Analyzing RADSeq Data

Jennifer Gardner

FISH 546: Bioinformatics

Question:

What are the steps to go from raw RADSeq data to data that could be input into a tree?

Question:

What are the steps to go from raw RADSeq data to data that could be input into a tree?

- Can I perform those steps following along from the methods section of a paper?

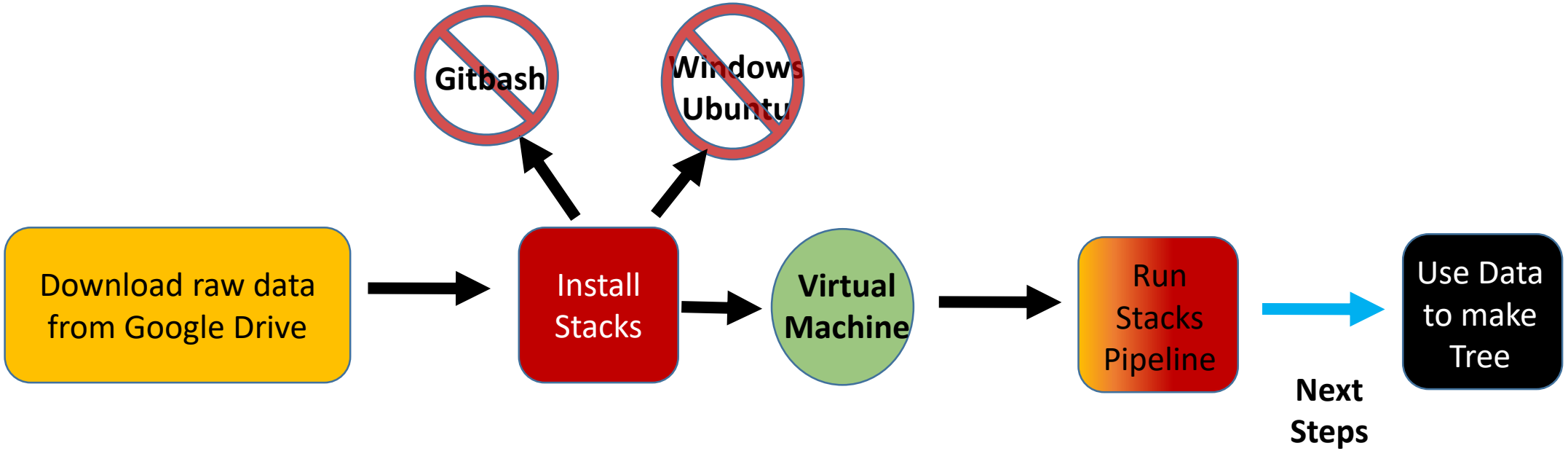
Question:

What are the steps to go from raw RADSeq data to data that could be input into a tree?

- Can I perform those steps following along from the methods section of a paper?

  Spoiler alert: NO

RADSeq Workflow using Stacks v 2.2



Raw data (37 gb fq.gz)
Barcode file (.txt)

Demultiplex samples with `process_radtags`

Directory with 48 fq.gz files, 1 per specimen

Filter and cap samples by read number with my script

Directory with 44 fq.gz files. 1e5<n-reads<2e6

Run `ustacks` on each sample to find loci

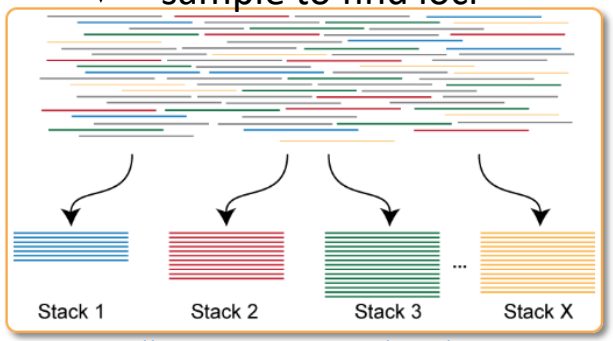


Figure http://catchenlab.life.illinois.edu/stacks/param_tut.php

Directory with ustacks output (Name.tags.tsv.gz, Name.snps.tsv.gz, Name.alleles.tsv.gz

Run `cstacks` on all samples to build catalog of loci (de novo genome)

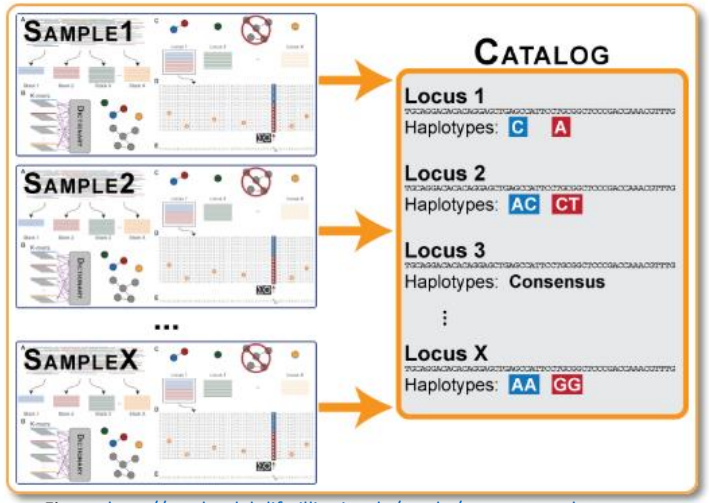


Figure http://catchenlab.life.illinois.edu/stacks/param_tut.php

Directory with ustacks and cstacks output data (ustack files and catalog files)

Run `sstacks` on all samples to align against the catalog

Directory with sstacks, ustacks and ccstacks data (name.matches.tsv.gz)

Directory with ustack, cstacks, and sstacks outputs

Run `tsv2bam` to transpose data to a bam alignment by loci instead of by specimen

Directory with tsv2bam output and ustacks, cstacks, and cstacks outputs (name.matches.bam)

Run `gstacks` to genotype

Directory with gstacks outputs?

Run `populations` to get population summary statistics such as F_{ST} and output data into a fasta or phylip that can be used for analysis

Fasta or phylip files that can now be used to make a tree!!

Raw data (37 gb fq.gz)
Barcode file (.txt)

Demultiplex samples with `process_radtags`

Directory with 48 fq.gz files, 1 per specimen

Filter and cap samples by read number with my script

Directory with 44 fq.gz files. 1e5<n-reads<2e6

Run `ustacks` on each sample to find loci

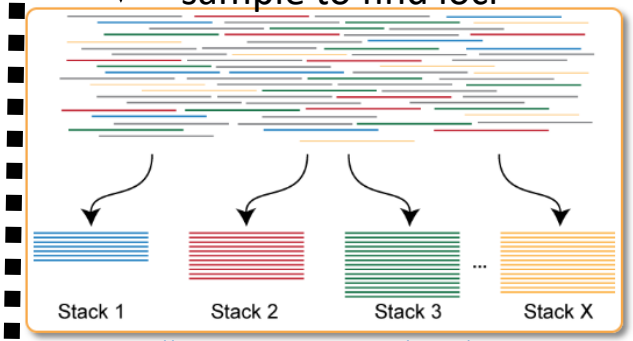


Figure http://catchenlab.life.illinois.edu/stacks/param_tut.php

Directory with ustacks output (Name.tags.tsv.gz, Name.snps.tsv.gz, Name.alleles.tsv.gz

Run `cstacks` on all samples to build catalog of loci (de novo genome)

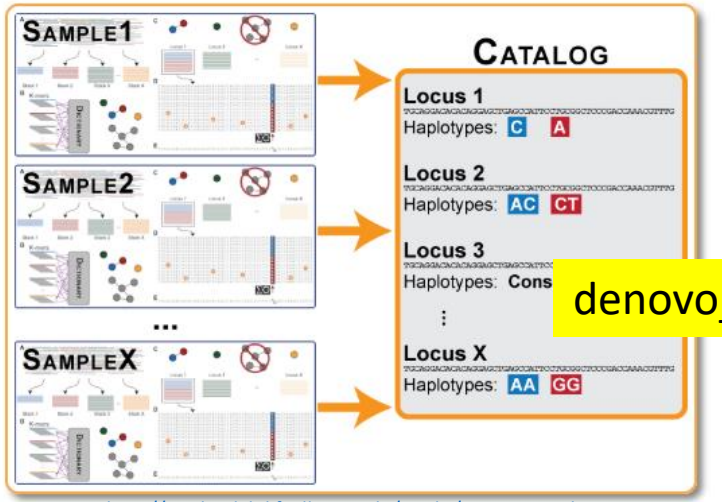


Figure http://catchenlab.life.illinois.edu/stacks/param_tut.php

Directory with ustacks and cstacks output data (ustack files and catalog files)

Run `sstacks` on all samples to align against the catalog

Directory with sstacks, ustacks and ccstacks data (name.matches.tsv.gz)

denovo_map.pl

Directory with ustack, cstacks, and sstacks outputs

Run `tsv2bam` to transpose data to a bam alignment by loci instead of by specimen

Directory with tsv2bam output and ustacks, cstacks, and cstacks outputs (name.matches.bam)

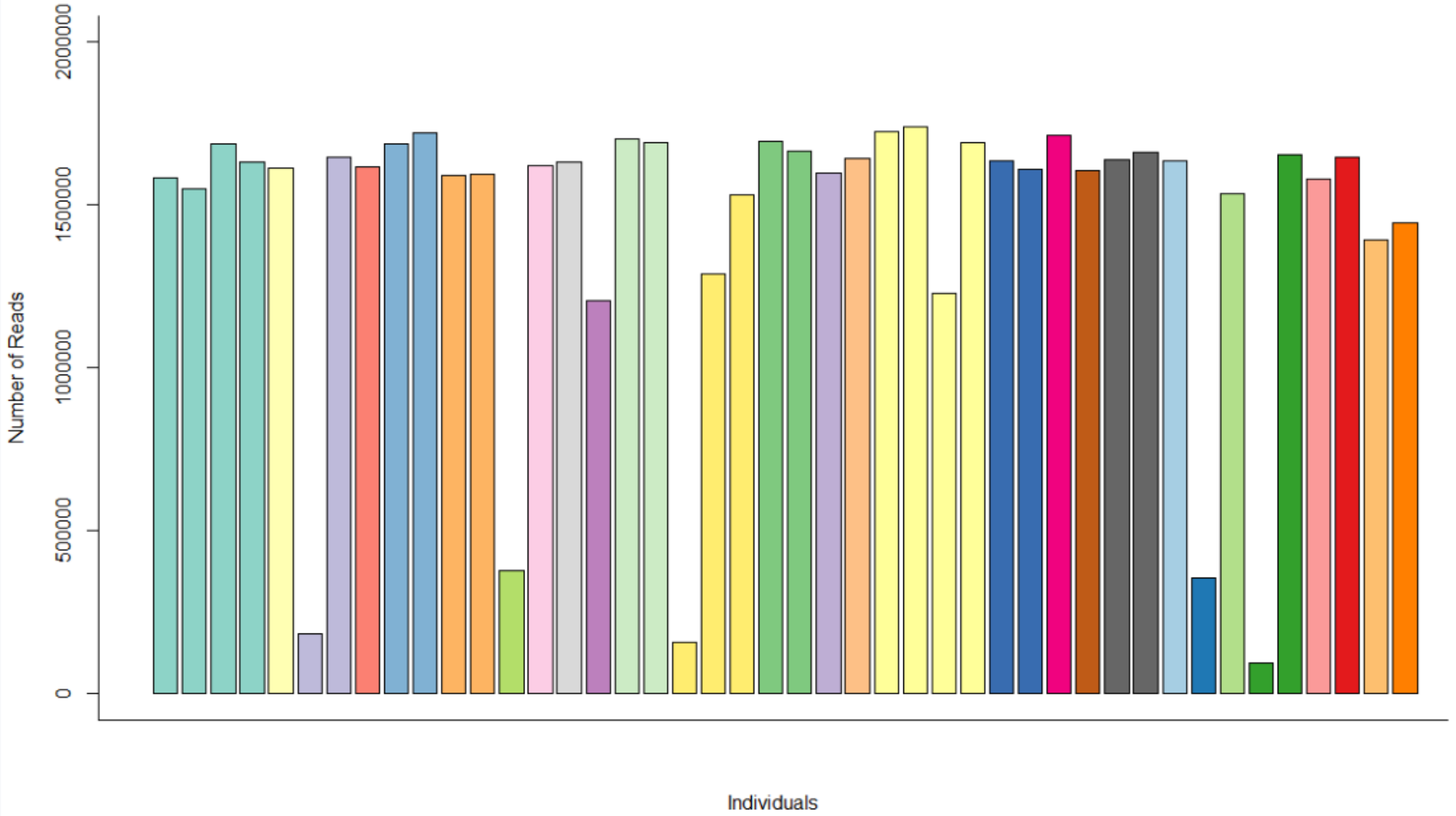
Run `gstacks` to genotype

Directory with gstacks outputs?

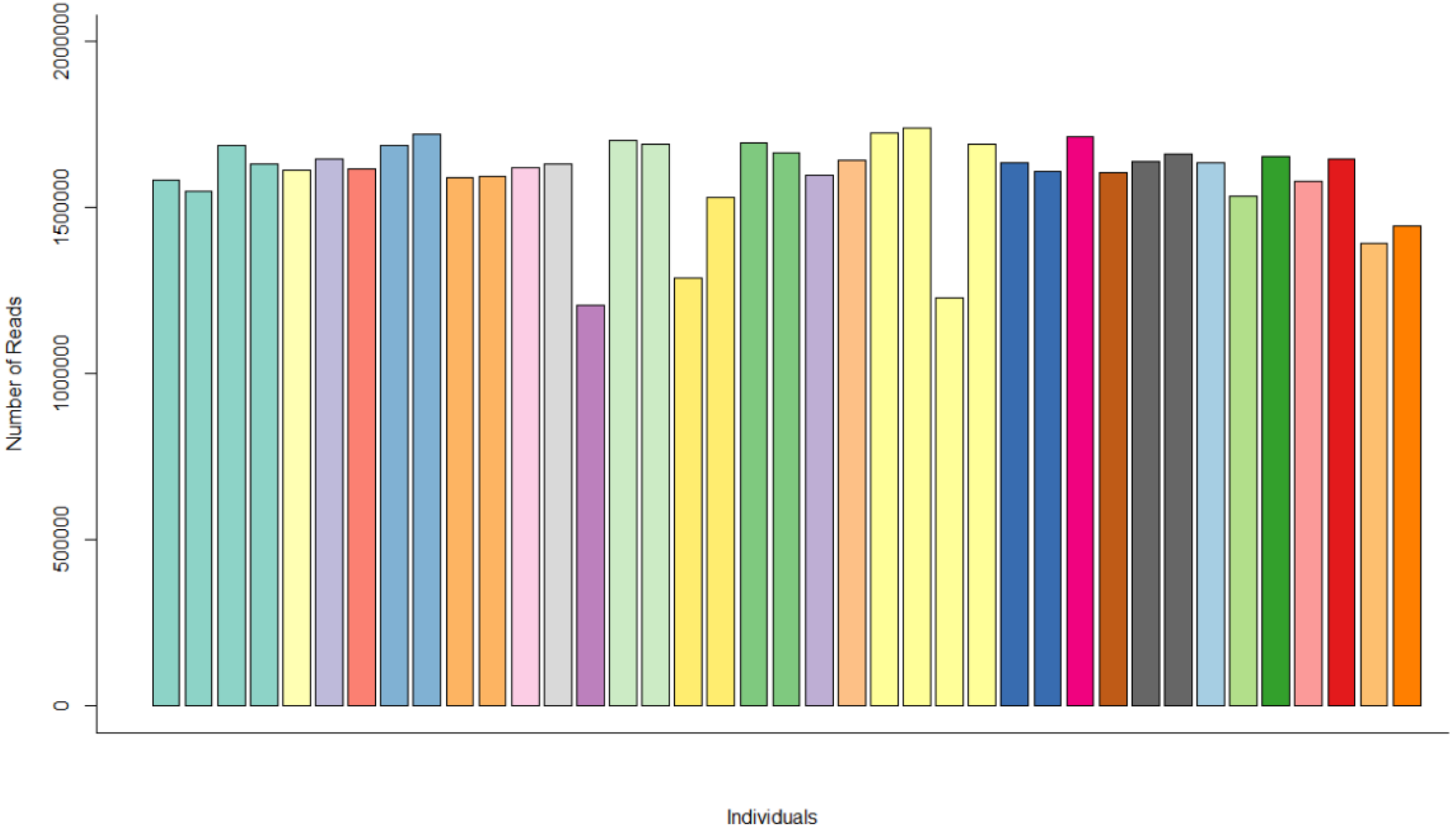
Run `populations` to get population summary statistics such as F_{ST} and output data into a fasta or phylip that can be used for analysis

Fasta or phylip files that can now be used to make a tree!!

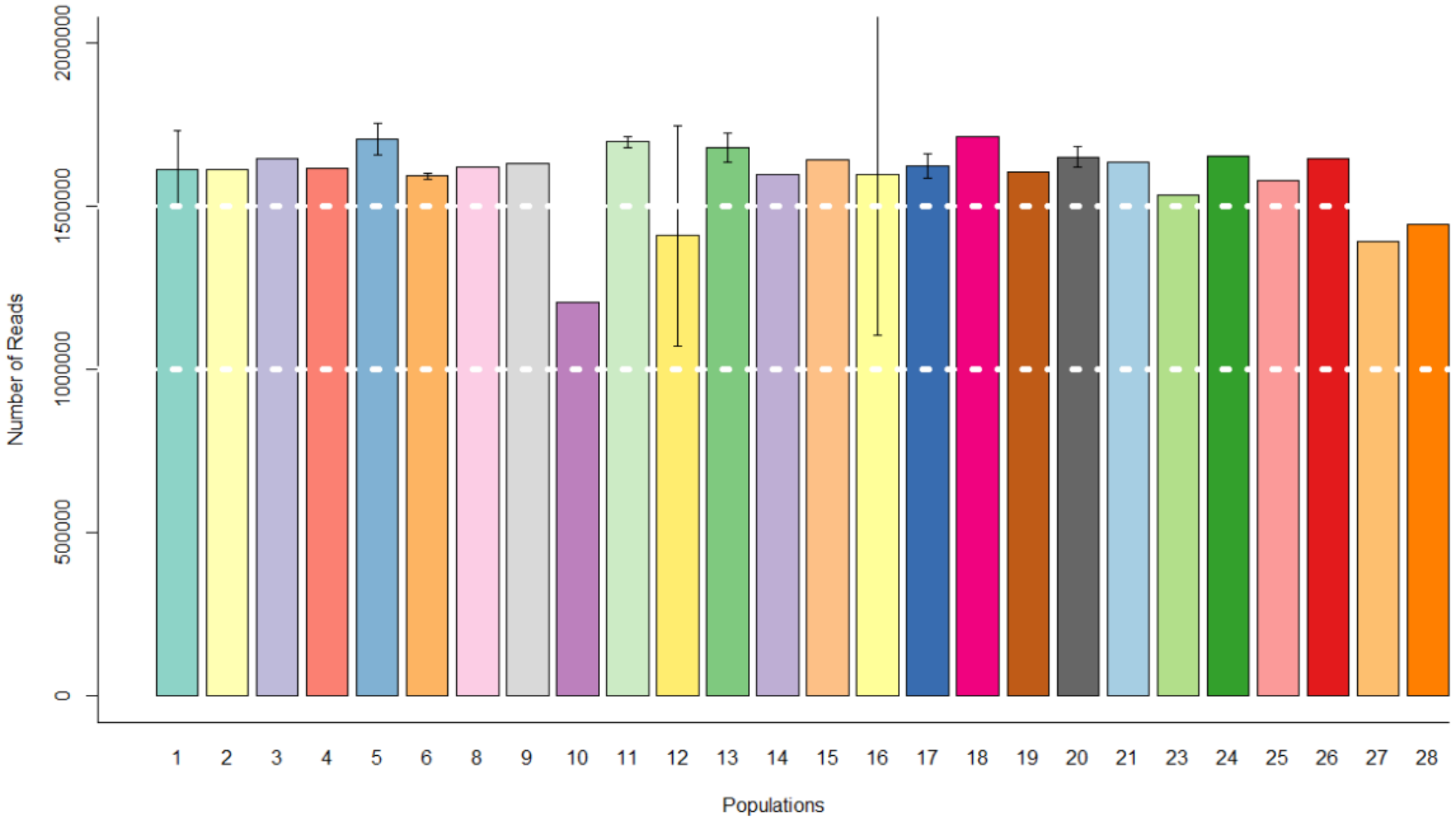
Visualization



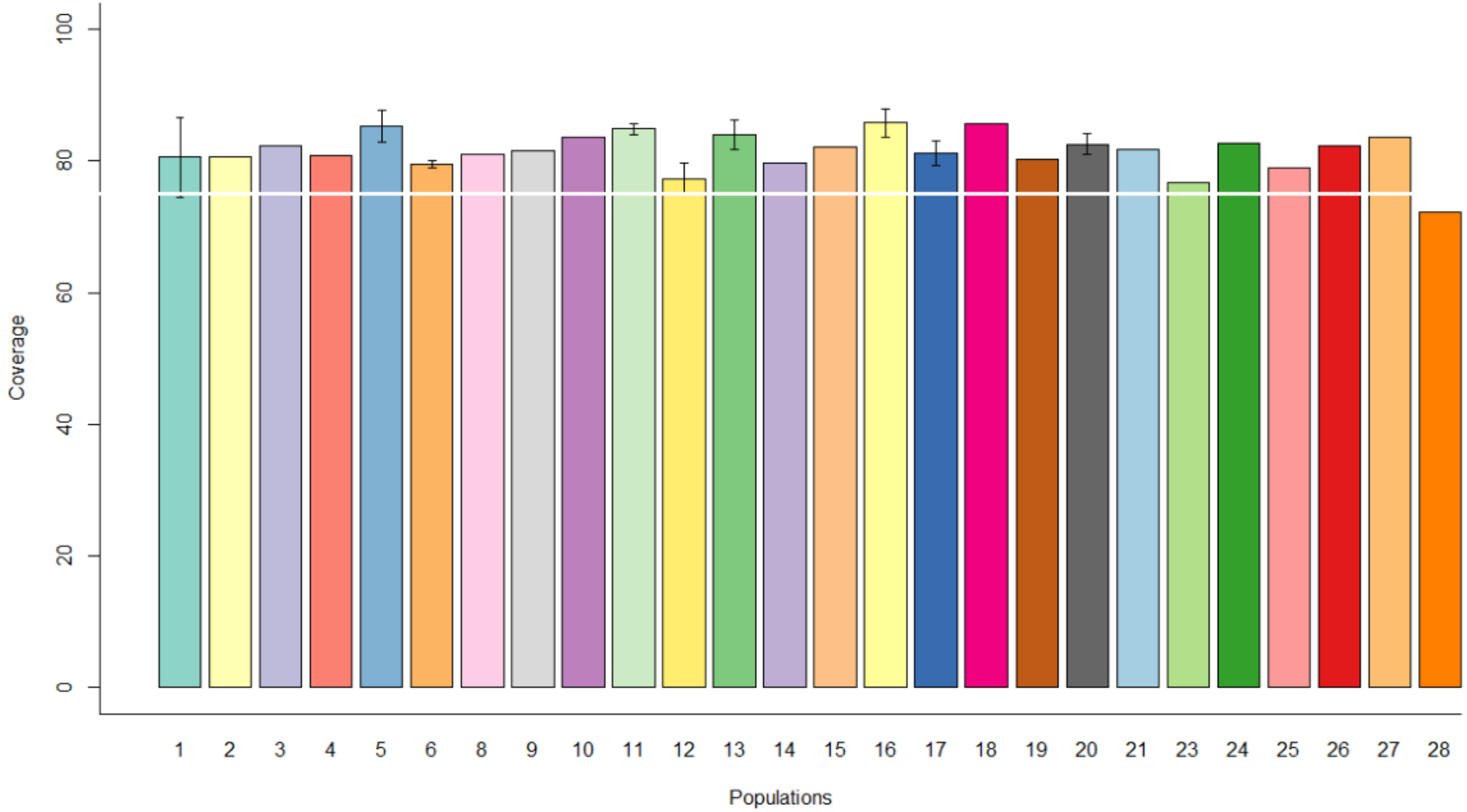
Visualization



Visualization



Visualization



Next Steps RADSeq

- Analyze output from denovo_map.pl
- Determine why it was failing outside the pipeline
- Take the data and make a tree
- Compare that tree to one generated in the paper I’m following
- Try it all again with Ipyrad and see if it gives similar trees

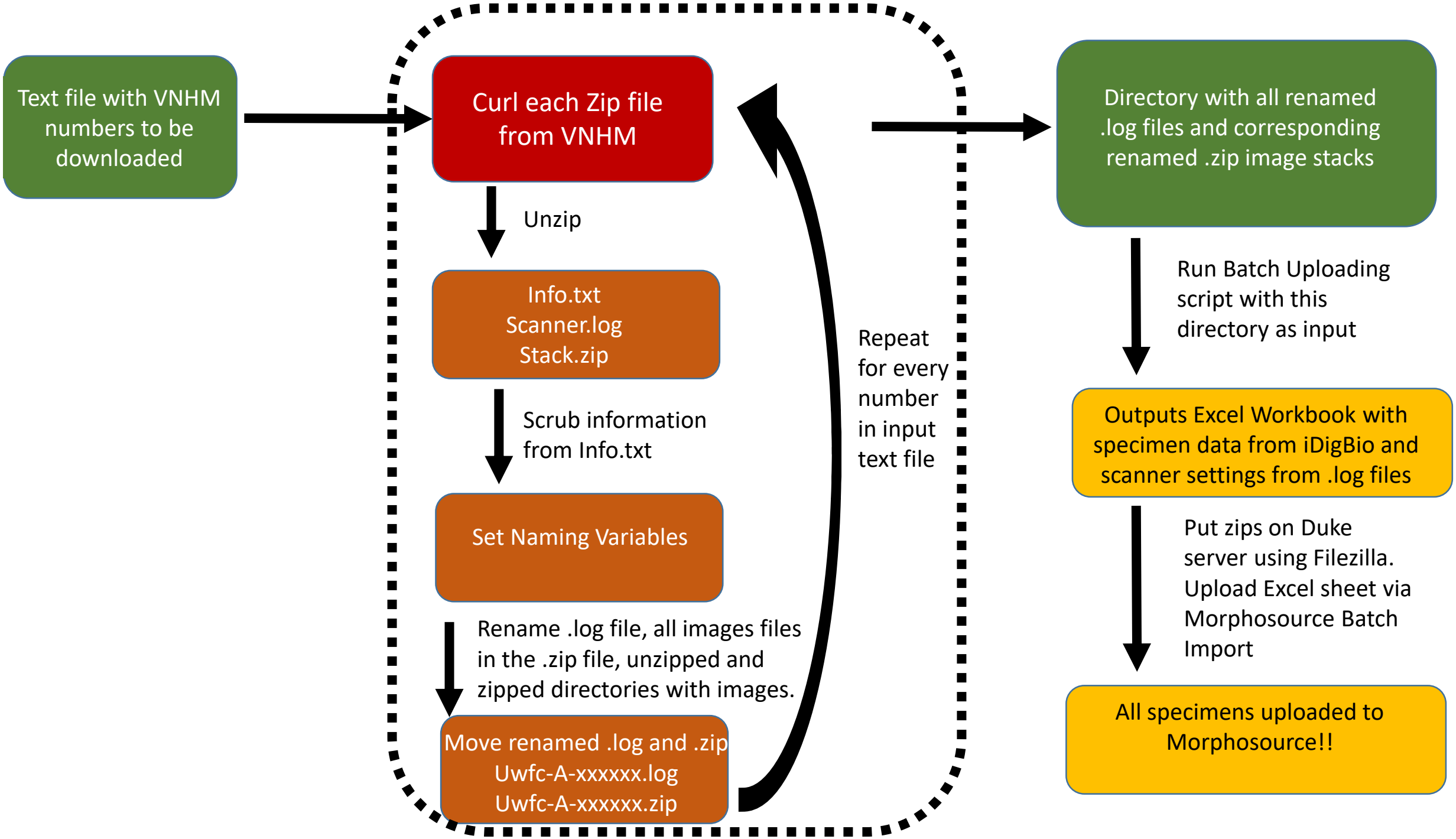
Bonus Question:

Can I use the command line to automate downloading and renaming CT scanning files to save time and remove user error?

Bonus Question:

Can I use the command line to automate downloading and renaming CT scanning files to save time and remove user error?

Spoiler Alert! YES!!


Text file with VNHM numbers to be downloaded
Curl each Zip file from VNHM
Unzip
Info.txt
Scanner.log
Stack.zip
Scrub information from Info.txt
Set Naming Variables
Rename .log file, all images files in the .zip file, unzipped and zipped directories with images.
Move renamed .log and .zip
Uwfc-A-xxxxxx.log
Uwfc-A-xxxxxx.zip
Repeat for every number in input text file
Directory with all renamed .log files and corresponding renamed .zip image stacks
Run Batch Uploading script with this directory as input
Outputs Excel Workbook with specimen data from iDigBio and scanner settings from .log files
Put zips on Duke server using Filezilla. Upload Excel sheet via Morphosource Batch Import
All specimens uploaded to Morphosource!!