Quack

Cruz Dávalos, Diana I.

30.10.2018

Summary of sequencing statistics

DNA was extracted from one tooth and built into two libraries. These libraries will be referred to as MA2776 and MA2777.

Library	Sequenced reads	Trashed reads after trimming (percentage)	Retained reads after trimming	Length retained reads	Duplicate reads	Unique endogenous reads	Unique endogenous (fraction)	Unique nuclear reads	Genome coverage	Length unique nuclear reads	Unique MT reads	MT coverage	Length unique MT reads
MA2776	392,956,393	12.79%	342,678,242	54	64.5%	2,409,896	0.00703	2,400,864	0.0351	45	9,032	29.4	53.9
MA2777	411,691,382	20.48%	327,366,561	47	95.9%	541,233	0.00165	535,492	0.00775	44.2	5,741	18.3	52.8
Quack	804,647,775	16.73%	670,044,803	50.6	85.2%	2,951,129	0.0044	2,936,356	0.0428	44.9	14,773	47.7	53.5

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 sequence using ANGSD and calling 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 of the merged dataset matches 23 out of 30 defining markers for the haplogroup **G3a1'2**. The remaining positions included 4 mismatches and 3 sites without calls.

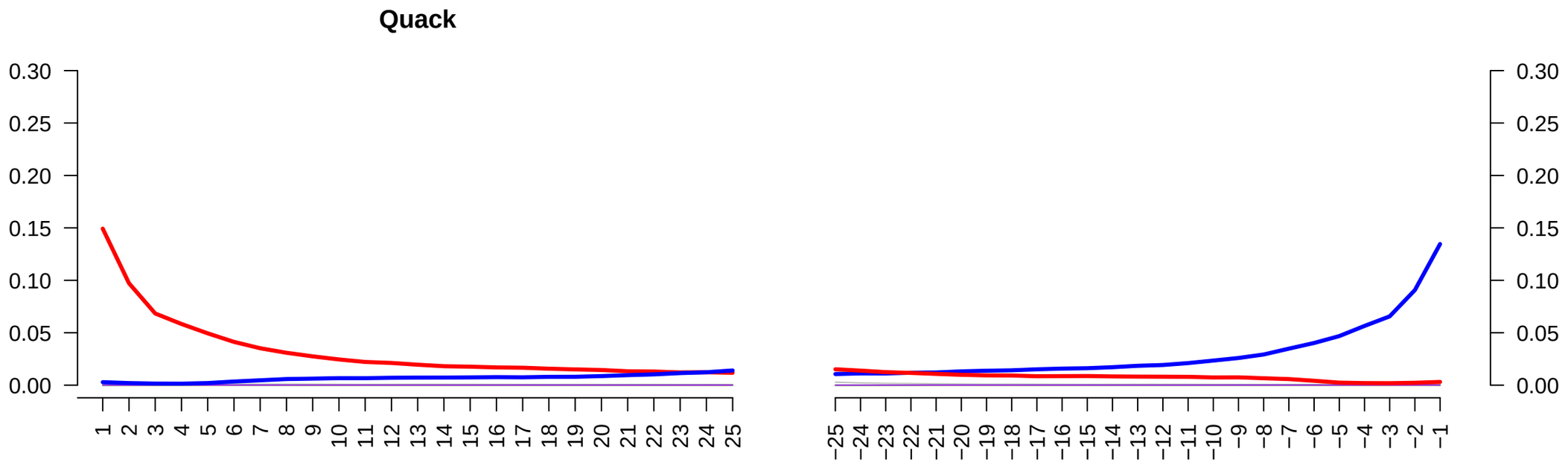
Position	Defining marker	Consensus in Quack	Consensus in MA2776	Consensus in MA2777
1438	G	G	G	A
4769	G	A	A	A
8701	G	A	A	A
11719	A	G	G	A
16362	C	T	T	T

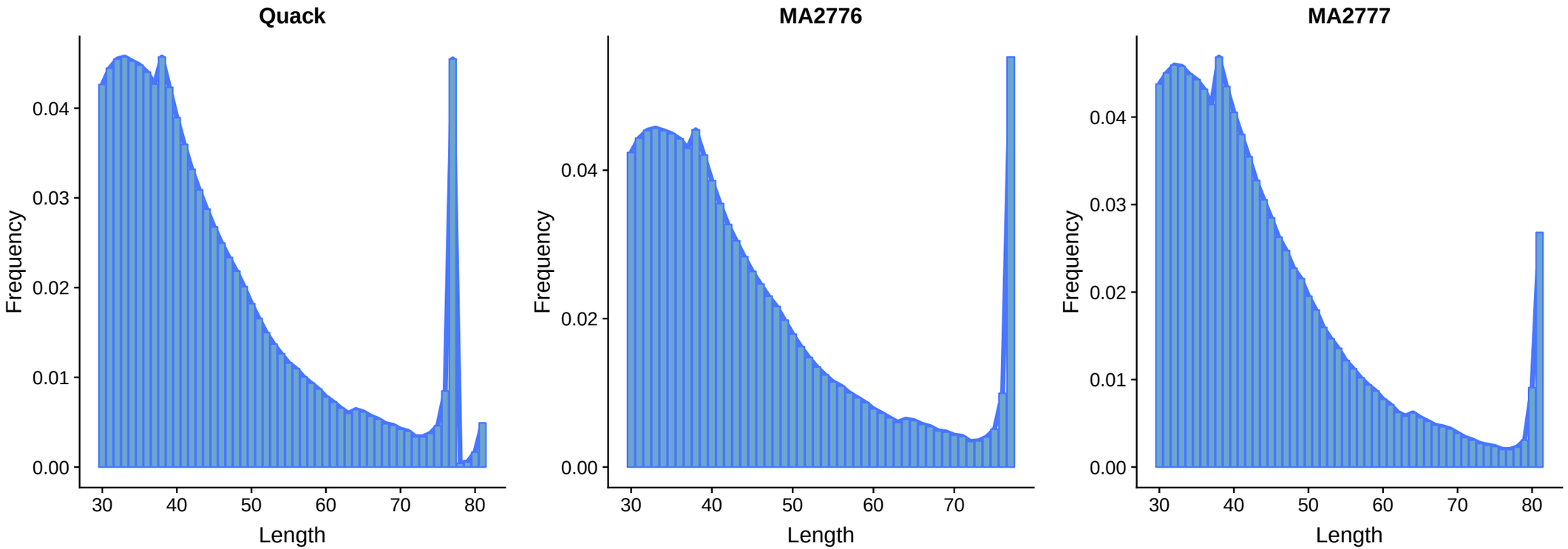
About haplogroup G

> 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. We also plotted the read length distribution of the mapped reads. The molecular damage patterns and short read lengths suggest that we recovered ancient DNA fragments.


Quack
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
−25 −24 −23 −22 −21 −20 −19 −18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1



Contamination estimates

We estimated the contamination levels per library using the method introduced in Fu et al. (2013). The results suggest that the libraries have contamination levels of 7.2% (2.5th percentile = 4.6%; 97.5th percentile = 10.5%) for MA2776, and 5.5% (2.5th percentile = 3.1%; 97.5th percentile = 9.7%) for MA2777.

Quack

Contamination: ~5% (2.5%, 7.0%)

Coverage on MT: 47x

```
Crude contamination upper bound: 995 out of 12701 reads (7.8%) match another
```

```
genome better than the consensus (error rate is 0.02).

quantiles from n = 99829 samples (after discarding burnin of 171)
     2.5%     97.5%
0.9305844 0.9743480
MAP authentic: 0.9559791
Potential scale reduction factors:

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

MA2776

Contamination: ~7% (4.7%, 10.4%)

Coverage on MT: 29x

```
Crude contamination upper bound: 628 out of 7771 reads (8.1%) match another
genome better than the consensus (error rate is 0.0199).

quantiles from n = 99805 samples (after discarding burnin of 195)
     2.5%     97.5%
0.8957751 0.9530261
MAP authentic: 0.9284921
Potential scale reduction factors:

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

MA2777

Contamination: ~5% (3.1%, 9.7%)

Coverage on MT: 18x

```
Crude contamination upper bound: 382 out of 4929 reads (7.8%) match another
genome better than the consensus (error rate is 0.02).

quantiles from n = 99867 samples (after discarding burnin of 133)
```

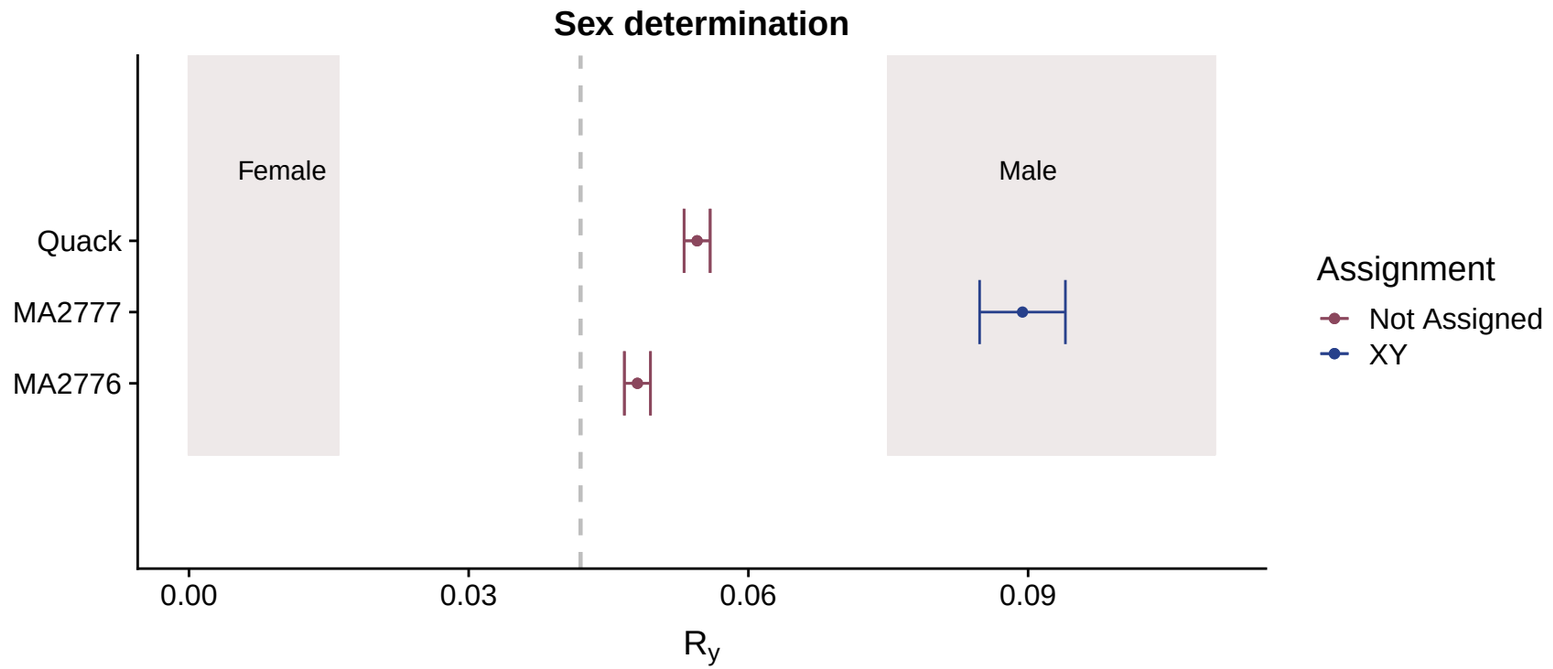
```
     2.5%     97.5%
0.9030789 0.9696461
MAP authentic: 0.9441228
Potential scale reduction factors:

     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 for the MA2777 library and the sex is not determined for MA2776 library, according to the thresholds defined in Skoglund et al. (2013) and shown in the figure below.

Library	Total reads	Reads (X + Y)	Reads (Y)	R_y	SE	95% CI	Assignment
MA2776	2409896	82632	3975	0.0481	0.0007	0.0466-0.0496	Not Assigned
MA2777	541187	15057	1346	0.0894	0.0023	0.0848-0.094	XY
Quack	2951083	97689	5321	0.0545	0.0007	0.053-0.0559	Not Assigned



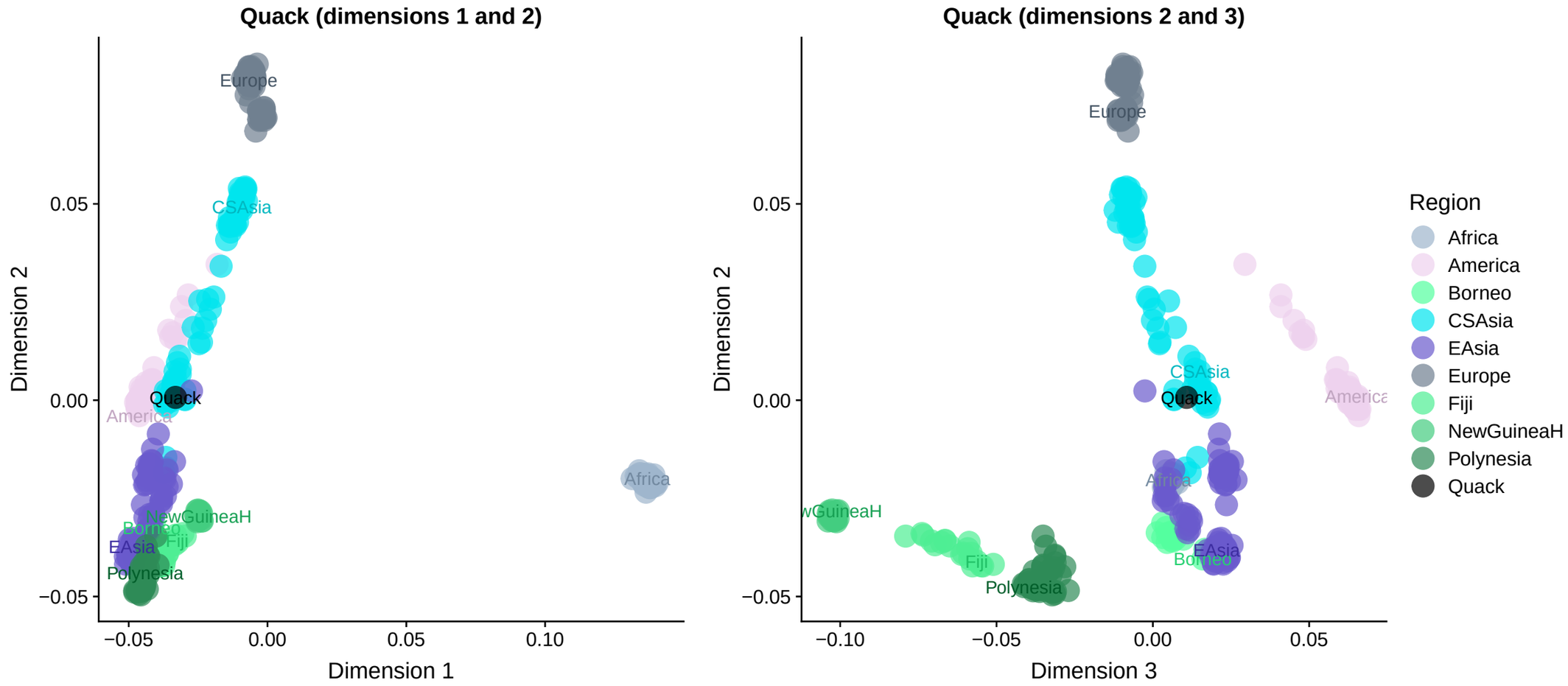
MDS

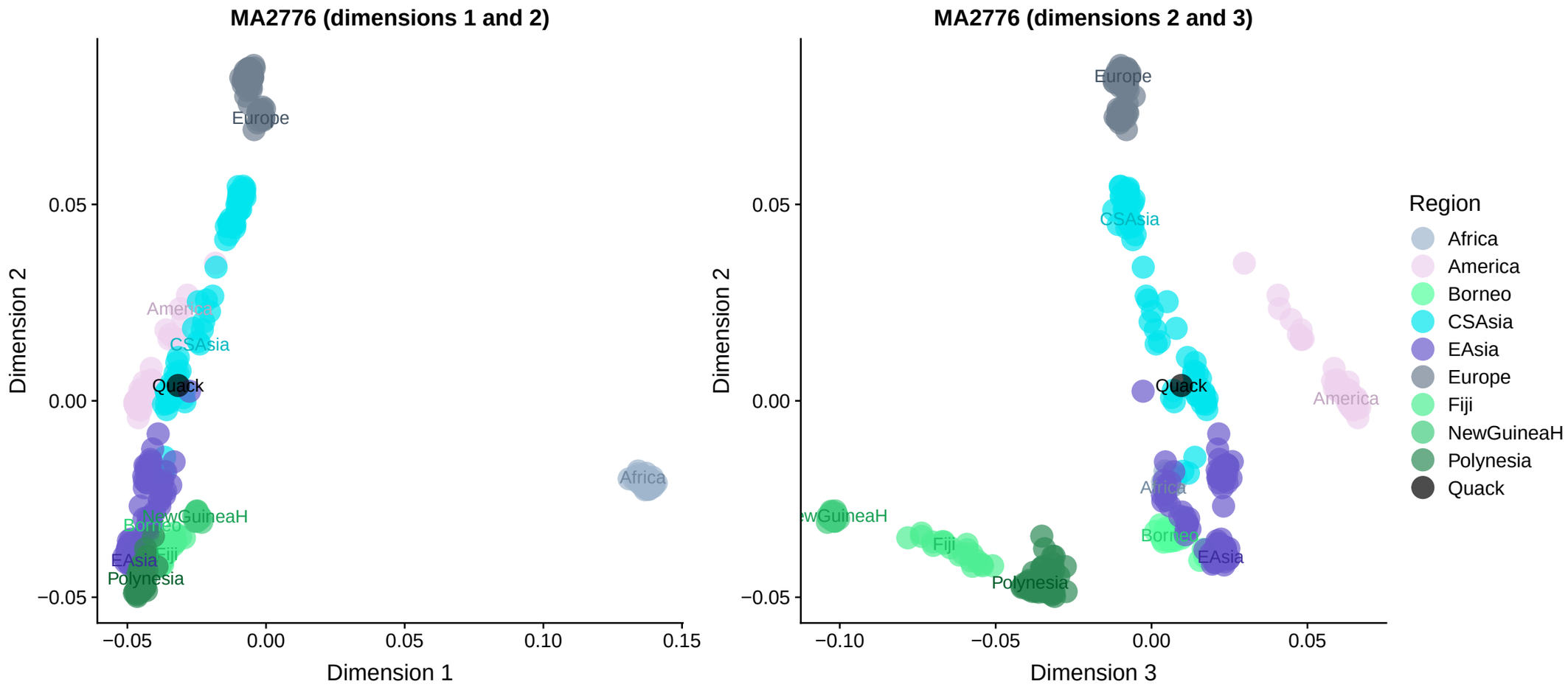
We performed a multidimensional scale analysis to visualize the genetic affinities of Quack. For this purpose, we used bammds (Malaspinas et al. 2014) and a panel with worldwide populations (Wollstein et al. 2010). We observe that Quack is not grouping with Native American populations, but with Central-South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

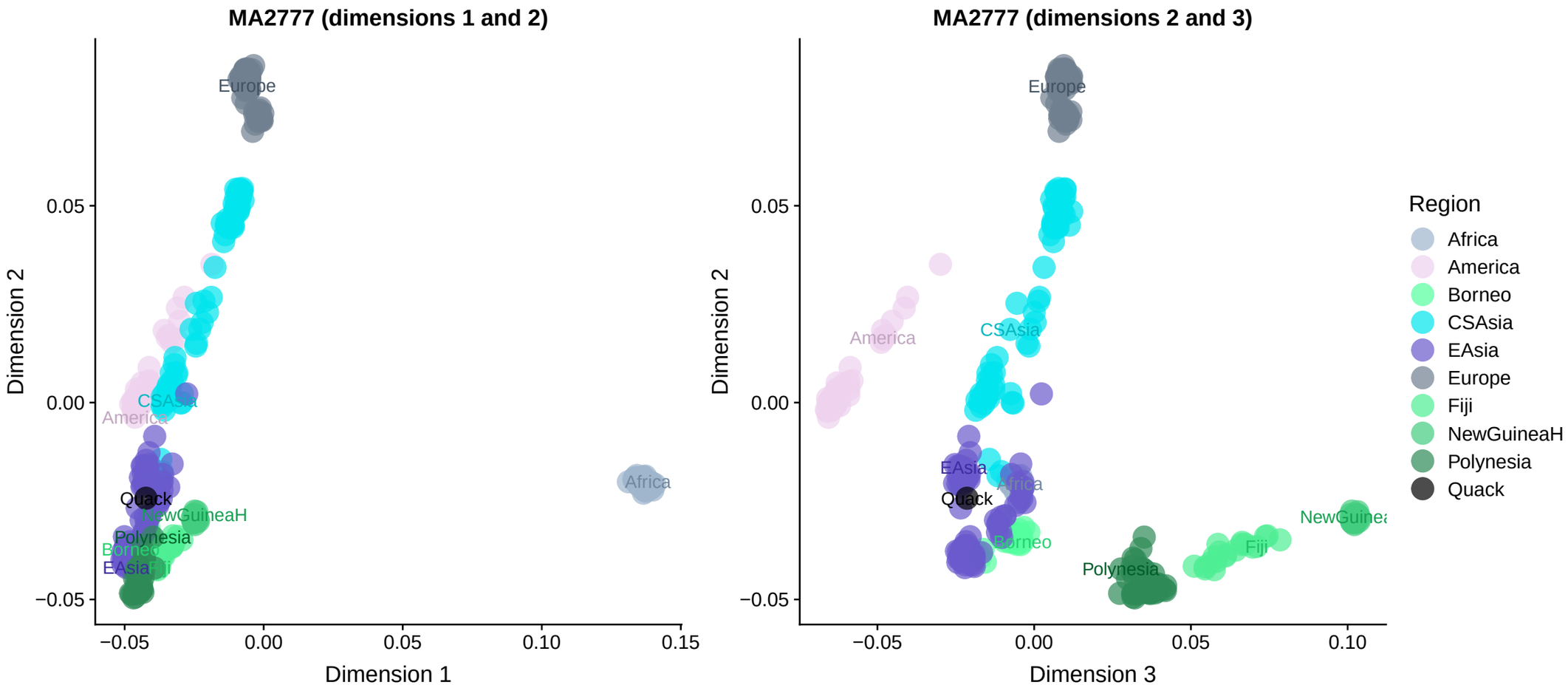
Quack: 40,144 out of 823,541 markers

MA2776: 34,022 out of 823,586

MA2777: 6,384 out of 823,760


Quack (dimensions 1 and 2)
Europe
CSAsia
Quack
America
Africa
NewGuineaH
Borneo
Fiji
EAsia
Polynesia
0.05
0.00
−0.05
Dimension 2
−0.05
0.00
0.05
0.10
Dimension 1
Quack (dimensions 2 and 3)
Europe
CSAsia
Quack
America
Africa
wGuineaH
Fiji
Polynesia
EAsia
Borneo
0.05
0.00
−0.05
Dimension 2
−0.10
−0.05
0.00
0.05
Dimension 3
Region
Africa
America
Borneo
CSAsia
EAsia
Europe
Fiji
NewGuineaH
Polynesia
Quack


MA2776 (dimensions 1 and 2)
Dimension 2
Dimension 1
0.05
0.00
−0.05
−0.05
0.00
0.05
0.10
0.15
Europe
America
CSAsia
Quack
Africa
NewGuineaH
Borneo
Fiji
EAsia
Polynesia
MA2776 (dimensions 2 and 3)
Dimension 2
Dimension 3
0.05
0.00
−0.05
−0.10
−0.05
0.00
0.05
Europe
CSAsia
Quack
America
Africa
Fiji
Polynesia
EAsia
Region
Africa
America
Borneo
CSAsia
EAsia
Europe
Fiji
NewGuineaH
Polynesia
Quack

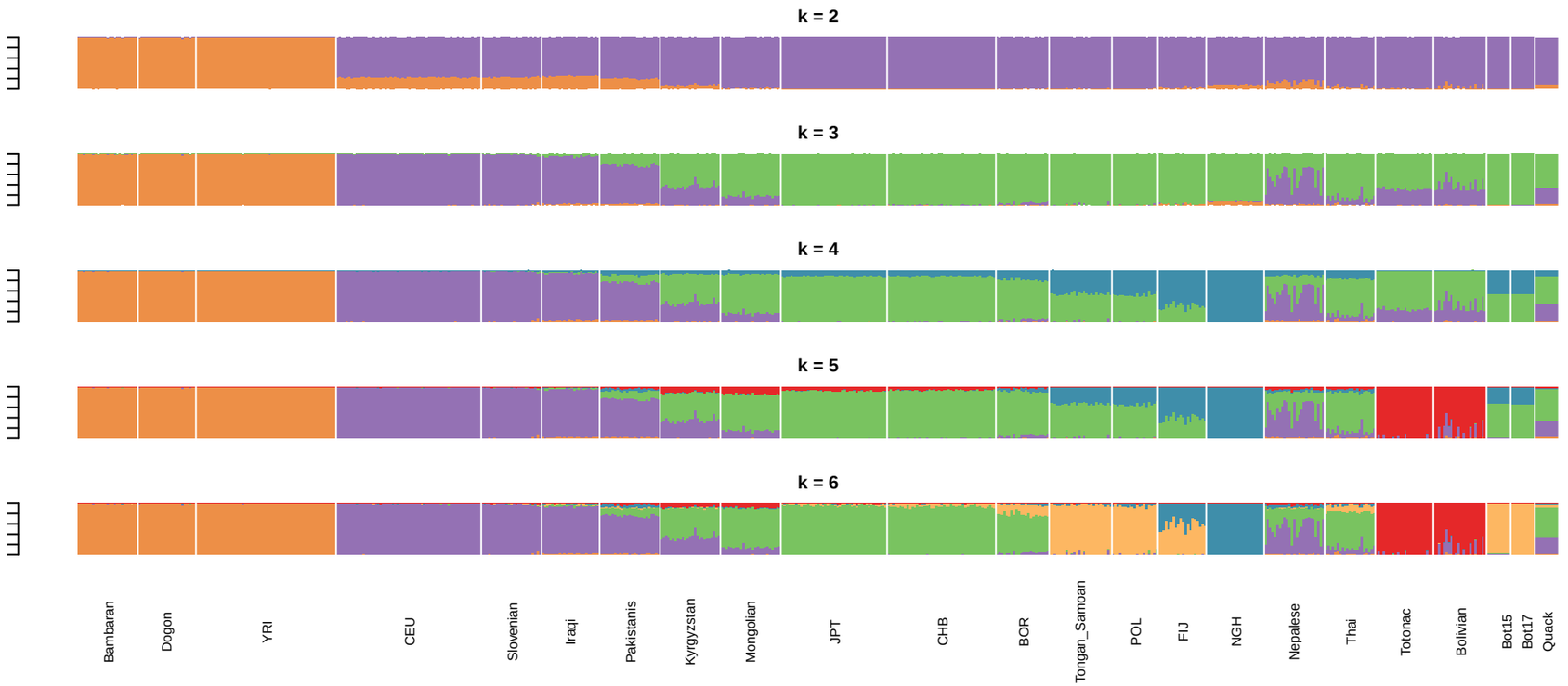


Ancestry estimates

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 combined (MA2776 and MA2777) libraries as well as that of Bot15 and Bot17 (Malaspinas et al. 2014) were merged to the panel on at most 31,373 sites.

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



We repeated the analysis using only a subset of the populations from the Wollstein panel, and a total of 9,701 sites.


k = 2
k = 3
k = 4
k = 5
k = 6
CEU
JPT
Bolivian
POL
Thai
Nepalese
Kyrgyzstan
Mongolian
Bot15
Bot17
Quack