Chapter 7 – Population genetics.

Block 1:

PZA00219_6: p=0.4532 and q=0.5468

Ho=0.2959

He=0.4956

F=1-(Ho/He)=0.4030

A departure from expected heterozygozity is observed, suggesting that there is an underlying structure in the population. Also note the relative high fixation index values.

Chisquare test statistic is 86.7 and 28.9 for marker 1 which suggests a strong deviation from expectation, and that the population is not in HW (at least from the information on this marker).

Block 2:

PZA00219_6	Guerrero	Jalisco	Mexico	Michoacan	Entire Pop
Ho	0.2711	0.0538	0.3718	0.4565	0.2959
He	0.3932	0.267	0.4861	0.4998	0.4956
F	0.3105	0.7519	0.2352	0.0865	0.403
Chisquare	21.6864	52.5826	4.3136	1.0331	86.7289

When considering the four subpopulations we can observe:

Lower F values (the difference between Ho and He decreases).

There is less evidence of departure from HW in e.g. Mexico, Michoacan

It seems that at least a good part of the deviation from HW observed when considering the population as a whole is partly dissipated when accounting for the sub-populations as defined by the geographical origin – an example of the Wahlund effect.

Neverthless, note that some of the sub-populations (most notably the Jalisco) still seem to have considerable structure.

Block 3:

559/592 results are significant overall. By chance we would expect 5% of 592= 29.6. The population as a whole is clearly out of HW equilibrium.

Block 4:

For t-tests, GUE 420/592, JAL 398/592, MEX 173/592, MIC 252/592 results are significant. By chance we would expect 5% of 592= 29.6. Individual populations are out of HW overall but substantially less so than the overall population.

The average fixation index across all markers and populations is 0.233 (se 0.006). If the subpopulations are to some extent responsible for that we should observe a decrease of F inside the subpopulations. A decrease of F is observed in 3 out of 4 populations, confirming what has been suggested by the analysis of two markers. The population "Jalisco" however, shows an increase of F, which might reflect the fact that there is considerable structure inside this group (we will come back to this later).

Block 5:

The F statistics confirm that the inbreeding within subpopulations is lower (0.205) than overall (0.246). The difference is accounted for by divergence between the subpopulations (Fst=0.052). However, F_{ST} is not especially large and F_{IS} is still quite large. So either there is more substructure within states, or there is true inbreeding within these subpopulations.

Block 6:

The eigenanalysis finds shows that the first 10 PCAs account for from 5.5% (PCA1) to 1.0% (PCA10) of the variation (total 20.2%). The graph is also smoothly decreasing. This confirms the suspicion that the population substructure is more complex than the four groups considered so far.

Block 7:

The scatter plot of the first two axes show indeed some correspondence with the classification according with the origin (blue, purple, and red dots are in general together in the plot).

However, some further sub-structure is evident, for example the green dots (group "Jalisco") seems to be subdivided into at least three subgroups. Also some genotypes from the blue set (Mexico) are quite different from the rest.

The observation for Jalisco is in agreement with the high F values we calculated earlier.

Block 8:

Dividing Jalisco, Mexico and Michoacan into two groups each had a very large effect on the F values and F statistics.

F_{IS} is now 0.185 instead of 0.205. Population differentiation as measured by F_{ST} is now 0.081 instead of 0.052. Much more of the overall inbreeding is now seen to be a result of population divergence.

The F values also decrease for all these subpopulations. For the smaller Mexico and Michoacan populations, F ~ 0 now.

However in Jalisco 2 especially F is still high – maybe there is even more substructure to be uncovered?

Block 9:

There appear to be two main groups of individuals with a few intermediates, although the bootstrap values near the centre of the network are mostly low. There is some more well-supported structure nearer to the tips of the tree but overall there is little higher-level structure evident beyond the two groups.

Block 10:

From the plot of Axis 1 vs Axis 2, there seems to be clearer structure with the PCoA than with the tree-based method on the barley data set. Axis 3 does not appear to add much meaningful information, when we compare it to axis 1 and axis 2

The plot is quite comparable to the tree except, with the weakly defined 3rd group towards the bottom being the intermediates in the middle of the tree

There are certainly 2 main groups. Probably we would count the individuals at the bottom centre as a third group, but it is not very clearly delineated.

Block 11:

Firstly, a clarification. Structure uses the genetic information in the data to define "natural populations" as groups of individuals which are in both Hardy-Weinberg equilibrium and linkage disequilibrium. A "population" in Structure is defined by a set of allele frequencies at all loci. Individuals are then assigned to populations, Because we allowed admixture (gene flow) between populations, an individual can be assigned to more than one population (e.g. 60% to Population 1, 40% to Population 2). We use the admixture model because in most realistic natural and artificial settings in which Structure is used, there will have been admixture.

Key to Structure Populations by state

K=3 Red – mostly found in GUE, few in MEX,MIC

Green – mostly found in MEX-MIC, some in JAL and GUE

Blue – only found in JAL

K=5 Red – mostly GUE, slightly in MEX

Green – only found in JAL

Blue – mostly found in MEX-MIC, some in JAL and GUE

Yellow – mostly GUE, slightly in MEX,MIC

Pink – only found in JAL

K=10 Red – only found in GUE

Green – only found in GUE

Blue – mostly found in GUE, slightly in MEX

Yellow – only found in JAL

Pink – mostly found in GUE, slightly in MIC

Light blue – only found in JAL

Orange – mostly found in MEX-MIC, some in JAL and GUE

Purple(wine colour) - mostly found in GUE, slightly in MEX

Salmon – only found in JAL

Grey-blue – only found in JAL

i) The likelihood increases but is starting to plateau at K=10. The *change* in likelihood is levelling off as well (when the change in likelihood is zero or very low we have reached the plateau) at around K=5. After a peak at K=3, alpha is steadily decreasing. Importantly, for all parameters the two replicate runs are very consistent for all values of K

ii) As we increase K, Mexico and Michocan change little; most Mexico-Michocan individuals are members of a single population. In Jalisco, the number of inferred populations is steadily rising and these populations are relatively "clean" = most individuals are >90% assigned to a single population. Guerrero is interesting – at an early stage, a large number of Guerrero individuals have mixed ancestry – as K increases some of these individuals are found to belong to distinct populations. Overall, as shown by the histograms, the number of individuals inferred to have mixed ancestry goes down. This is what we would like to see.

iii) At K=6, the ΔK drops to zero, suggesting this may be the best choice. From K=6 to K=10 the Ln P(D) is still improving but the gains may be marginal. The barplots suggest there may some limited gain from K=6 to K=10.

iv) If you want to capture maximum genetic diversity you should probably sample at least one individual from all the 6 (or 10) populations which structure has found. Once you have done this, your last few individuals might be selected based on the traits you are interested in – e.g. sample more from populations in dry regions if interested in drought tolerance. Overall you would use all information which you have to hand.

Block 12: Barley data set

i) The likelihood increases rapidly from 1 to 2 populations but then evens off more quickly than in the teosinte data set. If we look at the rate of change of the likelihood, it hits zero from K= 3 onwards. Alpha peaks at 2, but then becomes unstable – replicates with the same K have very different values of alpha. Clearly Structure is having difficulty fitting a model to this data set. This implies 2 possibilities – either K=2 for this barley data set or, for all K, the Structure model is a poor fit to the barley data set. This is similar to the results from PCoA and the tree drawing exercise, although in the PCoA and tree drawing, K=3 was possible; here it is unlikely.

ii) When K=3, one of the populations is intermediate between the other two, and on a very short branch. This is not the pattern we would expect for three natural populations.

iii) As we move from K=3 to K=5 and K=10, fewer individuals are being assigned to a population with high probability, i.e. there appear to be more admixed individuals, the higher the number of K. This is not really what we would expect if Structure is finding real population structure with higher K values.

iv) Given what we have seen here, Structure may not be the best tool to analyse collections of inbred crop lines. This is not very surprising. The model in Structure assumes that the individuals fall into natural populations in linkage and Hardy-Weinberg equilibrium. This model is not a good fit to collections of lines, although it may capture some basic aspects, if for example the material comes from 2 long-separated breeding pools (as it does here). In general, for collections of breeding material, other techniques such as PCoA or PCA, may be better for quantifying population structure for use in association mapping.

Block 13:

LD decays very fast as function of genetic distance

However, the LD matrix still show weak LD between markers further apart (darker red dots in the plot). Note that one marker shows moderate LD with almost all others – this marker may have a low minor allele frequency (MAF)

Block 14:

While still clear that LD decays very fast as function of genetic distance, there is only a very small difference in the magnitude of values.

In the LD matrix, it is clearer that there has been a general reduction in "background" LD (relatively more grey in the plot than before), especially off-diagonal LD. Most of the stronger red are close to the diagonal (where are expected to be) and the rogue marker is less clear.

Block 15:

Once again, the LD is confirmed to decay very fast.

The most remarkable thing is that most of the high LD between distant markers has disappeared (the LD matrix show red mostly close to the diagonals, the rest mostly grey).In the upper plot, it is also clear that the "background" low-level LD has reduced

It seems as if all (or most) of the population effects have been removed (expected as from previous results it was clear the best way to capture the population structure).

Block 16:

There are two major differences between the previous Teosinte population and the barley population to consider here:

Wild versus breeding material

Outcross versus self-pollinated

A priori one would expect a higher degree of LD in the barley population because of the breeding system (selfer) and because it is a population with a shorter recombination history (these are varieties).

Regarding population structure, it is hard to anticipate, but perhaps the breeding material (as elite material) will have a more compact pedigree structure (higher degree of relatedness) than the Teosinte diversity panel. Therefore we expect perhaps a stronger population effect in the Teosinte set in than in the barley set.

Block 17:

The LD decay is less pronounced than in the Teosinte set.

Also in general the r^2 values are quite high.

The higher r^2 values are expected because of the type of material (breeding material with longer LD blocks due to shorter recombination history, selection, etc).

It is interesting to see what the effect of the population structure would be

Block 18:

LD still decays considerably more slowly than than in the Teosinte data set, as well as being more variable.

In general the r^2 values are lower (especially visible in the LD matrix plot) than when population structure is not considered, which points to some effect of population structure in the original assessment.

Block 19:

As predicted, barley LD is much higher overall:

shorter recombination history

Inbreeding

Population structure correction more effective in teosinte:

Pop structure more important than kinship in teosinte data set, opposite in barley

For both populations it is clear that failure to account for population structure may lead to erroneous results (false positives) in association mapping

Complex population structure important in teosinte

In barley, only simple population structure (2-3 groups) but accounting for kinship very important

Block 20:

The recombination rate varies hugely along the chromosome from almost 0 in the middle to 1.4 at the ends. As the genetic map is based on LD, it will not lengthen at all over regions of maximum LD

This is the region around the centromere of the chromosome. All organisms have low recombination around the centromere but in wheat this effect is extreme; about 25% of the chromosome is compeltely protected from recombination

LD is inversely proportional to recombination frequency so LD would be high in the middle and other flat bits, lower at the ends. The result is that it will be impossible to fine map a QTL which is found in the max LD regions – your haplotype will be 100s of Mb long, with 1000s of genes in the centromeric area.

High LD and HW disequilibrium distinct to a genetic region suggests that there is some form of selection in these regions. In fact, they appear to demonsrate underdominance. Our data strongly suggests that these regions of the chromosome are segregating for blocks of DNA introgressed from other grass species. Several of these blocks were introduced into the bread wheat genome by breeders to bring in favourable traits from other species (diseases resistance, yield). In this population, one or more parents have an introgressed fragment that cannot recombine with the native wheat chromosome in this region (carried by the other parents).

Further exercises: pedigree analysis

There are 7 possible paths

Common

Ancestor	Path	N_i	f (anc)	Contribution to f		
A	L-I-F-C-**A**-D-G-J-M		9	0	$(1/2)^9$	=0.0020
B	L-I-G-D-**B**-E-H-K-M		9	0	$(1/2)^9$	=0.0020
B	L-J-G-D-**B**-E-H-K-M		9	0	$(1/2)^9$	=0.0020
B	L-I-G-D-**B**-E-J-M		8	0	$(1/2)^8$	=0.0039
C	L-I-F-**C**-G-J-M		7	0	$(1/2)^7$	=0.0078
J	L-J-M		3	$(1/2)^4$	$(1/2)^3(1.0625)$	=0.1328
G	L-I-G-J-M		5	$(1/2)^3$	$(1/2)^5(1.125)$	=0.0352
					f	=0.1856

In the penultimate path, J is itself inbred through 4 loops (G-D-B-E) so has an inbreeding coefficient of $(1/2)^4 = 0.0625$ so we add this much extra inbreeding to the whole calculation i.e. multiply by (1+0.0625) = 1.0625

In the last path, G is itself inbred through 3 loops (C-A-D) so has an inbreeding coefficient of $(1/2)^3 = 0.125$ so we add this much extra inbreeding to the whole calculation i.e. multiply by (1+0.125) = 1.125

Chapter 9 – Association mapping.

Block 1:

The whole marker data set has some structure in it, which is absent after thinning. Therefore this structure comes from clustering of markers in one region (in this case they are in a well-known introgression from rye on chromosome 1B that is present in many wheat varieties). The thinned data is more representative of relationships over the whole genome.

Block 2:

a: fdr threshold<per<bonf. Bonferroni = too conservative.

b: fdr – fewer hits found with unthinned data. Using in accurate kinship results in loss of power

Block 3:

Now a large proportion of the genome appears to be highly significant. Clearly it is necessary to adjust for kinship to avoid false positives.

Block 4:

Rht2 vanishes, as might be expected, but so do several other weak SNPS. All but one of these were unmapped – they may have been linked to Rht2 – although DArT 1472 on chromosome 16, which was only just significant previously, also vanishes. On the other hand, Rht1 appears as a strong, highly significant effect, having previously been undetected. Also PpdD1 is more significant than previously and 2 novel hits have also appeared, on chromosomes 8 (c. 136cm) and 16 (c. 8cM). These new hits may have previously been masked by Rht2. This is certainly the case for Rht1, as it is strongly negatively correlated with Rht2 in the datasets: breeders use either Rht2 or Rht1 to achieve semi-dwarf height in wheat but never both, as these individuals are fully dwarf and of no use to a farmer.

Block 5:

It makes a small difference here – the QQ plots are cleaner and a few minor hits are no longer found when population structure is also adjusted for. It might be a problem that we are double adjusting here, although a minor one.

Block 6:

It made a very small difference to the hits detected for FT.

Block 7:

We take the following approach:

1. Take the GENOTYPIC dataset and assign a random marker to be the exact site of a major QTL, and 100 random markers to be the sites of minor QTL (i.e. very minor). This major QTL starts out as perfect so a 0 allele has a phenotype value of 0, a 1 allele has a phenotype value of 1
2. Adjust these new phenotypes to account for %variation explained (as chosen at start). The total variation to be explained is simply the variance of the 1+101 markers with QTLs to begin with.

3. Adjust the phenotypes for heritability by adding some random normal variation to all of them proportional to the heritability (i.e. h2=1 would add 0 environmental variation, h2=0.1 would add a number from a normal distribution with 10 times the variance of the trait at the end of step 2.
4. Repeat steps 1-3 x 1000. You now have 1000 phenotypes, each with the name of a marker, indicating where the QTL is.
5. Run the QTL mapping on all 1000 phenotypes and calculate how often you detect the focal QTL. It is fairer if you ignore the focal marker in doing this, i.e. effectively use the adjacent markers.

If you run 5 heritabilities x 5 % vars x 1000 simulations, you will run 25,000 QTL analyses, which is one reason why people are lazy about running power analyses in the literature……