Introduction
Consider two parental populations: $P_1$ and $P_2$. Within each there are population allele frequencies given by:

\[ P_{1i} = (f_{11}, f_{12}, \ldots, f_{1n}) \]  

and

\[ P_{2i} = (f_{21}, f_{22}, \ldots, f_{2n}) \]

Where $i$ denotes the population. From each population leave that population to form a third population $P_h$.

\[ P_h = (f_{h1}, f_{h2}, \ldots, f_{hn}) \]

Each allele in $P_h$ came either from $P_1$ with probability $p_1$ or from $P_2$ with probability $p_2$ (i.e. $f_{hi} = p_{1} f_{1i} + p_{2} f_{2i}$).

From samples of $N_h$ alleles in population $P_h$ ($k = 1, 2, H$), we assume that we can observe allele frequencies of $n_1 = (n_{11}, n_{12}, \ldots, n_{1k})$ and $n_2 = (n_{21}, n_{22}, \ldots, n_{2k})$.

Analytical Method
The probability of coming to the observed configuration can be evaluated analytically by considering all possible ways that alleles in $P_h$ could have come from $P_1$ and $P_2$ and all possible parental configurations under the hybridisation occurred:

\[ P\left(n_{1i} + 1 \right) \sum_{k=1}^{N_h} n_{1i} + 1 \]

provided there is an allele of type $i$ in the Hybrid population. This method effectively picks both the population and the allele to be assigned both at the same time. This process continues (with new probabilities calculated each time) until there are no more alleles remaining in $n_h$.

Method 1
Send any allele $i$ to population $j$ with probability proportional to

\[ p_j (n_{ji} + 1) \sum_{k=1}^{N_h} n_{ji} + 1 \]

The probability that the parental configuration $n_{1i}$, $n_{2i}$, $n_{1j}$, $n_{2j}$ leads to the observed configuration $n_{1i}$, $n_{2i}$, $n_{1j}$, $n_{2j}$ is $\alpha_i \cdot \frac{\binom{n_{1i}}{n_{1j}} \binom{n_{2i}}{n_{2j}}}{\binom{N_h}{k}}$.

Computational Methods
The problem with the analytical method is that the sum has many $(\sum_{k=1}^{N_h} n_{1i} + 1)$ terms. This means that the algorithm for the analytical method is order $(N_h)^3$. It is preferable to generate a few parental configurations at random and average over the probabilities determined from these in a consistent way. This is done by assigning each allele $i$ in the Hybrid to a parental population $k$.

The intuitive way of doing this involves iterating through the alleles in $P_h$ and assigning them to $P_k$ based upon $p_k$. This can give terrible estimates, so other methods should be considered.

Three Candidate Methods
Three methods have been proposed to more efficiently assign the alleles in the Hybrid population to the parental populations:

Method 1
Send any allele $i$ to population $j$ with probability proportional to

\[ p_j (n_{ji} + 1) \sum_{k=1}^{N_h} n_{ji} + 1 \]

Method 2
Pick a population with probability $p_h$ then an allele $n_{ji} + 1$ for this fixed $k$.

Method 3
Pick an allele $i$ from among those present in the Hybrid with probability $\propto n_{1i}$, then a population proportional to

\[ \frac{p_k (n_{1i} + 1)}{\sum_{k=1}^{N_h} n_{1i} + 1} \]

and then the population which it is to be allocated to. This process continues (with new probabilities calculated each time) until there are no more alleles remaining in $n_h$.

Simulation
These methods (Method 0 along with Methods 1, 2 and 3) were simulated using C++ over a number of different inputs $(n_1 = (1, 2, H))$. From each method for each iteration a log-probability was returned based upon the multinomial-Dirichlet distribution and the appropriate importance-sampling weights. The system was simulated with

- the set of data given in the example used in slides by Jean-Marie Cornuet.
- five sets of data generated from the negative-binomial distribution.
- systematically chosen sets of data where $n_1$ and $n_2$ have fixed numbers of empty lineages (i.e. $n_{1i} = 0$ ($k = 1, 2$)).

An example of this type of data:

\[ n_{1i} = (0, 0, 0, 12, 12) \]
\[ n_{2i} = (0, 0, 12, 12) \]
\[ n_{1j} = (10, 10, 10, 10, 10) \]

Results
When considering the variation of the generated probabilities, the most useful results are in the form of figures showing the log-probabilities for all 4 methods over a number of iterations (Variance can be calculated as part of the simulation, but is easier to see/analyse via this sort of figure; narrower bands indicate higher efficiency).

Method 3 is more efficient than Method 0, 1 and 2. Method 3 is so good is somewhat surprising considering that it is the simplest method. Since Methods 0 and 3 are superior, it is possible that the lower variance of these methods can be attributed to the process of first selecting the allele $i$ from $n_h$ to be allocated, as Methods 0 and 3 both define which allele will be allocated before choosing which population to allocate to.

In general, the fewer zeroes $n_1$ and $n_2$ have, the closer the mean log-probabilities of Methods 0 and 3 are to the analytical solution. Method 3 with the very nearly full $n_1$ and $n_2$ is particularly close.

In terms of variance, Method 3 is generally slightly better than Method 0. The variance in the negative binomial datasets is similar, but the difference is somewhat noticeable in the Cornuet’s example data and in the mostly-full input data, except when $n_1$ and $n_2$ are both non-zero (in this simulation, the value 12 was used). It should be noted that these log-probabilities are weighted, and hence their average is calculated via the sum of their Importance Sampling weights, so all averages end up very close to each other/the analytical solution.

The difference between the analytical probability and the simulated mean probability for each method can also be considered.

Acknowledgments
- Thanks to Martin O’Hely for supervising the project. Also for the code $\text{log}\_\text{pdf}$ and $\text{rep} \_\text{sample}\_\text{dir}$ functions which were used.
- Also requiring acknowledgment are the slides by Jean-Marie Cornuet defining the 3 methods of allocating the Hybrid alleles to parental populations.

For further information, please contact Robert Cope at 42097544@student.uq.edu.au

Analysis
It is clear from both observing the figures showing variance in log-probability, and the differences between the analytical solution and the mean solutions, that in the vast majority of cases Methods 0 and 3 are clearly superior to Methods 1 and 2. The fact that Method 0 is so good is somewhat surprising considering that it is the simplest method.