**How to use STRUCTURE FROM vcf FILE**

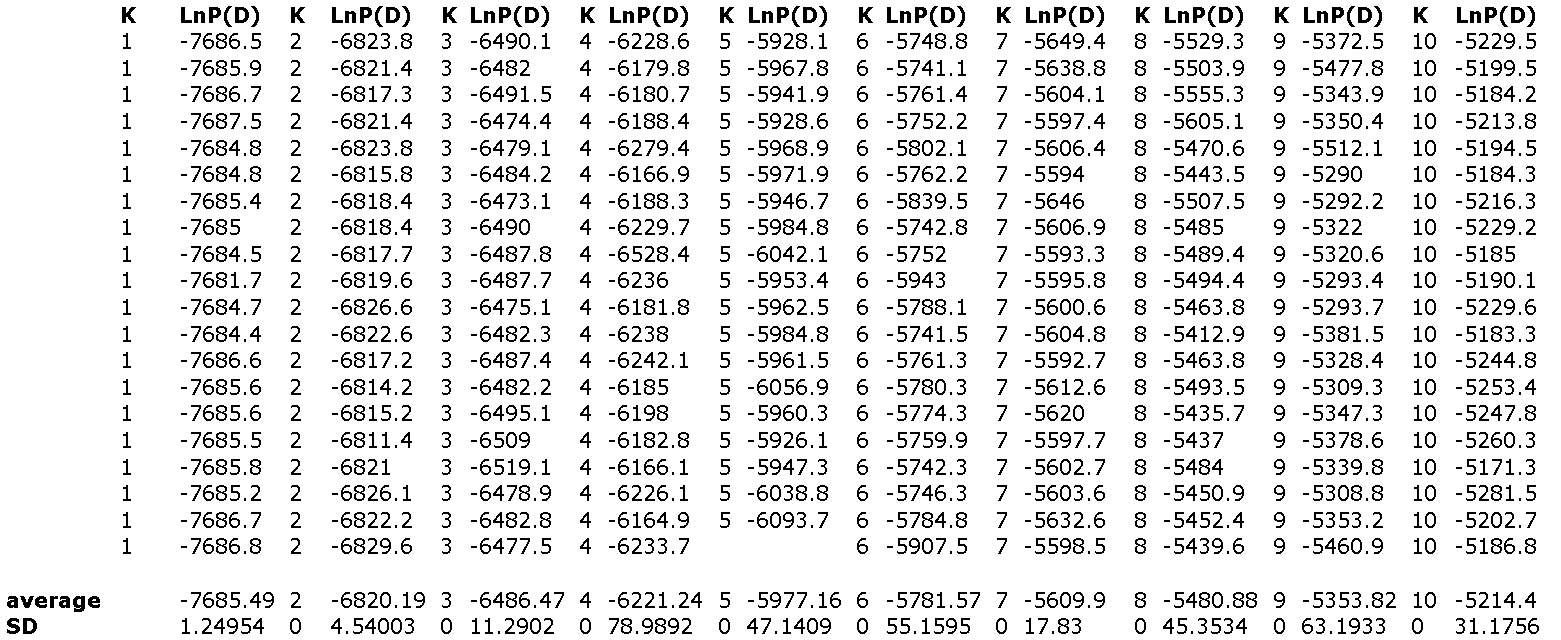
1. Generate STRUCTURE input file using PDGSpider
2. Edite
3. Go to FILE and select NEW PROJECT
4. Name the project eg. Zambia
5. Select the directory where the STRUCTURE input file is kept
6. Select the actual STRUCTURE input file
7. Press NEXT
8. Click on SHOW DATA FORMAT and use this information to fill out rest of project wizard eg. SHOW DATA FORMAT shows:

414 lines with 13

1 lines with 11 columns

This means there are 207 individuals (file in diploid format so 414/2 = 207) and there are 11 loci. Fill in this data in Project Wizard window

1. The missing data value is “-9” for STRUCTURE
2. Press NEXT
3. Select ROW OF MARKER NAMES then NEXT
4. Select INDIVIDUAL ID FOR EACH INDIVIDUAL and PUTATIVE POPULATION ORIGIN FOR EACH INDIVIDUAL
5. Press FINISH
6. Press PROCEED
7. Your data is now loaded onto STRUCTURE.
8. Click on PARAMETER SET and select NEW
   1. Under RUN LENGTH tab use Burnin Period of 10,000 and 10,000 MCMC reps. You can use larger numbers but it doesn’t significantly alter the results and just takes longer. The critical parameter in STRUCTURE is the number of runs (should do 20 runs)
   2. Under ANCESTRY MODEL tab use Admixture Model
   3. Under ALLELE FREQUENCY MODEL tab use Allele Frequencies Correlated or Independent. When correlated is selected it can overestimate the number of clusters (K)
   4. Leave the ADVANCED tab at default
   5. Press OK and give parameter set a name eg. Param1
9. Go to PROJECT and select START JOB
   1. Select the parameter set (eg Param1)
   2. Specify the number of clusters (K) to test eg 1 to 20
   3. Specify the number of iterations (runs) – needs to be at least 20
   4. Press START
   5. Leave STRUCTURE to run overnight or in background if you’re in the lab. The time it takes to run STRUTURE depends on the number of samples and loci and number of K being tested. Typically STRUCTURE takes a few hours for microsats but longer for wgs data. It uses a lot of memory so it’s easiest to just run overnight.
10. When STRUCTURE has finished running a new window will appear saying “JOB finished”
11. **Click** on the SIMULATION SUMMARY on left hand side. A table displaying the STRUCTURE results will be displayed. Go to FILE and SAVE AS TEXT. Save this file as something like “LnP(D) raw data”. There is an issue with STRUCTURE in that it does not save the runs so if you don’t save this simulation summary table before closing STRUCTURE you won’t have the data need to plot the LnP(D) vs K graph.
12. Open the saved simulation summary text file in Excel and use text import wizard to save in Excel format. Arrange data into the following format and calculate the average LnP(D) for each value of K and the SD for each value of K.

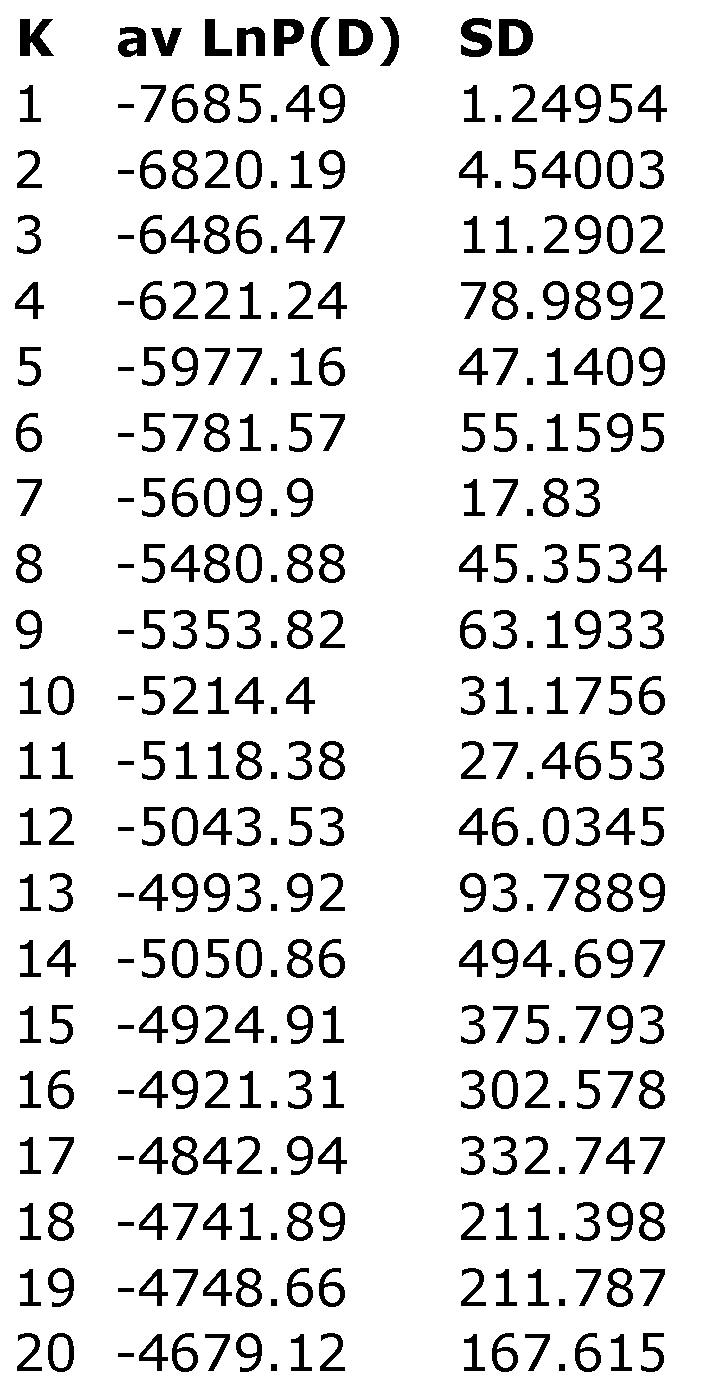


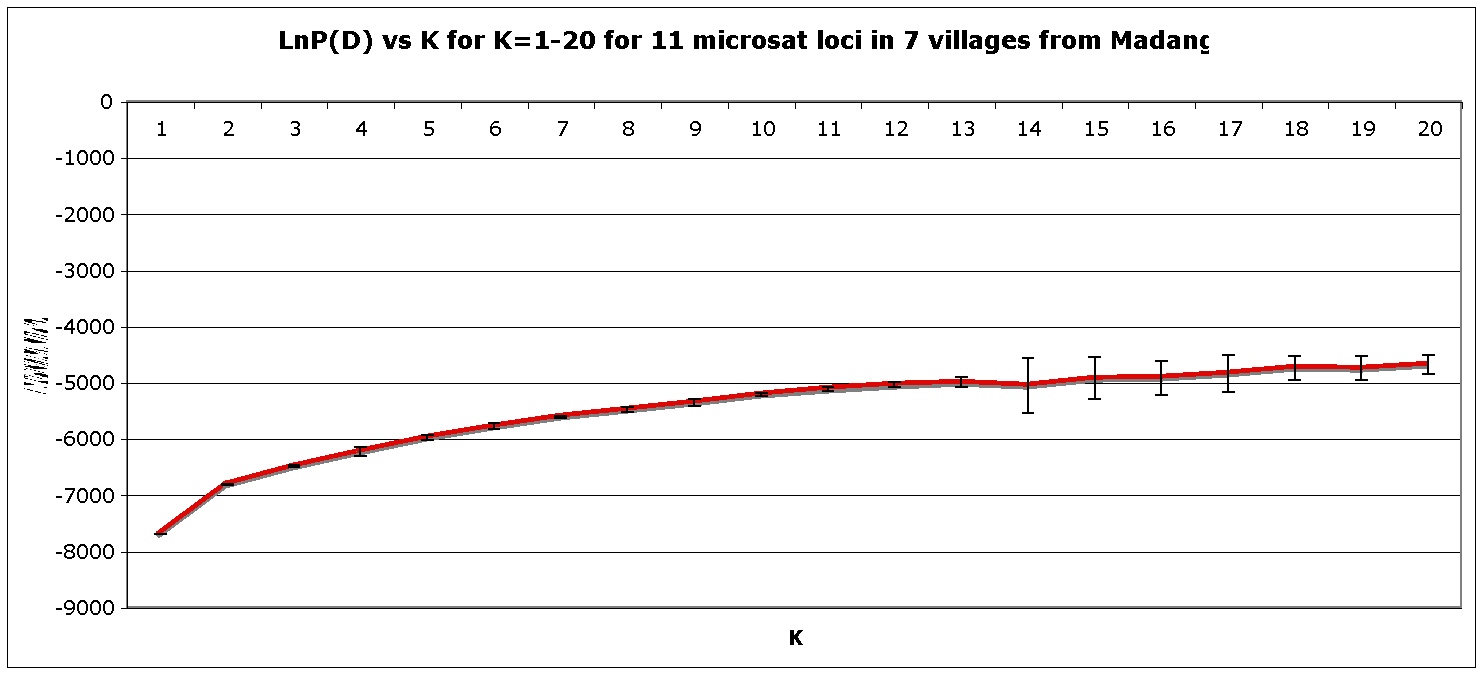
**COMMENT BY AB**

**From this data calculate deltaK as described in Evanno et al. 2005 (HIGHLY RECOMMENDED READING). The highest peak for deltaK is the best estimate of K, given the data.**

**There may also be heirachical migration schemes where strong population structure obscures weaker population structure. In this situation the uppermost hierarchical level is detected. Subsequent analysis should then be conducted with subsets of the data.**

**Also see Alyssa for explanation.**

1. Cut and paste the values for the average and SD for LnP(D) for each value of K into the following format:
2. Plot K vs average LnP(D) to determine optimal K



1. The optimal K occurs when the plot starts to plateau or decrease and the standard deviation increases. In this plot it is difficult to tell so one needs to look at the raw ancestry coefficient plots.
2. To plot the raw data of ancestry coefficients open the STRUCTURE results folder. In this case we’d called the STRUCTURE runs “Zambia”. Go to the directory where the “Zambia” STRUCTURE folder is. Open this folder and then open the Parameter name folder (eg Param1) and then open RESULTS folder. This folder contains results text files for each run. Randomly select a run for each value of K to plot i.e. If you choose to run STRUCTURE 20 times for each value of K from 1 to 20 then the following applies:

K=1, runs 1-20

K=2, runs 21-40

K=3, runs 41-60

K=4, runs 61-80 etc.

Randomly select a run for each a value of K, e.g run 30 for K=2, and then copy and paste the inferred ancestry of individuals for that run into an Excel spreadsheet:

FORMAT of RESULTS text file for run 30, K=2 for 20 individuals.

Inferred ancestry of individuals:

Label (%Miss) Pop: Inferred clusters

1 1353 (9) 1 : 0.989 0.011

2 1357 (9) 1 : 0.974 0.026

3 1380 (18) 1 : 0.010 0.990

4 1381 (27) 1 : 0.975 0.025

5 1386 (18) 1 : 0.989 0.011

6 1356 (18) 1 : 0.180 0.820

7 682 (9) 1 : 0.079 0.921

8 400 (9) 1 : 0.956 0.044

9 421 (9) 1 : 0.986 0.014

10 430 (0) 1 : 0.838 0.162

11 433 (9) 1 : 0.965 0.035

12 440 (18) 1 : 0.990 0.010

13 441 (9) 1 : 0.687 0.313

14 455 (18) 1 : 0.785 0.215

15 558 (9) 1 : 0.972 0.028

16 559 (9) 1 : 0.987 0.013

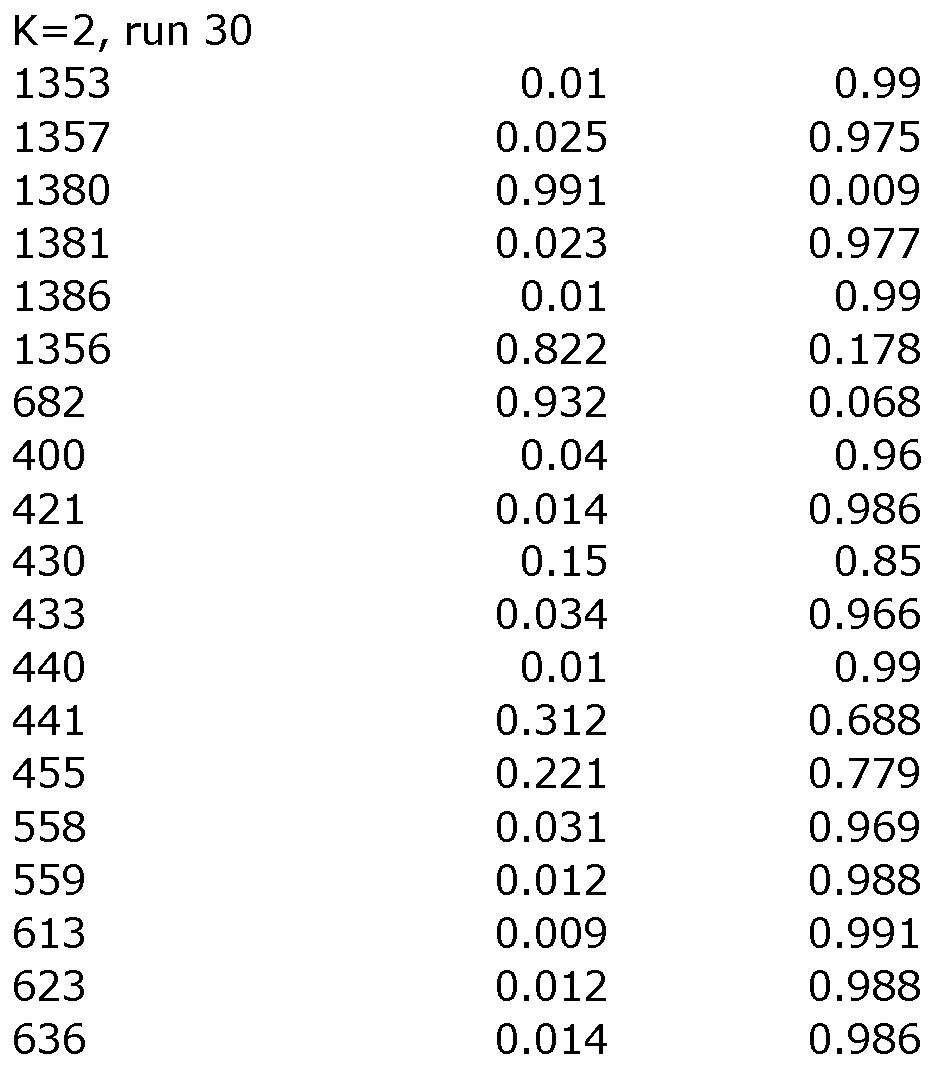
17 613 (9) 1 : 0.990 0.010

18 623 (18) 1 : 0.987 0.013

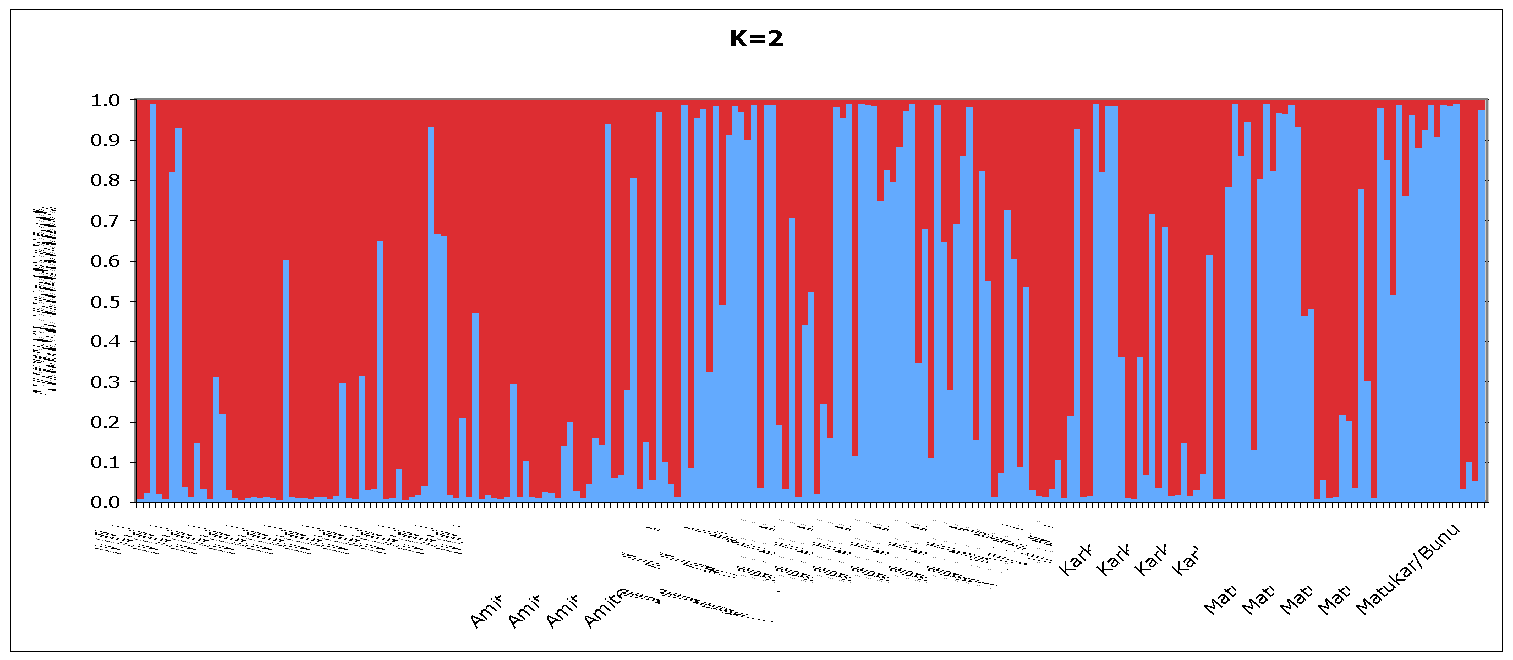
19 636 (9) 1 : 0.985 0.015

20 1325 (9) 1 : 0.988 0.012

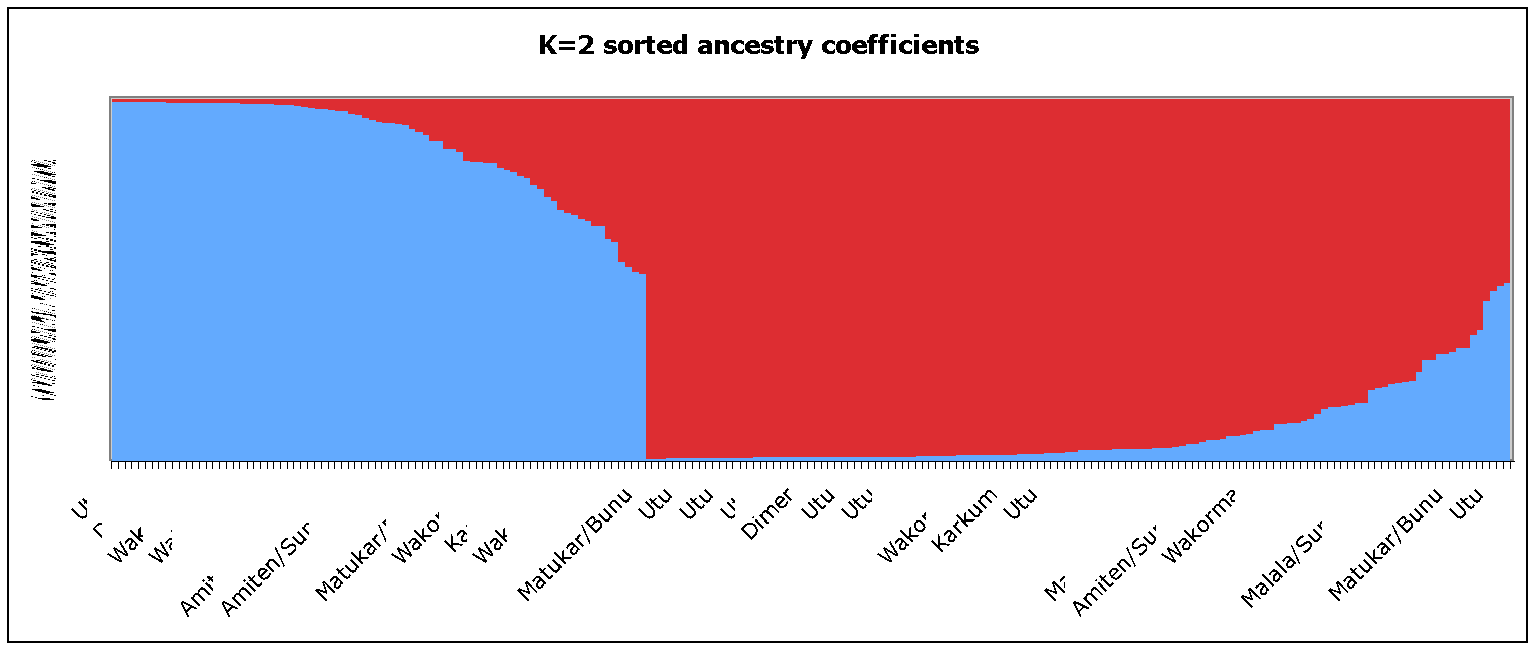
1. Delete the unwanted columns to produce format as follows:



1. Repeat steps 23 and 24 for each value of K you wish to plot
2. You can then plot the raw data for each value of K eg. K=2



1. Using the SORT DATA option in Excel, sort each ancestry coefficient column from largest to smallest and then plot the raw data:



1. Calculate the average ancestry coefficients for each village/population and plot to see if the different clusters are geographically structured:

