Population genetics of polyploids

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

Table of contents

Overview

In this workshop we will have a detailed look at the population genetics of polyploids. The main aim is to make you aware of the main theoretical and practical issues concerning polyploid data. We will start out with going over some basic concepts in population genetics: genetic diversity and population structure. Since your knowledge on these topics may be a bit rusty, we always first start out with explaining how these work in diploids and then develop the theory further for polyploids. Our main focus will be on autopolyploids (assuming polysomic inheritance), but later we will also explore some issues around allopolyploids.

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 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 and or species.

10. *Take a locus with four alleles with frequencies 0.55, 0.2, 0.15 and 0.1. 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 loses can no longer be called the "expected heterozygosity", but anyway it is common to still indicate this with H_S. So for every ploidy level the same equation as above is used:

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

Note that, 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?*

However, if we want to compare H_S (expected H) with H_O (observed H), for example for calculating F_{IS}, we also have to use a similar trick for calculating 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*. In that case, there is a probability of 2/3 that the second allele is a *B* (and the gamete is a heterozygote). The alternative way that a heterozygote gamete is formed is that the first allele is a *B* and the second an *A* and this 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 cytotypes that do not actually produce diploid gametes.

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. 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 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 is 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 "Part 2 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 its value depends also on the ploidy level. This makes comparisons among ploidy levels more difficult. 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. Here we will only look at the *rho*-statistic (Ronfort et al 1998), which has been developed especially for polyploids.

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 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 “Part 2 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?

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, even for highly variable markers such as microsatellites, 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 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 here 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 data is especially troublesome.

Simulated data

Assume a single population where diploids and tetraploids co-occur and freely interbreed (somewhat unrealistic, we know). Therefore, 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 alelle 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 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?*

Real data

Now let's look at real data - a dataset of 10,000 biallelic SNPs obtained for one diploid and one tetraploid population of *Arabidopsis arenosa*. Open the script called "RAD analysis arenosa.R". Set the working directory to the location of the data file "Arenosa_data.vcf.gz" and the file with the additional functions "custom functions". Then execute the first 30 lines of code. This will read the data and select two populations (one diploid and one tetraploid) from the dataset. It also calculates a PCA of the real dataset

43. *Is there a clear separation of the diploids and tetraploids?*

Execute the code from L63-95. This generate fully resolved genotypes (i.e., with known dosage) of autotetraploid individuals based on the allele frequencies of the diploids. It also calculates a PCA of those generated tetraploid genotypes and their source diploids.

44. *Is there a clear separation of the diploids and tetraploids?*

Execute the code from L100-123. This takes the above-generated genotypes and strips them of their dosage information. It then again calculates a PCA.

45. *Is there a clear separation of the diploids and tetraploids?*

46. *So overall, do the diploid populations differ from natural tetraploids? Is this differentiation a bias coming from comparison of populations with different ploidy?*

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 partly 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 the R command line and venture into the point-and-click user interface of Structure.

To work with Structure, we are going to have to get some data. For this we will, of course, use a script with simulated 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. However, after doing those generations there is quite a bit of code to create individuals with either dominant data, fully codominant data or codominant data with missing dosage information. Finally, these individual genotypes are then 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 get differentiated 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.

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

48. *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 also select the file with codominant data (start out easy).

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 this set a sensible name. Now click Run and set the number of assumed populations (k) to 2.

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

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

51. *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".*

Part 6: Segregation analysis

In this part, you will not do any assignments out of this manual, but Marc will give a Powerpoint presentation on segregation analysis of polyploids. In addition, he will guide you through the analysis using an Excel worksheet.

Part 7: Tripleurospermum

In the final part of this workshop you will look at some actual data from *Tripleurospermum inodorum*. This dataset was collected by Martin who will briefly give an overview of this species and his dataset. You will be divided into three groups who will analyse:

- The genetic diversity per population and per locus of diploid and tetraploids.
- The genetic differentiation among diploid populations, among tetraploid populations, and among the two cytotypes.
- A Structure analysis of the diploid populations, the tetraploid populations and the combined dataset.