Quack?

Cruz Dávalos, Diana I.

03.10.2018

Summary of sequencing statistics

DNA was extracted from two different teeth and built into one library per tooth. These libraries will be refered to as MA2776 and MA2777.

Library	Sequenced reads	Trashed reads after trimmming (fraction)	Retained reads after trimming	Length retained reads	Duplicate reads	Endogenous reads	Endogenous (fraction)	Nuclear reads	Genome coverage	Length nuclear reads	Mitochondri reads
PRI_44A7_MA2777	411,691,382	0.205	327,366,561	47	0	541,233	0.00165	535,492	0.00775	44.2	5,741
PRI_MPYN_MA2776	392,956,393	0.128	342,678,242	54	0	2,409,896	0.00703	2,400,864	0.0351	45	9,032
Quack	804,647,775	0.167	670,044,803	50.6	0	2,951,129	0.0044	2,936,356	0.0428	44.9	14,773

Reads mapping to HBV

A total of 142 reads mapped to the strain C of Hepatitis B virus (nomenclature from Mühlemann et al. 2018). All the mapped reads had a mapping quality of 37.

Strain ID	Average read depth	Sum of reads length	Chromosome length	Average read length
C: AB117758.1	2.405	7731	3215	54.4

Mitochondrial DNA analyses

The coverage on the mitochondrial chromosome is **47.7x** (total = 14,773 unique reads) We called the consensus secuqnece using ANGSD and caling the most frequent base among the reads mapping to a given position. This consensus sequence was later used to estimate contamination and recover the mitochondrial haplogroup.

Haplogroup

We uploaded the consensus sequence of the mitochondrial DNA to James Lick's website and to Haplogrep2. Both resources assigned the same haplogroup to the individual.

According to James Lick's website, the sequence matches 23 out of 30 defining markers for the haplogroup **G3a1'2**. The remaining positions included 4 mismatches and 3 sites without calls.

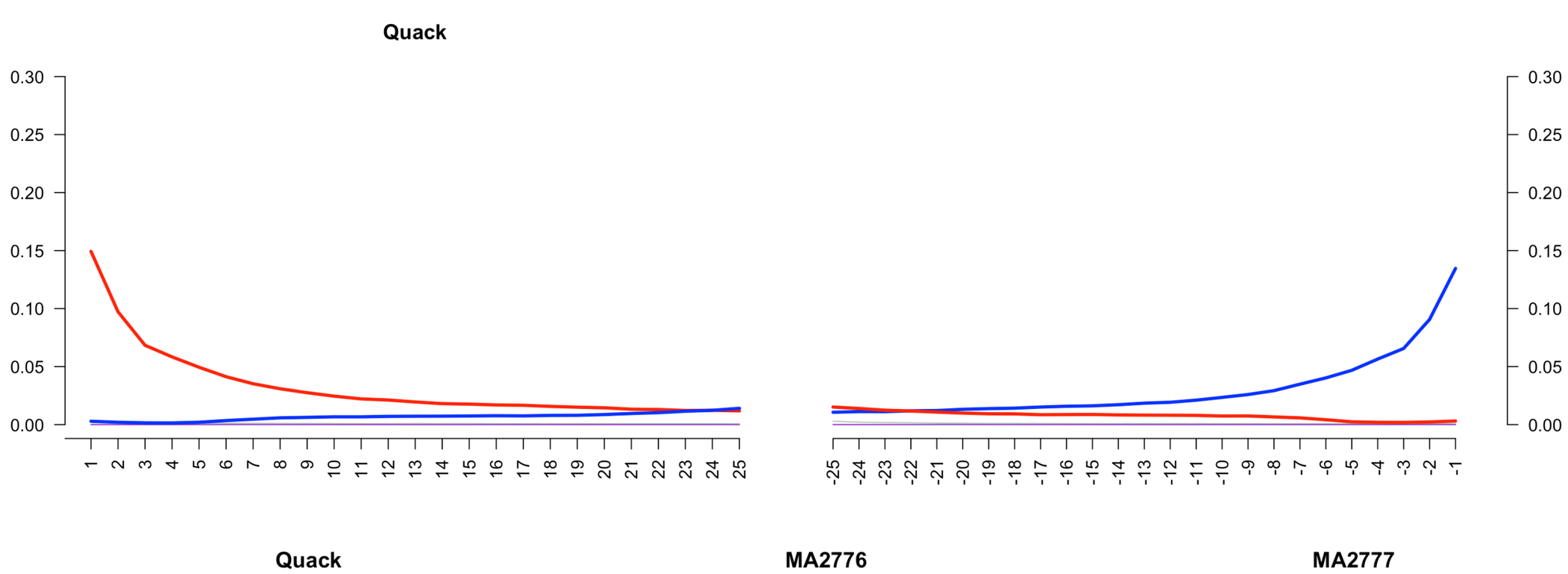
Position	Defining marker	Consensus in Quack	Consensus in MA2776	Consensus in MA2777
1438	G	G	G	A
4769	G	A	A	A
8701	G	A	A	A
11719	A	G	G	A
16362	C	T	T	T

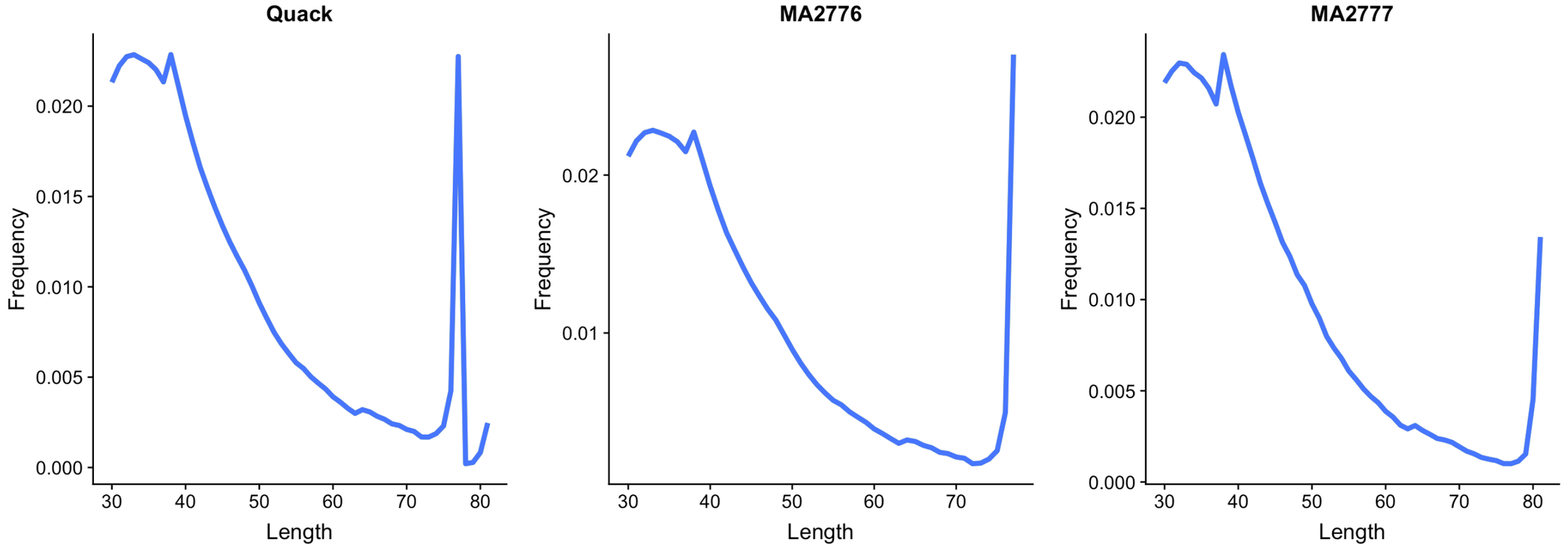
About haplogroup G

> Information extracted from Wikipedia It is an **East Asian haplogroup**. Today, haplogroup G is found at its highest frequency in indigenous populations of the lands surrounding the Sea of Okhotsk. Haplogroup G is one of the most common mtDNA haplogroups among modern Ainu, Japanese, Mongol, and Tibetan people (as well as among people of the prehistoric Jōmon culture in Hokkaidō). It is also found at a lower frequency among many other populations of East Asia, Central Asia, Bangladesh, Sri Lanka and Nepal. However, unlike other mitochondrial DNA haplogroups typical of populations of northeastern Asia, such as haplogroup A, haplogroup C, and haplogroup D, **haplogroup G has not been found among indigenous peoples of the Americas**.

Molecular damage

We used mapDamage2 to estimate the deamination rates across the reads mapping to the human genome.





Contamination estimates

We estimated the contamination levels per library using the method introduced in Fu et al. (201). The contamination varies depending on whether we ignore the transitions in the estimation. We present the results for both cases, considering the whole dataset and using only the reads without transitions.

When we use the whole data, the estimates for contamination are 9% (MA2776) and 15% (MA2777). If we filter out the transitions, the contamination estimates are <1% for both libraries.

The table below shows the number of polymorphisms (in the 311 humans dataset) relative to the Quack's consensus sequence. About 25% correspond to transitions.

Base in Quack	Polymorphism	Number of times found in dataset	Fraction of total polymorphisms
A	C	103	0.0057
A	G	6725	0.3734
A	N	187	0.0104
A	T	51	0.0028
C	A	101	0.0056
C	G	325	0.0180
C	N	464	0.0258
C	T	2137	0.1186
G	A	2563	0.1423
G	C	28	0.0016
G	N	142	0.0079
G	T	9	0.0005
T	A	59	0.0033
T	C	5099	0.2831
T	G	6	0.0003
T	N	13	0.0007

Notably, ~7.5% of the reads contain transitions that correspond to polymorphic sites found in the dataset.

Quack

Contamination: ~10%

Coverage on MT: 46x

```
Crude contamination upper bound: 12781 out of 179112 reads (7.1%) match another
genome better than the consensus (error rate is 0.0192).

quantiles from n = 99154 samples (after discarding burnin of 846)
     2.5%     97.5%
0.8887864 0.9059447
MAP authentic: 0.8972901
Potential scale reduction factors:

     Point est. Upper C.I.
[1,]          1          1
```

MA2776

Contamination: ~9%

Coverage on MT: 29x

```
Crude contamination upper bound: 2244 out of 31504 reads (7.1%) match another
genome better than the consensus (error rate is 0.0187).

quantiles from n = 99460 samples (after discarding burnin of 540)
     2.5%     97.5%
0.8908430 0.9277308
MAP authentic: 0.9110321
Potential scale reduction factors:

     Point est. Upper C.I.
[1,]          1          1
```

MA2777

Contamination: ~15%

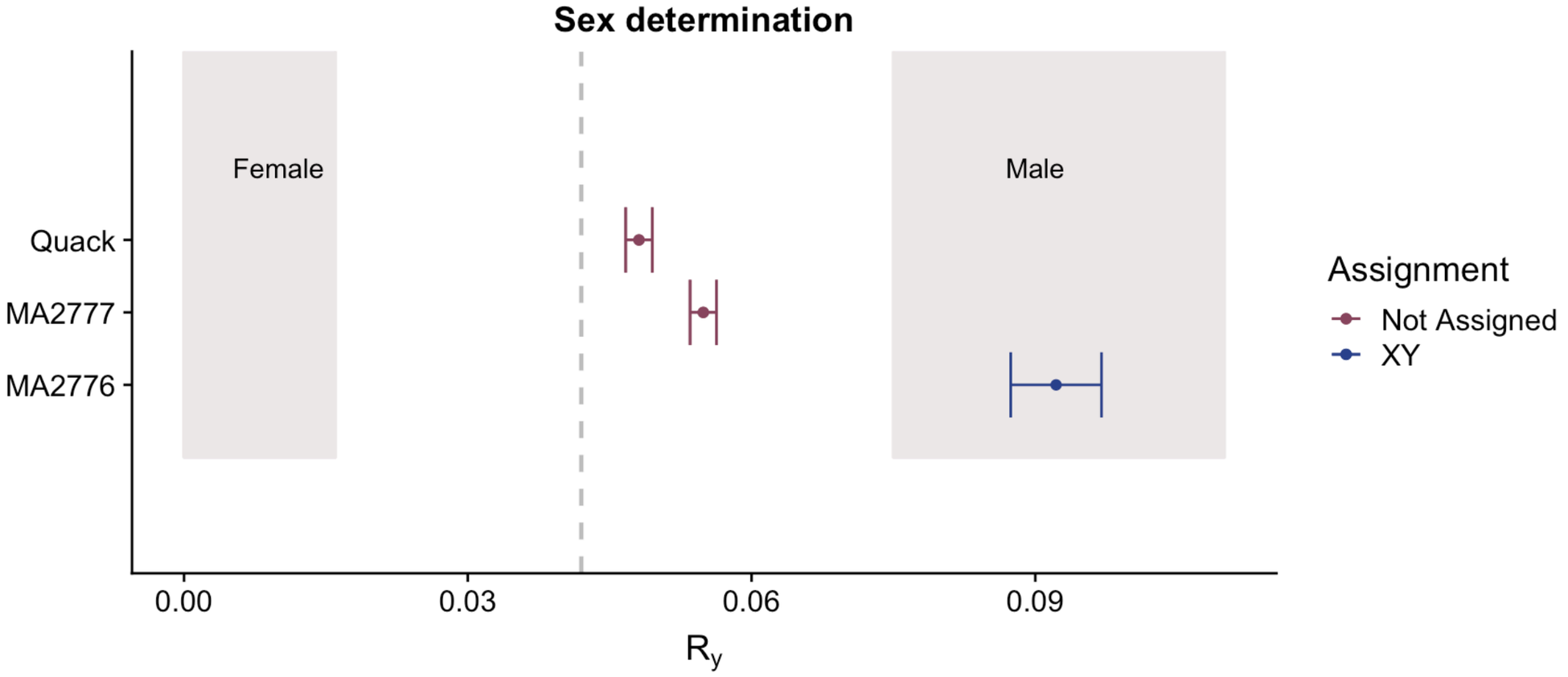
Coverage on MT: 18x

```
Crude contamination upper bound: 10801 out of 147627 reads (7.3%) match another
genome better than the consensus (error rate is 0.0192).

quantiles from n = 99096 samples (after discarding burnin of 904)
    2.5%    97.5%
0.847060 0.866473
MAP authentic: 0.8568024
Potential scale reduction factors:

     Point est. Upper C.I.
[1,]          1          1
```

Sex assignment

We determined the sex of the individual using the method presented in Skoglund et al. (2013). The ratio of reads mappping to the Y chromosome vs. those mapping to both sexual chromosomes is 0.0016 (95% CI = 0.0014 - 0.0018). The value is consistent with a **XY** karyotype, according to the thresholds defined in Skoglund et al. (2013) and shown in the figure below.

Nseqs	NchrY NchrX	NchrY	R_y	SE	X95 CI	Assignment
2951129	97735	5367	0.0549	7e-04	0.0535-0.0563	Not Assigned



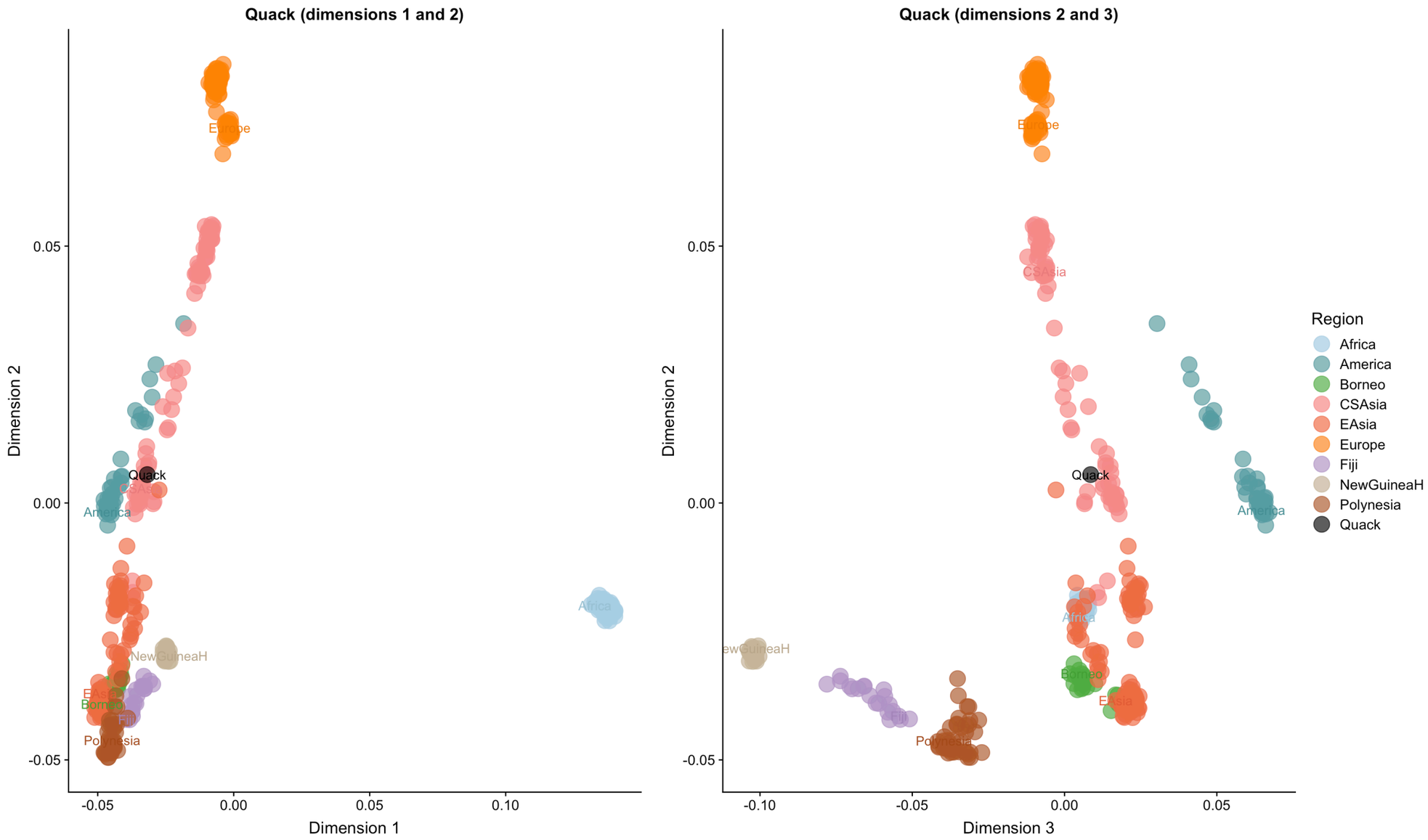
MDS

We performed a multidimensional scale analysis to visualize the genetic affinities of Quack. For this purpose, we used bammds (Malaspinas et al. 2014) and a panel with worldwide populations (Wollstein et al. 2010). We observe that Quack is not grouping with Native American populations, but with Central-South or East Asian populations in the panel.

Quack: 34,031 out of 823,581 markers

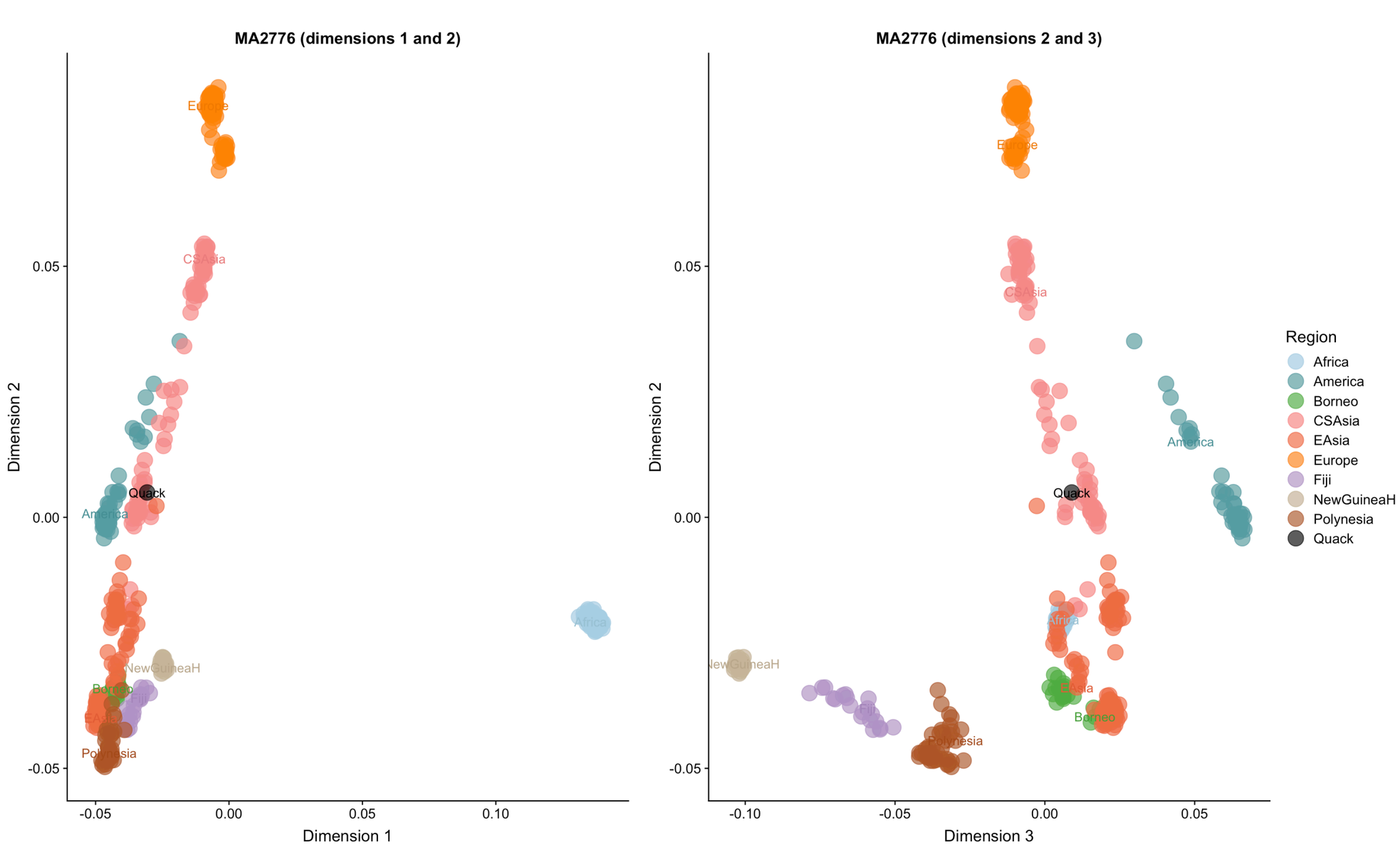
MA276: 34,038 out of 823,588

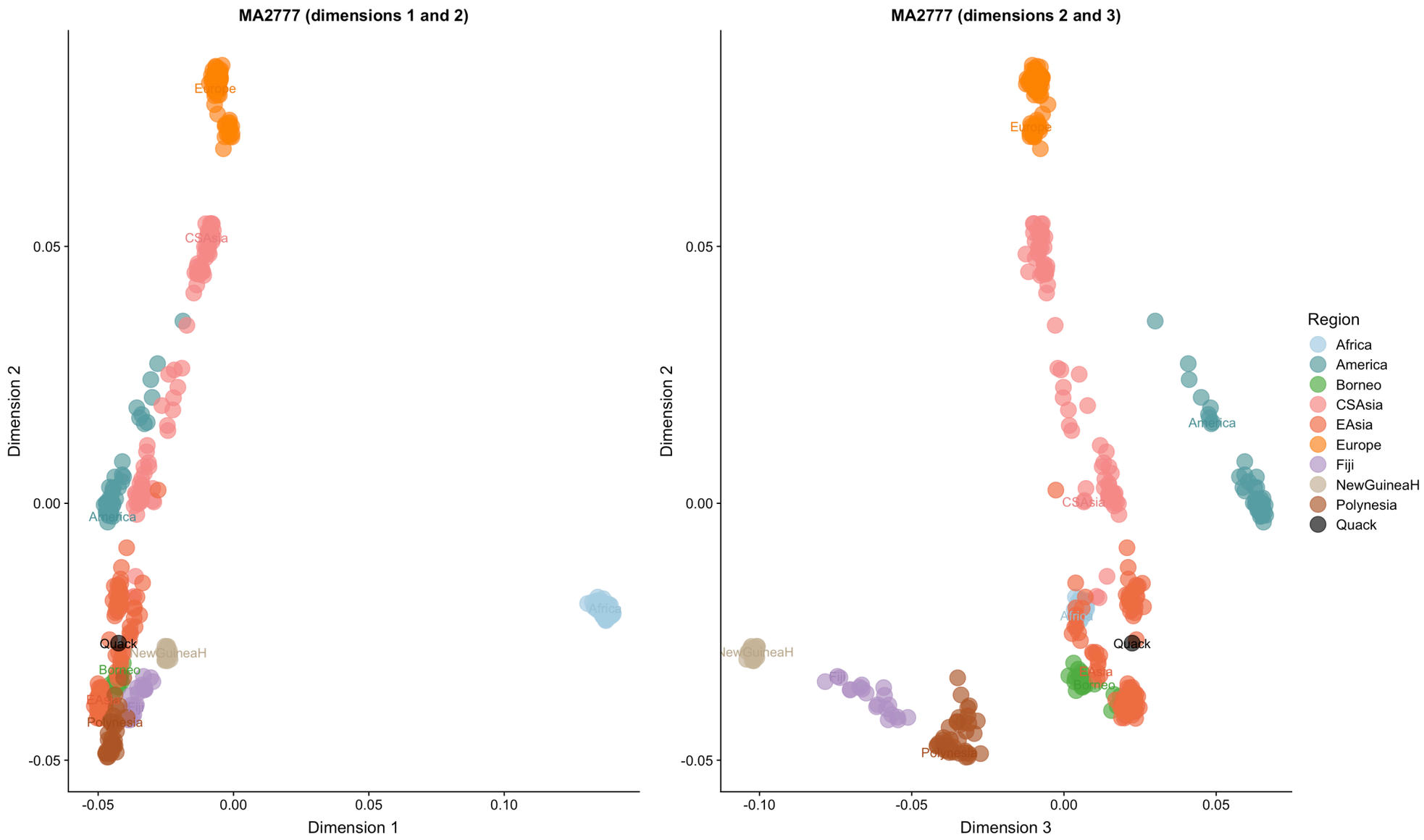
MA277: 6,381 out of 823,757



MA2776 (dimensions 1 and 2)

MA2776 (dimensions 2 and 3)





Ancestry estimates

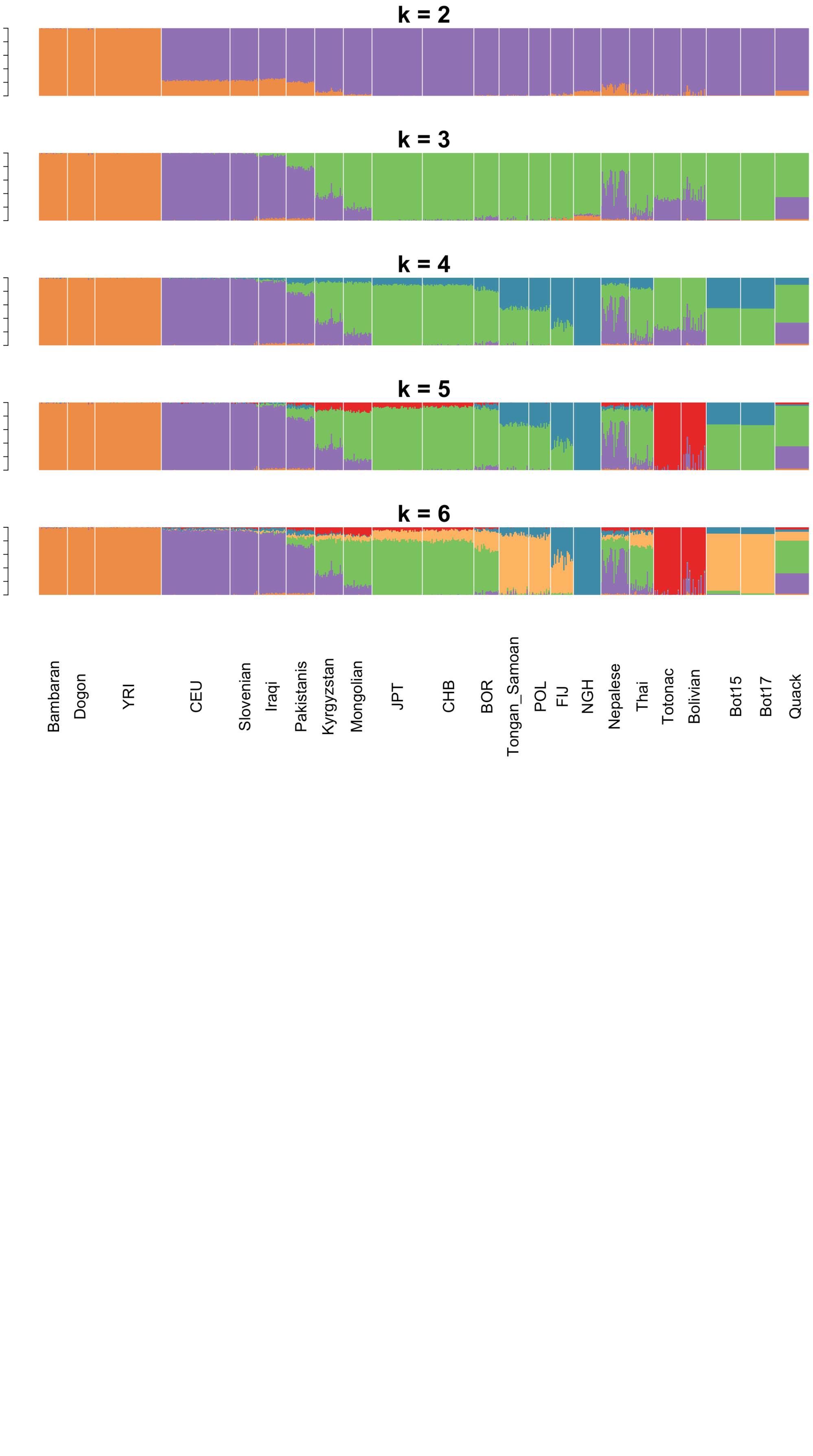
We used NGSAdmix to estimate the ancestry proportions for Quack. The comparison was made using the panel from Wollstein et al. (2010), used in Malaspinas et al. (2014). The panel comprises 583 individuals grouped in 20 worldwide populations/regions. Among these dataset, we have two Native American (Totonac and Bolivian), and several Remote Oceania (Polynesia, including various islands, and Fiji) populations.

The data from the single (MA2776 and MA2777) and combined (Quack) libraries, as well as that of Bot15 and Bot17 (Malaspinas et al. 2017) was merged to the panel on at most 28,207 sites.

We assumed 2 to 6 ancestral populations. The analysis reveals that Quack's ancestry corresponds mostly to that of the **European** (~30%, in purple) and **Asian** (~60%, in green) clusters.

Number of markers present in the libraries:

- MA2776: 23,228
- MA2777: 1,813
- Quack: 28,040


k = 2
k = 3
k = 4
k = 5
k = 6
Bambaran
Dogon
YRI
CEU
Slovenian
Iraqi
Pakistanis
Kyrgyzstan
Mongolian
JPT
CHB
BOR
ngan_Samoan
POL
FIJ
NGH
Nepalese
Thai
Totonac
Bolivian
Bot15
Bot17
Quack