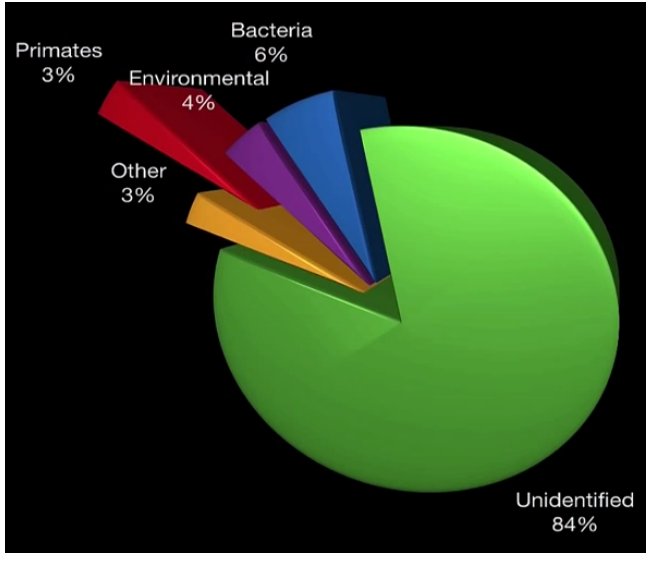
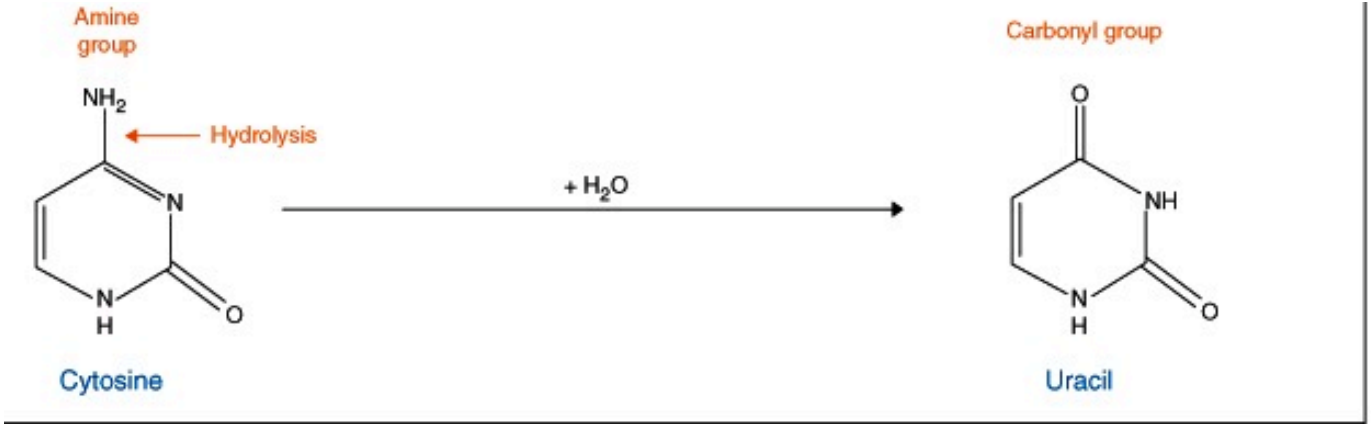
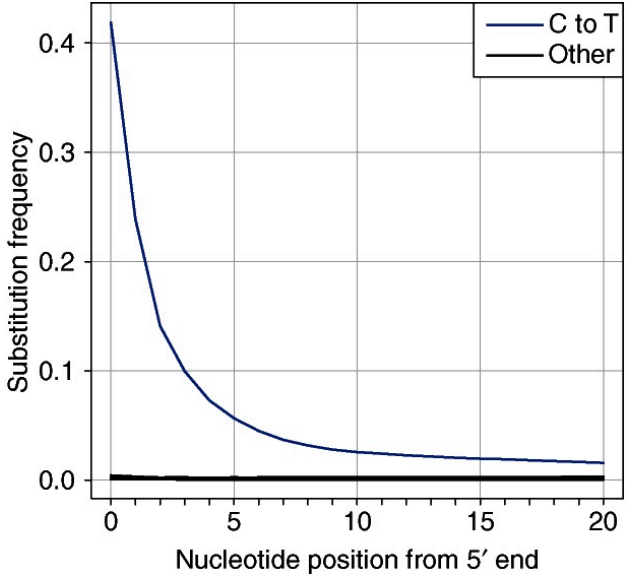
An R package for clustering and visualizing ancient dna signatures

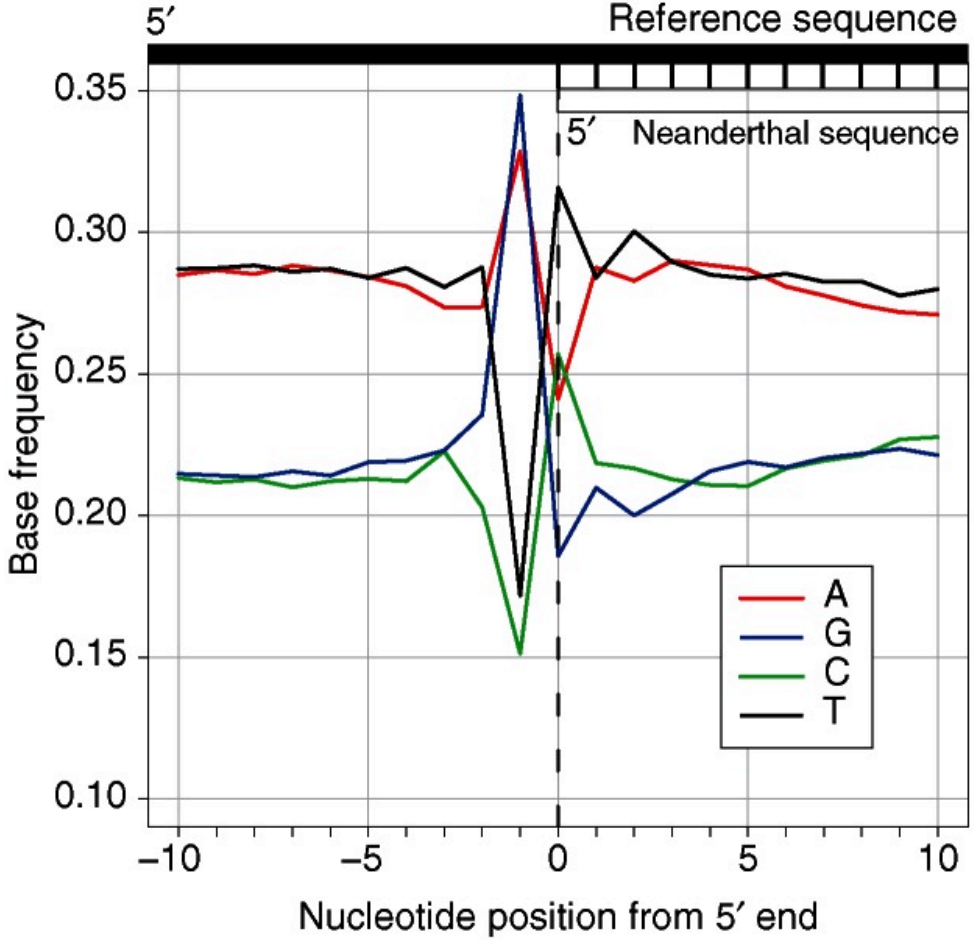
By Hussein Al-Asadi & Kushal K Dey



- DNA degradation and damage
- Bacterial DNA
- Human contamination

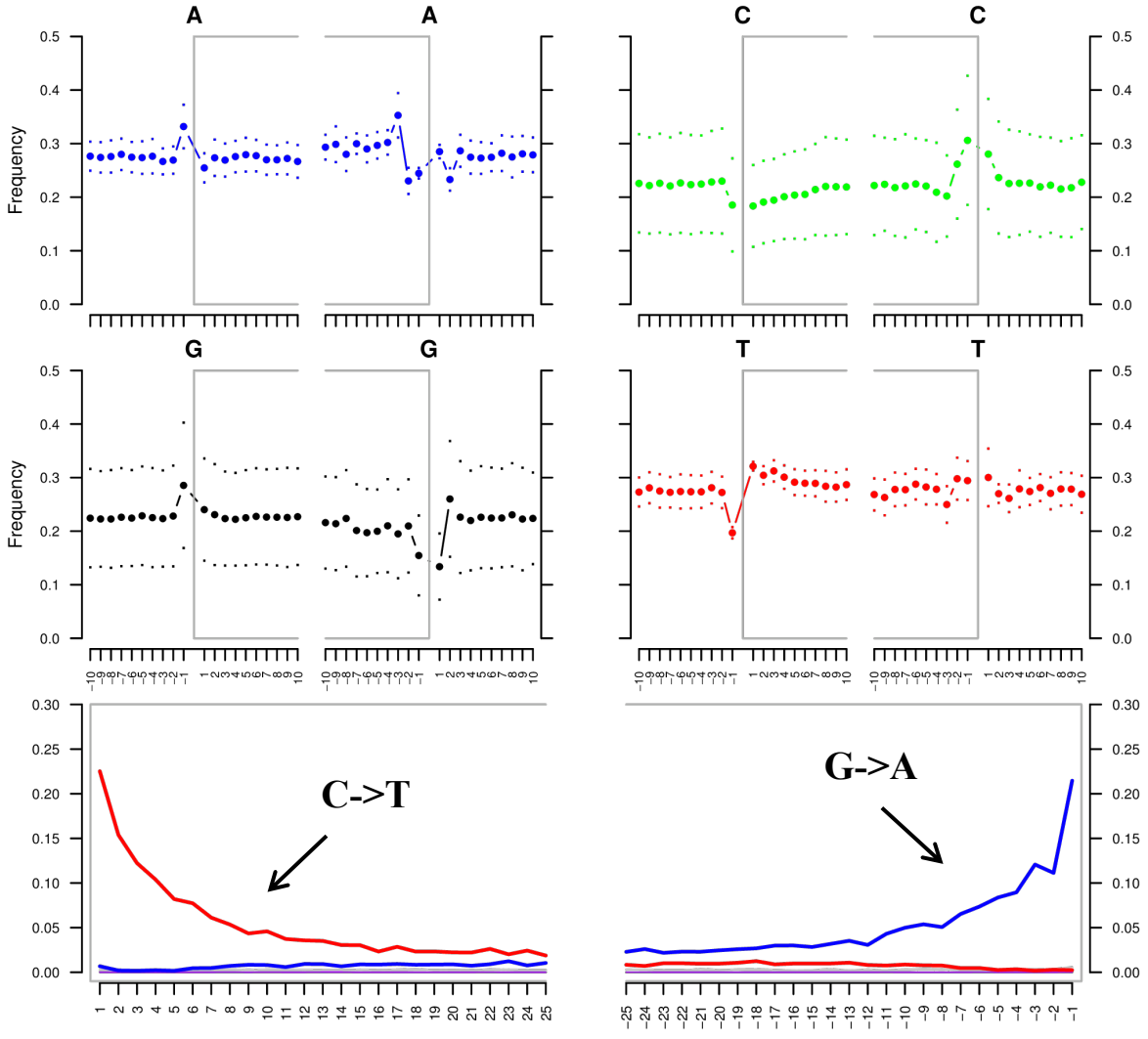

Amine group
Carbonyl group
NH2
Hydrolysis
+ H2O
N
N
H
O
NH
O
Cytosine
Uracil


C to T
Other
Substitution frequency
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
Nucleotide position from 5′ end

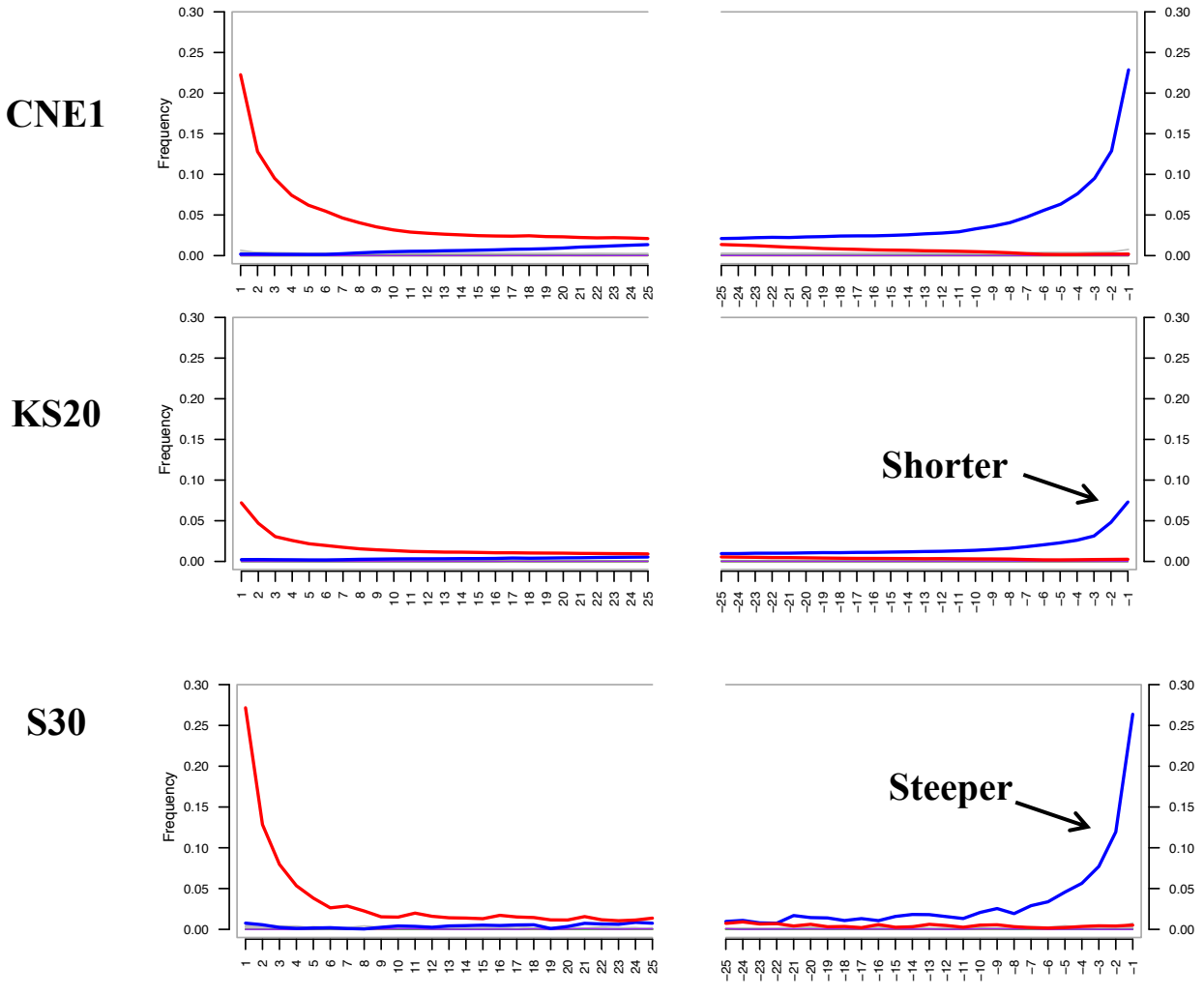

5′
Reference sequence
5′
Neanderthal sequence
0.35
0.30
0.25
0.20
0.15
0.10
Base frequency
A
G
C
T
−10
−5
0
5
10
Nucleotide position from 5′ end

mapDamage

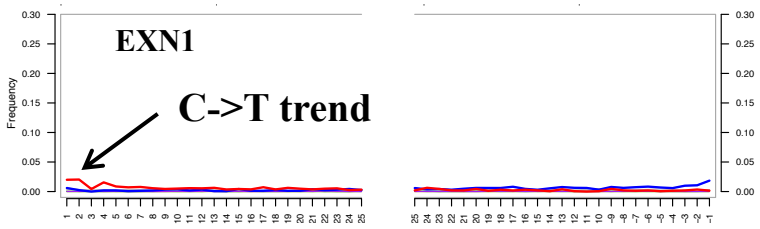
mapDamage: tracking and quantifying damage patterns in ancient DNA sequences

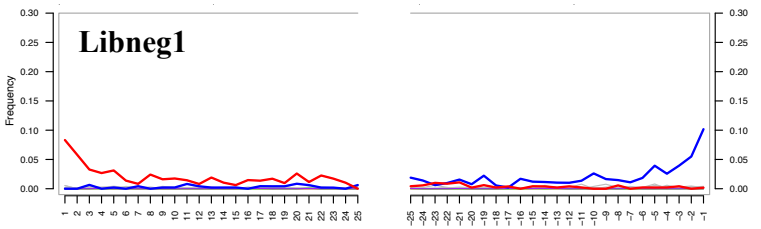


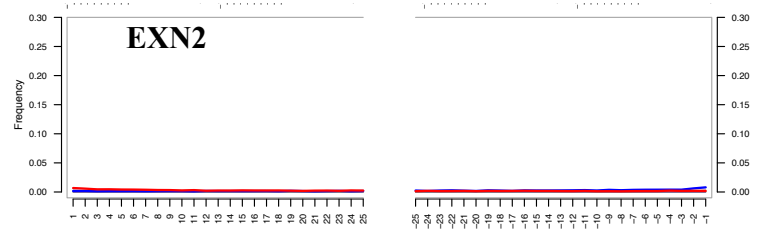
MapDamage on Ancients

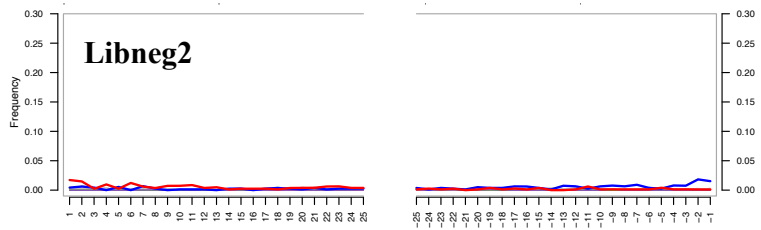


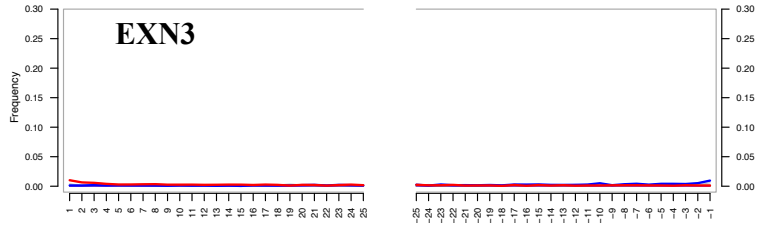
Control

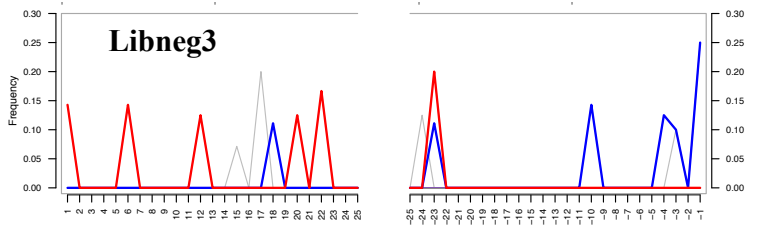

EXN1
C->T trend
Frequency

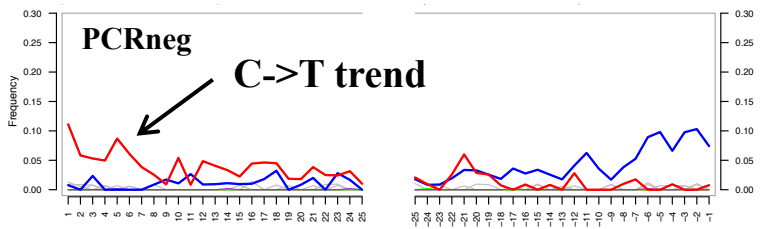

Libneg1
Frequency


EXN2
Frequency


Libneg2
Frequency

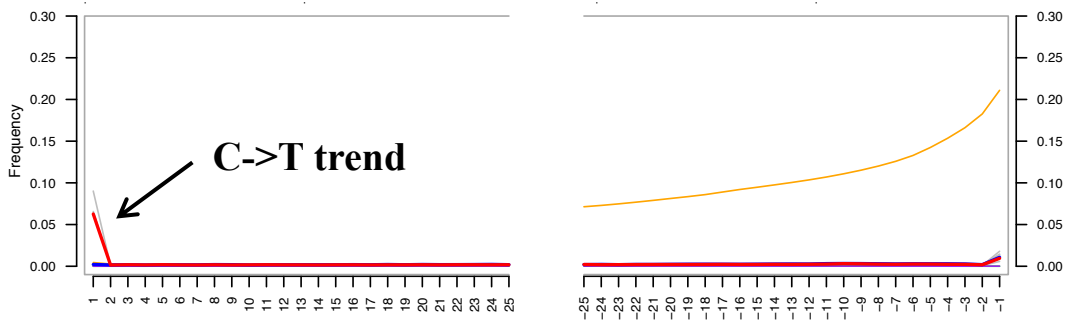

EXN3
Frequency


Libneg3
Frequency

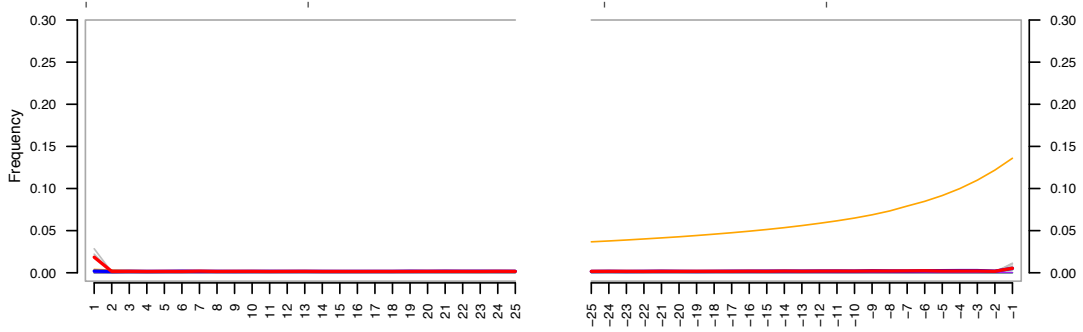

PCRneg
C->T trend
Frequency

MapDamage on Moderns

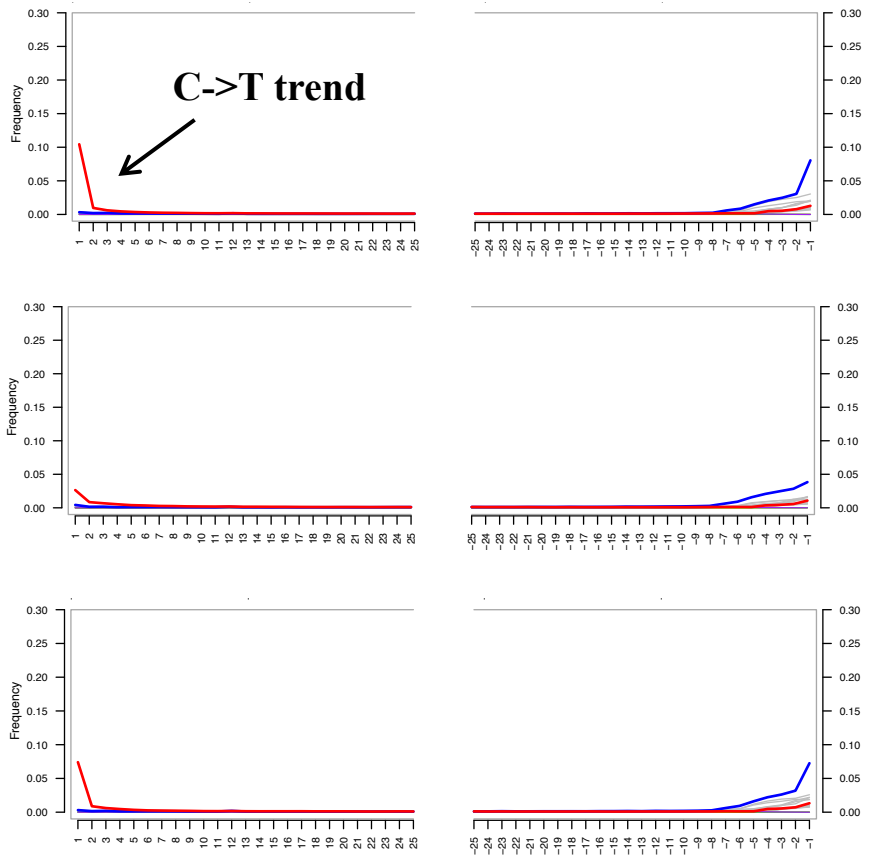
HG00097



HG00099



MapDamage on Ancient Sards. (different data-set)



Thoughts…

- It’s not clear what is a C->T damage profile is and what is a damage profile typical of a modern individual is.

- Difficult to perform joint 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 “mismatch patterns” (i.e. features) from BAM files, which are flexible

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

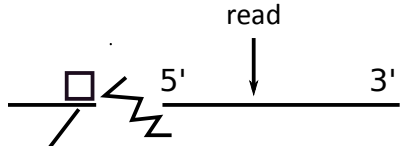
Step 1: Get mismatch patterns from BAM files

- Mutational pattern:
  - mutation
  - flanking bases (we use 1)
  - position from the end of the read
  - strand orientation
  - base pair immediately outside 5’ strand break.

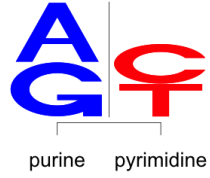
- Example output for Sample_1
  - A(T->C)A, 0, 50, +, A, 2
  - A(T->A)A, 10, 40, - , T, 1
  - C(C->T)T, 2, 48, +, G, 6

C → T	T → C	T → G	C → A	C → G	T → A
0.92	0.054	0.007	0.007	0.007	0.005

strand-orientation	Percentage
+	0.5
-	0.5

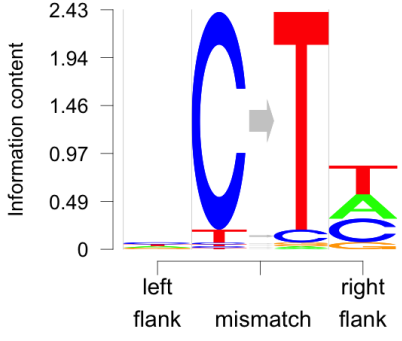


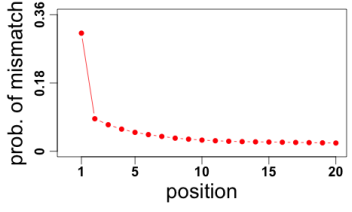
strand-break composition



Base	Percentage
A	0.36
C	0.18
G	0.32
T	0.14

strand orientation





Base	Percentage
A	0.24
C	0.28
G	0.23
T	0.25

Base	Percentage
A	0.29
C	0.28
G	0.08
T	0.35

Probability of mismatch across the read

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 a "damaged" ancestry and "un-damaged" ancestry where ancestries are defined by a probability vector on mutational patterns

Just Notation

- i = {1,….,I} I individuals (bam files)
- Each individual has J_i mismatch patterns.
- Let M denote the number of features, here M = 6.
- x_i,j,l where j={1,…,J_i}, l = {1,…M}
  - e.g. x_i,j,1 = A if 1 is the mutational feature corresponding to the left flanking base.

Generative Model

(1) Generate $z_{i,j} \sim Multinomial(1, q_i)$ where $z_{i,j} \in 1, ..., K$ and $i = 1, ..., I$ and $j = 1, ..., J_i$

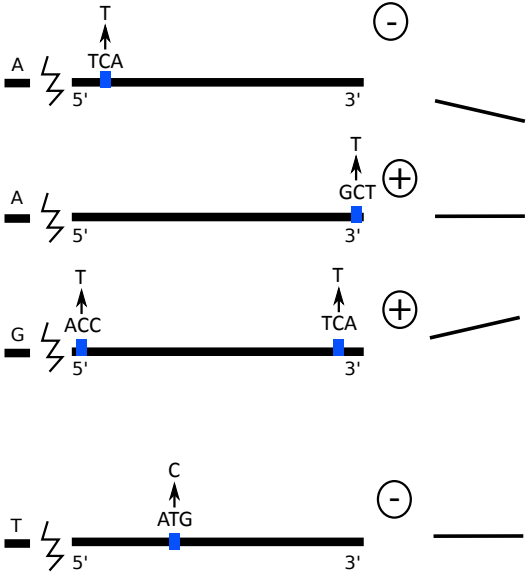
(2) For each $l = 1, ..., M$, generate $x_{i,j,l} \sim Multinomial(1, f_{z_{i,j},l})$

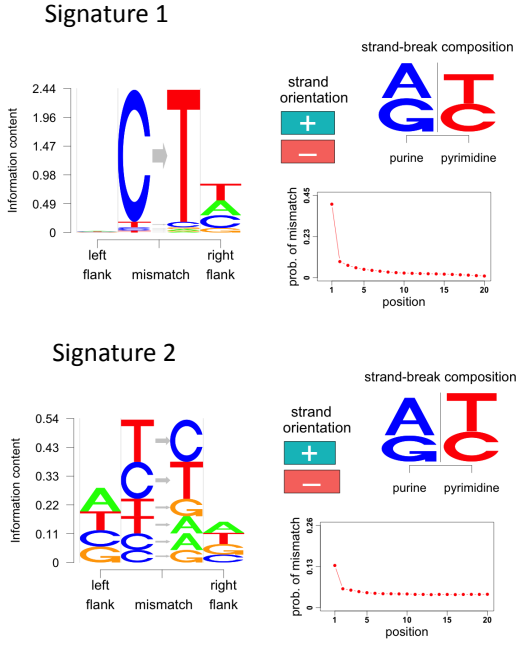
Graphical overview of aRchaic

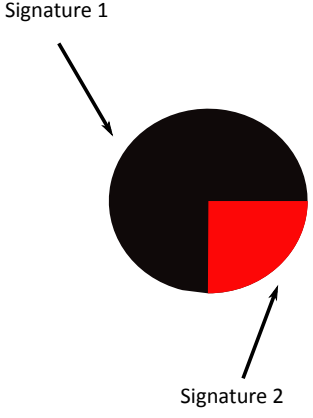
observed mismatch patterns

operative mismatch signatures

individuals are a mixture of signatures







Applications of aRchaIc

4 case studies

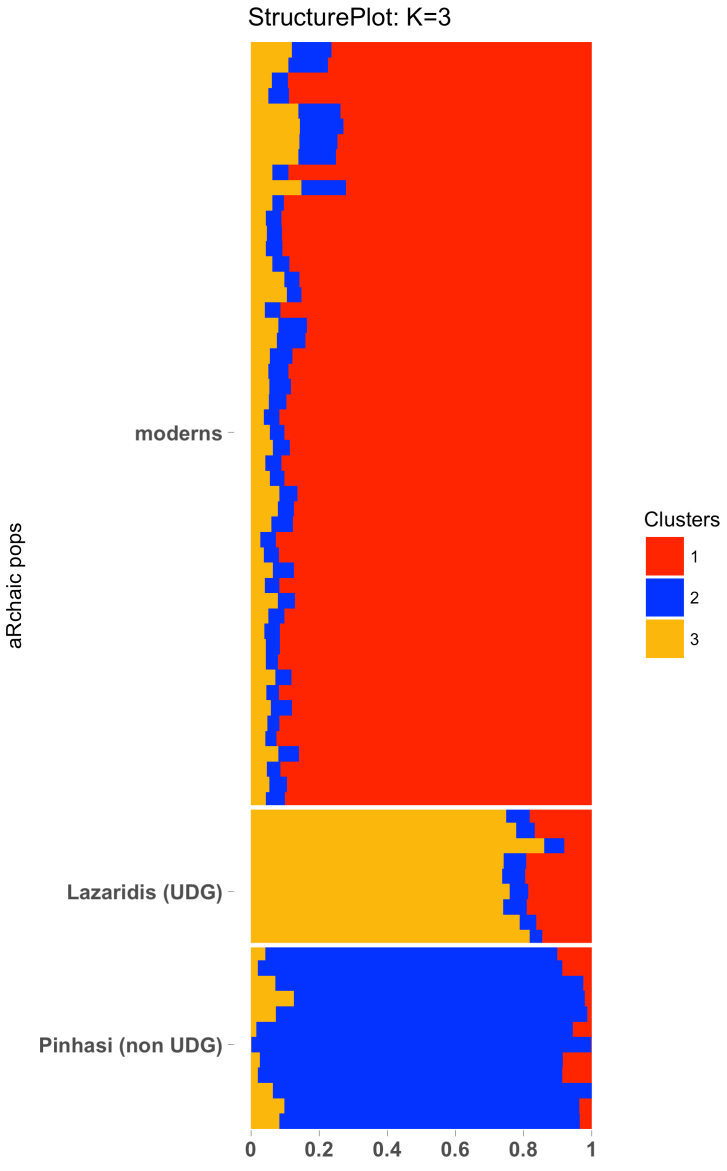
- Comparison of UDG treated aDNA (*Lazaridis+2014*), non UDG treated aDNA (Pinhasi et al) and moderns (1000G)

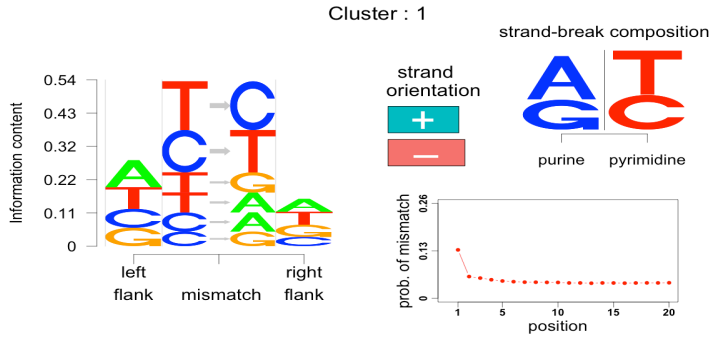
- Large scale aDNA due to *Mathieson+2015*

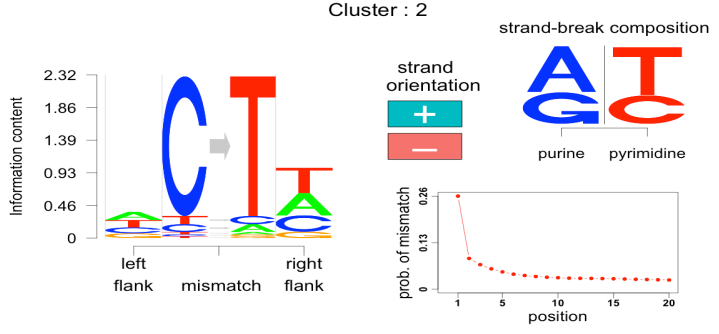
- *Lindo+2016* moderns and ancient DNA analysis

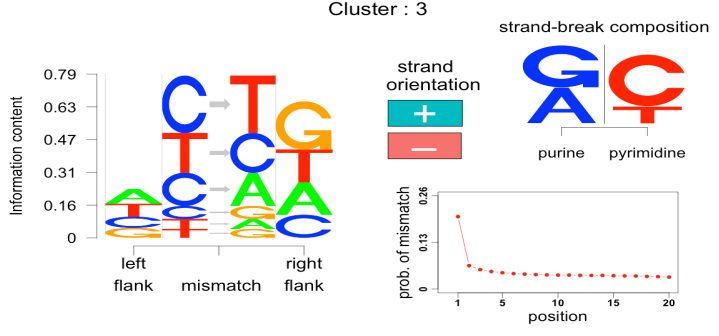
- Contamination study using simulations and *Mathieson+2015* data

Case Study 1 : 1000G moderns, Lazaridis and Pinhasi

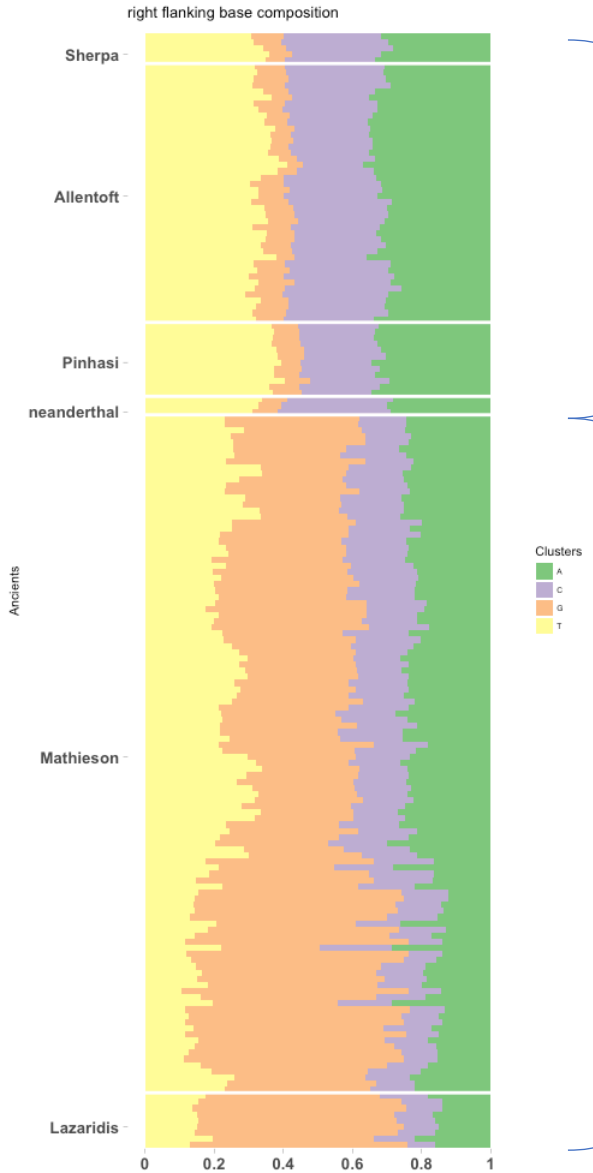

StructurePlot: K=3
aRchaic pops
moderns
Lazaridis (UDG)
Pinhasi (non UDG)
0
0.2
0.4
0.6
0.8
1
Clusters
1
2
3


Cluster : 1
strand-break composition
strand orientation
+
—
purine
pyrimidine
Information content
0.54
0.43
0.32
0.22
0.11
0
left flank
mismatch
right flank
prob. of mismatch
0.26
0.13
0
1
5
10
15
20
position


Cluster : 2
strand-break composition
strand orientation
+
—
purine
pyrimidine
Information content
2.32
1.86
1.39
0.93
0.46
0
left flank
mismatch
right flank
prob. of mismatch
0.26
0.13
0
1
5
10
15
20
position

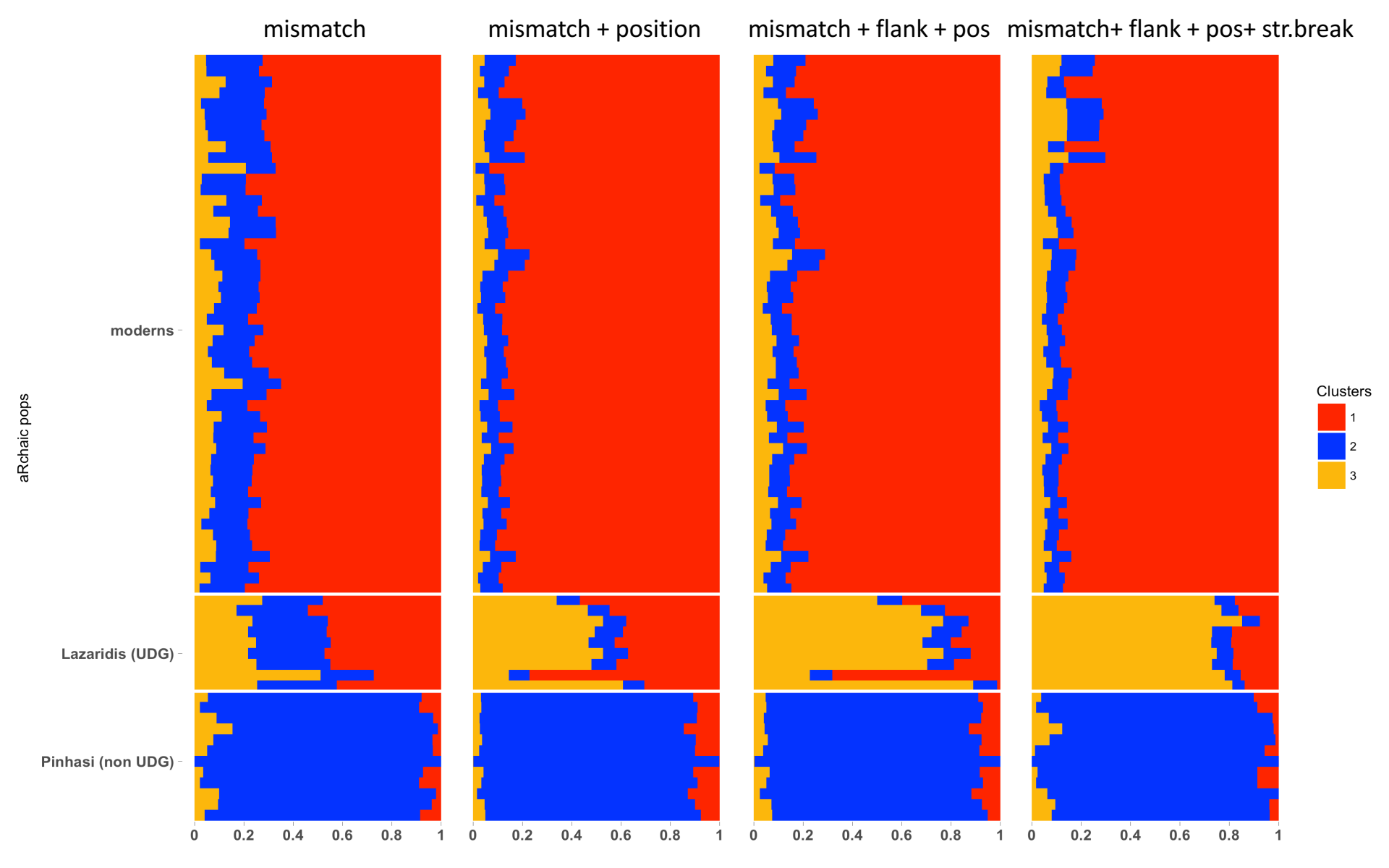

Cluster : 3
strand-break composition
strand orientation
+
—
purine
pyrimidine
Information content
0.79
0.63
0.47
0.31
0.16
0
left flank
mismatch
right flank
prob. of mismatch
0.26
0.13
0
1
5
10
15
20
position

We look at the right flanking base composition for all mutations of type C->T occurring inside 10 bases from start of the read on the “+” strand.

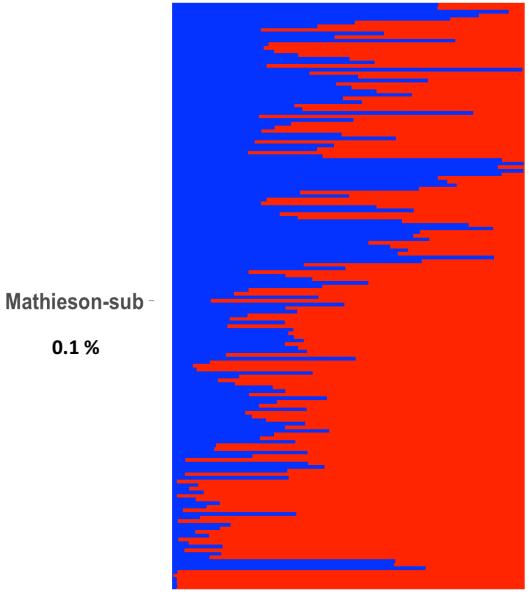


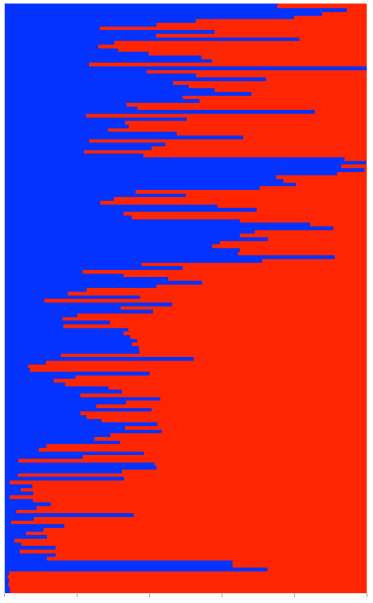
Non UDG treated samples (sherpa, Allentoft, Pinhasi, neanderthal)

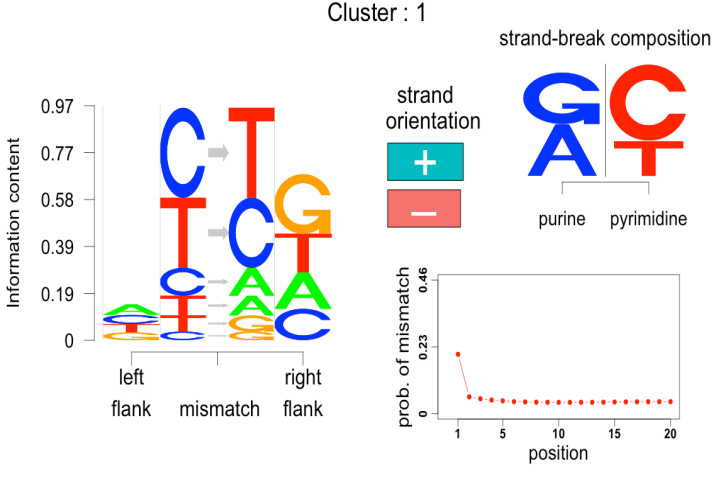
UDG treated samples (Mathieson, Lazaridis)

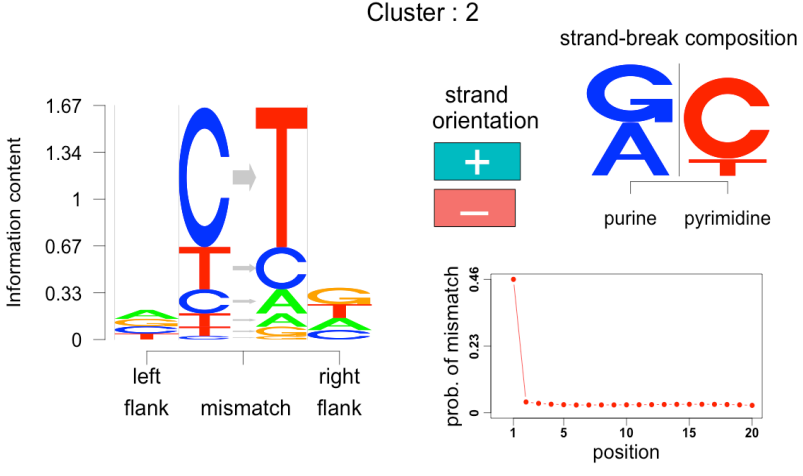

mismatch
mismatch + position
mismatch + flank + pos
mismatch+ flank + pos+ str.break
aRchaic pops
moderns
Lazaridis (UDG)
Pinhasi (non UDG)
0
0.2
0.4
0.6
0.8
1
Clusters
1
2
3

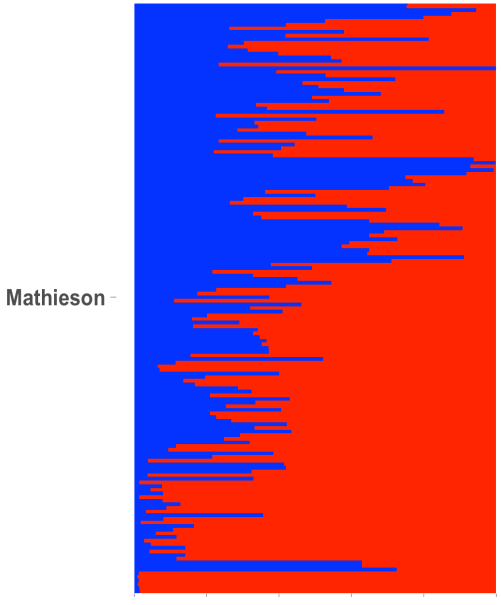
Case Study 2 : Mathieson data

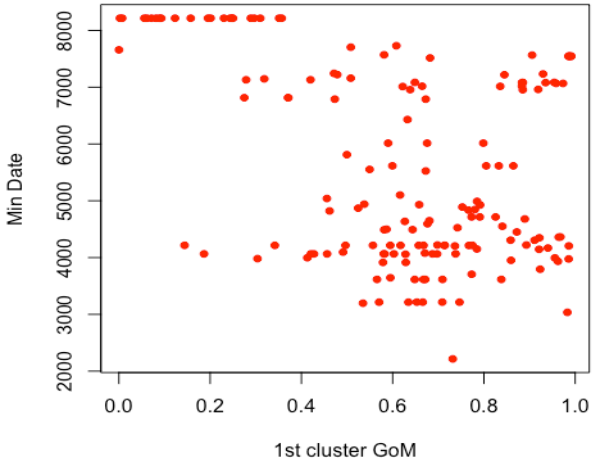

Mathieson-sub
0.1 %


Mathieson

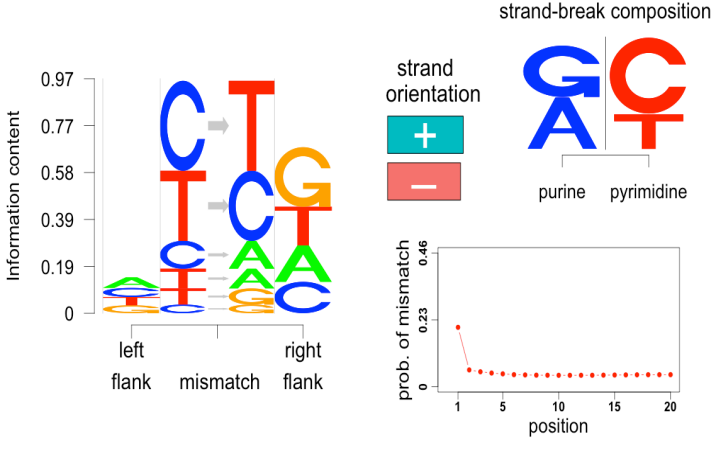

Cluster : 1
strand-break composition
strand orientation
+
−
purine
pyrimidine
Information content
0.97
0.77
0.58
0.39
0.19
0
left flank
mismatch
right flank
prob. of mismatch
0.46
0.23
0
1
5
10
15
20
position


Cluster : 2
strand-break composition
strand orientation
+
−
purine
pyrimidine
Information content
1.67
1.34
1
0.67
0.33
0
left flank
mismatch
right flank
prob. of mismatch
0.46
0.23
0
1
5
10
15
20
position

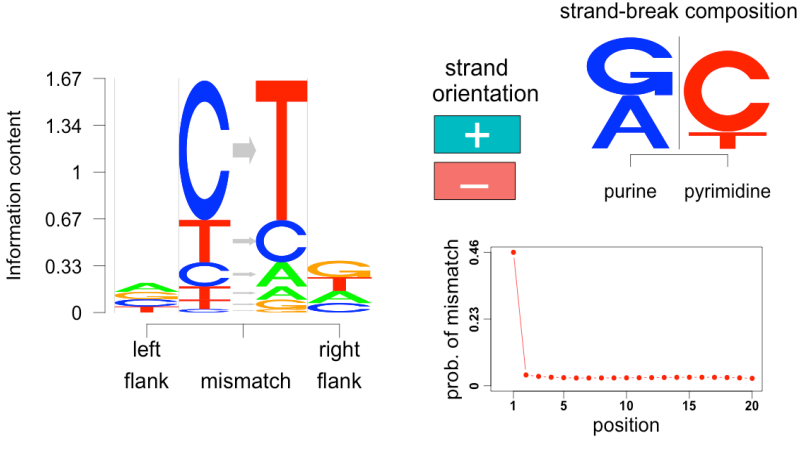




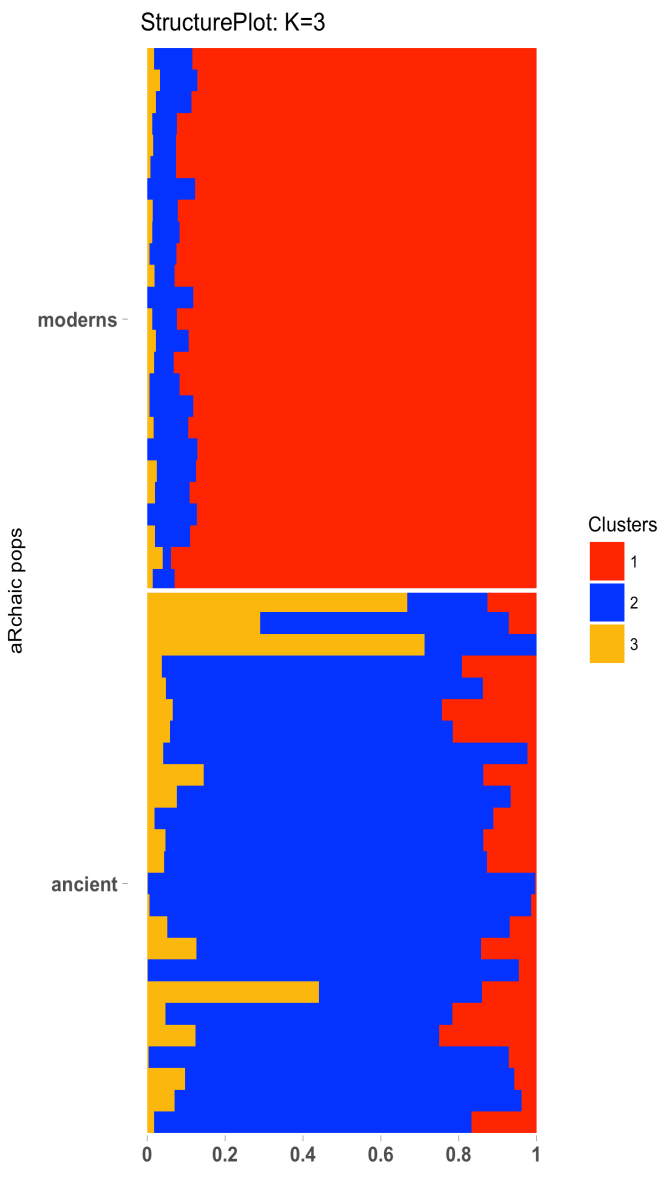
Cluster : 1

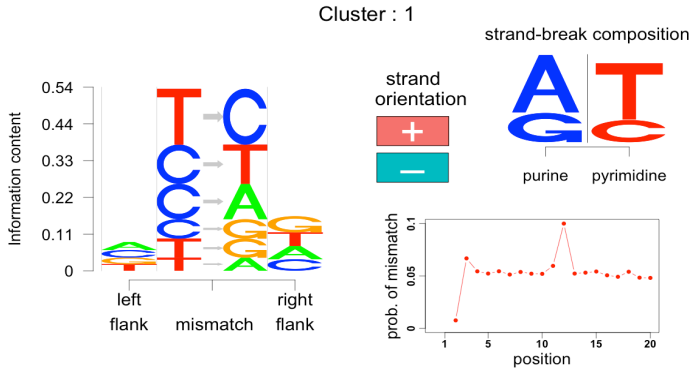


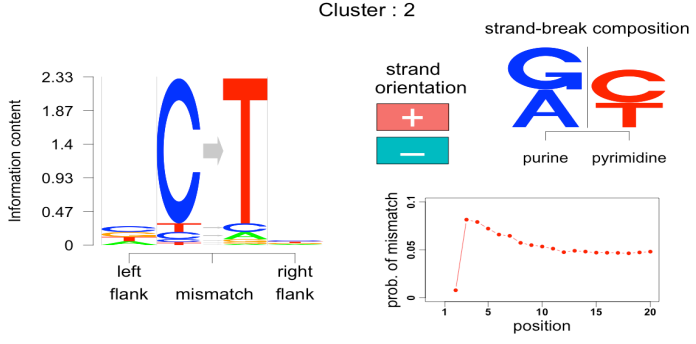
Cluster : 2

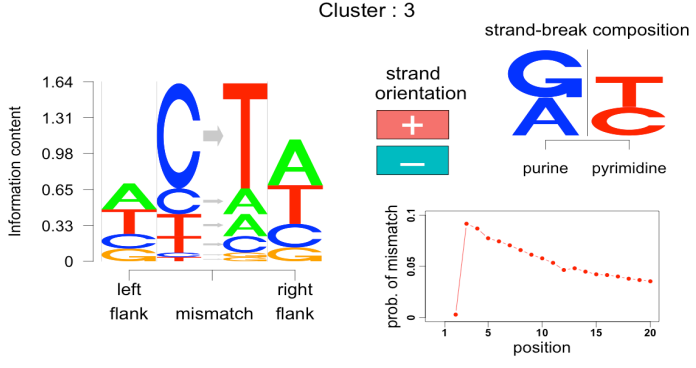


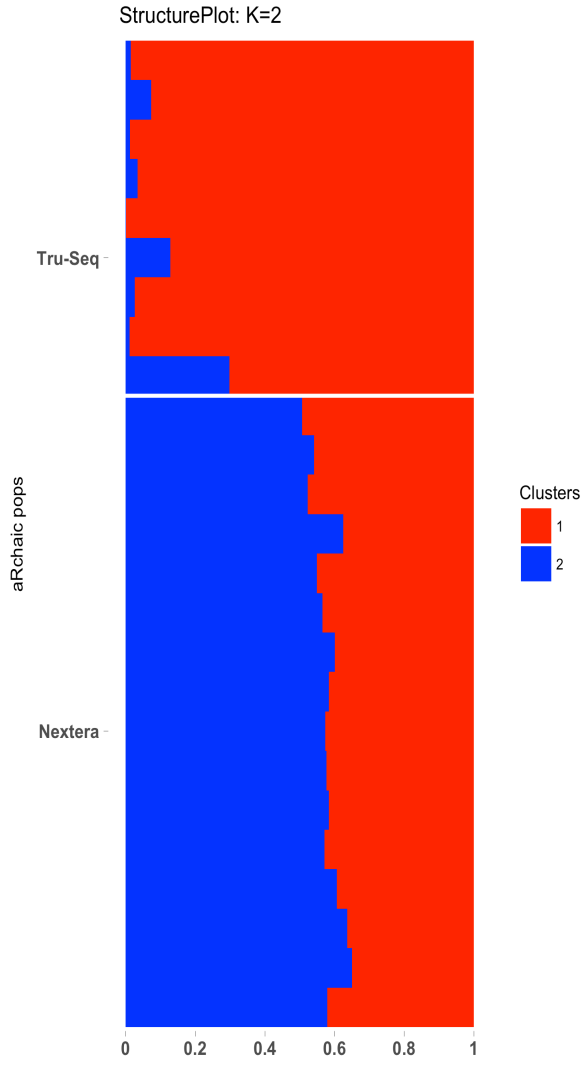
Case Study 3 : Lindo et al data

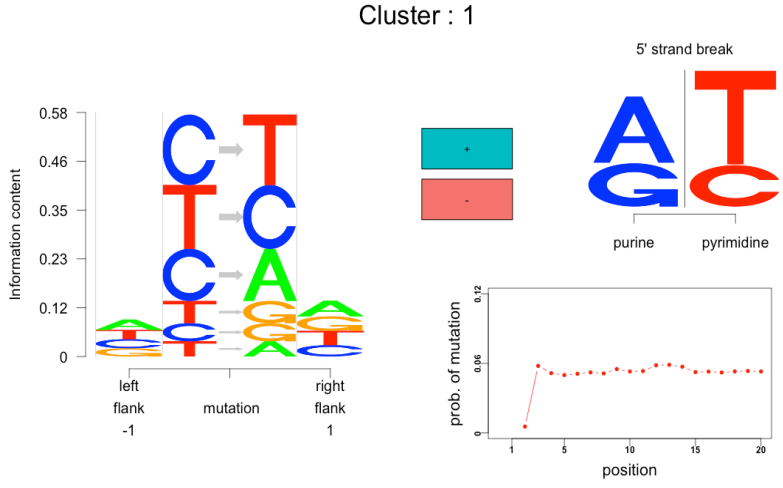

StructurePlot: K=3
moderns
ancient
aRchaic pops
0
0.2
0.4
0.6
0.8
1
Clusters
1
2
3

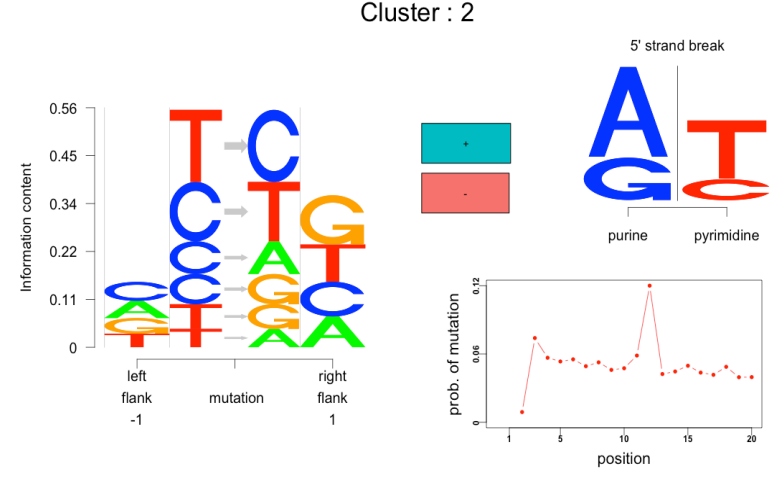

Cluster : 1
Information content
0.54
0.44
0.33
0.22
0.11
0
left flank
mismatch
right flank
strand orientation
+
−
strand-break composition
purine
pyrimidine
prob. of mismatch
0.1
0.05
0
1
5
10
15
20
position

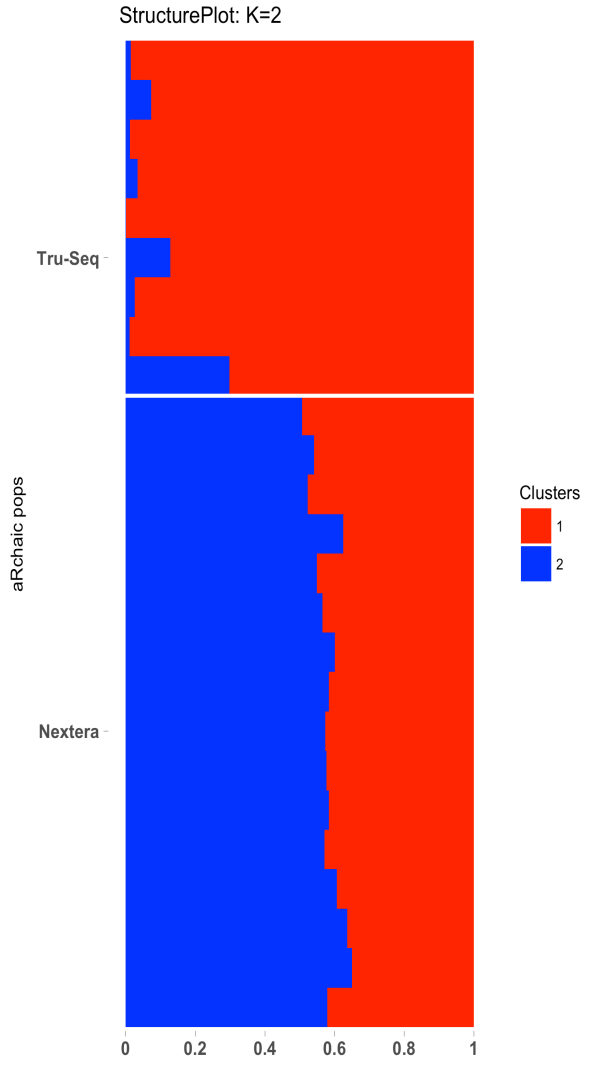

Cluster : 2
Information content
2.33
1.87
1.4
0.93
0.47
0
left flank
mismatch
right flank
strand orientation
+
−
strand-break composition
purine
pyrimidine
prob. of mismatch
0.1
0.05
0
1
5
10
15
20
position

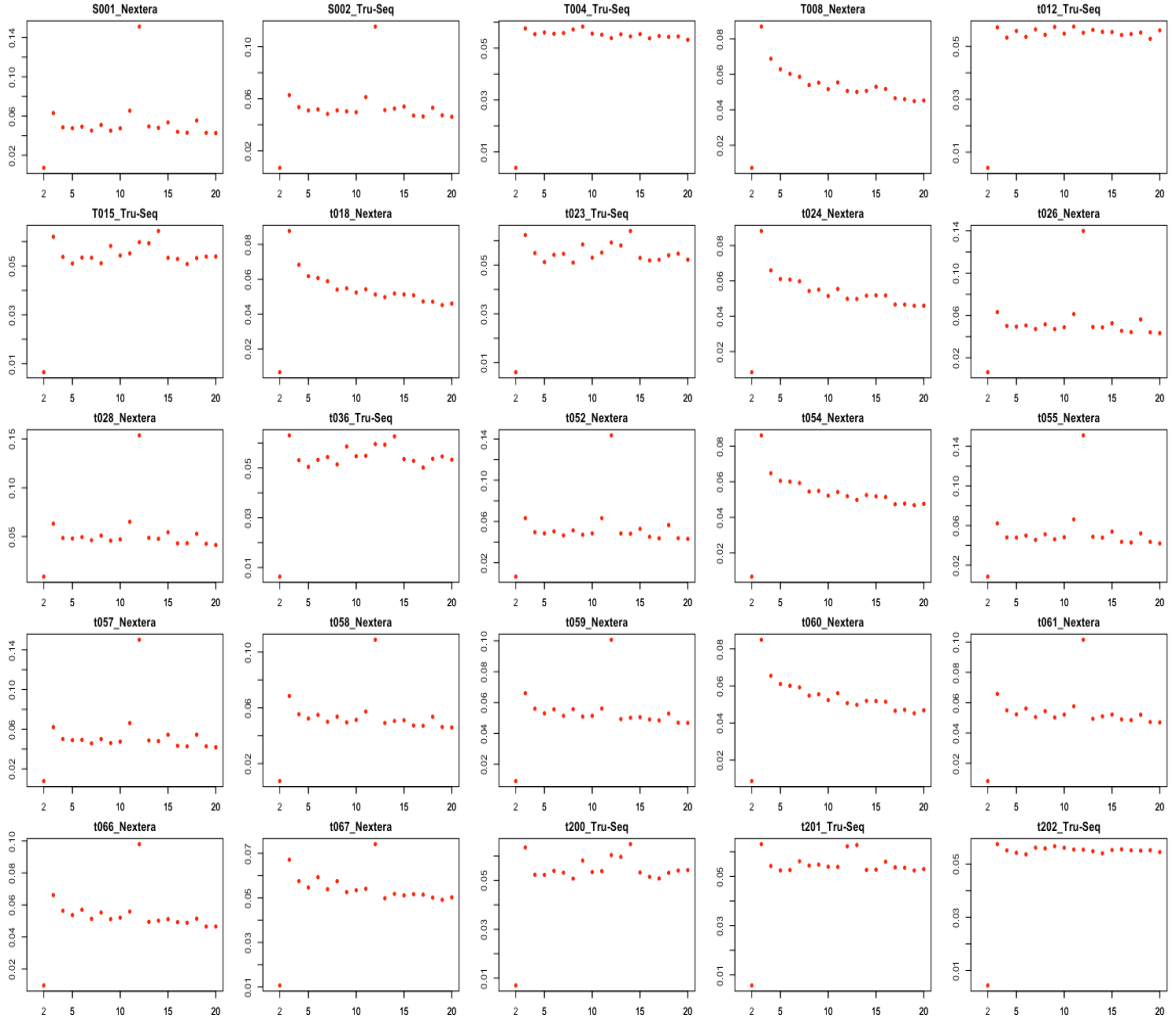

Cluster : 3
Information content
1.64
1.31
0.98
0.65
0.33
0
left flank
mismatch
right flank
strand orientation
+
−
strand-break composition
purine
pyrimidine
prob. of mismatch
0.1
0.05
0
1
5
10
15
20
position


StructurePlot: K=2
Tru-Seq
Nextera
aRchaic pops
Clusters
1
2
0
0.2
0.4
0.6
0.8
1


Cluster : 1
5' strand break
Information content
0.58
0.46
0.35
0.23
0.12
0
left flank -1
mutation
right flank 1
+
-
purine
pyrimidine
prob. of mutation
0.12
0.06
0
1
5
10
15
20
position


Cluster : 2
5' strand break
Information content
0.56
0.45
0.34
0.22
0.11
0
left flank -1
mutation
right flank 1
+
-
purine
pyrimidine
prob. of mutation
0.12
0.06
0
1
5
10
15
20
position


StructurePlot: K=2
Tru-Seq
Nextera
aRchaic pops
Clusters
1
2
0
0.2
0.4
0.6
0.8
1

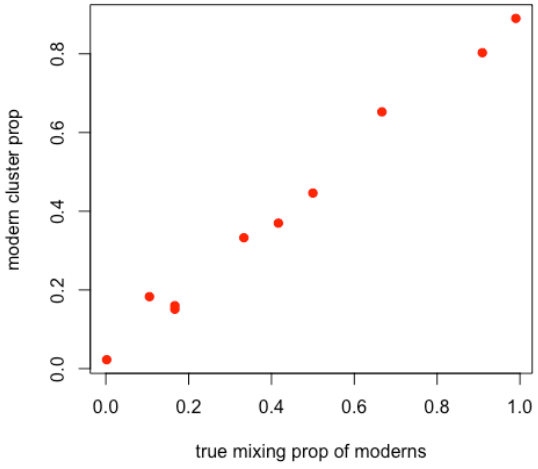

S001_Nextera
S002_Tru-Seq
T004_Tru-Seq
T008_Nextera
t012_Tru-Seq
T015_Tru-Seq
t018_Nextera
t023_Tru-Seq
t024_Nextera
t026_Nextera
t028_Nextera
t036_Tru-Seq
t052_Nextera
t054_Nextera
t055_Nextera
t057_Nextera
t058_Nextera
t059_Nextera
t060_Nextera
t061_Nextera
t066_Nextera
t067_Nextera
t200_Tru-Seq
t201_Tru-Seq
t202_Tru-Seq

Case Study 4 : Contamination

Pool all mismatch patterns from a modern sample and all mismatch patterns from an ancient sample and mix them in a certain ratio (normalized also by total number of mutations).

Then run **aRchaic** on these mixed samples with other modern and ancient samples and see whether the grades of membership match with the mixing proportion.

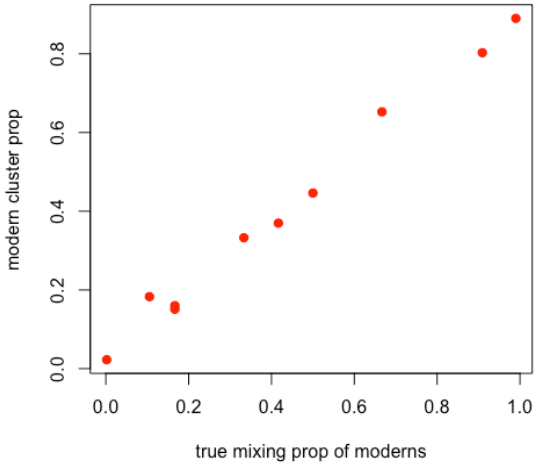
Pinhasi aDNA + moderns + their mixed samples



Pool all mismatch patterns from a modern sample and all mismatch patterns from an ancient sample and mix them in a certain ratio (normalized also by total number of mutations).

Then run **aRchaic** on these mixed samples with other modern and ancient samples and see whether the grades of membership match with the mixing proportion.

Pinhasi aDNA + moderns + their mixed samples

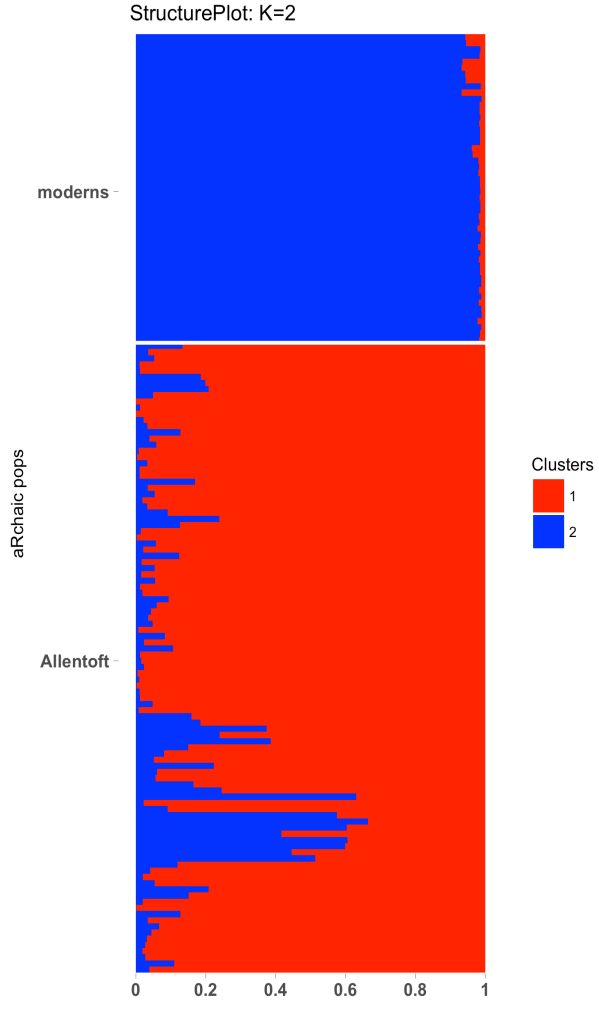


Issues with contamination:

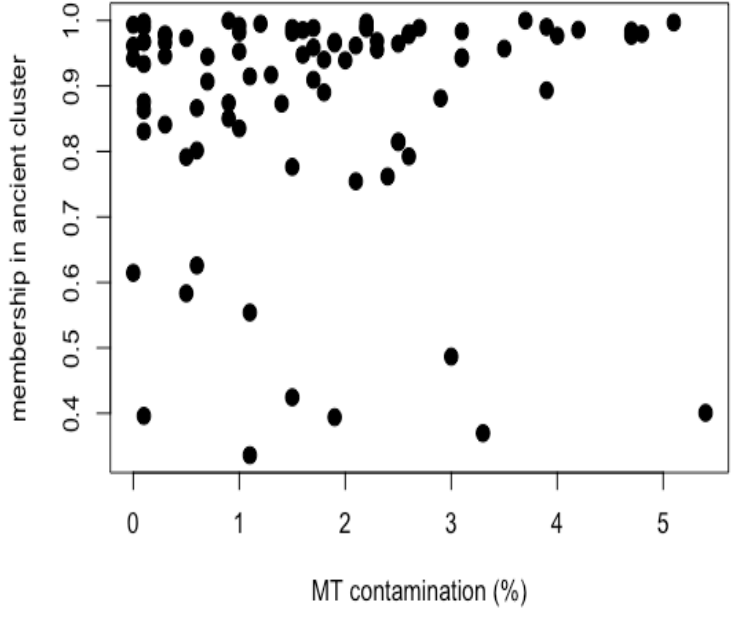
1) **aRchaic** only takes into account mismatches in calling mixing estimate and essentially throws away those reads that do not contain any mutations. General contamination models take care of that information

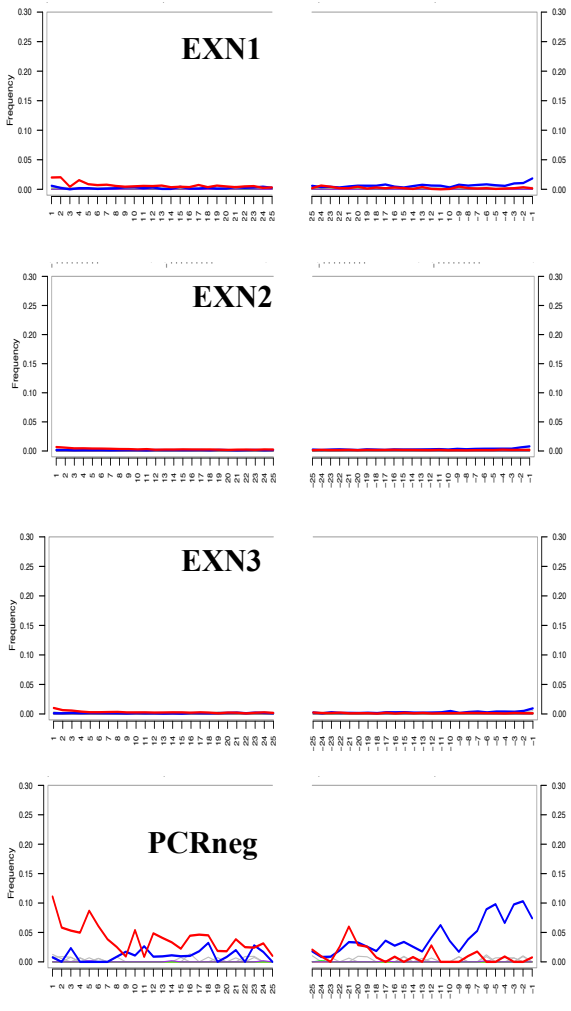
2) contemporary DNA that contaminates fossils also degrades to shorter length and accumulates mis-incorporations similar to ancient DNA (Briggs+2007, Malmström+2005, Sampietro+2006).

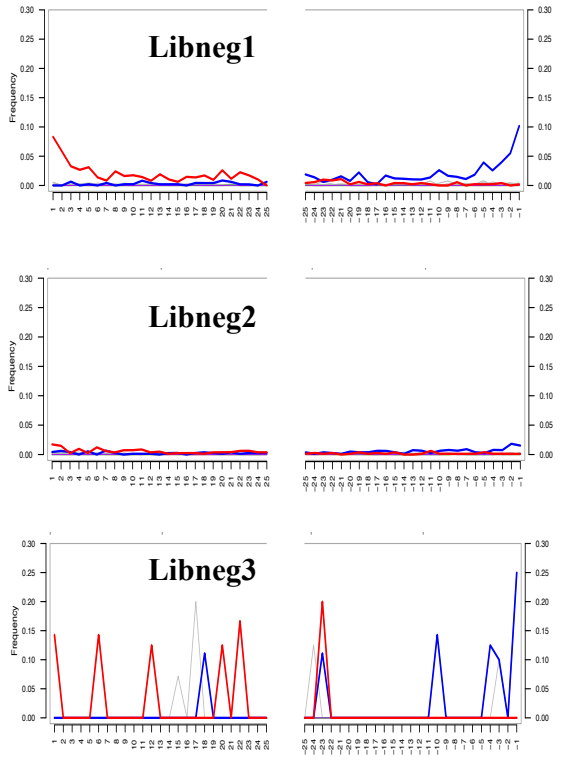
Moderns + Allentoft

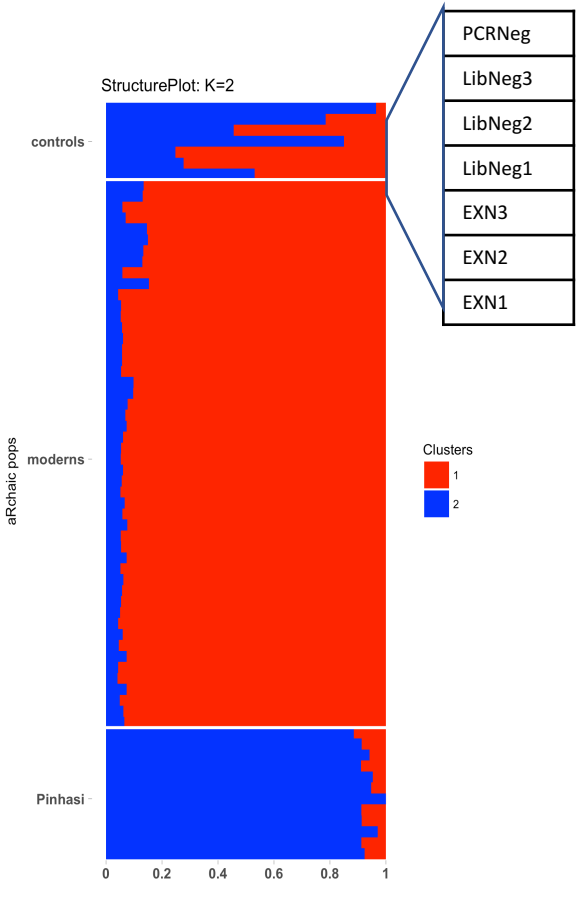


mtDNA contamination estimate vs aRchaic comparison for Allentoft samples




EXN1
EXN2
EXN3
PCRneg
Frequency


Libneg1
Libneg2
Libneg3
Frequency


StructurePlot: K=2
controls
moderns
Pinhasi
aRchaic pops
PCRNeg
LibNeg3
LibNeg2
LibNeg1
EXN3
EXN2
EXN1
Clusters
1
2
0
0.2
0.4
0.6
0.8
1

Acknowledgements

Matthew Stephens
John Novembre
Anna Di Rienzo

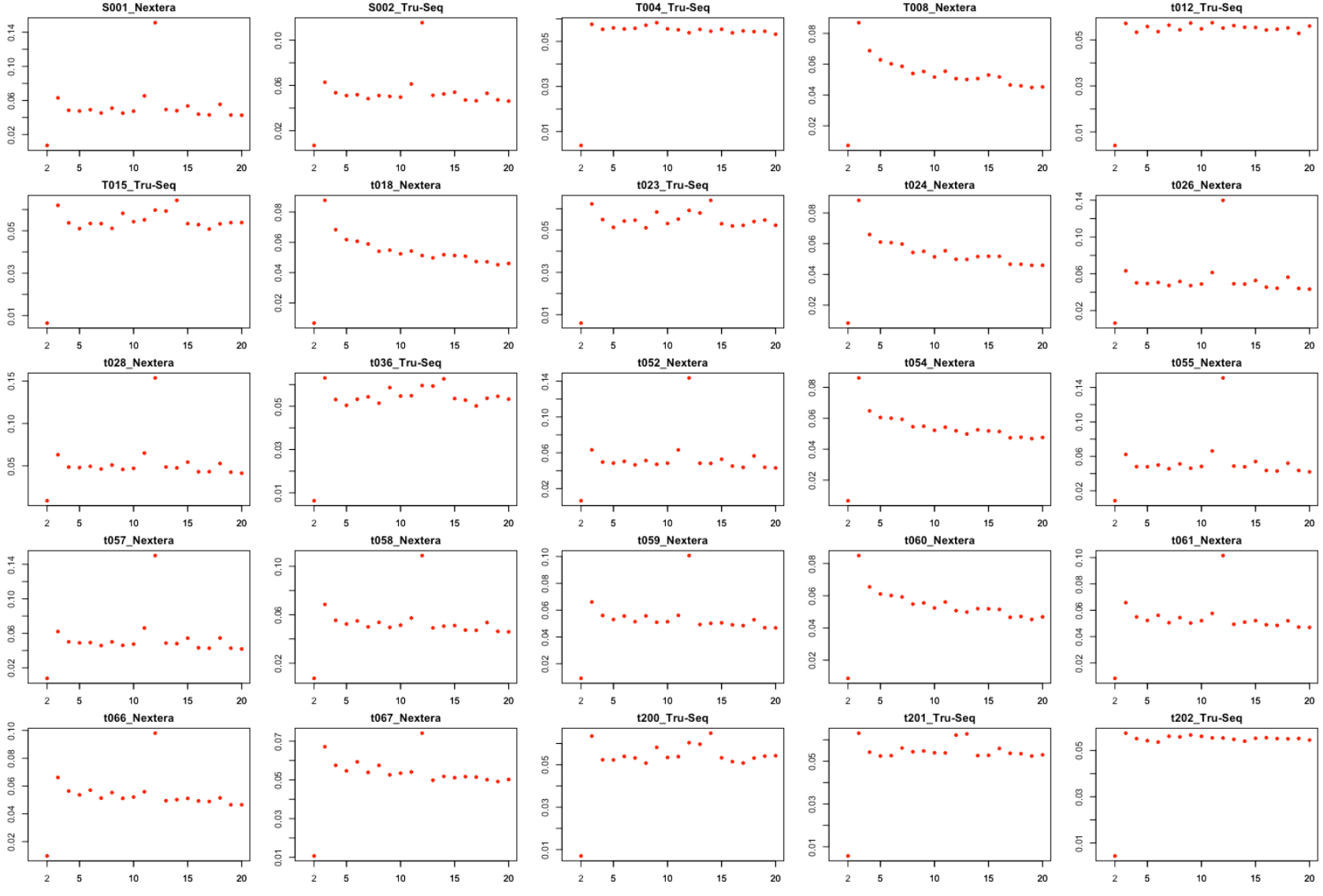
David Witonsky
John Lindo
Anna Gosling
Choongwong Jeong

John Blischak
Peter Carbonetto
Joe Marcus
Yuichi Shiraishi

Software

aRchaic is available as R package through
devtools::install_github("kkdey/aRchaic")

Webpage (still under work):
kkdey.github.io/aRchaic


S001_Nextera
S002_Tru-Seq
T004_Tru-Seq
T008_Nextera
t012_Tru-Seq
T015_Tru-Seq
t018_Nextera
t023_Tru-Seq
t024_Nextera
t026_Nextera
t028_Nextera
t036_Tru-Seq
t052_Nextera
t054_Nextera
t055_Nextera
t057_Nextera
t058_Nextera
t059_Nextera
t060_Nextera
t061_Nextera
t066_Nextera
t067_Nextera
t200_Tru-Seq
t201_Tru-Seq
t202_Tru-Seq