Compute phenotypic values and variances

Jinliang Yang

02-24-2022

Path Normalization

Select for shorter plants

- Tom collected a population of 20 maize landraces to select short plants.
- Then phenotypically characterized the 20 landraces and selected two of them as the founder lines to make an F2 population.
- After phenotyping and genotyping, Tom needs to determine which individual plants be more valuable to be selected and crossed.

```
geno <- read.table("data/geno.txt", header=FALSE)
dim(geno)
head(geno)
names(geno) <- c("chr", "pos", "ref", "alt", paste0("plant", 1:20))
```

Create an F2 population

- To make the F2 population, Tom chose `plant1` and `plant18` as the parents.
- `plant1` and `plant18` were selfed five generations to be pure inbreds.

```
### Just sample 100 markers from this Mt chr
set.seed(12579)
markers <- sample(1:nrow(geno), size=100)
f <- geno[sort(markers), c("chr", "pos", "ref", "alt", "plant1", "plant18")]
# select just one haplotype
f$plant1 <- gsub("/.*", "", f$plant1)
f$plant18 <- gsub("/.*", "", f$plant18)
# recoding to use -1,0,1
f[f==0] <- -1
# simulate the recombination rate
f$cM <- f$pos/5000
```

Create an F2 population

- To make the F2 population, Tom chose `plant1` and `plant18` as the parents.
- `plant1` and `plant18` were selfed five generations to be pure inbreds.

```
install.packages("devtools")
library(devtools)
install_github("lian0090/simuPoisson")
library(simuPoisson)
```

–

Then, simulate an F2 population with 200 individuals and 100 SNP markers.

```
pgeno <- t(f[, c("plant1", "plant18")])
pgeno <- apply(pgeno, 2, as.numeric)
f2 <- simuPoisson(pgeno, f$chr, f$cM, 200)

f2 <- as.data.frame(f2)
names(f2) <- paste0(f$chr, "_", f$pos)
write.table(f2, "data/f2_geno.csv", sep=",", quote=FALSE)
```

Compute genotype frq and allele frq

```
f2 <- read.csv("data/f2_geno.csv")
table(f2[,1])
```

Note that in this F2 population, the SNP coding is `-1, 0, 1` for A_1A1, A_1A_2, and A_2A_2.

```
table(f2[,6])
```

Frequencies

```
# For A1 allele
p <- (52*2+92)/((52+92+56)*2)
# For A2 allele
q <- (56*2+92)/((52+92+56)*2)
```

Observed allele frequency –

```
# For A1A1 genotype
A1A1 <- 52/(52+92+56)
# For A1A2 genotype
A1A2 <- 92/(52+92+56)
# For A2A2 genotype
A2A2 <- 56/(52+92+56)
```

Observed genotype frequency

Frequencies

###

Predicted genotype frequency

```
r
p^2
2pq
q^2
chisq.test(rbind(c(A1A1, A1A2, A2A2), c(p^2, 2pq, q^2)))
```

Phenotype

From this F2 population, Tom measured the plant height for each individual plant.

Phenotype in a population can be characterized in terms of its **mean** and **variance**.

Phenotype

```
pheno <- read.csv("data/f2_pheno.csv")
hist(pheno$height, main="Plant Height", xlab="Value (inch)", breaks=20)
```

Combine genotype and phenotype files

```
gp <- cbind(pheno, f2)
```

Genotypic value

$P = G + E$

$G = A + D$

Let's find out a and d at a specific Marker `Mt_24242`:

```
ggplot(gp, aes(x=as.factor(Mt_24242), y=height, color=as.factor(Mt_24242))) +
    geom_boxplot() +
    geom_jitter(color="black", size=1, alpha=0.9) +
    scale_color_manual(values=c("#E69F00", "#56B4E9", "#fe6f5e"))+
    labs(title="Mt_24242", y="Plant Height", x = "Genotype")+
    theme_classic() +
    guides(color=FALSE) +
    theme(plot.title = element_text(size=20, face = "bold"),
          axis.text=element_text(size=16, face="bold"),
          strip.text.y = element_text(size = 16, face = "bold"),
          axis.title=element_text(size=18, face="bold"),
          )
```

Genotypic value

Let's find out a and d at a specific Marker `Mt_24242`:

```
u <- mean(gp$height) # population mean
# A1A1
h1 <- mean(subset(gp, Mt_24242 == -1)$height)
# A1A2
h12 <- mean(subset(gp, Mt_24242 == 0)$height)
# A2A2
h2 <- mean(subset(gp, Mt_24242 == 1)$height)
```

```
a <- (h2 - h1)/2
midpoint <- h1+a
d <- h12 - midpoint
```

Allele Substitution Effect

The average effect of A1 and A2:

```
alpha <- a + d*(q - p)
alpha1 <- q*alpha
alpha2 <- -p*alpha
```

Breeding value

The **Breeding value** associated with Marker `Mt_24242` is defined as: the sum of α_i and α_j. - Breeding value is the value of an individual as a parent!

$$BV_{ij} = \mu + \alpha_i + \alpha_j$$

```
bv1 = u+alpha1 + alpha1
bv2 = u+alpha2 + alpha2
bv12 = u+alpha1 + alpha2
```

Genotypic value and breeding value

```
plot(c(0, 1, 2), c(h1, h12, h2), xlab="Genotype",ylab="", cex.lab=1.5, xaxt="n", pch=17, cex=2, col="red
axis(1, at=c(0, 1, 2), labels=c("A1A1", "A1A2", "A2A2"));
mtext("Breeding Value", side = 4, line = 1, cex=1.5, col="blue");
mtext("Genotypic Value", side = 2, line = 2, cex=1.5, col="red")
points(c(0, 1, 2), c(bv2, bv12, bv1), cex=2, col="blue")
lines(c(0, 1, 2), c(bv2, bv12, bv1), lwd=2, col="blue")
```

Additive and dominance variance

P = A + D + E

Phenotypic variance

```
Vp <- var(gp$height)
```

Additive genetic variance: V_A

These breeding values have a mean of zero, and their variance is the sum of the products of the genotype frequencies and the squared breeding values:

$$\begin{aligned} V_A &= p^2(2q\alpha)^2 + 2pq(q-p)^2\alpha^2 + q^2(-2p\alpha)^2 \\ &= 2pq\alpha^2(2pq + (q-p)^2 + 2pq) \\ &= 2pq\alpha^2(p+q)^2 \\ &= 2pq\alpha^2 \\ &= 2pq(a + d(q-p))^2 \end{aligned}$$

```
Va <- 2*p*q*(a + d*(q - p))^2
```

Dominance genetic variance: V_D

The variance due to dominance deviations is: the sum of the products of the genotype frequencies and the squared dominance deviation values.

$$\begin{aligned} V_D &= p^2(-2q^2d)^2 + 2pq(2pqd)^2 + q^2(-2p^2d)^2 \\ &= 4p^2q^2d^2(q^2 + 2pq + p^2) \\ &= 4p^2q^2d^2 \\ &= (2pqd)^2 \end{aligned}$$

```
Vd <- (2*p*q*d)^2
```

Additive and dominance variance

$G = A + D$

```
Vg <- Va + Vd
```

H^2 and h^2 due to this SNP marker

```
h2 <- Va/Vp
H2 <- Vg/Vp
```