FISH 546: Bioinformatics

Final presentation

Hyeon Jeong Kim

Devember 6, 2018

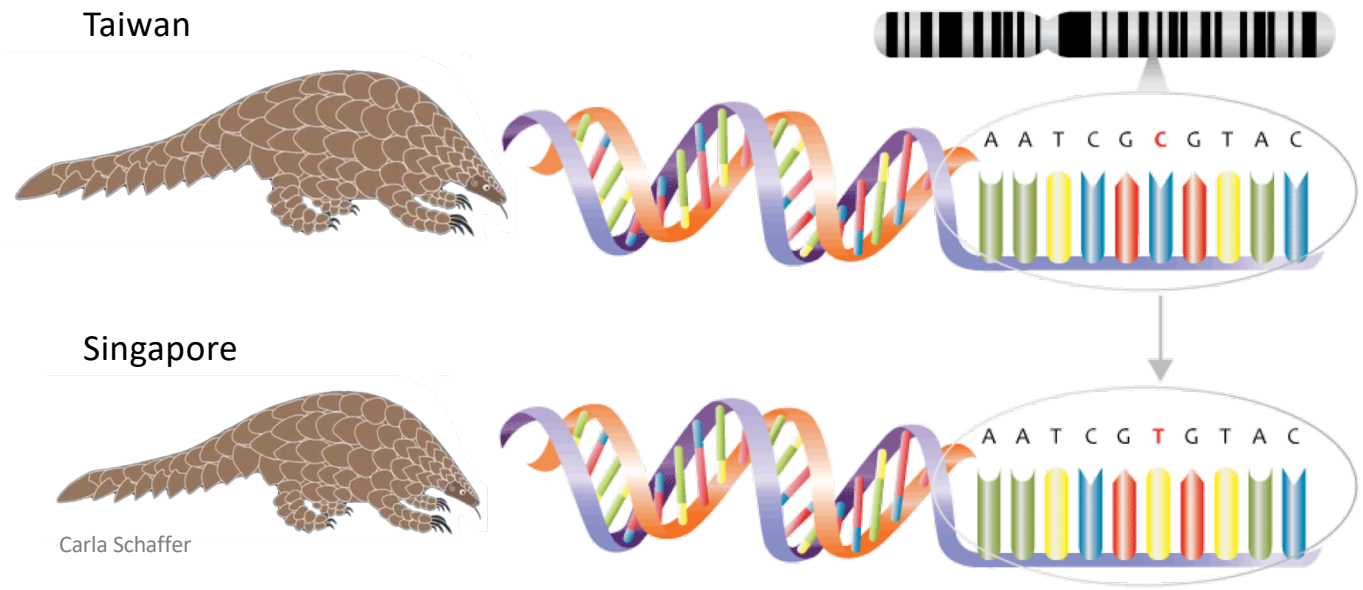
Question:

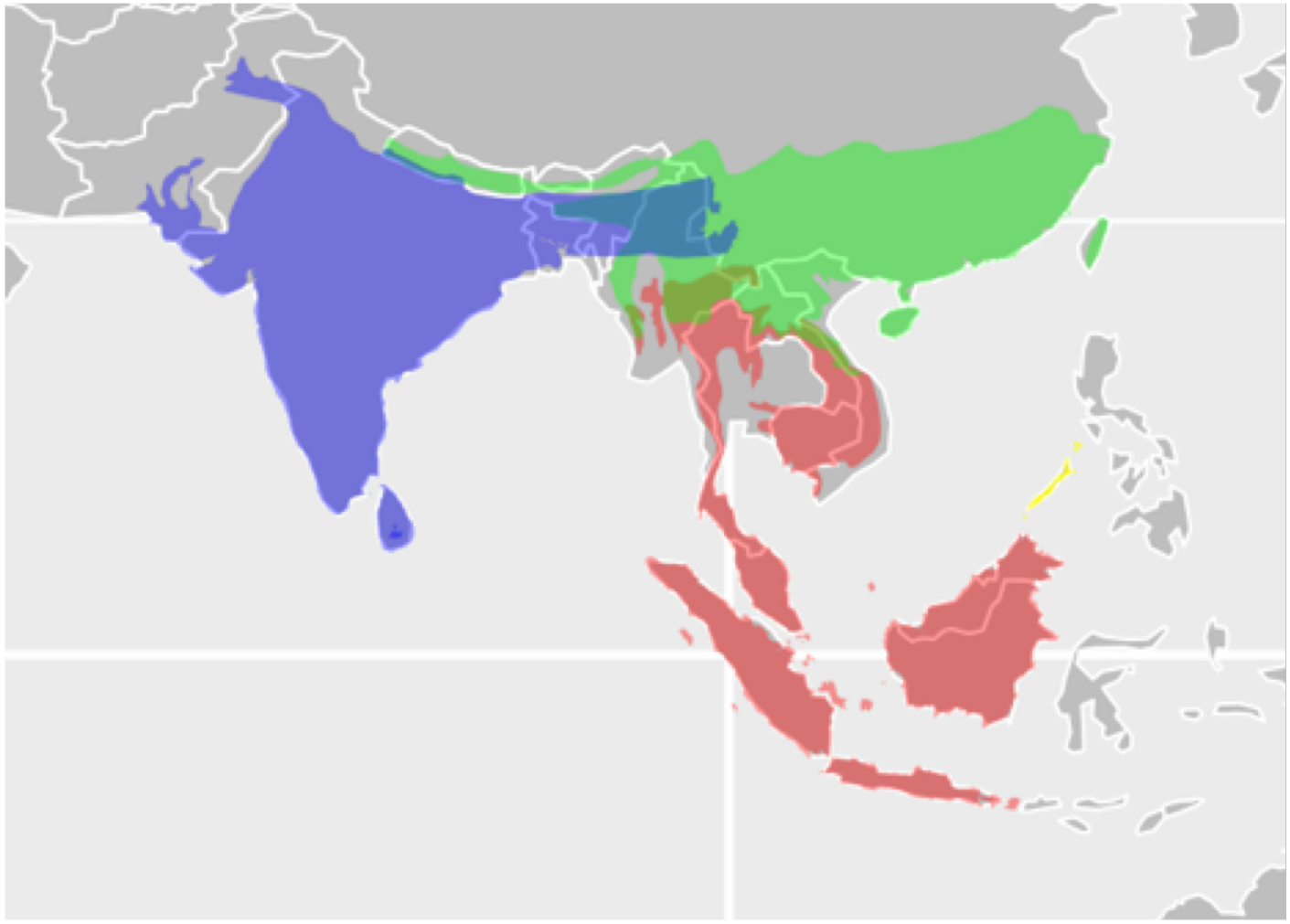
- How genetically distinct are pangolin populations within each species?
  - Can we use this information to conduct genetic assignment of seized pangolins?

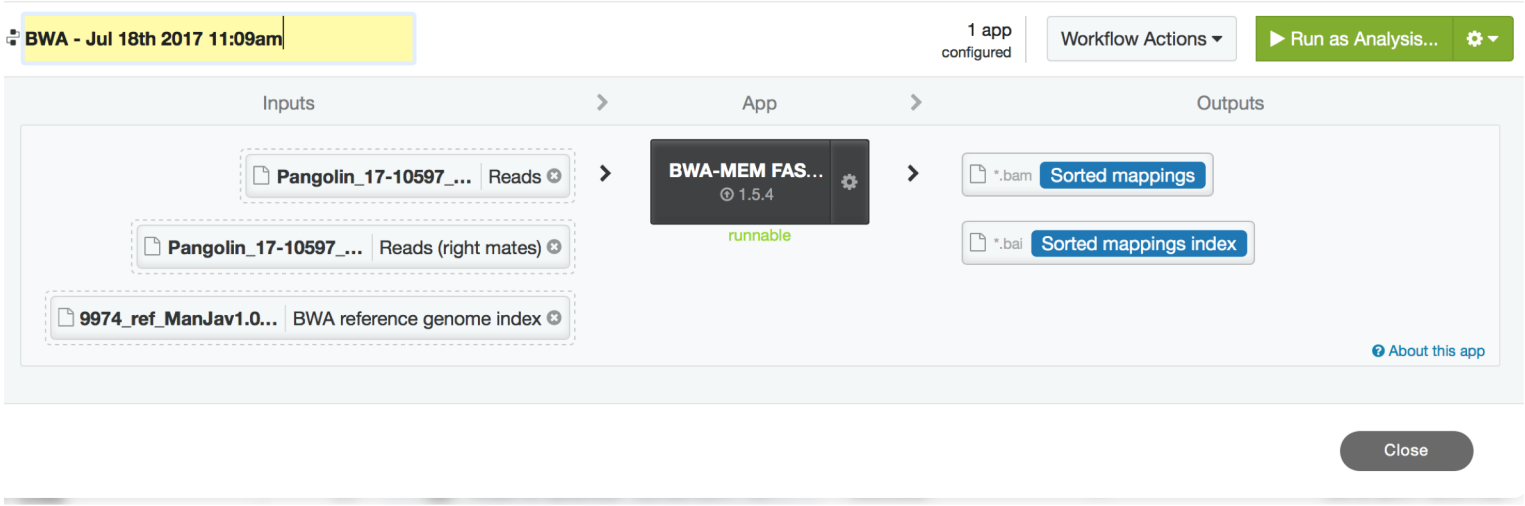
Pangolin scale burn in Cameroon.
Kenneth Cameron / USFWS / CC BY 2.0

Question:

Are Chinese and Sunda pangolins genetically distinct?





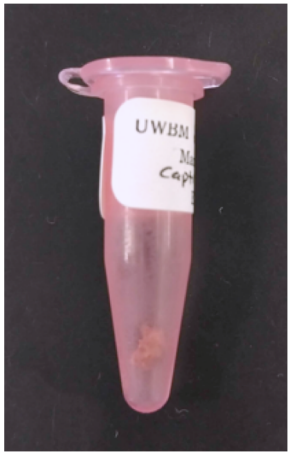


Sub-goal:

Create an easy to reproduce pipeline for variant calling

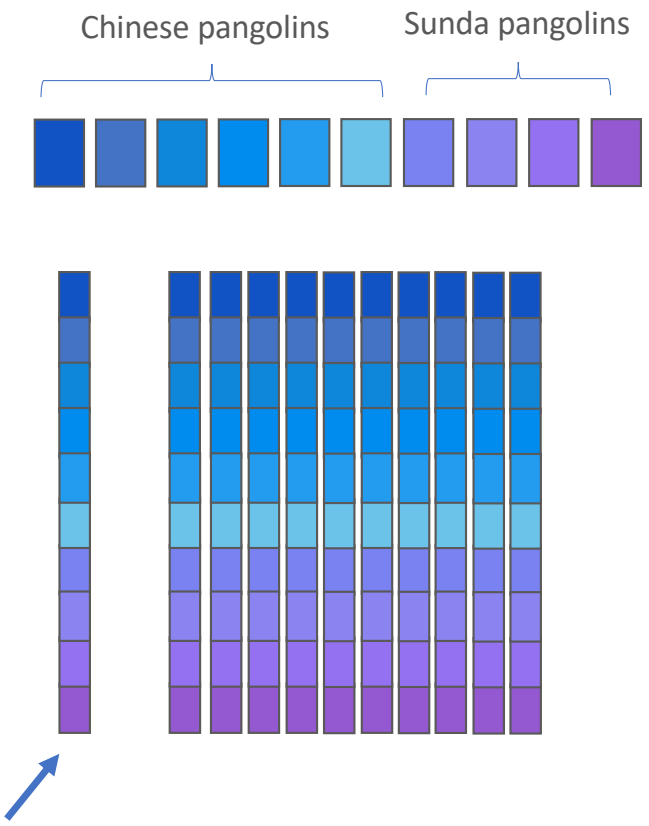
```
#!/bin/bash
## Job Name
#SBATCH --job-name=fish-angsd-run3-admix
## Allocation Definition
## On mox and ikt, the account and partition options should be the same.
#SBATCH --account=stf
#SBATCH --partition=stf
## Resources
## Nodes
#SBATCH --nodes=1
## Walltime (3 hours). Do not specify a walltime substantially more than your job needs.
#SBATCH --time=24:00:00
## Memory per node. It is important to specify the memory since the default memory is very small.
## For mox, --mem may be more than 100G depending on the memory of your nodes.
## For ikt, --mem may be 58G or more depending on the memory of your nodes.
## See above section on "Specifying memory" for choices for --mem.
#SBATCH --mem=100G
## Specify the working directory for this job
#SBATCH --workdir=/gscratch/stf/kimh11/bwa
##turn on e-mail notification
#SBATCH --mail-type=ALL
#SBATCH --mail-user=kimh11@uw.edu
## export all your environment variables to the batch job session
#SBATCH --export=all


module load contrib/angsd/0.921
FILTERS="-uniqueOnly 1 -remove_bads 1 -minMapQ 25 -minQ 25 -dosnpstat 1 -doHWE 1 -sb_pval 1e-5 -hetbias_pval 1e-5 -skipTriallelic 1 -minInd
 7 -snp_pval 1e-5 -minMaf 0.1"
TODO="-doMajorMinor 1 -doMaf 1 -doCounts 1 -makeMatrix 1 -doIBS 1 -doCov 1 -doGeno 8 -doVcf 1 -doPost 1 -doGlf 2"

angsd -b bams -GL 1 $FILTERS $TODO -P 1 -out angsd-run3-admix
~
```

Sequencing



Tissue n = 10

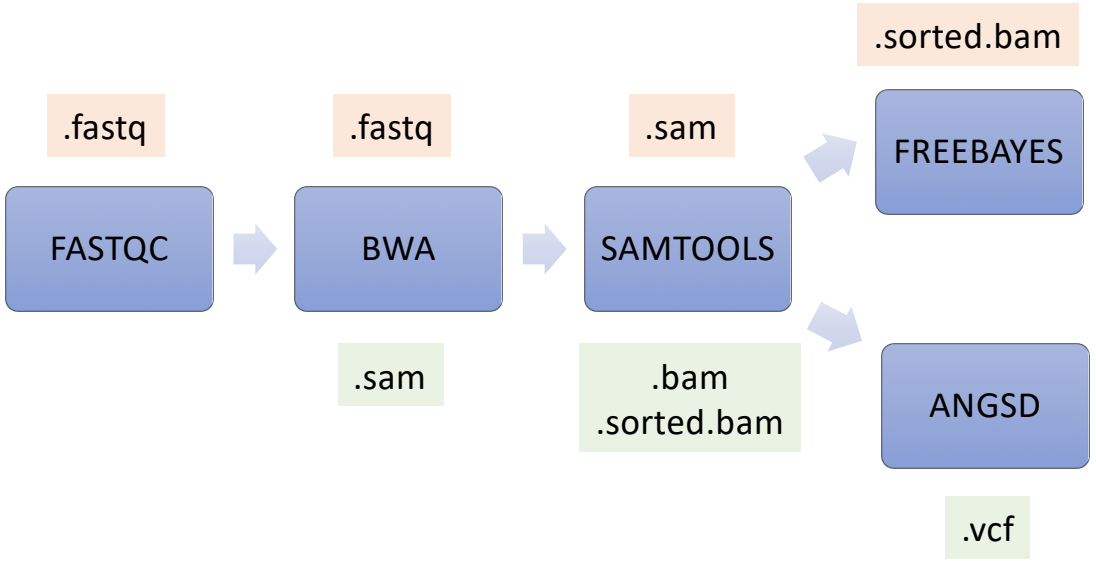
- 4 *Manis javanica* (Sunda pangolins)
- 6 *Manis pentadactyla* (Chinese pangolins)



Bioinformatic pipeline

Input:

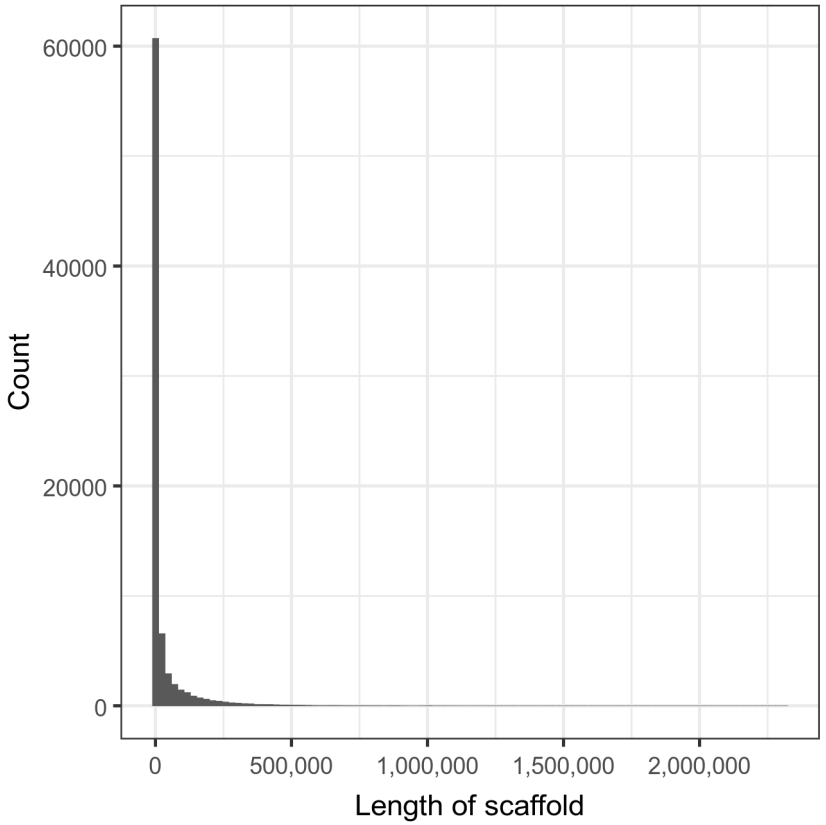
Output:



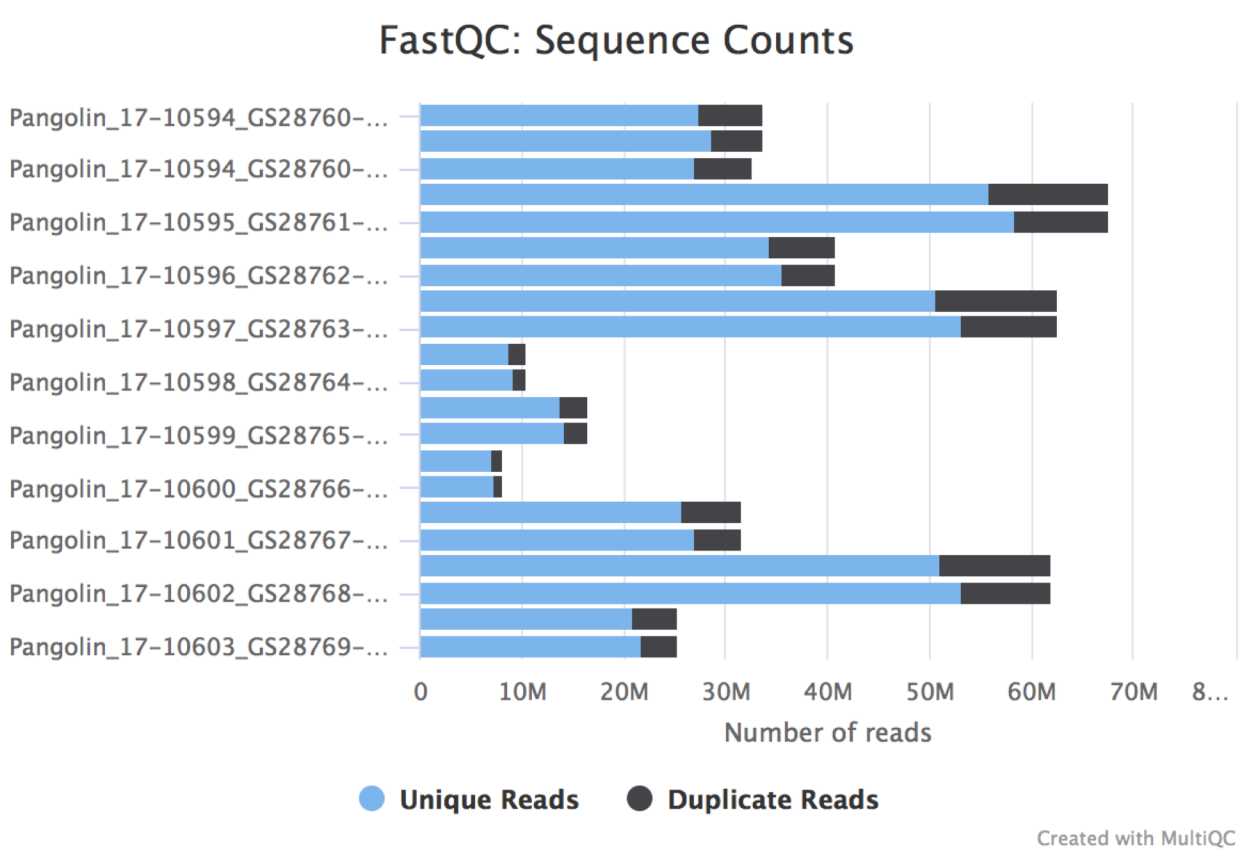
Manis javanica reference genome

- Number of scaffold : 80670

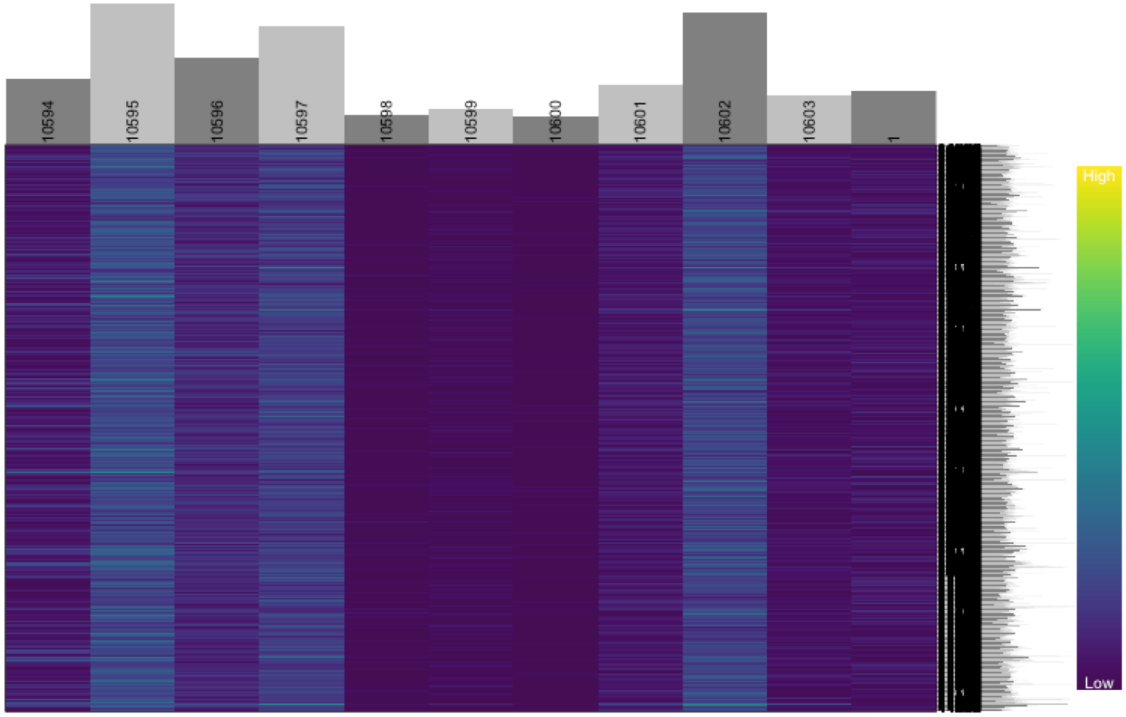
- Longest scaffold: 2314370
- Shortest scaffold: 1000



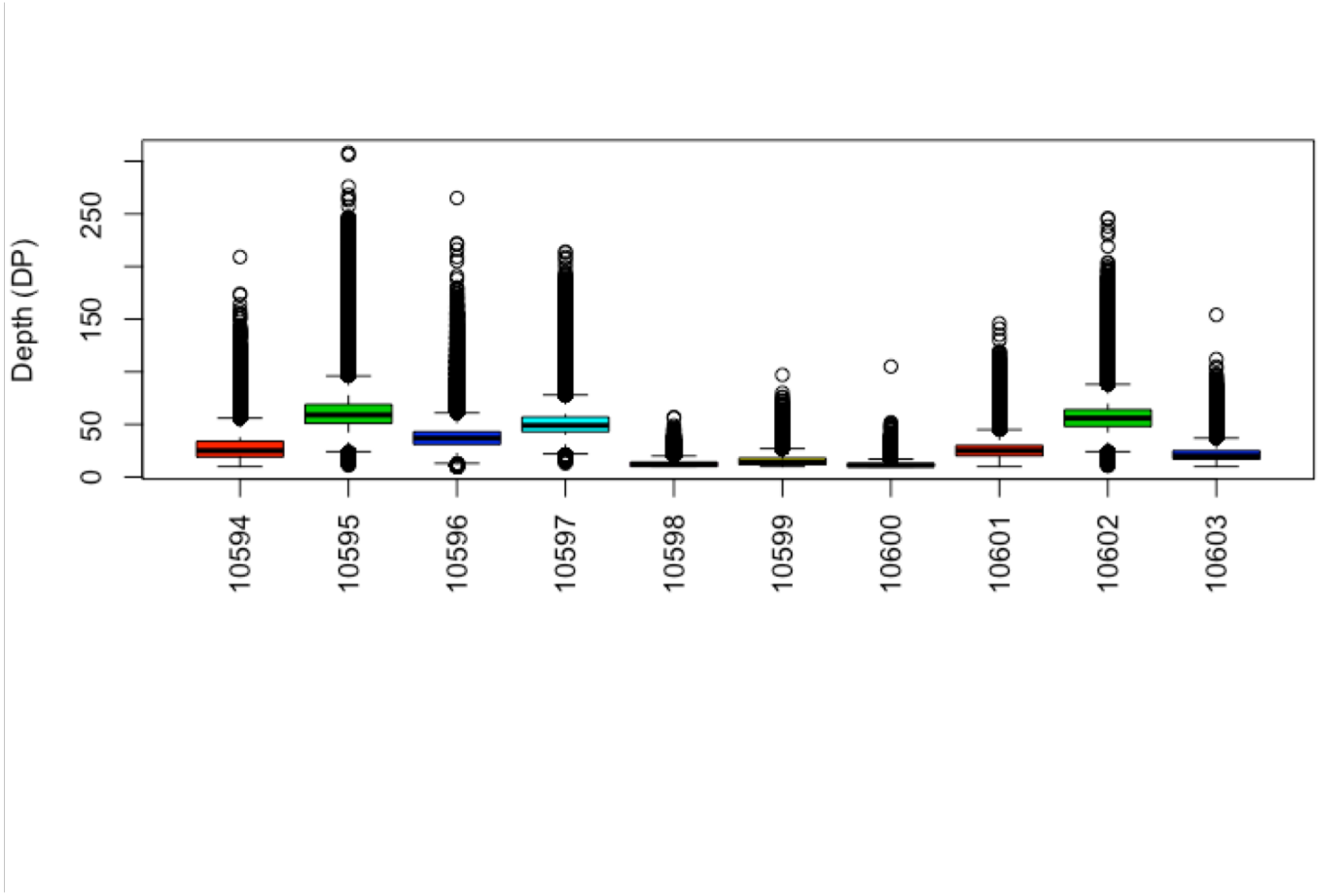
My sample quality



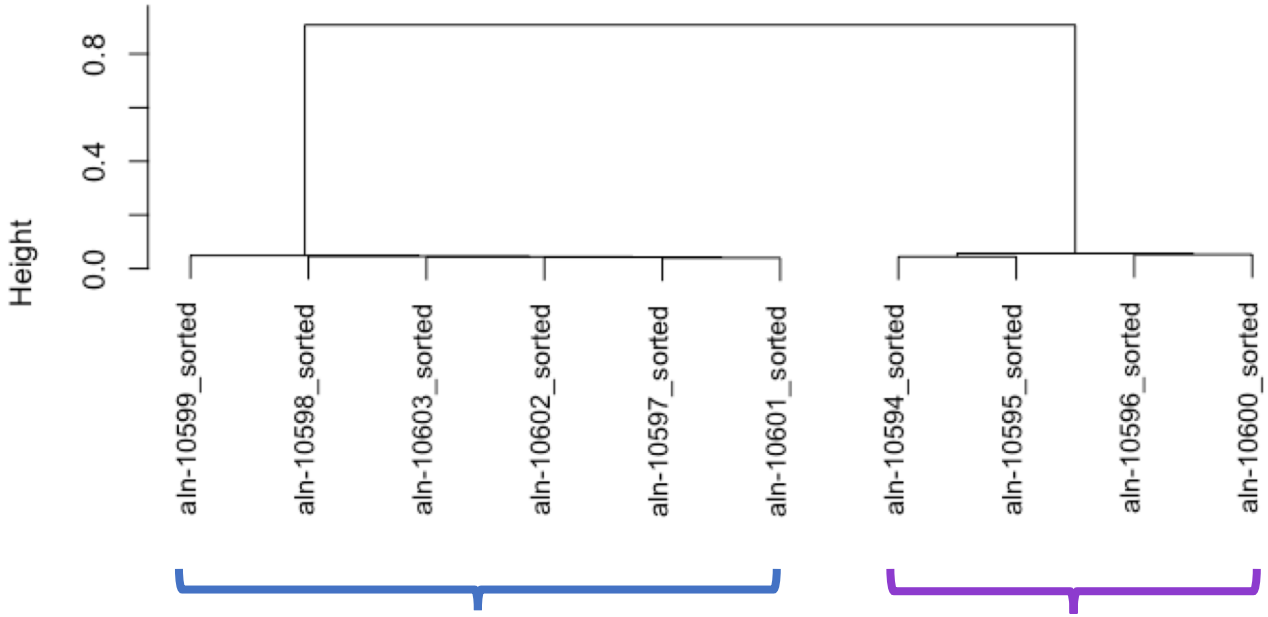
Results


10594
10595
10596
10597
10598
10599
10600
10601
10602
10603
1
High
Low

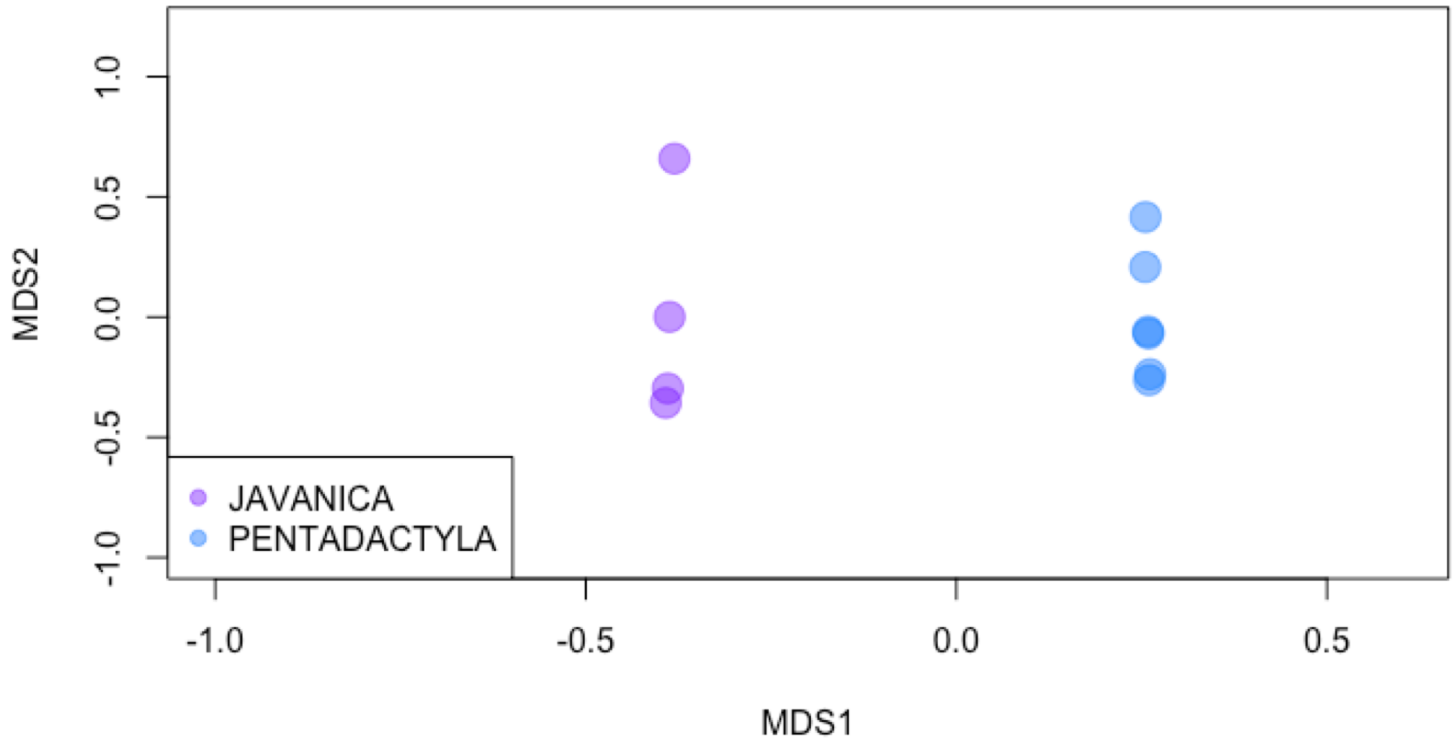
Results


Depth (DP)
0
50
150
250
10594
10595
10596
10597
10598
10599
10600
10601
10602
10603

Results


Height
0.0
0.4
0.8
aln-10599_sorted
aln-10598_sorted
aln-10603_sorted
aln-10602_sorted
aln-10597_sorted
aln-10601_sorted
aln-10594_sorted
aln-10595_sorted
aln-10596_sorted
aln-10600_sorted
Chinese pangolins
Sunda pangolins

Results



Next steps

1. Sequence museum samples
2. Re-run the analysis using the full dataset
3. Figure out a SNP filtering pipeline