Solutions to exercises about inference of admixture and population structure

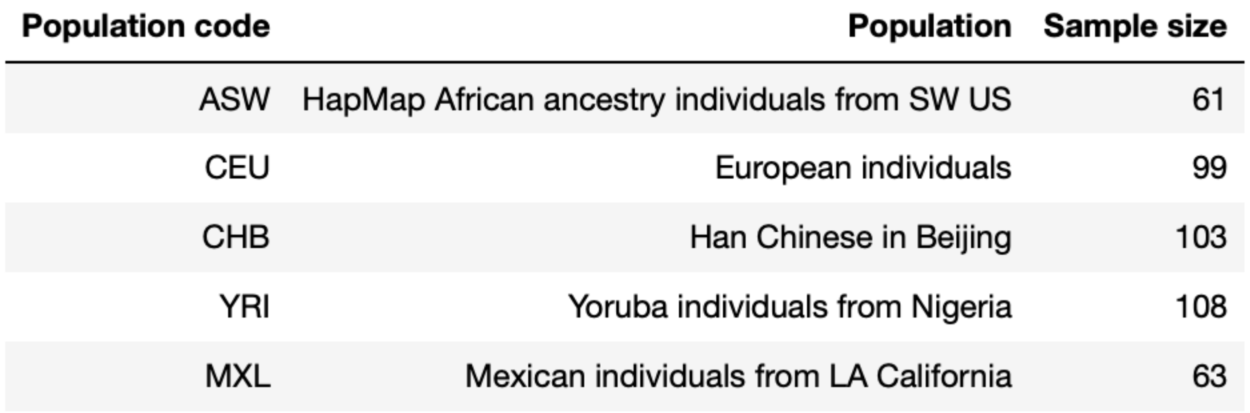
Summer 2024

Exercise A:

Use of NGSadmix to infer admixture proportions for numerous individuals

Small example: overview of dataset

Dataset consists of mapped data from:

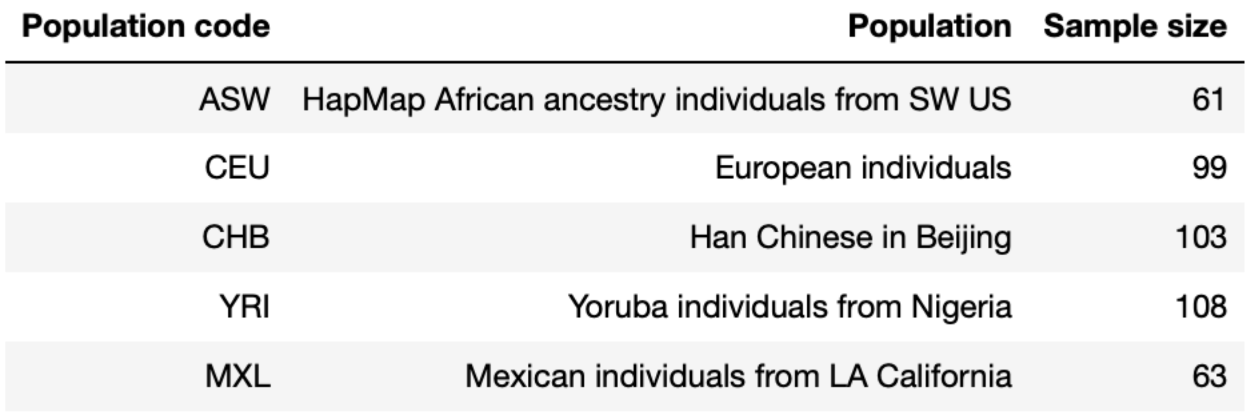


Population code	Population	Sample size
ASW	HapMap African ancestry individuals from SW US	61
CEU	European individuals	99
CHB	Han Chinese in Beijing	103
YRI	Yoruba individuals from Nigeria	108
MXL	Mexican individuals from LA California	63

Question:
How many loci do we have genotype likelihoods for?

Small example: overview of dataset

Dataset consists of mapped data from:



Population code	Population	Sample size
ASW	HapMap African ancestry individuals from SW US	61
CEU	European individuals	99
CHB	Han Chinese in Beijing	103
YRI	Yoruba individuals from Nigeria	108
MXL	Mexican individuals from LA California	63

Question:
How many loci do we have genotype likelihoods for?

Solution:

```
gunzip -c all.beagle.gz | wc -l
1307
```

I.e. 1307 lines of which 1 is a header line, so 1306 loci.

Small example: look at the GL input file

First 6 data lines including only the first 9 columns:

```
> gunzip -c all.beagle.gz | head -n 7 | cut -f1-9
marker      allele1  allele2  Ind0      Ind0      Ind0      Ind1      Ind1      Ind1
1_20018051  2        1        0.799979  0.200021  0.000000  0.799890  0.200110  0.000000
1_20018077  1        3        0.799951  0.200049  0.000000  0.940831  0.059169  0.000000
1_20018091  3        1        0.888773  0.111227  0.000000  0.984605  0.015395  0.000000
1_20018096  2        0        0.000000  0.997885  0.002115  0.000000  0.015392  0.984608
1_20018195  2        0        0.969690  0.030310  0.000000  0.999509  0.000491  0.000000
1_20050859  3        1        0.000000  0.030420  0.969580  0.888711  0.111289  0.000000
```

Question:

What is the most likely genotype for Ind1 in the 1st locus?

Small example: look at the GL input file

First 6 data lines including only the first 9 columns:

```
> gunzip -c all.beagle.gz | head -n 7 | cut -f1-9
marker      allele1  allele2  Ind0      Ind0      Ind0      Ind1      Ind1      Ind1
1_20018051  2        1        0.799979  0.200021  0.000000  0.799890  0.200110  0.000000
1_20018077  1        3        0.799951  0.200049  0.000000  0.940831  0.059169  0.000000
1_20018091  3        1        0.888773  0.111227  0.000000  0.984605  0.015395  0.000000
1_20018096  2        0        0.000000  0.997885  0.002115  0.000000  0.015392  0.984608
1_20018195  2        0        0.969690  0.030310  0.000000  0.999509  0.000491  0.000000
1_20050859  3        1        0.000000  0.030420  0.969580  0.888711  0.111289  0.000000
```

Question:

What is the most likely genotype for Ind1 in the 1st locus?

Solution:

Ind0 in locus 1: allele1 allele1 or 2 2, i.e. GG

Small example: running NGSadmix

Then we ran NGSadmix:

```
$NGSadmix -likes all.beagle.gz -K 3 -minMaf 0.05 -seed 1 -o all
```

Question:
Is it clear what that command means?

Small example: running NGSadmix

Then we ran NGSadmix:

```
$NGSadmix -likes all.beagle.gz -K 3 -minMaf 0.05 -seed 1 -o all
```

Question:
Is it clear what that command means?

Solution:

- ”-likes all.beagle.gz”: input file with GLs is called all.beagle.gz
- ”-K 3”: assume 3 ancestral populations
- ”-minMaf 0.05”: only use loci with minor allele frequency > 0.05
- ”-seed 1”: set seed to 1
- ”-o all”: give all output files the prefix ”all”

Small example: looking at NGSadmix output

Then we looked at the output files. First the log file:

```
Input: lname=all.beagle.gz nPop=3, fname=(null) qname=(null) outfiles=all
Setup: seed=1 nThreads=1 method=1
Convergence: maxIter=2000 tol=0.000010 tolLike50=0.100000 dymBound=0
Filters: misTol=0.050000 minMaf=0.050000 minLrt=0.000000 minInd=0
Input file has dim: nsites=1306 nind=435
Input file has dim (AFTER filtering): nsites=1306 nind=435
        [ALL done] cpu-time used =  12.26 sec
        [ALL done] walltime used =  12.00 sec
best like=-454474.751725 after 137 iterations
```

Question:
What is the log likelihood of the estimates achieved?

Small example: looking at NGSadmix output

Then we looked at the output files. First the log file:

```
Input: lname=all.beagle.gz nPop=3, fname=(null) qname=(null) outfiles=all
Setup: seed=1 nThreads=1 method=1
Convergence: maxIter=2000 tol=0.000010 tolLike50=0.100000 dymBound=0
Filters: misTol=0.050000 minMaf=0.050000 minLrt=0.000000 minInd=0
Input file has dim: nsites=1306 nind=435
Input file has dim (AFTER filtering): nsites=1306 nind=435
        [ALL done] cpu-time used =  12.26 sec
        [ALL done] walltime used =  12.00 sec
best like=-454474.751725 after 137 iterations
```

Question:
What is the log likelihood of the estimates achieved?

Small example: looking at NGSadmix output

Then we looked at the first line of the fopt output file:

```
> zcat all.fopt.gz | head -n1
0.29904643371021311093 0.38077976484864278772 0.74495838518322954336
```

Question:
What is the estimated allele frequency of 1st locus in the 3 assumed ancestral populations?

Small example: looking at NGSadmix output

Then we looked at the first line of the fopt output file:

```
> zcat all.fopt.gz | head -n1
0.29904643371021311093 0.38077976484864278772 0.74495838518322954336
```

Question:
What is the estimated allele frequency of 1st locus in the 3 assumed ancestral populations?

Solution:
0.299 0.381 0.745

Small example: looking at NGSadmix output

Next, we looked at the 6th line of the qopt output file:

```
>  head -n6 all.qopt | tail -n1
0.00000000099999999997 0.99999999799999994554 0.00000000099999999997
```

Question:

Based on this: does the individual look admixed?

Small example: looking at NGSadmix output

Next, we looked at the first line of the qopt output file:

```
>  head -n6 all.qopt | tail -n1
0.00000000099999999997 0.99999999799999994554 0.00000000099999999997
```

Question:

Based on this: does the individual look admixed?

Solution:

No because it has basically 100% ancestry from the second assumed ancestral population.

Small example: looking at NGSadmix output

Looking at the 6th line of the input bamfile list:

```
> head -n6 all.files | tail -n1
/course/popgen23/ida/admixexercise/smallbams/small.NA19121.mapped.ILLUMINA.bwa.YRI.low_coverage.20130415.bam
```

Question:

Which population does the individual come from? And what does NGSadmix estimate the allele frequency to be at the first locus in that population?

Small example: looking at NGSadmix output

Looking at the first line of the input bamfile list:

```
> head -n6 all.files | tail -n1
/course/popgen23/ida/admixexercise/smallbams/small.NA19121.mapped.ILLUMINA.bwa.YRI.low_coverage.20130415.bam
```

Question:

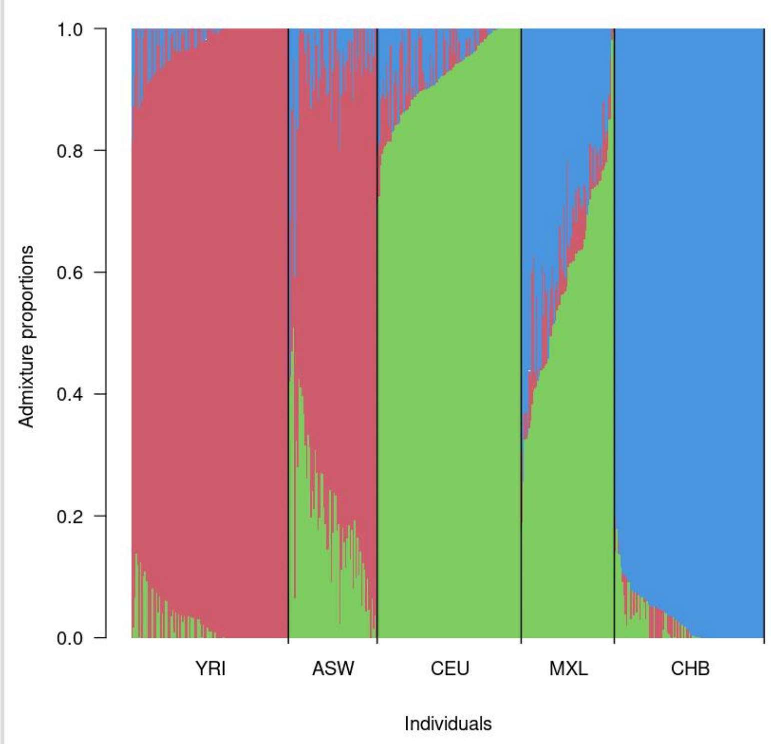
Which population does the individual come from? And what does NGSadmix estimate the allele frequency to be at the first locus in that population?

Solution:

It is from YRI (African). And since we just saw that the individual is estimated to have all its ancestry from the 2nd assumed ancestral population, this means that we can find NGSadmix' frequency estimates for YRI in the second column in the qopt file. Earlier we saw that the 2nd column of the first line in the qopt file (which contains the frequency estimates for the first locus) was 0.381. So, the solution is: 0.381.

Small example: plotting estimated admixture proportions

Finally, you were asked to plot the estimated admixture proportions:



Question:
Try to explain the plot. What does it suggest about whether the individuals are admixed?

Solution:
1 vertical line per sample w. color proportions showing ancestry proportion estimates
Several of the individuals look admixed (has more than one color)

Bigger example: overview of data

Same samples but with data 100000 sites (so still limited data but a bit more realistic…☺).

And you were asked to look at the results of 20 runs for K=3

Question:

Does it look convergence was reached?

Bigger example: overview of data

Same samples but with data 100000 sites (so still limited data but a bit more realistic…☺).

And you were asked to look at the results of 20 runs for K=3

Question:

Does it look convergence was reached?

Solution:

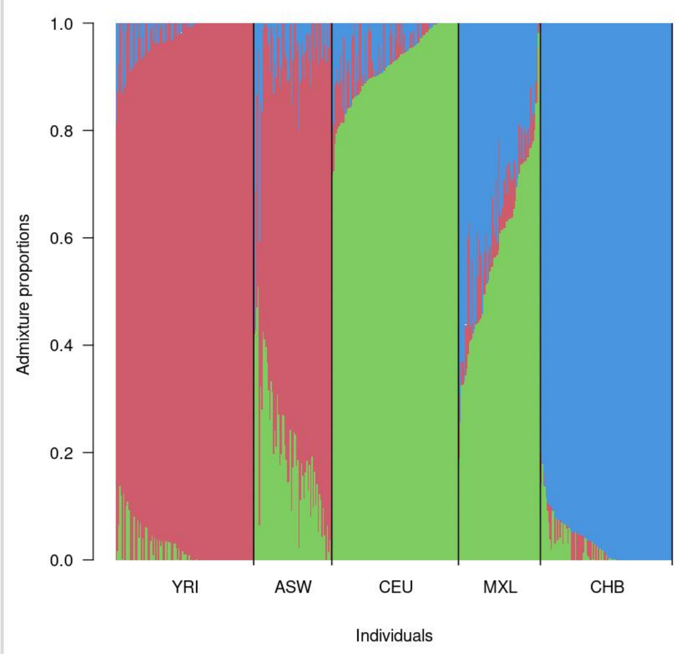
```
cat -n allK$K.likes | sort -rhk2
3      -34941707.471918
16     -34941707.578680
7      -34941707.737304
18     -34941707.948875
2      -34941707.989102
```

The top 5 solutions are all within 1 likelihood unit, so yes.

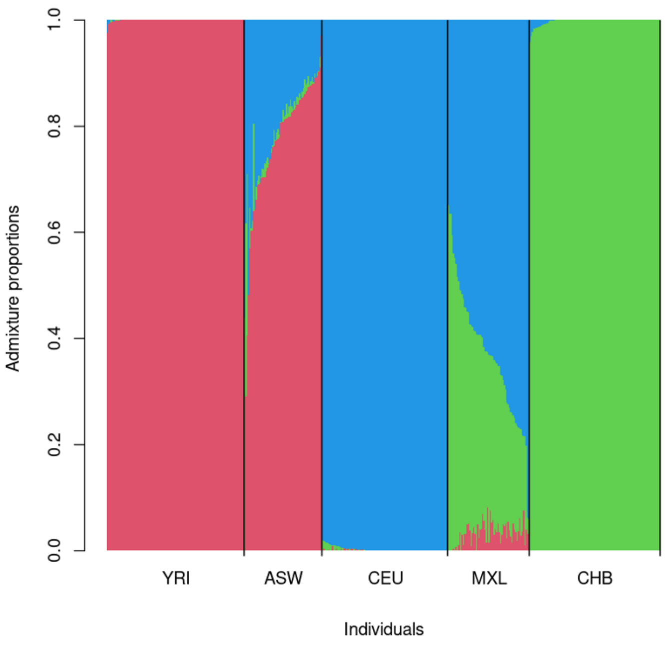
Bigger example: admixture proportion plot

Then you were asked to plot the one with the highest likelihood and compare to the 1st one

1st plot



New plot

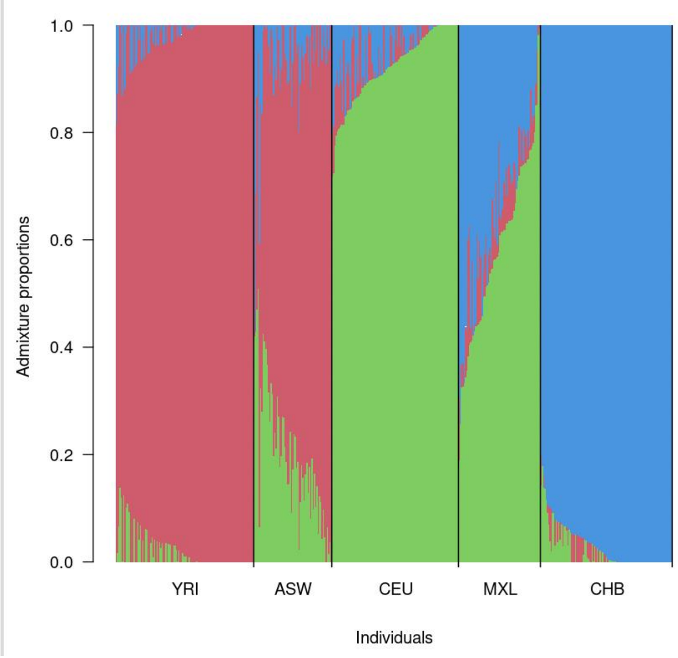


Question:
Why do you think it looks different than the previous admixture plot we visualized with the same individuals?

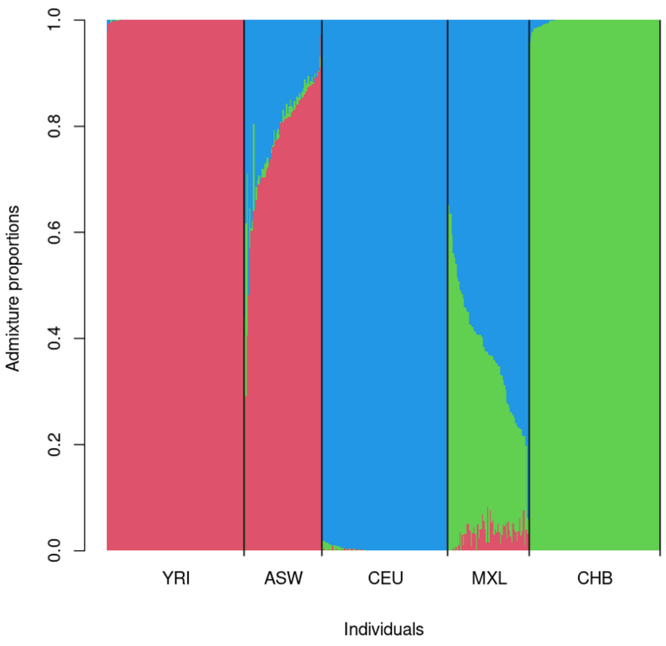
Bigger example: admixture proportion plot

Then you were asked to plot the one with the highest likelihood and compare to the 1st one

1st plot



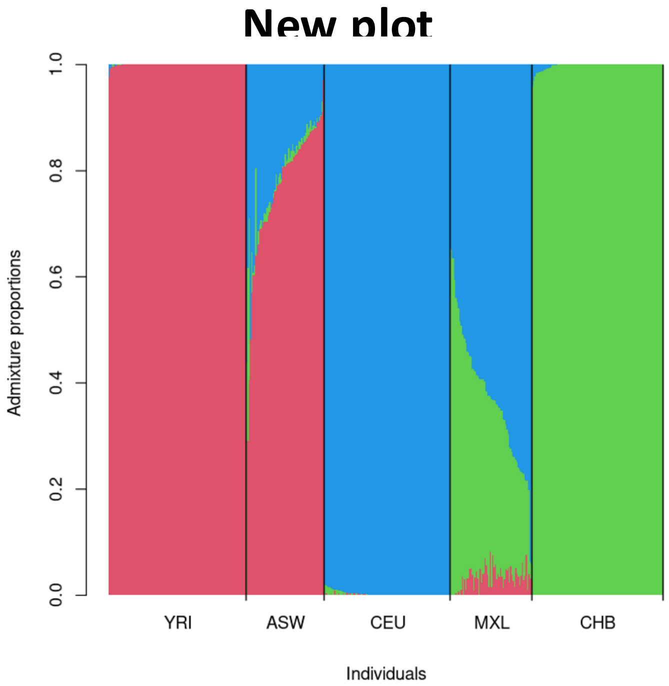
New plot



Question:
Why do you think it looks different than the previous admixture plot we visualized with the same individuals?

Solution:
Because we have many more SNPs! Also, we checked for convergence.

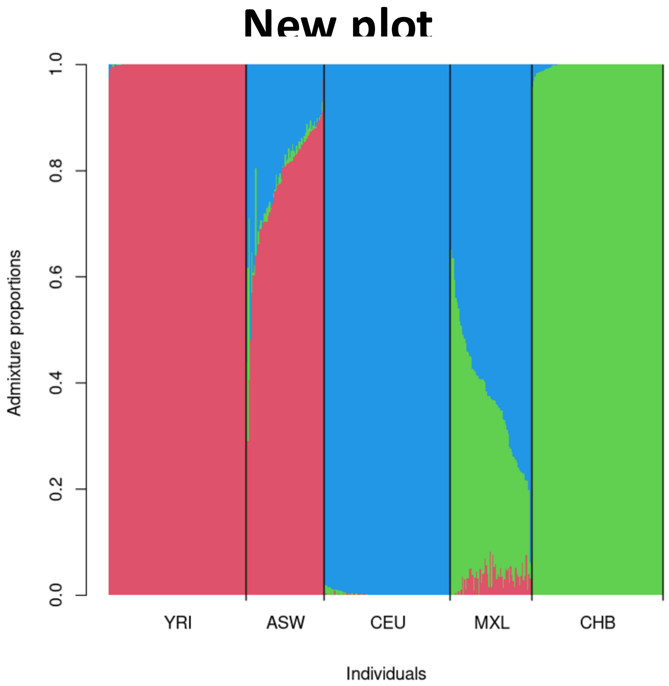
Bigger example: admixture proportion plot



Question:

How many populations would you say now are admixed? Which population seem to be the admixture source? Does that make sense given what you know of these populations?

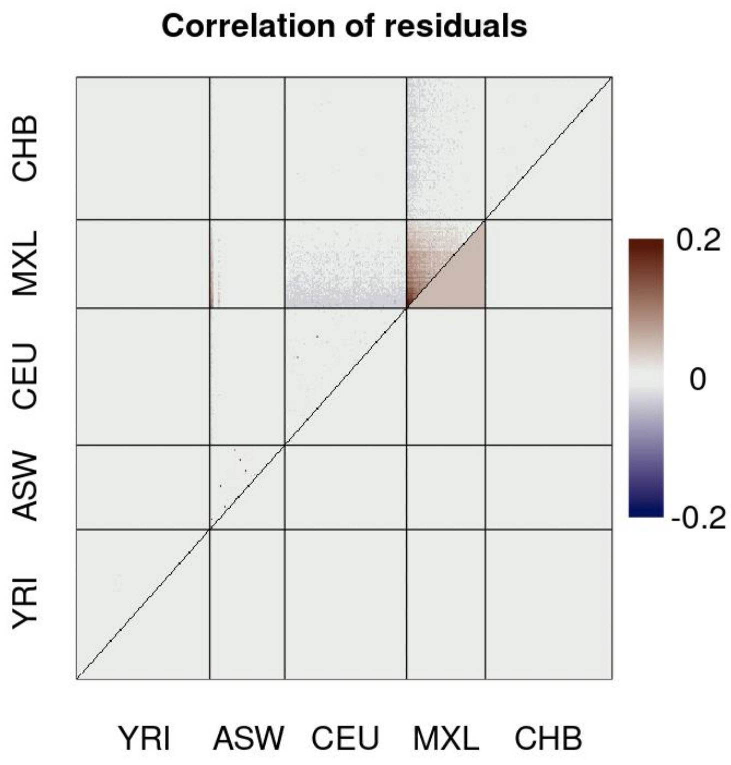
Bigger example: admixture proportion plot

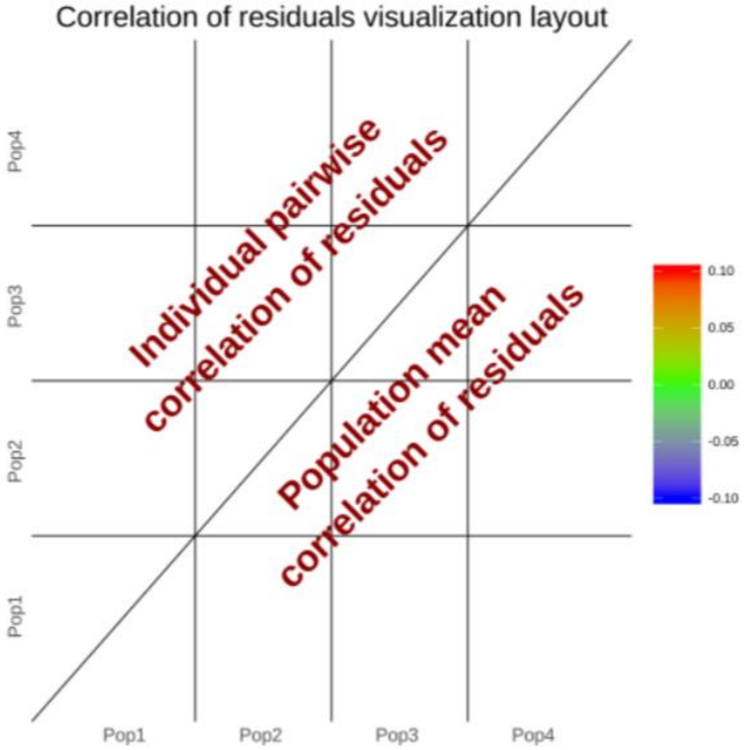


Question:
How many populations would you say now are admixed? Which population seem to be the admixture source? Does that make sense given what you know of these populations?

Solution:
ASW and MXL look admixed. It looks like the source populations are YRI and CEU for ASW and CEU, YRI and CHB for MXL. The latter (CHB) seems odd since MXL are Native Americans.

Bigger example: assessing fit



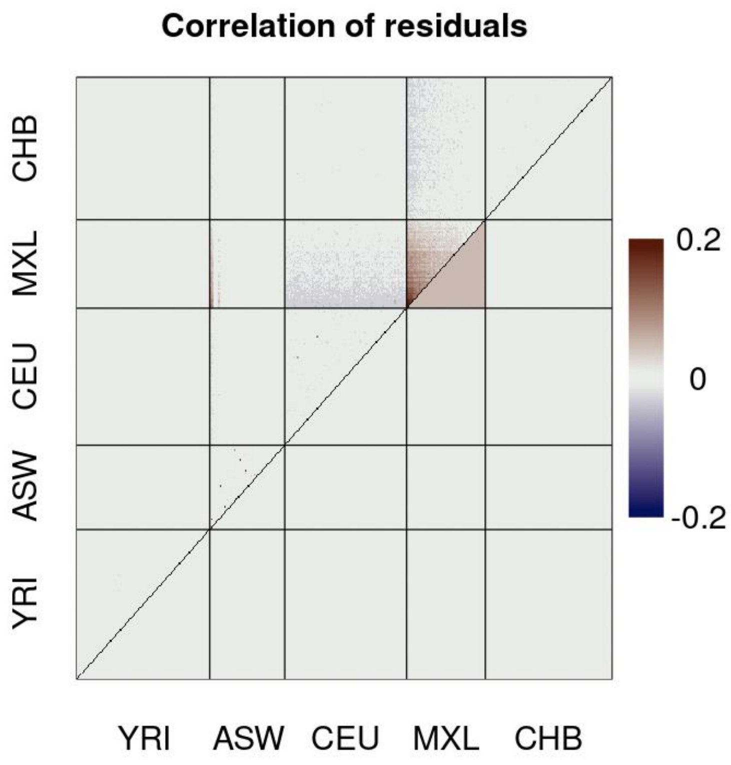


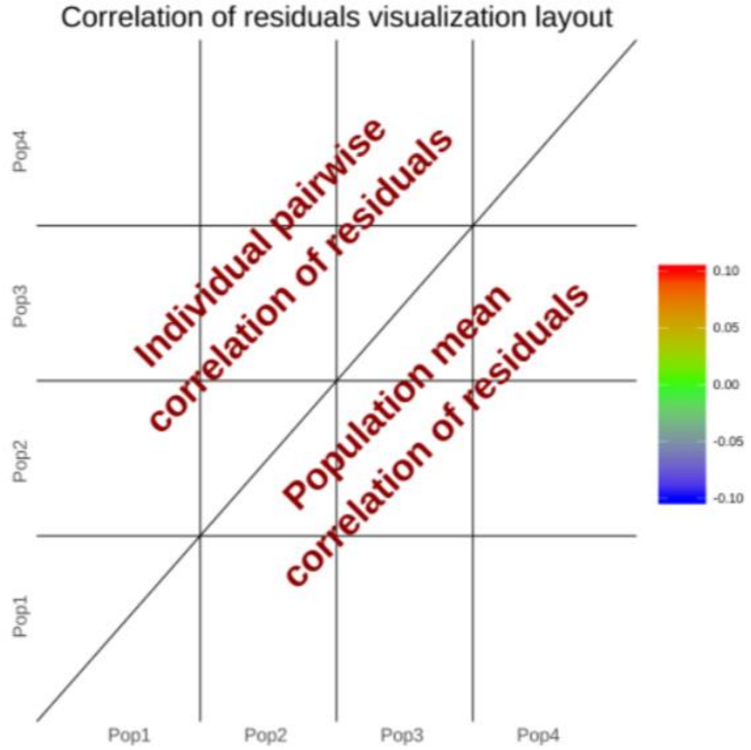
Correlation of residuals visualization layout
Individual pairwise correlation of residuals
Population mean correlation of residuals
Pop4
Pop3
Pop2
Pop1
Pop1
Pop2
Pop3
Pop4
0.10
0.05
0.00
-0.05
-0.10

Question:

Is there any population for which the estimated admixture proportions do not seem to be a good fit?

Bigger example: assessing fit




Correlation of residuals visualization layout
Pop4
Pop3
Pop2
Pop1
Pop1
Pop2
Pop3
Pop4
Individual pairwise correlation of residuals
Population mean correlation of residuals
0.10
0.05
0.00
-0.05
-0.10

Question:

Is there any population for which the estimated admixture proportions do not seem to be a good fit?

Solution:

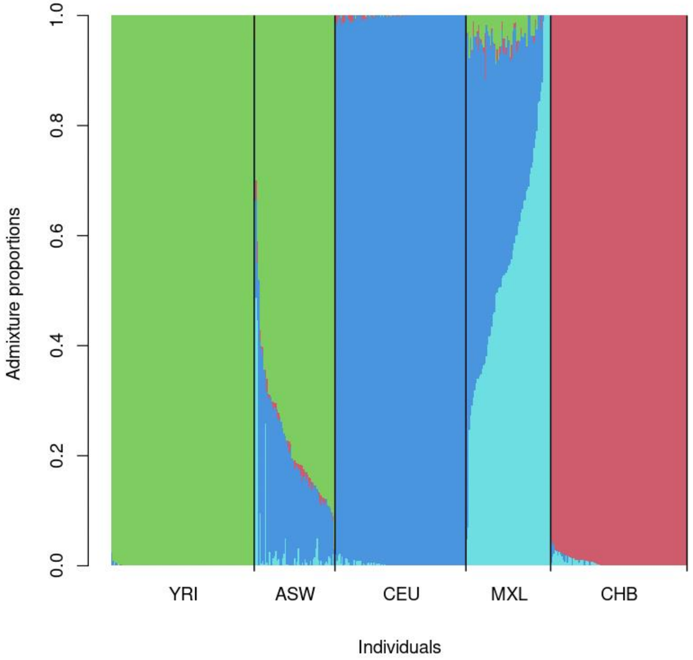
There is an avg. positive correlation within the Mexicans suggesting a bad fit for these

Bigger example: analysis with K=4

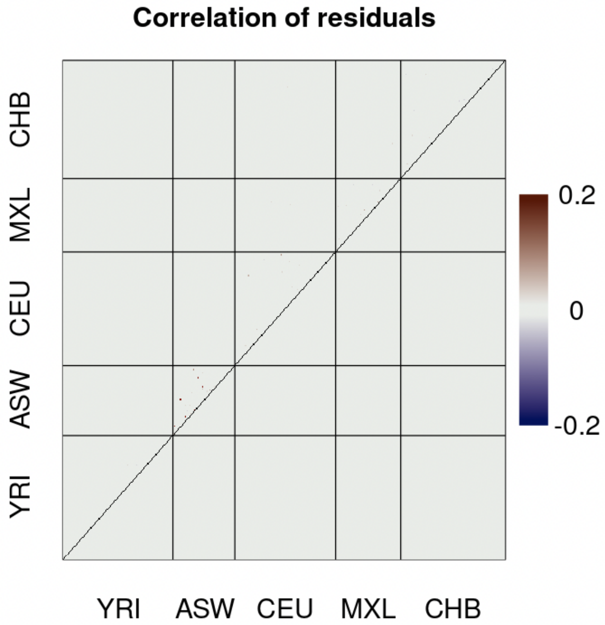
Top 5 likelihoods:

9	-34654597.394711
11	-34654597.515125
18	-34654597.670429
8	-34654597.952226
14	-34654598.071523

Admixture plot



evalAdmix plot



Question:

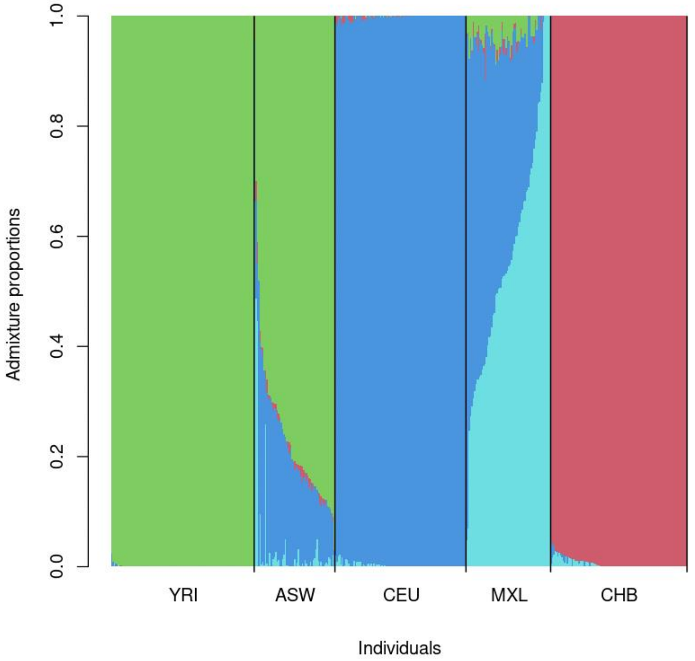
What population does the new cluster that we have added correspond to? Based on the correlation of residuals, would you say adding that cluster has given a significant improvement to the model fit?

Bigger example: analysis with K=4

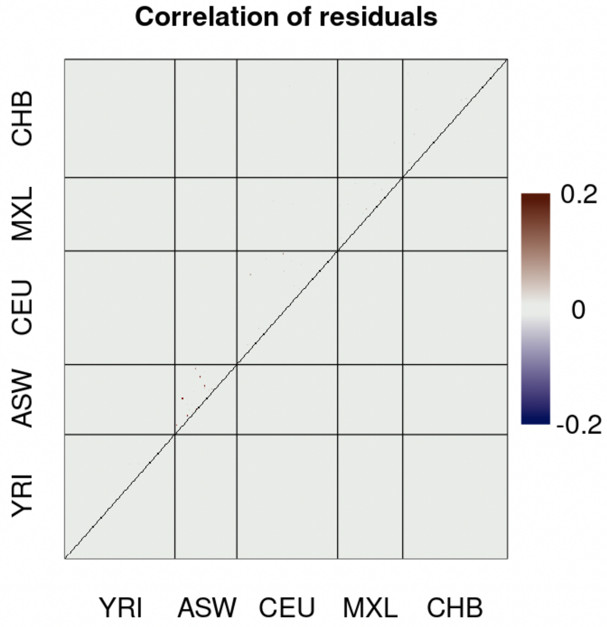
Top 5 likelihoods:

9	-34654597.394711
11	-34654597.515125
18	-34654597.670429
8	-34654597.952226
14	-34654598.071523

Admixture plot



evalAdmix plot



Question:

What population does the new cluster that we have added correspond to? Based on the correlation of residuals, would you say adding that cluster has given a significant improvement to the model fit?

Solution:

Native American ancestry and yes (useful especially in cases without prior knowledge!)

Exercise C (only if time allows):

Use of fastNGSadmix to infer admixture proportions for 3 samples

Overview of data

We have a **reference dataset** for worldwide 7 populations in two files

1) 1 file with allele frequencies:

```
id        chr  pos     name       A0_freq  A1  French  Han     Chukchi Karitiana Papuan  Sindhi  Yoruba
1_752566  1    752566  rs3094315  G        A   0.1667  0.0606  0.3696  0.0834    0.0714  0.3056  0.6714
```

2) 1 file with numbers of samples used to estimate the frequencies

```
French  Han  Chukchi  Karitiana  Papuan  Sindhi  Yoruba
25      33   23       12         14      18      70
```

And finally, we have **GLs for three samples**: sample1, sample2, sample3

Overview of data

We ran fastNGSadmix on the three samples.

Question:

From the log files: how many loci are the analyses based?

Overview of data

We ran fastNGSadmix on the three samples.

Question:

From the log files: how many loci are the analyses based?

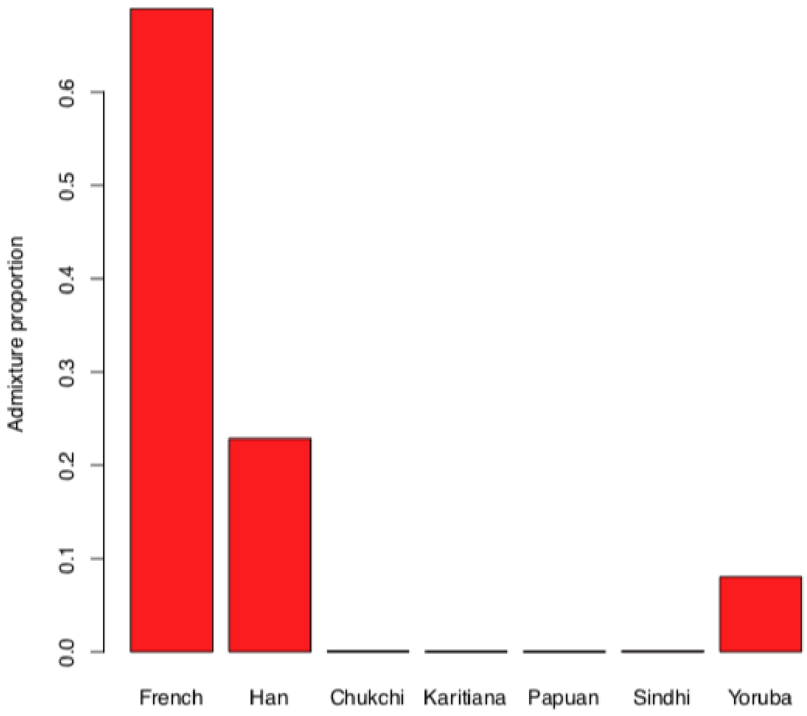
Solution:

- sample1: 49643
- sample2: 20903
- sample3: 91

Results: sample 1

Question:

What do you think of sample1 is based on the first analysis – does it look admixed?



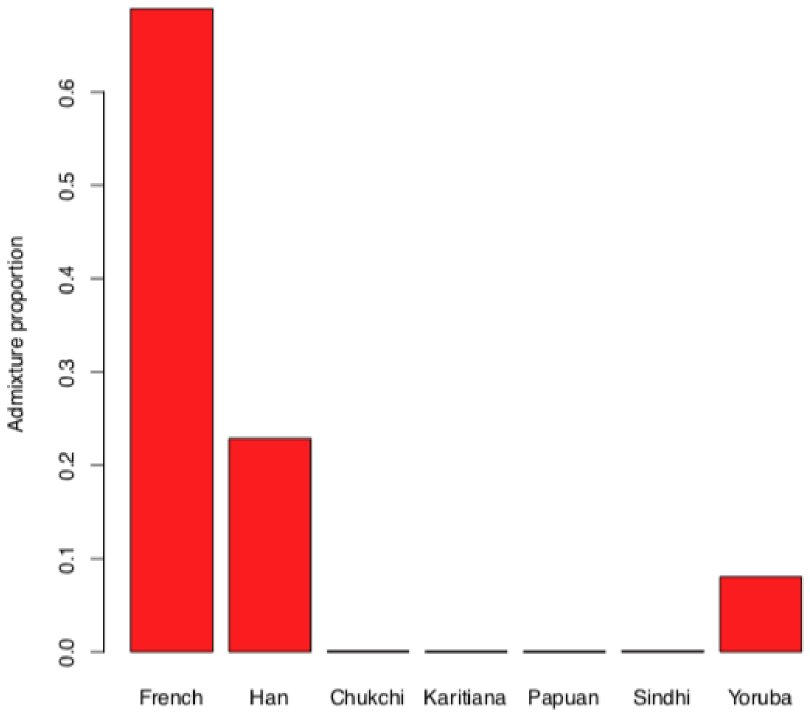
Solution:

Admixed with ca. 65% French, ca. 23% and ca. 12% Yoruba.

Results: sample 1

Question:

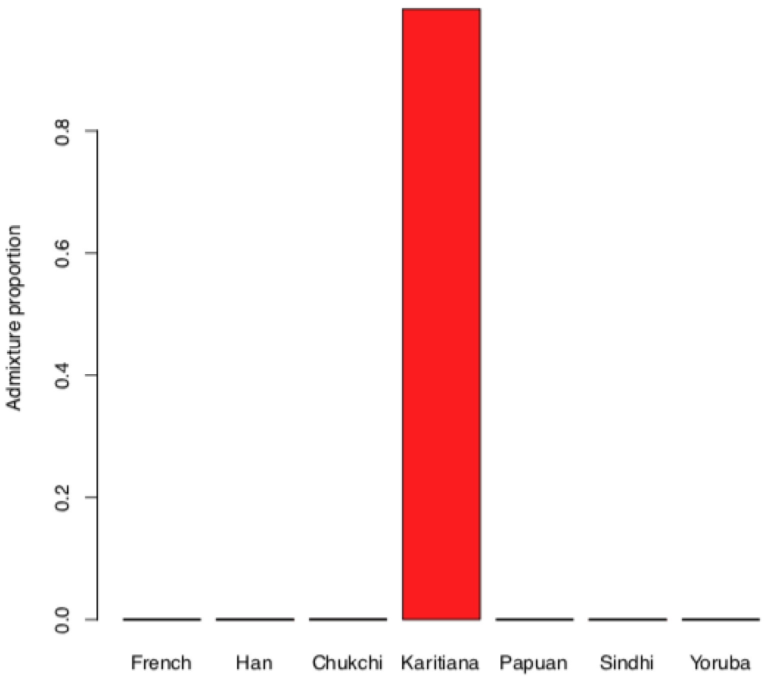
What do you think of sample1 is based on the first analysis – does it look admixed?



Results: sample 2

Question:

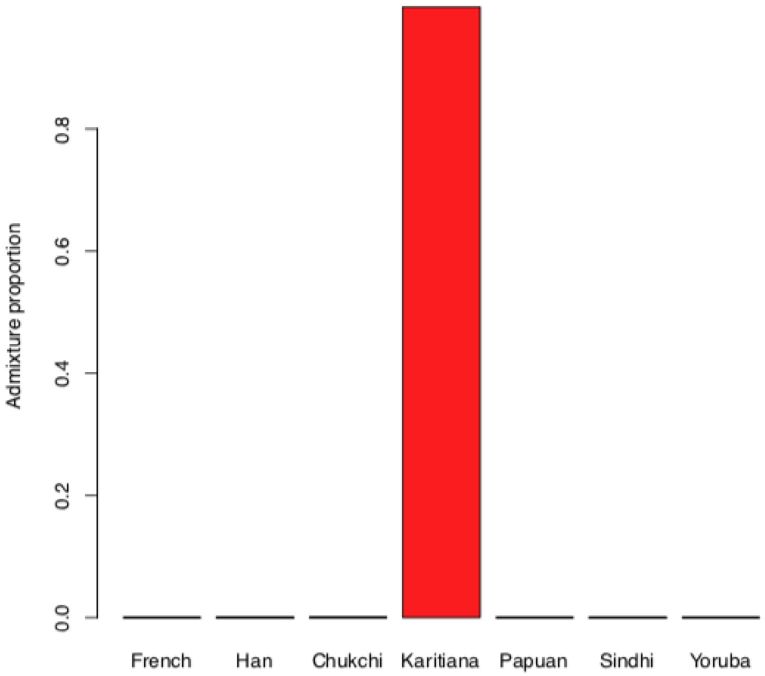
What do you think the ancestry of sample2 is?



Results: sample 2

Question:

What do you think the ancestry of sample2 is?



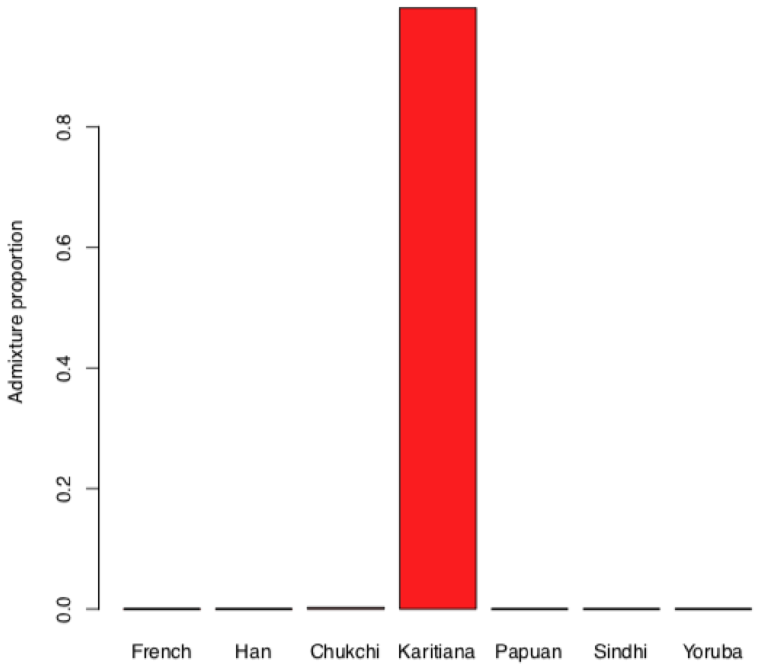
Solution:

Unadmixed Karitiana

Results: sample 3

Question:

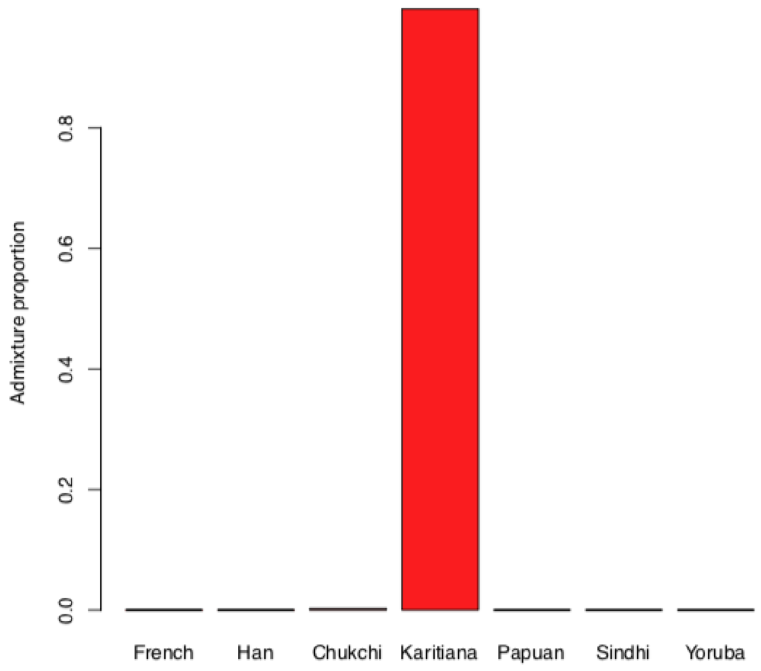
What do you think the ancestry of sample3 is?



Results: sample 3

Question:

What do you think the ancestry of sample3 is?



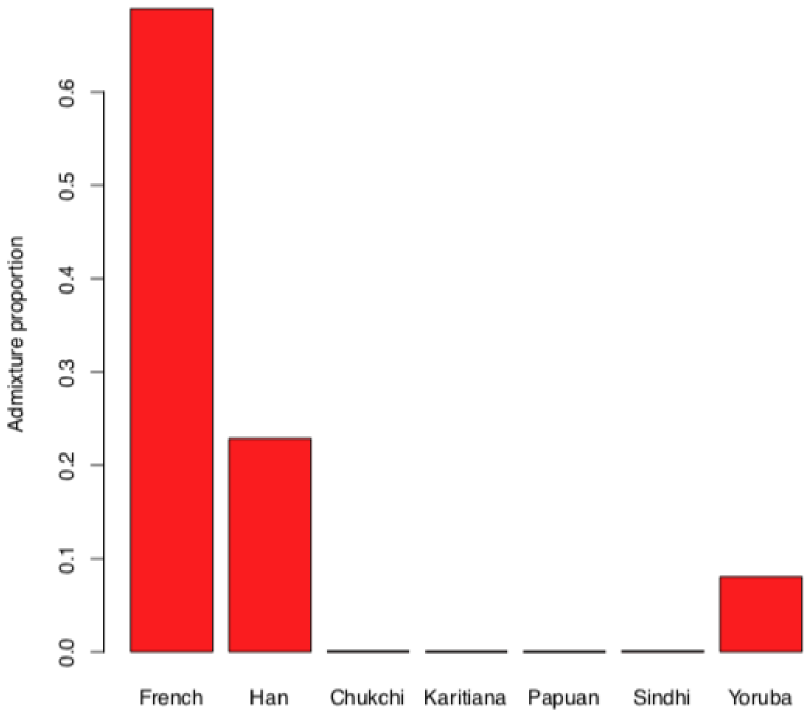
Solution:

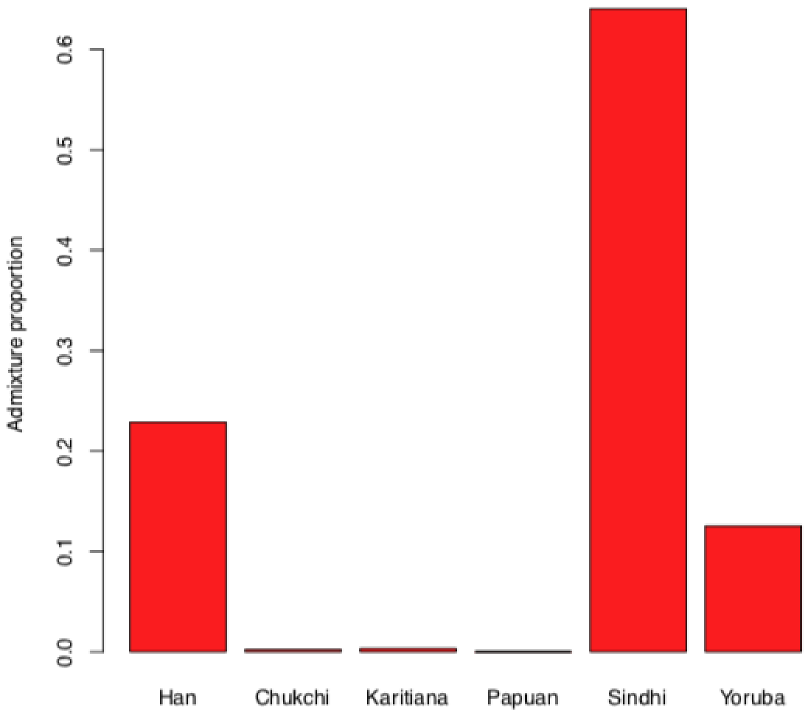
Unadmixed Karitiana

Returning to sample 1 – comparing the two analyses

Question:

What do you think the ancestry of sample1 is?





Solution:

Depends on the panel! (ca. 65% French or ca. 65% Sindhi)

Results: sample 1

Question:

Why does the result depend on reference panel and what are the consequences?

Results: sample 1

Question:

Why does the result depend on reference panel and what are the consequences?

Solution:

We simulated the data and know the truth is that the samples is 65% French and not Sindhi, but looks like Sindhi in the absence of French since Sindhi and French are genetically fairly similar – so be careful with what you conclude!

Returning to sample 3

Question:

Do you trust the results for sample3 given it’s only based on 92 loci?

Returning to sample 3

Question:

Do you trust the results for sample3 given it’s only based on 92 loci?

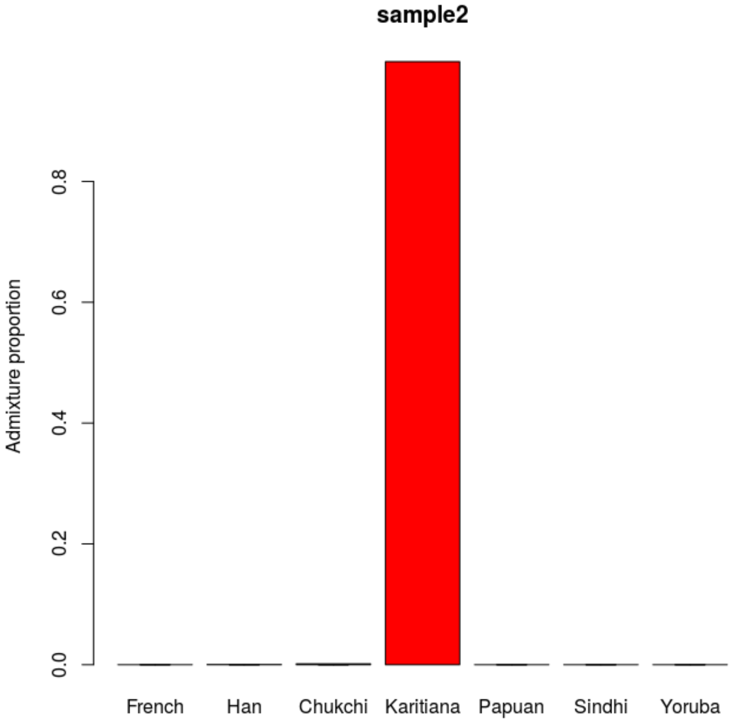
Solution:

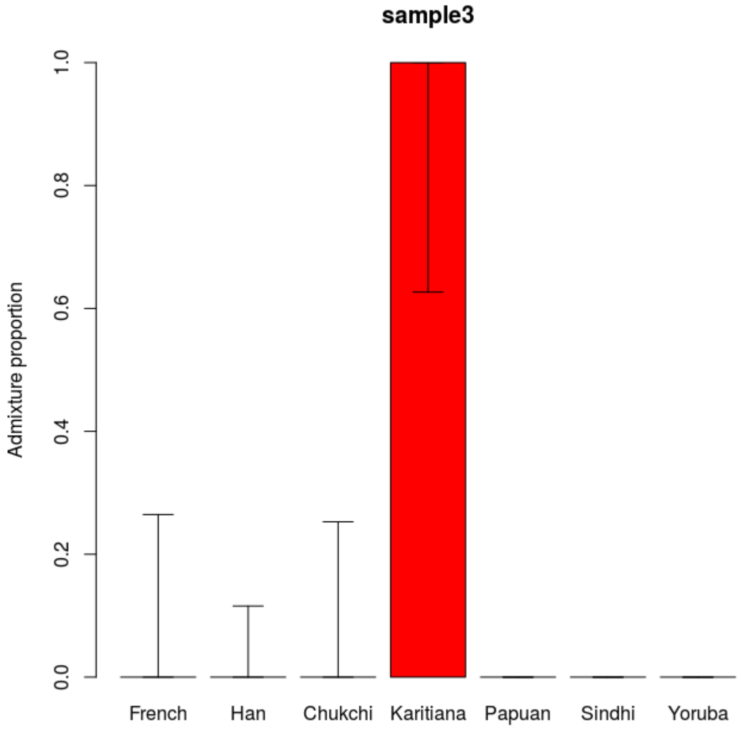
- In this case it is true! It’s a down-sampled version of sample2. So you can get very far with few loci when you have a reference panel!!
- However, in general difficult to say based on just the estimated proportions
- To address this we can do bootstrap

Bootstrap results: sample 2 and 3

Question:

What does that tell us?

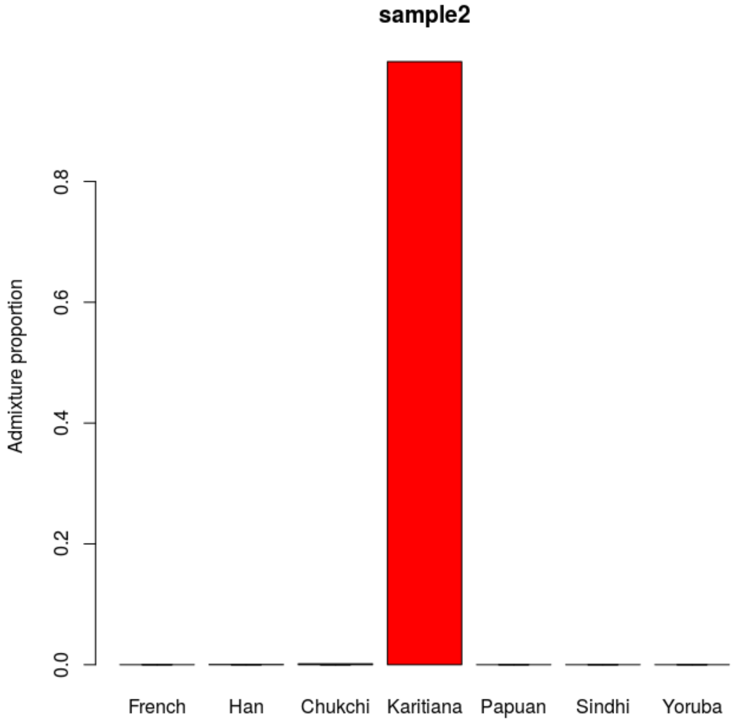


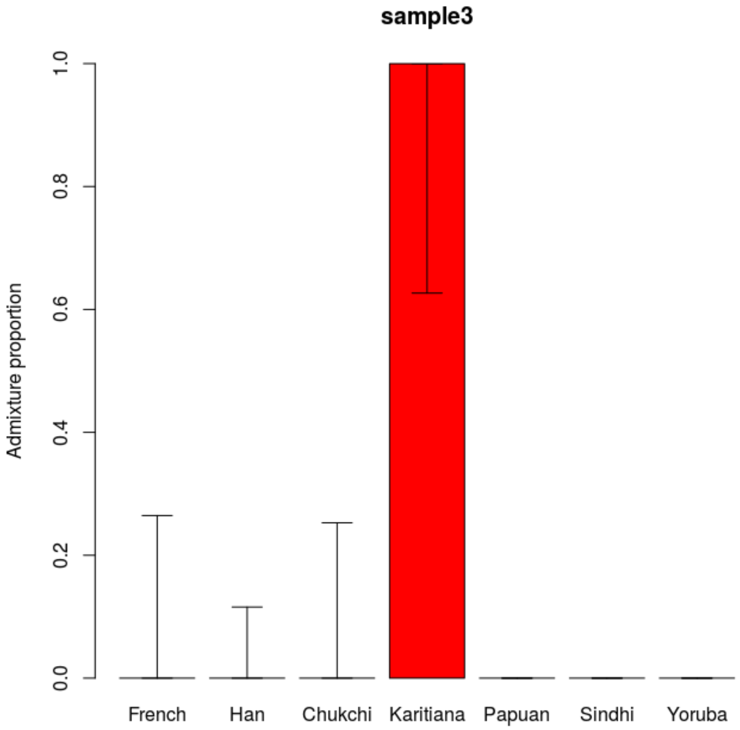


Bootstrap results: sample 2 and 3

Question:

What does that tell us?





Solution:

There is more uncertainty when there are fewer sites, and this uncertainty is worth reporting!