UNIL | Université de Lausanne
Département de biologie computationnelle

Swiss Institute of Bioinformatics

Reviews in Quantitative Biology

Computational methods to analyze ancient DNA data

Anna-Sapfo Malaspinas
Department of Computational Biology
University of Lausanne

Structure of the talk/review

- Introduction to ancient DNA
  - Definition of aDNA
  - A brief history of the field
  - Characteristics
    - DNA degradation
    - Contamination
  - Standardized workflow in the lab
  - What is it used for?

- Computational methods
  - Map/assemble the data
  - Assess authenticity
  - Population genetics:
    - Infer demography
    - [Infer selection]
  - [Phylogenetics]
  - [Environmental (eDNA)/metagenomics]

- [Future directions]
  - [Wet lab developments]
  - [Computational developments]

Structure of the talk/review

- Introduction to ancient DNA
  - Definition of aDNA
  - A brief history of the field
  - Characteristics
    - DNA degradation
    - Contamination
  - Standardized workflow in the lab
  - What is it used for?

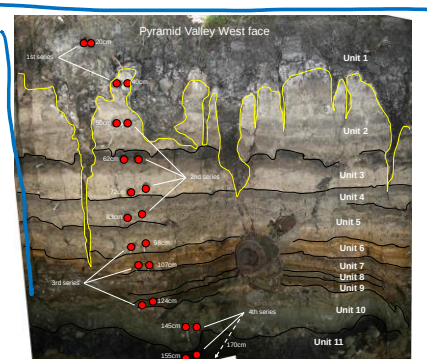
- Computational methods
  - Map/assemble the data
  - Assess authenticity
  - Population genetics:
    - Infer demography
    - [Infer selection]
  - [Phylogenetics]
  - [Environmental (eDNA)/metagenomics]

- [Future directions]
  - [Wet lab developments]
  - [Computational developments]

Please modify this structure as you see fit

Definition: DNA from “old stuff”

What is ancient DNA? DNA isolated from “ancient” specimens such as:



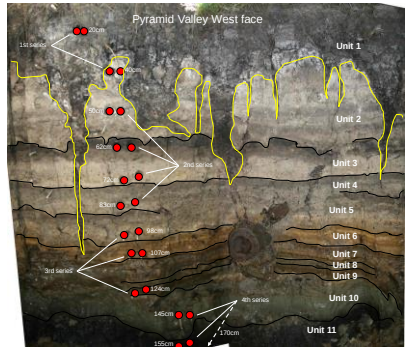
eDNA

Reviews:
1. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Pääbo, S. Ancient DNA. *Nat Rev Genet* **2**, 353–359 (2001).
2. Orlando, L. *et al.* Ancient DNA analysis. *Nat Rev Methods Primers* **1**, 1–26 (2021).
3. Slatkin, M. Statistical methods for analyzing ancient DNA from hominins. Current Opinion in Genetics & Development 41, 72–76 (2016).

Ancient DNA: DNA from “old stuff”

What is ancient DNA? DNA isolated from ancient specimens such as:



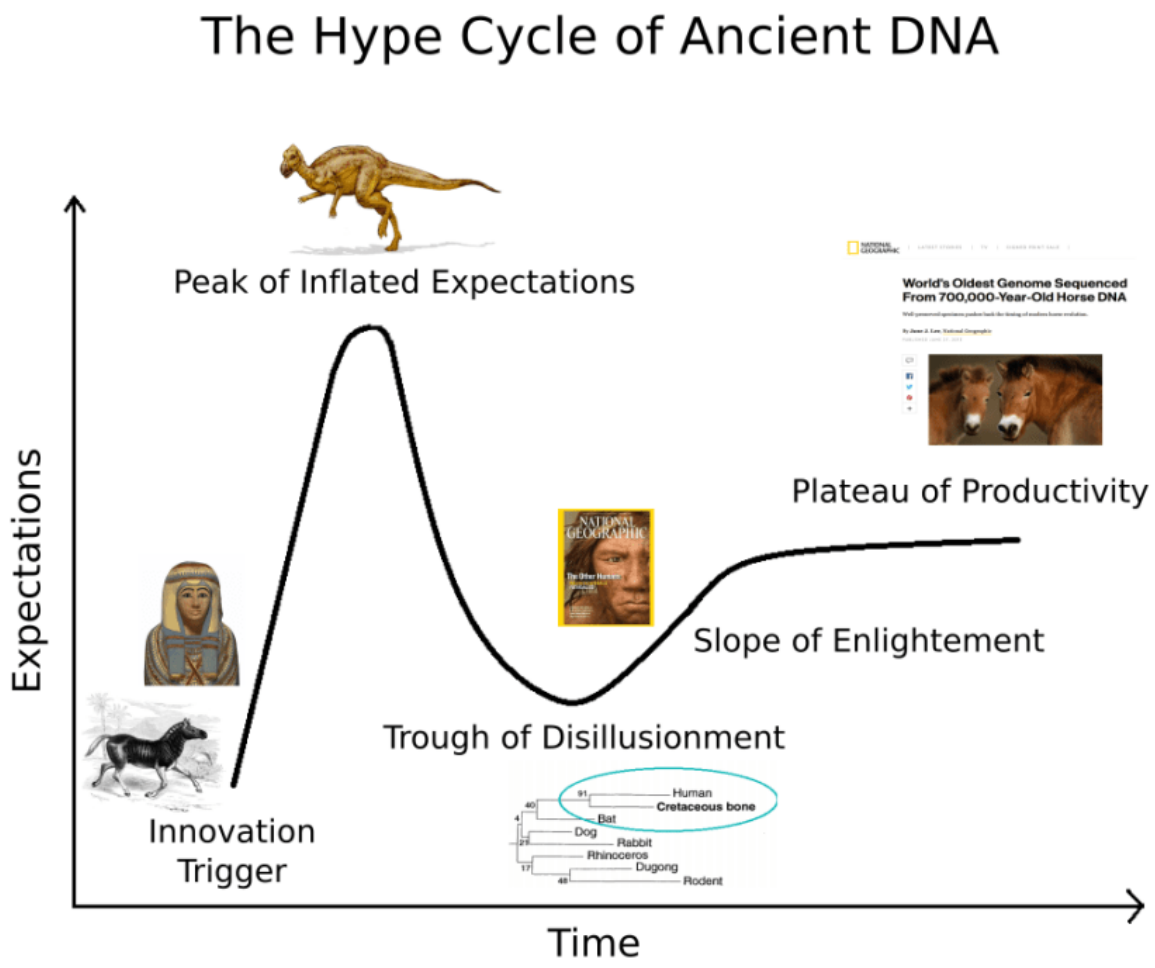
DNA can survive thousands of years in the remains of old organisms

After death: nucleases + microorganisms can freely operate and degrade the DNA

When “good” environmental conditions → some DNA left

A brief history “The Hype Cycle of Ancient DNA”

by Patrícia Chrzanová Pečnerová



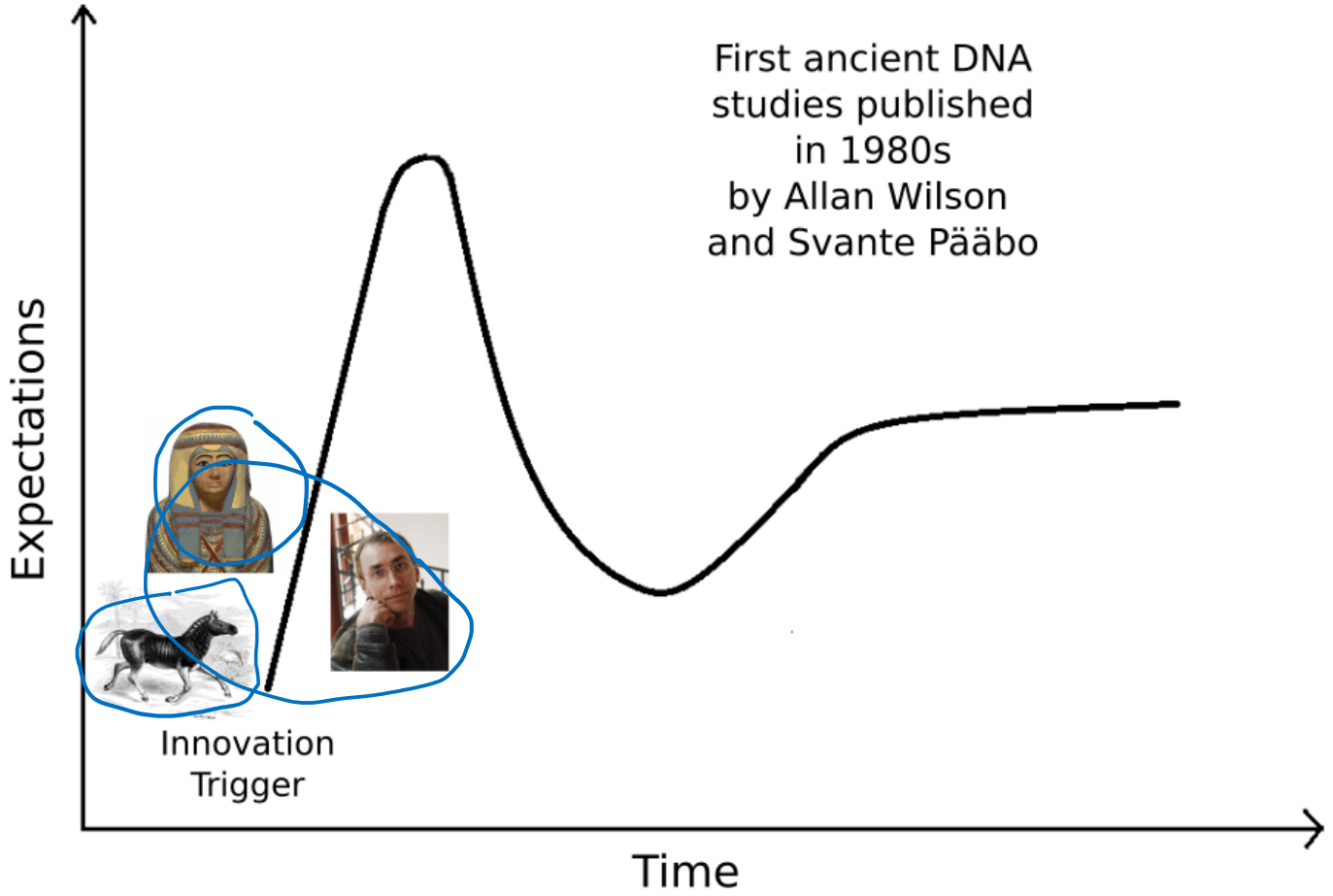
Hype cycle:

- “five phases of evolution of a **new** technology
- concentrating on the relationship between the hype* and real adoption of the technology”

*extravagant or intensive publicity or promotion.

The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová



Phase I: Innovation Trigger

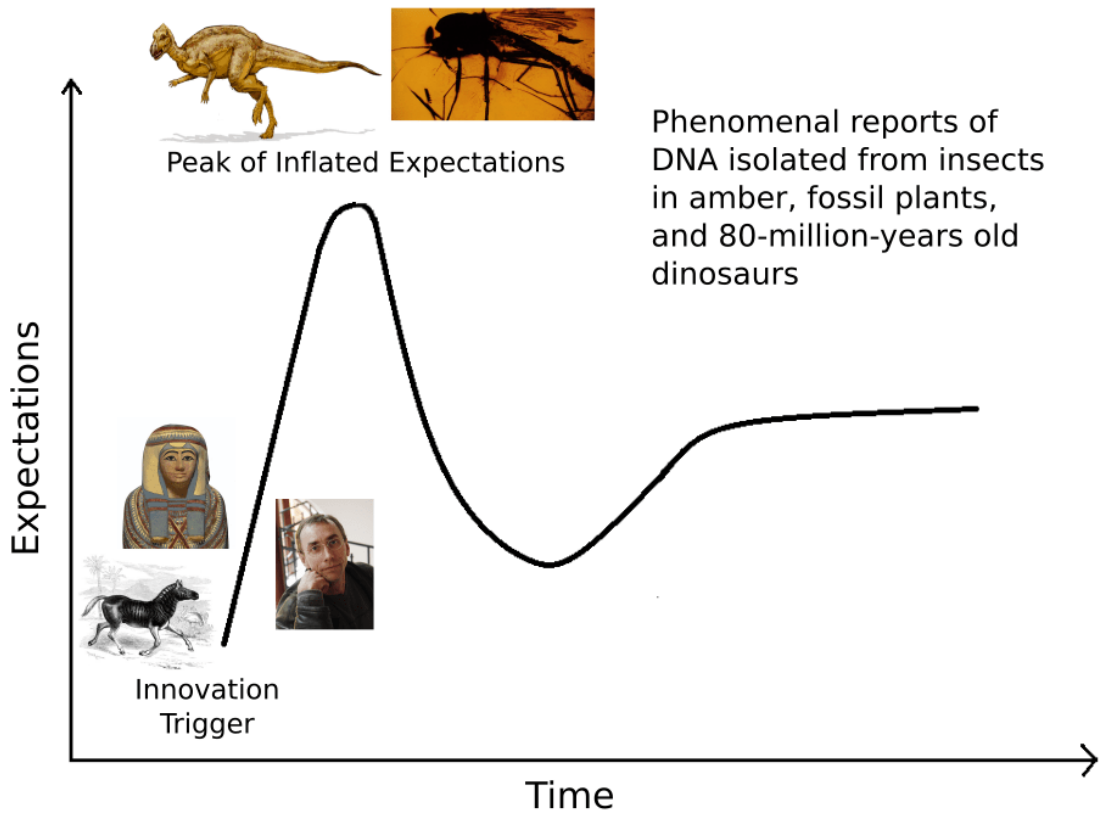
Quagga and Egyptian mummy

R. **Higuchi**, B. Bowman, M. Freiberger, O. A. Ryder, A. C. **Wilson**, *Nature*. **312**, 282–284 (1984).

S. **Pääbo**, *Nature*. **314**, 644–645 (1985).

The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová



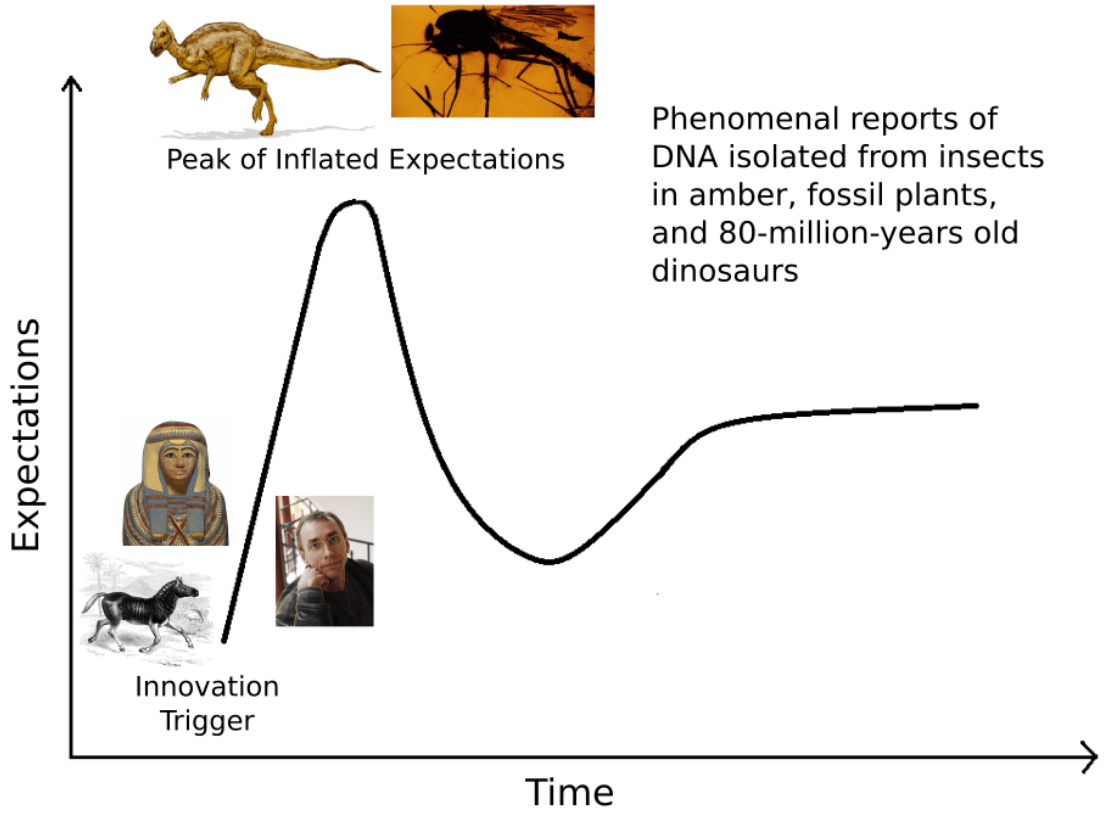
Phase II: Peak of Inflated Expectations

At first, some cool results:
Tasmanian wolf, moa, and even humans.

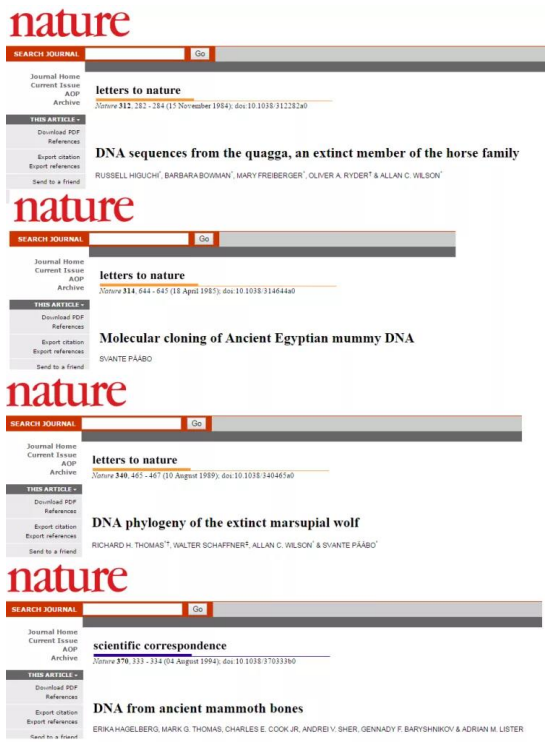
But then, “things got out of hands”.

The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová



Phase II: Peak of Inflated Expectations



“The downfall”

Science, 1994, Vol. 266, 1229-1232

DNA Sequence from Cretaceous Period Bone Fragments

Scott R. Woodward,* Nathan J. Weyand, Mark Bunnell

“The downfall”

Science, 1994, Vol. 266, 1229-1232

DNA Sequence from Cretaceous Period Bone Fragments

Scott R. Woodward,* Nathan J. Weyand, Mark Bunnell

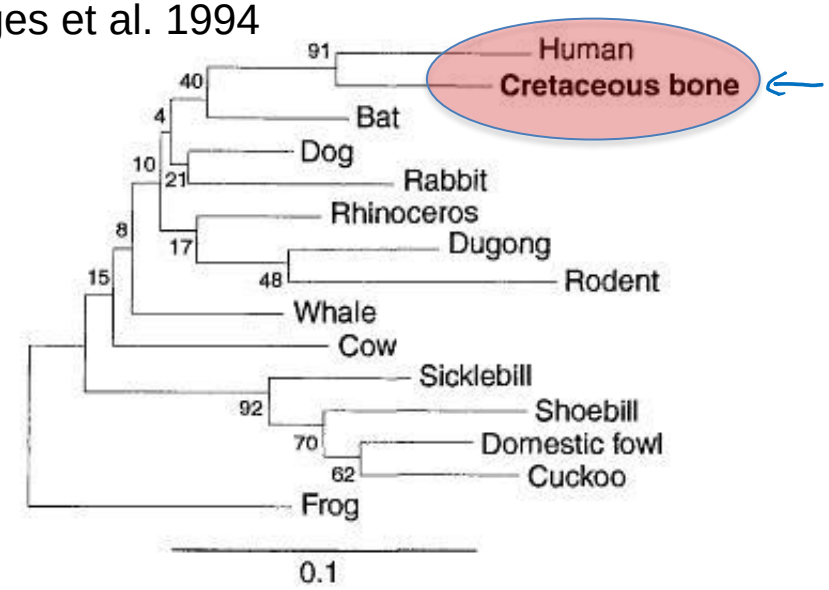
“The downfall”

Science, 1994, Vol. 266, 1229-1232

DNA Sequence from Cretaceous Period Bone Fragments

Scott R. Woodward,* Nathan J. Weyand, Mark Bunnell

Hedges et al. 1994



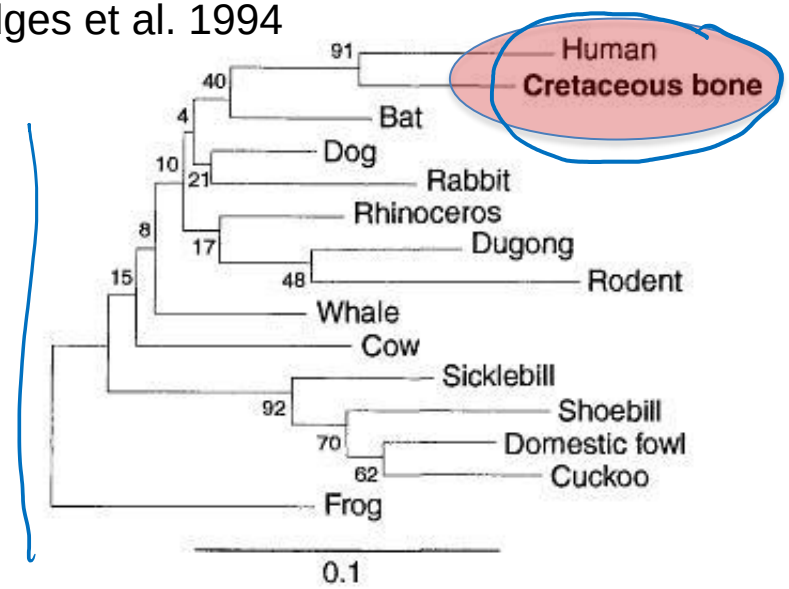
“The downfall”

Science, 1994, Vol. 266, 1229-1232

DNA Sequence from Cretaceous Period Bone Fragments

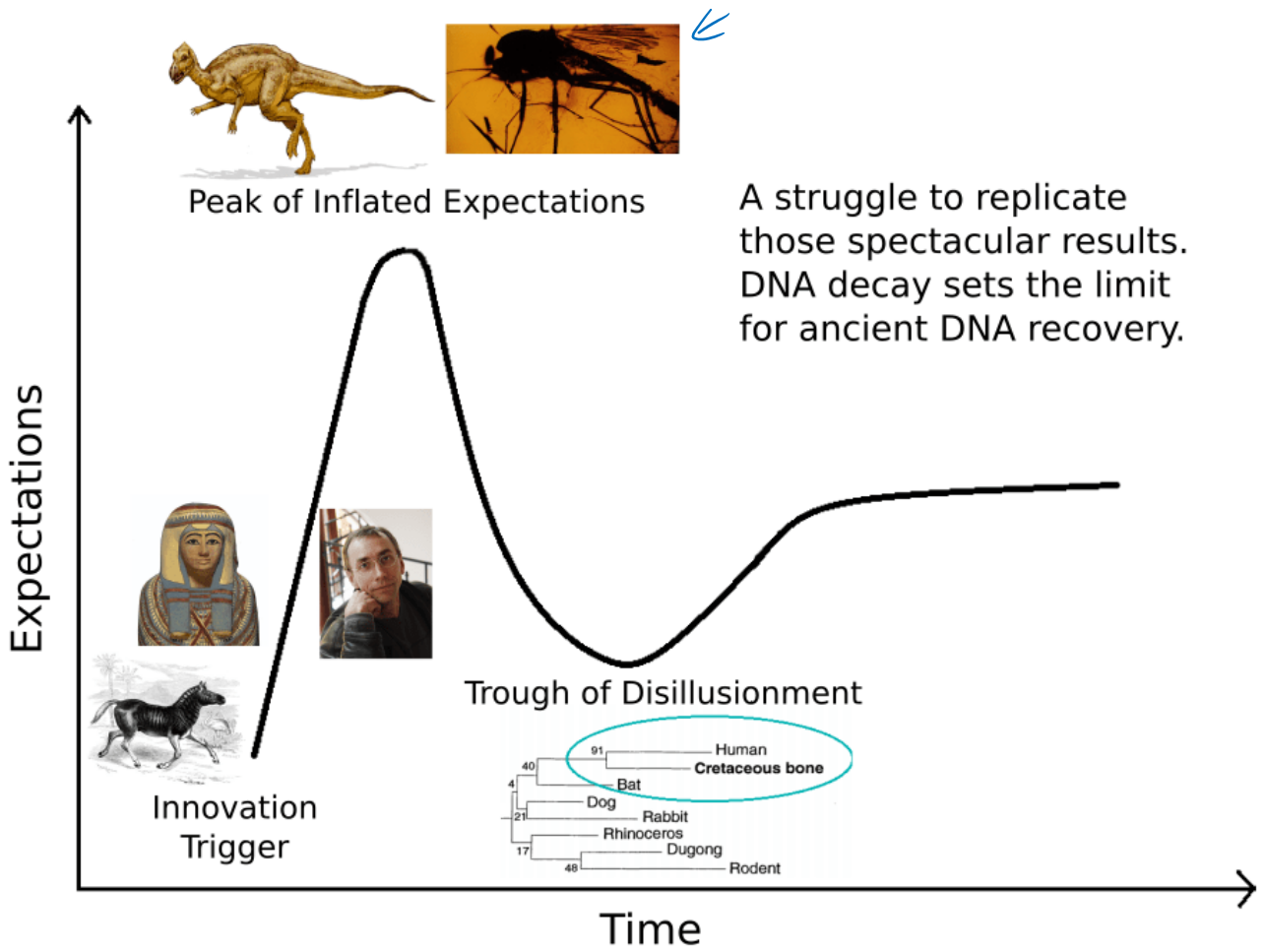
Scott R. Woodward,* Nathan J. Weyand, Mark Bunnell

Hedges et al. 1994



The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová

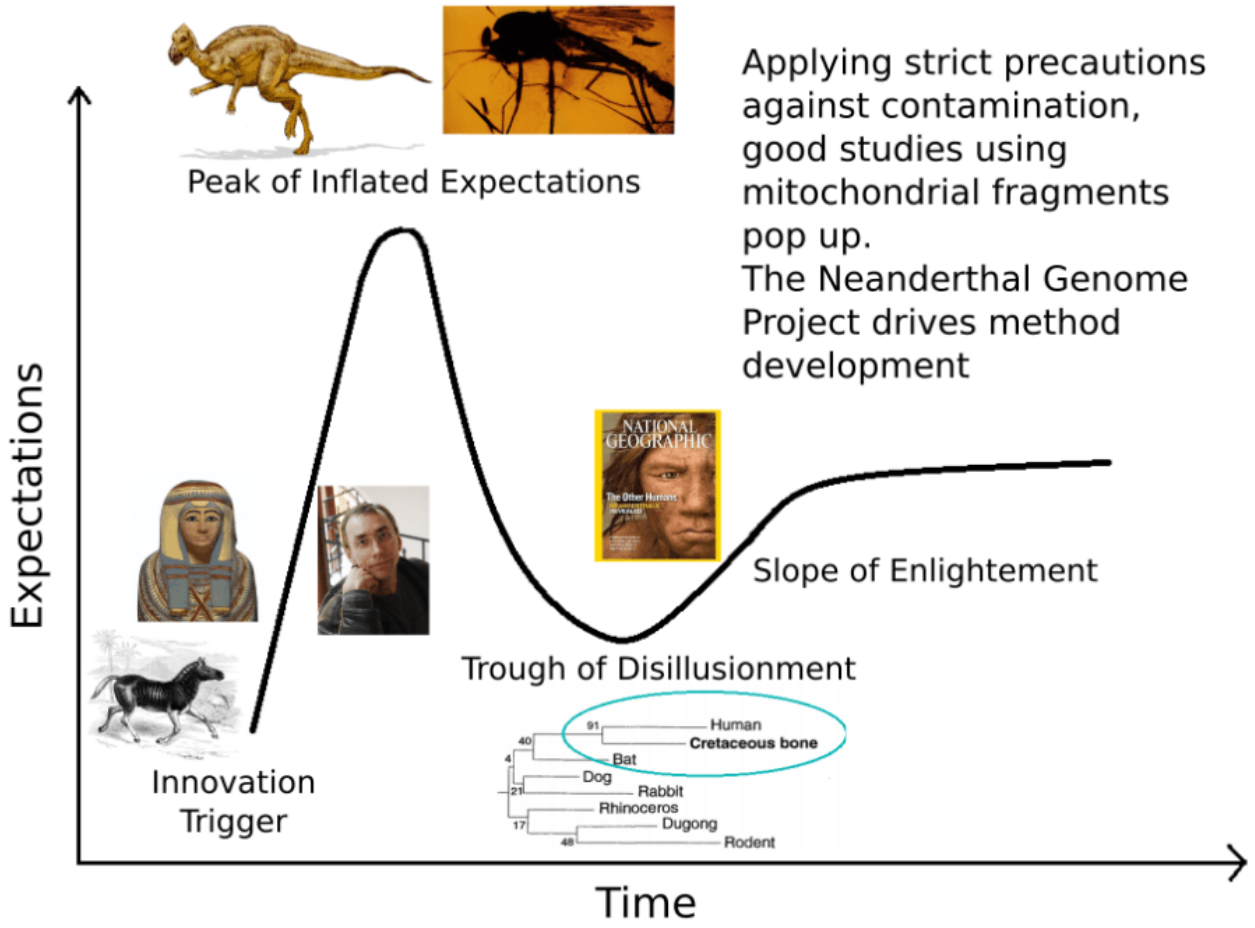


Phase III: Trough of Disillusionment

Struggle to replicate the results

The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová



Phase IV: Slope of Enlightenment

The period of “high profile contamination cases” is followed by

“do it right or not at all”

now standardized work flow and protocols

1. Cooper, A. & Poinar, H. N. Ancient DNA: Do It Right or Not at All. *Science* **289**, 1139–1139 (2000).

Standardized ancient DNA workflow

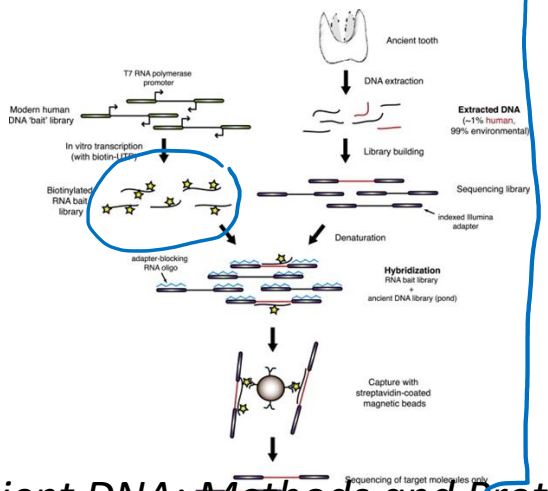
- Sample collection
- In an ancient DNA lab:
  - specific lab conditions
  - indexed library adapters
- DNA extraction/library build
- [Enrichment of specific regions]
- High-throughput sequencing

Sample collection

Lab work

Extraction, library build

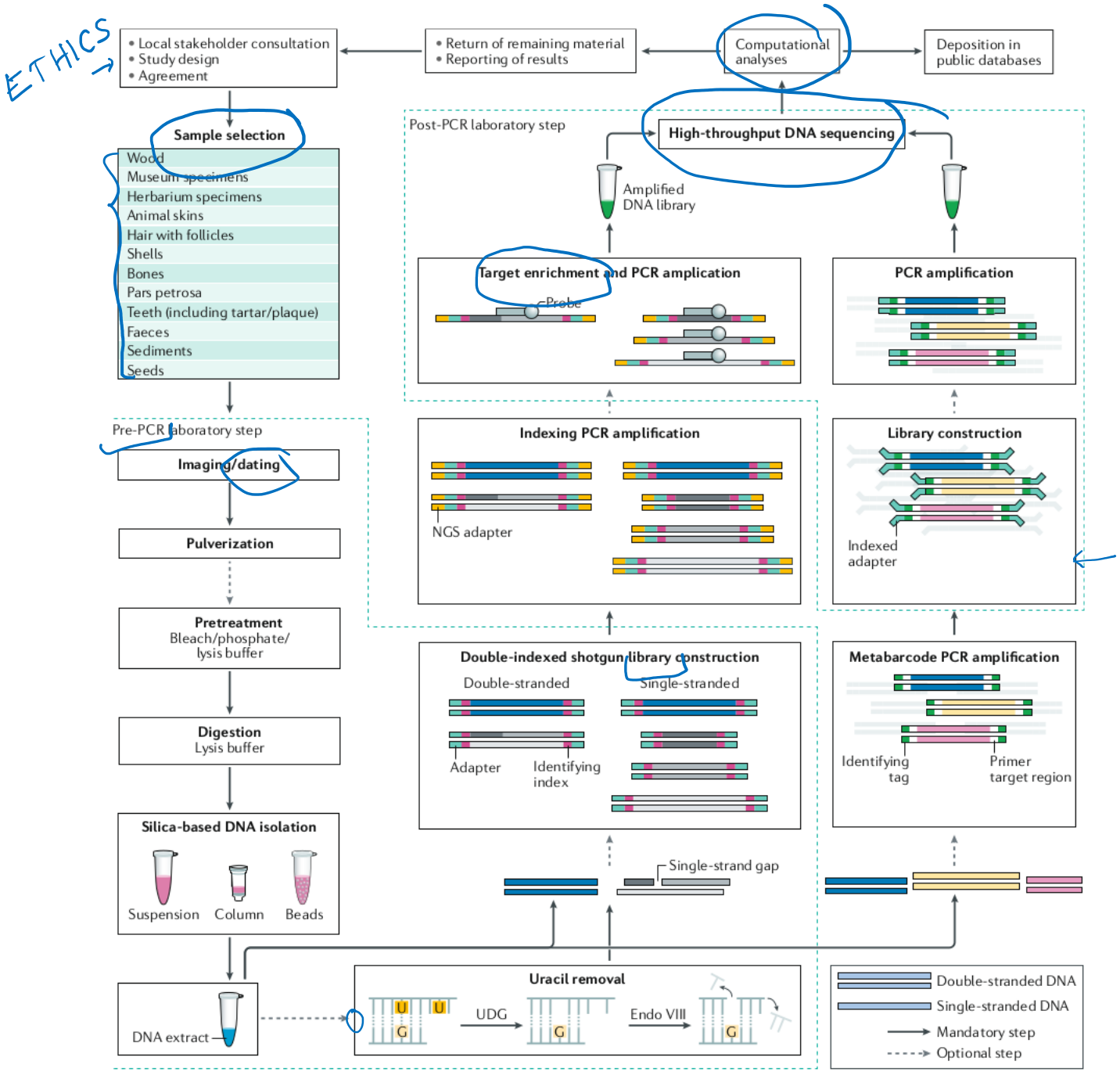
[Enrichmnent]



High-throughput sequencing
(in general Illumina)

Ancient DNA: Methods and Protocols. (Humana, 2019).
ISBN: 978-1-4939-9175-4

Standardized ancient DNA workflow

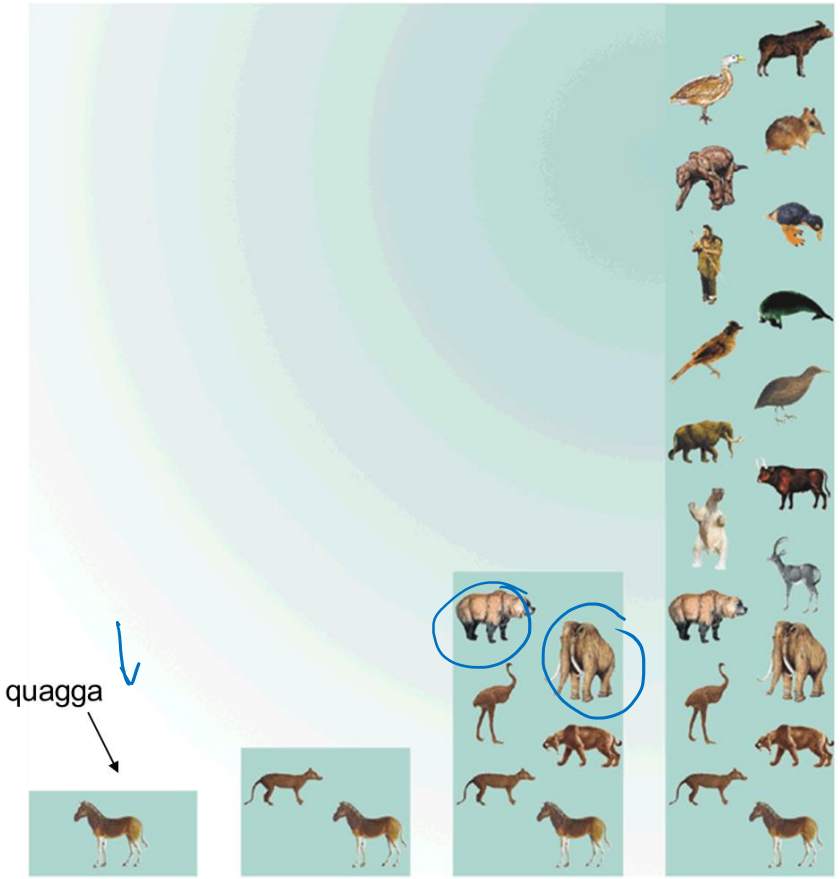


2. Orlando, L. *et al.* Ancient DNA analysis. *Nat Rev Methods Primers* **1**, 1–26 (2021).

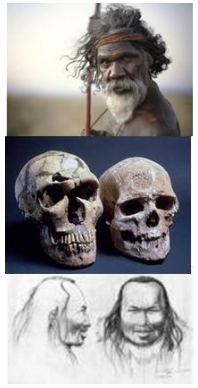
The Hype Cycle of Ancient DNA

by Patrícia Chrzanová Pečnerová

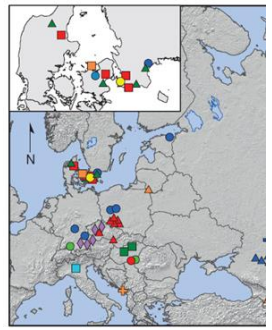
Phase V: "Plateau" of Productivity



Michael Hofreiter et al. Nature Reviews Genetics, 2001

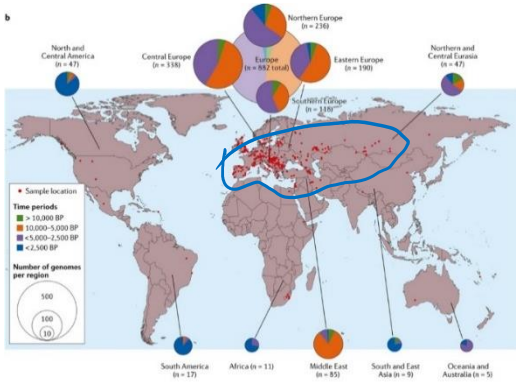


2010



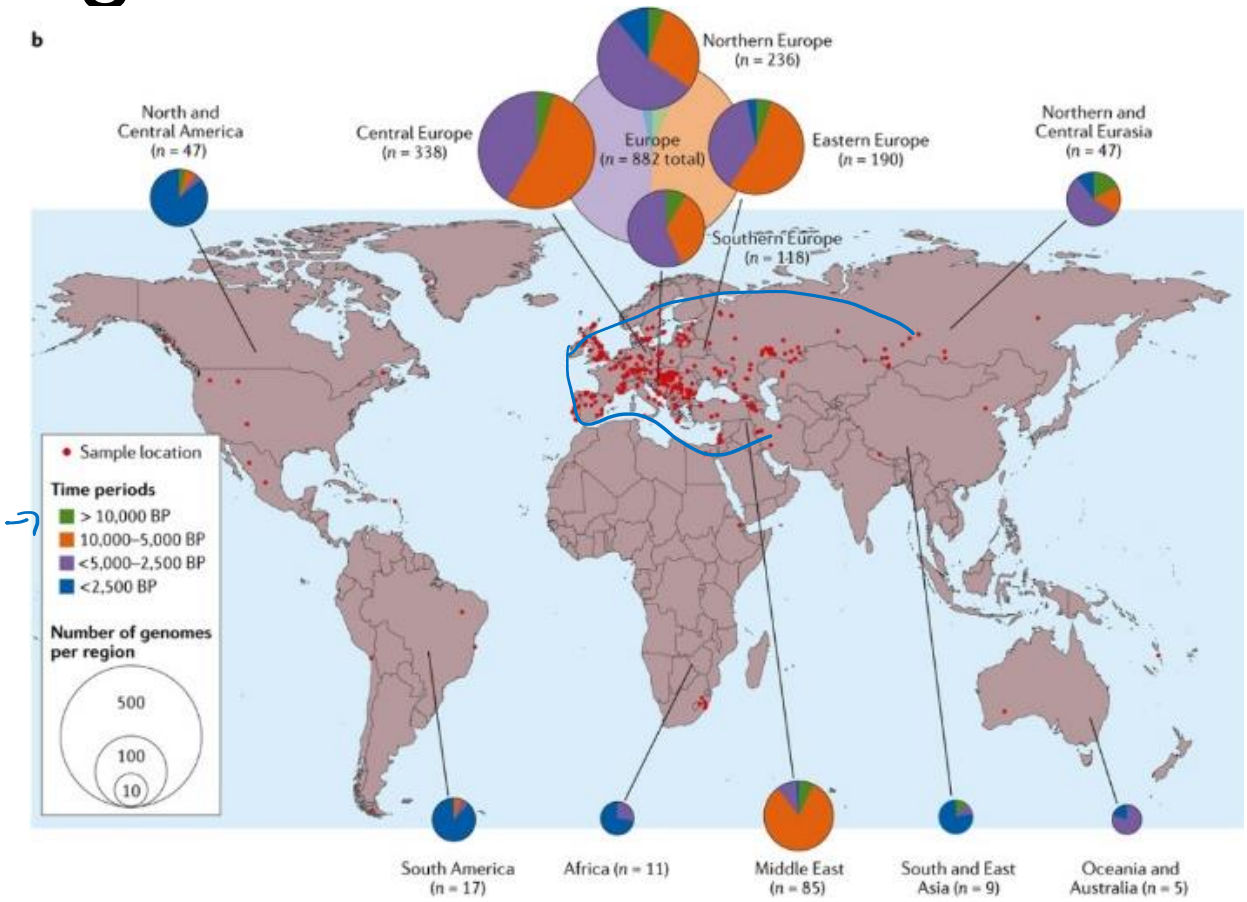
2015

"Paleogenomic"

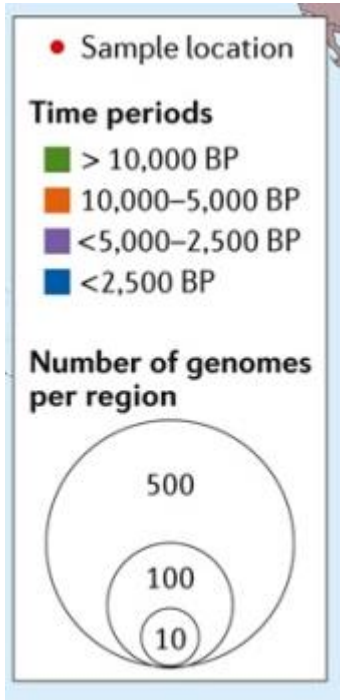


Marciniak, S. & Perry, G. H. (2017).

Effective protocols have allowed the sequencing of 1000s ancient human genomes from all over the world



And from all many time periods!



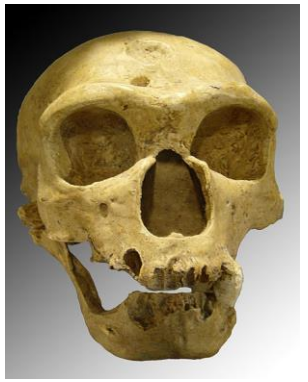
Ancient human genomes

Marciniak, S. & Perry, G. H. (2017).

What for?

E.g. Environmental DNA: getting clues about past environments

Studying extinct species



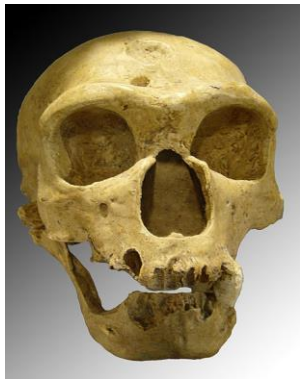
Characterizing the human past

Timing the spread of pathogens

What for?

E.g. Environmental DNA: getting clues about past environments

Studying extinct species



Characterizing the human past

Timing the spread of pathogens

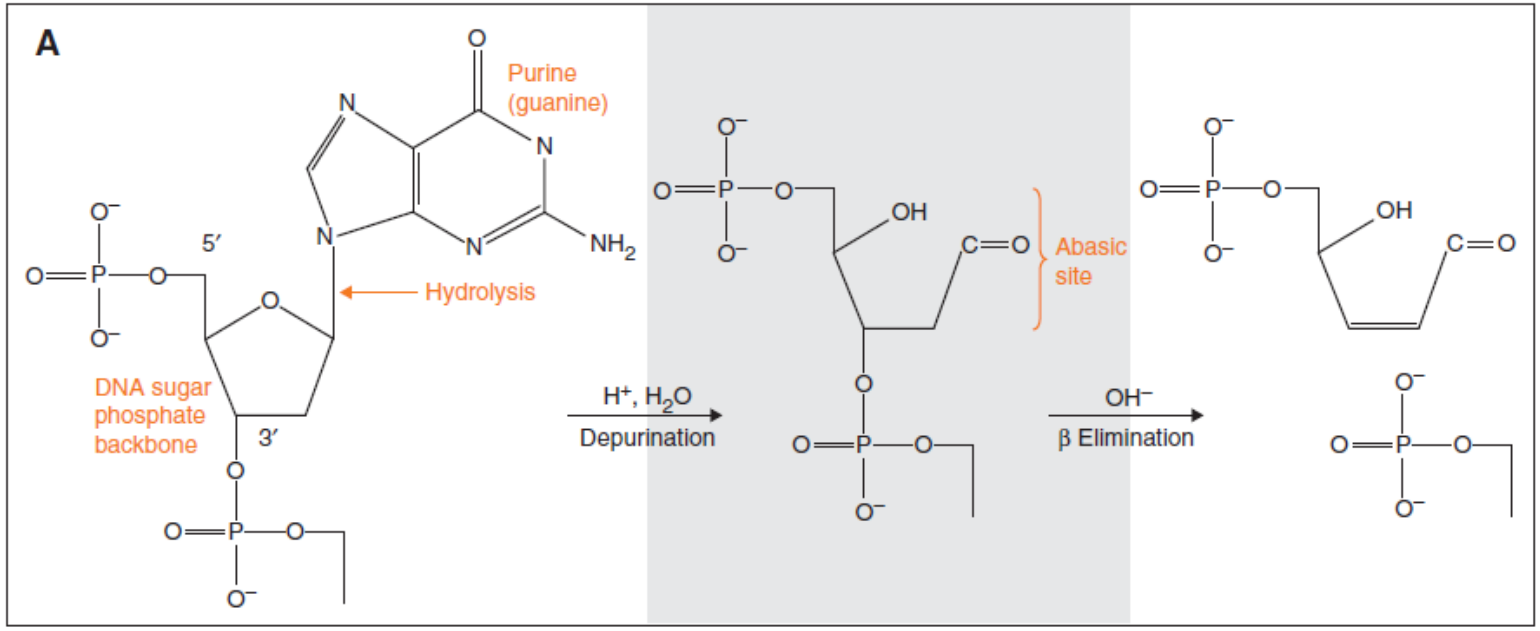
Molecular characteristics

aDNA

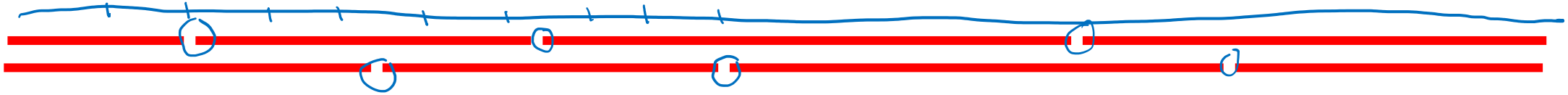
- **fragmentation** of the DNA
- DNA lesions **block the replication** of the DNA molecules by polymerases
- **damage:** incorrect nucleotide incorporated

post-mortem DNA fragmentation

Purines (G and As) are removed: depurination + beta – elimination



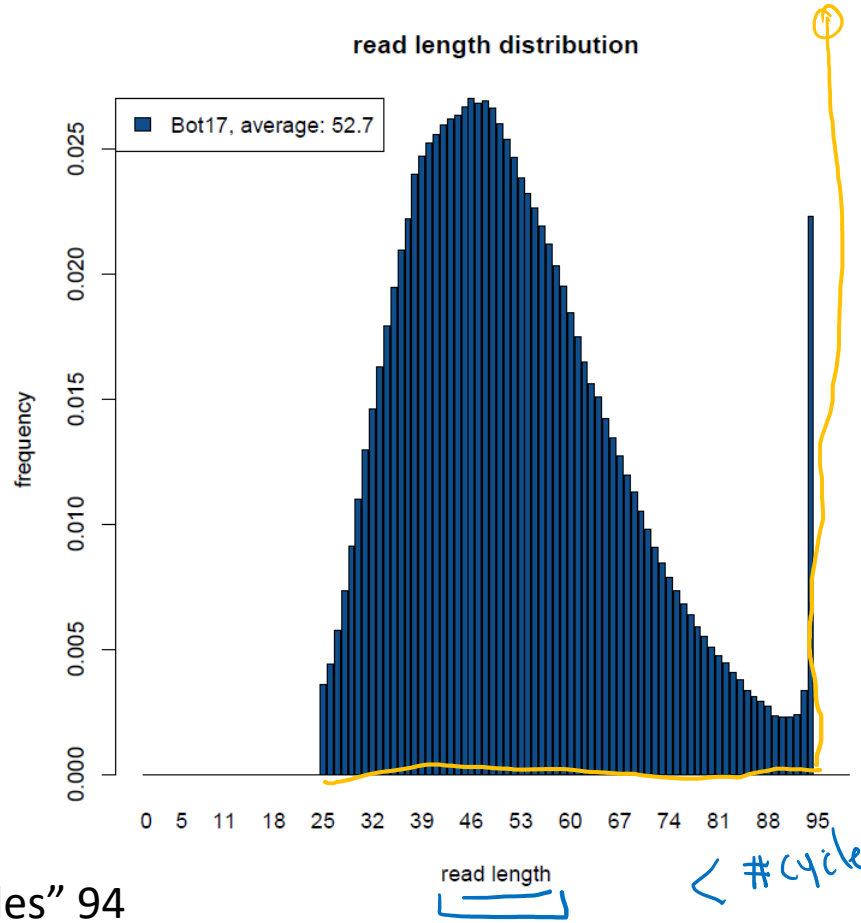
Dabney, M. Meyer, S. Pääbo, *Cold Spring Harb Perspect Biol*. **5**, a012567 (2013).



"DNA breaks"

post-mortem DNA fragmention

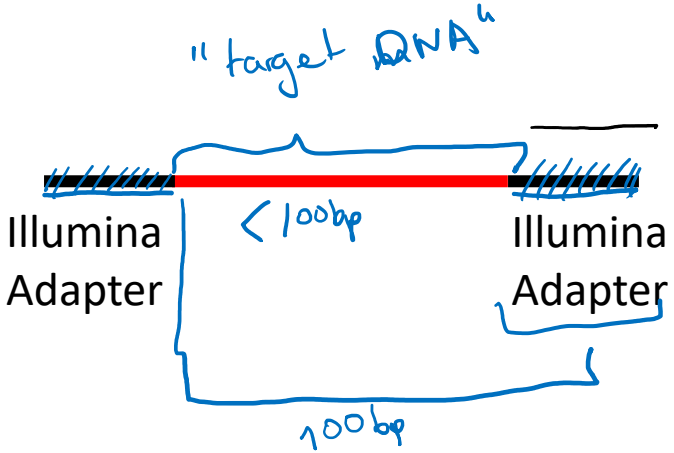
On average, reads shorter than the number of cycles used for an Illumina run



"#cycles" 94

< #cycles of sequencer

NB: If sequence of interest is short, you read into the adapters (AdapterRemoval*).



*S. Lindgreen, BMC Research Notes. 5, 337 (2012).
Schubert, M. et al. BMC Research Notes 9, 88 (2016).

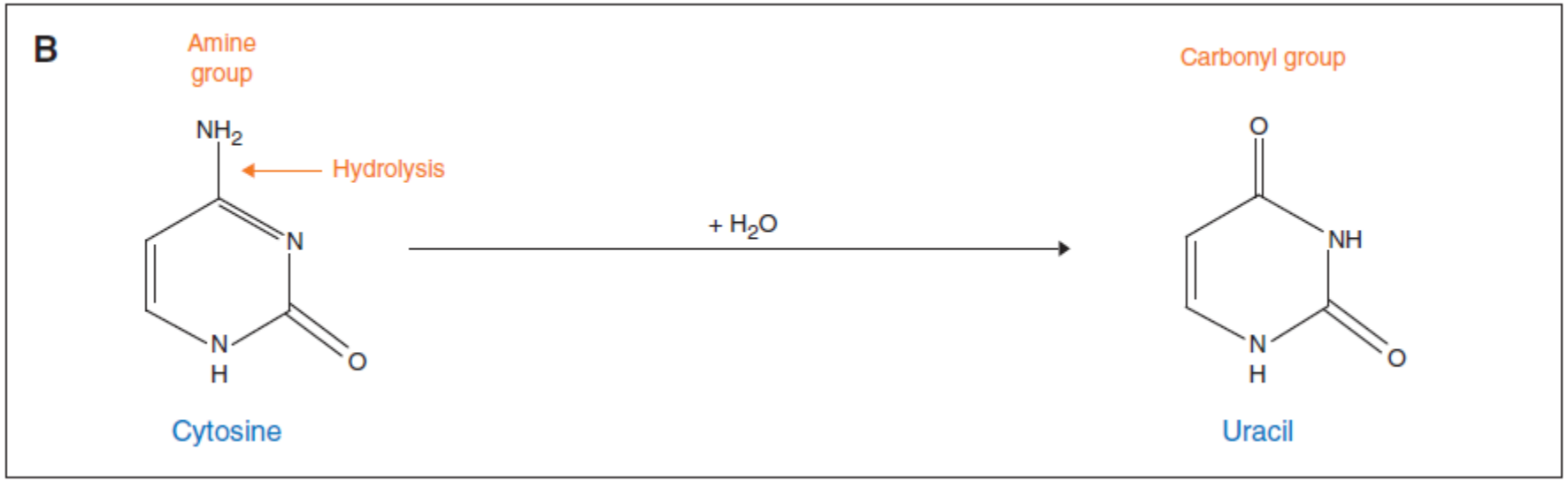
post-mortem DNA damage

DNA pairing

A = U

C ≡ G

C → U, especially at the **end of the molecules**



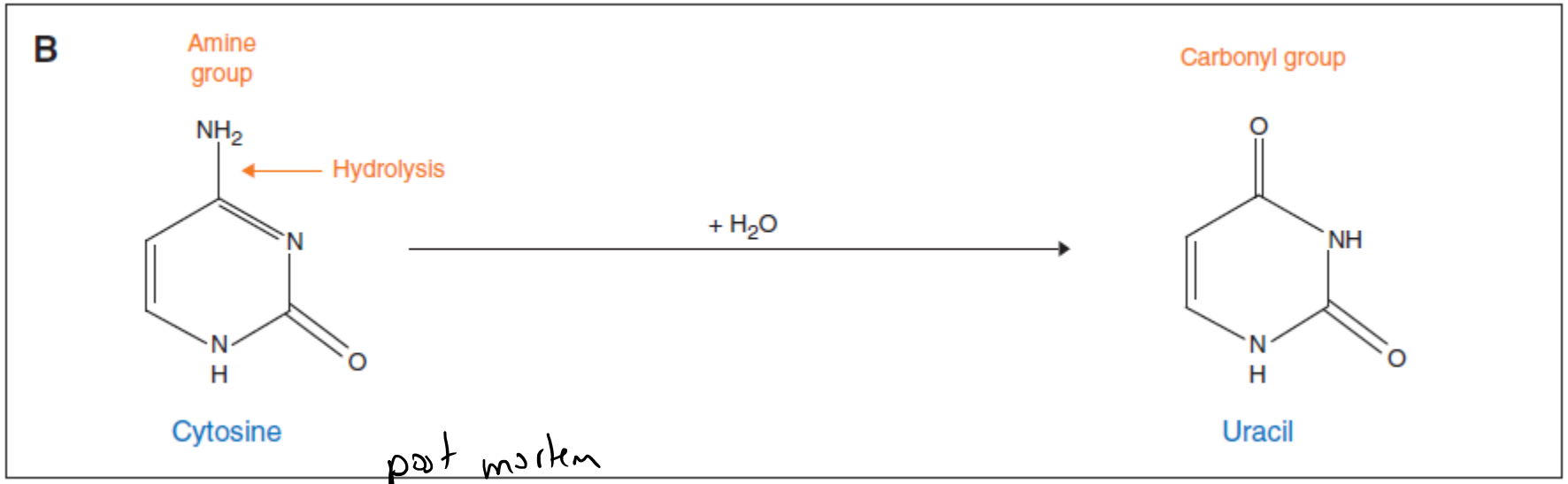
post-mortem DNA damage

DNA pairing

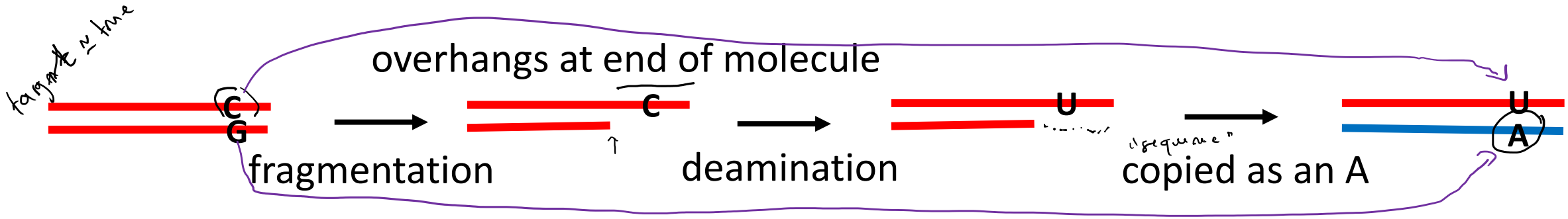
A = U

C ≡ G

C → U, especially at the **end of the molecules**



post mortem



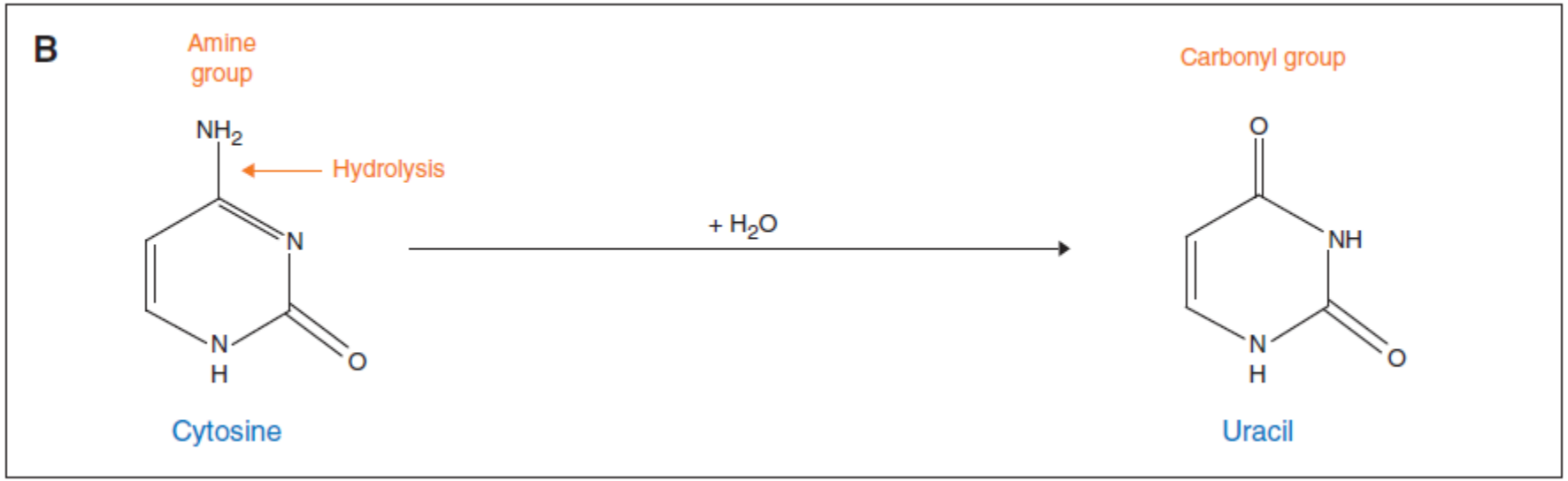
DNA pairing

A = U

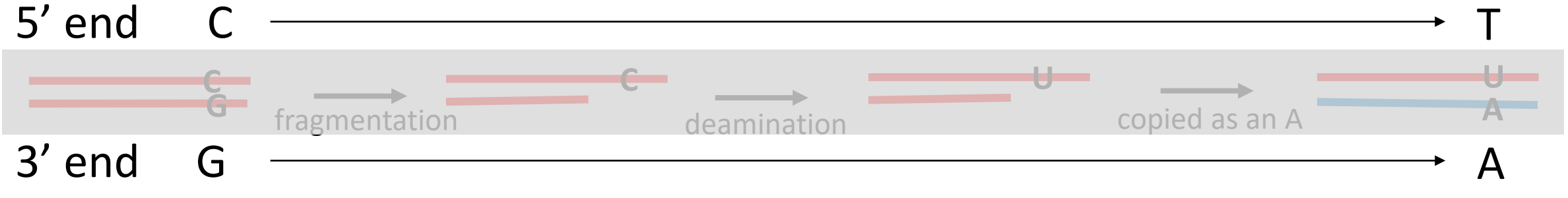
C ≡ G

post-mortem DNA damage

C → U, especially at the **end of the molecules**

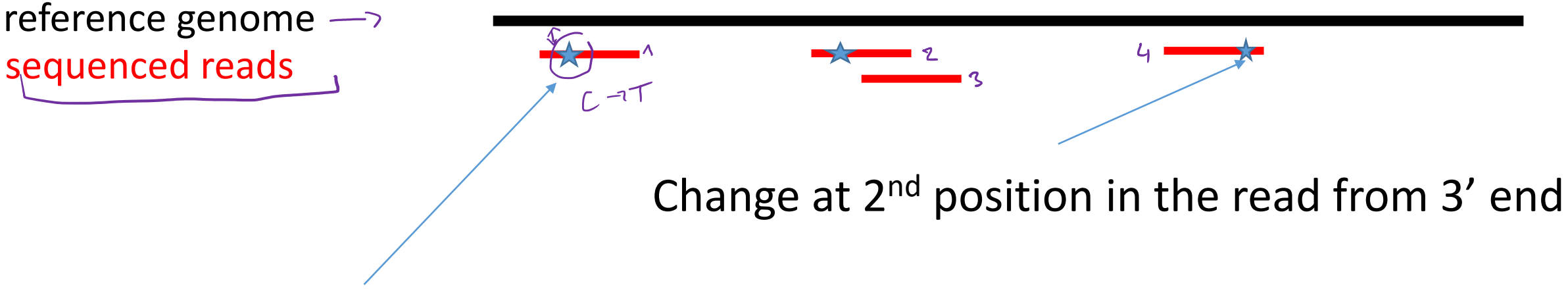


Dabney, M. Meyer, S. Pääbo, *Cold Spring Harb Perspect Biol*. **5**, a012567 (2013).



post-mortem DNA damage:

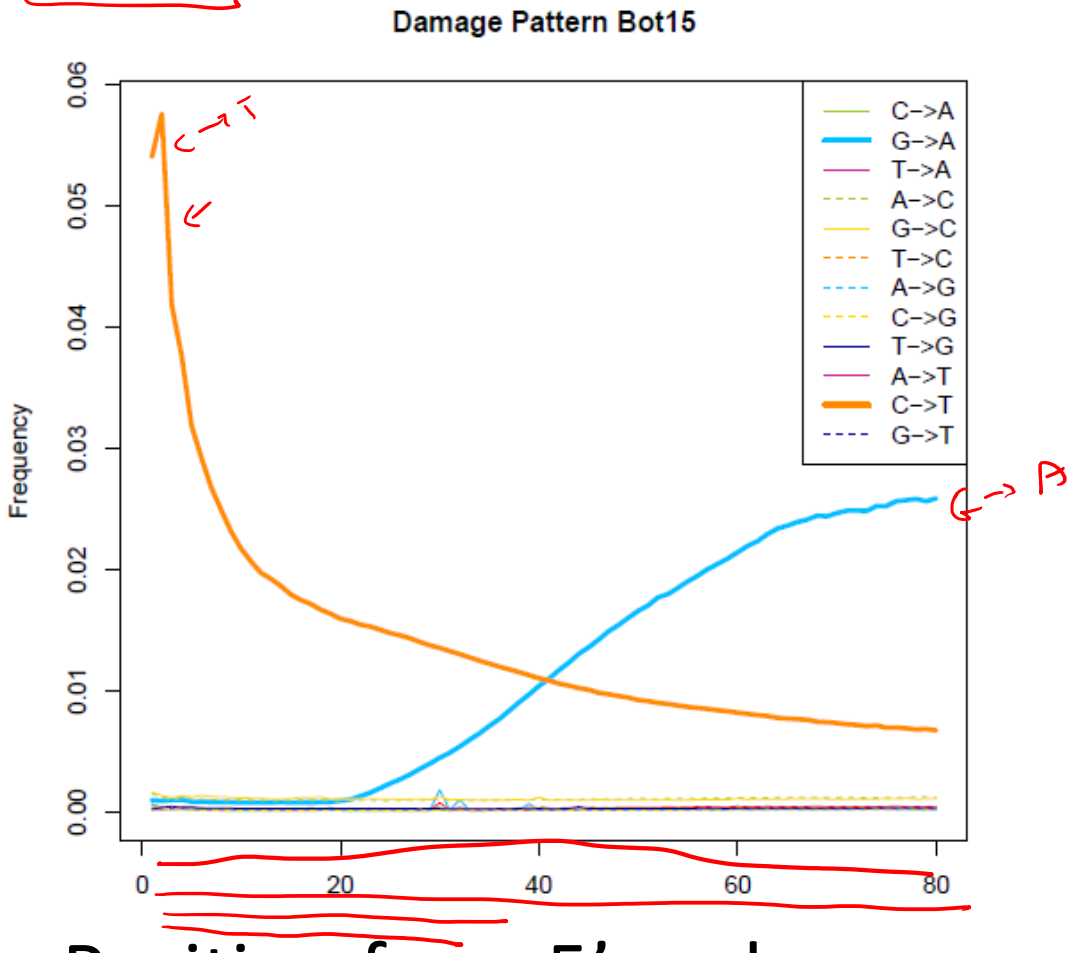
tabulate the number of differences between sequenced reads and **reference genome**



ref genome

post-mortem DNA damage

C → T (up to 40% on first base)



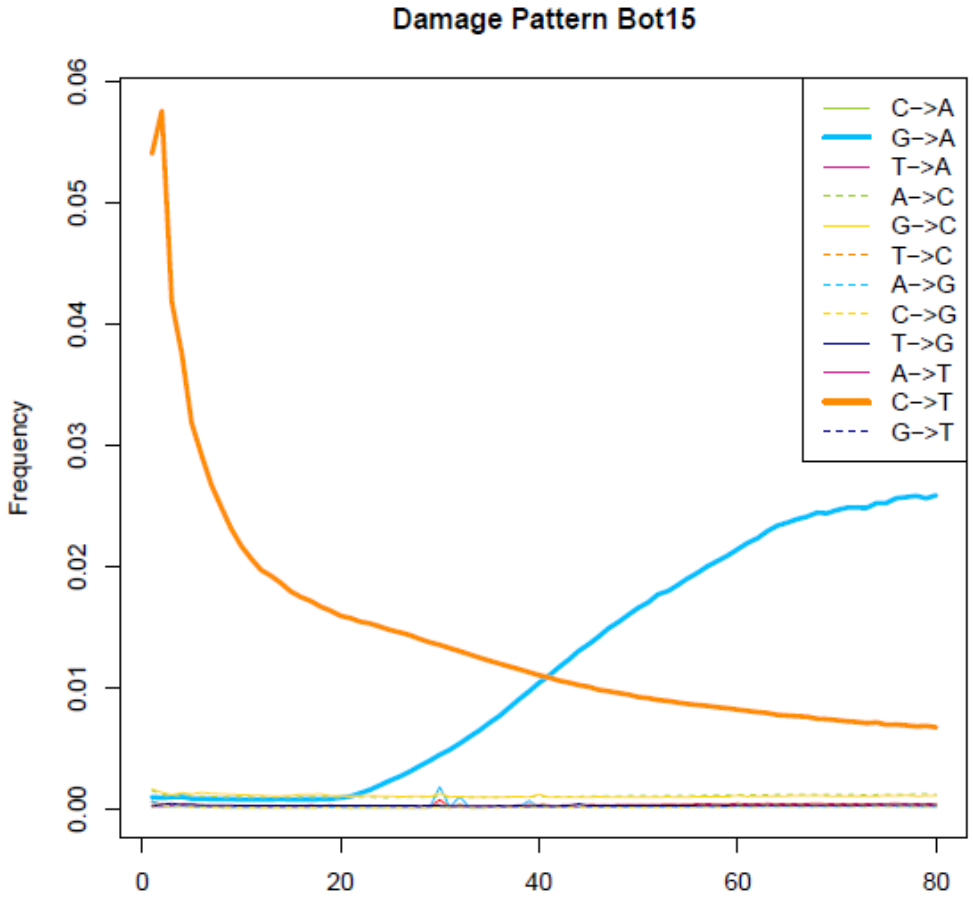
Position from 5’ end

ref genome

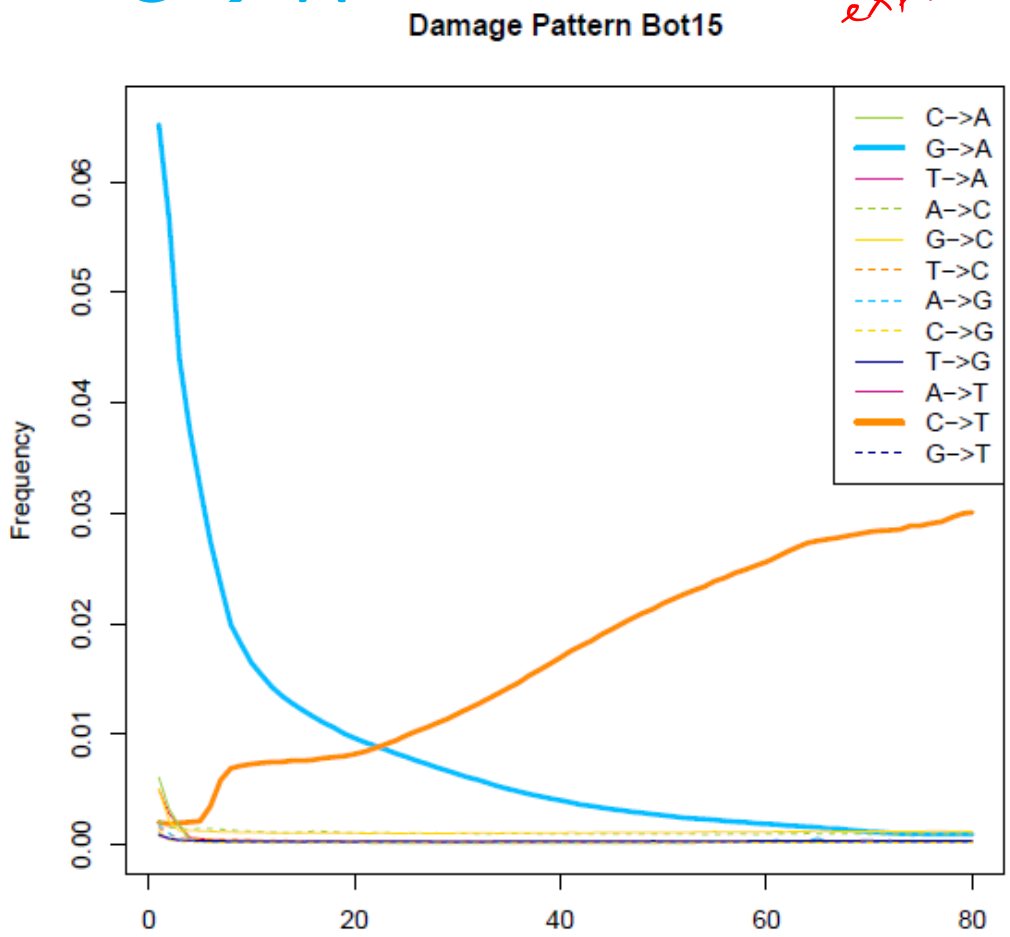
post-mortem DNA damage

C → T (up to 40% on first base)

G → A



Position from 5’ end



Position from 3’ end

Molecular damage is treated as a nuisance for inferring the evolutionary history

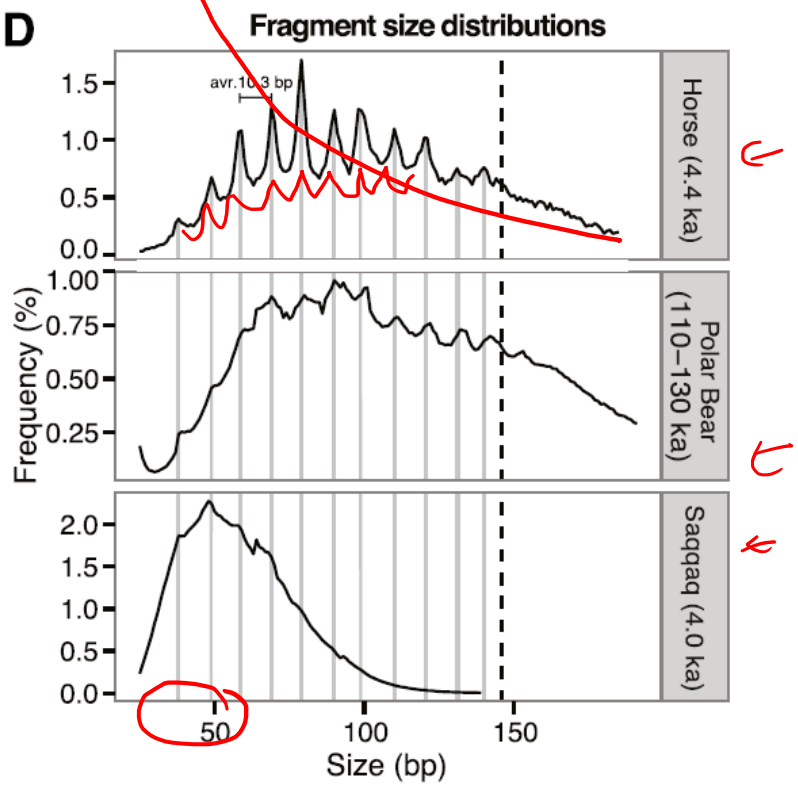
However it can also be used:

- To assess authenticity
- To extract meaningful biological features

Example 1:
Read length distribution might reflect nucleosome geometry

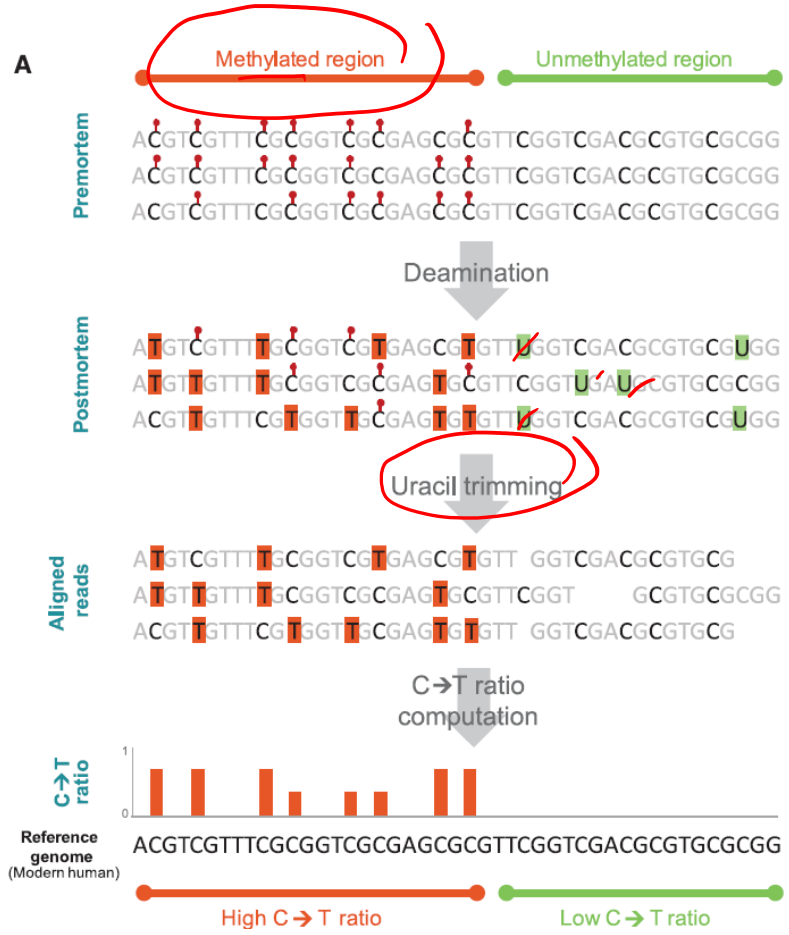
Distribution of fragment sizes from ancient samples :

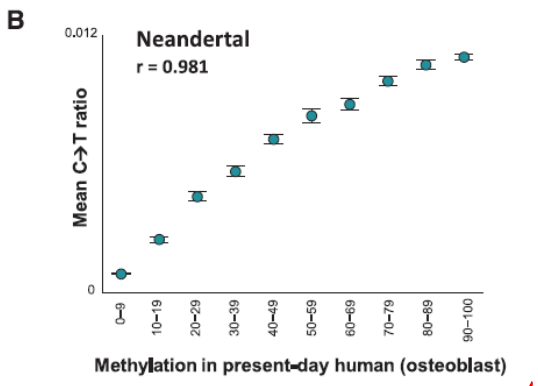
consistent cleavage at exposed nucleosome-wrapped DNA strands every 10 bp



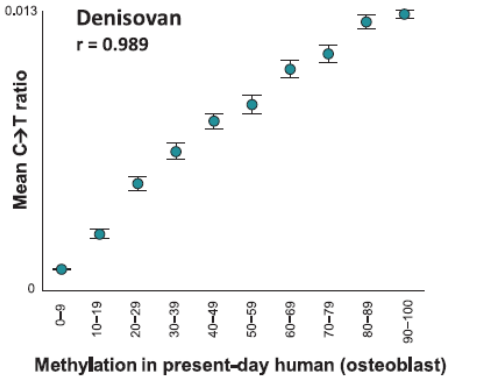
J. S. Pedersen *et al.*, *Genome Res.* **24**, 454–466 (2014).

Example 2: damage: reconstruct the DNA Methylation of ancient humans





fraction of CpG→TpG substitutions correlates with methylation in present-day human osteoblast



fraction of CpG→TpG substitutions may serve as a proxy for the levels of methylation in ancient DNA

D. Gokhman *et al.*, *Science*. **344**, 523–527 (2014).

Because of the lesions: DNA comes in low amounts. Example: Bronte Age Eurasia.

11 JUNE 2015 | VOL 522 | NATURE | 167

Population genomics of Bronze Age Eurasia

Morten E. Allentoft[1]*, Martin Sikora[1]*, Karl-Göran Sjögren[2], Simon Rasmussen[3], Morten Rasmussen[1], Jesper Stenderup[1], Peter B. Damgaard[1], Hannes Schroeder[1,4], Torbjörn Ahlström[5], Lasse Vinner[1], Anna-Sapfo Malaspinas[1], Ashot Margaryan[1], Tom Higham[6], David Chivall[6], Niels Lynnerup[7], Lise Harvig[7], Justyna Baron[8], Philippe Della Casa[9], Paweł Dąbrowski[10], Paul R. Duffy[11], Alexander V. Ebel[12], Andrey Epimakhov[13], Karin Frei[14], Miroslaw Furmanek[8], Tomasz Gralak[8], Andrey Gromov[15], Stanislaw Gronkiewicz[16], Gisela Grupe[17], Tamás Hajdu[18,19], Radoslaw Jarysz[20], Valeri Khartanovich[15], Alexandr Khokhlov[21], Viktória Kiss[22], Jan Kolář[23,24], Aivar Kriiska[25], Irena Lasak[8], Cristina Longhi[26], George McGlynn[17], Algimantas Merkevicius[27], Inga Merkyte[28], Mait Metspalu[29], Ruzan Mkrtchyan[30], Vyacheslav Moiseyev[15], László Paja[31,32], György Pálfi[32], Dalia Pokutta[2], Lukasz Pospieszny[33], T. Douglas Price[34], Lehti Saag[29], Mikhail Sablin[35], Natalia Shishlina[36], Václav Smrčka[37], Vasilii I. Soenov[38], Vajk Szeverényi[22], Gusztáv Tóth[39], Synaru V. Trifanova[38], Liivi Varul[25], Magdolna Vicze[40], Levon Yepiskoposyan[41], Vladislav Zhitenev[42], Ludovic Orlando[1], Thomas Sicheritz-Pontén[3], Søren Brunak[3,43], Rasmus Nielsen[44], Kristian Kristiansen[2] & Eske Willerslev[1]

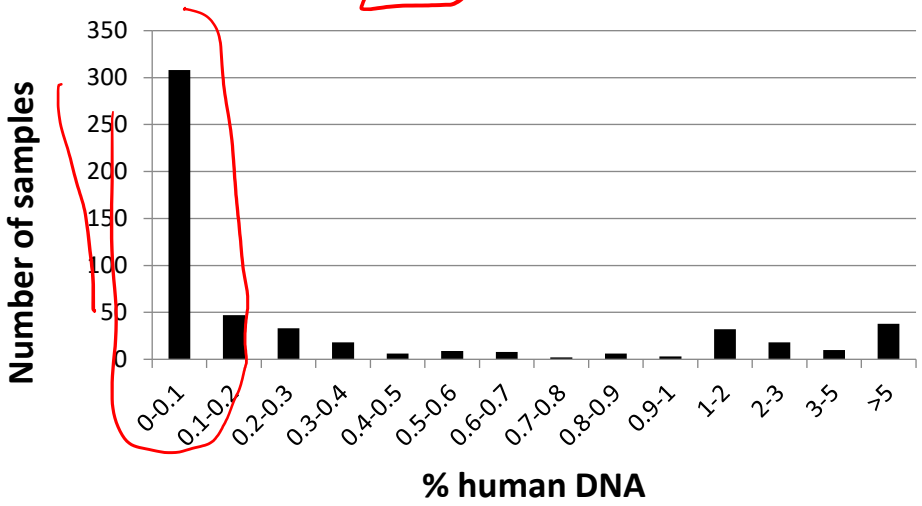
Because of the lesions: DNA comes in low amounts. Example: Bronte Age Eurasia.

11 JUNE 2015 | VOL 522 | NATURE | 167

Population genomics of Bronze Age Eurasia

Morten E. Allentoft[1]*, Martin Sikora[1]*, Karl-Göran Sjögren[2], Simon Rasmussen[3], Morten Rasmussen[1], Jesper Stenderup[1], Peter B. Damgaard[1], Hannes Schroeder[1,4], Torbjörn Ahlström[5], Lasse Vinner[1], Anna-Sapfo Malaspinas[1], Ashot Margaryan[1], Tom Higham[6], David Chivall[6], Niels Lynnerup[7], Lise Harvig[7], Justyna Baron[8], Philippe Della Casa[9], Paweł Dąbrowski[10], Paul R. Duffy[11], Alexander V. Ebel[12], Andrey Epimakhov[13], Karin Frei[14], Miroslaw Furmanek[8], Tomasz Gralak[8], Andrey Gromov[15], Stanislaw Gronkiewicz[16], Gisela Grupe[17], Tamás Hajdu[18,19], Radoslaw Jarysz[20], Valeri Khartanovich[15], Alexandr Khokhlov[21], Viktória Kiss[22], Jan Kolář[23,24], Aivar Kriiska[25], Irena Lasak[8], Cristina Longhi[26], George McGlynn[17], Algimantas Merkevicius[27], Inga Merkyte[28], Mait Metspalu[29], Ruzan Mkrtchyan[30], Vyacheslav Moiseyev[15], László Paja[31,32], György Pálfi[32], Dalia Pokutta[2], Lukasz Pospieszny[33], T. Douglas Price[34], Lehti Saag[29], Mikhail Sablin[35], Natalia Shishlina[36], Václav Smrčka[37], Vasilii I. Soenov[38], Vajk Szeverényi[22], Gusztáv Tóth[39], Synaru V. Trifanova[38], Liivi Varul[25], Magdolna Vicze[40], Levon Yepiskoposyan[41], Vladislav Zhitenev[42], Ludovic Orlando[1], Thomas Sicheritz-Pontén[3], Søren Brunak[3,43], Rasmus Nielsen[44], Kristian Kristiansen[2] & Eske Willerslev[1]

Screening of >600 (3'000 BC) Bronze Age samples



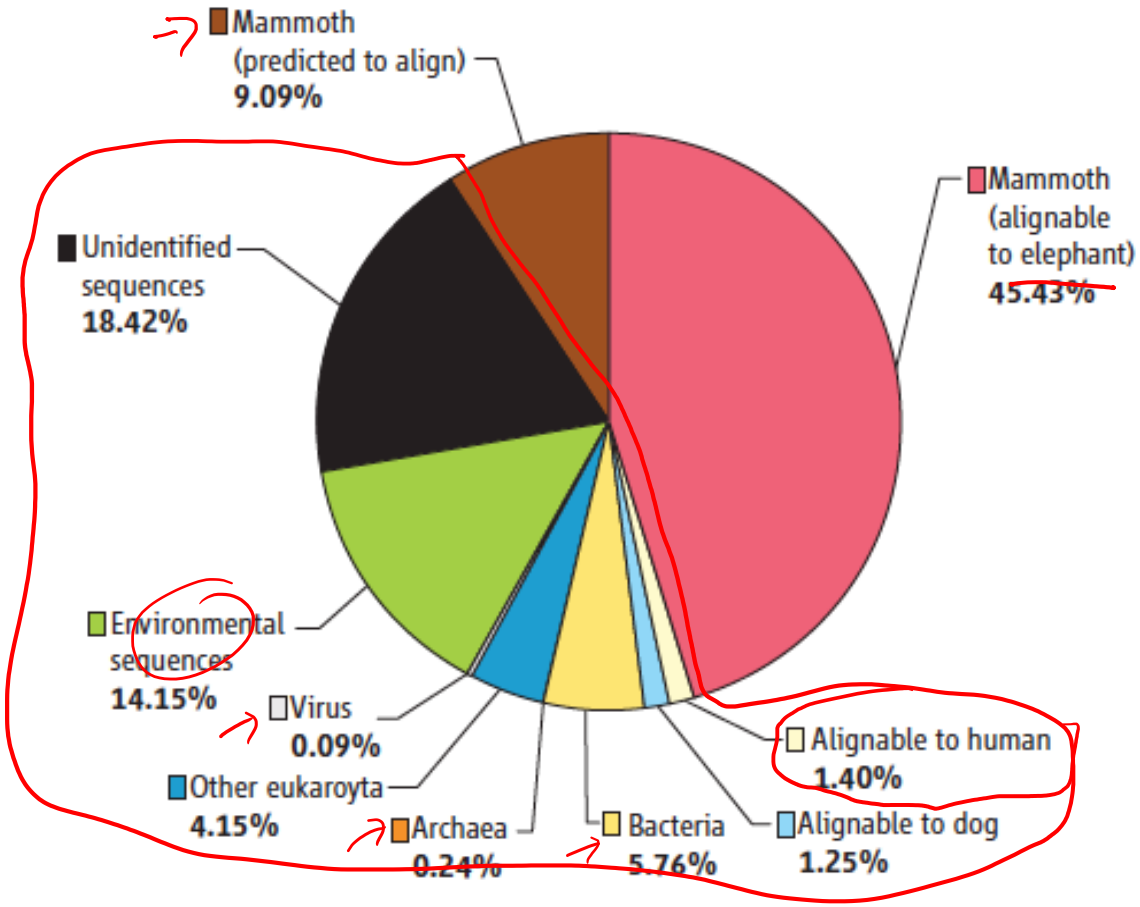
~ 1/2 samples with less **0.1%** "human DNA"

Ancient DNA comes in low amounts

Most of the DNA in an ancient DNA extract is of "exogenous" source

E.g., DNA extracted from a mammoth bone:

- 50% elephant/mammoth DNA
- 50% "**contamination**":
  - bacteria,
  - viruses,
  - and humans!



Source: Poinar *et al*., Science, 2006

Human contamination can lead to wrong evolutionary inferences (e.g. Egyptian mummy, Neanderthal, ancient humans).

It is hard to identify contamination when sequencing humans because the sequence identity is high.

human 1
" 2

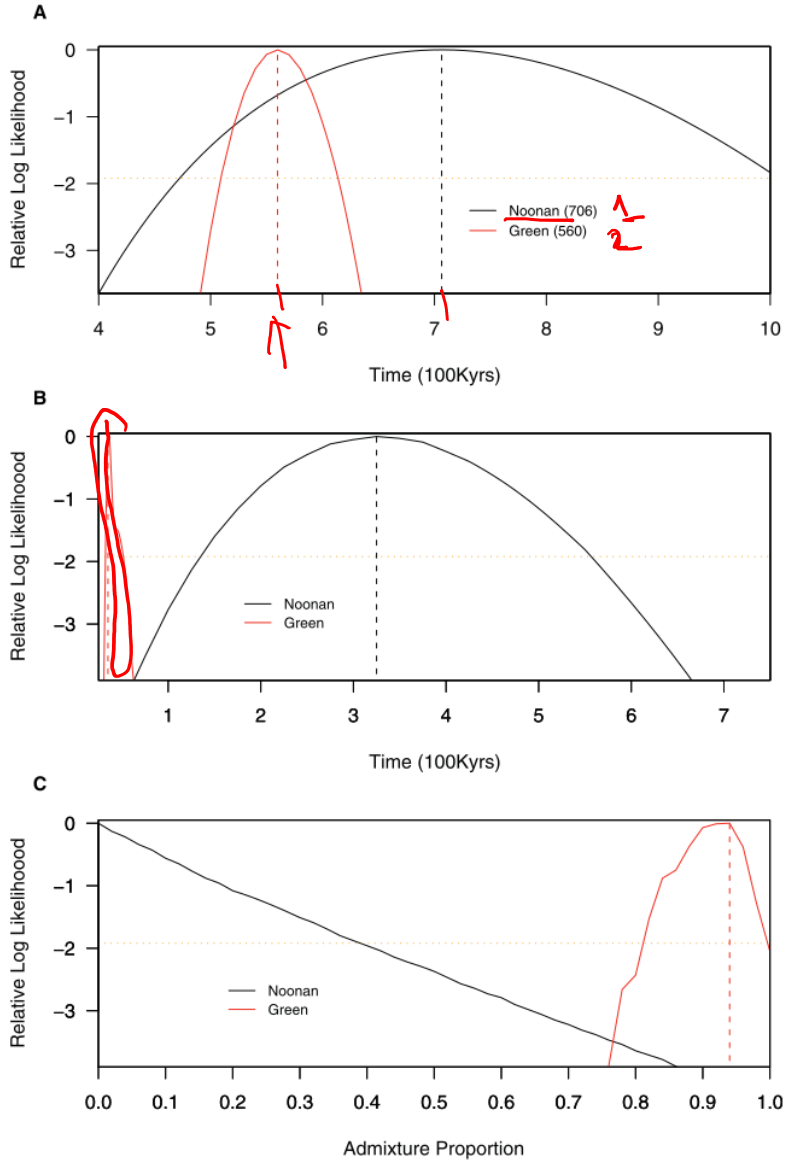


Figure 1. Likelihood Curves for (A) Human–Neanderthal Divergence Time, (B) Modern European–Neanderthal Split Time, and (C) Neanderthal Contribution to Modern European Ancestry for the Noonan et al. (1) and Green et al. (2) Data

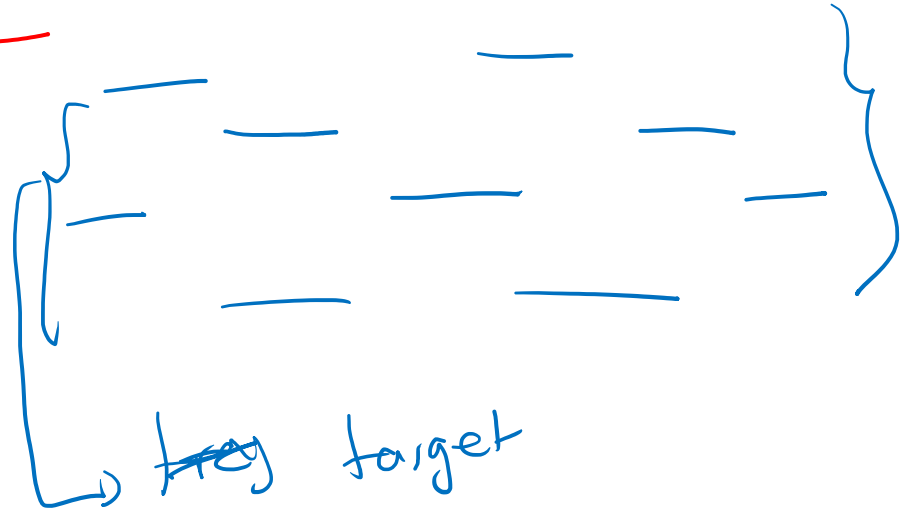
See Materials and Methods for details.

doi:10.1371/journal.pgen.0030175.g001

Wall, J. D. & Kim, S. K. PLoS Genet 3, e175 (2007).

(2) Computational methods

- Map/assemble the data
- Assess authenticity
- Variant calling
- Population genetics:
  - Infer demographic
  - [Infer selection]
- [Phylogenetics]
- [Environmental (eDNA)/metagenomics]



Map/assemble the data, “bam file”

Typical steps to assemble present-day genomes:

- mapping step (reference genome)

  e.g. bwa, Li, H., and Durbin, R. (2009). Bioinformatics 25, 1754–1760.

- remove duplicates

  e.g. picard, http://broadinstitute.github.io/picard/

- realignment step

  e.g. GATK, McKenna, A. et al. Genome Res. 20, 1297–1303 (2010).

“Extra steps” for ancient DNA:

(1) As DNA fragments are short, part of the illumina adapter sequenced as well

Adapter removal step:

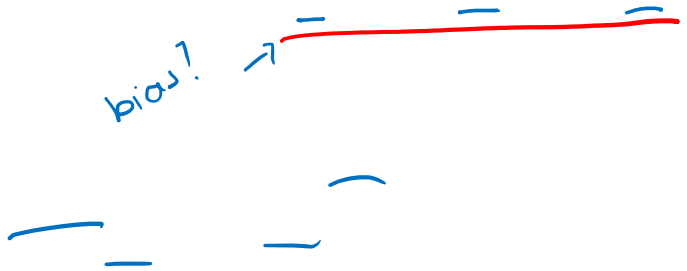
Schubert, M., Lindgreen, S. & Orlando, L. BMC Research Notes 9, 88 (2016).

(2) Many errors +short fragments

Disabling the seed in bwa:

Schubert, M. et al. BMC Genomics 13, 178 (2012).

Map/assemble the data, “bam file”



Typical steps to assemble present-day genomes:

- mapping step (reference genome)

  e.g. bwa, Li, H., and Durbin, R. (2009). Bioinformatics 25, 1754–1760.

- remove duplicates

  e.g. picard, http://broadinstitute.github.io/picard/

- realignment step

  e.g. GATK, McKenna, A. et al. Genome Res. 20, 1297–1303 (2010).

“Extra steps” for ancient DNA:

(1) As DNA fragments are short, part of the illumina adapter adapter sequenced as well

Adapter removal step:

Schubert, M., Lindgreen, S. & Orlando, L. BMC Research Notes 9, 88 (2016).

(2) Many errors +short fragments

Disabling the seed in bwa:
Schubert, M. et al. BMC Genomics 13, 178 (2012).

Reference bias in ancient DNA:
Günther, T. & Nettelblad, C. PLOS Genetics 15, e1008302 (2019).

“Removing it”:
Martiniano, R., et al. Genome Biology 21, 250 (2020).

Reference bias in ancient DNA

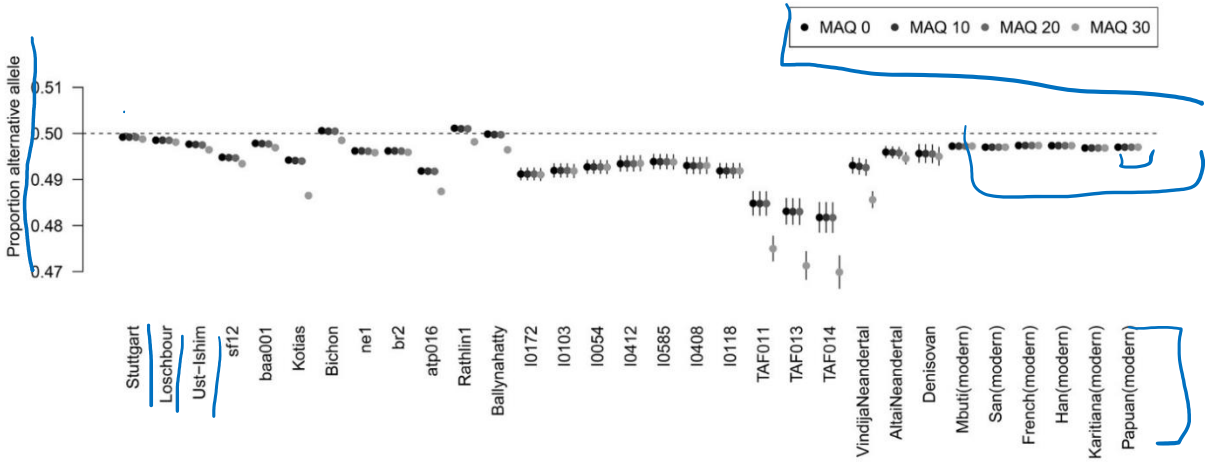


Fig 1. Reference bias in published genome-wide ancient DNA datasets for different minimum mapping quality thresholds. The plot shows the average proportion of reads at heterozygous transversion sites (see Methods) representing the alternative allele. Error bars indicate two standard errors of the mean.

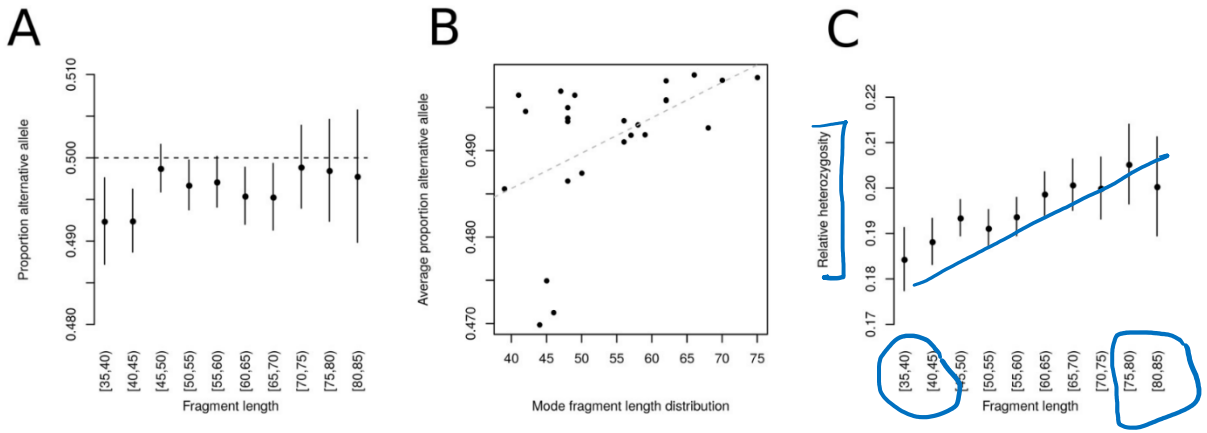


Fig 2. Connection between fragment length and reference bias. (A) Proportion of alternative allele for different fragment length bins in the high coverage individual sf12. (B) Correlation between average proportion of alternative alleles and the mode of the fragment size distribution across all investigated individuals. (C) Proportion of heterozygous sites among all sites with sufficient coverage for different fragment length bins in the high coverage individual sf12. All error bars indicate two standard errors.

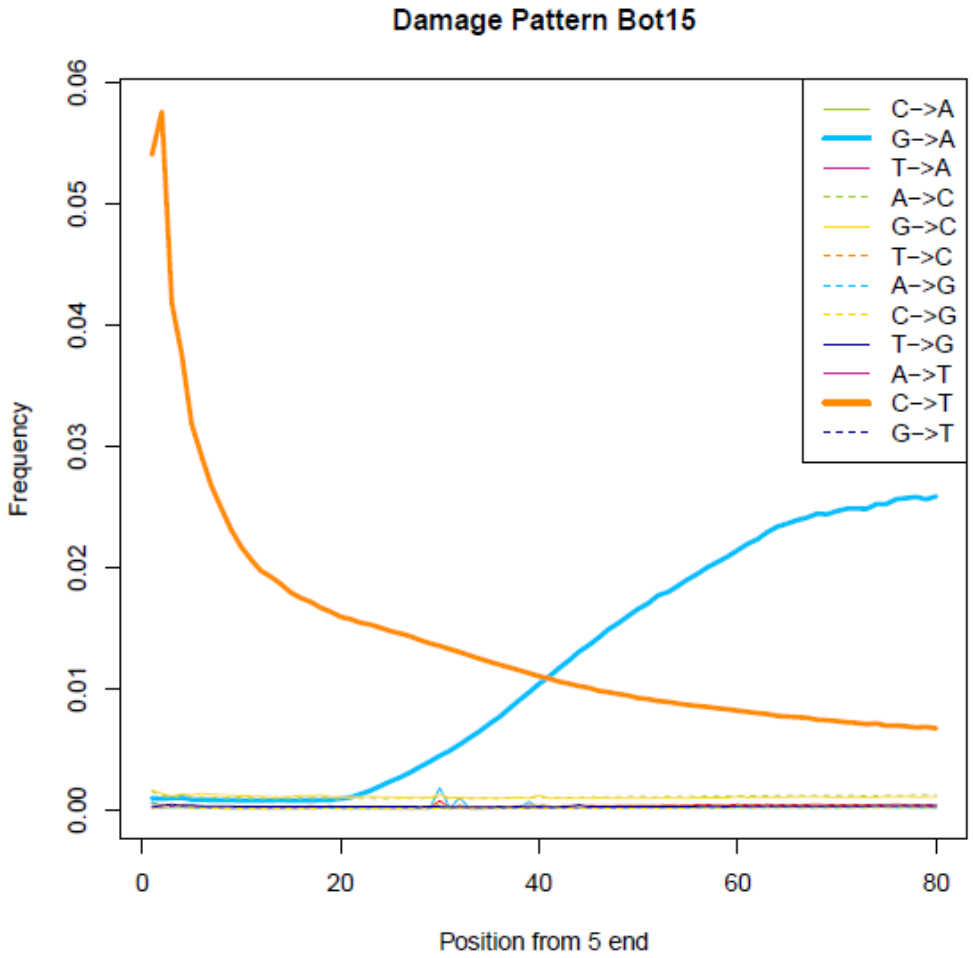
Günther, T. & Nettelblad, C. PLOS Genetics 15, e1008302 (2019).

Read alignment and processing pipelines

PALEOMIX: Schubert, M. et al. Nat Protoc 9, 1056–1082 (2014).

EAGER: Peltzer, A. et al. Genome Biology 17, 60 (2016).

Assess authenticity: damage estimate

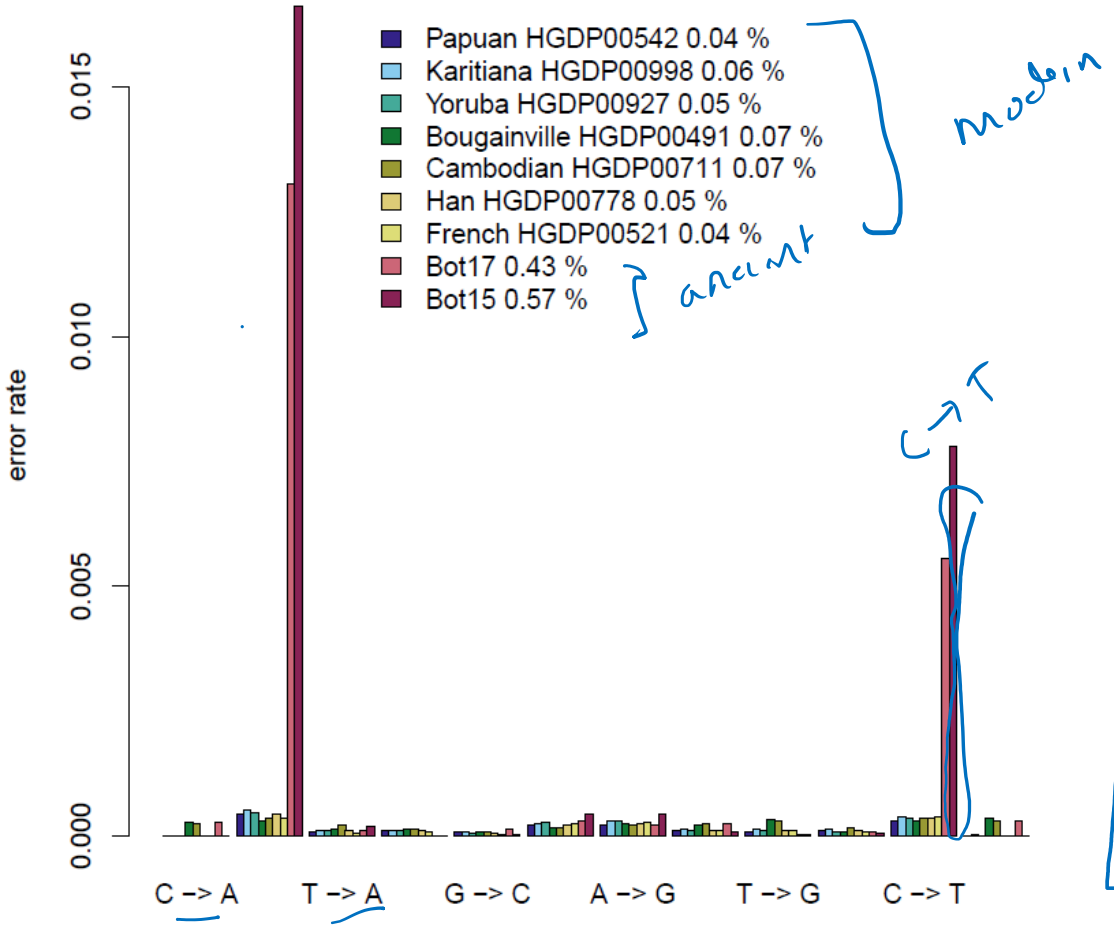


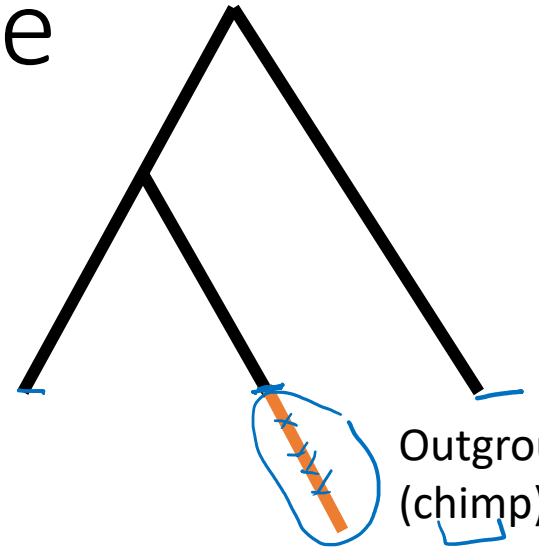
Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682 (2013).

Or simpler version "counting":
Malaspinas, A.-S. *et al.* bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics* **30**, 2962–2964 (2014).

Assess authenticity: damage estimate

overall error rate across the genome





Modern sample

Ancient sample

T. S. Korneliussen, A. Albrechtsen, R. Nielsen, BMC Bioinformatics. 15, 356 (2014). http://www.popgen.dk/**angsd**/index.php/ANGSD

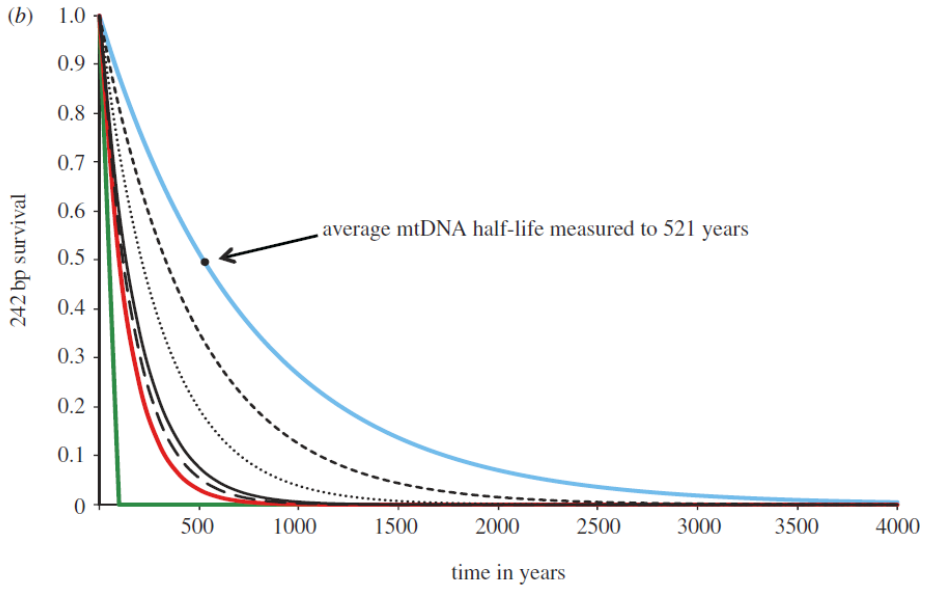
Orlando, L. et al. Nature 499, 74–78 (2013).

Read length distribution and decay rate for ancient DNA

- 1.

Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B* rspb20121745 (2012) doi:10.1098/rspb.2012.1745.

$$N_t = 3.61 \times e^{-0.0013t}$$



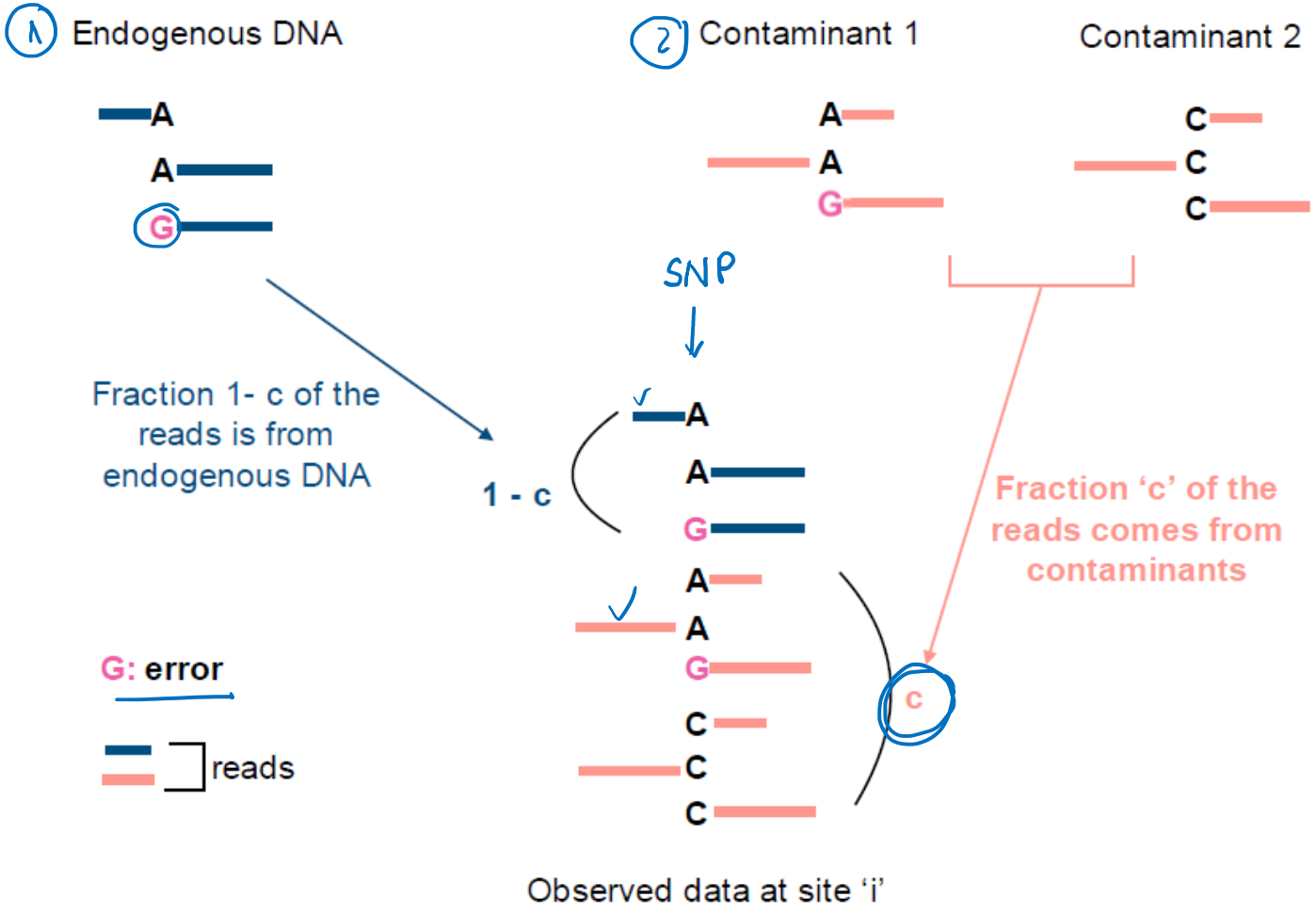
Assess authenticity: contamination estimate

- Based on haploid data (e.g. sexual chromosomes)
  - MT or X (or Y) chromosome based estimates
    - contamix: Fu, Q. et al. A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. Current Biology 23, 553–559 (2013).
    - schmutzi: Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. Genome Biology 16, (2015).
    - contaminationX: Moreno-Mayar, J. V. et al. A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. Bioinformatics (2019) doi:10.1093/bioinformatics/btz660.
- Based on diploid data
  - DICE: Racimo, F., Renaud, G. & Slatkin, M. Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. PLOS Genetics 12, e1005972 (2016).
  - ContamLD: Nakatsuka, N. et al. ContamLD: estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. Genome Biology 21, 199 (2020).

Estimating **human contamination** from (low depth) **haploid** data

Víctor Moreno-Mayar

Jyoti Dalal



X
Y
MT

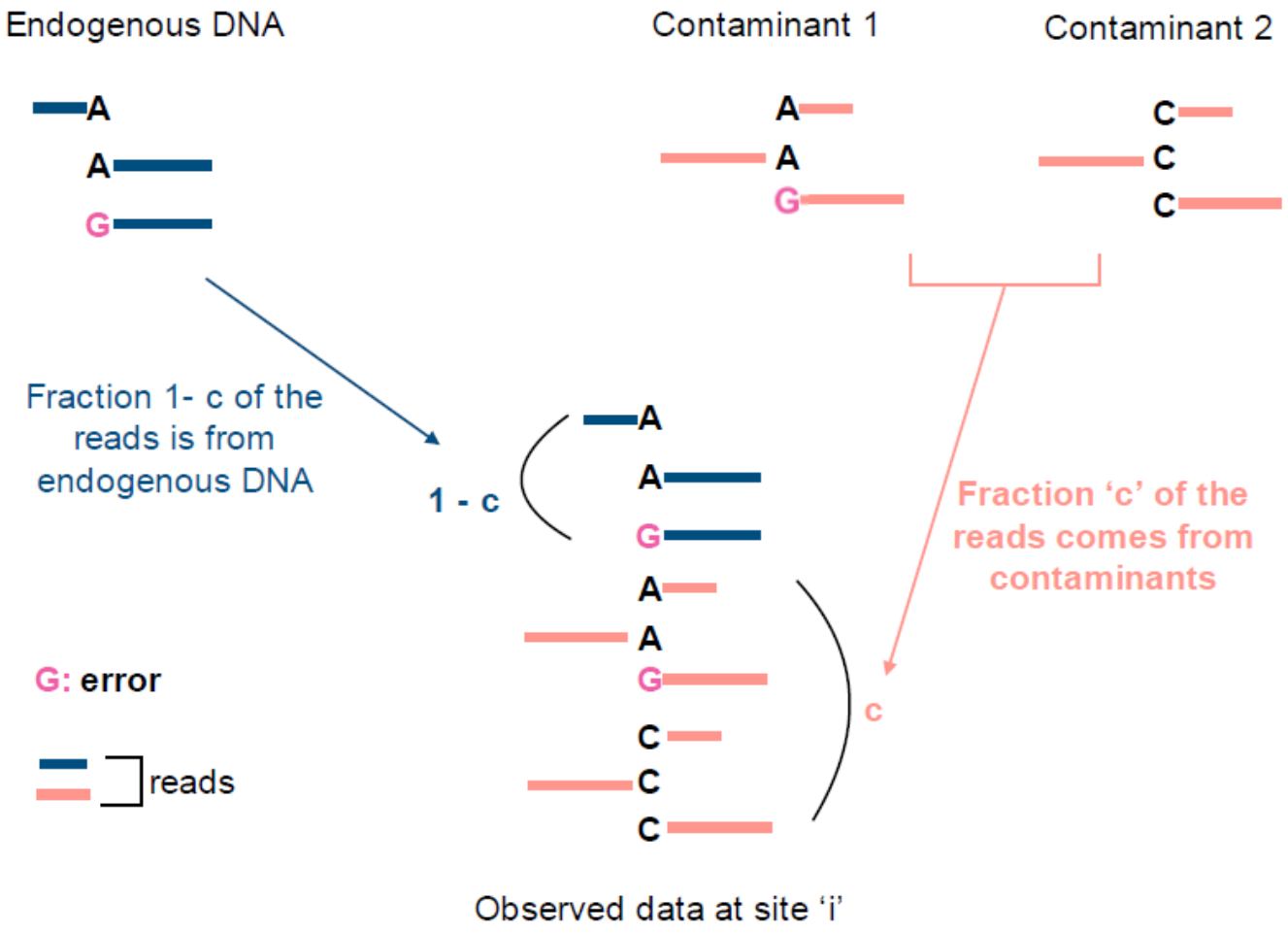
H1 : A
H2 : A
H3 : C

mixture A, C

Estimating **human contamination** from (low depth) **haploid** data

Víctor Moreno-Mayar

Jyoti Dalal



f_A^i: freq. of 'A' at site 'i' in contaminants

f_C^i: freq. of 'C' at site 'i' in contaminants

Estimating **human contamination** from haploid data

Víctor Moreno-Mayar

Jyoti Dalal

Assume:
Large pool of reads to draw from → sampling with replacement.
Reads covering only one site → independence.
Biallelic sites. No mapping error (!).

$$\ell(c) = p(X|c, \Gamma, F)$$

$$= \prod_{i=1}^{L} \left(\frac{1}{2} \binom{n_T^i}{n_1^i} (p_1^i)^{n_1^i} (1 - p_1^i)^{(n_T^i - n_1^i)} + \frac{1}{2} \binom{n_T^i}{n_1^i} (q_1^i)^{n_1^i} (1 - q_1^i)^{(n_T^i - n_1^i)} \right)$$

Total number of reads at position i

Assuming a true allele in the endogenous individual with equal probability

Estimating **human contamination** from haploid data

Víctor Moreno-Mayar

Jyoti Dalal

E.g.: Naturally segregating A, T alleles at position i
Condition on A being the endogenous read

$$p(\text{one A allele in a single draw}|c, \Gamma, F) =$$

$$c(f_A\gamma_{AA} + f_T\gamma_{TA}) + (1-c)\gamma_{AA}$$

no error

2 → 1 error

no error

contaminant read

frequency of *A* allele in the contaminating population

frequency of *T* allele in the contaminating population

endogenous read

Estimating **human contamination** from haploid data

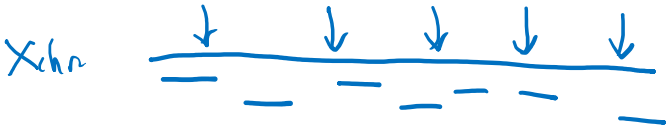
Víctor Moreno-Mayar

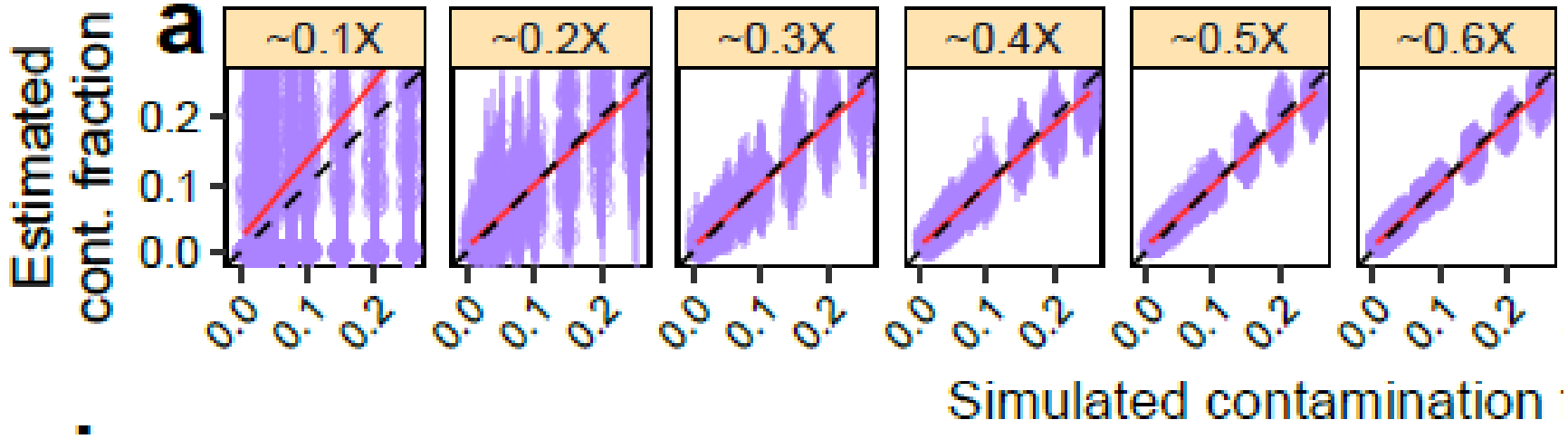
Jyoti Dalal

Simulations: mix of genomes/"bam files"

Maximum likelihood estimate for the contamination fraction **c** :

French (European) and Yoruba (African) samples:





Moreno-Mayar, J. V. *et al.*. *bioRxiv* (2019). doi:10.1101/594481

Variant calling, accounting for low depth & high error rate: examples

Variant calling: accounting for low depth & high error rate

Few of the popgen methods used are specifically developed for ancient DNA.

In practice, usually, the focus is on treating the data differently *prior* to downstream analyses.

For instance:

- sample a read at every position: e.g. *Green, R. E. et al. Science 328, 710–722 (2010).*
- trim the reads (remove 5 bp at the beginning and end): *Skoglund, P. et al. Science 344, 747–750 (2014).*
- filter our transitions (C/T and G/A) – or repeat the analyses
- compute the genotype likelihood and integrate over the values (see later)
- imputation: *GLIMPSE, Rubinacci, S. et al. Nature Genetics 53, 120–126 (2021).*
- call genotype ("pretend it is modern DNA"):
  ANGSD: http://www.popgen.dk/angsd/index.php/ANGSD
  ATLAS: Link, V. et al. , biorxiv, (2017) doi:10.1101/105346.
  snpAD: Prüfer, K. snpAD: an ancient DNA genotype caller. Bioinformatics 34, 4165–4171 (2018).

Population ~~genetics~~ structure: infer demography

- Model "free" methods
  - **PCA/MDS,** ADMIXTURE

    Assumes "nothing" about population genetics
  - **D-**, F- statistics:

    make some assumptions but no estimates of population genetic parameters

- Population genetic models
  - not specific to aDNA: e.g. dadi (Gutenkunst et al. (2009)., fastsimcoal (Excoffier et al. , *PLoS Genet*.), momi (Kamm, et al. 2017).

  (coalescent, numerical approximation to diffusion)
  - Test for direct ancestry : Rasmussen et al. 2014
  - Branch shortening : Fu et al. 2013
  - LD based methods : …

PCA/MDS: commonly used softwarec

genetic data:

Patterson, N., Price, A. L. & Reich, D. PLoS Genet 2, e190 (2006).

McVean, G. PLoS Genet 5, e1000686 (2009).

aDNA:

smartPCA: https://github.com/chrchang/eigensoft/

wiki/smartpca

PCAngsd: Meisner, J. & Albrechtsen, A. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics 210, 719–731 (2018).

bammds: Malaspinas, A.-S. et al. Bioinformatics 30, 2962–2964 (2014).

Classical Multidimensional Scaling

MDS

- Consider:
  - a set of *n* objects (*r,s*),
  - measurement of the dissimilarity between objects (δ_{rs}).

- Multidimensional Scaling (MDS): search for a low dimensional space (Euclidean) where
  - *points* represent *objects*
  - distances between the points (d_{rs}) match as well as possible the original distances (δ_{rs}).

Classical Multidimensional Scaling

- Consider:
  - a set of *n* objects (*r,s*),
  - measurement of the dissimilarity between objects (δ_{rs}).

- Multidimensional Scaling (MDS): search for a low dimensional space (Euclidean) where
  - *points* represent *objects*
  - distances between the points (d_{rs}) match as well as possible the original distances (δ_{rs}).

Background: example 1

Input: a matrix Δ of dissimilarities between pairs of chromosomes

Output: coordinates for each chromosomes

Background: example 1

Input: a matrix Δ of dissimilarities between pairs of chromosomes

Output: coordinates for each chromosomes
Example:
3 individuals, 4 sites:

A1: A A G
A2: A A C
B: T T T

Background: example 1

Input: a matrix Δ of dissimilarities between pairs of chromosomes

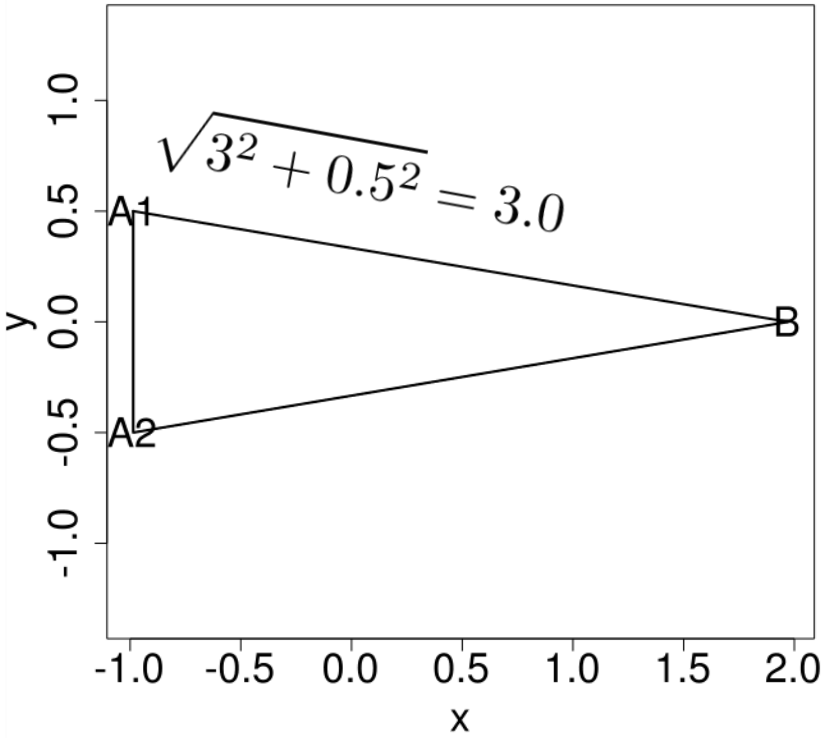
Output: coordinates for each chromosomes
Example:
3 individuals, 4 sites:

A1: A A G
A2: A A C
B: T T T

Distance matrix

	$A1$	$A2$	B
$A1$	0		
$A2$	1	0	
B	3	3	0

MDS



Background: example 1

Input: a matrix Δ of dissimilarities between pairs of chromosomes

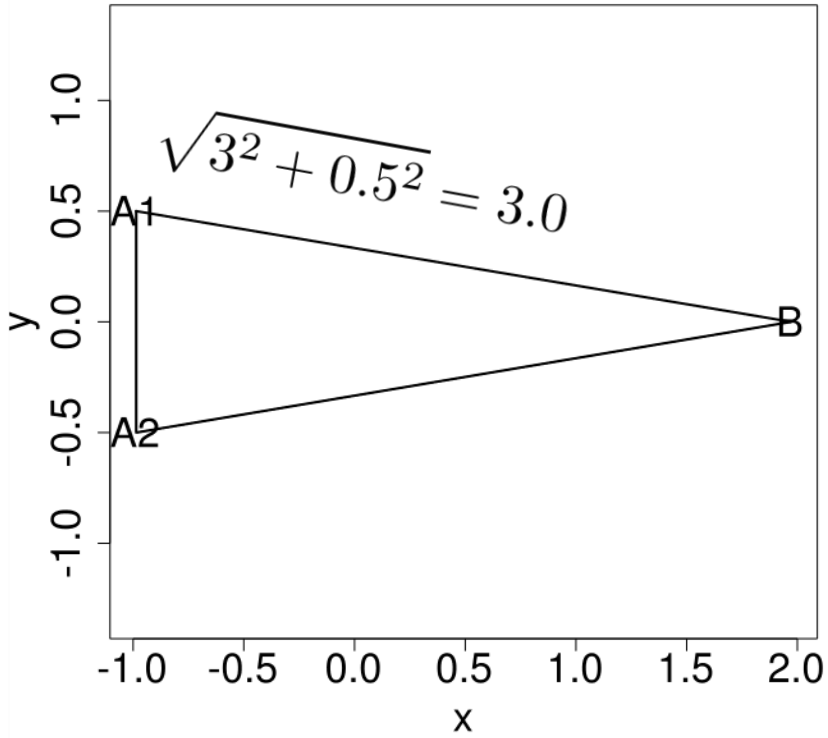
Output: coordinates for each chromosomes
Example:
3 individuals, 4 sites:

A1: A A G
A2: A A C
B: T T T

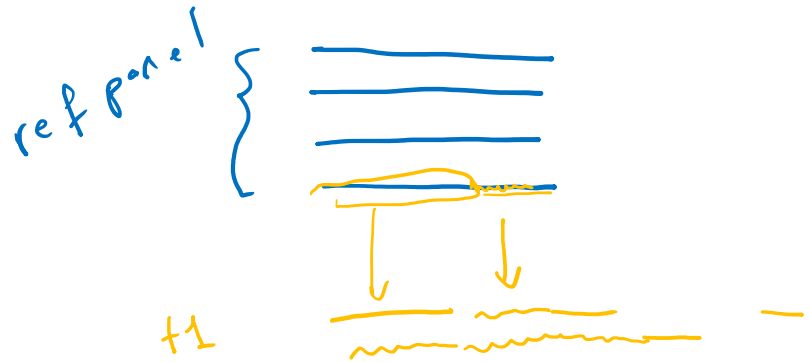
Distance matrix

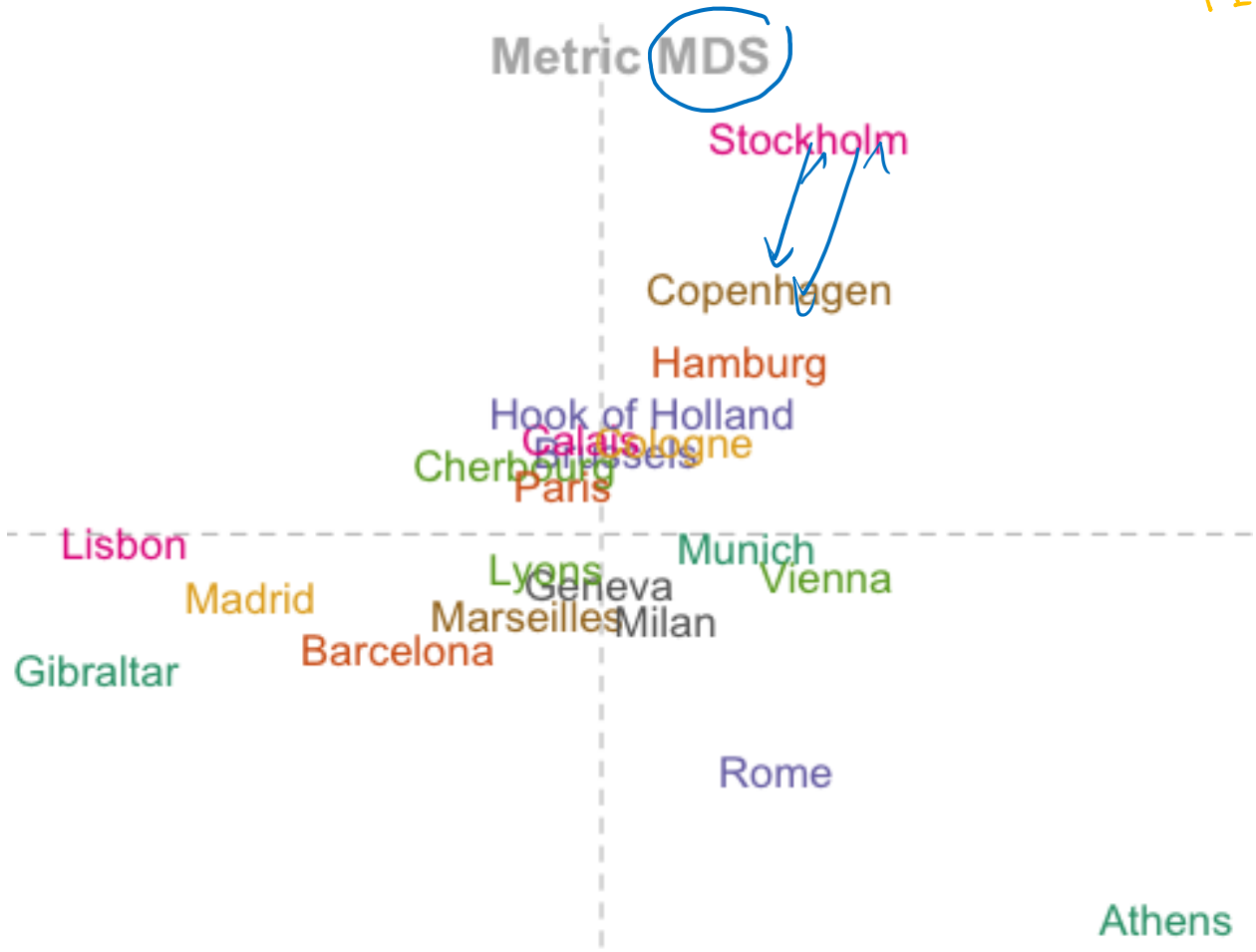
	$A1$	$A2$	B
$A1$	0		
$A2$	1	0	
B	3	3	0

MDS



Background: example 2

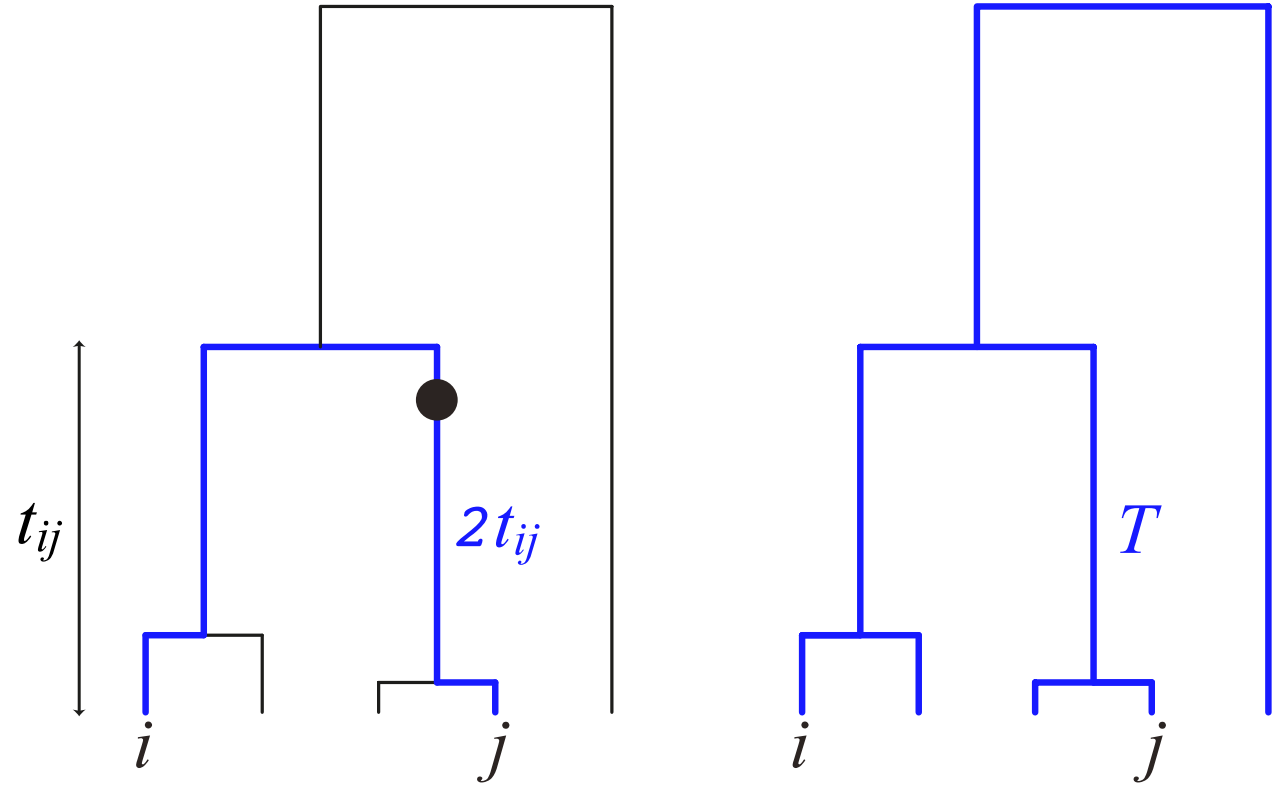




Genetic distance and coalescent theory

"Genetic distance": here we use "allele sharing distance"

$$\frac{\text{\# alleles mismatching between chrom } i \text{ and } j}{\text{\# SNPs}} \sim \frac{2E[t_{ij}]}{E[T]}$$



Kingman's coalescent

MDS and ancient DNA

There is a “natural”way to handle **low amounts** of data in MDS by adjusting the distance function:

$$\frac{\text{\# alleles mismatching between chrom } i \text{ and } j}{\text{\# SNPs (excl. sites missing in either i or j)}}$$

Damage: working on it

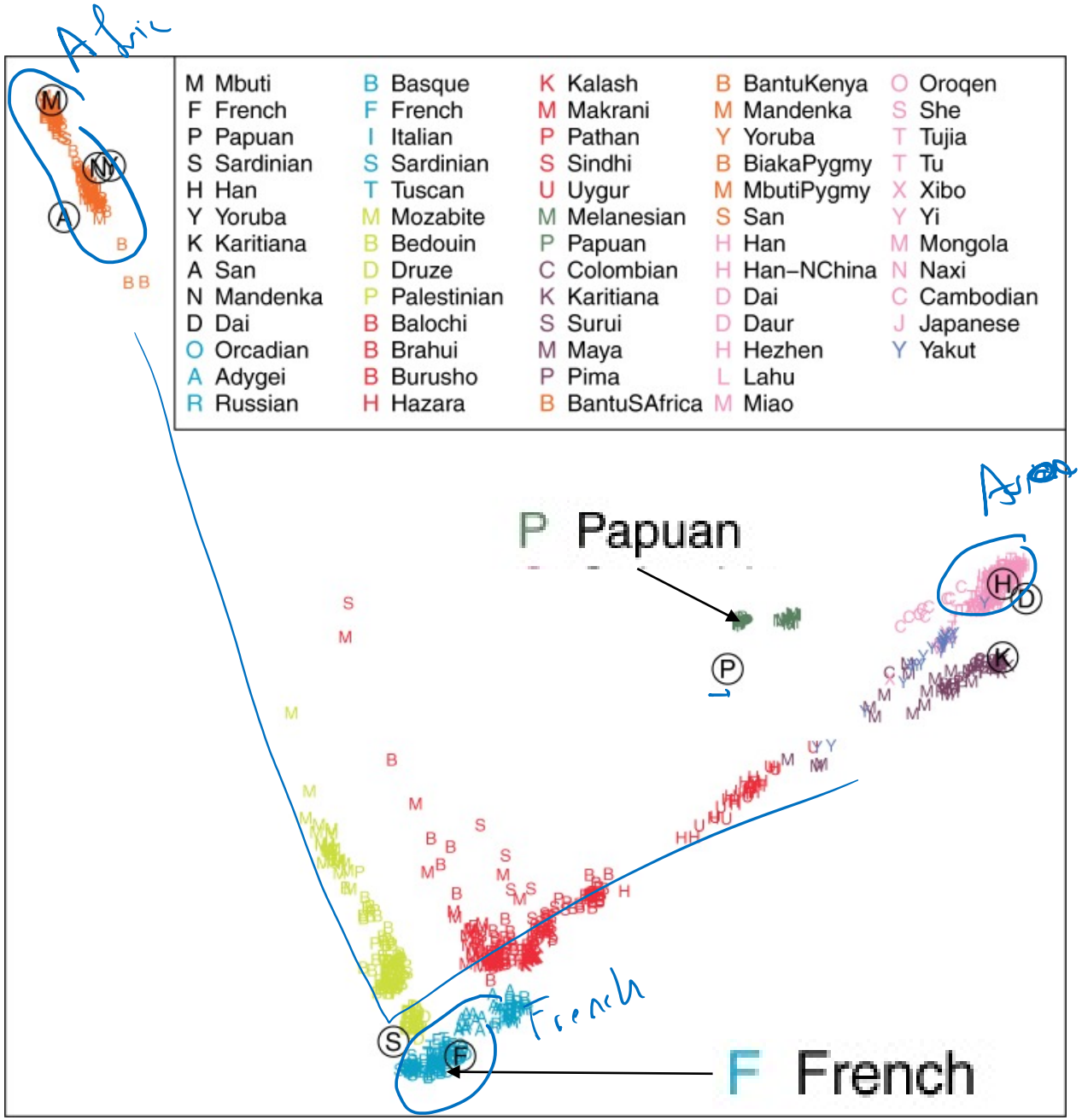
Malaspinas *et al.*, *Bioinformatics*. **30**, 2962–2964 (2014).

MDS and ancient DNA

There is a “natural”way to handle **low amounts** of data in MDS by adjusting the distance function:

$$\frac{\text{\# alleles mismatching between chrom } i \text{ and } j}{\text{\# SNPs (excl. sites missing in either i or j)}}$$

Damage: working on it



First two dimensions of an MDS plot including the ten 0.1X modern human genomes and the HGDP SNP data

Malaspinas *et al.*, *Bioinformatics*. **30**, 2962–2964 (2014).

MDS and ancient DNA

There is a “natural”way to handle **low amounts** of data in MDS by adjusting the distance function:

$$\frac{\text{\# alleles mismatching between chrom } i \text{ and } j}{\text{\# SNPs}}$$

(excl. sites missing in either i or j)

Table 1. Summary of the simulation results for the ten modern genomes. For more details, see Supplementary data

Min. approx. depth of coverage to…	…recover geographic region as closest centroid	…recover true population within three closest centroids	…be placed within population ellipse
Mbuti (Africa)	0.001	0.001	0.1
French (Europe)	0.001	0.01	0.1
Papuan (Oceania)	0.001	0.001	0.5
Sardinian (Europe)	0.1	0.01	0.5
Han (Eastern Asia)	0.001	0.1	0.01
Yoruba (Africa)	0.001	0.001	0.1
Karitiana (America)	0.01	0.01	0.1
San (Africa)	0.001	0.001	1
Mandenka (Africa)	0.001	0.1	0.1
Dai (Eastern Asia)	0.001	0.5	0.5

“ADMIXTURE”, clustering algorithm to detect population structure

- Original model “Structure”

Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. Genetics 155, 945–959 (2000).

- Draw a read and use standard method: ADMIXTURE

Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. (2009) doi:10.1101/gr.094052.109.

- Genotype likelihoods: ngsadmix

Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating Individual Admixture Proportions from Next Generation Sequencing Data. Genetics 195, 693–702 (2013).

NGSAdmix

$$p\left(G_{ij}\middle|Q,F\right)=p\left(G_{ij}\middle|h^{ij}\right)=\begin{cases}\left(h^{ij}\right)^2 & \text{if} \quad G_{ij}=0\\ 2h^{ij}\left(1-h^{ij}\right) & \text{if} \quad G_{ij}=1\\ \left(1-h^{ij}\right)^2 & \text{if} \quad G_{ij}=2.\end{cases}$$

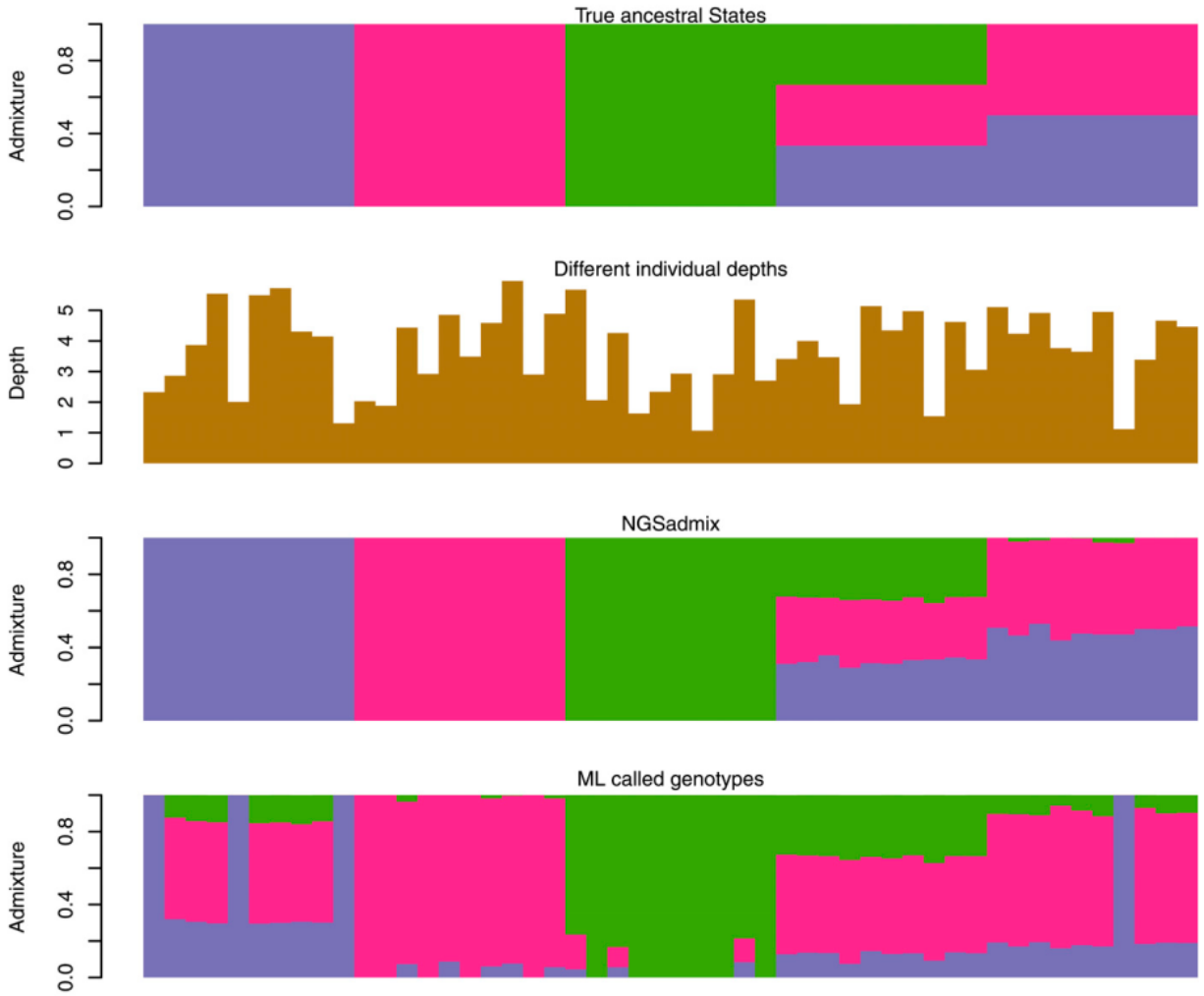
$$p(G|Q,F)=\prod_{j=1}^{M}\prod_{i=1}^{N}p\left(G_{ij}\middle|Q,F\right)=\prod_{j=1}^{M}\prod_{i=1}^{N}p\left(G_{ij}\middle|h^{ij}\right)$$

$$p(X|Q,F)=\prod_{j=1}^{M}\prod_{i=1}^{N}p\left(X_{ij}\middle|Q,F\right)=\prod_{j=1}^{M}\prod_{i=1}^{N}p\left(X_{ij}\middle|h^{ij}\right)$$

$$=\prod_{j=1}^{M}\prod_{i=1}^{N}\sum_{G_{ij}\in\{0,1,2\}}p\left(X_{ij}\middle|G_{ij}\right)p\left(G_{ij}\middle|h^{ij}\right).$$

NGSAdmix

NGSAdmix


True ancestral States
Admixture
0.0
0.4
0.8
Different individual depths
Depth
0
1
2
3
4
5
NGSadmix
Admixture
0.0
0.4
0.8
ML called genotypes
Admixture
0.0
0.4
0.8

F-stats/D-stats

- **Applications: used to argue for Homo Sapiens/Nanderthal gene flow**

Green, R. E. et al. A Draft Sequence of the Neandertal Genome. Science 328, 710–722 (2010).

- **For a theoretical/coalescent theory perspective:**

Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. Mol Biol Evol 28, 2239–2252 (2011).

Peter, B. M. Admixture, Population Structure and F-Statistics. Genetics genetics.115.183913 (2016) doi:10.1534/genetics.115.183913.

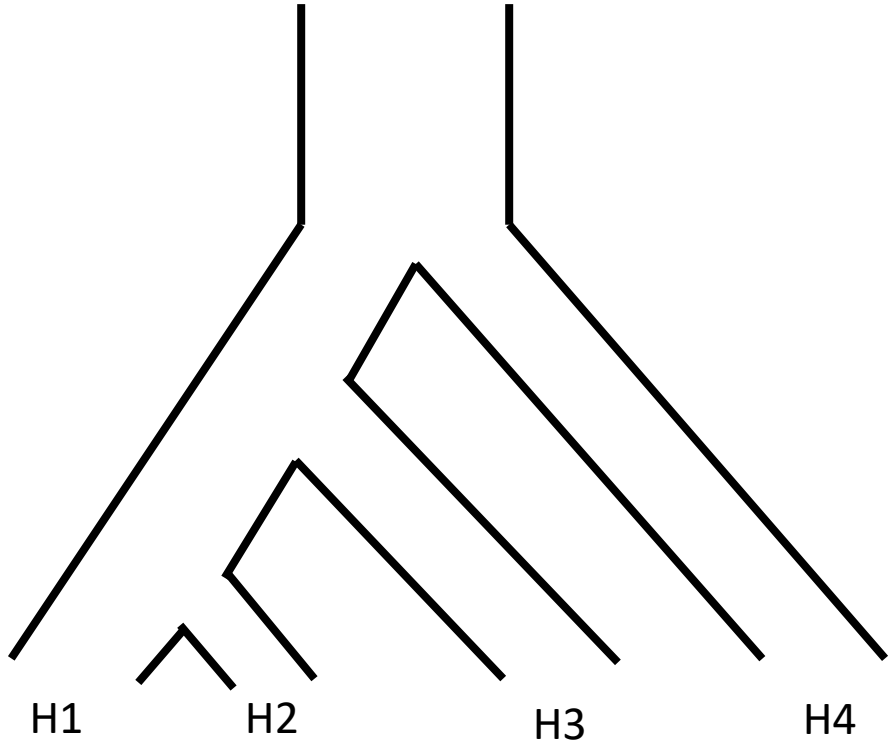
- **For implementations:**

ADMIXtools: Patterson, N. J. et al. Ancient Admixture in Human History. Genetics genetics.112.145037 (2012) doi:10.1534/genetics.112.145037.

FrAnTK: Moreno-Mayar, J. V. FrAnTK: a Frequency-based Analysis ToolKit for efficient exploration of allele sharing patterns in present-day and ancient genomic datasets. G3 Genes|Genomes|Genetics (2021) doi:10.1093/g3journal/jkab357.

Summary statistics: D-statistics, **D**avid Reich statistic

Assume 4 populations, 1 sample per population, (no migration, population structure or admixture), and the following topology:



Summary statistics: D-statistics, **D**avid Reich statistic

Assume 4 populations, 1 sample per population, (no migration, population structure or admixture), and the following topology:

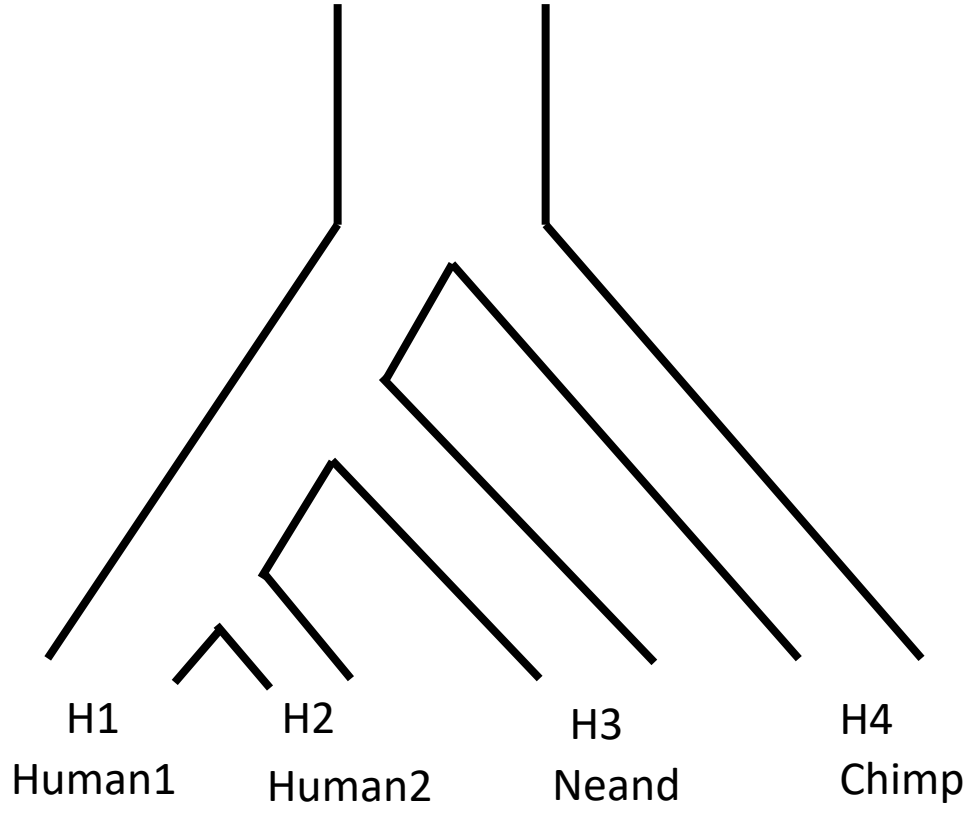
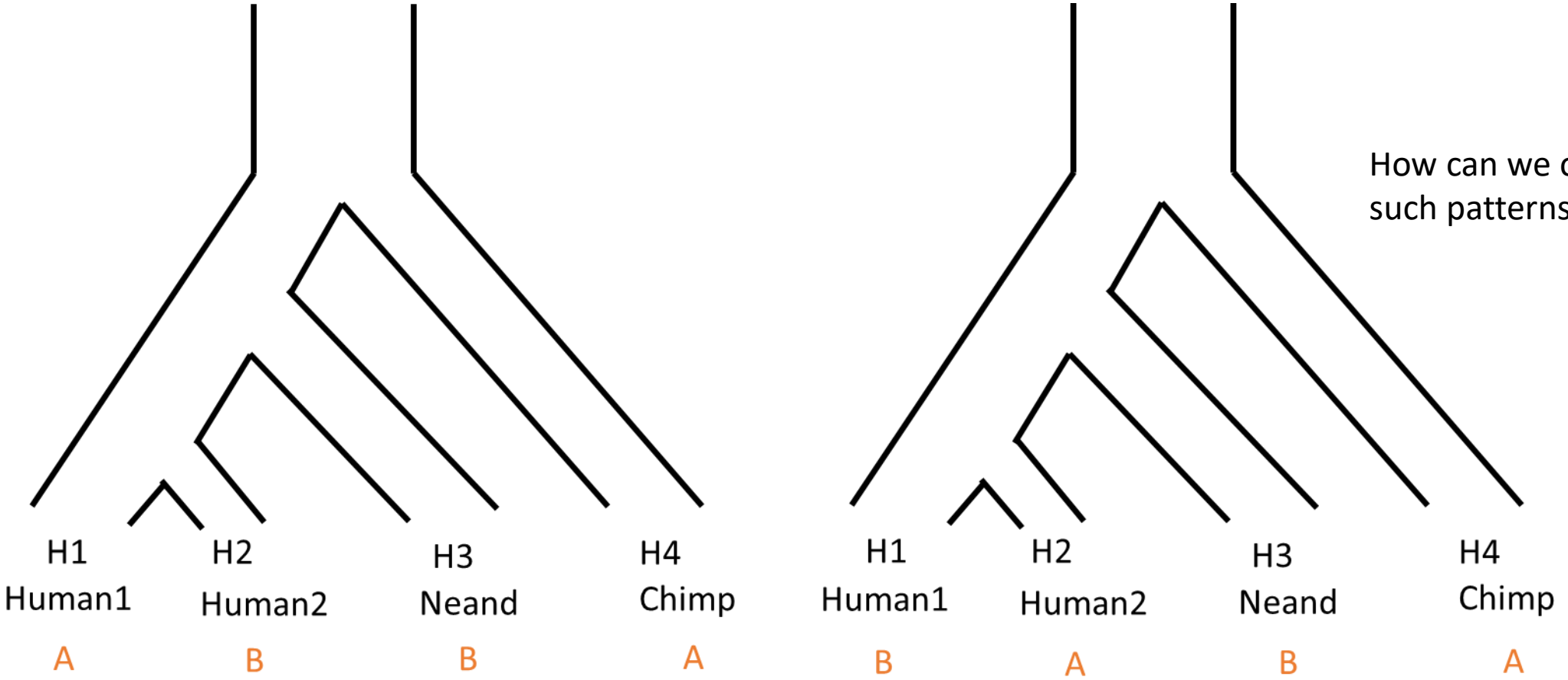


Illustration the statistics
for the human Neanderthal case

Summary statistics: D-statistics, **D**avid Reich statistic



How can we obtain such patterns?

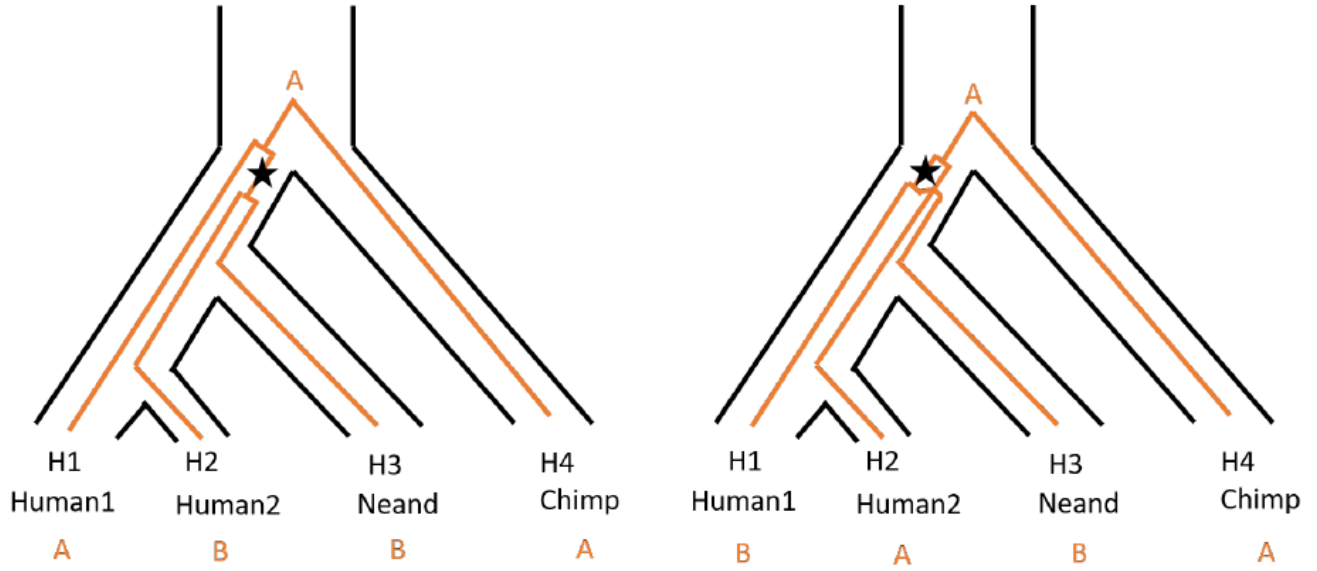
Denoting ABBA and BABA polymorphic sites that are the result of one mutation (infinite sites model) on the tree (as above).

Summary statistics: D-statistics, **D**avid Reich statistic

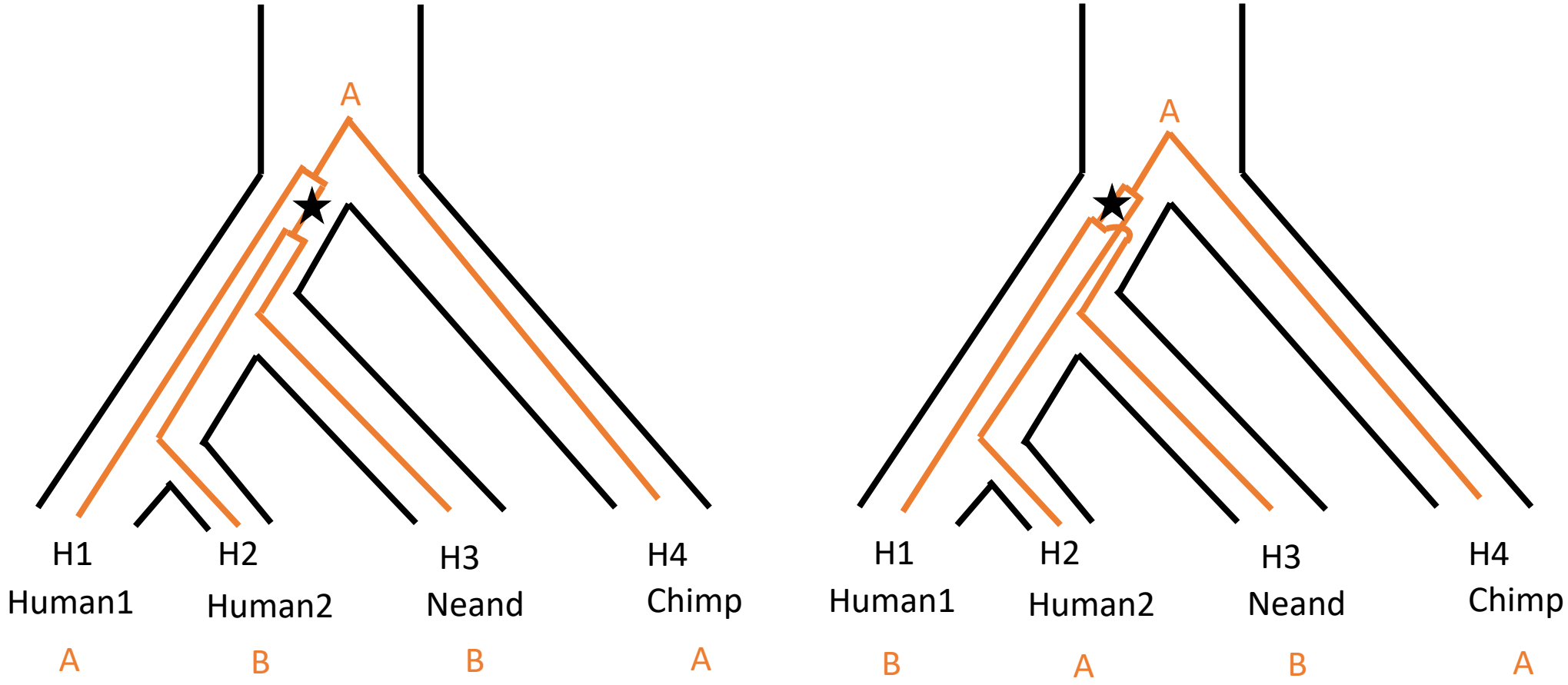
We define the D statistics as follows:

$$D(H_1, H_2; H_3, H_4) = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

where n_{ABBA} (resp. n_{BABA}) is the number of $ABBA$ (resp. $BABA$) across the genomes.



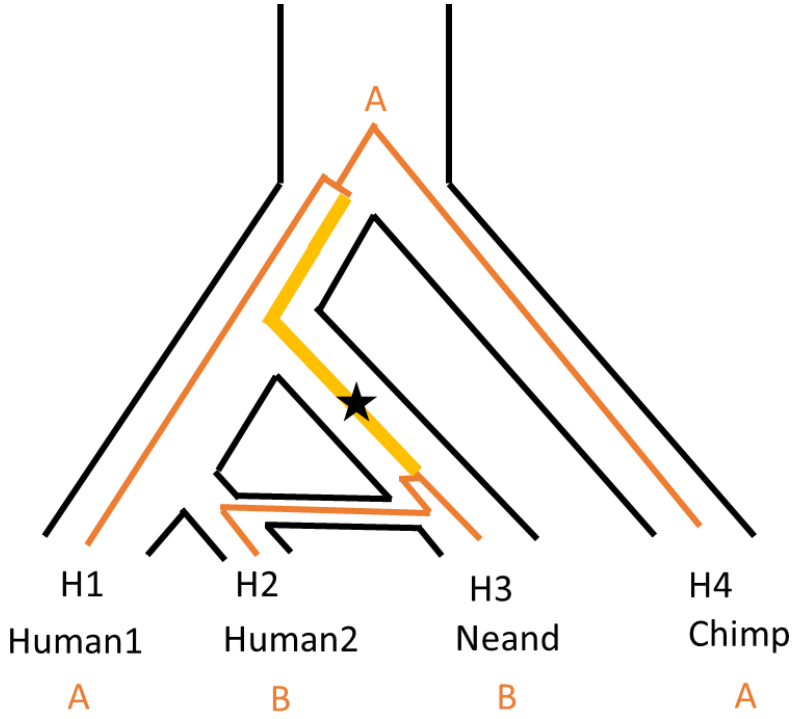
Which pattern is more frequent? I.e. is $D > 0$ or $D < 0$?


A
H1
Human1
A
H2
Human2
B
H3
Neand
B
H4
Chimp
A
A
H1
Human1
B
H2
Human2
A
H3
Neand
B
H4
Chimp
A

D>0? E.g., with admixture

$$D(H_1, H_2; H_3, H_4) = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

With some admixture:

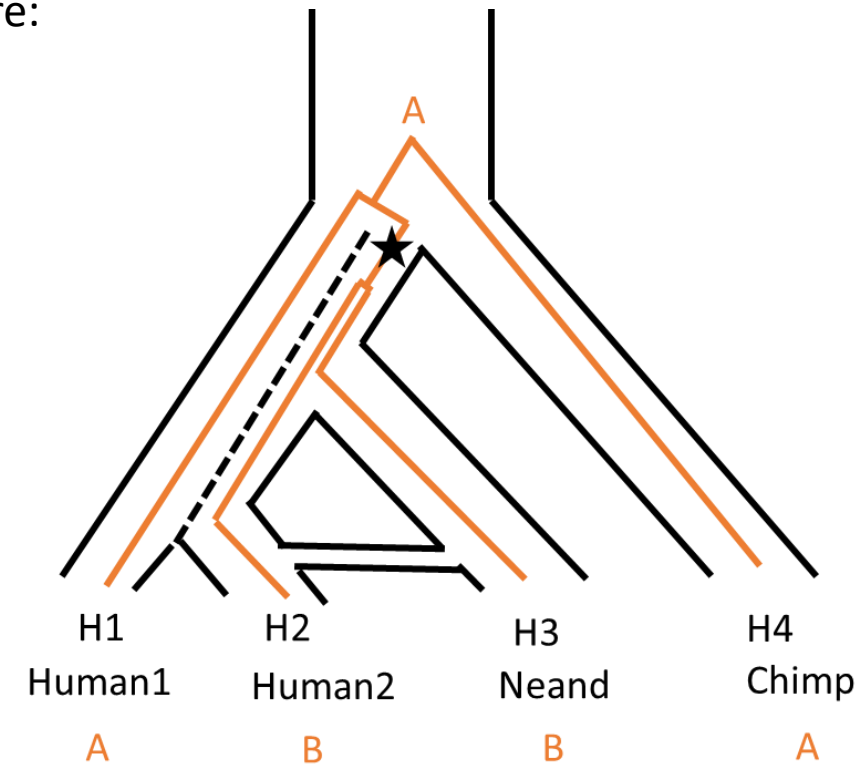


More ways to get ABBA than BABA: $P(\text{tree}_{ABBA}) > P(\text{tree}_{BABA})$

D>0? Or, population structure

$$D(H_1, H_2; H_3, H_4) = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

But also, with population structure:

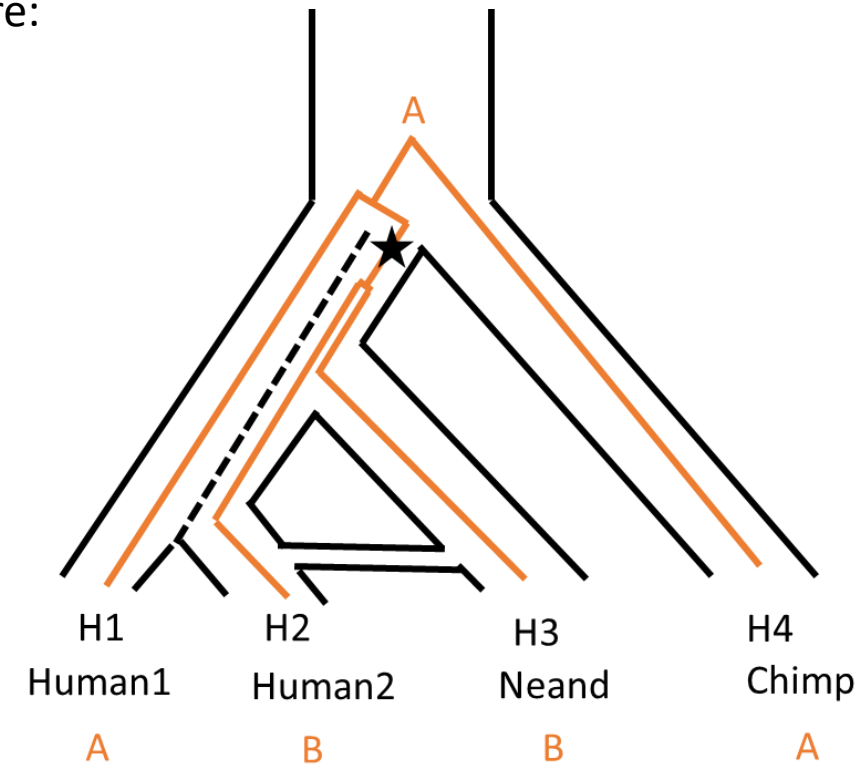


More ways to get ABBA than BABA: $P(\text{tree}_{ABBA}) > P(\text{tree}_{BABA})$

D>0? Or, population structure

$$D(H_1, H_2; H_3, H_4) = \frac{n_{ABBA} - n_{BABA}}{n_{ABBA} + n_{BABA}}$$

But also, with population structure:



More ways to get ABBA than BABA: $P(\text{tree}_{ABBA}) > P(\text{tree}_{BABA})$

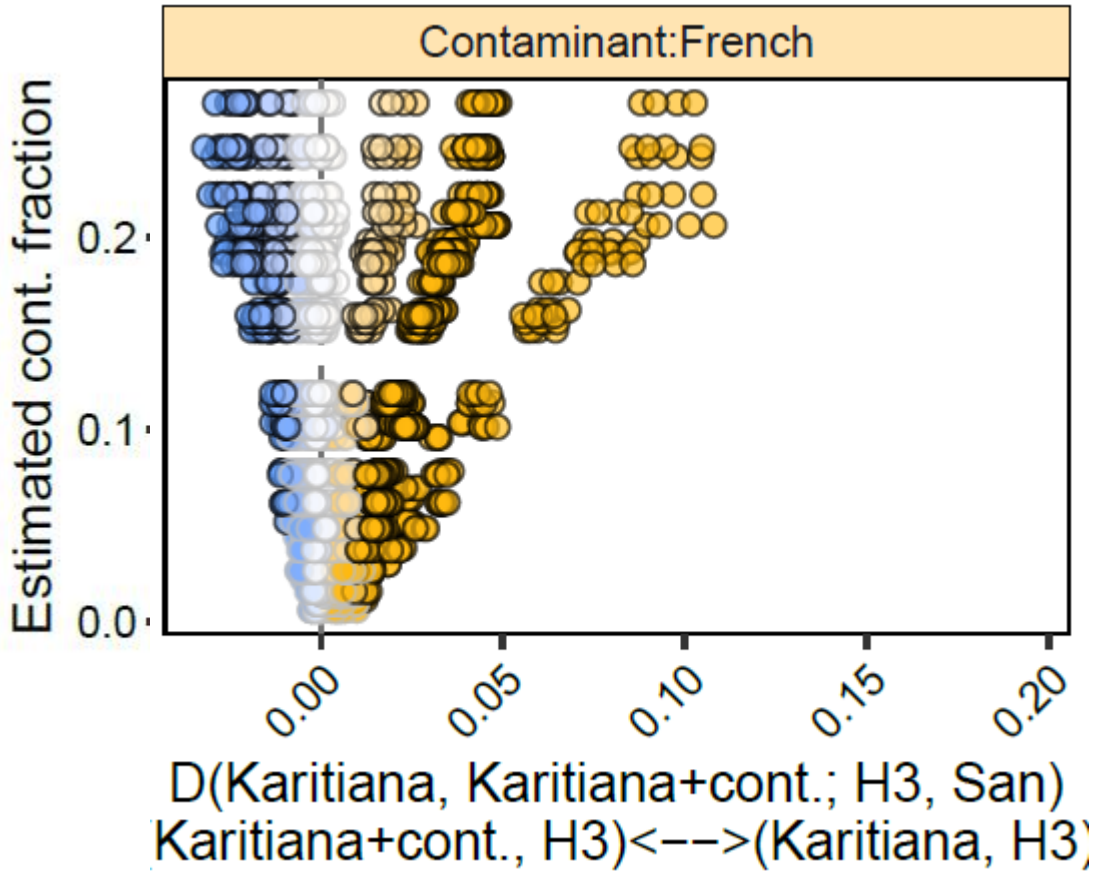
Dstats and ancient DNA

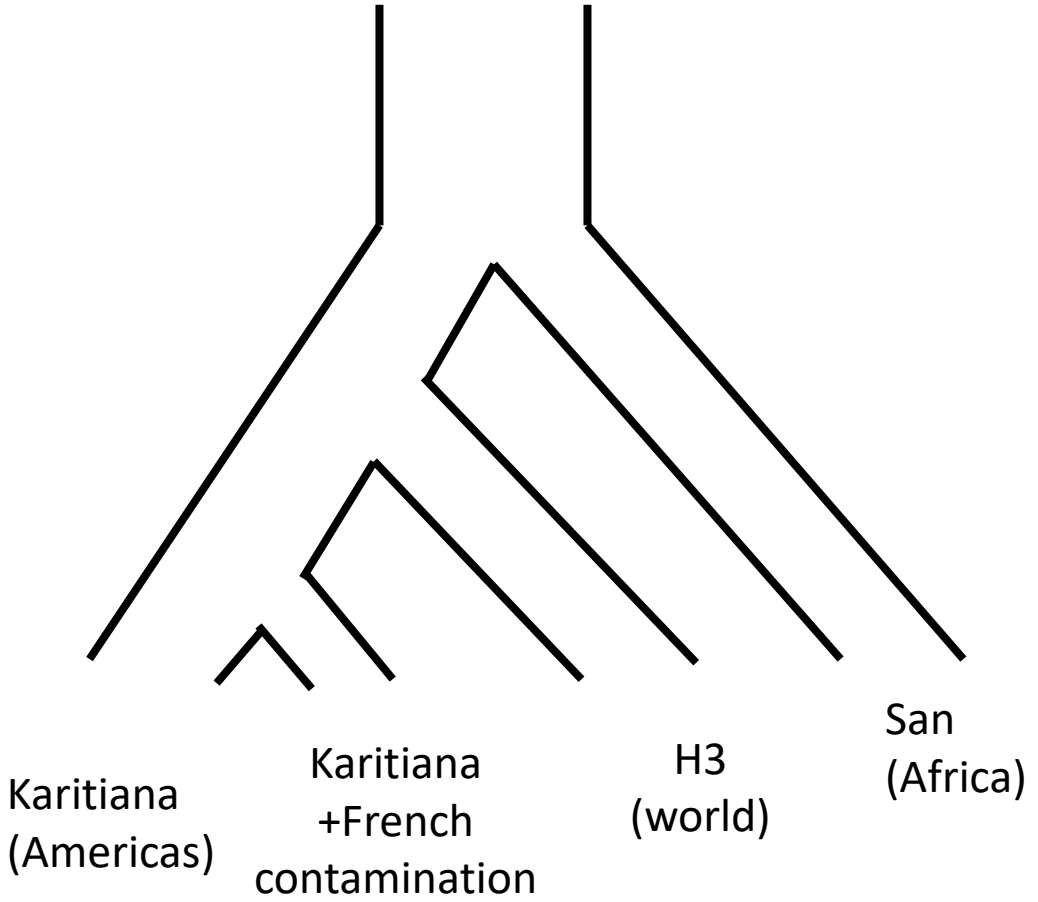
- Widely used for detecting gene flow
- Was (first) applied to ancient DNA to propose Neanderthal admixture (or population structure, or)

  Green et al. 2001

- Neat because: sampling a single read
- ABBA, BABA is less sensitive to error (would require to hit a polymorphic site or two errors)

Dstats and ancient DNA: back to contamination





contamination fraction as low as 2% could result in rejecting a true null hypothesis

Other options

[Population genetics: infer selection]

Ask Lucas ☺

[Phylogenetics]

[Environmental (eDNA)/metagenomics]

Ask Davide☺

Future directions

- [Wet lab developments]
- [Computational developments]

To wrap up:

Existing reviews☺!

1. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Pääbo, S. Ancient DNA. *Nat Rev Genet* **2**, 353–359 (2001).
2. Orlando, L. *et al.* Ancient DNA analysis. *Nat Rev Methods Primers* **1**, 1–26 (2021).
3. Slatkin, M. Statistical methods for analyzing ancient DNA from hominins. Current Opinion in Genetics & Development 41, 72–76 (2016)
4.Novembre, J. & Ramachandran, S. Perspectives on Human Population Structure at the Cusp of the Sequencing Era. Annual Review of Genomics and Human Genetics 12, 245–274 (2011).

I would suggest to only read them *after* have a first draft of your own to avoid unintentional “copy pastes”.

However, they are great and you should make sure to bring in a new angle if you want to publish your work (so eventually read them).