Ancient DNA

DTU Next Generation Sequencing Analysis

Martin Sikora, PhD

Globe Institute, University of Copenhagen

martin.sikora@sund.ku.dk

The past month in ancient DNA

DNA sleuths solve mystery of the 2,000-year old corpse

19 December 2023

BBC NEWS

DNA analysis showed that this young man travelled to Cambridgeshire from the furthest reaches of the Roman Empire 2,000 years ago

19th December 2023

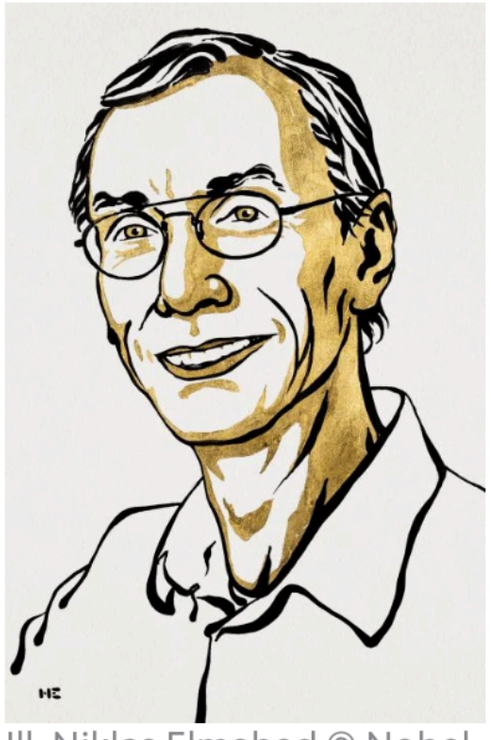
Nobel prize in Medicine 2022

The Nobel Prize in Physiology or Medicine 2022

Svante Pääbo

Share this

Svante Pääbo
Facts



Ill. Niklas Elmehed © Nobel Prize Outreach

Svante Pääbo
The Nobel Prize in Physiology or Medicine 2022

Born: 20 April 1955, Stockholm, Sweden

Affiliation at the time of the award: Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, Okinawa Institute of Science and Technology, Okinawa, Japan

Prize motivation: “for his discoveries concerning the genomes of extinct hominins and human evolution”

Prize share: 1/1

4th October 2022

1984 - the first 2 ancient DNA sequences

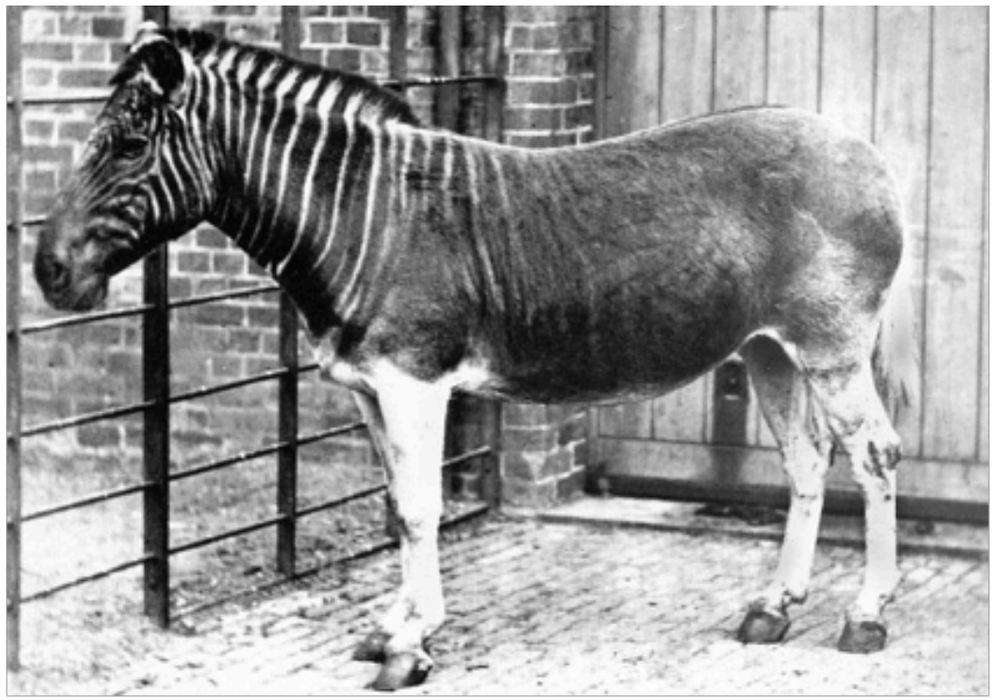
NATURE VOL. 312 15 NOVEMBER 1984

DNA sequences from the quagga, an extinct member of the horse family

Russell Higuchi*, Barbara Bowman*, Mary Freiberger*, Oliver A. Ryder† & Allan C. Wilson*

* Department of Biochemistry, University of California, Berkeley, California 94720, USA
† Research Department, San Diego Zoo, San Diego, California 92103, USA

To determine whether DNA survives and can be recovered from the remains of extinct creatures, we have examined dried muscle from a museum specimen of the quagga, a zebra-like species (*Equus quagga*) that became extinct in 1883 (ref. 1). We report that DNA



Unidentified reading frame 1

Quagga	C CCA ATC CTG CTC GCC GTA GCA TTC CTC ACA CTA GTT GAA CGA AAA GTC TTA GGC TAC ATA CAA CTT CGT AAA GGA CCC AAC ATC GTA GGC CCC TAT GGC CTA CTA CAA CGC ATT AC
Zebra	. . . . . T . . . . G . . . . . T . . . . . C . . . . . G*

Cytochrome oxidase I

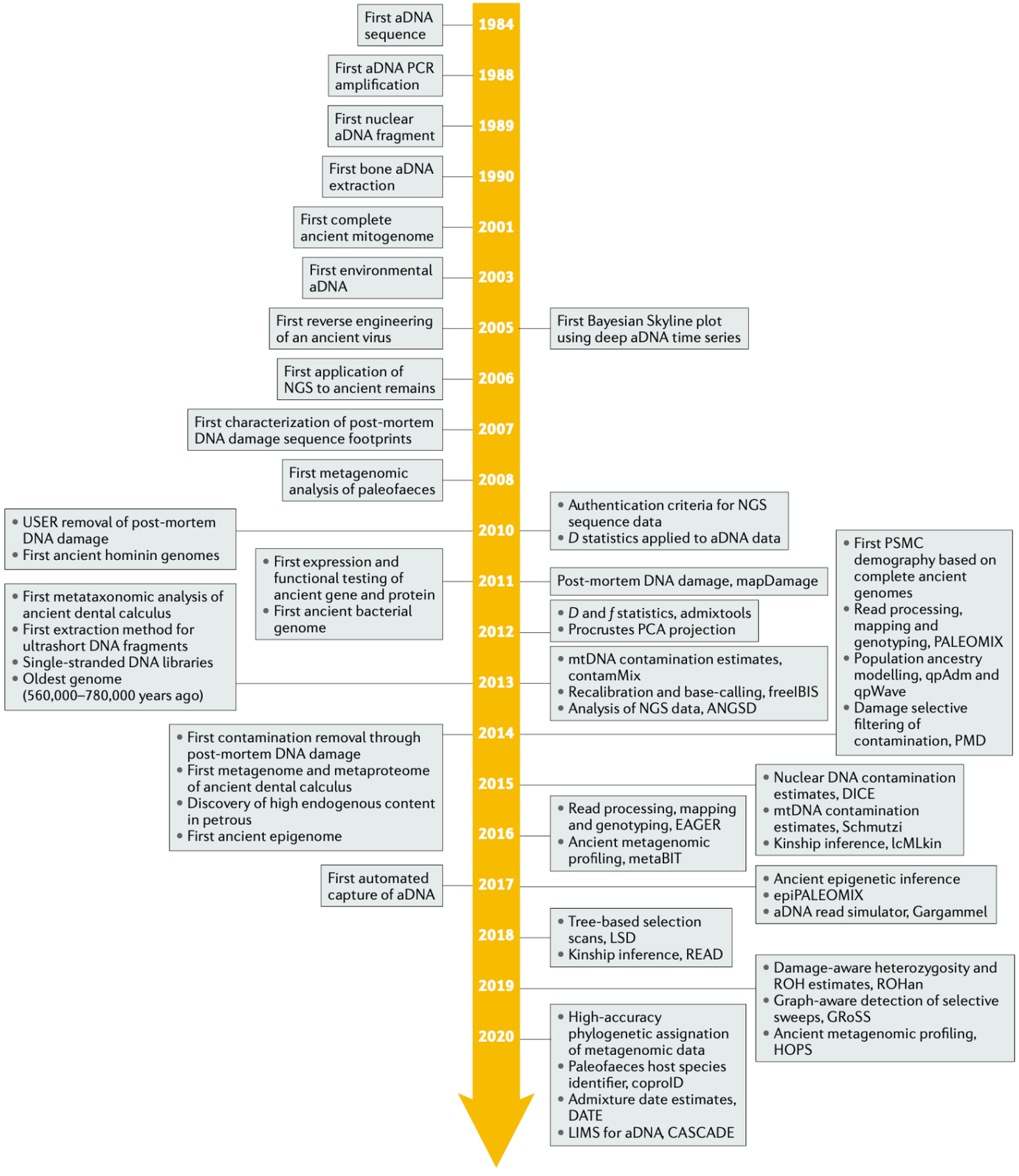
Quagga	A GGA GGA TTC GTT CAC TGA TTC CCT CTA TTC TCA GGA TAC ACA CTC AAC CAA ACC TGA GCA AAA ATT CAC TTT ACA ATT ATA TTC GTA GGG GTC AAC ATA ATT TTC TTC CCA
Zebra	G . . . . . T . . . . . G . . . . . C . . . . . A . . . . T . . . . C* . . . . .

Fig. 1 Sequences of the coding strands determined for two pieces of quagga mtDNA. The sequences are arranged in triplets corresponding to the amino acids that they encode. At 12 positions, the quagga sequences differ from those of mtDNA from a mountain zebra; only for these positions is the nature of the base specified for the zebra. The two asterisks identify triplets at which the zebra and quagga differ by an amino acid replacement.

Milestones in aDNA research

Wet lab

Dry lab



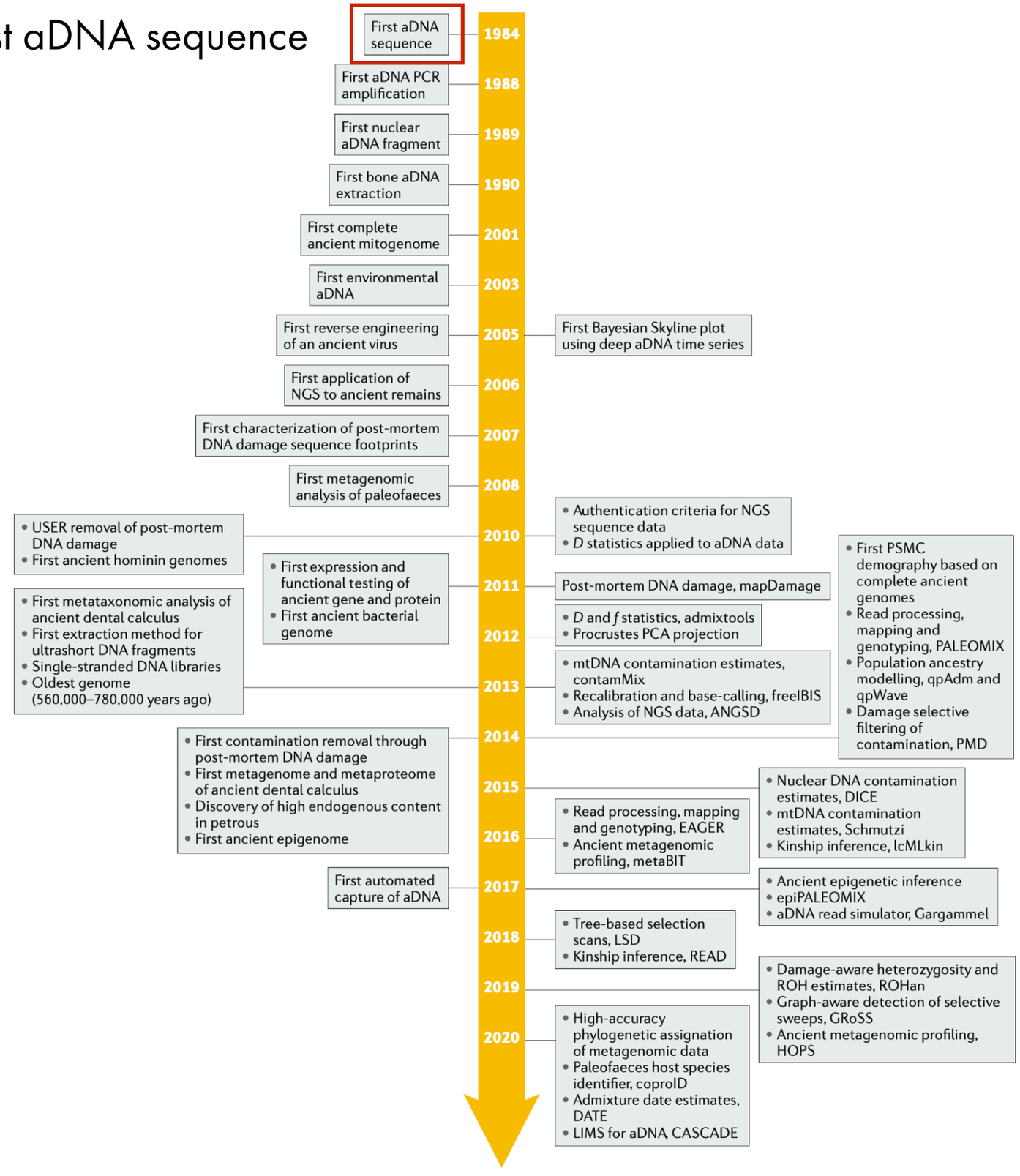
Orlando et al (2021) Nat Rev Methods Primers

Milestones in aDNA research

Wet lab

Dry lab

First aDNA sequence



Orlando et al (2021) Nat Rev Methods Primers

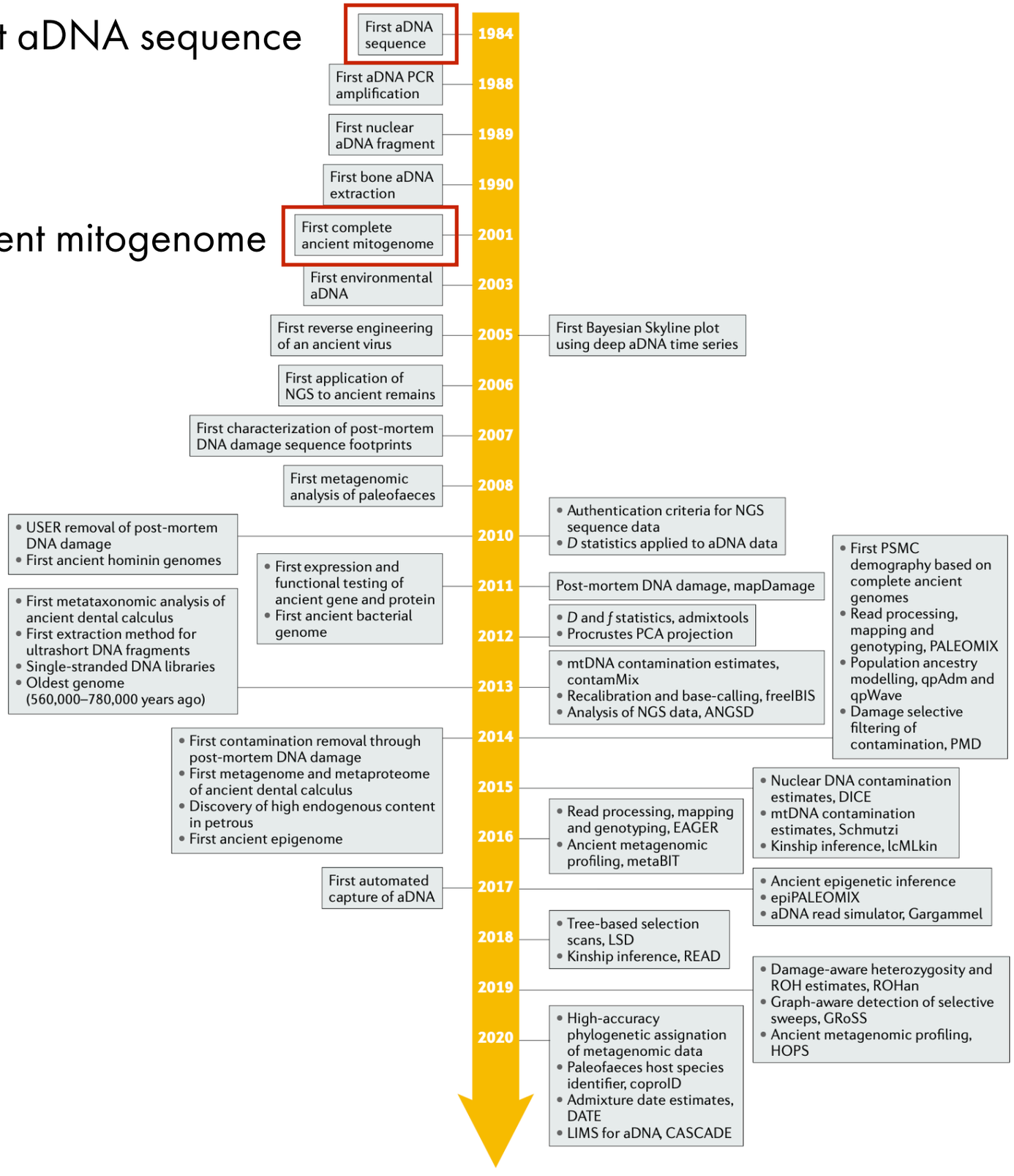
Milestones in aDNA research

Wet lab

Dry lab

First aDNA sequence

First ancient mitogenome



Orlando et al (2021) Nat Rev Methods Primers

Milestones in aDNA research

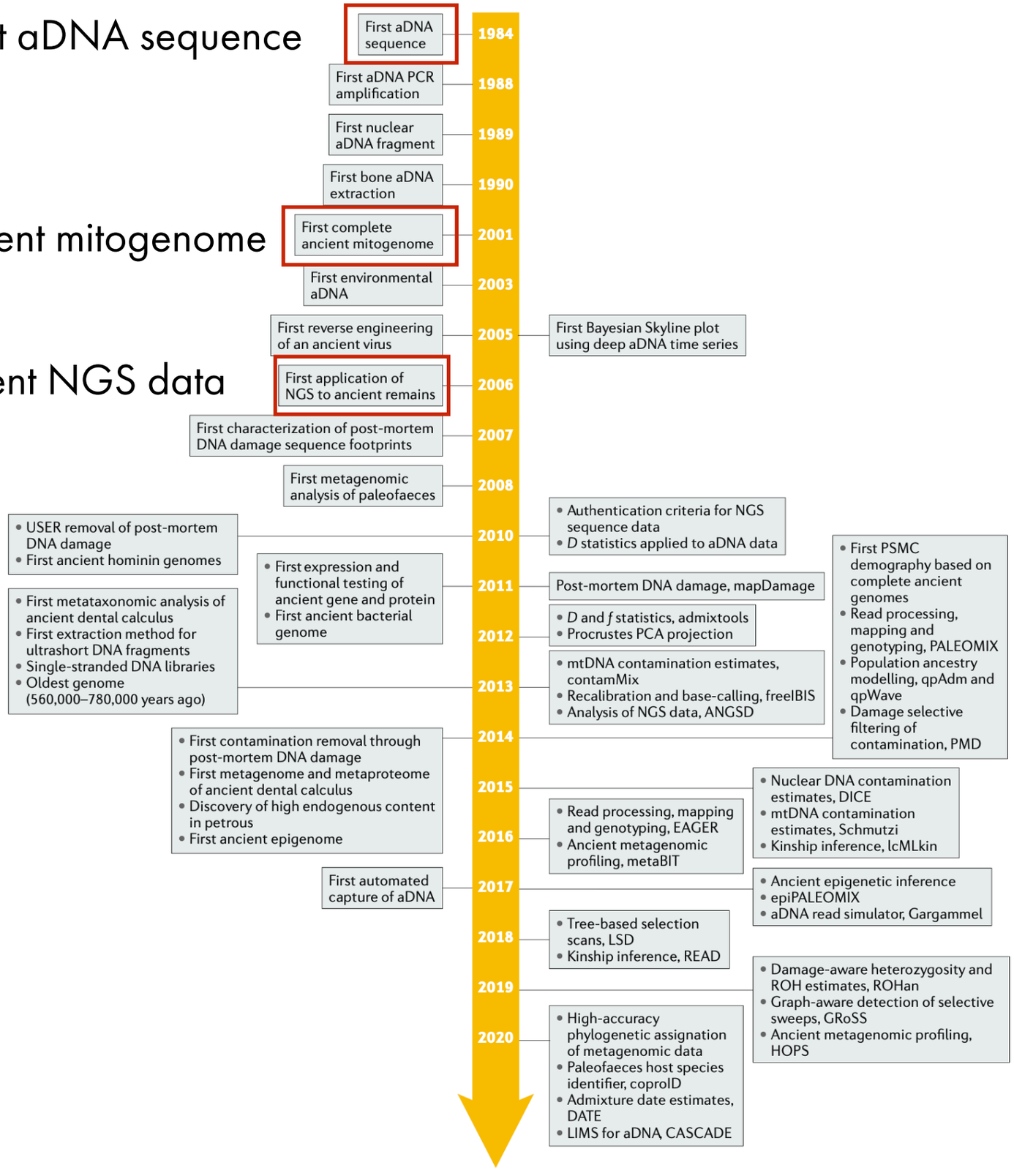
Wet lab

Dry lab

First aDNA sequence

First ancient mitogenome

First ancient NGS data



Orlando et al (2021) Nat Rev Methods Primers

Milestones in aDNA research

Wet lab

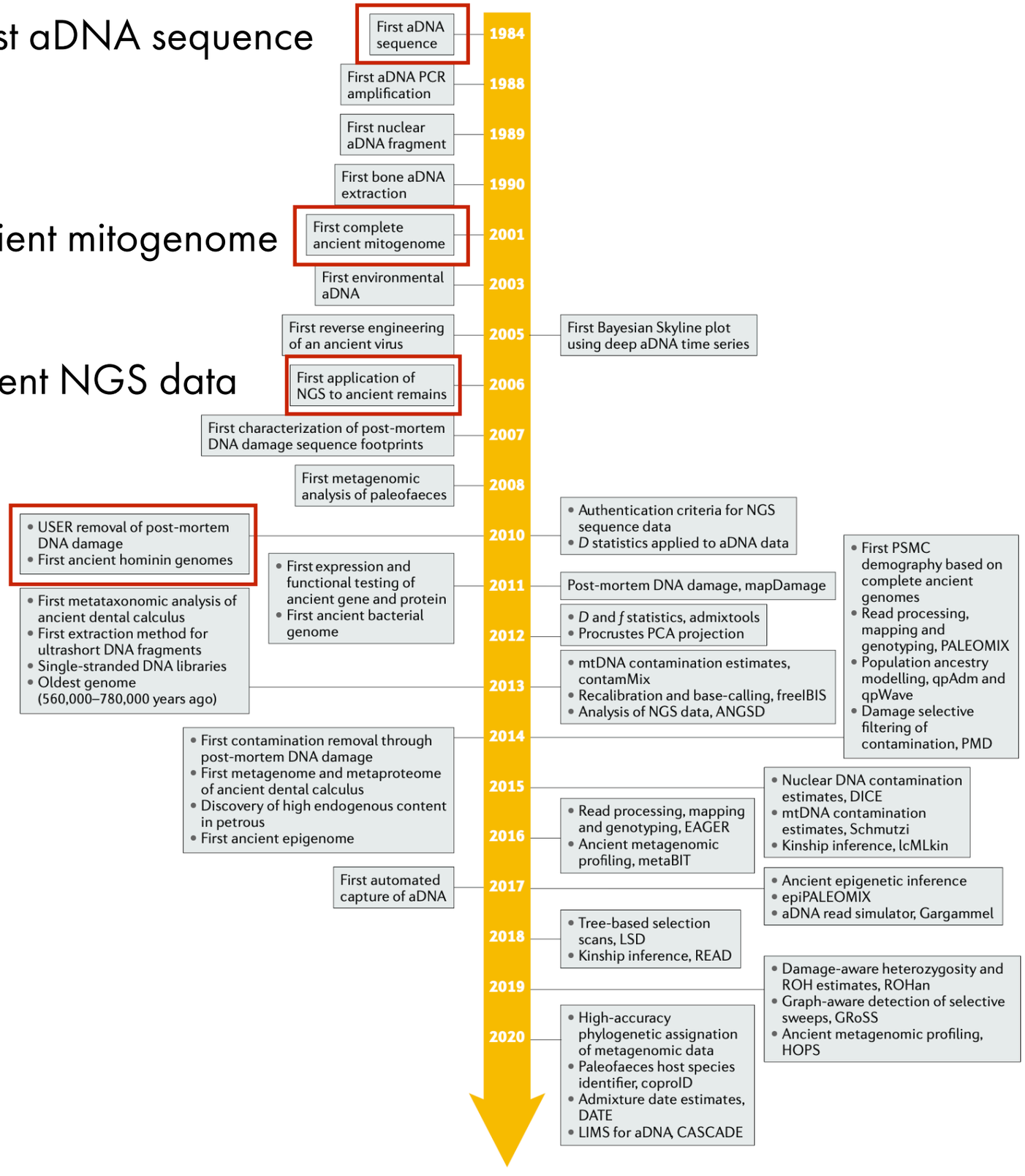
Dry lab

First aDNA sequence

First ancient mitogenome

First ancient NGS data

First ancient human genome



Orlando et al (2021) Nat Rev Methods Primers

Milestones in aDNA research

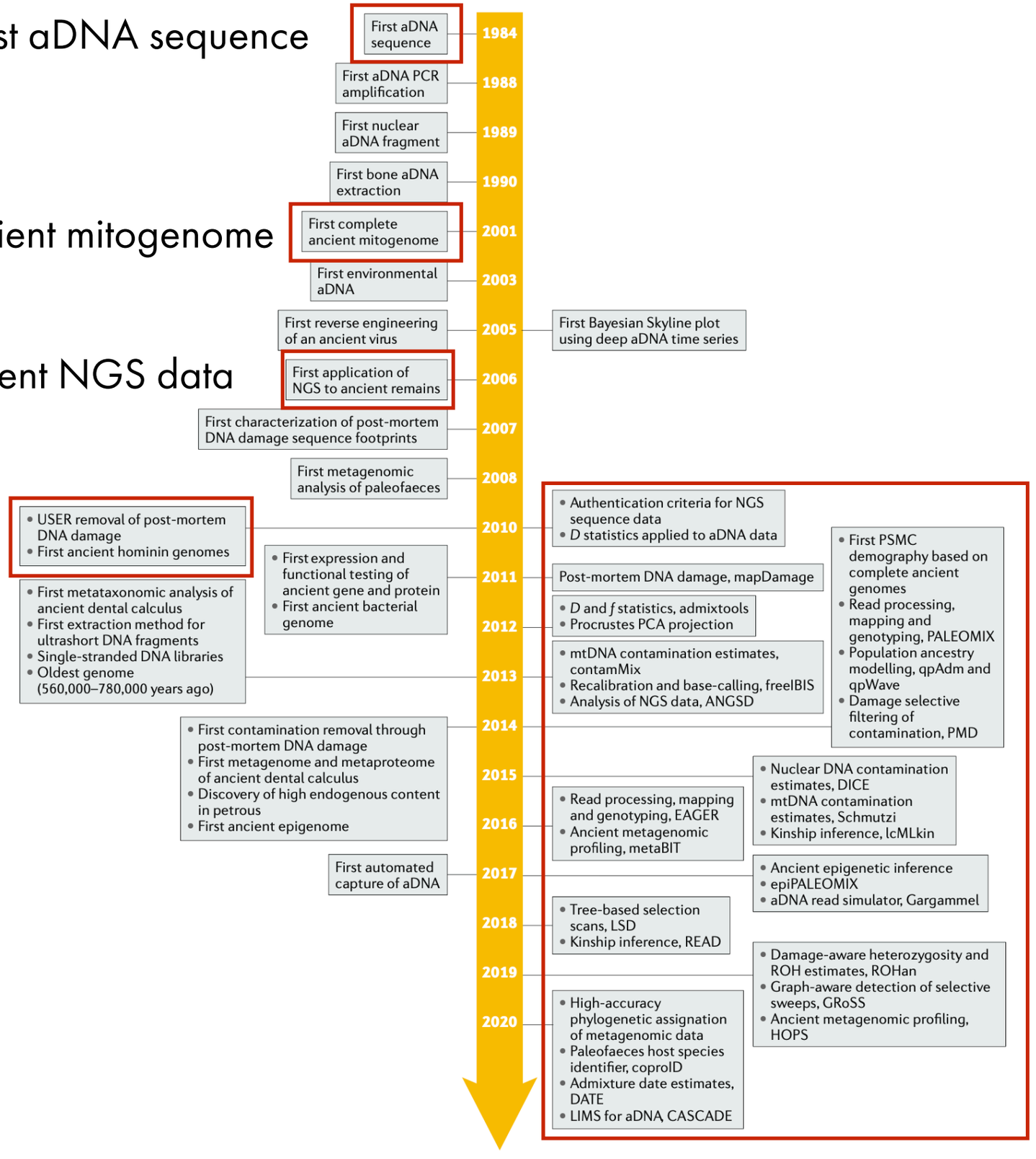
Wet lab

First aDNA sequence

First ancient mitogenome

First ancient NGS data

First ancient human genome

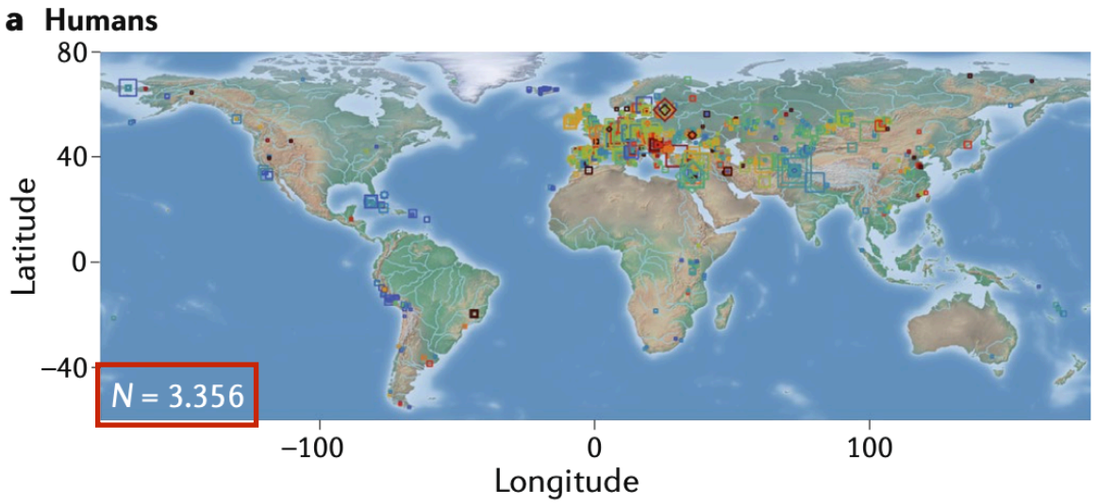


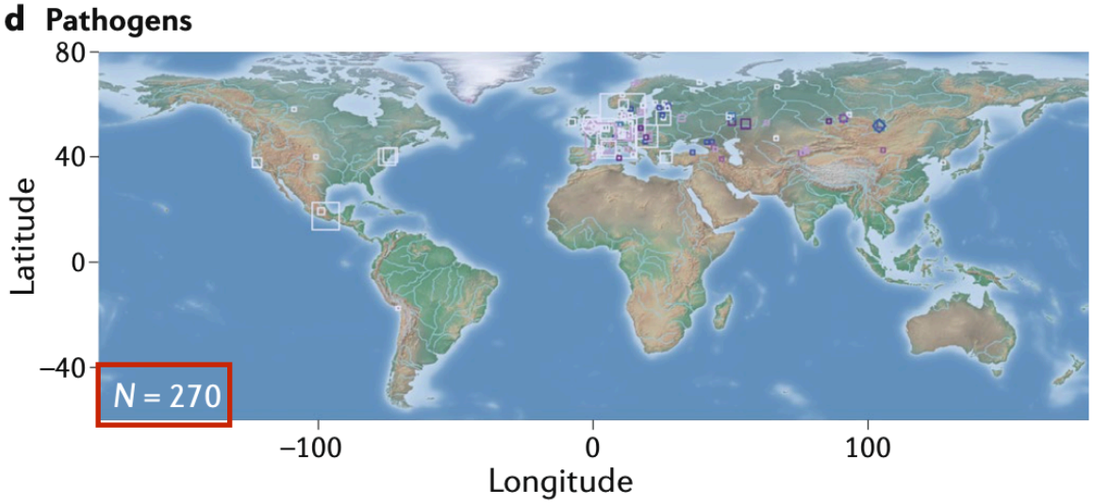
Dry lab

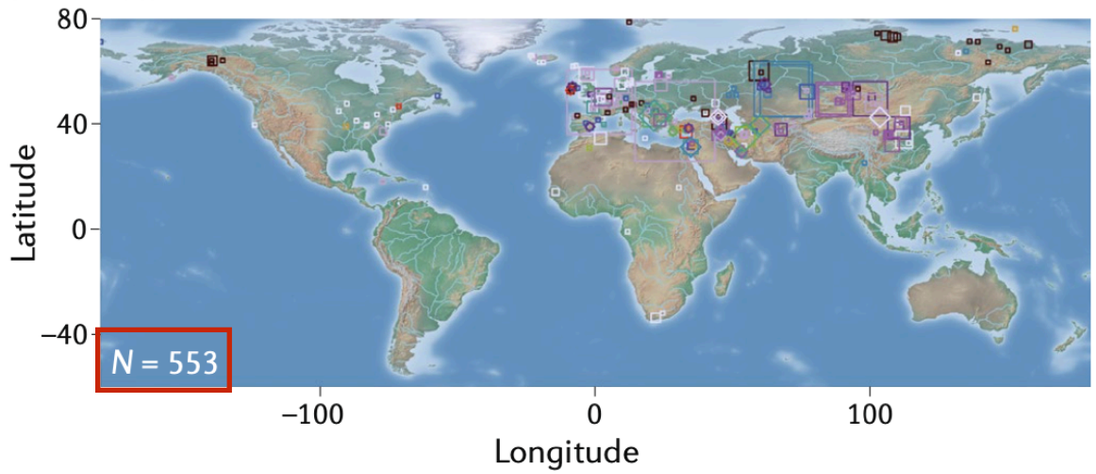
Rich analysis toolkit development after 2010

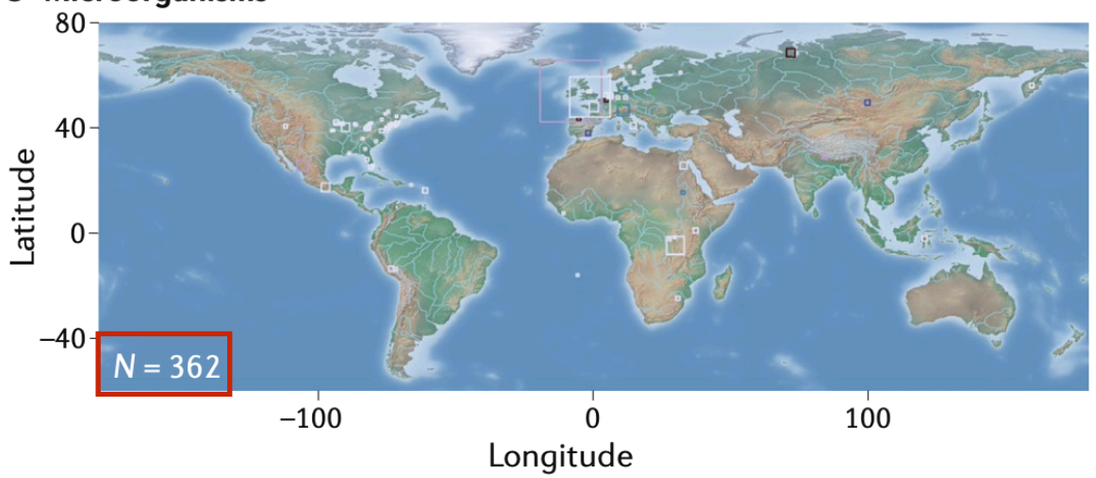
Orlando et al (2021) Nat Rev Methods Primers

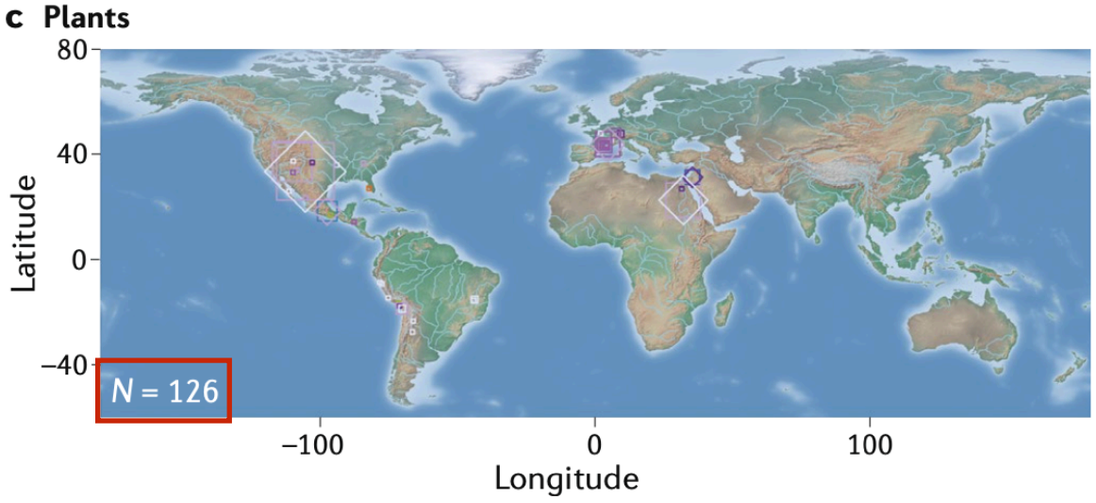
2021 - the floodgates are open





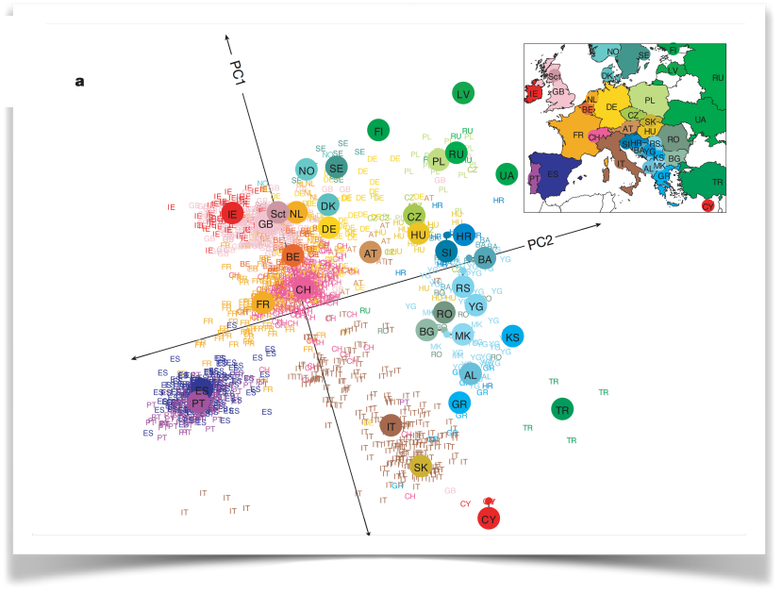




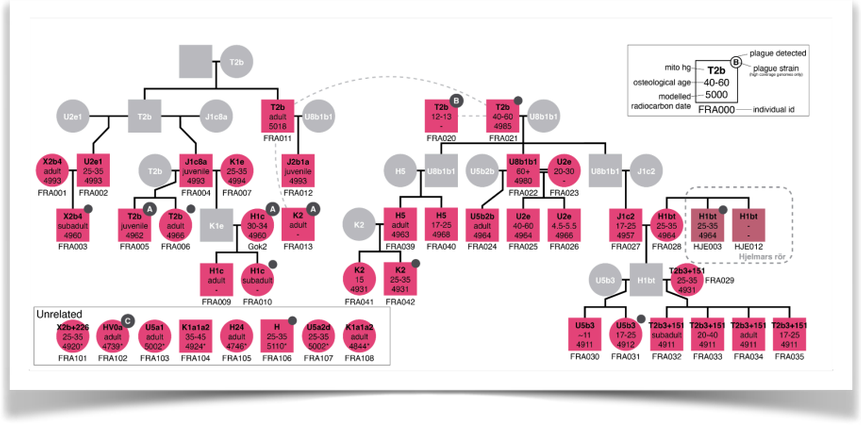


Orlando et al (2021) Nat Rev Methods Primers

Characteristics and challenges of ancient DNA



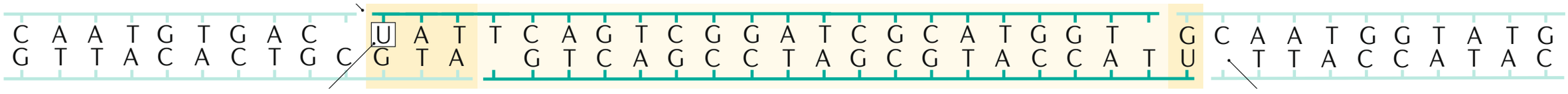
Ancient DNA population genomics



Ancient DNA studies showcase

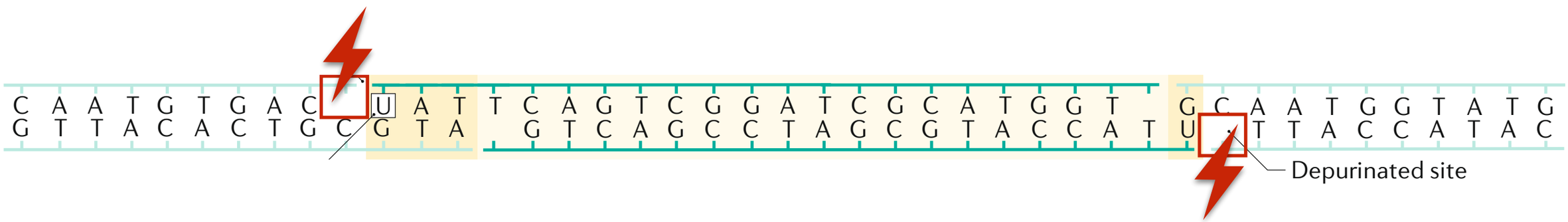
Characteristics and challenges of aDNA data

Characteristics of ancient DNA



Orlando et al (2021) Nat Rev Methods Primers

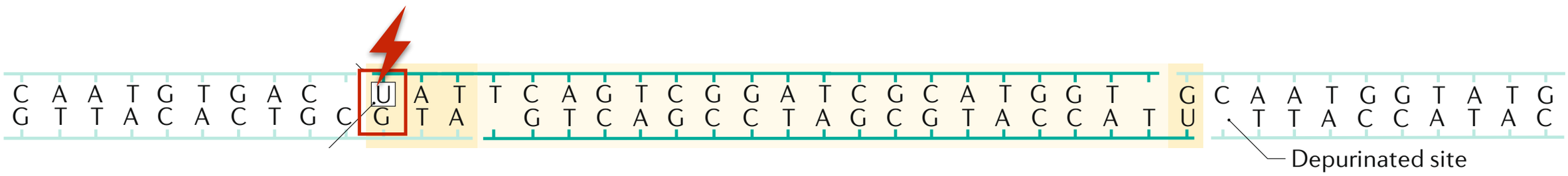
Characteristics of ancient DNA



Highly fragmented - short molecules (< 100bp)

Orlando et al (2021) Nat Rev Methods Primers

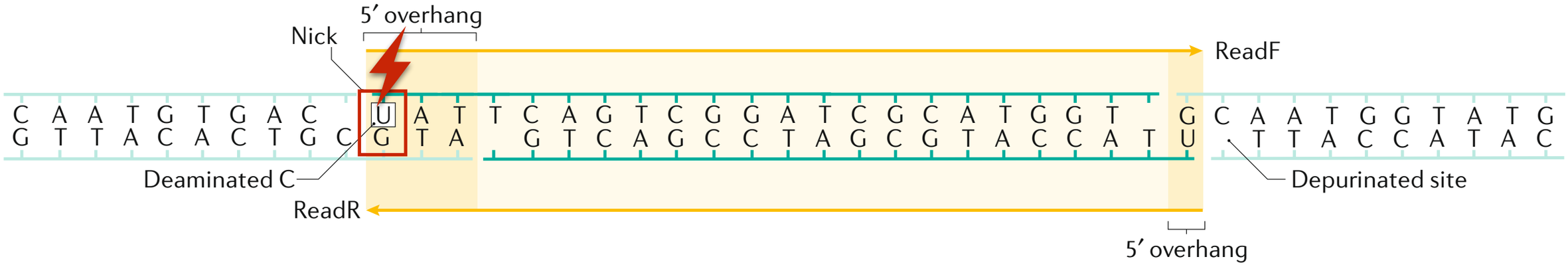
Characteristics of ancient DNA

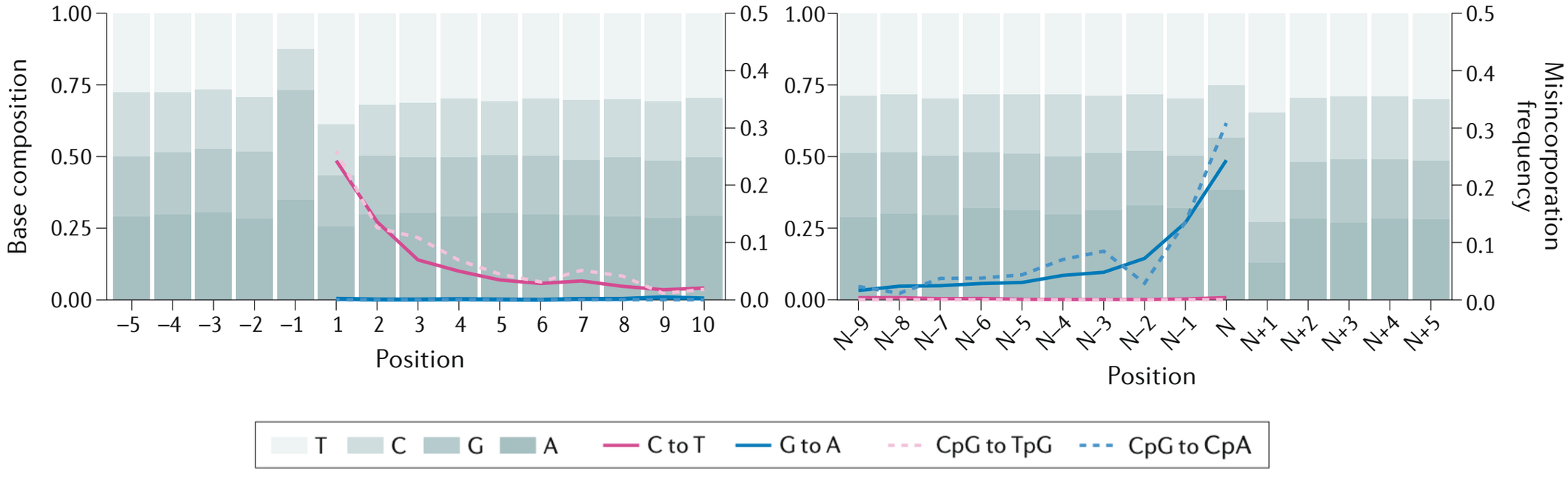


Post-mortem DNA damage

Orlando et al (2021) Nat Rev Methods Primers

Characteristics of ancient DNA





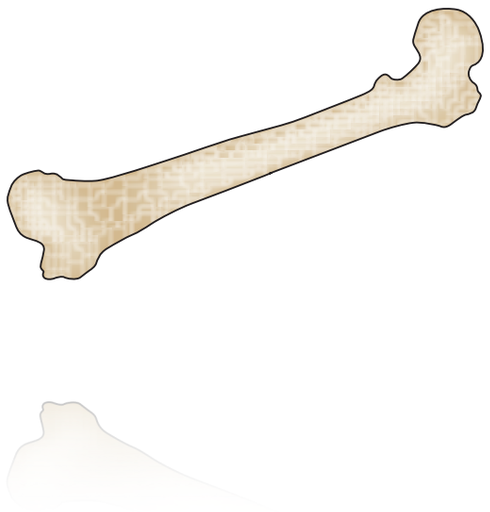
Excess of purine bases upstream of 5′ read start

Increased rates of C>T substitutions towards 5′ read ends

Orlando et al (2021) Nat Rev Methods Primers

Ancient DNA workflow

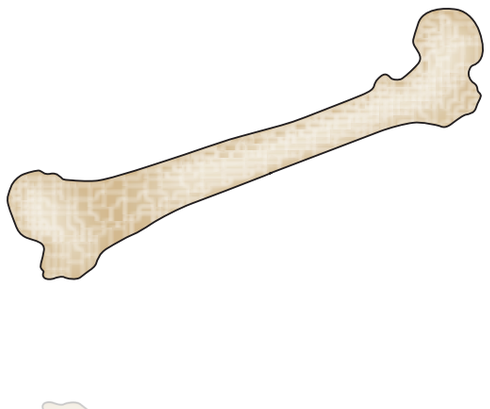
Sample



Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

Ancient DNA workflow

Sample



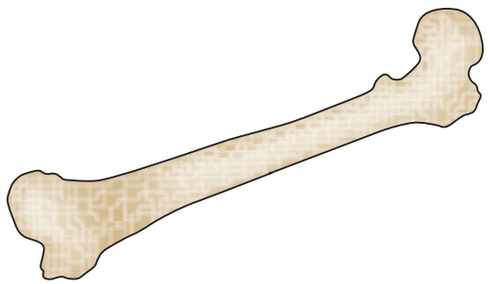
DNA sequences

Data generation (wet lab)

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

Ancient DNA workflow

Sample



DNA sequences

Analysis results

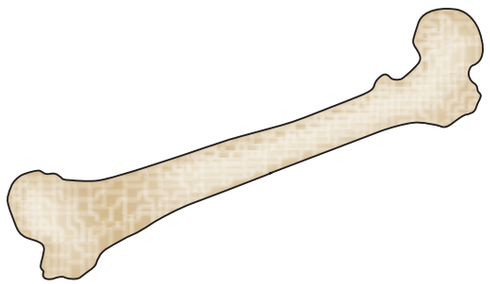
Data generation (wet lab)

Data analysis (dry lab)

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

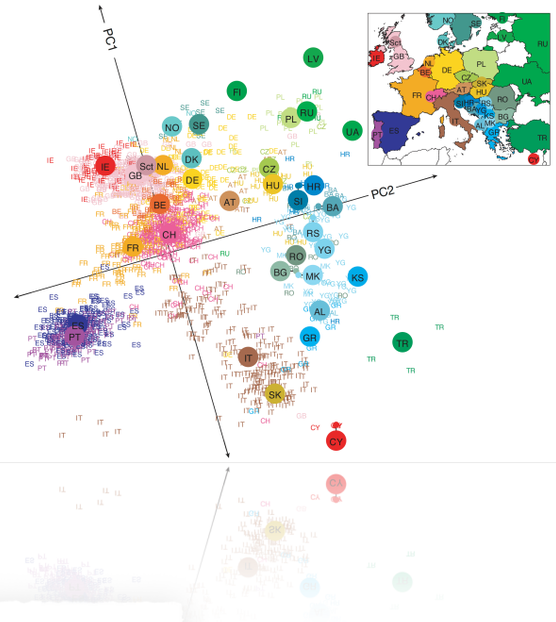
Ancient DNA workflow

Sample



DNA sequences

Analysis results



Data generation (wet lab)

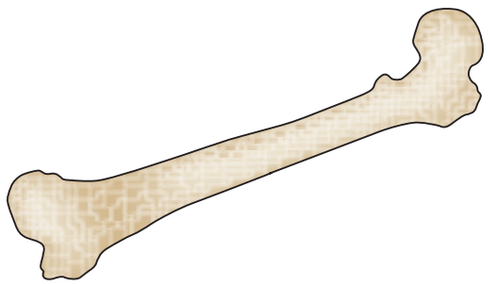
DNA preservation
Endogenous content
Sample / tissue availability

Data analysis (dry lab)

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

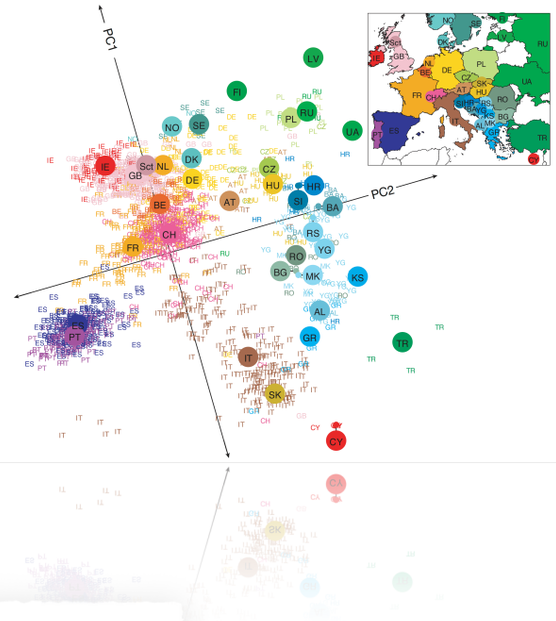
Ancient DNA workflow

Sample



DNA sequences

Analysis results



Data generation (wet lab)

DNA preservation
Endogenous content
Sample / tissue availability

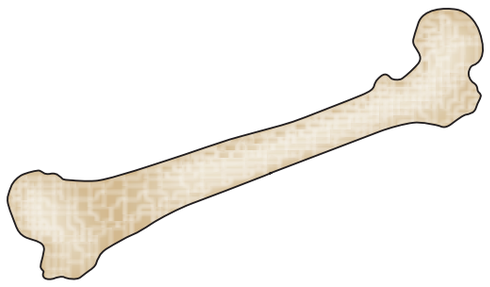
Data analysis (dry lab)

DNA mixtures / contamination
Genome coverage
Error rates

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

Ancient DNA workflow

Sample



DNA sequences

Analysis results

Data generation (wet lab)

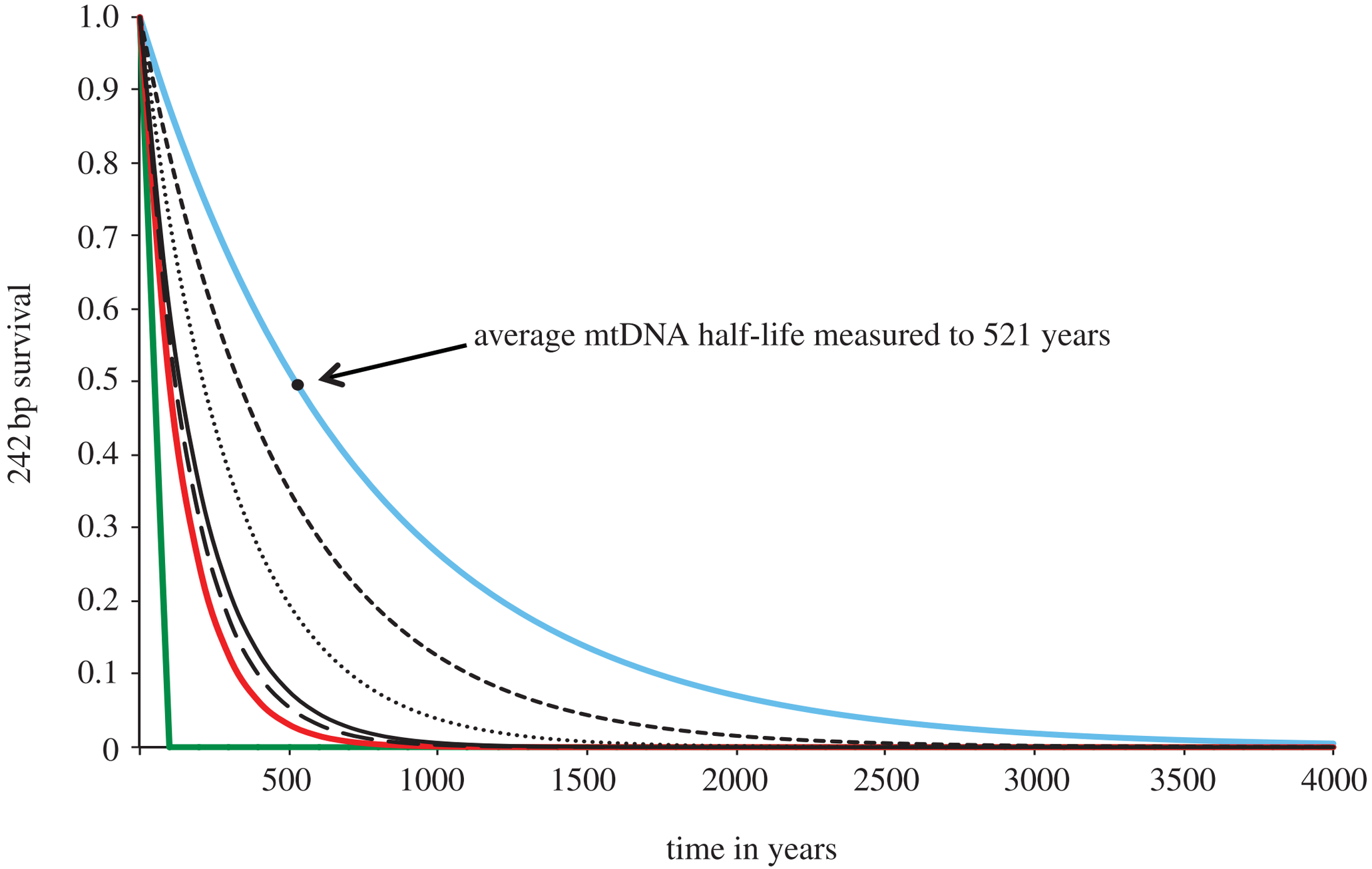
DNA preservation
Endogenous content
Sample / tissue availability

Data analysis (dry lab)

DNA mixtures / contamination
Genome coverage
Error rates

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

Wet lab challenges - DNA preservation



Ancient DNA fragmentation follows exponential decay

Allentoft et al (2012) Proc R Soc B

Wet lab challenges - endogenous DNA

Endogenous DNA

Other DNA

Wet lab challenges - endogenous DNA

Low endogenous

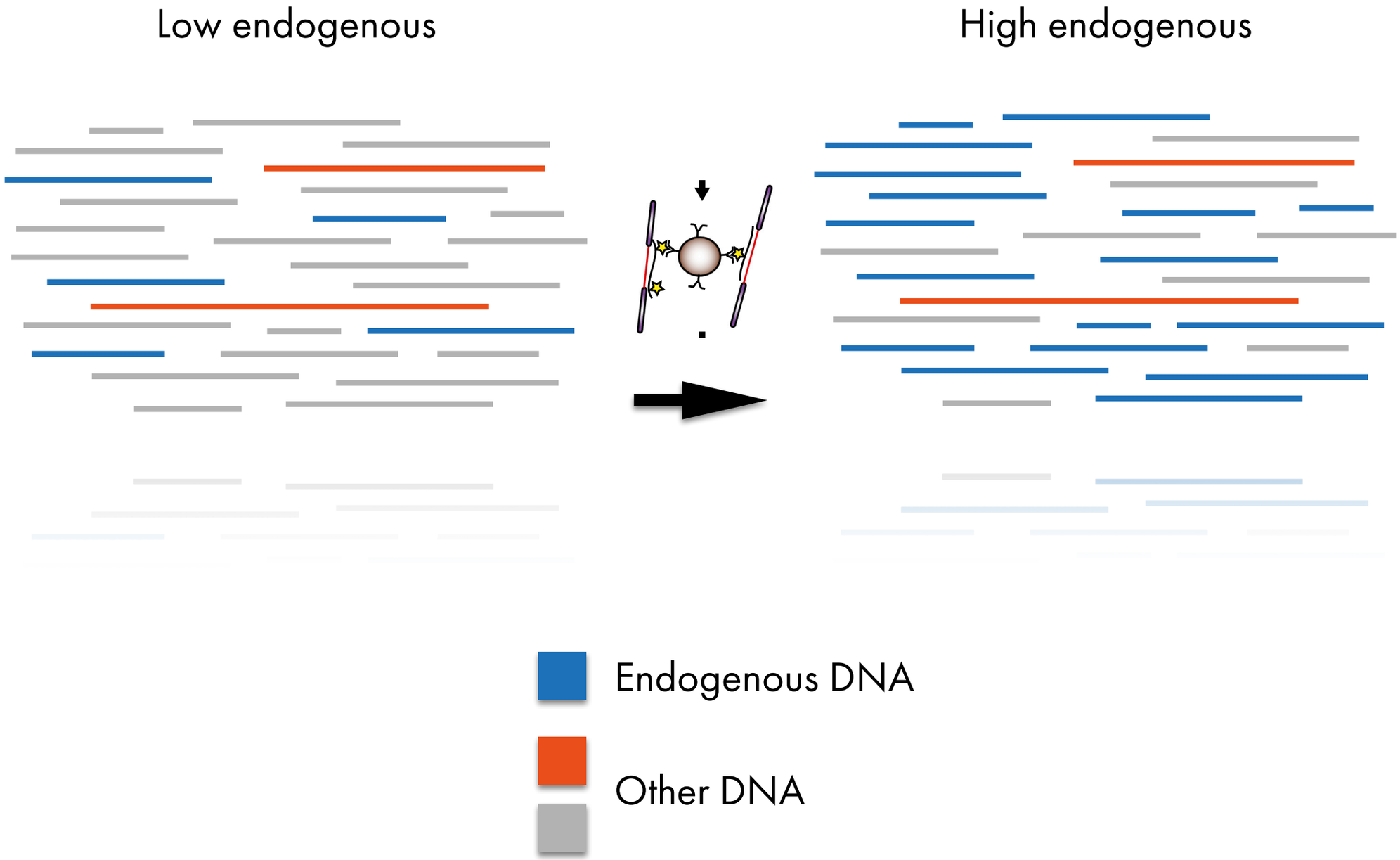
High endogenous

Endogenous DNA

Other DNA

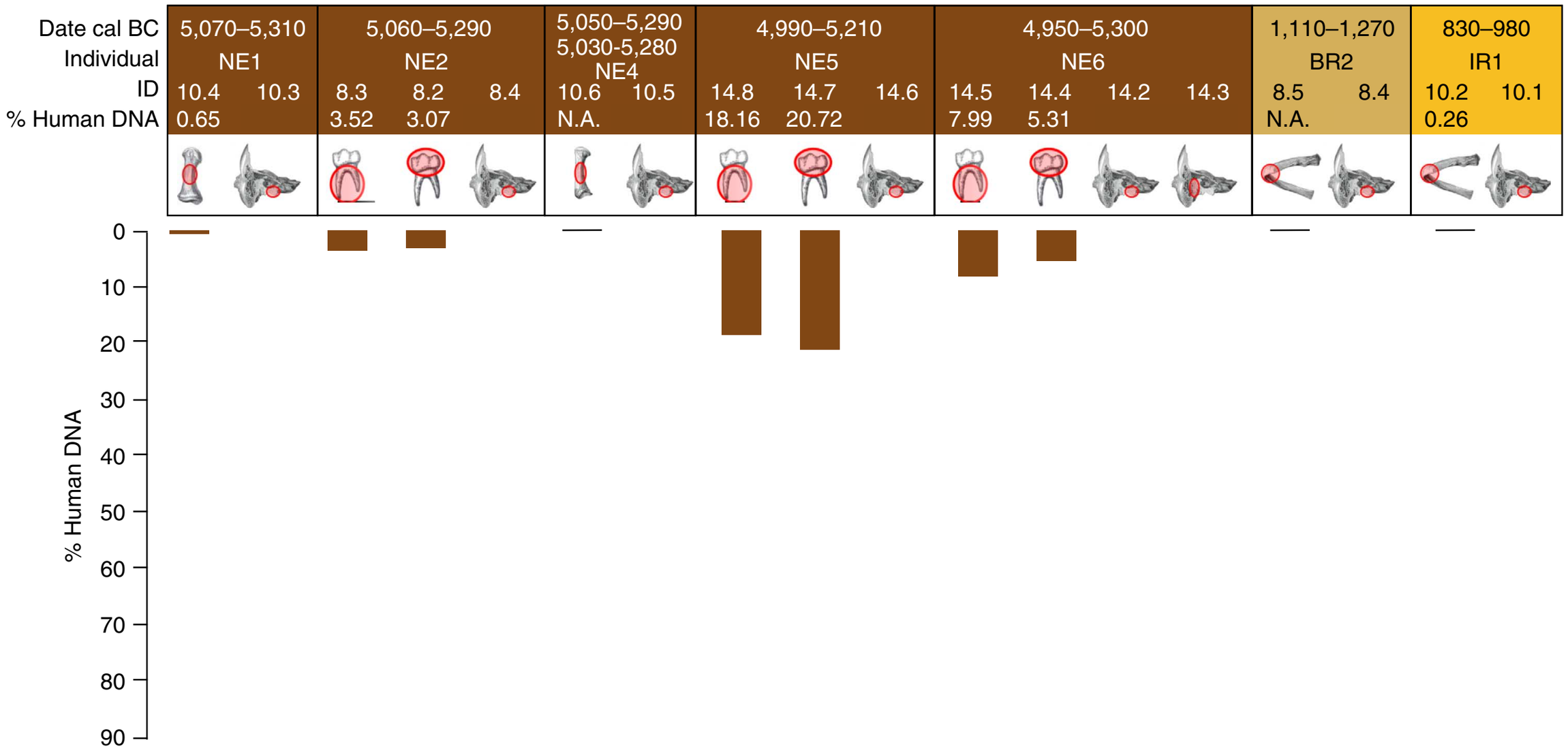
Endogenous DNA content varies substantially between samples

Wet lab challenges - endogenous DNA



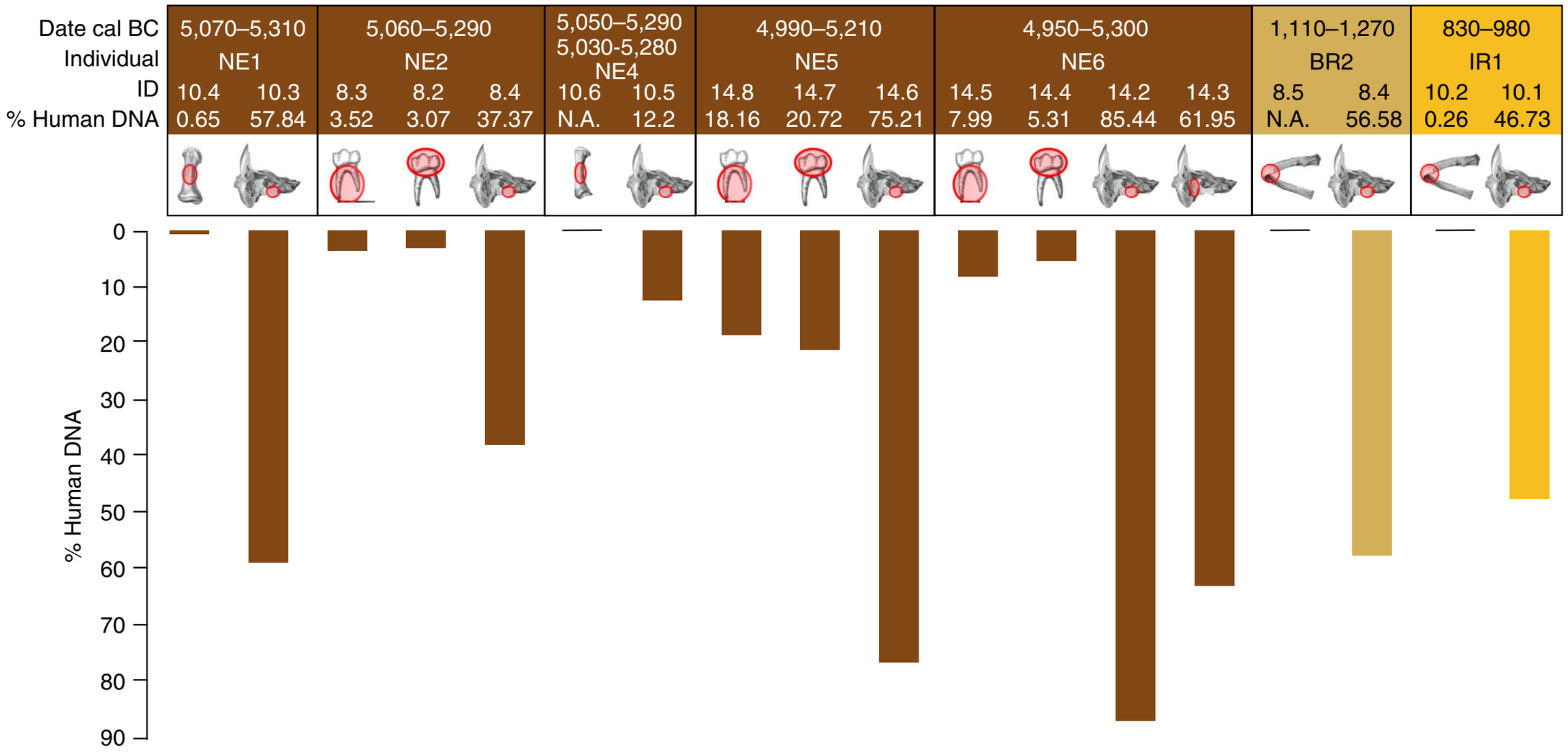
DNA capture enrichment (targeted or whole genome) to increase endogenous DNA

Wet lab challenges - endogenous DNA



Gamba et al (2014) Nat Comm

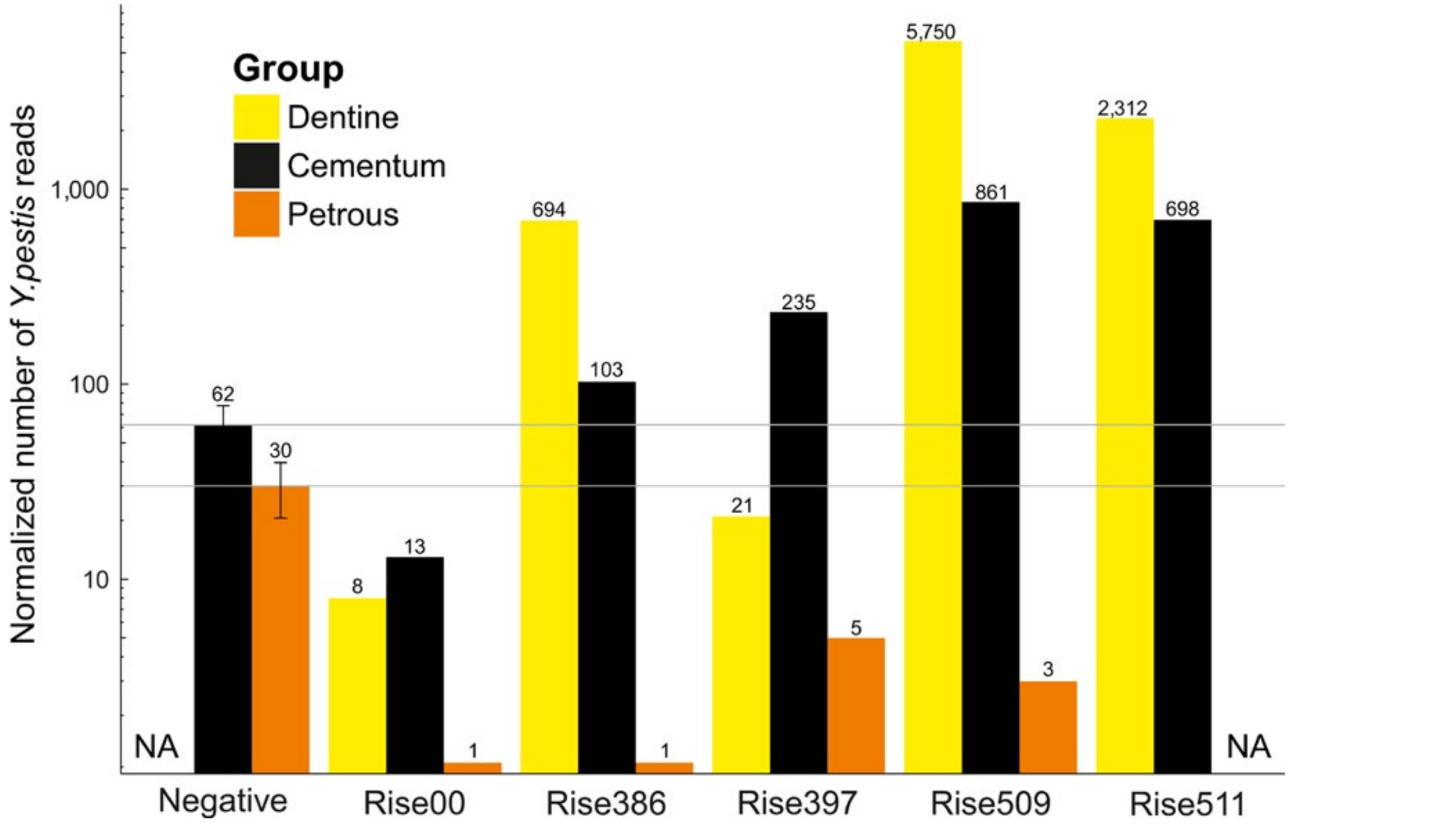
Wet lab challenges - endogenous DNA



High endogenous human DNA yield from petrous portion of temporal bone

Gamba et al (2014) Nat Comm

Wet lab challenges - tissue availability



Presence of pathogen DNA in different tissues

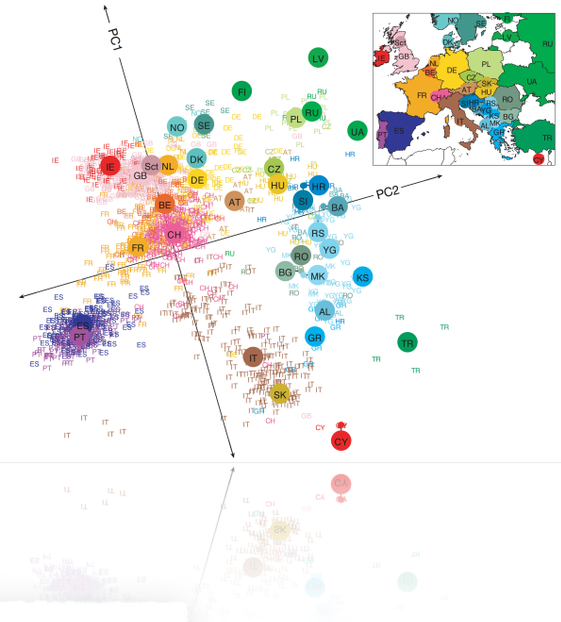
Margaryan et al (2018) Ecol Evol

Ancient DNA workflow

Sample

DNA sequences

Analysis results



Data generation (wet lab)

DNA preservation
Endogenous content
Sample / tissue availability

Data analysis (dry lab)

DNA mixtures / contamination
Genome coverage
Error rates

Stoneking and Krause (2011) Nat Rev Genet; Novembre et al (2008) Nature

Dry lab challenges - DNA mixture

Endogenous ancient DNA

Dry lab challenges - DNA mixture

Endogenous ancient DNA

Modern DNA contamination

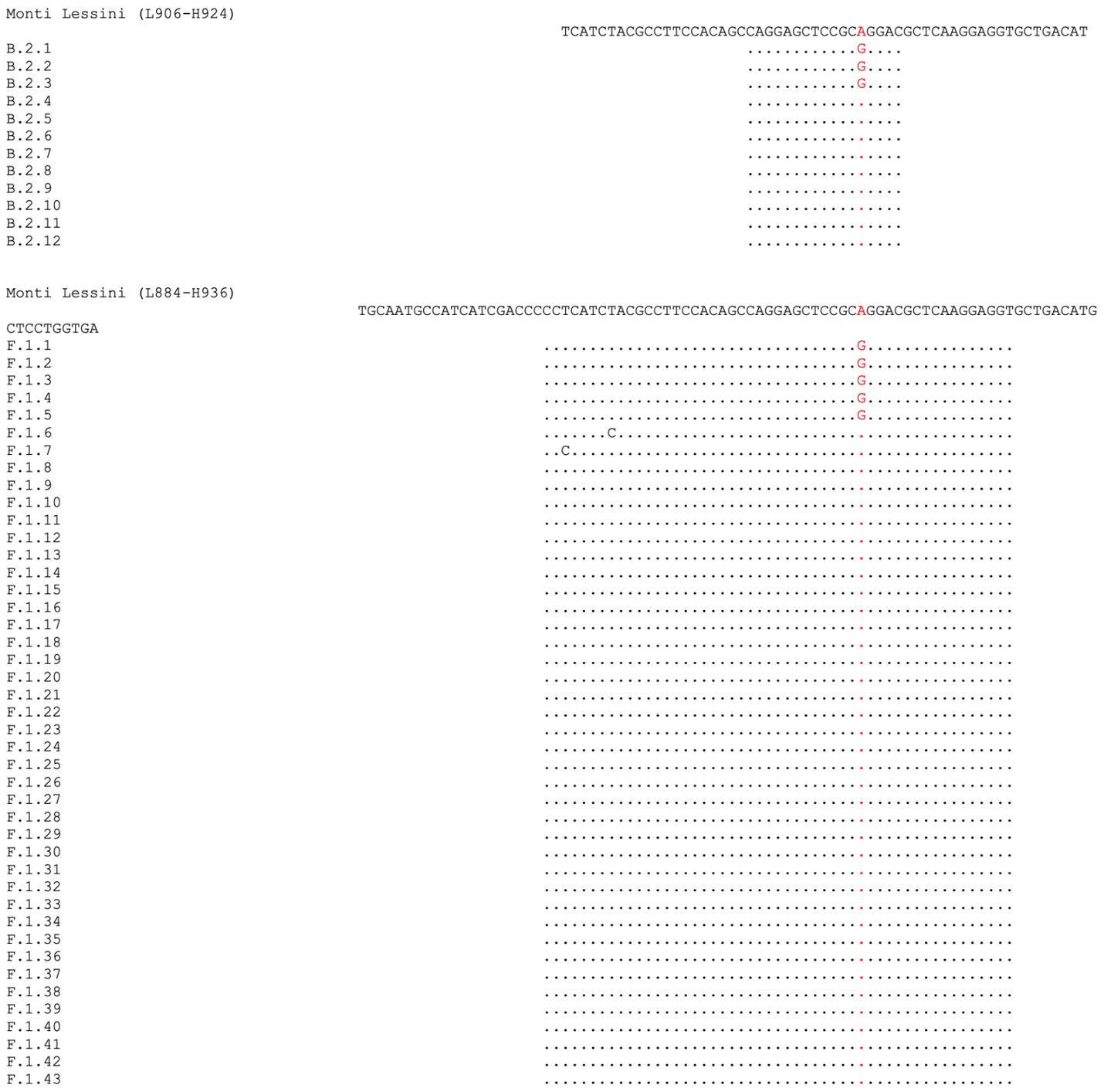
Dry lab challenges - DNA mixture

Ancient DNA data is metagenomic data

Dry lab challenges - DNA contamination

Barcelona lab

Florence lab

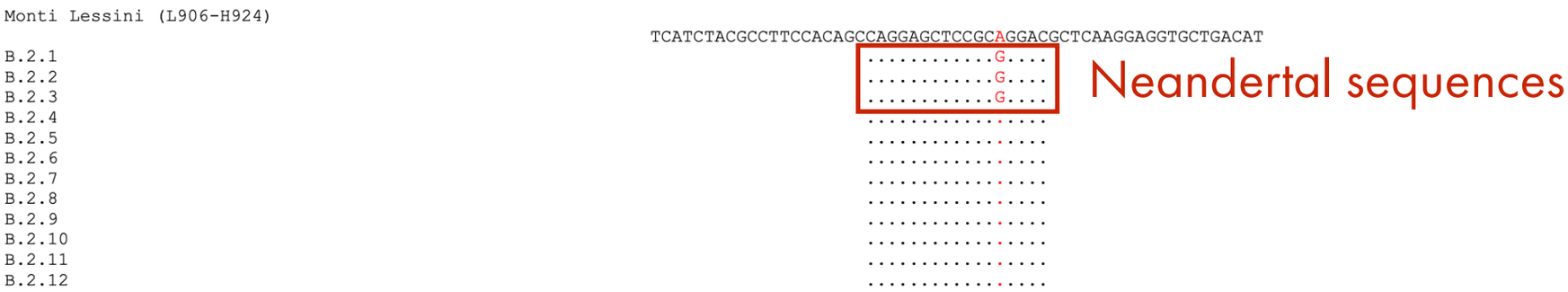


MC1R gene fragments amplified from Monti Lessini Neandertal DNA extract

Lalueza-Fox (2007) Science

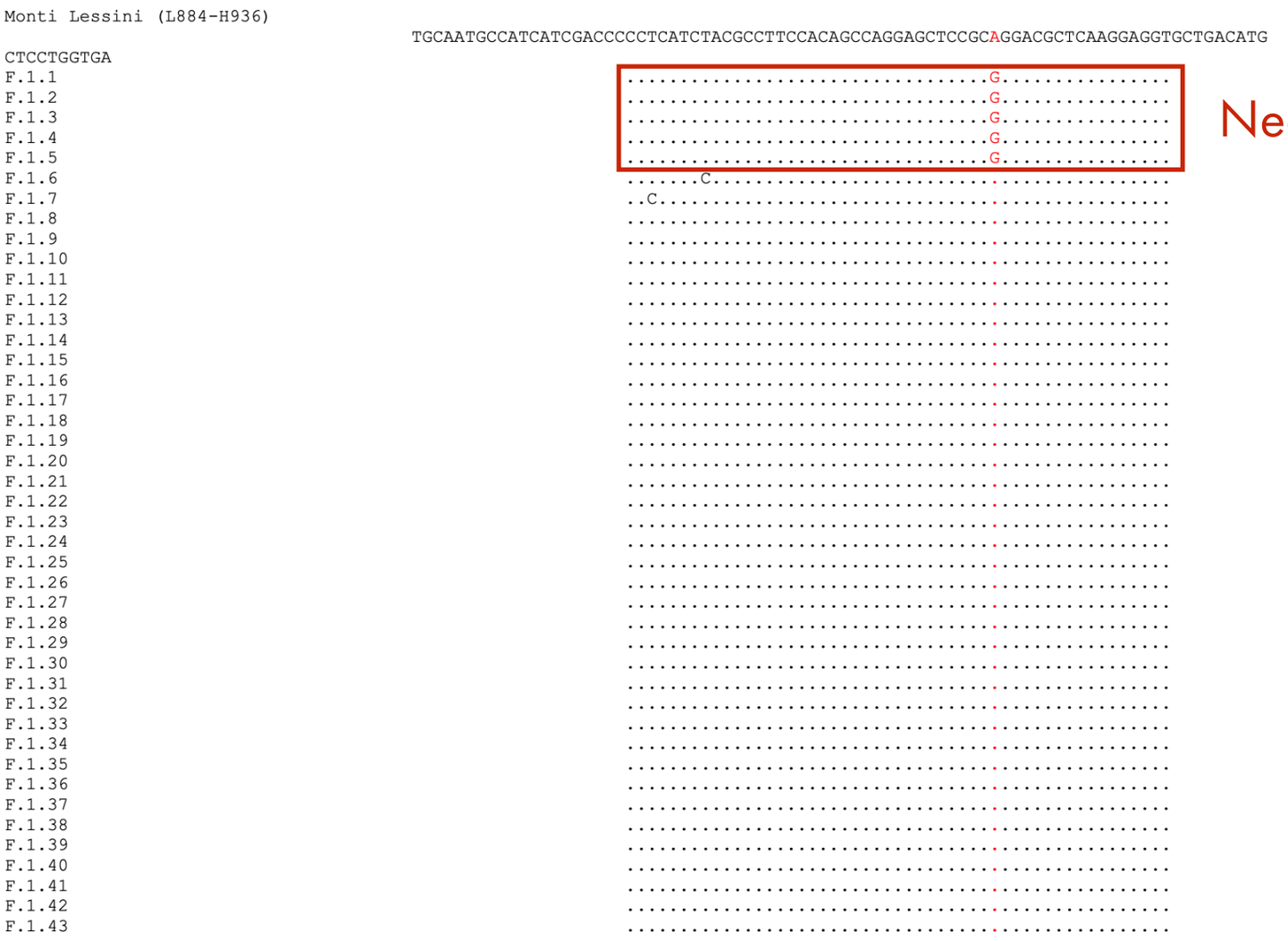
Dry lab challenges - DNA contamination

Barcelona lab



Neandertal sequences

Florence lab

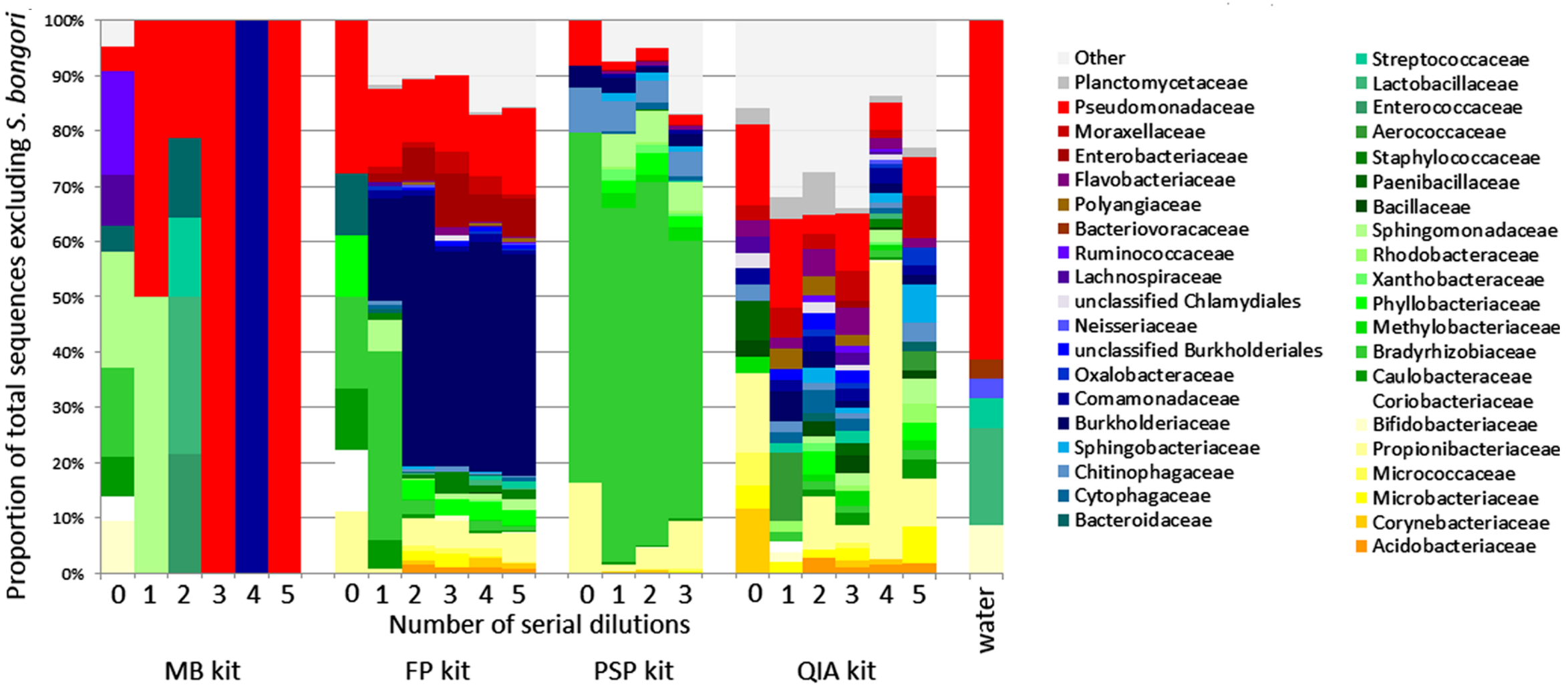


Neandertal sequences

MC1R gene fragments amplified from Monti Lessini Neandertal DNA extract

Lalueza-Fox (2007) Science

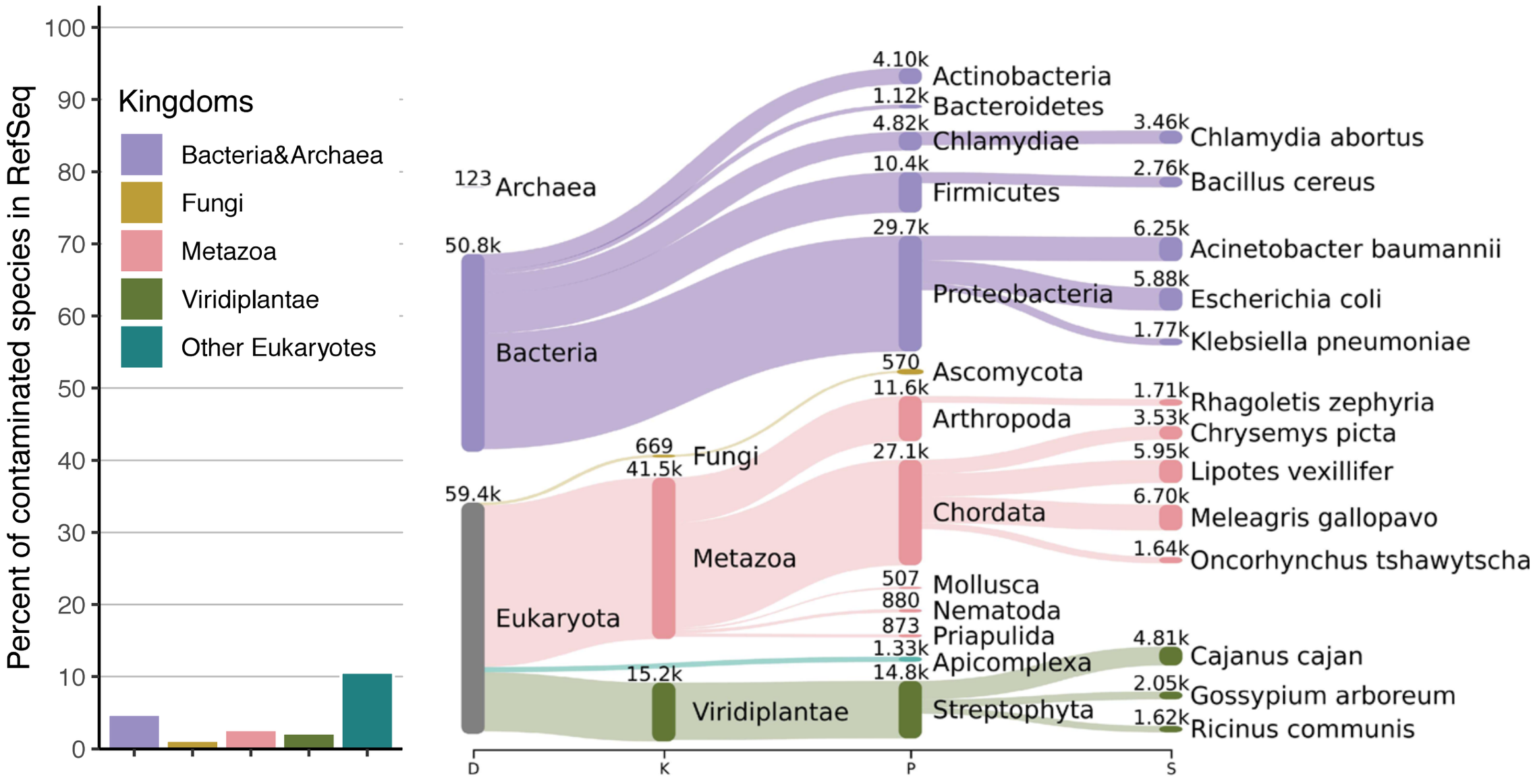
Dry lab challenges - DNA contamination



Commonly used laboratory reagents have distinct microbial contamination profiles

Salter et al (2014) BMC Biology

Dry lab challenges - DNA contamination

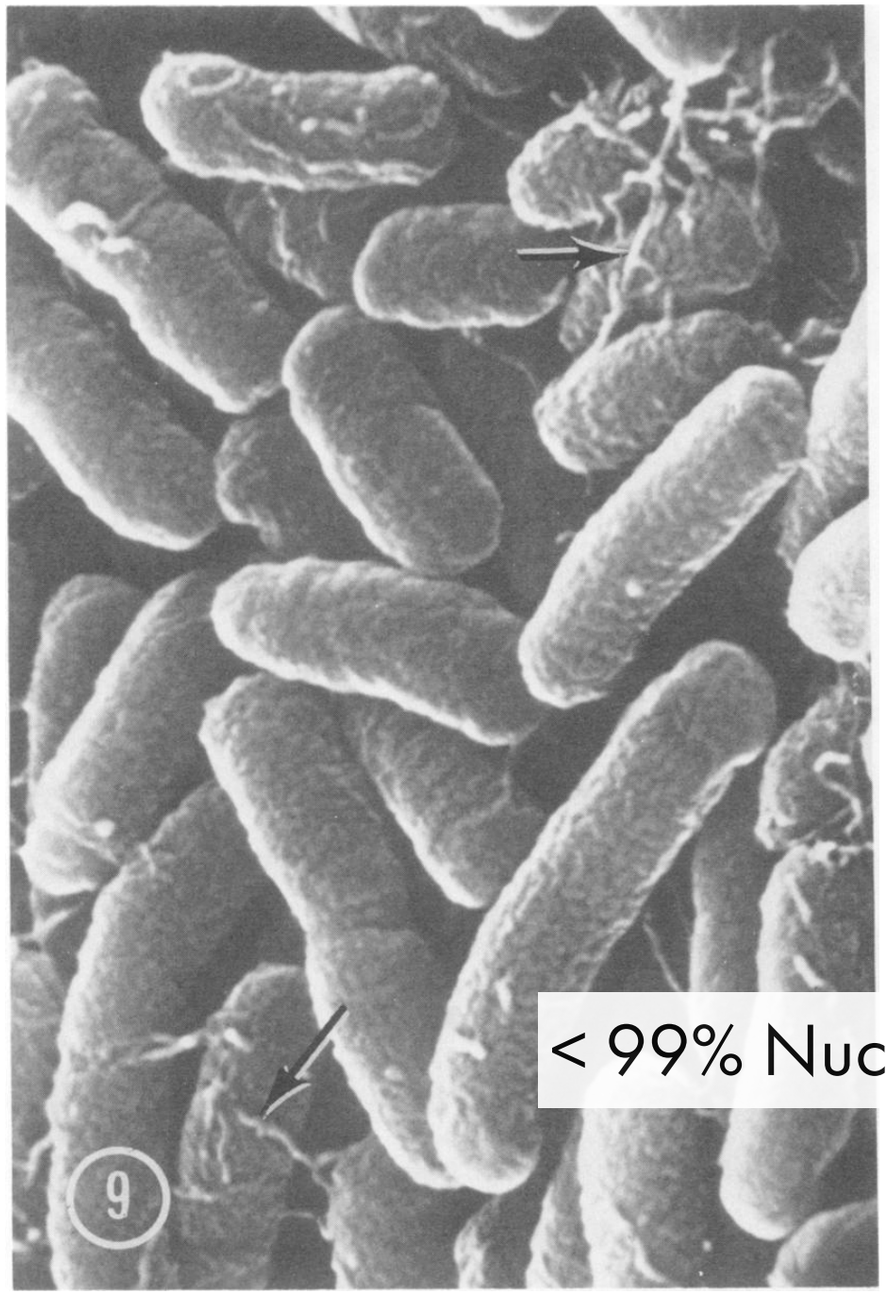


Public genome databases contain many contaminated entries

Steinegger and Salzberg (2020) Genome Biol

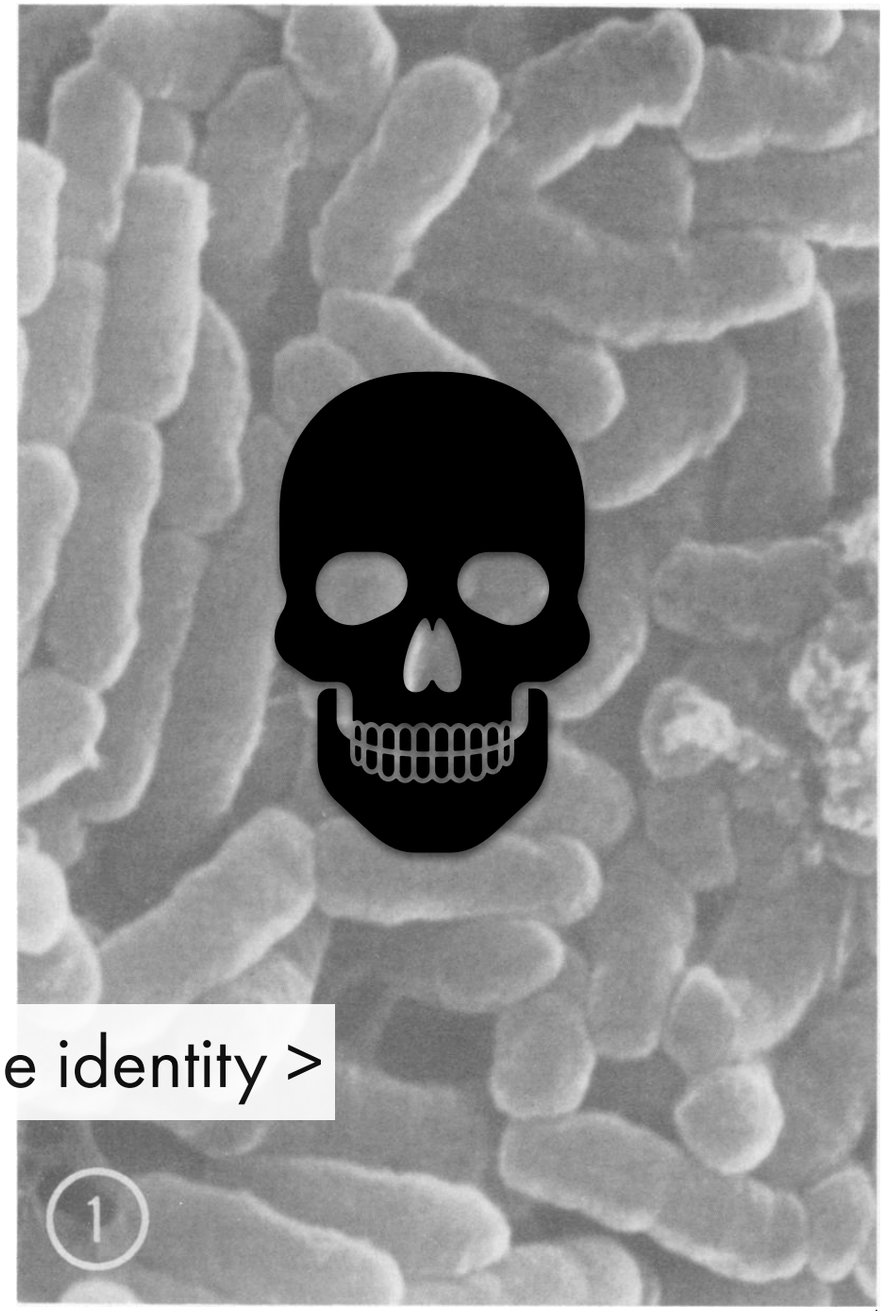
How can we authenticate our data?

Y. pseudotuberculosis



Commonly found in environment
occasional mild pathogen

Y. pestis

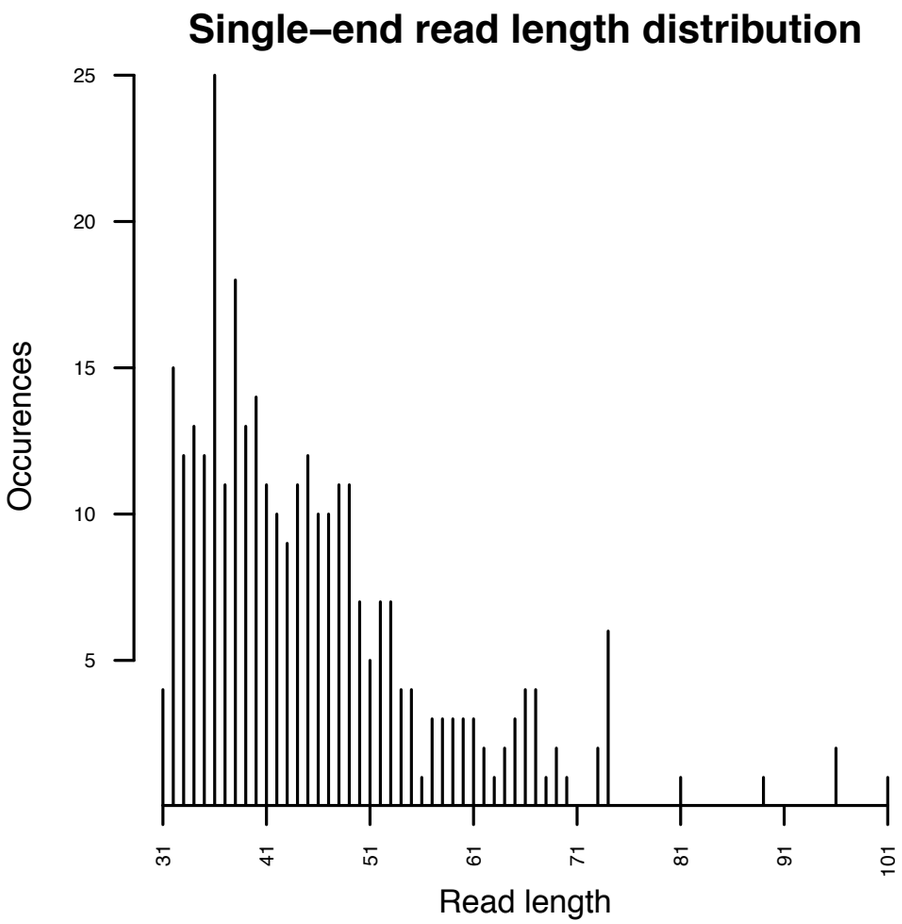


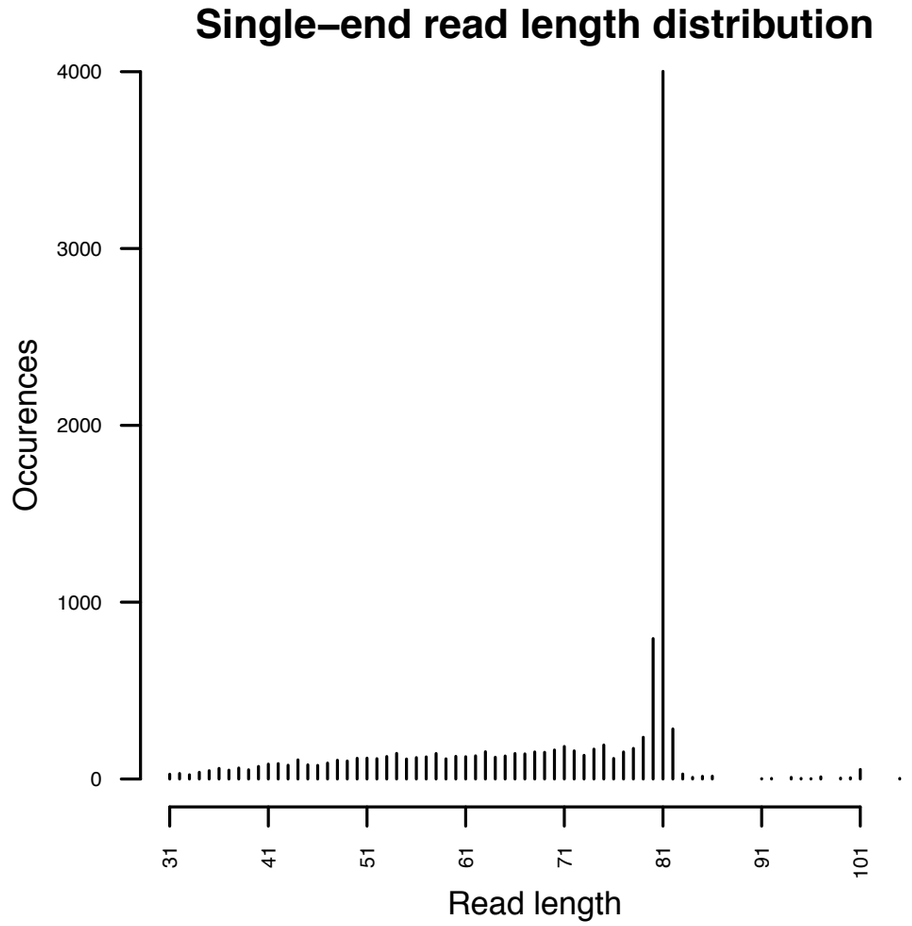
Vector-borne
Severe pandemic pathogen

< 99% Nucleotide identity >

Chen and Elberg (1977) Infection and Immunity

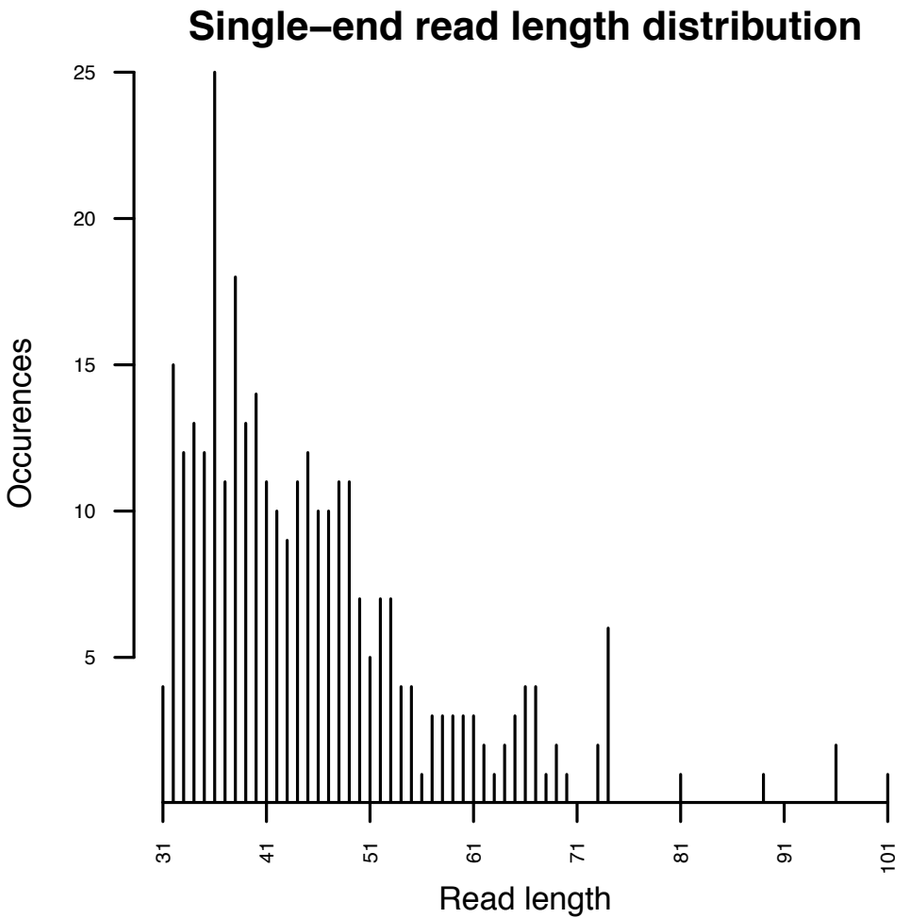
Authentication of ancient DNA

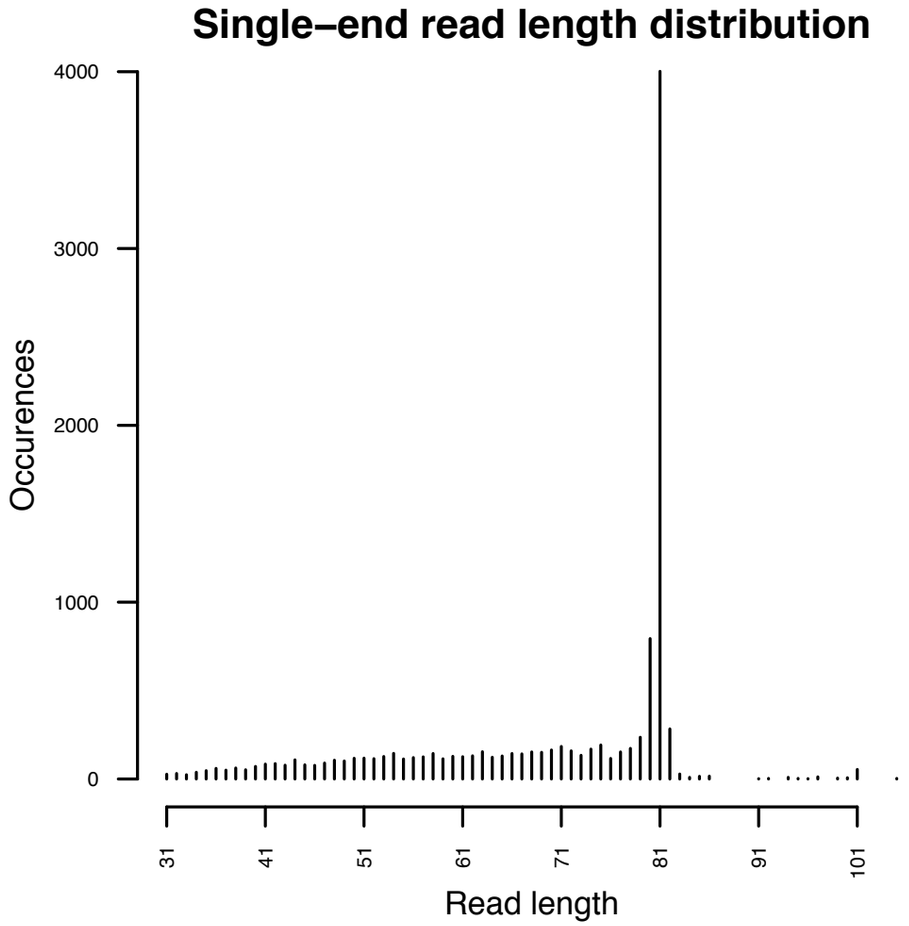




DNA fragment length distribution

Authentication of ancient DNA

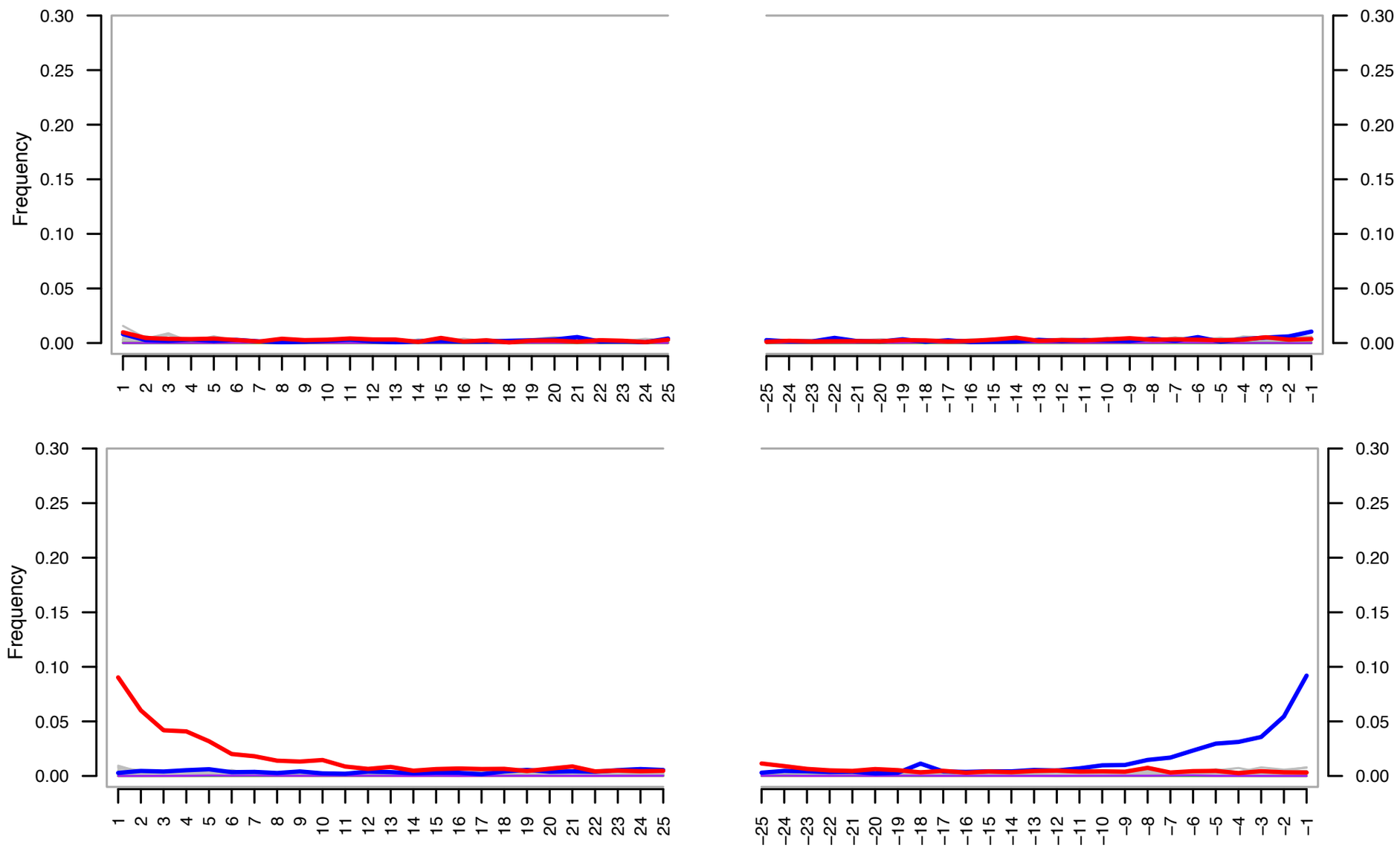




DNA fragment length distribution

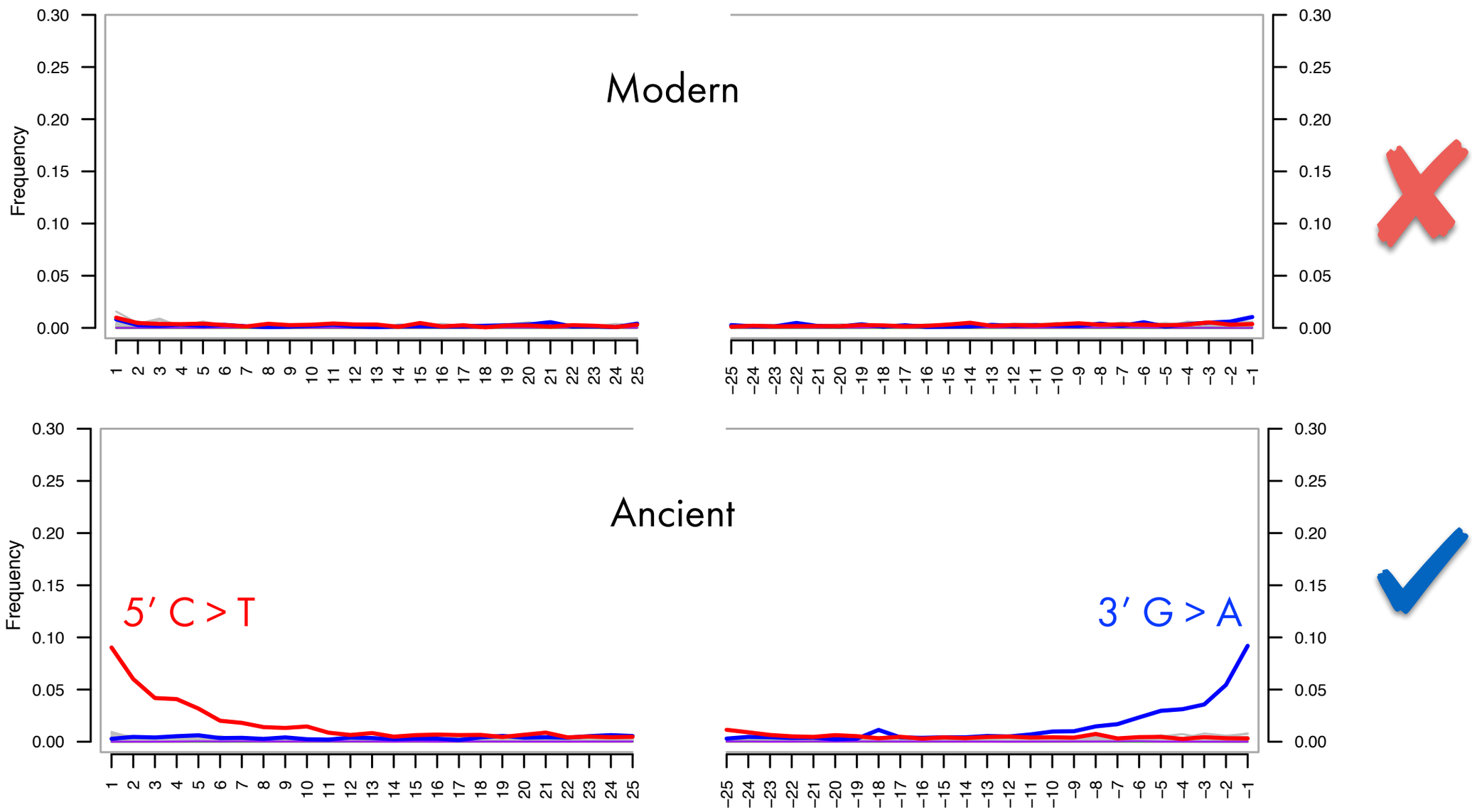
Ancient DNA molecules are usually short

Authentication of ancient DNA



Post-mortem DNA damage

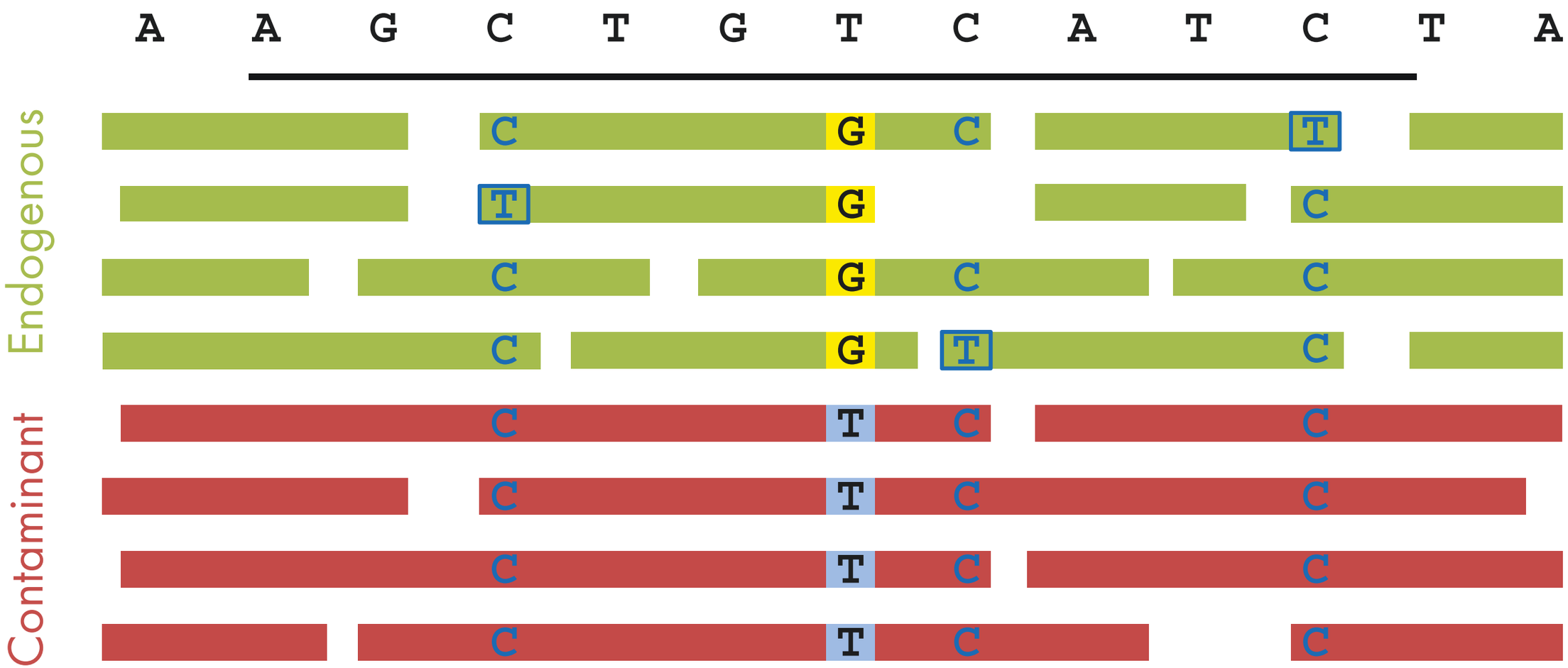
Authentication of ancient DNA



Post-mortem DNA damage

Increased rates of C>T and G>A substitution towards read ends

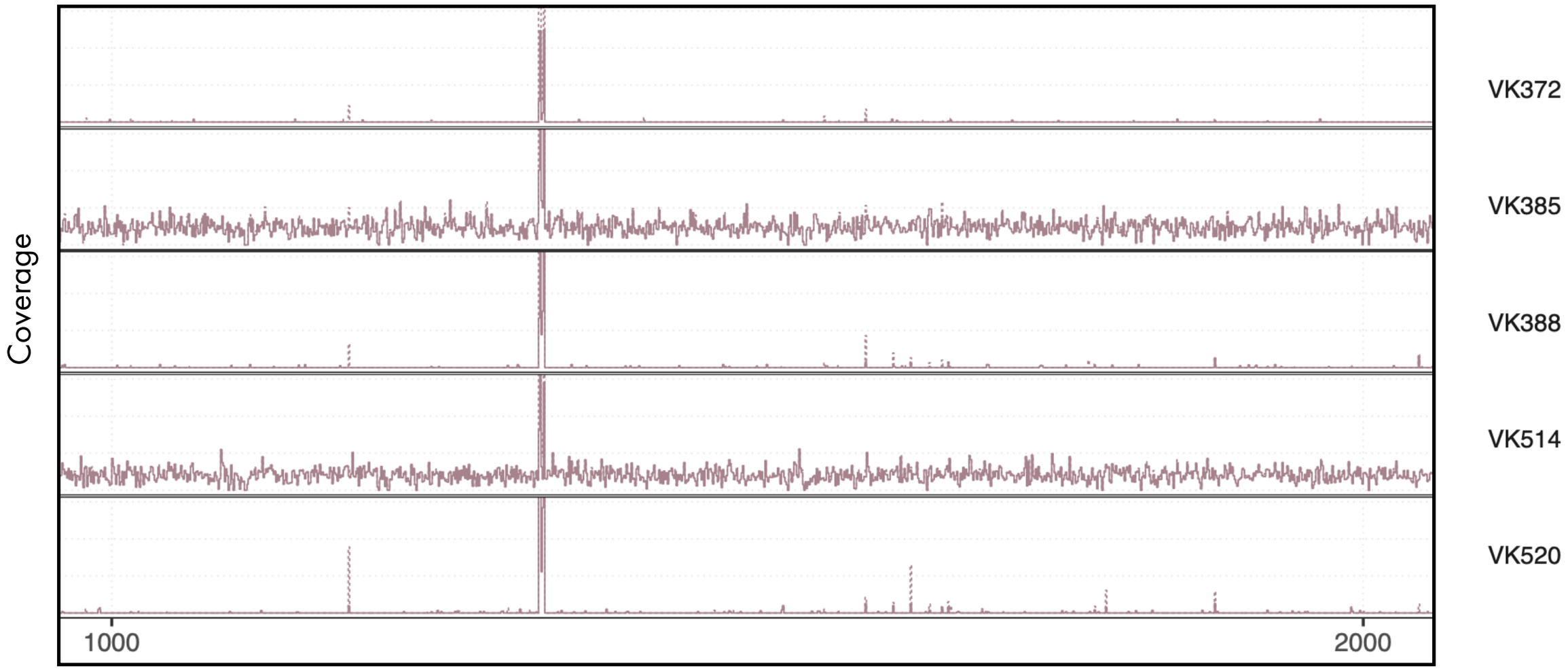
Authentication of ancient DNA



Quantify fraction of contaminating sequences

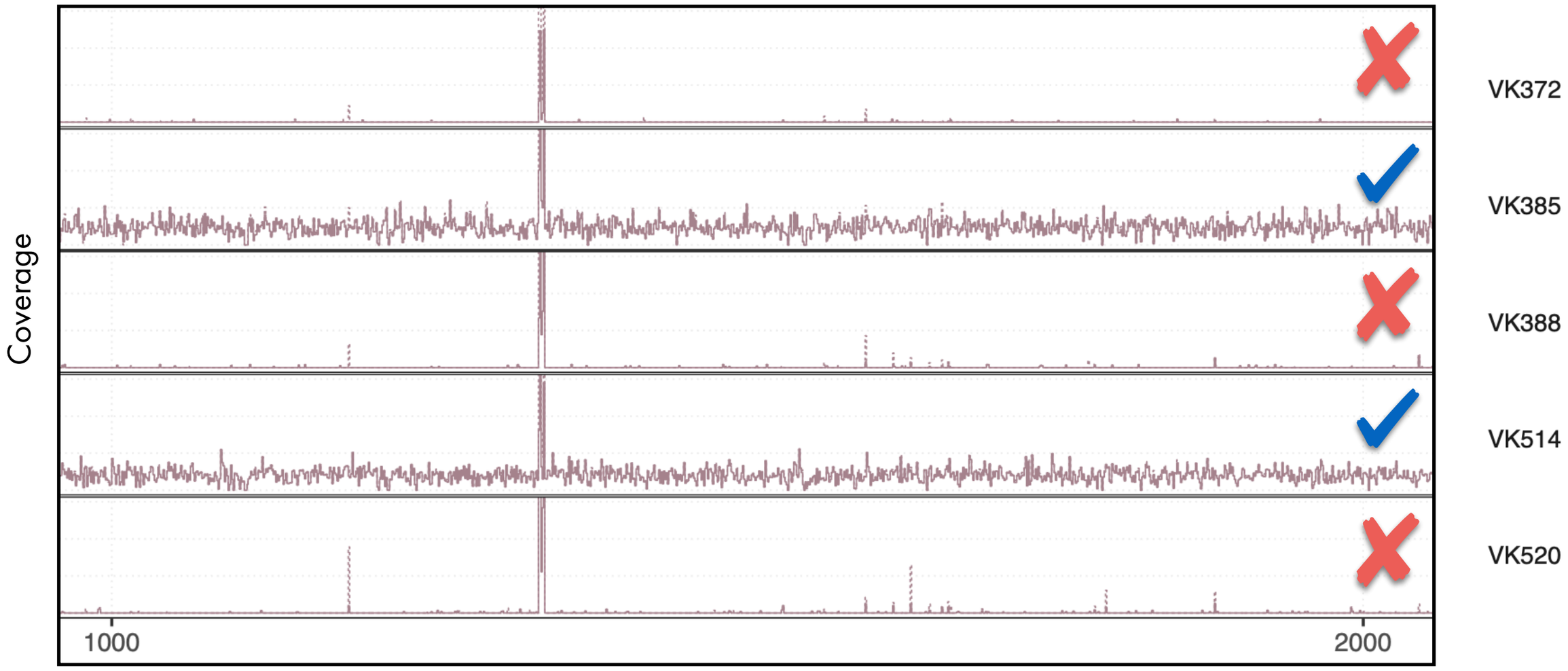
Renaud et al (2015) Bioinformatics

Authentication of ancient DNA



Genomic coverage

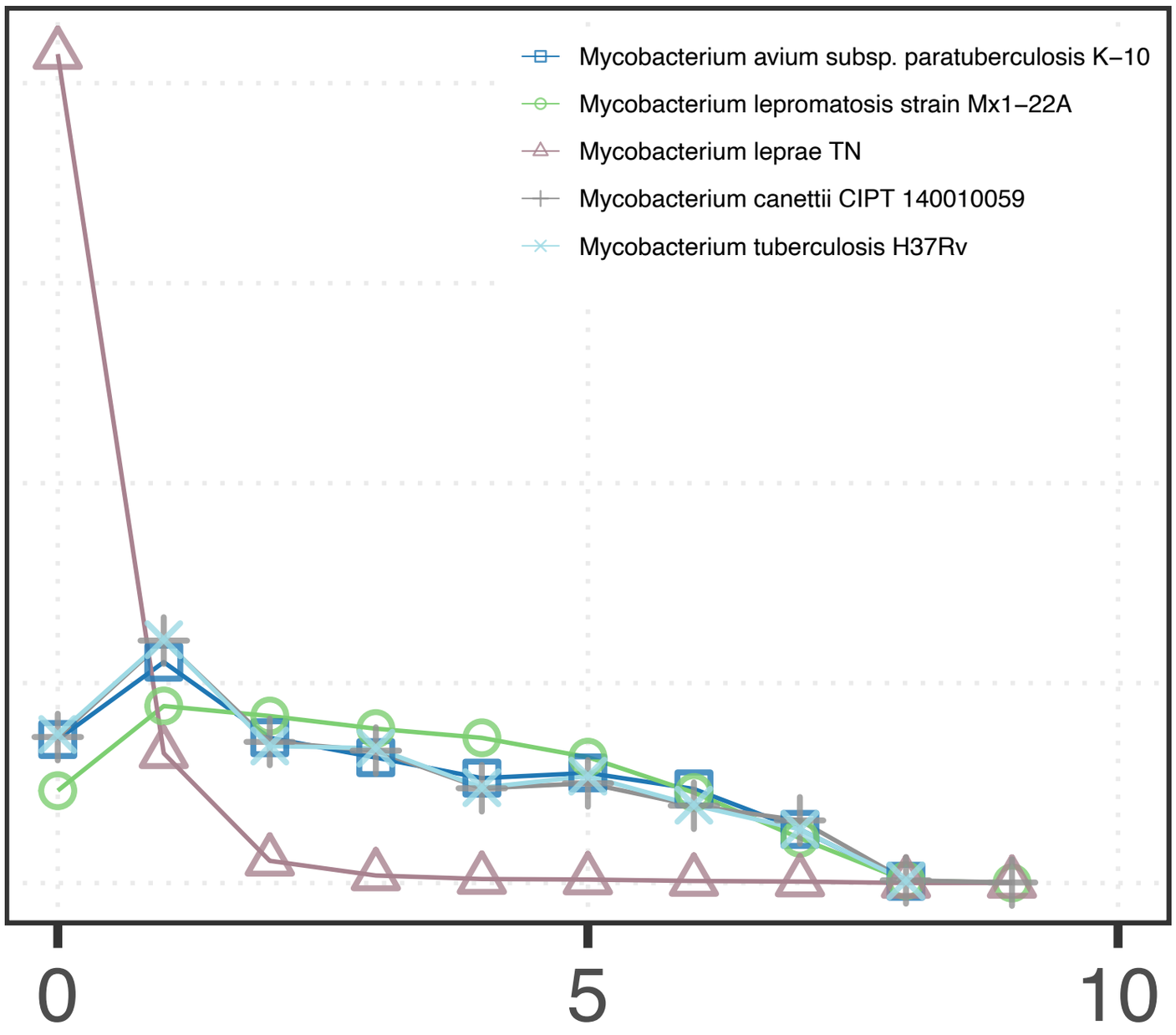
Authentication of ancient DNA



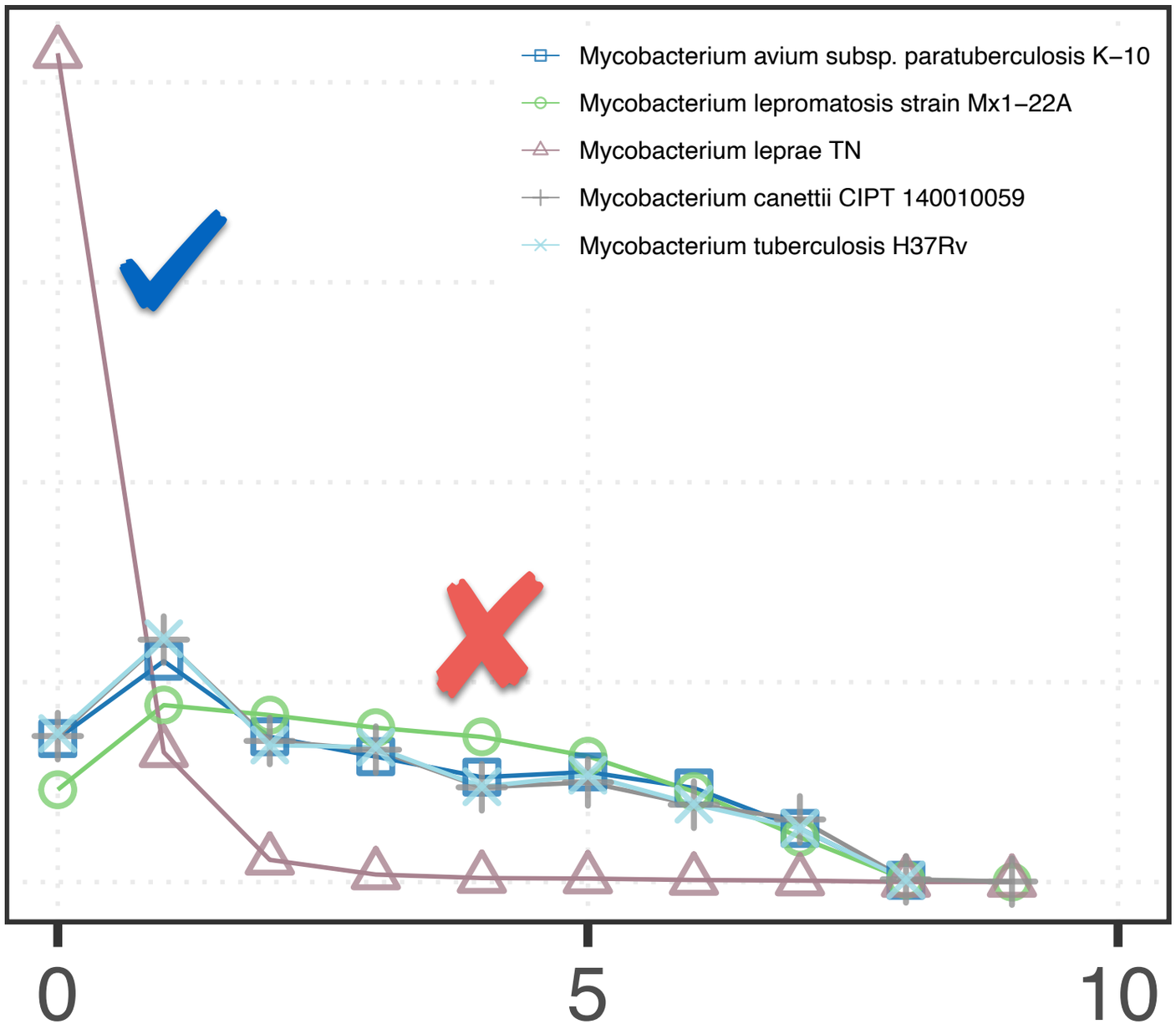
Genomic coverage

Mapped reads should be evenly distributed across the reference

Authentication of ancient DNA



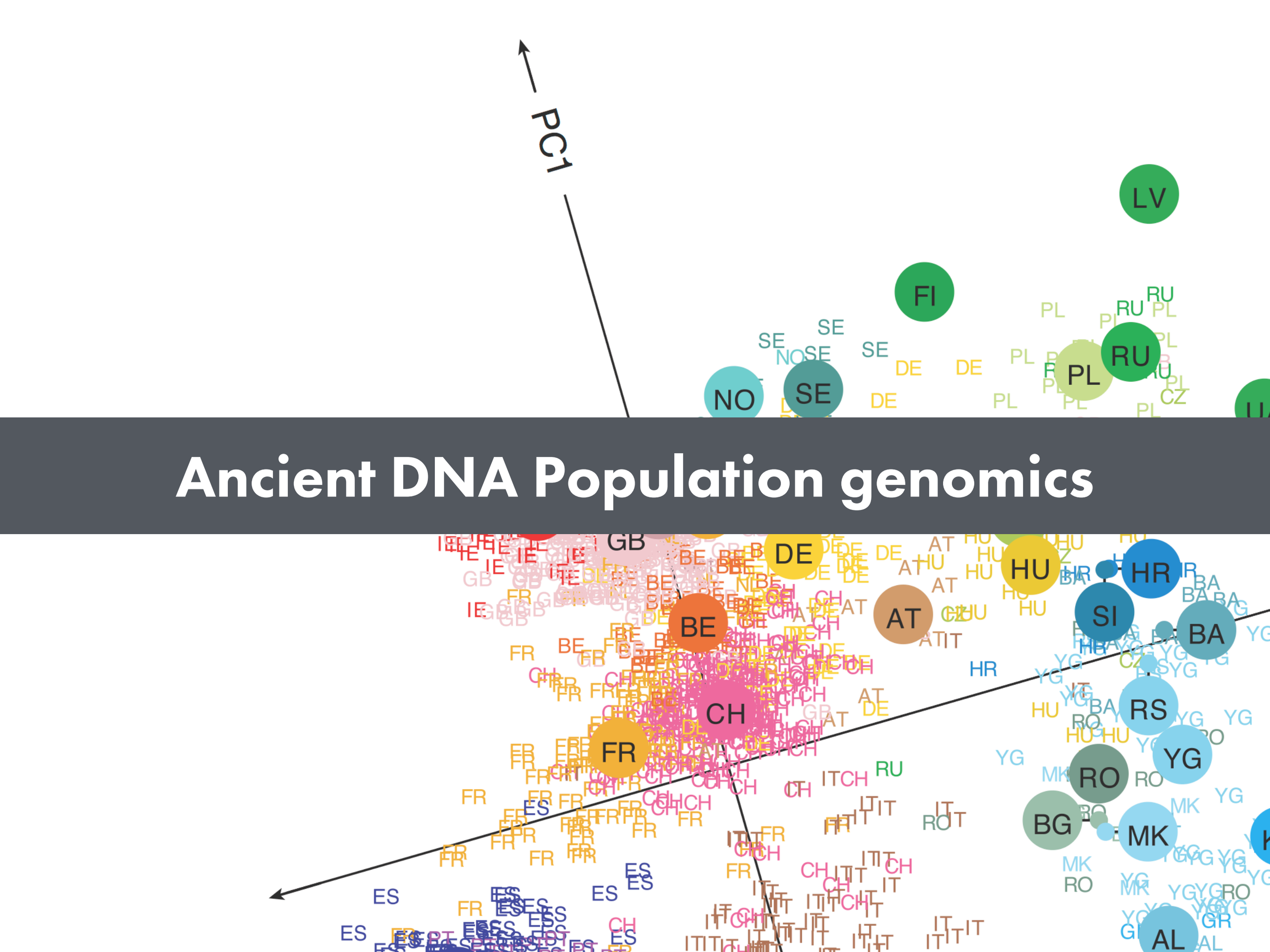
Authentication of ancient DNA



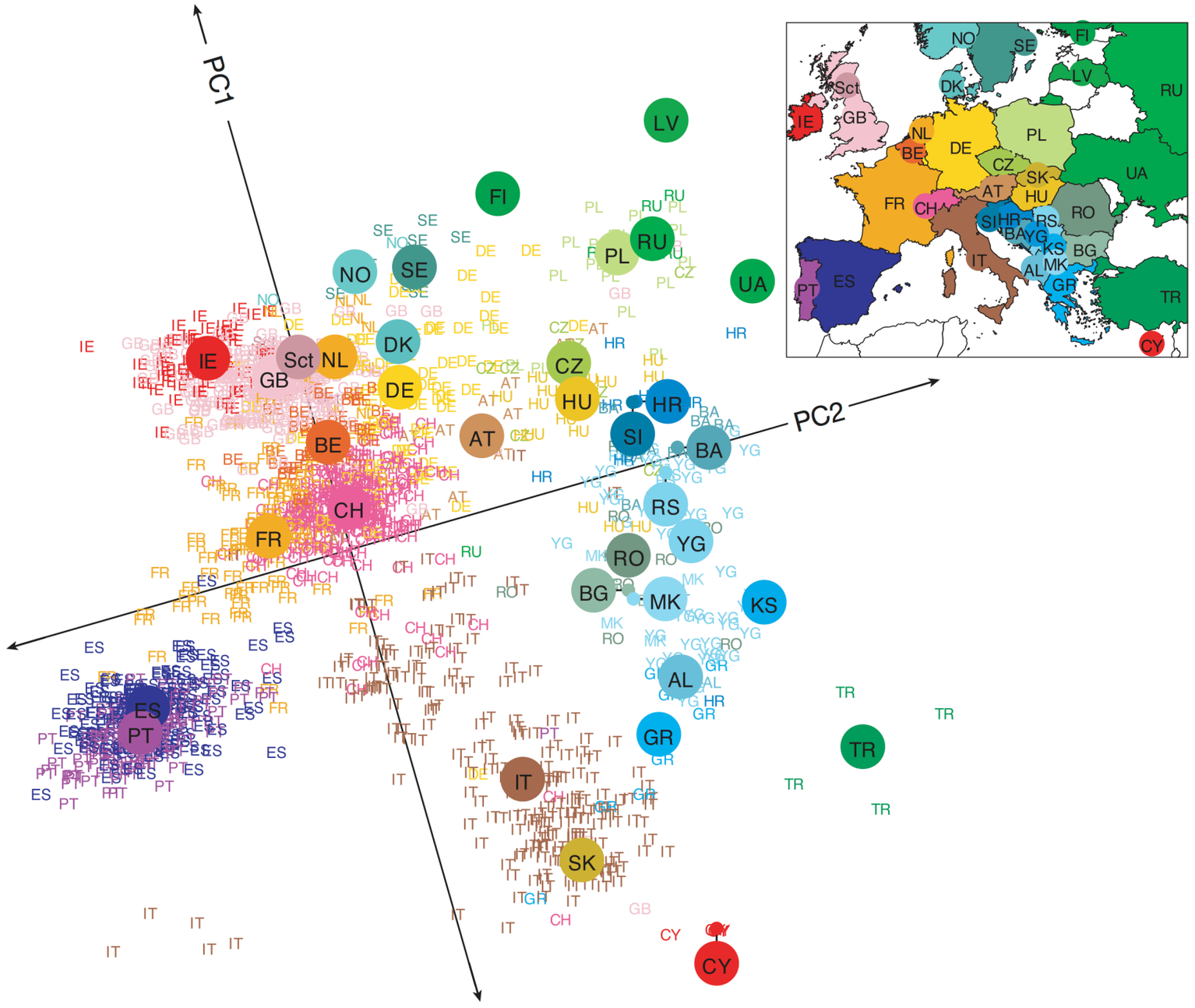
Edit distance distribution

Reads should have low divergence to the reference

Ancient DNA Population genomics



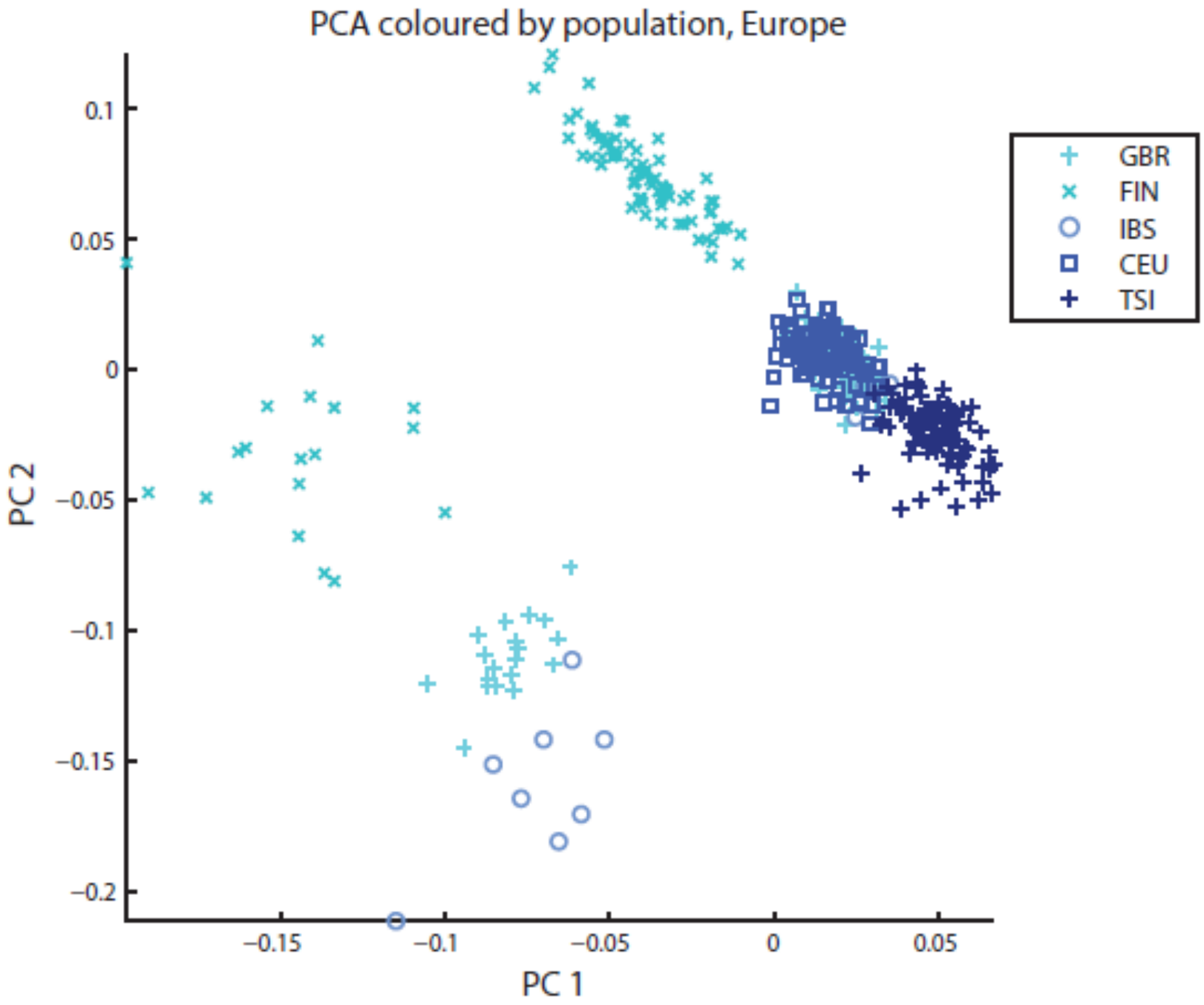
Principal component analysis



Genes mirror geography in Europe

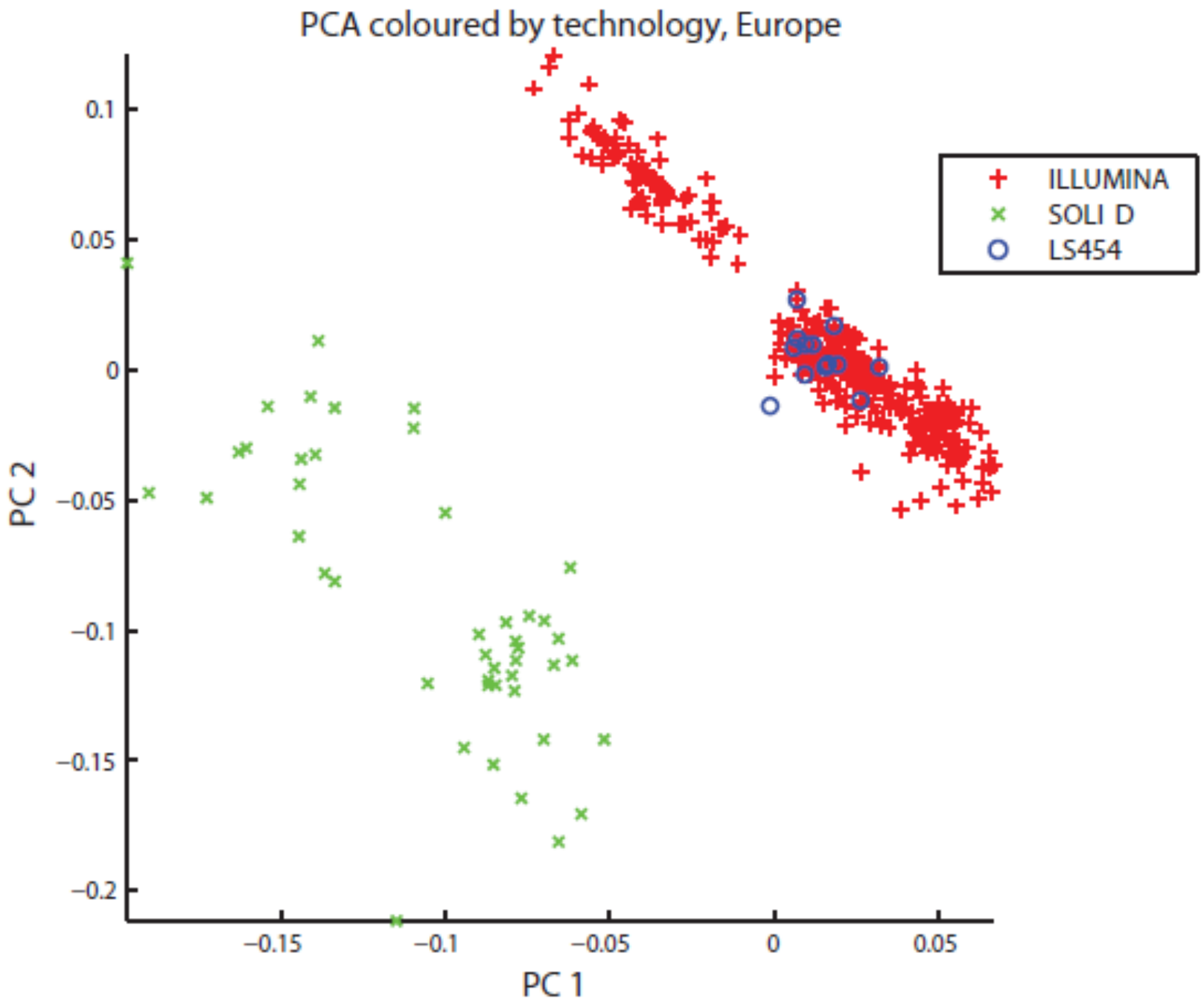
Novembre et al (2008) Nature

Usage case: QC



The 1000 Genomes Consortium (2012) Nature

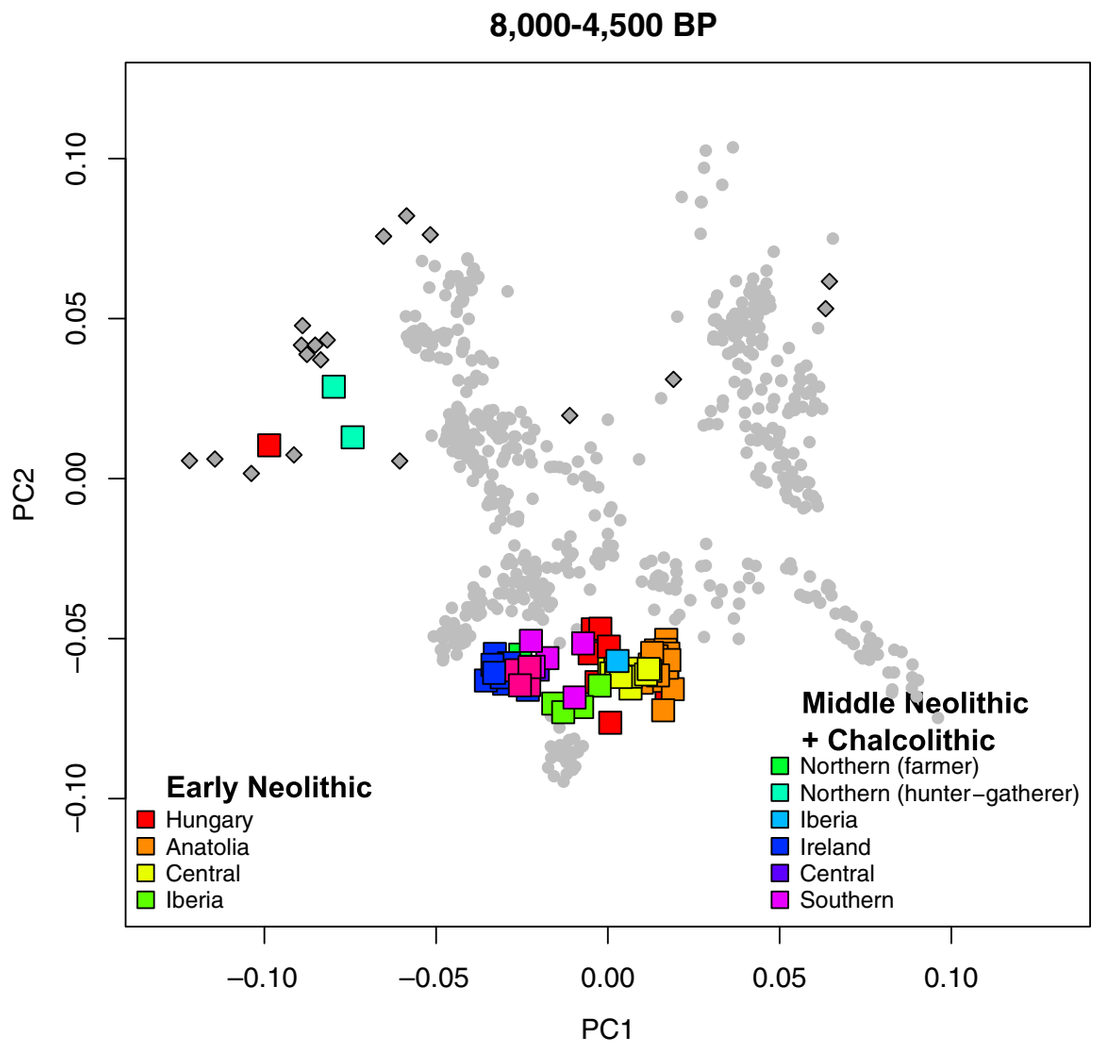
Usage case: QC



PCA can reveal batch effects in datasets

The 1000 Genomes Consortium (2012) Nature

PCA of ancient humans in West Eurasians

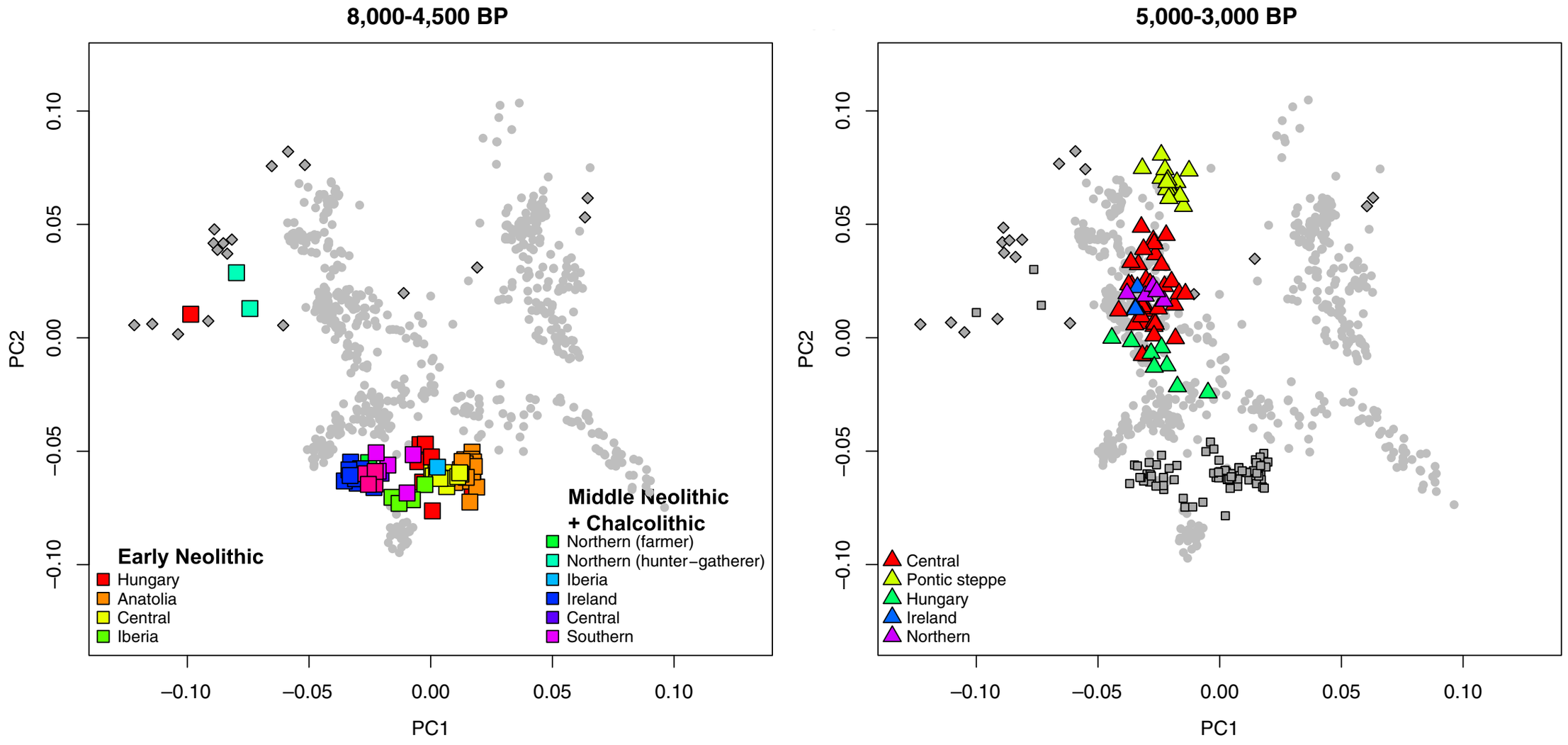


Infer principal components using high quality modern data

Project low coverage ancient samples onto inferred components

Guenther and Jakobbson (2016) Curr Op Genet Dev

PCA of ancient humans in West Eurasians

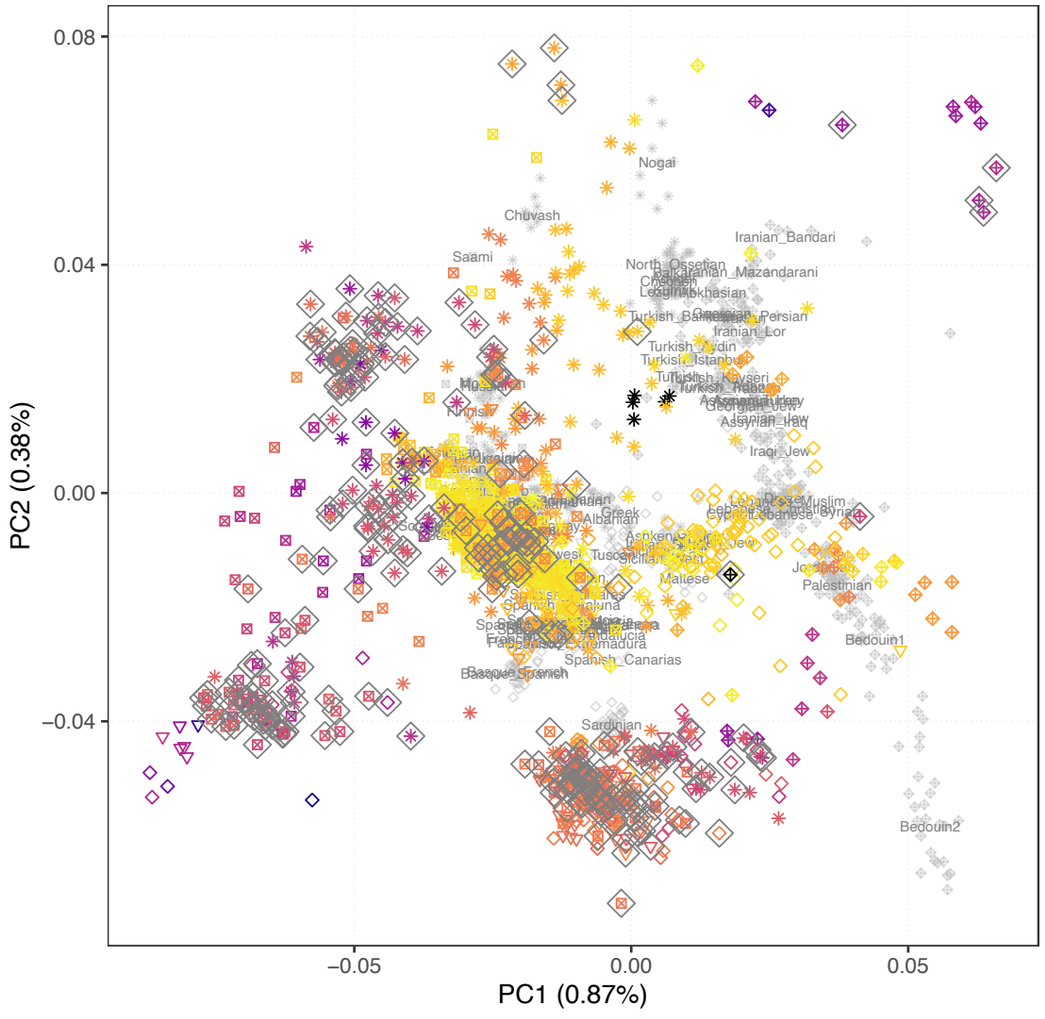


Infer principal components using high quality modern data

Project low coverage ancient samples onto inferred components

Guenther and Jakobbson (2016) Curr Op Genet Dev

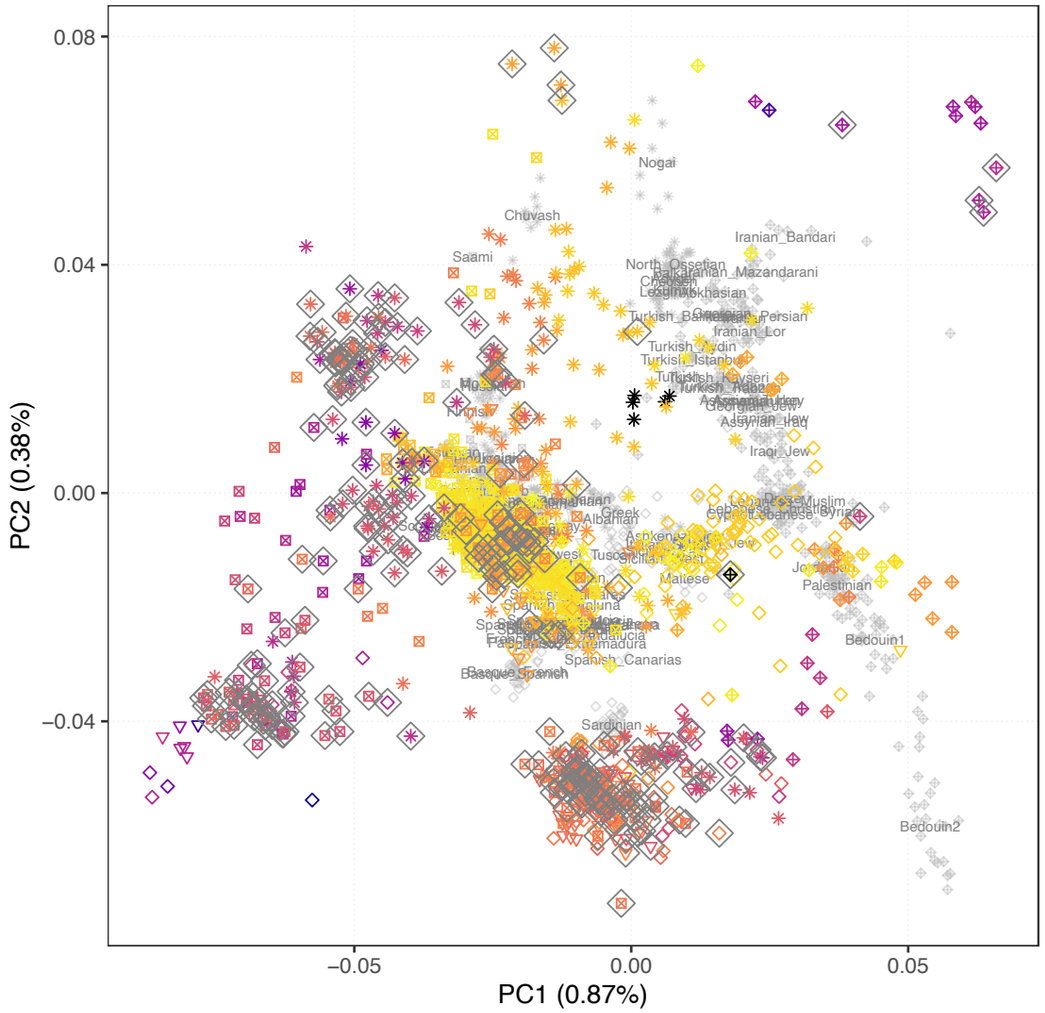
Drawbacks with PCA projections



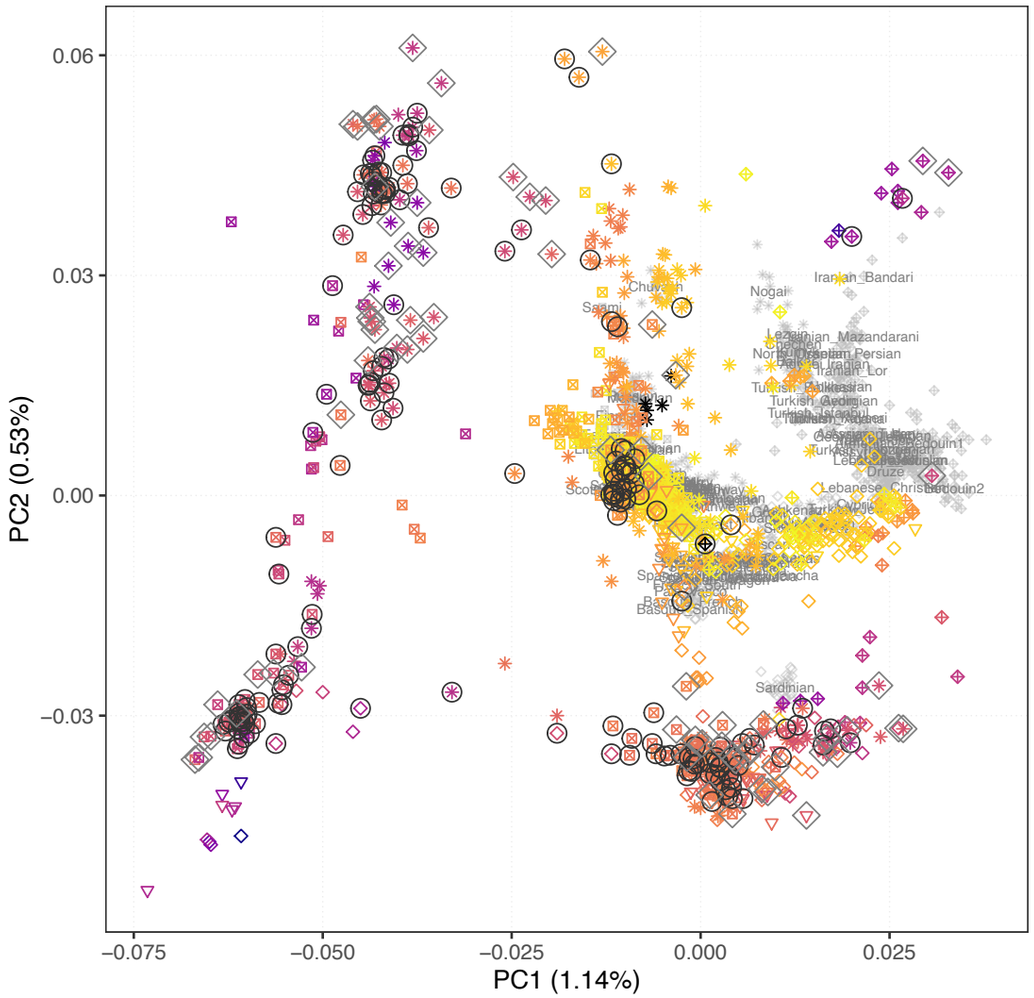
Ancient projected (pseudohaploid)

Allentoft, Sikora et al (2022) bioRxiv

Drawbacks with PCA projections



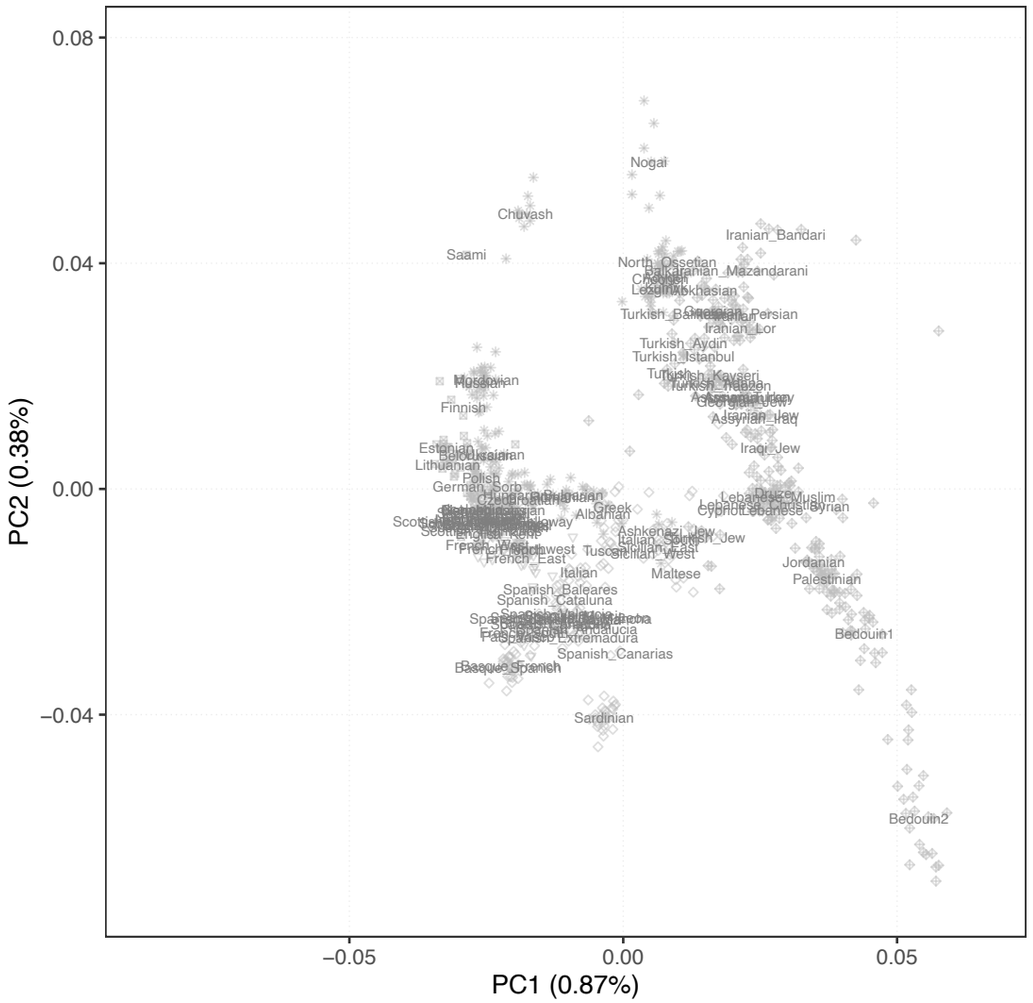
Ancient projected
(pseudohaploid)



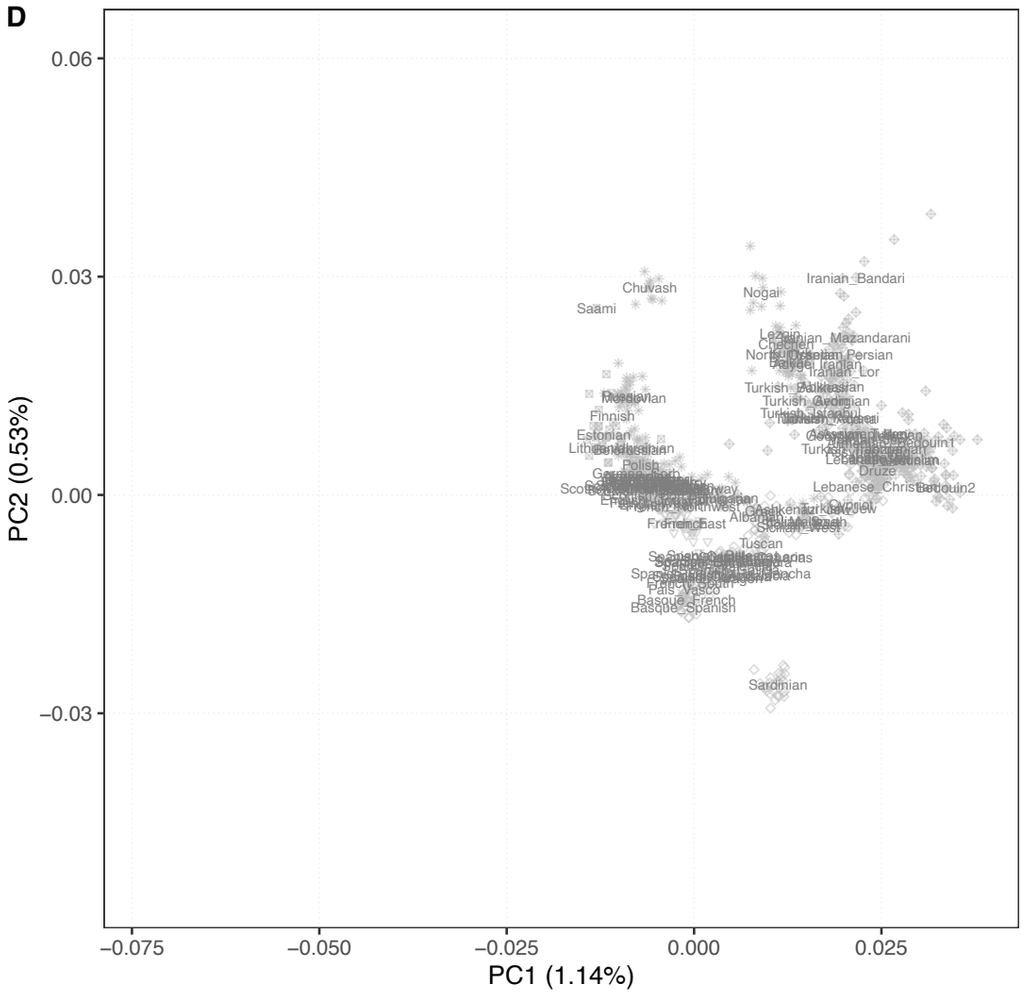
Ancient included
(diploid imputed)

Allentoft, Sikora et al (2022) bioRxiv

PCA of ancient West Eurasians



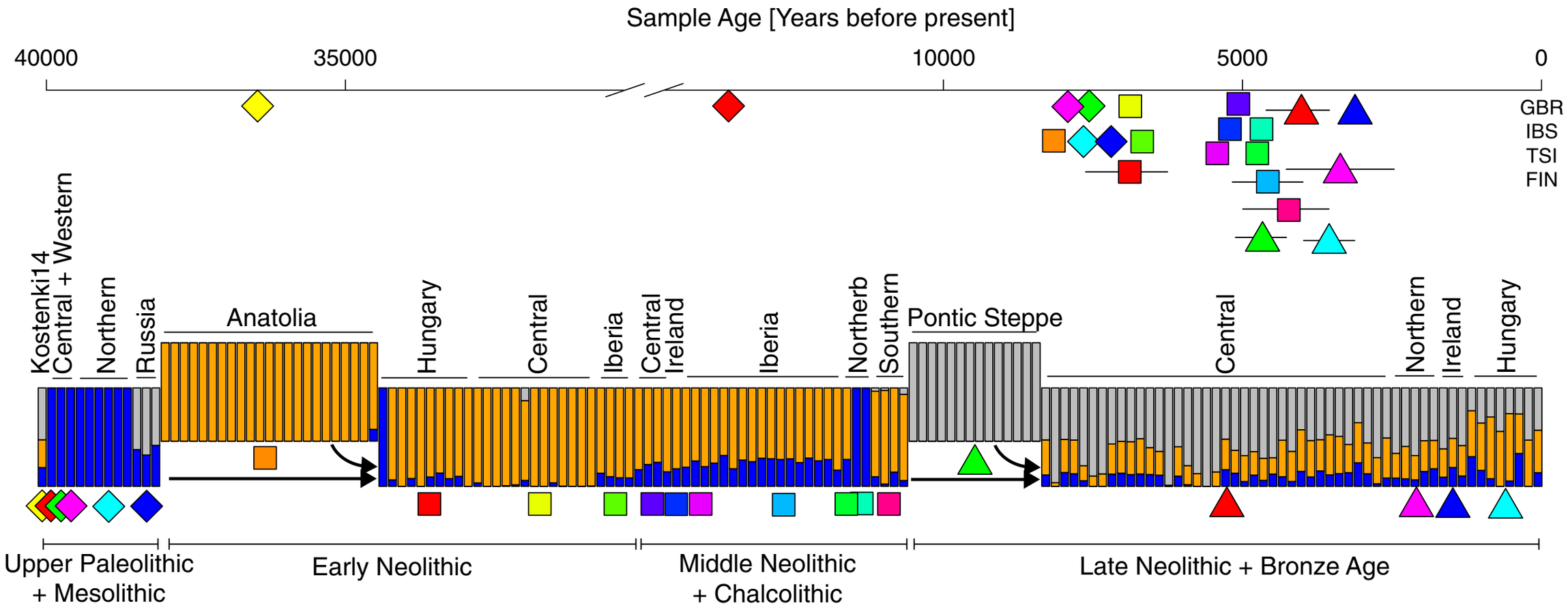
Ancient projected
(pseudohaploid)



Ancient included
(diploid imputed)

Allentoft, Sikora et al (2022) bioRxiv

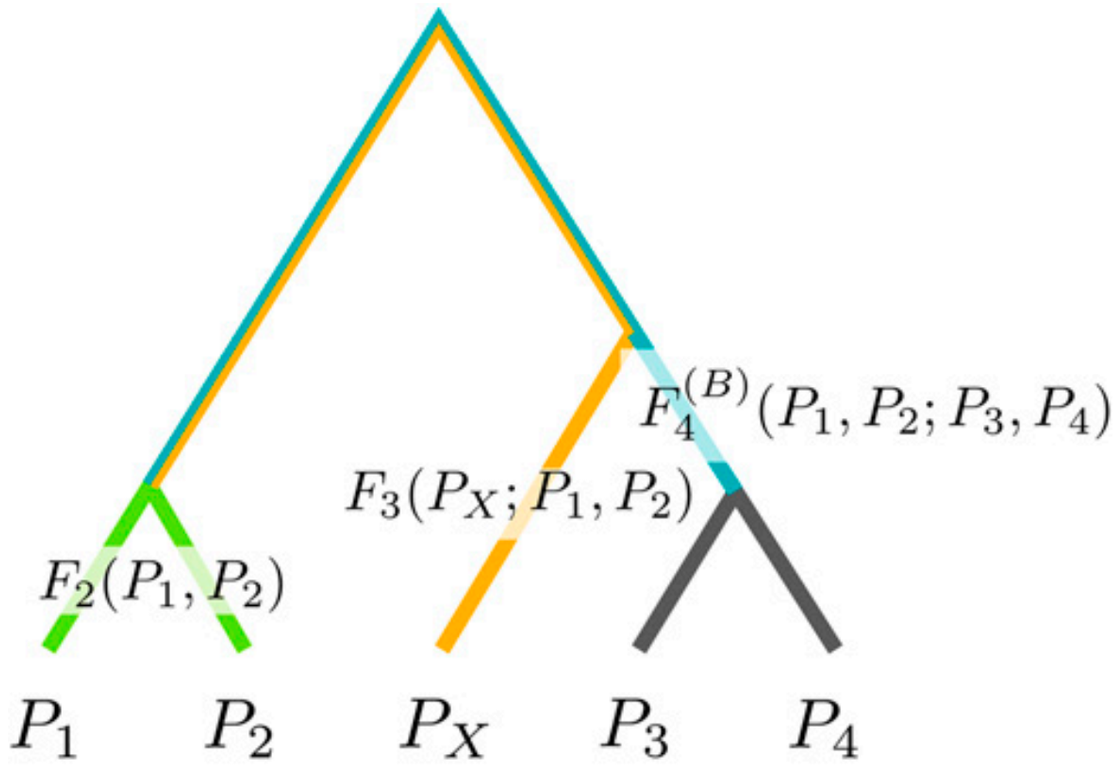
Model-based clustering



Individuals are modelled as mixtures of a pre-defined number of source populations

Guenther and Jakobbson (2016) Curr Op Genet Dev

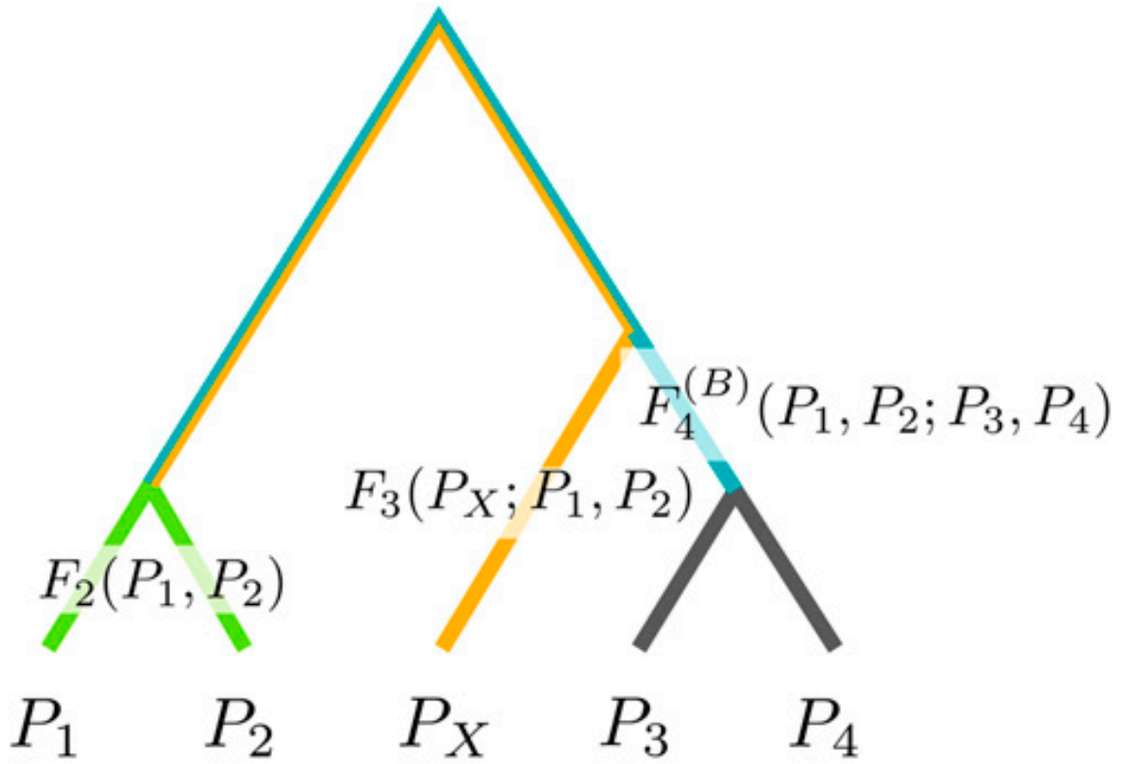
The F-statistic framework



Population phylogeny

Peter (2016) Genetics

The F-statistic framework



Population phylogeny

$$\frac{F_2(P_1, P_2)}{\mathbb{E}[(p_1 - p_2)^2]}$$

$$\frac{F_3(P_X; P_1, P_2)}{\mathbb{E}(p_X - p_1)(p_X - p_2)}$$

$$\frac{F_4(P_1, P_2, P_3, P_4)}{\mathbb{E}(p_1 - p_2)(p_3 - p_4)}$$

F-statistics can be interpreted as branch lengths between populations

Peter (2016) Genetics

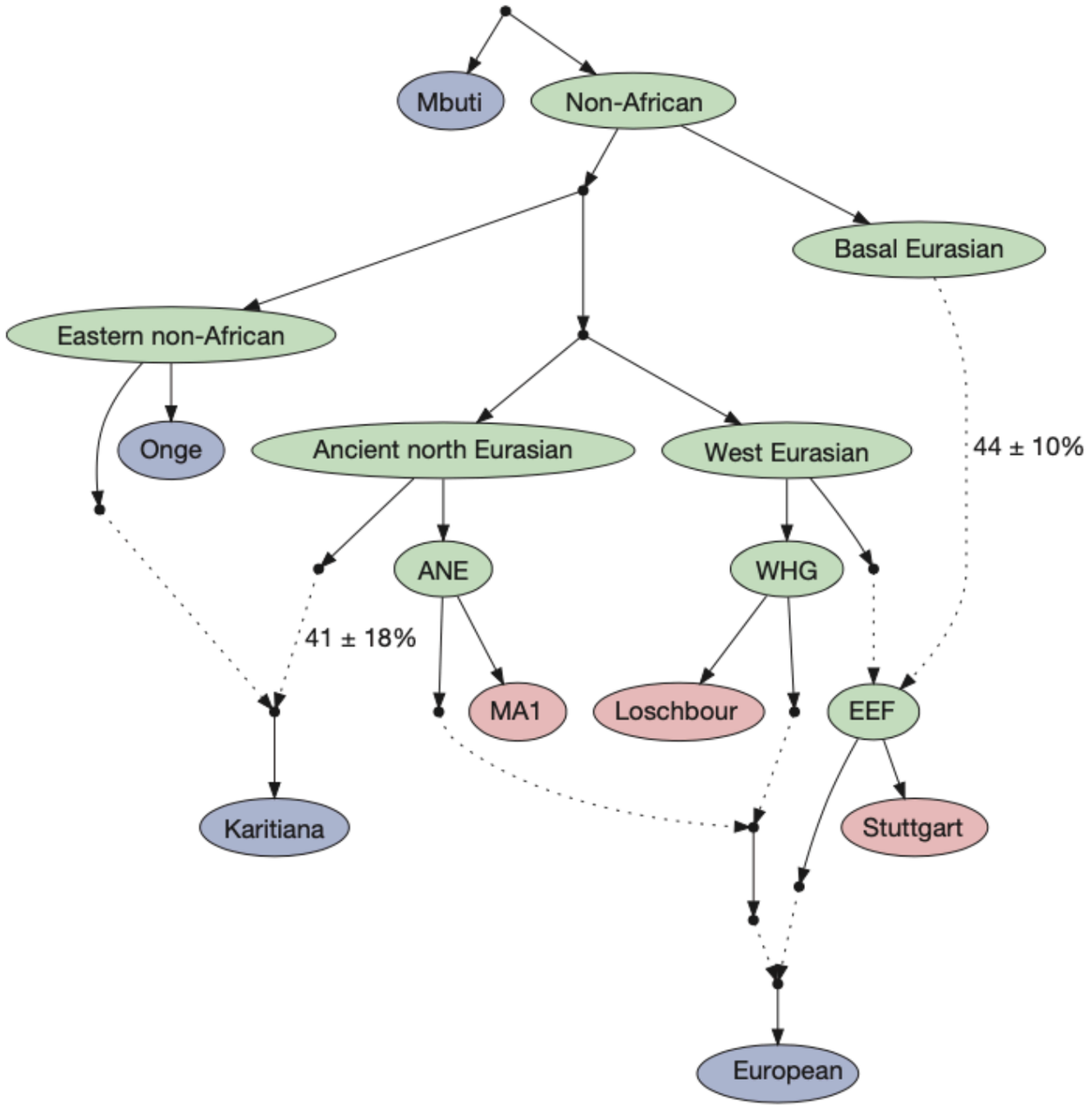
The F-statistic framework

F-statistic	Application	Test	Interpretation
f_2(A,B)	Branch length		
f_3(X;A,B)	Admixture f_3 - test	$f_3 < 0$	X is admixed related to A,B
	Outgroup - f_3		If X is outgroup to (A,B), f_3 proportional to shared drift between X and divergence of (A,B)
D(A,B;C,D)	D - test	D = 0	(A,B) form a clade with respect to (C,D)
D(O,B;C,D)	Symmetry test	D = 0	If O is outgroup to (B,C,D), tests for symmetry of B with respect to (C,D)
f_4(A,B;C,D)	f_4 - ratio test	$\alpha > 0$	Admixture proportion > 0
	Number of distinct ancestry streams between sets of outgroup and target populations (*qpWave*)		If rank of f4 - matrix is m, target populations are carry at least m + 1 streams of ancestry differentially related to the outgroup set
	Phylogeny-free estimation of admixture proportions (*qpAdm*)		Admixture proportions and fit for a target population as a mixture of N source populations
	Admixture graph fitting (*qpGraph*)		Goodness of fit of f-statistics predicted for specific graph topology

Wide range of applications to test hypotheses about admixture between populations

Peter (2016) Genetics

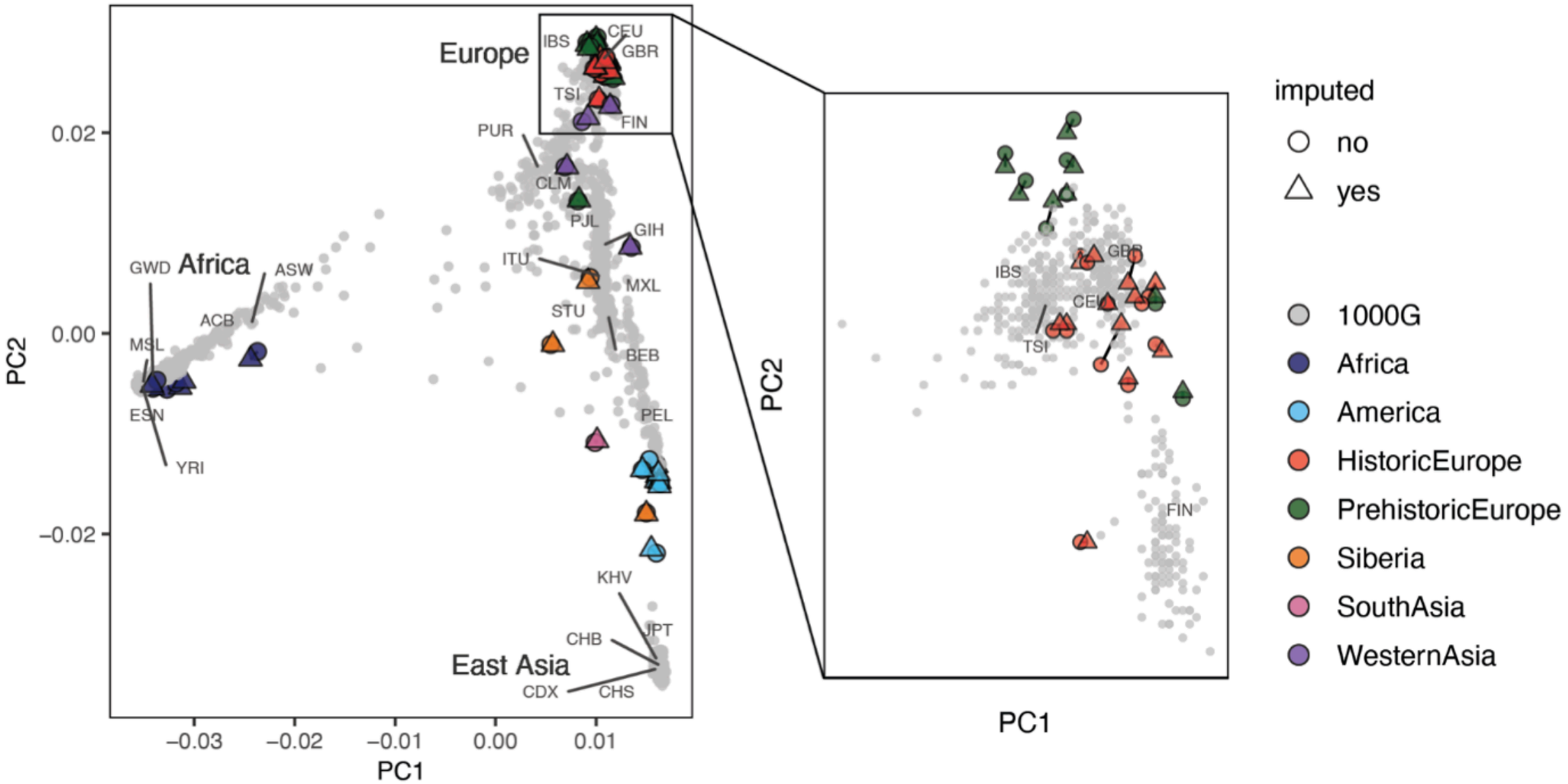
Admixture graphs



Admixture graph fitting suggests three ancestral populations for modern Europeans

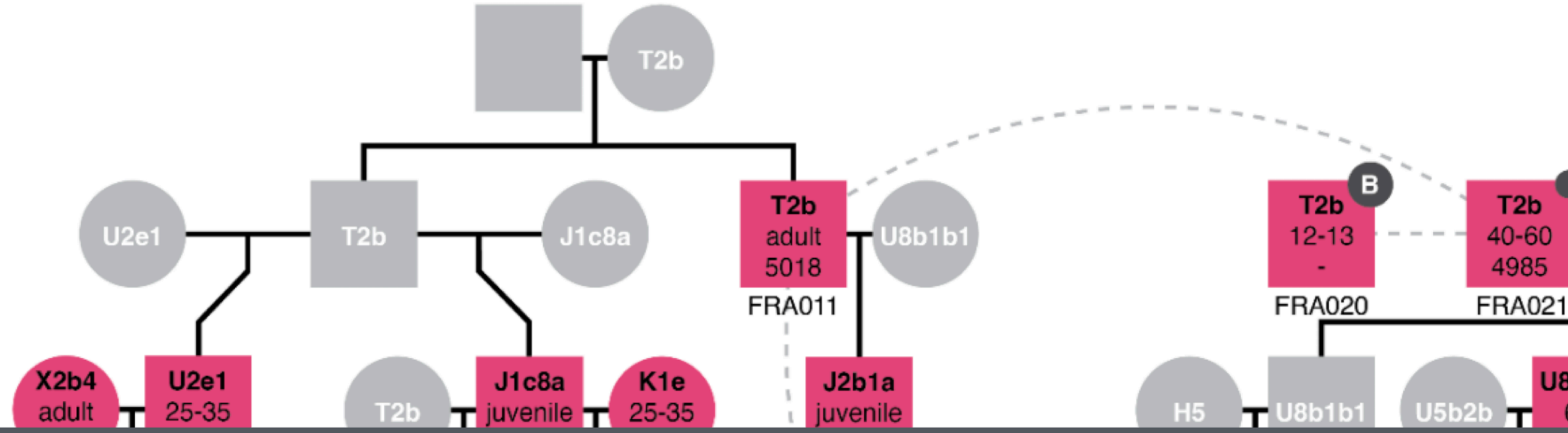
Lazaridis et al (2014) Nature

Imputation of ancient DNA



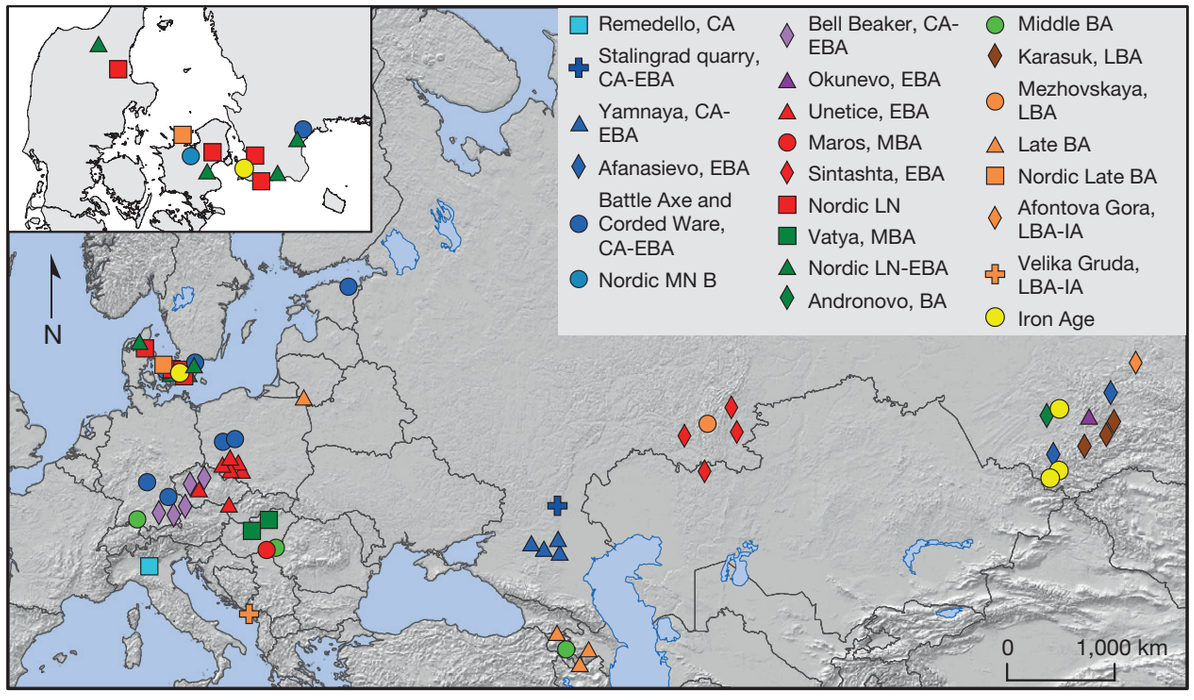
Ancient human genomes can be accurately imputed using modern reference panels

Lazaridis et al (2014) Nature

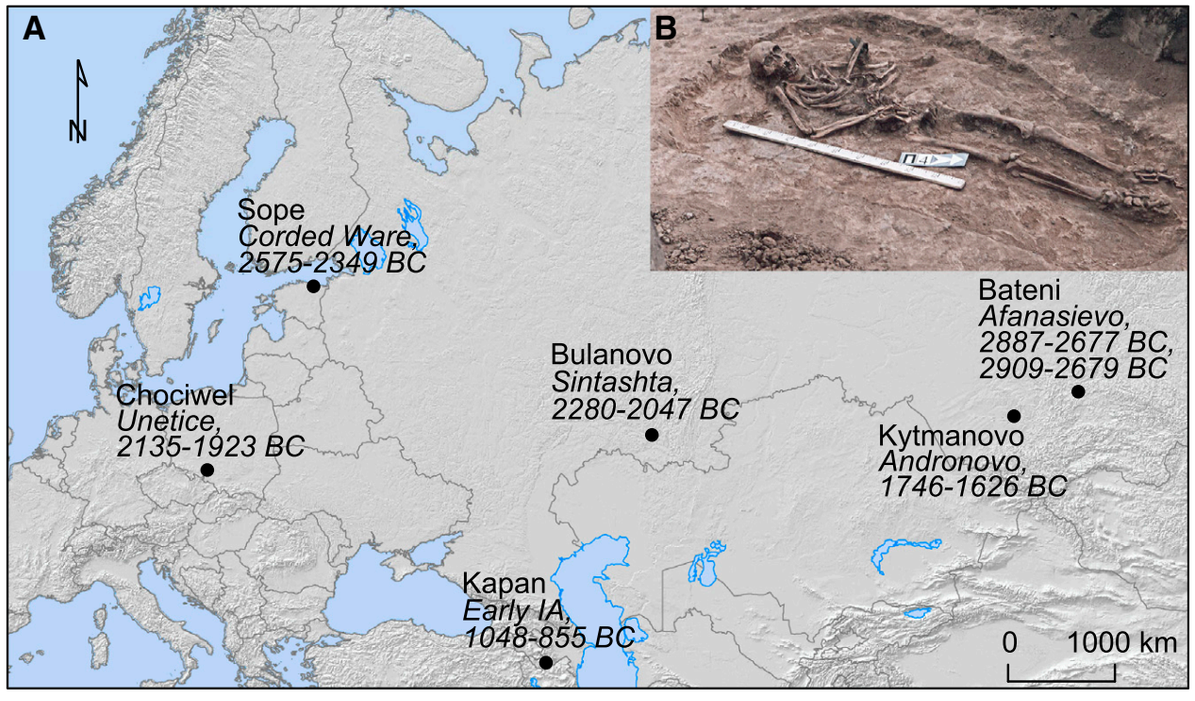


Ancient DNA studies showcase

Discovery of ancient plague strains in population-scale data



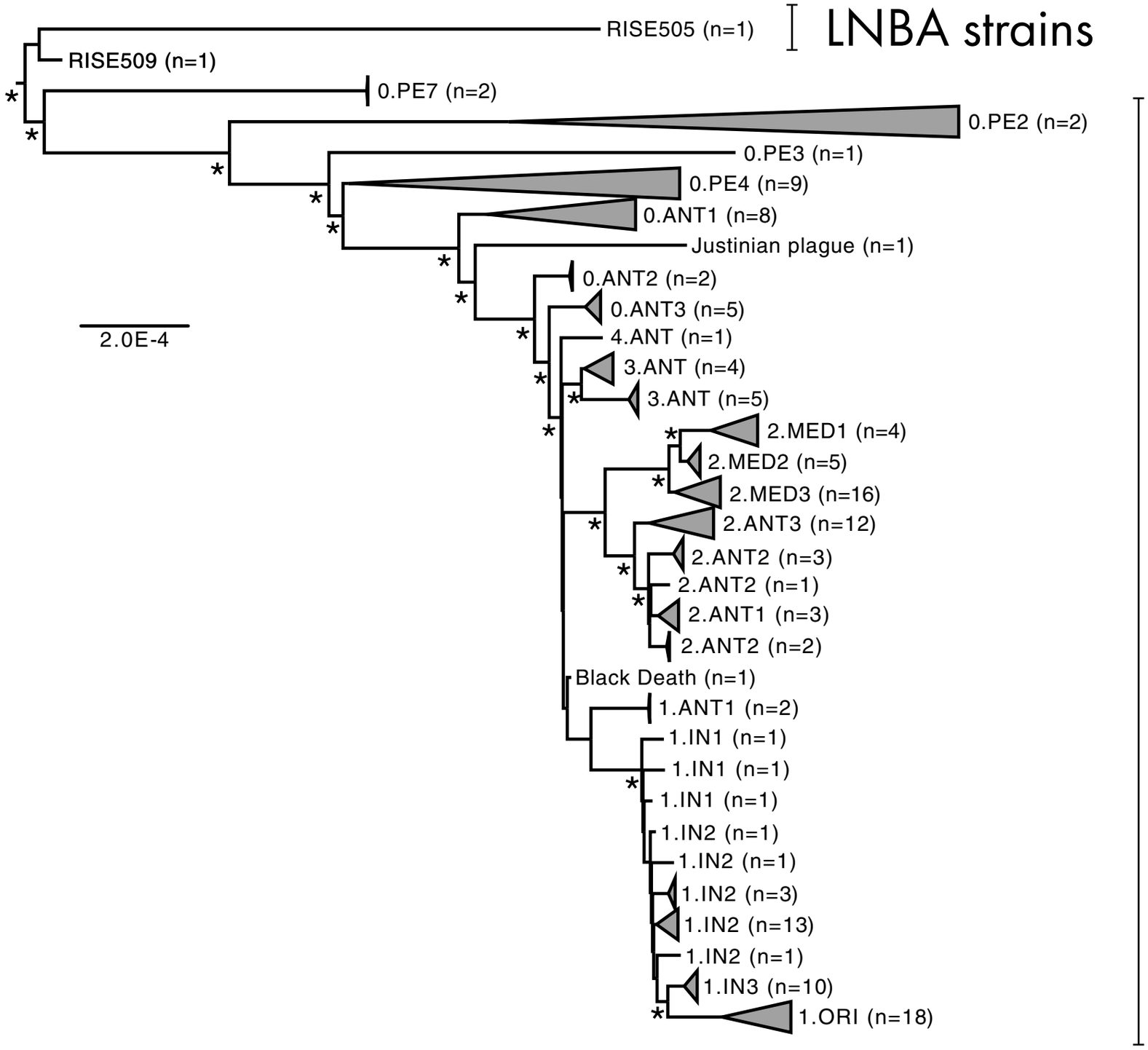
101 ancient human genomes from Late Neolithic / Bronze Age Eurasia



Ancient *Yersinia pestis* genomes recovered from 7 individuals

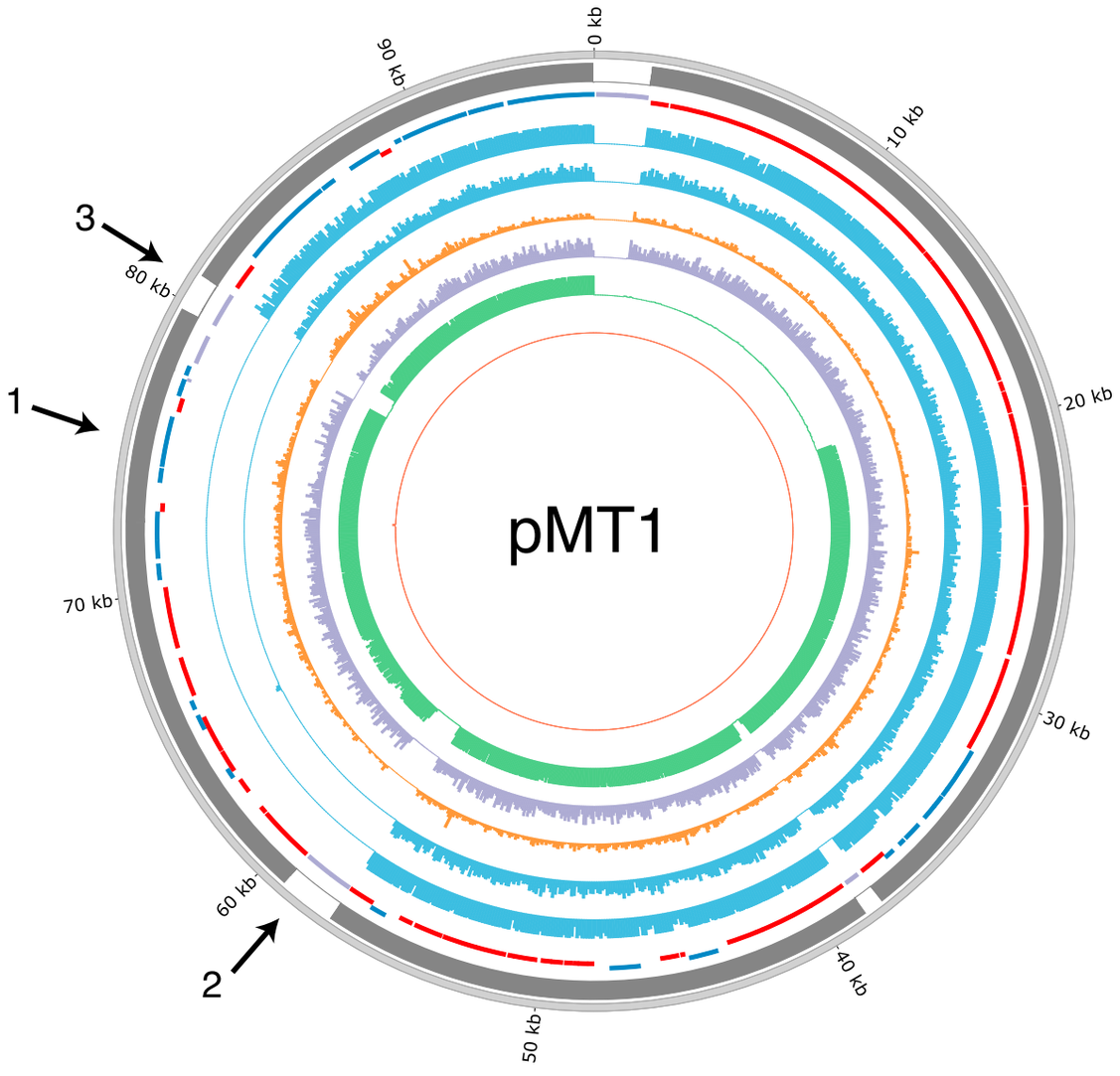
Allentoft*, Sikora* et al (2015) Nature; Rasmussen et al (2015) Cell

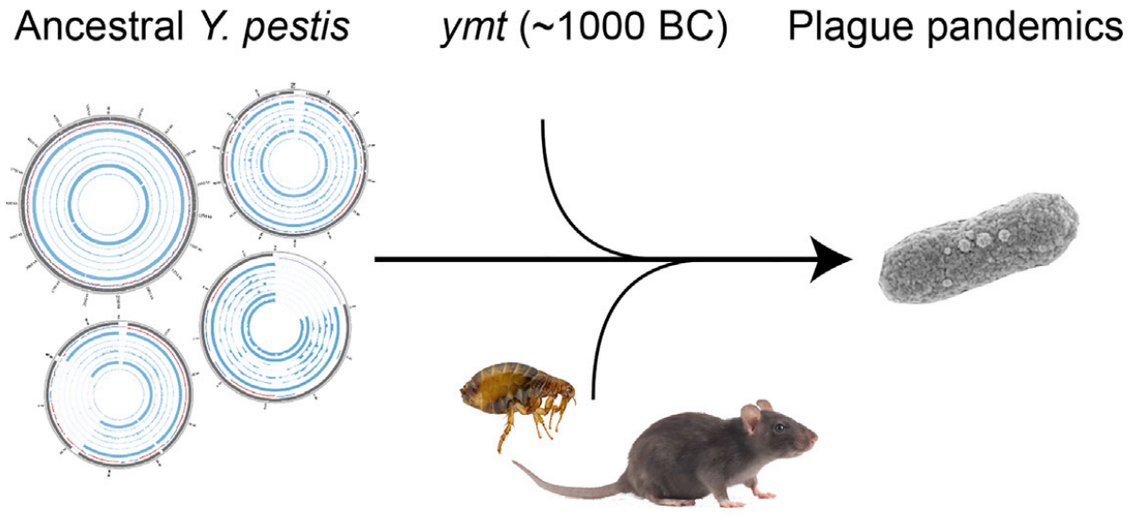
Ancient strains form extinct basal clade



Rasmussen et al (2015) Cell

Evolution of virulence factors

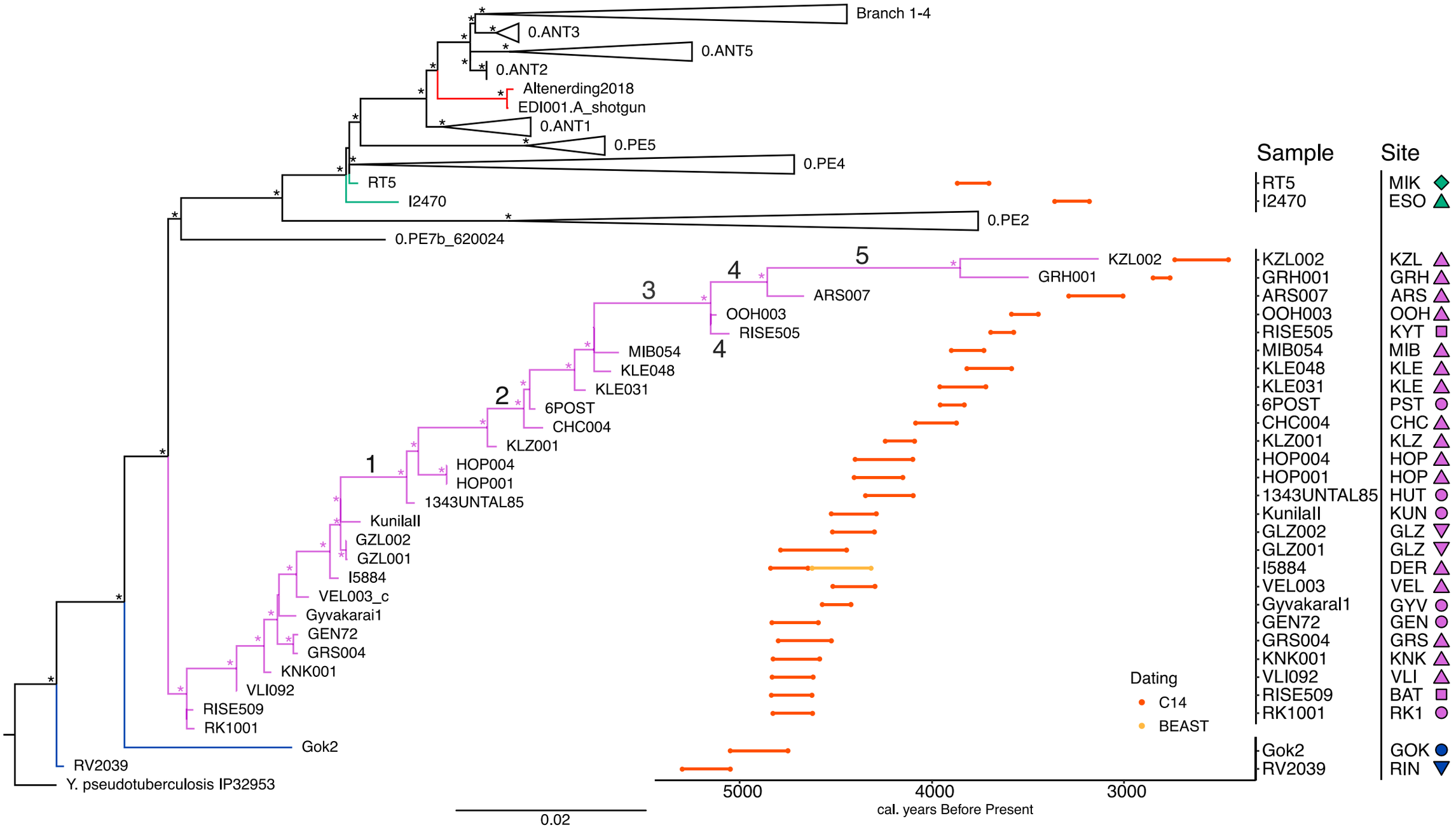




Bronze Age *Yersinia pestis* genomes lack *ymt* gene for efficient transmission in fleas

Rasmussen et al (2015) Cell

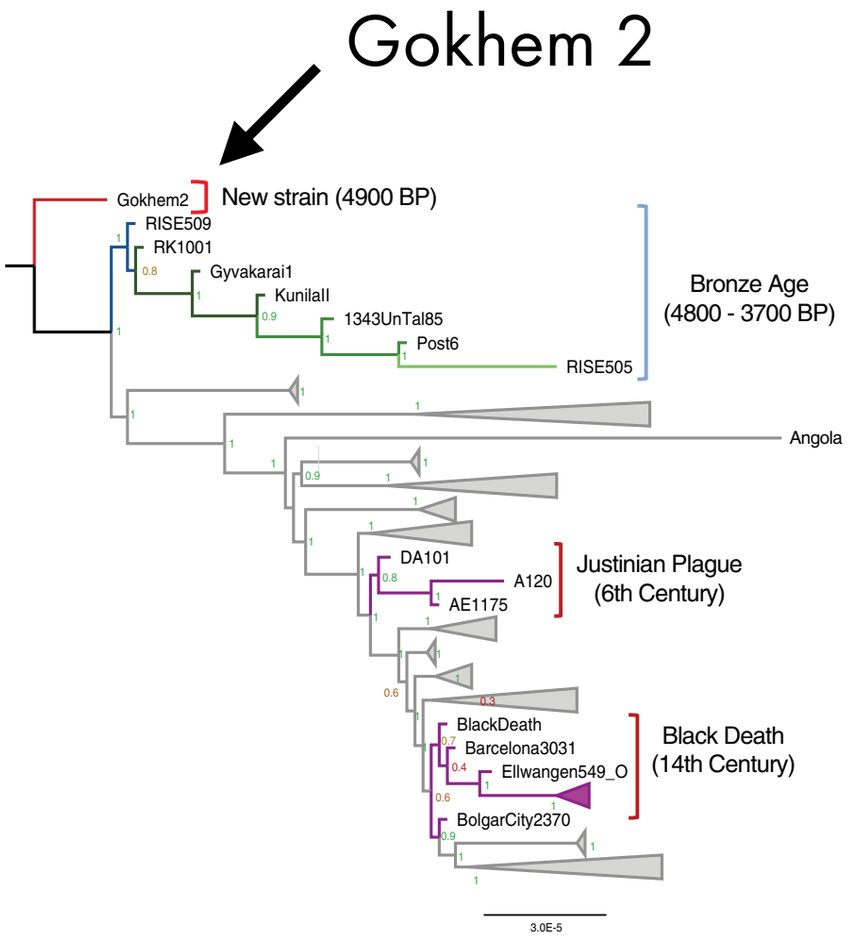
Rapid plague dispersals during the Stone Age



LNBA genomes derive from a single lineage

Andrades Valtuena et al (2022) PNAS

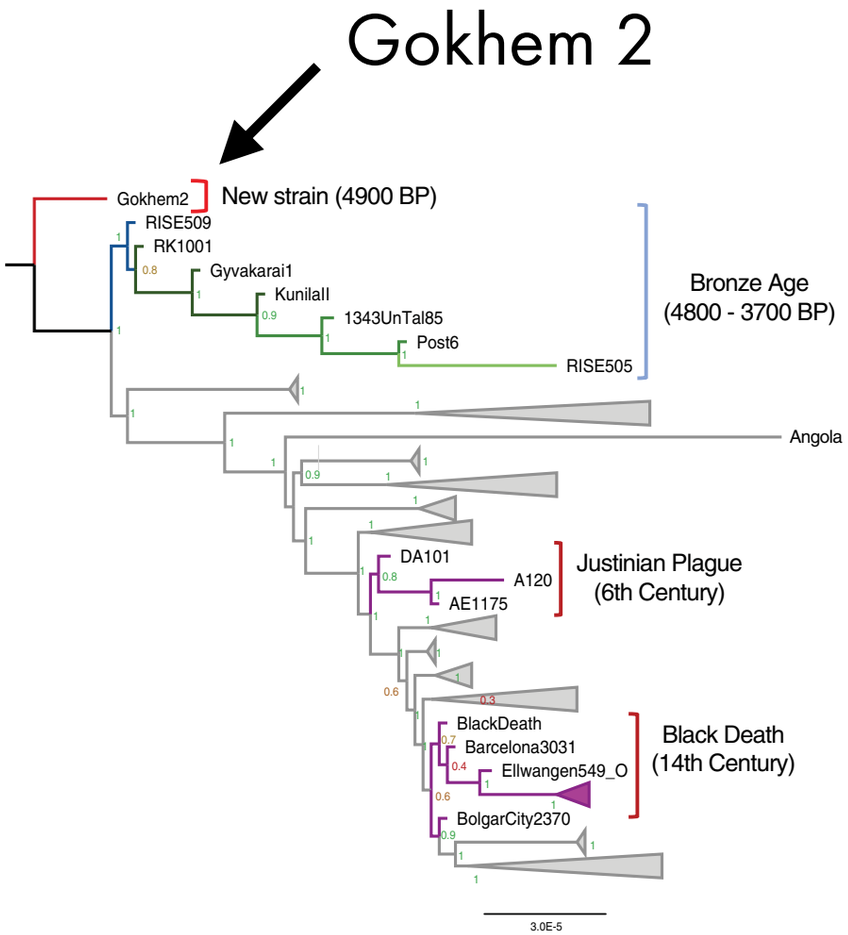
Plague and the Neolithic decline

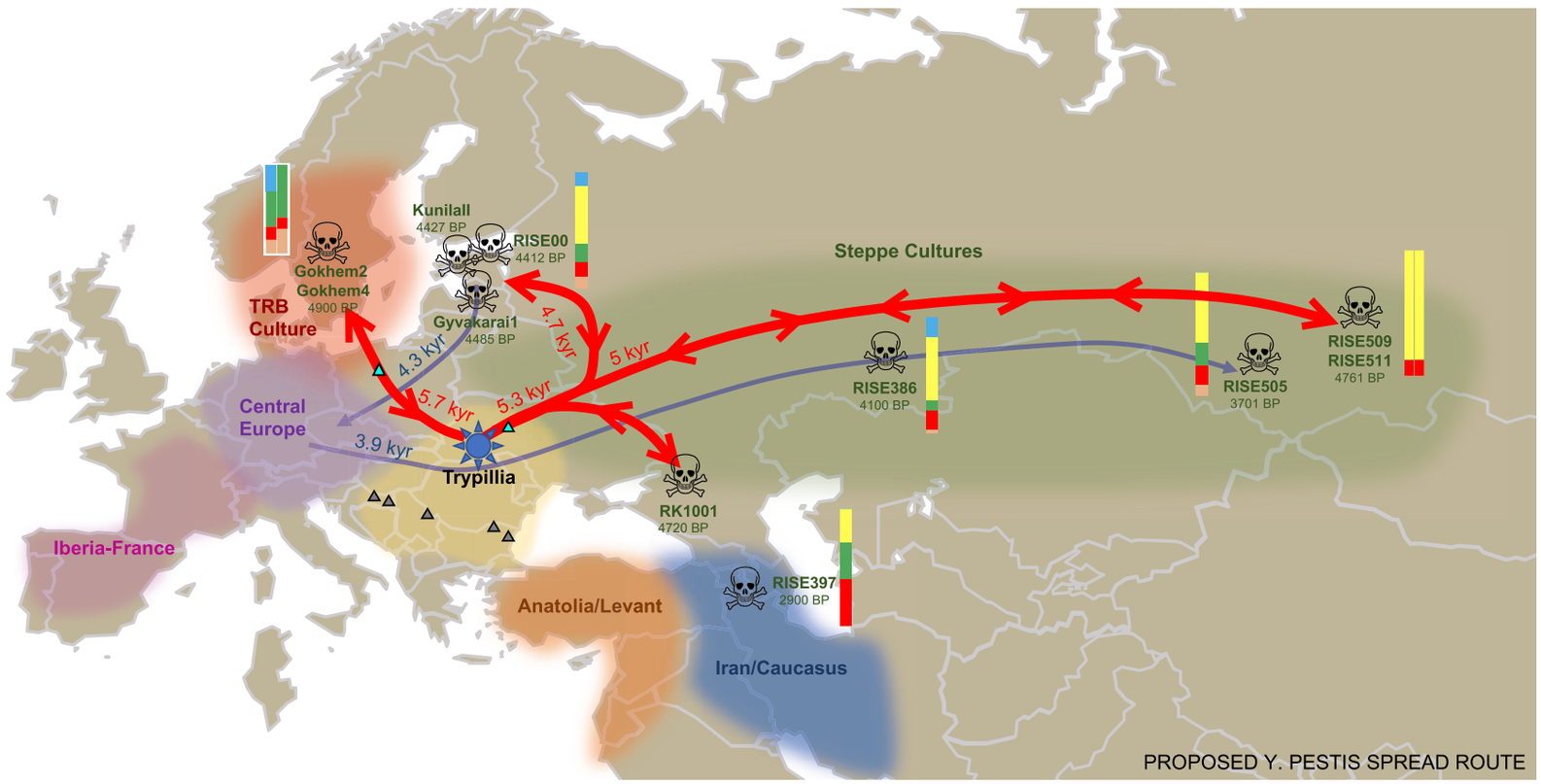


Discovery of a basal plague lineage in Late Neolithic Sweden

Rascovan et al (2018) Cell

Plague and the Neolithic decline

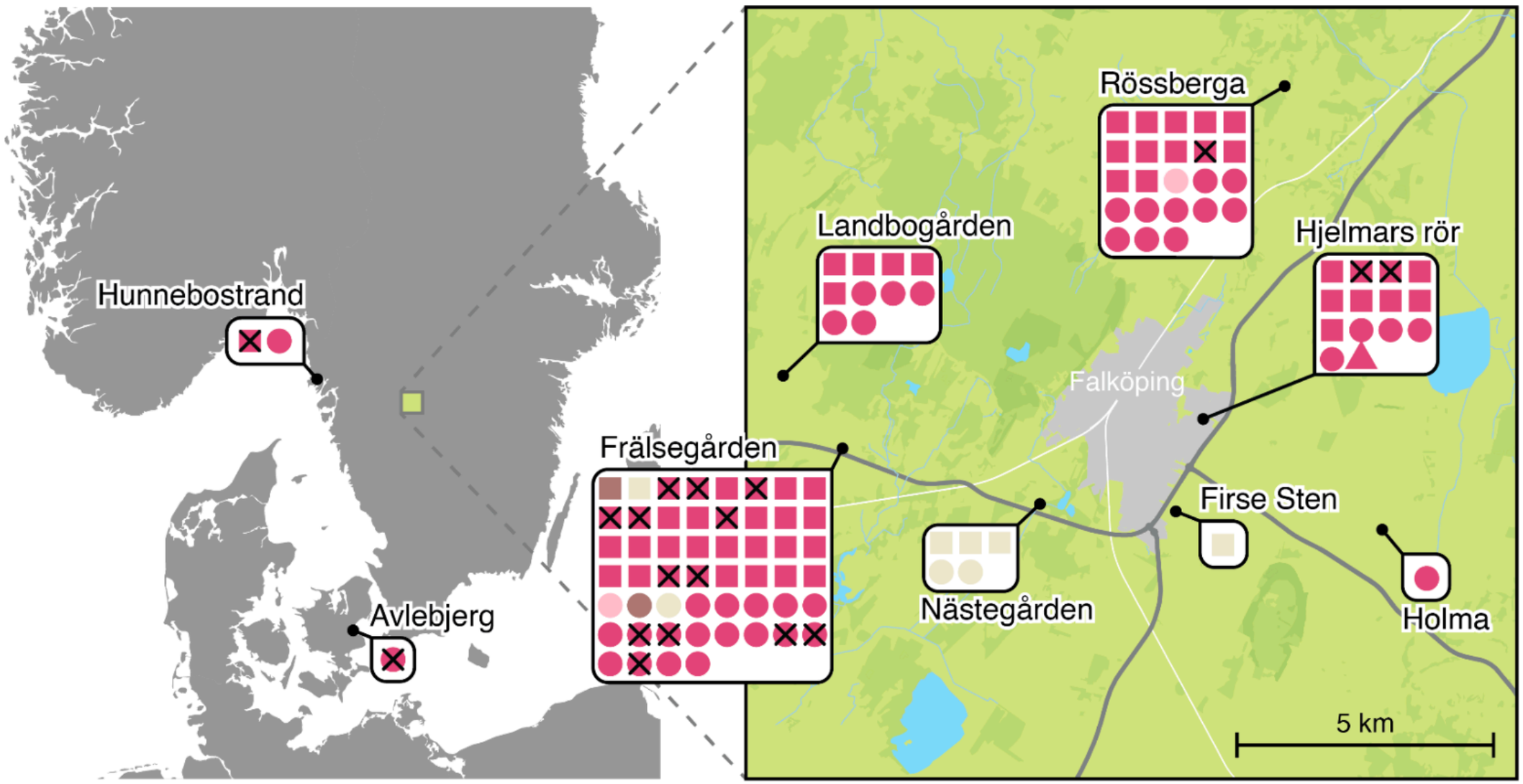




Discovery of a basal plague lineage in Late Neolithic Sweden

Rascovan et al (2018) Cell

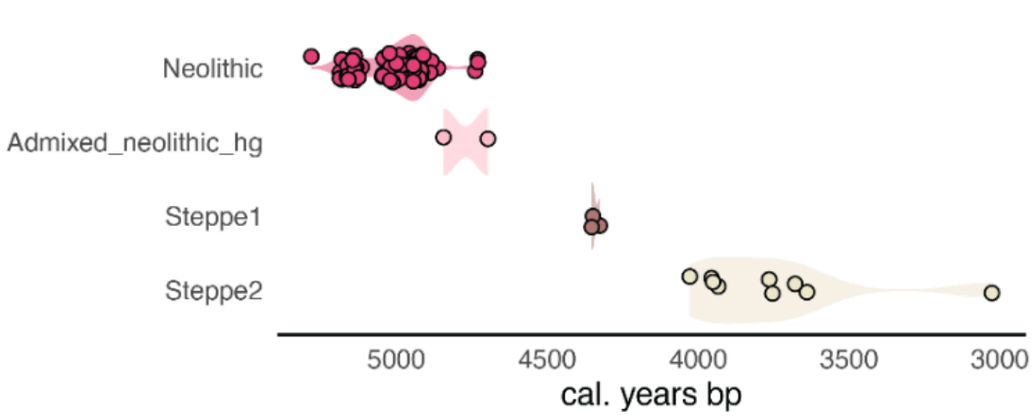
Population genomics of LN Scandinavia

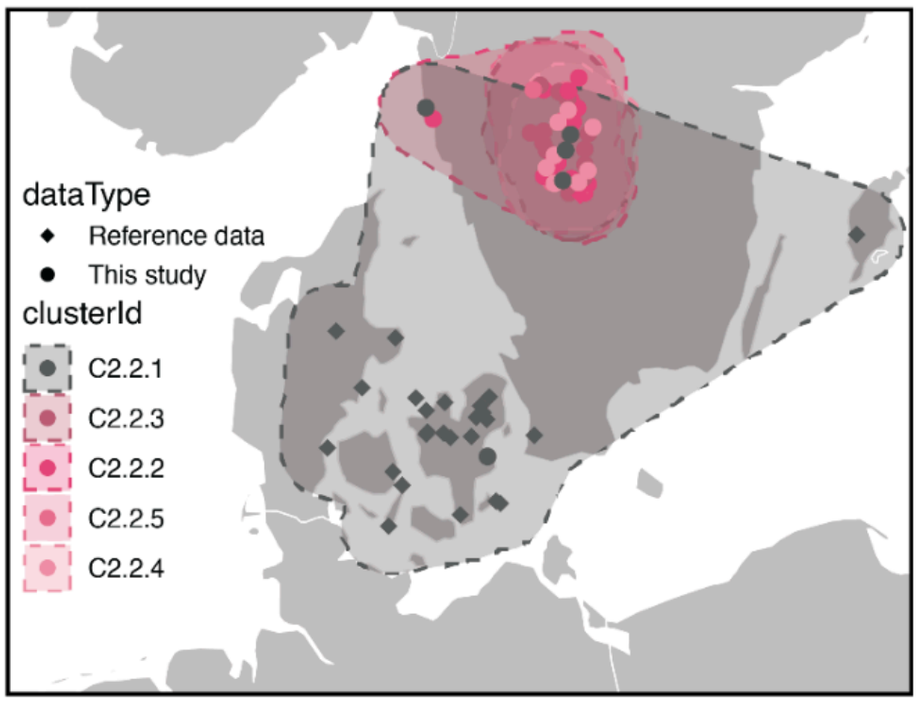


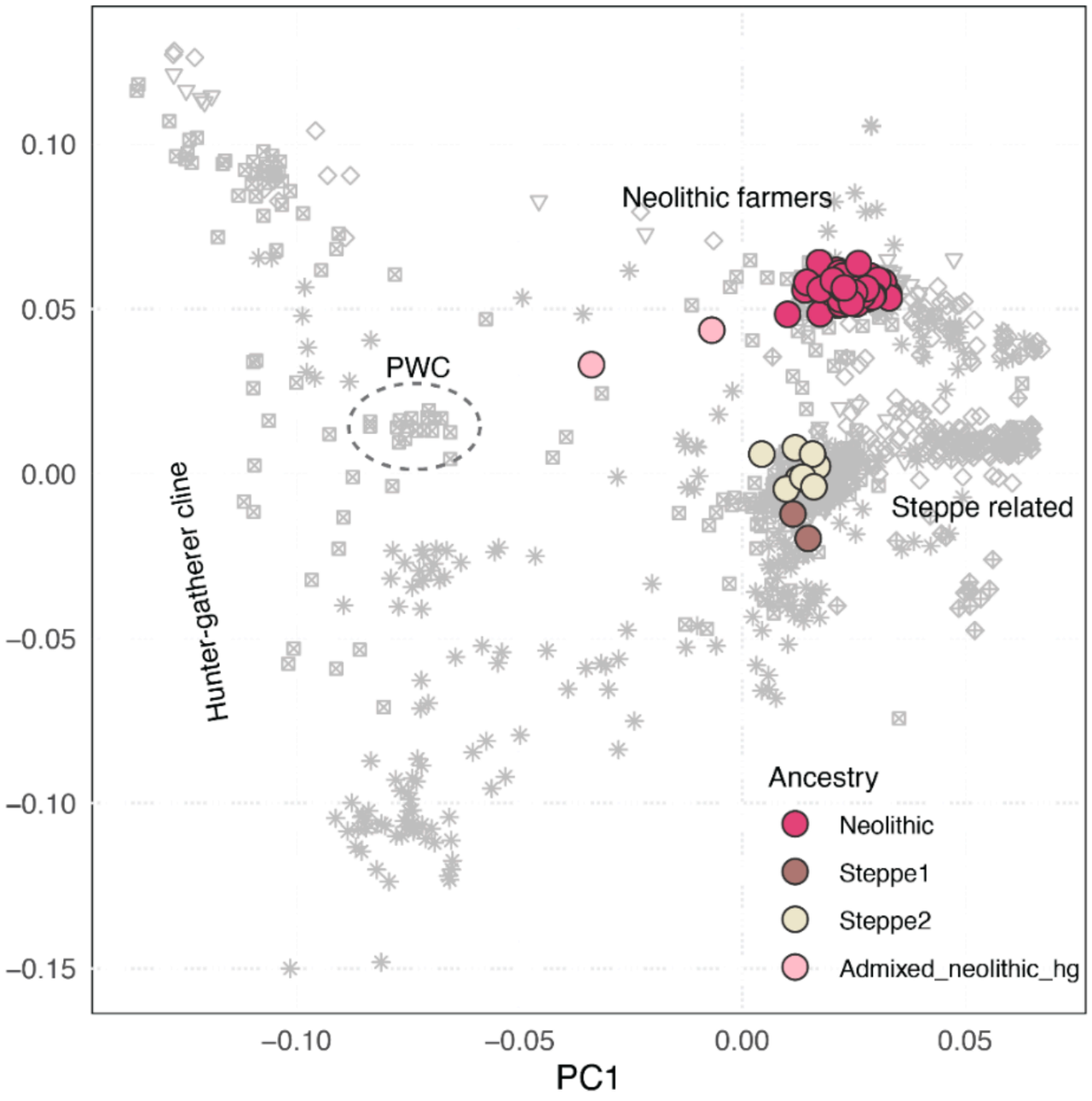
109 ancient human genomes from megalithic graves in southern Scandinavia

Seersholm et al submitted

Multiple distinct ancestries over time



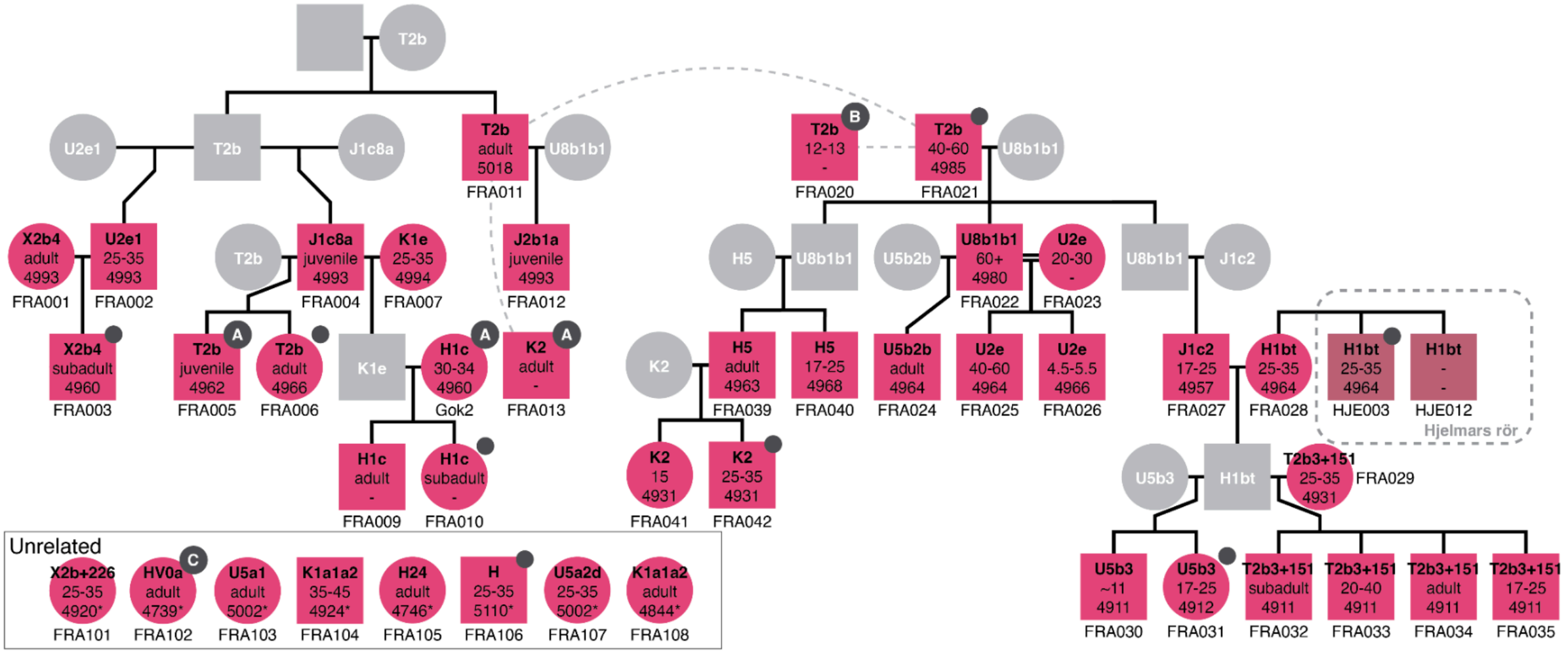




Repeated use of megalithic graves by peoples with different ancestries

Seersholm et al submitted

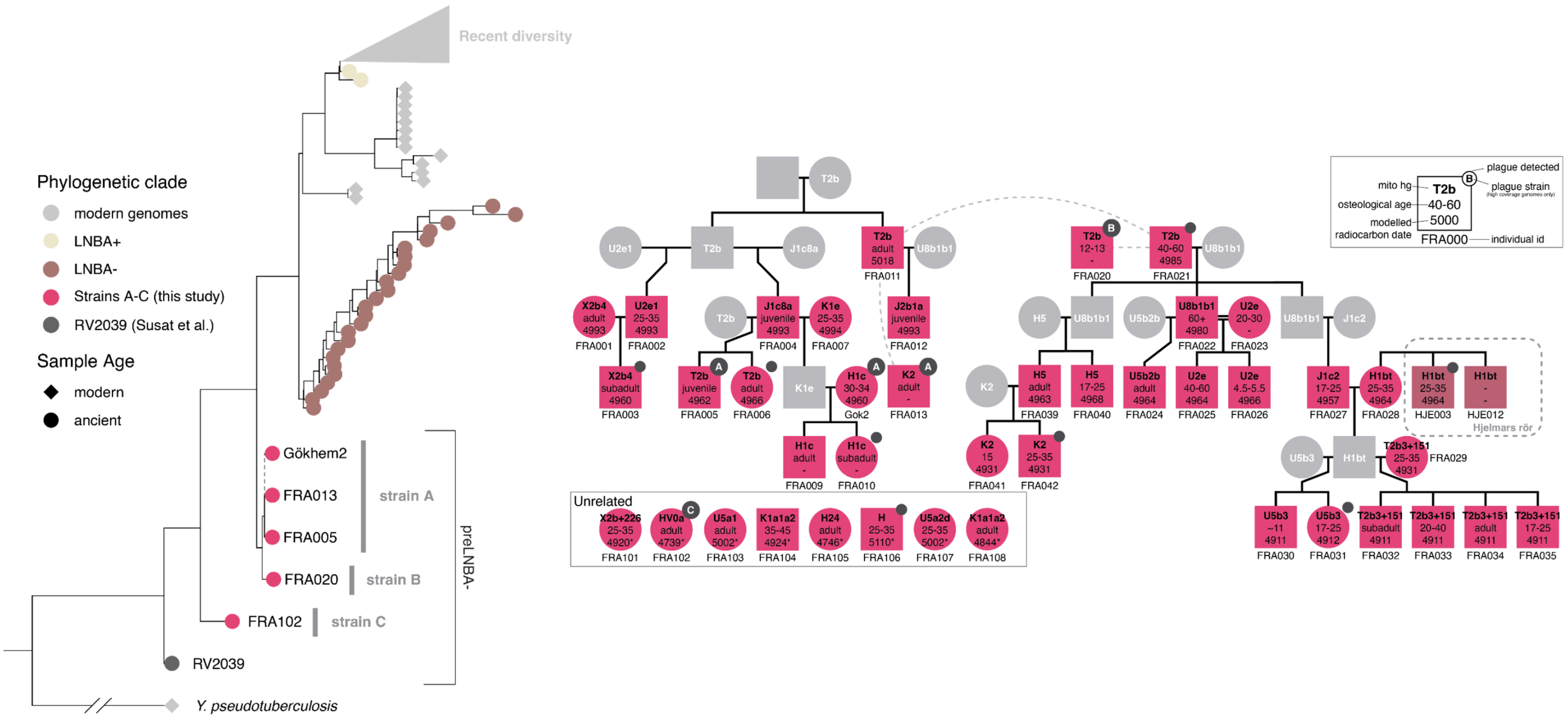
A five generation pedigree of Neolithic farmers



Pedigree of 52 individuals suggests patrilineal and patrilocal social organisation

Seersholm et al submitted

Multiple plague outbreaks across families



High prevalence of plague and multiple outbreaks of distinct strains

Seersholm et al submitted

A sneak peak of next week's Nature issue

Population Genomics of Stone Age Eurasia

Morten E. Allentoft, Martin Sikora, Alba Refoyo-Martínez, Evan K. Irving-Pease, Anders Fischer, William Barrie, Andrés Ingason, Jesper Stenderup, Karl-Göran Sjögren, Alice Pearson, Bárbara Sousa da Mota, Bettina Schulz Paulsson, Alma Halgren, Ruairidh Macleod, Marie Louise Schjellerup Jørkov, Fabrice Demeter, Maria Novosolov, Lasse Sørensen, Poul Otto Nielsen, Rasmus H.A. Henriksen, Tharsika Vimala, Hugh McColl, Ashot Margaryan, Melissa Ilardo, Andrew Vaughn, Morten Fischer Mortensen, Anne Birgitte Nielsen, Mikkel Ulfeldt Hede, Peter Rasmussen, Lasse Vinner, Gabriel Renaud, Aaron Stern, Theis Zetner Trolle Jensen, Niels Nørkjær Johannsen, Gabriele Scorrano, Hannes Schroeder, Per Lysdahl, Abigail Daisy Ramsøe, Andrei Skorobogatov, Andrew Joseph Schork, Anders Rosengren, Anthony Ruter, Alan Outram, Aleksey A. Timoshenko, Alexandra Buzhilova, Alfredo Coppa, Alisa Zubova, Ana Maria Silva, Anders J. Hansen, Andrey Gromov, Andrey Logvin, Anne Birgitte Gotfredsen, Bjarne Henning Nielsen, Borja González-Rabanal, Carles Lalueza-Fox, Catriona J. McKenzie, Charleen Gaunitz, Concepción Blasco, Corina Liesau, Cristina Martinez-Labarga, Dmitri V. Pozdnyakov, David Cuenca-Solana, David O. Lordkipanidze, Dmitri En'shin, Domingo C. Salazar-García, T. Douglas Price, Dušan Borić, Elena Kostyleva, Elizaveta V. Veselovskaya, Emma R. Usmanova, Enrico Cappellini, Erik Brinch Petersen, Esben Kannegaard, Francesca Radina, Fulya Eylem Yediay, Henri Duday, Igor Gutiérrez-Zugasti, Inna Potekhina, Irina Shevnina, Isin Altinkaya, Jean Guilaine, Jesper Hansen, Joan Emili Aura Tortosa, João Zilhão, Jorge Vega, Kristoffer Buck Pedersen, Krzysztof Tunia, Lei Zhao, Liudmila N. Mylnikova, Lars Larsson, Laure Metz, Levon Yepiskoposyan, Lisbeth Pedersen, Lucia Sarti, Ludovic Orlando, Ludovic Slimak, Lutz Klassen, Malou Blank, Manuel González-Morales, Mara Silvestrini, Maria Vretemark, Marina S. Nesterova, Marina Rykun, Mario Federico Rolfo, Marzena Szmyt, Marcin Przybyła, Mauro Calattini, Mikhail Sablin, Miluše Dobisíková, Morten Meldgaard, Morten Johansen, Natalia Berezina, Nick Card, Nikolai A. Saveliev, Olga Poshekhonova, Olga Rickards, Olga V. Lozovskaya, Olivér Gábor, Otto Christian Uldum, Paola Aurino, Pavel Kosintsev, Patrice Courtaud, Patricia Ríos, Peder Mortensen, Per Lotz, Per Persson, Pernille Bangsgaard, Peter de Barros Damgaard, Peter Vang Petersen, Pilar Prieto Martinez, Piotr Włodarczak, Roman V. Smolyaninov, Rikke Maring, Roberto Menduiña, Ruben Badalyan, Rune Iversen, Ruslan Turin, Sergey Vasilyiev, Sidsel Wåhlin, Svetlana Borutskaya, Svetlana Skochina, Søren Anker Sørensen, Søren H. Andersen, Thomas Jørgensen, Yuri B. Serikov, Vyacheslav I. Molodin, Vaclav Smrcka, Victor Merz, Vivek Appadurai, Vyacheslav Moiseyev, Yvonne Magnusson, Kurt H. Kjær, Niels Lynnerup, Daniel J. Lawson, Peter H. Sudmant, Simon Rasmussen, Thorfinn Korneliussen, Richard Durbin, Rasmus Nielsen, Olivier Delaneau, Thomas Werge, Fernando Racimo, Kristian Kristiansen, Eske Willerslev

The Selection Landscape and Genetic Legacy of Ancient Eurasians

Evan K. Irving-Pease, Alba Refoyo-Martínez, Andrés Ingason, Alice Pearson, Anders Fischer, William Barrie, Karl-Göran Sjögren, Alma S. Halgren, Ruairidh Macleod, Fabrice Demeter, Rasmus A. Henriksen, Tharsika Vimala, Hugh McColl, Andrew Vaughn, Aaron J. Stern, Leo Speidel, Gabriele Scorrano, Abigail Ramsøe, Andrew J. Schork, Anders Rosengren, Lei Zhao, Kristian Kristiansen, Peter H. Sudmant, Daniel J. Lawson, Richard Durbin, Thorfinn Korneliussen, Thomas Werge, Morten E. Allentoft, Martin Sikora, Rasmus Nielsen, Fernando Racimo, Eske Willerslev

Genetic risk for Multiple Sclerosis originated in Pastoralist Steppe populations

William Barrie, Yaoling Yang, Kathrine E. Attfield, Evan Irving-Pease, Gabriele Scorrano, Lise Torp Jensen, Angelos P. Armen, Evangelos Antonios Dimopoulos, Aaron Stern, Alba Refoyo-Martinez, Abigail Ramsøe, Charleen Gaunitz, Fabrice Demeter, Marie Louise S. Jørkov, Stig Bermann Møller, Bente Springborg, Lutz Klassen, Inger Marie Hyldgård, Niels Wickmann, Lasse Vinner, Thorfinn Sand Korneliussen, Martin Sikora, Kristian Kristiansen, Santiago Rodriguez, Rasmus Nielsen, Astrid K. N. Iversen, Daniel J. Lawson, Lars Fugger, Eske Willerslev

Questions? Interested in a project?

martin.sikora@sund.ku.dk