Basics of genomics & quality control: II

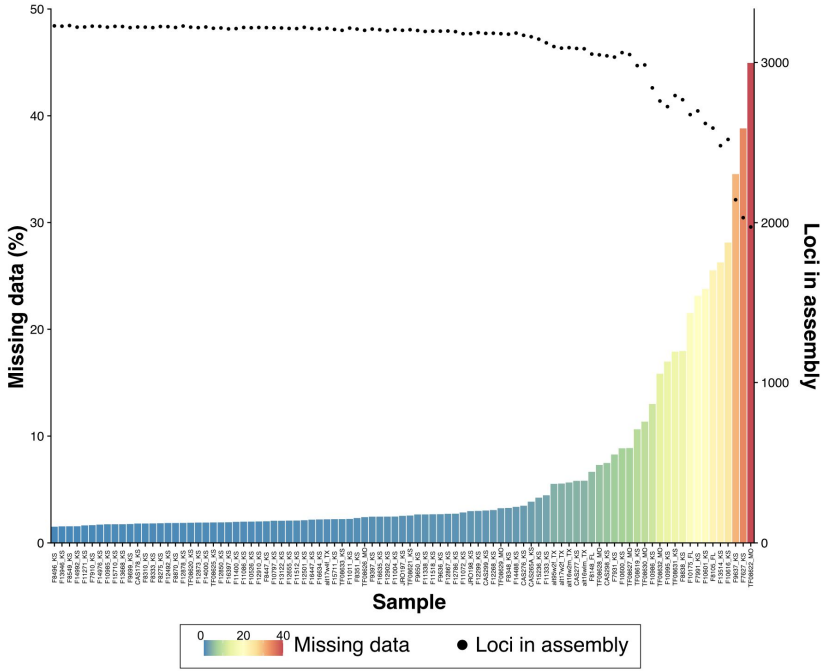
1st EvoGenomics Methods Workshop

Anne Chambers (eachambers@berkeley.edu)

Quality control

There are several key data visualizations that can help us troubleshoot or identify potential “red flags” in our data before moving forward with analyses

Distribution of **basic summary statistics**



Quality control

There are several key data visualizations that can help us troubleshoot or identify potential “red flags” in our data before moving forward with analyses

Distribution of **basic summary statistics**

PCA for dimensional reduction

What types of things might correspond to different PCA groupings?

Quality control

There are several key data visualizations that can help us troubleshoot or identify potential “red flags” in our data before moving forward with analyses

Distribution of **basic summary statistics**

PCA for dimensional reduction

What types of things might correspond to different PCA groupings?

- Libraries
- Samples with very high amounts of missing data
- Might have real biological significance: population structure, presence of inversions, etc.

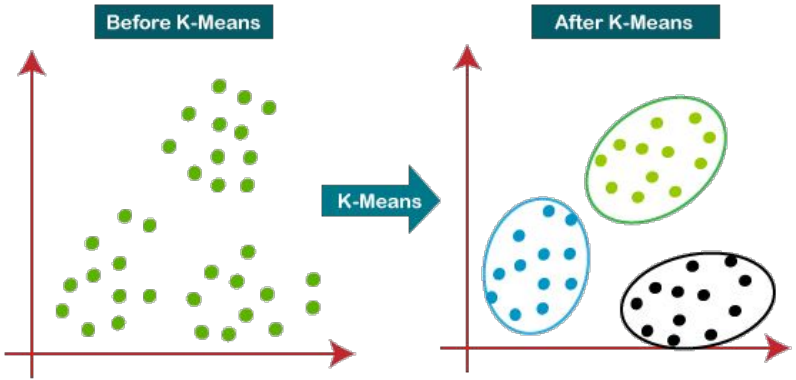
Quality control

There are several key data visualizations that can help us troubleshoot or identify potential “red flags” in our data before moving forward with analyses

Distribution of **basic summary statistics**

PCA for dimensional reduction

Basic **clustering** or population structure analyses



https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Admixture

admixture is similar to STRUCTURE, fastSTRUCTURE, etc.

Model-based clustering method, agnostic to sampling

User-specified number of clusters (*K*), method will cluster samples accordingly

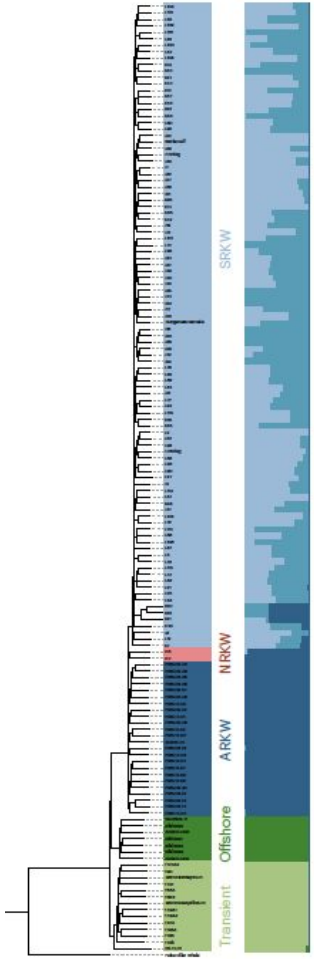


Figure from Kardos et al. (2023) https://doi.org/10.1038/s41559-023-01995-0

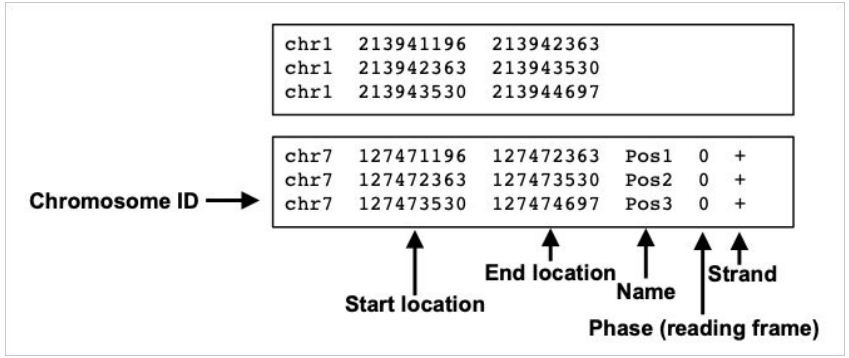
Admixture

admixture is similar to STRUCTURE, fastSTRUCTURE, etc.

Model-based clustering method, agnostic to sampling

User-specified number of clusters (*K*), method will cluster samples accordingly

```
admixture file.bed 5
```



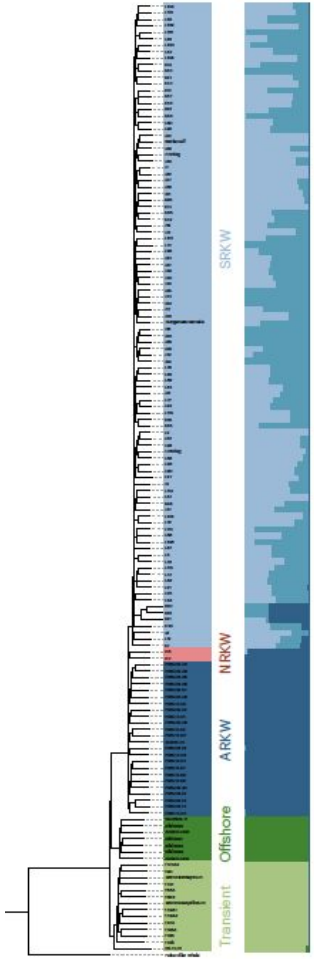
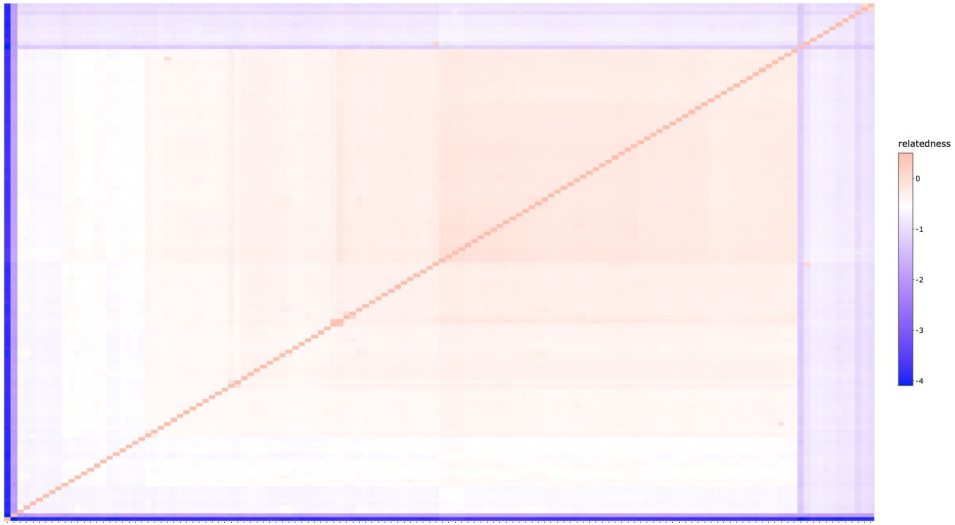


Figure from Kardos et al. (2023) https://doi.org/10.1038/s41559-023-01995-0

Other good QC visualizations that can be used

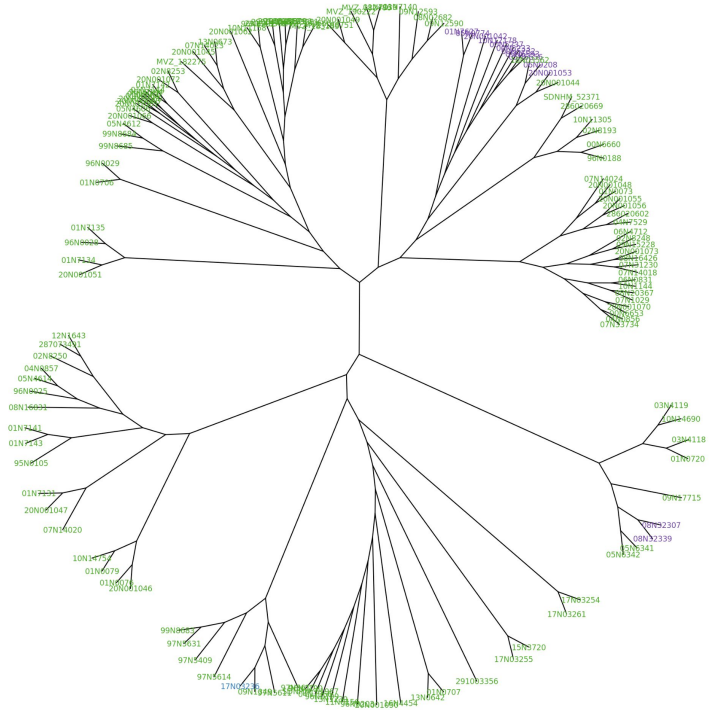
Relatedness plot




```
plink2 --vcf file.vcf --out outfile_name --make-king-square
```

https://www.cog-genomics.org/plink/2.0/distance

Neighbor-joining tree



EXERCISE 2: data visualization for quality control

Download the worksheet here:
https://github.com/eachambers/EvoGeno-Methods-Workshop/blob/main/Workshop1/Exercises/EvoGenomics_Ws1_Ex2.txt

Download the R script here:
https://github.com/eachambers/EvoGeno-Methods-Workshop/blob/main/Workshop1/Exercises/Workshop1_Exercise2.R

```
# ================================================================================
#                                    EXERCISE 2
# ================================================================================

# ================================ 1. RUNNING ADMIXTURE ================================

# Admixture takes in a bed file, which you should have created using plink. We're going to
# run it on both our LD-pruned and un-pruned datasets just to see how the results compare.
# Sometimes, chromosome names aren't supported by admixture, and this is the case with the
# Lampropeltis data. The problem is with the bim file, which is the file that accompanies
# a bed file. The first column of the bim file needs to be the chromosome code, which must
# be 'X'/'Y'/'XY'/'MT'; '0' indicates unknown. If you take a look at the bim files that were
# made by Plink in exercise 1, you'll notice that the first column has RAD tag numbers. The
# easiest way to get around this is to replace all of these RAD tag numbers with 0s. To do
# so, run the following on **BOTH** your pruned and unpruned bim files:

$ awk '{$1=0;print $0}' file.bim > file.bim.tmp # adds zeros to first column
$ mv file.bim.tmp file.bim # replaces original bim file

#       (1a) Run admixture with K=3 for both the pruned and unpruned bed files.
#               *** YOUR ANSWER HERE ***
```

```
# === 2. DATA VISUALIZATION OF SUMMARY STATISTICS ===

# Be sure to set your working directory (wherever you saved your output files!)
# and load relevant libraries. The following questions can be answered (and
# visualized using only the tidyverse package).

# 2a ----------------------------------------------------------------------

#   Using the missing data stats you generated in step 1.3a, plot the data however you
#   would like. What is the average proportion of missing data across the entire
#   dataset?
# TODO: **YOUR ANSWER HERE**


# 2b ----------------------------------------------------------------------

#   Using the depth statistics you generated in step 1.3b, plot the data in whichever
#   way you would like. What is the average read depth across the entire dataset?
# TODO: **YOUR ANSWER HERE**
```

Further processing for downstream population genetics analyses

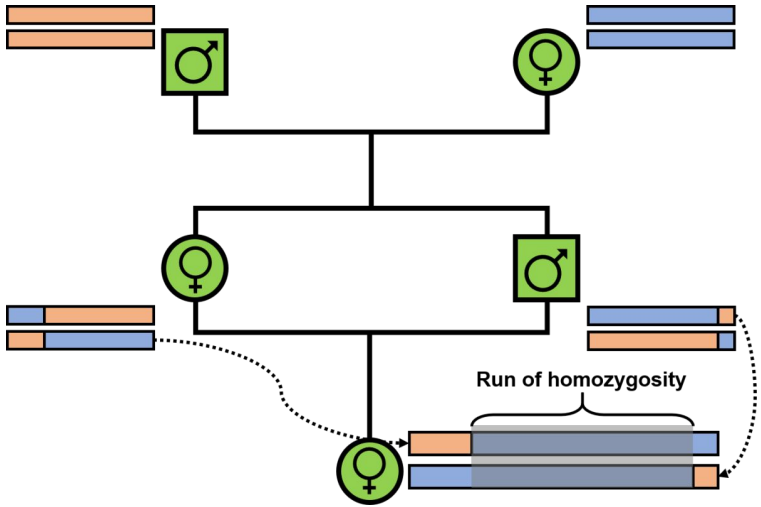
- Genetic distance matrix (plink)

```
plink --vcf file.vcf --out prefix_name --distance square
```

Further processing for downstream population genetics analyses

- Genetic distance matrix (plink)
- Runs of homozygosity (bcftools or plink)

```
bcftools roh --threads 10 -G30 --AF-dflt 0.4 -O z -o output.roh file.vcf
```

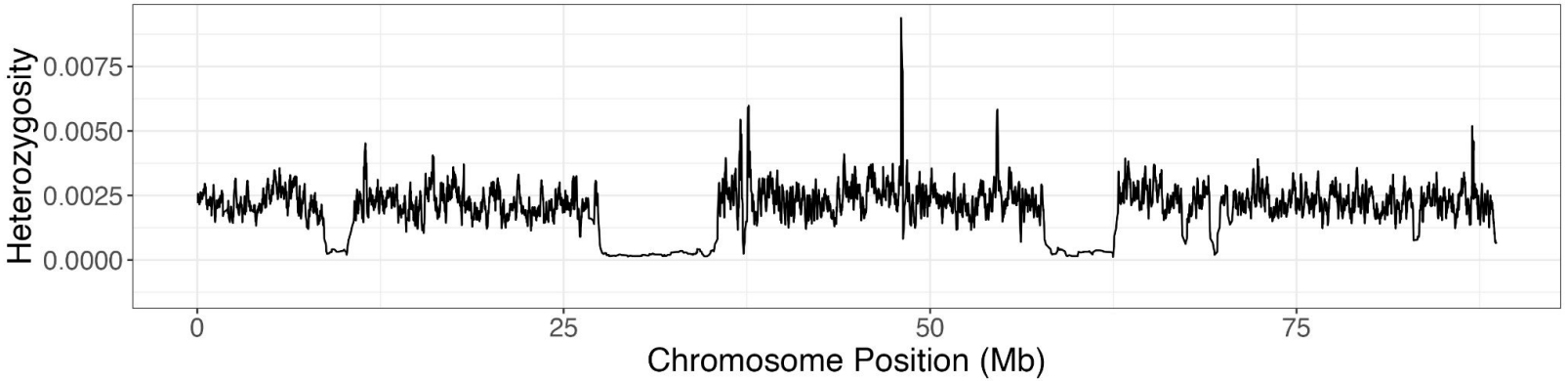


https://theg-cat.com/2022/06/02/runaway-inbreeding-how-runs-of-homozygosity-impact-conservation-management/

Further processing for downstream population genetics analyses

- Genetic distance matrix (plink)
- Runs of homozygosity (bcftools or plink)

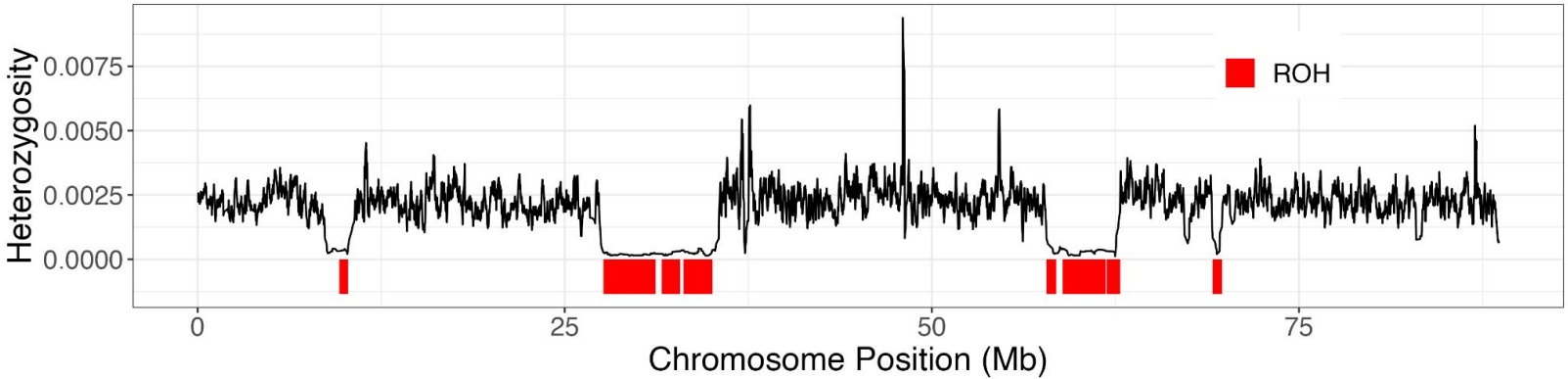
```
bcftools roh --threads 10 -G30 --AF-dflt 0.4 -O z -o output.roh file.vcf
```



Further processing for downstream population genetics analyses

- Genetic distance matrix (plink)
- Runs of homozygosity (bcftools or plink)

```
bcftools roh --threads 10 -G30 --AF-dflt 0.4 -O z -o output.roh file.vcf
```



Further processing for downstream population genetics analyses

- Genetic distance matrix (plink)
- Runs of homozygosity (bcftools or plink)
- Imputation of missing genotypes