Appendix 1:

Using LINKNE to Calculate N_e from Linkage Disequilibrium

The program LINKNE extends the method of calculating contemporary effective population size (N_e) from linkage disequilibrium (LD) by using unphased genotype data (Waples and Do 2008) to incorporate the effects of linkage and (i) estimate N_e in past generations, and (ii) remove bias caused by violating the assumption that all pairs of loci in the dataset are unlinked.

Following Waples (2006), the observed LD, as measured by r^2, for a given pair of loci can be broken down into two components: one due to the effects of genetic drift in finite populations, and one due to the effects of sampling a limited number of individuals from the population. Thus:

$$E(\hat{r}^2_{\text{total}}) = E(\hat{r}^2_{\text{drift}}) + E(\hat{r}^2_{\text{sample}}) \quad (1)$$

Effective population size (N_e) can be estimated by considering the effects of the component of LD due to drift alone, which can be obtained by rearrangement:

$$E(\hat{r}^2_{\text{drift}}) = E(\hat{r}^2_{\text{total}}) - E(\hat{r}^2_{\text{sample}}) \quad (2)$$

Both values on the right side of the above equation can be estimated from observed genotype data.

Estimating LD Component Due to Sampling Variation

Weir and Hill (1980) showed that for a randomly mating population, the contribution to LD of sampling a finite number of individuals could be estimated as

$$E(\hat{r}^2_{\text{sample}}) = \frac{1}{S} \quad (3)$$

where S is the number of individuals sampled. However, England et al. (2006) found a large downward bias in estimates of effective size when using this equation, particularly if sample size was small relative to the true N_e. To address this bias, Waples (2006) suggested a bias correction, based on simulated data, which depended on the sample size: For S > 30,

$$E(\hat{r}^2_{\text{sample}}) = \frac{1}{S} + \frac{3.19}{S^2} \quad (4)$$

and for S < 30:

$$E(\hat{r}^2_{\text{sample}}) = 0.0018 + \frac{0.907}{S} + \frac{4.44}{S^2} \quad (5)$$

Following Waples and Do (2008), r^2_{sample} is averaged across pairs of alleles and loci. To account for effects of different sample sizes (due to missing data) and number of alleles, a weighting factor is applied to each locus pair. The weight of locus pair (*i*, *j*) is calculated as

$$w_{ij} = n_{ij} S_{ij}^2 \quad (6)$$

where…

$$n_{ij} = (n_i - 1)(n_j - 1)$$

and n_i and n_j are number of alleles at loci *i* and *j*, respectively, and S_{ij} is the sample size of individuals genotypes at each locus.

The weighted arithmetic mean across loci is then calculated as…

$$\hat{r}^2_{\text{sample}} = \frac{1}{W} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} r^2_{ij,sample} \quad (7)$$

where…

$$W = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij}$$

Estimating Total LD

For each pair of alleles in pairwise comparisons of loci in the dataset, $\hat{r}^2_{\text{total}}$ is estimated using Burrow's composite linkage disequilibrium measure (*D*), as in Weir (1996), with the formula:

$$D = \hat{\Delta}_{AB} = \frac{n_{AB}}{n} - 2\hat{p}_A\hat{q}_B \quad (8)$$

where…

$n =$ number of individuals genotyped at both loci,

$n_{AB} = 2n_1 + n_2 + n_3 + \frac{n_4}{2}$,

$n_1 =$ count of double homozygous individuals,

n_2 and $n_3 =$ counts of individuals heterozygous for one or the other allele,

$n_4 =$ count of double heterozygous individuals, and

$\hat{p}_A$ and $\hat{q}_B =$ allele frequencies of each allele among individuals genotyped at both loci.

Burrow's *D* can be standardized by allele frequency to produce the correlation coefficient $\hat{r}_{AB}$…

$$\hat{r}_{AB} = \frac{\hat{\Delta}_{AB}}{\sqrt{(\hat{p}_A(1-\hat{p}_A) + (\hat{h}_{AA} - \hat{p}_A^2))(\hat{q}_B(1-\hat{q}_B) + (\hat{h}_{BB} - \hat{q}_B^2))}} \quad (9)$$

where…

$\hat{h}_{AA}$ and $\hat{h}_{BB}$ = the observed proportions of homozygous genotypes of each allele in the individuals genotyped at both loci, and

$\hat{p}_A$ and $\hat{q}_B =$ allele frequencies of each allele among individuals genotyped at both loci.

The square of this value ($\hat{r}_{AB}^2$) is calculated for each locus pair (averaging across all allelic combinations), weighted as in the previous section, and averaged across all pairs of loci as…

$$\hat{r}^2_{\text{total}} = \frac{1}{W} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} w_{ij} r^2_{ij,total} \quad (10)$$

Estimating N_e with LD from Markers with Known Linkage Relationships

Binning Estimates:

To obtain precise estimates of N_e at multiple points in time, estimates from multiple locus pairs must be binned together across a range of recombination rates. Hayes et al. (2003) derived an approximate relationship between recombination rate and time as…

$$t = \frac{1}{2c} \tag{11}$$

where…

t = the number of generations in the past to which the estimate applies, and

c = the recombination rate between a pair of loci (in Morgans)

Assuming this relationship holds for all possible recombination rates, this means that the majority of the spectrum of recombination rates reflects N_e at very recent generations. For example, while $t = 1$ when loci are completely unlinked ($c = 0.5$), t does not reach two generations in the past until $c = 0.25$ and does not reach ten generations in the past until $c = 0.1$. Because of this, binning locus pairs into equally sized windows, based on recombination rates, produces trend lines that are not very informative as they mostly reflect recent N_e and tend to collapse past estimates into a small region at the end of the trend line. A more informative way of binning locus pairs is by generation, which allows finer scale changes to be revealed but requires that bins are larger at larger recombination rates. LINKNE allows users to choose whether bins should be defined by equally sized recombination rate windows or by generations. If generation-based bins are specified, the program first produces recombination rate-based bins (based on a size specified by the user) and then iteratively merges bins when the bin midpoints refer to time points within two generations of each other (based on the above equation).

Calculating N_e for each bin

As stated above, given $\hat{r}^2_{\text{total}}$ and $\hat{r}^2_{\text{sample}}$, the component of LD due to drift can be calculated by Equation 2. Weir and Hill (1980) showed that under the assumption of random mating and if N_e and S are relatively large, r^2-drift can be written as a function of N_e and the recombination rate (c, in Morgans) between pairs of loci…

$$\hat{r}^2_{\text{drift}} = \frac{(1-c)^2 + c^2}{2N_e c(2-c))} \tag{12}$$

Hill (1981) simplified this equation by separating the term relating to recombination rate from the term relating to N_e…

$$\hat{r}^2_{\text{drift}} = \frac{\gamma}{N_e} \tag{13}$$

where…

$$\gamma = \frac{(1-c)^2+c^2}{2c(2-c)}, \text{ and}$$

c = the mean recombination rate of a bin.

This is rearranged to calculate N_e as…

$$N_e = \frac{\gamma}{\hat{r}^2_{\text{drift}}} \tag{14}$$

References

England PR, Cornuet J-M, Berthier P, Tallmon DA, Luikart G (2006). Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conserv Genet* **7**: 303–308.

Hayes BJ, Visscher PM, Mcpartlan HC, Goddard ME (2003). Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. *Genome Res* **13**: 635–643.

Hill WG (1981). Estimation of effective population size from data on linkage disequilibrium. *Genet Res* **38**: 209–216.

Waples RS (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conserv Genet* **7**: 167–184.

Waples RS, Do C (2008). Ldne: a program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* **8**: 753–6.

Weir BS, Hill WG (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**: 477–488.