# Chapter 8 <br> Outcrossing and Selection 

Plant data are often obtained from seeds collected from a set of maternal parent plants, leaving open question of paternal parentage. Attention is paid to the estimation of proportion of offspring produced by outcrossing as opposed to selfing.

## Estimation of Outcrossing

- Estimates of the proportion of offspring produced by outcrossing are based on the proportion of heterozygous offspring carrying one allele that could not be of maternal origin.
- Estimation from homozygous female parents
- Outcrossing rates can be estimated from progeny arrays derived from female homozygous parents
- Let $t$ be the probability that each offspring of any maternal plant $A u A u$ is an outcross, and $1 t$ be the probability that it is a self. The probability of heterozygous offspring is

$$
\sum_{v \neq u} t p_{v}=t\left(1-p_{u}\right)
$$

- The heterozygote numbers $\tilde{H} u$ in a sample of size $N u$ from $A u A u$ mother are binomially distributed

$$
\tilde{H}_{u} \sim B\left[N_{u}, \mathrm{t}\left(1-p_{u}\right)\right]
$$

- Combining over maternal genotypes gives the likelihood of $t$

$$
L(t) \propto \prod_{u}\left[t\left(1-p_{u}\right)\right]^{\tilde{H}_{u}}\left[1-t\left(1-p_{u}\right)\right]^{N_{u}-\tilde{H}_{u}}
$$

- MLE for $t(\hat{t})$ must satisfy

$$
\frac{1}{\hat{t}} \sum_{u} \tilde{H}_{u}=\sum_{u} \frac{\left(N_{u}-\tilde{H}_{u}\right)\left(1-p_{u}\right)}{1-\hat{t}\left(1-p_{u}\right)}
$$

- When only one maternal genotype $A A$ is used, the estimate can be found explicitly as

$$
\hat{t}=\frac{\tilde{H}_{A}}{N_{A}\left(1-p_{A}\right)}
$$

- Example of a barley population
- The variance of $\hat{t}$

$$
\operatorname{Var}(\hat{t})=\sum_{u}\left(\frac{\partial \hat{t}}{\partial \tilde{H}_{u}}\right)^{2} \operatorname{Var}\left(\tilde{H}_{u}\right)+\cdots
$$

- Estimation from equilibrium population
- A population with a constant proportion $t$ of outcrossing eventually reaches an equilibrium in which genotypic frequencies remain constant over time for neutral loci.
- For a locus with alleles $A u$ these frequencies are

$$
\begin{aligned}
& P_{u u}=p_{u}^{2}+p_{u}\left(1-p_{u}\right) \frac{1-t}{1+t}, \text { for } A_{u} A_{u} \\
& P_{u v}=2 p_{u} p_{v} \frac{2 t}{1+t}, \text { for } A_{u} A_{v}
\end{aligned}
$$

- Outcrossing and inbreeding coefficients are related by

$$
f=\frac{1-t}{1+t}
$$

- An estimate and its variance for $t$ from $\hat{f}$

$$
\begin{aligned}
& \hat{t}=\frac{1-\hat{f}}{1+\hat{f}} \\
& \operatorname{Var}(\hat{t})=\left(\frac{d \hat{t}}{d \hat{f}}\right) \operatorname{Var}(\hat{f})=\frac{4}{1+t} \operatorname{Var}(\hat{f})
\end{aligned}
$$

- Estimation from offspring of arbitrary female parents

| Maternal Genotype | Offspring Genotype | Observed <br> Count | Probability | Expected <br> Count |
| :---: | :---: | :---: | :---: | :---: |
| $A_{u}^{*} A_{u}^{*}$ | $A_{u}^{*} A_{u}^{*}$ | ${ }_{\text {un }} X_{u n}$ | $1-t$ | ${ }_{u u} S_{u u}$ |
|  |  |  | $t p_{u}$ | ${ }_{u ı} C_{u u}$ |
|  | $A_{u}^{*} A_{v}$ | ${ }_{u l} X_{u v}$ | $t p_{v}$ | ${ }_{u u} C_{u v}$ |
| $A_{u}^{*} A_{v}^{*}$ | $A_{u}^{*} A_{u}^{*}$ | ${ }_{u v} X_{u u}$ | $(1-t) / 4$ | ${ }_{u v} S_{u u}$ |
|  |  |  | $t p_{u} / 2$ | ${ }_{u v} C_{u u}$ |
|  | $A_{v}^{*} A_{v}^{*}$ | ${ }_{u v} X_{v v}$ | $(1-t) / 4$ | ${ }_{u v} S_{v v}$ |
|  | $A_{v}^{*} A_{v}$ |  | $t p_{v} / 2$ | ${ }_{u v} C_{v v}$ |
|  | $A_{u}^{*} A_{v}^{*}$ | ${ }_{u v} X_{u v}$ | $(1-t) / 2$ | ${ }_{u v} S_{u v}$ |
|  | $A_{u}^{*} A_{v}$ |  | $t p_{v} / 2$ | ${ }_{u v} C_{u v}$ |
|  | $A_{v}^{*} A_{u}$ |  | $t p_{u} / 2$ | ${ }_{u v} C_{u v}$ |
|  | $A_{u}^{*} A_{w}$ | ${ }_{u v} X_{u w}$ | $t p_{w} / 2$ | ${ }_{u v} C_{u w}$ |
|  | $A_{v}^{*} A_{w}$ | ${ }_{u v} X_{v w}$ | $t p_{w} / 2$ | ${ }_{v v} C_{v w}$ |

It may be that the actual numbers of offspring per parent are not large enough to infer the maternal genotype. It is then necessary to assign probabilities to each possible maternal genotype.

$$
\begin{aligned}
&{ }_{u u} s_{u u}=\left[\frac{1-t}{(1-t)+t p_{u}}\right]{ }_{u u} x_{u u} \\
&{ }_{u u} c_{u u}=\left[\frac{t p_{u}}{(1-t)+t p_{u}}\right]{ }_{u u} x_{u u} \\
&{ }_{u v} s_{u u}=\left[\frac{1-t}{(1-t)+2 t p_{u}}\right]{ }_{u v} x_{u u} \\
&{ }_{u v} c_{u u}=\left[\frac{2 t p_{u}}{(1-t)+2 t p_{u}}\right]{ }_{u v} x_{u u} \\
&{ }_{u v} s_{v v}=\left[\frac{1-t}{(1-t)+2 t p_{v}}\right]{ }_{u v} x_{v v} \\
& u_{v v} c_{v v}=\left[\frac{2 t p_{v}}{(1-t)+2 t p_{v}}\right]{ }_{u v} x_{v v} \\
&{ }_{u v} s_{u v}=\left[\frac{1-t}{(1-t)+t\left(p_{u}+p_{v}\right)}\right]{ }_{u v} x_{u v} \\
&{ }_{u v} c_{u v}=\left[\frac{t p_{v}}{(1-t)+t\left(p_{u}+p_{v}\right)}\right]{ }_{u v} x_{u v} \\
& \hat{p}_{u}=\frac{. . c_{u}}{. . c . .} \\
& \hat{t}= . . c . . \\
& . . \\
& t_{. .}+. . c . .
\end{aligned}
$$

## Example: pgm-1 in Lupinus albus data

Table 8.3 Notation for applying the EM algorithm to Pgm-1 in Lupinus albus dati of Green et al. (1980). Asterisks denote maternal alleles.

| Maternal Genotype | Offspring Genotype | Observed Count | Probability | Expected Count |
| :---: | :---: | :---: | :---: | :---: |
| $F^{*} F^{*}$ | $F^{*} F^{*}$ |  | $1-t$ | ${ }_{F F} S_{F F}$ |
|  | $F^{*} F$ |  | $t p_{F}$ | ${ }_{F F} C_{F F}$ |
|  | $F^{*} S$ | 5 | $t p_{S}$ | $\mathrm{FFF}_{F S}$ |
| $S^{*} S^{*}$ | $S^{*} S^{*}$ |  | $1-t$ | SS $s_{S S}$ |
|  | $S^{*} S$ |  | $t p_{S}$ | $S S S{ }_{S S}$ S |
|  | $S^{*} F$ | 9 | $t^{\prime}{ }_{F}$ | ${ }_{S S} c_{S F}$ |
| $S^{*} F^{*}$ | $S^{*} S^{*}$ |  | $(1-t) / 4$ | $S F F S S{ }^{\text {S }}$ |
|  | $S^{*} S$ |  | $t p_{S} / 2$ | SF ${ }^{\text {c }}$ SS |
|  | $F^{*} F^{*}$ |  | $(1-t) / 4$ | $S_{F} s_{F F}$ |
|  | $F^{*} F$ |  | $t p_{F} / 2$ | $S_{F} C_{F F}$ |
|  | $S^{*} F^{*}$ |  | $(1-t) / 2$ | $S_{F} S_{S F}$ |
|  | $S^{*} F$ | 28 | $t p_{F} / 2$ | $S_{F} c_{S F}$ |
|  | $F^{*} S$ |  | $t p_{S} / 2$ | $S F F F^{c_{F S}}$ |

Table 8.4 Parent-offspring genotype combinations for Lupinus albus data of Green et al. (1980).

|  |  | Locus |  |  |
| :---: | :---: | ---: | ---: | ---: |
| Maternal Genotype | Offspring Genotype | Pgm-1 | 6Pgd-2 | Aat-2 |
| $F F$ |  |  |  |  |
|  | $F F$ | 120 | 108 | 80 |
|  | FS | 5 | 4 | 7 |
| FS | FF | 11 | 17 | 14 |
|  | $F S$ | 28 | 35 | 30 |
|  | SS | 16 | 14 | 16 |
|  |  |  |  |  |
| SS | FS | 9 | 16 | 4 |
|  | SS | 104 | 263 | 283 |
|  |  |  |  |  |

Table 8.5 Outcrossing estimates for Lupinus albus data of Green et al. (1980). Initial iterates were 0.5 for $t$ and $p_{F}$.

|  | Pgm-1 |  |  | 6 Pgd-2 |  |  | Aat-2 |  |
| :---: | :---: | :---: | :--- | :--- | :--- | :--- | :--- | :--- |
| Iterate | $t$ | $p_{F}$ |  | $t$ | $p_{F}$ |  | $t$ | $p_{F}$ |
|  |  |  |  |  |  |  |  |  |
| 1 | 0.118 | 0.704 |  | 0.089 | 0.702 |  | 0.068 | 0.468 |
| 2 | 0.117 | 0.678 |  | 0.088 | 0.674 |  | 0.065 | 0.300 |
| 3 | 0.118 | 0.663 |  | 0.089 | 0.658 |  | 0.069 | 0.228 |
| 4 | 0.118 | 0.655 |  | 0.089 | 0.648 |  | 0.073 | 0.197 |
| 5 | 0.118 | 0.650 |  | 0.090 | 0.641 |  | 0.077 | 0.182 |
| 6 | 0.118 | 0.648 |  | 0.091 | 0.636 |  | 0.081 | 0.173 |
| 7 | 0.118 | 0.646 |  | 0.091 | 0.633 |  | 0.083 | 0.166 |
| 8 | 0.118 | 0.646 |  | 0.092 | 0.630 |  | 0.086 | 0.162 |
| 9 | 0.118 | 0.645 |  | 0.092 | 0.628 |  | 0.087 | 0.159 |
| 10 | 0.118 | 0.645 |  | 0.092 | 0.627 |  | 0.089 | 0.156 |
| 20 | 0.118 | 0.644 |  | 0.093 | 0.623 |  | 0.094 | 0.148 |
| 30 | 0.118 | 0.644 |  | 0.093 | 0.623 |  | 0.094 | 0.148 |
|  |  |  |  |  |  |  |  |  |

- Multilocus estimates
- The immediate benefit of multiple loci is that it is easier to identify outcross individuals.
- The number $n$ of discernible outcrosses in a sample of size $N$ is binomially distributed
$n \sim B[N, t(1-\alpha)]$,
where $\alpha$ is the probability that an outcross will not be discerned.
- For multiple loci, $\alpha=\prod_{1} \beta_{l}$ where $\beta_{l}$ is the probability that outcrosses can't be detected at lous $l$.
- Estimating number of paternal parents $P_{1}\left(n \mid p_{A}\right)$ : probability of mother-father pairs and offspring arrays when there is a single father for a family.
$P_{2}\left(n \mid p_{A}\right)$ : probability of mother-male trios and offspring arrays when two males inseminate the mother.


## - Most likely paternal plants

Table 8.10 One-locus, two-allele lod scores for possible paternal plants for one locus with two alleles.

| Mother | Offspring | Male |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $M$ | $C$ | $A$ | $\operatorname{Pr}(C \mid M)$ | $\operatorname{Pr}(C \mid M, A)$ | Ratio |
|  |  |  |  |  |  |
| $A a$ | $A A$ | $A A^{*}$ | $p_{A}$ | 1 | $1 / p_{A}$ |
|  |  | $A a$ | $p_{A}$ | $1 / 2$ | $1 /\left(2 p_{a}\right)$ |
|  |  | $a a$ | $p_{a}$ | 0 | 0 |

*Most likely father.
$R_{1}$ is the event that male $A$ is the father of offspring $C$ from female parent $M$, and $R_{2}$ is the event that male $A$ is simply a random male from the population not related to $C$, then the likelihoods are

$$
\begin{aligned}
& L\left(R_{1}\right)=\operatorname{Pr}(C \mid M, A) \operatorname{Pr}(M) \operatorname{Pr}(A) \\
& L\left(R_{2}\right)=\operatorname{Pr}(C \mid M) \operatorname{Pr}(M) \operatorname{Pr}(A) \\
& \operatorname{lod}\left(R_{1}, R_{2}\right)=\log \left[\frac{\operatorname{Pr}(C \mid M, A)}{\operatorname{Pr}(C \mid M)}\right]
\end{aligned}
$$

## Estimation of Selection

- The methods depend on estimating genotypic freguencies in successive generations, and comparing the results to those expected on the basis of no selection.
- Goodness-of-fit test for selection
- Before selection

| $A A$ | $A a$ | $a a$ |
| :---: | :---: | :---: |
| $P_{A}^{2}$ | $2 P_{A} P_{a}$ | $P_{a}^{2}$ |

- After selection

| $A A$ | $A a$ | $a a$ |
| :---: | :---: | :---: |
| $w_{A A} P_{A}^{2} / \bar{w}$ | $2 P_{A} P_{a} / \bar{w}$ | $w_{a a} P_{a}^{2} / \bar{w}$ |

- MLE for $P_{A}$ and $w_{A A}$

$$
\begin{aligned}
& \hat{P}_{A}=\frac{2 n_{A A}+n_{A a}}{2\left(n_{A A}+n_{A a}+n_{a a}\right)} \\
& \hat{w}_{A A}=\frac{2 n_{A A}\left(2 n_{a a}+n_{A a}\right)}{n_{A a}\left(2 n_{A A}+n_{A a}\right)}
\end{aligned}
$$

- Disequilibrium coefficient

$$
D_{A}=\frac{P_{A}^{2} P_{a}^{2}\left(w_{A A} w_{a a}-1\right)}{\bar{w}^{2}}
$$

- Estimation with one generation

$$
\begin{aligned}
& b \sim B\left(n, \frac{v}{v+1}\right) \\
& \hat{v}=\frac{b}{v+1-b} \\
& \operatorname{Var}(\hat{v}) \approx\left(\frac{\partial \hat{v}}{\partial b}\right)^{2} \operatorname{Var}(b)=\frac{v(v+1)^{2}}{n}
\end{aligned}
$$

- Components of selection
- The several types of selection that act at different stages of the life cyche

$$
\begin{aligned}
\text { Mating } & \longrightarrow \text { Zygote }_{t} \xrightarrow{\text { early }} \text { Adult }_{t} \xrightarrow{\text { late }} \text { Mating } \\
& \longrightarrow \text { Zygote }_{t+1} \xrightarrow{\text { early }} \text { Adult }_{t+1}
\end{aligned}
$$

- Genotypic frequencies under Prout's model

| $A A$ | $A a$ | $a a$ |
| ---: | ---: | ---: |

Generation $t$
Preselection $\quad P_{A}^{2} \quad 2 P_{A} P_{a} \quad P_{a}^{2}$
Early selection $\quad P_{A}^{2} E_{A A} \quad 2 P_{A} P_{a} \quad P_{a}^{2} E_{a a}$
Late selection $\quad P_{A}^{2} E_{A A} L_{A A} \quad 2 P_{A} P_{a} \quad P_{a}^{2} E_{a a} L_{a a}$
Generation $t+1$
Preselection $\quad X_{A}^{2} \quad 2 X_{A} X_{a} \quad X_{a}^{2}$
Early selection $\quad X_{A}^{2} E_{A A} \quad 2 X_{A} X_{a} \quad X_{a}^{2} E_{a a}$
$X_{A}=P_{A}^{2} E_{A A} L_{A A}+P_{A} P_{a}, X_{a}=P_{a}^{2} E_{a a} L_{a a}+P_{a} P_{A}$

- Maximum likelihood estimation of viability selection
- Selection for partical selfers
- Use of mother-offspring data


## Summary

Homework: Exercise 81. (pg 290)

