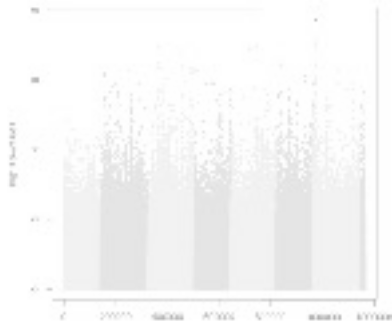


Got Variants? Downstream Analyses for PopGen and Evolution Studies

Jacob B. Landis

School of Integrative Plant Science

Cornell University

and

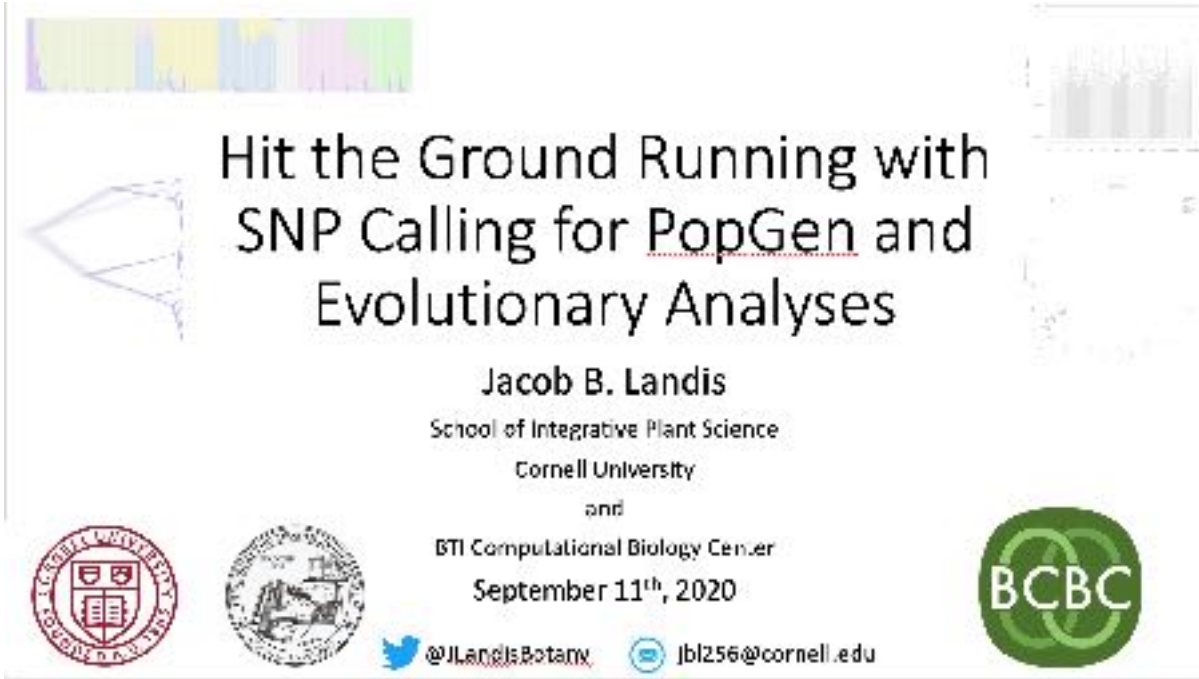
BTI Computational Biology Center (BCBC)

February 5th, 2021

@JLandisBotany jbl256@cornell.edu

So now you have variants.

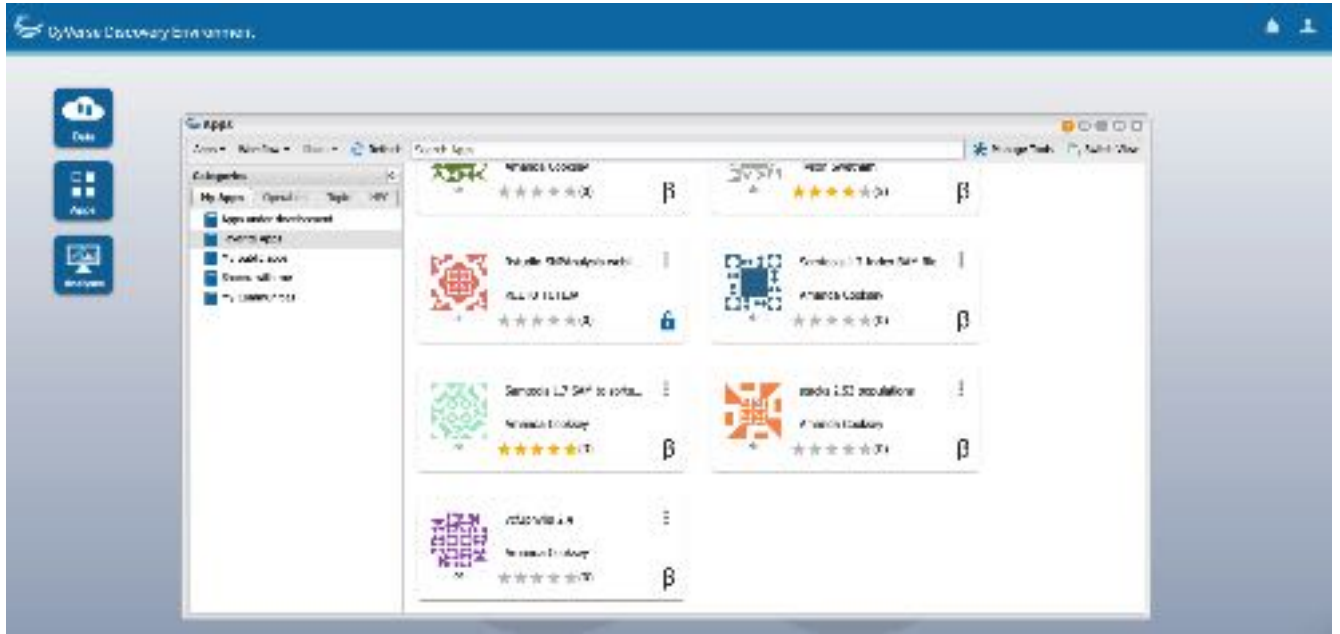
- Last September we walked through calling variants with Stacks and GATK
- Here we will pick up where we left off, cover downstream analyses
- Data and scripts on GitHub: https://github.com/bcbc-group/CyVerse_Variant_Analyses
- CyVerse Discovery Environment



https://cyverse.org/webinar_SNP_analysis

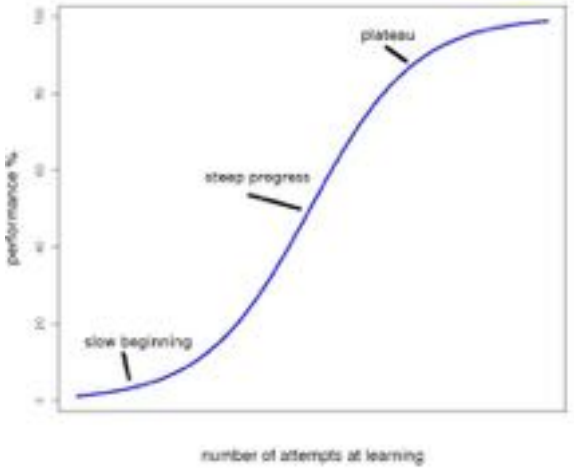
Discovery Environment

- Almost every analysis has its own App, though we will be using the same App for all the Rstudio steps
- Most programs have more options than included in the Apps, so if you want to explore in-depth I suggest using the command line versions



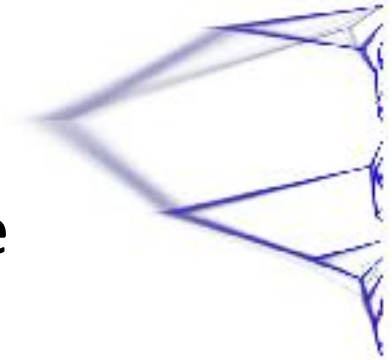
What can you expect after this?

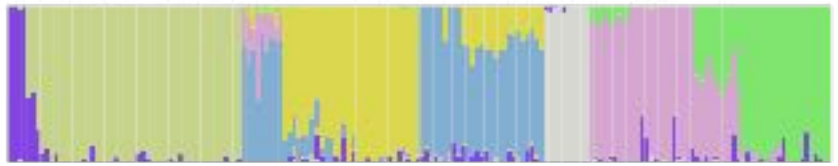
- Hard to design a tutorial for all skill-levels
- Do not expect anyone to be a pro, but hopefully decrease the learning curve for your own projects and data by showing what can work
- If you have never done these analyses before
  - Provide workflows with CyVerse resources and standalone programs
- If you have some experience with these methods
  - Provide command line options for analyses, provide a bit more flexibility and power
- If you are an advanced user
  - Provide some alternative methods and avenues to explore more

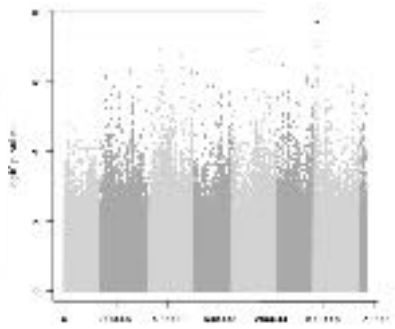


What we will cover

- Tree building
- PCA and Structure
- Isolation by Distance
- GWAS



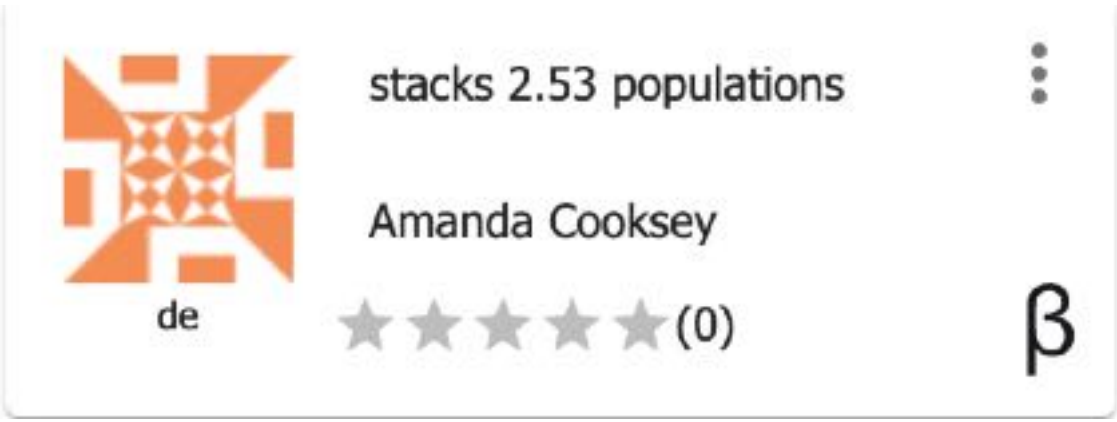




- Many options available for each of these, covering some examples
- Some data sets will work better for some analyses than others
  - Computational requirements, reference vs *de novo* approaches for SNP calling, number of contigs in the reference
  - Will try to highlight these when possible, both in slides and in the tutorial materials

Need to give thanks where it is due

- Throughout the webinar and tutorial you will see Apps for analyses
- Reetu Tuteja and Amanda Cooksey at CyVerse did an amazing job getting everything pulled together and operational
- Thank you for the amazing work

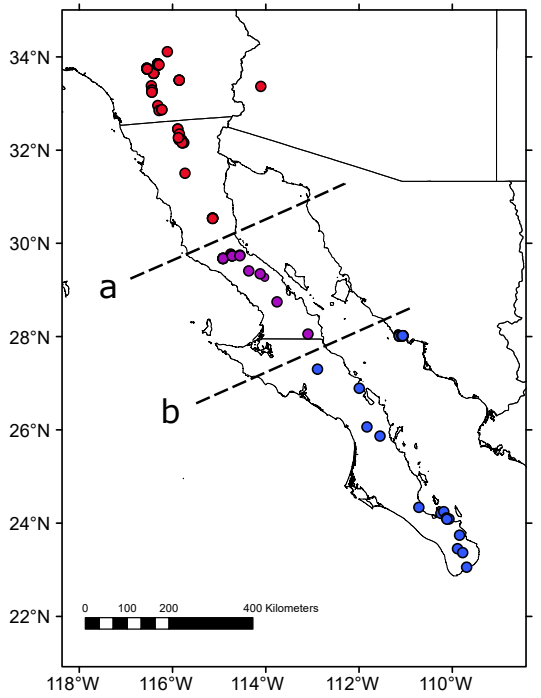


Where the data comes from

- Using a modified VCF (Variant calls) from a recently accepted publication on *Washingtonia* palms
  - Originally included 181 accessions
  - Some analyses here only have 48 accessions for ease
- Include known populations and geographic locations
- Mix of observed and simulated phenotype data for GWAS (not included in the original study)

Washingtonia filifera

Lorena Villanueva



Botanical Journal of the Linnean Society

Genetic and morphological differentiation in *Washingtonia* (Arecaceae):

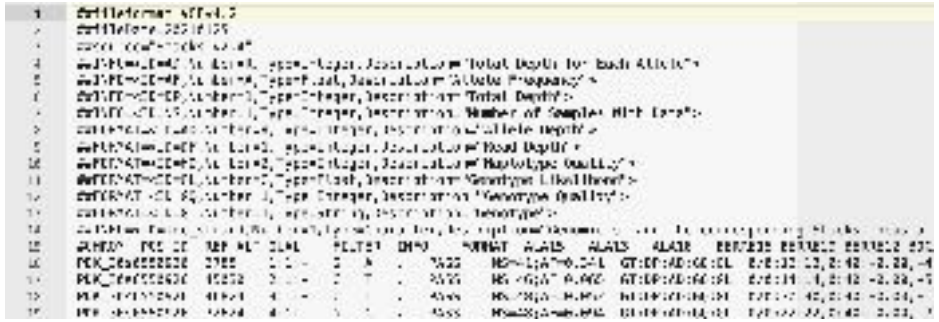
solving a century-old palm mystery

Lorena Villanueva-Almanza[1,4], Jacob B. Landis[1], Daniel Koenig[1,2,3], and Exequiel Ezcurra

File conversion

- One of the more frustrating aspects is that programs often do not take the same input files

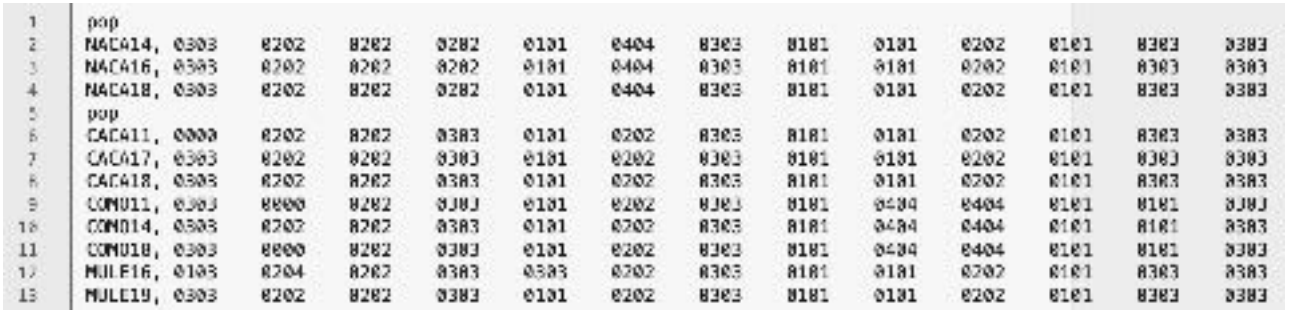
VCF



Structure

Fasta

GenePop



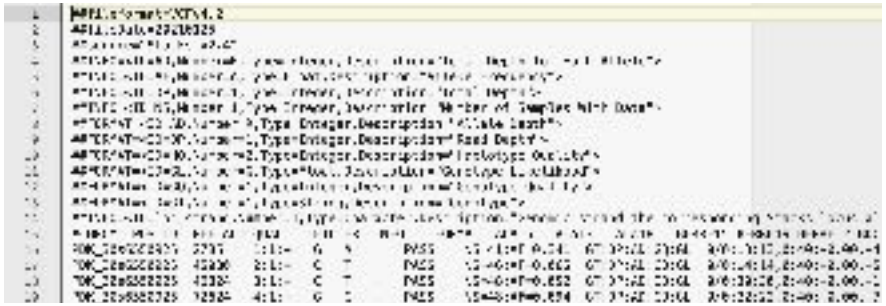
- A few tools exist to help with this problem

File conversion

- Can use Stacks (commonly used for SNP calling) to generate many different downstream files by specifying a VCF and a population map

DE APP

VCF



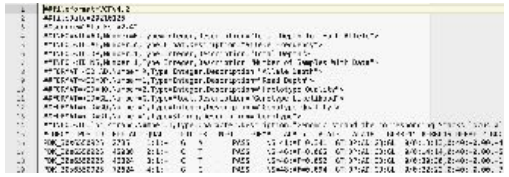
Population map

1	ALA15	ALA15
2	ALA13	ALA13
3	ALA18	ALA18
4	BERRE15	BERRE15
5	BERRE19	BERRE19
6	BERRE12	BERRE12
7	BOCA20	BOCA20
8	BOCA14	BOCA14
9	BOCA19	BOCA19

```
populations  -V Samples.vcf -O working_files/ -M pop_map.txt --ordered-export --vcf --genepop --structure
```

- VCF2Phylip is a python script that easily converts a VCF to nexus and fasta

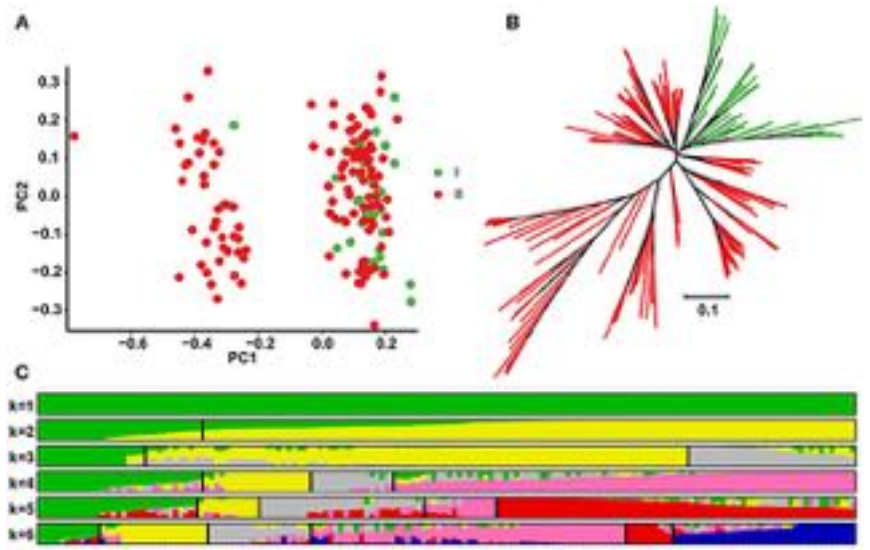
VCF



```
python3 vcf2phylip.py -i Samples.vcf -n -f
```

Showing relatedness

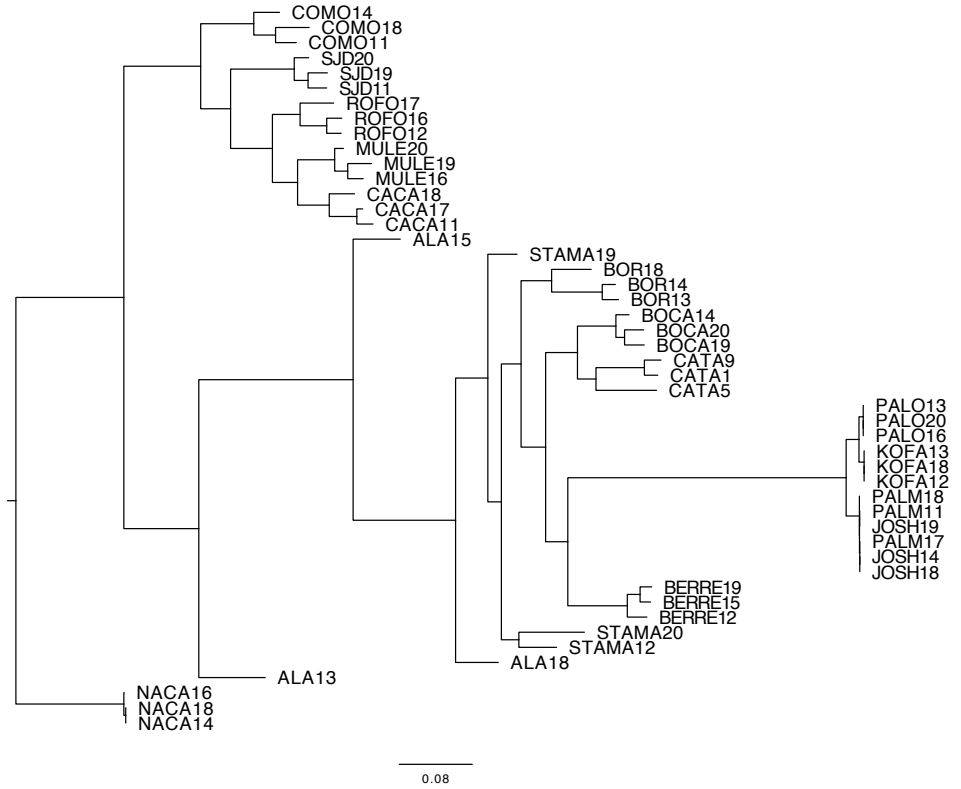
- Many different ways to show how your samples may be related
- Input data is slightly different between PCA, Structure, and phylogenetic inference
  - PCA and Structure you may have prior knowledge on which samples group together based on collection sites
- Pruning for Linkage Disequilibrium is important
  - Failure to do this will inflate bootstrap support values
  - Also speed requirements for additional loci that do not alter the story at all



Zhang et al. 2017; *Frontiers in Plant Science*

SNPhylo – Maximum likelihood

- Very important that chromosomes are represented by numbers only
  - Things like “contig” will cause the analysis to crash
- Default is only for 22 chromosomes (autosomes in humans), can specify any number with the -a flag (such as -a 600000)
  - This will not hamper the analysis in any way
- Since we have LD pruned previously, -l is set to 1.0 to ignore




```
snphylo.sh -v Samples.vcf -p 95 -c 5 -l 1.0 -m 0.05 -M 0.8 -P ML_tree -a 22
```

SNAPP - Bayesian Coalescence

- A Maximum-likelihood, concatenated data set may not be appropriate for population level analyses - gene flow between pops
- Explores the conflicting signal between SNPs
- Computationally very intensive and slow, maybe not the best option
- Ignores loci with missing data, so may want to further prune input file

```
vcftools --vcf Samples.vcf --keep SNAPP.txt --max-missing 1.0 --thin 5000 --recode --recode-INFO-all --out Samples_SNAPP
```

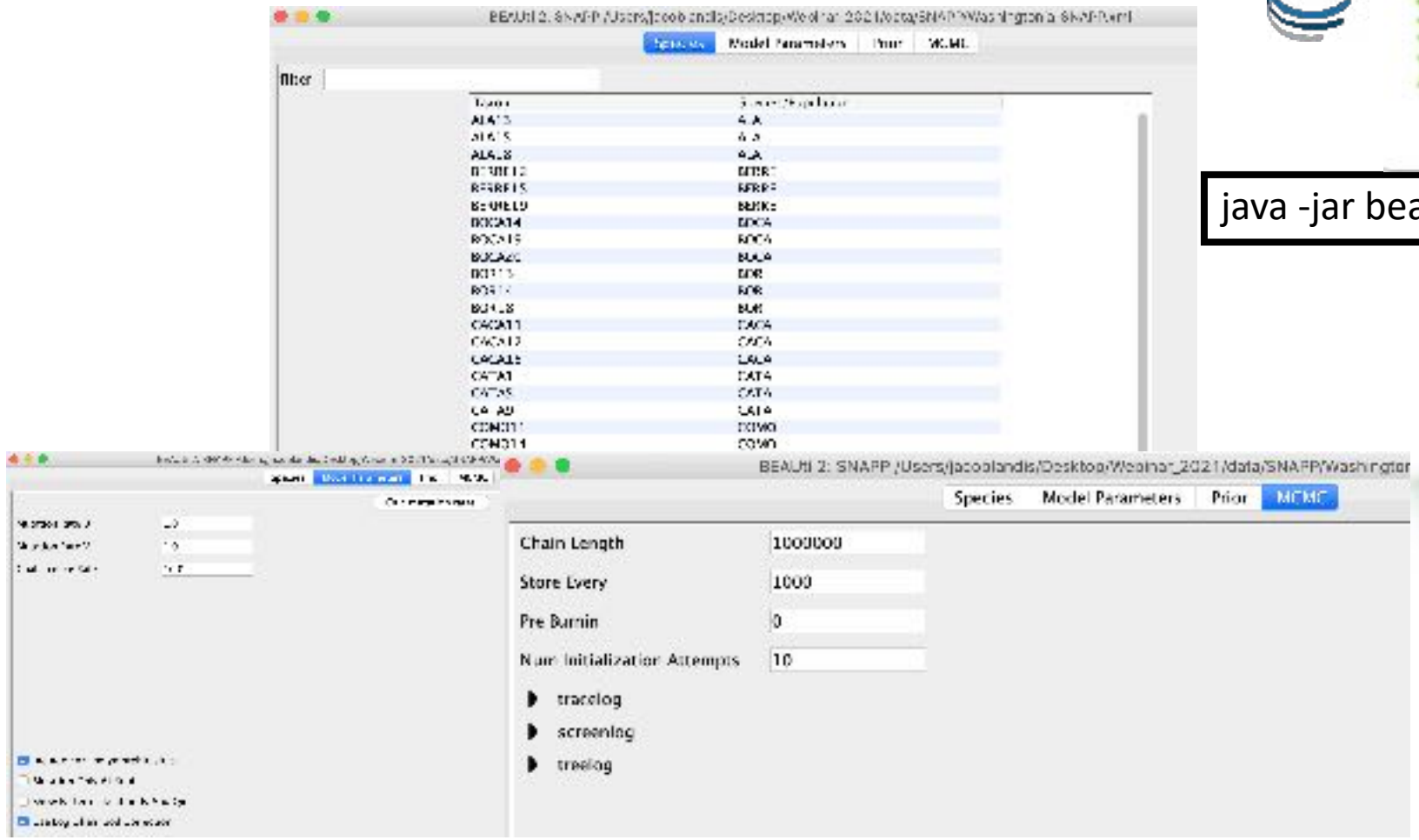
- Use the created Fasta file to generate the XML file to run in BEAST2
  - Standalone program BEAUTi packaged with BEAST2

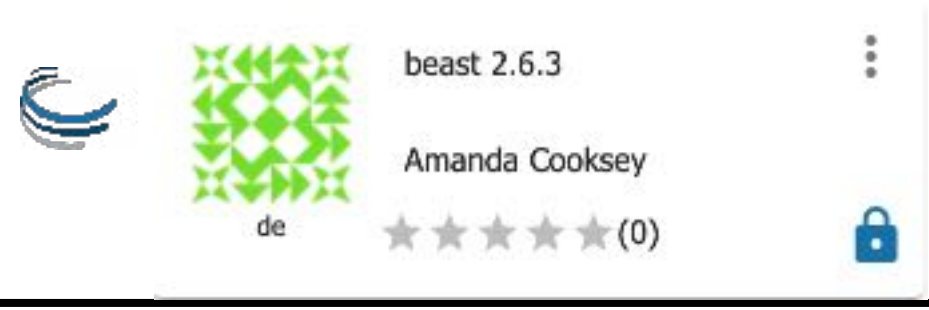
Beast2

Bayesian evolutionary analysis by sampling trees

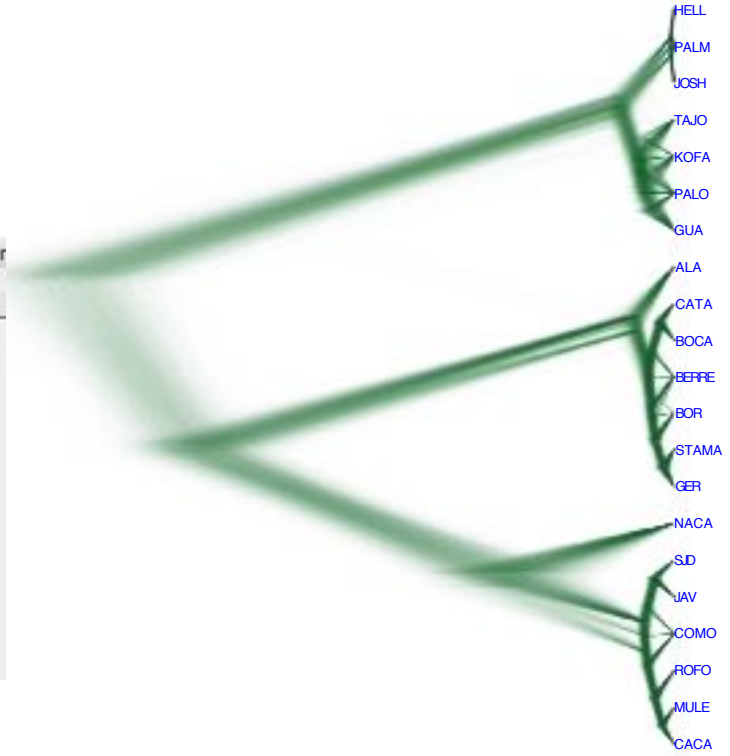
SNAPP - Bayesian Coalescence

BEAUTi input



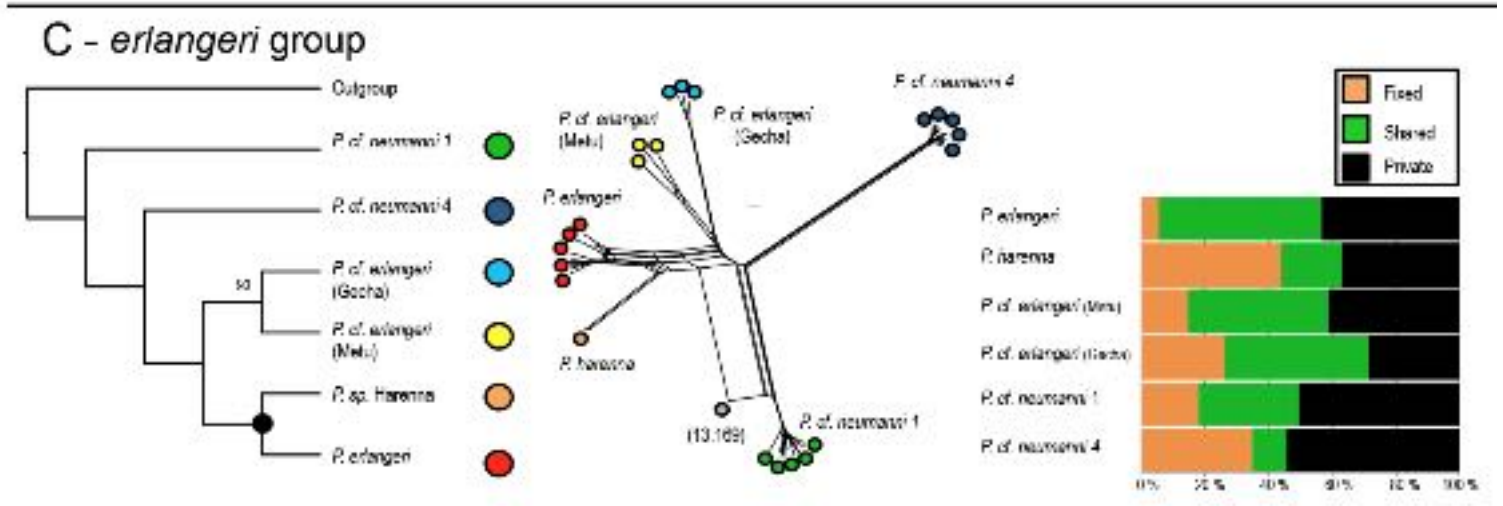


java -jar beast.jar -threads 24 Washingtonia_SNAPP.xml



SplitsTree – Neighbor Net

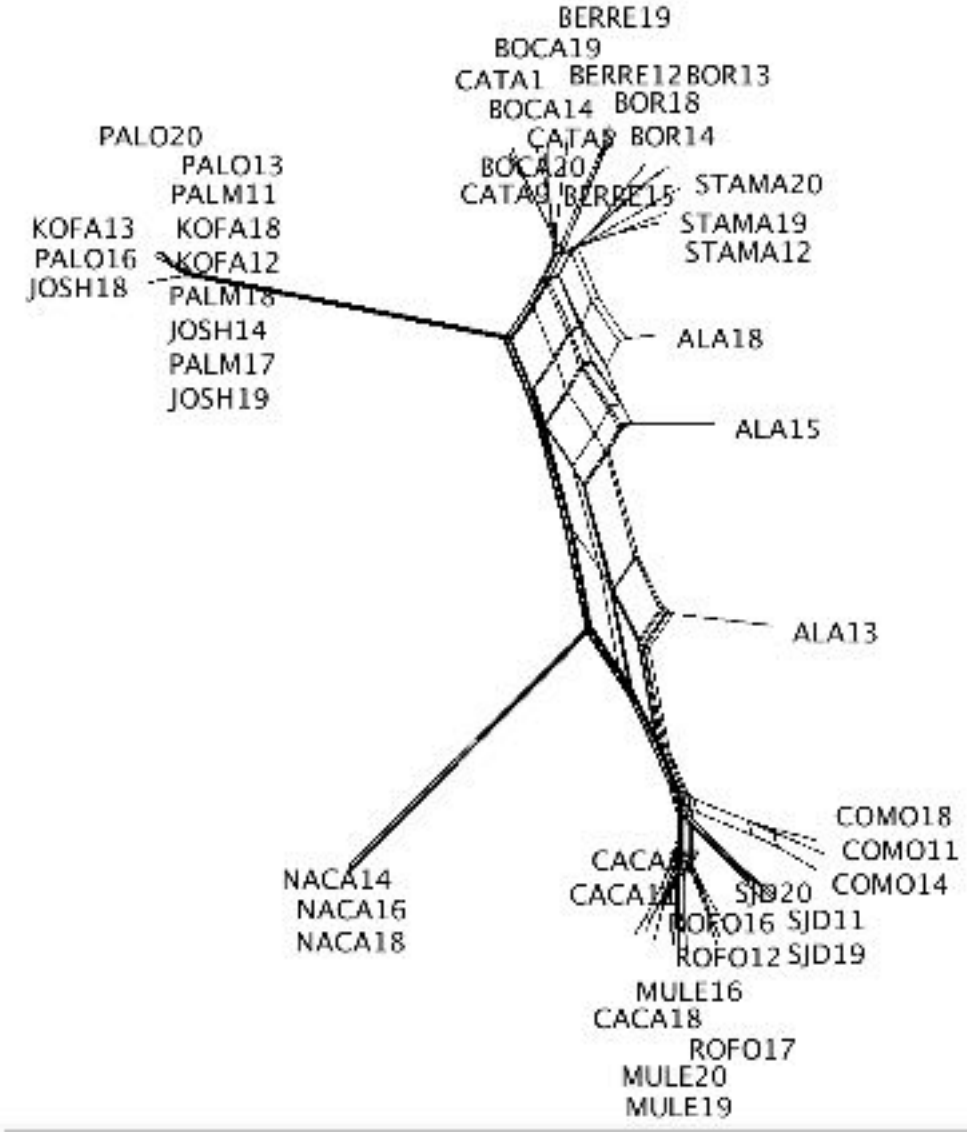
- This approach uses a distance matrix to cluster samples but does not rely on a bifurcating, hierarchical layout
- Collection of clusters represented with splits graphs, some splits may show conflicting signals
  - “Ideal” data will show a tree pattern
- Effective for computing hybridization or recombination networks



Reyes-Velasco et al. 2018; *PloS One*

SplitsTree – Neighbor Net

- SplitsTree runs pretty fast and has a graphical interface; run locally
  - SplitsTree4: https://software-ab.informatik.uni-tuebingen.de/download/splitstree4/welcome.html
  - SplitsTree5: https://software-ab.informatik.uni-tuebingen.de/download/splitstree5/welcome.html
- Load the nexus file created earlier
  - File -> Open -> Samples.nex

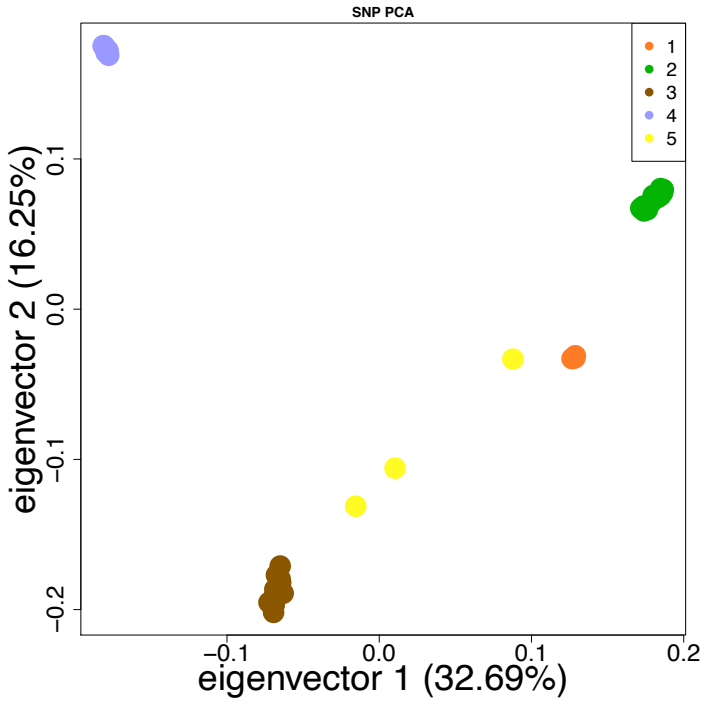


Principal Component Analysis (PCA)

- Way to take a high dimensionality data set (tens to hundreds of thousands of SNPs) and reduce to something informative
  - Usually the first handful of principal components (PC)
- PC1 is the direction that maximizes the variance of the projected data
- Do not need to designate *apriori* populations/clusters
  - Can color code for easier interpretation
- The top PCs of genetic PCA often explain a lower percentage of the variance than phenotypic PCA
  - 70-80% ideal for top PCs in phenotypic data, 20-30% is great for genetic data
  - Number of characters

Principal Component Analysis (PCA)

- There is a Rstudio App in the Discovery Environment with all the necessary R libraries
- Web browser of Rstudio, run through all the commands as if you were on your own machine
- Script will write PDFs of each plot, to view these click File -> Open File -> selected PDF
- Download to your local machine; when the Rstudio instance is cancelled, all plots will be lost
- Will create PC1 vs PC2, PC2 vs PC3, and PC3 vs PC4

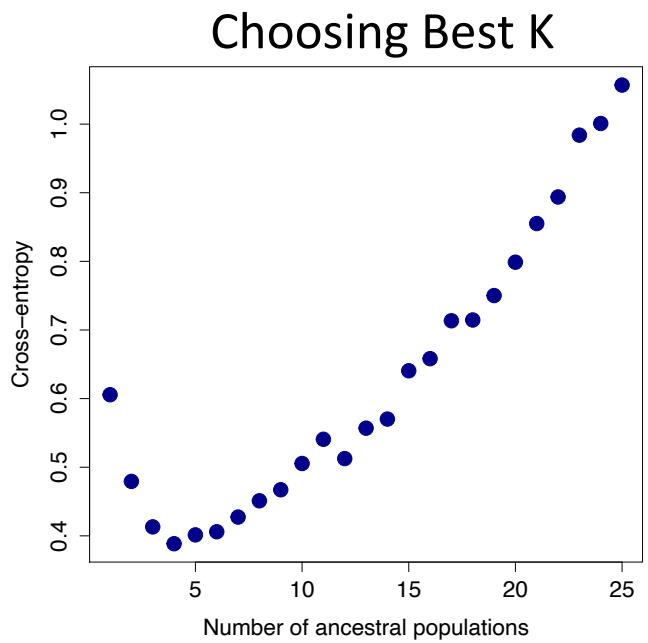


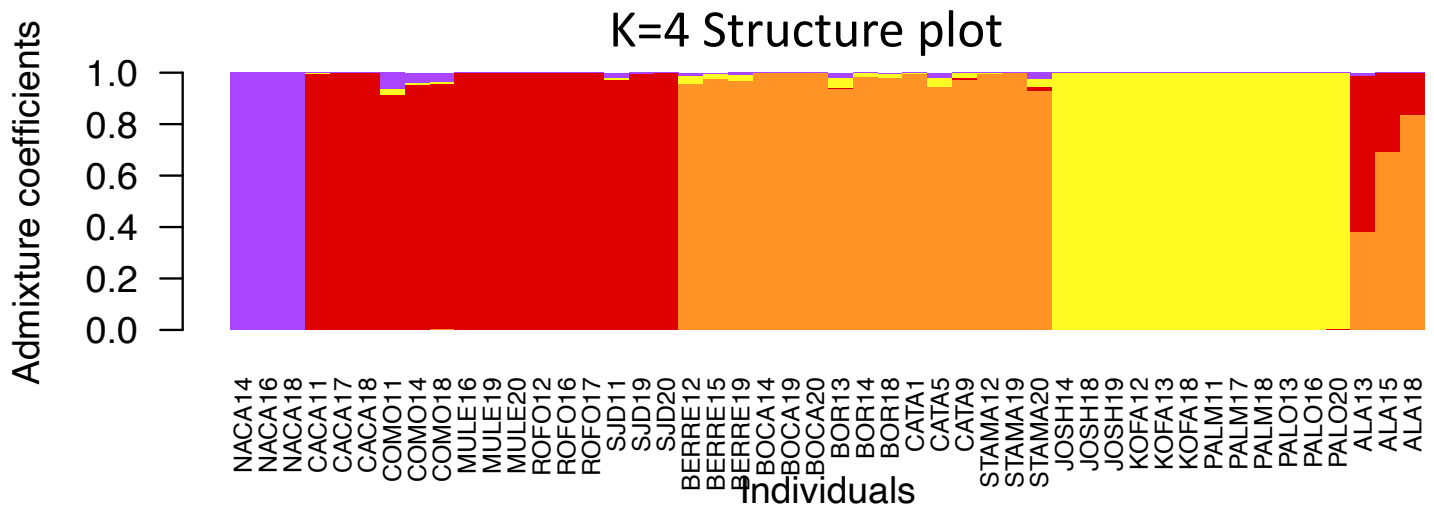
Structure/LEA

- Way to investigate ancestral populations (K) and mixture between samples
  - Can over interpret the best K value (Lawson et al. 2018)
- Picking the best K?
  - The published deltaK method tends to over pick K=2 (Janes et al. 2017)
  - LEA uses a cross-entropy criteria to select best K, low point on the graph
- As opposed to Structure and Admixture, the LEA R package does not rely on assumptions of no genetic drift, Hardy-Weinberg, or linkage equilibrium
  - More suitable for inbred lineages
- Stacks can create the input for Structure, however you will need to delete the first two header rows before running it (tutorial data has already been done)

Structure/LEA

- R script to use the LEA package to test for best K (2-25) and create Structure plots for K2-K8
- To make a cleaner plot, samples (2 lines per sample) can be reorganized in a text editor





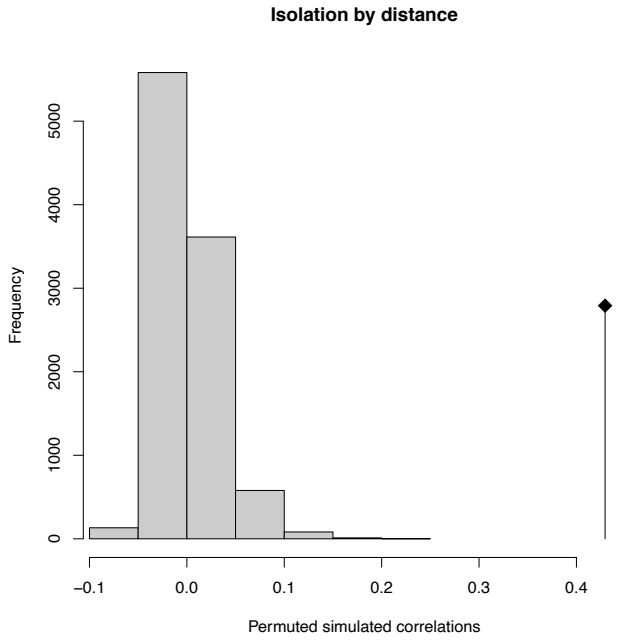
Isolation by Distance

- Isolation by Distance is a consequence of limited dispersal
  - Pairs of populations close to each will be more genetically similar than populations far away
  - Environmental barriers can inhibit dispersal which can be biologically interesting
- Strong presence of IBD can bias outlier loci scans; up to 30% of loci can be below the significance threshold but caused by nothing more than spatial patterns (Meirmans 2012)
- Use the GenePop file from Stacks, though we need to rename as .gen
  - This has already been done for the example data
- Input file of geographic location
  - Individuals or populations

	A	B	C
1	sample_id	LATITUDE	LONGITUDE
2	ALA15	27.29756	-112.88292
3	ALA13	27.29738	-112.88238
4	ALA18	27.29728	-112.88304
5	BERRE15	30.53677	-115.1359
6	BERRE19	30.538167	-115.13575
7	BERRE12	30.52792	-115.13335
8	BOCA20	29.67203	-114.91109

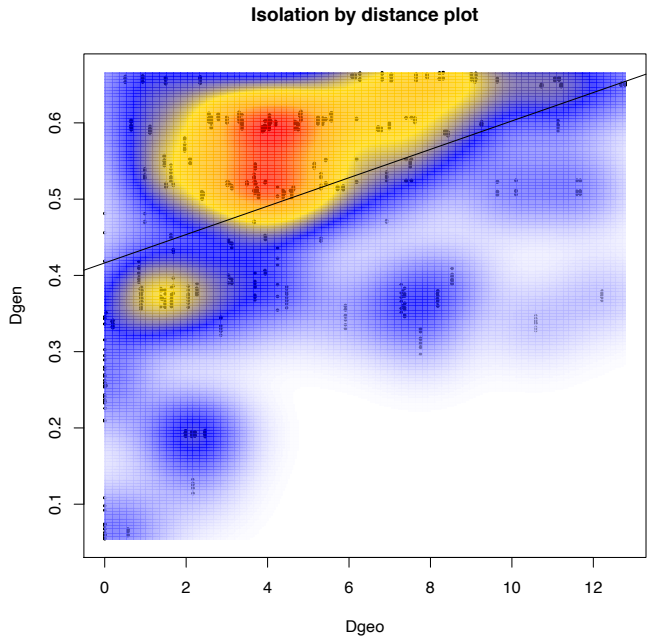
Isolation by Distance

- Rscript for Rstudio session 
- Performs 10,000 permutations to test for significance
  - Permutes one matrix (geographical distance) while the other is fixed (genetic)
- Produces a plot of genetic distance against geographic distance



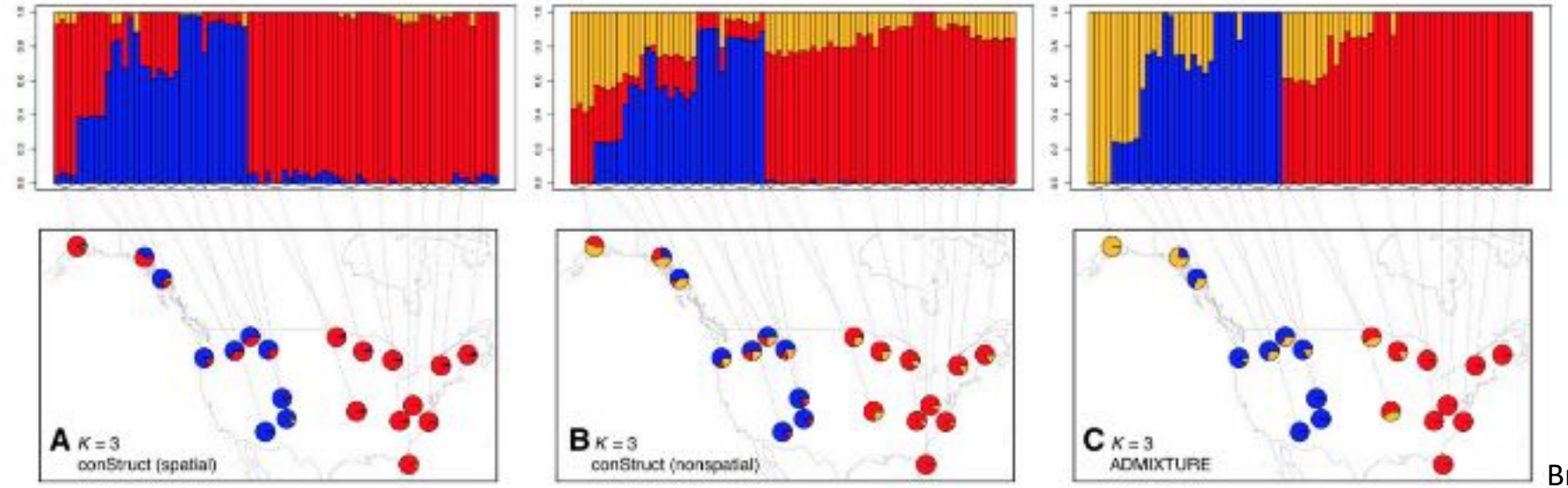
Monte-Carlo test
Observation: 0.4294366

Based on 10000 replicates
Simulated p-value: 9.999e-05



Alternatives

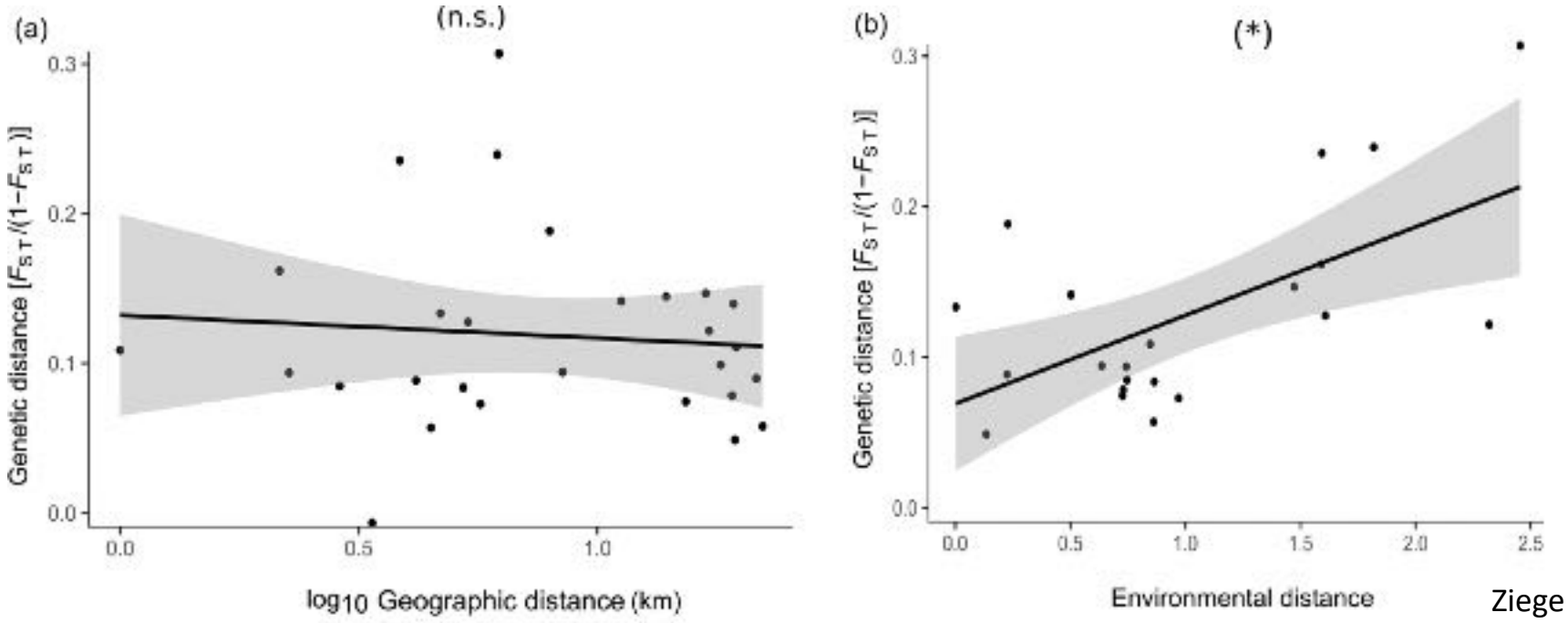
- Continuous patterns of differentiation may cause issues in analyses
  - Harder to assign discrete clusters
- A newer program called conStruct was recently released (2018)
- Categorizes natural genetic variation with both continuous and discrete patterns



Bradburd et al. 2018; *Genetics*

Isolation by Environment/Resistance

- IBE - Genetic and environmental distances are positively correlated independently of geographic distance (Wang and Bradburd 2014)
- IBR – Relationship between genetic distance and resistance distance, which better accounts for habitat heterogeneity (McRae 2006)



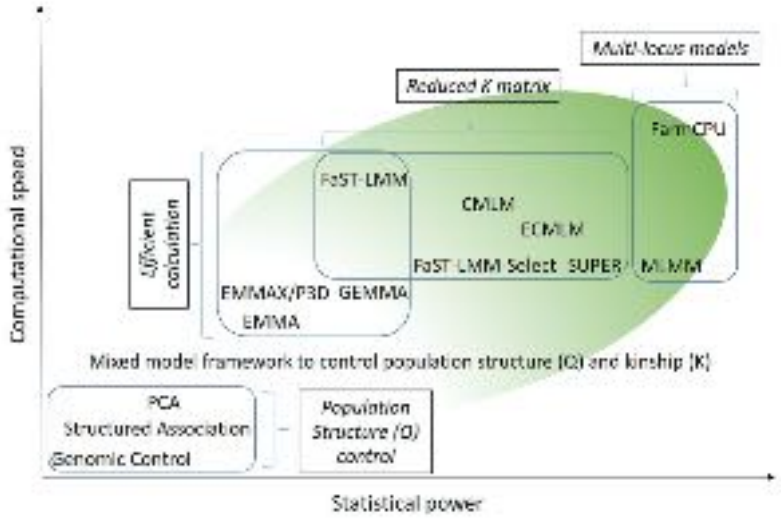
Ziege et al. 2020; *Scientific Reports*

Genome Wide Association

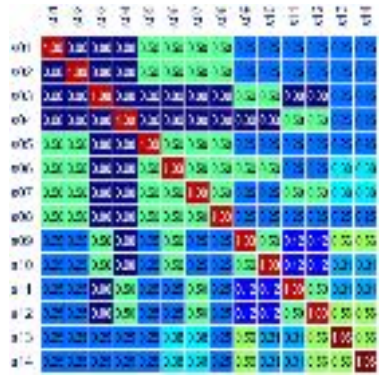
- Scan markers to look for association with SNPs and phenotypes of interest
- Considerations – normalize phenotype data, quantitative continuous data, make sure sample size is large enough, fairly dense sampling of SNPs
- Most methods are designed for reference genome data
  - Low number of contigs/chromosomes
  - *de novo* aspects have issues with LD and lack of coverage across the genome
- LD pruning important, will change the level of significance based on number of SNPs/tests

GWA/Models

- GLM – General Linear Model
  - Association using a least squares fixed effects linear model
  - Reduces false positives with the inclusion of population structure (PCA)
- MLM – Mixed Linear Model
  - Incorporates covariates for population structure and kinship to reduce false positives
  - Can be computationally challenging for large data sets
- Kinship matrix
  - Account for relationships among individuals.
  - Probabilistic estimate that a random gene from $subject_i$ is identical by descent (ibd) to a gene in the same locus from $subject_j$



Cortes et al. 2021; *Plant Genome*

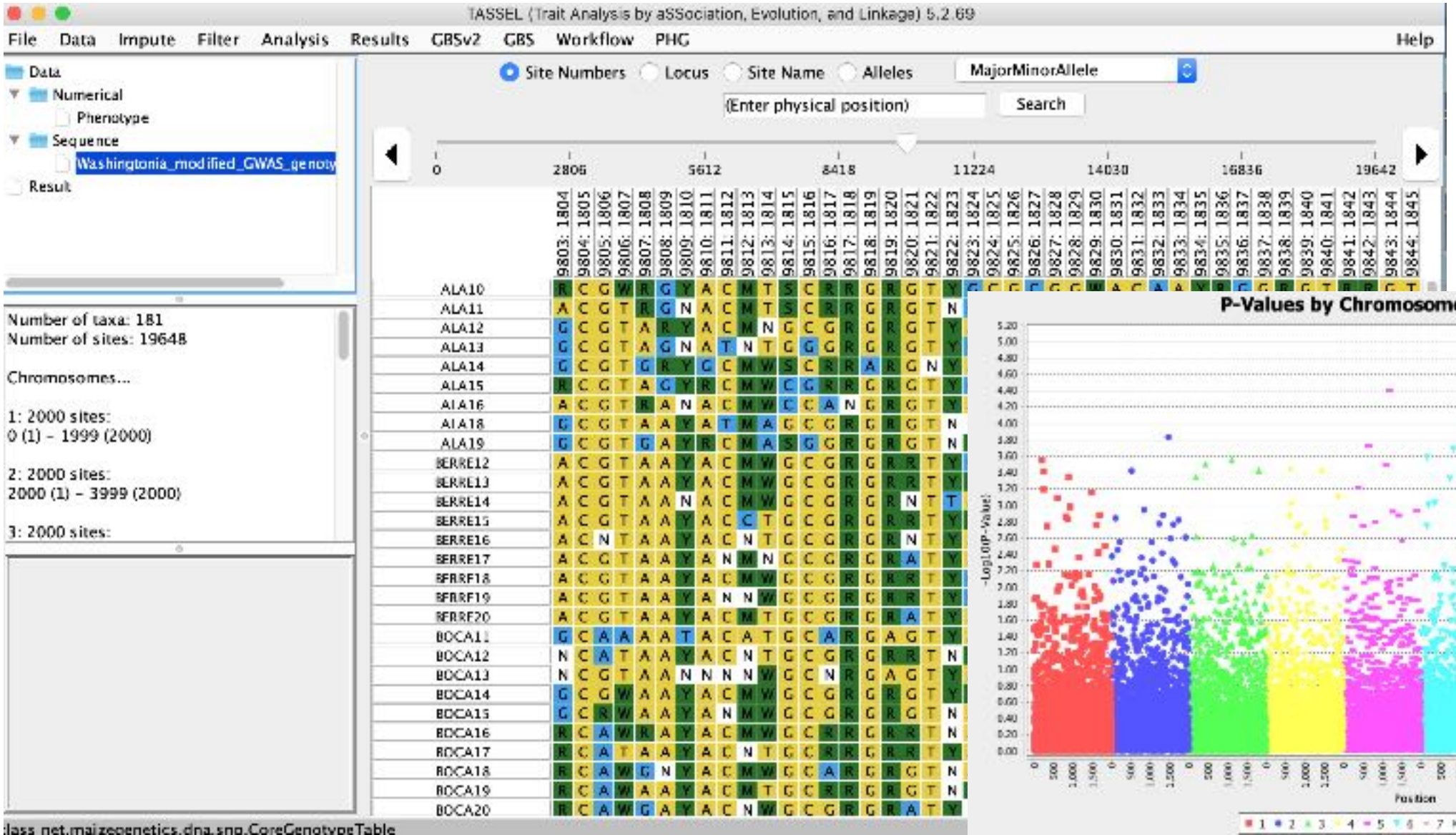


Tassel

- Graphical interface on your local machine
- Point and click options for almost everything you would want to do
- Pretty extensive tutorials online
- Fast, works on almost any platform
- Not picky about samples with missing data, genotype and phenotype in different orders, can use a VCF as direct input
- Not much control over plotting options, though will plot for you directly

TASSEL - Trait Analysis by aSSociation, Evolution and Linkage

Tassel


TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage) 5.2.69
File Data Impute Filter Analysis Results GBSv2 GBS Workflow PHG Help
Data
Numerical
Phenotype
Sequence
Washingtonia_modified_GWAS_genoty
Result
Site Numbers Locus Site Name Alleles
MajorMinorAllele
(Enter physical position)
Search
Number of taxa: 181
Number of sites: 19648
Chromosomes...
1: 2000 sites:
0 (1) - 1999 (2000)
2: 2000 sites:
2000 (1) - 3999 (2000)
3: 2000 sites:
class net.maizegenetics.dna.snp.CoreGenotypeTable

P-Values by Chromosome for Trait1
-Log10(P-Value)
Position

EMMAX

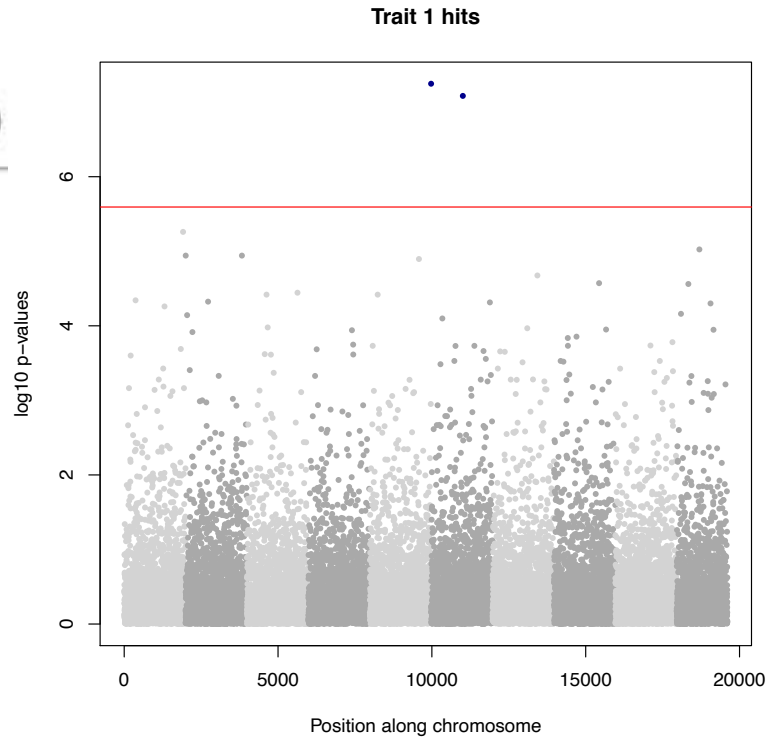
- Widely used for larger data sets, though new publications are using GEMMA
- Need to reformat VCF to Plink
- Plink is very picky about number of chromosomes (give it --allow-extra-chr for more than 22 chromosomes)
- Two steps to run GWA: first generate kinship matrix, then actual test
- Will not plot results for you, need to use R to plot (example script provided in tutorial)

EMMAX

- Reformat the VCF to Plink

- Generate the kinship matrix and test for associations all in the same step
  - Kinship matrix is not saved unfortunately

- Use the resulting manhattan.plot file and the file with the chromosome/bp positions to plot in Rstudio
  - Default is a gray alternation of chromosomes and highlight significant peaks; all this is customizable



Wrap up

- Covered many aspects of population genetics and evolutionary analyses
  - Tree building, population structuring, and genome wide association
- Most analyses can be done with a wide variety of data: RAD-Seq, Hyb-Seq, RNA-Seq, and genome resequencing
- Some analyses have requirements with specific contig/chromosome formatting or number of contigs/chromosomes
- Point and click options in the CyVerse Discovery Environment as well as command line options for local machines

References

- Bradburd et al. 2018. Inferring continuous and discrete population genetic structure across space. *Genetics*
- Frichot and Francois. 2015. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*
- Cortes et al. 2021. Status and prospects of genome-wide association studies in plants. *The Plant Genome*
- Frichot et al. 2014. Fast and efficient estimation of individual ancestry coefficients. *Genetics*
- Janes et al. 2017. The K = 2 conundrum. *Molecular Ecology*
- Kloepper and Huson. 2008. Drawing explicit phylogenetic networks and their integration into SplitsTree. *BMC Evolutionary Biology*
- Lawson et al. 2018. A tutorial on how not to over-interpret Structure and Admixture bar plots. *Nature Communications.*
- McRae. 2006. Isolation by resistance. *Evolution*
- Merirmans. 2012. The trouble with isolation by distance. *Molecular Ecology*
- Reyes-Velasco et al. 2018. Revisiting the phylogeography, demography and taxonomy of the frog genus *Ptychadena* in the Ethiopian highlands with the use of genome-wide SNP data. *PLoS One*
- Wang and Bradburd. 2014. Isolation by environment. *Molecular Ecology*
- Zhang et al. 2017. Genome-Wide Association Study of Major Agronomic Traits Related to Domestication in Peanut. *Frontiers in Plant Science*
- Ziege et al. 2020. Population genetics of the European rabbit along a rural-to-urban gradient. *Scientific Reports*

CyVerse is supported by the National Science Foundation under Grants No. DBI-0735191, DBI-1265383 and DBI-1743442.