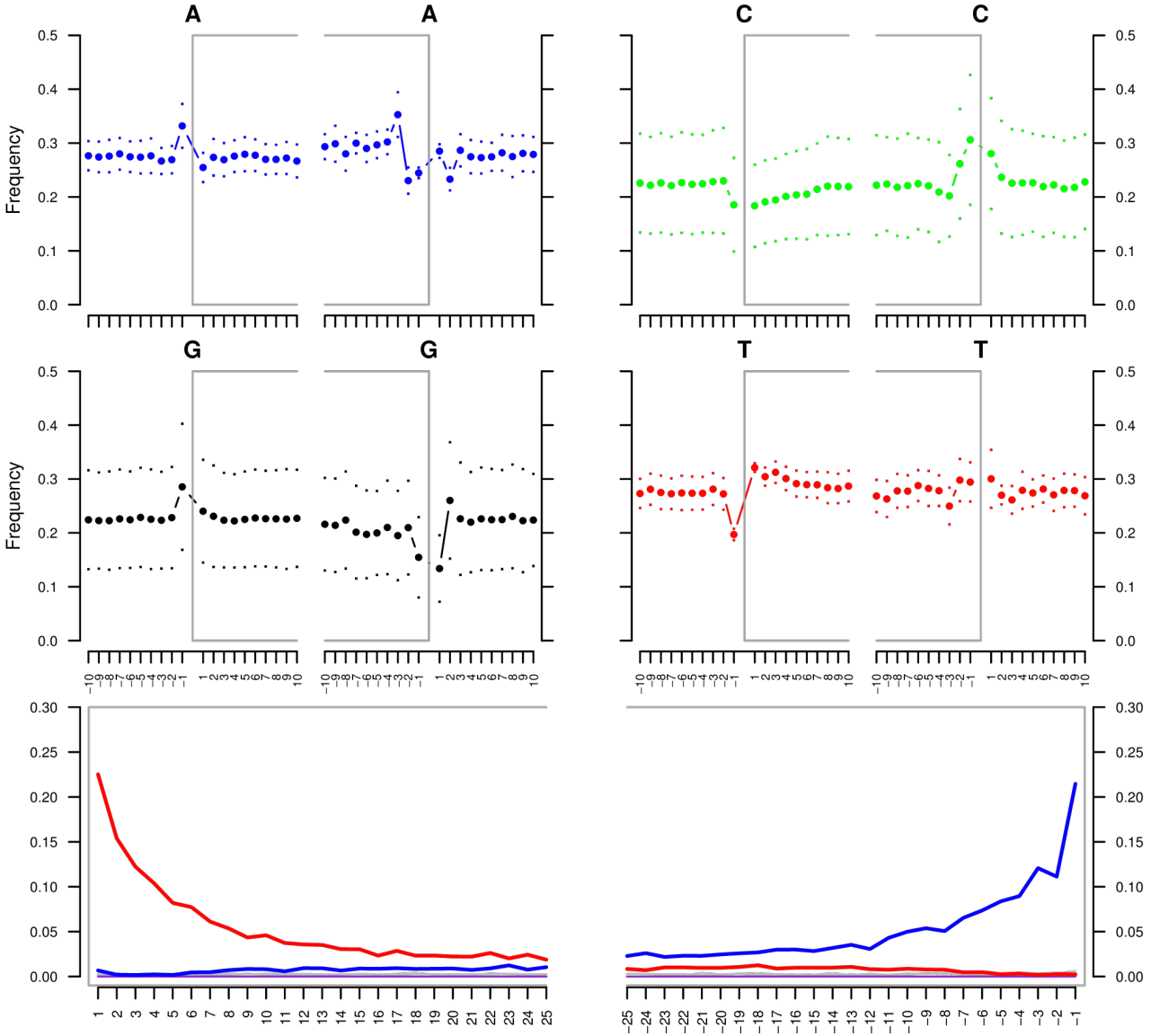
An R package for clustering and visualizing ancient dna signatures

By Hussein Al-Asadi & Kushal Dey

MapDamage


A
A
C
C
G
G
T
T
Frequency
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.00
0.05
0.10
0.15
0.20
0.25
0.30

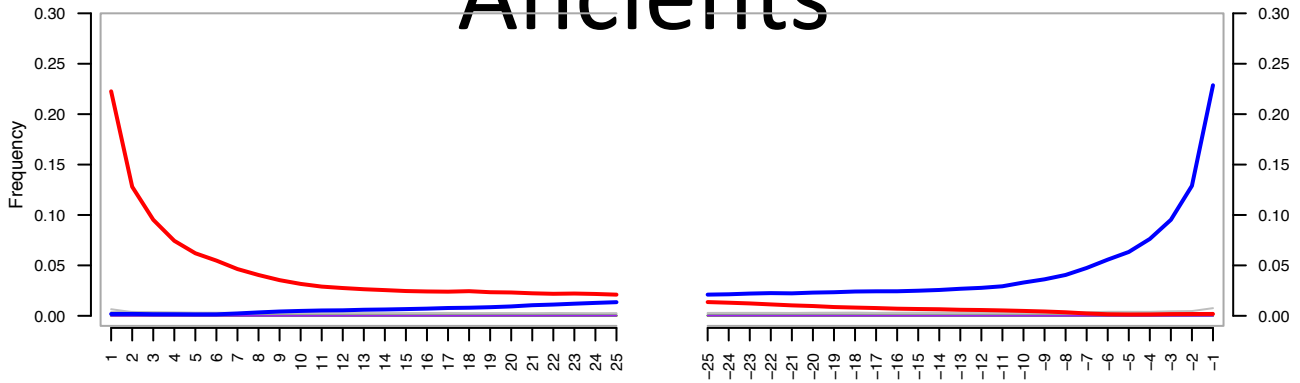
Data-set # 1

- DiRenzio Lab: 47 Ancients and 6 control samples

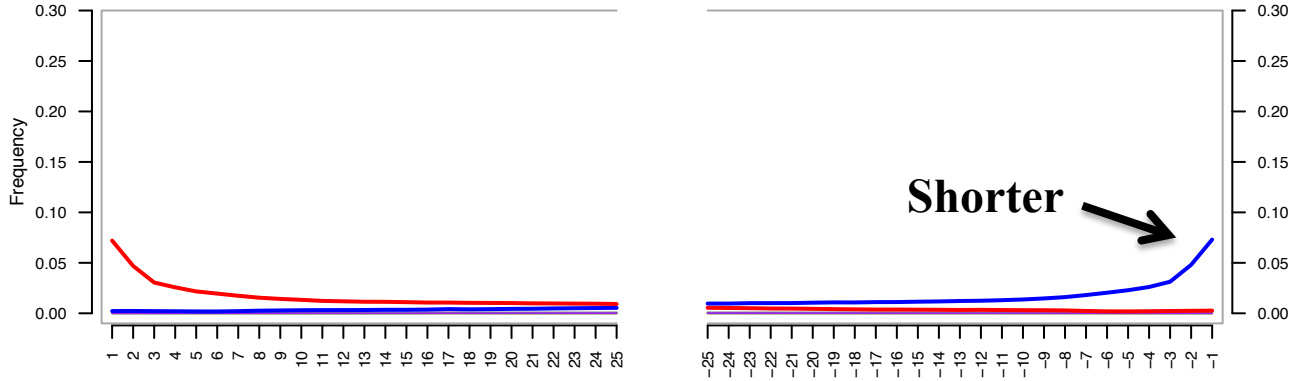
- Questions:
  - Do the ancients look “ancient”?
  - Is there any DNA in the the controls?
    - If so, is the DNA modern or ancient?

MapDamage on Ancients

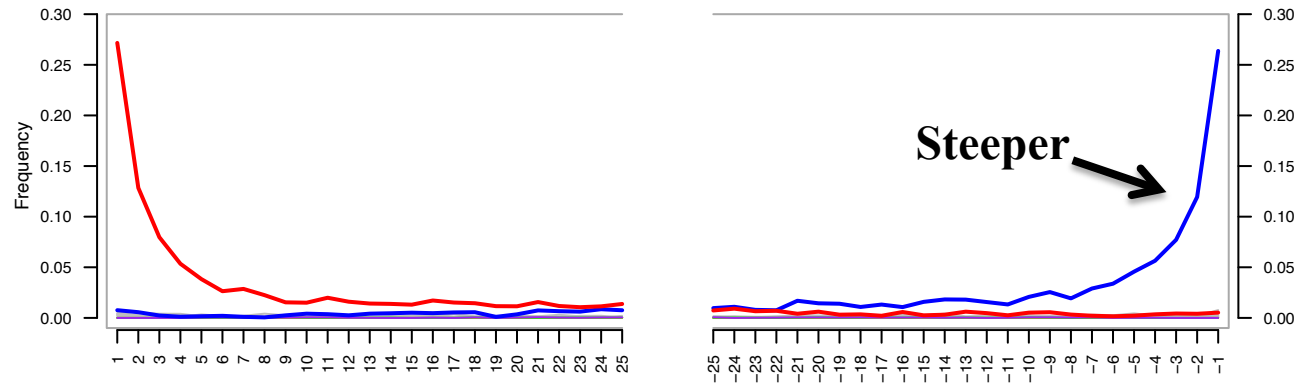
CNE1



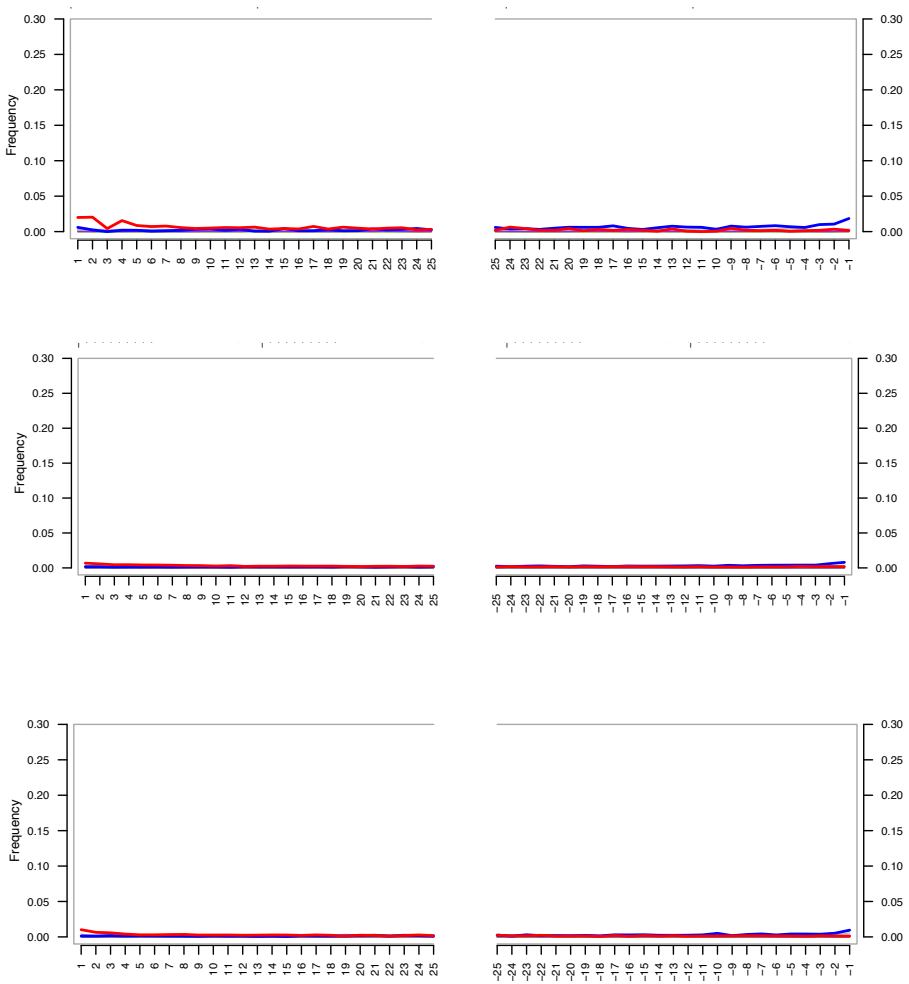
KS20

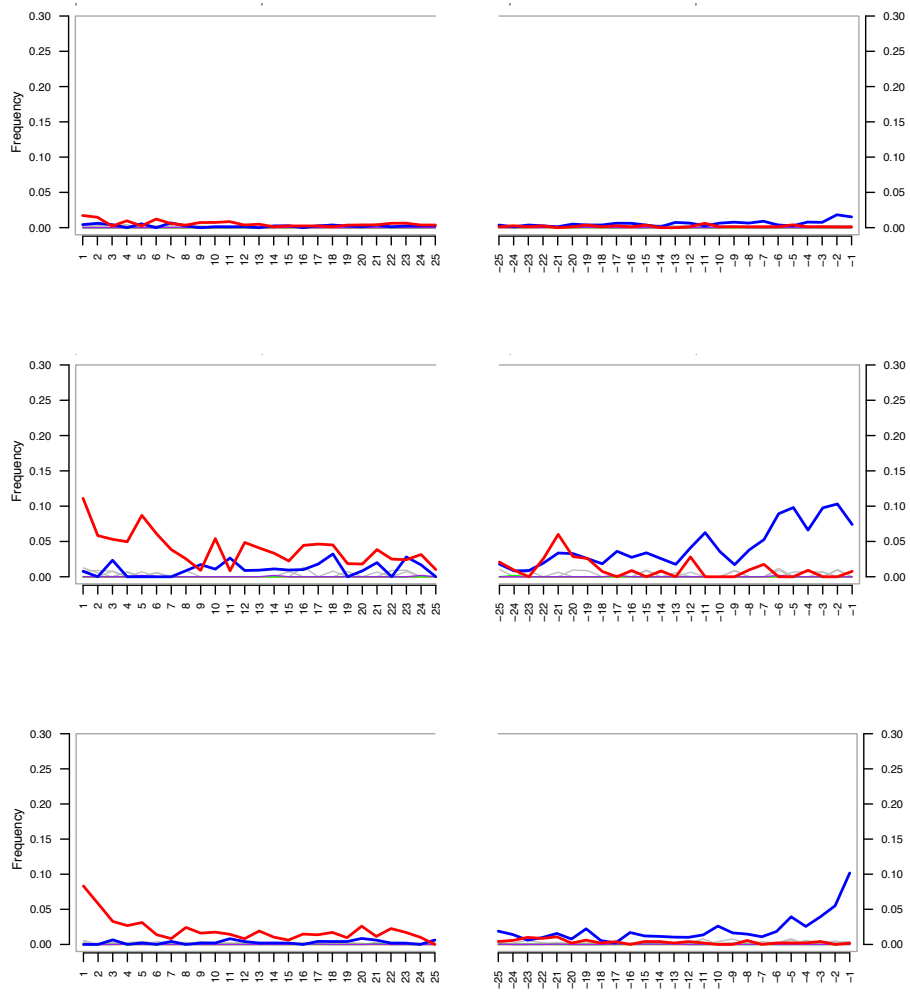


S30



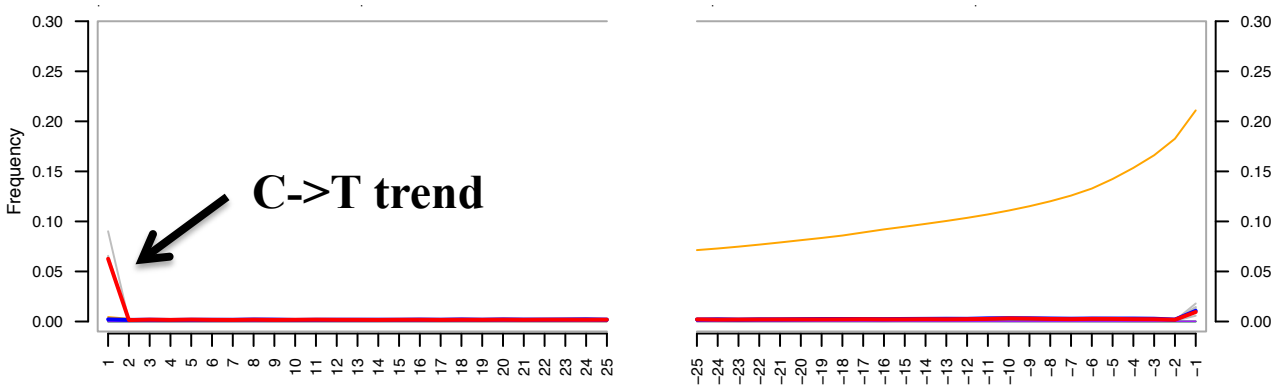
MapDamage on Controls



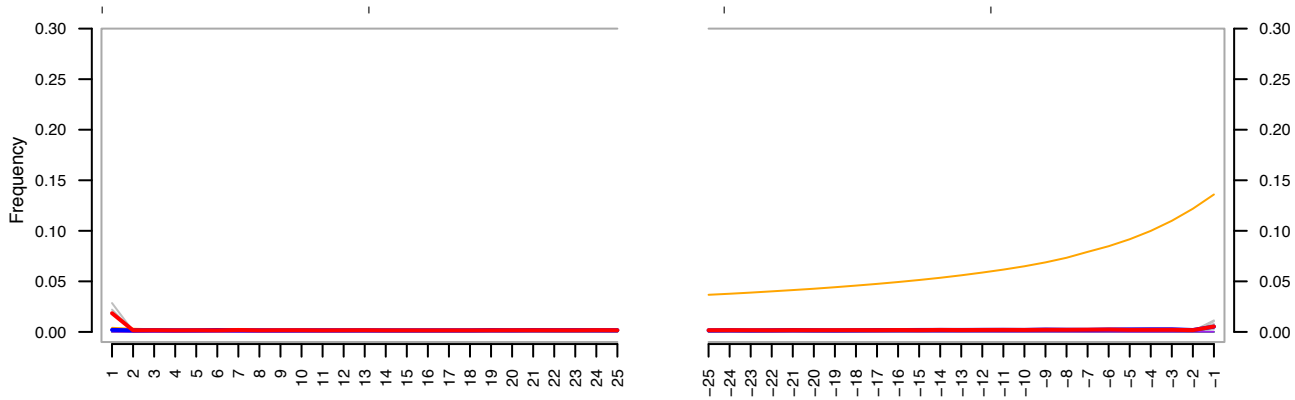


MapDamage on Moderns

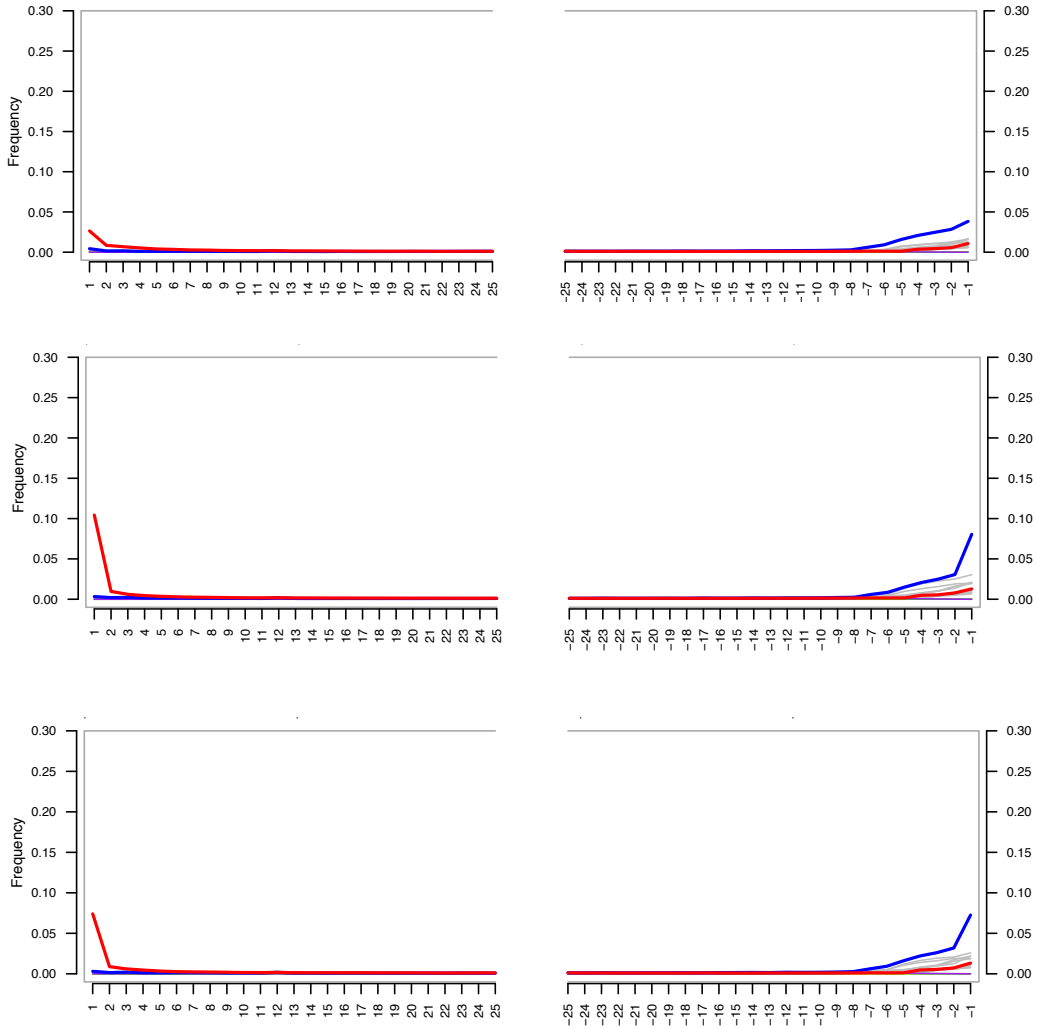
HG00097



HG00099



MapDamage on Ancient Sards. (different data-set)



Thoughts…

- Difficult to perform comparative analysis
  - Hard to look at all plots simultaneously
- Ancient samples have different C->T damage profiles.
  - Could be caused by different preservation conditions
- Are we missing other patterns?

High-level overview of aRchiac

- Step 1: Gets summary data from BAM files. Summary data is flexible. Not only C->T. Incorporates flanking bases.

- Step 2: Cluster and visualize samples based on signatures.
  - Allows you to visualize everything in one plot.
  - Allows you to compare samples
  - Clusters are flexible.

Step 1: Get signatures from BAM files

- Loop through BAM file. Record mutation, flanking bases (we use 2), and position on read.

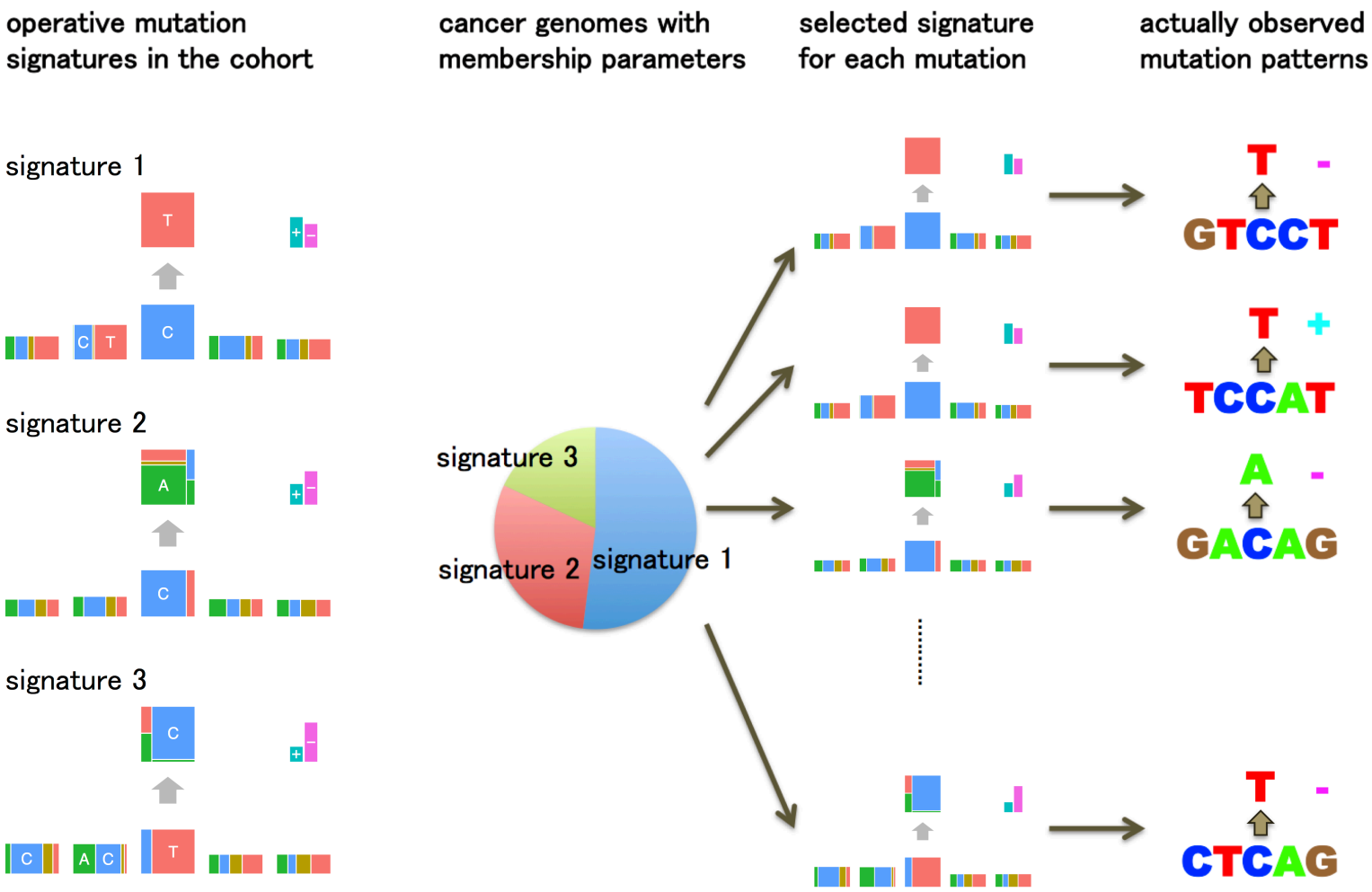
- Example Output
  - Sample_1, AA(T->A)AA, 0, 20
  - Sample_2, AA(T->A)AA, 10, 12
  - Sample_2, GC(C->T)TT, 2, 323

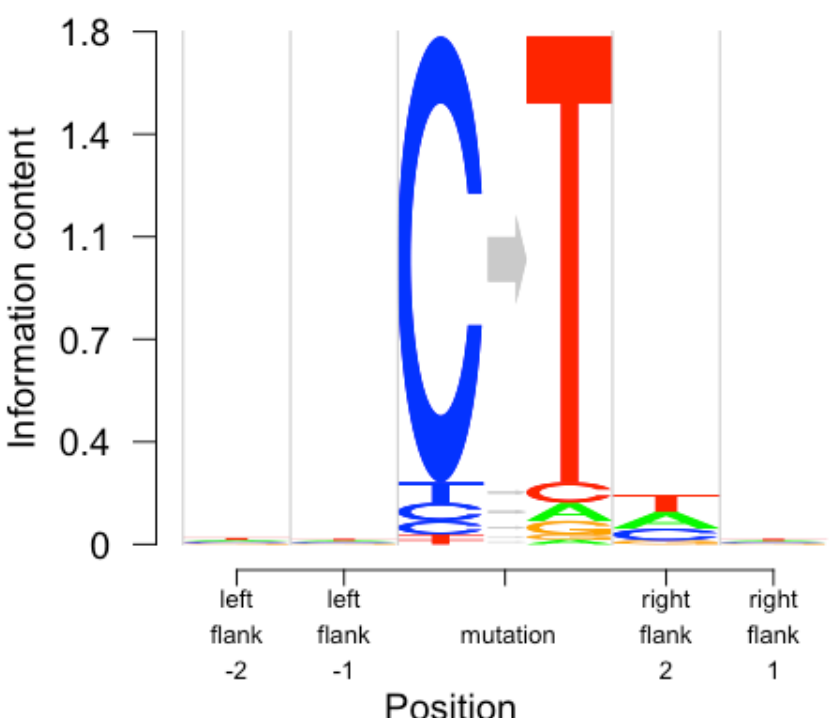
Step 2: Clustering and visualization

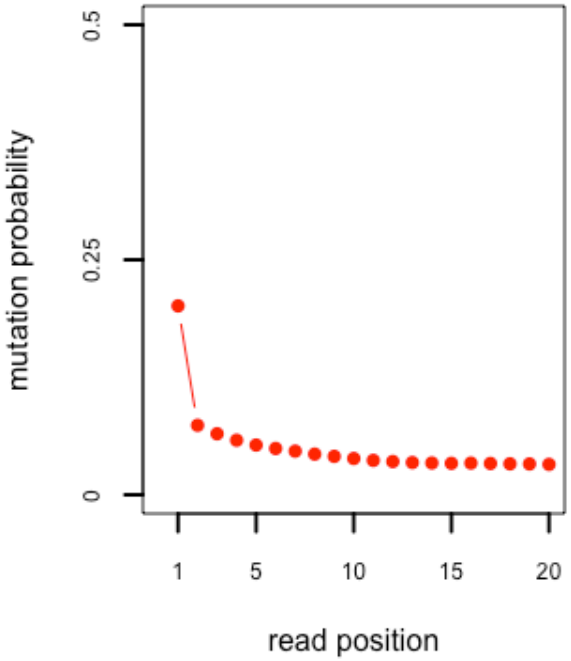
- STRUCTURE. Clusters are "source" allele frequencies. Each individual is a mixture of source allele frequencies. For example, if K=2, African-American will be mixture of "African" cluster and "European" cluster.

- Here, clusters are mutational signatures. Each individual has a membership for each cluster.
  - For example, if K=2, contaminated individual will be mixture of ancient cluster (defined by high C->T) and modern cluster (defined by modern transition/ transversion ratio)

Model based on paper by Yuichi & Matthew




1.8
1.4
1.1
0.7
0.4
0
Information content
left flank -2
left flank -1
mutation
right flank 2
right flank 1
Position


0.5
0.25
0
mutation probability
1
5
10
15
20
read position

Modifications

- Scaling
- Positional Information
- Implemented new code. Fixed optimization issues with Yuichi's code.

Back to Data-Set #1.

- Anna Gosling
- Anna Gosling + 1000g
- There is contamination

Data-set # 2

- Sards + Sherpa
- Sards + Sherpa + moderns.
- There is contamination

Data-set # 3

- IO + RISE + GOSLING + MODERN
- Batch effect or substructure. Be careful with interpretation.

- C->T is most important. Filter out C->T and red position
- Show PCA plot.

- Fix two clusters, one representative of ancient and one of moderns.
- Get an idea of contamination.
- Show them plot for Anna Gosling's and Sards data.