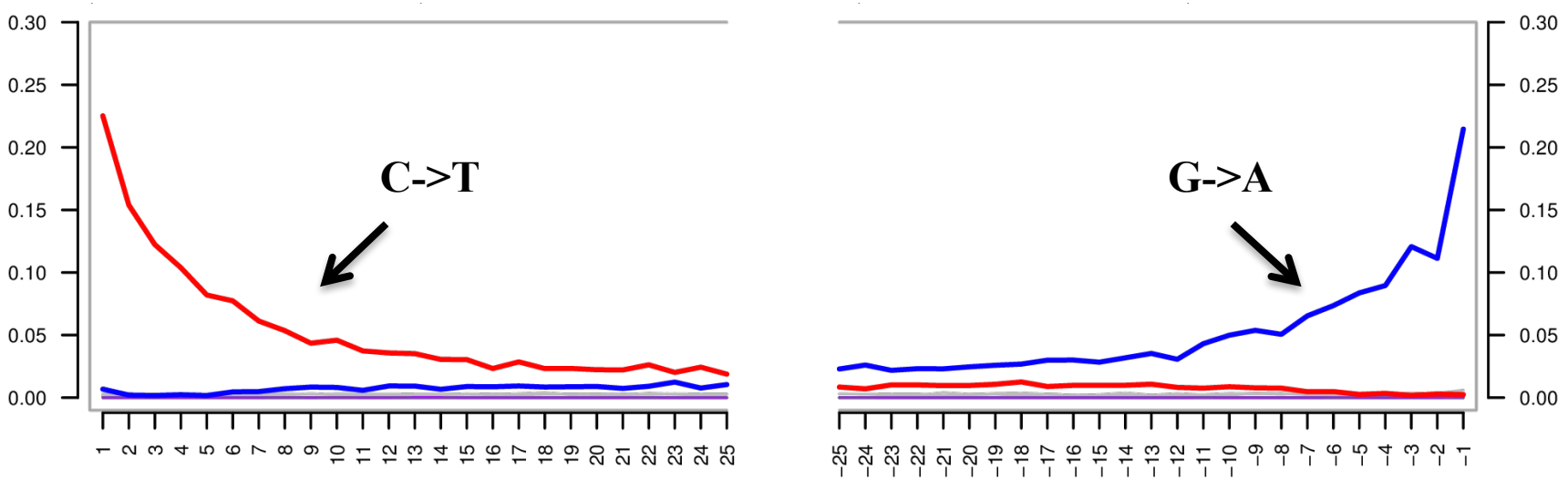
An R package for clustering and visualizing ancient dna signatures

By Hussein Al-Asadi & Kushal Dey

mapDamage

mapDamage: tracking and quantifying damage patterns in ancient DNA sequences



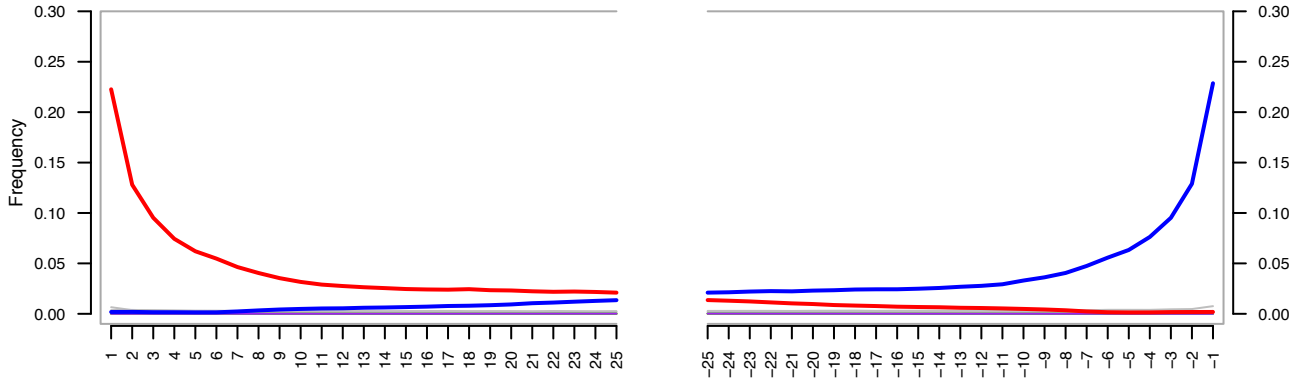
Data-set # 1

- DiRenzio Lab: 48 Ancients and 7 control samples

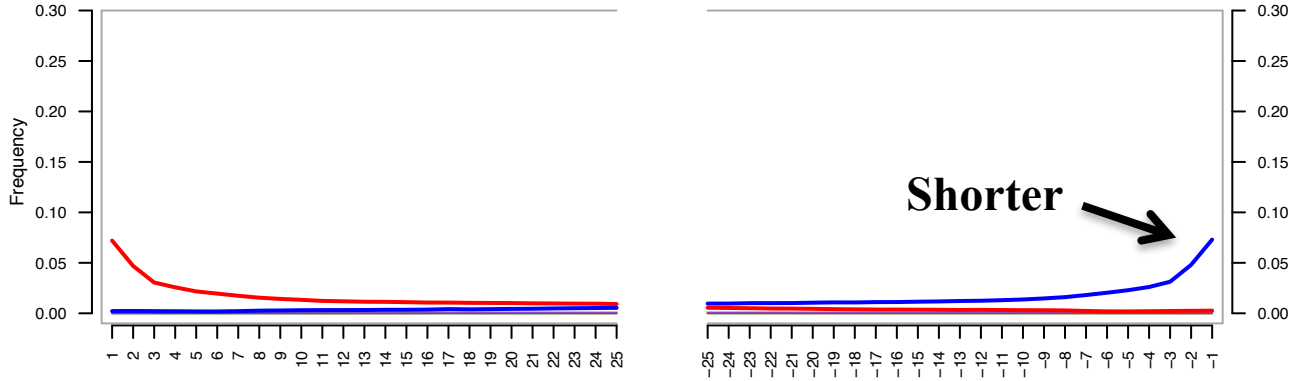
- Questions:
  - Do the ancients look “ancient”?
  - Is there any DNA in the the controls?
    - If so, is the DNA modern or ancient?

MapDamage on Ancients

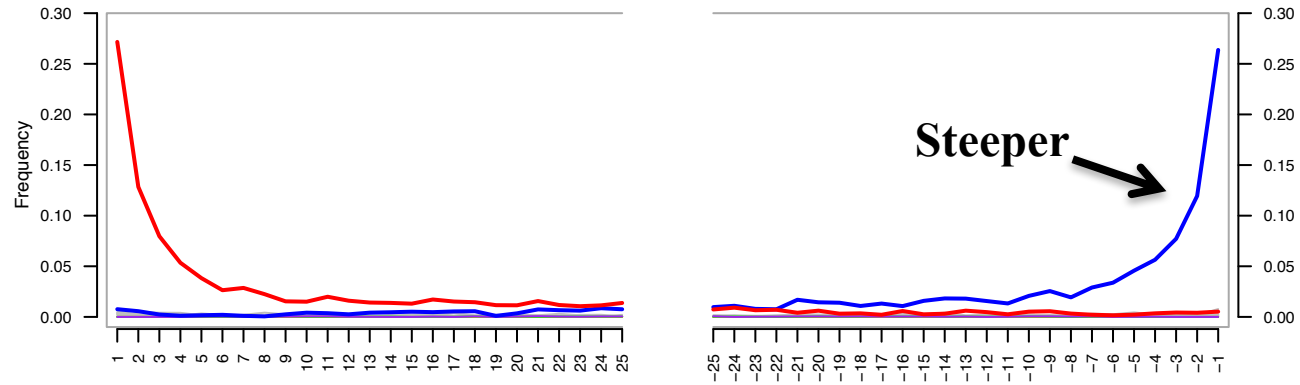
CNE1



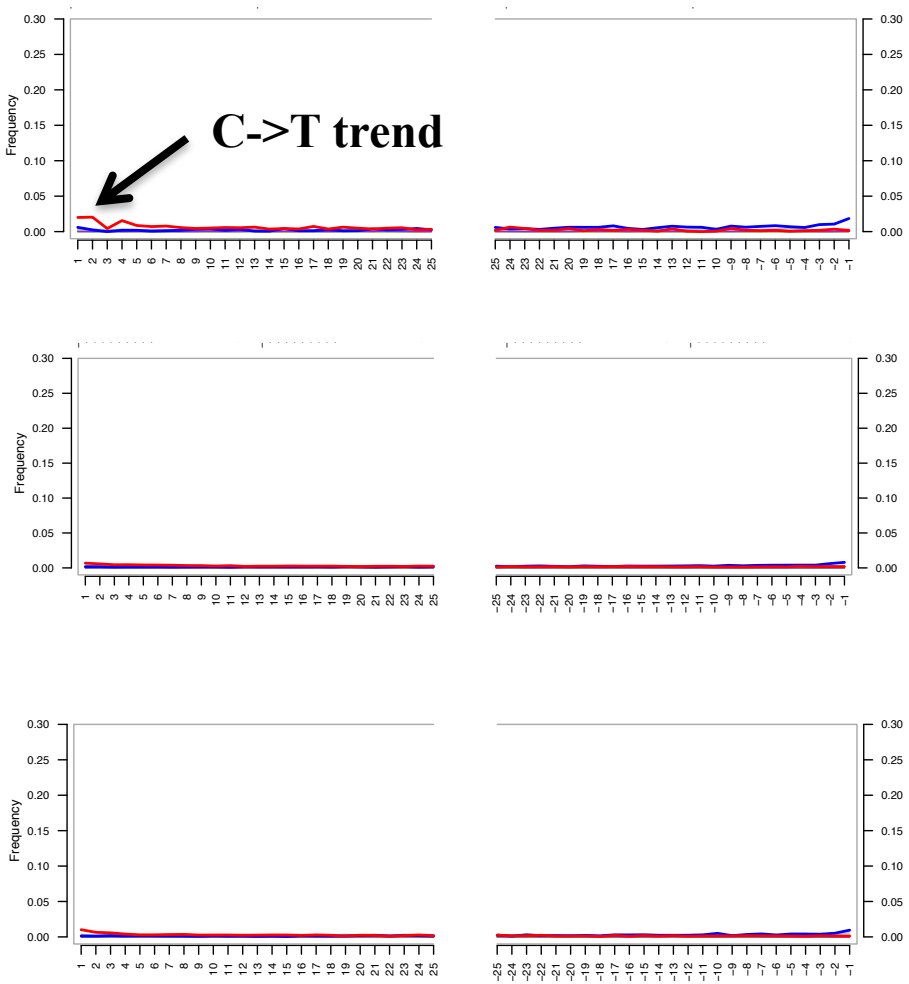
KS20

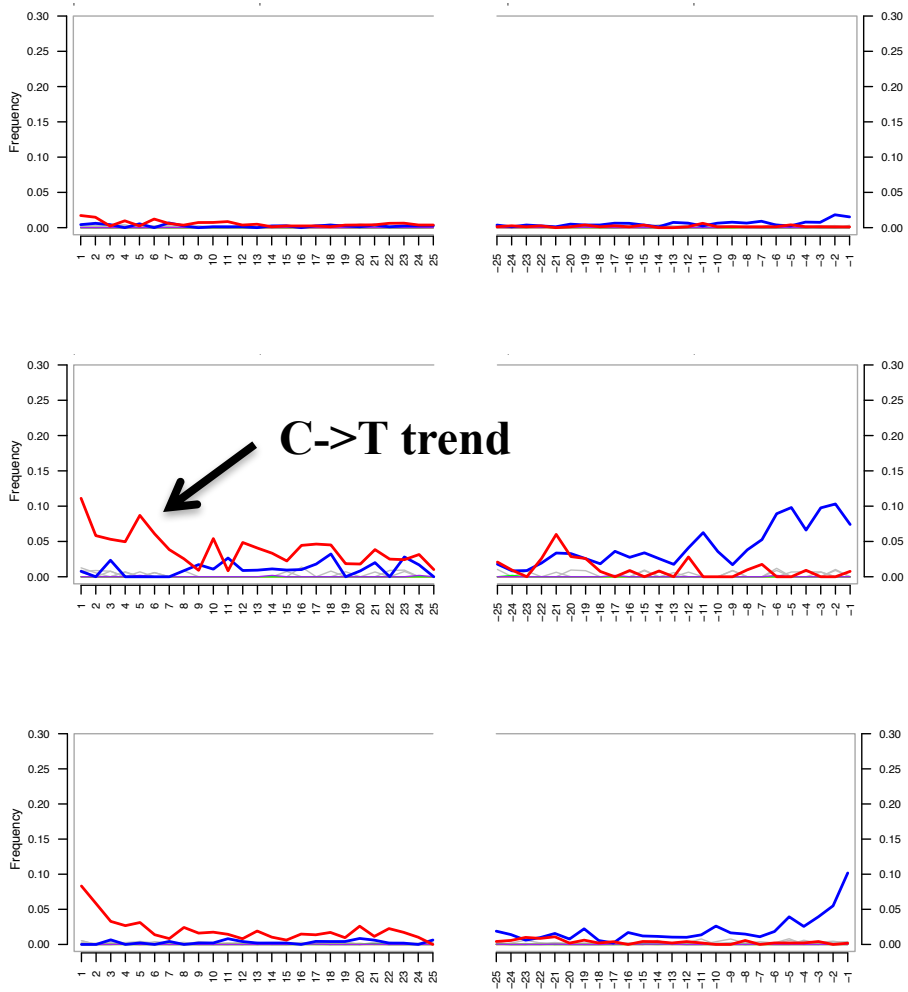


S30



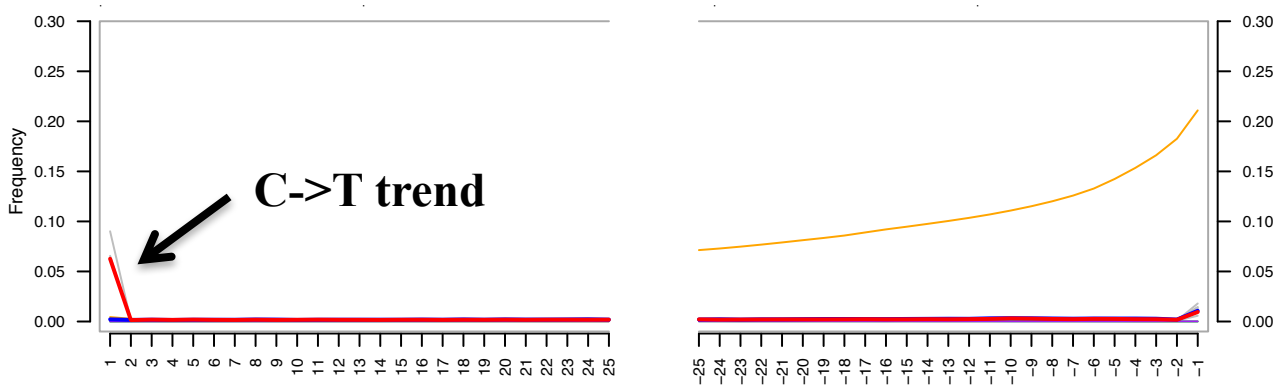
MapDamage on Controls



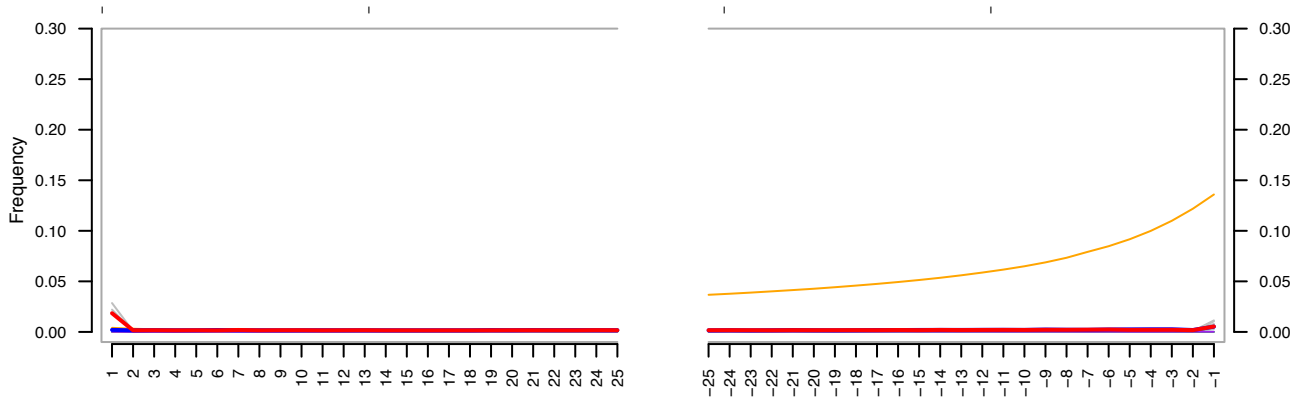


MapDamage on Moderns

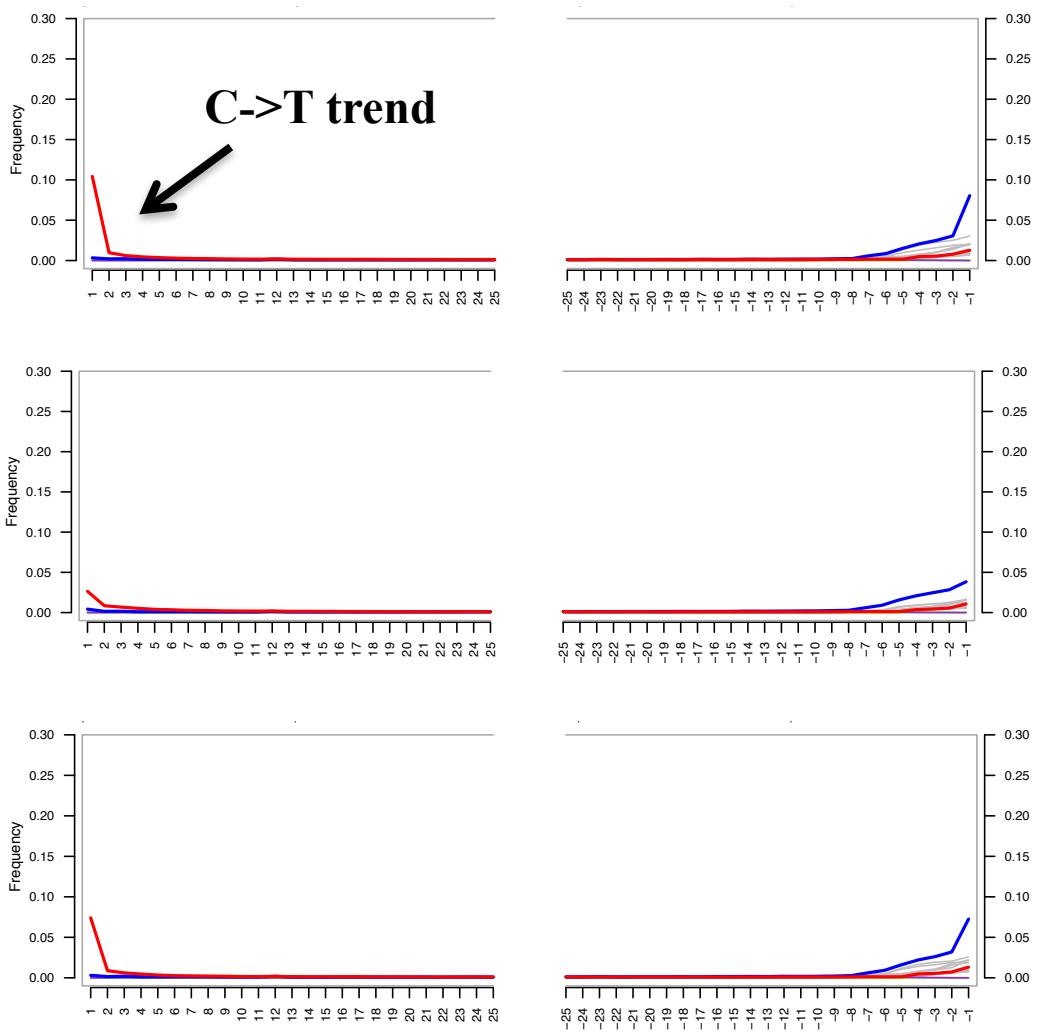
HG00097



HG00099



MapDamage on Ancient Sards. (different data-set)



Thoughts…

It’s not clear what is an “ancient” C->T damage profile and what is an “modern” profile?

- Difficult to perform comparative analysis
  - MapDamage is an individual-by-individual analysis
  - Hard to look at all plots simultaneously

- Are we missing other patterns?

High-level overview of aRchaIc

- **Step 1:** Gets “mutational patterns” from BAM files, which are flexible

- **Step 2:** Cluster and visualize samples based on “mutational patterns”.

Step 1: Get mutational patterns from BAM files

- Mutational pattern: mutation, flanking bases (we use 2), and position from the end of the read.
  - Loop through bam file and record the number of all such patterns

- Example output for Sample_1
  - AA(T->A)AA, 0, 50, 7
  - AA(T->A)AA, 10, 40, 3
  - GC(C->T)TT, 2, 48, 6

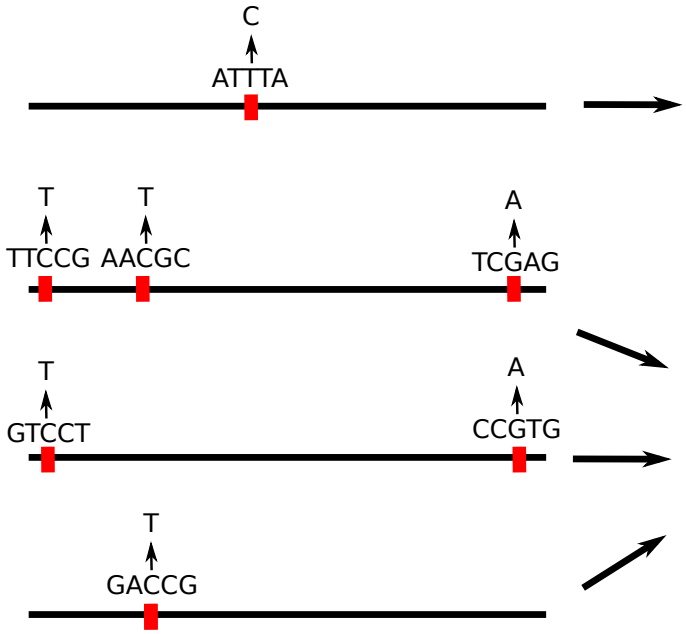
Step 2: Clustering and visualization

- STRUCTURE: if K=2, a African-American individual will be mixture of African ancestry and European" ancestry where ancestries are defined by a probability vector on allele frequencies

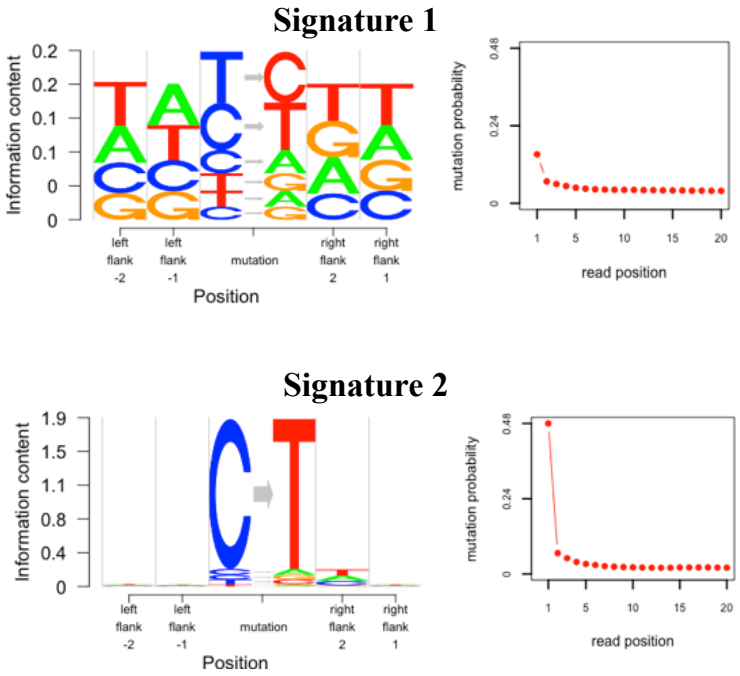
- aRchaic: if K=2, a contaminated individual will be mixture of ancient ancestry and modern ancestry where ancestries are defined by a probability vector on mutational patterns

Graphical overview of aRchaic

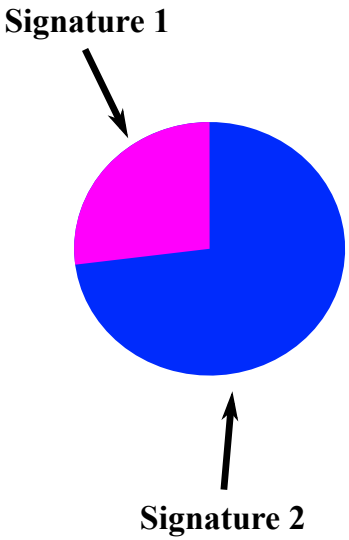
actual observed mutational patterns

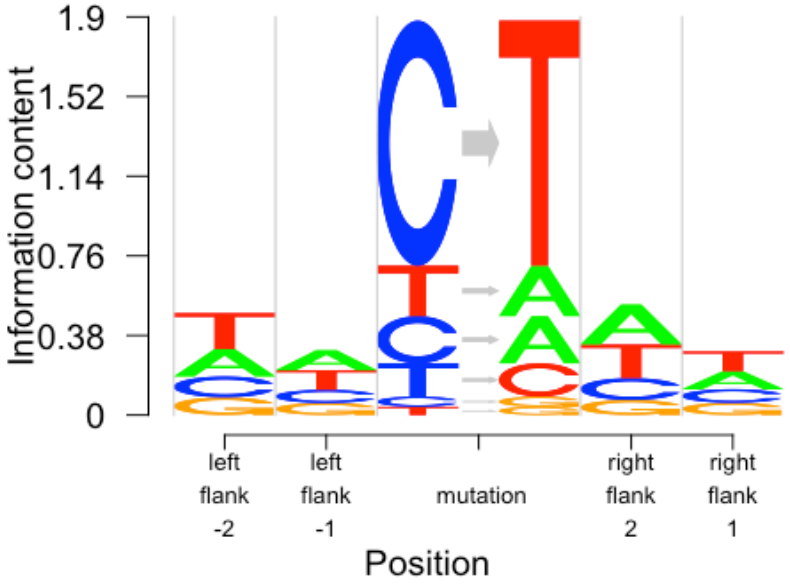


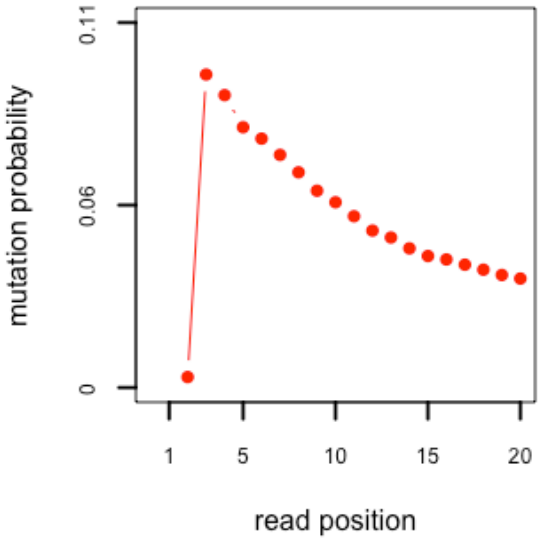
operative mutation signatures



individuals are mixtures of signatures





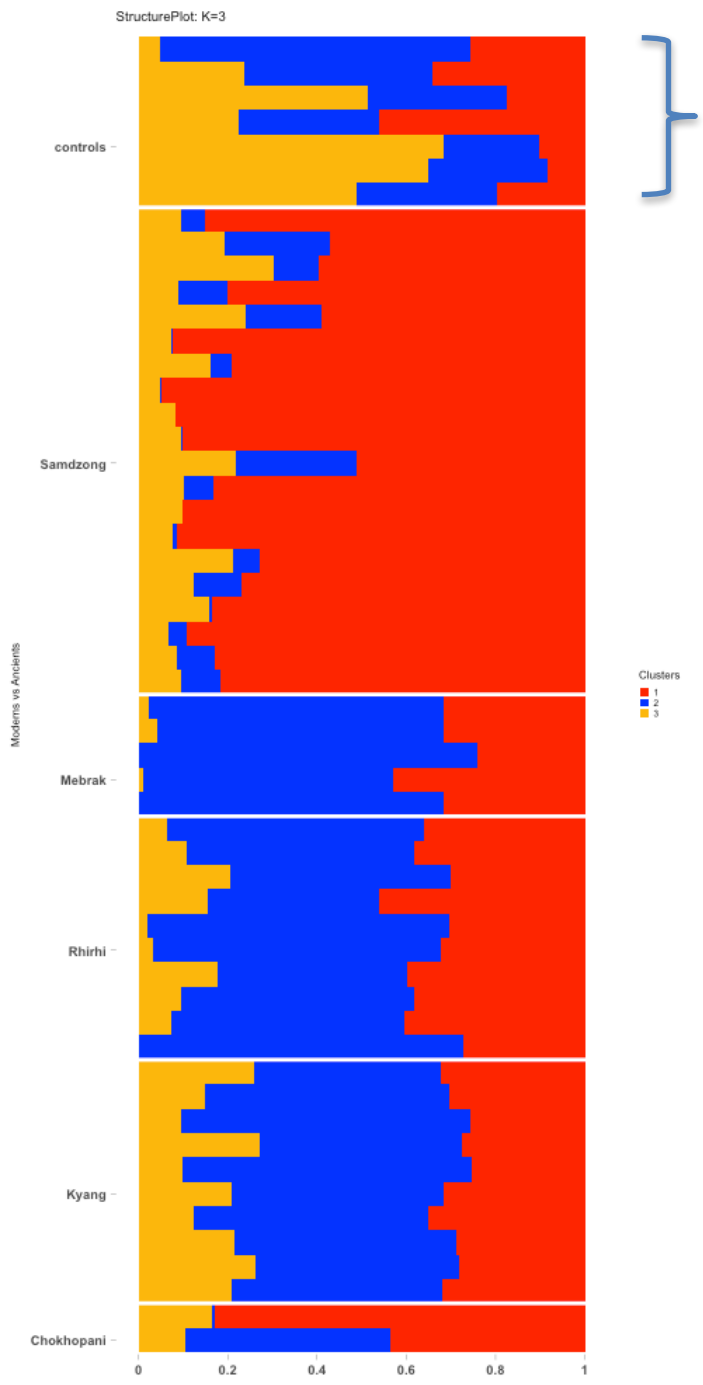


	Left flank2	Left flank 1	Right flank 1	Right flank 2
A	0.27	0.31	0.36	0.28
C	0.20	0.20	0.19	0.21
G	0.17	0.18	0.13	0.19
T	0.35	0.30	0.30	0.31

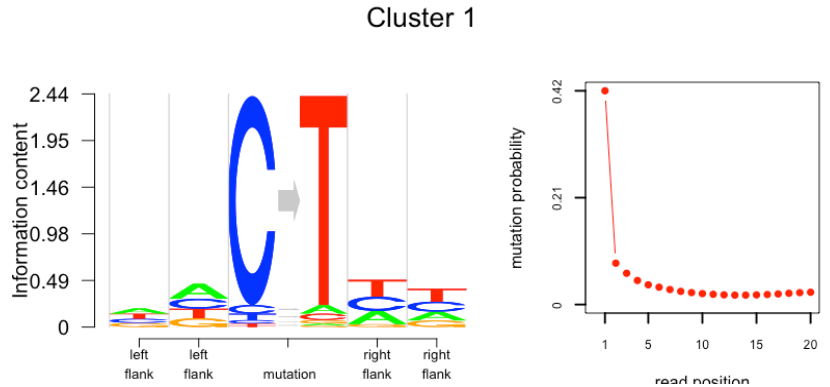
mutations	prob
C -> T	0.62
C -> A	0.12
C -> G	0.03
T -> A	0.13
T -> G	0.08
T -> C	0.02

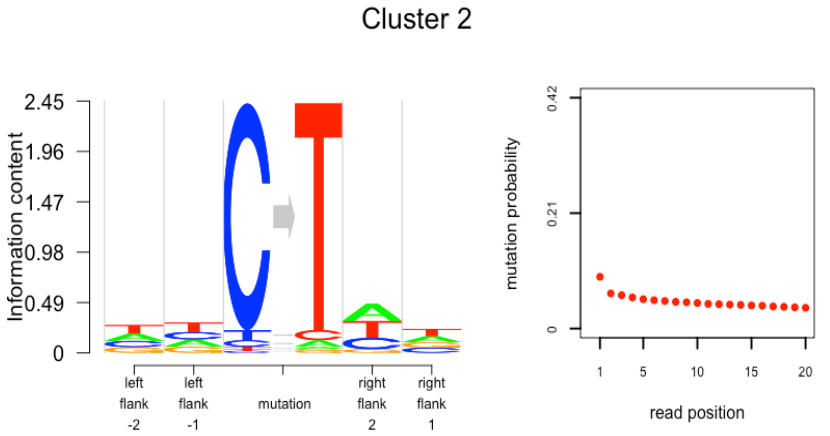
Applications of aRchaIc on real data

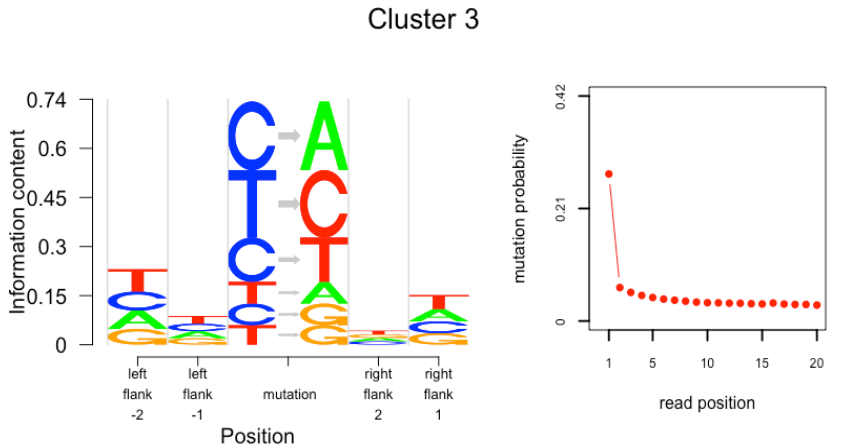
47 ancients + 7 controls from Di Rienzo lab



PCRNeg
Libneg3
Libneg2
Libneg1
EXN3
EXN2
EXN1

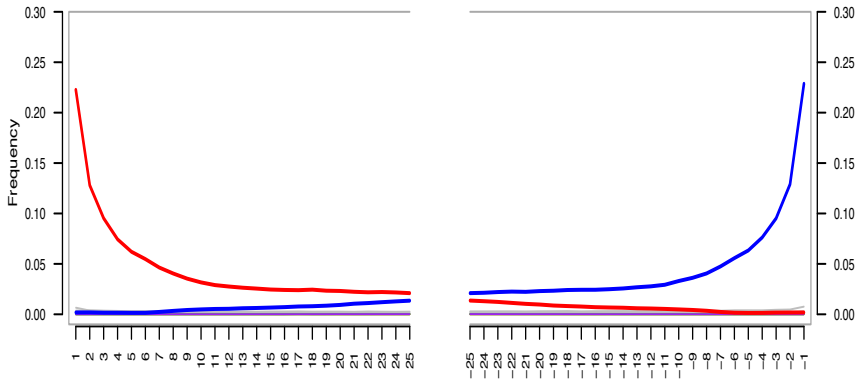




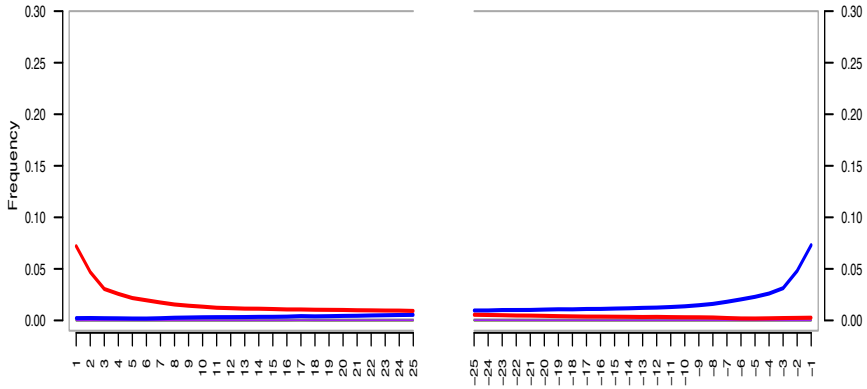


MapDamage on Ancients

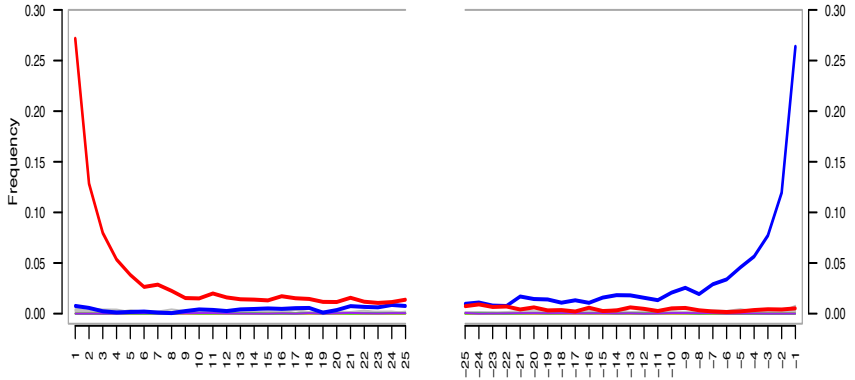
CNE1

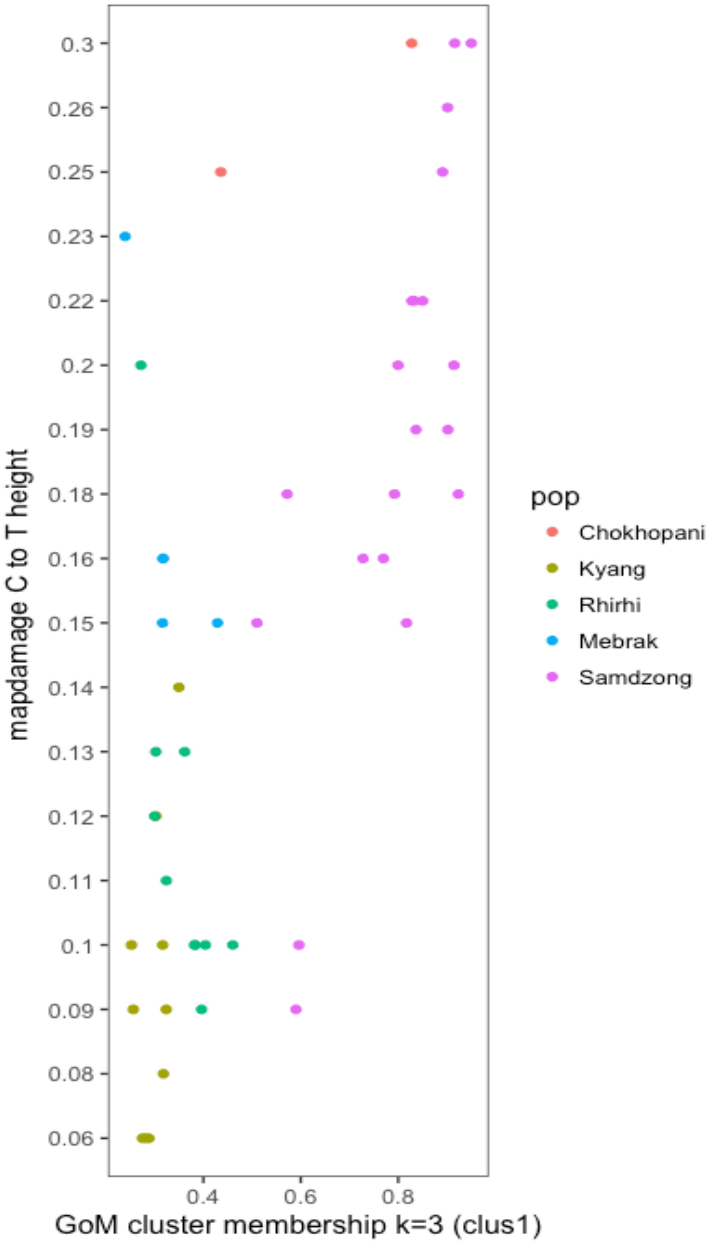


KS20



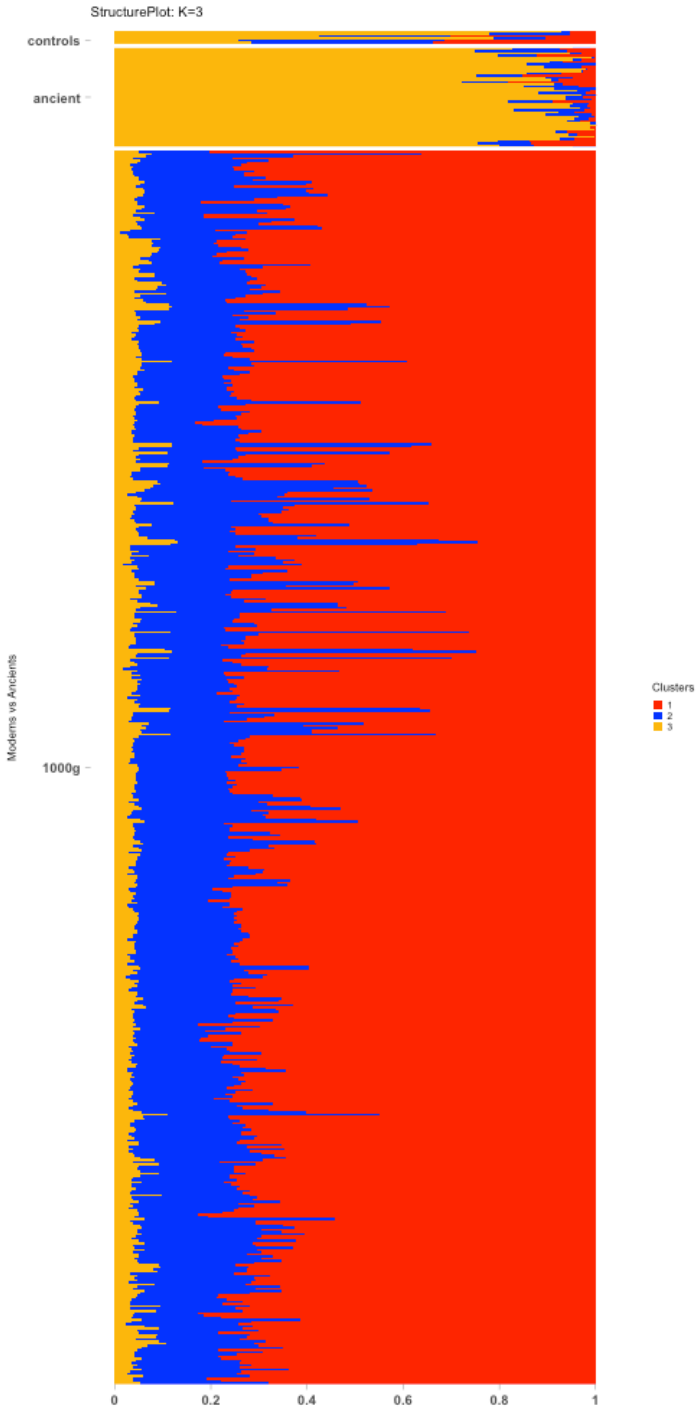
S30

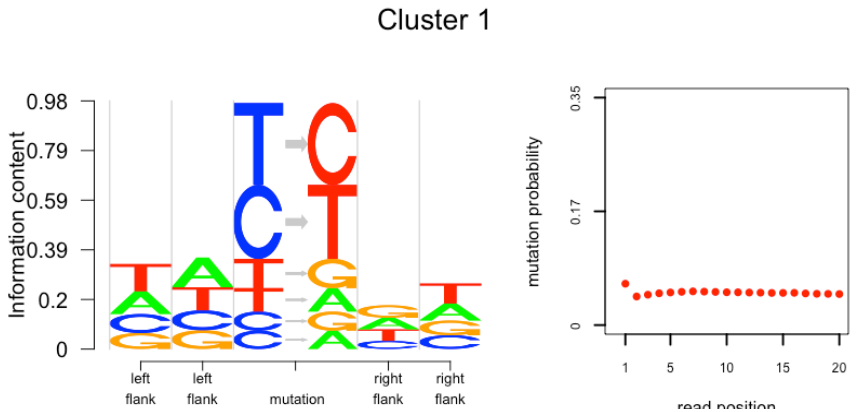


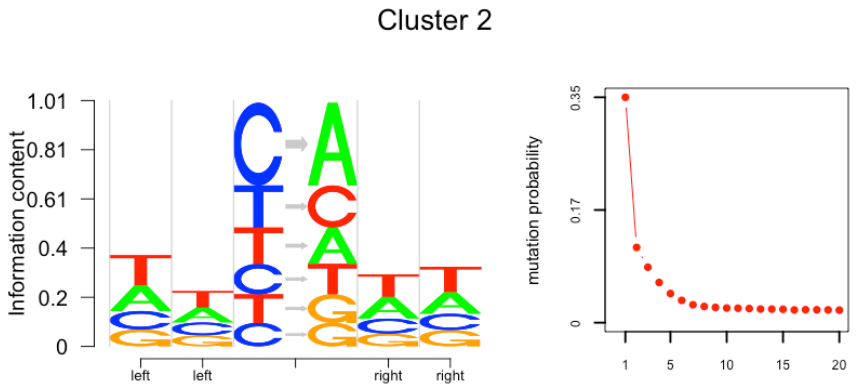


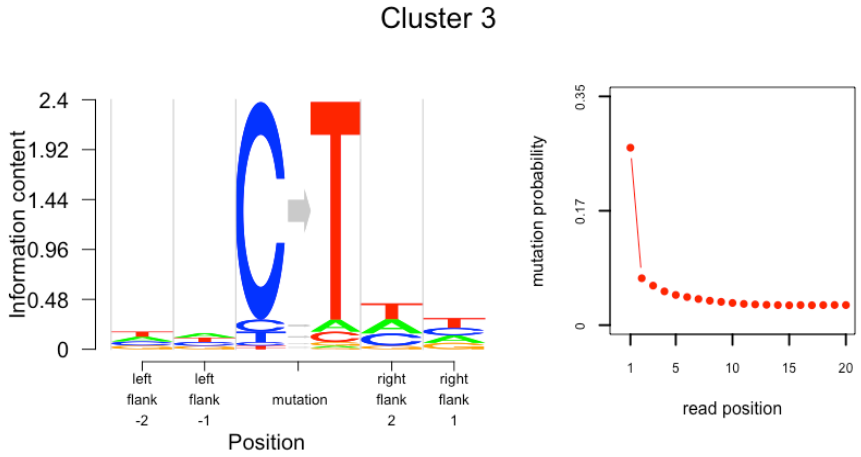
Applications of aRchaIc on real data

47 ancients + 7 controls from Di Rienzo lab + 500 randomly selected 1000 Genome modern samples of European and American ancestry


StructurePlot: K=3
controls
ancient
1000g
Moderns vs Ancients
Clusters
1
2
3
0
0.2
0.4
0.6
0.8
1

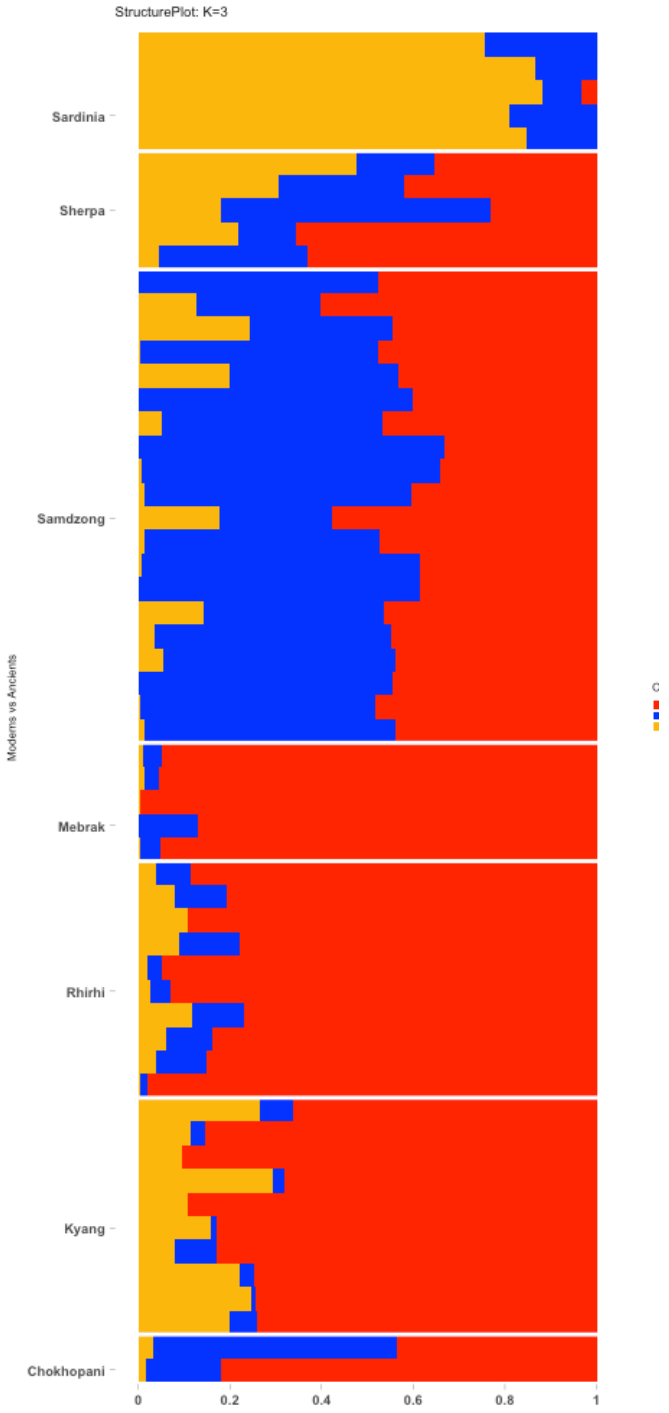

Cluster 1
Information content
0.98
0.79
0.59
0.39
0.2
0
left flank
left flank
mutation
right flank
right flank
mutation probability
0.35
0.17
0
1
5
10
15
20
read position


Cluster 2
Information content
1.01
0.81
0.61
0.4
0.2
0
left
left
right
right
mutation probability
0.35
0.17
0
1
5
10
15
20

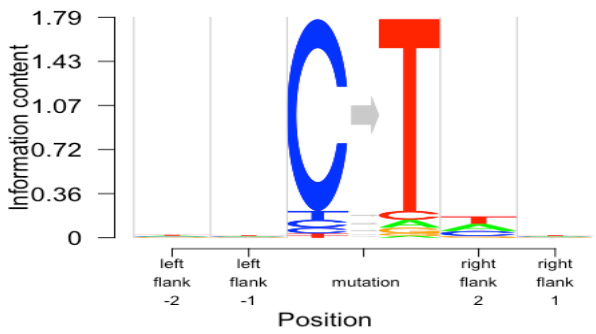

Cluster 3
Information content
2.4
1.92
1.44
0.96
0.48
0
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position
mutation probability
0.35
0.17
0
1
5
10
15
20
read position

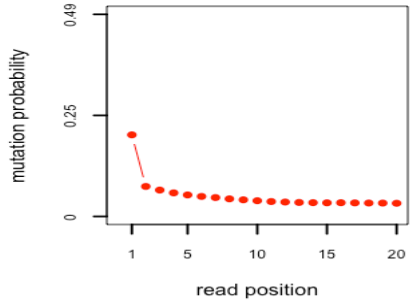
Applications of aRchaIc on real data

47 ancients + 5 Sherpa samples + 5 ancient Sardinia samples recently collected

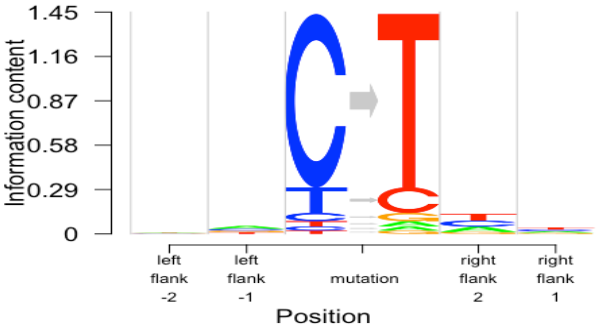


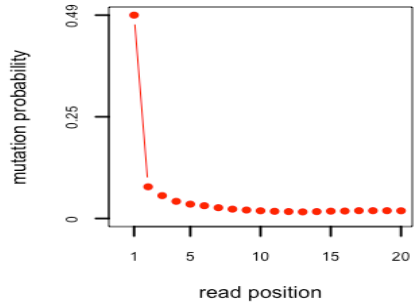
Cluster 1



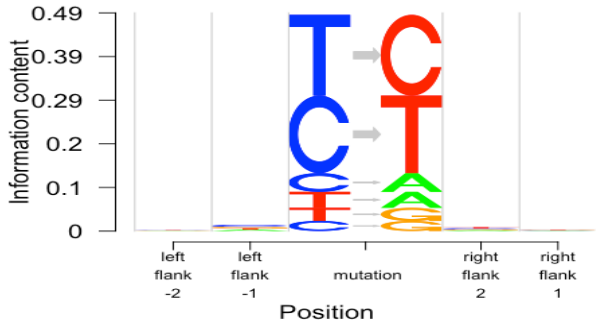


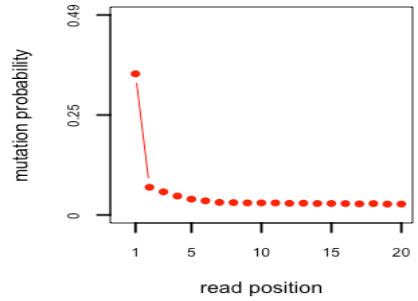
Cluster 2





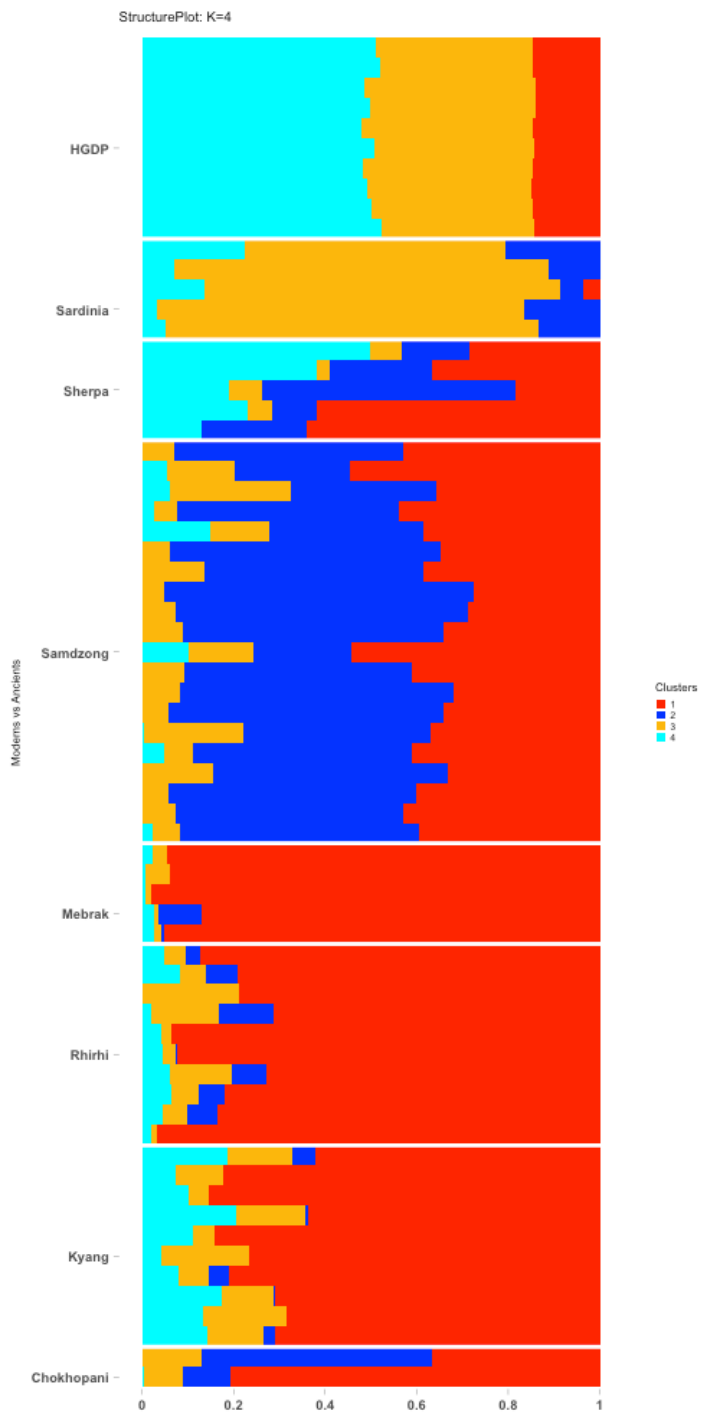
Cluster 3

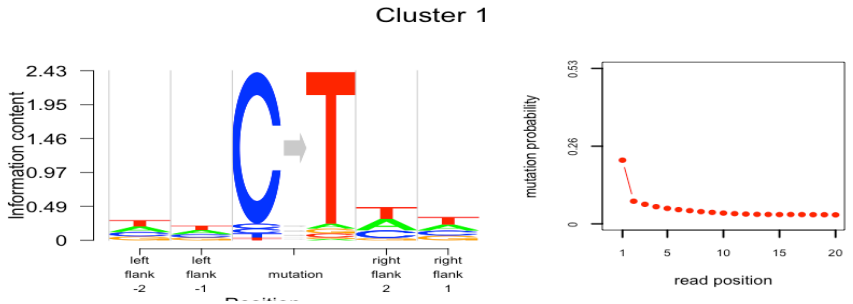


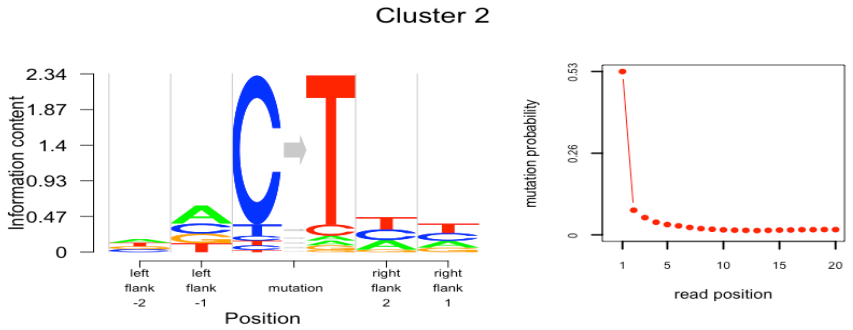


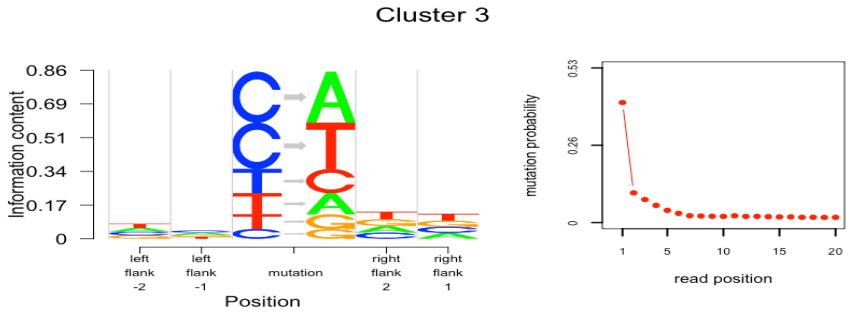
Applications of aRchaIc on real data

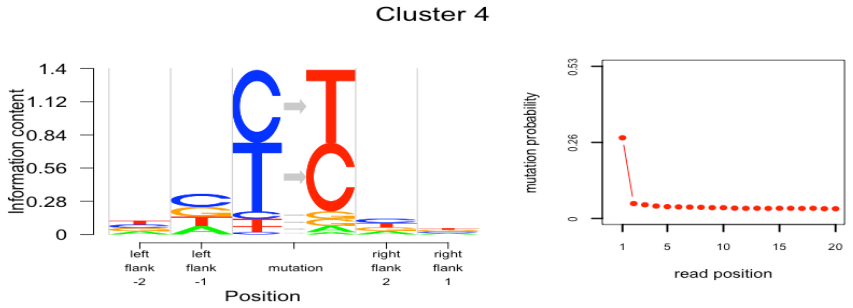
47 ancients + 5 Sherpa samples + 5 ancient Sardinia samples recently collected + 10 HGDP modern samples


StructurePlot: K=4
HGDP
Sardinia
Sherpa
Samdzong
Mebrak
Rhirhi
Kyang
Chokhopani
Moderns vs Ancients
0
0.2
0.4
0.6
0.8
1
Clusters
1
2
3
4


Cluster 1
Information content
2.43
1.95
1.46
0.97
0.49
0
left flank -2
left flank -1
mutation
right flank 2
right flank 1
mutation probability
0.53
0.26
0
1
5
10
15
20
read position

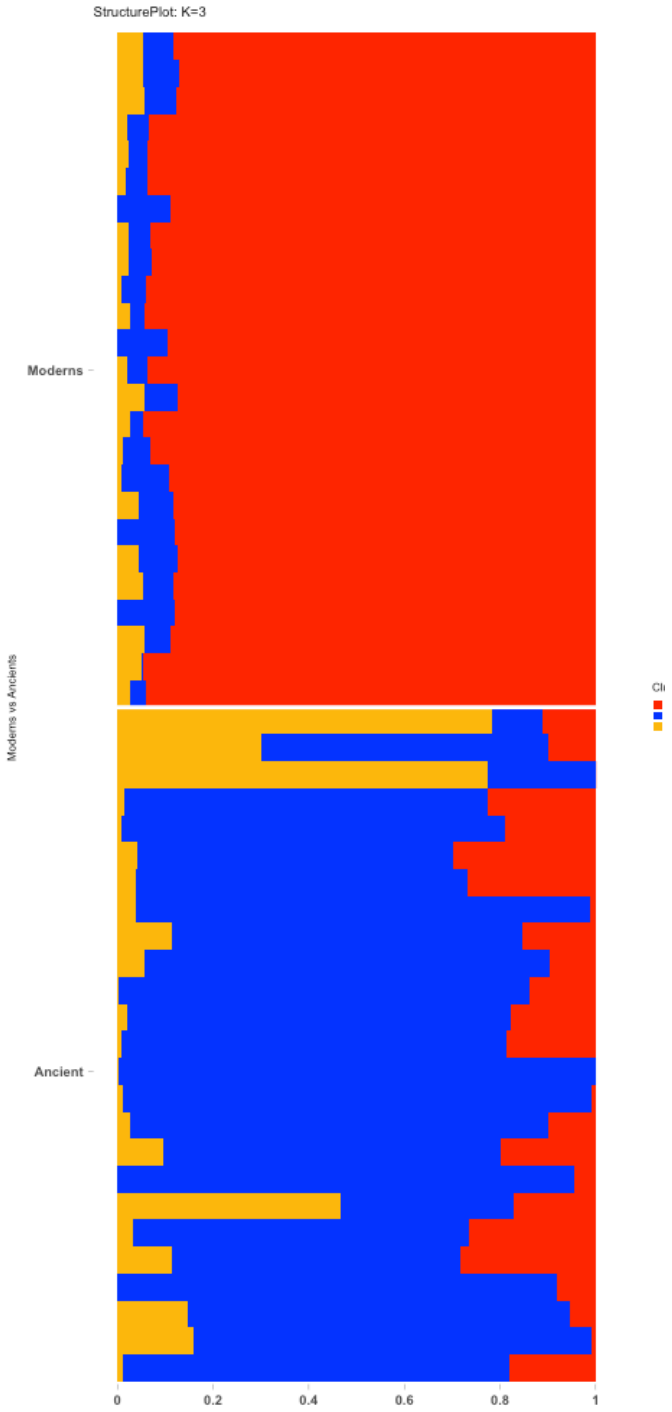

Cluster 2
Information content
2.34
1.87
1.4
0.93
0.47
0
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position
mutation probability
0.53
0.26
0
1
5
10
15
20
read position


Cluster 3
Information content
0.86
0.69
0.51
0.34
0.17
0
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position
mutation probability
0.53
0.26
0
1
5
10
15
20
read position

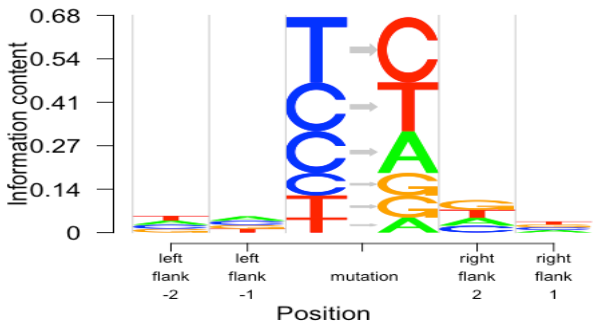

Cluster 4
Information content
1.4
1.12
0.84
0.56
0.28
0
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position
mutation probability
0.53
0.26
0
1
5
10
15
20
read position

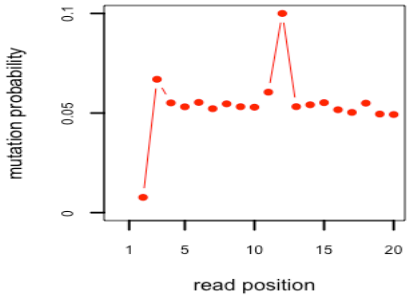
Applications of aRchaIc on real data

25 ancients + 25 moderns from the ancient data collected by John Lindo (DiRienzo Lab)

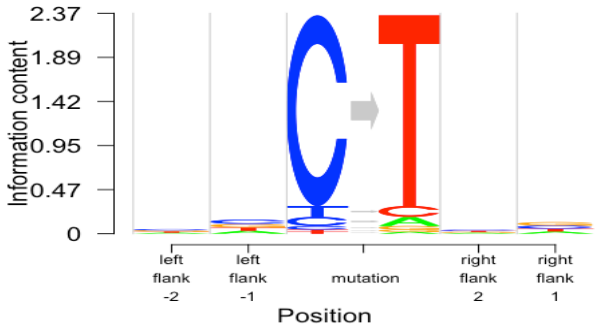


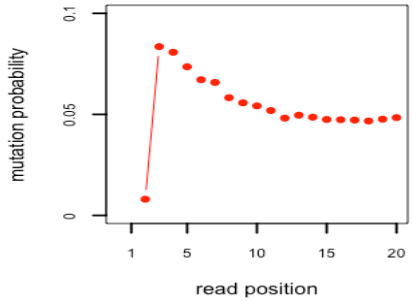
Cluster 1



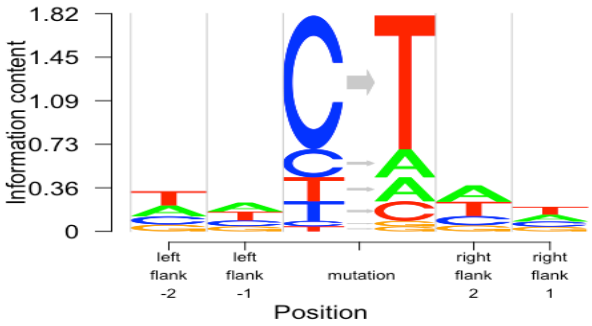


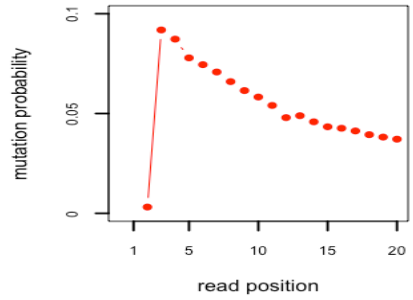
Cluster 2

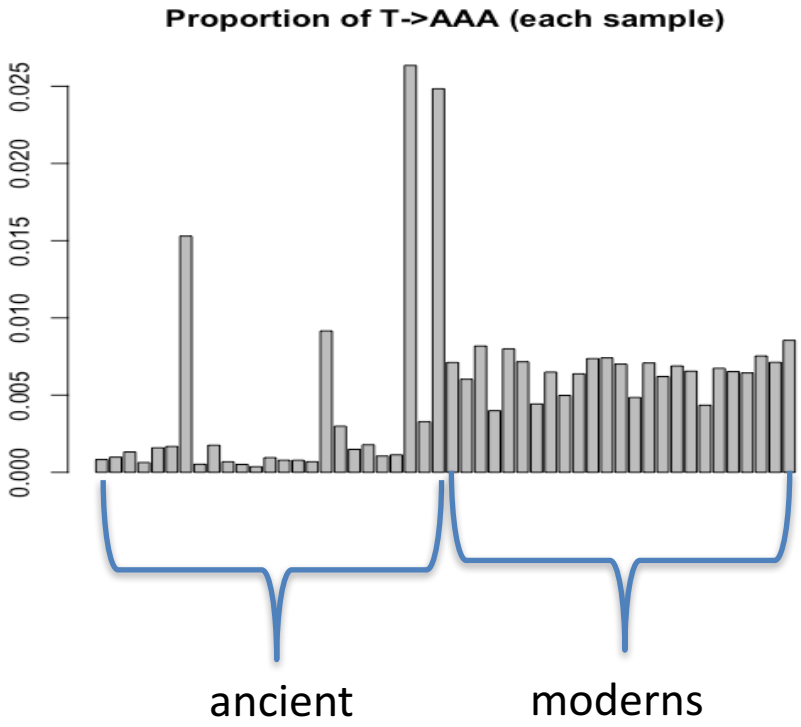


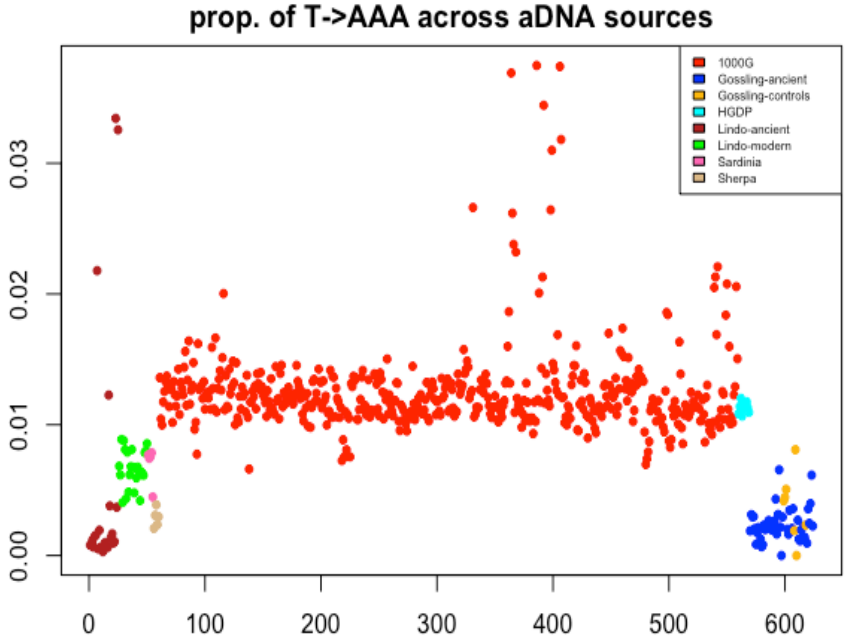


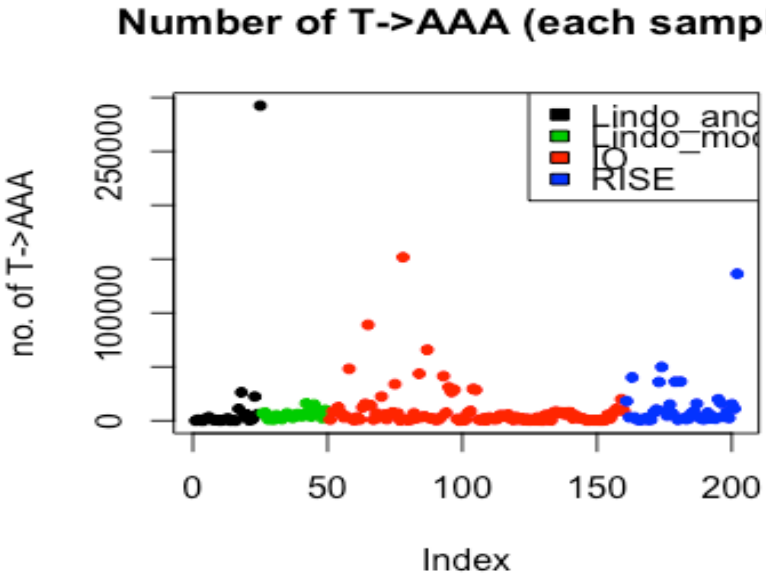
Cluster 3






Proportion of T->AAA (each sample)
0.000
0.005
0.010
0.015
0.020
0.025
ancient
moderns


prop. of T->AAA across aDNA sources
1000G
Gossling-ancient
Gossling-controls
HGDP
Lindo-ancient
Lindo-modern
Sardinia
Sherpa
0.00
0.01
0.02
0.03
0
100
200
300
400
500
600

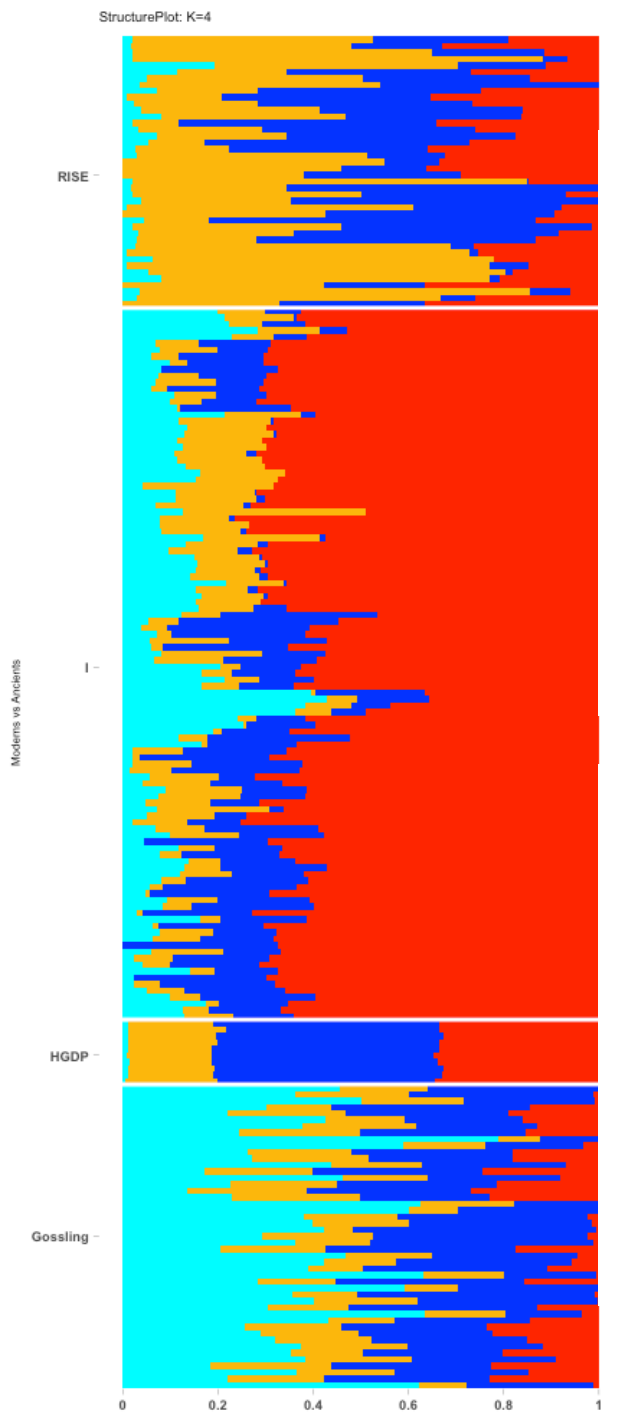

Number of T->AAA (each sample)
Lindo_anc
Lindo_mod
IO
RISE
no. of T->AAA
0
100000
250000
0
50
100
150
200
Index

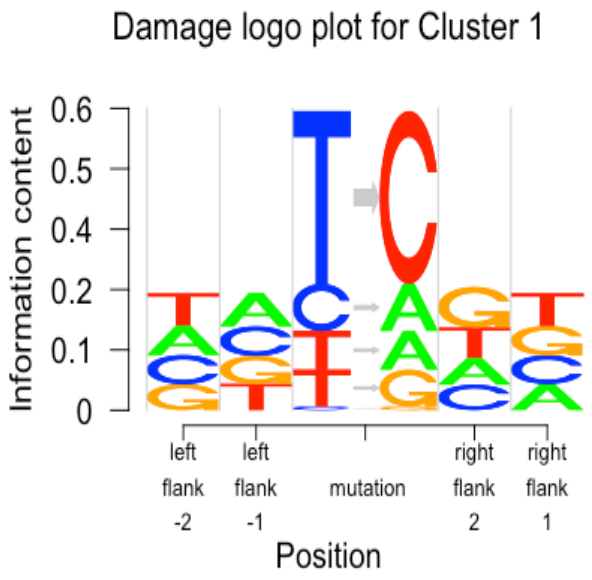
Applications of aRchaic on real data

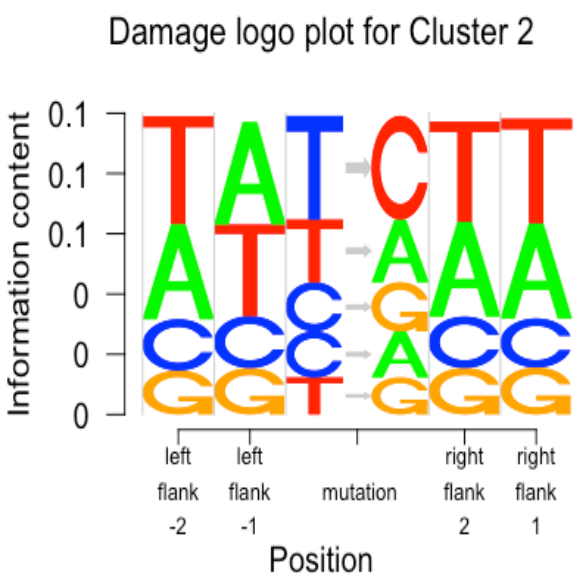
Allentoft + Haak et al published ancient data + Di Rienzo data + 10 HGDP modern samples

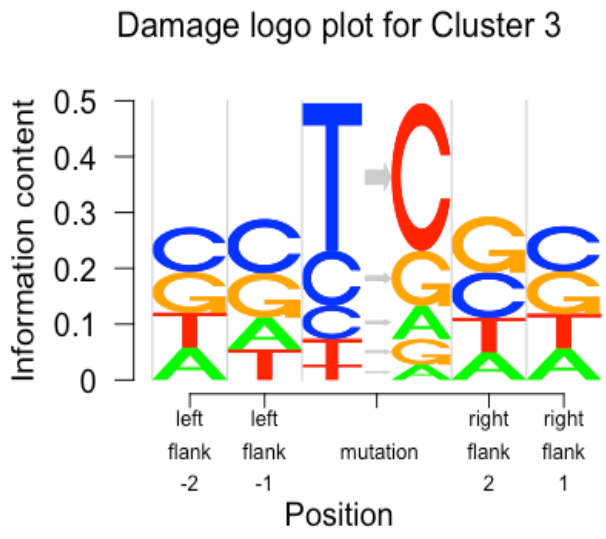
We show the lab specific effects on mutational patterns.

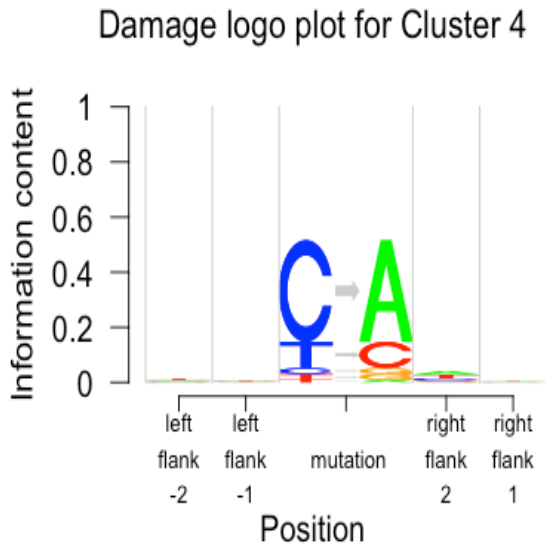
We remove the C-> T patterns and the read position information and ran the model on remaining mutational patterns.


StructurePlot: K=4
RISE
I
HGDP
Gossling
Moderns vs Ancients
0
0.2
0.4
0.6
0.8
1


Damage logo plot for Cluster 1
Information content
0
0.1
0.2
0.3
0.4
0.5
0.6
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position


Damage logo plot for Cluster 2
Information content
0
0
0
0.1
0.1
0.1
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position

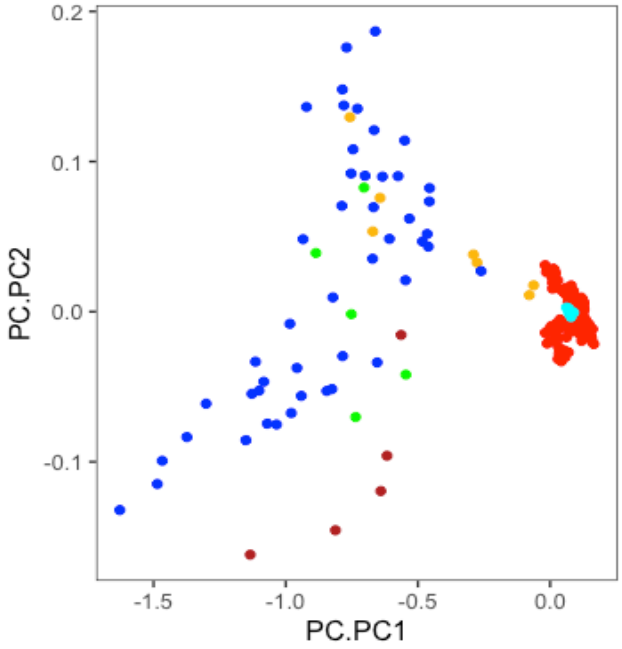

Damage logo plot for Cluster 3
Information content
0
0.1
0.2
0.3
0.4
0.5
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position


Damage logo plot for Cluster 4
Information content
0
0.2
0.4
0.6
0.8
1
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position

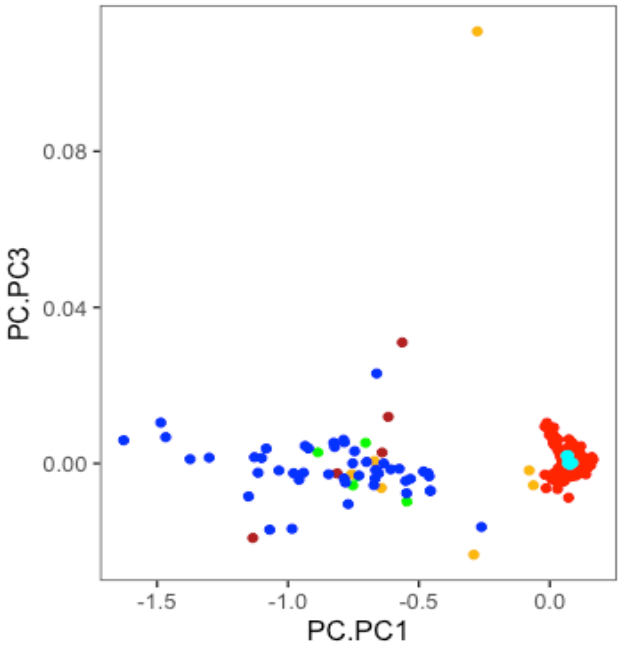
Applications of aRchaIc on real data

Sherpa + Gossling + HGDP + 100 randomly selected samples from 1000 Genomes + ancient Sardinian data

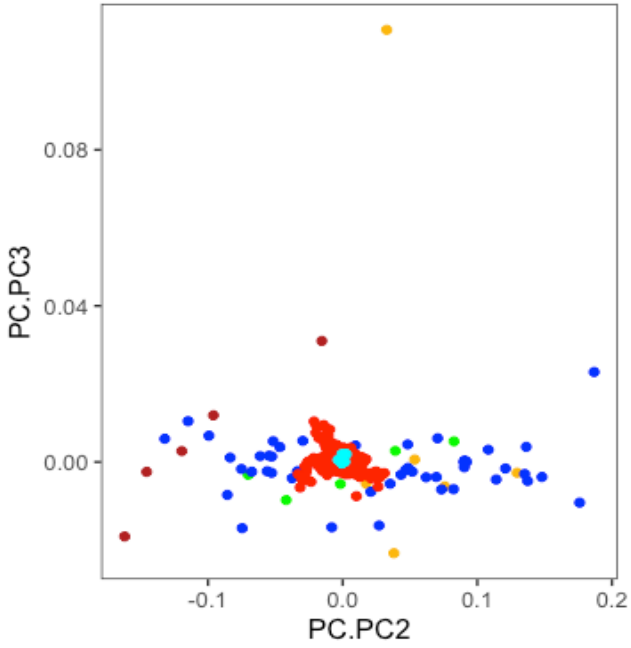
We focus on only known damage signatures – C->T along with the position of the mutation from end of the read (from position 0 to position 10)


0.2
0.1
0.0
-0.1
PC.PC2
-1.5
-1.0
-0.5
0.0
PC.PC1


factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
sardinia
sherpa

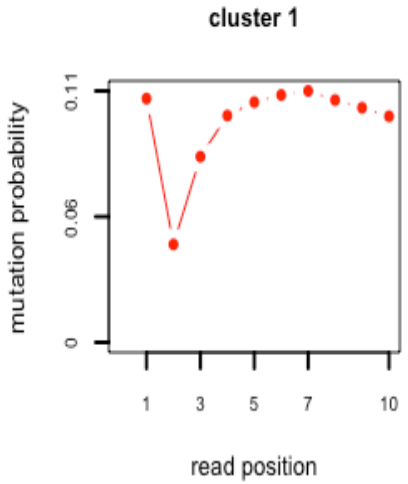

0.08
0.04
0.00
PC.PC3
-1.5
-1.0
-0.5
0.0
PC.PC1

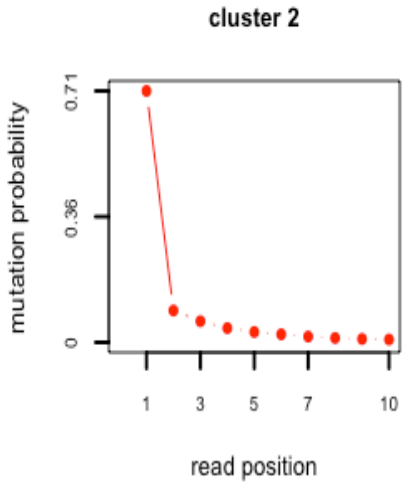

factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
sardinia
sherpa

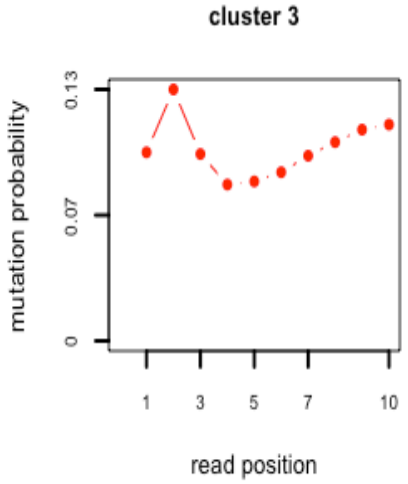

0.08
0.04
0.00
PC.PC3
-0.1
0.0
0.1
0.2
PC.PC2

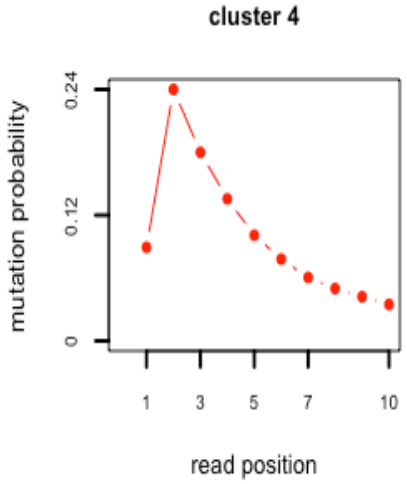

factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
sardinia
sherpa


sherpa
sardinia
hgdp
gosling-control
gosling-ancient
1000G
Tissue type
0
0.2
0.4
0.6
0.8
1
Cluster
1
2
3
4


cluster 1
mutation probability
0
0.06
0.11
1
3
5
7
10
read position


cluster 2
mutation probability
0
0.36
0.71
1
3
5
7
10
read position


cluster 3
mutation probability
0
0.07
0.13
1
3
5
7
10
read position


cluster 4
mutation probability
0
0.12
0.24
1
3
5
7
10
read position

Applications of aRchaIc

In classifying between moderns and ancient samples

Consensus ancient data panel : Sherpa + the Gossling ancient.

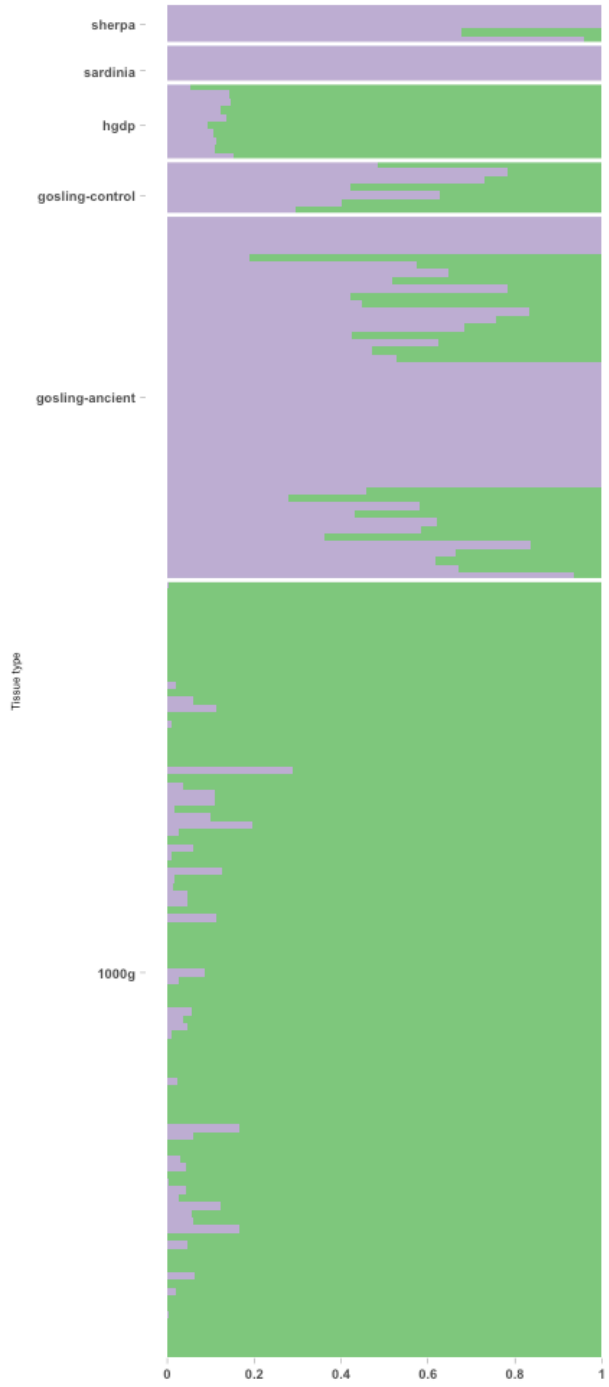
Consensus modern data panel: 1000 Genomes + HGDP moderns.

- We train the moderns and the ancient panels using the above data and test on the Sardinian ancients and the Gossling controls data.

- We apply Support Vector machines for classifying the test samples and generate a probability for each test sample to come from consensus moderns and equivalently, from consensus ancient panels.

- For Sardinian samples (total 5), 2 samples were classified as ancients and 3 as moderns.

- For Gossling controls data (total 7), 2 were classified stronglt as ancients (EXN3 and Libneg3) and the rest as moderns.

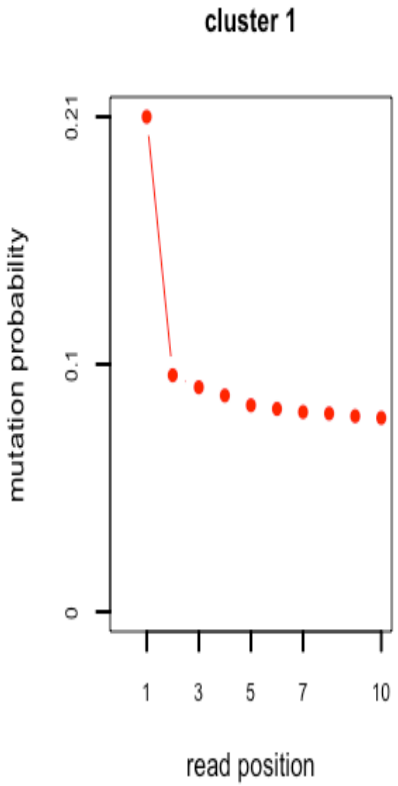


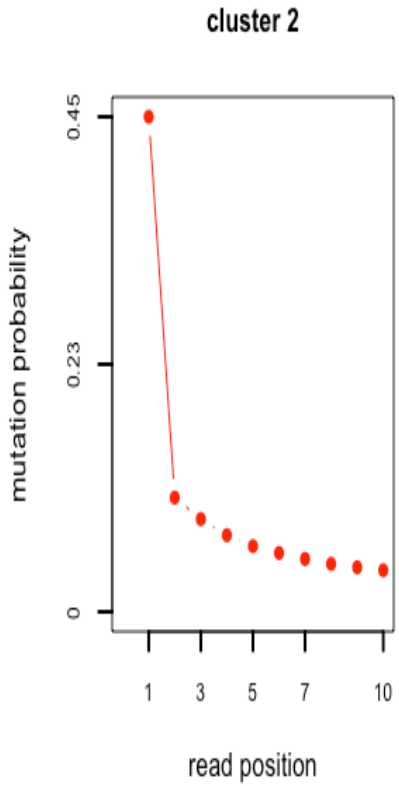
Train: ancient
Test

Train: moderns
Test

Train: ancient

Train: moderns





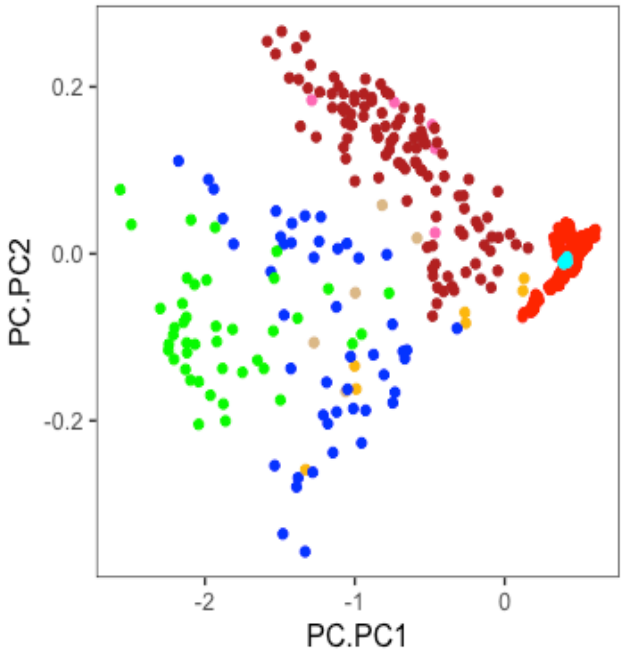
R package aRchaic

Developmental version :

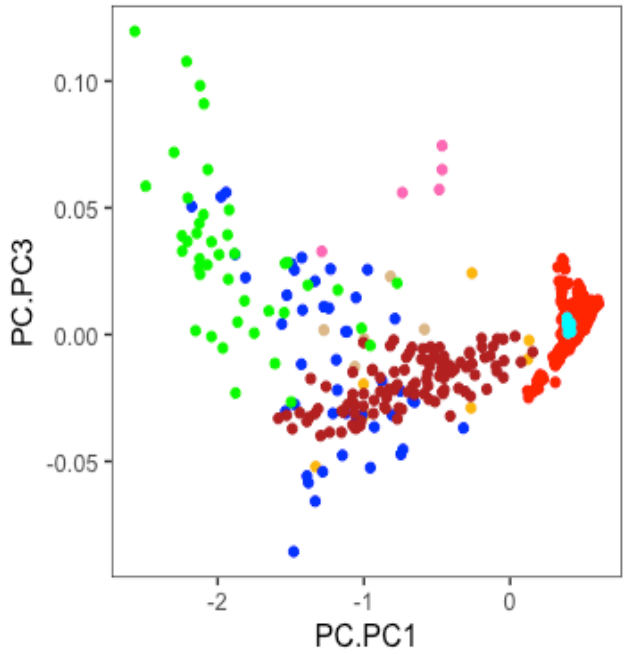
https://github.com/kkdey/aRchaic

The codes and the scripts for generating the plots and also additional plots for other values of K can be found at

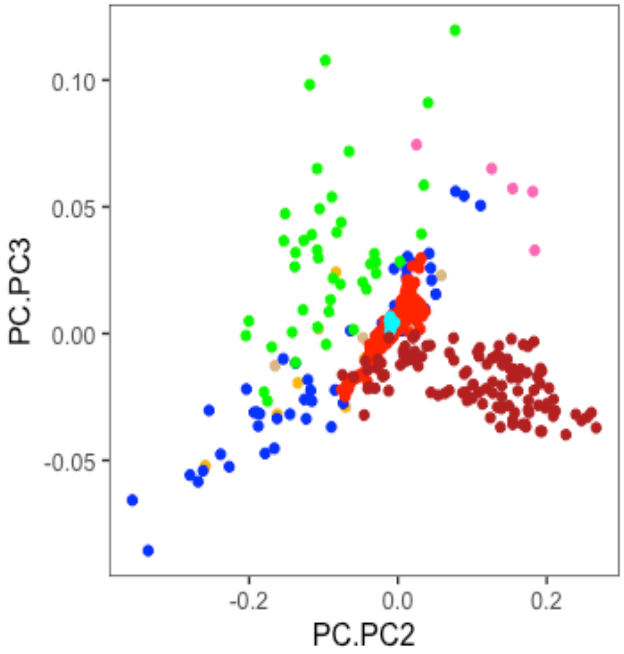
https://github.com/halasadi/ancient-damage


0.2
0.0
-0.2
PC.PC2
-2
-1
0
PC.PC1


factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
I
RISE
sardinia
sherpa


0.10
0.05
0.00
-0.05
PC.PC3
-2
-1
0
PC.PC1


factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
I
RISE
sardinia
sherpa


0.10
0.05
0.00
-0.05
PC.PC3
-0.2
0.0
0.2
PC.PC2


factor(labels)
1000G
gosling-ancient
gosling-control
hgdp
I
RISE
sardinia
sherpa