Mahdi_Anvari.compG_HW1_Shell

March 24, 2025

1 Computational Genomics - Homework 1

1.1 Shell Section:

[7]:

```
pwd
```

```
/mnt/d/Daneshga/Term 8/Computational Genomics/HW1
```

[8]:

```
ls
```

```
Mahdi_Anvari.compG_HW1_R.ipynb
Mahdi_Anvari.compG_HW1_Shell.ipynb
```

1.2 1.

[9]:

```
wget https://dalexander.github.io/admixture/binaries/admixture_linux-1.3.0.tar.
  ↪gz
wget https://dalexander.github.io/admixture/admixture-manual.pdf
wget https://dalexander.github.io/admixture/hapmap3-files.tar.gz
```

```
--2025-03-23 15:59:35--
https://dalexander.github.io/admixture/binaries/admixture_linux-1.3.0.tar.gz
Resolving dalexander.github.io (dalexander.github.io)… 185.199.111.153,
185.199.109.153, 185.199.110.153, …
Connecting to dalexander.github.io
(dalexander.github.io)|185.199.111.153|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 1916705 (1.8M) [application/gzip]
Saving to: ‘admixture_linux-1.3.0.tar.gz’

admixture_linux-1.3 100%[===================>]   1.83M   407KB/s    in 4.6s

2025-03-23 15:59:44 (407 KB/s) - ‘admixture_linux-1.3.0.tar.gz’ saved
[1916705/1916705]

--2025-03-23 15:59:44--  https://dalexander.github.io/admixture/admixture-
manual.pdf
```

```
Resolving dalexander.github.io (dalexander.github.io)… 185.199.111.153,
185.199.109.153, 185.199.110.153, …
Connecting to dalexander.github.io
(dalexander.github.io)|185.199.111.153|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 301291 (294K) [application/pdf]
Saving to: ‘admixture-manual.pdf’

admixture-manual.pd 100%[===================>] 294.23K  91.2KB/s    in 3.2s

2025-03-23 15:59:50 (91.2 KB/s) - ‘admixture-manual.pdf’ saved [301291/301291]

--2025-03-23 15:59:50--
https://dalexander.github.io/admixture/hapmap3-files.tar.gz
Resolving dalexander.github.io (dalexander.github.io)… 185.199.111.153,
185.199.109.153, 185.199.110.153, …
Connecting to dalexander.github.io
(dalexander.github.io)|185.199.111.153|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 1865216 (1.8M) [application/gzip]
Saving to: ‘hapmap3-files.tar.gz’

hapmap3-files.tar.g 100%[===================>]   1.78M   348KB/s    in 5.2s

2025-03-23 15:59:57 (348 KB/s) - ‘hapmap3-files.tar.gz’ saved [1865216/1865216]
```

[47]:

```
mkdir admixture_linux-1.3.0 hapmap3-files
tar -xzf admixture_linux-1.3.0.tar.gz -C admixture_linux-1.3.0/
tar -xf hapmap3-files.tar.gz -C hapmap3-files/
```

[48]:

```
ls
```

```
Mahdi_Anvari.compG_HW1_R.ipynb
admixture_linux-1.3.0.tar.gz
Mahdi_Anvari.compG_HW1_Shell.ipynb  hapmap3-files
admixture-manual.pdf                hapmap3-files.tar.gz
admixture_linux-1.3.0
```

[49]:

```
ls hapmap3-files
```

```
hapmap3.bed  hapmap3.bim  hapmap3.fam
hapmap3.map
```

[52]:

```
head hapmap3-files/hapmap3.bim
```

```
1       rs10458597      0       554484  0       2
1       rs12562034      0       758311  1       2
```

```
1       rs2710875       0       967643  1       2
1       rs11260566      0       1168108 1       2
1       rs1312568       0       1375074 1       2
1       rs35154105      0       1588771 0       2
1       rs16824508      0       1789051 1       2
1       rs2678939       0       1990452 1       2
1       rs7553178       0       2194615 1       2
1       rs13376356      0       2396747 1       2
```

[53]:
```
head hapmap3-files/hapmap3.fam
```

```
2431 NA19916 0 0 1 -9
2424 NA19835 0 0 2 -9
2469 NA20282 0 0 2 -9
2368 NA19703 0 0 1 -9
2425 NA19901 0 0 2 -9
2427 NA19908 0 0 1 -9
2430 NA19914 0 0 2 -9
2470 NA20287 0 0 2 -9
2436 NA19713 0 0 2 -9
2426 NA19904 0 0 1 -9
```

1.3 2.

[50]:
```
ls admixture_linux-1.3.0/dist/admixture_linux-1.3.0/
```

```
README.32.txt  admixture  admixture-

manual.pdf  admixture32
```

[51]:
```
cat admixture_linux-1.3.0/dist/admixture_linux-1.3.0/README.32.txt
```

```

As of version 1.2, admixture is compiled as a 64-bit binary.
admixture32 is a 32-bit version provided for compatibility with older
systems.

--Dave
```

[56]:
```
admixture_linux-1.3.0/dist/admixture_linux-1.3.0/admixture32 --help
```

```
****                   ADMIXTURE Version 1.3.0                  ****
****                    Copyright 2008-2015                     ****
****           David Alexander, Suyash Shringarpure,            ****
****                John  Novembre, Ken Lange                   ****
****                                                            ****
****                  Please cite our paper!                    ****
****   Information at www.genetics.ucla.edu/software/admixture  ****
```

```
ADMIXTURE basic usage:  (see manual for complete reference)
  % admixture [options] inputFile K

where:
  K is the number of populations; and
  inputFile may be:
    - a PLINK .bed file
    - a PLINK "12" coded .ped file

Output will be in files inputBasename.K.Q, inputBasename.K.P

General options:
  -jX          : do computation on X threads
  --seed=X     : use random seed X for initialization

Algorithm options:
   -m=
  --method=[em|block]     : set method.  block is default

   -a=
  --acceleration=none   |
                 sqs<X> |
                 qn<X>      : set acceleration

Convergence criteria:
  -C=X : set major convergence criterion (for point estimation)
  -c=x : set minor convergence criterion (for bootstrap and CV reestimates)

Bootstrap standard errors:
  -B[X]      : do bootstrapping [with X replicates]
```

1.4 3.

[59]:

```
cd hapmap3-files/
```

[60]:

```
ls
```

```
hapmap3.bed  hapmap3.bim  hapmap3.fam
hapmap3.map
```

[61]:

```
../admixture_linux-1.3.0/dist/admixture_linux-1.3.0/admixture32 hapmap3.bed 3
```

```
****                   ADMIXTURE Version 1.3.0                    ****
****                    Copyright 2008-2015                       ****
****           David Alexander, Suyash Shringarpure,              ****
```

```
****                John  Novembre, Ken Lange                 ****
****                                                          ****
****                 Please cite our paper!                   ****
****   Information at www.genetics.ucla.edu/software/admixture  ****

Random seed: 43
Point estimation method: Block relaxation algorithm
Convergence acceleration algorithm: QuasiNewton, 3 secant conditions
Point estimation will terminate when objective function delta < 0.0001
Estimation of standard errors disabled; will compute point estimates only.
Size of G: 324x13928
Performing five EM steps to prime main algorithm
1 (EM)  Elapsed: 0.19   Loglikelihood: -4.38757e+06     (delta): 2.87325e+06
2 (EM)  Elapsed: 0.18   Loglikelihood: -4.25681e+06     (delta): 130762
3 (EM)  Elapsed: 0.179  Loglikelihood: -4.21622e+06     (delta): 40582.9
4 (EM)  Elapsed: 0.189  Loglikelihood: -4.19347e+06     (delta): 22748.2
5 (EM)  Elapsed: 0.193  Loglikelihood: -4.17881e+06     (delta): 14663.1
Initial loglikelihood: -4.17881e+06
Starting main algorithm
1 (QN/Block)    Elapsed: 0.591  Loglikelihood: -3.94775e+06     (delta): 231058
2 (QN/Block)    Elapsed: 0.544  Loglikelihood: -3.8802e+06      (delta): 67554.6
3 (QN/Block)    Elapsed: 0.567  Loglikelihood: -3.83232e+06     (delta): 47883.8
4 (QN/Block)    Elapsed: 0.661  Loglikelihood: -3.81118e+06     (delta): 21138.2
5 (QN/Block)    Elapsed: 0.749  Loglikelihood: -3.80682e+06     (delta): 4354.36
6 (QN/Block)    Elapsed: 0.707  Loglikelihood: -3.80474e+06     (delta): 2085.65
7 (QN/Block)    Elapsed: 0.695  Loglikelihood: -3.80362e+06     (delta): 1112.58
8 (QN/Block)    Elapsed: 0.62   Loglikelihood: -3.80276e+06     (delta): 865.01
9 (QN/Block)    Elapsed: 0.543  Loglikelihood: -3.80209e+06     (delta): 666.662
10 (QN/Block)   Elapsed: 0.722  Loglikelihood: -3.80151e+06     (delta): 579.49
11 (QN/Block)   Elapsed: 0.848  Loglikelihood: -3.80097e+06     (delta): 548.156
12 (QN/Block)   Elapsed: 0.76   Loglikelihood: -3.80049e+06     (delta): 473.565
13 (QN/Block)   Elapsed: 0.703  Loglikelihood: -3.80023e+06     (delta): 258.61
14 (QN/Block)   Elapsed: 0.767  Loglikelihood: -3.80005e+06     (delta): 179.949
15 (QN/Block)   Elapsed: 0.803  Loglikelihood: -3.79991e+06     (delta): 146.707
16 (QN/Block)   Elapsed: 0.736  Loglikelihood: -3.79989e+06     (delta): 13.1942
17 (QN/Block)   Elapsed: 0.804  Loglikelihood: -3.79989e+06     (delta): 4.60747
18 (QN/Block)   Elapsed: 0.716  Loglikelihood: -3.79989e+06     (delta): 1.50012
19 (QN/Block)   Elapsed: 0.701  Loglikelihood: -3.79989e+06     (delta):
0.128916
20 (QN/Block)   Elapsed: 0.694  Loglikelihood: -3.79989e+06     (delta):
0.00182983
21 (QN/Block)   Elapsed: 0.694  Loglikelihood: -3.79989e+06     (delta):
4.33787e-05
Summary:
Converged in 21 iterations (16.556 sec)
Loglikelihood: -3799887.171935
Fst divergences between estimated populations:
        Pop0    Pop1
```

```
Pop0
Pop1    0.163
Pop2    0.073   0.156
Writing output files.
```

What information does the input .bed file contain? The .bed file used by ADMIXTURE is a binary genotype file in PLINK format. It encodes the genotype calls (0, 1, or 2 copies of the reference allele) for all individuals at each SNP in a compact binary form. This file is accompanied by a .bim file (with SNP info) and a .fam file (with sample metadata). Together, these files describe:

1. Which individuals were genotyped
2. Which SNPs were genotyped
3. The genotype of each individual at each SNP (e.g., AA, AB, or BB)

ADMIXTURE reads the .bed file to learn how allele frequencies vary across individuals and populations.

What does K represent in ADMIXTURE? K represents the number of ancestral populations (clusters) you want ADMIXTURE to infer. It is a user-defined input that tells the program how many genetic components to decompose the dataset into.

Why do you need to provide K? ADMIXTURE doesn't automatically determine the optimal number of populations. You must provide K because:

1. The model it fits is unsupervised — it doesn't know in advance how many populations are biologically meaningful.
2. Different values of K may reveal different patterns of structure, and you may compare them using cross-validation error to choose the best K.

So, setting K is a hypothesis about how many distinct ancestral groups may have contributed to the observed genotypes.

1.5 4.

[63]:

```
ls
```

```
hapmap3.3.P  hapmap3.3.Q  hapmap3.bed
hapmap3.bim  hapmap3.fam  hapmap3.map
```

[64]:

```
head hapmap3.3.P
```

```
0.999990 0.999990 0.999990
0.946581 0.934992 0.901852
0.989626 0.382598 0.918612
0.973109 0.682057 0.907595
0.678695 0.918927 0.129153
0.999990 0.999990 0.999990
```

```
0.999990 0.990119 0.999990
0.841989 0.203466 0.851233
0.967501 0.860690 0.622157
0.870693 0.862778 0.842376
```

[65]:

```
head hapmap3.3.Q
```

```
0.000010 0.896321 0.103669
0.009659 0.830876 0.159465
0.055770 0.725441 0.218790
0.000010 0.866447 0.133543
0.029255 0.888970 0.081775
0.009302 0.859576 0.131122
0.000010 0.715624 0.284366
0.013736 0.810352 0.175913
0.000010 0.727122 0.272868
0.034870 0.821125 0.144004
```

Mahdi_Anvari.compG_HW1_R

March 24, 2025

0.1 R section:

[8]:
```r
getwd()
```

'/mnt/d/Daneshga/Term 8/Computational Genomics/HW1'

0.2 5.

[9]:
```r
table_Q = read.table("hapmap3-files/hapmap3.3.Q")
```

[10]:
```r
head(table_Q)
```

A data.frame: 6 × 3

	V1 <dbl>	V2 <dbl>	V3 <dbl>
1	0.000010	0.896321	0.103669
2	0.009659	0.830876	0.159465
3	0.055770	0.725441	0.218790
4	0.000010	0.866447	0.133543
5	0.029255	0.888970	0.081775
6	0.009302	0.859576	0.131122

[11]:
```r
print(dim(table_Q))
```

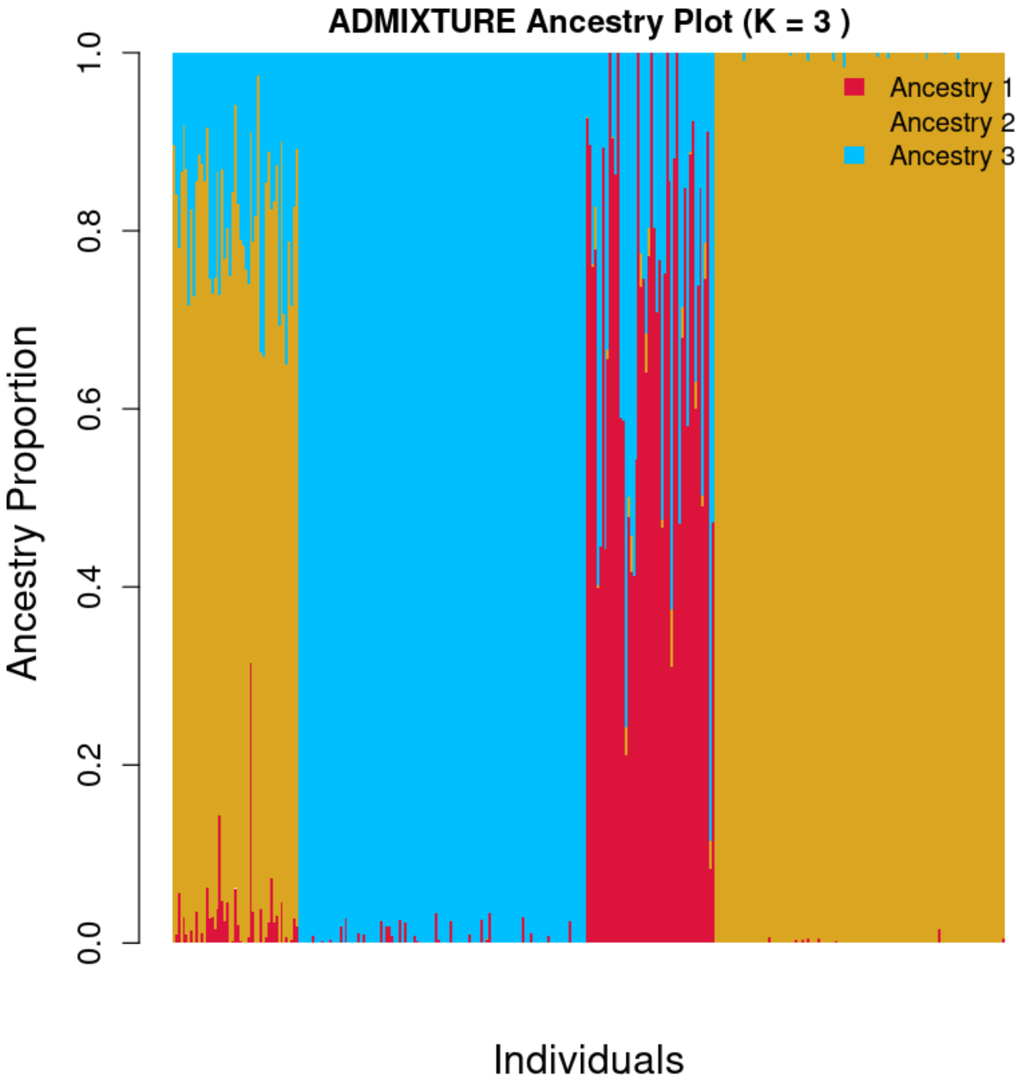
```
[1] 324   3
```

[12]:
```r
colors <- c("#DC143C", "#DAA520", "#00BFFF")

par(mar = c(5, 5, 2, 2), cex.axis = 1.2, cex.lab = 1.5, cex.main = 1.2)

barplot(t(as.matrix(table_Q)), col=colors,
        xlab="Individuals", ylab="Ancestry Proportion", space=0,
        border=NA, main = paste("ADMIXTURE Ancestry Plot (K =", 3, ")"))

legend("topright", legend=paste("Ancestry", 1:3),
       fill=colors, bty="n",
       border=NA, cex=1)
```



The Q plot generated by ADMIXTURE shows individual-level ancestry proportions across the inferred ancestral populations (in this case, K = 3). Each individual is represented by a stacked bar, where the colors represent the estimated proportion of ancestry from each of the three populations.

0.3 6.

```
[13]: table_P = read.table("hapmap3-files/hapmap3.3.P")
```

```
[14]: head(table_P)
```

A data.frame: 6 × 3

	V1 <dbl>	V2 <dbl>	V3 <dbl>
1	0.999990	0.999990	0.999990
2	0.946581	0.934992	0.901852
3	0.989626	0.382598	0.918612
4	0.973109	0.682057	0.907595
5	0.678695	0.918927	0.129153
6	0.999990	0.999990	0.999990

[15]:

```
cat("Number of SNPs used in the ADMIXTURE analysis:", nrow(table_P), "\n")
```

```
Number of SNPs used in the ADMIXTURE analysis: 13928
```

How many SNPs were used in the ADMIXTURE analysis? The .P file contains the population-specific allele frequencies for each SNP across K ancestral populations. So, the number of rows in the .P file corresponds to the number of SNPs used in the analysis.

0.4 7.

[16]:

```
P_var <- apply(table_P, 1, var)
```

[17]:

```
print(head(P_var))
```

```
[1] 0.0000000000 0.0005388747 0.1101394982 0.0233117901 0.1639084658
[6] 0.0000000000
```

[18]:

```
length(P_var)
```

13928

[19]:

```
max_var_index <- which.max(P_var)
min_var_index <- which.min(P_var)
```

[20]:

```
max_var_afs <- table_P[max_var_index, ]
min_var_afs <- table_P[min_var_index, ]
```

highest cross-population variance:

[21]:

```
max_var_index
```

13465

[22]:

```
P_var[max_var_index]
```

0.309251993194333

[23]:

```
max_var_afs
```

A data.frame: 1 × 3

	V1 <dbl>	V2 <dbl>	V3 <dbl>
13465	0.931699	0.004461	0.99999

lowest cross-population variance:

[24]:
```
min_var_index
```

1

[25]:
```
P_var[min_var_index]
```

0

[26]:
```
min_var_afs
```

A data.frame: 1 × 3

	V1 <dbl>	V2 <dbl>	V3 <dbl>
1	0.99999	0.99999	0.99999

By calculating the variance of allele frequencies across populations (per SNP), we can:

Identify SNPs that differ strongly in AF across populations → high variance (likely ancestry-informative markers).

Identify SNPs that are similarly distributed across populations → low variance (likely shared variation).