psych

Supplementary File 1

Appendix: Admixture Simulations

In the following, the effect of admixture will be simulated, so as to quantify predictions of evolutionary models.

We begin with a simple scenario. Suppose there is a trait which is 100% heritable and which is conditioned by 10 variants of equal effect which can be either A_0 or A_1, with effect sizes of 0 and 1, respectively. Grant further that each variant has a frequency of 50%, which is unrelated to the frequency of the other variants (i.e. complete linkage equilibrium; no correlations between variant frequencies between populations). Additionally, grant complete additivity — that there are no interactions. Furthermore, grant there is just a single long genome with one copy for each locus (haploidity). Finally, suppose we have two populations both with 50,000 people. If we simulate data based on these assumptions and calculate each individual's polygenic score (sum of the variants that they have), then the resulting distributions will look as presented in Figure A1. The resulting distributions are approximately normal (per the central limit theorem). Since we did not introduce any group differences in the assumptions, there are none to be seen aside from minor chance variation.

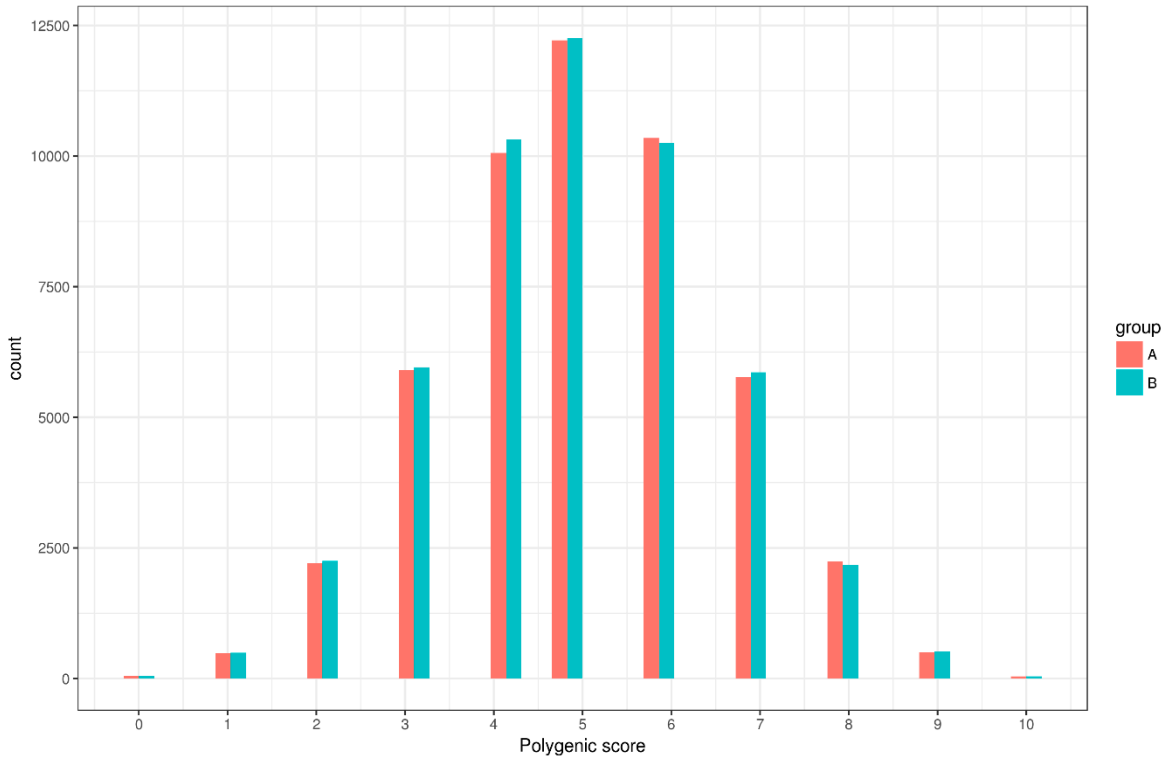

Figure A1. Simulation 1 - distribution of polygenic scores by group

In the above, we implicitly assumed that genotype fully corresponds to phenotype. In real life, however, no trait is exactly 100% heritable (though some come close), and there is always some developmental noise. The amount of noise, as well as other factors not shared by siblings brought up together, collectively form the unshared environment variance component (E) estimated in biometric analysis. We can add this to our simulation by adding a small unshared environment component that consists of a standard normal distribution (mean of 0, standard deviation [sd] of 1). If we run the simulation with this change, the histogram will look as it does in Figure A2. Note that we are now plotting the phenotypic score.

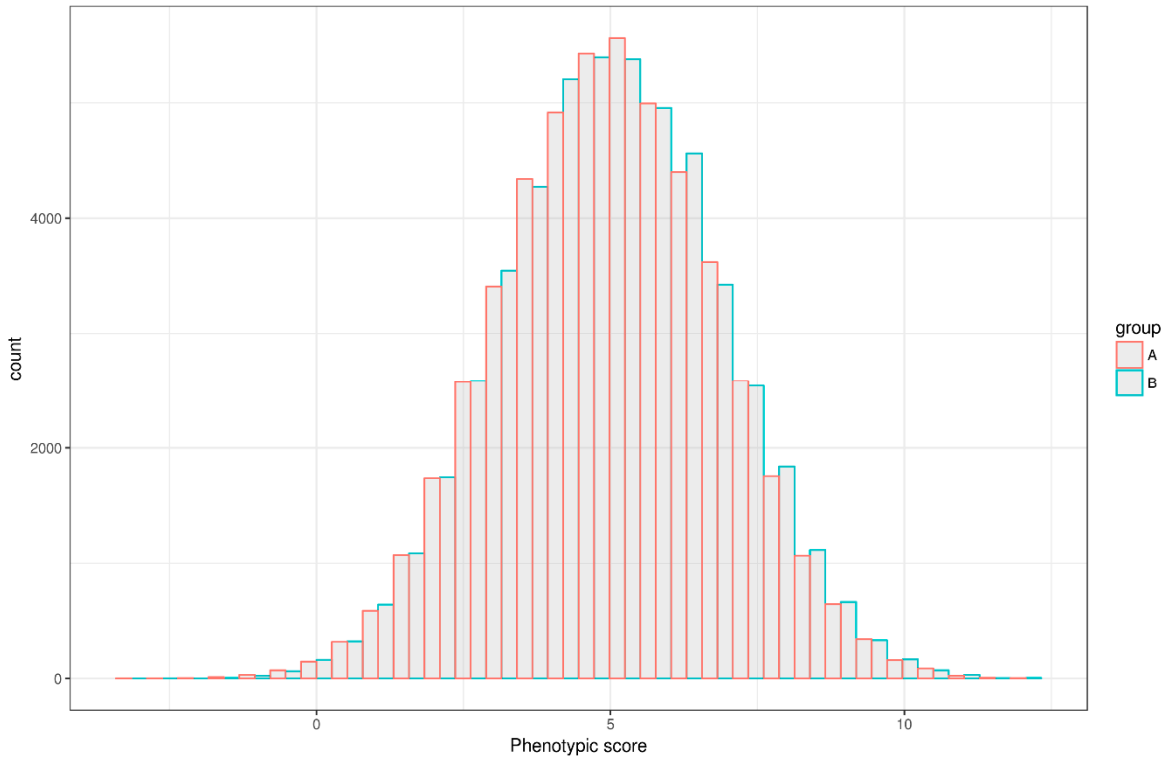

Figure A2. Simulation 2 - histogram of phenotypic scores by group.

Based on these simple assumptions, we have recreated the normal distribution so often seen in real data. However, the distributions have also become complex and not easy to visualize from the histogram. In light of this, we take another step and switch to density fits. This consists of estimating the underlying distribution of values and plotting the distribution instead of the data. The fitted distributions are shown in Figure A3.

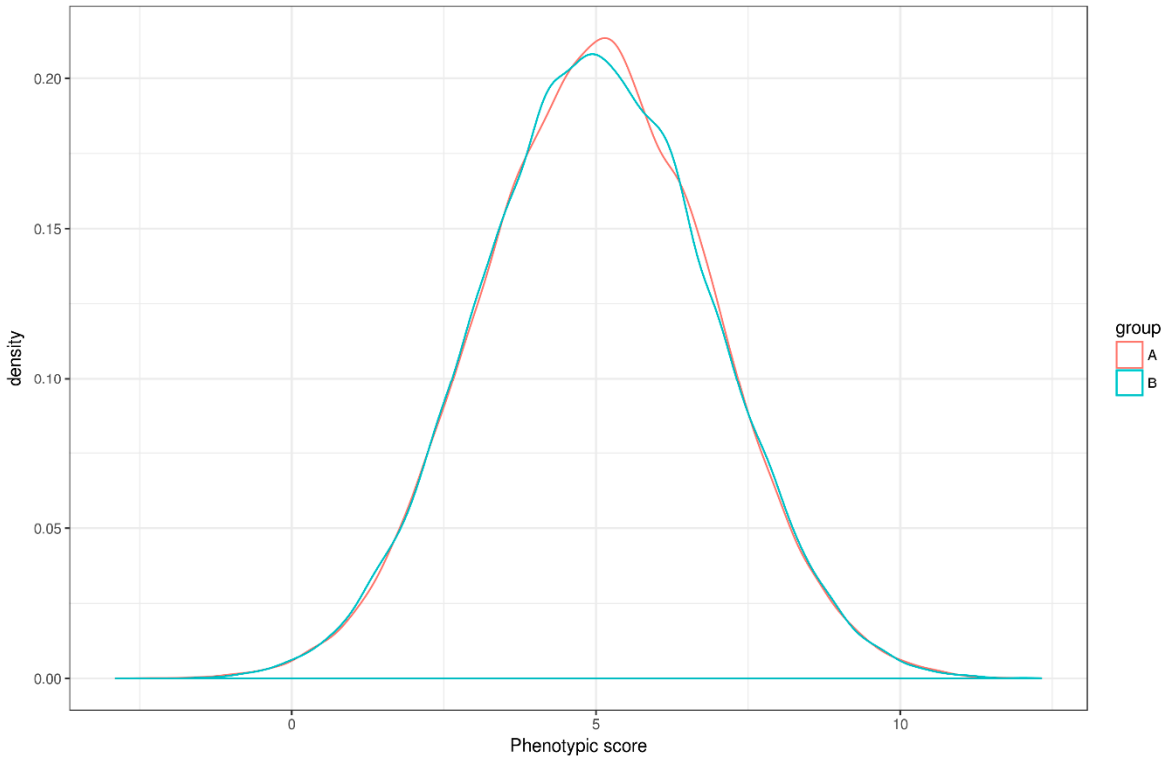

Figure A3. Simulation 2 - distribution of phenotypic scores by group.

We can now more clearly see the underlying distributions. Because we introduced a noise component, the polygenic scores are no longer identical with the phenotypic scores and do not correlate perfectly with them. The observed regression lines are of course identical by group, and the corresponding correlation is 0.85.

So far we have been assuming that there are no non-chance group differences in their polygenic scores. However, in real life this is unlikely. A group difference in polygenic scores consists of differences in the frequencies of one of more of the variants. To show this, we change the frequencies of the variants to 0.55 and 0.45 for group A and B, respectively. The resulting distributions are shown in Figure A4.

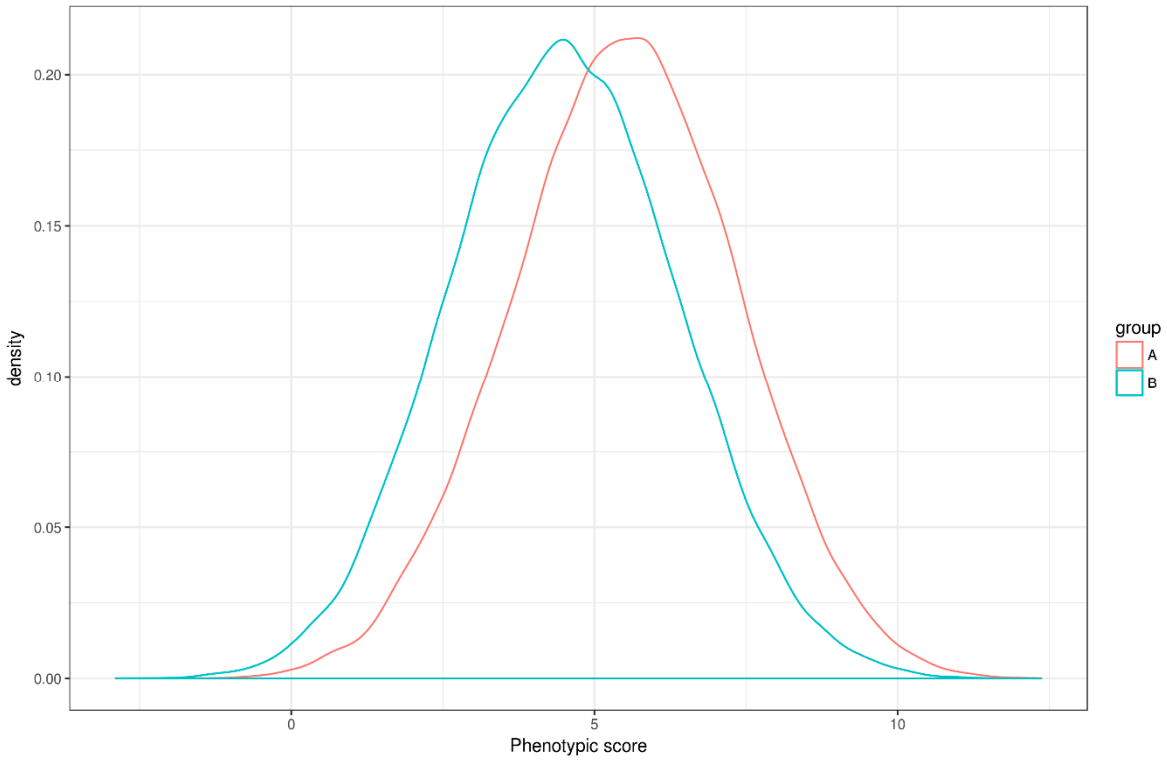

Figure A4. Simulation 3 - Distributions of phenotypic scores by group.

As expected, we now see a difference in the phenotypic scores as a result of the genetic difference we created. Unsurprisingly, if we calculate the mean polygenic scores by group, these are 5.5 and 4.5 for A and B respectively (number of variants times their probability by group). In terms of effect size, the phenotypic gap is 0.64 d (standardized) or 1 in raw units, representing a medium to large gap by conventional standards.

If we repeat the scatterplots that were done in simulation 3, we still find that the lines are identical, as seen in Figure A5.

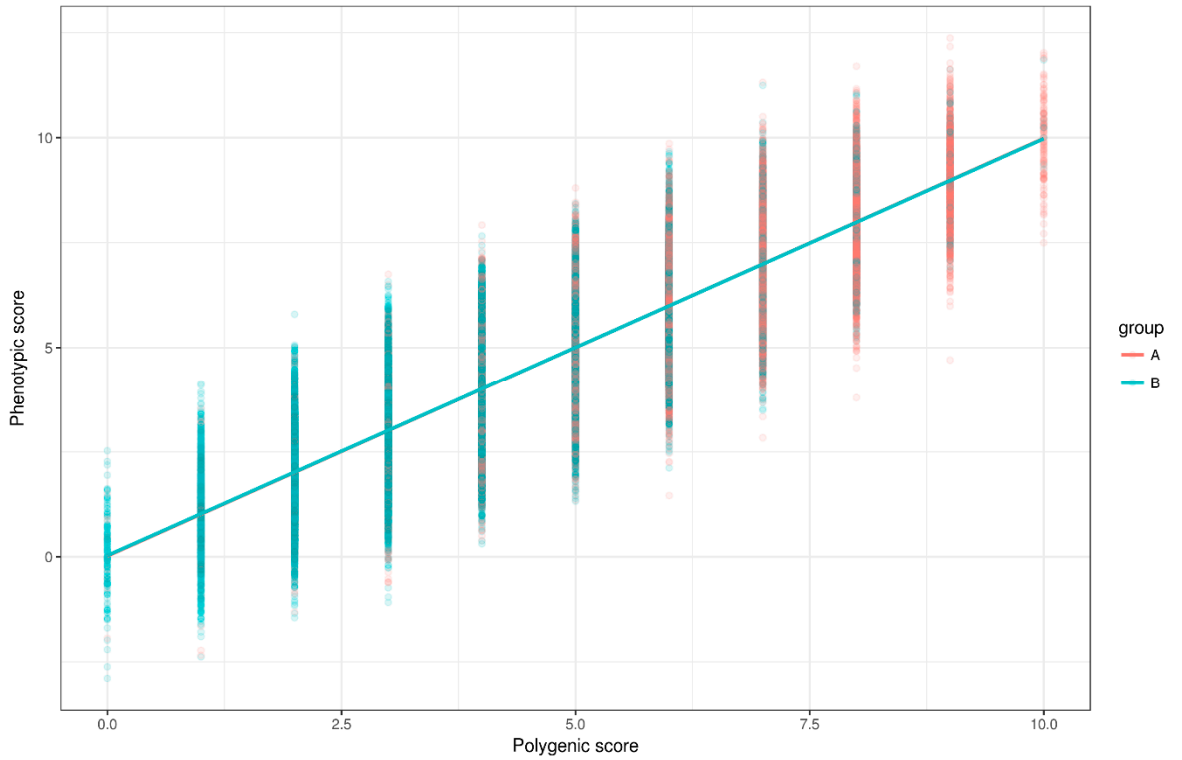

Figure A5. Simulation 3 - Scatterplot of polygenic and phenotypic scores by group.

The absence of any difference in slope or intercept indicates that the group difference was completely due to genetic effects, or in other words, that the between group heritability was 100%. This is despite the fact that the within group heritability was not 100%. If we model the data, we see that group membership becomes irrelevant once we take polygenic scores into account, as seen in Table A1. In other words, controlling for true cause, the non-causal correlate no longer had any validity.

Table A1. Simulation 3 - Models of group differences. Unstandardized betas.

Model	Group membership	Polygenic score
1	1.00	
2	.00	1.00

Individual differences in many traits are likely to be multifactorial and to include common environmental effects. We can introduce individual variation into our simulation by adding another normally distributed environmental factor (analogous to shared environment in biometric analysis). Note that for it to contribute to the group difference, the distributions must have different means by group. In our case, we use means of 1 and 0 for A and B, respectively. Figure A6 shows the phenotypic distributions.

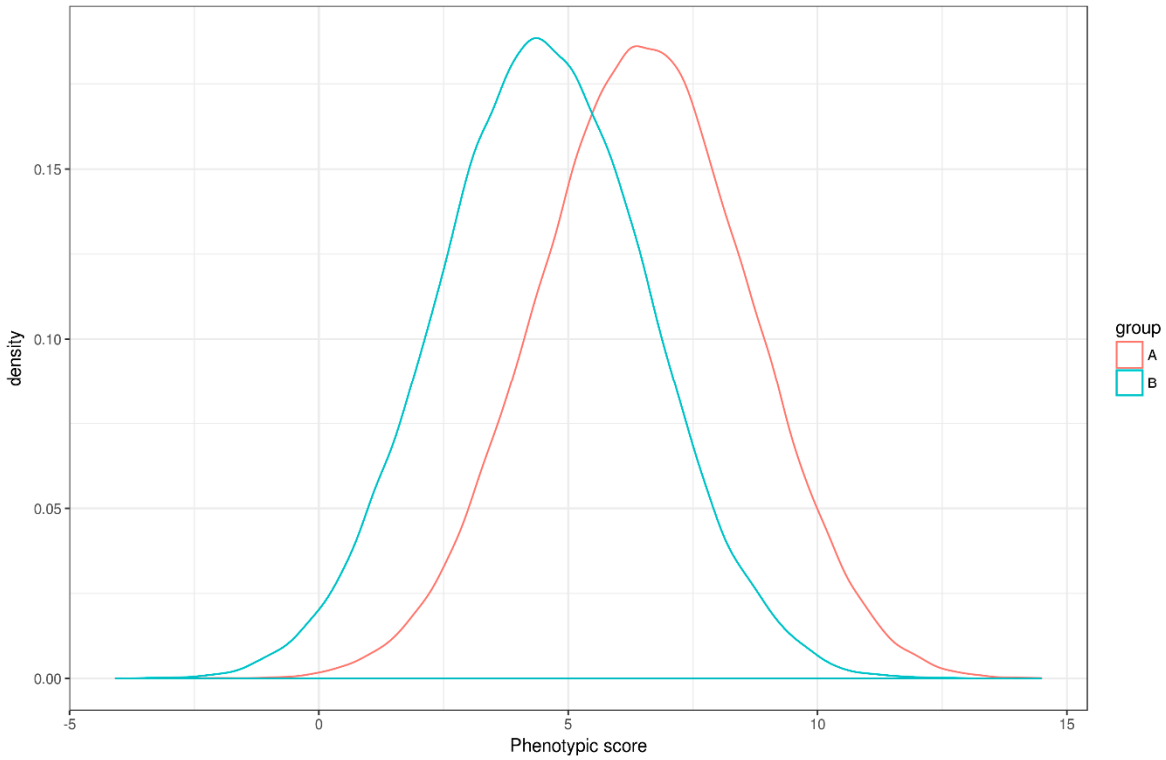

Figure A6. Simulation 4 - Distribution of phenotypic scores by group.

The group difference has increased due to the addition of this environmental factor, and is now 2 in raw units. The standardized difference, however, remains at 0.64 d because the standard deviation (pooled) also increased. Table A2 shows the correlations between the variables in the simulation.

Table A2. Simulation 4 - Correlations between variables.

	Polygenic score	**Shared environment**	**Unshared environment**	**Phenotypic score**
Polygenic score	1.00	0.14	0.00	0.77
Shared environment	0.14	1.00	0.01	0.58
Unshared environment	0.00	0.01	1.00	0.43
Phenotypic score	0.77	0.58	0.43	1.00

There is now a small positive correlation, r = 14, between polygenic score and shared environment, or in other words, there is a gene-environment correlation (rGE). This results from the common cause, namely group membership. Scatterplots of the relationships between the two causes and the phenotypic scores are shown in Figure A7.

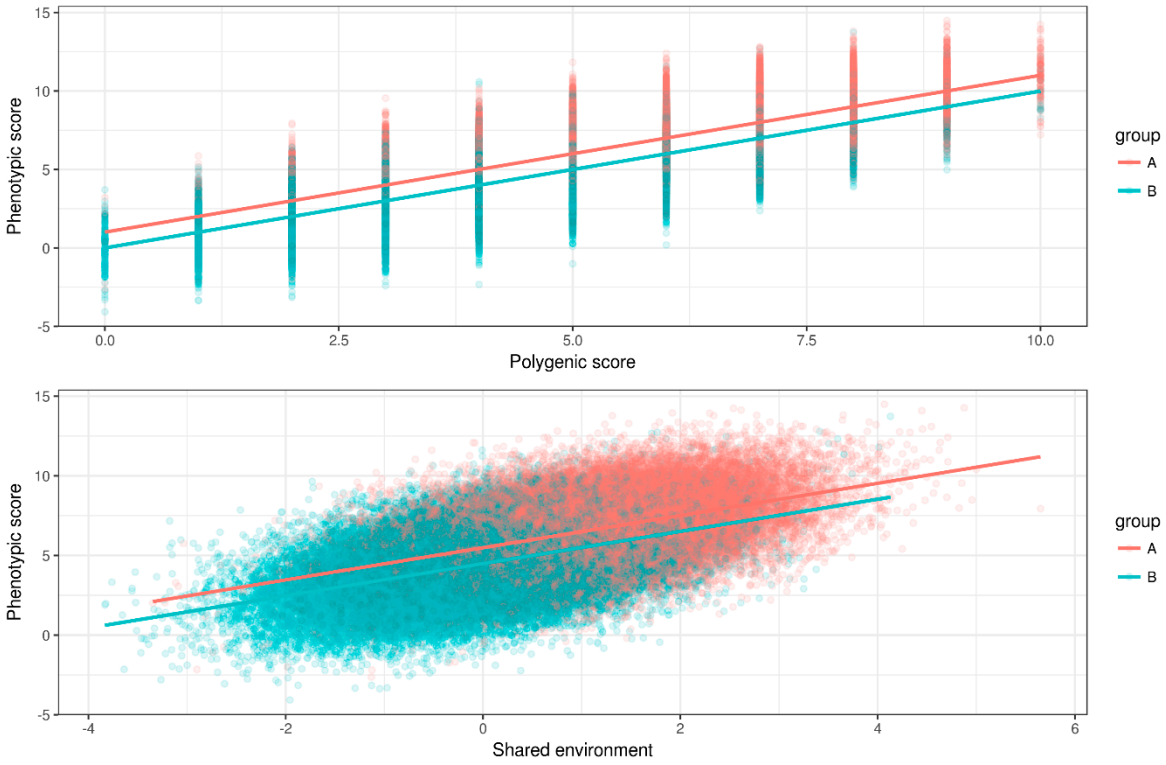

Figure A7. Simulation 4 - Scatterplots of polygenic score/shared environment and phenotypic scores by group.

Unlike in simulation 3, there is now a difference in the intercepts of the regression lines, indicating some unmodeled cause. The slopes, however, are the same. What is the between-group heritability in simulation 4? Despite the unstandardized betas of the predictors being identical, the variance proportions (and correlations) are not the same. This is because the predictors do not have the same variance. Formally, the between group heritability is given by the equation below:

$$h_B{}^2 \approx h_W{}^2 \frac{(1-t)\mathrm{r}}{(1-r)\mathrm{t}}$$

Where h^2 refers to the heritability, subcripts B and W to between and within groups, respectively, r is the genotypic intraclass correlation, and t is the phenotypic intraclass correlation.

The equation was presented by DeFries (1972), and relies upon a number of approximations. If we apply it to the data from simulation 4, we find that the between group heritability is 25%, and the shared environmental part is thus 75% (since non-shared environment does not contribute in this models). In real life, we do not yet have accurate estimates of the true polygenic scores of persons. Thus, unless we can account for the measurement error in our modeling, we need to use another method. One such method is the admixture analysis.

Simulating data for this method is more complicated than the previous cases. For more reliable results, we need to increase the number of causal variants, and 100 was chosen as a reasonable compromise between computation time and realism. Second, we need to add non-causal loci that are used to estimate the ancestry of a given person. For this, we add 1000 variants (also with just 0s and 1s). These are only used to score the ancestry of a person and do not affect the phenotype (i.e. we assume no pleiotropy). Third, we assume random mating, no sexes/hermaphrodites (everybody can mate with anybody), zero fatality (everyone lives to breed), perfect recombination (recombination happens at every locus), and perfect generations (everyone reproduces at the same time with 2 offspring and then dies). These assumptions are needed to simplify the mating process, but do not affect the results in a relevant way. To avoid very long computation times, we also decrease the group sizes to 5,000 for each. We then simulate our founder populations for each group, with the same

situation as in simulation 3, 100% between-group heritability combined with substantial but non-perfect within-group heritability (genetics + noise).

To get admixed persons, we need to mate our total population. Since we assumed random mating, this is done by picking pairs at random until no one is left. Each pair then breeds 2 new individuals by picking a random half of their variants from each parent. This represents the next generation. We repeat this 5 times, so that we have 6 generations (0 through 5). The relationship between ancestry and phenotypic score by generation is shown in Figure A8.

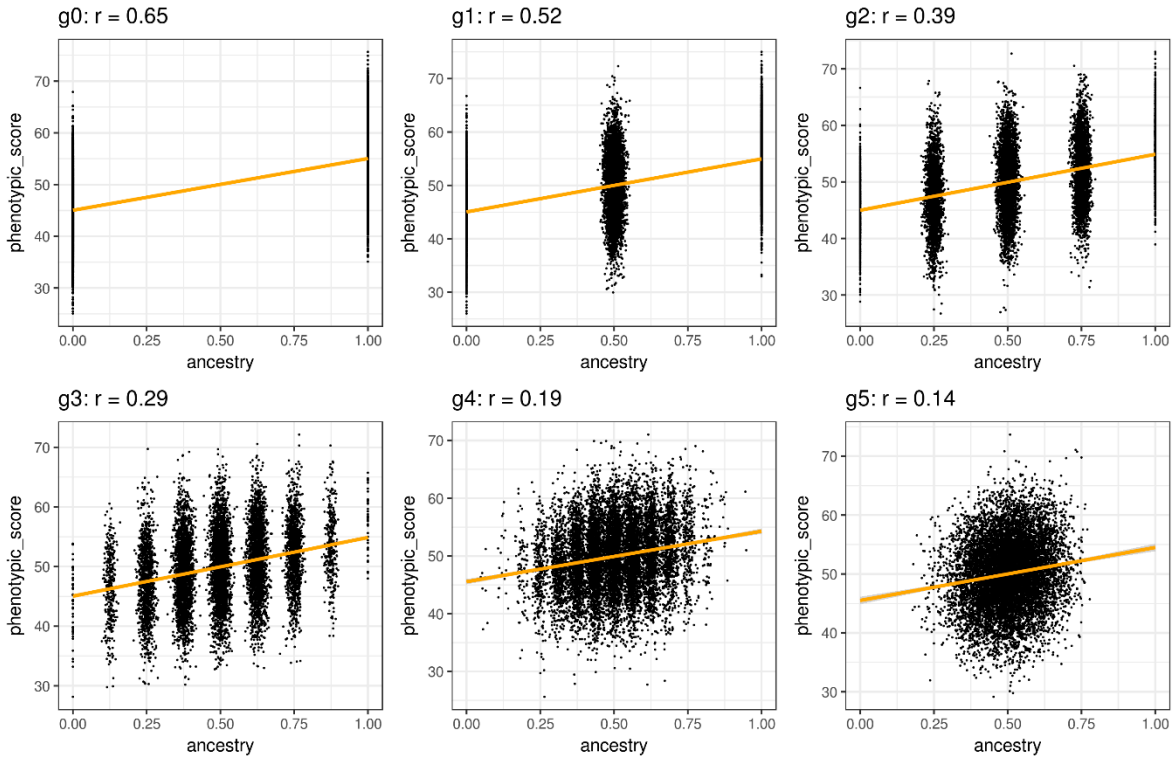

Figure A8. Simulation 5 - Scatterplots of ancestry and phenotypic score for 6 generations.

There are two things of note here. First, that the predicted values at 0 and 100% ancestry stay the same despite the increasing admixture: they are always 45 and 55, which are the ancestry mean polygenic scores. Second, the correlations do not stay the same because they depend on the variation in ancestry. Since variation in ancestry decreases with each generation, the correlations must go down. As the size of correlations (and standardized betas) cannot be used to estimate the importance of genetic causation for a given difference, one should use the model predicted values at 100% ancestry for a given group.

Simulation 6 – Non-random mating

Since in simulation 5 we assumed random mating, it takes only a few generations to have no visible groups left. As such, there are no two fairly distinct groups we can calculate a between-group heritability for, even though the genetic causes are still related to ancestry. Humans do not pick mates at random, and in fact, there is evidence that they prefer mates similar in ancestry, even within racial groups and when controlling for educational attainment. As a simple solution, mates were matched for ancestry imperfectly such that some outbreeding would occur. The relationship between ancestry and phenotypic score is shown in Figure A9.

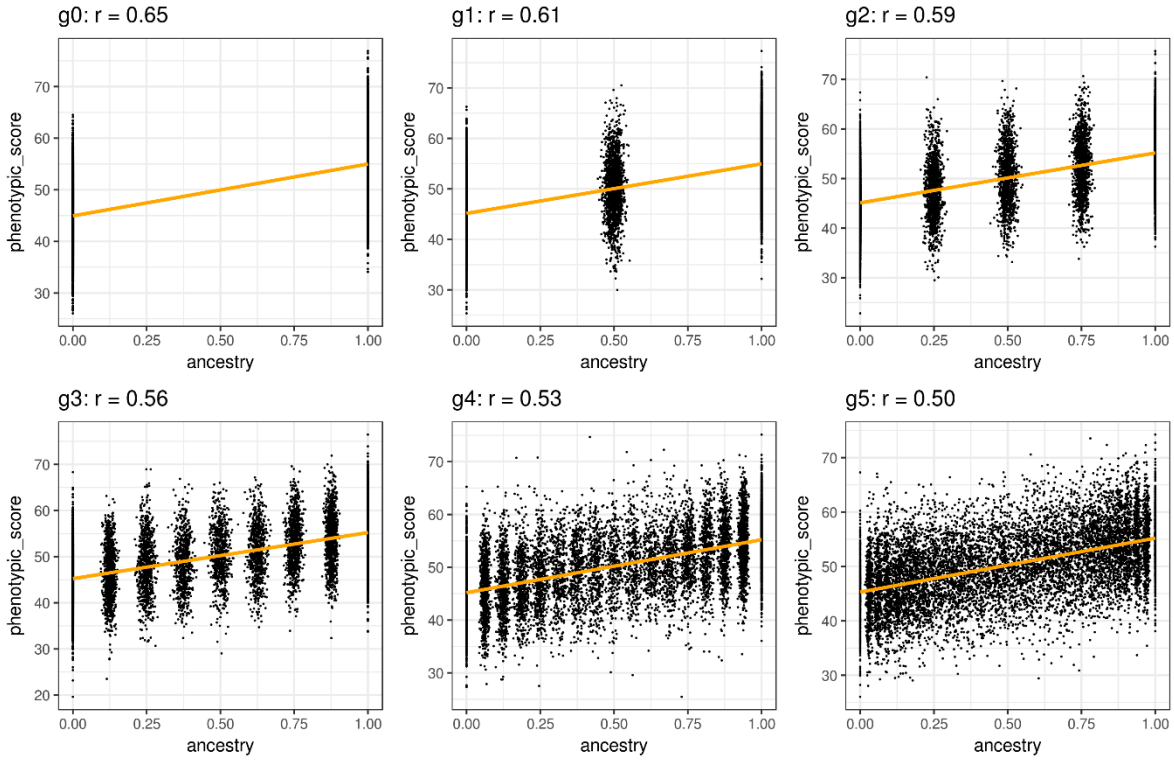

Figure A9. Simulation 6 - Scatterplots of ancestry and phenotypic score for 6 generations.

It can be seen that there is vastly more variation in ancestry in simulation 6 compared to 5. The standard deviations of ancestry in generation 5 (g5) were .34 and .09, for simulation 6 and 5, respectively. A simpler way to see this is to inspect the distribution of ancestry in each simulation side by side, as shown in Figure A10.

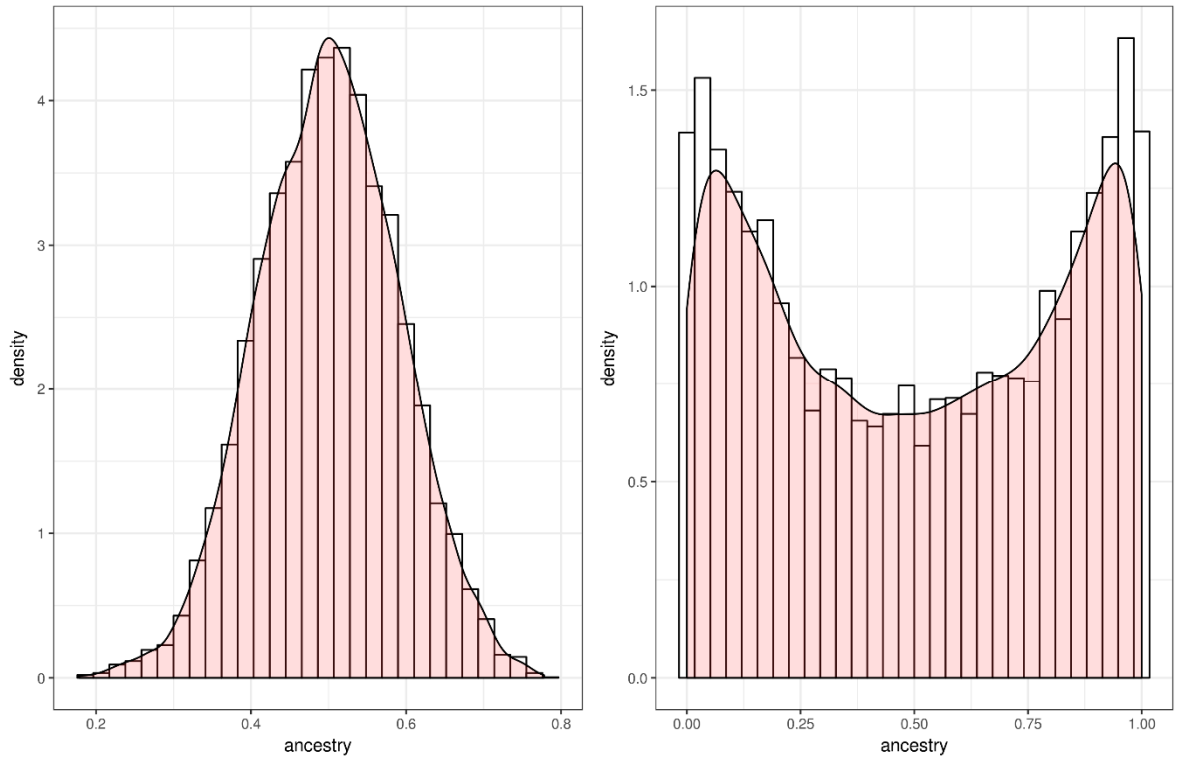

Figure A10. Comparison of ancestry distributions for simulation 5 (left) and 6 (right).

While in simulation the distribution is unimodal, it is still bimodal in simulation 6 after 5 generations of interbreeding. With our bimodal ancestry distribution in hand, we can now simulate results for self-identified racial/ethnic (SIRE) groups. This is done simply by using the ancestry as the probability of being in group A, i.e. a person with 99% (A) ancestry would have a 99% chance of self-identifying as A. The results of this are shown in Figure A11.

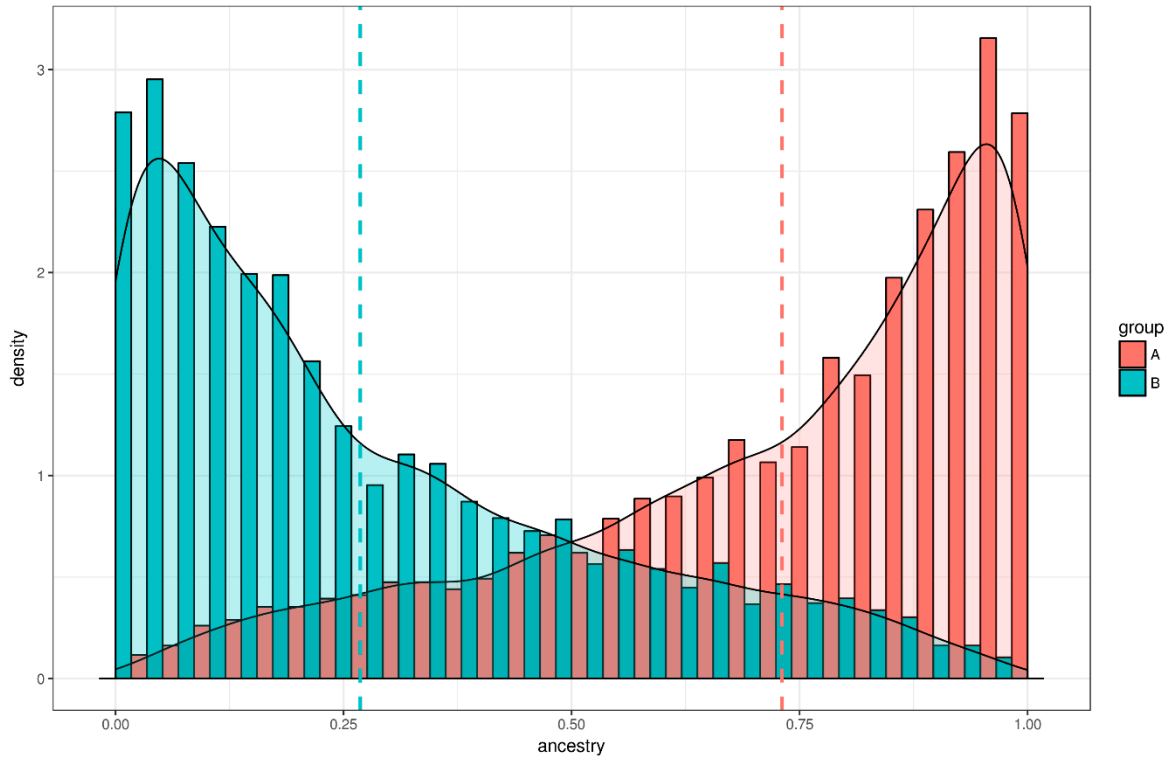

Figure A11. Simulation 6 - Ancestry distribution by SIRE. Vertical lines show the mean.

As expected, each group was higher on its own ancestry and there was significant overlap between them. Finally, we draw the regression lines by SIRE to see if we can correctly deduce the ancestry mean phenotypic scores. The scatterplot is shown in Figure A12.

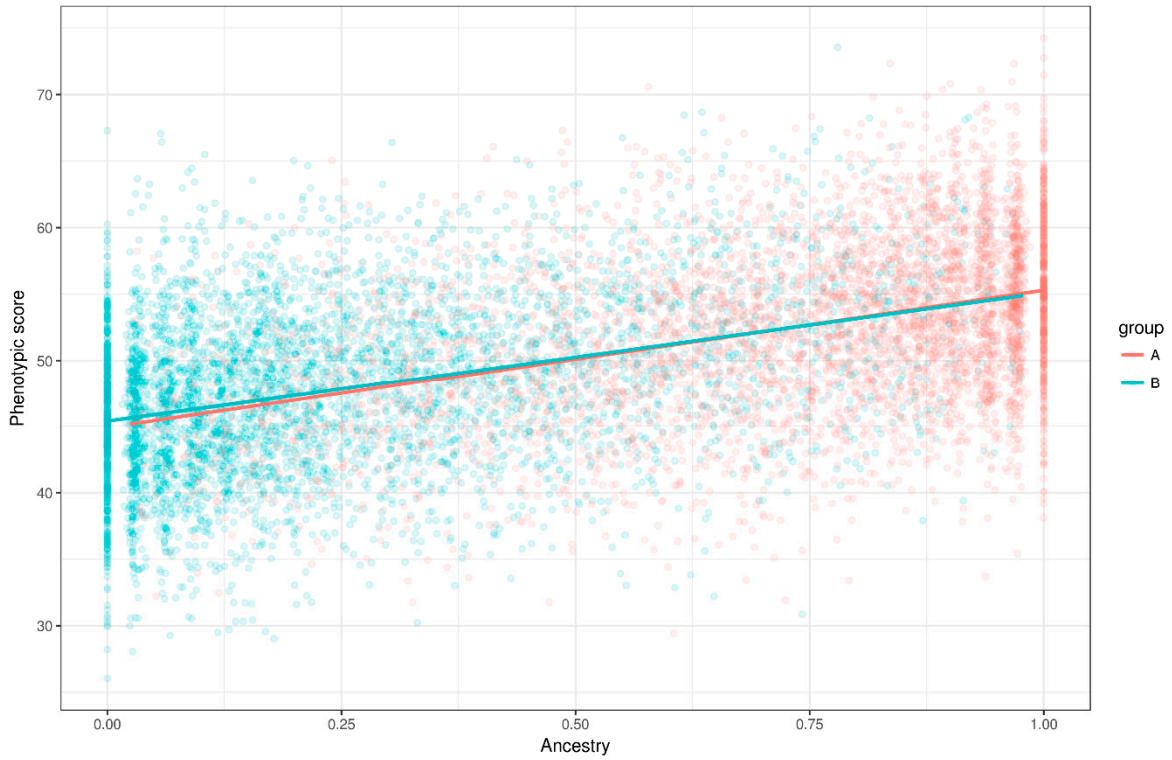

Figure A12. Simulation 6 - Scatterplot of ancestry and Phenotypic score by SIRE.

As expected, the predicted values for 0% and 100% ancestry are the same for each SIRE group (45 and 55). Finally, we apply the equation mentioned in Section **Error! Bookmark not defined.** (African Americans). It correctly reached the result that 100% of the between-group variance was genetic in origin.

One can further introduce more complexities such as multiple ancestries and non-genetic causation (<100% between group heritability). However, we think it is clear at this point that admixture analysis can be used to find indirect evidence of genetic causation. Certainly, it can produce evidence incongruent with genetic models and thus offers a test of them.

Supplementary. File 2

An online supplementary file 2 for this paper is available through Open Science Frame: https://osf.io/64y8m/