha(1-\alpha) - \delta = Qq_1(1-q_1) + (1-Q)q_0(1-q_0)$.

The random variables for genotypic values (G), local ancestry (A) and global ancestry ($\bar{A}$) are defined as in Appendix A. For the stratified population model, we assume that the trait value depends on both the QTL genotype and which subpopulation the individual is from. Specifically, we assume that for subpopulation 0, the trait follows this distribution:

$$Y \sim \begin{cases} N(a, \sigma^2) & \text{if } T_1T_1 \\ N(0, \sigma^2) & \text{if } T_1T_2 \\ N(-a, \sigma^2) & \text{if } T_2T_2 \end{cases}.$$

For subpopulation 1, we assume that the mean values of the trait are shifted by a constant c, such that the trait distribution is the following:

$$Y \sim \begin{cases} N(a+c, \sigma^2) & \text{if } T_1T_1 \\ N(c, \sigma^2) & \text{if } T_1T_2 \\ N(-a+c, \sigma^2) & \text{if } T_2T_2 \end{cases}.$$

B.1.2 Means and Variances

As before, we write $G = X_1 + X_2 - 1$, where X_1 and X_2 are Bernoulli random variables (scoring the presence of the T_1 allele on each haplotype) with probability p'. However, because there is random mating within subpopulations, but not between subpopulations, the haplotypes are not independent. It is straightforward to show that $\mu_{X1} = \mu_{X2} = p'$, $V_{X1} = V_{X2} = p'(1-p')$ and $Cov(X_1, X_2) = \delta\Delta^2$. This gives:

$$\mu_G = 2p' - 1 \qquad \text{and} \qquad V_G = 2p'(1-p') + 2\delta\Delta^2$$

$$= 2(p_0 + \Delta\alpha) - 1 \qquad\qquad = 2[\gamma' + (\alpha(1-\alpha) + \delta)\Delta^2]\,.$$

Y follows a mixture distribution, dependent on G and subpopulation membership. If we let $S = 1$ if a random individual is from subpopulation 1, then S is Bernoulli with probability parameter Q. Then $Y = aG + cS + N(0, \sigma^2)$. This gives:

$$\mu_Y = a\mu_G + cQ$$
$$= a(2p' - 1) + cQ$$

and

$$V_Y = a^2 V_G + c^2 V_S + 2acCov(G, S) + \sigma^2$$
$$= 2a^2[p'(1-p') + \delta\Delta^2] + c^2 Q(1-Q) + 4ac(q_1 - q_0)Q(1-Q)\Delta + \sigma^2.$$

Local ancestry at any locus is the average of two correlated random Bernoulli random variables (A1 and A2, for haplotypes 1 and 2) with $\mu_{A1} = \mu_{A2} = \alpha$, $V_{A1} = V_{A2} = \alpha(1-\alpha)$ and $Cov(A_1, A_2) = \delta$. This gives

$$\mu_A = \alpha \qquad \text{and} \qquad V_A = \frac{[\alpha(1-\alpha) + \delta]}{2}.$$

For global ancestry, we assume as before that there are m loci on each of K chromosomes, and that r_{jk} and r_U are the recombination probabilities between the jth and kth locus on the same chromosome and unlinked loci on different chromosomes, respectively. We write global ancestry as an average over the two haplotypes:

$$\bar{A} = 1/2\left(\frac{\sum_{j=1}^{M} A_{1j}}{M} + \frac{\sum_{j=1}^{M} A_{2j}}{M}\right).$$

Unlike the single admixed population model (Appendix A), A_1 and A_2 are correlated due to the population structure. For each locus we have:

$$\mu_{A1} = \mu_{A2} = \alpha\,, \qquad V_{A1} = V_{A2} = \alpha(1-\alpha), \quad Cov(A_{1j}, A_{2j}) = \delta\,.$$

For two loci j and k on the same chromosome:

$$Cov(A_{1j}, A_{1k}) = \alpha(1-\alpha)(1-r_{ij}) + \delta r_{jk}\,.$$

For two loci j and j' on different chromosomes (but the same haplotype):

$$Cov(A_{1j}, A_{1j'}) = \alpha(1-\alpha)(1-r_U) + \delta r_U\,.$$

For loci on different haplotypes:

$$Cov(A_{1j}, A_{2k}) = Cov(A_{1j}, A_{2j'}) = \delta\,.$$

The mean of global ancestry is unaffected by the correlation and is the same as for local ancestry:

$$\mu_{\bar{A}} = \alpha\,.$$

The variance is more complicated:

$$V_{\bar{A}} = Var\left(\frac{\sum_{j=1}^{M} A_{1j} + \sum_{j=1}^{M} A_{2j}}{2M}\right)$$

$$= Var\left(\frac{\sum_{j=1}^{M} A_{1j}}{2M}\right) + Var\left(\frac{\sum_{j=1}^{M} A_{2j}}{2M}\right) + \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} Cov(A_{1i}, A_{2j})\,.$$

Following the same expansion from Appendix A, we can show

$$Var\left(\frac{\sum_{j=1}^{M} A_{1j}}{2M}\right) = Var\left(\frac{\sum_{j=1}^{M} A_{2j}}{2M}\right) = \frac{1}{4}[(\alpha(1-\alpha)-\delta)\varphi + \delta]\,,$$

where $\varphi = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq j}^{m} r_{jk}$ as previously defined. We also easily see because of equality of ancestry covariances between different haplotypes that

$$\frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} Cov(A_{1i}, A_{2j}) = \frac{\delta}{2}.$$

It follows that

$$V_{\bar{A}} = \frac{1}{2}[(\alpha(1-\alpha) - \delta)\varphi + 2\delta] .$$

The population membership variable, S, is binomial with

$$\mu_S = Q \qquad \text{and} \qquad V_S = Q(1-Q) .$$

B.1.3 Covariances

Next, we derive the covariances between the random variables. The covariance between the trait and QTL genotype is

$$C_{YG} = aV_G + cC_{GS} .$$

With some algebra we can show that

$$C_{GS} = 2Q(1-Q)(q_1 - q_0)\Delta ,$$

which gives

$$C_{YG} = 2[a(p'(1-p') + \delta\Delta^2) + 2cQ(1-Q)(q_1 - q_0)\Delta] .$$

To derive the covariance between the QTL genotype and the local ancestry at the QTL write G and A in terms of the two haplotypes as in Appendix A. It follows that

$$C_{GA} = Cov\left((X_1 + X_2 - 1), \frac{A_1 + A_2}{2}\right)$$

$$= \frac{1}{2}\big(Cov(X_1, A_1) + Cov(X_2, A_2) + Cov(X_1, A_2) + Cov(X_2, A_1)\big) .$$

It is not hard to show that

$$Cov(X_1, A_1) = Cov(X_2, A_2) = \alpha(1-\alpha)\Delta$$

and

$$Cov(X_1, A_2) = Cov(X_1, A_2) = \delta\Delta ,$$

which gives

$$C_{GA} = (\alpha(1-\alpha) + \delta)\Delta$$

and

$$\begin{aligned} C_{YA} &= aC_{GA} + cC_{SA} \\ &= a(\alpha(1-\alpha) + \delta)\Delta + cQ(1-Q)(q_1 - q_0)\,. \end{aligned}$$

For covariances with global ancestry, we assume, without loss of generality, that the QTL is at the gth locus on the first chromosome. Then

$$\begin{aligned} C_{G\bar{A}} &= \frac{1}{M}\sum_{j=1}^{M} Cov(G, A_{.j}) \\ &= \frac{1}{M}[C_{GA} + \sum_{j\neq g}^{M} Cov(G, A_{.j})]\,. \end{aligned}$$

The second term sums over loci on the same chromosome as the QTL and different chromosomes from the QTL. For locus j on the same chromosome with the QTL,

$$Cov\big(G, A_{.j}\big) = (\alpha(1-\alpha) + \delta)\Delta\big(1 - r_{gj}\big) + 2\delta\Delta r_{gj}.$$

For locus j on a different chromosome from the QTL,

$$Cov\big(G, A_{.j}\big) = (\alpha(1-\alpha) + \delta)\Delta(1 - r_u) + 2\delta\Delta r_u.$$

This gives

$$\begin{aligned} C_{G\bar{A}} &= (\alpha(1-\alpha) + \delta)\Delta\varphi_g + 2\delta\Delta\big(1 - \varphi_g\big) \\ &= (\omega\varphi_g + 2\delta)\Delta\,. \end{aligned}$$

The covariance between the trait and global ancestry is a function of the covariance between the QTL genotype and global ancestry:

$$C_{Y\bar{A}} = aC_{G\bar{A}} + cC_{S\bar{A}}\,.$$

It is not hard to show that $C_{S\bar{A}} = Q(1-Q)\,(q_1 - q_0)$, which gives:

$$C_{Y\bar{A}} = a\left[(\alpha(1-\alpha)+\delta)\Delta\varphi_g + 2\delta\Delta\left(1-\varphi_g\right)\right] + cQ(1-Q)(q_1-q_0)$$

$$= a(\omega\varphi_g + 2\delta)\Delta + cQ(1-Q)(q_1-q_0)\,.$$

B.1.4 Correlations

With variances and covariances derived, we now derive relevant correlations to assess confounding. First, we consider local ancestry. From the main text, we saw that we need to examine two correlations to assess confounding, ρ^2_{GA} and $\rho^2_{Y\bar{A},G}$. It is straightforward to show

$$\rho^2_{GA} = \frac{(\alpha(1-\alpha)+\delta)\Delta^2}{(\alpha(1-\alpha)+\delta)\Delta^2+\gamma'}$$

$$= \frac{(\omega+2\delta)\Delta^2}{(\omega+2\delta)\Delta^2+\gamma'}\,.$$

The partial correlation is

$$\rho^2_{Y\bar{A},G} = \frac{(C_{YA}V_G - C_{YG}C_{GA})^2}{(V_Y V_G - C^2_{YG})(V_A V_G - C^2_{GA})}$$

$$= \frac{2c^2Q(1-Q)\delta\gamma'}{2c^2Q(1-Q)\delta\gamma' + \left((\omega+2\delta)\Delta^2+\gamma'\right)(c^2Q(1-Q)\omega + (\omega+2\delta)\sigma^2)}\,.$$

For global ancestry, the correlations are the following:

$$\rho^2_{G\bar{A}} = \frac{\left(\omega\varphi_g+2\delta\right)^2\Delta^2}{[(\omega+2\delta)\Delta^2+\gamma'](\omega\varphi+2\delta)}$$

$$\rho^2_{Y\bar{A},G}$$

$$= \frac{2c^2Q(1-Q)\delta[\omega\Delta^2(1-\varphi_g)+\gamma']^2}{\left[(\omega\varphi+2\delta)[(\omega+2\delta)\Delta^2+\gamma'] - \Delta^2\left(\omega\varphi_g+2\delta\right)^2\right]\left[c^2Q(1-Q)(\omega\Delta^2+\gamma') + \sigma^2[(\omega+2\delta)\Delta^2+\gamma']\right]}\,.$$

For population membership, we have

$$\rho_{GS}^2 = \frac{2\Delta^2\delta}{2\Delta^2\delta + (\omega\Delta^2 + \gamma')}$$

$$\rho_{YS,G}^2 = \frac{c^2Q(1-Q)(\omega\Delta^2+\gamma')}{c^2Q(1-Q)(\omega\Delta^2+\gamma') + [(\omega+2\delta)\Delta^2 + \gamma']\sigma^2}.$$

B.1.5 Regression Parameters

We first consider the unadjusted from Appendix A.1.5. Given the calculations in B.1.3, we have

$$\beta_G = \frac{C_{YG}}{V_G}$$

$$= a + \frac{cQ(1-Q)(q_1-q_0)\Delta}{(\omega+2\delta)\Delta^2+\gamma'}.$$

For the model adjusted by local ancestry, we have the unstandardized multiple regression coefficient for genotype:

$$\beta_G^* = \frac{C_{YG}V_A - C_{YA}C_{GA}}{(V_GV_A - C_{GA}^2)}$$

$$= \frac{(aV_G + cC_{GS})V_A - (aC_{GA}+cC_{SA})C_{GA}}{(V_GV_A - C_{GA}^2)}$$

$$= a + \frac{c(C_{GS}V_A - C_{SA}C_{GA})}{(V_GV_A - C_{GA}^2)}.$$

Recognizing that $C_{GS} = 2\Delta C_{SA}$ and $C_{GA} = 2\Delta V_A$, shows that the second term is 0, and thus $\beta_G^* = a$.

For the model adjusted by global ancestry,

$$\beta_G' = a + c\frac{Q(1-Q)(q_1-q_0)\omega\Delta\left(\varphi-\varphi_g\right)}{\left[(\omega\varphi+2\delta)(\omega+2\delta)-\left(\omega\varphi_g+2\delta\right)^2\right]\Delta^2+\gamma'(\omega\varphi+2\delta)}$$

$$= a + c\frac{Q(1-Q)(q_1-q_0)\omega\Delta\dfrac{(\varphi-\varphi_g)}{(\omega\varphi+2\delta)}}{\left[(\omega+2\delta)-\dfrac{\left(\omega\varphi_g+2\delta\right)^2}{(\omega\varphi+2\delta)}\right]\Delta^2+\gamma'} \quad .$$

For the model adjusted for membership: $E(Y|G,S) = \beta_0'' + \beta_G''G + \beta_S''S$, it is not hard to show that

$$\beta_G'' = \frac{C_{YG}V_S - C_{YS}C_{GS}}{(V_GV_S - C_{GS}^2)}$$

$$= \frac{(aV_G + cC_{GS})V_S - (aC_{GS}+cV_S)C_{GS}}{(V_GV_S - C_{GS}^2)}$$

$$= a\,.$$

Appendix B.2. Variances, covariances, correlations and regression parameters for testing at a marker in a stratified admixed population model (2 strata).

B.2.1 Definitions

Consider the stratified population defined in Appendix B.1 but now suppose that we are testing a marker locus that is not the QTL. Parameters for the QTL are defined in Appendix B.1. As in Appendix A.2, we let the marker locus have two alleles, L_1, L_2, with

p_{Li}=the frequency of allele L_1 in ancestral population i, for $i = 0,1$.

and let $\Delta_L = p_{L1} - p_{L0}$.

The allele frequency at the marker in subpopulation k is

$$p^*_{Lk} = q_k p_{L1} + (1 - q_k) p_{L0}$$

and the allele frequency in the stratified population as a whole is

$$p^{**}_L = Q p^*_{L1} + (1 - Q) p^*_{L0}$$
$$= \alpha p_{L1} + (1 - \alpha) p_{L0} ,$$

where $\alpha = Q q_1 + (1 - Q) q_0$ as before.

The disequilibrium coefficient in ancestral population i,D_i, is defined in Appendix A.2.Following the derivations in Appendix A.2, we can write the disequilibrium coefficient in the kth admixed subpopulation ($j = 0,1$) as:

$$D^*_k = (1 - r_{gl})(W_k + q_k(1 - q_k)\Delta\Delta_L) ,$$

where $W_k = q_k D_1 + (1 - q_k) D_0$. Finally considering the stratified population as a whole, we can write the disequilibrium coefficient as:

$$D^{**} = Q D^*_1 + (1 - Q) D^*_0 + Q(1 - Q)(q_1 - q_0)^2 \Delta\Delta_L$$

$$= \left(1 - r_{gl}\right)(W^* + \omega\Delta\Delta_L) + \delta\Delta\Delta_L \,,$$

where $W^* = QW_1 + (1 - Q)W_0$ and other parameters are defined in previous appendices.

Finally, we assume that the trait follows the same conditional normal distribution defined in Appendix B.1.1 dependent on the QTL genotype and the subpopulation membership.

B.2.2 Means and Variances

The mean and variance for the genotype random variable at the marker, L, are analogous to those derived for the QTL in B.1.2:

$$\mu_L = 2p'_L - 1 \qquad \text{and} \qquad V_L = 2p'_L(1 - p'_L) + 2\Delta_L^2\delta$$

$$= 2(p_{L0} + \Delta_L\alpha) - 1 \qquad = 2[\gamma'_L + (\alpha(1 - \alpha) + \delta)\Delta_L^2] \,,$$

where $\gamma'_L = \alpha p_{L1}(1 - p_{L1}) + (1 - \alpha)p_{L0}(1 - p_{L0})$.

The means and variances for the trait, Y, and global ancestry, $\bar{A}$, are the same as derived in Appendix B.1.2.

The mean and variance for the measure of local ancestry at the marker, $A_{\cdot l}$, are the same as for the QTL:

$$\mu_{A_l} = \alpha \qquad \text{and} \qquad V_{A_l} = \frac{[\alpha(1 - \alpha) + \delta]}{2}.$$

B.2.3 Covariances

To derive the covariance between the QTL and the marker genotype random variables, we write G and L as a sum of Bernoulli random variables as in A.2.3:

$$G = X_1 + X_2 - 1 \text{ and } L = X_{L1} + X_{L2} - 1,$$

where X_1 and X_{L1} and Bernoulli random variables with probabilities p^* and p_L^*, respectively. It follows that

$$C_{GL} = Cov(X_1, X_{L1}) + Cov(X_2, X_{L2}) + Cov(X_1, X_{L2}) + Cov(X_2, X_{L1}) .$$

The first two terms represent the covariance between alleles on the same haplotype and is equivalent to the disequilibrium coefficient defined in B.2.1.

$$Cov(X_1, X_{L1}) = Cov(X_2, X_{L2}) = D^{**} .$$

The second two terms represent the covariance between two alleles on different haplotypes in the same individual. We can show that

$$Cov(X_1, X_{L2}) = Cov(X_2, X_{L1}) = \delta\Delta\Delta_L ,$$

which gives

$$C_{GL} = 2(D^{**} + \delta\Delta\Delta_L) .$$

To derive the covariance between the trait and marker genotype, we write $Y = aG + cS + N(0, \sigma^2)$, which implies that $C_{YL} = aCov(G, L) + cCov(S, L)$. It is not difficult to show that $Cov(S, L) = 2Q(1 - Q)(q_1 - q_0)\Delta_L$ so that

$$C_{YL} = 2a(D^{**} + \delta\Delta\Delta_L) + 2cQ(1 - Q)(q_1 - q_0)\Delta_L .$$

The covariance between the marker genotype and local ancestry at the marker is the same as for the QTL, replacing QTL parameters with marker parameters:

$$C_{LA_l} = (\omega + 2\delta)\Delta_L .$$

The same is true for global ancestry:

$$C_{L\bar{A}} = (\omega\varphi_l + 2\delta)\Delta_L ,$$

where $\varphi_l = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{M}\sum_{j \neq l}^{m} r_{lj}$ (the final term is summed over all markers on the same chromosome with the marker being tested).

Finally, we can show that the covariance between the QTL genotype and the local ancestry at the marker is

$$C_{GA_l} = \left[\omega(1 - r_{gl}) + 2\delta\right]\Delta \, .$$

Writing $Y = aG + cS + N(0, \sigma^2)$ as before, it then follows that $C_{YA_l} = aCov(G, A_l) + cCov(S, A_l)$. The first term is defined immediately above and the second term is the same as $Cov(S, A)$ derived in B.1.3.

$$C_{YA_l} = a\left[\omega(1 - r_{gl}) + 2\delta\right]\Delta + cQ(1 - Q)(q_1 - q_0) \, .$$

B.2.4 Correlations

We now derive the correlations necessary to assess confounding and validity. For local ancestry, we examine the following two correlations to assess confounding:

$$\rho^2_{LA} = \frac{(\omega + 2\delta){\Delta_L}^2}{(\omega + 2\delta){\Delta_L}^2 + \gamma'_L} \, .$$

The partial correlation is more complicated. After some algebra, we can show that the numerator of the partial correlation is

$$numerator\left(\rho^2_{YA,L}\right) \propto \Phi^2 \, ,$$

where

$$\Phi = a\left\{\left(1 - r_{gl}\right)(\omega + 2\delta)(\Delta\gamma'_L - W^*\Delta_L) + 2\delta\Delta r_{gl}\gamma'_L\right\} + cQ(1 - Q)(q_1 - q_0)\gamma'_L \, .$$

For global ancestry, we can show

$$\rho^2_{L\bar{A}} = \frac{\left((\omega\varphi_l + 2\delta)\Delta_L\right)^2}{2[(\alpha(1-\alpha)+\delta)\Delta_L^2 + \gamma'_L]\frac{1}{2}[(\alpha(1-\alpha)-\delta)\varphi + 2\delta]}$$

$$= \frac{\left((\omega\varphi_l + 2\delta)\Delta_L\right)^2}{\left[(\omega + 2\delta)\Delta_L{}^2 + \gamma'_L\right][\omega\varphi + 2\delta]} .$$

The numerator of $\rho^2_{Y\bar{A},L}$ becomes

$$numerator(\rho^2_{Y\bar{A},L}) \propto (\Phi + \Upsilon)^2 ,$$

where

$$\Upsilon = 2\omega\Big\{a\Big\{\big(1 - r_{gl}\big)(W^*\Delta_L(1-\varphi_l) - \Delta\gamma'_L) + \Delta\gamma'_L\varphi_g$$

$$+ \Delta\Delta_L^2\big\{r_{gl}(\omega\varphi_l + 2\delta) + (\varphi_g - \varphi_l)(\omega + 2\delta)\big\}\Big\}$$

$$- [cQ(1-Q)(q_1 - q_0)\Delta_L^2][(\varphi_l - 1)]\Big\} .$$

The denominator is more complex and is not necessary to evaluate confounding, since we need only evaluate whether the numerator is 0, so we do not derive the denominator here.

B.2.5 Regression Parameters

Regression parameters for adjusted and unadjusted models can be derived as a function of the correlation coefficients. For the unadjusted model we have

$$\beta_L = \frac{C_{YL}}{V_L}$$

$$= \frac{a(D^{**} + \delta\Delta\Delta_L) + cQ(1-Q)(q_1 - q_0)\Delta_L}{(\omega + 2\delta)\Delta_L^2 + \gamma'_L} .$$

For the model adjusted for local ancestry we have

$$\beta_L^* = \frac{C_{YL}V_A - C_{YA}C_{LA}}{(V_L V_A - C_{LA}^2)}$$

$$= \frac{a\left[D^{**} + \Delta\Delta_L\left(\omega\left(1 - r_{gl}\right) + \delta\right)\right]}{\gamma_L'}$$

$$= \frac{a(1 - r_{gl})W^*}{\gamma_L'} .$$

For the model adjusted for global ancestry the genotype coefficient is

$$\beta_L' = \frac{C_{YL}V_{\bar{A}} - C_{Y\bar{A}}C_{L\bar{A}}}{\left(V_L V_{\bar{A}} - C_{L\bar{A}}^2\right)}$$

$$= \frac{a\left[(D^{**} + \delta\Delta\Delta_L)(\omega\varphi + 2\delta) - \Delta\Delta_L\left(\omega\varphi_g + 2\delta\right)(\omega\varphi_l + 2\delta)\right] + cQ(1-Q)(q_1 - q_0)\Delta_L\omega(\varphi - \varphi_l)}{\Delta_L^2[(\omega + 2\delta)(\varphi\omega + 2\delta) - (\varphi_l\omega + 2\delta)^2] + \gamma_L'(\varphi\omega + 2\delta)}$$

$$= \frac{a\left[(D^{**} + \delta\Delta\Delta_L) - \Delta\Delta_L\frac{\left(\omega\varphi_g + 2\delta\right)(\omega\varphi_l + 2\delta)}{(\omega\varphi + 2\delta)}\right] + cQ(1-Q)(q_1 - q_0)\Delta_L\omega\frac{(\varphi - \varphi_l)}{(\omega\varphi + 2\delta)}}{\Delta_L^2\left[(\omega + 2\delta) - \frac{(\omega\varphi_l + 2\delta)^2}{(\omega\varphi + 2\delta)}\right] + \gamma_L'} .$$

Finally, for the model adjusted for subpopulation membership:

$$E(Y|L,S) = \beta_{L0}'' + \beta_L''L + \beta_{LS}''S ,$$

we have

$$\beta_L'' = \frac{C_{YL}V_S - C_{YS}C_{LS}}{(V_L V_S - C_{LS}^2)}$$

$$= \frac{a\left(1 - r_{gl}\right)(W^* + \omega\Delta\Delta_L)}{\omega\Delta_L^2 + \gamma_L'}$$

$$= \frac{a(D^{**} - \delta\Delta\Delta_L)}{\omega\Delta_L^2 + \gamma_L'} .$$

Appendix B.3. Generalization of variances, covariances, correlations and regression parameters for testing at the QTL in a stratified admixed population model with >2 strata.

B.3.1 Definitions

Here we extend the definitions in Appendix B.1.1 to >2 subpopulations. Consider a stratified population that is composed of Π admixed subpopulations, where each admixed subpopulation is derived from two ancestral populations (ancestry 0 and 1). As before, we assume random mating within admixed subpopulations but not between. We define the following probabilities:

Q_s =the probability that a random individual from the whole population is from subpopulation s.

q_s = probability a locus is from ancestral population 1, for subpopulation $s = 0, \dots, \Pi - 1$.

Consider the QTL defined previously with alleles T_1, T_2 and with

p_i=the frequency of allele T_1 from ancestral population i, for $i = 0,1$. It follows that the allele frequency in subpopulation s is

$$p_s^* = q_s p_1 + (1 - q_s) p_0;$$

and the allele frequency in the stratified population as a whole is

$$p' = \sum_{s=0}^{\Pi-1} Q_s p_s^*$$

$$= \alpha p_1 + (1 - \alpha) p_0 ,$$

where $\alpha = \sum_{s=0}^{\Pi-1} Q_s\, q_s$ is the probability that a haplotype (at any specific position) sampled from the stratified population is from ancestral population 1.

We define $\Delta = p_1 - p_0$ and $\gamma' = \alpha p_1(1-p_1) + (1-\alpha)p_0(1-p_0)$ as before, and generalize the following definitions:

$$\delta = \sum_{s=0}^{\Pi-1} Q_s\,(q_s - \alpha)^2$$

$\omega = \alpha(1-\alpha) - \delta = \sum_{s=0}^{\Pi-1} Q_s\, q_s(1-q_s)$.

The random variables for genotypic values (G), local ancestry (A) and global ancestry ($\bar{A}$) are defined as before. We extend the trait distribution as follows for subpopulation s:

$$Y \sim \begin{cases} N(a + c_s, \sigma^2) & \text{if } T_1T_1 \\ N(c_s, \sigma^2) & \text{if } T_1T_2 \\ N(-a + c_s, \sigma^2) & \text{if } T_2T_2 \end{cases},$$

where c_s is a constant shift in the trait mean for the *s*th subpopulation. For our previous example (Appendix B.1) with two subpopulations $c_0 = 0$ and $c_1 = c$.

B.3.2 Means and Variances

The mean and variance of the genotype variable G take the same form as before:

$$\mu_G = 2(p_0 + \Delta\alpha) - 1 \qquad \text{and} \qquad V_G = 2[\gamma' + (\omega + 2\delta)\Delta^2]$$

If we let $S_s = 1$ if a random individual is from subpopulation s and 0 otherwise, then we can write $Y = aG + \sum_{s=0}^{\Pi-1} c_s S_s + N(0, \sigma^2)$. This gives

$$\mu_Y = a\mu_G + \sum_{s=0}^{\Pi-1} c_s\, Q_s$$

and

$$V_Y = a^2 V_G + \sum_{s=0}^{\Pi-1} c_s^2 Q_s(1-Q_s) - \sum_{s=0}^{\Pi-1}\sum_{j\neq s} c_j c_s Q_j Q_s + 4a\Delta \sum_{s-0}^{\Pi-1} c_s Q_s\,(q_s - \alpha) + \sigma^2$$

The means and variances for local and global ancestry take the same forms as before, with the extended parameter definitions above:

$$\mu_A = \alpha \qquad \text{and} \qquad V_A = \frac{[\omega + 2\delta]}{2},$$

$$\mu_{\bar{A}} = \alpha \qquad \text{and} \qquad V_{\bar{A}} = \frac{[\omega\varphi + 2\delta]}{2}.$$

Note that we use the relationship $\alpha(1-\alpha) + \delta = \omega + 2\delta$ to make the expressions more uniform, but it is not hard to show that these reduce to the same forms in the previous appendices.

Finally, the population membership variable, $\boldsymbol{S} = (S_0, \ldots S_{\Pi-1})$, is multinomial with

$$\mu_S = (Q, \ldots Q_{\Pi-1}) \qquad \text{and} \qquad V_{S_i} = Q_i(1-Q_i)\,, Cov\big(S_i, S_J\big) = -Q_i Q_j.$$

B.3.3 Covariances

The covariances involving the trait random variable for this generalized model change form as follows:

$$C_{YG} = aV_G + 2\Delta \textstyle\sum_{s=0}^{\Pi-1} c_s Q_s(q_s - \alpha)\,,$$

$$C_{YA} = a(\omega + 2\delta)\Delta + \sum_{s=0}^{\Pi-1} c_s Q_s(q_s - \alpha)\,,$$

$$C_{Y\bar{A}} = a(\omega\varphi_g + 2\delta)\Delta + \sum_{s=0}^{\Pi-1} c_s Q_s(q_s - \alpha)\ .$$

The forms of the covariances between genotype and ancestry variables remain the same as derived before:

$$C_{GA} = (\omega + 2\delta)\Delta,$$

$$C_{G\bar{A}} = (\omega\varphi_g + 2\delta)\Delta\,.$$

B.3.4 Correlations

The correlations necessary to address confounding generalize as follow. For local ancestry,

$$\rho^2_{GA} = \frac{(\omega + 2\delta)\Delta^2}{(\omega + 2\delta)\Delta^2 + \gamma'}\,.$$

$$numerator(\rho^2_{YA,G}) \propto 2\gamma' \left[\sum_{s=0}^{\Pi-1} c_s Q_s (q_s - \alpha)\right]^2 .$$

For global ancestry, the correlations are the following:

$$\rho^2_{G\bar{A}} = \frac{\left(\omega\varphi_g + 2\delta\right)^2\Delta^2}{[(\omega + 2\delta)\Delta^2 + \gamma'](\omega\varphi + 2\delta)}$$

$$numerator(\rho^2_{Y\bar{A},G}) \propto 2\left[\sum_{s=0}^{\Pi-1} c_s Q_s (q_s - \alpha)\right]^2 [\omega\Delta^2(1 - \varphi_g) + \gamma']^2\,.$$

As in the case of two strata, these equations show that ancestry (global and local) is generally a confounder unless $\Delta = 0$, $c_i = c_j\ \forall i, j$ or $q_i = q_j\ \forall i, j$.

B.3.5 Regression Parameters

The regression parameters also generalize as we would expect: For the unadjusted model, we have

$$\beta_G = a + \frac{\Delta \sum_{s=0}^{\Pi-1} c_s Q_s (q_s - \alpha)}{(\omega + 2\delta)\Delta^2 + \gamma'} \ .$$

For the model adjusted by local ancestry, we have

$$\beta_G^* = a$$

For the model adjusted by global ancestry,

$$\beta_G' = a + \frac{\omega\Delta\left(\varphi - \varphi_g\right) \sum_{s=0}^{\Pi-1} c_s Q_s (q_s - \alpha)}{\left[(\omega\varphi + 2\delta)(\omega + 2\delta) - \left(\omega\varphi_g + 2\delta\right)^2\right]\Delta^2 + \gamma'(\omega\varphi + 2\delta)}$$

These formulas lead to the same conclusion as the two-strata case that only the local-ancestry adjusted model estimates the genetic effect a. Notably, though if $\varphi \approx \varphi_g$, the bias in the global-ancestry adjusted model becomes small.

It is not as simple to derive the regression coefficient for the model adjusted for subpopulation membership because this requires partialling out multiple indicator variables for each $\Pi - 1$ strata. Intuitively however, it makes sense that with full adjustment for membership, the genotype term will estimate the true genetic effect a as it does in the two-stratum case.

Appendix C. Derivations for φ and φ_l.

C.1 We begin with the function φ:

$$\varphi = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{Mm} \sum_{j=1}^{m} \sum_{k \neq i}^{m} r_{jk} \,,$$

where r_{jk} is the recombination probability between the jth and kth loci on the same chromosome and r_U is the recombination probability between loci on different chromosomes. Specifically, in our context the recombination probability is the probability that on a random haplotype there has been at least one recombination event since the initial migration (admixture). For example, with a single generation of random mating following migration, the recombination probability is the same as the usual recombination fraction (the per-meiosis probability of recombination); the recombination probability that we define captures cumulative recombination over generations. We assume the number of chromosomes in K and the number of markers is the same for each chromosome, m, with $M = mK$. Let us assume that the expected number of recombination events on a chromosome is

$$s = (m-1)R \,,$$

where $R = r_{j,j+1}$ is the probability that a recombination event has occurred between two adjacent loci (assumed to be the same for all adjacent loci). Assuming independence of recombination events, we can then write:

$$r_{jk} = 1 - (1-R)^{|k-j|}$$

$$= 1 - \left(1 - \frac{s}{m-1}\right)^{|i-j|}.$$

We note that the term $\frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq j}^{m} r_{jk}$ is $1/K$ times the average of all pairwise recombination probabilities. With some algebra and using the properties of a geometric series, $\sum_{i=1}^{n} x^i = x\left(\frac{1-x^n}{1-x}\right)$, we can derive the following form:

$$\frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq j}^{m} r_{ij} = \frac{2}{Mm}\sum_{j=1}^{m-1}\sum_{k=1}^{m-i}(1-(1-R)^k)$$

$$= \frac{m-1}{M} - \frac{2}{Mm}\sum_{j=1}^{m-1}\sum_{k=1}^{m-i}(1-R)^k$$

$$= \left(\frac{m-1}{M}\right) - \frac{2}{Mm}\sum_{j=1}^{m-1}\left(\frac{1-R}{R}\right)\left(1-(1-R)^{m-j}\right)$$

$$= \left(\frac{m-1}{M}\right) - \frac{2(m-1)}{Mm}\left(\frac{1-R}{R}\right) + \frac{2}{Mm}\left(\frac{1-R}{R}\right)\sum_{j=1}^{m-1}(1-R)^{m-j}$$

$$= \left(\frac{m-1}{M}\right) - \frac{2(m-1)}{Mm}\left(\frac{1-R}{R}\right) + \frac{2}{Mm}\left(\frac{1-R}{R}\right)^2(1-(1-R)^{m-1})$$

$$= \left(\frac{m-1}{M}\right) - \frac{2(1-R)}{MmR^2}(mR-1+(1-R)^m)$$

$$= \frac{1}{K}\left[\frac{m-1}{m} - 2\left(\frac{m-1}{sm}\right)^2\left(\frac{m-1-s}{m-1}\right)\left(\frac{sm}{m-1} - 1 + \left(\frac{m-1-s}{m-1}\right)^m\right)\right].$$

We can simplify this by considering the limiting condition $m \to \infty$ and using the property $\lim_{n\to\infty}(1+\frac{1}{n})^n = e$, which then gives

$$\lim_{m\to\infty}\frac{1}{Mm}\sum_{j=1}^{m}\sum_{k\neq i}^{m} r_{jk} = \frac{1}{K}\left[1-\frac{2}{s^2}(s-1+e^{-s})\right].$$

Finally, for large m, we have

$$\varphi = 1-\left(\frac{K-1}{K}\right)r_U - \frac{1}{K}\left[1-\frac{2}{s^2}(s-1+e^{-s})\right].$$

Supplemental Figure 1 shows how this changes as a function of s for different values of K and r_u. We note φ decreases as s increases, with bounds $\varphi \in \left(\left(\frac{K-1}{K}\right)(1-r_U), 1-\left(\frac{K-1}{K}\right)r_U,\right)$.

C.2 We next consider the function φ_l for a marker at some fixed position l:

$$\varphi_l = 1-\left(\frac{K-1}{K}\right)r_U - \frac{1}{M}\sum_{j\neq l}^{m} r_{lj}\,,$$

where all variables are defined as above. The sum in the final term is the sum of recombination probabilities between the marker locus at position l and all loci on the same chromosome as the marker. We can write

$$\begin{aligned}
\frac{1}{M}\sum_{j\neq l}^{m} r_{lj} &= \frac{1}{M}\left(\sum_{j=1}^{l-1}\left(1-(1-R)^j\right)+\sum_{j=1}^{m-l}\left(1-(1-R)^j\right)\right)\\
&= \left(\frac{m-1}{M}\right)-\frac{1}{M}\left(\sum_{j=1}^{l-1}(1-R)^j+\sum_{j=1}^{m-l}(1-R)^j\right)\\
&= \left(\frac{m-1}{M}\right)-\frac{1}{M}\left(\frac{1-R}{R}\right)\left(\left(1-(1-R)^{l-1}\right)+\left(1-(1-R)^{m-l}\right)\right)\\
&= \frac{1}{K}\left[\left(\frac{m-1}{m}\right)-\left(\frac{m-1-s}{sm}\right)\left(2-\left(\frac{m-1-s}{m-1}\right)^{l-1}-\left(\frac{m-1-s}{m-1}\right)^{m-l}\right)\right].
\end{aligned}$$

To take the limit as $m \to \infty$, we write $l = mf$ where $f \in \left[\frac{1}{m}, 1\right]$ and takes on values such that l is an integer. Then, we have

$$\lim_{m\to\infty} \frac{1}{M} \sum_{j\neq l}^{m} r_{lj} = \frac{1}{K}\left[1 - \frac{1}{s}\left(2 - e^{-sf} - e^{-s(1-f)}\right)\right].$$

Finally, for large m, we have:

$$\varphi_l = 1 - \left(\frac{K-1}{K}\right) r_U - \frac{1}{K}\left[1 - \frac{1}{s}\left(2 - e^{-sf} - e^{-s(1-f)}\right)\right].$$

Supplemental Figure 1 shows a plots of as a function of s for l and the end ($f = 0$) and middle ($f = 1/2$) of the chromosome. The bounds for φ_l are $\left(\left(\frac{K-1}{K}\right)(1 - r_U), 1 - \left(\frac{K-1}{K}\right) r_U, \right)$, the same as for φ and independent of l for large m. The function is maximized when l is in the middle of the chromosome map and minimized when l at either end of the map, with φ falling between these two functions.