Stephen (Alex) Townsend
ISC5935 – Practical Genetic Inference
10/16/2015
Report on Structure

I. Introduction

Structure is a very commonly used software package used in population genetics to study the structure of genetic data samples and to study differences between population [1]. The basic idea behind the program is the Markov Chain Monte Carlo with Bayesian Inference to determine genetic difference between individuals [1]. Individual differences are then clustered into distinct populations in a way that very much resembles k-Means Clustering [1]. The program's input data is in the form of a data file with population information and locus information for n individuals and m loci where n and m need not be equal [2]. It is often used with diploid data, such as microsatellites, but can also be used with haploid data.

This report details the use of Structure on a set of simulated tutorial data where the true number of populations was known beforehand. The goal was to see if Structure could infer the correct number of populations from the simulated test data.

II. Methods and System Details

Before the beginning of the experiment, a data set was generated by the instructor who wrote the tutorial in which it was known that there should only have been 4 populations (See Figure 1). Once the data set was downloaded from the instructor's website along with the rest of the tutorial, we proceeded to load the loci data into structure. This data had been converted into a Structure-readable format prior to the beginning of this experiment using a software system called PGDSpider, which is a freely available, though closed-source file converter software [5].

Figure 1: Simulated Data Set

Population	# of Individuals
1	40
2	20
3	10
4	5

Once the data was loaded into a new project folder created by the Structure program, the parameter set was created. This parameter set included the following parameters: 75 individuals, 10 loci, 1 assumed population, a burn-in period of 1000 and 10000 repetitions. With the data and parameter set now generated, the Structure program was run on all of the data and created an output file which was then uploaded into the Structure Harvester program. Structure Harvester is a free to use, web-based program distributed by the Taylor Lab of UCLA which is designed to take the data from Structure and create plots from it such that determining the number of inferred populations is easier and more human-friendly [2]. Structure Harvester was used to create figure 1, which makes the process of determining the number of populations that Structure inferred much simpler since Structure will not simply give an output saying that the inference has determined that there are N populations [1].

Once the output data had been run through Structure Harvester, the output data from Structure Harvester was run through the CLUMPP program, which is designed to take averages of the replicates from Structure Harvester program's output data and is used to generate a file readable by the DISTRUCT program [3]. CLUMPP is distributed by the Rosenberg Lab of Stanford University. The program will take data and permute the replicates until they have been optimally aligned [3].

Lastly, the output data from CLUMPP was run through the DISTRUCT1.1 Program which is also put out by Rosenberg Lab of Stanford University. The DISTRUCT program is designed to create a graph of the CLUMPP output or Structure Harvester output [4]. In order to use Structure Harvester output directly, only one replicate could be used [2]. Though, since CLUMPP was used in this case, the CLUMPP data could be used directly and was [2]. This program's graphics were used to create figure 2 in the results section. The program makes figuring out the true number of populations even easier than Structure Harvester and also can show allele admixture and population interchange [4].

III. Results

Figure 2: Maximum Likelihood Graph from Structure Harvester

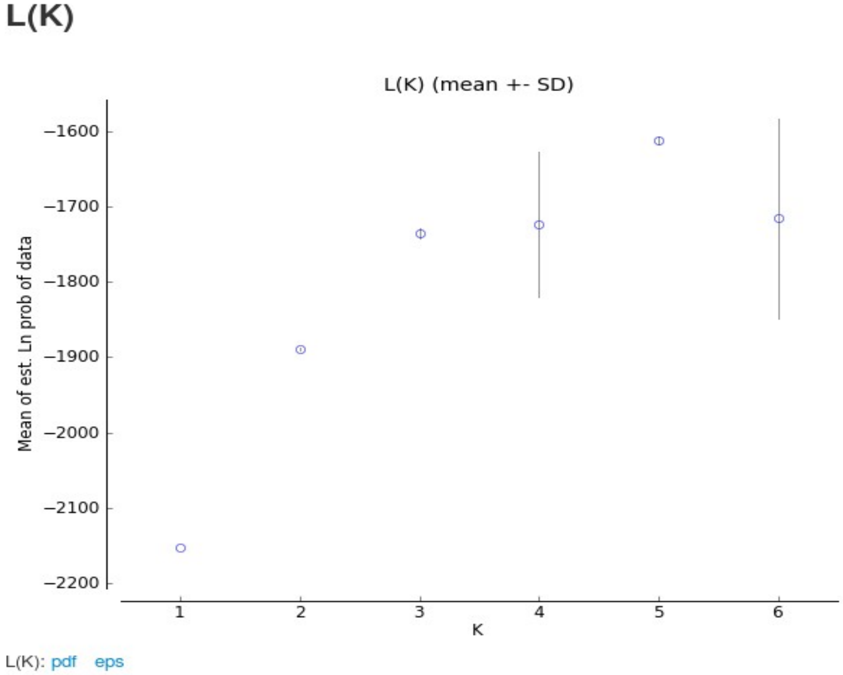
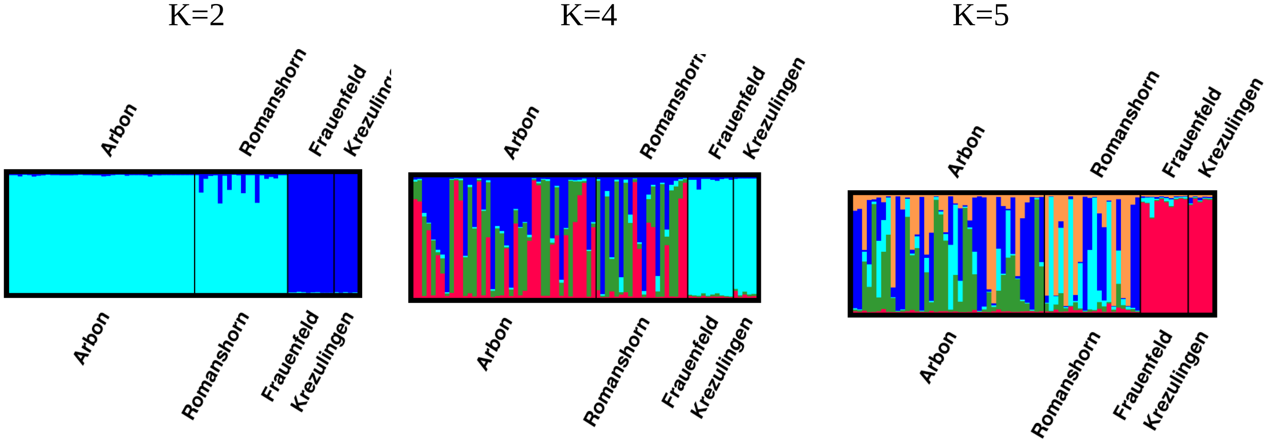


Figure 2 shows the results of running Structure Harvester in the form of a maximum likelihood graph. From this, we can see that there is most likely to be 5 populations though the error bars on the point at K=6 indicate that there COULD be 6 populations since K=5 falls inside the K=6 error bars. The point at K=4 could also confound the results since K=5 lies at the extreme of the error bounds for K=4. Thus, K=5 seems the more reliable due to it's high likelihood but small error bars.

Figure 3: Population Inference Graph from DISTRUCT for K=2, 4 & 5 Populations


K=2

K=4

K=5

Figure 3 shows the genomic structure of the samples at the 4 locations if there are 2, 4 or 5 populations present. Each bar on the graph indicates an individual within the data set. The colored areas on each bar indicate which part of that individual's genome would be assigned to a different population. The graph of the K=2 data shows that the relative proportions of the genes of the light blue colored Arbon and Romanshorn population that would get assigned to the dark blue population is the same across all three graphs. However, the K=4 and K=5 graphs show that the Arbon and Romanshorn populations are very highly subdivided with large proportions of each individual getting assigned to different populations.

IV. Discussion

The results above are quite interesting as the show what may be interpreted as two different answers. Structure, as shown through the graph by Structure Harvester, seems to believe that there are actually 5 populations, not 4, as we know the true answer should have been. Similarly, when we look at the graphs generated by DISTRUCT from the CLUMPP and Harvester data, we see that it clearly looks like there should only be two populations since 4 looks like there is far too much admixture between the other two locations for them to be distinct and the two locations on the far right seem to constitute only one population. Similarly, when you look at the K=5 graph, the situation, as far as DISTRUCT is concerned, only gets worse if you assume there are 5 populations. These programs seem to be convinced that there are only two populations.

V. Conclusion

To conclude, it is quite clear that both answers cannot be correct. In fact, based on the fact that this is trial data, we know that neither are correct, though it could be argued that Structure Harvester gives an answer that is closer to the correct one than DISTRUCT does. We know that there should be 4 populations due to the fact that our dataset is simulated and the true number of populations was known. There very likely has been a lot of gene flow between Arbon and Romanshorn recently and that is making their population structure more complex and interwoven and thus preventing them from being distinguished by the program. There could be several reasons for this. The most likely of which is simply sample size. Were the sample size hundreds of individuals from each population instead of less than 100 for all four combined, the populations may have been much more clearly distinct and much easier for the programs to separate out. In the end though, all three programs are very useful for population genetics and provide an invaluable resource for working with and visualizing large data sets and drawing reasonable conclusions from them. It would be a good idea to investigate this same dataset further with another method that may be sensitive to more complex population structures.

VI. Works Cited

1. Beerli, Peter, Ph.D. "Peter Beerli's Classes." *Structure Theory*. Peterbeerli.com, 9 Sept. 2015. Web. 16 Sept. 2015. <http://www.peterbeerli.com/classes/index.php?title=StructureTheory>.

2. Beerli, Peter, Ph.D. "Peter Beerli's Classes." *Structure Tutorial*. Peterbeerli.com, 9 Sept. 2015. Web. 16 Sept. 2015. <http://www.peterbeerli.com/classes/index.php?title=StructureTutorial>.

3. Rosenberg, Noah, Ph.D. "CLUMPP." *CLUMPP: CLUster Maching and Permutation Program*. Stanford University, 15 Oct. 2015. Web. 16 Sept. 2015. <https://rosenberglab.stanford.edu/clumpp.html>.

4. Rosenberg, Noah, Ph.D. "Distruct: Graphical Display of Population Structure." *Distruct: Graphical Display of Population Structure*. Stanford University, 15 Oct. 2014. Web. 16 Sept. 2015. <https://rosenberglab.stanford.edu/distruct.html>.

5. Beerli, Peter, Ph.D. "Structure Tutorial." ISC5935: Practical Genetic Inference. FSU Department of Scientific Computing, Tallahassee. 9 Sept. 2015. Lecture.