Selection on correlated genotypes and the change in frequency in response to selection

Moi Exposito-Alonso

These analyses are a proof of concept of how different genome-wide association methods on relative fitness capture better or worse the consequences of natural selection on the genetic makeup of a population. Specificially, both direct and indirect natural selection will change the frequency of genetic variants of a population after a single selection event. Virtually all the methods and discussions below have already been published by Gompert et al 2017 Molecular Ecology ("Multilocus approaches for the measurement of selection on correlated genetic loci"), so we refer to them for a much more extense discussion.

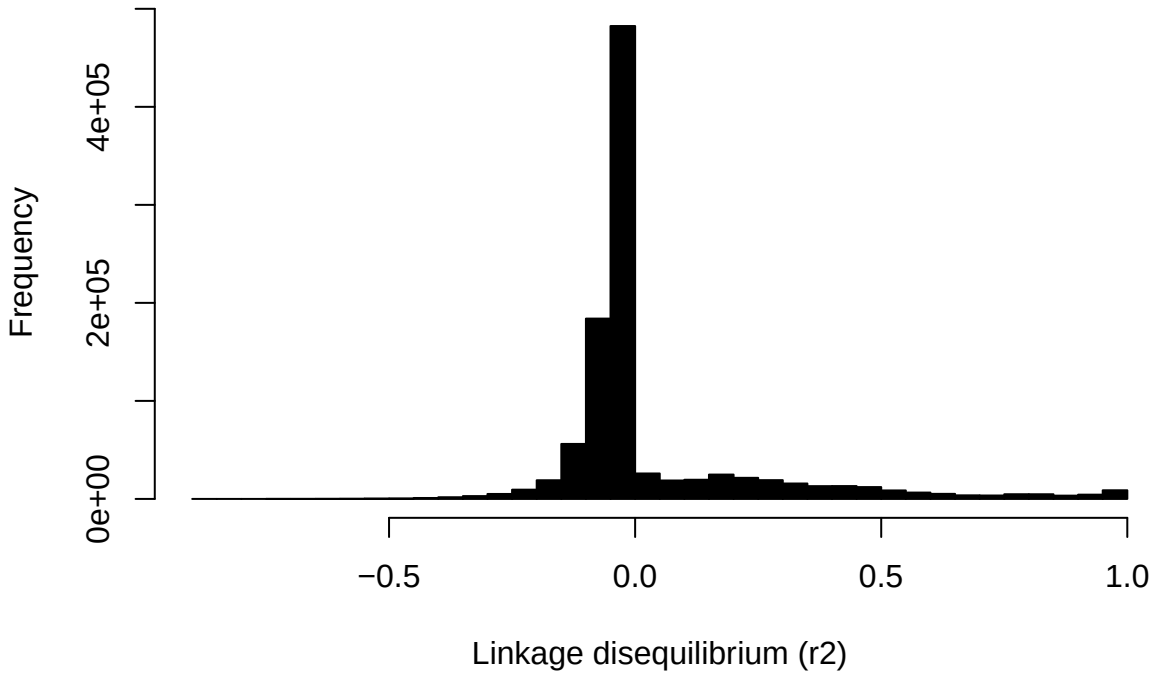
To summarize them, Gompert thinks of the association of genotypes with fitness in analogy to Arnold & Lande's classic Evolution paper "The Measurement of Selection on Correlated Characters" that tries to address direct and direct selection on multiple phenotypes. From their manuscript, the formulation of total selction over a trait zi is represented by: $s = Cov[w, zi]$; where w is the relative fitness. Because other phenotypic traits can covary with zi, one cannot be sure fromt his approach that all the selection experienced by zi is from direct effects, but a sum of indirects effects from n other traits: $s = \sum_{j=1}^{n} Cov[zi, zj]\beta_j$; where β represents the direct effects of seleciton and can be calculated as $\beta = P^1 s$, where P is the n dimensional variance-covariance matrix.

Gompert et al discuss that the same approach can be applied to genetic loci instead of phenotypes (although with some extra nuances as the large amount of predictors that can be solved with modern GWA approaches).

The SNPs in the genome of A thaliana are correlated

For simplicity in the analyses below, we use 1000 contiguous SNPs from chromosome one. This is the same patrix used in the GWA of Exposito-Alonso, Burbano, Bossdorf, Nielsen & Weigel 2018 bioRxiv https://doi.org/10.1101/321133. The genome matrix contains only biallelic SNPs oriented based on minimum allele frequency (alternative allele = minor frequency allele)

```
## g<-readplink("data-raw/515g_1000.012") ## run 1 time
genomes<-readRDS('data-big/genomes.rda')
genomes<-attachgenomes(genomes)

X<-genomes$g[,] # originally coded as 0 and 2
X1=X/(X+1e-32) # transformed to 0 and 1
Xs<-X-1 # transform to -1 +1 format for one software

# LD r
V=cor(X1)
hist(V,xlab="Linkage disequilibrium (r2)", main="",col="black", breaks=50)
```

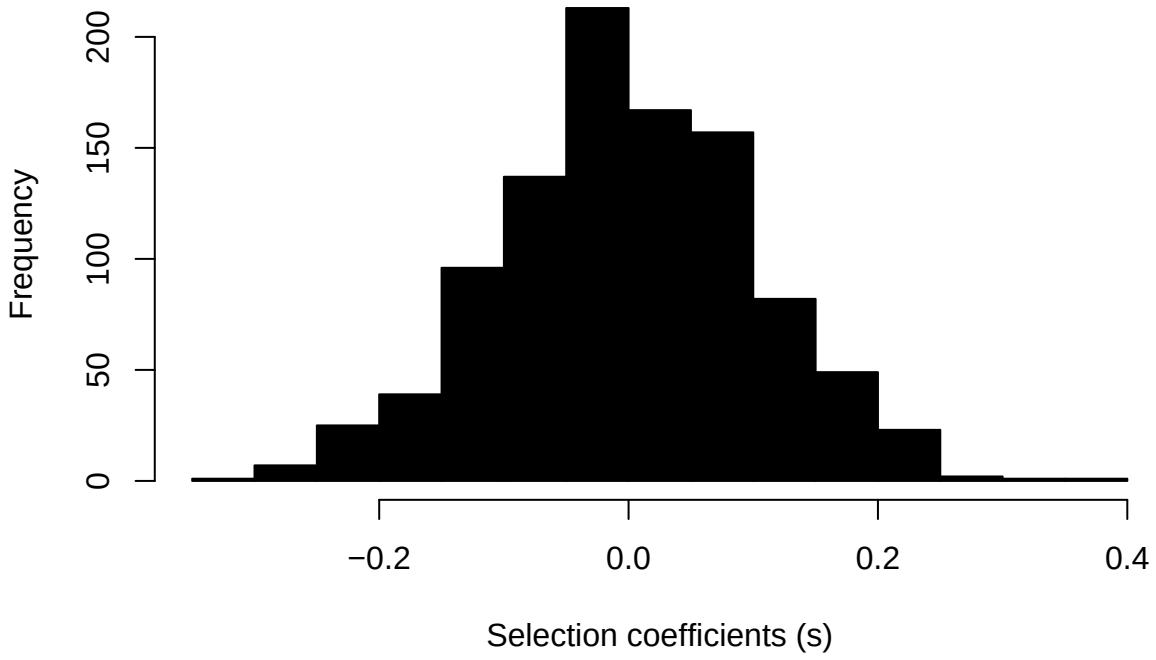


We can see that alternative (minor) alleles are biased to have positive linkage disequilibrium

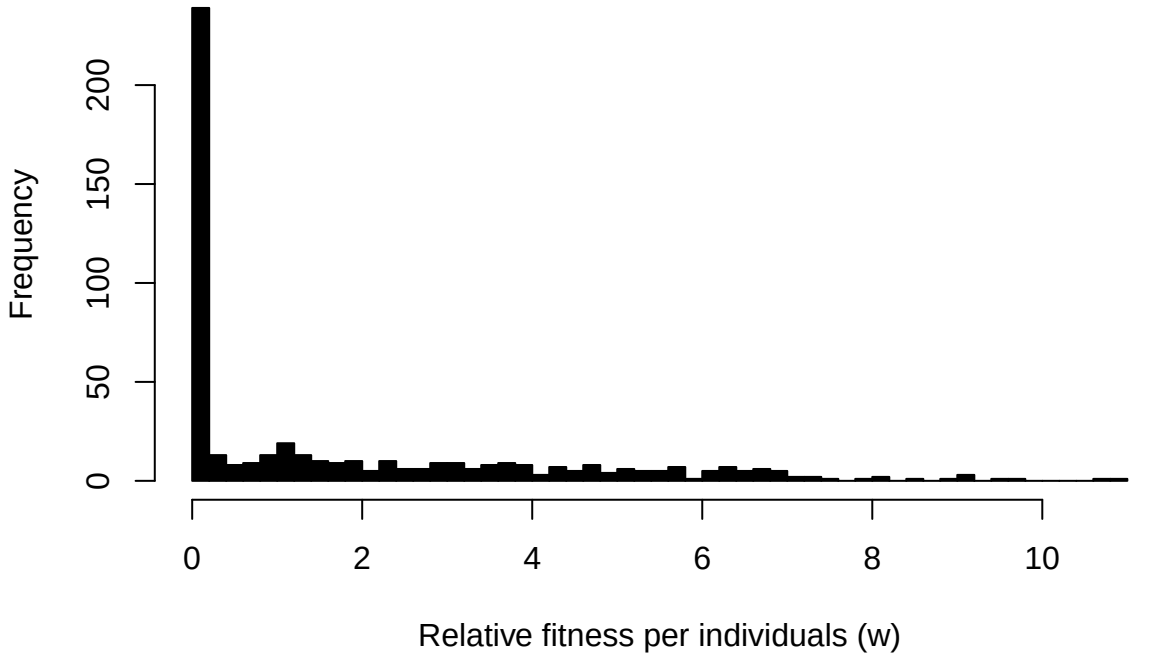
Simulate selection coefficients and fitness

In order to evaluate later whether GWA captures well true selection coefficients, we simulate selection coefficient and calculate the overall relative fitness of the different Arabidopsis accessions (based on an additive model).

```
set.seed(1)

s<-rnorm(1000,mean = 0,sd = 0.1) # polygenic

# s<-rep(0,1000) # for monogenic architecture, results are comparable
# s[sample(1:1000,1)]<-1 * sample(c(-1,1),1)
# s[sample(1:1000,1)]<-0.5 * sample(c(-1,1),1)
# s[sample(1:1000,1)]<-0.05 * sample(c(-1,1),1)


hist(s, xlab="Selection coefficients (s)",main="",col="black")
```




```r
y<-(X1 %*% s)
h2 <- 0.9 # high heritability for simplicity
y <- y + rnorm(515,mean=0,sd=sqrt((1-h2)/h2*var(y)))
y<-y/mean(y)
y[y<0]<-0 # fitness cannot be negative
hist(y, xlab="Relative fitness per individuals (w)",main="",col="black",breaks=50)
```



Example of association with marginal GWA

One approach used in Exposito-Alonso (2018) to understand the total selection acting on a genetic variant is simply the difference in relative fitness between genotypes carrying alternative and reference alleles,

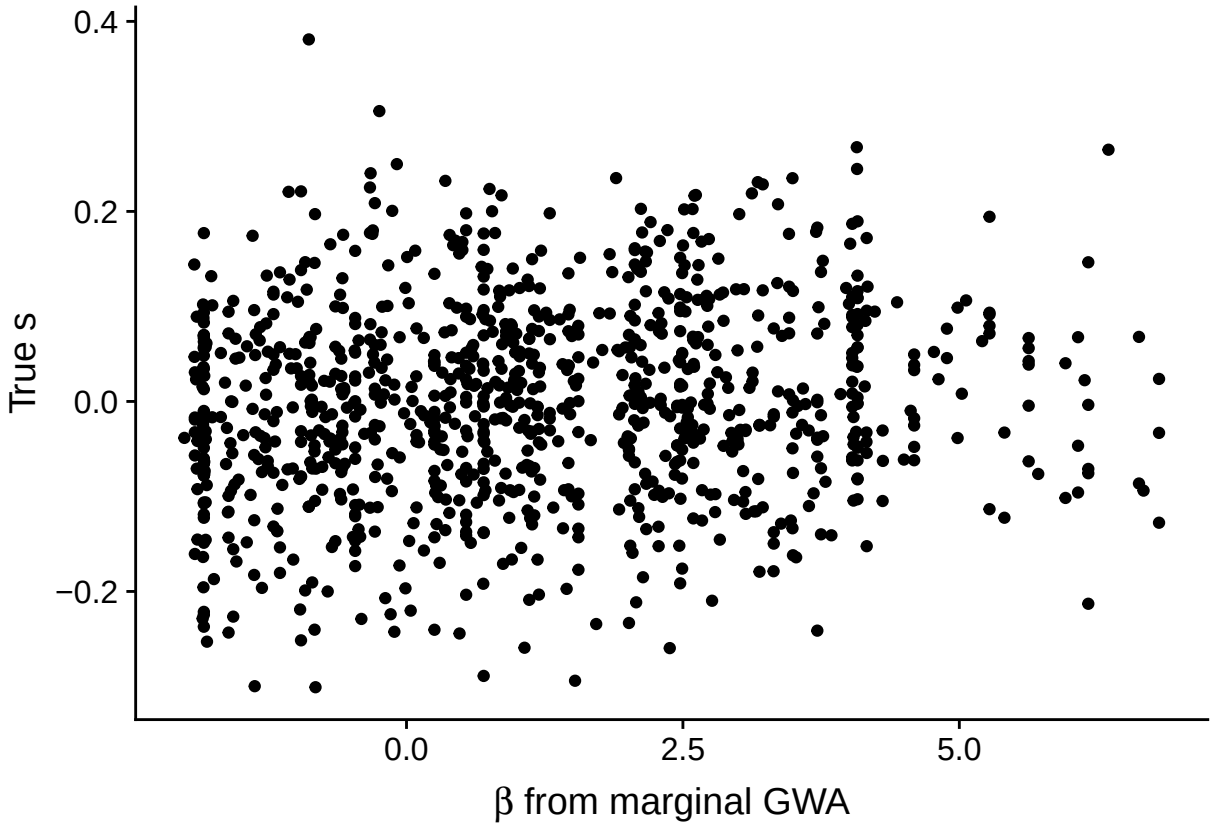
$\beta = s = w_{11} - w_{00}$, what we called allelic selection differential (in analogy to phenotypic selection differential, see Gompert et al 2017). This can be calculated with a marginal or single marker GWA that is carried iteratively along the genome and would take the form:

$$y = X_i\beta_i + \epsilon$$

,

where i refers to a specific SNP, and the linear model is repeated separately for each SNP. This does not include a kinship matrix nor fits multiple markers at a time, therefore linked effects from other markers will be also incorporated in the β. ϵ is the Gaussian residual error.

```
marglm<-gwamarg(y, X1)
b=c(marglm)
# hist(b, xlab="Inferred allelic selection differentials (b=s)",main="",col="black")
# plot(b ~ genomes$map$physical.pos, xlab="Chromosome position (bp)", pch=16)
pl1<-ggdotscolor(b,s,
            xlab = TeX("$\\beta$ from marginal GWA"),
            ylab = TeX("True $s$"))  #%>% addggregression()
pl1
```



Due to the extent of LD that we observed before, the marginal GWA correlates with the true selection coefficients only slightly.

Example of association of kinship GWA

The now most common GWA approaches try to remove effects of SNPs that are correlated with other SNPs. This is done in a number of ways. Perhaps accounting for a relationship or kinship matrix between all

genotypes of the GWA panel is the most accepted one (Yu et al 2006 Nat Gen). The model would take the form:

$$y = X\beta + Zg + \epsilon$$

Where Z is the design matrix, g is the genotype background random effect with $Var(g) = K\sigma^2$, and K is the kinship matrix. This has been widely used in Arabidopsis litterature, e.g. Exposito-Alonso 2017 Nat Ecol Evol, Atwell et al 2011 Nature, 1001 Genomes Consortium 2016 Cell, etc.

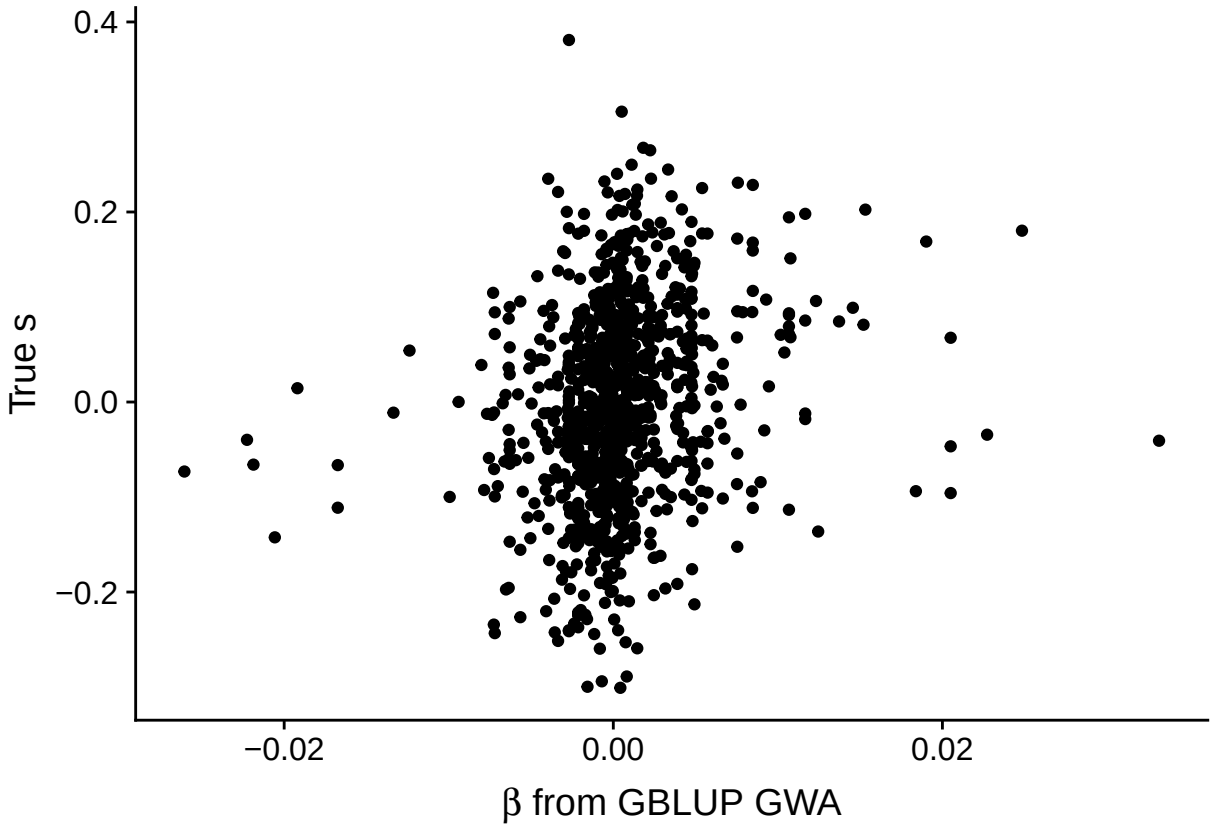
```
library(rrBLUP)
```

```
## Warning: package 'rrBLUP' was built under R version 3.4.3
```

```
K=A.mat(Xs)
colnames(K)=rownames(K)=genomes$fam$sample.ID
pheno=data.frame(genomes$fam$sample.ID,y)
geno_<-t(Xs)
colnames(geno_)<-genomes$fam$sample.ID
geno=cbind(data.frame(genomes$map$marker.ID,
                      genomes$map$chromosome,
                      genomes$map$physical.pos), geno_)



## Two step. GBLUP and gwa on residuals (for simplicity)
blup<-kin.blup(data=data.frame(y=y, Genotype=genomes$fam$sample.ID) ,
               pheno="y",
               geno="Genotype",
               K=K)
message("Variance assigned to the kinship term: ", blup$Vg / (blup$Vg+blup$Ve)
)
```

```
## Variance assigned to the kinship term: 0.893363004226154
```

```
resvarprop=blup$Ve / (blup$Vg+blup$Ve)

myres<-blup$resid
b_gblup<-gwamarg(myres, X1)
bg=b_gblup * resvarprop
pl2<-ggdotscolor(bg,s,
            xlab = TeX("$\\beta$ from GBLUP GWA"),
            ylab = TeX("True $s$"))
pl2
```



Because the effects of the kinship matrix have been removed from the trait, the effect sizes of β are very smal (one result in Exposito-Alonso 2018 from the similar BSLMM).

Note: The commented approach of ridge regression is similar to the BSLMM approach used in Exposito-Alonso 2018 bioRxiv, but one would reach the same conclusion as with the kinship matrix. In fact, under standard Kinship matrices, BSLMM can be identical to GBLUP (see "Polygenic modeling with bayesian sparse linear mixed models" from Zhou Carbonetto & Stephens 2013 PLOS Gen).

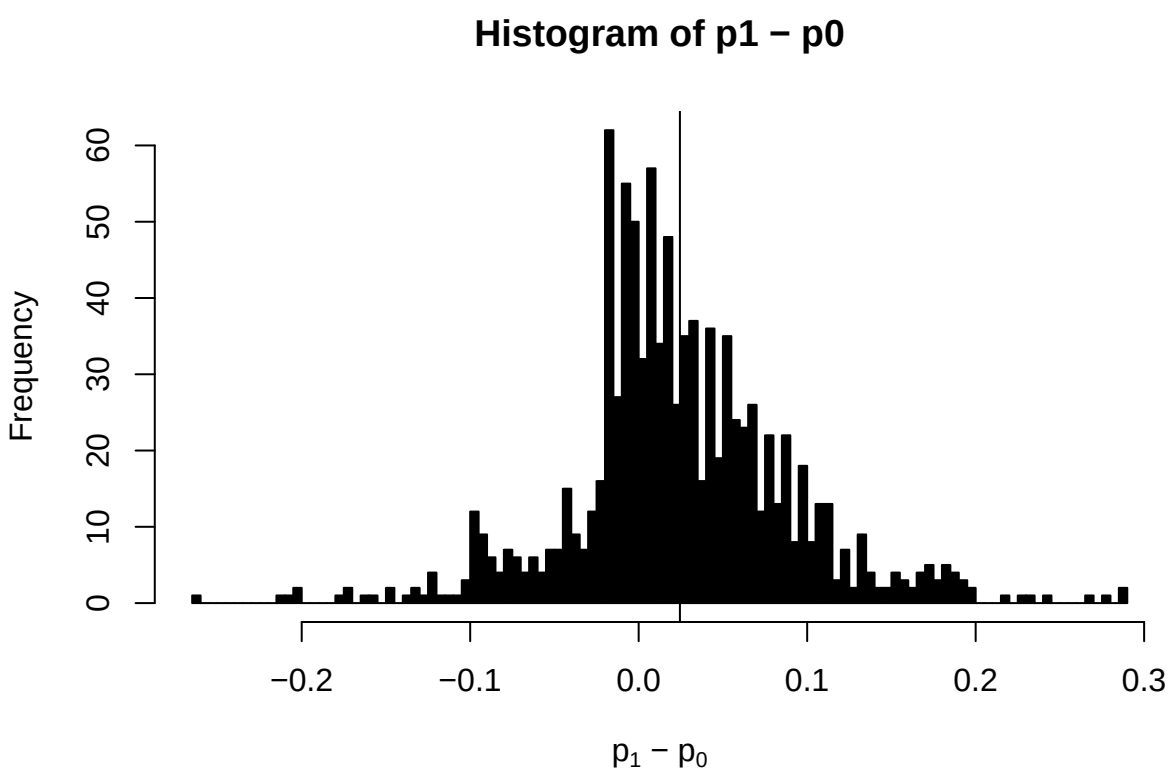
The relationship of β with allele population frequency change

The above examples showcase that estimating selection coefficients is not an easy task, and GWA methods need to be extended specifically for fitness associations, where a number of complications arise, namely zero inflated response variables, linked effects between a large number of parameters (more than observations), and potentially non-additive effects (Exposito-Alonso & Nielsen, to be submitted).

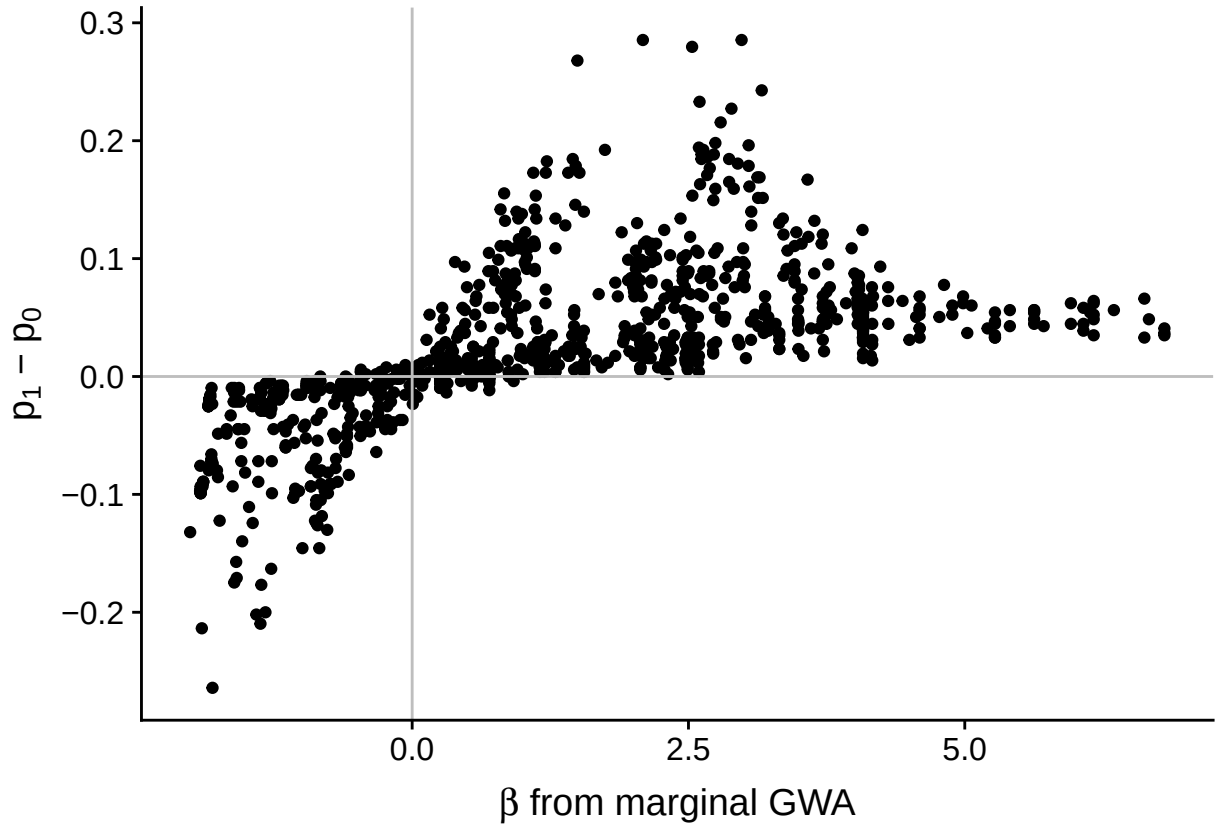
Regardless of the above complications, when the aim of characterizing the impact of selection on a population with high linkage (as the highly diverse species-wide Arabidopsis 515 panel), a marginal GWA might capture what we aim: how much natural drives frequency changes in a selection event. Particularly, in selfer species with strong linkage structure (population structure) and with past histories of local adaptation that might geneate positive linkage between multiple adaptive alleles and adaptive alleles with private alleles of populations, not correcting out the effects of genetic draft (linked selection) might be desirable.

To provide a visual example, we can compare estimates from kinship vs marginal GWA with allele frequency changes in a simple individual-based population simulation approach. In such approach, we sample the 515 genotypes that will comprise the next generation proportionally to their relative fitness. Then we calculate the allele frequencies of the 515 genotypes before and after one generation of selection.

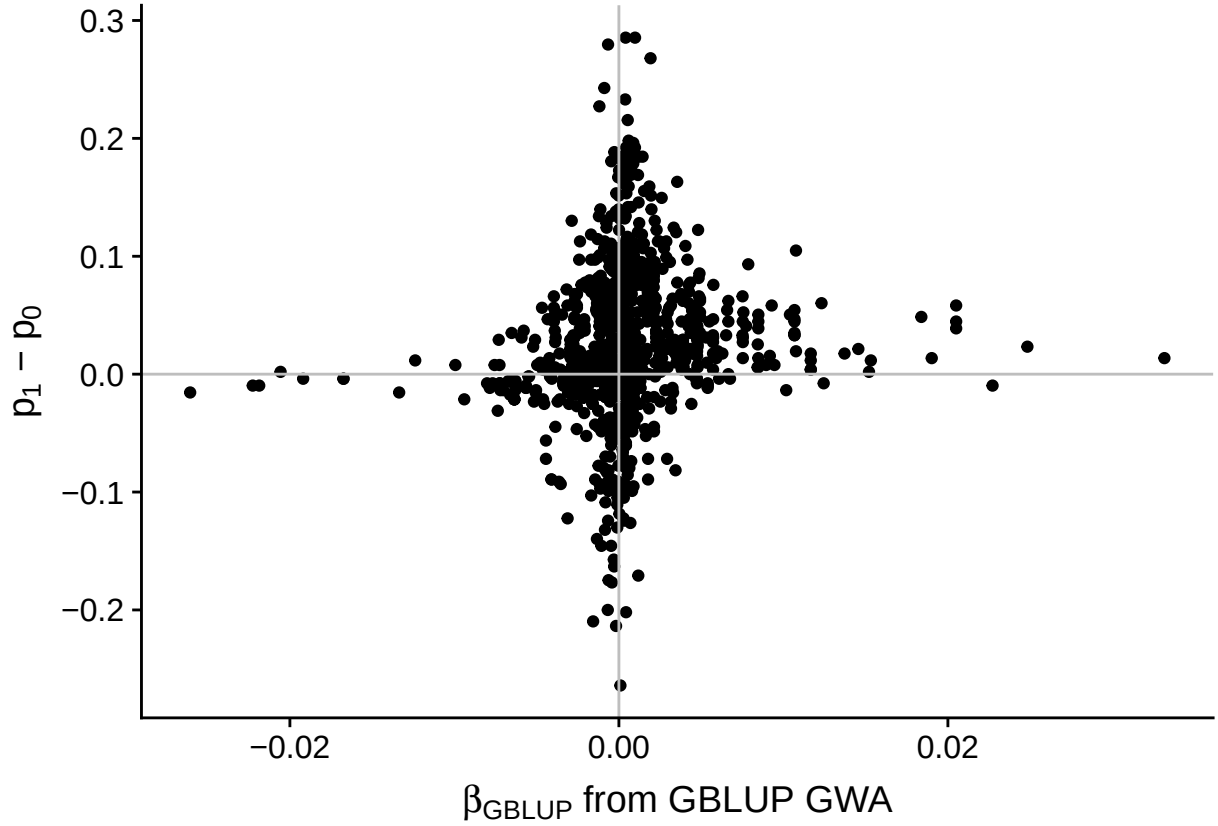
```r
set.seed(1)

p0=apply(X1,2,mean)

p1<-sapply(1:ncol(X1), function(i){
        sample(x = X1[,i],
           size = 515,
           prob = c(y),
           replace=T
           ) %>% mean
        }
      )

hist(p1-p0, col="black",breaks=100,xlab=TeX("p_1 - p_0"))
abline(v=mean(p1-p0))
```




```r
pl3<-ggdotscolor(b,p1-p0,
             ylab=TeX("p_1 - p_0"),
             xlab=TeX("$\\beta$ from marginal GWA") ) +
              geom_hline(yintercept=0,color="grey") +
              geom_vline(xintercept=0,color="grey")

pl4<-ggdotscolor(c(bg),p1-p0,
             ylab=TeX("p_1 - p_0"),
             xlab=TeX("$\\beta_{GBLUP}$ from GBLUP GWA")) +
              geom_hline(yintercept=0,color="grey") +
              geom_vline(xintercept=0,color="grey")
pl3
```


0.3
0.2
0.1
0.0
−0.1
−0.2
$p_1 - p_0$
0.0
2.5
5.0
β from marginal GWA

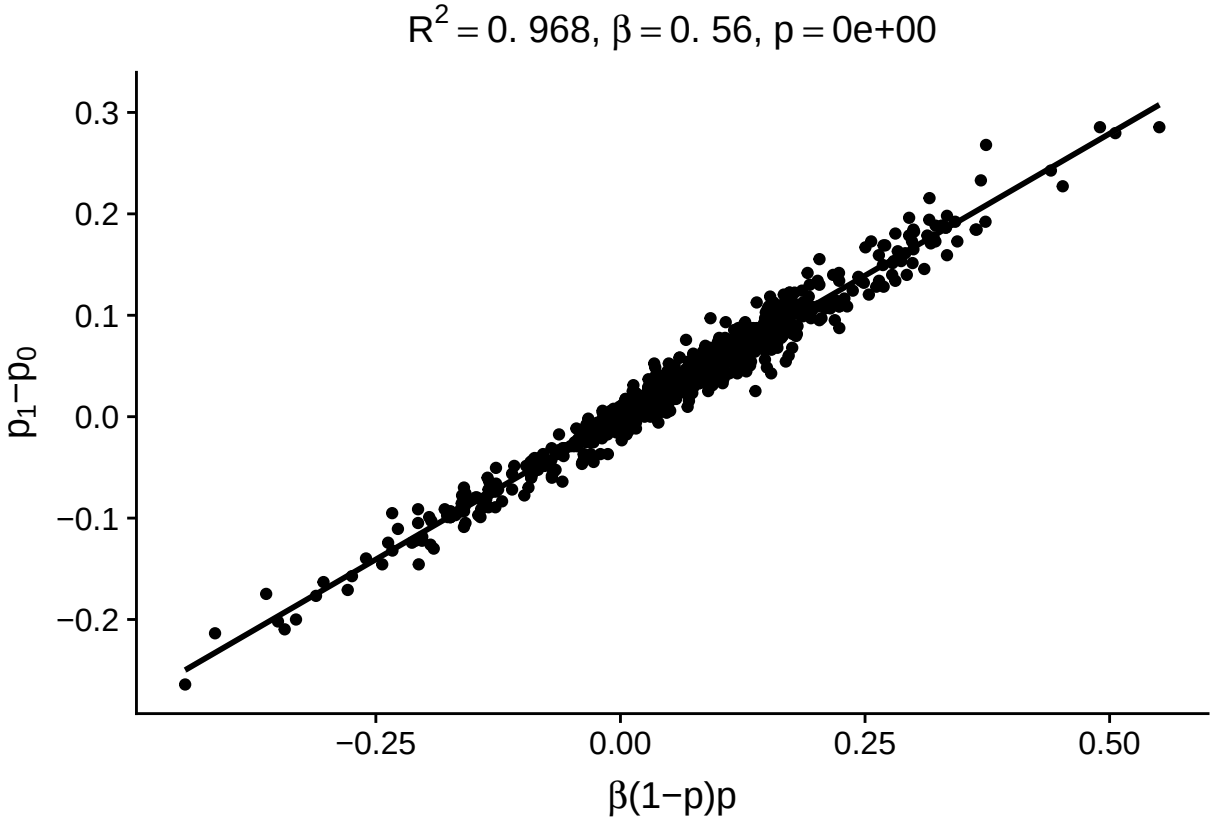
```
p14
```


0.3
0.2
0.1
0.0
−0.1
−0.2
$p_1 - p_0$
−0.02
0.00
0.02
β_{GBLUP} from GBLUP GWA

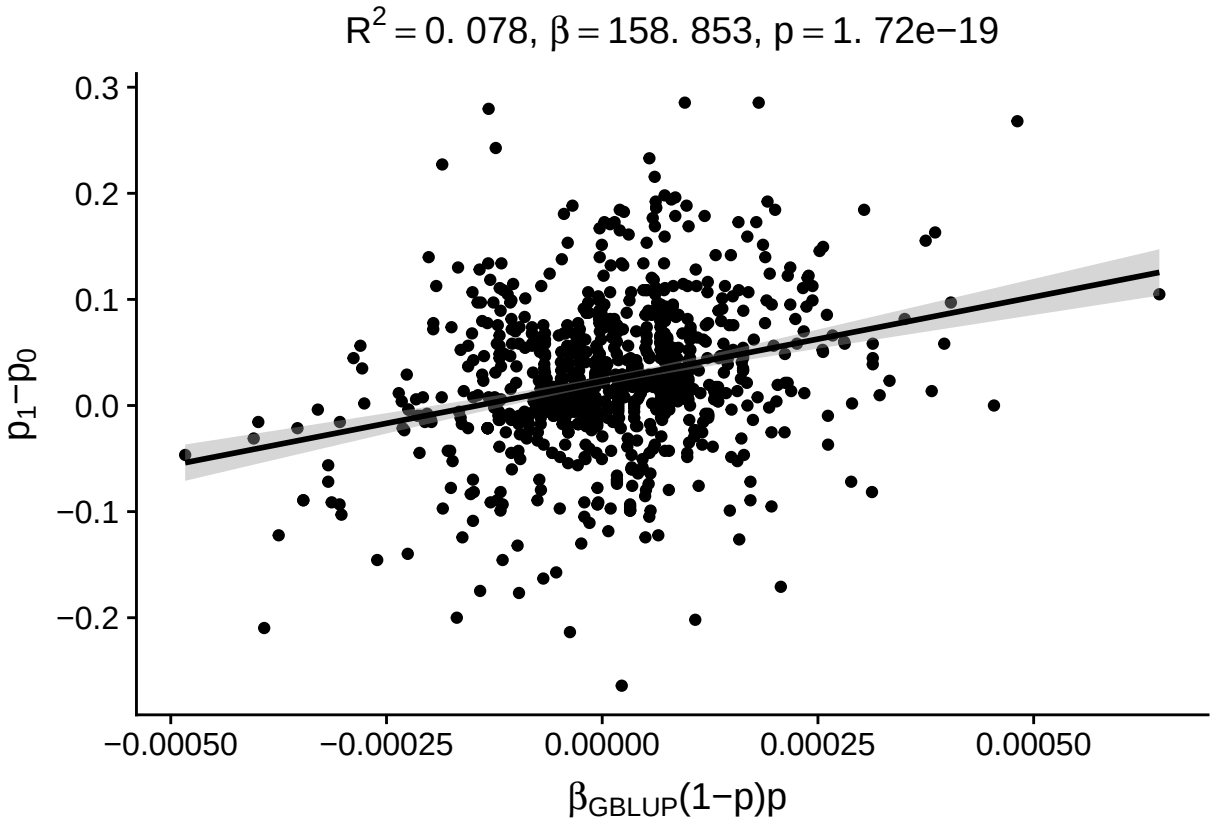
This is because the simple calculation of mean fitness between genotypes carrying alternative and reference alleles, $s = w_{11} - w_{00}$, which is essentially what a marginal GWA provides, is directly proportional to the change in frequency of the allele in the population, no matter whether this is due to causal effects in fitness of the allele or linked effects of other alleles. The equation is $\delta p = s \times p(1-p)$ (Charlesworth & Charleswroth 2010, page 53).

Below is the comparison of a marginal GWA β estimate (aka. realized $s = w_{11} - w_{00}$), which match perfectly.

```
pl5<-ggdotscolor( b*(1-p0)*p0 ,p1-p0,
              xlab = TeX("$\\beta (1-p)p$"),
              ylab=TeX("p_1-p_0")
              ) %>% moiR::addggregression()
pl6<-ggdotscolor( c(bg)*(1-p0)*p0 ,p1-p0,
              xlab = TeX(" $\\beta_{GBLUP} (1-p)p$"),
              ylab=TeX("p_1-p_0")
              ) %>% moiR::addggregression()
pl5
```




```
pl6
```

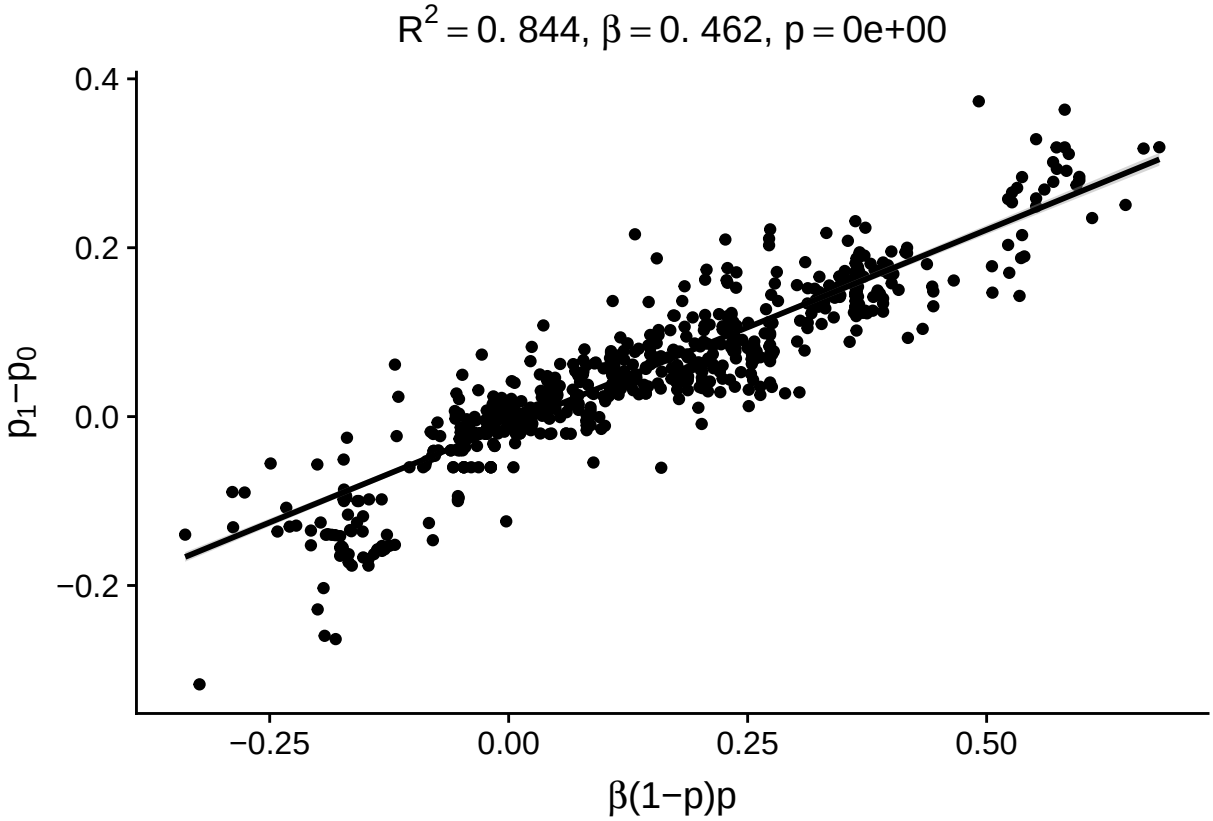


Extrapolability to other groups of genotypes?

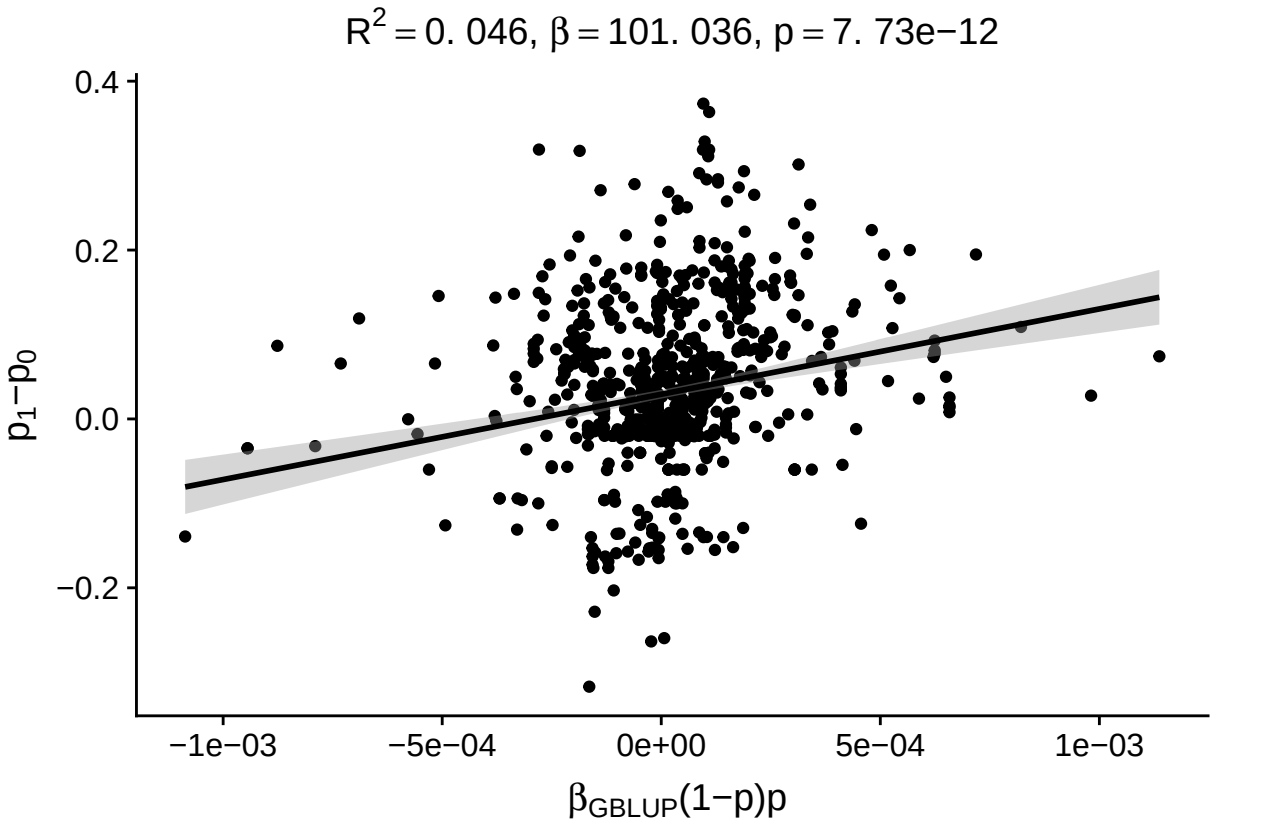
In order to check whether the previous two predictions of allele frequency are just dependent on this population of 515 genotypes or not, the individual based simulations are run in a subset population of over 50 genotypes, instead of 515 genotypes. These 50 were not randomly selected (otherwise allele frequencies would not be much alterred), but we chosen to be only from Spain – an area where populations are known to be distinct genetic groups.

```
set.seed(1)
# Just to try with a subset of the species
# sampled<-sample(1:515,250)
# sampled<-1:100
data(acc515)
rownames(acc515) <- acc515$id
# acc515_backup<-acc515
acc515<-acc515[as.character(genomes$fam$sample.ID),]

sampled<-which(acc515$country =="ESP")
# sampled<-which(acc515$country =="ESP") %>% sample(.,10)
sampled<-which(acc515$country =="ESP") %>% sample(.,50)
# sampled<-which(acc515$country =="SWE")  %>% sample(.,50)
# sampled<-which(acc515$country =="USA") %>% sample(.,10)

p0_=apply(X1[sampled,],2,mean)
p1_<-sapply(1:ncol(X1[sampled,]), function(i){
        sample(x = X1[sampled,i],
           size = 515,
```

```
            prob = c(y[sampled]),
            replace=T
            ) %>% mean
          }
        )
# ggdotscolor(b,p1_-p0_,
#               ylab=TeX("p_1 - p_0"),
#               xlab=TeX("$\\beta$ from marginal GWA") ) +
#                geom_hline(yintercept=0,color="grey") +
#                geom_vline(xintercept=0,color="grey")
# ggdotscolor(c(bg),p1_-p0_,
#               ylab=TeX("p_1 - p_0"),
#               xlab=TeX("$\\beta_{GBLUP}$ from GBLUP GWA")) +
#                geom_hline(yintercept=0,color="grey") +
#                geom_vline(xintercept=0,color="grey")
pl7<-ggdotscolor( b*(1-p0_)*p0_ ,p1_-p0_,
              xlab = TeX("$\\beta (1-p)p$"),
              ylab=TeX("p_1-p_0")
              ) %>% moiR::addggregression()
pl8<-ggdotscolor( c(bg)*(1-p0_)*p0_ ,p1_-p0_,
              xlab = TeX(" $\\beta_{GBLUP} (1-p)p$"),
              ylab=TeX("p_1-p_0")
              ) %>% moiR::addggregression()

pl7
```



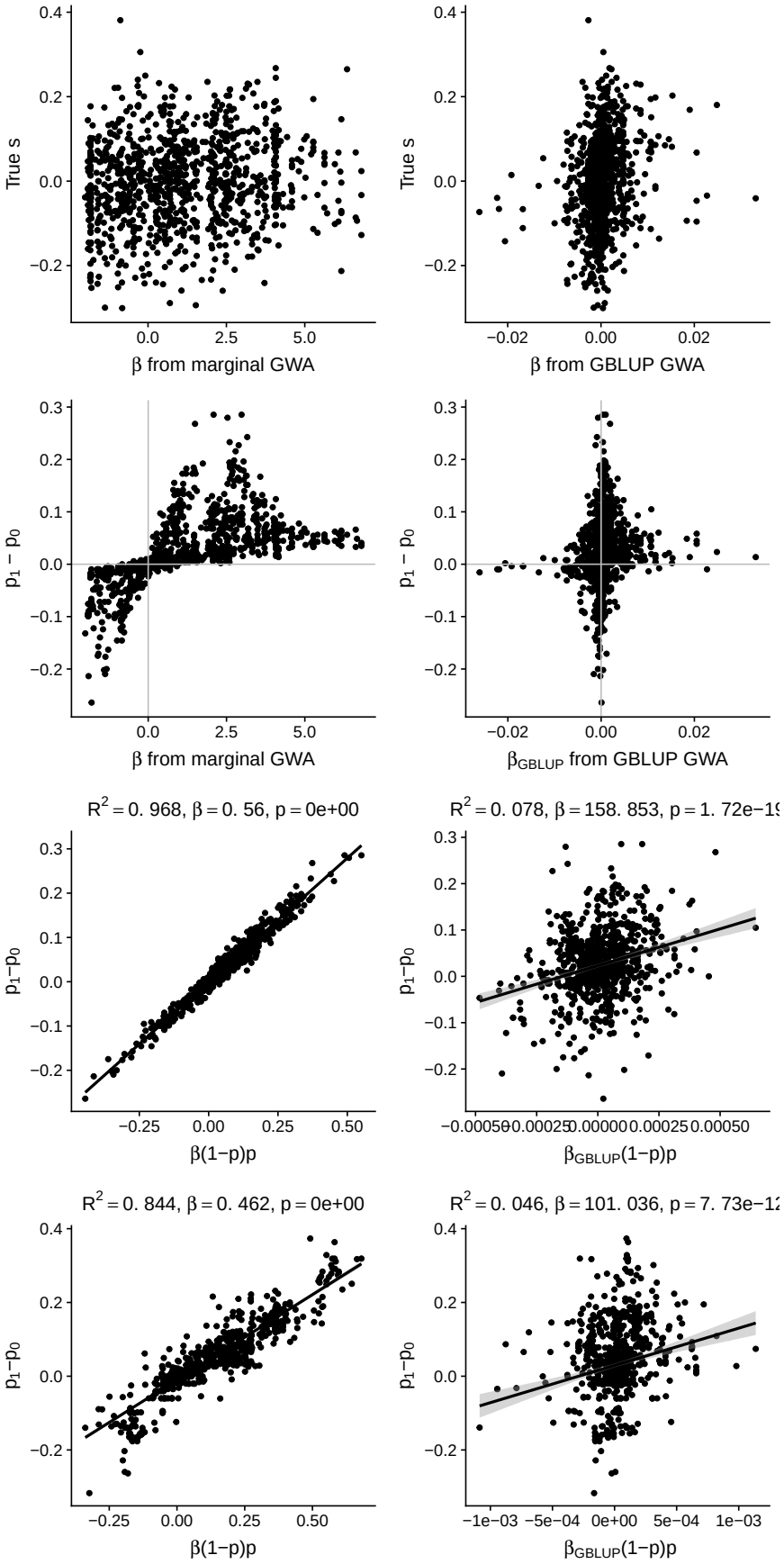
```
p18
```



This shows that the rankings of SNPs from marginal GWA are still useful to understand selection in another population with a different genetic makeup (both in diversity and linkage).

All figures together

```
plot_grid(p11,p12,p13,p14,p15,p16,p17,p18,ncol=2)
```


True s
β from marginal GWA
True s
β from GBLUP GWA
p1 − p0
β from marginal GWA
p1 − p0
βGBLUP from GBLUP GWA
R2 = 0. 968, β = 0. 56, p = 0e+00
p1−p0
β(1−p)p
R2 = 0. 078, β = 158. 853, p = 1. 72e−19
p1−p0
βGBLUP(1−p)p
R2 = 0. 844, β = 0. 462, p = 0e+00
p1−p0
β(1−p)p
R2 = 0. 046, β = 101. 036, p = 7. 73e−12
p1−p0
βGBLUP(1−p)p

NOTES

NOTES: To make sure that that this extrapolability is not because the GBLUP GWA is not very powerful, we also check whether a ridge regression or lasso (a likelihood approach analog of BSLMM) works better, but still only marginal GWA measured in the 515 population correlates significantly with allele frequency changes.

```
#############################################################################
# We can also use ridge regression, in analogy to the
# Bayesian Sparse Linear Mixed Model -- this is a multiloci gwa, an is probably the best estimating tru
# library(MASS)
# bridge<-lm.ridge(y~X1)
# b_ridge=bridge$coef
# ggdotscolor(b_ridge,s,
#             xlab = TeX("$\\beta$ from Multilocus Ridge GWA"),
#             ylab = TeX("$s$")) %>% moiR::addggregression()
# ggdotscolor( c(b_ridge)*(1-p0_)*p0_ ,p1_-p0_,
#              xlab = TeX(" $\\beta_{ridge} (1-p)p$"),
#              ylab=TeX("p_1-p_0")
#              ) %>% moiR::addggregression()
```

NOTES:Another GWA approach would be to simply fit the

```
#############################################################################
# PC<-princomp(K)
# # plot(PC)
# PCs<-PC$scores[,1:10]
# b<-gwamarg(y,X1)
# b_pca<-gwapca(y,X1,PCs)
# cor(b,b_pca)
#
# ggdotscolor(b_pca,s,
#             xlab = TeX("$\\beta_{pca}$ from PCA corrected GWA"),
#             ylab = TeX("$s$")) %>% moiR::addggregression()
# ggdotscolor( c(b_pca)*(1-p0_)*p0_ ,p1_-p0_,
#              xlab = TeX(" $\\beta_{pca} (1-p)p$"),
#              ylab=TeX("p_1-p_0")
#              ) %>% moiR::addggregression()
```

NOTES: We have also randomized the SNP direction to remove LD structure, and the conclusions hold the same

```
# # Randomize direction of SNPs
# Xs2<-apply(Xs,2,function(i) i* sample(c(-1,1),size=1,prob = c(1,1)))
#
#
# # LD r2
# V=cor(Xs2)
# # dim(V)
# hist(V,xlab="Linkage disequilibrium (r2)", main="",col="black",breaks=100)
#
# # Calculate phenotypes
# y<-(Xs2 %*% s)
# h2 <- 0.9
```

```
# y <- y + rnorm(515,mean=0,sd=sqrt((1-h2)/h2*var(y)))
# y<-y/mean(y)
# y[y<0]<-0
# # hist(y, xlab="Relative fitness per individuals (w)",main="",col="black",breaks=50)
#
# # GWA marg
# marglm<-gwamarg(y, Xs2)
# b=marglm
#
# # GWA after kinship
# blup<-kin.blup(data=data.frame(y=y, Genotype=genomes$fam$sample.ID) ,
#                pheno="y",
#                geno="Genotype",
#                K=K)
# blup$Vg / (blup$Vg+blup$Ve)
# myres<-blup$resid
# bg<-gwamarg(myres, Xs2)
#
# # Allele frequency change
# Xs2=Xs2+1
# X01=(Xs2)/( (Xs2)+1e-10)
#
# p0=apply(X01,2,mean)
# # hist(p0,xlab="alternative allele frequency",main="",col="black", breaks=100,xlim=c(0,1))
#
# p1<-sapply(1:ncol(X01), function(i){
#         sample(x = X01[,i],
#            size = 515,
#            prob = c(y),
#            replace=T
#            ) %>% mean
#         }
#       )
#
# ggdotscolor(b,s,
#             xlab = TeX("$\\beta$ from marginal GWA"),
#             ylab = TeX("True $s$"))
# ggdotscolor(c(bg),s,
#             xlab = TeX("$\\beta_{GBLUP}$ from GBLUP GWA"),
#             ylab = TeX("True $s$"))
# (ggdotscolor(b,p1-p0,
#              ylab=TeX("p_1 - p_0"),
#              xlab=TeX("$\\beta$ from marginal GWA")) %>% moiR::addggregression())
# (ggdotscolor(c(bg),p1-p0,
#              ylab=TeX("p_1 - p_0"),
#              xlab=TeX("$\\beta_{GBLUP}$ from GBLUP GWA")) %>% moiR::addggregression())
```