Population genetics of polyploids

ForBio
Forskerskole i Biosystematikk
Research School in Biosystematics

Oslo, Norway, December 3-7, 2018

Patrick Meirmans, University of Amsterdam, p.g.meirmans@uva.nl

Marc Stift, University of Konstanz, marcstift@gmail.com

Filip Kolar, Charles University Prague, fillip.kolar@gmail.com

Olivier Hardy, Université Libre de Bruxelles, ohardy@ulb.ac.be

Patrick Monnahan, University of Minnesota, pmonnahan@gmail.com

Camille Roux, Université de Lausanne, camille.roux.1983@gmail.com

CONTENTS

Day 1:
Diversity and Differentiation

Part 1: HWE and genetic diversity

Hardy-Weinberg equilibrium

The Hardy-Weinberg principle is one of the cornerstones of population genetics, and is something you have probably heard of. It deals with the expected genotype frequencies under random mating, in a single infinitely large population (without selection and other disturbing factors). However, it was originally formulated for diploids and works somewhat differently for polyploids.

Imagine a locus with two alleles, A and B, that have the frequencies p=0.8 and q=0.2, respectively.

1. *To freshen up your memory: for a diploid population what are the expected genotype frequencies under HW-equilibrium?*

Now imagine a locus with the same properties (biallelic, p=0.8 and q=0.2), but now in a population of autotetraploids (so assuming tetrasomic inheritance). Let's address the same problem as above, but in somewhat smaller steps.

2. *Which genotypes can be formed?*

The easiest way to derive the genotype frequencies for an autotetraploid is to look at the gametes produced by this population.

3. *Which diploid gamete genotypes (allelic combinations) can be produced?*

Assume that combining of alleles to form diploid gametes is completely random.

4. *Given the population allele frequencies above, what will be the frequencies of these gametes?*

Assume that combining of gametes to form tetraploid zygotes is completely random.

5. *What will be the frequency of each genotype in the progeny? Check if the frequencies that you calculated sum up to one. HINT: make a cross table of all pairs of gametes.*

In diploids, HWE is reached in a single generation of random mating. Let's evaluate if this also holds for autopolyploids. Imagine that two previously separated populations of tetraploids have fused and the new population consists for 50% of genotype *AAAA* and for 50% of genotype *BBBB*.

6. *Under random mating, which gametes are produced by this population and at what frequencies?*

7. *Assume these gametes unite at random – so undergo a single generation of random mating. How can you immediately see that this population is NOT in Hardy Weinberg equilibrium?*

The inbreeding coefficient F_{IS}

To quantify deviation from HWE it is common to calculate the summary statistic F_{IS}, which is calculated by comparing the observed heterozygosity (H_O) with the expected heterozygosity (H_S) under HWE. The heterozygosity is calculated as the frequency of heterozygotes in the population. Note that though expected heterozygosity is often abbreviated as H_E, we prefer to use H_S here, which stands for 'H_E in the Subpopulation'. Later, this will allow us to distinguish it from H_T, which stands for 'H_E in the Total population', i.e., a collection of multiple subpopulations. For now, we only deal with a single population, for which the inbreeding coefficient can be calculated as:

$$F_{IS} = (H_S - H_O) / H_S$$

The value of F_{IS} ranges from -1, indicating a complete lack of homozygotes, to 1, indicating a complete lack of heterozygotes. A value of 0 means that there are exactly as many heterozygotes as expected under HWE.

8. *For a diploid population with genotype frequencies AA=0.3, AB=0.2, and BB=0.5, what are the values of H_O, H_S and F_{IS}?*

When there are multiple alleles, the number of heterozygous genotypes increases quickly with the number of alleles. Therefore, it is easier to look at the homozygotes for calculating H_S. So, for a diploid, H_S can be calculated by summing over all the different alleles as:

$$H_S = 1 - \sum p_i^2$$

where p_i is the frequency of allele *i*.

9. *How would you then calculate the expected heterozygosity for a single multi-allelic locus in a sample of* tetraploids*?*

Genetic diversity

H_S is generally used as a measure of genetic diversity (then called the gene diversity, Nei 1987), which allows comparison of genetic diversity among populations or species.

10. *Take a locus with four alleles with frequencies 0.55, 0.2, 0.15 and 0.1. Using the above equations, what would be the expected heterozygosity for diploids and for tetraploids?*

11. *So, are polyploids more diverse than diploids despite the same allele frequencies?*

So you (supposedly) saw that calculating H_S for polyploids by only looking at the expected homozygote frequencies will make comparisons among ploidy levels impossible. Therefore, it is necessary to now switch off your biological common sense.

Generally, it is agreed upon to calculate the gene diversity for polyploids in exactly the same way as for diploids. This does mean that the gene diversity can no longer be called the “expected heterozygosity”. Nevertheless it is common to still indicate this with H_S. So, to calculate H_S, regardless of ploidy level, the same equation introduced for diploids (see above) is used:

$$H_S = 1 - \sum p_i^2$$

So to recap: no matter what the ploidy level is, the square term in the equation is never replaced with a higher (or lower) power.

12. *Take a locus with four alleles with frequencies 0.39, 0.28, 0.27 and 0.06. What would be the gene diversity for diploids and for tetraploids?*

13. *Write down some possible tetraploid genotypes for a locus with alleles A, B, C, and D (being exhaustive is not necessary). Are the heterozygotes all equally heterozygous?*

If we want to compare H_S (expected H) with H_O (observed H) in a tetraploid, for example for calculating F_{IS}, we have to use a trick (similar to the one we used for H_S) to calculate H_O. One way to calculate H_O for a tetraploid is by calculating the so-called "gametic heterozygosity". For a given tetraploid genotype, the gametic heterozygosity can be calculated by randomly combining its alleles into diploid gametes and assessing the frequency of heterozygous gametes. Take, for example, genotype $AABB$. There is a probability of 0.5 that the first drawn allele is an *A*. To obtain a heterozygous gamete, the second allele drawn should be a *B* (which has a probability of 2/3). Alternatively, a heterozygous gamete can be formed by first drawing allele *B* and then *A*, which has the same probability (for this genotype).

14. *What is then the combined probability of drawing a heterozygote diploid gamete for genotype AABB? In other words, what is the gametic heterozygosity for this genotype?*

15. *Which would you (intuitively) think has a higher heterozygosity: AAAB or AABB?*

16. *What is the gametic heterozygosity of genotypes AAAA, AAAB, CCDD, AABC, and ABCD? It helps to enumerate all the possible ways to draw gametes from a genotype.*

Once we have gametic heterozygosity for each genotype, we can average over all genotypes in the population to calculate H_O for a sample of tetraploids. The same concept of gametic heterozygosity can be applied to other ploidy levels. Note that for this we would still use conceptual diploid gametes, even for ploidy levels that do not actually produce diploid gametes (just like we never replace the square term in the equation for H_S for ploidies other than diploid).

17. *What is the gametic heterozygosity for the octoploid genotype AAAAAAAB?*

18. *What would be the reason why the gametic heterozygosity for an octoploid is calculated assuming diploid gametes and not tetraploid gametes?*

The approach of gametic heterozygosity allows comparison of H_O to H_S when it is calculated as if the species were diploid (as explained above), meaning we can calculate a more meaningful estimate of F_{IS}. Doing this by hand is a bit too tedious for this practical, but luckily there is software

for this. For now, just remember that calculating these summary statistics for polyploids requires some additional steps.

Simulating genetic diversity

The genetic diversity of a population depends on a number of factors such as the mutation rate, population size, etc. Here, we will use simulations to see how polyploidy affects the level of diversity.

For these simulations, you will use the R-script "Heterozygosity.R". The script first establishes a single population consisting of a specified number of individuals, all of the same ploidy level. However, individuals are not modelled explicitly; there is just an array containing the allele frequencies at a given number of loci. Genetic drift is simulated by drawing random numbers from a multinomial distribution. The expectation for these random draws are based on the current allele frequencies, with a bit of mutation sprinkled on top. Every random draw represents a single generation of random mating.

19. *Run the script with the default settings and study the resulting graph. Describe what is happening here.*

20. *Now start with maximum diversity (equal initial frequencies for all alleles at a locus), what changes?*

21. *Change the ploidy level to several different values and run the script for each (again start with maximum diversity). Describe what happens to the equilibrium level of genetic diversity (for which we here take the average value over last 1000 generations).*

22. *Now run the model for a tetraploid population and write down the equilibrium value of H_S. Now set the ploidy level back to diploid. Which other parameter do you need to change to get –approximately– the same level of diversity as for the tetraploids?*

Part 2: Population differentiation

The fixation coefficient F_{ST}

Up to now we have discussed only a single population, but most population genetic analyses actually focus on multiple populations. Analysing the differentiation among populations allows us to make inferences on a range of topics such as migration, historical processes, conservation, and adaptation. In a way, looking at genetic differentiation amounts to looking at genetic diversity, but then how it is distributed within and among populations. To quantify the degree of population differentiation, the summary statistic F_{ST} is used –a close relative of F_{IS} that we used above. F_{ST} is calculated as a comparison of the expected heterozygosity within populations (H_S) and the expected heterozygosity of the total population (H_T):

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

The values of F_{ST} range from 0, indicating no differentiation, to 1, indicating fixation in all populations: all populations only have a single allele left, but this is not always the same allele in all populations.

The simplest model of population structure is the island model, which is widely used in population genetics. Under the island model there is a set of populations that all have the same number of individuals (N). Mating within populations is completely random and the species are hermaphroditic annuals. Population connectivity is modelled as equal migration among all populations, meaning that there is no spatial structure. As above, there is also some mutation. Under the island model, the equilibrium value of F_{ST} depends on the balance of drift, mutation, and migration.

Here, we will perform some simulations of a set of populations under the island model. The first simulations we will look at are in the script “Genetic Differentiation Fst”. This script is a bit more complicated than the previous, so take your time to go through it. Start by looking at the function called *sum.stats* that is defined at the top, which will calculate the summary statistics.

In contrast to the previous simulations we did for the heterozygosity, we now use a locus with only two alleles. This means that the populations can be represented by a simple array with every row representing a population and every column a locus. The cell value then represents the number of copies of allele A in the population. The allele frequency can thus be obtained by dividing this value by the total number of chromosome copies in the population (the number of individuals times the ploidy level).

23. ***Extra for R-gurus:*** *Find the spot where the populations are being initialised. Do all populations have exactly the same allele frequencies? If not, why is there variation?*

The core mechanics of the model are very similar to that of the previous model: all stochasticity –in migration, mutation and drift– derives from drawing random numbers. Last time we used a multinomial distribution since we had multi-allelic loci, but this time a binomial distribution will suffice since we have biallelic data now.

24. ***Extra for R-gurus:*** *The spots in the code where mutation and migration are implemented should be easy to find, but can you also pinpoint where drift is implemented?*

25. *Run the script with the default settings and study the resulting graph. Describe what is happening here.*

26. *Change the ploidy level to several different values and run the script for each. Describe what happens to the equilibrium level of F_{ST}.*

The script "Compare F-stats.R" will create a plot of the value of F_{ST} as a function of the migration rate for different ploidy levels (based on theoretical expectations, which we will not go into here). Run the script and study the output.

27. *At what migration rate do you see the largest difference in value between the ploidy levels?*

Alternatives to F_{ST}

Despite its wide use, there are some serious problems with F_{ST}: its value depends on the mutation rate (not specific to polyploids). This is annoying, as we prefer it to describe population connectivity. A problem that is important for this workshop is that F_{ST} also depends on the ploidy level. This complicates comparisons among ploidy levels . To overcome these problems, several alternative statistics have been proposed. The statistics F'_{ST} (Meirmans & Hedrick 2011) and D (Jost 2008) are supposed to solve the dependence on the mutation rate. The *rho*-statistic (Ronfort et al 1998) was developed to circumvent the problems related to polyploids, and is therefore of particular relevance for this workshop.

28. ***Extra for R-gurus:*** *Open the script "Genetic Differentiation Rho.R", and have a look at it. What are the differences with the previous script?*

29. *Run the script "Genetic Differentiation Rho.R" for different ploidy levels. What happens with the value of rho?*

30. *For what ploidy level is the value of rho equal to that of F_{ST}?*

Run the script "Compare F-stats.R" again, but now modify it to plot *rho* instead of F_{ST}.

31. *What difference do you see with the plot you made previously for F_{ST}?*

32. *Which statistic is preferable for comparing population differentiation across ploidy levels, based on these results?*

Part 3: The missing dosage information

In diploids, the distinction between homozygotes and heterozygotes is very straightforward: individuals with one allele have two copies of that allele (they are homozygote); individuals with two alleles have one copy of each (they are heterozygote). However, for polyploid individuals, it is unlikely that you can obtain fully resolved genotypes (i.e., to determine the dosage). When the two alleles *A* and *B* are found in a tetraploid, this could be any of the partially heterozygous genotypes *AAAB*, *AABB*, or *ABBB*. In practice, it is often impossible to distinguish between these based on the number of reads or gel intensities, so such individuals are usually simply coded in a partially

dominant way, e.g. as *AB*. This missing dosage information introduces a bias in the calculation of the allele frequencies and the summary statistics for differentiation and diversity.

Assume a sample from a tetraploid population with the following genotype frequencies:

Genotype	Individuals	Partially dominant genotype
AAAA	10	
AAAB	20	
AABB	70	
ABBB	80	
BBBB	20	

33. *What are the allele frequencies when the dosage for the genotype is known (as in the first column)?*

34. *Fill in the genotypes you see when the dosages are not known. What are the allele frequencies when those values are used?*

35. *Why don´t we just forget about dosage, and analyse markers in a partially dominant fashion?*

The next generation sequencing revolution has greatly facilitated genotyping and thus population genetics.

36. *Does NGS data also suffer from problems in the determination of dosage?*

For SNPs called from NGS genotyping, the strength of the problem of missing dosage depends on the sequencing coverage. Actually, when the coverage is very low the missing dosage problem also applies to diploids. Imagine a ridiculously low coverage of 2.

37. *What is the probability of correctly inferring the genotype for a diploid individual that is heterozygous for a certain SNP?*

With reasonable sequencing depth, dosage can be inferred from the number of times the different alleles are encountered at a locus. Since polyploids are more complex than diploids, a higher sequencing depth is needed for a good scoring of heterozygotes.

The script titled "Overlap.R" addresses how well different genotypes at a biallelic locus can be separated for tetraploids given a specified sequencing depth. Note that the `rbinom` function is used here, as that lends itself better to the creation of histograms.

38. ***Extra for R-gurus:*** *Run the script with multiple sequencing depths. What is the lowest sequencing depth at which you think the results are still acceptable?*

Part 4: Mixing ploidy levels

Possibly the most interesting evolutionary questions that can be analysed using population genetics involve the analysis of multiple ploidy levels in a single species. Unfortunately, this is also when the problem of missing dosage information is especially troublesome.

Simulated data

Assume a single population where diploids and tetraploids co-occur and freely interbreed (somewhat unrealistic, we know). In other words, the diploids and tetraploids form a single gene pool and have exactly the same allele frequencies. A sample of 100 individuals is taken from each cytotype and analysed using 100 SNP markers. However, the dosage information is missing for the tetraploids.

39. Why don´t you expect any separation between these two cytotypes when performing a PCA?

Such a PCA is simulated in the script "SNP simulation.R". Note that in this script we do not simulate generations of random mating with mutation and drift. Instead, we directly draw random allele frequencies and use these to construct diploid and tetraploid genotypes. These genotypes are then stripped of their dosage information.

40. Did your above expectation turn out correct? What is happening here?

41. Do you get better results when you increase the number of loci?

Set the number of loci back to the default of 100 and change the cytotype of ploidy2 to hexadecaploid (2n=16x), and run the model.

42. *What happens to the spread of points of the hexadecaploids along the second PCA axis? Is this because they have less genetic diversity?*

Part 5: Structure

The program Structure by Pritchard et al (2000) is a widely used method to detect clustering in population genetic data. We will not go into the exact working of the program in this workshop. In short, the program uses assignment methods to assign individuals to a specified number of populations (k). It then uses Bayesian methods, including a Monte Carlo Markov Chain to find the optimal distribution of individuals over clusters. One of the most interesting aspects is that individuals are allowed to be admixed, meaning that they can be assigned to multiple clusters.

Structure has special provisions for polyploid data, so it is worthwhile to have a look at that. This means that we will soon have to leave the comfy confines of R and venture into the point-and-click user interface of Structure. Fortunately, to work with Structure, we need data and for this we will –of course– use an R-script to simulate genetic data. The simulation part of this script ("Structure.R") works mostly like the one for genetic differentiation we used above, with a number of generations of sampling from a binomial distribution, with expectations that include mutation and migration. This is followed by quite a bit of code to create individuals with either dominant, fully codominant or codominant genotype data with missing dosage information. Finally, the simulated genotype data is written to a file in the correct format for Structure.

The interesting thing about the used script is that there is a number of populations that are allowed to differentiate from each other, but there are also two ploidy levels in each population. This allows us to simultaneously analyse true structure among populations and spurious separation between cytotypes that arises from the missing dosage.

Run the script to just before the point where the data gets written to a file, and look at the PCA plot based on dominant data.

43. *Is the separation between the ploidy levels or the differentiation between the populations the most important aspect in the PCA plot?*

44. *Which parameter(s) do you have to change to place the separation between the ploidy levels on the first PCA axis?*

Now it's time to get the data into *Structure*. Reset the script to its default values and run it completely. Now open the program *Structure* and select "New Project" from the "File" menu. Give the project a name and browse to the directory that contains the files that you just created. Then select the file with codominant data (let's start simple).

You then will be asked a number of questions. The answer to most of these you should be able to figure out from the settings of the script. Leave the checkboxes that you don't understand unchecked, except "Individual ID for each individual" and "Putative population origin for each individual", which should be checked. You will undoubtedly make an error somewhere and it will complain; in that case go back and see if you can fix it (see it as a training for working with your own data).

When you finally managed to get your data in, you have to create a parameter set. Here, use 1000 steps for the burn-in and 10000 for the MCMC (these are short, but will suffice for the first run). Give the parameter set a sensible name. Now click Run and set the number of assumed populations (k) to 2.

45. *Does Structure split the dataset by ploidy level, or by population?*

46. *Try different levels of k and check the results for bias due to populations consisting of a mixture of diploids and tetraploids.*

47. *Now do the same for the dominant data and the data with the missing dosage. Note that for the dominant data, you have to check the box labelled "Row of recessive alleles".*

Day 2:
Mixed Mating Model

Part 1: Introducing the Mixed mating model (MMM)

The mixed mating model (MMM) assumes that a population made of hermaphrodites generates offspring by selfing at rate *s* and by outcrossing (random mating) at rate 1-*s*. This model is relevant for many plant species and a few animal species. Selfing results in more inbred offspring, causing a heterozygote deficit, hence $H_O < H_S$ and $F_{IS} > 0$ (recall that $F_{IS} = 1 - H_O / H_S$). Outcrossing generates offspring for which $H_O = H_S$ and $F_{IS} = 0$, at least under random mating. Therefore, under the MMM, F_{IS} asymptotically reaches a value that depends only on *s*, and the ploidy level. We will see how to derive this asymptotic value because it is of interest to estimate *s* from a marker-based estimate of F_{IS}.

48. *When a diploid of genotype AB and a tetraploid of genotype ABCD self-fertilize, what are the offspring genotypes and their relative frequencies? (Hint: For tetraploids, first write down all the possible gametes, that can be formed, and their frequencies)*

49. *By which ratio does the observed heterozygosity decrease per generation after selfing of i) a diploid heterozygote (AB) and ii) a fully heterozygous tetraploid (ABCD)? Calculate H_O for the parent as well as for the offspring outlined in the previous answer. Remember that the (gametic) heterozygosity of an autopolyploid is the fraction of pairs of alleles that are not identical in state (possible values of H_O in a 4x are 0, 1/2, 2/3, 5/6, 1).*

50. *Imagine a large <u>diploid</u> population (so with negligible drift) under the MMM. Derive the observed heterozygosity in the offspring generation (H_O') according to the observed heterozygosity in the parental generation (H_O), the selfing rate (s) and the gene diversity (H_S). Hint: While $H_O = H_S$ when there is no drift and no selfing, think which fraction of the offspring population contributes to a reduced heterozygosity as calculated above, and which fraction does not.*

51. *Same question for a <u>tetraploid</u>. This is more subtle because among the 6 possible pairs of alleles from an individual, 2 pairs correspond to the alleles found within the uniting gametes and therefore do not contribute at all to any change in heterozygosity (the heterozygosity for these pairs corresponds to the heterozygosity in the parental generation). The remaining 4 pairs correspond to alleles sampled between uniting gametes (heterozygosity for these pairs is affected by selfing vs outcrossing). If you succeed, you can try to generalize the transition equation for a k-ploid.*

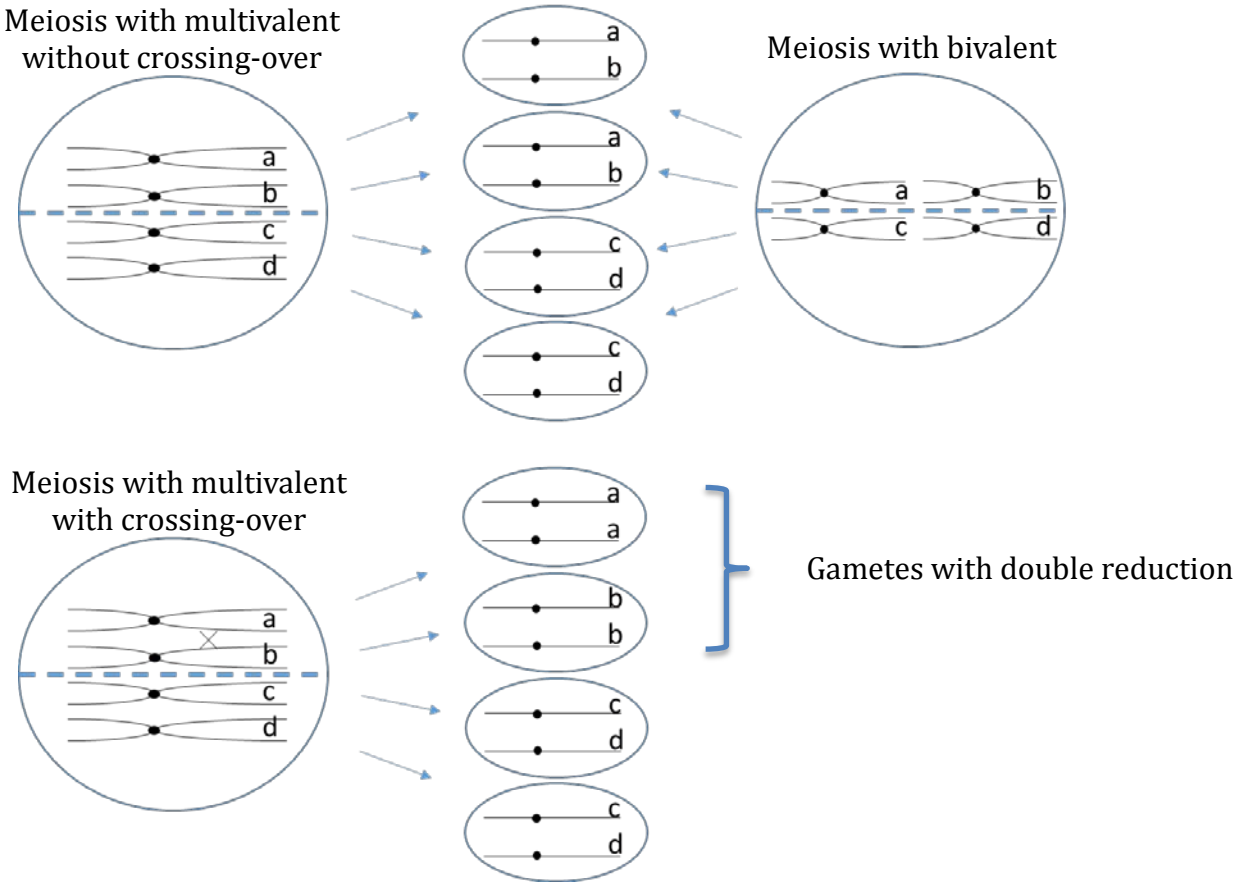
52. *Now, consider that an equilibrium has been reached, so that H_O'=H_O. What is the equilibrium F_{IS} in a diploid, tetraploid, or k-ploid? Hint: Solve the equation and re-arrange it to have on one side the fraction Ho/Hs.*

The resulting equations are easily inverted, allowing to estimate the selfing rate (s) from the level of inbreeding (F_{IS}): $\hat{s} = \frac{k F_{IS}}{1+(k-1)F_{IS}}$

However, estimating F_{IS} from heterozygote deficit is prone to different kinds of biases, especially in polyploids. A first source of bias –independent of the ploidy level, but possibly exacerbated by it– is the possible presence of *null alleles*, i.e. an allele that is not detected by the genotyping procedure used. For example, the first step in genotyping a microsatellite marker is to amplify a DNA fragment including the microsatellite by PCR using specific primers. However, if mutations or deletions have occurred in a primer region, the PCR may fail, leaving no PCR product (if the individual was homozygote for this null allele) or produce a genotype that appears more homozygous than it actually is, because only amplifiable alleles will be detected. Hence, a heterozygote containing a null allele will be misinterpreted as homozygote for the visible allele in a diploid, and as a more homozygous genotype in a polyploid. A second source of bias, specific to polyploids, is that *allele dosage can be difficult to assess*, so that a tetraploid showing alleles A and B could be ABBB, AABB or ABBB (with a null allele, 0, it could also be 0AAB, 0ABB or 00AB). A third source of bias (also specific to autopolyploids) that can increase the heterozygote deficit independently of selfing is related to meiosis. This phenomenon is called *double reduction* and is now addressed.

Part 2: Double reduction and consequences for MMM predictions

So far, we considered that a tetraploid genotype ABCD could only produce 6 types of gametes, corresponding to the random sampling of 2 alleles without replacement among the 4 alleles of the tetraploid (i.e. AB, AC, AD, BC, BD, CD). However, under certain conditions, gametes of type AA, BB, CC or DD can also be produced in autopolyploids, a phenomenon called ‘double reduction’. The latter can occur when multivalents are formed during meiosis, that is when all homologous chromosomes align in parallel during the metaphase of the first meiotic division, allowing crossing-over to occur between chromatids that will migrate in the same pole of the first division. After the second division, it is then possible that sections of sister chromatids (which are perfectly identical) end in the same gamete.



So, double reduction requires (i) the formation of multivalents at meiosis, and (ii) a locus sufficiently far from the centromere to have crossing-over between the centromere and the locus. The rate of double reduction, denoted α, is the probability that a pair of alleles in a gamete are copies of the same chromosome in the parent. Thus α varies among loci. If the locus is far enough from the centromere so that there is a high probability that at least one crossing-over occurs, gametic genotypes correspond to the random sampling of 2 chromatids without replacement among the $2 \cdot k$ chromatids present when the meiosis starts in a k-ploid (no chromatid segregation, only chromosome segregation, happens during first meiotic division if $\alpha = 0$). In this situation, we can consider that gametic genotypes result from chromatid segregation, while in the absence of double reduction ( ) they result from chromosome segr

53. *What is the probability that an ABCD tetraploid produces an AA gamete under chromatid segregation? In other words, what is the probability of sampling AA by randomly picking 2 alleles without replacement among the 8 following copies: AABBCCDD? What is the maximal rate of double reduction in a k-ploid?*

Double reduction could thus be seen as a form of inbreeding occurring within gametes. A general transition equation for observed heterozygosity in a k-ploid where double reduction occurs at rate α is: $H_O' = (1-\alpha)\cdot[(k/2-1)/(k-1)]\cdot H_O + [(k/2)/(k-1)]\cdot[s\cdot H_O\cdot(k-1)/k + (1-s)\cdot H_S]$

54. *For tetraploids without double reduction, is this equation equivalent to the equation you obtained in question 52?*

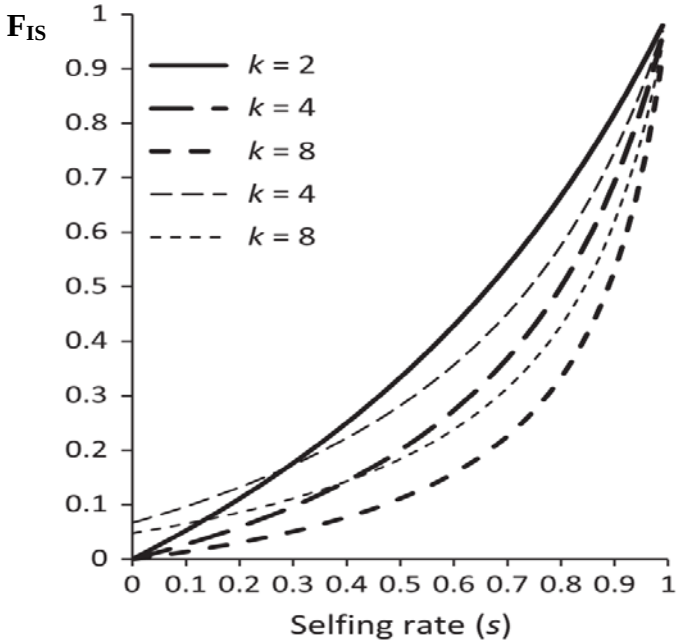
55. ***Extra for math-gurus:*** *Can you explain all terms in the transition-equation of H_O, in particular what represent the fractions involving k?*

*At equilibrium, the heterozygo*te deficit is : $F_{IS} = [s + (k\text{-}2)\cdot\alpha]/ [k – (k\text{-}1)\cdot s + (k\text{-}2)\cdot\alpha]$

Inverting this equation leads to an estimator of the selfing rate:

$$\hat{s} = \frac{k\,F_{IS} - (k-2)(1-F_{IS})\alpha}{1 + (k-1)\,F_{IS}}$$

The following graph shows how F_{IS} varies as a function of the selfing rate in 2x, 4x and 8x, without double reduction ($\alpha = 0$, chromosome segregation, thick lines) or with maximal double reduction ($\alpha = 1/(2k\text{-}1)$, chromatid segregation, thin lines).



56. Are higher-level autopolyploids more (or less) prone to inbreeding when consanguineous mating occurs than lower-level autopolyploids or diploids?

57. In autopolyploids, does significant heterozygote deficit demonstrate that mating is non-random?

Part 3: Identity disequilibrium and MMM predictions

F_{IS} is closely related to the selfing rate but it is also affected by other factors (double reduction, null alleles) limiting the accuracy at which selfing rates can be deduced from F_{IS} estimates. Therefore, the so-called 'standardized identity disequilibrium' is a better alternative, because it is strongly affected by selfing but less affected by null alleles and double reduction. We will see this below, and prove that it is more robust for estimating selfing rates than F_{IS}. The 'standardized identity disequilibrium' is a measure of the correlation between loci of the observed heterozygosity at the individual scale. We have seen above that for a single locus selfed individuals are more likely to be more homozygous than outcrossed individuals; however this effect occurs at all loci in the genome. Hence, if there is a mix of selfed and outcrossed individuals, we expect to see a correlation between loci of their respective levels of homozygosity across individuals.

Consider two loci, 1 and 2, and denote the observed heterozygosity at these loci at the population level h_1 and h_2, and the observed double heterozygosity at these loci h_{12}. Double heterozygosity is the probability of sampling (within an individual) a pair of non-identical alleles at locus 1 and at locus 2 ,. The standardized identity disequilibrium, g_{12}, is then defined as:

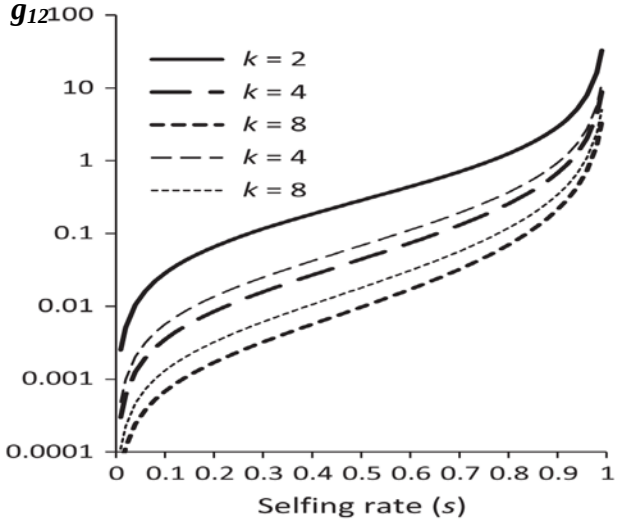
$$g_{12} = \frac{h_{12}}{h_1 h_2} - 1 \, .$$

If heterozygosity is independent between loci, we expect that the probability of double heterozygosity at loci 1 and 2 is simply the product of single-locus heterozygosities at these loci ($h_{12} = h_1.h_2$) so that $g_{12} = 0$. This is what happens in a random mating population. But in a MMM, because selfed individuals are more homozygous at all loci and outcrossed individuals are more heterozygous, heterozygosity is not independent between loci and double heterozygosity $h_{12} > h_1 \cdot h_2$, so that $g_{12} > 0$. A transition equation for h_{12} can be constructed as we did for single-locus heterozygosity but this is a bit more complex (it is detailed in Hardy 2016, Mol. Ecol. Resour.

16:103-117). At equilibrium, g_{12} (denoted γ^*_{2z}) can be expressed as a function of the selfing rate (s) and rates of double reduction at loci 1 and 2 (α_1, α_2):

$$\gamma^*_{2z} = \frac{2s(1+(k-2)\alpha_1)(1+(k-2)\alpha_2)}{(1-s)\left[(4-12k+7k^2)-(14-20k+7k^2)s+(k-2)(k-2+(3k-4)s)(\alpha_1+\alpha_2)-(k-2)^2(1+s)\alpha_1\alpha_2\right]}$$

Graphically, for 2x, 4x and 8x, with (thin line) or without (thick line) double reduction, this gives:



58. *If selfing was estimated from F_{IS} or from g_{12} but without knowing the rate of double reduction at the loci investigated, which estimate should be more accurate (compare the last two figures)?*

The previous equation looks complex but it simplifies considerably for particular cases. For example, for a tetraploid, under chromosome segregation (no double reduction, i.e., $\alpha_1=\alpha_2=0$), it reduces to:

$$\gamma^*_{2z} = \frac{s}{(1-s)(34-23s)} = \frac{s}{34-57s+23s^2}\,.$$

To estimate the selfing rate from an estimate of the standardized identity disequilibrium g_{12} (denoted g_{2z}), one can invert the previous $g_{12}(s)$ equation and consider either no double reduction, thus chromosome segregation (α=0):

$$\hat{s} = \frac{1 + (9 - 16k + 7k^2)\hat{g}_{2z} - \sqrt{1 + 2(9 - 16k + 7k^2)\hat{g}_{2z} + (4k - 5)^2\hat{g}_{2z}^2}}{(14 - 20k + 7k^2)\hat{g}_{2z}}$$

or maximal double reduction, thus chromatid segregation (α=1/(2k-1)):

$$\hat{s} = \frac{9 + (13 - 40k + 28k^2)\hat{g}_{2z} - 3\sqrt{9 + 2(13 - 40k + 28k^2)\hat{g}_{2z} + (8k - 7)^2\hat{g}_{2z}^2}}{(34 - 64k + 28k^2)\hat{g}_{2z}}$$

In tetraploids under chromosome segregation, it simplifies to: $\hat{s} = \frac{1+57\hat{g}_{2z} - \sqrt{1+114\hat{g}_{2z}+121\hat{g}_{2z}^2}}{46\,\hat{g}_{2z}}$

Now, how to estimate g_{12}? It seems natural to apply the definition: $g_{12} = \frac{h_{12}}{h_1 h_2} - 1$. However, a better estimator (less biased when computed from a limited sample of genotypes) is:

$$\hat{g}_{2z} = \frac{\sum_i h_{i1}h_{i2}\,/N}{\sum_i \sum_{j\neq i} h_{i1}h_{j2}\;/(N(N-1))} - 1 = \frac{(N-1)\sum_i h_{i1}h_{i2}}{(\sum_i h_{i1})(\sum_i h_{i2}) - \sum_i h_{i1}h_{i2}} - 1$$

where h_{i1} and h_{i2} are the degrees of heterozygosity observed in individual i at loci 1 and 2, respectively, and N is the total number of genotypes individuals sampled. The same notation can be used to define an unbiased estimator of the inbreeding coefficient:

$$\hat{F}_{IS} = 1 - \frac{\sum_l \sum_i h_{il}}{\frac{N^2}{(N-1)}\sum_l \left(1 - \sum_a p_{al}^2 - \frac{k-1}{kN^2}\sum_i h_{il}\right)}$$

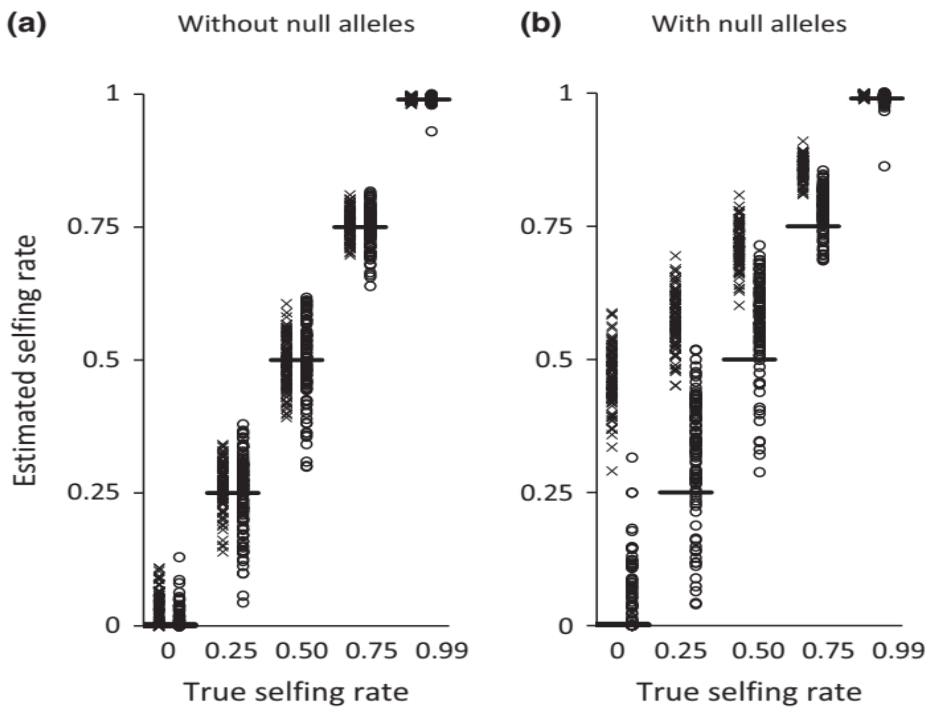
where h_{il} is the observed heterozygosity of individual i at locus l, and p_{al} is the observed frequency of allele a at locus l in the sample of N individuals.

59. *Consider a sample of 5 tetraploid individuals from a population with their genotypes at two loci as given below. For each locus, try to estimate the coefficients of inbreeding (F_{IS}) and coefficient of standardized identity disequilibrium (g_{12}).*

Individu	**Locus1**	**Locus2**	h_{i1}	h_{i2}	$h_{i12} = h_{i1} \cdot h_{i2}$
Ind1:	aabc	pqrr			
Ind2:	aaac	qqqr			
Ind3:	aabb	ppqr			
Ind4:	abbc	pqrr			
Ind5:	abbb	pprr			
Sum					

60. *The 5 genotypes above had known allele dosage. However, allele dosage is often difficult to assess. How could you estimate g_{12} if you had only phenotypes? Would it cause substantial bias? Hint: In the absence of information of allele dosage, you could consider the possible alternative genotypes corresponding to a phenotype and take the mean.*

Individu	**Locus1**	**Locus2**	**h_{i1}**	**h_{i2}**	**$h_{i12} = h_{i1}.h_{i2}$**
Ind1:	abc	pqr			
Ind2:	ac	qr			
Ind3:	ab	pqr			
Ind4:	abc	pqr			
Ind5:	ab	pr			
Sum					



Selfing rate estimates (100 replicates) in a simulated tetraploid population for each of five true selfing rate values (0, 0.25, 0.5, 0.75, 0.99; represented by horizontal segments), based on estimates of the inbreeding coefficient (F_z, crosses) or of the standardized identity disequilibrium coefficient (g_{2z}, circles). A mixed mating model population of 2500 individuals was simulated with 10 loci containing each five alleles with a triangular allele frequency distribution where (a) all alleles were assumed visible or (b) one allele was assumed to be a null allele. Selfing rate was estimated using the phenotypes of 50 randomly sampled individuals, considering in turn 100 independent resampling. Chromosome segregation (i.e. rate of double reduction $\alpha = 0$) was simulated for all loci, and accordingly, equations (15) and (17) were applied to estimate the selfing rate when F_z or $g_{2z} \geq 0$, while the estimator was set to 0 when Fz or $g_{2z} < 0$. It is worth noting that estimates are less precise (higher variance) when the true selfing rate is intermediate and that while F_z-based selfing estimates display less variance than g_{2z}-based estimates, they are much more biased in the presence of null alleles (Fig. from Hardy 2016).

61. *The figure above shows estimates of selfing rate in simulated tetraploid populations using F_{IS} (x) or g_{12} (o). When null alleles occur, estimates based on F_{IS} are much more biased (upwards) than those based on g. Can you explain this difference? Consider the definition of F_{IS} and of g_{12}, and how null alleles are expected to affect the different terms.*

Part 4: Estimating the selfing rate using SPAGeDi

The method to estimate the selfing rate in a MMM based on identity disequilibrium is implemented in the software SPAGeDi (since version 1.5). We will now see how to use it.

The file "DataAllozymes(diploids_and_tetraploids).txt" –which is one of the datasets available from the SPAGeDi webpage– contains a dataset formatted for SPAGeDi and including comments. Open it in a spreadsheet like Excel (columns are separated by tabs) to understand how it is formatted (pay attention to the meaning of the 6 format numbers). This is a 20 years old dataset from the "allozyme" era! It gives the genotypes at 5 loci of 65 diploid and 74 tetraploid individuals of *Centaurea jacea* coexisting in a meadow. An advantage of allozyme genotyping was that allele dosage could be assessed quite reliably, at least in tetraploids.

Then, open "inSpagediAfzelia155ind.txt" which is a recent dataset of 10 nuclear microsatellites genotyped for 155 individuals of a tetraploid tree of the genus *Afzelia*. These individuals were sampled in a 3 km by 3 km forest stand and the coordinates of each individual (in meters) is given in the second and third columns. Here the allele numbers represent the size of PCR products as estimated using a capillary sequencer. The four alleles of each single-locus genotype are separated by dots to read them more easily (any non-numerical character, including a space, could be used instead), but this is not compulsory (alleles can be adjoined, as in the allozyme dataset, but they must then always contain the same number of digits, 1, 2 or 3, as defined by the fifth format number). Note that for the Afzelia data set, the degree of PCR amplification was used to try to assess allele dosage: when 2 or 3 alleles were detected, they were entered by decreasing order of amplification and the remaining places (2 or 3 respectively) were filled by these alleles in this order. Hence, if a genotype show three alleles of decreasing intensities 135, 145, 123, the genotype is coded 135.145.123.135, assuming thus that there were 2 copies of allele 135.

To open the file in SPAGeDi launch the executable and follow the instructions displayed. There are a number of options to select in successive panels. To keep the output file simple, in the panel "STATISTICS…", select option "0" (no pairwise statistics). To test for statistical significance (by data permutation) and/or to obtain standard errors (by jack-knifing over loci), you must select respectively options 3 and/or 4 in the panel "COMPUTATIONAL OPTIONS". To obtain selfing rate estimates, you must select option 7 in the last panel of options: "OUTPUT OPTIONS". All results are output to a unique text file (default name is "out.txt") that can be opened as a spreadsheet. When you have categories (defined in the second column if the 2nd format number is >0), selfing rate is estimated for each category. There is one estimate using the allele dosage given in the dataset, and one estimate based on phenotypes and considering a mean individual heterozygosity across the different possible genotypes that can correspond to a phenotype (see

Table 1 of Hardy 2016 for details in 4x, 6x and 8x). These selfing rate estimators assume chromosome segregation, and thus ignore double reduction.

62. *What are the estimates of selfing rate in diploid and tetraploid Centaurea populations, and in the tetraploid Afzelia population.*

Consider now the file "inSpagediAfzelia179ind(noCat).txt". It is the same dataset as "inSpagediAfzelia155ind.txt" but more complete, as it contains 24 additional individuals of *Afzelia*.

63. *What is the selfing rate estimated in the Afzelia population with 179 sampled individuals?*

Well, it isn't normal that adding a few individuals has such a strong effects on the selfing rate estimation, is it? While the estimation method based on identity disequilibrium is fairly robust in many aspects (allele dosage, null alleles, double reduction, biparental inbreeding,…), it is unfortunately very sensitive to a form of the Wahlund effect occurring when mixing differentiated populations. To find out the origin of this problem, use STRUCTURE to assess whether there is a substructure within the dataset of 179 Afzelia individuals. You will find two datasets already formatted for STRUCTURE:
"inStructureAfzelia179indNull.txt" and "inStructureAfzelia179indNoNull.txt".

64. *Open these datasets and identify how they differ.*

Now create two projects in STRUCTURE, one for each dataset, and launch a series of runs for K=1 to K=5 with 5 replicates. When preparing the projects in STRUCTURE, note that: "-9" is the missing data value, the individuals are tetraploid, both data files contain a "Row of markers names" and for both "Data file stores data for individuals in a single line". However, the data file "inStructureAfzelia179indNull.txt" is the only one containing a "Row of recessive alleles", and thus the only STRUCTURE project that will consider the possible occurrence of null alleles in the given genotypes (which are interpreted as phenotypes corresponding to multiple genotypes differing by allele dosage, including null alleles). Finally you must check "Individual ID for each individual", declare two extra columns (they contain spatial coordinates not used by STRUCTURE but useful to map the resulting genetic clusters) and press just ok when asking "Indicator code for locus without ambiguity" (we consider that all loci can be ambiguous). For the model (select menu "Parameter set" > "New…"), you can use default parameters and "Length of Burnin period" = 10 000, "Number of MCMC reps after Burnin" = 10 000. Then (select menu "Project" > "Start a job") run 5 replicates for each value of K ranging from 1 to 5. Note that it is generally not recommended to use such a low number of replicates, but we do this to save time here in the practical.

65. *How does the likelihood of the data vary with the assumed number of cluster K? You can assess this by looking at "Simulation summary" (in the left panel), or using for example the online tool "STRUCTURE HARVESTER" if you are already familiar.*

66. *Compare now the "Bar plot" for different runs (and Sort by Q). Are there runs for which all individuals are put in well-delimited clusters without intermediates?*

67. *Were null alleles detected (scroll to the bottom of the "Simulation result" window)? At which frequency?*

Now, classify each individual as belonging to cluster 1 or cluster 2 following the results of STRUCTURE, and edit the file "inSpagediAfzelia179ind(noCat).txt" to add a column just after the names of individuals where you will indicate the cluster (note that the 179 samples in the STRUCTURE and SPAGeDi data files are in the same order). Change the 2nd format number at the first line of the file (2 instead of 0) to indicate that the data file contains 2 categories of individuals. Estimate again the selfing rate using SPAGeDi. Also select option 1 of the panel of "OUTPUT OPTIONS" to display in the result file the diversity parameters and allele frequencies per category (cluster).

68. *What are the selfing rates in the two clusters of Afzelia?*

69. *Why did the mixing of the two clusters of Afzelia lead to so biased estimates of selfing rate? Compare the diversity parameters of the two clusters.*

70. *What could be the origin of the existence of two clusters in the Afzelia population sampled?*

71. *Use SPAGeDi to estimate differentiation parameters between the two clusters. How do F_{ST} and Rho differ from each other?*

72. *Use SPAGeDi to estimate R_{ST}, a measure of differentiation similar to F_{ST} but taking microsatellite allele sizes into account. We expect $R_{ST} > F_{ST}$ if clusters tend to diverge in their mean allele sizes, a situation that can occur under stepwise mutations in the absence of gene flow and if clusters are isolated for long (for more generations than the reciprocal of the mutation rate). SPAGeDi implements a test permuting allele sizes among allelic states to assess if $R_{ST} > F_{ST}$. Does this test show a significant difference?*

Part 5: Kinship coefficients and spatial genetic structure in polyploids

Relatedness between individuals is a key notion in population genetics, that can be tackled from a genealogical perspective (a genealogy conveys information of the probabilities that genes sampled in different individuals are identical by descent, meaning that they derive from the same allele of a common ancestor) or using genetic markers (markers allow to compute probabilities of identity in states between genes sampled in different individuals). There are many measures of relatedness, among which two are more commonly used.

(1) The “coefficient of relationship”, R_{ij}, is related to the probability that an allele sampled in individual *i* is identical by descent to one of the alleles sampled in individual *j*.

(2) The “coefficient of kinship” or “coefficient of coancestry”, F_{ij}, is related to the probability that an allele sampled in individual *i* is identical by descent to an allele randomly sampled in individual *j*.

The nuance may seem subtle but, basically, between outbred *k*-ploids, $R_{ij} = k \cdot F_{ij}$ because the allele in *i* is compared to *k* alleles in *j* in the case of R_{ij}, increasing by *k* times the chance that it is identical by descent to at least one of them (assuming that alleles are independent within each individual, thus that individuals are outbred).

F_{ij} starts with an “*F*” like F_{ST} or F_{IS}, and this reflects that these statistics are of the same nature because they always correspond to probabilities of identity between pairs of alleles. They differ by the way alleles are sampled: within an individual (without replacement) in the case of F_{IS}, within a population in the case of F_{ST} and between a pair of individuals in the case of F_{ij}. It is not uncommon to find in the population genetics literature that these “*F*-statistics” ARE measures of probabilities of identity by descent, but this is an oversimplification, only marginally valid when considering infinitely large theoretical populations. They are rather ratios of differences of probabilities of identities (by descent or in state): one can define $F_{ij} = (Q_{ij} - Q) / (1 - Q)$ where Q_{ij} is probability of identity between alleles sampled in individuals *i* and *j*, and Q is probability of identity between alleles sampled in the overall population ($Q = \Sigma p_i^2$). Similarly, F_{ST} can be defined as $F_{ST} = (Q_S - Q_T) / (1 - Q_T) = (H_T - H_S) / H_T$ where Q_S and Q_T are probabilities of identity between alleles sampled in a same subpopulation, and between alleles sampled between subpopulation, respectively, and correspond to the complement of the expected heterozygosities at the subpopulation and between population scales.

73. Consider a population of tetraploid individuals where the frequencies of alleles A, B, C and D at a given locus equal 0.2, 0.2, 0.3, and 0.3. Estimate the kinship, F_{ij}, between genotypes AABC vs. ACCD and between genotypes AABC and AABB. Hint: to estimate Q_{ij} , take in turn each allele within the first individual i *and check the proportion of alleles in* j *that are identical in state and then sum over all compared alleles (e.g. in AABC vs. ACCD, the first allele, A, is in ¼ frequency in the second individual)*

74. What is the mean F_{ij} over all pairs of individuals from a population?

F_{ij} is thus a RELATIVE measure of relatedness, relative to a sample of individuals that are considered as "unrelated" on average, or relative to given allele frequencies. There is in fact no "absolute" measure of kinship because we can argue that all organisms on earth are related to some extent.

F_{ij} is expected to have particular values for particular pairs of relatives. For example, in diploids, in the absence of selfing or other forms of inbreeding, we expect F_{ij} = ¼ between a parent and its offspring or between full-sibs, F_{ij} = 1/8 between half-sibs, F_{ij} = 1/16 between first cousins,… In tetraploids, all these values must be divided by two! That's where the R_{ij} coefficient makes sense: whatever the ploidy level, in an outbred population, R_{ij} = ½ between a parent and its offspring or between full-sibs, R_{ij} = ¼ between half-sibs, R_{ij} = 1/8 between first cousins, etc… R_{ij} is also the relevant parameter in the theory of kin selection and inclusive fitness. Note that R_{ij} is to F_{ij} what *Rho* is to F_{ST}.

75. What are the expected values of F_{ij} and R_{ij} for identical twins in humans? Or for two ramets from an octoploid plant with clonal propagation?

Estimates of F_{ij} based on genetic markers are notoriously imprecise, unless you have lots of markers (e.g. thousands of SNPs). To compensate for this measurement error, F_{ij} can be averaged over pairs of individuals that could share similar values, like individuals separated by similar spatial distances. Indeed, under limited gene dispersal, you expect that the relatedness between individuals will decay with the distance separating them, what is called "isolation by distance". Averaging F_{ij} in a spatial context is a prime application of SPAGeDi (the second line of each data set defines the distances intervals over which pairwise F_{ij} between individuals, or F_{ST} between populations, will be averaged).

To characterise the spatial genetic structure of the *Afzelia* population, launch again in SPAGeDi the data file "inSpagediAfzelia155ind.txt". This time, in the panel "STATISTICS for individual level analyses", select option 1 "KINSHIP coefficient (Loiselle et al., 1995)". You can

again select options for jack-knife and permutation tests. Look at the output file and make a graph representing how the kinship coefficient varies according to the distance separating the individuals.

76. Is there evidence of isolation by distance in Afzelia?

77. SPAGeDi also allows to compute F_{ij} for pairs of individuals belonging to the same category or to distinct categories. Try to use the file you created with the two Afzelia clusters identified as distinct categories to compare kinship within and between clusters.

Day 3:
Approximate Bayesian Computation

Part 1: Approximate Bayesian Computation (ABC) for analysing polyploid genomic data

Frequentist- and Bayesian inference represent two very different statistical frameworks. They are two different schools of thoughts, using different logics of probability:

1) frequentist thinking, which seeks the probability of events according to a certain theory.

2) the Bayesian thought, which seeks the probability of alternative theories in view of certain events.

Let's illustrate these two frameworks with a roll of the die. Imagine that I am hiding behind a screen, holding a 4-sided, 6-sided, 8-sided, 12-sided, and a 20-sided die in my hand. Then, I randomly pick a die, roll it, and get 7. This event obviously has a certain probability.

You probably would not do it like this, but one way to assess this probability would be to not consider how many sides each die has, and thus assess the probability of the event to be 1 out of 5 (20%)so there is a 20% probability for each die. But of course it is impossible to roll a 7 with a 4-sided die or a 6-sided die. In fact , only 3 dice can give 7, so the probability 1 out of 3 is more realistic. This would be the purely frequentist inference. We are nowadays used to think beyond this, and can come up with more realistic probabilities by using the prior information that we have about the dice:

4-sided die: $P(die\_4 \mid obs=7)$

6-sided die: $P(die\_6 \mid obs=7)$

8-sided die: $P(die\_8 \mid obs=7)$

12-sided die: $P(die\_12 \mid obs=7)$

20-sided die: $P(die\_20 \mid obs=7)$

78. Can you calculate these probabilities?

These probabilities multiplied by the prior probability will allow us to calculate posterior probabilities for each die using the Bayes-formula

$P(\text{ die,i} \mid 7) = [\ P(\ 7 \mid \text{die,i}\ ) \times P(\ \text{die,i}\ )\ ] / P(7)$

In this formula, $P(\ 7 \mid \text{die,i}\ )$ is the probability fo each die i of giving the observation 7. $P(\ \text{die,i}\ )$ is the prior probability of each die (so based on previous information). $P(7)$ is the probability to observe 7 (the sum of the probabilities of all the events that could lead to 7).

Of course, we would obtain more information about which die I am rolling with, if I would roll it again and tell you the number I got. Let's say I rolled 8. If I now repeat all the previous steps, but with the updated prior (so the posterior probability we calculated after my first roll of the die), I will get a new (and more realistic) posterior probability.

Now play with the script located at:
https://github.com/popgenomics/ABC_WGD/blob/master/comparison_ABC_bayes.R

79. Try to specifically see how the sample size (the number of times I roll the die) affects the posterior probability. What do you see, how many rolls are sufficient to allow you to decide with a comfortable certitude which die I have been rolling.

In evolutionary biology, we often also want to distinguish between alternative hypotheses: are my populations genetically connected by migration? Did the population recently go through a bottleneck? Are my individuals diploids or tetaploids? Is my species auto- or allopolyploid? As in the example with the dice above, to make a reasonable comparison, we have to calculate the posterior probability of observing the data under each of the hypotheses. However, for many complex evolutionary hypotheses, it is not possible to calculate these posterior probabilities in a quantitative way.

In the die-example, the posterior probabilities of the different dice could be calculated easily because we knew the exact rules leading to the probabilities of the observation for each die. In population genetics, it is often impossible to calculate the exact probability of an observation under a given model *P*(observation | model), where observation represents a series of statistics such as the Fst, theta, π, Tajima's D, etc.., and the model represents an evolutionary scenario that you want to test as auto versus allopolyploid), meaning that we cannot calculate the exact probability of a model from a given observation (i.e., *P*(model | observation)) which is what we as biologists are trying to infer.

During this practical, we will calculate these probabilities by approximating *P*(observation | model) using a method called ABC (Approximate Bayesian Computation). ABC forms a very flexible inferential statistical framework that can be used to a wide variety of problems, not limited to population genetics. Their main objective is to distinguish between a set of proposed models based on their ability to produce simulated data that resembles the observed data. The first step in ABC is to describe the observed data using a set of summary statistics (e.g. H_S, F_{ST}, theta, etc). The second step of ABC consists of formulating models to simulate data and choosing a relevant set of parameters. For each of the models to be compared, the following steps are performed:

1) Choose a set of parameters for the model; the parameters are randomly drawn following "*a priori*" rules defined by the experimenter. For example, for a model (or scenario) describing the subdivision of an ancestral population into two daughter populations that are allopatric, we would have 4 parameters: the time of split (in number of generations), the size of the ancestral population (in number of individuals), and the sizes of the 2 daughter populations.
2) Simulate a population genetic dataset under the selected model parameters. For instance, simulations for the combination of random parameters {T_{split} = 142,020; $N_{ancestral}$ = 1,678,123; N_{pop}_A = 652; N_{pop}_B = 896,182}, then simulation of another random combination of parameters {T_{split} = 7,981,727; $N_{ancestral}$ = 378; N_{pop}_A = 917,812; N_{pop}_B = 10,120}, and this, about 10,000 times.

3) Calculate the summary statistics for the simulated data (the same statistics as those used to describe the observed data). Thus, for each model, we have 10,000 simulated values of Fst, theta, Tajima's D, etc....

4) By iterating these steps many times (note that in the practical we keep the number of steps low at 10,000 for time reasons), it is possible to approximate the expected joint distribution of the summary statistics under each of the proposed models. The last step consists of statistically comparing the observed statistics with the simulated statistics in order to identify the model that best matches the observations (detailed below).

To practice with the above procedures of ABC, we will not walk you through step-by-step, but rather we will perform the different steps of ABC together. This way, you will get hands-on experience. Emphasis will be placed on the flexibility of the ABC to evaluate a wide range of hypotheses, rather than being trained in the tools used during the course. After this joint part, the idea is that everybody will try to apply ABC methods to analyse real genomic data from a tetraploid species/populations. More specifically, the goal will be to make inferences about:

i) the mode of gamete transmission in the tetraploid (disomic *vs.* tetrasomic);

ii) the putative origin of the tetraploid (allo- vs. autopolyploid).

iii) your own scenarios.

We will mainly deal with the point (4) from the previous section. How to decide which of the models best matches the observed data? There is a wide variety of ABC algorithms in the literature that compare an observed dataset with multiple simulated data for different models. Their main goal is always the same: to calculate the relative posterior probabilities for each model compared. Here, we will practice doing ABC using a novel approach based on a machine learning tool named Random Forests to conduct selection among the highly complex demographic models. In brief, Random Forests are forests of decision trees. Individually, a tree is a set of binary allocation rules ("at this node go right / go left"). The successive binary criteria ($X_{obs,j} > X_j$?) at each node of the tree will discriminate between the compared models.

As an example of a decision tree, one can imagine a doctor, called "Dr. A". If you see Dr. A, he will always ask you the first question: "Do you have a temperature higher or lower than 38°C"?

I) If it is "yes", he will ask you if the tongue is white.

I-1) If the tongue is white, he will take your blood pressure and then come up with a diagnosis.

I-2) If the tongue is not white, he will look at the ears and then come up with a diagnosis.

II) If the temperature is not higher than 38°C, he will ask you if you have a headache.

II-1) If you have a headache, he will look at the whites of your eyes and then come up with a diagnosis.

II-2) If you don't have a headache, he'll touch the liver and then come up with a diagnosis.

In other words, Dr. A followed a decision tree with a series of 3 dichotomous steps, more or less relevant in the ability to make a good diagnosis. The outcome of the last steps (for node I-1, I-2, II-1 and II-2) is called a leaf of the decision tree. The currently applied tree was actually trained empirically on the basis of past observations (it was based on what the Dr. learned in his particular medical school, which in turn represents the medical knowledge/views of the teachers)..

However, a doctor B, from another academy, from another country, will apply another decision tree (How are the intestines? Are we constipated? Do we have pimples? Etc...). The same patient may get a different diagnosis from Dr. B than from Dr. A.

You could imagine a Random Forest as a set of thousands of doctors, each predicting a diagnosis based on their own decision trees. One would then apply a majority rule over all trees in the forest to establish a better diagnosis than any of the individual diagnoses.

Together, we will build such a forest to predict (infer) a past evolutionary scenario given a set of currently observed data (i.e. genetic data). The advantage we have over Dr. A and Dr. B is that we can simulate our evolutionary scenarios, while they cannot simulate symptoms to train their decision trees. We can do this by producing trees from simulated data or by taking sub-samples from previous datasets by bootstrapping (i.e., drawing by replacement).

Now, we will use the R package ‘abcrf’ to:

1) train a Random Forest using simulated datasets.
2) make a prediction (inference) about the more likely scenario best explaining the observed dataset. Specifically, we will try to:
    - Discriminate between disomic inheritance and tetrasomic inheritance
    - Discriminate an autopolyploid origin from an allopolyploid origin

Day 4:
Selective Sweeps and Linked Selection

Overview

The objective of today's workshop is to (re)acquaint ourselves with the basic principles of allele frequency change in response to selection with a focus on how/where this process depends on ploidy. We will start slow, introducing each idea in the context of diploidy, and then elaborate the ideas for higher ploidy. The latter half of the day will be dedicated to the topic of linked selection, which concerns the effect of selection on linked, neutral diversity. As this is currently a major focus in population genomics, we will spend ample time reviewing the historic and current observations surrounding this phenomenon in diploids before discussing shifting our focus to the expectations and analysis of polyploid data. We will focus primarily on autopolyploids, but will take time to consider allopolyploids at each step or where students are interested.

Today you will work with help of a set of custom R functions that are available at GitHub. To download, type the following command or download ZIP archive from the GitHub webpage

```
git clone https://github.com/pmonnahan/PolyploidyWorkshop.git
```

Part 1: The Response to Selection

Allele frequency change: Diploids

As you may recall from an introductory course in evolutionary biology, there are four major evolutionary forces that can change allele frequency in a population: migration, genetic drift, mutation, and **selection**. Today, we will focus on the last of these forces. Typically, the basic principles governing how selection changes allele frequency is introduced for the simplest case (i.e. bi-allelic locus, infinite population size, random mating, etc.) in the following table structure (taken from Hartl & Clark *Principles of Population Genetics [1]*):

	Genotype			Total
Generation $t-1$	AA	Aa	aa	
Frequency before selection	p^2	$2pq$	q^2	$1 = p^2 + 2pq + q^2$
Relative fitness (viability)	w_{11}	w_{12}	w_{22}	
After selection	p^2w_{11}	$2pqw_{12}$	q^2w_{22}	$\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$
Normalized	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$	
Generation t	$p' = \frac{p^2w_{11} + pqw_{12}}{\bar{w}}$			
	$q' = \frac{pqw_{12} + q^2w_{22}}{\bar{w}}$			

80. If $p = 0.5$, $w_{11} = 1$, $w_{12} = 0.95$, and $w_{22} = 0.9$, what will p be in the next generation?

81. What if $w_{11} = 1$, $w_{12} = 0.9$ and $w_{22} = 0.8$?

82. Recalculate above two problems, but with $p = 0.05$. Consider the change in allele frequency (Δp) when $p = 0.05$ versus when $p = 0.5$. At which frequency is selection more effective at changing allele frequency? Is Δp a linear function of p?

The *w* values above are in terms of *relative* fitness, where the genotype with the highest probability of surviving is given the value of 1, and all other fitness values are scaled relative to this genotype. For example, if the *absolute* fitness values of AA, Aa, and aa are 0.75, 0.75, and 0.5 (i.e. an AA individual has a 0.75 probability of surviving), then the relative fitness values are 0.75/0.75 = 1, 0.75/0.75 = 1, and 0.5/0.75 = 0.67.

83. If instead the absolute fitness values exemplified above were 0.0075, 0.0075, and 0.005 (i.e. each genotype is much less likely to survive), what are the relative fitness values of each genotype? How does this affect the rate of allele frequency change?

It is often convenient to re-parameterize the *relative* fitness values in terms of *s* (the selection coefficient; ranging from 0 to 1) and *h* (the dominance coefficient; also ranging from 0 to 1). If 'A' is the beneficial allele, then the three genotypes, AA, Aa, and aa, will have fitness 1 + s, 1 + *hs*, and 1. However, we need to divide each of these by (1 + s) in order for the most-fit genotype to equal 1.

84. What is s and h in Q 81?

85. What would the heterozygote fitness be if the mutation was additive (h = 0.5)?

86. *If p = 0.5, what value of h will result in the highest allele frequency change in the following generation? What about if p = 0.05? Hint: use dipTraj() function in the set of R commands you downloaded from GitHub (Day7_PolyploidyWorkshop2018.R) and look for dp1 in the output. Try h = 0, and then repeat with h = 1 and h = 0.5. You can use an arbitrary value for s.*

87. *If s is equivalent for a beneficial allele in both diploids and tetraploids (i.e. that is $w_{22} - w_{11} = w_{2222} - w_{1111}$), and the allele is completely dominant, do you expect the allele frequency change to be greater, less, or the same in tetraploids? Why?*

Allele frequency change: Tetraploids

Here, we will be assuming autotetraploidy and tetrasomic inheritance. The process of selection that we have just outlined for diploids is slightly more complicated for tetraploids. This is primarily due to the increased number of genotypes that we have to consider. Using the diploid example (copied below), lets determine the relevant formulae for tetraploids.

	Genotype			Total
Generation $t-1$	AA	Aa	aa	
Frequency before selection	p^2	$2pq$	q^2	$1 = p^2 + 2pq + q^2$
Relative fitness (viability)	w_{11}	w_{12}	w_{22}	
After selection	p^2w_{11}	$2pqw_{12}$	q^2w_{22}	$\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$
Normalized	$\frac{p^2w_{11}}{\bar{w}}$	$\frac{2pqw_{12}}{\bar{w}}$	$\frac{q^2w_{22}}{\bar{w}}$	
Generation t	$p' = \frac{p^2w_{11} + pqw_{12}}{\bar{w}}$			
	$q' = \frac{pqw_{12} + q^2w_{22}}{\bar{w}}$			

88. Fill in the table below with the corresponding formulae for tetraploids

	Genotype				
Gen. t-1					
Freq. before selection					
Relative fitness					
After selection					
Average fitness ($\bar{w}$)					
Normalized					

89. What is now the equation for expressing p′ in terms of p, q, and the relative fitnesses?

As before, we can write the relative fitnesses in terms of *s*, but we now require 3 dominance coefficients: h_1, h_2, and h_3, which will modify the effect of the selection coefficient when the mutant allele is present in one, two, or three copies, respectively.

90. Consider an "additive" locus with s = 0.1. For diploids, h = 0.5. For tetraploids, $h_1 = 0.25$, $h_2 = 0.5$, and $h_3 = 0.75$. Compare mean fitness in diploids vs tetraploids when p = 0.5. Hint: you can use getFits() function in R.

91. Plot mean fitness against allele frequency for a dominant, recessive, and additive locus for both diploids and tetraploids. Under what scenarios do diploids display higher mean fitness for a given allele frequency? Hint: look for where 'mean_fit' is used for plotting in the R script.

92. *Now, let us consider allele frequency change instead of mean fitness. If p = 0.5 and s = 0.1, find the allele frequency in the following generation for diploids and tetraploids. Do this for each scenario of dominance: additivity (h = 0.5 in diploids and $h_1 = 0.25$, $h_2 = 0.5$, and $h_3 = 0.75$ in tetraploids), dominant (h = 1 for all) and recessivity (h = 0 for all). In which case does selection happen more quickly in tetraploids? Hint: use dipTraj() and tetTraj() to avoid hand-calculation.*

The formula for p′ that we have derived above for both diploids and tetraploids is known as a recursive formula; it gives the value in the next generation as a function of the current value. Given an initial starting value, we can use this formula to calculate the frequency at each subsequent generation by recursively substituting p′ for p.

93. *For both diploids and tetraploids, plot the frequency at each generation for an additive beneficial (s = 0.1) allele beginning at a frequency of 0.05. Hint: Search for "Q89" in R script. This section uses dipTraj() and tetTraj() to generate data and ggplot() to visualize. Does the beneficial allele reach fixation more quickly in diploids or tetraploids? Is this always the case for different scenarios of dominance (h) and s? Hint: For the last part question, modify the h values in the code for Q89.*

Part 2: Fixation probability

Based on our above *deterministic* model of selection, the fixation of a beneficial allele, would seem to be virtually guaranteed (i.e. Δp will always be positive for any value of p), ***regardless*** of how strongly the beneficial allele is favoured (i.e. the selection coefficient, s). For weakly selected alleles, however, as p approaches 0, so too does Δp, such that even in very large populations, there comes a point at which a beneficial allele can be stochastically lost due to **genetic drift**.

In 1927, Haldane [2] famously showed that, for a weakly beneficial allele beginning with a single copy in a large haploid population, the probability that the allele is ultimately fixed in the population is approximately $2s$. In other words, the majority of beneficial alleles will be randomly lost due to drift unless selection is strong. For an arbitrary ploidy number, this probability becomes $2h_c s$, where h_c is the dominance of the allele **when in single copy** [3]. In our discussions above, this (h_c) was h for diploids and h_1 for tetraploids.

94. *Some refresher questions. What is the fixation probability of a neutral mutation that arises as a single copy in a population? Does genetic drift operate more influentially in smaller or larger populations?*

95. *For a beneficial mutation, if the selection coefficient of an allele is equal for diploids and tetraploids and h = 0.5 in diploids and h_1 = 0.25 in tetraploids (i.e. additive), is the fixation probability higher in a diploid or tetraploid population?*

Haldane's rule of $2h_c s$ is a good approximation to the fixation probability for large populations, but it breaks down when the population size (N) gets small. Kimura [4, 5] showed that the more general approximation for the fixation probability of a beneficial, **additive (h = 0.5)** allele is:

$$\Psi \approx \frac{1 - \exp[-2cN_e sp]}{1 - \exp[-2cN_e s]} \quad \text{(eqn. 1; [6])}$$

where N_e is the effective population size, and c is the ploidy, where c = 1 for haploids, c = 2 for diploids, c = 4 for tetraploids, etc.

96. *Based on the above equation, is the fixation of a beneficial allele more likely with an effective population size of 100, 1000, or 10,000? Here, assume diploidy and that the allele is present in a single copy in the population (i.e. p = 1 / 2Ne). Also, assume s = 0.01. Hint: look at the R function getFix().*

97. *For an equivalent population size and selection coefficient, is the fixation probability greater or lesser for tetraploids? Use s = 0.01, and Ne = 100. Assume the beneficial allele arises as a single copy. What about when Ne = 1000?*

VERY IMPORTANT: Equation 1 is taken from Otto and Whitlock [6], which uses a different parameterization of *s*. From the source, "If the average fitness of individuals who carry one copy of A is **(1 + s)** times greater than the average fitness of individuals who carry no copies of A, then the fitness effect of allele A is measured by s, the selection coefficient." This differs from our previous treatment of selection in which the fitness of Aa was **(1 + *hs*)** and the fitness of Aaaa was **(1 + $h_1 s$).** Such re-parameterizations are admittedly confusing, yet are unfortunately common in population genetics. This is an important lesson that is particularly apt for considering how ploidy effects selective processes. Whether or not we believe that ploidy "matters" can depend entirely on how we parameterize *s* and/or *h.*

98. Repeat Q93, but replace s with hs (for diploids) or $h_1 s$ (for tetraploids). Here, use h = 0.5 and $h_1 = 0.25$.

Otto and Whitlock [6] also provide a convenient formula for the expected or average time to fixation for finite populations, **conditional on the allele eventually fixing** (they cite Ewen's 1979 textbook, *Mathematical Population Genetics*). Recall that this is only true of an **additive** allele.

$$\bar{t} \approx \frac{2 \ln(cN_e - 1)}{s} \quad \text{(eqn 2)}$$

99. What is the expected time to fixation for diploids and tetraploids if Ne = 1000 and s = 0.1? Does this match what you expect based on your answer to Q89? Why or why not? Hint: consider using hs or $h_1 s$ as in Q94.

Part 3: Rate of adaptation versus rate of fixation

We demonstrated in the first section that the rate of allele frequency change tends to be higher in diploids (also noted by [3, 7, 8]). This is distinct from the rate of *adaptation*, which typically refers to the change in mean fitness of the population over time. In specific cases, the former will result in the latter. For example, if there is a single *additive* ($h = 0.5$) beneficial allele starting at the same frequency in a diploid and tetraploid population, then we would expect a faster increase in mean fitness in the diploid population purely by virtue of the faster allele frequency change.

100. *For an additive, beneficial allele ($h = 0.5$; $h_1 = 0.25$, $h_2 = 0.5$, and $h_3 = 0.75$), plot mean fitness versus time (generations) for diploids and tetraploids when $s = 0.1$. Hint: The R functions that we used to calculate the allele frequency trajectories in Q89 will also report the mean fitness of the population over time. Search for Q96 in the R script for help with plotting.*

It is arguably more likely, however, that tetraploid populations will more quickly increase in mean fitness, despite a slower time to fixation of individual beneficial alleles. We will begin with a simple case, as before, and then move on to more nuanced explanations.

101. *Plot the allele frequency trajectory for a completely dominant allele (all h values are 1) starting at a frequency of 0.05. Then, plot the mean fitness. Does fitness increase more quickly in diploids or tetraploids? Hint: search for Q97*

We have just seen that tetraploids do initially adapt more quickly for selection on a single dominant locus, although diploids will quickly catch up as the allele approaches fixation. There are at least two more general reasons that tetraploids may ultimately adapt more quickly than diploids if all else is equal.

(1) In all of the previous examples looking at allele frequency change, we have stipulated that the allele is at the same starting frequency in both the diploid and the tetraploid population. If beneficial alleles are starting at a higher frequency on average in tetraploid populations, then these alleles may fix more quickly in tetraploid populations resulting in an overall faster rate of adaptation. One scenario in which beneficial alleles may be expected to start at higher frequency in tetraploids would be if a population is adapting to a novel environment. Alleles that were deleterious in the prior environment may now be beneficial in the new environment. For example, consider an allele that confers resistance to a pest, but requires a lot of energy to produce the functional protein. This allele would be deleterious in a pest-free environment, but highly beneficial in an environment where the pest is present.

The expected frequency of this allele when deleterious can differ dramatically between diploids and tetraploids, depending on the dominance of the allele. The equilibrium allele frequency is a balance between input from recurrent mutation and the removal from purifying selection. For a completely recessive allele, the equilibrium allele frequency will be:

$$\widehat{q_D} = \sqrt[2]{\frac{\mu}{s}} \qquad \widehat{q_T} = \sqrt[4]{\frac{\mu}{s}} \qquad \text{(eqns 3a-b)}$$

where 'D' is for diploids and 'T' is for tetraploids.

102. If mutation rate (μ) = $1x10^{-8}$ and s = 0.01, which has a higher equilibrium allele frequency: diploids or tetraploids?

This difference is less dramatic for partially recessive mutations. In that case, the equilibrium allele frequencies at mutation-selection balance become:

$$\hat{q} = \frac{\mu}{sh_c}$$

where, again, h_c is the dominance of the allele in single copy (i.e., Aa in diploids and Aaaa in tetraploids).

103. Repeat 98. for a partially recessive mutation where h = 0.5 in diploids and h_1 = 0.25 in tetraploids.

As long as $h_1 < h$ (i.e. single-copy dominance of Aaaa in tetraploids < dominance in Aa in diploids), the equilibrium allele frequency will be higher in tetraploids. If there is an environmental change such that these previously deleterious alleles are now beneficial, the tetraploids will have a "head-start" over diploids since they will be at a higher starting frequency. Whether or not this results in faster adaptation following introduction to a new environment will depend on the dominance of these mutations (particularly the dominance while in single-copy) and how dominance changes across ploidy.

104. Using your results from 98 and the dipTraj() and tetTraj() functions, determine if fixation will occur more quickly in diploids or tetraploids? Let h_2 = 0.5 and h_3 = 0.75 and keep all other variables as they were in 98. Repeat for 99. Is your prior answer still true?

(2) The other reason that tetraploids may ultimately adapt faster than diploids results from the fact that beneficial mutations are expected to be introduced at twice the rate into tetraploids when compared to diploids. This description is best laid out by Otto and Whitton (2000). Using their notation, the introduction of beneficial mutations into a population will occur at rate *cNv*, where:

- *c* is the ploidy level
- *v* is the per-base mutation rate of beneficial alleles
- and *N* is the effective population size.

If we recall that $2h_cs$ is the probability that a new beneficial mutation survives stochastic loss, and s is the percent increase in fitness upon fixation of the beneficial allele, then the rate of fitness increase (i.e. adaptation) in the population is $cNv(2h_cs)s$. As long as, $h_1 > {}^{h}/_{2}$, then the rate of adaptation will be greater in tetraploids than in diploids.

105. *For what types of mutations would we expect $h_1 > {}^{h}/_{2}$? Do we think this will be true for the majority of mutations?*

106. *What about the selection coefficient, s? Is it reasonable to expect that $w_{22} - w_{11} = w_{2222} - w_{1111}$? Or, would you expect $w_{22} - w_{11} < w_{2222} - w_{1111}$? Why or why not?*

In summary, although the rate of fixation of particular beneficial alleles is generally expected to be faster in diploids, it is reasonable to expect that the overall rate of adaptation may be greater in tetraploids. In the previous section, we demonstrated how this might be true both for when adaptation occurs from *standing genetic variation* or if it relies on the input of new mutation. This distinction between the rate of fixation and the rate of adaptation is important to keep in mind whenever we consider the population genomic signal that selection will leave behind in either diploids or tetraploids.

Part 4: Linked Selection

Linked selection generally refers to the effect of selection at a causal site on the diversity of neutral mutations surrounding that site. There are two major forms: background selection and genetic hitchhiking, which correspond to the effects of deleterious and beneficial mutations, respectively. Initially, we will consider genetic hitchhiking and then briefly go over background selection if time allows.

The introduction of genetic hitchhiking is largely credited to Maynard-Smith and Haigh (1974) [9]. In this classic paper, they derive the expected reduction in frequency at a linked, neutral locus following the fixation of a beneficial allele. Unfortunately, they only derive these values for haploids and diploids. Beginning with haploids, they organize their derivation starting with a table like the one we used initially for understanding allele frequency change in response to selection. However, here we are considering two biallelic loci: *A/a* and *B/b*.

Genotype ...	AB	aB	Ab	ab
Frequency	p_nQ_n	$p_n(1-Q_n)$	$(1-p_n)R_n$	$(1-p_n)\,(1-R_n)$
Fitness	$1+s$	$1+s$	1	1

- *B* is the beneficial allele (i.e., A/a is the neutral locus),
    - and p_n is its frequency at generation *n*.
- Q_n is the proportion of *B* haplotypes that also contain *A* at generation *n*,
    - while R_n is the proportion of *b* haplotypes containing *A*.

We will assume that the beneficial mutation initially occurs on an '*a*' background (i.e. *aB*), such that initial proportion $Q_0 = 0$. As the *B* allele increases in frequency, so too will the '*a*' allele. Recombination is required in order for *Q* to increase above 0. We are interested in observing what happens to Q_n and R_n as the beneficial allele goes to fixation (p_n goes to 1). To capture this behaviour into a single value, Maynard Smith and Haigh focused on the ratio of Q_∞ to R_0, which corresponds to the proportional reduction in frequency of *A* upon fixation of the *B* allele.

Again, similar to how we calculated allele frequency change in response to selection, they initially derived formula for what *p, Q,* and *R* will be in the following generation as a function of these values in the current generation. They then derive a formula for *Q* as time goes to infinity (i.e. Q_∞).

$$Q_\infty = cR_0(1-p_0)\sum_{n=0}^{\infty}(1-c)^n/\{1-p_0+p_0(1+s)^{n+1}\},$$

where ***c* is the recombination fraction**; the ratio of the number of recombined gametes to the total number of gametes produced [10]. NOTE: ***c*** was previously used to denote ploidy in the section on fixation probability. From now on, *c* will refer to the recombination fraction.

The math quickly gets worse from here. Consider expanding the Σ in the above equation as time (i.e *n*; number of generations) increases. Every generation adds a new term to the equation.

Not surprisingly, this equation does not simplify to a convenient form for arbitrary *n*. Instead, they derive an approximation for Q_∞ / R_0, using differential equations to produce the following equation.

$$\frac{dQ}{dp} = \frac{cR_0}{sp} e^{-ct}.$$

If c is so small that, over the major time that p increases from p_0 to 1, e^{-ct} remains effectively at 1, integration of (13) gives

$$\frac{Q_\infty}{R_0} \simeq \frac{c}{s} \log \frac{1}{p_0}.$$

Unfortunately, this derivation for diploids becomes substantially more complicated. A "simple" deterministic equation for Q_∞ , as we showed above for haploids, does not exist. However, we can still use differential equations to derive an approximation. Note the added complexity in the equation below relative to what we saw for haploids.

$$\frac{1}{R_0}\frac{dQ}{dp} \simeq \frac{c(1+hs)}{s} \frac{1+sp(2h+p(1-2h))}{p(h+p(1-2h))\{1+s(h+p(1-h))\}}.$$

By integrating the above equation, Maynard Smith and Haigh found the following equations for h = 0 and h = 1, respectively:

$$\frac{Q_\infty}{R_0} \sim \frac{c}{s}\frac{1}{p_0}. \qquad \frac{Q_\infty}{R_0} \sim \frac{2c}{s} \log \frac{1}{p_0},$$

For all other values of *h*, they find:

$$\frac{Q_\infty}{R_0} \simeq \frac{c}{hs} \log \frac{1}{p_0}.$$

107. Will selection for B have a larger impact on the frequency of A (i.e. will Q_∞ / R_0 decrease more) if the B is dominant, recessive, or additive (i.e. h = 1, h = 0, or h = 0.5)? Hint: When thinking about this problem, pick an arbitrary value for p_o , the exact value is not important for answering this question.

108. Do you expect the impact of linked selection (Q_∞ / R_0) to generally be smaller or larger for diploids relative to haploids? Why? Which values in the above equations will depend on ploidy?

Theoretically, it ought to be possible to extend the logic of Maynard Smith and Haigh for tetraploids. However, recall how the math became substantially more complicated when we went from haploids to diploids. Although approximations like the ones above have not yet been made for tetraploids, they will likely depend on the same crucial parameters: selection, recombination, dominance, and initial frequency.

109. Based on our reasoning that we developed in the prior question, do we expect the impact of linked selection (Q_∞ / R_0) to increase or decrease for tetraploids?

There is a growing body of evidence that suggests much of the genome may be impacted by linked selection [11], with some advocating that it should replace neutral theory as the null expectation in population genetic and genomic inference [12]. Currently, we are lacking both the theory and resources to fully explore these processes in tetraploids. For example, a classic hallmark of linked selection is a correlation between levels of polymorphism and recombination rate along a chromosome [13]. In order to estimate recombination rates using next-generation sequencing data, it is typically necessary to provide "phased" genotype calls, i.e. reconstruct haplotypes of particular alleles over the loci in a genome (two in diploid, four in tetraploid). Phasing algorithms capable of handling tetraploid genotypes do exist, but are challenging to implement in my experience. Even if we were able to accurately phase tetraploid genotypes, any comparisons relative to diploids would be suspect without a clear theoretical understanding of the relationship between recombination and selection in tetraploids.

Additionally, we need to consider the effect of polyploidy on levels of polymorphism (i.e. the mutational process) in a population. The derivation above showed how a neutral locus is affected by selection on a newly arisen beneficial allele at a separate, linked locus, but this was for a "simple" two-locus case. In tetraploids, both beneficial and neutral mutations are expected to be introduced at twice the rate in a tetraploid population (all else equal; population size and mutation rate).

110. How might this increased mutational input alter the signals of linked selection in tetraploids?

In the absence of further theory regarding linked selection in tetraploids, it is difficult to elaborate much further on the potential effects of polyploidy on genetic hitchhiking. Clearly, it is a complicated issue that is difficult to reduce to single metrics. In place of this, we can use simulations to help develop our intuition. The remaining functions in the R code supplied for today's session can be used to simulate, analyse, and visualize the effect of polyploidy on linked selection. These functions are under the subheading labelled "Part 4: Linked Selection" in the R script:

- *PloidyForSim()* performs stochastic simulations of selection on a beneficial allele for a given population size. It returns a vector of the allele frequency at each generation as the beneficial allele proceeds to fixation.
- *msselRun()* will use this trajectory along with the coalescent simulator, ***mssel*** [14]**,** to simulate patterns of diversity along a sequence of a given length following a selective sweep (i.e. hitchhiking).
- *msselCalc()* will analyse the output of *mssel* using the R package "PopGenome" and produce a data table containing a number of population genetic metrics calculated in windows along the simulated sequence. I have included a number of plotting functions to visualize these results.

Let's spend a good deal (perhaps the rest) of time exploring the effects of the various factors that we have discussed on signals of linked selection. I have included additional detail on running these scripts and the relevant parameters that may be interesting to adjust in the R script.

References

1. Hartl DL, Clark AG, Clark AG. Principles of population genetics: Sinauer associates Sunderland; 1997.
2. Haldane JBS, editor A mathematical theory of natural and artificial selection. Mathematical Proceedings of the Cambridge Philosophical Society; 1927: Cambridge University Press.
3. Otto SP, Whitton J. Polyploid incidence and evolution. Annual review of genetics. 2000;34(1):401-37.
4. Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(6):713.
5. Kimura M. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics. 1957:882-901.
6. Otto SP, Whitlock MC. Fixation probabilities and times. e LS. 2001.
7. Bever JD, Felber F. The theoretical population genetics of autopolyploidy. Oxford surveys in evolutionary biology. 1992;8:185-.
8. Hill R. Selection in autotetraploids. TAG Theoretical and Applied Genetics. 1971;41(4):181-6.
9. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genetical Research. 1974;23(1):23-35. Epub 2009/04/01. doi: 10.1017/S0016672300014634.
10. Xu S. Principles of statistical genomics: Springer; 2013.
11. Slotte T. The impact of linked selection on plant genomic variation. Briefings in functional genomics. 2014;13(4):268-75.
12. Hahn MW. Toward a selection theory of molecular evolution. Evolution. 2008;62(2):255-65.
13. Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356:519. doi: 10.1038/356519a0.
14. Berg JJ, Coop G. A coalescent model for a sweep of a unique standing variant. Genetics. 2015:genetics. 115.178962.
15. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289.
16. Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 2012;190(1):5.
17. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics. 1995;141(4):1619.
18. Ronfort J. The mutation load under tetrasomic inheritance and its consequences for the evolution of the selfing rate in autotetraploid species. Genetics Research. 1999;74(1):31-42.