# Chapter 8 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 *AuAu* is an outcross, and 1 *t* be the probability that it is a self. The probability of heterozygous offspring is

$$\sum_{v \neq u} t p_v = t(1 - p_u)$$

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

$$\widetilde{H}_{u} \sim B[N_{u}, t(1-p_{u})]$$

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

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

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

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

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

$$\hat{t} = \frac{\tilde{H}_A}{N_A(1-p_A)}$$

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

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

- 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 Au these frequencies are

$$P_{uu} = p_u^2 + p_u (1 - p_u) \frac{1 - t}{1 + t}, \text{ for } A_u A_u$$
$$P_{uv} = 2p_u p_v \frac{2t}{1 + t}, \text{ for } A_u A_v$$

Outcrossing and inbreeding coefficients are related by

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

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

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

$$Var(\hat{t}) = \left(\frac{d\hat{t}}{d\hat{f}}\right)Var(\hat{f}) = \frac{4}{1 + t}Var(\hat{f})$$

• Estimation from offspring of arbitrary female parents

Maternal Genotype	Offspring Genotype	Observed Count	Probability	Expected Count
$A_u^*A_u^*$	$A_u^*A_u^*$	V	1 - t	$_{uu}S_{uu}$
	$A_u^*A_u$	uuAuu	$tp_u$	$_{uu}C_{uu}$
	$A_u^*A_v$	$_{uu}X_{uv}$	$t p_v$	$_{uu}C_{uv}$
$A_u^*A_v^*$	$A_u^*A_u^*$	V	(1-t)/4	$_{uv}S_{uu}$
	$A_u^*A_u$	$uvX_{uu}$	$tp_u/2$	$_{uv}C_{uu}$
	$A_v^*A_v^*$		(1-t)/4	$_{uv}S_{vv}$
	$A_v^*A_v$	$_{uv}X_{vv}$	$tp_v/2$	$_{uv}C_{vv}$
	$\begin{array}{c} A_u^*A_v^*\\ A_u^*A_v\\ A_v^*A_u \end{array}$	uvXuv	(1-t)/2 $t p_v/2$ $t p_u/2$	$uvS_{uv}$ $uvC_{uv}$ $uvC_{uv}$
	$A_u^*A_w$	uvXuw	$tp_w/2$	$_{uv}C_{uw}$
	$A_v^*A_w$	$_{uv}X_{vw}$	$tp_w/2$	$_{uv}C_{vw}$

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.

$$uu S_{uu} = \left[\frac{1-t}{(1-t)+tp_u}\right] uu X_{uu}$$

$$uu C_{uu} = \left[\frac{tp_u}{(1-t)+tp_u}\right] uu X_{uu}$$

$$uv S_{uu} = \left[\frac{1-t}{(1-t)+2tp_u}\right] uv X_{uu}$$

$$uv C_{uu} = \left[\frac{2tp_u}{(1-t)+2tp_u}\right] uv X_{uu}$$

$$uv S_{vv} = \left[\frac{1-t}{(1-t)+2tp_v}\right] uv X_{vv}$$

$$uv C_{vv} = \left[\frac{2tp_v}{(1-t)+2tp_v}\right] uv X_{vv}$$

$$uv S_{uv} = \left[\frac{1-t}{(1-t)+2tp_v}\right] uv X_{vv}$$

$$uv S_{uv} = \left[\frac{1-t}{(1-t)+t(p_u+p_v)}\right] uv X_{uv}$$

$$\hat{p}_u = \frac{..C_u}{..C_u}$$

$$\hat{t} = \frac{..C_u}{..S_u+..C_u}$$

### Example: pgm-1 in Lupinus albus data

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

Maternal	Offspring	Observed		Expected
Genotype	Genotype	Count	Probability	Count
$F^*F^*$	$F^*F^*$	120	1-t	$_{FF}s_{FF}$
	$F^*F$	120	$tp_F$	FFCFF
	$F^*S$	5	$tp_S$	FFCFS
$S^*S^*$	$S^*S^*$	104	1-t	SS <sup>8</sup> SS
	$S^*S$	104	$tp_S$	SSCSS
	$S^*F$	9	$tp_F$	SSCSF
$S^*F^*$	$S^*S^*$	16	(1-t)/4	SFSSS
	$S^*S$	10	$tp_S/2$	$SF^{C}SS$
	$F^*F^*$	11	(1-t)/4	$SF^{S}FF$
	$F^*F$	11	$tp_F/2$	$SF^{C}FF$
	$S^*F^*$		(1-t)/2	SFSSF
	$S^*F$	28	$tp_F/2$	$SF^{C}SF$
	$F^*S$		$tp_S/2$	$SF^{C}FS$

		-	Locus	
Maternal Genotype	Offspring Genotype	Pgm-1	6Pgd-2	Aat-2
FF	FF	120	108	80
	FS	5	4	7
FS	FF	11	17	14
	FS	28	35	30
	SS	16	14	16
SS	FS	9	16	4
	SS	104	263	283
and the second sec				

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

**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	6Pgd	-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,  $\alpha = \prod_{l} \beta_{l}$  where  $\beta_{l}$  is the probability that outcrosses can't be detected at lous l.
- Estimating number of paternal parents  $P_1(n \mid p_A)$ : probability of mother-father pairs and offspring arrays when there is a single father for a family.

 $P_2(n \mid p_A)$ : 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	Ċ	A	$\Pr(C M)$	$\Pr(C M,A)$	Ratio
Aa	AA	AA* Aa	$p_A$ $p_A$	$1 \\ 1/2$	$\frac{1/p_A}{1/(2p_a)}$
		aa	$p_a$	0	0
*Most li	kely 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

$$L(R_1) = \Pr(C \mid M, A) \Pr(M) \Pr(A)$$
  

$$L(R_2) = \Pr(C \mid M) \Pr(M) \Pr(A)$$
  

$$lod(R_1, R_2) = log \left[ \frac{\Pr(C \mid M, A)}{\Pr(C \mid M)} \right]$$

## **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

$$\begin{array}{c|ccc} AA & Aa & aa \\ \hline P_A^2 & 2P_A P_a & P_a^2 \end{array}$$

- After selection

$$\begin{array}{c|ccc} AA & Aa & aa \\ \hline w_{AA}P_A^2/\overline{w} & 2P_AP_a/\overline{w} & w_{aa}P_a^2/\overline{w} \end{array}$$

MLE for 
$$P_A$$
 and  $w_{AA}$   
 $\hat{P}_A = \frac{2n_{AA} + n_{Aa}}{2(n_{AA} + n_{Aa} + n_{aa})}$   
 $\hat{w}_{AA} = \frac{2n_{AA}(2n_{aa} + n_{Aa})}{n_{Aa}(2n_{AA} + n_{Aa})}$ 

- Disequilibrium coefficient  $D_A = \frac{P_A^2 P_a^2 (w_{AA} w_{aa} - 1)}{\overline{w}^2}$
- Estimation with one generation

$$b \sim B(n, \frac{v}{v+1})$$
$$\hat{v} = \frac{b}{v+1-b}$$
$$Var(\hat{v}) \approx (\frac{\partial \hat{v}}{\partial b})^2 Var(b) = \frac{v(v+1)^2}{n}$$

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

 $\begin{array}{l} \text{Mating} \longrightarrow \text{Zygote}_t \stackrel{\text{early}}{\longrightarrow} \text{Adult}_t \stackrel{\text{late}}{\longrightarrow} \text{Mating} \\ \longrightarrow \text{Zygote}_{t+1} \stackrel{\text{early}}{\longrightarrow} \text{Adult}_{t+1} \end{array}$ 

	AA	Aa	аа	
Generation t				
Preselection	$P_A^2$	$2P_A P_a$	$P_a^2$	
Early selection	$P_A^2 E_{AA}$	$2P_A P_a$	$P_a^2 E_{aa}$	
Late selection	$P_A^2 E_{AA} L_{AA}$	$2P_A P_a$	$P_a^2 E_{aa} L_{aa}$	
Generation $t + 1$				
Preselection	$X_A^2$	$2X_A X_a$	$X_a^2$	
Early selection	$X_A^2 E_{AA}$	$2X_A X_a$	$X_a^2 E_{aa}$	
$X_A = P_A^2 E_{AA} L_{AA} + P_A P_a, X_a = P_a^2 E_{aa} L_{aa} + P_a P_A$				

- Genotypic frequencies under Prout's model

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

## Summary

Homework: Exercise 81. (pg 290)