Means and Variances

Gustavo de los Campos (gustavoc@msu.edu)

October 2nd, 2025

1) Single-locus model

This section presents the math underlying a single-locus model.

Throughout the document we use a standard quantiative genetics model in which a quantitative phenotype (Y) is represented as the sum of a genetic value (G) and an environmental effect (E)

$$Y = G + E.$$

Above, G is the expected value of the phenotype given the genotype of the individual; therefore, $[E] = 0$.

In a bi-allelic locus with alleles $(A_1; A_2)$, for a diploid organism, we can have three genotypes; therefore, G can take on three values

- $E[Y|A_1A_1] = G(A_1A_1)$,
- $E[Y|A_1A_2] = G(A_1A_2)$, and
- $E[Y|A_2A_2] = G(A_2A_2)$.

We will study the statistical/quantitative genetics properties of this model (the means, the variances, and heritability) using two equivalent approaches. First, we will derive the mean and variance under Hardy-Weinberg assumptions (HW). Second, we will derive the same parameters using a regression approach. Both approaches are equivalent; however, the regression approach will pave the way towards the models commonly used in GWAS and genomic prediction.

1.1) Means and Variance under HWE

Consider a single locus with two alleles $(A_1\ ;\ A_2)$ and let p denote the frequency of allele A_1, that is $P(A_1) = p$, implying, $P(A_2) = 1 - p = q$. For this locus we have three possible genotypes (A_1A_1, A_1A_2, A_2A_2), Table 1 gives the expected genotype frequencies under Hardy-Weinberg Equilibrium (HWE).

Table 1: Genotypes and expected frequencies under Hardy-Weinberg Equilibrium

Genotype	HWE Frequency
A_1A_1	p^2
A_1A_2	$2pq$
A_2A_2	q^2

Genetic values

As noted eaerlier, since we have three possible genotypes we can have three genetic values (or means):

- $G(A_1A_1)$,
- $G(A_1A_2)$, and
- $G(A_2A_2)$.

Following the parameterization used in Falconer & Mackay (1996), for the single locus model we represent the three possible genetic values using

- $G(A_1A_1) = \mu + a$,
- $G(A_1A_2) = \mu + d$, and
- $G(A_2A_2) = \mu - a$.

These three means can be simplified by subtracting μ such that

- $G(A_1A_1) = a$,
- $G(A_1A_2) = d$, and
- $G(A_2A_2) = -a$.

With this parameterizations, we model the three means (all represented as deviations from μ) using two parameters (a, and d); this is represented graphically in Figure 1 which includes four cases: additive model ($d = 0$), partial dominance ($|d| < |a|$), complete dominance ($d = a$), and over-dominance ($|d| > |a|$).

Technically, a is one-half of the difference between the genetic value of the two homozygous, $a = [G(A_1A_1) - G(A_2A_2)]/2$, the sign of a depends on the allele chosen as reference.

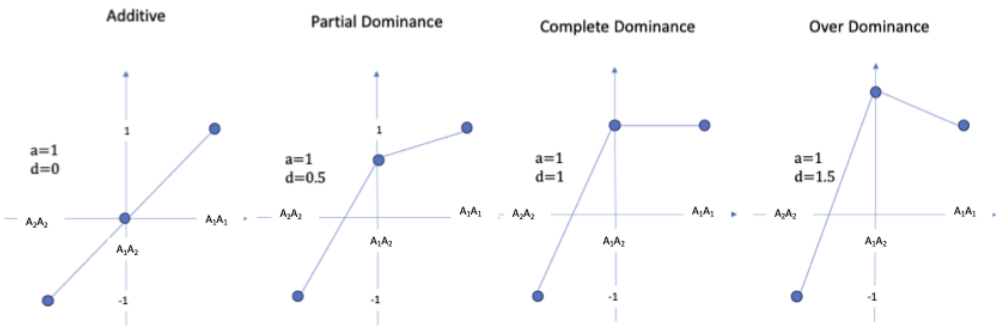

Figure 1: Figure 2: Means in the Single Locus Model

The population mean under Hardy-Weinberg equilibrium

To compute the population mean, we update **Table 1** adding a column for the genetic values. After some simplifications, including using $p^2 - q^2 = p - q$, we get $E(G) = a(p - q) + 2pqd$.

Table 2: Computing the population mean

Genotype	HWE Frequency	G	Freq×G
A_1A_1	p^2	a	p^2a
A_1A_2	$2pq$	d	$2pqd$
A_2A_2	q^2	-a	$-q^2a$
Expected value			$a(p - q) + 2pqd$

Additive action: As noted, if $d = 0$ (strictly additive action) the genetic value of the heterozygous is exactly the average of the genetic value of the two homozygous.

Genetic variance in the single locus model under Hardy-Weinberg Equilibirum

Recall that $\sigma^2_G = Var(G) = E(G^2) - E(G)^2$. We already have an expression for $E(G)$; to derive $E(G^2)$, we can update Table 3 by adding a column for G^2, multiplying by frequency, and taking column-wise sums to obtain the expected values.

Table 3: Computing the Genetic Variance

Genotype	HWE Frequency	G	$Freq \times G$	$Freq \times G^2$
A_1A_1	p^2	a	p^2a	p^2a^2
A_1A_2	$2pq$	d	$2pqd$	$2pqd^2$
A_2A_2	q^2	-a	$-q^2a$	q^2a^2
Expected Value			$a(p-q) + 2pqd$	$a^2(p^2+q^2) + 2pqd^2$

Using $E[G^2] = a^2(p^2+q^2) + 2pqd^2$ and $E(G) = a(p-q) + 2pqd$ in $Var(G) = E(G^2) - E(G)^2$ leads, after simplifications, to the following expression for the genetic variance of a single locus:

$\sigma^2_G = 2pq(a + d[q-p])^2 + (2pqd)^2$.

Broad-sense heritability (in a single locus model)

The broad sense heritability of a trait is the proportion of the phenotypic variance explained by genetic factors. Applying the variance function to each side of the phenotypic equation, $Y = G + E$, we get that phenotypic variance is $Var(Y) = Var(G) + Var(E) + 2Cov(G, E)$. If we assume that genetic and environmental factors are un-correlated, we have

$$Var(Y) = Var(G) + Var(E)$$

In this setting, the proportion of the phenotypic variance explained by genetic factors (or broad sense heritability) is

$$H^2 = \frac{Var(G)}{Var(G) + Var(E)}$$

Where, $Var(G) = 2pq(a + d[q-p])^2 + (2pqd)^2$,

Departures from HWE

Above, we derive the mean, vairance, and heritability under HWE. The same framework can be applied without making HWE assumptions by replacing the predicted probabilities of each of the genotypes with the observed ones, i.e., replacing in the tables presented above p^2, $2pq$ and q^2 with $p(A_1A_1)$, $p(A_2A_2)$, and $p(A_2A_2)$–these probabilities can be estimated from genotype data.

1.2) Regression approach

In this section we recast the model presented above as a linear regression problem.

Let Z_m and Z_m be dummy variables for the maternal and paternal gametes received by a subject ($Z_* = 1$ if A_1, 0 if the gamete carries A_2). The number of copies of the A_1 allele carried by a subject is $Z = Z_m + Z_p$. To implement the regression model matching the parameterization used above we will introduce $X = Z - 1$ such that

- $X = -1$ if the genotype is A_2A_2,
- $X = 0$ if the genotype is A_1A_2, and
- $X = 1$ if the phenotype is A_1A_1

defines a *contrast* between the means of A_1A_1 and A_2A_2. Additionally, we will introduce H, a dummy variable for the heterozygous genotype $H = \{1, if\ heterozygous\ ;\ 0, otherwise\}$.

Using X and H we can write the one-locus model in regression form as follows

$$Y = \mu + Xa + Hd + E$$

The conditional means in this model are

- $E[Y|A_2A_2] = E[Y|X = -1, H = 0] = \mu - a$
- $E[Y|A_1A_2] = E[Y|X = 0, H = 1] = \mu + d$
- $E[Y|A_1A_1] = E[Y|X = 1, H = 0] = \mu + a$

Which is equivalent to the parameterization presented earlier in this document.

Example 1

Consider a single locus under HWE with the following parameters

```
# Parameters
p=0.05;   mu=3; a=1; d=0.8
```

Using the formulas derived above, $E[G] = \mu + a(p - q) + 2pqd$ and $\sigma^2_G = 2pq(a + d[q - p])^2 + (2pqd)^2$, the mean and genetic variance at the locus are:

```
q=1-p

# Formula derived above
mG=mu+a*(p-q)+2*p*q*d
vG=2*p*q*((a+d*(q-p))^2)+(2*p*q*d)^2
message(' Mean: ',mG, ' Variance: ',vG,'.')
```

```
##  Mean: 2.176 Variance: 0.286824.
```

We now simulate a large number of genotypes under HWE, compute the genetic values using the parameters defined above and verify that the realized mean and variance are very close to the ones derived above.

```
# Genotypes and genetic values
n=1e6 # large n
X= rbinom(n=n,size=2,prob=p)-1 # X is -1,0,1
G=ifelse(X== -1,mu-a,ifelse(X==0,mu+d,mu+a))

mG.b=mean(G)
vG.b=var(G)

message('Means: mG=',mG,' ; realized mean=', round(mG.b,4),'.')
```

```
## Means: mG=2.176 ; realized mean=2.1755.
```

```
message('Variance: vG=',vG,' ; realized var.=', round(vG.b,4),'.')
```

```
## Variance: vG=0.286824 ; realized var.=0.2861.
```

Recovering μ, a and d using regression analysis

```r
H=ifelse(X==0,1,0)
fmAD=lm(G~X+H)
tmp=cbind(c(a,d),unname(coef(fmAD)[-1]))
rownames(tmp)=c('a','d')
colnames(tmp)=c('Moments', 'Regression')
knitr:::kable(tmp)
```

	Moments	Regression
a	1.0	1.0
d	0.8	0.8

Average effect of an allele substitution (α)

The average effect of the A_1 allele is defined as the expected change in genetic value of an individual that you would observe if you:

- Sample at random an individual from the population,
- Sample at random one allele from the sampled individual, and
- Replace the sampled allele with the allele in question (e.g., A_1).

For allele A_1 the allele substitution effect is denoted as α_1. Likewise, we can define α_2 to denote the average effect of substituting an allele sample at random from the population with an A_2 allele.

The **average allele substitution effect** is the difference between α_1 and α_2 and is denoted as $\alpha = \alpha_1 - \alpha_2$. This concept plays a central role in the definition of breeding values and the additive variance.

We can also arrive at the average allele substitution effect from a linear regression perspective. Consider regressing the genetic values on allele dosages (i.e., a covaraite that counts the number of copies of a reference allele (e.g., $A1$), that is

$G = \mu + X\alpha + \delta$

The regression coefficient is

$$\alpha = Cov(X, G)/Var(X)$$

.

We can also derive α under HWE using rules of expectations, variances and (co)variances. It can be shown that under HWE $\alpha = Cov(X, G)/Var(X) = a + d(q - p)$.

Regressing Y on X yields an unbiased estimate of α.

Example 2

```r
aHat=fmAD$coef[2]
dHat=fmAD$coef[3]
cbind(c(a,d),c(aHat,dHat))
```

```
##   [,1] [,2]
## X  1.0  1.0
## H  0.8  0.8
```

```r
pHat=mean(X)/2
qHat=1-pHat

alpha=a+d*(q-p)
# compares true alpha, versus two estimates one derived from fmA, the other derived from fmAD
fmA<-lm(G~X)

c(alpha,fmA$coef[2],aHat+dHat*(qHat-pHat))
```

```
##                  X        X
## 1.720000 1.720489 2.520234
```

Remarks

- Regressing a phenotype on a contrast for additive effect (e.g. X={-1;0;1} or $X+1$={0,1,2}) plus a contrast for dominance, H, leads yo estimates for a and d.
- Regressing a phenotype on only allele dosage (either on X$={-1;0;1} or $(X+1)$={0,1,2}) yields an estimate of $\alpha = a + d(q-p)$.
- α can be estimated either using an additive model (`fmA` in the above example) or by estimating a and d from the regression $Y = \mu + Xa + Zd + \varepsilon$ and then plugging those into $\alpha = a + d * (q-p)$.

Additive variance in a single-locus model

The additive variance, σ^2_α is the amount of variance that can be explained by regression on an additive contrast (X). Thus

$$\sigma^2_\alpha = Var(X\alpha) = Var(X)\alpha^2$$

Under HWE $Var(X) = 2pq$; therefore,

$$\sigma^2_\alpha = 2pq\alpha^2.$$

Dominance variance:

Using a sequential decomposition of the genetic variance, we can define the dominance variance as:

$\sigma^2_D = \sigma^2_G - \sigma^2_\alpha = (2pqd)^2$

Single-locus Heritability

As noted earlier, using $Y = G + E$ and assuming there is no (co)variance between G and E, we can define the (single-locus) **broad-sense heritability** as the proportion of the phenotypic variance that can be explained by inter-individual differences in genetic values, that is

$H^2 = Var(G)/Var(Y) = Var(G)/[Var(G) + Var(E)] = \sigma^2_G/[\sigma^2_G + \sigma^2_E]$.

Likewise, using the 'additive model' we can define the **narrow-sense heritability** as

$h^2 = Var(X\alpha)/Var(Y) = \sigma^2_\alpha/[\sigma^2_\alpha + \sigma^2_\gamma + Var(E)]$.

Example 3

Here, we add an environmental effect with a variance such that $H^2 = 0.1$, to the simulation of Examples 1 and 2, and use linear models to estimate narrow and broad sense heritabilities for a single-locus mode. Although we have a sizable dominance effect (0.8), the linear model is able to capture more than 95% of the total genetic variance (compare H^2 with h^2), this happens because the allele frequency is far from 0.5.

```r
H2=0.1
vG=var(G)
vE=vG/H2*(1-H2)
vA=2*p*q*(alpha^2)
vG/(vG+vE)
```

```
## [1] 0.1
```

```r
# 'true' narrow sense heritability
h2=vA/(vG+vE)


c(H2,h2,h2/H2)
```

```
## [1] 0.10000000 0.09823265 0.98232654
```

```r
# Simulating a phenotype
E=rnorm(n=n,sd=sqrt(vE))
Y=100+G+E

# Fitting Additive and A+D models
fmA=lm(Y~X)
fmAD=lm(Y~X+H)

# ANOVA
SSy=sum((Y-mean(Y))^2)
RSSa=sum(residuals(fmA)^2)
RSSad=sum(residuals(fmAD)^2)

H2Hat=(SSy-RSSad)/SSy

h2Hat=(SSy-RSSa)/SSy

cbind('Parameter'=c('H2'=H2,'h2'=h2),'Estimate'=round(c('H2'=H2Hat,'h'=h2Hat),5))
```

```
##     Parameter Estimate
## H2 0.10000000  0.09831
## h2 0.09823265  0.09646
```

2) Two-locus models

In a single locus with two alleles we have three possible genotypes. We model these genetic values using three parameters: intercept, additive and dominance effects, that is $G = \mu + Xa + Hd$. Strictly speaking, dominance is an interaction of alleles within a locus.

When extending this model to two loci, we will need to consider **interactions of alleles between loci (i.e., epistasis)** and, for the analyses of variance, (co)variance of alleles between loci (i.e., linkage disequilibrium).

2.1) A two-locus additive model

Define α_A and α_B as the average effect of an allele substitution at locus A and B, respectively. With this, we can represent the nine means using: $G = \mu + X_A\alpha_A + X_B\alpha_B$, where each X_A={-1 if A_2A_2, 0 if A_1A_2, and 1

if A_1A_1} and X_B{-1 if B_2B_2, 0 if B_1B_2, and 1 if B_1B_1} are additive contrasts for locus A and B, respectively. This additive model approximates the nine means that we can have with a plane, where the slope along the columns and rows are α_A and α_B, respectively.

Table 6: Genetic values for a two-locus additive model

Locus 1/Locus2	B_2B_2	B_1B_2	B_1B_1
A_2A_2	$\mu - \alpha_A - \alpha_B$	$\mu - \alpha_A$	$\mu - \alpha_A + \alpha_B$
A_1A_2	$\mu - \alpha_B$	μ	$\mu + \alpha_B$
A_1A_1	$\mu + \alpha_A - \alpha_B$	$\mu + \alpha_A$	$\mu + \alpha_A + \alpha_B$

We can estimate the above model by regressing a phenotype on additive contrasts for each of the loci (X_A and X_B),

$Y = \mu + X_A\alpha_A + X_B\alpha_B + \varepsilon$

This additive model involves three parameters $\{\mu; \alpha_A; \alpha_B\}$. Recall that in this 2-locus model we have nine means; therefore, we still have six more DFs that can be fitted.

2.2) Two-locus model with additive and dominance effects

Dominance can be defined as interactions of alleles within a locus. Using H_A and H_B as dummy variables for heterozygous genotypes at locus A and B, and letting d_A and d_B be the dominance effects, we can write the nine means as a function of five parameters $\{\mu, a_A, a_B, d_A, d_B\}$: $G = \mu + X_Aa_A + X_Ba_B + H_Ad_A + H_Bd_B$

Table 6: Genetic values for a two-locus system with additive and dominance effects

Locus 1/Locus2	B_2B_2	B_1B_2	B_1B_1
A_2A_2	$\mu - a_A - a_B$	$\mu - a_A + d_B$	$\mu - a_A + a_B$
A_1A_2	$\mu - a_B + d_A$	$\mu + d_A + d_B$	$\mu + \alpha_B + d_A$
A_1A_1	$\mu + a_A - a_B$	$\mu + a_A + d_B$	$\mu + a_A + a_B$

We can estimate the parameters of the model in Table 6 by regressing phenotypes on four contrasts

$Y = \mu + X_Aa_A + X_Ba_B + H_Ad_A + H_Bd_B + \varepsilon$

2.3) Epistatic interactions

The A+D model of Table 6 involves 5 parameters. We can think of this model as one having two additive contrasts (X_A and X_B) and two within-locus interactions between the maternal and paternal alleles (H_A and H_B). Since there are nine means, we can still add four more parameters. What interactions are we still missing? We are missing the interactions of alleles between loci, called epistatic interactions. The following model adds those interactions in the previous regression:

$$Y = \mu + X_Aa_A + X_Ba_B + H_Ad_A + H_Bd_B + (X_A \times X_B)b_{aa} + (X_A \times H_B)b_{ad} + (H_A \times X_B)b_{da} + (H_A \times H_B)b_{dd} + \varepsilon.$$

Above, b_{aa}, b_{ad}, b_{da}, and b_{dd} represent additive-by-additive, additive-by-dominance, and dominance-by-additive interactions, respectively. For a two-locus model we cannot model higher order interactions (e.g., additive-by-additive-by-additive); this becomes possible when more loci are involved because the number of estimable means grows exponentially with the number of loci involved.

Example 4: A model for 2 loci under linkage equilibrium

The following R-code simulates two independent loci and a purely genetic trait ($Y = G$, i.e., no environmental effect). After simulating data, we fit an additive model, an additive + dominance, and the full epistatic model.

```
# Parameters
b=c(mu=10, aA=1,aB=.5,dA=1,dB=.5,aa=.4,dd=0,ad=.3,da=-.3)
pA=0.1
pB=0.15

# Sample size
n=10000

# 0/1/2 GENOTYPES
X1=rbinom(n,size=2,prob=pA)
X2=rbinom(n,size=2,prob=pB)

XA=X1-1 # Mapping from 0/1/2 to -1/0/1
XB=X2-1 # Mapping from 0/1/2 to -1/0/1

HA=as.integer(X1==1)
HB=as.integer(X2==1)

# incidence matrix for the full model
XF=cbind(1,XA,XB,HA,HB,XA*XB,HA*HB,XA*HB,HA*XB)
G=XF%*%b

fmA=lm(G~XA+XB)
fmAD=lm(G~XA+XB+HA+HB)
fmF=lm(G~XA+XB+HA+HB+XA*XB+HA*HB+XA*HB+HA*XB)

round(coef(fmA),5)
```

```
## (Intercept)          XA          XB
##    10.95092     1.73431     0.29918
```

```
round(coef(fmAD),5)
```

```
## (Intercept)          XA          XB          HA          HB
##     9.83757     0.83105     0.13005     1.16957     0.25484
```

```
cbind(b=b,bHat=round(coef(fmF),5))
```

```
##       b bHat
## mu 10.0 10.0
## aA  1.0  1.0
## aB  0.5  0.5
## dA  1.0  1.0
## dB  0.5  0.5
## aa  0.4  0.4
## dd  0.0  0.0
## ad  0.3  0.3
## da -0.3 -0.3
```

2.5) Variance decomposition in the two-locus model

The total genetic variance and the broad-sense heritability can be estimated from the full epistatic model. To illustrate this, we will add environmental effects to the previous simulation, using an environmental variance that generates a broad-sense heritability of 0.5.

Example 5

```
H2=0.5
Y=G+rnorm(n=n,sd=sqrt(var(G)*(1-H2)/H2))
fmFull=lm(Y~XA+XB+HA+HB+XA*XB+HA*HB+XA*HB+HA*XB)

vY=var(Y)
vEHat=sum(residuals(fmFull)^2)/(n-ncol(XF))
vGHat=vY-vEHat
```

2.6) Additive and Non-Additive Variance in multi-locus models

Recall that the variance of a sum is $Var(a+Xb, c+Wd) = Var(X)b^2 + Var(W)d^2 + bdCov(X,W)$. Thus, the variance of a sum can be (uniquely) decomposed into term-specific components if the terms are uncorrelated. Otherwise, co-variances between terms would have a non-null contribution to the variance of the sum. The contrasts for additive, dominance, and epistatic interactions are independent only under highly stylized conditions, including linkage equilibrium between the two loci. This means that in most real cases, there is no unique decomposition of the genetic variance into independent additive, dominance, and epistatic components.

Therefore, to decompose the total genetic variance into separate components, we need to use a sequential decomposition of variance. This requires defining what terms to enter into the model first. A traditional approach is to first introduce the additive components, then dominance, and finally epistatic interactions. The following example decomposes the total genetic variance into additive, dominance, and epistatic components using sequential ANOVA.

Example 6

```
fmA=lm(Y~XA+XB)
fmAD=lm(Y~XA+XB+HA+HB)
fmFull=lm(Y~XA+XB+HA+HB+XA*XB+HA*HB+XA*HB+HA*XB)

vY=var(Y)

# Sum of squares
RSSA=sum(residuals(fmA)^2)
RSSAD=sum(residuals(fmAD)^2)
RSSF=sum(residuals(fmFull)^2)
SSY=sum((Y-mean(Y))^2)

vGHat=SSY/(n-1)-RSSF/(n-9)

vAHat=SSY/(n-1)-RSSA/(n-3)
vDHat=RSSA/(n-3)-RSSAD/(n-5)
vEPisHat=RSSAD/(n-5)-RSSF/(n-9)

c('Additive'=vAHat,'Dominance'=vDHat,'Epistasis'=vEPisHat,'Total Genetic'=vGHat)
```

```
##      Additive     Dominance     Epistasis Total Genetic
##    0.57372011    0.05721918    0.01037561    0.64131490
```

```
c(vAHat,vAHat+vDHat)/vGHat
```

```
## [1] 0.8945997 0.9838213
```

Note that although we are simulating substantial dominance and epistatic interactions, an additive model captures ~90% of the genetic variance, and modeling additive plus dominance effects captures 99% of the total genetic variance.

2.7) Partition of the genetic variance into locus specific components when the two loci are in Linkage Equilibirum

Epistasis implies that the effects of alleles at one locus are modulated by alleles at another locus. Under these conditions we cannot decompose the total genetic variance into locus-specific components.

If we don't have epistasis and genotypes at different loci are in linkage equilibrium we can decompose the total genetic variance into locus-specific components, specifically, for a two locus model involving additive and dominance effects only we have

$$\begin{aligned} Var(G) &= Var(\mu + X_1a_1 + H_1d_1 + X_2a_2 + H_2d_2) \\ &= Var(X_1a_1 + H_1d_1) + Var(X_2a_2 + H_2d_2) \\ &= a_1^2Var(X_1) + d_1^2Var(H_1) + a_2^2Var(X_2) + d_2^2Var(H_2) \end{aligned}$$

Which provides a decomposition of the total variance into locus specific additive and dominance components. Note that above we are able to decompose the variance at each locus into additive and dominance terms because in the characterization we used $X \in \{-1, 0, 1\}$ and $H \in \{0, 1, 0\}$ are orthogonal.

2.8) Contribution of linkage disequilibirum (LD) in a 2-locus additive model

Let's consider now a situation where the loci are in LD, and, to keep things simple, lets focus on an additive model, that is

$$G = \mu + X_1a_1 + X_2a_2$$

The genetic variance in this model is

$$Var(G) = a_1^2Var(X_1) + a_2^2Var(X_2) + a_1a_2Cov(X_1, X_2)$$

The term $a_1a_2Cov(X_1, X_2)$ captures the contribution of LD to genetic (additive in the above model) variance. This term can be positive, null, or negative.

3) Extension to multiple loci

The extension of the two-locus model to multiple loci is conceptually straightforward. However, for 10 bi-allelic loci we have $3^{10} = 59,049$ possible means. The number of possible contrasts that can be included in the model grows exponentially and the math gets untraceable very quickly. Still, the additive and additive + dominance models are often easy to implement since the number of parameters involved is linear on the number of loci. For an additive model, for L loci we need L+1 parameters, and for an additive + dominance model we need 2L+1 parameters.

3.1) A matrix formulation of an Additive+Dominance multi-locus model

An additive+dominance effects model for an arbitrary number of loci can be written ans

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{H}\boldsymbol{d} + \boldsymbol{\varepsilon}$$

where:

- $\boldsymbol{y} = (y_l, y_2, ..., y_n)'$ is an n-dimensional vector with phenotypic measurements,
- $\mathbf{1} = (1, 1, ..., 1)'$ is a vector of 1's,
- μ is an intercept,
- $\boldsymbol{X} = \{X_{ij}\}$ is a matrix of additive contrasts (one column per loci),
- $\boldsymbol{H} = \{H_{ij}\}$ is a matrix of contrasts for dominance (one column per loci), and,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_1, ..., \varepsilon_n)'$ is a vector with environmental effects.

3.2) The contribution of linkage disequilibirum to additive variance

The additive variance in the above model is

$$Var(\boldsymbol{\alpha'}\boldsymbol{x}_i) = \boldsymbol{\alpha'}\boldsymbol{\Sigma_x}\boldsymbol{\alpha}$$

where, $\boldsymbol{\Sigma_x}$ is a (co)variance matrix containing in the diagonals the variances of each of the loci and in the off-diagonals co-variances between loci. We can write $\boldsymbol{\Sigma_x = D - (\Sigma_x - D)}$ where $\boldsymbol{D} = Diag\{Var(x_i j)\}$ is a diagonal matrix with the variances of genotypes at each of the loci and $(\Sigma_x - D)$ is a matrix with zeroes in the diagonals and co-variances between gentoypes at pairs of loci. Recall that under HWE, the diagonal elements are $Var(x_{ij} = 2p_j(1 - p_j)$ and under random mating $Cov(x_{ij}, x_{ij'}) = 2D_{jj'}$, where $D_{jj'}$ is the covariance of the refernce alleles at locus j and j' within gametes.

Using the above, we can decompose the additive variance into a component equal to the sum of the additive variances at each loci (the **genic variance** $\boldsymbol{\alpha'D\alpha} = \Sigma_j\{Var(x_{ij})\alpha_j^2\}$) and a component capturing the contribution of linkage disequilibrium to genetic variance ($\boldsymbol{\alpha'\Sigma_x\alpha - \alpha'D\alpha}$).

Acknowledgments: I Alexa Lupi (PhD) for reading and commenting an early version of this handout.

References

Falconer, D.S. and Mackay, T.F.C. (1996) Introduction to Quantitative Genetics. 4th Edition, Addison Wesley Longman, Harlow.