Population Genetics Problem #1

Microsatellite loci are widely used in population genetics for analysis of population structure, for inferring parentage, and estimating relationships. They have two features that make them very attractive for such purposes: (1) They have a high rate of mutation, leading to a large amount of variability within populations. (2) They are co-dominant, meaning that alleles contributed by both parents can be seen in their offspring – or so we usually assume.

It turns out that microsatellite loci often have "null" alleles, i.e., alleles that fail to amplify in a PCR reaction. If you're not aware that these alleles are present, then you may mistakenly think an individual is homozygous for a particular allele, when she's actually heterozygous for that allele and a null allele. de Sousa et al. [1] investigated this problem in Norway spruce, *Picea abies*. The following table shows a portion of the data they collected in their analysis. Using these data, answer the following questions:

Maternal Genotype	Seed genotype				
A_0A_{123}	A_0A_0	A_0A_x	A_0A_{123}	$A_{123}A_x$	$A_{123}A_{123}$
	1	10	0	10	3
$A_{113}A_{119}$	$A_{113}A_{113}$	$A_{113}A_y$	$A_{119}A_y$	$A_{113}A_{119}$	$A_{119}A_{119}$
	1	12	10	0	0

1. Does meiosis appear to be fair in both mothers?[1]

   Both mothers are heterozygous, and meiosis is fair if they produce each of the two possible gametes in equal proportions. For the first mother

$$\begin{aligned} N_0 &= 11 \\ N_{123} &= 13 \quad . \end{aligned}$$

   For the second mother

$$\begin{aligned} N_{113} &= 13 \\ N_{119} &= 10 \quad . \end{aligned}$$

[1] **Warning**: One of the things you'll learn about me is that when I ask questions like this I'm not looking for a simple "yes" or "no" answer. You'll have to explain why you came to the conclusion, and to do that you'll probably find it useful to do some statistical calculations. In this case, those calculations are pretty simple.

Comparing the first outcome to the expected 12:12 with a χ^2 goodness of fit test gives us $\chi^2 = 0.1667$ with 1 degree of freedom: $P = 0.6831$. The second outcome gives us $\chi^2 = 0.3913$, $P = 0.5316$.

In short, there's no evidence that meiosis is unfair, which isn't quite the same thing as evidence that it *is* fair, but it will do. A better approach would be to estimate the segregation ratio and provide confidence intervals (classical statistics) or credible intervals (Bayesian statistics) for it.

2. Given that A_0 is a null allele, what are the estimated frequencies of A_0, A_{113}, A_{119}, A_{123}, and all alleles other than these four in pollen that fertilized the ovules on these trees? You can group all of the other alleles together into a single "pseudoallele" and call it A_u.[2]

   The allele denoted as A_x for the first mother refers to all alleles other than A_0 or A_{123}, including A_{113} and A_{119}. The allele denoted as A_y for the second mother refers to all alleles other than A_0, A_{113} or A_{119}, including A_{123}.

   The key to answering this question is to construct a segregation table for each maternal plant.[3] Specifically,

		Offspring phenotype										
Maternal plant	gamete	0	113	119	123	u	113/119	113/123	113/u	119/123	119/u	123/u
$A_0 A_{123}$	0	p_0	p_{113}	p_{119}	p_{123}	p_u	0	0	0	0	0	0
	123	0	0	0	$p_{123} + p_0$	0	0	p_{113}	0	p_{119}	0	p_u
	Total	$p_0/2$	$p_{113}/2$	$p_{119}/2$	$p_{123} + p_0/2$	$p_u/2$	0	$p_{113}/2$	0	$p_{119}/2$	0	$p_u/2$
$A_{113} A_{119}$	113	0	$p_{113} + p_0$	0	0	0	p_{119}	p_{123}	p_u	0	0	0
	119	0	0	$p_{119} + p_0$	0	0	p_{113}	0	0	p_{123}	p_u	0
	Total	0	$(p_{113} + p_0)/2$	$(p_{119} + p_0)/2$	0	0	$(p_{113} + p_{119})/2$	$p_{123}/2$	$p_u/2$	$p_{123}/2$	$p_u/2$	0

   Now given those frequencies, we can calculate the expected frequency for each offspring class in the original table.

Maternal Genotype	Seed genotype				
$A_0 A_{123}$	$A_0 A_0$ $p_0/2$	$A_0 A_x$ $(p_{113} + p_{119} + p_u)/2$	$A_{123} A_x$ $(p_{113} + p_{119} + p_u)/2$	$A_{123} A_{123}$ $p_{123} + p_0/2$	
$A_{113} A_{119}$	$A_{113} A_{113}$ $(p_{113} + p_0)/2$	$A_{113} A_y$ $(p_{123} + p_u)/2$	$A_{119} A_y$ $(p_{123} + p_u)/2$	$A_{113} A_{119}$ $(p_{113} + p_{119})/2$	$A_{119} A_{119}$ $(p_{119} + p_0)/2$

[2]To answer this question you will almost certainly need to use `WinBUGS`. You could design an E-M algorithm to calculate the maximum-likelihood estimate, but getting the Bayesian estimate with `WinBUGS` is a lot easier. In the unlikely event that you construct an E-M algorithm to solve the problem, describe the algorithm in enough detail that we can give you partial credit if you make any mistakes. In the likely event that you use `WinBUGS`, include your source code in your answer.

[3]Or at least it's the key for me. Your mileage may vary.

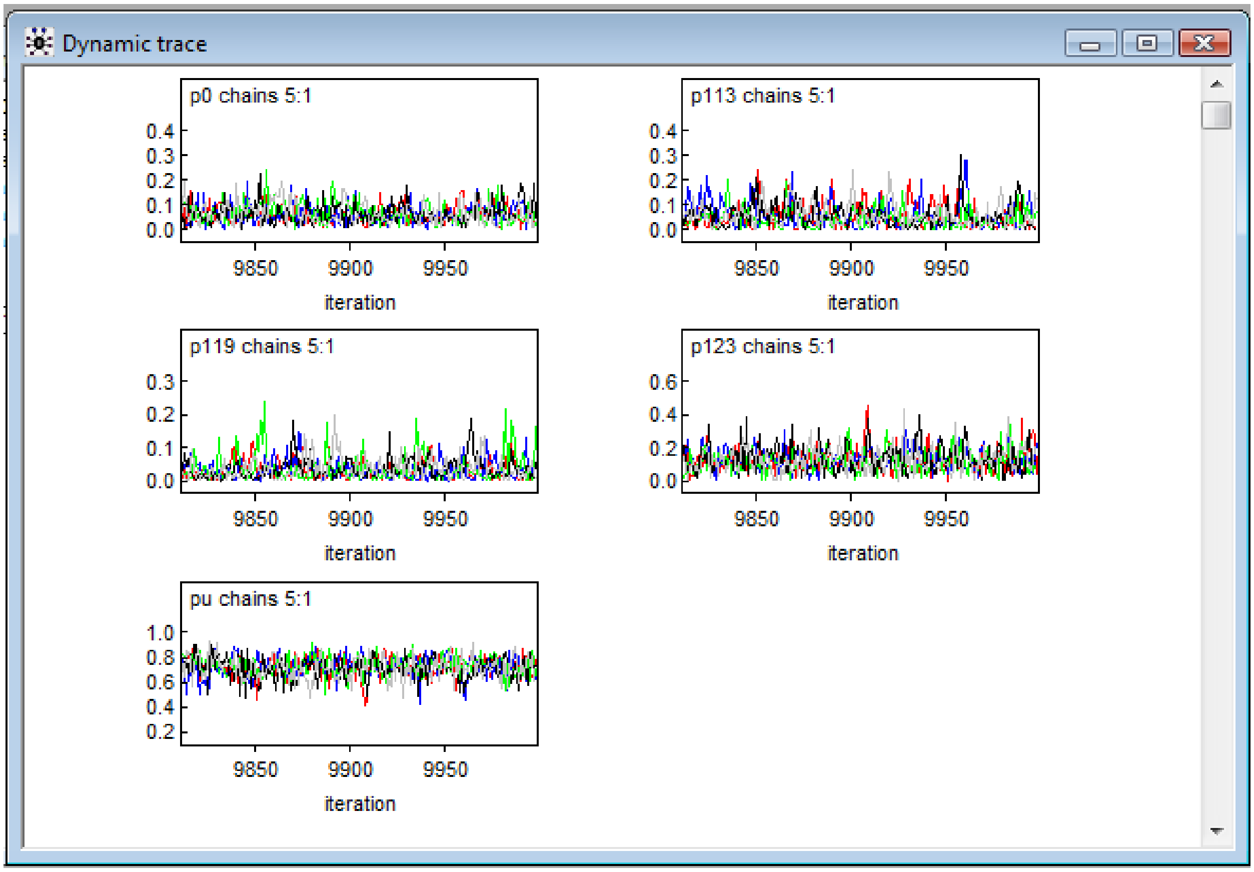
Now we're ready to set up a WinBUGS model to get estimates of the allele frequencies. Here's the code:

```
model {
  # mother 1
  mother1[1:4] ~ dmulti(p1[], n1)
  p1[1] <- p0/2
  p1[2] <- (p113 + p119 + pu)/2
  p1[3] <- (p113 + p119 + pu)/2
  p1[4] <- p123 + p0/2
  # mother 2
  mother2[1:5] ~ dmulti(p2[], n2)
  p2[1] <- (p113 + p0)/2
  p2[2] <- (p123 + pu)/2
  p2[3] <- (p123 + pu)/2
  p2[4] <- (p113 + p119)/2
  p2[5] <- (p119 + p0)/2

  # priors for p
  a0 ~ dexp(1)
  au ~ dexp(1)
  a113 ~ dexp(1)
  a119 ~ dexp(1)
  a123 ~ dexp(1)
  sum <- a0 + au + a113 + a119 + a123
  p0 <- a0/sum
  pu <- au/sum
  p113 <- a113/sum
  p119 <- a119/sum
  p123 <- a123/sum
}

list(mother1=c(1, 10, 10, 3), mother2=c(1, 12, 10, 0, 0), n1=24, n2=23)
```

The trace looks good, so the statistics provide reasonable posterior estimates.



Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
p0	0.06396	0.03918	3.159E-4	0.009246	0.05736	0.1571	5001	25000
p113	0.05267	0.04603	4.606E-4	0.001654	0.04019	0.1707	5001	25000
p119	0.03698	0.03575	3.807E-4	9.398E-4	0.02633	0.1334	5001	25000
p123	0.1254	0.07144	5.17E-4	0.01506	0.1161	0.2901	5001	25000
pu	0.721	0.089	7.448E-4	0.5212	0.7294	0.8696	5001	25000

3. **Bonus question**: What frequency of homozygotes would you expect given these estimates of allele frequencies? Now suppose that you didn't know there was a null allele and that you calculated allele frequencies from these data.[4] What frequency of homozygotes would you expect?

[4]The null allele homozygote wouldn't be in these data, so exclude it. To get the frequency of the remaining alleles, divide their frequency by $1 - p_u$.

Given all five allele frequencies

$$\begin{aligned}\text{homozygosity} &= 0.06396^2 + 0.05267^2 + 0.03698^2 + 0.1254^2 + 0.721^2 \\ &= 0.5438 \quad .\end{aligned}$$

Allele frequencies and homozygosity that would be estimated if we didn't know about null alleles

$$\begin{aligned}p_{113} &= 0.05267/(1 - 0.06396) = 0.05627 \\ p_{119} &= 0.03698/(1 - 0.06396) = 0.03951 \\ p_{123} &= 0.1254/(1 - 0.06396) = 0.1340 \\ p_u &= 0.721/(1 - 0.06396) = 0.7703 \\ \text{homozygosity} &= 0.05627^2 + 0.03951^2 + 0.1340^2 + 0.7703^2 \\ &= 0.6160 \quad .\end{aligned}$$

So if we have null alleles in a sample and don't know about it, we'll think that homozygotes are more common than they actually are.

References

[1] S. N. de Sousa, R. Finkeldy, and O. Gailing. Experimental verification of microsatellite null alleles in Norway spruce (*Picea abies* [L.] Karst.): implications for population genetic studies. *Plant Molecular Biology Reporter*, 23:113–119, 2005.