

Week 2 –Population genetic structure & size

Tuesday

Consequences of creating sub-populations

Quantifying genetic differences among sub-populations

Genotype frequencies in sub-populations

Effective genetic size of sub-populations

Estimating N_e from molecular data

Estimating individual membership to populations with molecular data

Thursday

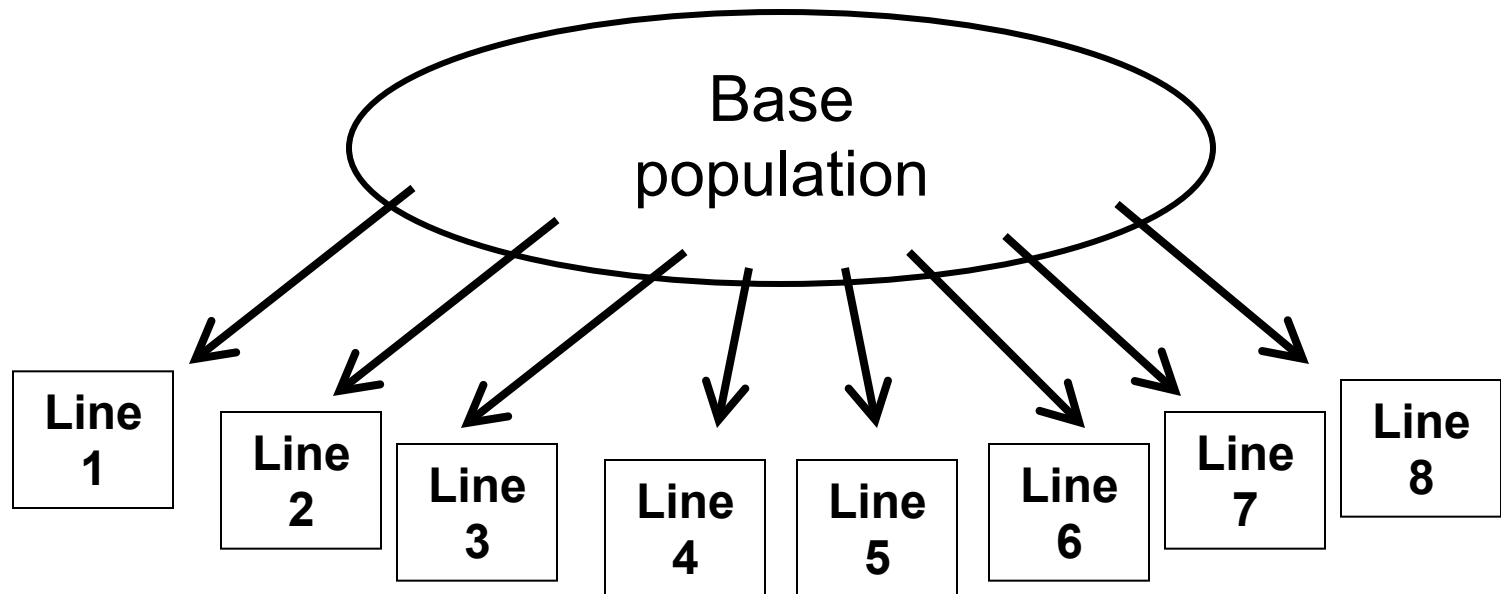
Review of assignments

Quiz



Consequences of creating sub-populations

- Remember that random genetic drift arises from the effect of sampling
- The smaller the sub-population, the larger the effect of random genetic drift
- Obviously, the smaller the sub-population, the higher the likelihood of two alleles at an individual being IBD



Let's consider that these sub-population had the same allele frequency as the base population for and allele A1 at a frequency $p = 0.5$

Generation	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8
0	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

We can calculate the mean frequency of allele A1 at the first generation $p(A1_0)$:

$$\begin{aligned} \text{Mean } p_0 &= \sum p_{i_0} / n \\ &= 4 / 8 = 0.5 \end{aligned}$$

And the variance of the frequency of A1:


$$\begin{aligned} \text{Var } p_0 &= \sum (p_{i_0} - p_{\text{founder}})^2 / (n-1) \\ &= (0.5-0.5)^2 + \dots / (8-1) = 0 \end{aligned}$$

Generation	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8
0	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
1	0.60	0.40	0.60	0.40	0.60	0.40	0.60	0.40

Mean p_1

$= \sum p_{i_1} / n$

$= 4 / 8 = 0.5$



Var p_1

$= \sum (p_{i_1} - p_{\text{founder}})^2 / (n-1)$

$= (0.6-0.5)^2 + / (8-1) = 0.01$

Generation	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8
0	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
1	0.60	0.40	0.60	0.40	0.60	0.40	0.60	0.40
2	0.80	0.20	0.80	0.20	0.80	0.20	0.80	0.20

Mean p_2

$= \Sigma p_{i_2} / n$

$= 4 / 8 = 0.5$

Var p_2

$= \Sigma (p_{i_2} - p_{\text{founder}})^2 / (n-1)$

$= (0.8-0.5)^2 + / (8-1) = 0.10$

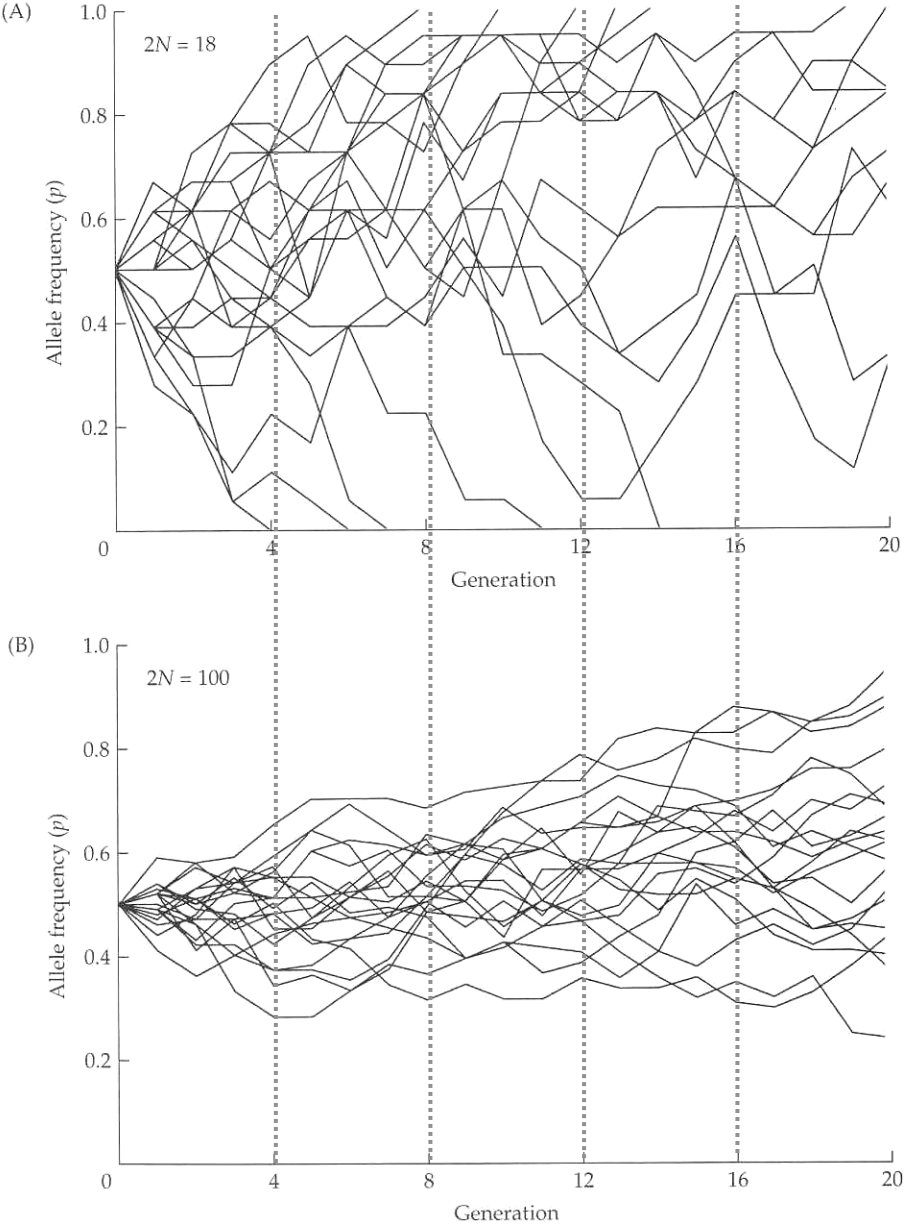
Generation	Line 1	Line 2	Line 3	Line 4	Line 5	Line 6	Line 7	Line 8
0	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
1	0.60	0.40	0.60	0.40	0.60	0.40	0.60	0.40
2	0.80	0.20	0.80	0.20	0.80	0.20	0.80	0.20
3	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

Mean $p_0 = \sum p_{i_3} / n$
 $= 4 / 8 = 0.5$

Var $p_3 = \sum (p_{i_3} - p_{\text{founder}})^2 / (n-1)$
 $= (1.0-0.5)^2 + / (8-1) = 0.28$

- In summary, the expected **mean allele frequency across all sub-populations or lines does not change**.
- However, **there is an increase in allele frequency variance** that increases as the number of generations increase in a sub-population.
- The **allele frequency variance also increases with smaller subpopulation sizes**. Consider an extreme example where each subpopulations are created from different single individual sampled from the founder population. The consequence is that maximum variance is achieved in the first generation.
- Consider the figures in the next slide to illustrate the effect of sub-population size and number of generations on the allele frequency variance.

Week 2 – Consequences of creating sub-populations



In summary:

- The final consequence of random genetic drift is fixation or loss of alleles from the founder population, in the subpopulations.
- Fixation or loss occur at all loci.
- The proportion of lines that will have one or the other allele fixed is equal to the frequency of alleles in the founder population.
- Random genetic drift leads to loss of variation within sublines and higher variation among sublines.
- Increase of homozygotes in the entire population.

Quantifying genetic differences among sub-populations

Random genetic drift, mutation, migration and selection, and other forces that change gene frequencies in populations lead to genetic structure – i.e. unequal gene and genotype frequencies among populations, with variable levels of similarity among them.

To know if there is population structure, and what level of similarity exists among them, is highly relevant in breeding. Common questions that need to be addressed include:

Quantifying genetic differences among sub-populations

Random genetic drift, mutation, migration and selection, and other forces that change gene frequencies in populations lead to genetic structure – i.e. unequal gene and genotype frequencies among populations, with variable levels of similarity among them.

To know if there is population structure, and what level of similarity exists among them, is highly relevant in breeding. Common questions that need to be addressed include:

- Are all breeding lines genetically similar?
- If lines are not similar, should they be combined or introgressed?
- Does genetic differentiation reflect the history of the breeding program?
- And if that is the case, are we missing critical sources of genetic variation?

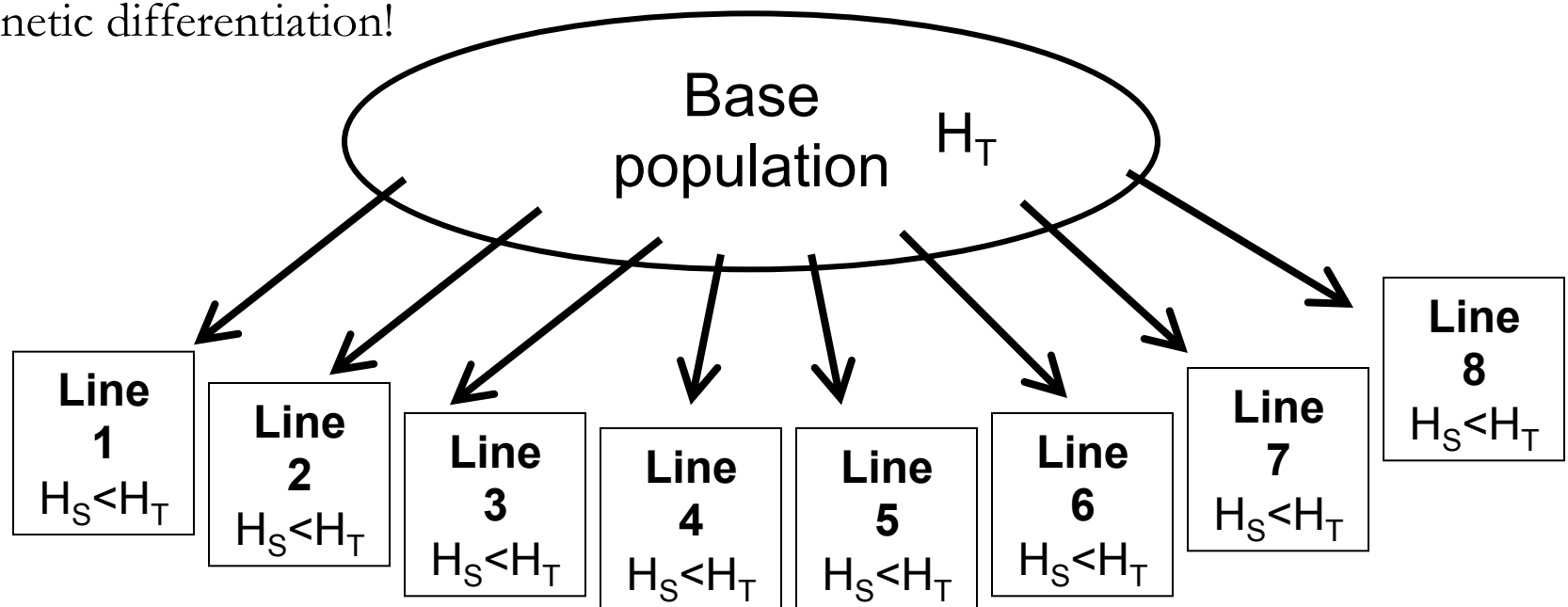
Let's think about a population in Hardy-Weinberg Equilibrium:

$$f(A_1A_1) = p^2$$

$$f(A_1A_2) = H = 2pq$$

$$f(A_2A_2) = q^2$$

When we reviewed Random Genetic Drift, we saw that population subdivision causes a reduction in heterozygosity. When measuring the population structure we are measuring nothing more than differences between subpopulations that originated from an ancestral population that gave rise to all individuals of one species. So the reduction in heterozygosity can be our measure of the population genetic differentiation!



The relative proportion of expected number of heterozygotes (H_S) in the lines relative to the expected number of heterozygotes in the total population (H_T) is quantified by F_{ST} , the fixation index. The fixation index provides a measure of the proportion of heterozygotes in the **S**ubpopulation relative to the **T**otal population, and is estimated by:

$$F_{ST} = (H_T - H_S) / H_T$$

$$H_S = H_T - F_{ST} * H_T$$

In summary, the heterozygosity of lines is smaller than if they were combined into a single, large, randomly mating base population. F_{ST} measures the extent of the reduction in heterozygosity.

Another useful interpretation is that F_{ST} measures the amount of variation in the whole population that is due to genetic differentiation among subpopulations.

Let’s look at some hypothetical and extreme data:

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Mean		
q	0.50	0.50	0.50	0.50	0.50	=	q
2q(1-q)	0.50	0.50	0.50	0.50	0.50	=	H_s


$H_T =$ 

$F_{ST} =$

The entire population (Pop. 1, 2, 3 and 4 combined) is genetically variable for the locus above. However there is no genetic variation among populations.

Let’s look at some hypothetical and extreme data:

	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Mean		
q	1.00	1.00	0.00	0.00	0.50	=	q
2q(1-q)	0.00	0.00	0.00	0.00	0.00	=	H_s

H_T = 

F_{ST} =

Populations (Pop. 1, 2, 3 and 4 combined) are genetically variable for the locus above. At least one population differs completely from the others in terms of gene frequency (**F_{ST}** = 1).

Example:

Eucalyptus 9 SSR 5 sub-populations

The value estimated for F_{ST} (0.32) was interpreted by Sewall Wright the following way:

< 0.05	Little genetic differentiation
$0.05 - 0.15$	Moderate genetic differentiation
$0.15 - 0.25$	High genetic differentiation
> 0.25	Very high genetic differentiation

In reality, genetic structure studies generally sample many loci to obtain an averaged estimate. Single loci may be influenced by the effect of sampling, selection when linked to genes associated with fitness, and other factors. Also, one has to avoid loci that may be under any type of selection.

Effective genetic size of sub-populations

So, you discovered that the sub-populations in your breeding program are genetically distinct from each other.

Now what?

Probably one of the next things you would like to do is know how different the individuals are, within each sub-population. If they are all highly similar genetically you may want to keep only part of them. On the other hand, if they are highly diverse, then you may consider keeping the population intact to preserve its genetic variation.

A census size of each sub-population may provide an idea of the extent of the genetic diversity that is present there. But what if the individuals in the sub-population are highly related to each other? In that scenario, the “true” size from a genetics standpoint, would likely be less than the census size.

We call this the effective population size (N_e).

Effective genetic size of sub-populations

In population genetics it is often assumed that populations are in HWE – that is, they have an equal number of individuals at each generation, among other requirements. In reality, breeding populations rarely (if ever) have properties of a population in HWE.

As a result the census number of a population (its actual population size N) does not define it well from a genetic perspective. Therefore, another parameter is used, the effective population size (N_e). N_e is essentially a weighted average number of individuals that takes into consideration individual relatedness.

Not surprisingly, N_e is directly related to the coefficient of inbreeding F :



$$N_e = N / (1+F)$$

Unequal number of individuals in successive generations

Several factors will impact the effective population size. One of most significant is when changes in population size occur.

In natural and breeding populations, the population size varies throughout successive generations. The effective population size can be estimated by:

$$1/N_e = (1/t) (1/N_1 + \dots + 1/N_t)$$

Where N represents the number of individuals at each successive generations (t).

In the example on the right:

$$\begin{aligned} 1/N_e &= (1/3) (1/10 + 1/3 + 1/30) \\ &= (1/3) (14/30) \\ &= 14/90 \end{aligned}$$

$$N_e = 6.4$$

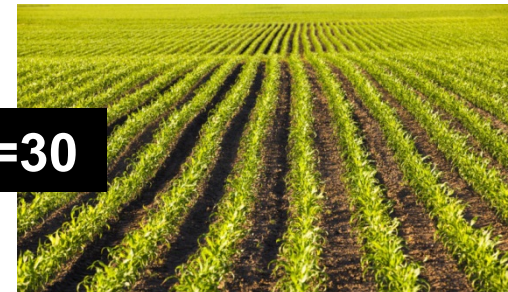
$N_1=10$



$N_2=3$



$N_3=30$



Unequal number of individuals in successive generations

Now let's consider a more extreme example:

$$1/N_e = (1/t) (1/N_1 + \dots + 1/N_t)$$

$N_1=10$



$N_2=100$



$N_3=1000$



This examples shows well the effect of having one or few generations with few individuals. Such events are called **bottlenecks**, and the drift in gene frequency that occurs in new populations that are formed after such events are called **founder effects**.

Unequal number of individuals in successive generations

Bottleneck events are common in the evolutionary history of plants, as well as in breeding programs.

Therefore, knowledge about the date (how many generations ago did it happen?) and the magnitude (what fraction of the total population was selected?) of the reduction in genetic diversity can be very useful in predicting how much genetic variation there truly is in the population.

Note that other factors, particularly life-history traits (e.g. mating system), can greatly impact how the effective population size will change over time.

Unequal distribution of family size

In an ideal population, the number of progeny that survive to become breeding individuals in the next generation is equal to the number parents (2).

Real populations that have a stable population size are likely to have an average of 2 offspring per cross. But the variance in family size tends to increase because individuals have different fertility and progeny viability differs from family to family. As a result, a **larger** number of individuals that become breeders in the next generation will be originated from a **smaller** number of parents, and N_e is smaller than the actual population size N .

The effect of family size on the effective population size N_e can be estimated by:

$$N_e = 4N / (2 + \sigma_k^2)$$

Unequal distribution of family size

$$N_e = 4N / (2 + \sigma_k^2)$$

Where σ_k^2 is the variance of family size. To estimate this variance one needs to know the distribution of the number of offspring that will survive from a cross and make it to the next generation as breeders. Considering that most parental crosses produce large numbers of zygotes but only few survive, this is considered to follow a Poisson distribution, or a distribution of rate events. For the Poisson distribution, the mean is equal to its variance (mean = $\sigma_k^2 = 2$).

Unequal distribution of family size

$$N_e = 4N / (2 + \sigma_k^2)$$

Therefore, in an ideal population where the variance of the family size is 2, the effective population size N_e will be the same as the actual population size.

For most populations, however, $\sigma_k^2 > 2$, and therefore the effective population size $N_e < N$.

There are a significant number of situations in breeding populations where knowing the effective population size can be relevant for conservation efforts and progressing breeding programs.

Estimating N_e from molecular data

Very often it is difficult or impossible to estimate the exact effective size of a population.

Most methods typically required data from multiple generations, in order to estimate N_e .

New methods that use molecular data now can be applied to obtain estimates of the real effective population size.

For example – the CCLONES breeding population has ~ 900 individuals.


Assume you became the director of this breeding program, and there was no information about the pedigree that led to this population.

However, you know that ~ 4800 SNP markers have been genotyped in this population.

NEESTIMATOR v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data

C. DO,* R. S. WAPLES,† D. PEEL,‡ G. M. MACBETH,§ B. J. TILLET¶ and J. R. OVENDEN**

Implements multiple methods to estimate N_e from molecular data.

-  linkage disequilibrium method (Waples & Do 2008)
- heterozygote-excess method (Zhdanova & Pudovkin 2008)
- molecular coancestry method (Nomura 2008)

NEESTIMATOR v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data

C. DO,* R. S. WAPLES,† D. PEEL,‡ G. M. MACBETH,§ B. J. TILLET¶ and J. R. OVENDEN**

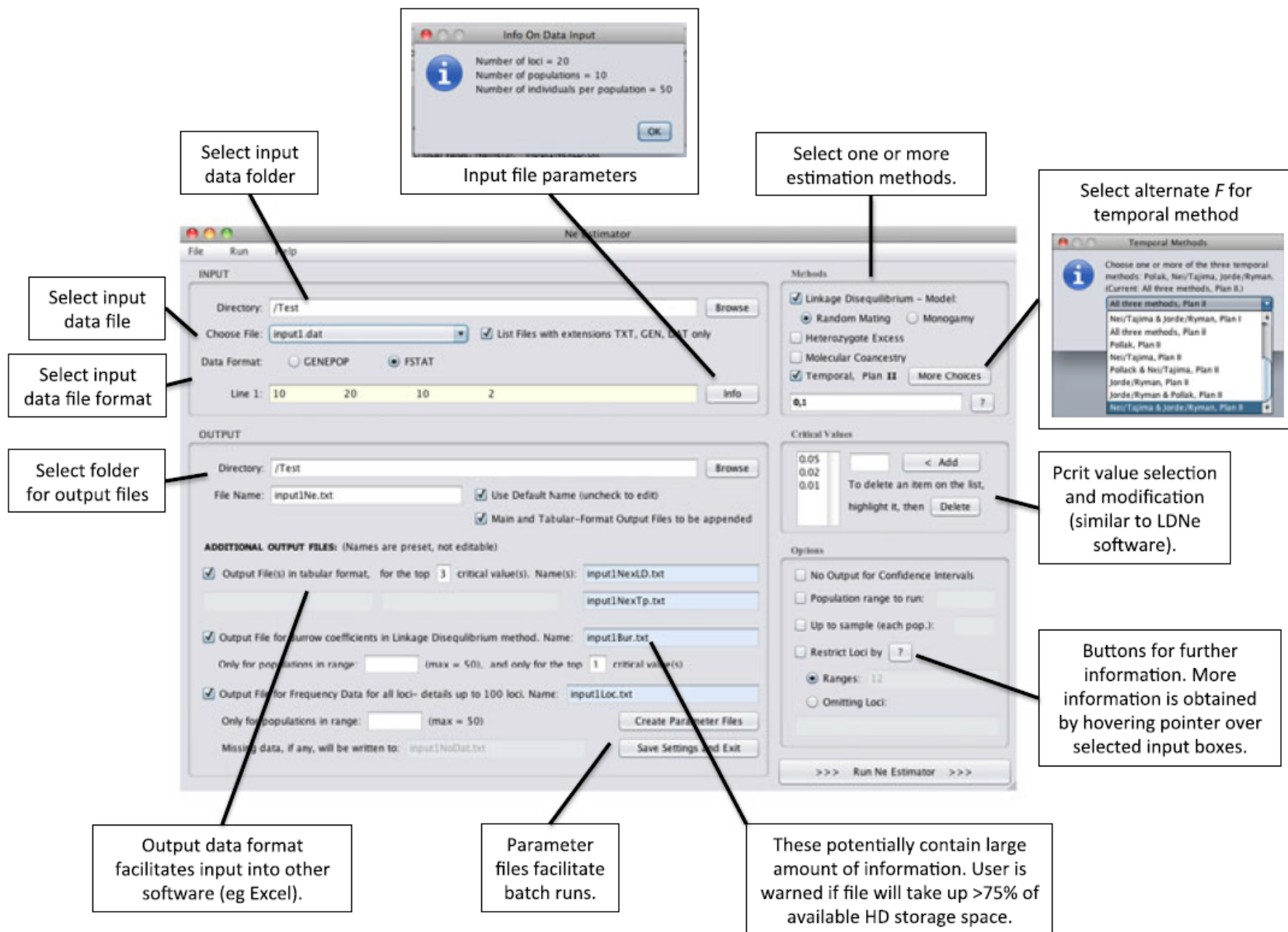
Download it from <http://molecularfisherieslaboratory.com.au/neestimator-software/>

Takes datasets in STAT and GENEPOP format

Sample (reduced and full) datasets available for CCLONES

923 individuals

4854 SNP loci



Output from NeEstimator v.2
Starting time: Fri Aug 22 19:25:39 2014
Input File: "cclonesNe.gen"

Number of Loci = 4854

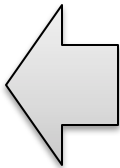
LD mating model: Random

Population 1 [140099] (Number of Individuals = 923)

Lowest Allele Frequency Used 0.050 0.020 0.010

LINKAGE DISEQUILIBRIUM METHOD

Harmonic Mean Sample Size =	641.8	676.5	692.9
Independent Comparisons =	5112973	7483872	8826262
OverAll r^2 =	0.01091	0.01071	0.01038
Expected r^2 Sample =	0.00120	0.00119	0.00119
Estimated Ne^ =	32.1	32.8	34.1
95% CIs for Ne^			
* Parametric	32.1	32.8	34.0
	32.2	32.9	34.1



* CIs by Jackknife are skipped when number of polymorphic loci > 100

Ending time: Fri Aug 22 21:03:56 2014

Estimating individual membership to populations with molecular data

Very often, breeding programs have a long history, that includes a founding population and introgressions along breeding cycles.

Breeders often don't have documentation that describes the origin of each material.

Furthermore, the genetic relatedness among materials introgressed into breeding programs is frequently unknown.

Estimating individual membership to populations with molecular data

It can be very valuable to recognize if there are groups of individuals that share recent ancestry – that is, that are more genetically similar or come from the same population.

It is also valuable to, for any new material introduced in the breeding program, to recognize if that material is more or less genetically similar to other germplasm in the breeding program.

Until recently, most methods were based on the genetic distance between individuals and populations. These methods did not provide proper confidence intervals, and did not permit the inclusion of prior information.

More recently, new methods that are based on modeling the data to certain population profiles, and then comparing the likelihood of each profile, have been developed and implemented.

Estimating individual membership to populations with molecular data

- Model-based clustering methods
- Assumes that the data represents a random sample drawn from a parametric model – inferences for the parameters are made by trying to model the ideal clustering membership of each sample
- In this generation, each allele copy originated in one of the founding populations
- Want to figure out the probability alleles in each individual originated in population k : \mathbf{Q} vector
- Each clustering is defined by sets of individuals that minimize the deviation from HWE
- Allows for inclusion of prior information

By far the most popular approach has been developed by Pritchard:

<http://pritch.bsd.uchicago.edu/structure.html>

Population 1



$$f(A_1) = p_1$$

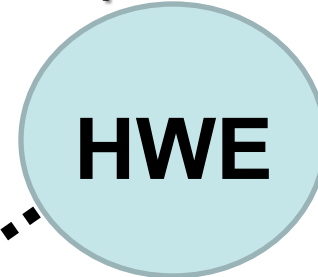
$$f(a_1) = q_1$$

$$f(A_1A_1) = p_1^2$$

$$f(Aa_1) = 2p_1q_1$$

$$f(a_1a_1) = q_1^2$$

Population 2



Population 5



$$f(A_5) = p_5$$

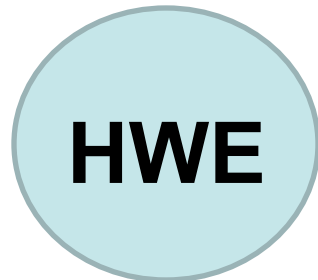
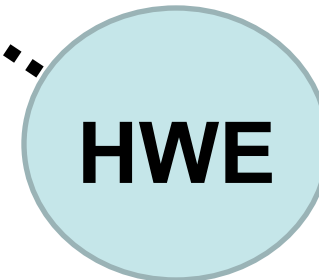
$$f(a_5) = q_5$$

$$f(A_5A_5) = p_5^2$$

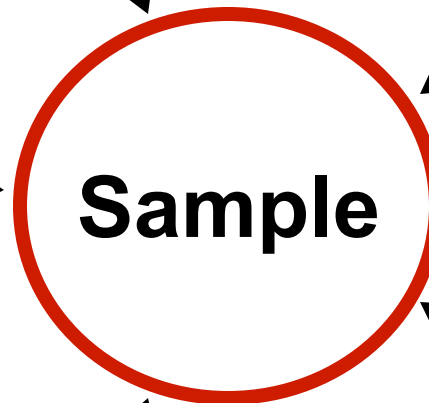
$$f(A_5a_5) = 2p_5q_5$$

$$f(a_5a_5) = q_5^2$$

Population 3



Population 4



Estimating K

Problem: Difficult to estimate allele frequencies, admixture proportions and number of groups simultaneously

Suggested by Pritchard:

Phase 1: Define number of k populations to test.

Phase 2: Examine clustering of individuals to evaluate appropriateness of the number of selected populations.

Input data file:

- Open files *Euccalyptus.txt*;

- Component of the file:

Row: marker name (optional)
 inter-marker distance (optional)
 individual data (required)
 phase information (optional)

Column: **label (individual information - optional)**
 PopData (prior population information - optional)
 PopFlag (use/not use pop. data-optional), Phenotype
 (optional), extra columns (optional)
 Genotype data (individual genotypes - required)

101	12476	144113	18156	24123	27120
101	124104	144121	18166	24131	27120
102	124102	144134	18150	-9	27120
102	124106	144134	18150	24131	27138

Modeling decisions:

- How long to run the program
 - burnin length (how long to run before start collecting data): 10,000-100,000 adequate, but verify summary statistics
 - run length: several runs at each K , at different lengths, and verify if answers are consistent through runs – 10,000-100,000 adequate for estimating number of sub-populations – 1,000,000 or more for good estimates of $\Pr(X | K)$

Modeling decisions:

- Ancestry model
 - No admixture – i.e. each individual comes from one discrete population – output is the posterior probability of one individual i belonging to population K .
 - Admixture – i.e. ancestry may be present – output is the posterior mean estimate of proportion of the individuals genome that originated from population K – good starting point.
 - Using prior information – i.e. information from geographic sampling is available, or may want to use samples of known sub-population origin to define where others came from

Running STRUCTURE:

- Data input:
 - File → New Project: enter name of project, directory where the output should be saved, and data file → Next
 - Enter number of individuals (234), number of loci (9) and value of missing data (-9) → Next
 - Check box if data contains following rows (leave all blank) → Next
 - Check box if data contains following columns (check: individual id for each individual) → Finish → Proceed

Running STRUCTURE:

- Parameter set:
 - Parameter set → New
 - Run length: burnin (20,000) and number of MCMC steps after burnin (20,000-100,000)
 - Ancestry model: Admixture with population data
 - Allele frequency model: Allele frequency independent
 - Please name parameter set: Enter name for set of parameters
 - Parameter set → Run
 - Set number of assumed populations

Running STRUCTURE:

- Two groups run analysis independently:
 - Run length: burnin (10,000) and number of MCMC steps after burnin (20,000 and 100,000)
 - Ancestry model: Admixture with population data
 - Allele frequency model: Allele frequency independent
 - Run model for $K = 1, 2, 3, 4$ and 5
 - After complete let's compare results