Making the Most of Ancient DNA

Homa Amini

Max-Planck Institute for Evolutionary Anthropology

Leipzig, Germany

October 27, 2016

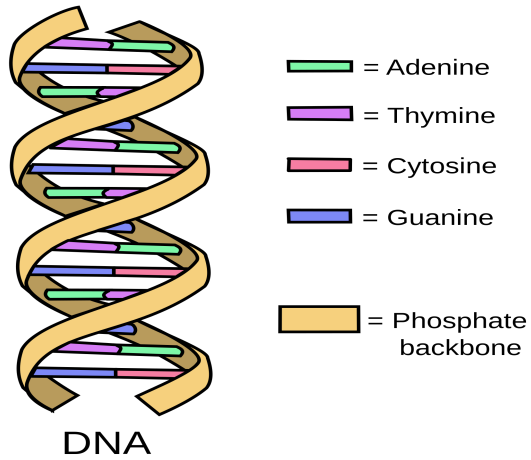
UPPSALA
UNIVERSITET

Why Study Ancient DNA?

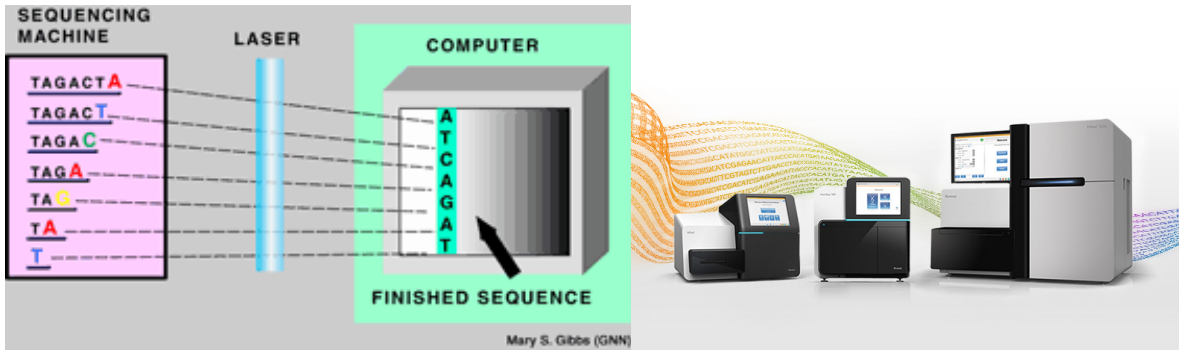
Source: www.sciencemag.org

What is a DNA Sequence?

- Let X be a sequence of λ random variables.
  $X = \{X_1, X_2, ..., X_\lambda\}$
  which takes the values from the alphabet
  $\{A, C, G, T\}$.



- e.g. ACGGCTAAGTCAGATACAGCTCGCA



Source: medivizor.com
http://www.genomenewsnetwork.org

What is Ancient DNA?

Let lowercase $x = \{x_1, x_2, ..., x_\lambda\}$ represent a specific DNA sequence.

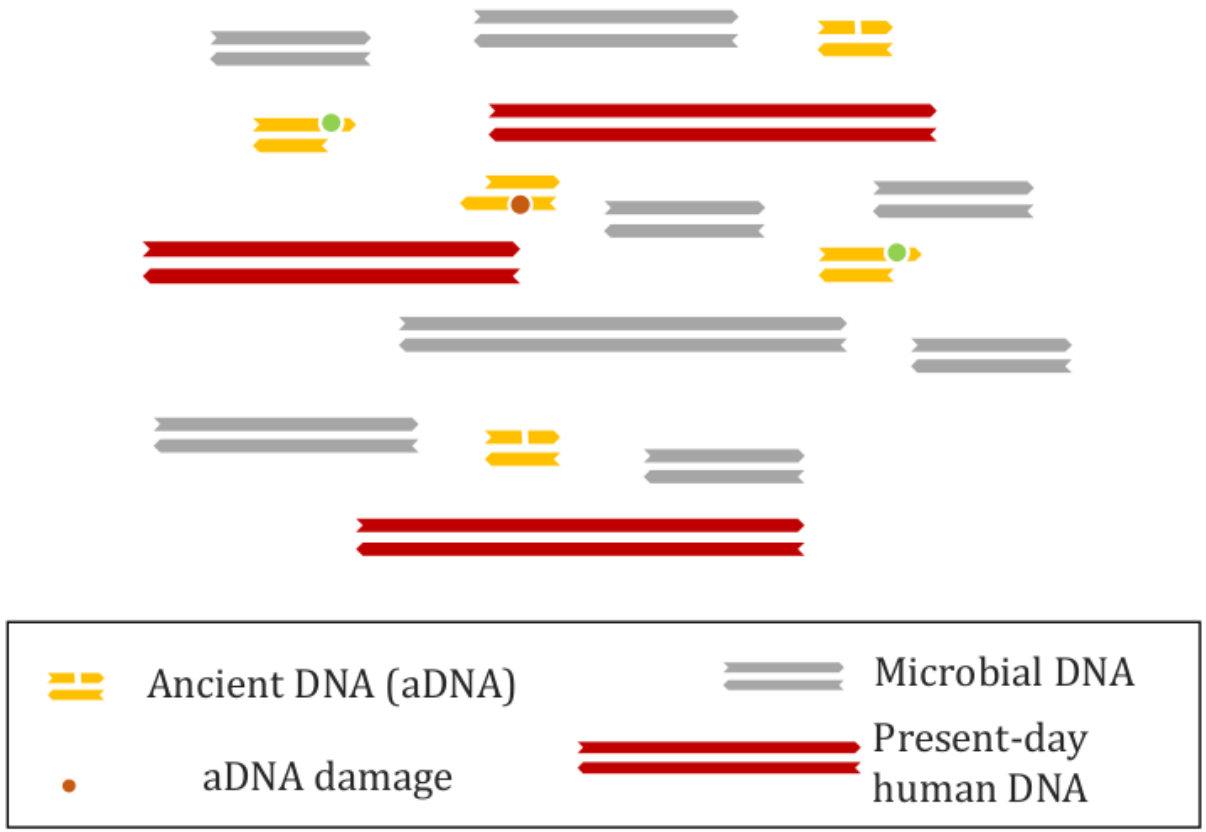
- e.g. ATTGTTACAGATATT

Characteristics

- Hard to retrieve.
- Short molecule $\leq$ 50 bp.
- *Post-mortem* molecule damage: deamination (substitutions of C$\rightarrow$T).
- Microbial contamination.

Source: Meyer M,et al. Science 2012 & http://www.ourdailyread.com/2014/02/

Sequencing Ancient DNA



DNA Similarities

- DNA sequences may have changed from a common ancestor through various reasons:
  - Change of a letter

    ACGCTATGCA
    ACCCTATGCA
  - Insertion of a letter

    ACGCTATGC–A
    ACGCTATGCTA
  - Deletion of a letter

    ACGCTATGCA
    ACGCT–TGCA
- A few mutations make sequences different, but still "similar" and we are looking for these similarities.

DNA Sequence Alignment

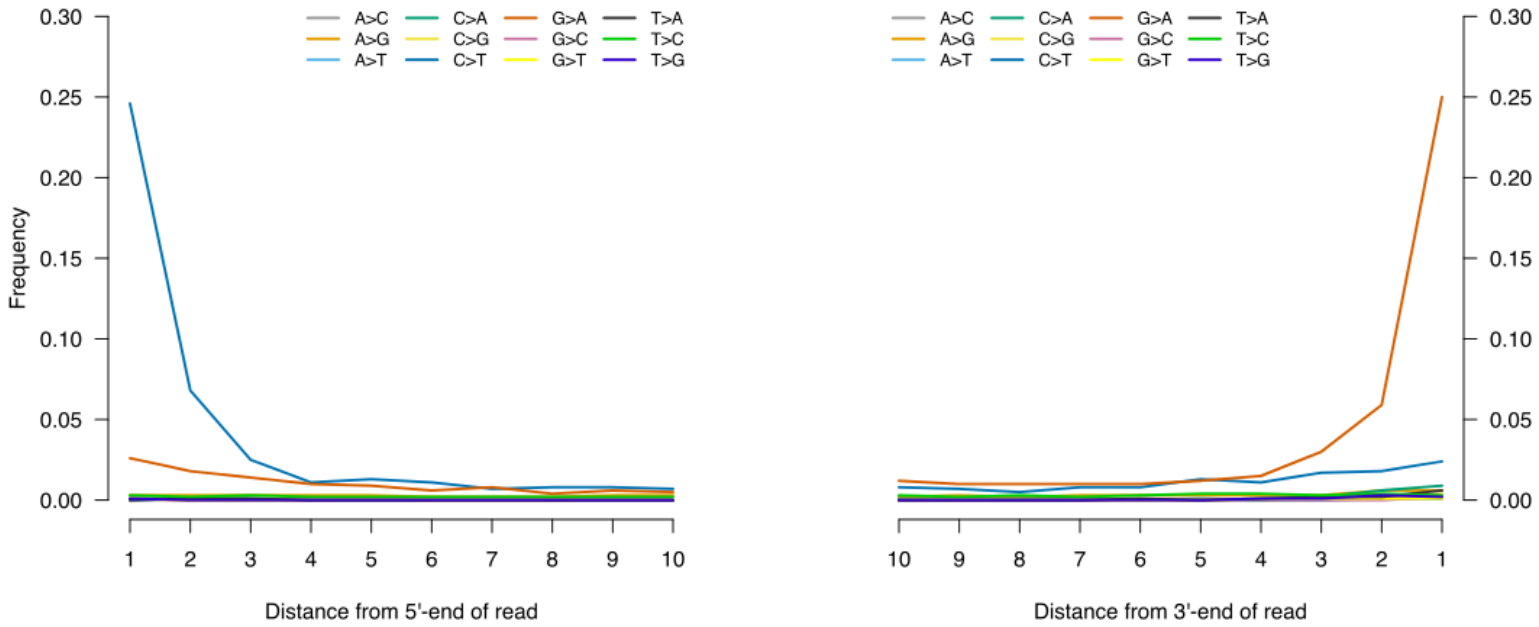
- Alignment to the reference genome to identify the molecules we are interested in.
- Approximate pattern matching.
- e.g.

S_1 : **ACGCTATAGCA**
S_2 : **CGATGAC**

A plausible alignment of S_1 and S_2:

S_1 : **ACGCTATAGCA-**
S_2 : **-CG--AT-G-AC**

Post-mortem Deamination Damage



A high nucleotide misincorporation rate of thymine(T) in place of cytosine(C) near the ends of ancient DNA reads.

Source: http://mitosuite.com/example/results.html

What Sort of Alignment?

- **Global alignment:**

S_1 : ACCGTCGCTACTGCTGT.....CAGATCGCTCATCGCATACTGTCT
S_2 : ACCGTGTA ..... CAGATCGCTCATCATCGCAGTCGATAGCCTGTCT

- **Local alignment:**

S_1 : ACCGTCGCTACTGCTGTCAGATCGCTCAGTTCGATCTG
S_2 : GTCTGTCAGAGATCGCTCATCGCATACTGTCTCGTCC

- **Semi-Global alignment:**

S_1 : ACCGTCGCTACTGCTGTCAGATCGCTCAGTTCGCTGTCAAGTCACT
S_2 : AGATCGCTCA

Scoring an Alignment

- The alignment score is the sum of all the scores of paired aligned characters plus the gap scores.
- e.g. substitution matrix

$S(S_{1i}, S_{2j})$	A	C	G	T
A	1	-1	-1	-1
C	-1	1	-1	-1
G	-1	-1	1	-1
T	-1	-1	-1	1

If g $= -1$

S_1 :	C	C	G	A	-	T	A
S_2 :	T	C	G	-	C	T	A

we score it by:

$$S(C,T) + S(C,C) + S(G,G) + 2g + S(T,T) + S(A,A)$$

$$-1 + 1 + 1 - 1 - 1 + 1 + 1 = 1$$

How to Find the Optimal Alignment?

- Simple approach: Compute and score all possible alignments.

  There are:

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \simeq \frac{2^{2n}}{\sqrt{2\pi n}}$$

  global alignments.
  - Time complexity $\mathcal{O}(2^{2n})$.
- Dynamic programming
  - Calculate the best alignment of all prefixes of the sequences instead of the best alignment of the two sequences.
  - Time complexity $\mathcal{O}(n^2)$.

Overview of BWA Aligner

- Based on FM-index (Burrows-Wheeler Transform plus auxiliary data structures) which enables fast matching.
- Fast and moderate memory footprint (<4GB).
- Backtracking algorithm with heuristics.
- Seed heuristic.
- Does not take deamination into consideration.

Why is a New Aligner Needed?

- Alignment 1:

  ACCCACTCTACCTCATCGACC

  ATTTACTCTACCTCATCGATT

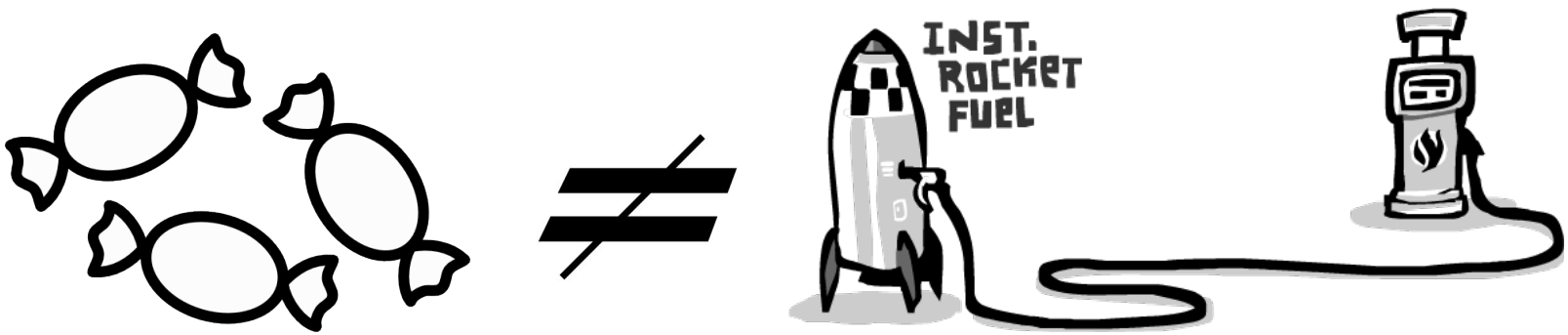
- Alignment 2:

  ACCCACTCTACCTCATCGACC

  CCCCTGTTTACCACCTCGACC

R-Candy: New Aligner for Ancient DNA

- Aligns ultra-short reads of length $\leq$ 35bp.
- Copes with high levels of specific damage expected in ancient DNA.



Source: http://www.5ideas.in

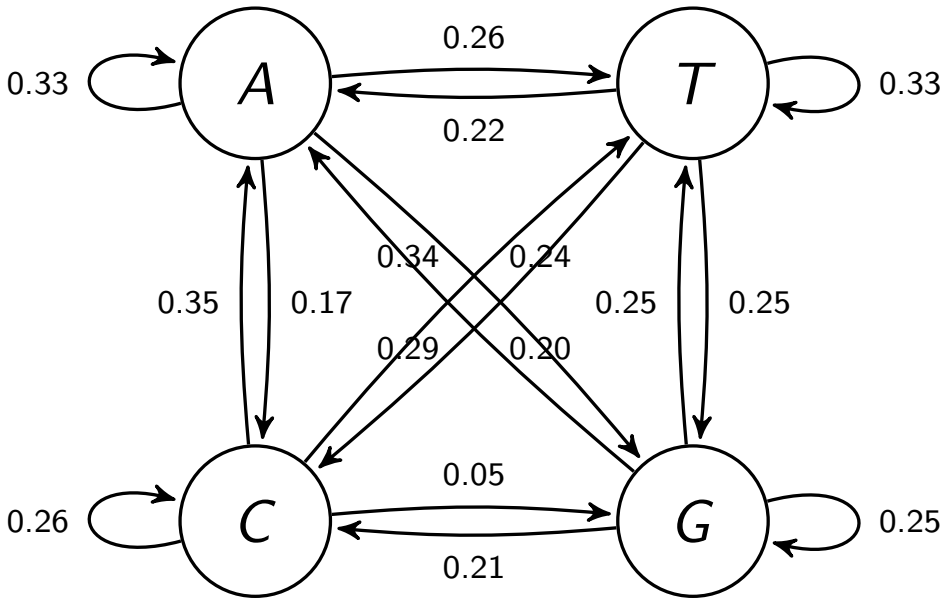
Overview of R-Candy's Algorithm

- Based on FM-index (Burrows-Wheeler Transform plus auxiliary data structures) which enables fast matching.
- Backtracking on FM-index.
- Implements a model for damage expected on ancient DNA.
- Tailored for short reads of ancient DNA.

Test Data

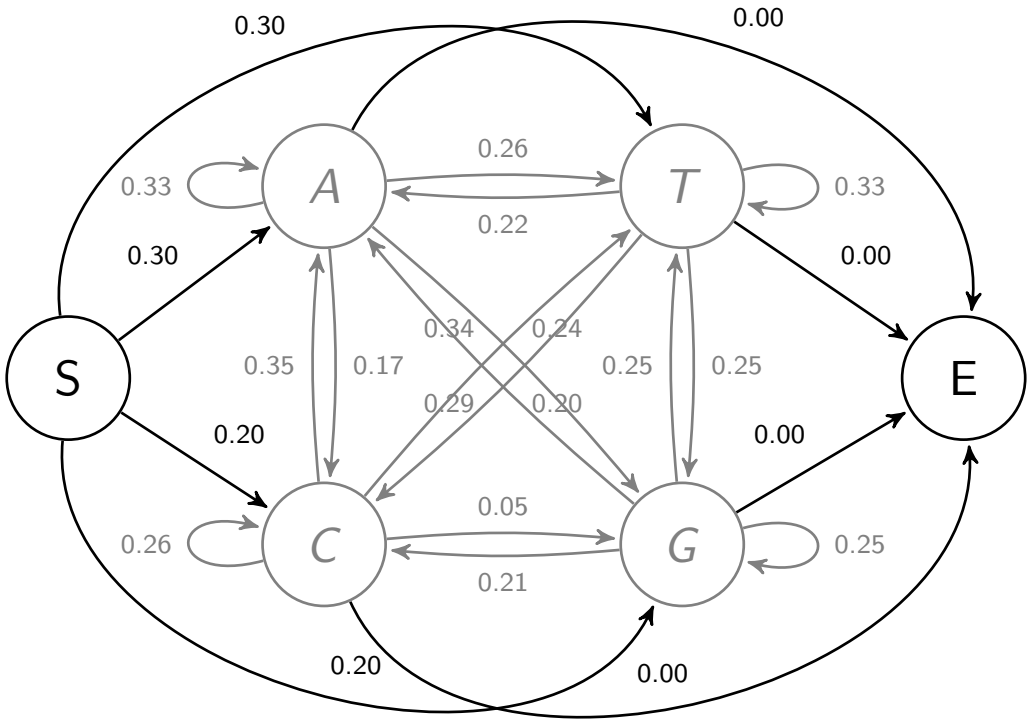
- To compare sensitivity of R-Candy to BWA aligner.
  - Simulated modern and ancient DNA reads originated from a simulated genome.
- To rule out possible artifacts of genome simulation.
  - Simulated modern and ancient DNA reads originated from a real genome.
- To evaluate the rate of false positive alignments.
  - Simulated random reads to mimic exogenous DNA.

Genome Simulation



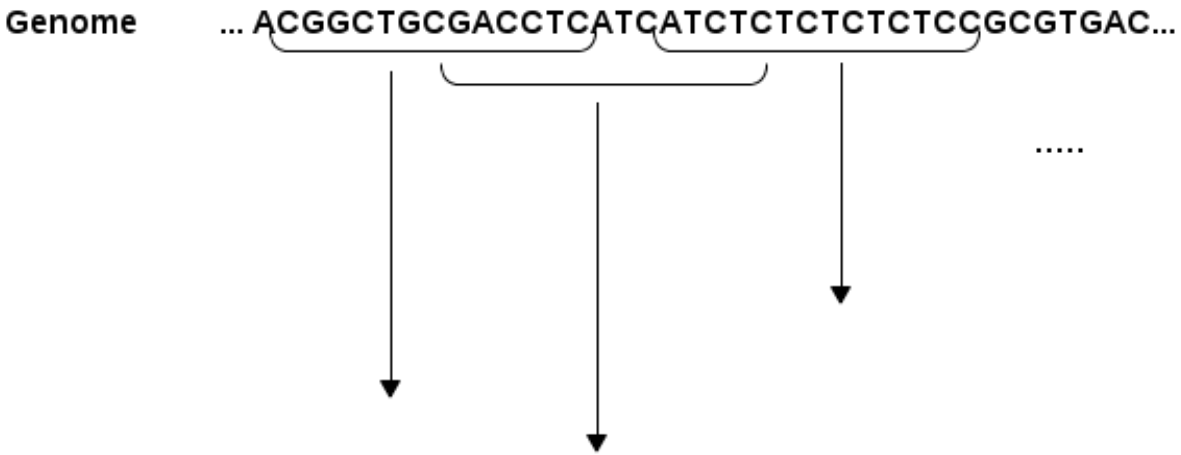
Trained 1st order Markov chain based on the human reference genome.

1st Order Markov Chain



Trained 1st Order Markov-Chain based on the human reference genome with start and end states

Read Simulation

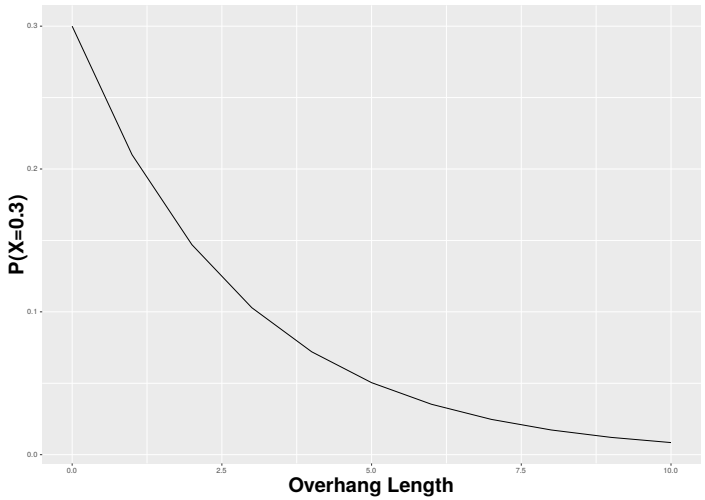


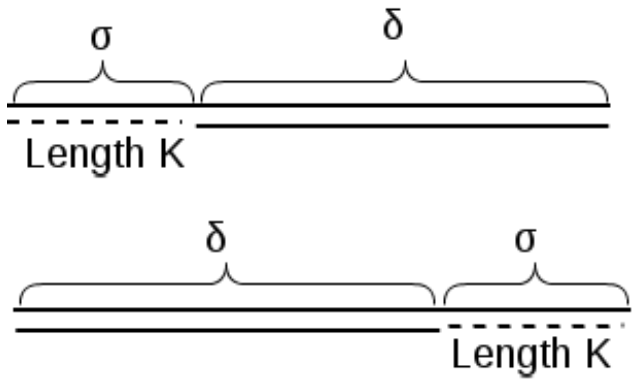
Reads.sam

```
Genomic_0  0  Simulated_Genome  8657611  0  *  *  0  0   CGGCTGACCTCATCT
Genomic_1  0  Simulated_Genome  1668761  0  *  *  0  0   TGACGGGGTCGTTCA
```

Read Simulation (Cont.)

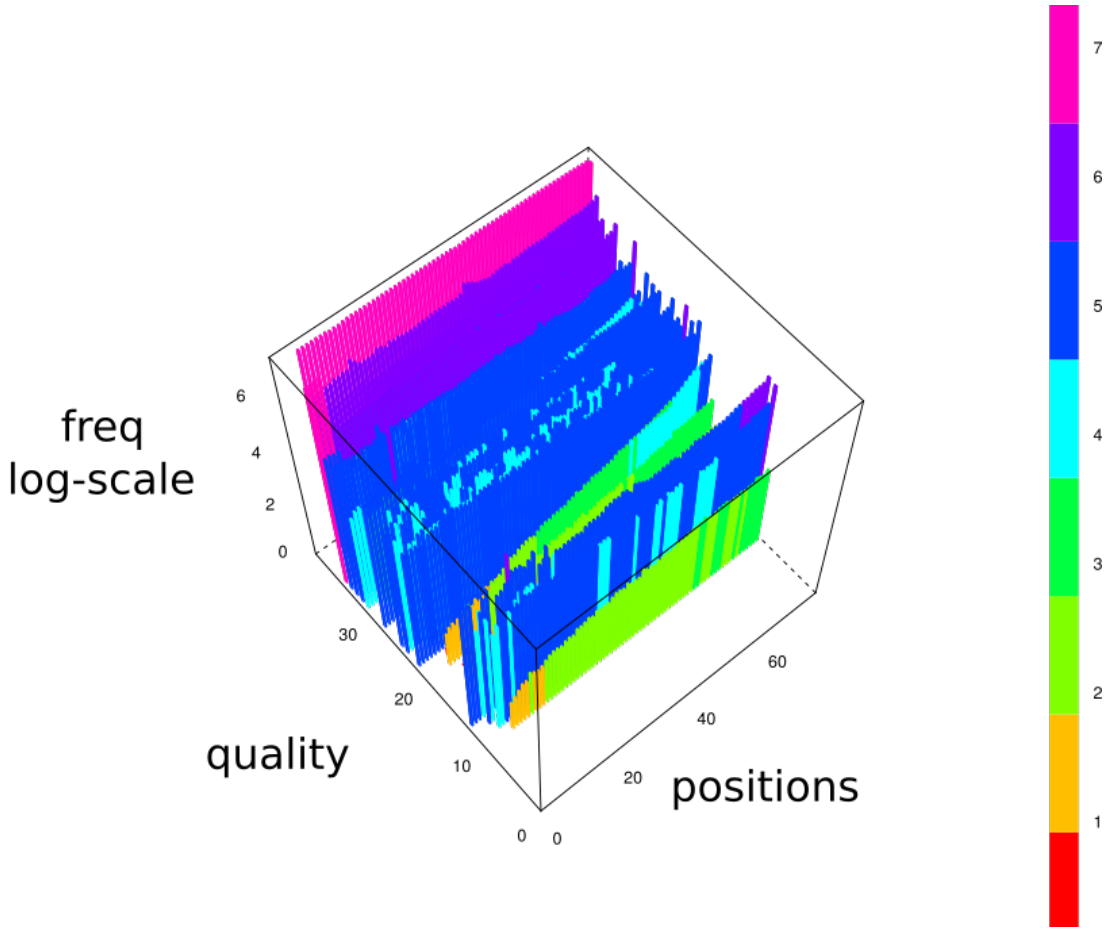
- **Divergence**: different divergence rates are simulated by a given number of mismatches per read.
- **Deamination damage**: based on deamination parameters (overhang (σ) and double stranded (δ) deamination rates plus the probability of being in overhangs) provided by users.





Read Simulation (Cont.)

- **Sequencing error**: specific types of errors introduced by the sequencing machine based on the distribution of quality scores on an actual sequencing run.



Read Simulation (Cont.)

	Divergence	Deamination Damage	Sequencing Error
Modern DNA	✓		✓
Ancient DNA	✓	✓	✓
Exogenous DNA		✓	✓

Evaluation Criteria

- **Mapping accuracy**:

$$\textbf{Sensitivity}(TPR) = \frac{\#TP}{\#TP + \#FN}$$

$$\textbf{Specificity}(TNR) = \frac{\#TN}{\#TN + \#FP}$$

- **Throughput**: the number of mapped reads per second.
- **Memory footprint**: required memory for indexing, processing and storing.

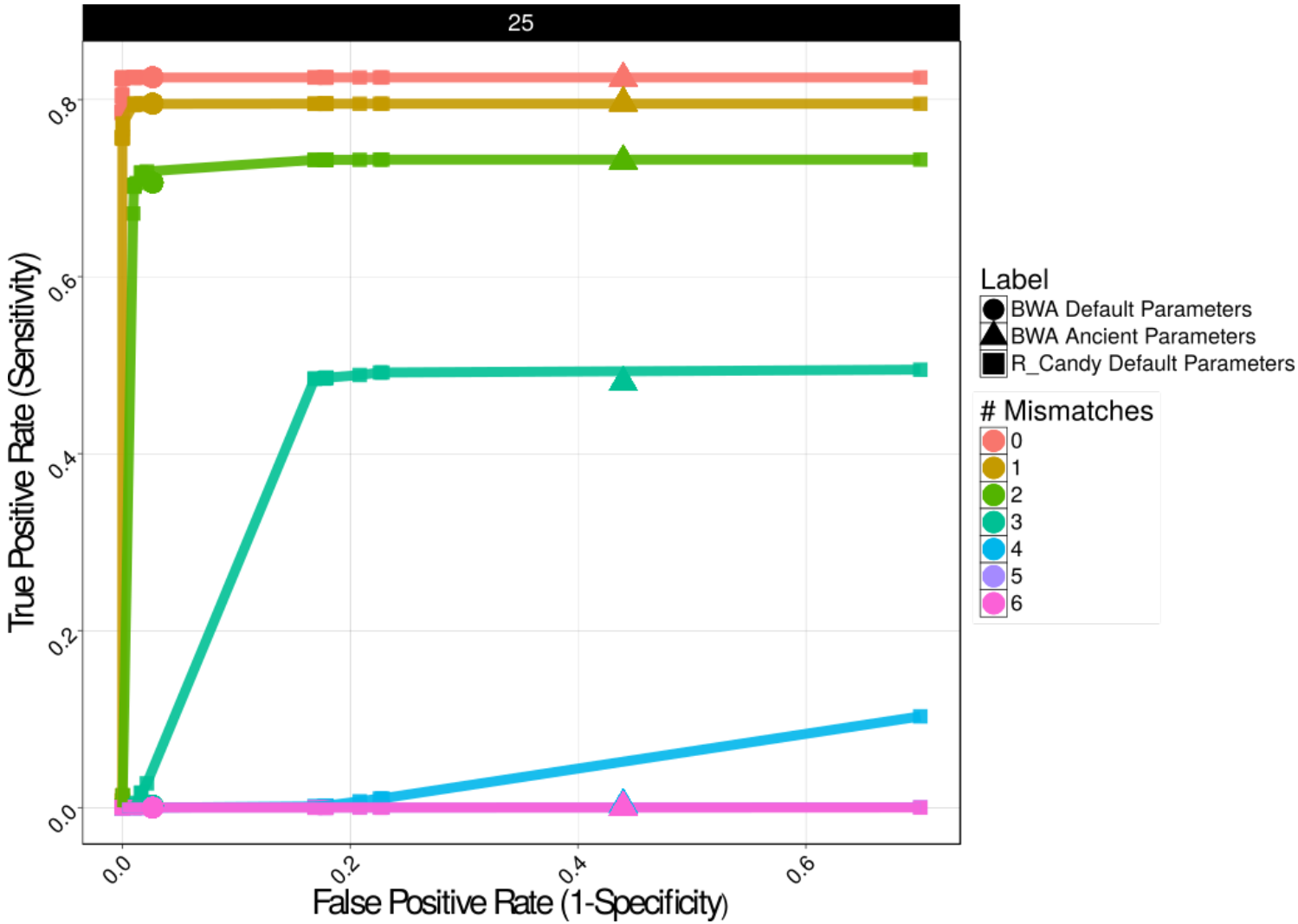
Source: alazycowboy.com

Evaluation Scenarios

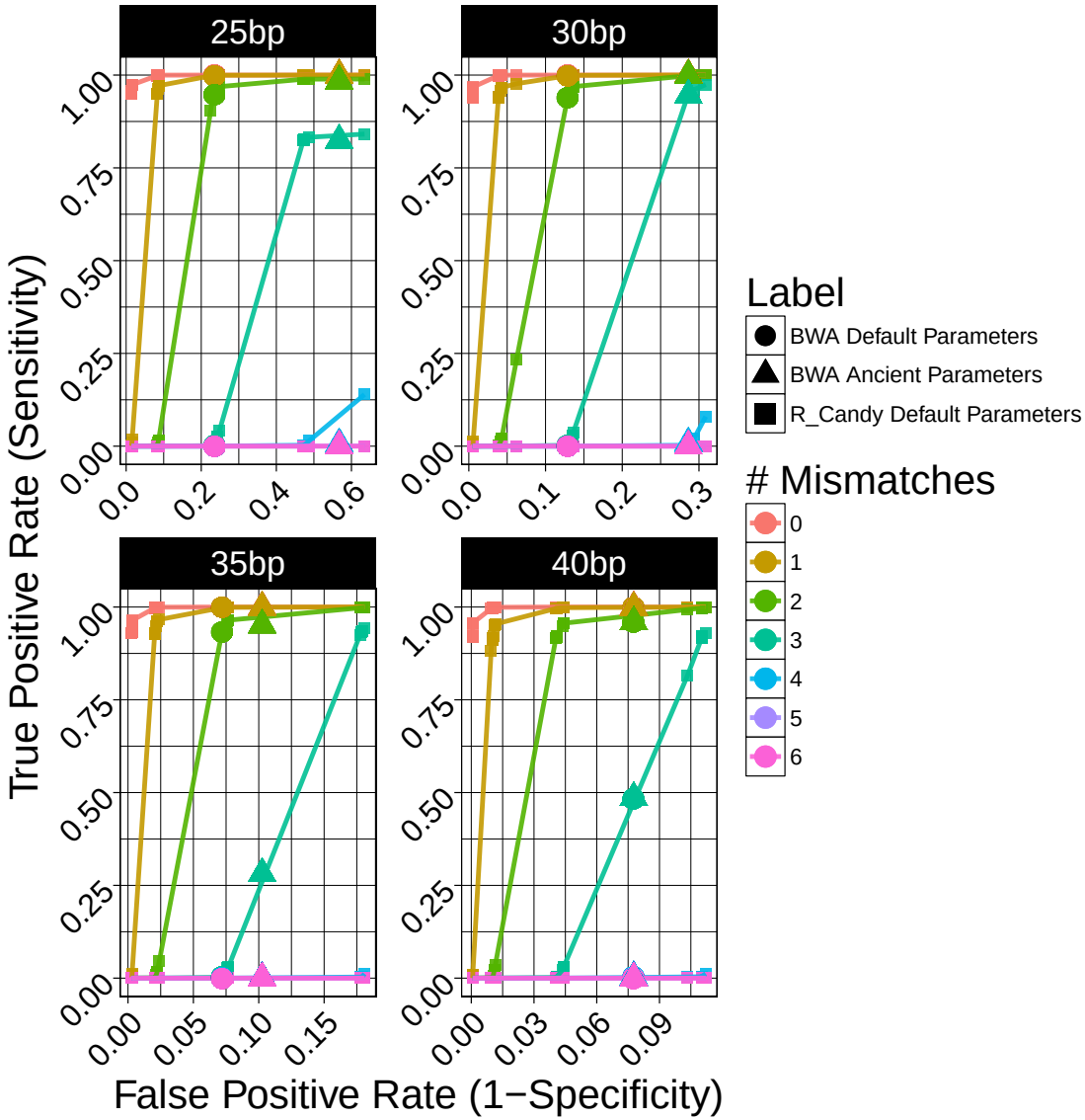
	DNA Reads		Aligned To		Parameters	
	Modern	Ancient	Simulated Genome	Real Genome	Default (No Deamination)	Ancient
MSD	✓		✓		✓	
MRD	✓			✓	✓	
MSA	✓		✓			✓
MRA	✓			✓		✓
ASA		✓	✓			✓
ARA		✓		✓		✓

M = Modern DNA
A = Ancient DNA/parameters
S = Simulated genome
R = Real genome
D = Default parameters

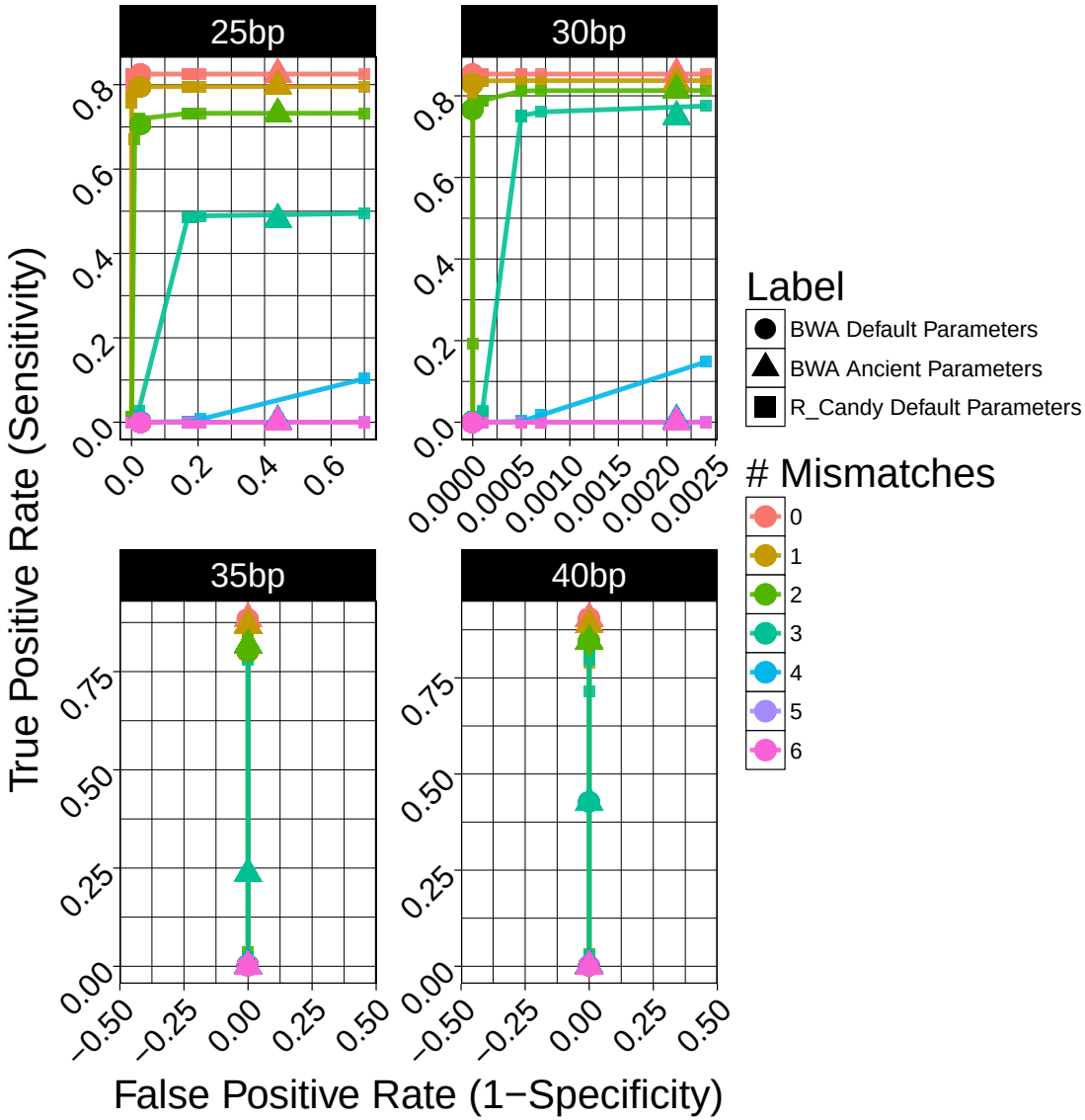
ROC Curve



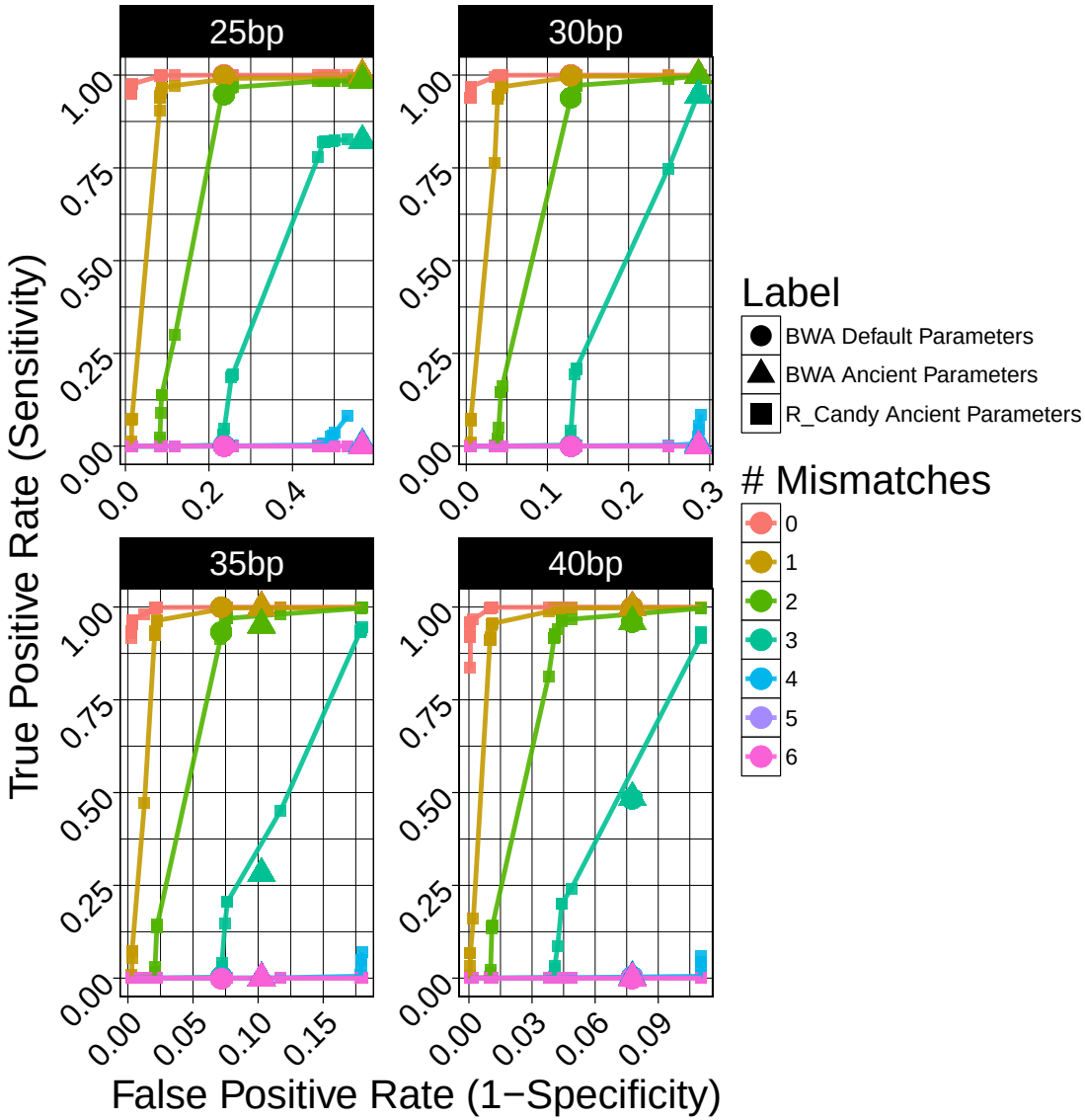
Modern DNA Simulated Genome, Default Parameters



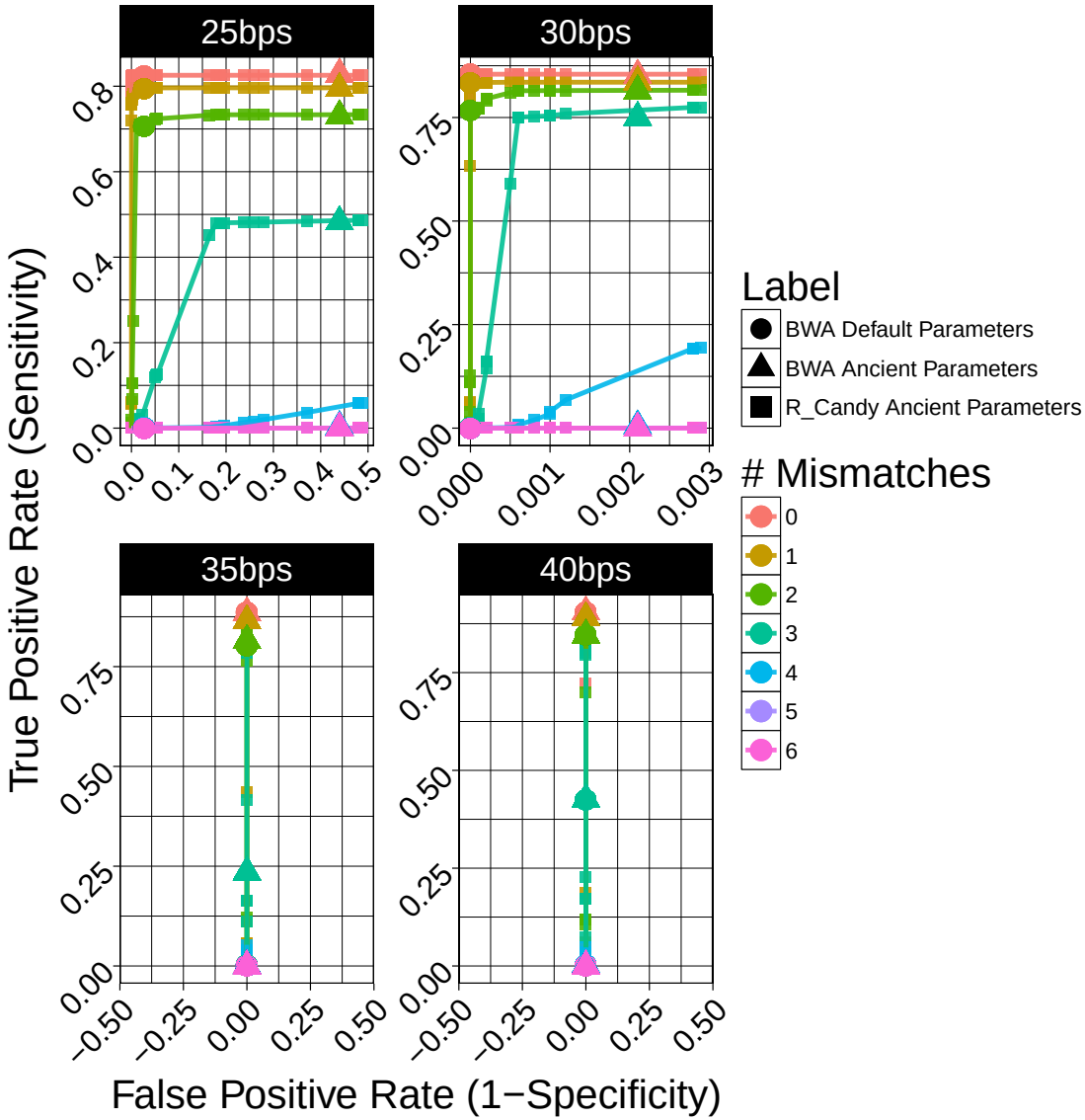
Modern DNA Human Ref Genome, Default Parameters



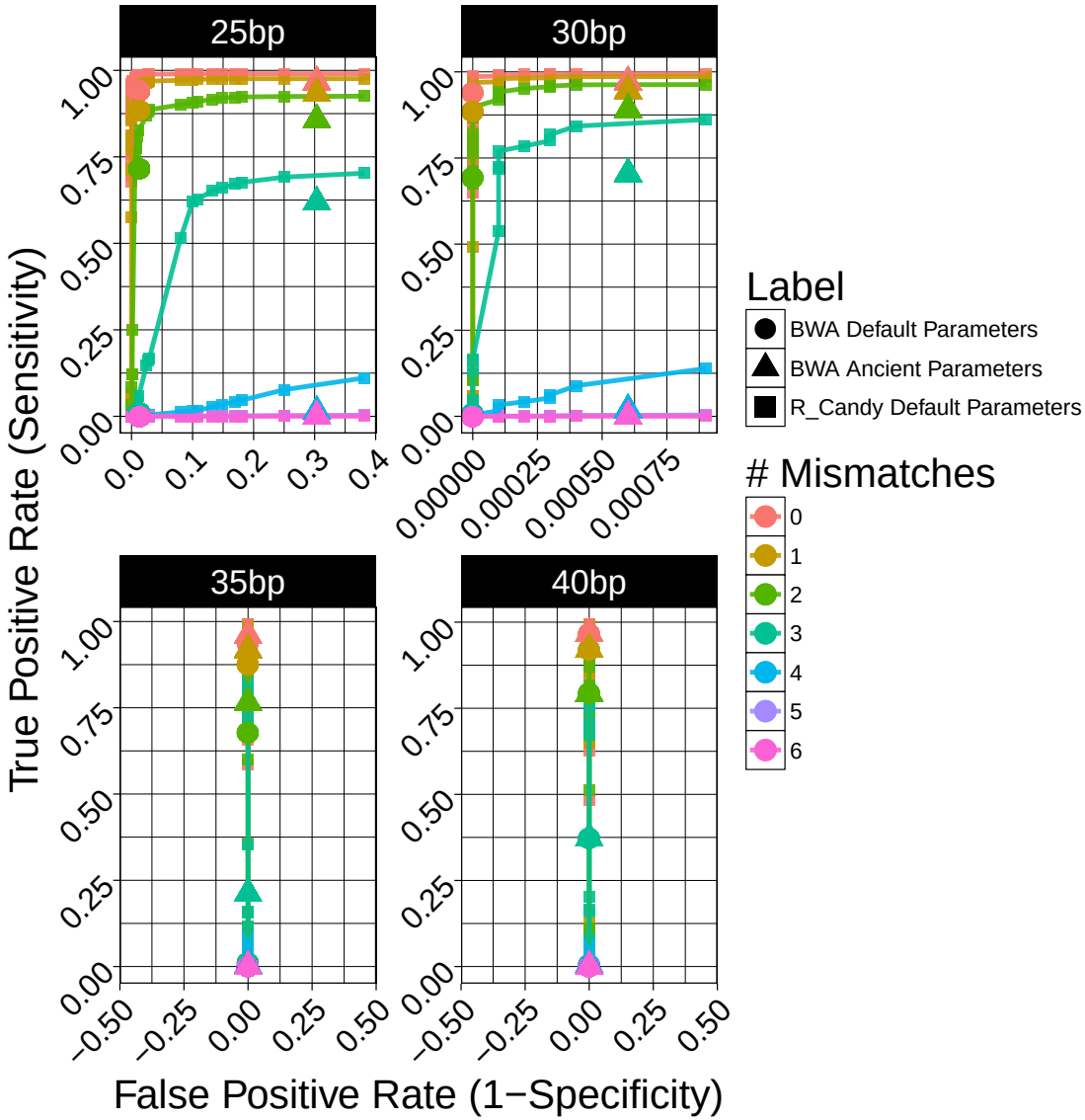
Modern DNA Simulated Genome, Ancient Parameters



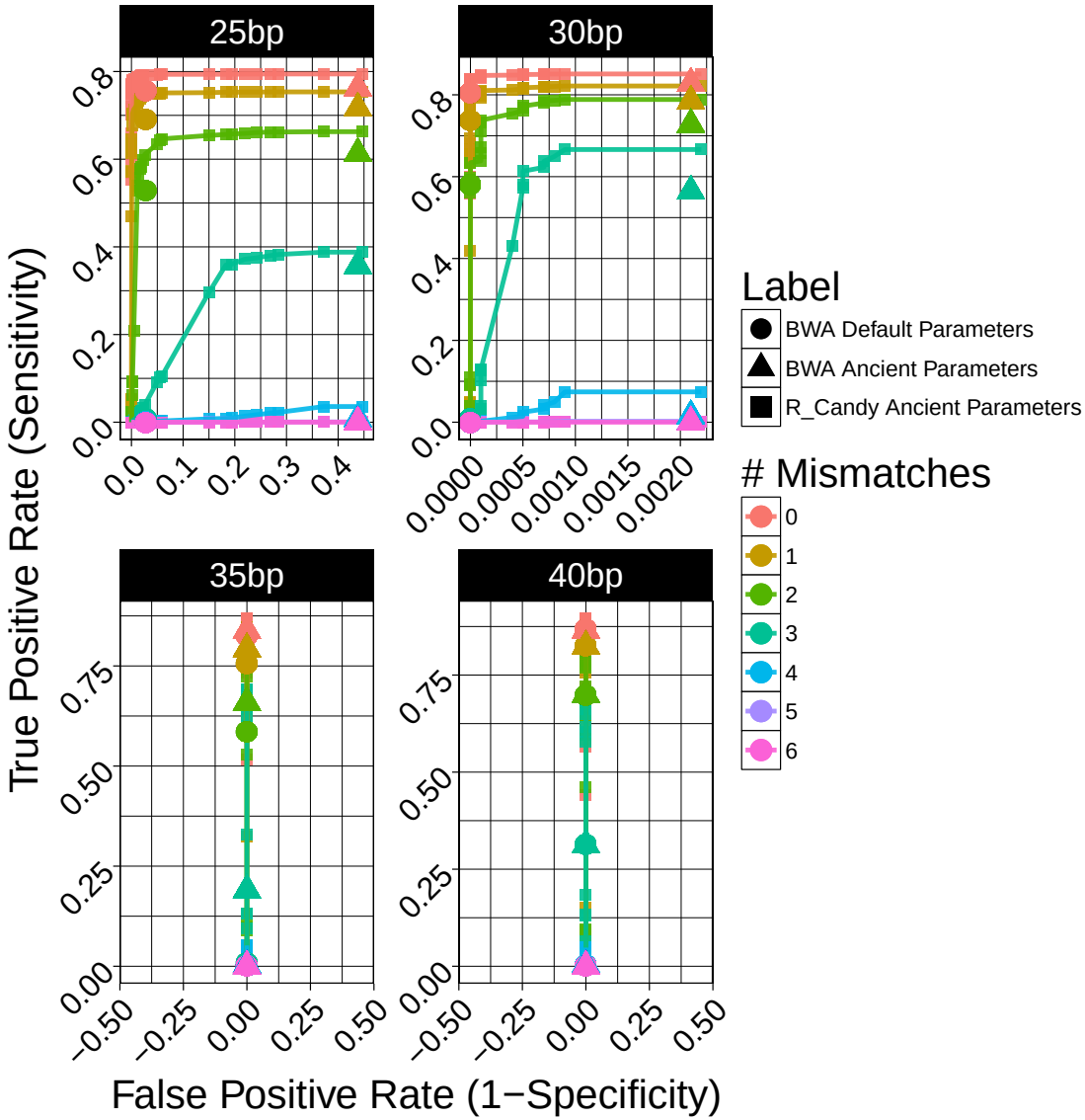
Modern DNA Human Ref Genome, Ancient Parameters



Ancient DNA Simulated Genome, Ancient Parameters

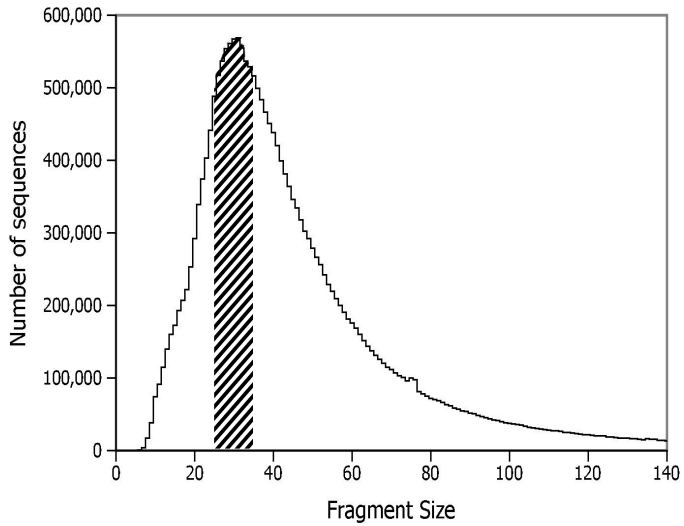


Ancient DNA Human Ref Genome, Ancient Parameters



Conclusion

- R-Candy reduces the minimum read length that can be used for ancient DNA alignment from 35 bp to about 25 bp.



- Can be used to align both ancient and modern DNA.
- Memory efficient (4 GB).
- Very low throughput rate (speed) therefore unusable in the case of big-data.

Source: Meyer M, et al. Nature 2014

Future Work

- Different search algorithm starting from the middle part of the reads.
- Require a different index structure (Bi-directional Wavelet Tree).
- Enable a seed strategy where the middle part of a read would serve as seed.
- Dynamic programming with a Full-Text index might extend the usefulness of R-Candy to longer reads.

Acknowledgement

Udo Stenzel

Janet Kelso

Availability

R-Candy: https://bitbucket.org/ustenzel/r-candy.

readSim:(genome and read simulator)
https://github.com/Homa1127/simulateGenome.git.

R-Candy's Speed Performance

Type	Read length	Speed BWA default (reads/s)	Speed BWA ancient (reads/s)	Speed R-Candy ancient (reads/s)
Genomic	25	222	34	1.79
Genomic	30	526	52	2.22
Genomic	35	625	434	2.26
Genomic	40	500	357	2.45
exogenous	25	147	13	1.68
exogenous	30	217	42	2.19
exogenous	35	232	153	1.67
exogenous	40	178	144	3.10

The alignment speed for ancient reads aligned to the human reference genome.

R-Candy's Memory Usage

Type	Read length	BWA default memory usage (MB)	BWA ancient memory usage (MB)	R-Candy memory usage (MB)
Genomic	25	945	947	1181
Genomic	30	945	949	1182
Genomic	35	945	945	1183
Genomic	40	945	945	1181
exogenous	25	945	947	814
exogenous	30	945	947	815
exogenous	35	945	945	825
exogenous	40	945	945	828

The alignment speed for ancient reads aligned to the human reference genome.

Scoring Matrix for ancient DNA

$$\begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon+p_G-4\epsilon p_G & \epsilon \\ \epsilon & 1-3\epsilon-p_C+4\epsilon p_C & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon-p_G+4\epsilon p_G & \epsilon \\ \epsilon & \epsilon+p_C-4\epsilon p_C & \epsilon & 1-3\epsilon \end{pmatrix}$$