The Effect of Directional Dominance on Additive Effect Sizes

Kyelee Fitts, Jeremy Berg, Yuval Simons, Guy Sella
Systems Biology Department, Columbia University

December 12, 2017

Abstract

The averge effect size α is the estimated effect size for an allele for a polygenic trait. These effect sizes are estimated using Genome-Wide Association Studies, and we can estimate signals of natural selection for polygenic traits by weighting by α in statistical tests for polygenic adaptation. However, in the presence of directional dominance, we hypothesized that such statistical tests could be biased. For example, alleles which are systematically dominant and have increased in frequency in the recent past will have deflated average effect sizes, and vice versa. Simulations of the test statistic Q_x used to detect polygenic adaptation, expanded to account for dominance effects, show that the effect that directional dominance has on the statistic is to shift the distribution to the right. Furthermore, directional dominance causes α to systematically stretch or shrink genetic value over generations, further pointing to inflations of deflations of α in accordance with directional dominance. Preliminary investigations of the UK Biobank data for height show that height displays both polygenic adaptation and directional dominance.

Introduction

The diversity of life on Earth is due in large part to the adaptation of organisms to their varying environments. We now know that adaptation has a large genetic basis. Thus, understanding how and why such genetic variation occurs is a major goal of evolutionary biology. Quantifying genetic adaptation is an important aim of population geneticists, and to this end finding methods to detect signals of selection within the genome is of particular interest. Change in the mean phenotype of a population over time can be attributed to many causes including genetic drift or gene flow, so detecting signals of selection in particular is no trivial task.

When natural selection occurs with polygenic traits, the signal of selection get spread out over many loci, such that no individual loci displays large frequency changes against the background of genetic drift. Understanding how to find signals of natural selection for such polygenic traits has been the subject of much study in recent years. Until recently, the population genetic methods to detect polygenic adaptation have been limited to large-effect alleles by finding individual loci with significant allele frequency changes in the recent past– this is an indication of strong, recent national selection.

Lewontin and Krakauer developed a novel statistical test for this kind of selection based on the parameter F_{st}, which was first introduced by Wright, who defined it as a measure comparing the genetic variation of an allele at a specific site in a subpopulation to that of the entire population [6]. Lewontin and Krakauer used the fact that under no selection, the expected F_{st} at a given site will be the same as the population F_{st}. Further, they concluded that the distribution of F_{st} across all sites will be chi-squared F_{st} [11]. The natural conclusion is that sites with statistically different F_{st} values are candidates for loci that have been acted upon by selection. These conclusions were extended by Spitze, who coined the parameter Q_{st} as a measure of how the variation of all loci contributing to a phenotype of a subpopulation compares to that of the entire population [3]. Essentially, Q_{st} is analogous to F_{st}, except that instead of looking at the variation of one allele at a locus, it measures the ratio of the variation of all loci contributing to a phenotype within a subpopulation to that of the entire population. These early tests for selection utilized the ratio $\frac{Q_{st}}{F_{st}}$ as a test statistic for detecting signals of selection, where $F_{st} = Q_{st}$ is the null model, where no selection occurs.

However, these early tests were limited by the inability to access the many thousands of loci that affect polygenic traits. The advent of modern computational tools have given us the ability to do just that, allowing us to better evaluate selection in highly polygenic traits such as height. One of the most important tools developed are Genome-Wide Association Studies (GWAS) which determine which loci across the genome have significant effects on a certain polygenic trait [10] and what the average effect sizes of those loci are. In its simplest form, GWAS works by estimating a linear regression based on the following model [9]:

$$\vec{Y} = \mu + \alpha\vec{g} + \vec{\epsilon} \tag{1}$$

Where $\vec{Y}$ is a vector of the measurements of the phenotype of interest over a population of genetically unrelated individuals, $\vec{g}$ is a vector of the genotypes of each individual, μ is the mean phenotype in the sample, and α is the average effect in the sample of switching the allele with the other allele at the given locus. $\vec{\epsilon}$ is the vector of error terms for each individual. GWAS performs a linear regression using this model, where under the null model, the allele does not contribute to polygenic selection and is only subject to drift. If the allele is statistically significant, GWAS will estimate its nonzero effect size.

Turchin et al. [4] were the first to show how this GWAS data could be used to detect signals of selection in polygenic traits. The idea is to look for coordinated shifts in frequency over many different loci associated with a polygenic trait to find statistically significant changes in allele frequency that are unlikely to have been caused by drift alone. These tests depend on the fact that under drift (the null hypothesis), the effect sizes α and the allele frequencies of the alleles are independent. In otherwords, the test detects signals of selection by looking for significant, coordinated changes in allele frequencies that correlates with a change in the effect sizes of the alleles in question. If allele frequencies and effect sizes are not independent in the null model, however, then the test will start to over or underestimate signals of selection. We hypothesize that one effect that would confound the test for polygenic adaptation in this way is directional dominance.

Dominance in population genetics refers to when the effect of one allele depends on the presence of another allele at the locus. We can account for the effect of dominance in the effect size by parametrizing α in terms of the homozygous effect size (A), or effect size of the most frequent homozygote, and the dominance deviation (D), which is the difference in effect size for the heterozygote deviating from $\frac{1}{2}$ of the homozygous effect size [8]:

$$\alpha_\ell = \frac{1}{2}A_\ell + D_\ell\left(1 - 2p_{1\ell}\right). \tag{2}$$

In particular, directional dominance is when alleles with a postive effect size on the trait are systematically dominant and alleles with a negative effect size on the trait are systematically recessive or vice versa (i.e., dominant alleles have a negative effect or recessive have a positive effect). In terms of (2), this means that the average D over all loci is different from zero.

We hypothesized that directional dominance is problematic in tests of polygenic adaptation because the effect size we estimate in the presence of directional dominance depends on how the allele frequency p has changed in the recent past. In particular, recessive alleles which have decreased in frequency in the recent past will tend to have larger effect sizes, as shown in Figure 1 in the supplement. Notice that when there is no dominance, α is constant, but in the presence of directional dominance (average $D \neq 0$) for alleles with lower frequency will have a higher effect size.

The particular test statistic Q_x analyzed here to evaluate this hypothesis was introduced by Berg and Coop [2]. This statistic depends on F_{st} and a generalized analogy to Q_{st}, expressed in terms of α and the allele frequencies p over loci l and l', summed over populations m. V_a is the additive genetic variance of the entire population, and $\overline{p}_l$ is the mean frequency.

$$Q_X = \frac{1}{V_A F_{ST}} \sum_{m=1}^{M} \sum_{\ell=1}^{L} \sum_{\ell\prime=1}^{L} \alpha_\ell \alpha_{\ell'} \left(p_{m\ell} - \overline{p}_\ell\right) \left(p_{m\ell\prime} - \overline{p}_{\ell\prime}\right) \tag{3}$$

The distribution of this statistic is expected to be chi-squared. Notice that the additive effect size α is the weighting factor for loci in the expression for Q_x.

In the presence of directional dominance, we hypothesized a bias in tests for polygenic adaptation that depends on α. For example, alleles which are dominant and which have recently increased in frequency (large p, positive D, for positive A) will tend to have smaller effect sizes, while alleles which are dominant which have recently decreased in frequency (small p, negative D, for positive A) will tend to have larger effect sizes. In the latter case, the test for polygenic selection advanced by Berg and Coop (3) will tend to make false positive judgements for selection, because the statistic will be calculated over alleles which do not actually increase the effect size.

Height is one polygenic trait with many well-defined signficant alleles via GWAS [4] that also shows evidence of polygenic selection. It has also been shown to exhibit directional dominance [5]. In otherwords, we suspect that the non-independence between allele frequency and effect size caused by directional dominance will cause tests of polygenic adaptation to over or underestimate signals of polygenic selection. We aim to quantify the hypothesized bias, first with simulated populations, then with height genotype data from the UK Biobank.

Theory/Methods

Using the expression for the test statistic Q_x (3), we can substitute the expression for α (2) and manipulate the expression algebraically to derive the following expansion for the test statistic, in terms of the homozygous effect (A) and the dominance deviation (D):

$$
\begin{aligned}
&\sum_{l=1}^{L}(\frac{1}{2}A_l(p_{1l}-\epsilon_l))^2+\sum_{l=1}^{L}\sum_{l\neq l'}^{L}(\frac{1}{4}A_l(p_{1l}-\epsilon_l)A_{l'}(p_{1l'}-\epsilon_{l'}))\\
&+\sum_{l=1}^{L}A_lD_l(1-2p_{1l})(p_{1l}-\epsilon_l)^2+\sum_{l=1}^{L}\sum_{l\neq l'}^{L}(\frac{1}{2}A_lD_{l'}(1-2_{p1l'})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'})\\
&\qquad+\frac{1}{2}D_lA_{l'}(1-2_{p1l})(p_{1l}-\epsilon_l)(p_{1l'}-\epsilon_{l'}))\\
&+\sum_{l=1}^{L}(D_l(1-2p_l)(p_{1l}-\epsilon_1))^2+\sum_{l=1}^{L}\sum_{l\neq l'}^{L}D_l(1-2p_l)(p_{1l}-\epsilon_1)D_{l'}(1-2p_{l'})(p_{1l'}-\epsilon_{l'})
\end{aligned}
\tag{4}
$$

We can loosely consider the single summation terms to be variances corresponding to the expansion of additive effects multiplied by additive effects, additive times dominance, and dominance times dominance, and the double summation terms as covariances of these quantities. We expect the inflation of the test statistic due to dominance effects to come from the last two terms. Note that when dominance is not present ($D = 0$) the expansion reduces to the expression Berg and Coop present when α is treated as a constant [2].

Using this expansion, we have created simulations to characterize our hypothesized dominance bias. We used simulated populations under a couple of assumptions: first, that the values of the dominance deviations and homozygous effects are constant throughout the population. While this is not true in general, we expect that the distributions of these parameters are roughly normal [1], and noise around the expected values of the dominance deviations or homozygous effects should not affect the distribution of the test statistic too much. Second, that F_{st} for these simulated approximations can be roughly estimated by the number of generations elapsed over the population size [7]. Third, that the distribution of allele frequencies after one generation can be approximated by a normal distribution centered at the ancestral frequency with variance $F_{st} * \epsilon * (1-\epsilon)$ where ϵ is the ancestral frequency [8].

All simulations were performed in R.

Results/Discussion

Simulations of the expansion (4) were performed with population size 10000 over 100 generations, starting at an ancestral frequency of 0.5, an homozygous effect size of 0.5, and a dominance deviation value of 0 (Figure 2). The distribution of Q_x was simulated under these conditions for 1000 replicates. The distribution was roughly chi-squared, as expected, with mean of 0.99 and variance of 2.2. Adding directional dominance effects to simulation led to a shift of the distribution, as shown in Figure 3.

To further explore this shift, we plotted the expected cumulative distribution function of the distribution of the test statistic over several different dominance deviation values. Figure 4 shows that with no dominance, the cdf of the statistic is very close to the expected cdf of a chi-squared distributed function (in red), with mean 1 and variance 2 (as expected for a statistic taken over one population). For dominance deviations greater than 0 (Figure 5 shows the expected cdf of a statistic with dominance deviation of 0.5), the distribution is strongly shifted to the right. This observation is taken further in Figure 6, where we plot the proportion of the test statistic over the threshold of the statistic such that p = 0.05. In otherwords, we plot the proportion of false positives for the test statistic in the presence of directional dominance. For dominance deviation values of 0, this proportion is very close to 0.05. However, as dominance effects increase, there is a corresponding increase in the proportion of the statistic over this threshold. Notice that as the dominance deviation approaches 1 (complete dominance), almost all of the values of Q_x are over the expected threshold.

Next, we were interested in seeing how the effect of dominance affects the average effect size over time. We simulated the frequency of an allele over 1000 generations under the Normal approximation to drift both without dominance (Figure 7) and with a dominance value of 0.2 (Figure 8), replicating this simulation 10000 times, representing 10000 simulated populations. The red line shows the average allele frequency over time. As expected for no dominance the average allele frequency stays at 0.5, or the ancestral allele frequency, because alleles are equally likely to be fixed or lost among many populations in this case. Interestingly, for a dominance deviation of 0.2, the average allele frequency also appears to stay very close to the ancestral frequency, indicating that for this low of a dominance deviation it is difficult to qualitatively distinguish between the proportion of populations that fix the allele and the proportion of populations for which the allele is lost.

However, by plotting the genetic value (allele frequency multiplied by average effect size) of a simulated allele over time, we expect to see that dominance will have a nontrivial effect on the trajectory of the average genetic value. This is because while there is still on average an even split between populations that fix the allele and populations that lose it, the multiplication of the average effect will have the effect of stretching the genetic value of alleles that have recently increased in frequency in the case of negative dominance deviations (Figure 10) or decreased in frequency in the case of positive dominance deviations (Figure 11), with a corresponding shrinkage in the opposite direction. Indeed, we see that the average genetic value (red line). Figure 9 shows the genetic value over time of populations with no dominance, which as expected, with no stretching or shrinking of the genetic value in either direction.

This last plot is particularly interesting because it shows that in the presence of directional dominance, the stretching and shrinking effect of α has the potential to skew the genetic value of the allele in question, meaning not only that GWAS is more likely to choose these alleles as significant for selection, but also that when these alleles are used in the test statistic Q_x, their contribution will be inflated or deflated, because the average effect is what weighs alleles in the statistic.

Finally, in order to look for the effects of directional dominance on the average effect size in alpha, we wanted to show that height is one polygenic trait that shows signals of directional dominance. To this end, taking GWAS data from the UK Biobank, we created a qq plot of the dominance deviation p-values, choosing SNPs that were most significant ($p <= 10^{-8}$) p-values for the average effect size over all chromosomes. Figure 12 shows that the p-values are skewed– in otherwords, we see that for a large number of significant sites, the dominance deviation p-values are not distributed as we would expect, indicating that there are signs of dominance in height.

Furthermore, we verified that height shows signals of directional dominance in particular by bootstrapping the dominance deviation values to see if on average they are nonzero across all significant loci (Figure 13). We chose alleles with average effect size p-value $<= 10^{-8}$ and used sampled the most significant alleles within blocks of approximately independent alleles in the genome (to account for linkage effects). With 10000 replicates, we found a mean of $7.9 * 10^{-4}$, which suggests that dominance deviations are at least slightly nonzero and the height displays directional dominance.

Conclusion

These simulated results and the preliminary UKBiobank verification of directional dominance in height suggest that, as hypothesized, the effect of directional dominance on the average effect size is nontrivial and in fact could result in false positives or negatives in the statistical tests for polygenic adaptation. Although we have no reason to suspect this effect is very large, as dominance itself does not an particularly large effect on polygenic traits in general, quantifying this effect in height could open many potential avenues of research in understanding how tests for polygenic adaptation could be biased.

Next steps for this project in particular are delving into the UKBiobank data in earnest, performing an in-house GWAS and substracting dominance deviation and homozygous effect size data from alleles that show signs of directional dominance.

Acknowledgements

Thank you to Jeremy Berg, my mentor in the Sella Lab, as well as to Yuval Simons, who initially proposed this project, and Guy Sella, my PI. Thank you also to Hakhamanesh Mostafavi for providing GWAS data for the preliminary analysis of the UK Biobank data.

References

[1] Meuwissen Th Bennewitz J. The distribution of additive and dominance effects in porcine f2 crosses. *Journal of Animal Breeding Genetics*, 2010.

[2] Jeremy J. Berg et al. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. *bioRxiv*, 2017.

[3] K Spitze et al. Population structure in Daphnia obtusa: quantitative genetic and allozymic variation. *Nature Genetics*, 1993.

[4] Michael C Turchin et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, 44:1015–1019, 2012.

[5] Peter K Joshi et al. Directional dominance on stature and cognition in diverse human populations. *Nature*, 523:459–462, 2015.

[6] Sewall Wright et al. The Genetical Structure of Populations. *Annals of Eugenetics*, 1949.

[7] Albert V. Smith George Nicholson et al. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society*, 2002.

[8] John H. Gillespie. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 1998.

[9] Ben Hayes. Overview of statistical methods for genome-wide association studies (gwas). *Genome-Wide Association Studies and Genomic Prediction*, 2013.

[10] Matthew Brown Peter Visscher et al. Five years of gwas discovery. *American Journal of Human Genetics*, 2012.

[11] Jesse Krakauer R.C. Lewontin et al. Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. *Nature Genetics*, 1973.

Supplement

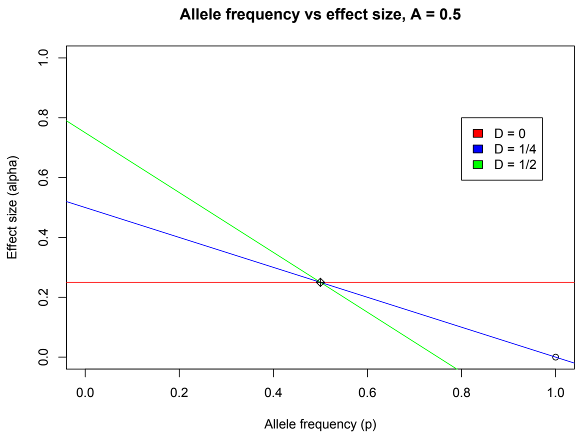

Fig. 1: Effect sizes for various dominance deviation values under directional dominance

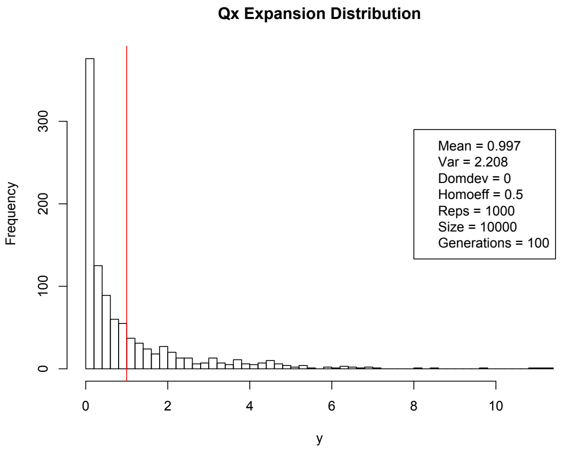

Fig. 2: Q_x Expansion Distribution

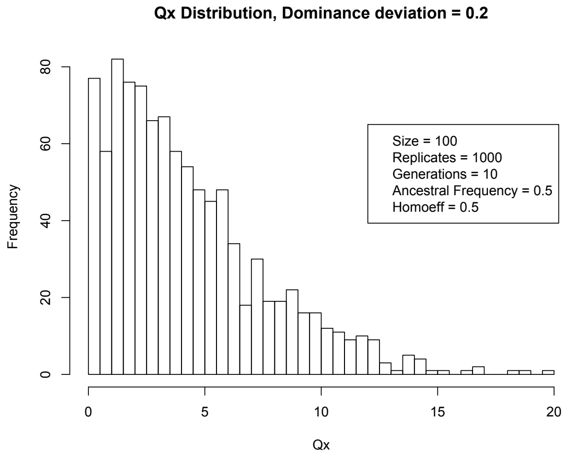

Fig. 3: Q_x Distribution with dominance

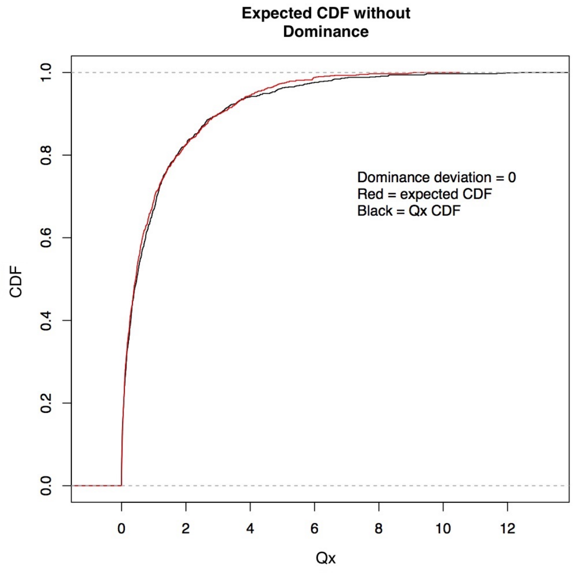

Fig. 4: Expected CDF of Q_x with dominance $= 0$

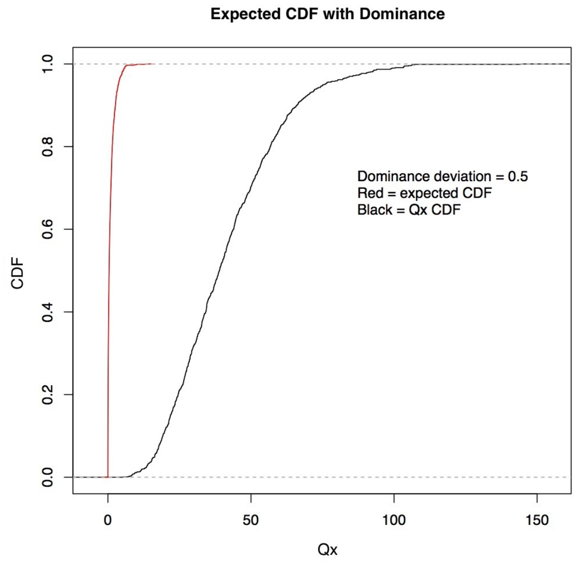

Fig. 5: Q_x Distribution with dominance $= 0.5$

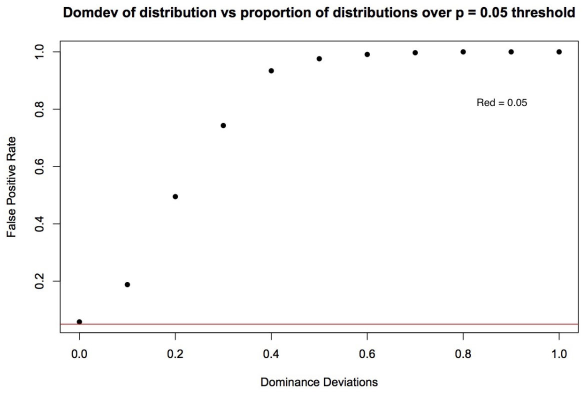

Fig. 6: Proportion of Q_x over expected value for p $= 0.5$

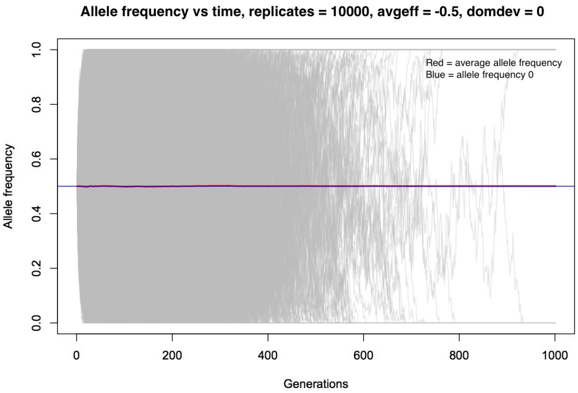

Fig. 7: Allele frequency over time for dominance = 0

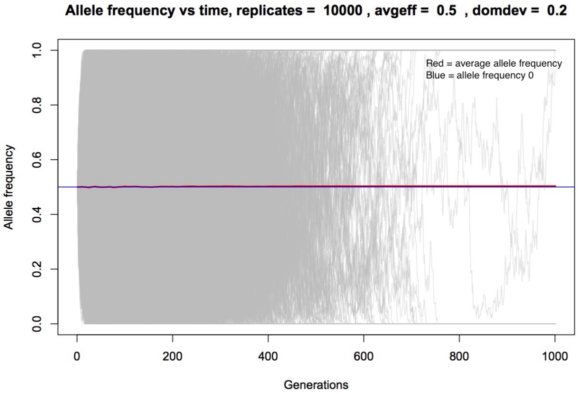

Fig. 8: Allele frequency over time for dominance = 0.2

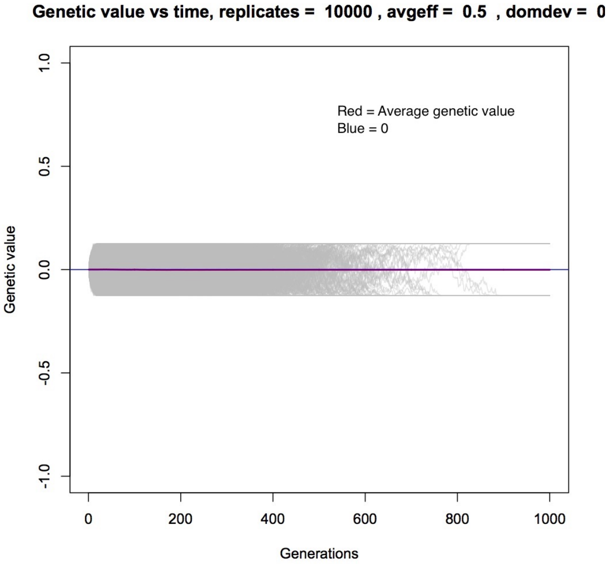

Fig. 9: Genetic value over time for dominance = 0

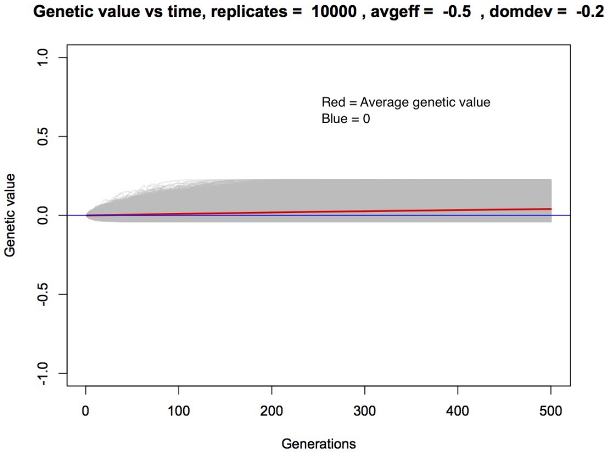

Fig. 10: Genetic value over time for dominance = -0.2

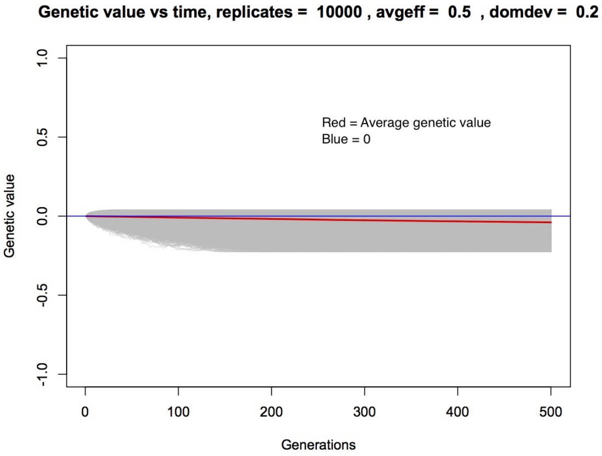

Fig. 11: Genetic value over time for dominance = 0.2

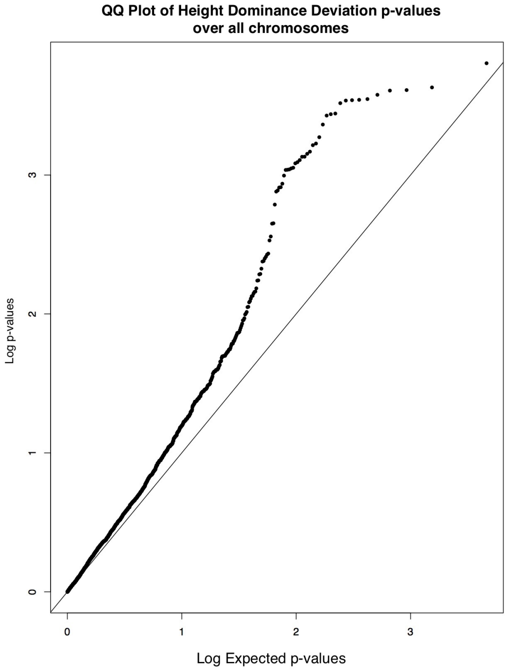

Fig. 12: Height Domiance Deviation p-value qq-plot

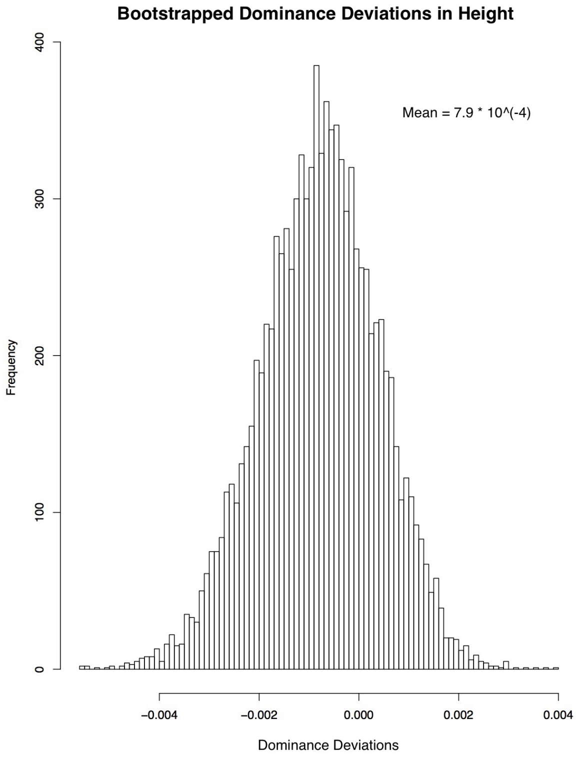

Fig. 13: Bootstrapped Dominance Deviations for height for significant alleles