STRUCTURE and Problem #2

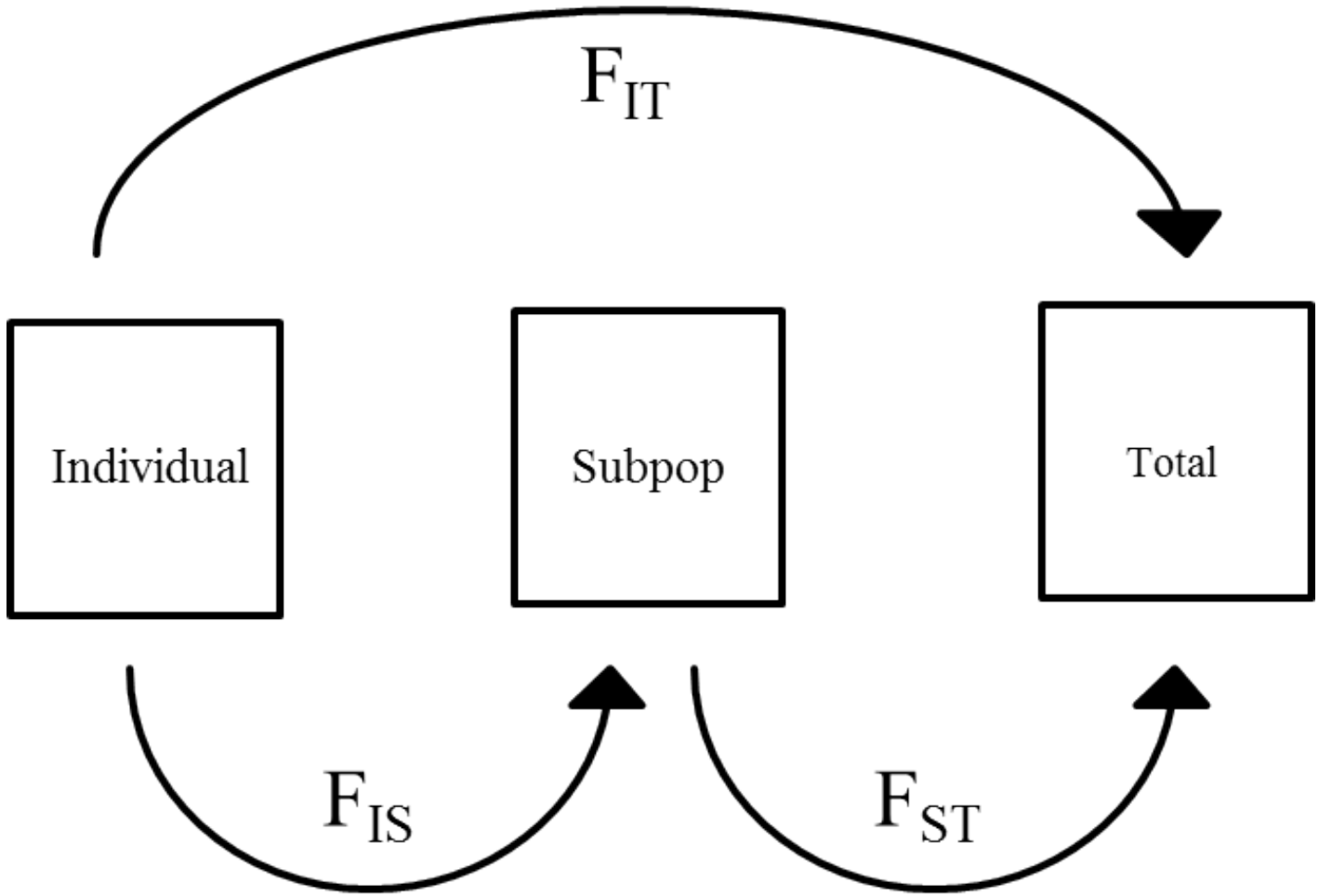
Nora Mitchell

February 7, 2017

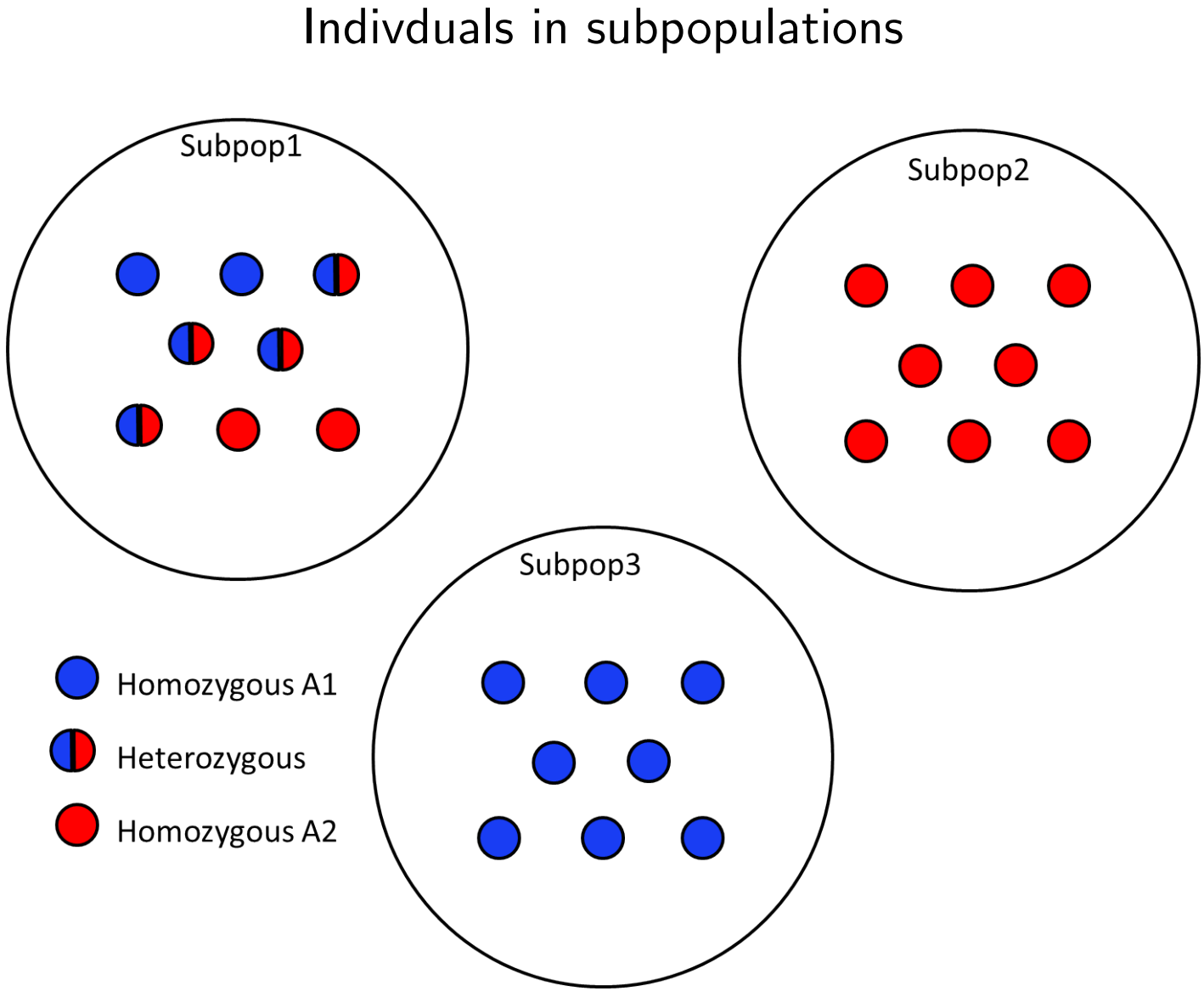
Goals for Today's Lab

- Review F-statistics conceptually
- Install and learn how to use STRUCTURE
- Introduce Problem #2

Hierarchical F-statistics

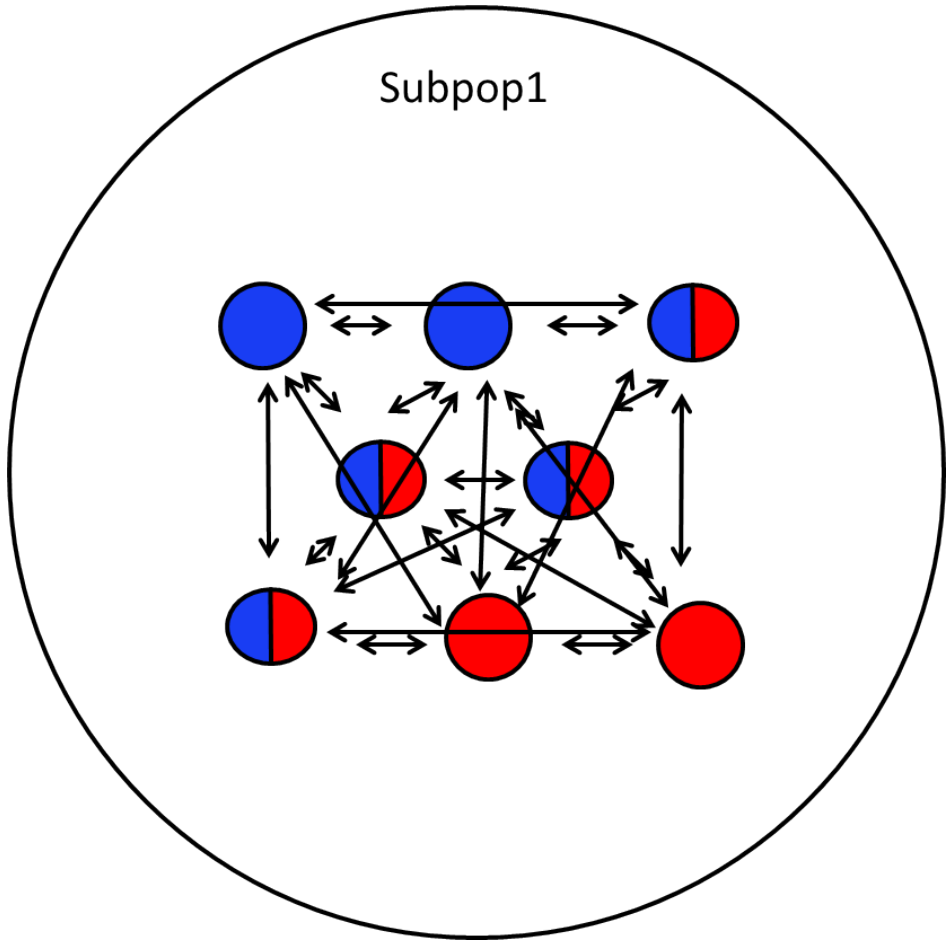


Toy Example


Indivduals in subpopulations
Subpop1
Subpop2
Subpop3
Homozygous A1
Heterozygous
Homozygous A2

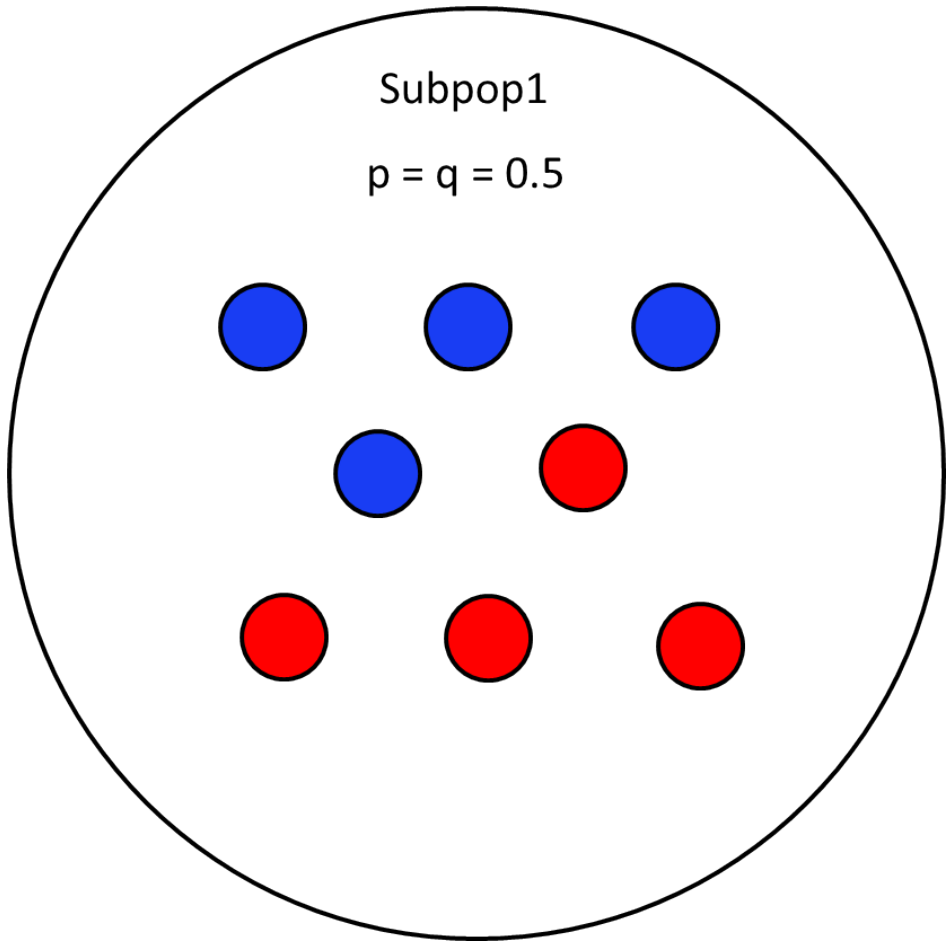
Fis

Fis is the variation of individuals within subpopulations (f, inbreeding)
Fis is a measure of departure from H-W



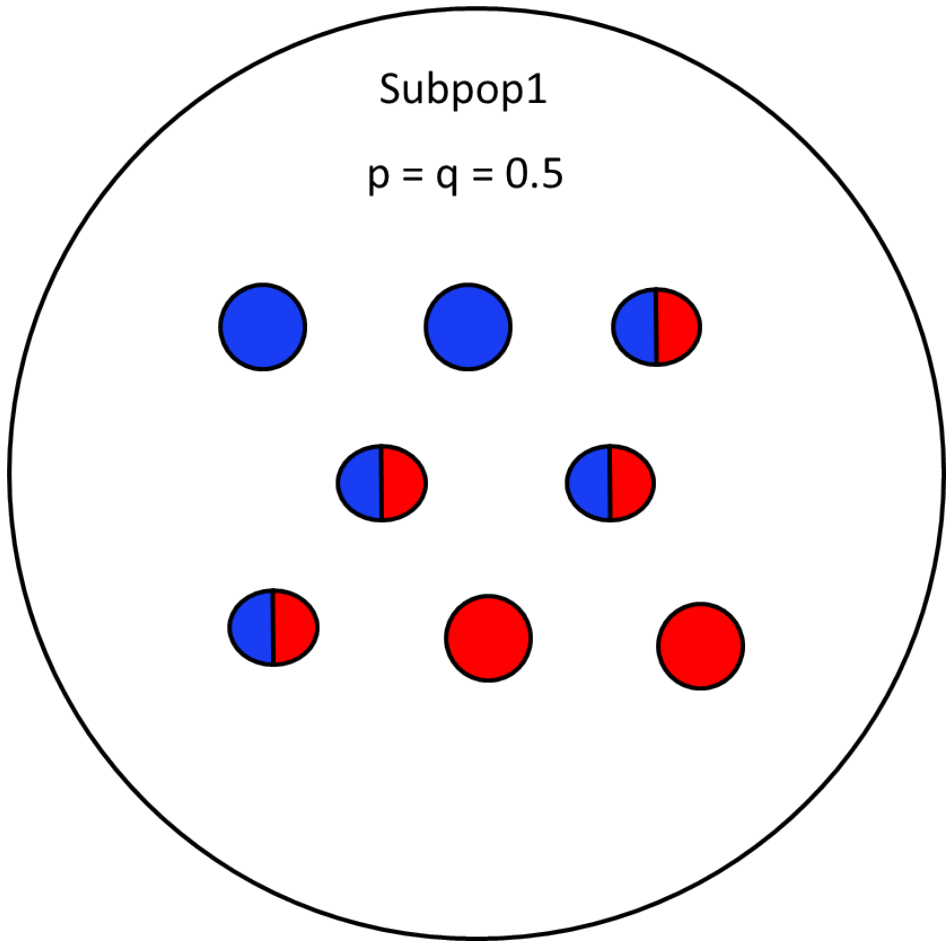
Socrative

Is this figure an example of high or low inbreeding (f, Fis)?



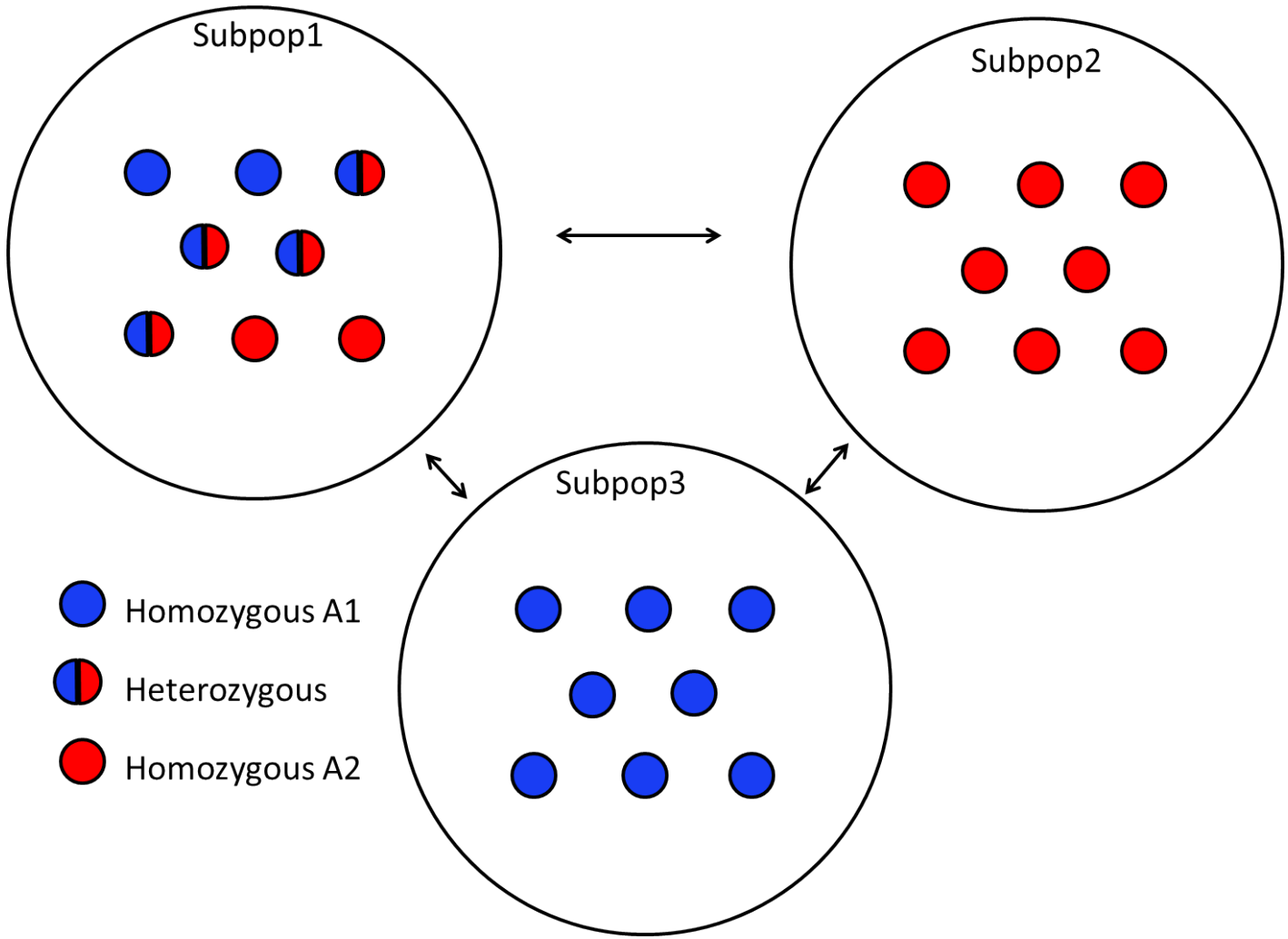
Fis

Is this figure an example of high or low inbreeding (f, Fis)?



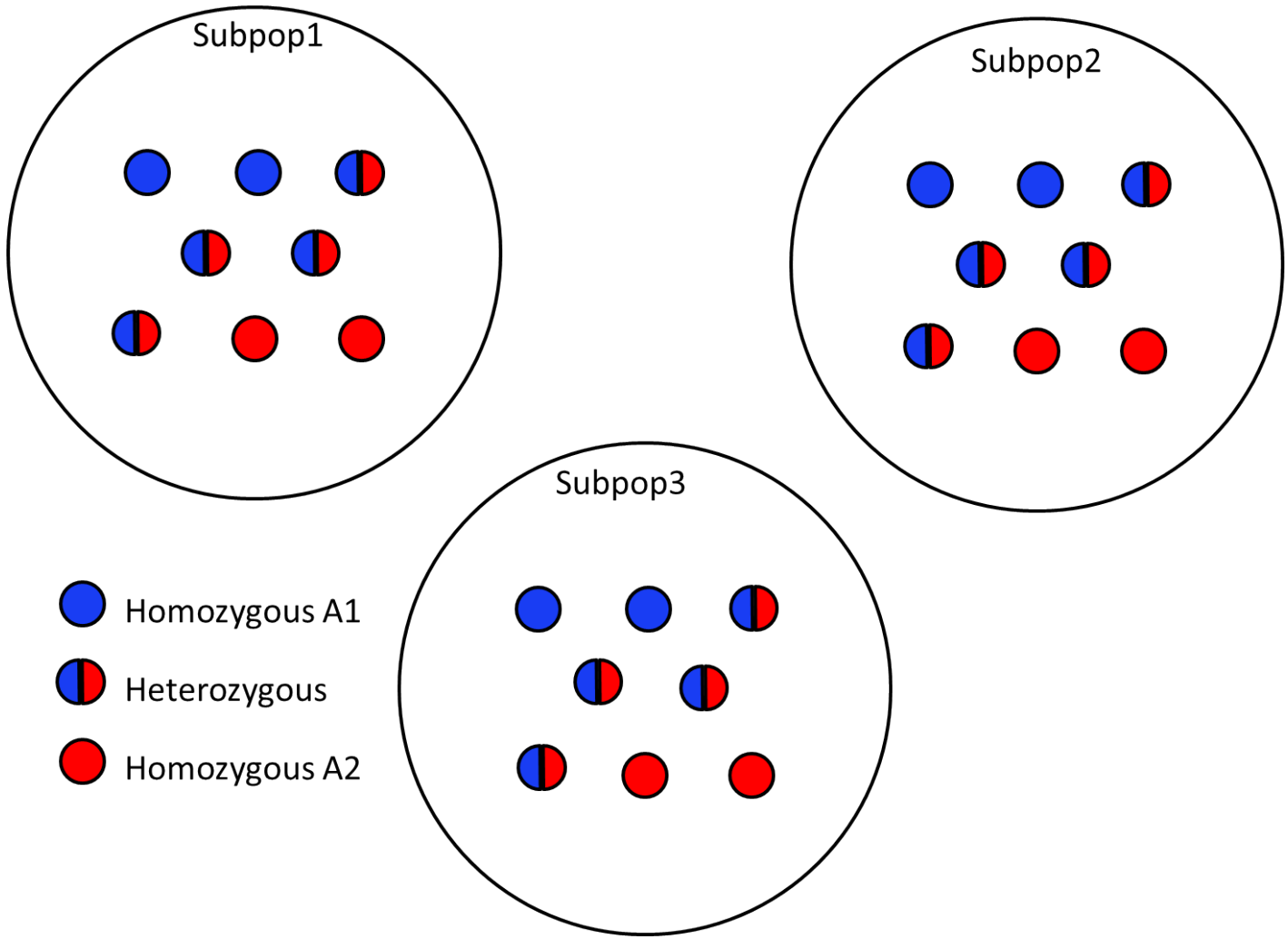
Fst

Fst is the variation among subpopulations within the total (θ)



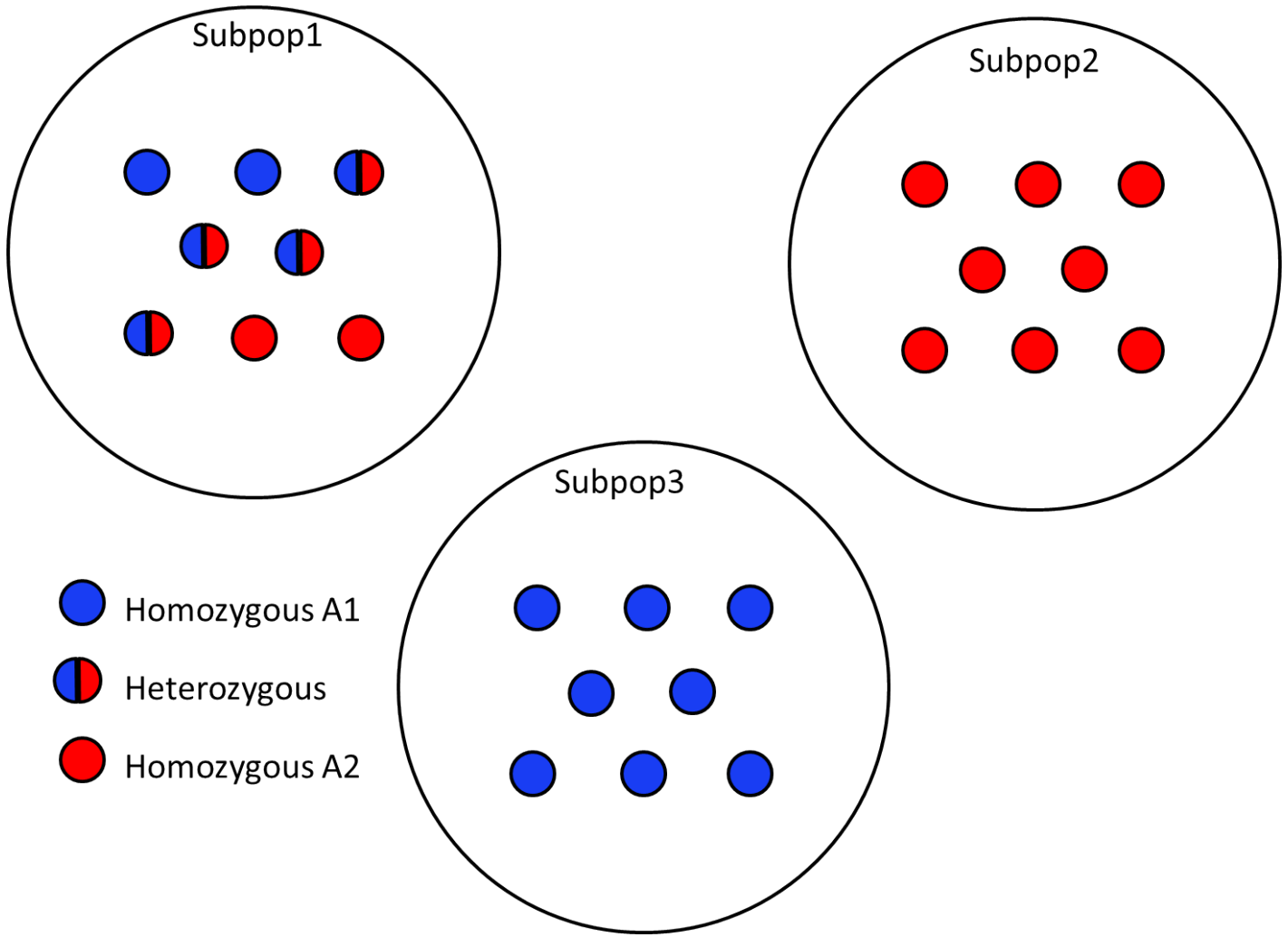
Socrative

Is this figure an example of high or low population differentiation (θ, Fst)?



Fst

Is this figure an example of high or low population differentiation (θ, Fst)?



Human Example

Rosenberg et al. (2002, Science) looked at diversity in humans!

- 377 autosomal microsatellite loci
- 1056 individuals
- 52 populations
- 8 regions

Human Example

AMOVA (to look at variance components, slightly different, analogous to F-statistics)

Where is most of the variation?

Sample	Re-gions	Pops	Within pops	Among pops within regions	Among regions
World	5	52	93.2	2.5	4.3
Africa	1	6	96.9	3.1	
Eurasia	3	21	98.3	1.2	0.5
Europe	1	8	99.3	0.7	
Middle East	1	4	98.6	1.3	
Central/South America	1	9	98.6	1.4	
East Asia	1	17	98.7	1.3	
Oceania	1	2	93.6	6.4	
America	1	5	88.4	11.6	

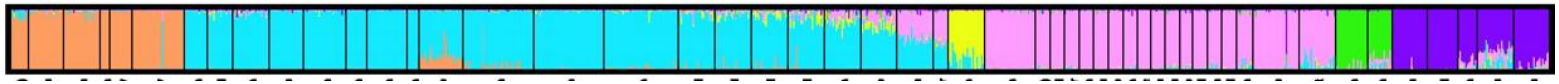
From Rosenberg et al. (2002)

Individual Assignment

How many distinct groups are there?

What groups do *individuals* belong to?

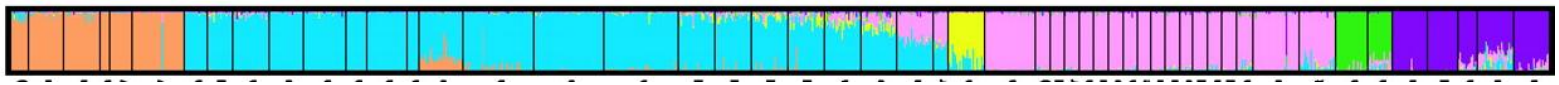
STRUCTURE



STRUCTURE is a free software package from Pritchard et al. (2000)

- Uses multi-locus genotype data to investigate population structure
- Assigns individuals to "K" number of clusters
- Can be used to identify distinct populations, hybrids, migrants, etc.
- Can use different genetic markers (microsats, SNPs, RFLPs, AFLPs)
- Takes an MCMC approach
- `http://pritchardlab.stanford.edu/structure.html`

STRUCTURE

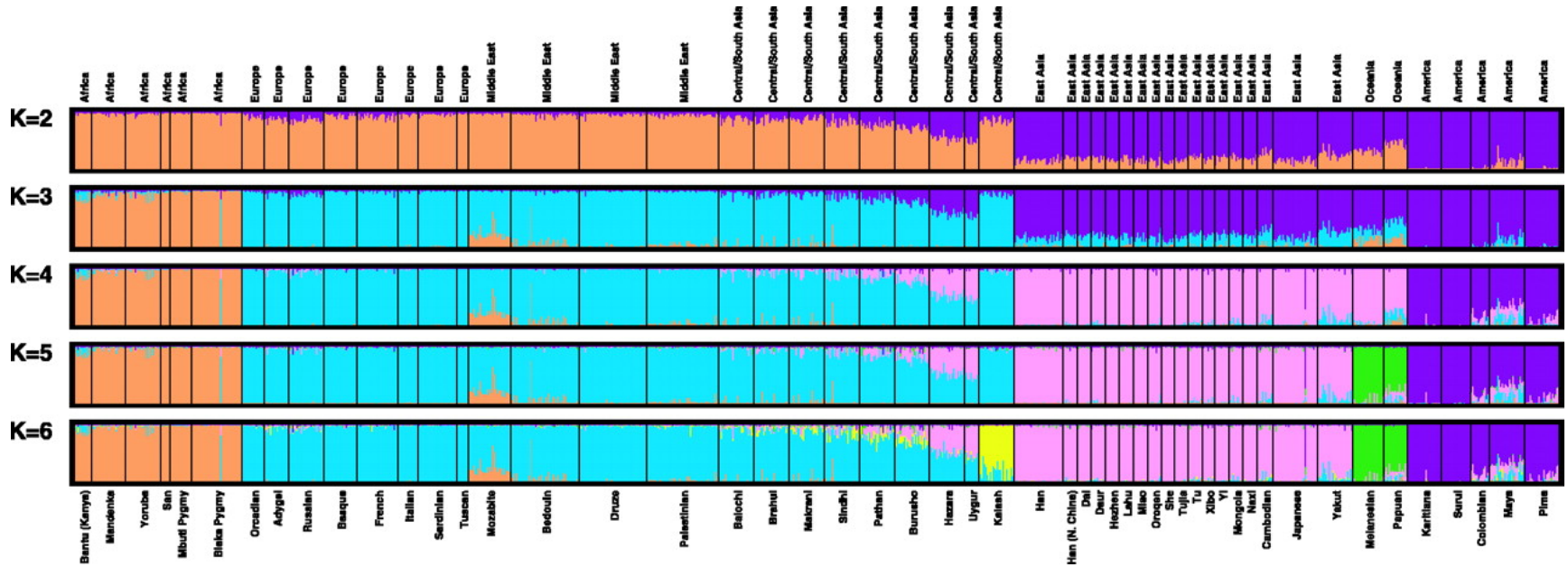


Install Structure now and we will walk through an example of how to use it!

`http://pritchardlab.stanford.edu/structure_software/release_versions/v2.3.4/html/structure.html`

If you are having trouble on Mac, see pages 4 & 5 of project 2

Human Example



From Rosenberg et al. (2002)

Interpreting Structure Output

After you run Structure for K=1:N, there are two ways to choose the "right K"

1. Look at DeltaK output from Structure Harvester (measures rate of change of probability density of data given that K-value)
   Choose highest DeltaK
2. Look at the mean log posterior probability of the data LnP(D), also known as L(K)
   Choose a value where this seems to level off

There may be more than one "correct" answer regarding the K chosen! Justify your choice!

Structure Harvester

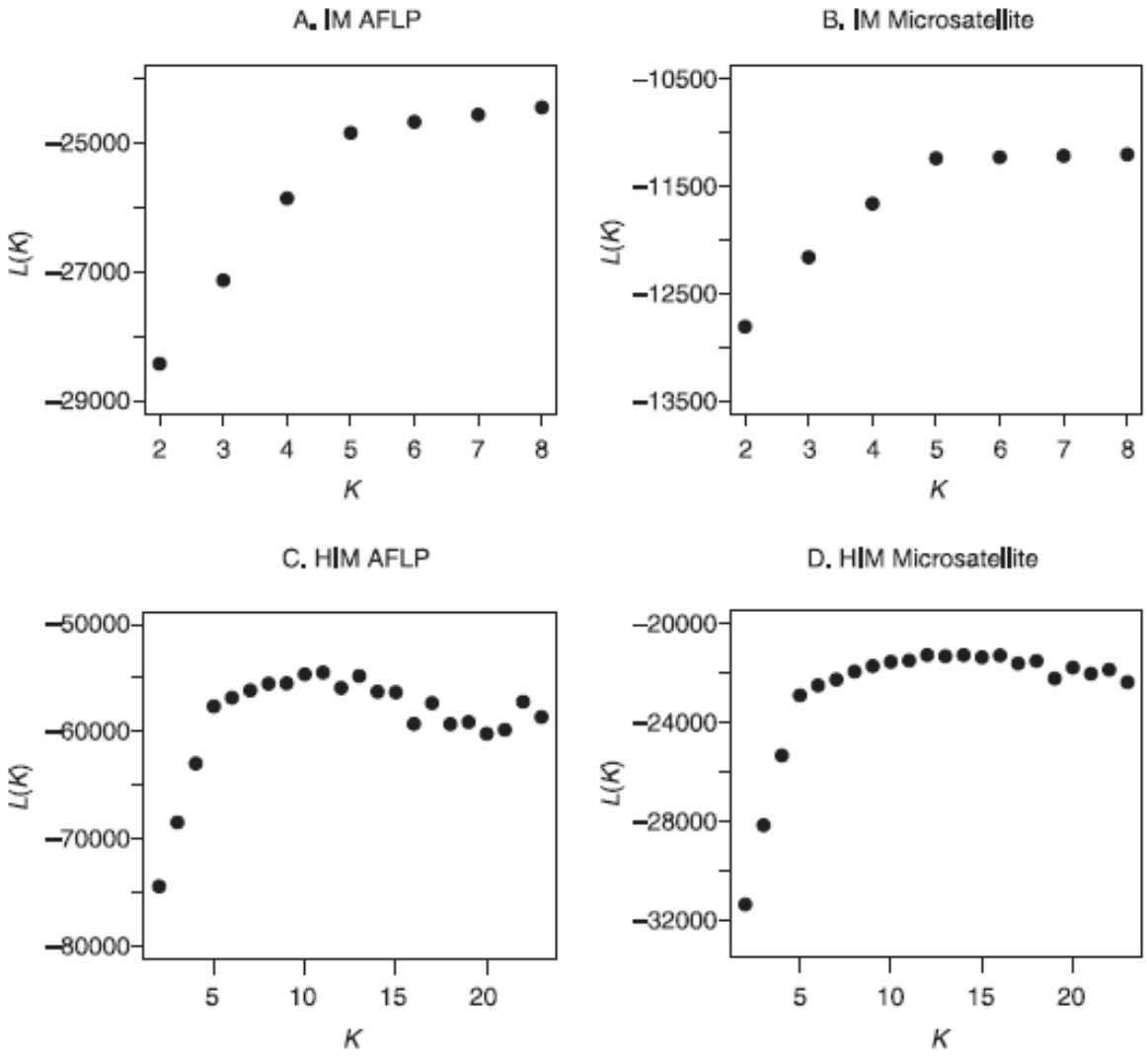
Structure Harvester is a web-based program that takes the output from multiple runs of structure (in zip file format) to calculate DeltaK from Evanno et al.

DeltaK is a measure of the rate of chage in the log probability of the data betwee successive K values

`http://taylor0.biology.ucla.edu/structureHarvester/`

Structure Harvester

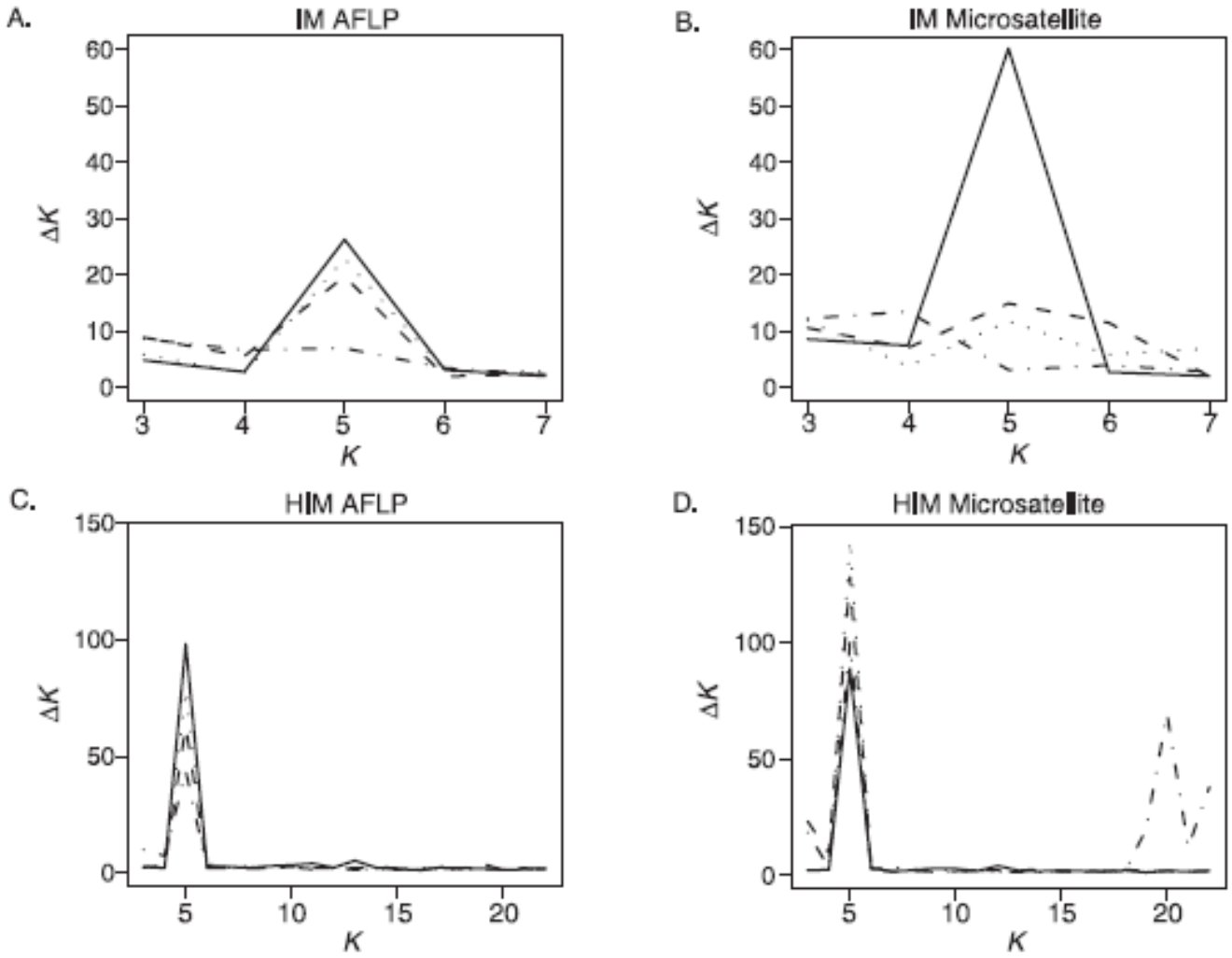
LnK



From Evanno et al. (2005)

Structure Harvester

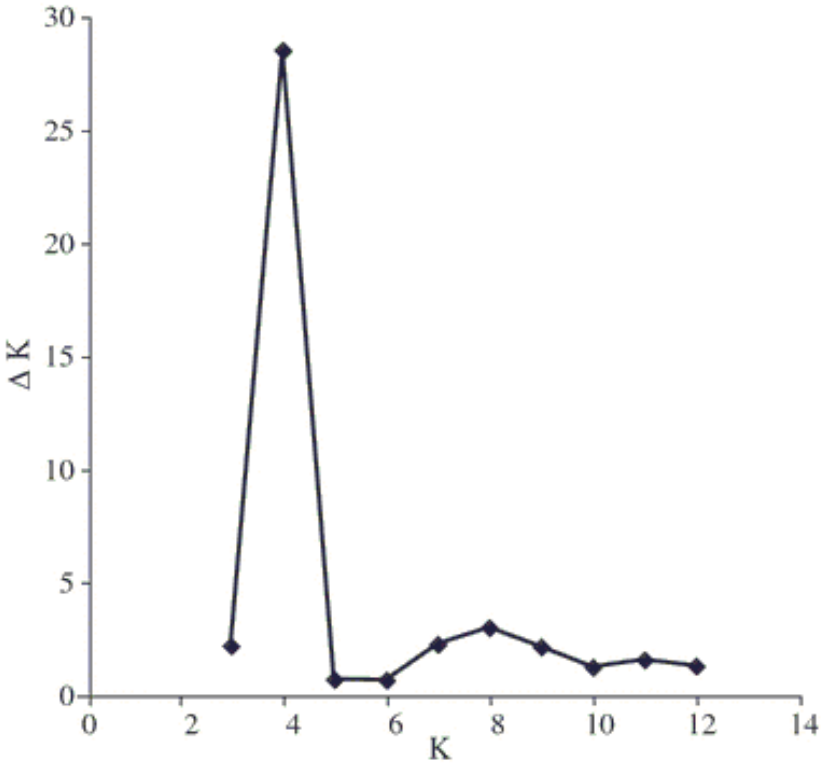
DeltaK

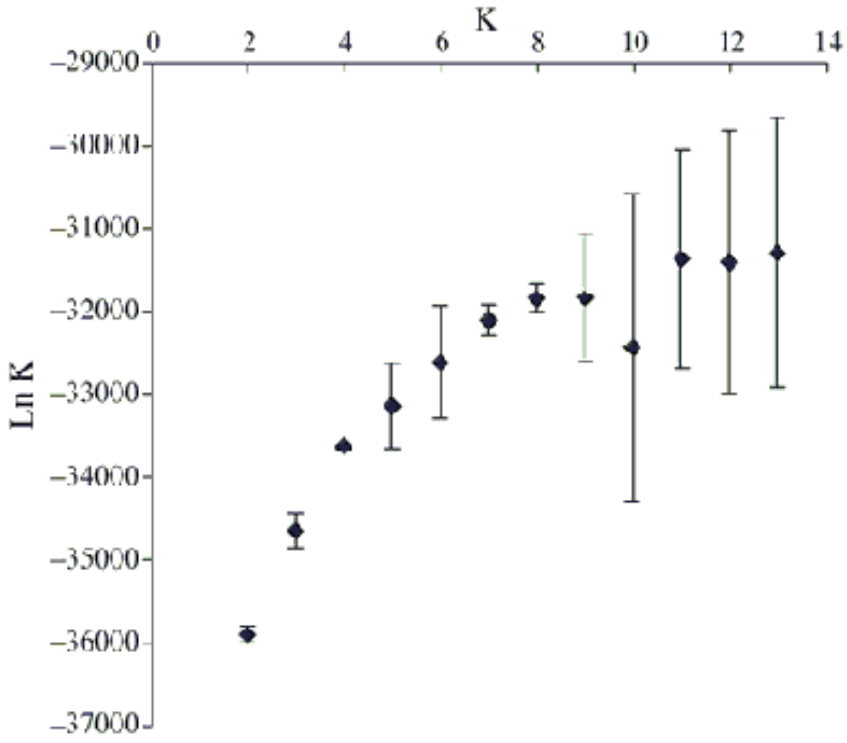


From Evanno et al. (2005)

Socrative

What is a reasonable estimate of K given these plots?

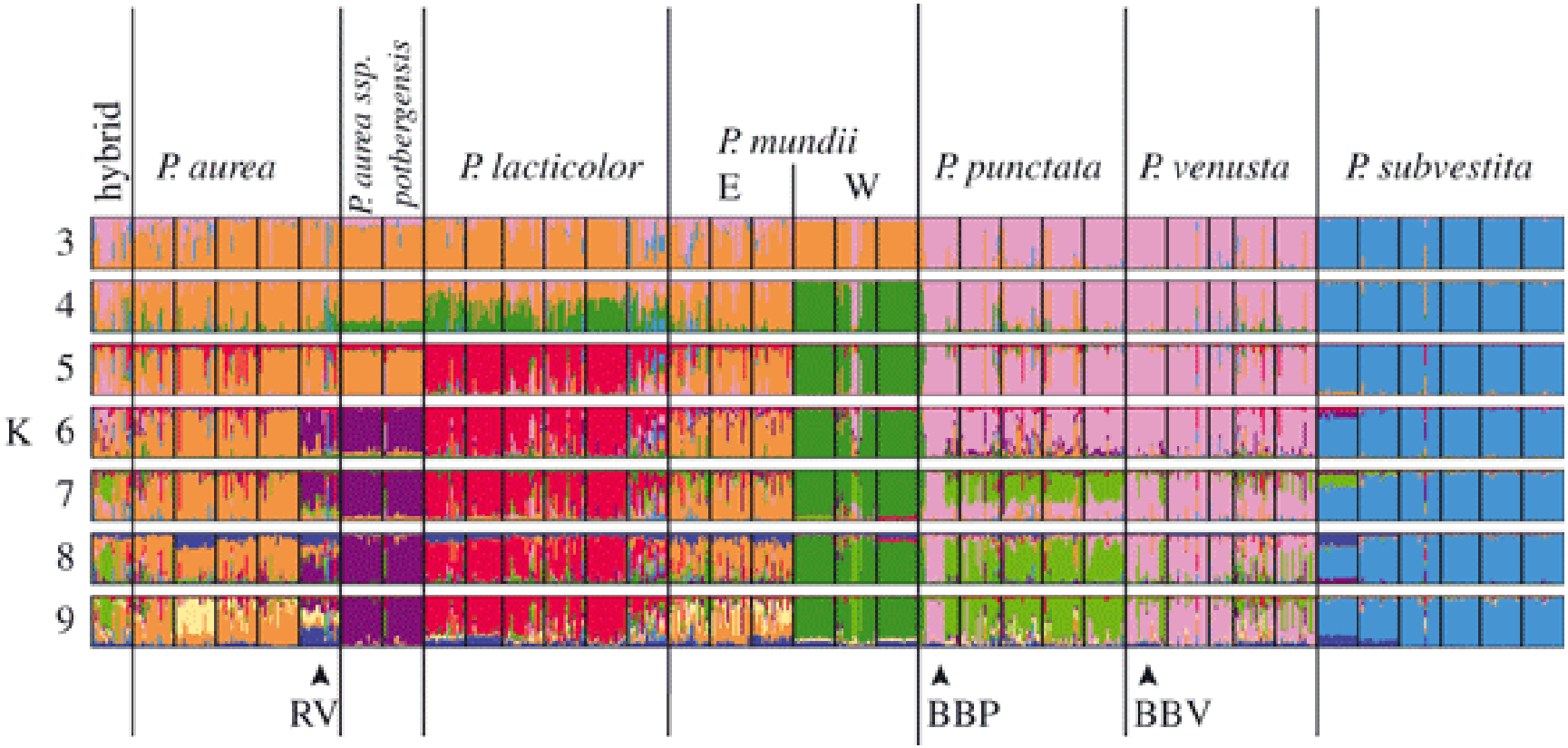




From Prunier and Holsinger (2010)

White proteas

Look at Structure barplots for different Ks



From Prunier and Holsinger (2010)

TeraStructure

What about large datasets?

TeraStructure is a shortcut scalable approach for giant datasets: For instance: 10^{12} observed genotypes, 1 million individuals at 1 million SNPs

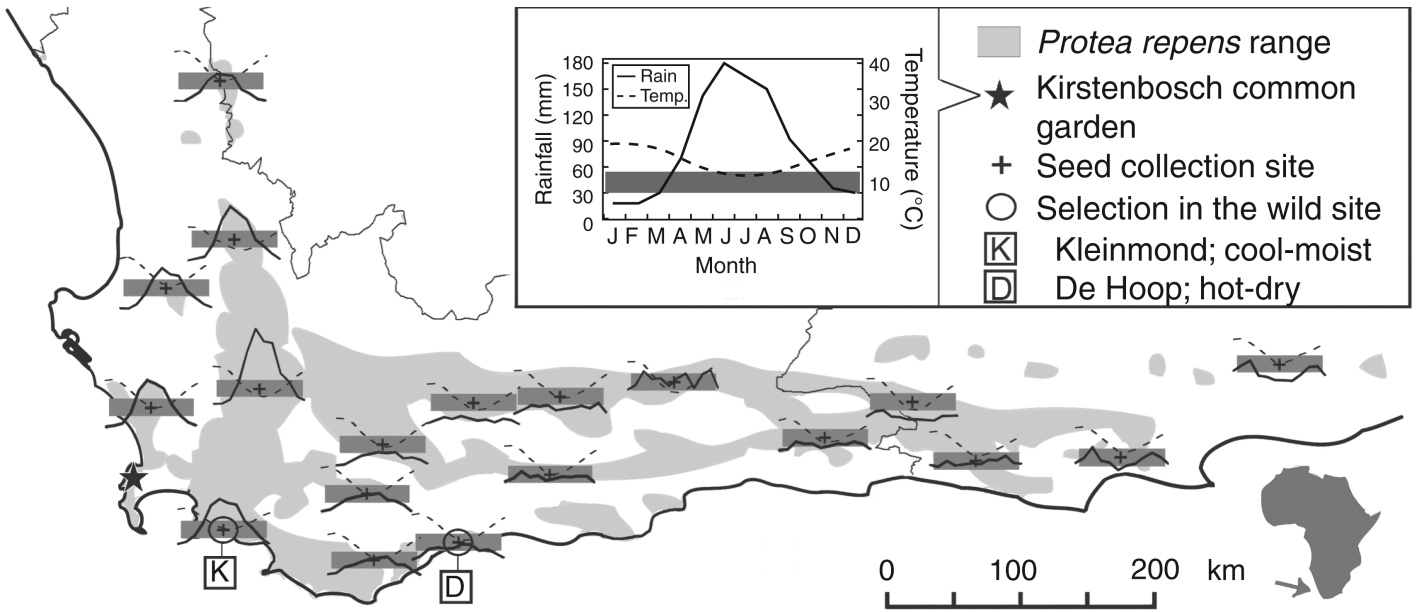
From Gopalan et al. (2016)

Project 2

Protea repens is a widespread South African shrub

Project 2

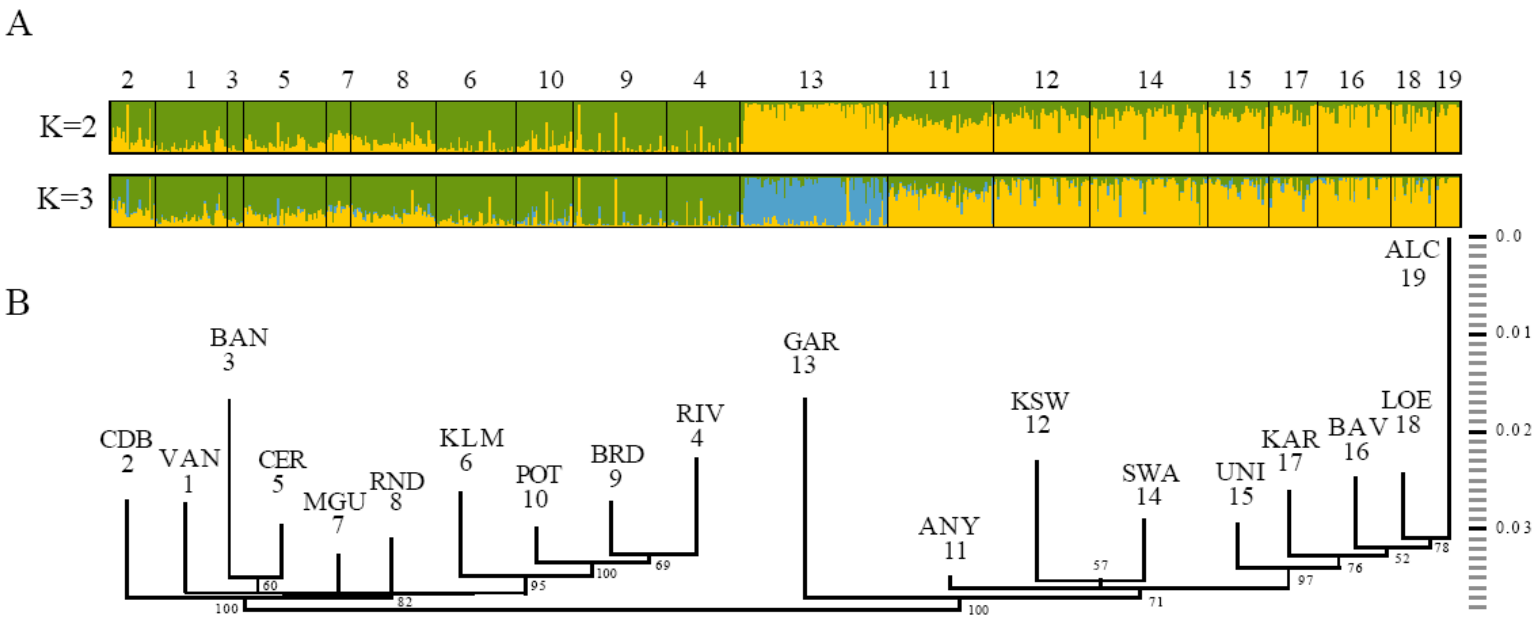
Samples from 19 populations across its range
Originally 2006 polymorphic loci



From Carlson et al. (2015)

Project 2

Samples from 19 populations across its range
Originally 2006 polymorphic loci



Prunier et al. Accepted

Project 2

For this project, analyzing Fst outlier loci

- 662 individuals
- 19 populations
- 173 SNP loci

From Prunier et al. Accepted

Project 2

Questions

- What are estimates of Fis and Fst using Weir and Cockerham's approach?
- What are estimates using Kent's Bayesian approach? How do they compare with the above?
- Is there evidence for inbreeding in *Protea repens*?
- How similar or different is the genetic structure for these loci compared with the publication based on individual assignment?

From Prunier et al. Accepted

Project 2

Methods Hints

- Use adegenet in R to estimate Weir and Cockerham's F-stats.
- Use Kent's code for Bayesian estimates of theta and f. Means and credible intervals!
- Compare models using DIC to see if there is evidence for inbreeding! (Set DIC to TRUE in code!)
- Is a higher or lower DIC indicative of a "better" model?
- In Structure, run for K $= 2$ to K $= 19$. Follow instructions in tutorial.
- Bayesian code and Structure will take a chunk of time to run!

Project 2

Write-up Hints

- What are Fst outliers? Why might they be different? (Outside source...?)
- Write-up shoud include appropriate figures
- Answer questions as if they were main questions/hypotheses in introduction of a paper. Your write-up is a condensed results and/or discussion section.

Project 2

IMPORTANT

Send me zip file with your Structure results for K = 2 to K = 19 by Thursday at midnight!

I will compile class data and run it through Structure Harvester and send you the results!

Write-up due to me via e-mail next Tuesday Feb 13th, 9:30am

Works Cited

- Carlson, J.E., C.A. Adams, and K.E. Holsinger (2015). Intraspecific variation in stomatal traits, leaf traits and physiology reflects adaptation along aridity gradients in a South African shrub. Annals of Botany 117(1); 195-207.
- Dent, A., and vonHoldt, B.M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4(2):359-361.
- Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611-2620.
- Gopalan, P., W. Hao, D.M. Blei, and J.D. Storey. 2016. Scaling probabilistic models of genetic variation to millions of humans. Nature Genetics 48:1587-1590.
- Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.
- Prunier, R., and K. E. Holsinger. 2010. Was it an explosion? Using population genetics to explore the dynamics of a recent radiation within Protea (Proteaceae L.). Molecular Ecology 19(18): 3968-3980.
- Prunier, R., M. Akman, N. Aitken, C. Kremer, A. Chuah, J. Borevitz, and K.E.Holsinger. Accepted. Isolation by distance and isolation by environment contribute to population differentiation in *Protea repens* (Proteaceae L.), a widespread South African species. American Journal of Botany.
- Rosenberg, N.A., J.K. Pritcharad, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. 2002. Genetic Structure of Human Populations. Science 298(5602): 2381-2385.