

Making the Most of Ancient DNA

Homa Amini

Max-Planck Institute for Evolutionary Anthropology
Leipzig, Germany

October 27, 2016



MAX-PLANCK-GESELLSCHAFT



UPPSALA
UNIVERSITET



MAX-PLANCK-GESELLSCHAFT

Why Study Ancient DNA



UPPSALA
UNIVERSITET

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



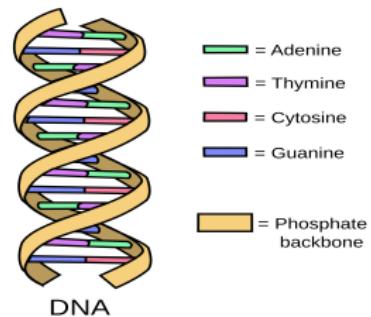
Source: www.sciencemag.org

What is a DNA Sequence

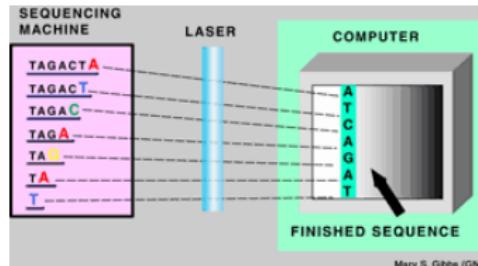
- Let X be a sequence of λ random variables.

$$X = \{X_1, X_2, \dots, X_\lambda\}$$

which takes the values from the alphabet $\{A, C, G, T\}$.



- e.g. ACGGCTAACAGTCAGATACAGCTCGCA

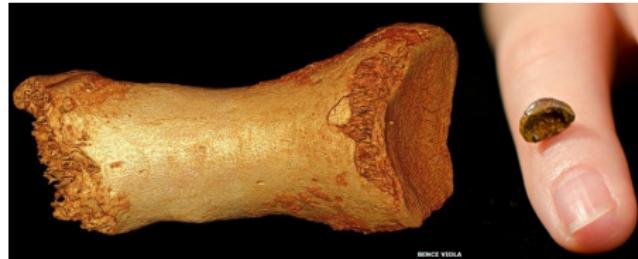


Mary S. Gibbs (GNN)

Source: medivizor.com

<http://www.genomenewsnetwork.org>

What is Ancient DNA



Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

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

- e.g. ATTGTTACAGATATT

Characteristics

- Hard to retrieve.
- Short molecule ≤ 50 bp.
- *Post-Mortem* damage: deamination (substitutions of C→T).
- Microbial contamination.

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

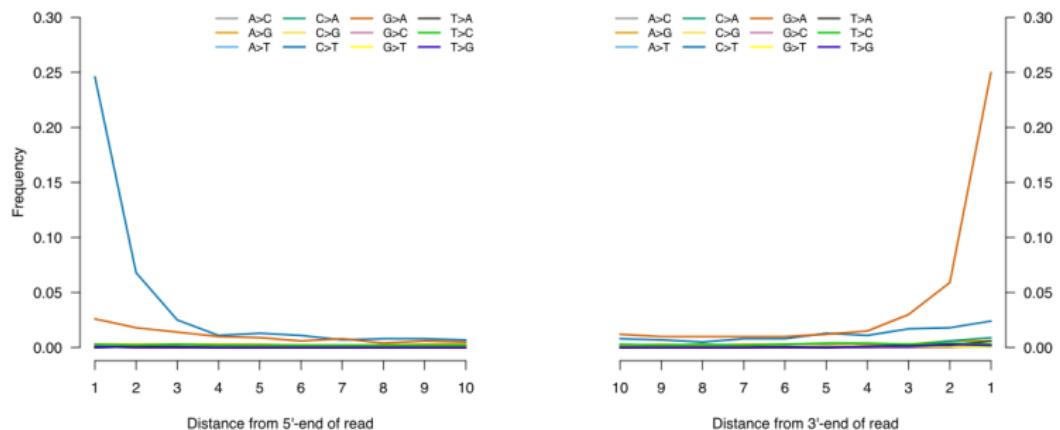
Post-Mortem Deamination Damage

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



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>

Sequencer

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

- A sequencer is a machine used to determine the order of DNA bases.



Source: <http://www.illumina.com>



MÄX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

DNA Similarity



UPPSALA
UNIVERSITET

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

- Change of a letter

ACGCTATGCA
ACC**TATGCA**

- Insertion of a letter

ACGCTATGC-A
ACGCTATGC**TA**

- Deletion of a letter

ACGC**TATGCA**
ACGC**T-TGCA**

- A few mutations make sequences different, but still “similar” and we are looking for these similarities.



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

DNA Sequence Alignment



UPPSALA
UNIVERSITET

- 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**



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

What Sort of Alignment?



UPPSALA
UNIVERSITET

■ Global alignment:

S_1 : ACCGTCGCTACTGCTGT CAGATCGCTCATCGCATACTGTCT
 S_2 : ACCGTGTA CAGATCGCTCATC~~ATCGC~~AGTCGATAGCCTGTCT

■ Local alignment:

S_1 : ACCGTCGCTACTGCTGTC~~AGATCGCTCA~~GTTCGATCTG
 S_2 : GTCTGTCAG~~AGATCGCTCA~~TCGCATACTGTCTCGTCC

■ Semi-Global alignment:

S_1 : ACCGTCGCTACTGCTGTC~~AGATCGCTCA~~GTTCGCTGTCAAGTCACT
 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 $g = -1$

$$\begin{array}{l} S_1 : \quad C \quad C \quad G \quad A \quad - \quad T \quad A \\ S_2 : \quad T \quad C \quad G \quad - \quad C \quad T \quad A \end{array}$$

we score it by:

$$\begin{aligned} S(C, T) + S(C, C) + S(G, G) + 2g + S(T, T) + S(A, A) \\ -1 + 1 + 1 - 1 - 1 + 1 + 1 = 1 \end{aligned}$$



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

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 best alignment of the two sequences.
 - Time complexity $\mathcal{O}(n^2)$.



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

Overview of BWA Aligner



UPPSALA
UNIVERSITET

- 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.



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

Why is a New Aligner Needed?

■ Alignment 1:

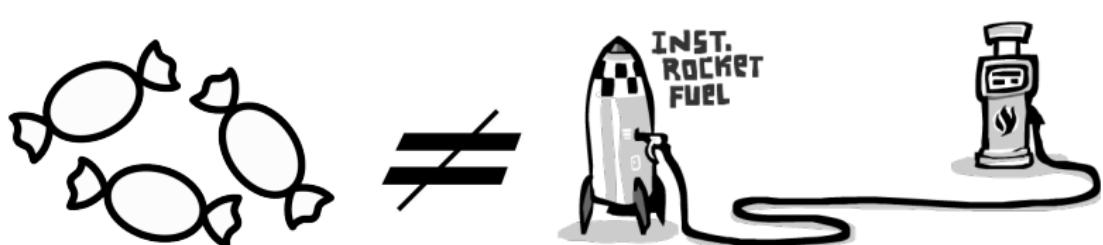
A**CCC****A****C****T****C****T****A****C****C****T****C****A****T****C****G****A****C****C**
A**T****T****T****A****C****T****C****T****A****C****C****T****C****A****T****C****G****A****T****T**

■ Alignment 2:

A**CC****C****A****C****T****C****T****A****C****C****T****C****A****T****C****G****A****C****C**
C**CC****C****T****G****T****T****A****C****C****A****C****T****C****G****A****C****C**

R-Candy: New Aligner for Ancient DNA

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



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



MAX-PLANCK-GESSELLSCHAFT



UPPSALA
UNIVERSITET

Overview of R-Candy's Algorithm

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

- 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.



MÄX-PLANCK-GESELLSCHAFT

Test Data



UPPSALA
UNIVERSITET

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

- To compare sensitivity of R-Candy to BWA aligner.
 - Simulated modern and ancient DNA reads originated from a simulated genome.
- To roll 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 as stands for exogenous DNA.

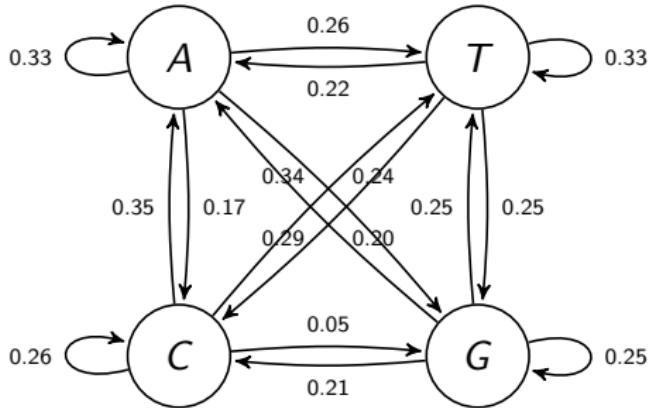
Genome Simulation

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

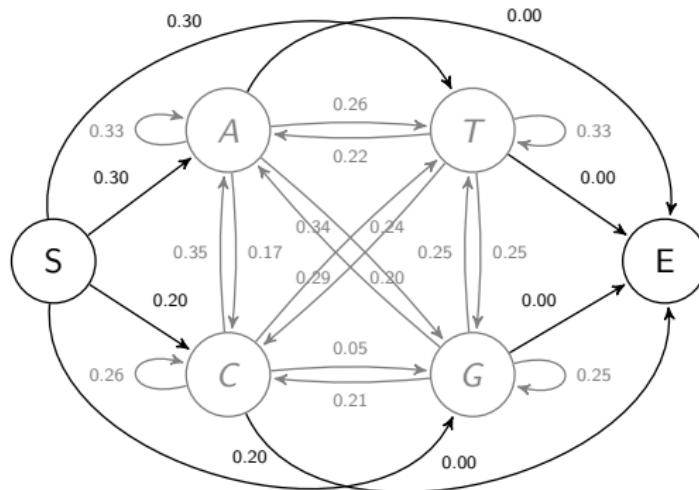
Evaluation of
the New
Approach

Conclusion &
Future Work



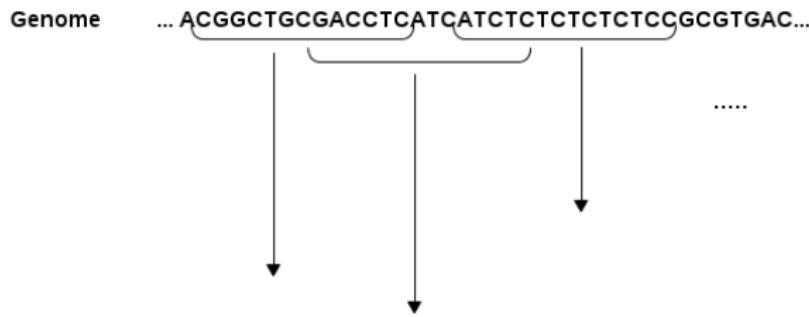
Trained 1st order Markov chain based on the human reference genome.

1st Order Markov Chain



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

Read Simulation (Cont.)

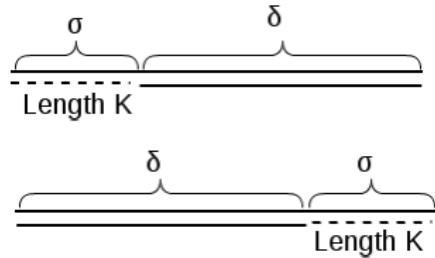
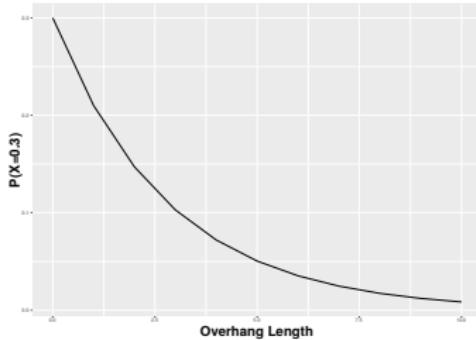


Reads.sam

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

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.)

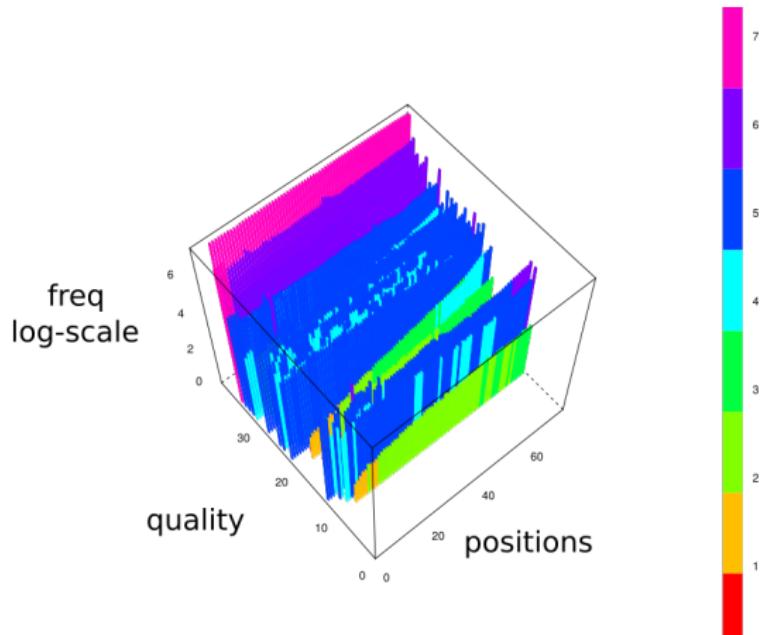
Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

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





MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

Read Simulation (Cont.)

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

Evaluation Criteria

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

■ Mapping accuracy:

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

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



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



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the DataCommon
Approaches
to String
AlignmentsEvaluation of
the New
ApproachConclusion &
Future WorkUPPSALA
UNIVERSITET

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/Fresh DNA

A = Ancient DNA/parameters

S = Simulated genome

R = Real genome

D = Default parameters

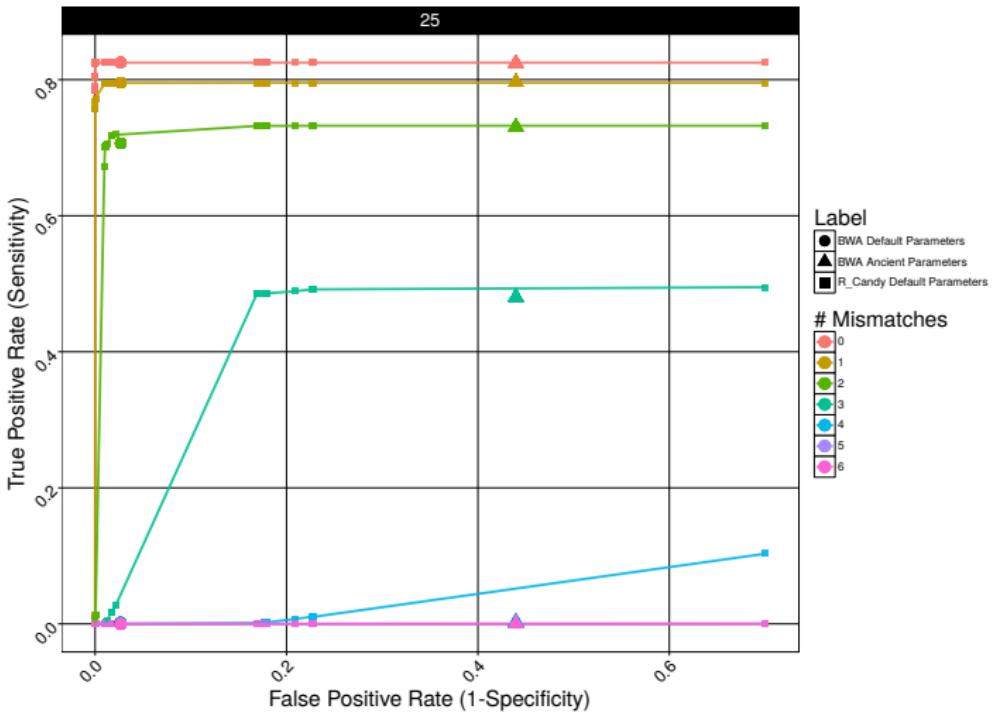
ROC Curve

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



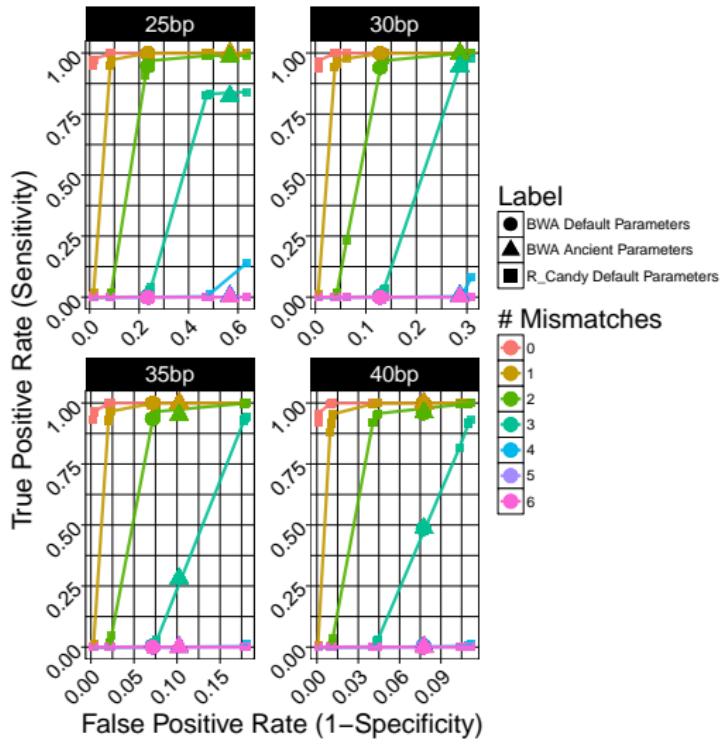
Modern DNA Simulated Genome, Default Parameters

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



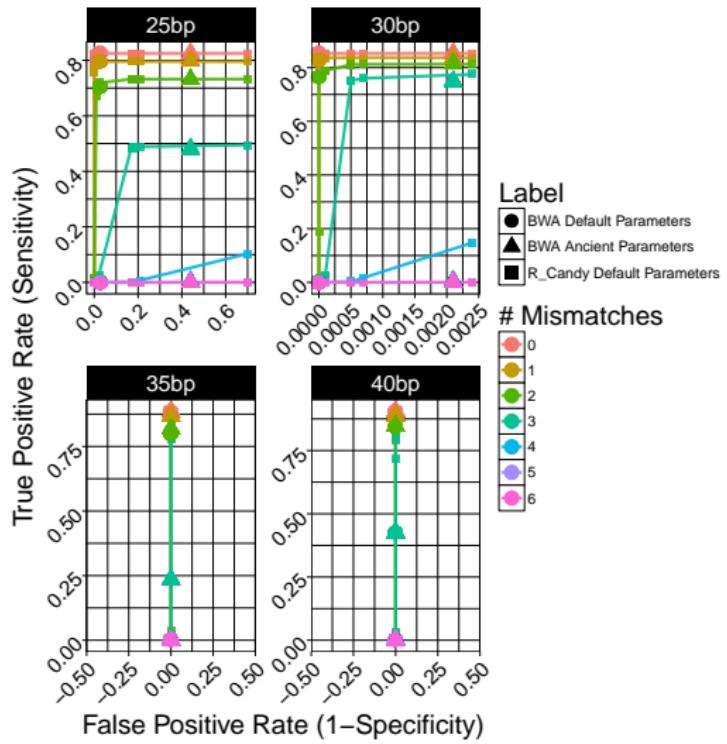
Modern DNA Human Ref Genome, Default Parameters

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



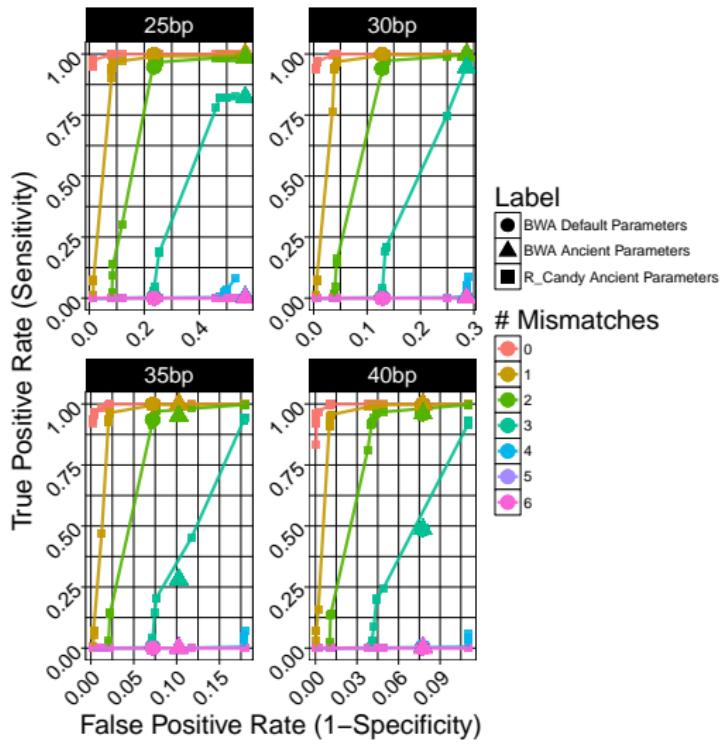
Modern DNA Simulated Genome, Ancient Parameters

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



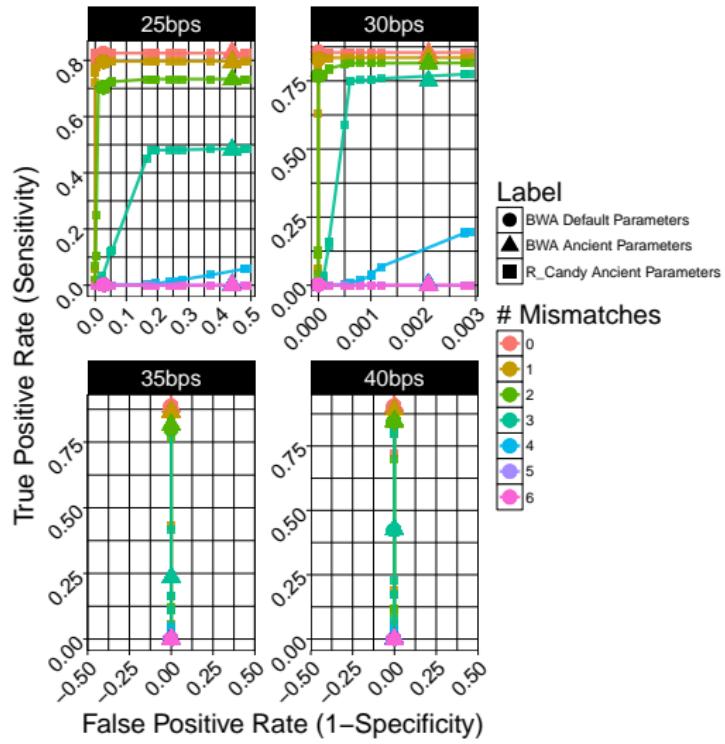
Modern DNA Human Ref Genome, Ancient Parameters

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



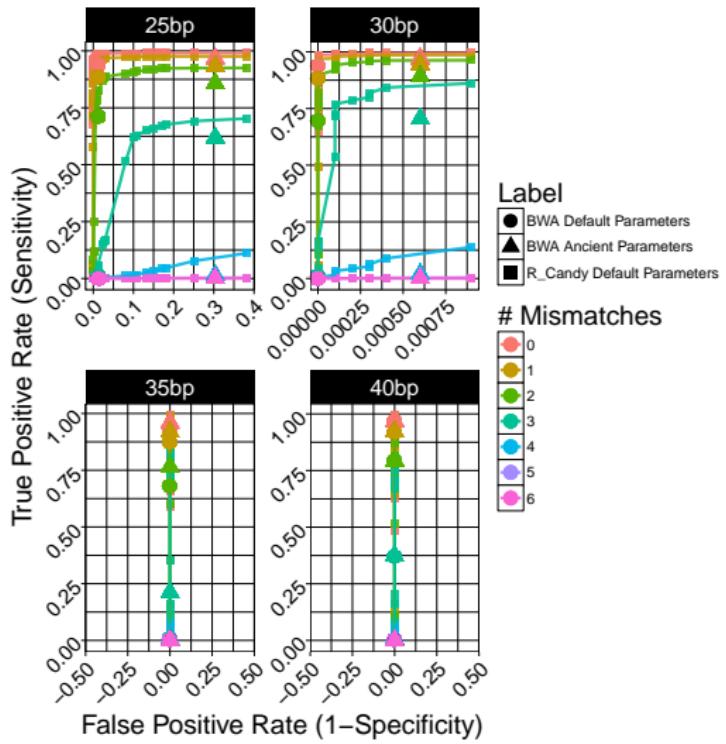
Ancient DNA Simulated Genome, Ancient Parameters

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

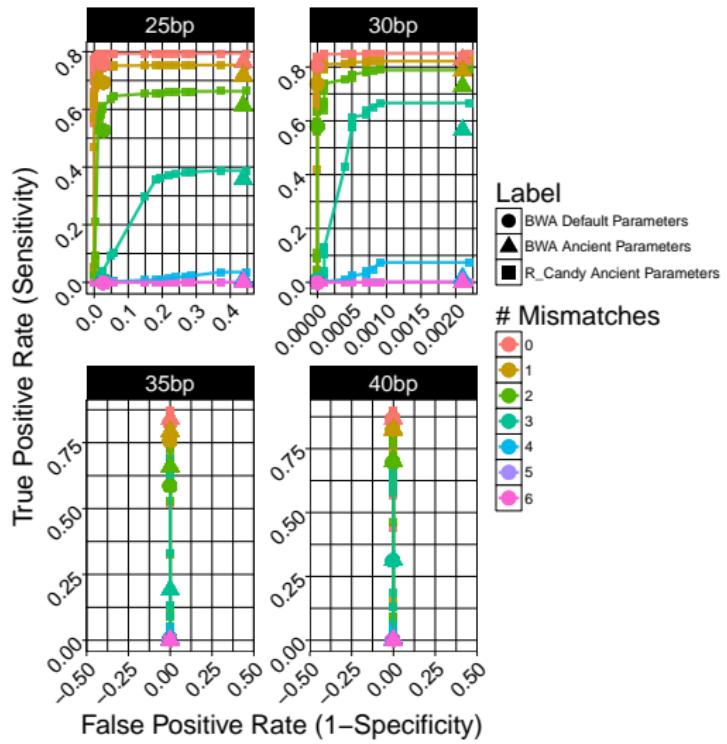


Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

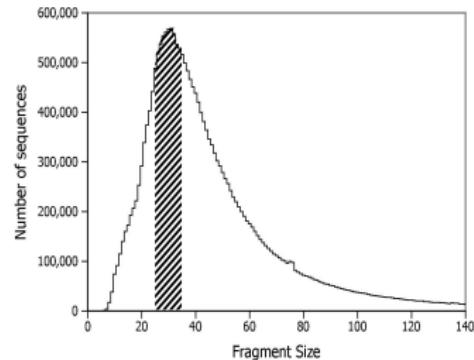
Evaluation of
the New
Approach

Conclusion &
Future Work



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 2013



MÄX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

Future Work

- Different search algorithm starting from the middle part of the reads.
- Requires 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.



MAX-PLANCK-GESSELLSCHAFT

Acknowledgement

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work



UPPSALA
UNIVERSITET



Udo Stenzel



Janet Kelso



MAX-PLANCK-GESELLSCHAFT

Availability



UPPSALA
UNIVERSITET

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

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

readSim:(genome and read simulator)

<https://github.com/Homa1127/simulateGenome.git>.



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the DataCommon
Approaches
to String
AlignmentsEvaluation of
the New
ApproachConclusion &
Future Work

R-Candy's Speed Performance

UPPSALA
UNIVERSITET

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.



MAX-PLANCK-GESELLSCHAFT

Specification
and
Properties of
the DataCommon
Approaches
to String
AlignmentsEvaluation of
the New
ApproachConclusion &
Future Work

R-Candy's Memory Usage

UPPSALA
UNIVERSITET

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.



MAX-PLANCK-GESELLSCHAFT



UPPSALA
UNIVERSITET

Scoring Matrix for ancient DNA

Specification
and
Properties of
the Data

Common
Approaches
to String
Alignments

Evaluation of
the New
Approach

Conclusion &
Future Work

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