

---

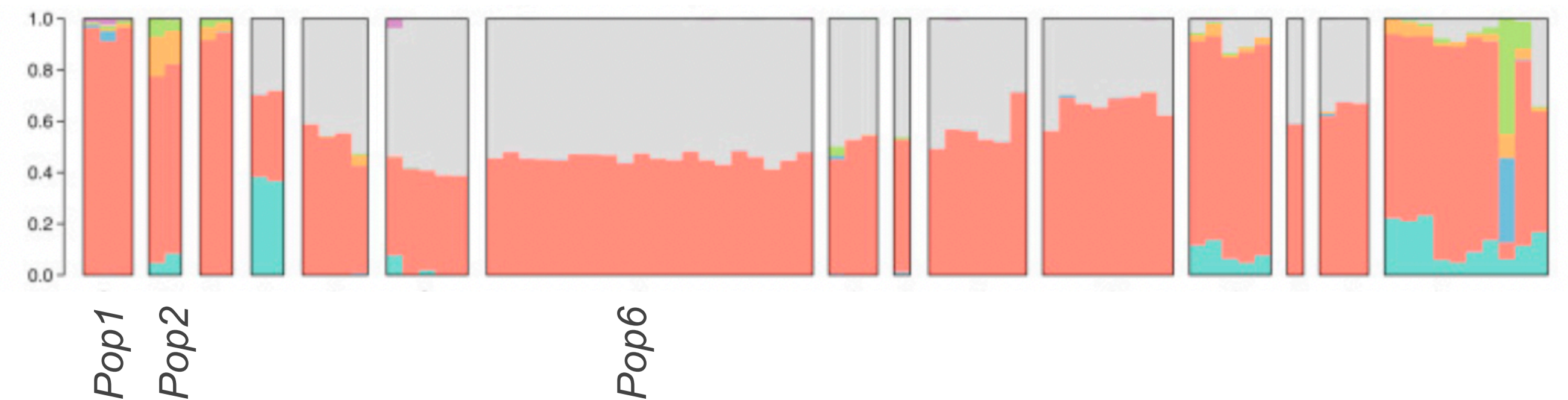
## Week-5

# ADMIXTURE

---

**Aim:** Learning how to run an ADMIXTURE analysis, running a simple analysis, introduction to Pong and Clumpp

**Hands-on:** Running ADMIXTURE software and plotting the results in R.



---

**Reading suggestions:** Alexander, Novembre and Lange (2009) [10.1101/gr.094052.109](https://doi.org/10.1101/gr.094052.109)

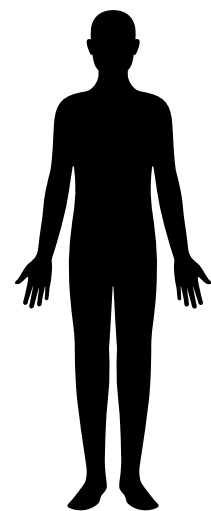
# Genetic admixture

*Combinations of two or more previously isolated ancestry sources*



## ADMIXTURE software

*Maximum likelihood estimation of ancestries*



- Individual is admixed: has genetic ancestry from multiple populations
- Estimate the ancestry: Model individual's genotype as a mixture of “K” clusters of populations : How?
- Define the clusters by genotype frequencies + use an estimator to discover contribution of each cluster to the individual genotype

# ADMIXTURE software

## *Running the software*

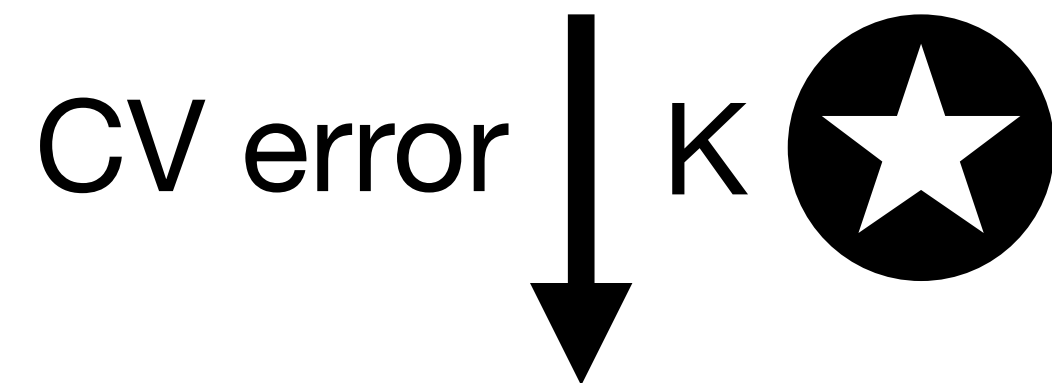
***How do we run the analysis?*** *admixture*

```
admixture filename.bed #Number_of_clusters
```

***How to decide for correct number of K?*** *cross validation*

```
admixture filename.bed #Number_of_clusters --cv
```

Default: 5-fold



```
for i in {1..10}; do admixture filename.bed ${i} --cv | tee log${i}.out; done
```

```
grep -h CV log*.out
```

```
CV error (K=1): XXX CV error (K=2): XXX
```

# ADMIXTURE

## *Plotting the results*

---

```
R  
> x=read.table("result.Q")  
> barplot(t(as.matrix(x)),xlab="Individual #", ylab="Ancestry", border=NA)
```



# ADMIXTURE - cross validation to pick the best K

*Example to cross-validation error*

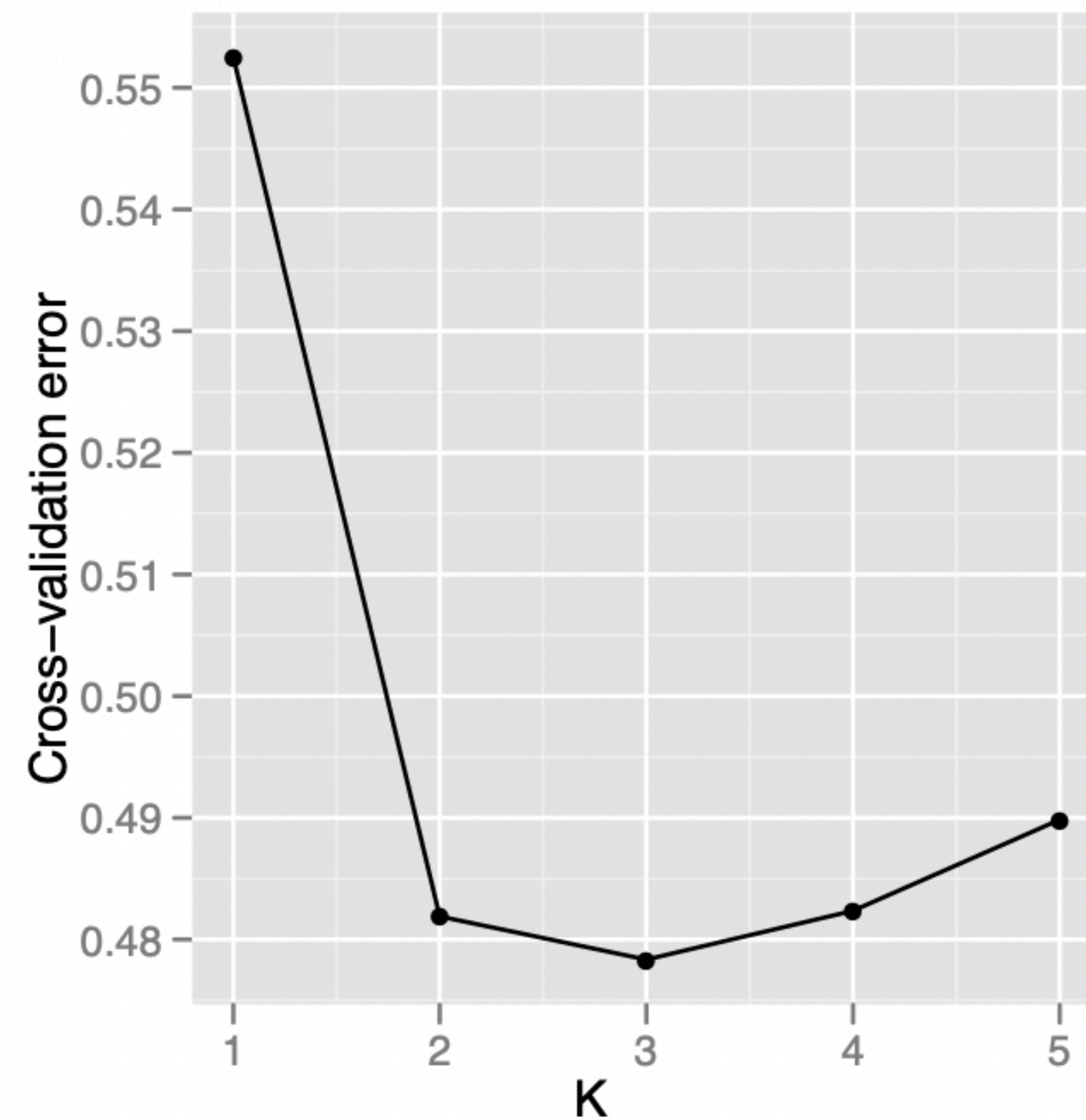
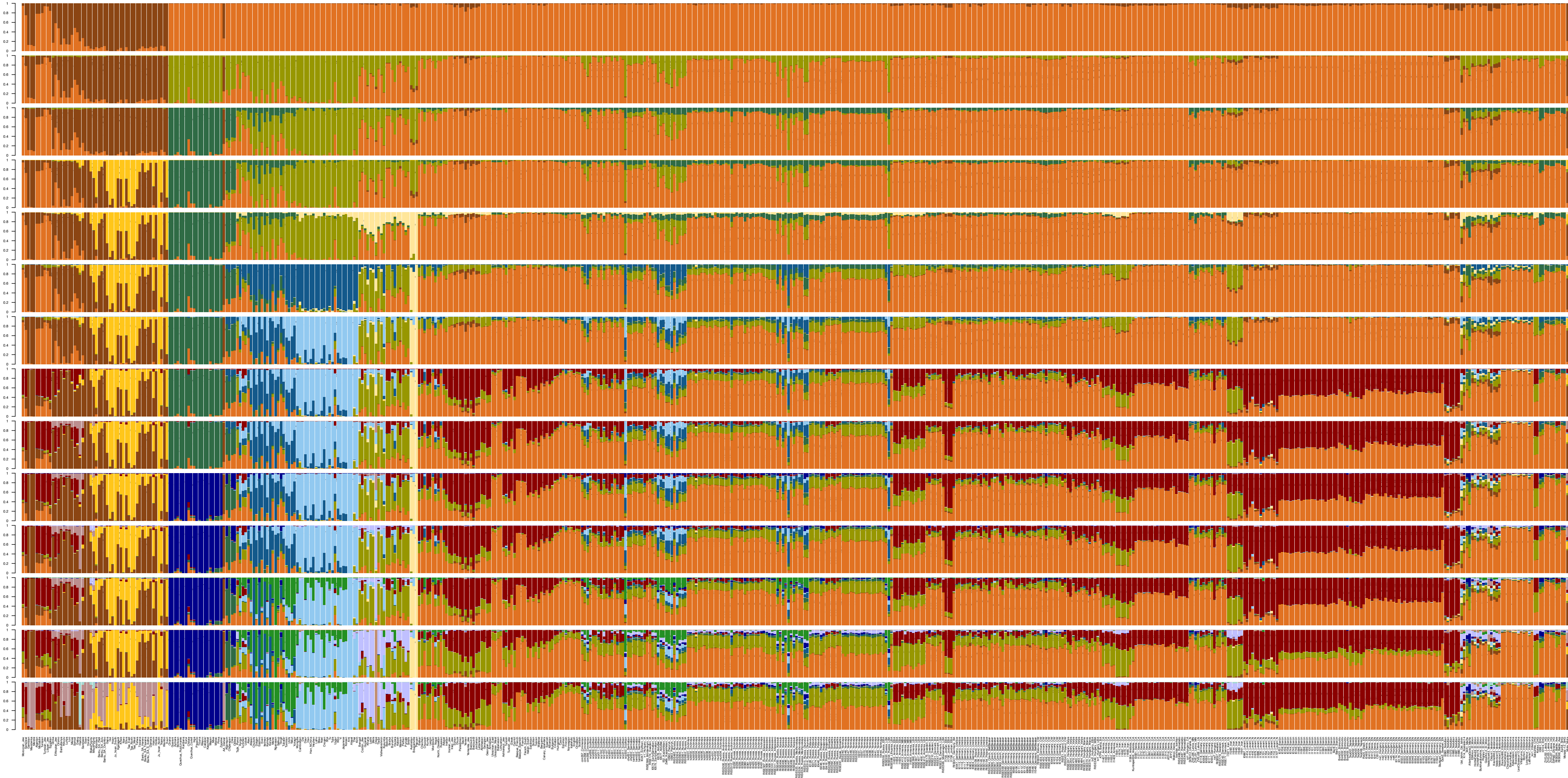


Figure 1: Cross-validation plot for the hapmap3 dataset

<http://dalexander.github.io/admixture/admixture-manual.pdf>

# ADMIXTURE

*From  $K=2$  to  $K=N$  Run for each  $K$   $N$  times using random seed + find common signals Clumpp & PONG*



# ADMIXTURE

*From  $K=2$  to  $K=N$  Run for each  $K$   $N$  times using random seed + find common signals Clumpp & PONG*

## *How to run with a random seed -s*

admixture filename.bed #Number\_of\_clusters -s \${RANDOM}

## *How to run it multiple times?*

```
for i in {2..K}; do mkdir K${i}; done
for k in {2..K}; do (for i in {1..N}; do mkdir ./K${k}/run${i}; done);done
#K= Number of Ks
#N= Number of runs
```

**for k in {2..K}; do (for i in {1..N}; do cp admixture.sh ./K\${k}/run\${i}; done);done # copy your script under each directory**

**#Run ADMIXTURE**

**#Update K**

**#in each K subdirectory you created above run:**

**find . -type d -exec sh -c '(cd {} && ./admixture.sh K)' ';'**

#!/bin/bash

**#Update based on your file name**

K=\$1

**/path/to/admixture/software\*/admixture -s \$RANDOM Admixturedataset.geno\* \$K**

echo "----End of the Run---"



# ADMIXTURE

*Assumes linkage equilibrium between markers, pruning dataset based on LD is necessary*

---

***How to prune the dataset for LD? Using PLINK [LD: non-random association of alleles]***

plink --bfile yourfileprefix --indep-pairwise parameters

3 parameters:

1- window size

2- SNP number to shift the window

3-  $r^2$  value

Produces pruned set of markers in plink.prune.in file

plink --bfile yourfileprefix - -extract plink.prune.in --make-bed -- out XXX

measure of LD: +/- of a particular allele at locus X - +/- of a particular allele at locus Y



# CLUMPP and PONG

*Finding common signals between each run*

---

[https://rosenberglab.stanford.edu/software/CLUMPP\\_Manual.pdf](https://rosenberglab.stanford.edu/software/CLUMPP_Manual.pdf)

[https://link.springer.com/content/pdf/10.1007/978-1-0716-0199-0\\_4.pdf](https://link.springer.com/content/pdf/10.1007/978-1-0716-0199-0_4.pdf)