

BIOL334 Computer Practical:

Analyse real genetic data for your Practical Report



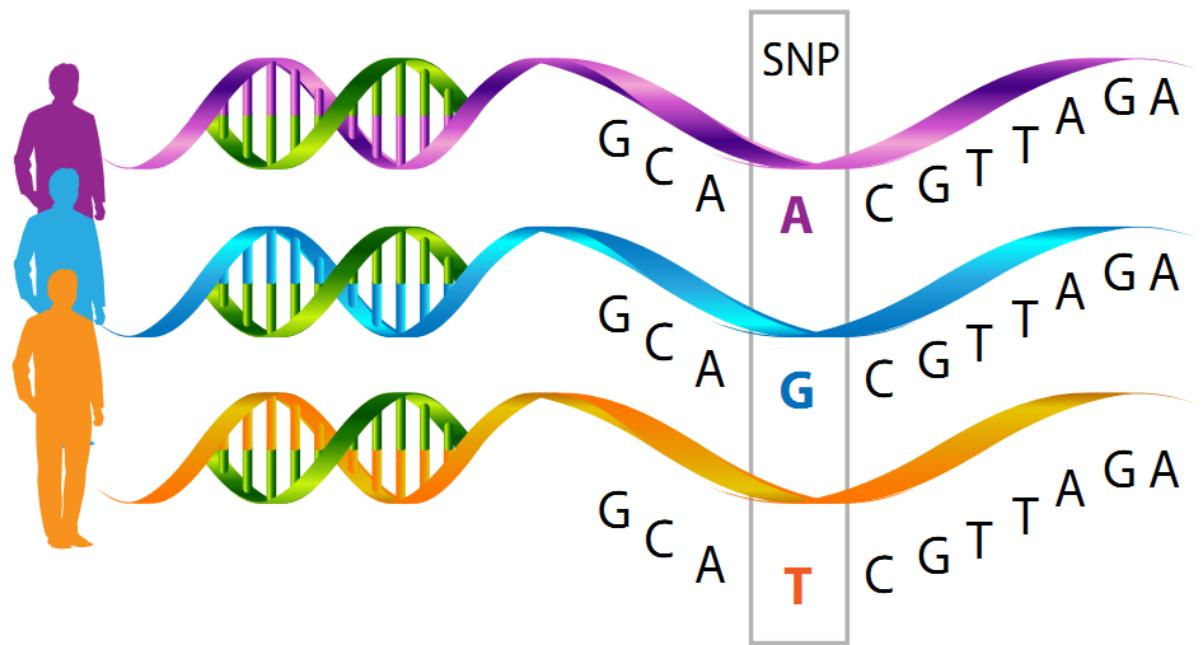
Practical Report

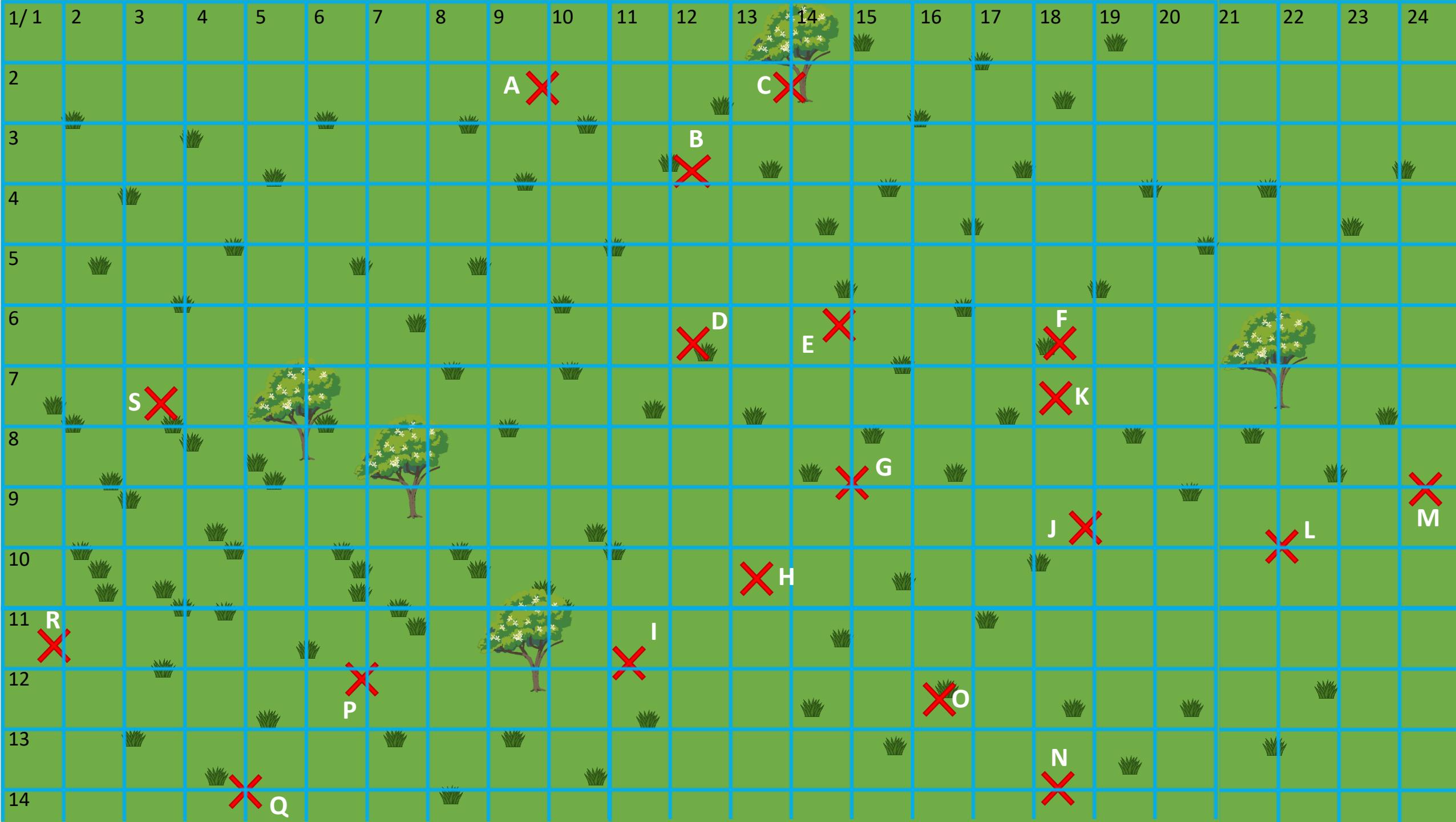
- Consult Unit Guide for Due Date
- Questions to be addressed are in these slides
- Rubric and instructions available online
 - under Wk 7 Tutorial on iLearn and in the Unit Guide

Make sure you save all of your results today!!!

The data:

- We have collected SNP data from 7 individuals at 19 locations
- We will assess:
 - Genetic summary statistics
 - Genetic structure
 - Spatial structure



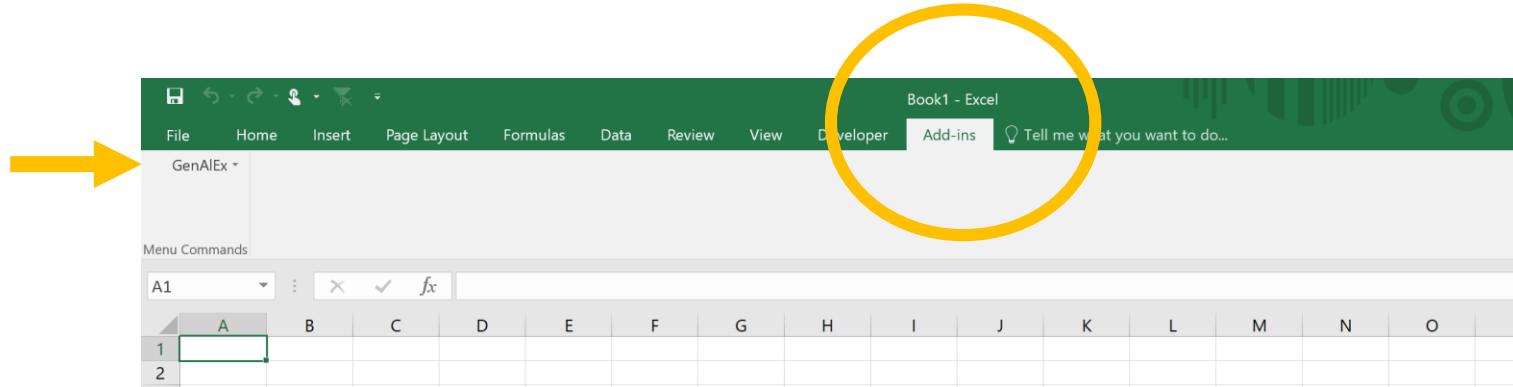


The Questions

- 1/ Is there any evidence for genetic structure among the sampling localities?
- 2/ Does genetic variation follow a pattern of isolation-by-distance?
- 3/ Are there environmental features that are implicated in patterns of genetic structure?
- 4/ Is there a risk of inbreeding?

Download Datasets and GenAIEx

- We will be using GenAIEx – an add-in for Excel
- Download GenAIEx and the Datasets from iLearn
 - Located under Week 7 Tutorial in the folder “Computer Practical materials”
 - There will be 3 Excel files with genetic data (Datasets 1, 2a and 2b)
- Open Excel and then open GenAIEx
 - A tab called “add-ins” should appear in Excel, with GenAIEx within that tab



Open Dataset 1 – Get to know the data

The screenshot shows a Microsoft Excel spreadsheet titled "Biol334_Dataset1 - Ex". The ribbon menu is visible at the top, and the formula bar shows "A1" and "1800". The data starts with a header row (Row 1) containing sample counts: 1800, 133, 19, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7. Rows 2 through 33 contain individual data points for each sample across 1800 SNPs. The columns are labeled A through N, and the rows are labeled 1 through 33. The first few rows of data are as follows:

1	1800	133	19	7	7	7	7	7	7	7	7	7	7	7	7	7
2		Pops	SiteA	SiteB	SiteC	SiteD	SiteE	SiteF	SiteG	SiteH	Sitel	SiteJ	SiteK			
3	Sample	Pop	SNP_1		SNP_2		SNP_3		SNP_4		SNP_5		SNP_6			
4	A1	SiteA	100	120	120	120	120	120	120	110	110	130	130			
5	A2	SiteA	100	100	110	120	120	120	120	0	0	130	130			
6	A3	SiteA	100	100	120	120	120	120	120	110	110	130	130			
7	A4	SiteA	120	120	110	120	120	120	120	110	110	130	130			
8	A5	SiteA	100	100	120	120	120	120	120	110	110	100	130			
9	A6	SiteA	100	100	120	120	100	120	120	0	0	100	130			
10	A7	SiteA	100	100	120	120	120	120	120	110	110	130	130			
11	B1	SiteB	100	100	110	120	120	120	120	110	110	130	130			
12	B2	SiteB	100	100	120	120	120	120	120	110	110	100	130			
13	B3	SiteB	100	100	120	120	100	120	120	110	110	100	130			
14	B4	SiteB	100	100	110	120	120	120	120	110	110	130	130			
15	B5	SiteB	100	100	120	120	120	120	120	110	110	130	130			
16	B6	SiteB	100	100	120	120	120	120	120	110	110	130	130			
17	B7	SiteB	100	100	120	120	120	120	120	110	110	100	130			
18	C1	SiteC	100	100	120	120	120	120	120	110	110	100	100			
19	C2	SiteC	100	100	120	120	120	120	120	110	110	130	130			
20	C3	SiteC	100	100	120	120	120	120	120	110	110	100	100			
21	C4	SiteC	100	100	120	120	100	120	120	110	110	130	130			
22	C5	SiteC	100	100	120	120	120	120	120	110	110	130	130			
23	C6	SiteC	100	100	110	120	120	120	120	110	110	130	130			
24	C7	SiteC	100	100	120	120	120	120	120	110	110	130	130			
25	D1	SiteD	0	0	120	120	120	120	120	110	110	130	130			
26	D2	SiteD	120	120	120	120	120	100	100	110	120	130	130			
27	D3	SiteD	100	100	120	120	120	120	120	110	110	130	130			
28	D4	SiteD	100	100	120	120	120	120	0	0	110	110	130	130		
29	D5	SiteD	100	100	120	120	120	120	120	110	110	130	130			
30	D6	SiteD	100	120	120	120	120	120	0	0	110	110	130	130		
31	D7	SiteD	100	100	120	120	120	120	120	110	110	130	130			
32	E1	SiteE	100	120	0	0	100	120	120	110	110	130	130			
33	F1	SiteF	100	120	120	120	120	120	120	110	110	130	130			

19 “Populations” (= “Sampling Sites”), named A → S

7 Samples (= “Individual animals”) per site, named with the Sample name and 1 → 7
= Total 133 samples

1800 SNPs (= “single nucleotide polymorphism”), sequenced for each individual:

100 = A

110 = C

120 = G

130 = T

0 = Missing data

Open Dataset 1 – Get to know the data

Biol334_Dataset1 - Ex													
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Developer	Add-ins	GenAlEx	Menu Commands		
1	1800	133	19	7	7	7	7	7	7	7	7	7	7
2	Pops	SiteA	SiteB	SiteC	SiteD	SiteE	SiteF	SiteG	SiteH	Sitel	SiteJ	SiteK	
3	Sample	Pop	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6					
4	A1	SiteA	100	120	120	120	120	120	110	110	130	130	
5	A2	SiteA	100	100	110	120	120	120	0	0	130	130	
6	A3	SiteA	100	100	120	120	120	120	110	110	130	130	
7	A4	SiteA	120	120	110	120	120	120	110	110	130	130	
8	A5	SiteA	100	100	120	120	120	120	110	110	100	130	
9	A6	SiteA	100	100	120	120	100	120	120	0	0	100	130
10	A7	SiteA	100	100	120	120	120	120	110	110	130	130	
11	B1	SiteB	100	100	110	120	120	120	110	110	130	130	
12	B2	SiteB	100	100	120	120	120	120	110	110	100	130	
13	B3	SiteB	100	100	120	120	100	120	120	110	110	100	130
14	B4	SiteB	100	100	110	120	120	120	110	110	130	130	
15	B5	SiteB	100	100	120	120	120	120	110	110	130	130	
16	B6	SiteB	100	100	120	120	120	120	110	110	130	130	
17	B7	SiteB	100	100	120	120	120	120	110	110	100	130	
18	C1	SiteC	100	100	120	120	120	120	110	110	100	100	
19	C2	SiteC	100	100	120	120	120	120	110	110	130	130	
20	C3	SiteC	100	100	120	120	120	120	110	110	100	100	
21	C4	SiteC	100	100	120	120	100	120	120	110	110	130	130
22	C5	SiteC	100	100	120	120	120	120	110	110	130	130	
23	C6	SiteC	100	100	110	120	120	120	110	110	130	130	
24	C7	SiteC	100	100	120	120	120	120	110	110	130	130	
25	D1	SiteD	0	0	120	120	120	120	110	110	130	130	
26	D2	SiteD	120	120	120	120	120	100	110	120	130	130	
27	D3	SiteD	100	100	120	120	120	120	110	110	130	130	
28	D4	SiteD	100	100	120	120	120	0	0	110	110	130	130
29	D5	SiteD	100	100	120	120	120	120	110	110	130	130	
30	D6	SiteD	100	120	120	120	120	0	0	110	110	130	130
31	D7	SiteD	100	100	120	120	120	120	110	110	130	130	
32	E1	SiteE	100	120	0	0	100	120	120	110	110	130	130

Data for each individual are in one row

Data for each SNP locus are in two columns,
= one column for each allele at that locus!

So individual E1:

Heterozygous for SNP 1 (A/G) and SNP 3 (A/G)

Has missing data for SNP 2

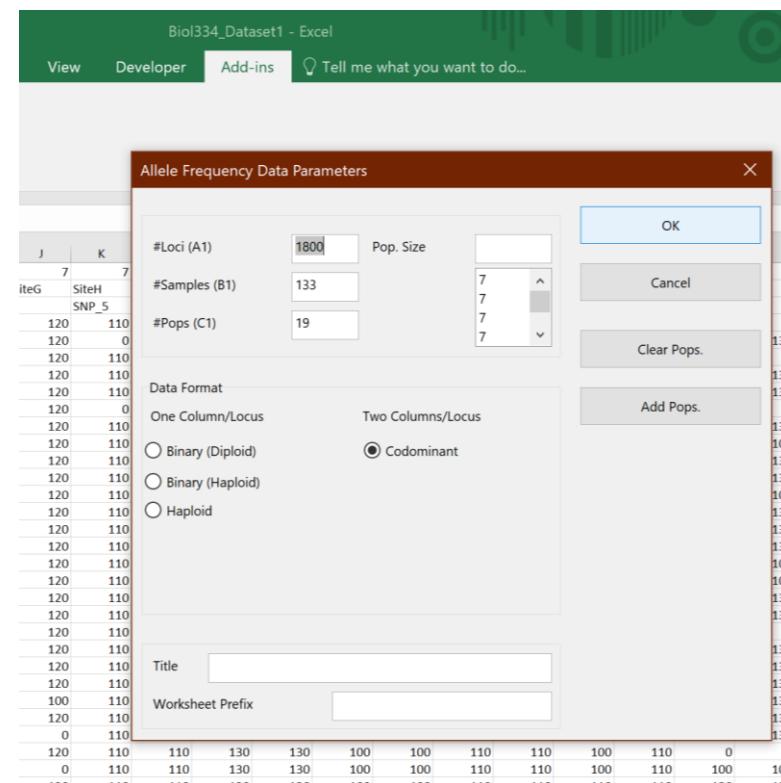
Homozygous for SNP 4 (G/G)

Calculate Genetic Summary Statistics:

Click GenAIEx, then Frequency

The screenshot shows a Microsoft Excel interface with the 'Biol334_Dataset1' workbook open. The 'GenAIEx' tab is selected in the ribbon. A context menu is open under the 'GenAIEx' tab, with 'Frequency...' highlighted. The main worksheet, 'Dataset1_SNPs', contains a table of genetic data with columns labeled C through M and rows labeled SiteA through SiteK. The cell at F120 contains the value '120'. The 'Data' tab is also visible in the ribbon.

In the pop-up window, check that the numbers for Loci, Samples and Pops are correct. Check the **Codominant** data format and click **OK**



Calculate Genetic Summary Statistics:

In the pop-up window:

UN-check Frequency by Pop

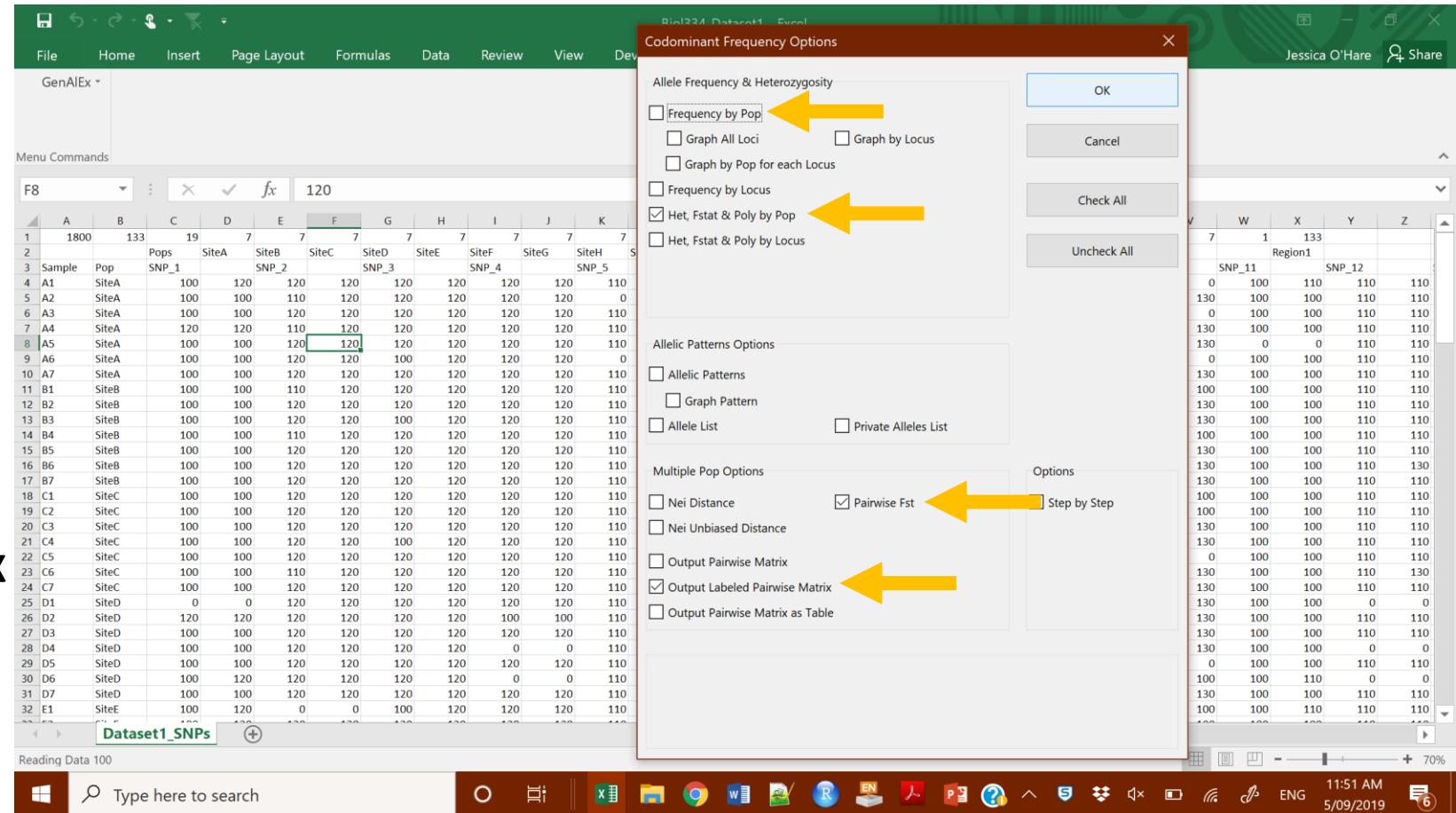
Check:

Het, Fstat & Poly by Pop

Pairwise Fst

Output Labelled Pairwise Matrix

Then click OK



You will get two results sheets:

HFP = Het, Fstat & Poly by Pop

FstL = Labeled Pairwise Fst Matrix

***Be patient – the calculations
will take a few minutes!!!***

HFP – Het, Fstat, Poly by Pop

This sheet has info for each locus within each pop



At the end, there is a table
with an overall mean per pop

(starting at row ~34214)

Biol334_Dataset1 - Excel

File Home Insert Page Layout Formulas Data Review View Developer Add-ins Tell me what you want to do... Jessica O'Hare Share GenAIEx

Menu Commands

C34211

A B C D E F G H I J K L M N O P Q

34213

34214 Mean and SE over Loci for each Pop

34215

34216 Pop

34217 SiteA Mean

34218 SE

34219

34220 SiteB Mean

34221 SE

34222

34223 SiteC Mean

34224 SE

34225

34226 SiteD Mean

34227 SE

34228

34229 SiteE Mean

34230 SE

34231

34232 SiteF Mean

34233 SE

34234

34235 SiteG Mean

Dataset1_SNPs HFP FstL

Ready

HFP – Het, Fstat, Poly by Pop



N = No. of samples with data

Na = No. of Different Alleles

Ne = No. of Effective Alleles = $1 / (\text{Sum } \pi^2)$

I = Shannon's Information Index
= $-1 * \text{Sum}(\pi * \ln(\pi))$

Ho = Observed Heterozygosity = No. of Hets / N

He = Expected Heterozygosity = $1 - \text{Sum } \pi^2$

uHe = Unbiased Expected Heterozygosity
= $(2N / (2N-1)) * He$

F = Fixation Index = $(He - Ho) / He = 1 - (Ho / He)$

Where π is the frequency of the i th allele for the population
& $\text{Sum } \pi^2$ is the sum of the squared population allele frequencies.

Can you see any major differences between pops?

A	B	C	D	E	F	G	H	I	J
Pop		N	Na	Ne	I	Ho	He	uHe	F
SiteA	Mean	6.876	1.732	1.389	0.360	0.213	0.236	0.254	0.072
	SE	0.009	0.010	0.008	0.006	0.005	0.004	0.005	0.010
SiteB	Mean	6.878	1.733	1.400	0.368	0.214	0.242	0.261	0.088
	SE	0.010	0.010	0.008	0.006	0.005	0.004	0.005	0.010
SiteC	Mean	6.912	1.745	1.401	0.370	0.215	0.242	0.261	0.081
	SE	0.008	0.010	0.008	0.006	0.005	0.004	0.005	0.010
SiteD	Mean	6.944	1.719	1.401	0.364	0.265	0.240	0.259	-0.090
	SE	0.006	0.011	0.008	0.006	0.006	0.004	0.005	0.010
SiteE	Mean	6.958	1.686	1.413	0.361	0.287	0.241	0.260	-0.150
	SE	0.005	0.011	0.009	0.006	0.007	0.005	0.005	0.010
SiteF	Mean	6.901	1.780	1.440	0.397	0.266	0.262	0.283	-0.019
	SE	0.008	0.010	0.008	0.006	0.005	0.004	0.005	0.009
SiteG	Mean	6.924	1.844	1.448	0.414	0.267	0.271	0.292	0.001
	SE	0.007	0.009	0.008	0.005	0.005	0.004	0.004	0.009
SiteH	Mean	6.808	1.790	1.436	0.398	0.260	0.262	0.283	0.001
	SE	0.012	0.010	0.008	0.006	0.005	0.004	0.004	0.010
SiteI	Mean	6.920	1.787	1.431	0.396	0.256	0.261	0.281	0.006
	SE	0.007	0.010	0.008	0.006	0.005	0.004	0.004	0.009
SiteJ	Mean	6.926	1.836	1.446	0.413	0.270	0.270	0.292	-0.011
	SE	0.007	0.009	0.008	0.005	0.005	0.004	0.004	0.009
SiteK	Mean	6.876	1.813	1.441	0.406	0.257	0.266	0.287	0.017
	SE	0.009	0.009	0.008	0.006	0.005	0.004	0.004	0.009
SiteL	Mean	6.929	1.809	1.439	0.404	0.262	0.265	0.286	-0.003
	SE	0.007	0.009	0.008	0.006	0.005	0.004	0.004	0.009
SiteM	Mean	6.932	1.743	1.412	0.374	0.248	0.247	0.266	-0.012
	SE	0.007	0.010	0.008	0.006	0.005	0.004	0.005	0.009
SiteN	Mean	6.932	1.792	1.432	0.396	0.259	0.260	0.281	-0.005
	SE	0.007	0.010	0.008	0.006	0.005	0.004	0.004	0.009
SiteO	Mean	6.923	1.752	1.411	0.375	0.259	0.247	0.266	-0.047
	SE	0.007	0.010	0.008	0.006	0.006	0.004	0.005	0.009
SiteP	Mean	6.938	1.816	1.444	0.407	0.268	0.267	0.288	-0.013
	SE	0.006	0.009	0.008	0.006	0.005	0.004	0.004	0.009
SiteQ	Mean	6.921	1.784	1.413	0.384	0.261	0.251	0.270	-0.044
	SE	0.007	0.010	0.008	0.006	0.005	0.004	0.004	0.009
SiteR	Mean	6.901	1.673	1.390	0.349	0.253	0.232	0.250	-0.081
	SE	0.009	0.011	0.009	0.006	0.006	0.005	0.005	0.010
SiteS	Mean	6.936	1.741	1.425	0.382	0.263	0.253	0.272	-0.037
	SE	0.006	0.010	0.008	0.006	0.006	0.004	0.005	0.009

FstL = Pairwise FST matrix

This sheet has a matrix for FST between each possible pair of pops

(with zero along the diagonal,
where it is a comparison
of each pop with itself)

Can you see any patterns?

W20

11 Pairwise Population Fst Values

12

13 SiteA SiteB SiteC SiteD SiteE SiteF SiteG SiteH SiteI SiteJ SiteK SiteL SiteM SiteN SiteO SiteP SiteQ SiteR SiteS

14 SiteA

15 SiteB

16 SiteC

17 SiteD

18 SiteE

19 SiteF

20 SiteG

21 SiteH

22 SiteI

23 SiteJ

24 SiteK

25 SiteL

26 SiteM

27 SiteN

28 SiteO

29 SiteP

30 SiteQ

31 SiteR

32 SiteS

33

Dataset1_SNPs HFP FstL



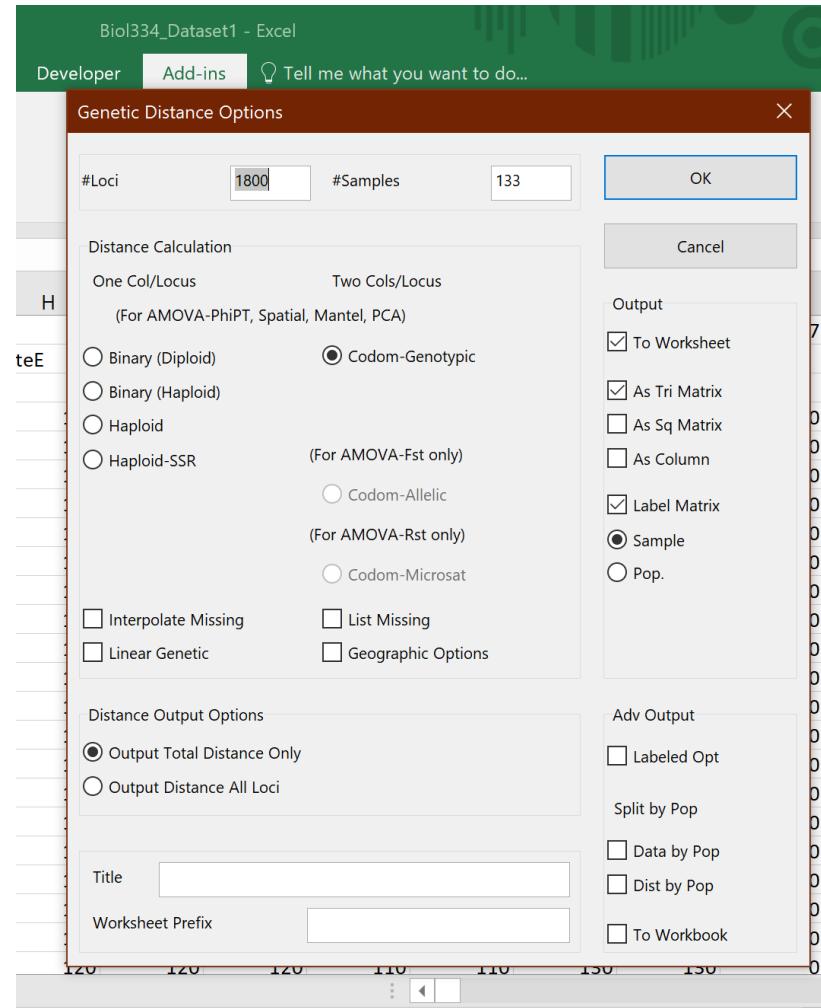
	SiteA	SiteB	SiteC	SiteD	SiteE	SiteF	SiteG	SiteH	Sitel	SiteJ	SiteK	SiteL	SiteM	SiteN	SiteO	SiteP	SiteQ	SiteR	SiteS	
SiteA	0.000																			SiteA
SiteB	0.050	0.000																		SiteB
SiteC	0.047	0.048	0.000																	SiteC
SiteD	0.135	0.132	0.130	0.000																SiteD
SiteE	0.135	0.133	0.132	0.053	0.000															SiteE
SiteF	0.111	0.108	0.105	0.043	0.058	0.000														SiteF
SiteG	0.101	0.100	0.098	0.050	0.067	0.039	0.000													SiteG
SiteH	0.110	0.109	0.106	0.044	0.046	0.041	0.039	0.000												SiteH
Sitel	0.107	0.105	0.102	0.056	0.046	0.048	0.044	0.034	0.000											Sitel
SiteJ	0.097	0.095	0.092	0.074	0.077	0.051	0.041	0.054	0.053	0.000										SiteJ
SiteK	0.103	0.103	0.102	0.064	0.063	0.048	0.042	0.045	0.045	0.038	0.000									SiteK
SiteL	0.102	0.099	0.096	0.068	0.067	0.052	0.043	0.047	0.042	0.039	0.043	0.000								SiteL
SiteM	0.105	0.104	0.102	0.090	0.080	0.070	0.063	0.059	0.047	0.057	0.060	0.048	0.000							SiteM
SiteN	0.104	0.104	0.100	0.088	0.098	0.070	0.061	0.069	0.075	0.061	0.067	0.067	0.080	0.000						SiteN
SiteO	0.115	0.112	0.108	0.100	0.113	0.084	0.072	0.086	0.085	0.073	0.078	0.075	0.090	0.052	0.000					SiteO
SiteP	0.103	0.100	0.098	0.082	0.097	0.067	0.053	0.070	0.072	0.056	0.066	0.062	0.079	0.055	0.053	0.000				SiteP
SiteQ	0.113	0.110	0.108	0.093	0.111	0.079	0.064	0.083	0.083	0.067	0.075	0.073	0.090	0.061	0.046	0.040	0.000			SiteQ
SiteR	0.136	0.135	0.132	0.101	0.118	0.091	0.076	0.095	0.095	0.083	0.093	0.086	0.109	0.094	0.091	0.057	0.065	0.000		SiteR
SiteS	0.115	0.114	0.109	0.091	0.108	0.076	0.062	0.081	0.083	0.069	0.077	0.072	0.091	0.070	0.066	0.043	0.048	0.046	0.000	Sites
	SiteA	SiteB	SiteC	SiteD	SiteE	SiteF	SiteG	SiteH	Sitel	SiteJ	SiteK	SiteL	SiteM	SiteN	SiteO	SiteP	SiteQ	SiteR	SiteS	

Genetic distance between individuals

Click GenAlEx > Distance > Genetic

The screenshot shows the Microsoft Excel ribbon with the 'Biol334_Dataset1' tab selected. The 'GenAlEx' tab is active. Under the 'Genetic...' command in the 'Distance' section, the 'Genetic...' option is highlighted.

In the pop-up window,
make sure the defaults match those shown here



Click OK

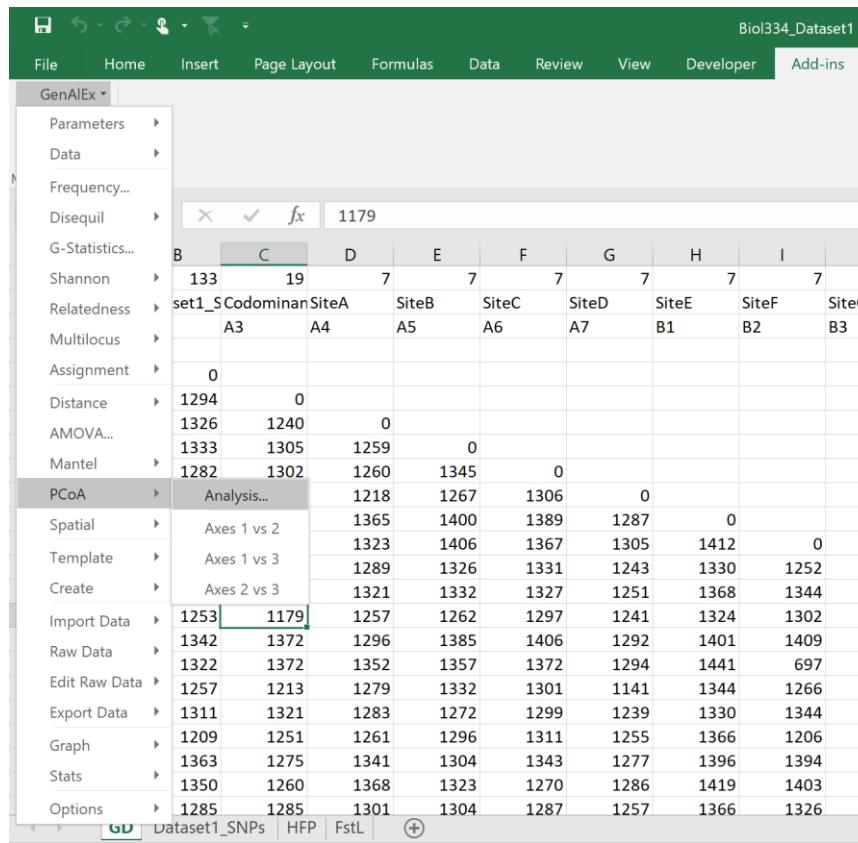
You will get a sheet called
GD (Genetic distance)

Genetic distance between individuals

Complete a Principal Coordinates Analysis (PCoA) using the Genetic Distance data

Within the GD sheet,

Click Genalex > PCoA > Analysis



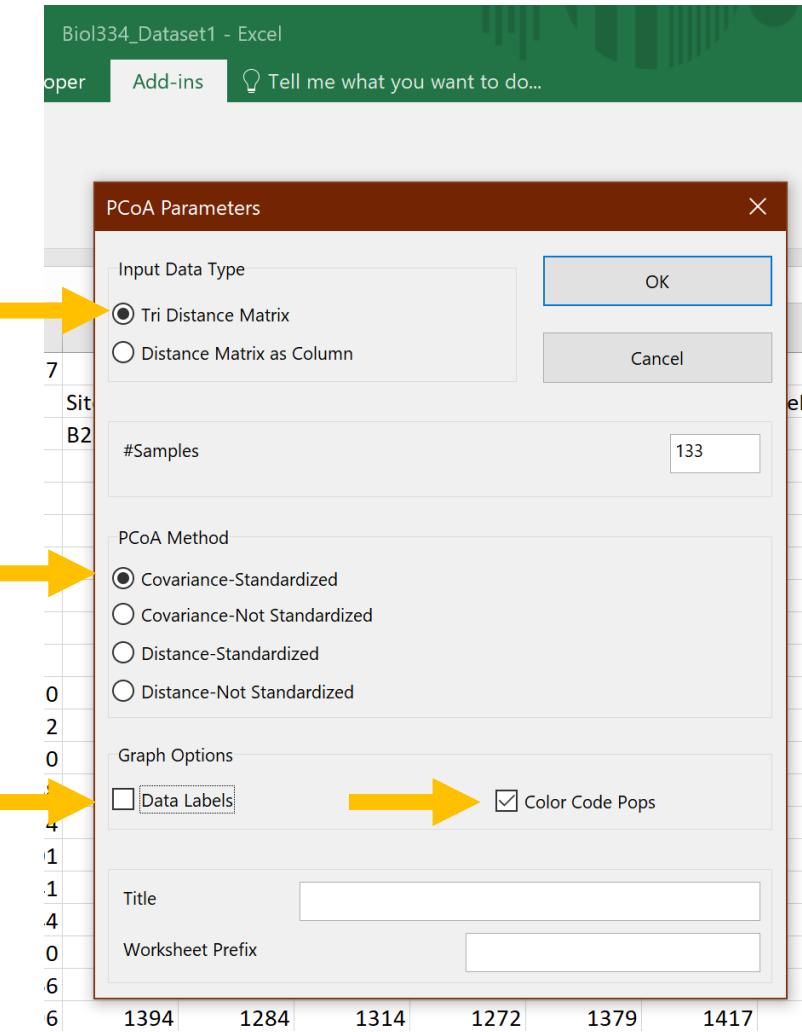
In the pop-up window,
click:

**Tri Distance Matrix
Covariance-Standardised
Color Code Pops**

UN-check Data Labels

Click OK

You will get a new sheet
called **PCoA** with a plot



Genetic distance between individuals

In the PCoA, each dot represents an individual
The closer together the dots (individuals),
the more closely related they are

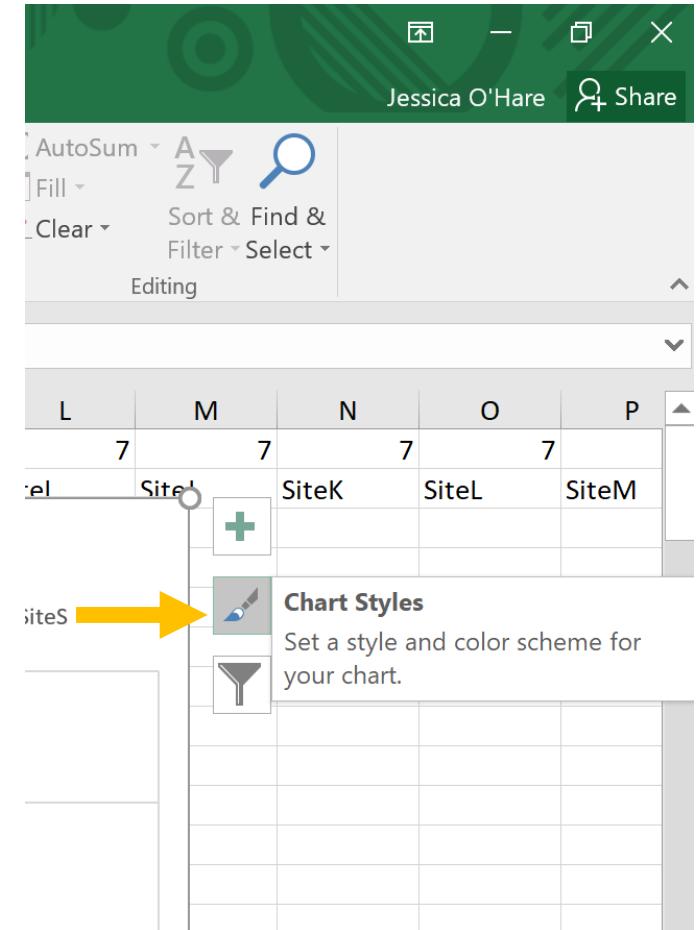
Make the PCoA plot bigger so that you can
clearly see all the Pop category labels

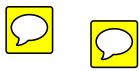
If the colour/symbol theme for the labels
isn't very clear, change the style and/or colours →

Are there any clear patterns in the PCoA?

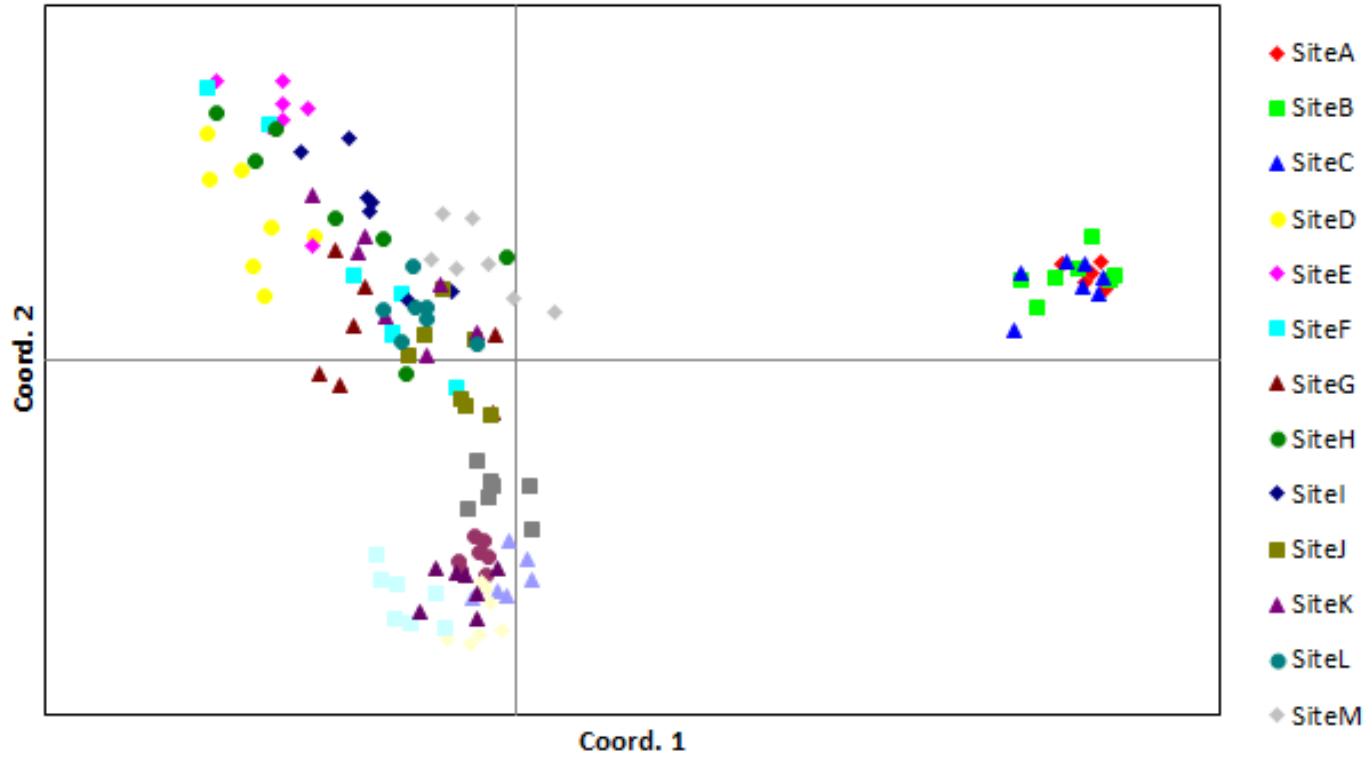
Look at the sampling map..

How does the PCoA match with the sampling scheme?

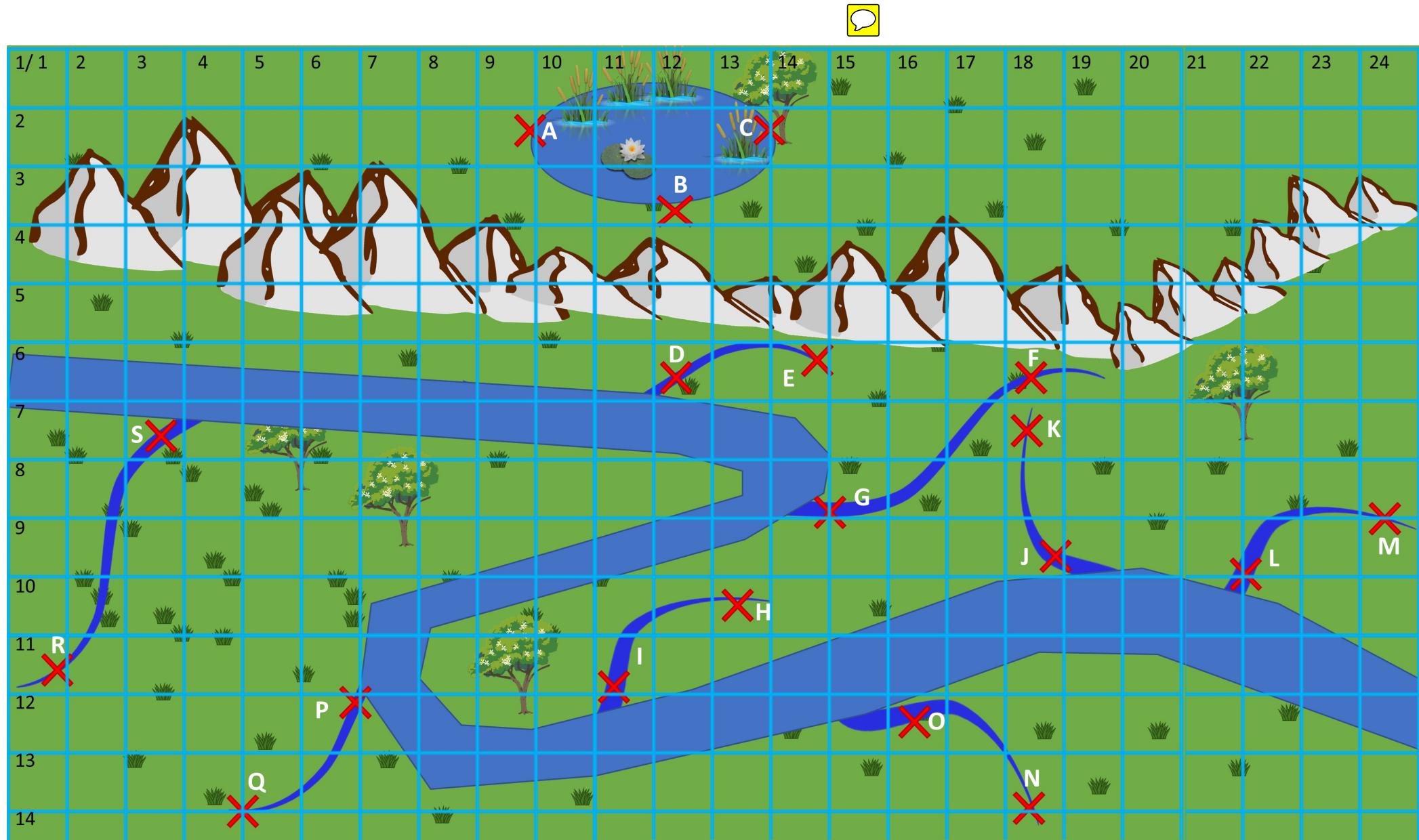




Principal Coordinates (PCoA)







Have another look at the Pairwise FST (FstL)

Do you see any patterns that reflect the PCoA results?

Let's now include some landscape information

Can this explain any patterns we may have observed?

Spatial auto-correlation

Test whether there is a relationship between relatedness and geographic distance

Open the other datasets from iLearn:

Dataset 2a = North pop

Dataset 2b = South pop

These datasets have a SNP data sheet, and GeoDist sheet with a matrix (in KM).

We used the grid numbers to generate geographic distance between sampling sites (“pops”)

Calculate Genetic Distance for both Datasets

GenAIEx > Distance > Genetic

Bio334_Dataset2a - Excel														
File	Home	Insert	Page Layout	Formulas	Data	Review	View	Developer	Add-ins	?	Tell me what you want to do...			
GenAlEx	Parameters													
	Data													
	Frequency...													
	Disequil													
	G-Statistics...													
Shannon	B	C	D	E	F	G	H	I	J	K	L	M	N	1
	70	10	7	7	7	7	7	7	7	7	7	7	7	1
Relatedness	Pops	SiteD	SiteE	SiteF	SiteG	SiteH	Sitel	SiteJ	SiteK	SiteL	SiteM			
	SNP_1		SNP_2		SNP_3		SNP_4		SNP_5		SNP_6			
Multilocus		100	120	120	120	120	120	120	120	110	110	130		130
Assignment		100	100	120	120	120	120	120	120	110	110	130		130
Distance	Genetic...			110	120	120	120	120	120	110	120	130		130
AMOVA...	Geographic...			120	120	120	120	120	120	110	120	0		0
Mantel	Genetic by Pop...			110	120	120	120	120	120	110	110	130		130
PCoA	Tri->Table			120	120	120	120	120	120	110	110	130		130
Spatial	Col->Table			110	120	100	120	120	120	110	110	100		130
Template	Tri->Labeled			120	120	120	120	120	120	110	110	130		130
Create	Sq->Labeled			110	120	120	120	120	120	110	110	130		130
Import Data	Tri->Extract Pops...			110	120	120	120	120	120	120	120	130		130
Raw Data	Col->Extract Pops...			110	120	120	120	120	120	110	110	130		130
Edit Raw Data	Tri->Extract Pops+Regions...			110	120	120	120	120	120	110	110	130		130
Export Data	Col->Extract Pops+Regions...			120	120	120	120	120	120	110	110	130		130
Graph	100			100	0	0	100	120	100	100	110	110	130	130
Stats	120			120	110	120	120	120	100	120	110	110	130	130
Options	100			100	110	120	120	120	120	120	110	110	130	130
	Dataset2a_SNPs			Dataset2a_GeoDist			(+)			:			<	

Leave defaults

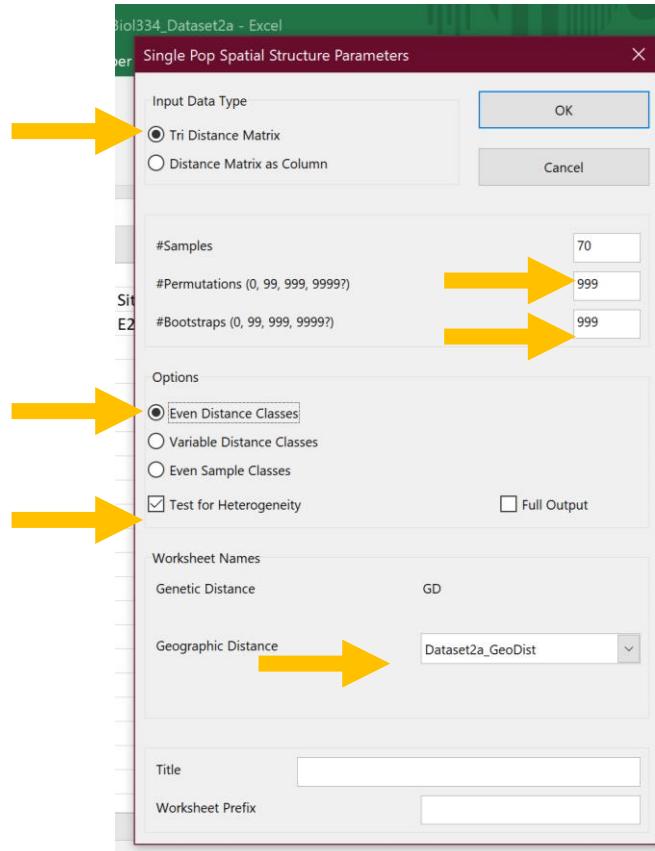
Compare Genetic and Geographic Distance

From within the **GD** sheet:

GenAIEx > Spatial > Single-pop

The screenshot shows the GenAIEx software interface. The 'Spatial' menu is open, and the 'Single Pop...' option is highlighted. The main workspace shows a portion of a dataset with columns labeled B through F and rows labeled SiteA through SiteF.

	B	C	D	E	F
Shannon	70	10	7	7	7
Relatedness	set2a_Codominant	SiteD	SiteE	SiteF	
Multilocus	D3	D4	D5	D6	
Assignment	0				
Distance	1097	0			
AMOVA...	983	1200	0		
Mantel	831	1138	1210	0	
PCoA	786	919	1053	843	0
	1050	1065	1129	729	976
Spatial	Single Pop...				



Tri-distance Matrix

999 Permutations

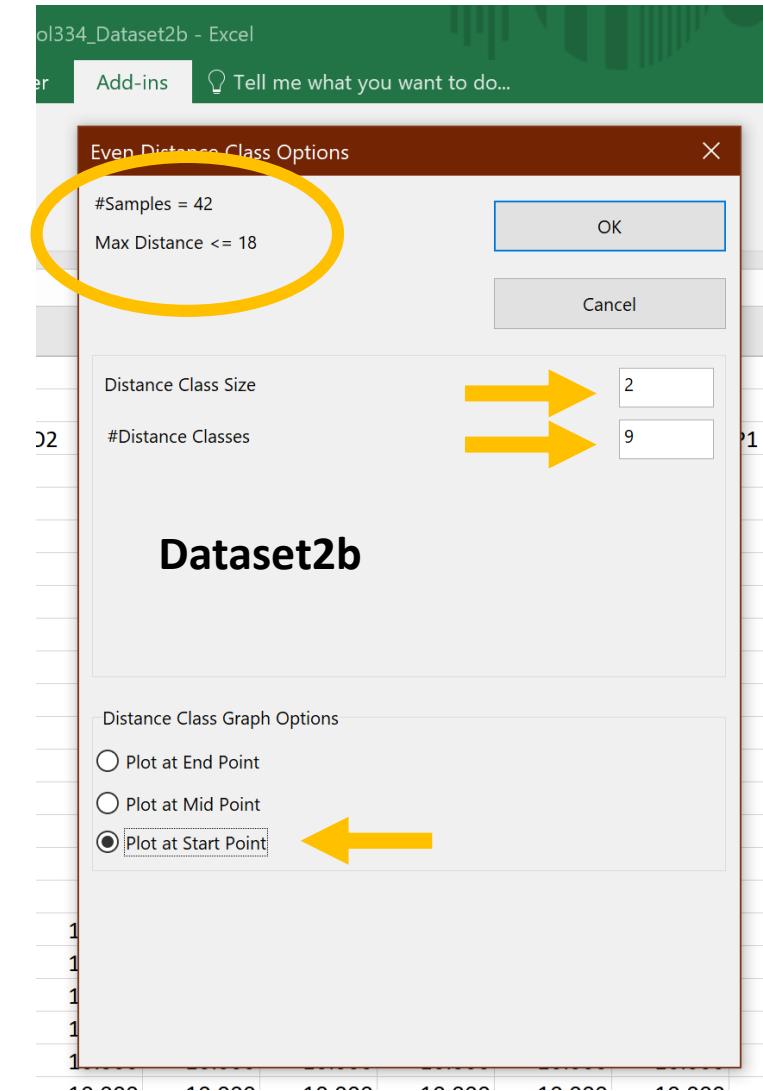
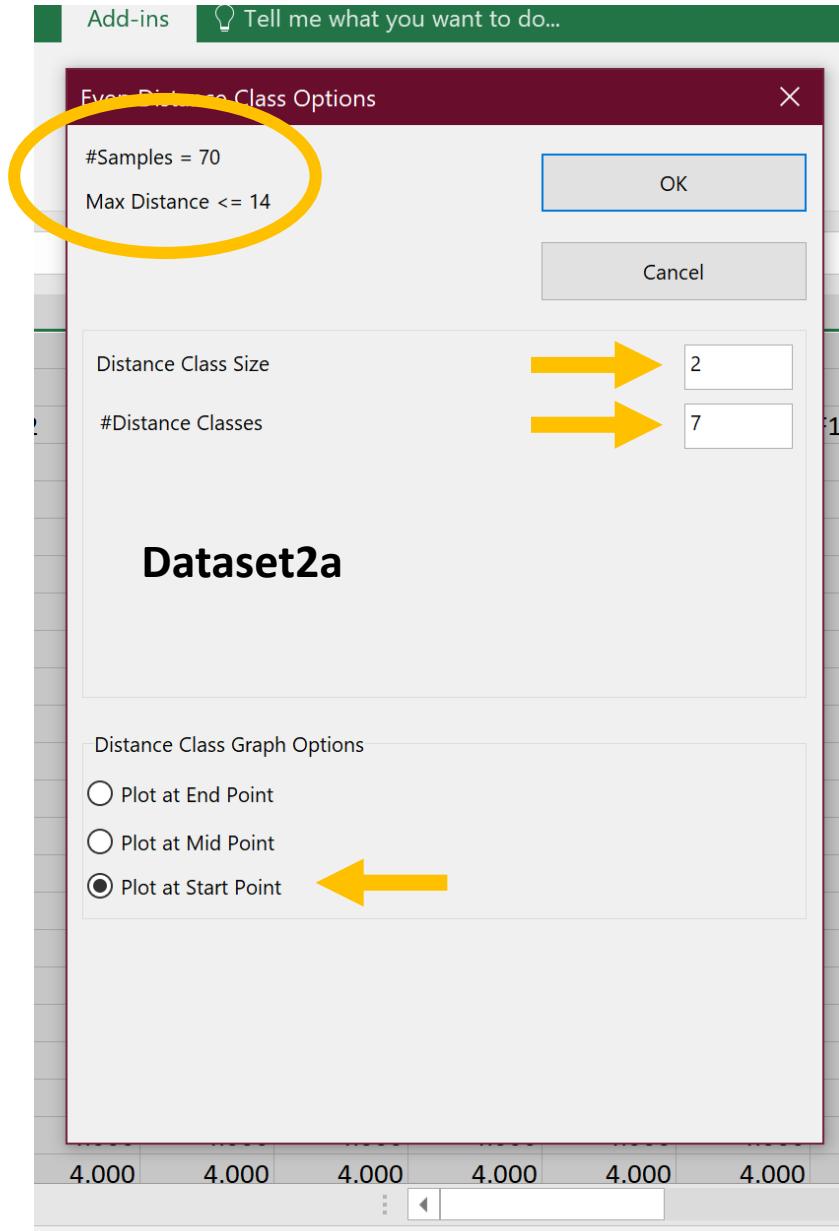
999 Bootstraps

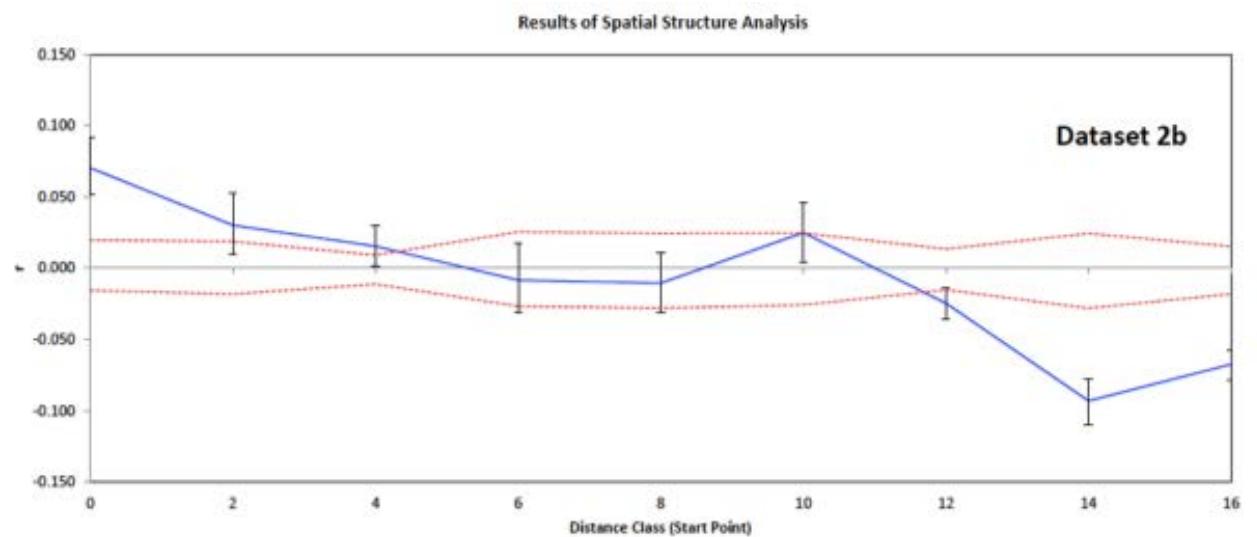
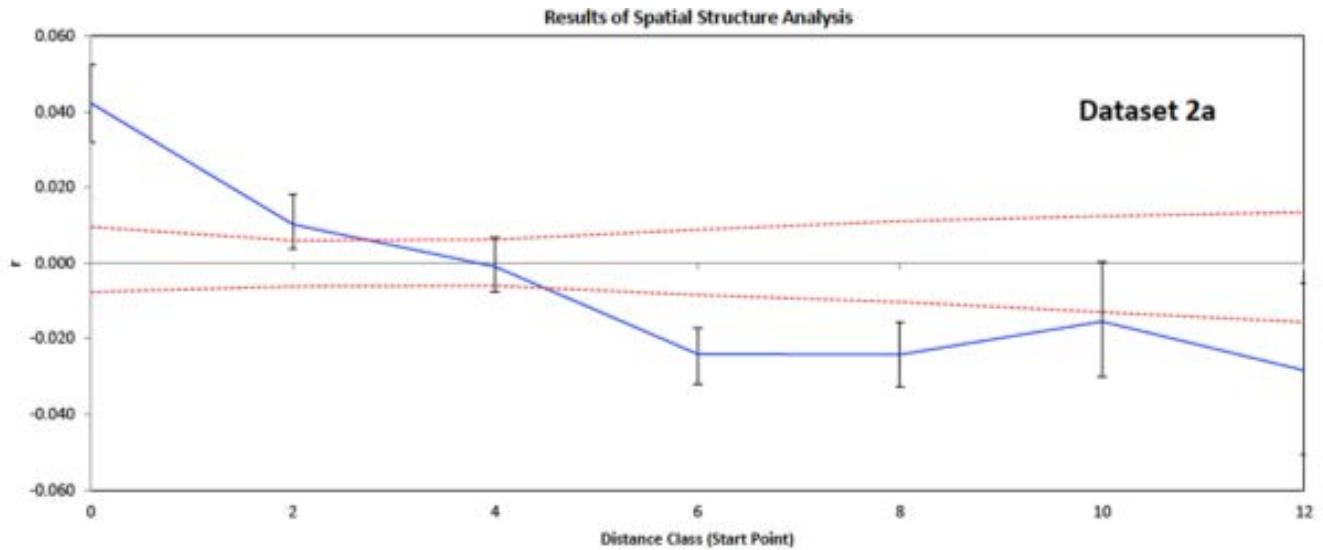
Even Distance Classes

Test for Heterogeneity

Match Geo Dist with the correct Geo Dist Sheet

Compare Genetic and Geographic Distance





Compare Genetic and Geographic Distance

What can you conclude from the correlogram (spatial autocorrelation plot)?

What are the red dashed lines?

What does the heterogeneity test tell us? What is the null hypothesis?

What does a significant result indicate?

Which distance classes are significantly different from what we expect under a null model?

Possible organism?

- Given the landscape patterns and genetic structure observed, what traits do you expect this organism to have?

Possible future work



- What are some other approaches that could help us better understand how environmental variables shape genetic variation?
- These can be suggested in the Discussion

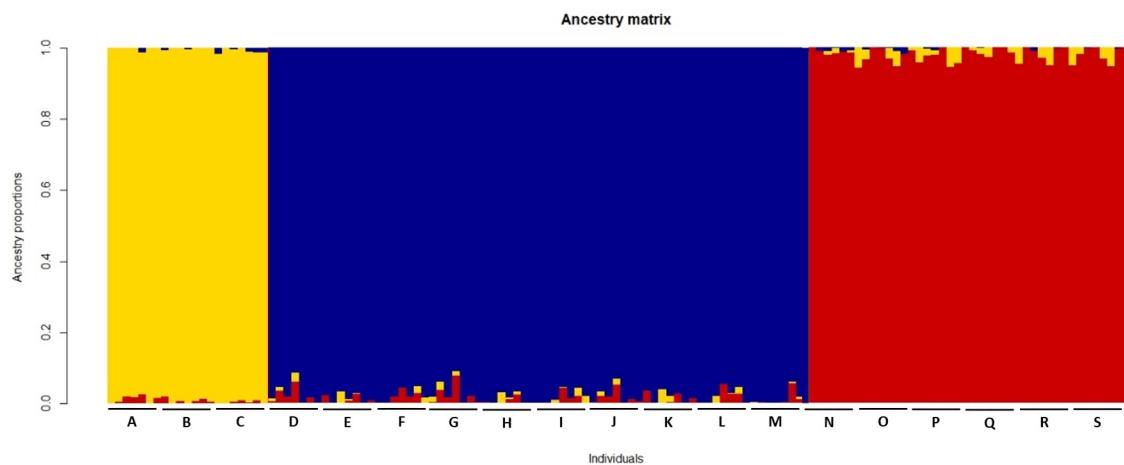


The table

Calculate using linkage disequilibrium

As the population gets smaller, the incidence of linkage disequilibrium increases

The mountain group is isolated by mountain range - Elevated homozygosity - effective pop size is lower than the other 2 groups



3 genetic groups

Those groups are evident in PCoA and Fst

Admixture - group individuals into genetic group based on HW

Genetic approach

Look at genotypes that would be expected in HW

3 different colours representing 3 different groups -

Proportion of genetic heritage that can be associated with those 3 groups

Individuals from each different group are pretty much linked to the group and not related with another group

Blue top - yellow middle - red bottom

Make sure you have saved all your results!!

Good luck preparing your Practical Report ☺