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

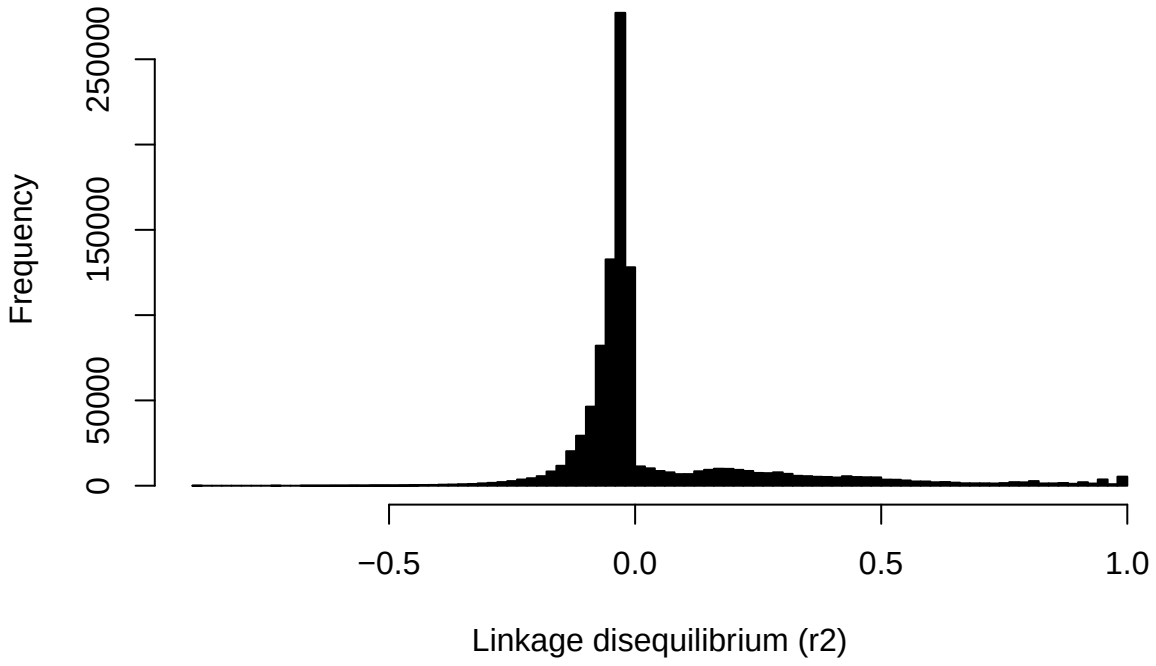
Moi Exposito-Alonso

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. Much of the methods and discussions below have already been addressed 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.

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[,]
Xs<-X-1 # transform to -1 +1 format

# LD r
V=cor(Xs)
hist(V,xlab="Linkage disequilibrium (r2)", main="",col="black", breaks=100)
```

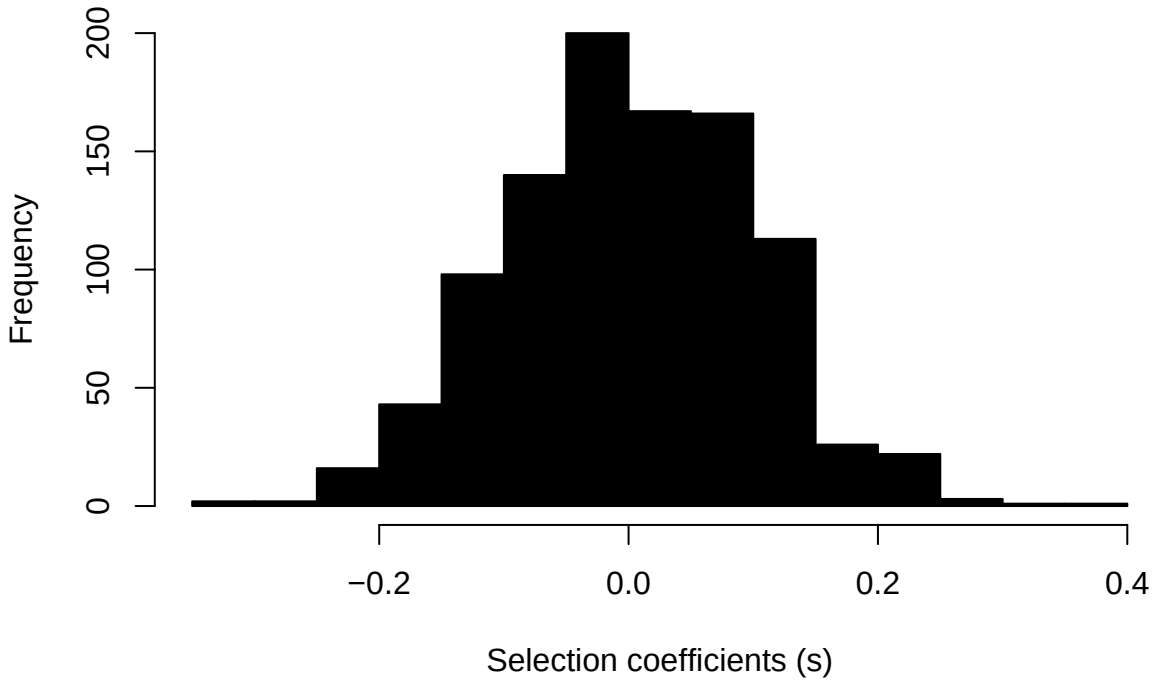


We can see that alternative (minor) alleles are on average in positive linkage disequilibrium

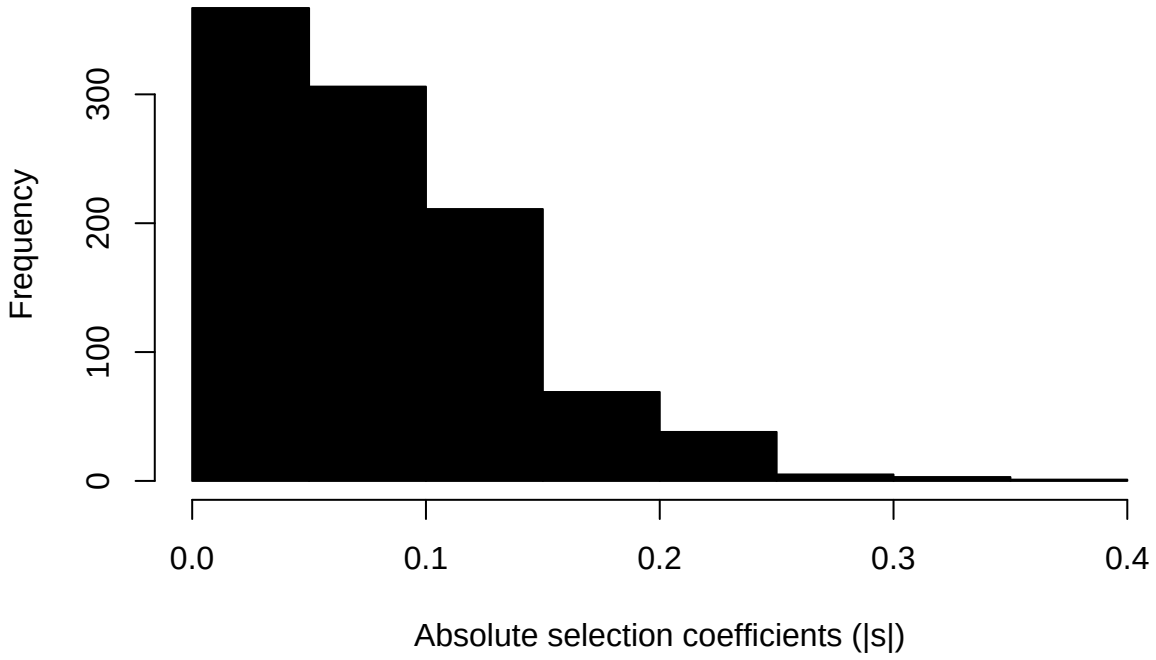
Simulate selection coefficients and fitness

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

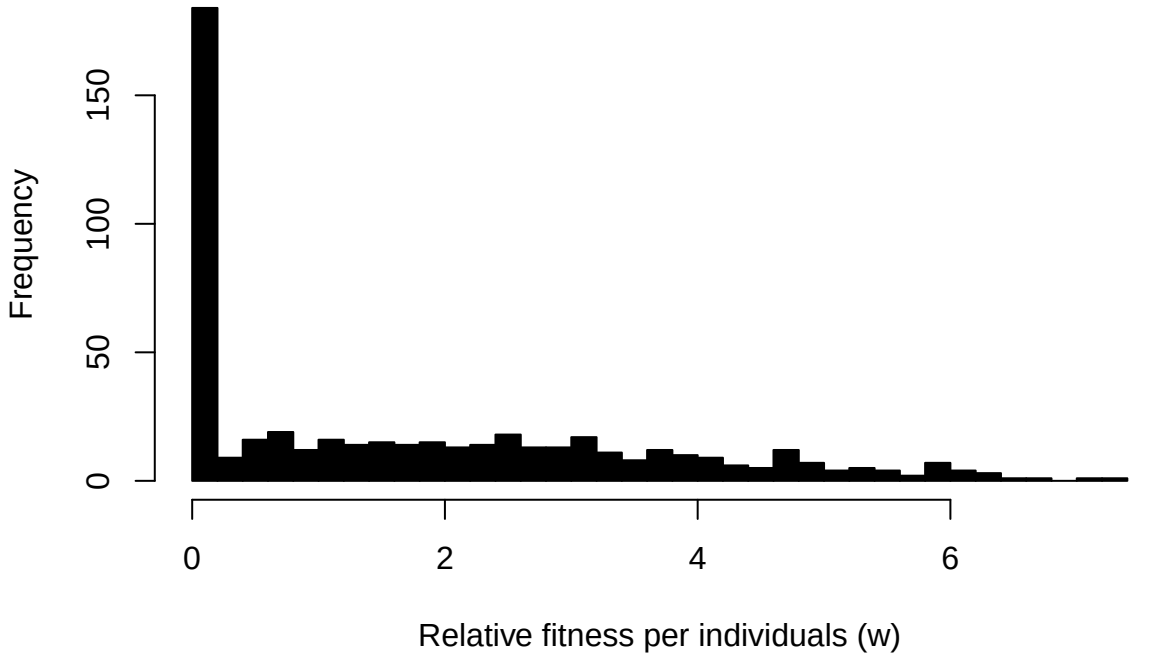
```
s<-rnorm(1000,mean = 0,sd = 0.1)
hist(s, xlab="Selection coefficients (s)",main="",col="black")
```




```
hist(abs(s), xlab="Absolute selection coefficients (|s|)",main="",col="black")
```




```r
y<-(Xs %*% 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)
```



This distribution very much looks like the distribution of fitness of Arabidopsis accessions in the drought treatment of the Spainish field site in Exposito-Alonso 2018 bioRxiv.

Example of association with marginal GWA

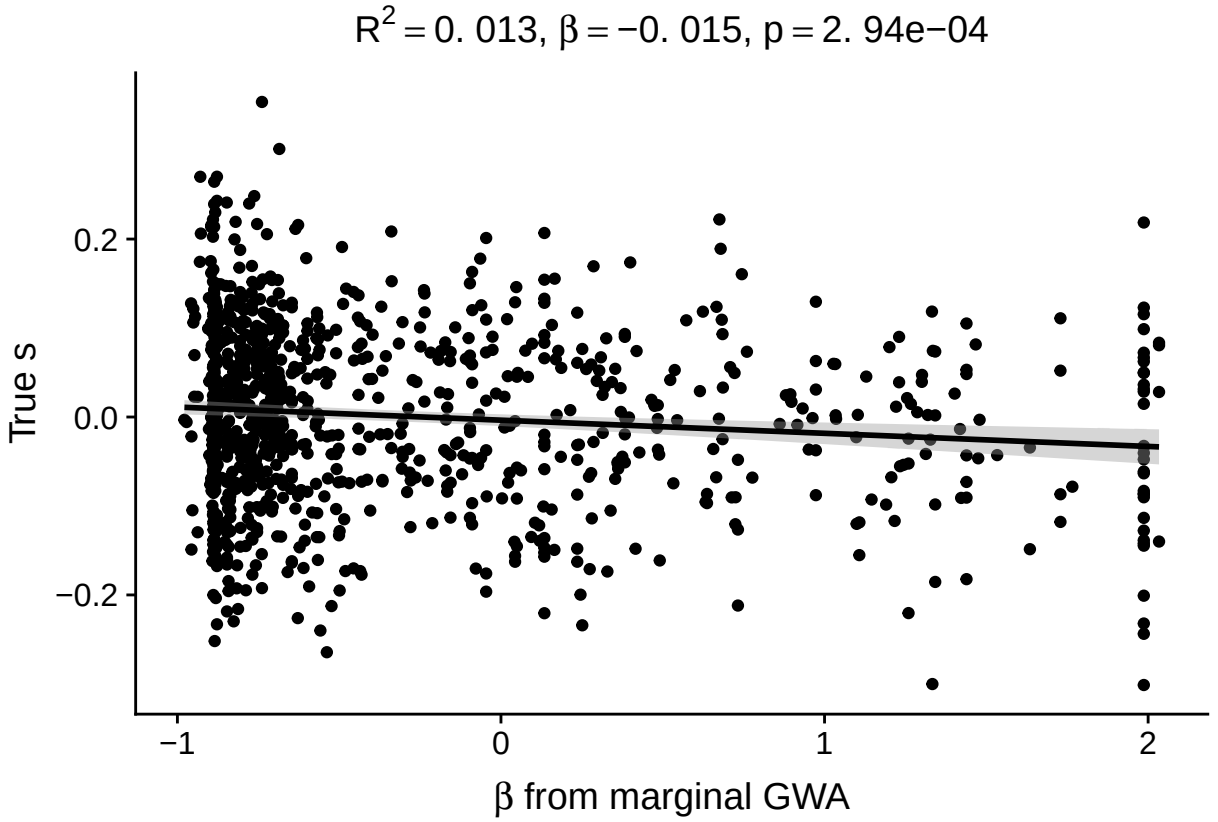
One approach used in Exposito-Alonso (2018) to understand the total selection acting on a genetic variant is simply the difference in relative fitness between genotypes carrying alternative and reference alleles, $\beta = s = w_{11} - w_{00}$, what we called allelic selection differential (in analogy 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, Xs)
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)
ggdotscolor(b,s,
            xlab = TeX("$\\beta$ from marginal GWA"),
            ylab = TeX("True $s$")) %>% moiR::addggregression()
```



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.

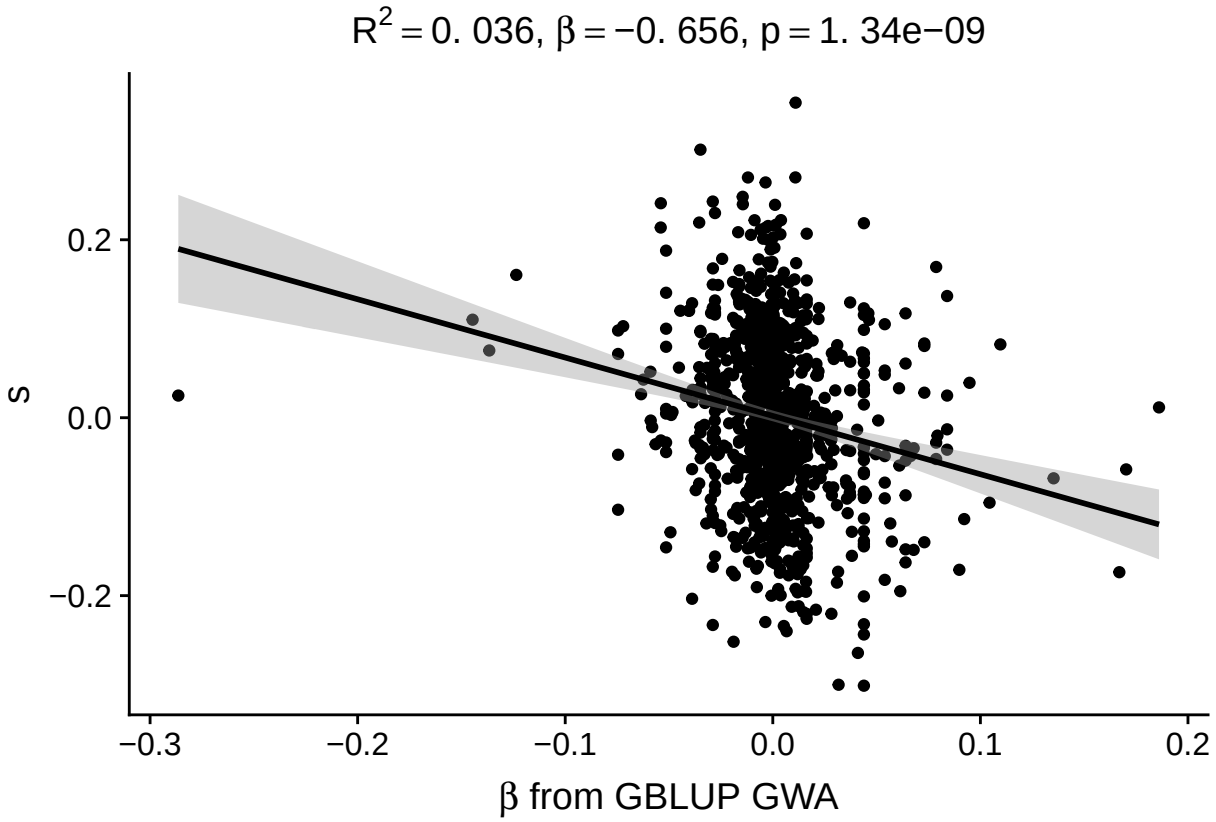
```
library(rrBLUP)
```

```
## 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)
blup$Vg / (blup$Vg+blup$Ve)
```

```
## [1] 0.7121466
```

```
myres<-blup$resid
b_gblup<-gwamarg(myres, Xs)
bg=b_gblup
ggdotscolor(bg,s,
            xlab = TeX("$\\beta$ from GBLUP GWA"),
            ylab = TeX("$s$")) %>% moiR::addggregression()
```




```
#############################################################################
# ## One step using the standard gwa in package, although do not povide effect estimate
# gwanormal<-GWAS(pheno,geno,plot = F,K=K)
# byu<-gwanormal$y # this is the proportion of variance explained
# maf=apply((Xs+1)/( (Xs+1)+1e-10) ,2,mean)
# direction<- (cor(y,X)/abs(cor(y,X)))
# bk<- sqrt(byu/(2*maf*(1-maf))) * direction # Transform percentage of variance explained to effect siz
# ggdotscolor(c(bk),s,
#             xlab = TeX("$\\beta$ from GBLUP GWA"),
#             ylab = TeX("$s$")) %>% moiR::addggregression()

#############################################################################
## We can also use ridge regression, in analogy to the Bayesian Sparse Linear Mixed Model -- this is a r
# library(MASS)
# bridge<-lm.ridge(y~Xs)
# b_ridge=bridge$coef
# ggdotscolor(b_ridge,s,
#             xlab = TeX("$\\beta$ from Multilocus Ridge GWA"),
#             ylab = TeX("$s$")) %>% moiR::addggregression()
```

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).

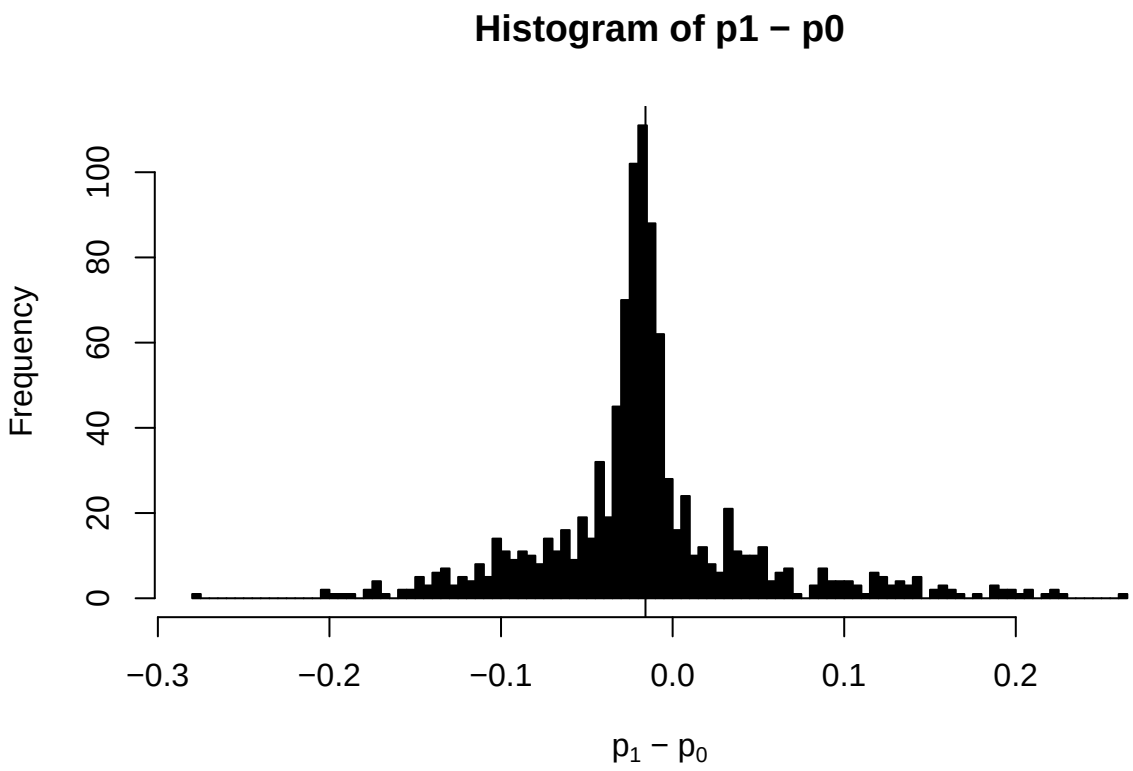
The relationship of β with allele population frequency change

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).

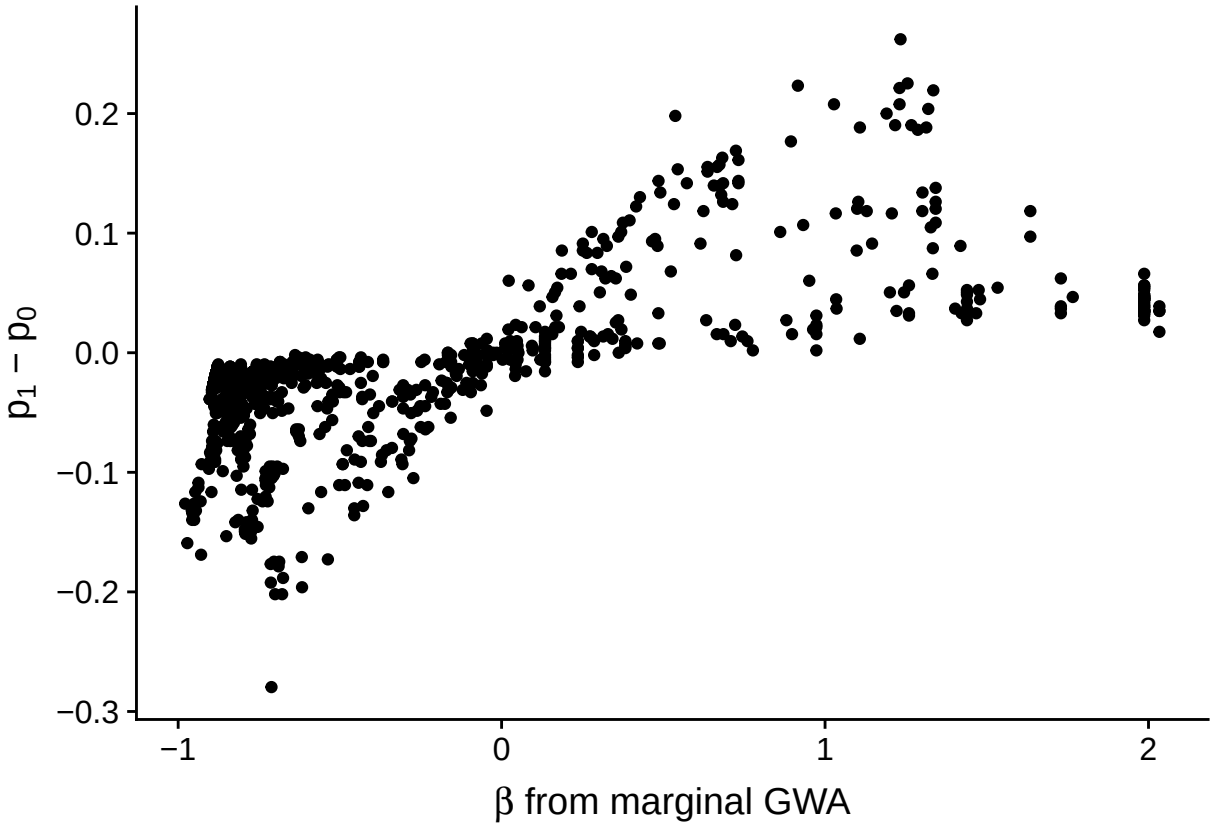
Regardless of the above complications, when the aim of characterizing the impact of selection on a population with high linkage (as the highly diverse species-wide Arabidopsis 515 panel), a marginal GWA might capture what we aim: how much natural drives frequency changes in a selection event. Particularly, 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.

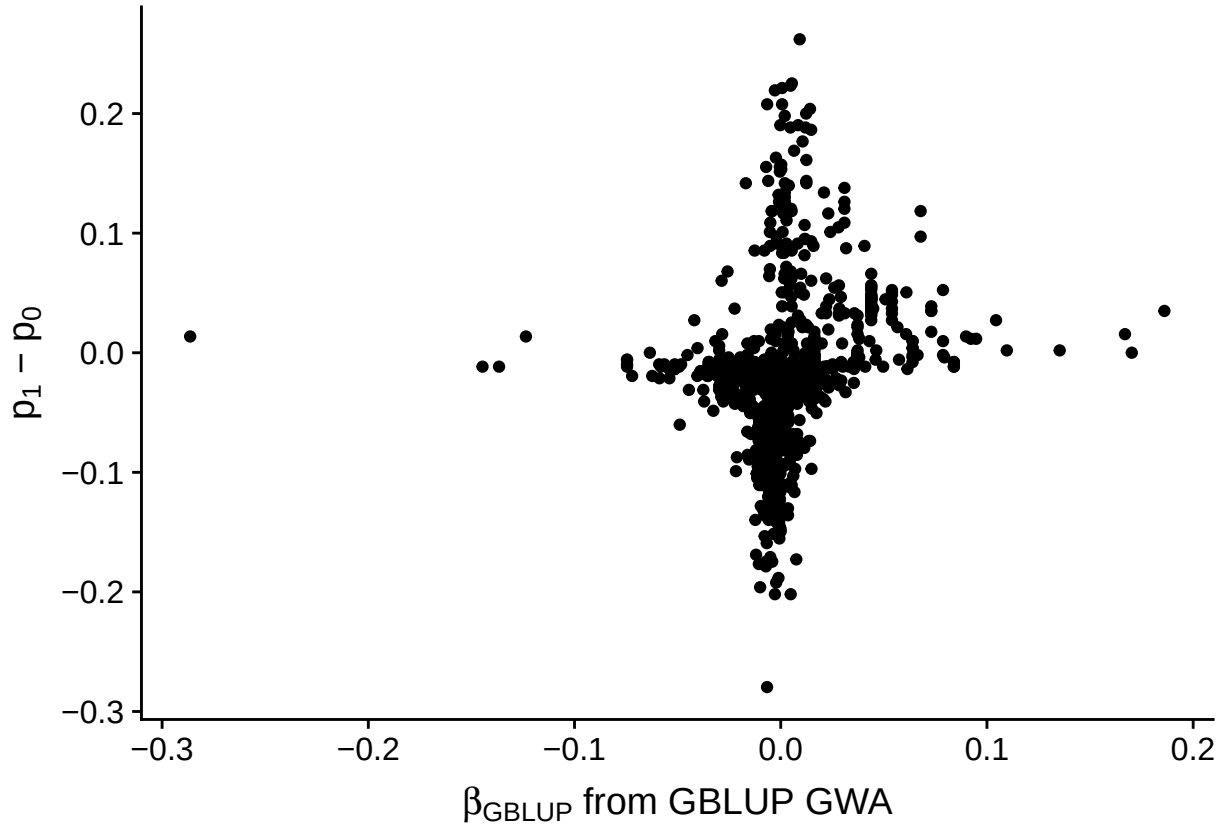
```
X01=(X/(X+1e-10))

p0=apply(X01,2,mean)

p1<-sapply(1:ncol(X01), function(i){
        sample(x = X01[,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))
```




```
ggdotscolor(b,p1-p0,
            ylab=TeX("p_1 - p_0"),
            xlab=TeX("$\\beta$ from marginal GWA")) # %>% moiR::addggregression()
```



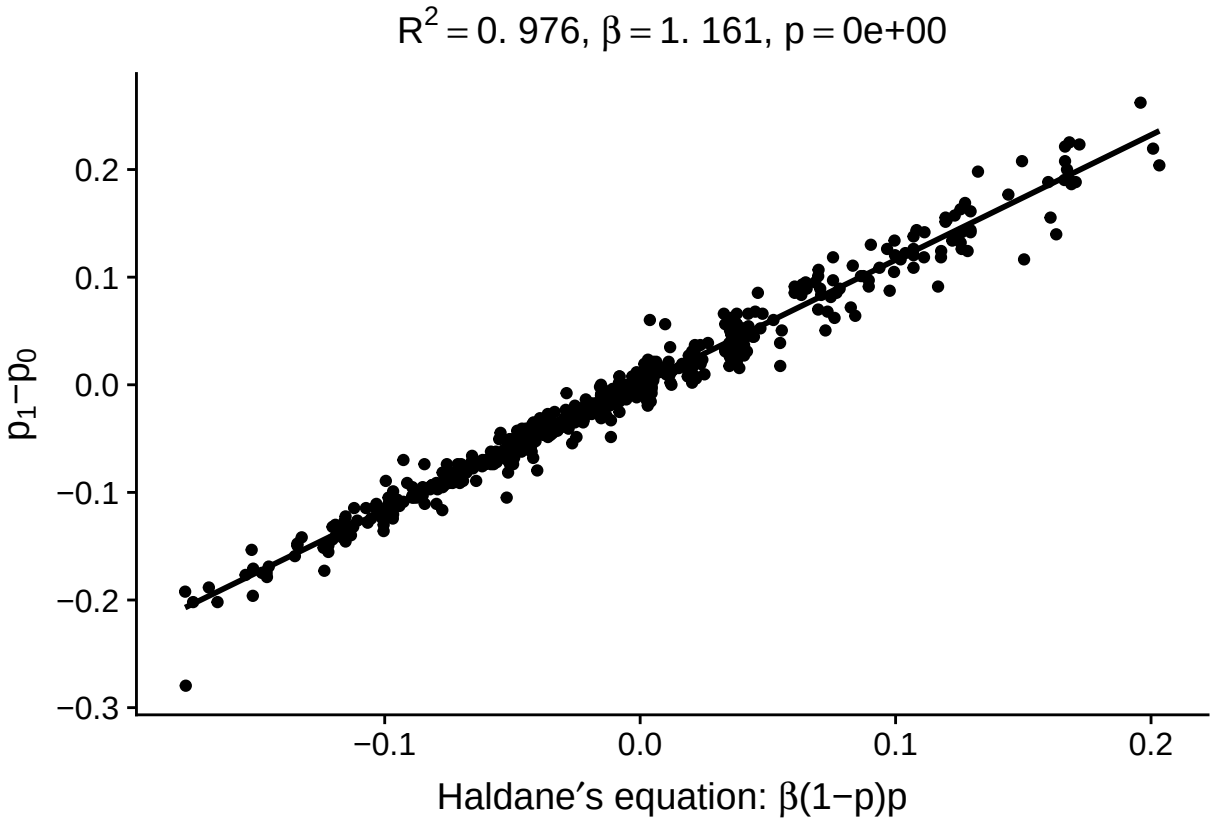
```
ggdotscolor(c(bg),p1-p0,
            ylab=TeX("p_1 - p_0"),
            xlab=TeX("$\\beta_{GBLUP}$ from GBLUP GWA")) # %>% moiR::addggregression()
```



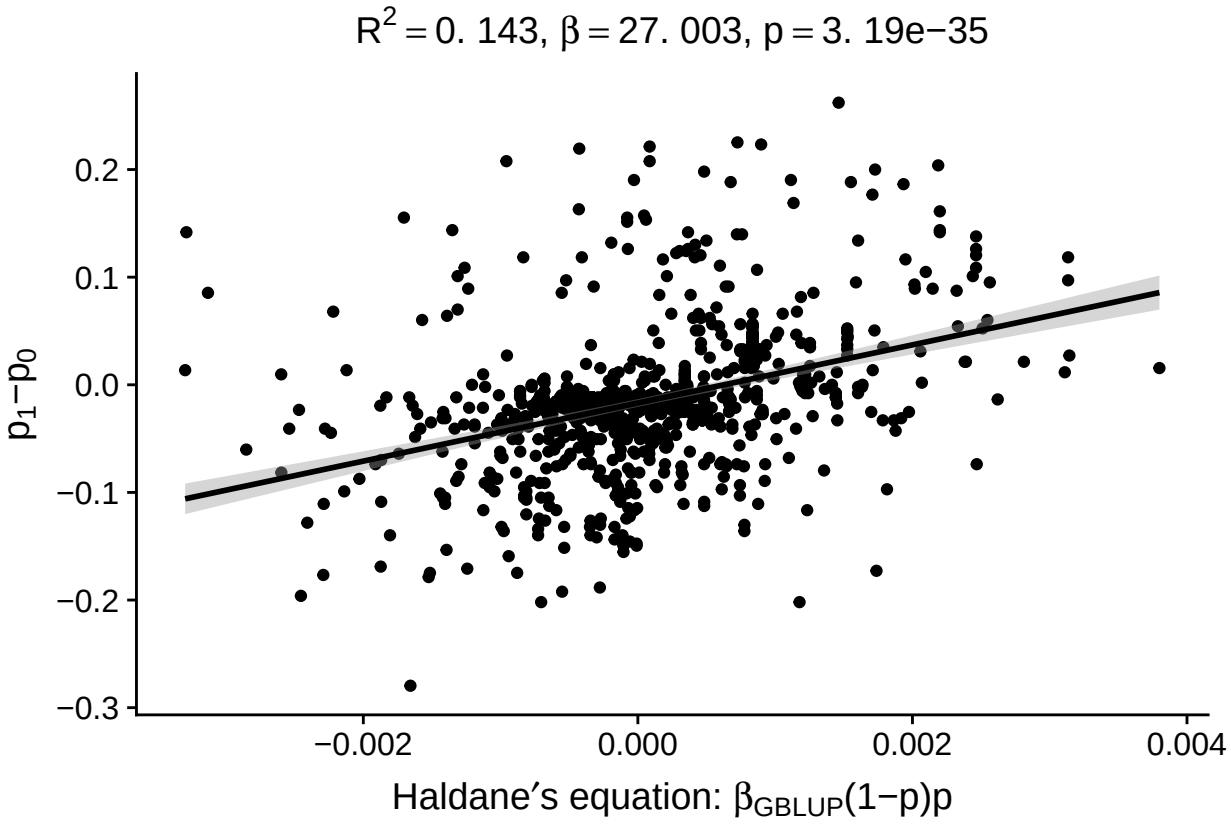
This is because the simple calculation of mean fitness between genotypes carrying alternative and reference alleles, $s = w_{11} - w_{00}$, which is essentially what a marginal GWA provides, is directly proportional to the change in frequency of the allele in the population, no matter whether this is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

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

```
ggdotscolor( b*(1-p0)*p0 ,p1-p0,
            xlab = TeX("Haldane's equation: $\\beta (1-p)p$"),
            ylab=TeX("p_1-p_0")
            ) %>% moiR::addggregression()
```



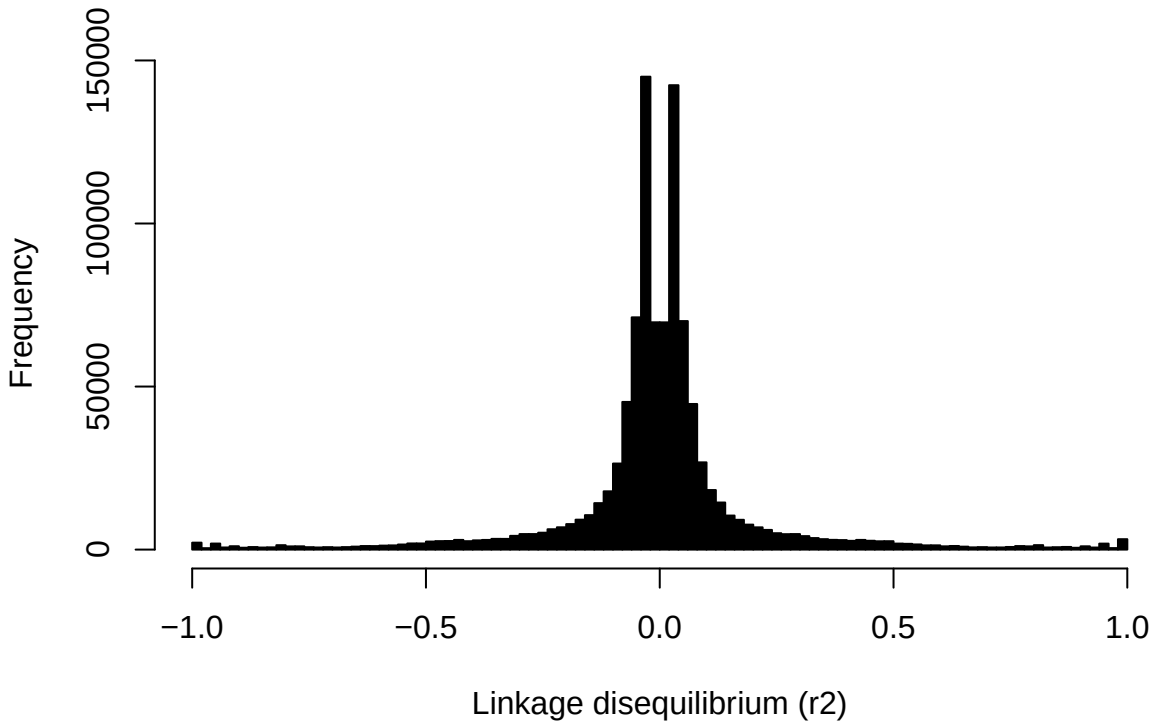

```
ggdotscolor( c(bg)*(1-p0)*p0 ,p1-p0,
             xlab = TeX("Haldane's equation: $\\beta_{GBLUP} (1-p)p$"),
             ylab=TeX("p_1-p_0")
             ) %>% moiR::addggregression()
```



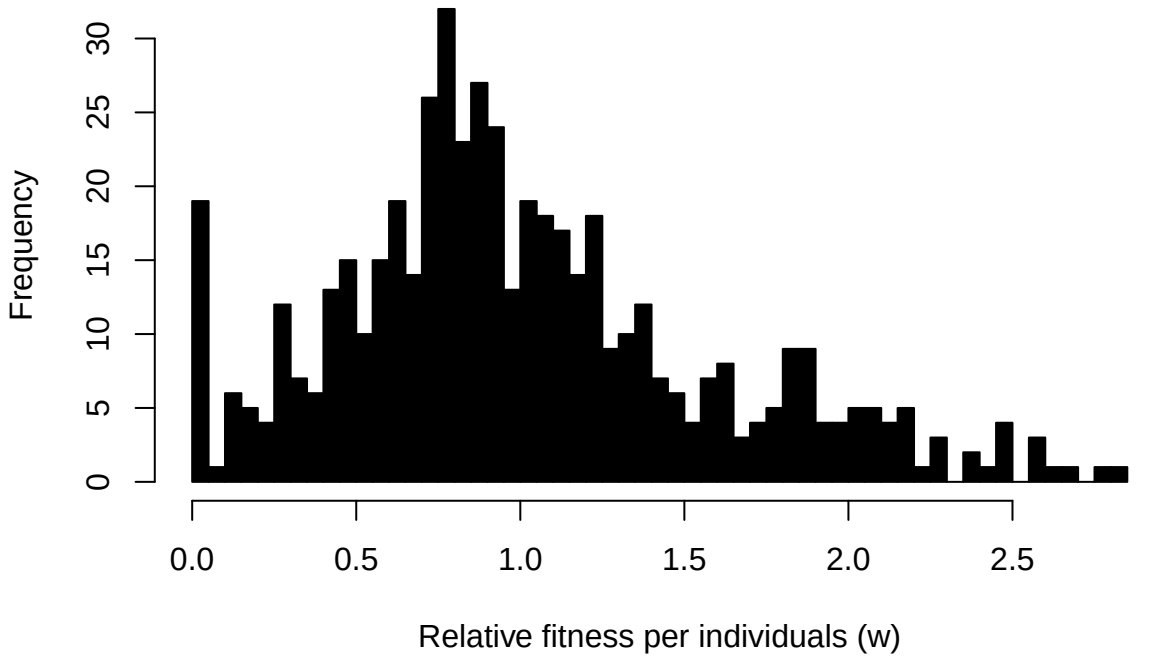
Artificially remove linkage structure and re-calculate

By shuffling the direction of the alternative alleles, we can have a random LD structure.

```
# 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)
```

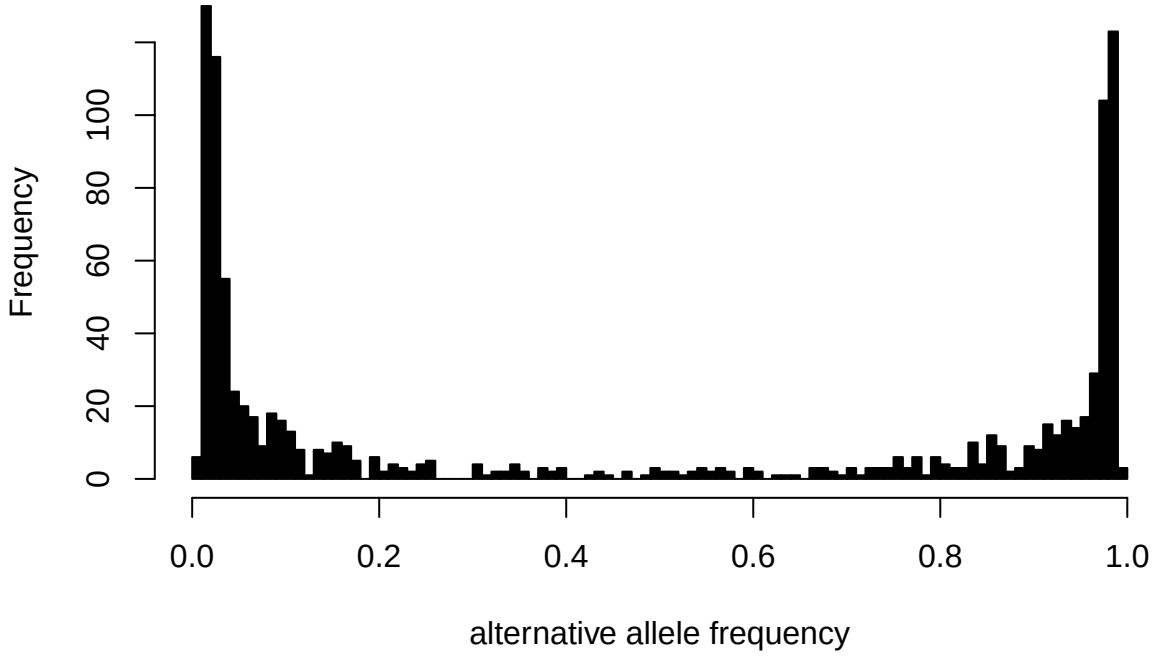



```
# GWA marg
marglm<-gwamarg(y, Xs2)
```

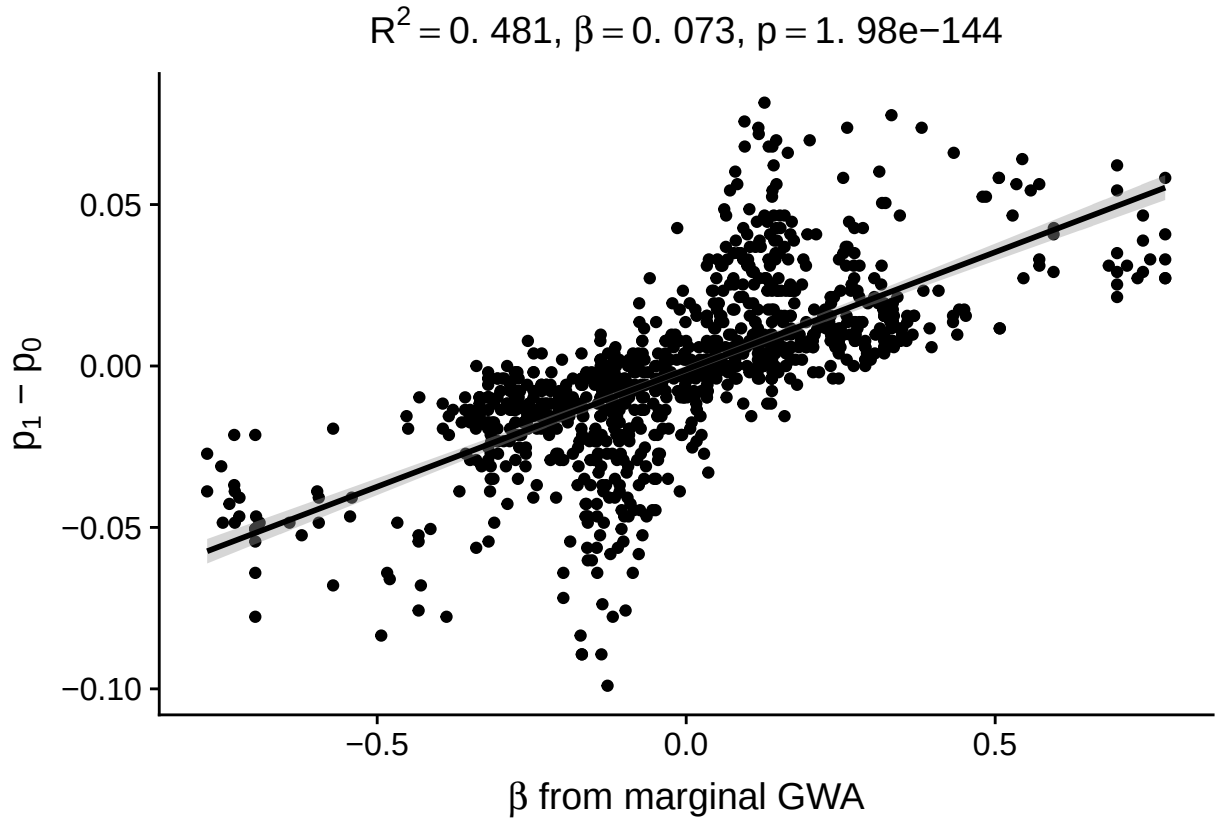
```
b=marglm

# 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)
```

```
## [1] 0.891565
```

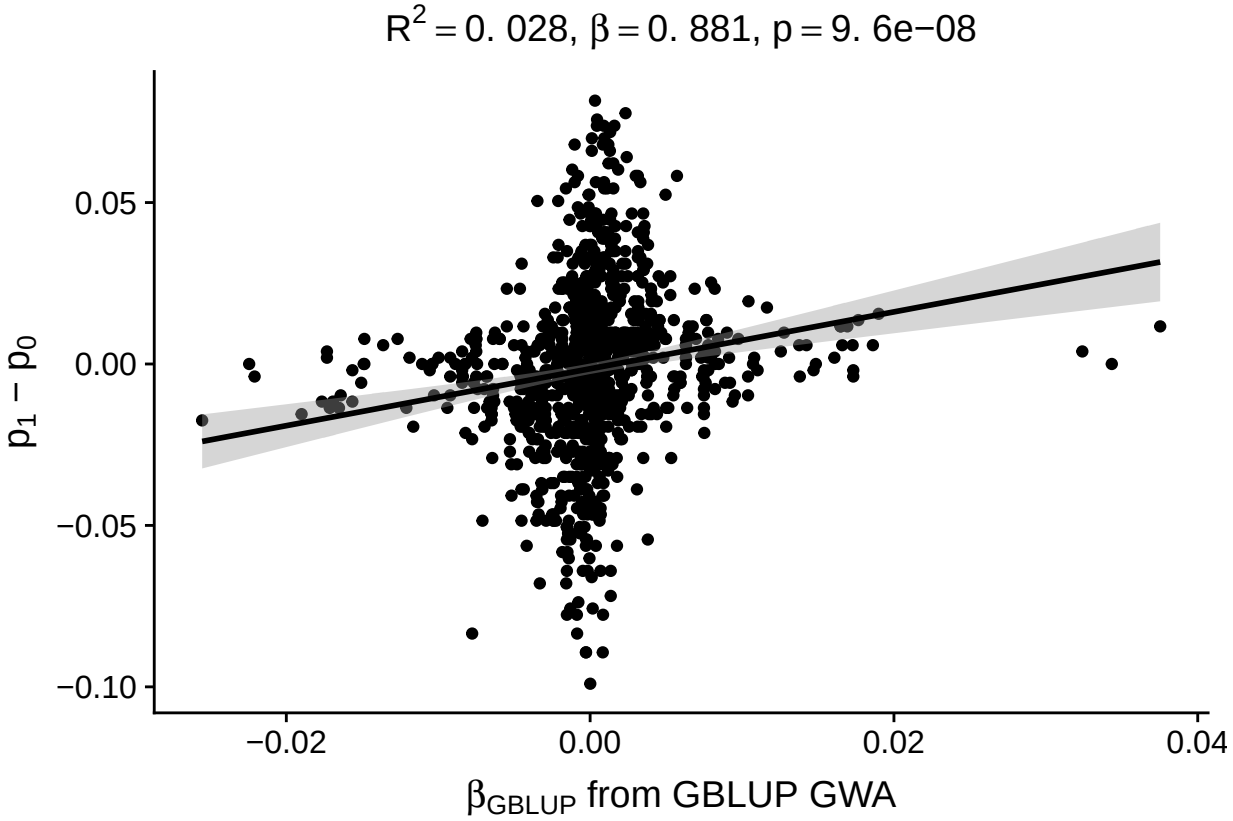
```
myres<-blup$resid
bg<-gwamarg(myres, Xs2)

# 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,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())
```



We observe the same pattern. Because what reflects best allele frequency change is ultimately the difference of genotypes of alternative alleles.