



Mass spectrometry-based computational identification of ancient protein sequences to unravel evolutionary history

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Geowissenschaften
der Ludwig-Maximilians-Universität
München

vorgelegt von
Petra Gutenbrunner

München, den 7. September 2021

Erstgutachter: PD. Dr. Gertrud E. Rößner

Zweitgutachter: Dr. Jürgen Cox

Tag der mündlichen Prüfung: 17. Dezember 2021

Contents

Summary	vii
Abbreviations	ix
1 Introduction	1
1.1 Biological systematics and taxonomy	1
1.2 Ancient biomolecules and their evolutionary context	2
1.2.1 Deoxyribonucleic acid (DNA)	3
1.2.1.1 Phylogenetic trees construction	4
1.2.1.2 Ancient DNA limitations	8
1.2.2 Proteins	9
1.2.2.1 From the genome to the transcriptome to the proteome	9
1.2.2.2 The complexity of a proteome	10
1.2.3 Ancient proteins	11
1.3 History of ancient protein sequence identification	12
1.4 Mass spectrometry	13
1.4.1 Ion source	14
1.4.2 Mass analyser	15
1.4.3 Fragmentation methods	15
1.5 Mass spectrometry-based proteomic strategies for peptide and protein identification	19
1.6 Peptide and protein sequence identification based on shotgun proteomics data	21
1.6.1 Spectral library search	21
1.6.2 Sequence database search	22
1.6.2.1 Andromeda	23
1.6.2.2 False discovery rate	24
1.6.3 De-novo sequencing	26
1.7 Applications of ancient proteins	28
1.8 Challenges of ancient protein sequence identification	30
2 Purpose and structure of the thesis	33
3 High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis	35
3.1 Abstract	35
3.2 Introduction	36
3.3 Methods	37

3.4	Results	43
3.5	Discussion	52
3.6	Author's contribution	53
3.7	Additional information	54
3.8	Supplementary information	55
4	Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra	69
4.1	Abstract	69
4.2	Introduction	71
4.3	Methods	72
4.3.1	HeLa dataset	72
4.3.2	Tims-TOF pro dataset	72
4.3.3	Ancient dataset	72
4.3.4	Pre-processing of MS/MS spectra for MaxNovo search	73
4.3.5	NOVOR data preparation and analysis	73
4.3.6	PEAKS data analysis	73
4.3.7	Benchmark based on the HeLa datasets	73
4.3.8	BLAST search	74
4.3.9	Software and data availability	74
4.4	Results and discussion	75
4.5	Conclusion	87
4.6	Author's contribution	88
4.7	Additional information	89
4.8	Supplementary information	90
4.8.1	A user guide on how to run MaxNovo in MaxQuant	90
5	The dental proteome of <i>Homo antecessor</i>	99
5.1	Abstract	99
5.2	Introduction	101
5.3	Methods	102
5.4	Results	110
5.5	Discussion	115
5.6	Author's contribution	116
5.7	Additional information	117
5.8	Extended data figures	119
5.9	Extended data tables	126
5.10	Supplementary information	128
5.10.1	Anthropological background	128
5.10.1.1	Atapuerca	128
5.10.1.2	Dmanisi	131
5.10.2	Supplementary methods and results	132
5.10.2.1	Amino Acid Racemization	132
5.10.2.2	Proteomic Analysis	134
5.10.2.3	Phylogenetic analysis	150

6	Enamel proteome shows that <i>Gigantopithecus</i> was an early diverging pongine	169
6.1	Abstract	169
6.2	Introduction	171
6.3	Methods	172
6.4	Results	180
6.5	Discussion	184
6.6	Author's contribution	185
6.7	Additional information	186
6.8	Extended data figures	187
6.9	Supplementary information	197
7	Conclusion and outlook	213
	Acknowledgement	247

Summary

Mass spectrometry-based ancient protein studies offer a window into the evolutionary past and allow us to deepen our knowledge in the fields of cultural heritage, archaeology and palaeontology. To unravel evolutionary history, traditionally phenotypic traits of fossils are investigated to resolve phylogenetic placement of extinct or unknown organisms. However, phylogenetic analysis is limited for highly fragmented fossil remains due to their lack of species-specific morphological features. In recent years, with advancements of analytical technologies, the study of ancient biomolecules—DNA and proteins—in organic fossil remains has aided phylogenetic reconstruction of several species. Like deoxyribonucleic acids (DNA), proteins comprise genetic information, but are preserved over longer timescales compared to DNA. Therefore, ancient proteins are a valuable resource for very old specimens (1+ million years), especially when ancient DNA (aDNA) is already fully degraded.

Mass spectrometry has evolved to analyse high-throughput data rapidly, accurately and with high sensitivity, making it the key method to study ancient proteins. However, despite recent developments and improvements of high-resolution mass spectrometers to detect proteins in complex mixtures, identification of ancient proteins is still challenging due to several characteristics specific to ancient proteins. Ancient proteins, but also aDNA, are typically low abundant, heavily degraded and highly modified due to post-mortem decay and are accompanied by highly abundant contaminants. These characteristics lead to low quality measurements of surviving proteins in organic tissues. Additionally, protein sequences of unsequenced or extinct species can contain unknown sequence variations due to their evolutionary distance to extant relatives. Commonly used database search approaches are not capable of detecting such sequence variations, as they rely on lookup databases of known protein sequences. The aforementioned challenges cause many acquired protein measurements to remain unidentified. However, high identification rates of accurately identified protein sequences is crucial to determine a species phylogenetic placement.

To address this issue, as part of my PhD, I worked on two complementary computational protein sequence identification approaches: the extension of the database search engine Andromeda and the development of MaxNovo, a novel spectrum graph-based de-novo sequencing algorithm, with the aim to explore novel peptide sequences, which is specifically important for ancient protein studies. Andromeda was extended to include predicted fragmentation intensities, which adds another layer of sequence context into the scoring. We successfully demonstrated an increase in peptide identifications by using the intensity informed Andromeda score in comparison to the conventional Andromeda. To provide highly accurate peptide fragment intensities, we developed in two machine

learning models: DeepMass:Prism, based on deep learning, and wiNNer, based on neural networks. Identification by de-novo sequencing and its possibility to explore novel peptide sequences is specifically important for ancient protein studies with unknown sequence variations. We therefore implemented MaxNovo using expert domain knowledge. Whereas state-of-the-art de-novo algorithms are based on deep learning algorithms, we can show that MaxNovo performs as well as or better than leading deep learning-based algorithms.

Additionally, I contributed to two ancient protein studies by performing the computational data analysis, which required additional implementation of a post-processing workflow for ancient protein sequence reconstruction to obtain sequence variations to perform phylogenetic analysis to resolve evolutionary relationships. In the first study, we provided evidence that *Homo antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins, including modern humans, Neanderthals and Denisovans. In the second study, with proteomic analysis of dental enamel from a 1.9 million year old molar, we revealed that *Gigantopithecus blacki* is a sister clade to orangutans with a common ancestor about 12-10 million years ago.

Abbreviations

Abbreviation	Description
aDNA	Ancient DNA
BI	Bayesian inference
BP	Before present
bp	Base pairs
BUP	Bottom-up proteomics
CID	Collisional induced dissociation
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
EI	Electron ionisation
ESI	Electrospray ionisation
ETD	Electron transfer dissociation
FDR	False discovery rate
FHWM	Full Width at Half Maximum
FP	False positive
HCD	Higher energy collision induced dissociation
HPLC	High-performance liquid chromatography
HTUs	Hypothetical taxonomic units
LC	Liquid chromatography
LCA	Last common ancestor
LC-MS/MS	(High performance) liquid chromatography tandem mass spectrometry
LUCA	Last universal common ancestor
m/z	Mass to charge ratio
MCMC	Markov chain Monte Carlo
MALDI	Matrix-assisted laser desorption ionisation
ML	Maximum likelihood
MP	Maximum parsimony
mRNA	Messenger RNA
MS	Mass spectrometry
MS1	MS spectrum
MSA	Multiple sequence alignment
mtDNA	Mitochondrial DNA
nDNA	Nuclear DNA
NGS	Next generation sequencing

NJ	Neighbour joining
OC	Osteocalcin
OTUs	Operational taxonomic units
PCR	Polymerase chain reaction
PEP	Posterior error probability
PMF	Peptide mass fingerprinting
PSM	Peptide spectrum match
PTM	Post-translational modification
RF	Radio frequencies
RNA	Ribonucleic acid
SAPs	Single-amino acid polymorphisms
SNPs	Single-nucleotide polymorphisms
TDP	Top-down proteomics
TOF	Time-of-flight
TP	True positive
tRNA	Transfer RNA
ZooMS	Zooarchaeology by Mass Spectrometry

Chapter 1

Introduction

1.1 Biological systematics and taxonomy

Taxonomy and biological systematics are related concepts and are crucial to studying the diversification of living forms and their relationships in the past and present. Taxonomy is the study of identification, naming and classification of biological organisms. Biological systematics, on the other hand, is the study of resolving evolutionary relationships of different organisms and determining common ancestry [Panawala, 2017]. Biological classification and phylogeny contribute to biological systematics. The biological classification system is necessary to explain and name all known extinct and extant species.

Around 2,000 BP, Aristotle developed the first biological classification system, which continued to be the dominating classification method until the 19th century [Cain, 2020]. On the island of Lesbos, he observed the sea and marine life, based on which he derived a ranking system from simple to complex organisms independent of any evolutionary aspect. Furthermore, Aristotle distinguished invertebrate from vertebrate animals and differentiated groups further based on their features, for example assigning whales, dolphins and porpoises to the more closely related mammals rather than to fishes.

In the 1750s, Carl Linnaeus, known as the “father of modern taxonomy”, formalised the modern and current system of naming organisms also called the binomial nomenclature [Cain, 2020, Calisher, 2007]. The Linnaeus system creates a taxonomic hierarchy by grouping organisms to taxa based on shared characteristics. Such taxa are assigned a certain rank based on their divergence level. Groups with the same rank can then be aggregated to form a higher-level group. At all levels latinised names are used e.g. *Homo sapiens* for the humans, whereof *Homo* is the genus and *sapiens* is the species name (Figure 1.1).

Linnaeus and other early naturalists used ranks to explain taxonomic relationships, rather than using it for nomenclature purposes, thus, taxon names were more closely associated with taxa than with ranks [Queiroz, 2012]. However, during the middle of the 19th century the rank-based nomenclature emerged. For example names of certain ranks are extended by a rank-specific suffix such as the *-idae* ending for taxa ranked as families. This nomenclature approach is inefficient for hierarchy changes, because the rank change of a taxon requires its name changes as well as of all its descendent taxa. Consequently,

in April 2000, a new phylogenetic nomenclature the International Code of Phylogenetic Nomenclature, also called PhyloCode has been introduced [Cantino and Queiroz, 2003]. Naming under the PhyloCode is based on phylogenetic relationships rather than on taxonomic ranks, thus, changes in rank do not affect names.

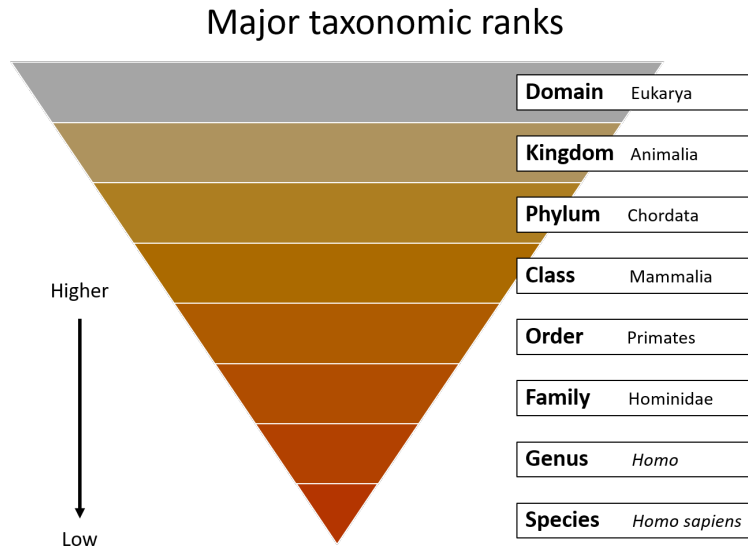


Figure 1.1: Illustration of the major taxonomic ranks based on the example of human. The highest taxonomic rank is domain and the lowest is species (e.g., *Homo sapiens*). At all levels latinised names are used.

On the other hand, phylogeny is the study of evolutionary history and relationships of organisms and reveals subsequently the taxonomic identity of an organism. The relationships are proposed as a result of phylogenetic inference methods that are based on behavioural, morphological, genetic and biochemical features and are visualized as evolutionary trees, also called phylogenetic trees [Choudhuri, 2014]. Traditionally, evolutionary relationships of organisms were resolved by molecular analysis of living organisms and the investigation of phenotypic traits in fossils of extinct organism [Cappellini et al., 2018]. Improvements of analytical technology over the last few decades have enabled the study of ancient biomolecules—molecules that are thousands of years old—directly obtained from remains of extinct organisms or archaeological objects.

1.2 Ancient biomolecules and their evolutionary context

Compared to traditional morphological approaches, molecular analysis of surviving biomolecules in ancient tissues allows deeper insights into evolutionary history as they are not limited to physical shapes. Moreover, biomolecular analysis has provided new insights for example into the genetic divergence from human and apes as well as genetic relationships between human populations [Cann et al., 1987]. For evolutionary studies, molecular species identification without shape restriction is especially advantageous because biomolecules relevant for phylogenetic placement can be still recovered even in heavily fragmented material. Nucleic acids and proteins have provided the highest contribution

in elucidating evolutionary history. Out of these two, the analysis of deoxyribonucleic acids (DNA) is a particularly well-known and widely employed approach, albeit not without limitations. The following sections provide an overview of DNA and proteins along with their ancient aspects.

1.2.1 Deoxyribonucleic acid (DNA)

The DNA contains the genetic instructions for an entire life cycle of cells including their development, maintenance, reproductions and destruction of cells and is unique for an organism. DNA consists of a series of nucleotides with one of four nucleobases: cytosine (C), guanine (G), adenine (A) and thymine (T). Complementary nucleobases (A/T and C/G) of two polynucleotides bind to each other into base pairs (bp), forming a double helix. DNA is denoted as a sequence of the single letters of the nucleobases, encompassing all genetic information of an organism for both, genes and non-coding DNA. A genome is the compendium of all genetic material of an organism and is stored in the cell's nucleus and mitochondria [McWilliams and Suomalainen, 2019]. The mitochondria are responsible for the energy production and harbour the mitochondrial DNA (mtDNA). While the DNA that is stored in the nucleus (nDNA) contains to equal parts the maternal and paternal DNA, the mtDNA was thought to contain solely the maternal DNA. However, McWilliams et al. found evidence that in rare cases the paternal mtDNA is also passed on to the offspring. Compared to nDNA, mtDNA is much smaller. The human nDNA is roughly 3 billion bp long, comprising approximately 20,000 protein-coding genes [Muñoz and Heck, 2014] while the human mtDNA is roughly 16,600 bp long and contains 37 coding genes [Taanman, 1999]. The study of the genome or genomics includes structural, functional and evolutionary analysis, but also covers DNA sequencing and its analysis.

By comparing DNA sequences—nuclear as well as mitochondrial—, evolutionary relationships between different organisms can be determined (Figure 1.2). Differences in the DNA, such as insertions, deletions and substitutions are indications of evolutionary changes [Tajima and Nei, 1984]. Substitutions in form of point mutations in the genome are also called single-nucleotide polymorphisms (SNPs).

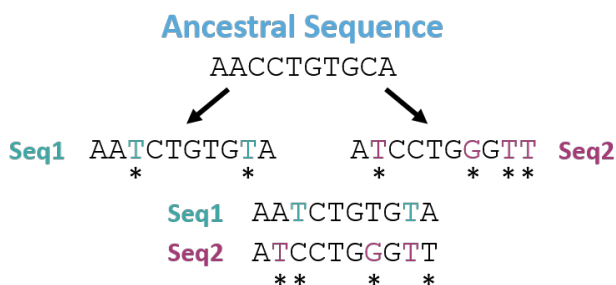


Figure 1.2: Two DNA sequences that derived from the same ancestral sequence. Both offspring sequences developed a different set of point mutations indicated by an asterisk. Adapted from [Salter, 2004]

The higher the similarity of the organisms' DNA, the closer the relationship between those organisms. The mtDNA is the preferred biomolecule for evolutionary studies, because the

mostly maternal genetic information allows to trace back the maternal lineage without the disturbance of the maternal and paternal recombination [Cann et al., 1987]. Moreover, by analysing DNA survival in bone, Allentoft et al. showed that nuclear DNA had degraded at least twice as fast as mtDNA [Allentoft et al., 2012].

Differences in DNA can be used for phylogenetic analyses to resolve evolutionary relationships amongst different taxa or species [Choudhuri, 2014]. The following sections explain the basics of and DNA's role in phylogenetic tree construction and visualisation.

1.2.1.1 Phylogenetic trees construction

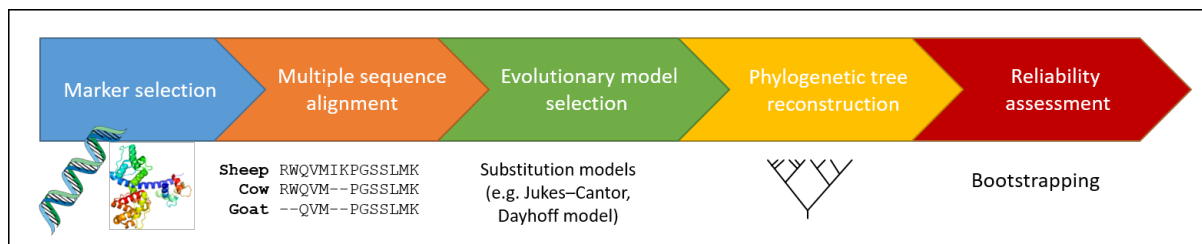


Figure 1.3: Main workflow steps to calculate a phylogenetic tree to determine the systematic position of an organism.

The construction of phylogenetic trees requires five main steps (Figure 1.3) [Choudhuri, 2014]. First, an appropriate molecular marker is selected such as DNA, or specifically genes or proteins. Second, the sequences of the chosen molecular marker are aligned by a multiple sequence alignment (MSA) algorithm to identify blocks of conserved residues and sequence variations.

Next, an evolutionary model is selected, which describes the process of genetic variations by fixed mutations [Arenas, 2015]. Evolutionary models suggest a substitution probability for nucleotides or amino acids. The simplest substitution model for DNA is Jukes-Cantor [Jukes, T.H. and Cantor, 1969], which is based on the assumption that all nucleotides occur and substitute with an equal frequency of 25%. However, mutations between chemically more similar nucleobases (transitions) are more likely than nucleobases with different chemical properties (transversions) [Collins and Jukes, 1994]. To account for these different frequencies, more complex models have been developed such as K80 developed by Kimura et al. [Kimura, 1980]. The F81 model developed by Felsenstein is based on the assumption that nucleotides occur in different frequencies [Felsenstein, 1981]. For proteins, the most well-known model is the Dayhoff model [Dayhoff et al., 1978]. Dayhoff et al. developed an empirical model by determining accepted point mutations, which are observed amino acid substitution frequencies derived from closely related protein sequences. The calculated mutation frequencies of each amino acid are provided as so-called point accepted mutation (PAM) matrix, where each row and column represents one of the twenty amino acids.

After the evolutionary model selection, the phylogenetic tree can be calculated based on the two main approaches: distance-based methods or character-based methods

[Scott and Gras, 2012]. Distance-based methods are applicable to different distance measures such as distances based on Euclidean distance calculations and genetic distances from sequences. Typically, by pairwise comparison a distance matrix is calculated which is used for phylogenetic tree construction [Choudhuri, 2014]. The two most common distance-based algorithms are: the neighbor-joining (NJ) algorithm and "the distance-based algorithm is the unweighted pair group method with arithmetic mean" (UPGMA), whereof UPGMA is the simplest one, but NJ is more popular [Kapli et al., 2020]. Overall, distance-based methods perform poorly for distantly related species.

Character-based methods rely on heritable traits, so-called characters, which are comparable across organisms. Examples for such characters are genetic, morphological, behavioural and molecular attributes [Scott and Gras, 2012]. Character-based algorithms are more complex than distance-based algorithms. In contrast to distance-based algorithms, which are based on pairwise sequence comparison, character-based algorithms simultaneously compare all sequences by looking at one substitution or site at a time. Additionally, character-based algorithms include optimisation methods [Choudhuri, 2014].

Common character-based methods are maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI) [Kapli et al., 2020]. The MP approach minimises the total number of evolutionary steps necessary to explain the given dataset of the organisms for which the phylogenetic position is questioned [Kannan and Wheeler, 2012]. In contrast to MP, ML and BI are based on an explicitly stated model, which allows the incorporation of known features of the process of sequence evolution (e.g. different transition and transversion rates) [Kapli et al., 2020]. The ML method calculates the probability that the selected evolutionary model predicts the observed sequences and for each tree topology the model parameters are optimised to maximise the likelihood [Kapli et al., 2020]. The tree topology with the highest likelihood represents the best estimate of the true phylogenetic tree. BI has been introduced in the 1990s for phylogenetic tree reconstruction and has become very popular [Nascimento et al., 2017]. By using the Markov chain Monte Carlo (MCMC) algorithm, it generates tree topologies and parameters from their posterior. For each tree topology a posterior probability is estimated based on the frequency of a certain topology is suggested by the algorithm. The topology with the highest posterior probability is chosen as the best estimate of phylogeny [Rannala and Yang, 1996, Kapli et al., 2020]. The ML and BI are widely used for phylogenetic tree construction.

Finally, the reliability of the estimated tree is assessed by bootstrapping (Figure 1.4), which provides a variability measure to samples' estimates [Choudhuri, 2014, Efron and Tibshirani, 1986, Efron, 2000, Kapli et al., 2020]. Bayesian methods do not require bootstrapping, because the calculated posterior probability of the estimated phylogeny provides the natural measure of confidence [Rannala and Yang, 1996, Kapli et al., 2020]. Phylogenetic analysis bootstrapping is conducted by repeated resampling of the original data to generate multiple different new subsets. Resampling is performed on random basis, therefore sequences can occur multiple times in different subsets, whereas others might not appear. Next, based on all subsets and the original dataset phylogenetic trees are calculated. The variability is obtained by comparing the topology of the original tree to all bootstrap trees to assess the reliability of the original phylogenetic tree.

Results of phylogenetic analyses can be visually presented as phylogenetic trees. The topology describes the structure and composition of all edges and nodes of a tree (Figure 1.5), whereof the nodes represent the taxonomic units that can be species or populations as well as genes or proteins [Choudhuri, 2014, Scott and Gras, 2012]. An edge, also referred to as branch, connects two nodes (i.e., taxonomic units) and when displaying the phylogenetic tree in form of a phylogram (Figure 1.5, A), the branch length is proportional to the evolutionary divergence and consequently, to the time of the development of evolutionary relationships between taxonomic units. For example, in a phylogenetic tree constructed based on DNA sequences, the branch length is a result of occurring nucleotide substitutions that arise between branching points.

Nodes can be internal or terminal and terminal nodes are also referred to as leaves or operational taxonomic units (OTUs) [Choudhuri, 2014]. OTUs are the taxonomic units of interest that are compared to each other to determine their evolutionary relationship. Internal nodes are hypothetical taxonomic units (HTUs) representing the last common ancestor (LCA) of the subsequent nodes.

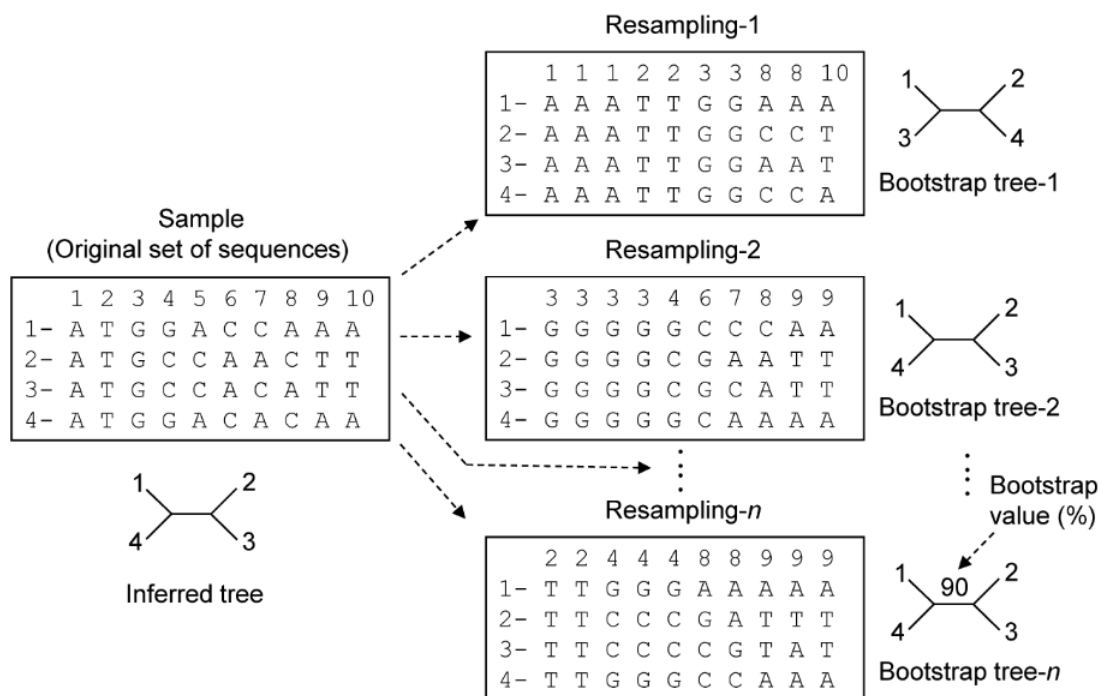


Figure 1.4: Illustration of the bootstrap method performed on common phylogenetic inference methods such as maximum parsimony (MP) and maximum likelihood (ML). The original data is multiple times randomly resampled by allowing repetition of data or characters. For each of the created new bootstrap datasets a phylogenetic tree is calculated. The tree topology of the bootstrap trees and the tree calculated based on the original data is compared to assess the reliability of the original phylogenetic tree [Choudhuri, 2014].

The emerging descendants or taxa of a single node are classified as sister group or sister taxa. When they share specific characteristics that uniquely distinguish them from other taxa, they can be grouped as clade. A clade is composed of a last common ancestor and its descendants. A taxonomic unit that is exclusive and not part of any clade is called outgroup. While phylogenetic trees can be presented as unrooted (Figure 1.5, B), an outgroup is typically added to generate rooted trees. Rooted trees have a last universal common ancestor (LUCA) functioning as a root node of which all descendants originate from.

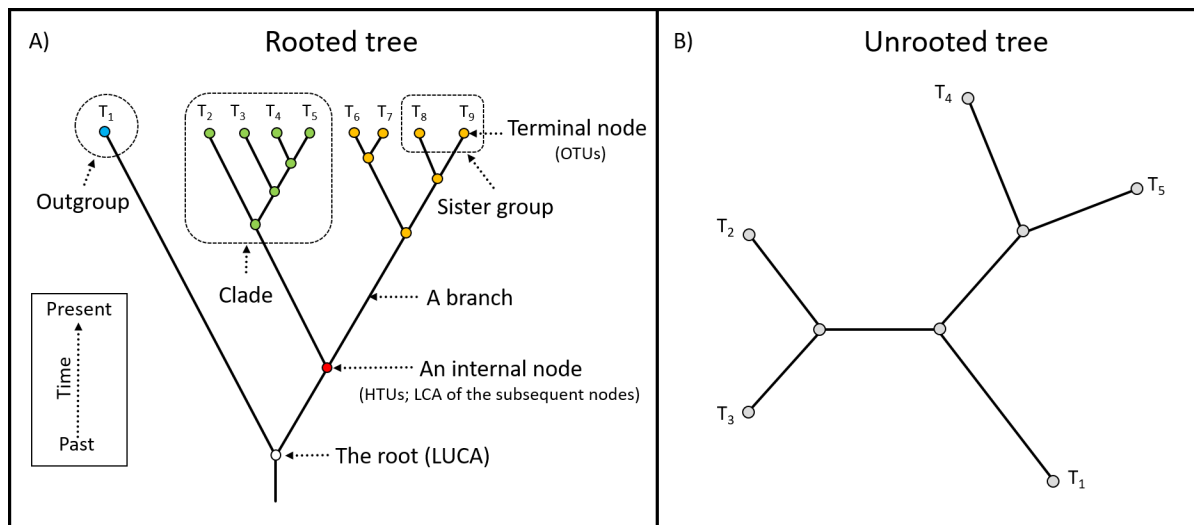


Figure 1.5: The topology of a phylogenetic tree is described by the composition of all nodes and branches. A tree can either be A) rooted or B) unrooted. Rooted trees can be enforced by adding an outgroup that is evolutionary distant enough to fall outside of the main clade, which introduces a last universal common ancestor (LUCA). Nodes can be internal and terminal also referred to as hypothetical taxonomic units (HTUs) and operational taxonomic units (OTUs), respectively. Adapted from [Choudhuri, 2014].

The first ancient DNA based phylogenetic tree was successfully constructed in 1984. More specifically, ancient mtDNA was sequenced from a quagga, an extinct zebra-like equid species. Based on mtDNA comparisons of the quagga and three other mammals, the calculated phylogenetic tree correctly resolves that quagga is most closely related to the zebra and more closely related to cow than to human [Higuchi et al., 1984]. Since then, many aDNA studies have followed. Next generation sequencing (NGS), the state-of-the-art DNA sequencing technology, provides untapped potential for ancient DNA based phylogenetic analysis [Behjati and Tarpey, 2013, Schuster, 2008].

1.2.1.2 Ancient DNA limitations

While ancient DNA harbours significant potential to understand evolutionary history, its characteristics pose several limitations on its applicability for ancient studies. The success of retrieving aDNA is heavily dependent on the preservation conditions such as moisture, oxygen and temperature during deposition [Allentoft et al., 2012, Dabney et al., 2013]. The decay of organic tissue as well as of DNA immediately starts after death due to the lack of enzymatic repair mechanisms of living cells [Allentoft et al., 2012, Dent et al., 2004]. Soft tissue undergoes the post-mortem process of autolysis and putrefaction followed by decomposition until the hard tissues such as bone, teeth and cartilage remain. After cell death, the DNA is cleaved into fragments by nucleases and micro-organisms further fragment the DNA during decomposition [Allentoft et al., 2012]. Hence, aDNA extracted from remains of dead organisms is without exception degraded [Dabney et al., 2013].

Following the pioneering work of Higuichi et al., where phylogenetic relationships were resolved based on ancient DNA, the potential of aDNA was further explored [Dorsey and Dill, 1989, Higuichi et al., 1984, Pääbo et al., 1989]. However, damage to the molecular structure often hindered or limited a detailed aDNA analysis [Dabney et al., 2013, Francalacci, 1995]. The invention of the polymerase chain reaction (PCR) has abet such studies due to selective amplification of short DNA fragments despite the risk of amplifying DNA of modern contaminants [Pääbo et al., 1989]. In 1992, DeSalle et al. claimed to have recovered aDNA from termites [DeSalle et al., 1992] preserved in amber that are estimated to be 25 million to 30 million years old. One year later, Cano et al. reported the extraction of aDNA from a 120–135-million-year-old weevil [Cano et al., 1993]. However, follow up studies have raised the concern that the extracted aDNA is a result of modern environmental DNA contamination (e.g., from human tissue, microbes, plants) [Pääbo et al., 2004, Pedersen et al., 2015, Peris et al., 2020, Willerslev and Cooper, 2005]. Until 2021, the oldest recovered aDNA was obtained from a horse specimen dated to be 780-560 thousand years old [Orlando et al., 2013], which was then replaced by aDNA successfully retrieved from mammoth specimen, which are more than one million years old [van der Valk et al., 2021].

As long as at least partially intact aDNA can be obtained, it is the preferred biomolecule for ancient studies, since it provides information in the highest resolution. However, ancient proteins can provide information on longer temporal scales, when aDNA is already fully degraded [Wadsworth and Buckley, 2014].

1.2.2 Proteins

As previously mentioned, the genome contains the information for an organism's proteins, encoding amino acids as building blocks of proteins. The term proteome was introduced, in analogy to the genome, to describe the entire set of proteins expressed by a genome, organelle, cell, tissue or organism at a certain time under defined conditions [Calvete et al., 2014]. The large-scale characterisation of entire proteomes is called proteomics [Graves and Haystead, 2002].

Based on the standard genetic code there are 20 different amino acids that can be assembled to a sequence by so called peptide bonds (Figure 1.6). The N-terminus of the resulting sequence is at the amino group and the C-terminus is at the carboxyl group.

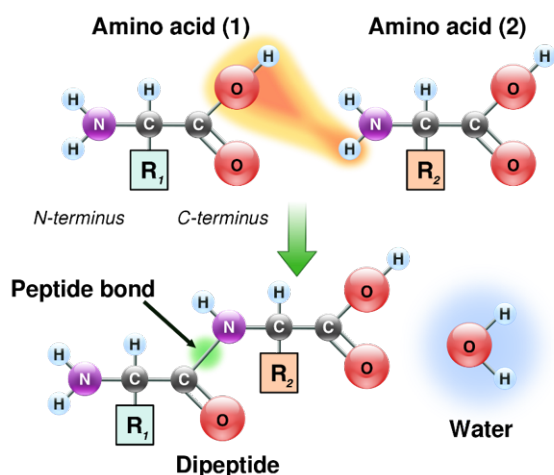


Figure 1.6: The chemical reaction between two amino acids results in the formation of a peptide bond with exclusion of water. The amino group (NH₂) represents the N-terminus and the carboxyl group (COOH) represents the C-terminus of a chain of amino acids. (https://en.wikipedia.org/wiki/Peptide_bond/media/File:Peptidformationball.svg, Access date: 2021-08-28)

Short sequences are typically called peptides and sequences with a length of 20 or more are called proteins. The following sections present a brief overview of gene expression that synthesises proteins and its resulting cellular complexity.

1.2.2.1 From the genome to the transcriptome to the proteome

Proteins are synthesized by a process called gene expression that transcribes a protein-coding gene to ribonucleic acid (RNA). Subsequently, the RNA is part of the translation of proteins at ribosomes (Figure 1.7). The composition of all RNA transcripts is called transcriptome. Like DNA, RNA consists of a chain of nucleotides with four nucleobases: cytosine (C), guanine (G), adenosine (A) and uracil (U), replacing thymine (T) in the DNA. RNA is single stranded, as opposed to double stranded DNA, and folded onto itself. Different types of RNA are involved in gene expression, such as messenger RNA (mRNA) and transfer RNA (tRNA).

mRNA is the result of a process called RNA splicing that removes non-coding regions (Figure 1.7). Consequently, only protein encoding regions (exons) remain, thus providing the information on the amino acids in the encoded protein. However, not all exons have to be included in the resulting mRNA. Different combinations of exons can be synthesised via alternative splicing, hence giving rise to different protein variations – also called isoforms – that originate from the same gene. Based on the information in the mRNA, the tRNA facilitates protein assembly by providing the required amino acids.

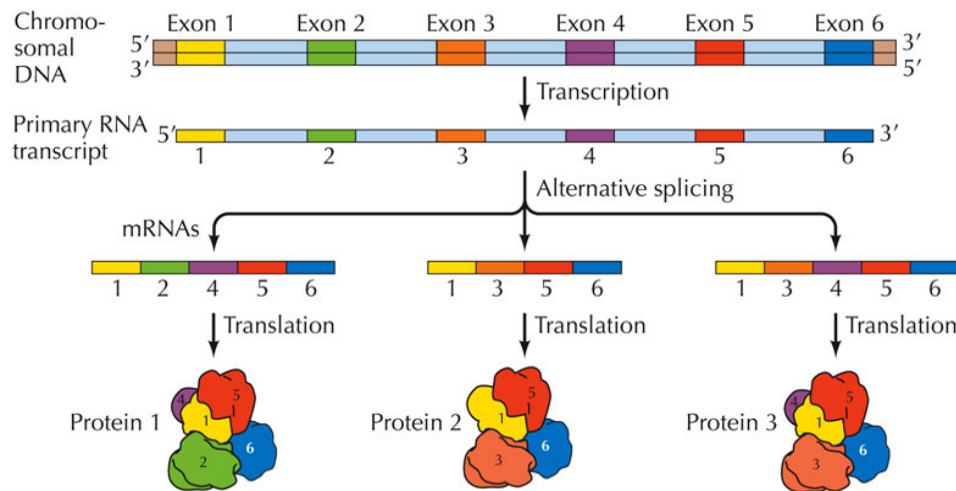


Figure 1.7: Brief overview of the gene expression machinery. First, the DNA is transcribed into RNA, which is then converted to mRNA by alternative splicing, which only consists of exons. Finally, the resulting mRNAs are translated into proteins (<https://sites.google.com/site/bio1040genbio2/chapter-18-regulation-of-gene-expression/alternative-rna-splicing-and-mrna-degradation>, Access date: 2021-08-28).

1.2.2.2 The complexity of a proteome

The functional diversity and its resulting phenotypes are not uniquely assigned to the genome. Rather, the genome is the starting point of an exponentially growing complexity driven not only by the transcriptome and proteome, but also the interactions among molecules in a cell (interactome) (Figure 1.8).

As mentioned previously, alternative splicing increases cellular complexity by giving rise to a range of protein isoforms based on the same protein-coding gene. Post-translational modifications (PTMs) of proteins into unique proteoforms further increases the complexity. There are two different types of PTMs in proteins: the covalent modifications and the covalent cleavage of peptide backbones [Walsh et al., 2005]. Covalent modifications unite all covalent additions of other molecules to a protein that can either occur on the N- or C-terminal of a protein or on an amino acid side chain. Furthermore, the attachment of covalent functional

groups to a protein can be mediated by specific enzymes (*in vivo*) or they can occur during the sample preparation (*in vitro*) as artefactual modifications [Liu et al., 2013]. Covalent modification can also be used to attach specific molecules as tags in relative protein quantification [Thompson et al., 2003]. The covalent cleavage of peptide backbones causes the breakdown of a protein into peptides (proteolysis) and consequently is involved in activity and lifetime control mechanisms of each protein in a cell [Walsh et al., 2005].

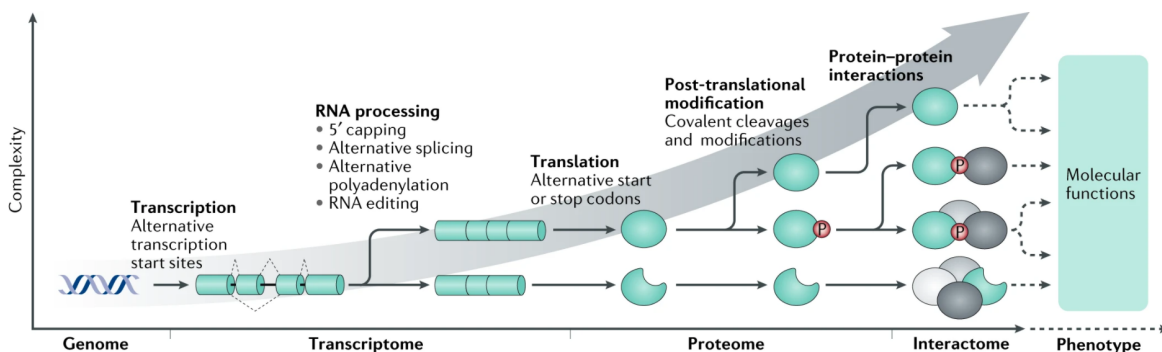


Figure 1.8: The exponentially growing cellular complexity [Bludau and Aebersold, 2020]

1.2.3 Ancient proteins

Like DNA, protein molecules contain genetic information that allow phylogenetic reconstruction. Given the temporal constraints of aDNA, ancient proteins are essential to investigating time scales beyond one million years, as they can survive in archaeological and palaeontological materials for such a long period of time [Andrews et al., 1985]. Biological substrates with a high mineral component such as bone, dentine or enamel are the best-preserved tissues and are therefore the focus of most ancient DNA and protein studies.

The calcified tissue bone is found in vertebrates. Of this tissue, 30% are organic components comprising more than 30 proteins of which type 1 collagen (COL1) is most abundant [Feng, 2009]. Overall, collagen is the most abundant structural protein in animals and constitutes in humans one third of all expressed proteins. The stability of collagen proteins is due to its unique feature: a right-handed bundle of three parallel, left-handed polyproline II-type helices held together by hydrogen bonds [Shoulders and Raines, 2009, Cappellini et al., 2018]. However, COL1 is highly conserved, thus not an ideal phylogenetic marker to distinguish species [Cappellini et al., 2019, Marks, 1988]. Dentine consists of 20% organic matter (by weight) of which 85-90% are a collagenous matrix. As in bone, COL1 is the most abundant protein in dentine [Jágr et al., 2012, Salmon et al., 2013].

Dental enamel is the hardest substance in the human body and the most extreme case of mammalian biomineralization. Measured by weight, only 1-1.5% are organic matter of which the protein amelogenin constitutes approximately 90%. Amelogenin is expressed by the gene AMELX, located on the X chromosome. In males, it is also expressed by the gene AMELY on the Y-chromosome. Hence, it is possible to determine the sex of a specimen

which is essential when studying the human past [Stewart et al., 2017]. Besides amelogenin, the enamel proteome also comprises enamelin, ameloblastin and tuftelin as well as the protease matrix metalloproteinase 20 (MMP20) [Chen and Liu, 2014].

1.3 History of ancient protein sequence identification

First attempts to study ancient protein residues have been made in 1954, about two decades before the advent of DNA sequencing [Abelson P.H., 1954, Cappellini et al., 2018]. With the application of immunoassay approaches it became possible to not only identify single amino acids, but to also target specific epitopes targeted to fossil proteins (Lowenstein, 1981). Nevertheless, results obtained from immunodetection methods were insufficient to infer evolutionary relationships as they were able to target but not sequence ancient proteins.

An important milestone for ancient protein sequencing was set with the invention of Edman degradation in 1967 [Edman and Begg, 1967]. However, the success of protein sequencing depended on high concentrations of purified and undamaged proteins in the sample, limiting the applicability in ancient protein studies due to the small material quantity in fossils.

In 2000, mass spectrometry (MS) was first applied to ancient protein studies [Ostrom et al., 2000] and has since then revolutionized ancient protein studies into a rapidly emerging field evolving in multiple different directions. The following section will introduce mass spectrometry, a powerful analytical technique that facilitates sequencing, identification and quantification of peptides and proteins.

1.4 Mass spectrometry

The aim of mass spectrometry (MS) analysis is to identify the analytes present in the sample on a large scale [Han et al., 2008]. In proteomics, these analytes are typically proteins or peptides that are extracted in a sample preparation step from a biological material and are identified by measuring their mass-to-charge (m/z) ratio. MS is currently the method of choice for proteomics studies as it allows for high-throughput analysis of complex protein mixtures.

A mass spectrometer includes at least three main components (Figure 1.9):

- an *ion source* that converts the analyte of interest in its intact form into gas phase ions
- a *mass analyser* that separates ions according to their m/z ratio
- a *detector* that measures and amplifies the signal of the separated ions to obtain and the abundance (intensity) of each ion at a certain m/z value

The measured ions are provided as two-dimensional mass spectrum (MS1) with the intensity as a function of the m/z ratios (Figure 1.9), A).

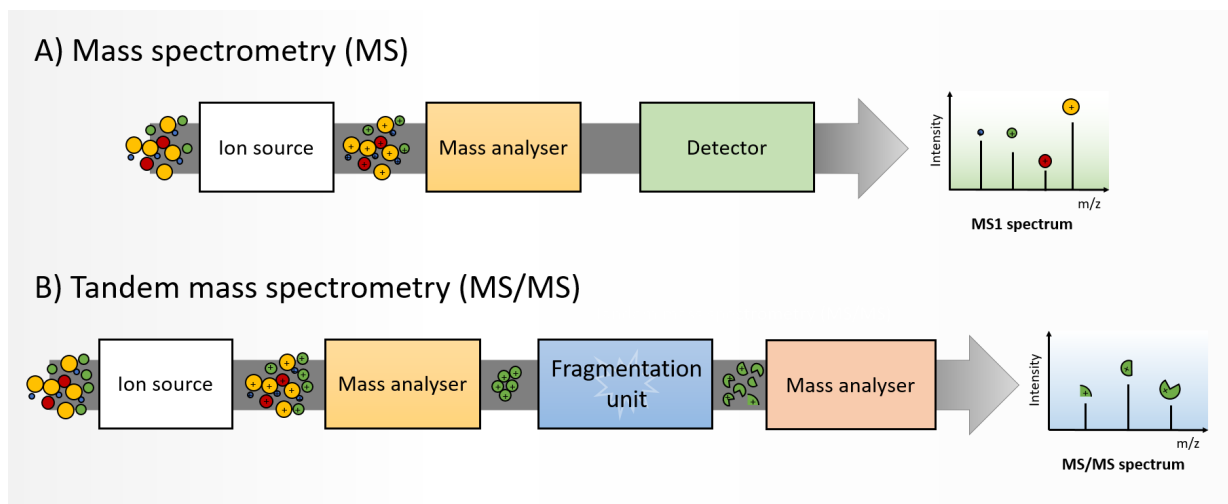


Figure 1.9: Main steps of a A) mass spectrometry (MS) analysis and a B) tandem mass spectrometry (MS/MS) analysis. MS/MS analysis performs an additional fragmentation step, where the selected analytes are further fragmented to generate a fragmentation pattern that allows sequence identification.

Identification of an analyte based on the measurement of its m/z is often ambiguous because several different amino acid sequences can have the same mass and charge state. To enable complete sequence characterization, the ions are further fragmented in a separate fragmentation unit (Figure 1.9), B; commonly used fragmentation methods will be explained in more detail in Section 1.4.3 Fragmentation methods). Like the intact ions, the fragmented ions are separated by a mass analyser and eventually detected. The output of this so-called tandem MS (MS/MS) analysis is an MS/MS spectrum, each peak representing a unique fragment ion along with its intensity.

Complex mixtures of analytes are typically not subjected to the MS all at once. Instead, the mixture is first subjected to a reversed-phase high-performance liquid chromatography (HPLC) that is directly coupled to the MS instrument. HPLC is an analytical method used for separation, identification and quantification of each component and it involves two phases: a non-polar stationary and a polar mobile phase [Dorsey and Dill, 1989, Thammana, 2016]. A pump pushes a solvent gradient (mobile phase) of increasing organic content (e.g., acetonitrile, methanol) including the sample mixture through a column that is filled with solid absorbent material (stationary phase). The separation is based on hydrophobic interaction strength of the solvent to the hydrophobic stationary phase. As a result, hydrophilic compounds are eluted first and more hydrophobic peptides are gradually released [Dorsey and Dill, 1989, Steen and Mann, 2004, Thammana, 2016]. By coupling HPLC to the mass spectrometer the two-dimensional output is turned into a three-dimensional output by adding the retention time information to a peak with a certain m/z ratio and intensity. The following sections provide a detailed description of the main elements of MS.

1.4.1 Ion source

The ion source is the component of a MS that ionises an analyte by adding one or more charges, in proteomics often by adding protons (protonation), to its molecules to make them measurable and detectable by electric and magnetic fields. Ion sources can be categorised in hard ioniser and soft ioniser. Hard ionisers such as the electron ionisation (EI) cause high fragmentation due to high quantities of residual energy hindering the analysis of labile biomolecules [Banerjee and Mazumdar, 2012]. This issue was solved with the invention of the two soft ionisation techniques: electrospray ionisation (ESI) [Yamashita and Fenn, 1984] and matrix-assisted laser desorption ionisation (MALDI) [Karas and Hillenkamp, 1988]. These techniques are currently the most common ionisation approaches used in proteomics.

In ESI, a solution containing the analyte flows through a small capillary where high voltage is applied resulting in a fine spray of charged droplets and rapid solvent evaporation. The solvent evaporation increases the electrical charge density at the surface as the droplet size decreases causing a further subdivision until on average one droplet contains one macromolecule. In their final stage, the stream ions are submitted to the vacuum of the mass spectrometer.

Unlike ESI, where analytes are directly vaporised and ionised from a liquid phase, MALDI sublimates and ionises the analyte out of a dry crystalline matrix directly from solid into gas phase using high energy pulsed laser beams [Aebbersold and Mann, 2003]. The analyte co-crystallises with the matrix material, which strongly absorbs the laser. The irradiation of the matrix causes rapid heating leading to sublimation of the matrix crystals including the analyte. By transfer of energy the analyte can be submitted to the mass spectrometer intact.

1.4.2 Mass analyser

After the ions, commonly generated by soft ionisers, enter the vacuum of the mass spectrometer, the mass analyser separates them based on their m/z values. The most common mass analysers are time-of-flight, quadrupole, ion trap and Orbitrap. The instruments used for the studies covered in this thesis are Orbitrap mass spectrometers produced by Thermo Fisher Scientific.

The time-of-flight (TOF) mass analyser determines the m/z ratios of ions by measuring the time that ions take to cross a field-free flight tube with a certain length after being accelerated by an electric field of known strength. As equally charged ions have the same kinetic energy, the m/z value determines the velocity of an ion. Hence, ions with different m/z values travel with different speed and consequently detected at different time points [Lane, 2005]. TOF systems are frequently coupled to MALDI.

The quadrupole mass analyser consists of four metal rods equidistantly arranged around a central axis of which two pairs are equally charged [Lane, 2005]. A continuous ion beam enters the quadrupole along the central axis. By applying different radio frequencies (RF) to the rods only ions with a specific m/z ratio will travel through the electric and magnetic field at the centre of rods and are eventually detected. Ions with other m/z ratios, however, cannot pass through and will collide with the rods, and thus, are not detected.

In the ion trap mass analyser, ions enter and exit a ring-shaped electrode via separate cap electrodes located at the top and at the bottom and are trapped until the separation process within the sandwiched electrode [Ho et al., 2003]. Like the quadrupole mass analyser, the ion trap separates the ions by steadily increasing the RF. In contrast to the quadrupole, however, the ions with the corresponding m/z ratios undergo unstable oscillation and are released to be detected while the other ions remain trapped in form of a stable oscillation.

The Orbitrap (Figure 1.10) was invented by Makarov in 1999 [Makarov, 2000] and is the most recent and, in proteomics, most widely used mass analyser. It gained immediate popularity and impact due to its high resolution and mass accuracy. It consists of a barrel-like outer electrode and a spindle-like central electrode, in which the ions are trapped by an orbital motion around the central spindle and oscillate along the length of the electrode. The frequency of this oscillation is related to the m/z ratio of the ions and can be detected on the outer electrodes as so-called image current. This image current is finally converted into a mass spectrum based on Fourier-transformation.

1.4.3 Fragmentation methods

As mentioned previously, MS/MS analysis has an additional fragmentation step to enable sequence identification based on the analytes' fragmentation pattern [Sinitcyn et al., 2018a]. In the so-called data-dependent acquisition mode (DDA), the top n most intense MS1 peaks (precursors) are selected for fragmentation at a given retention time. Alternative acquisition modes are data-independent acquisition (DIA) and targeted acquisition. DIA is based on the selection of a constant mass range independent of the measured MS1 spectra. Targeted acquisition monitors the analysed peptides and only certain peptides based on a target list are selected for fragmentation.

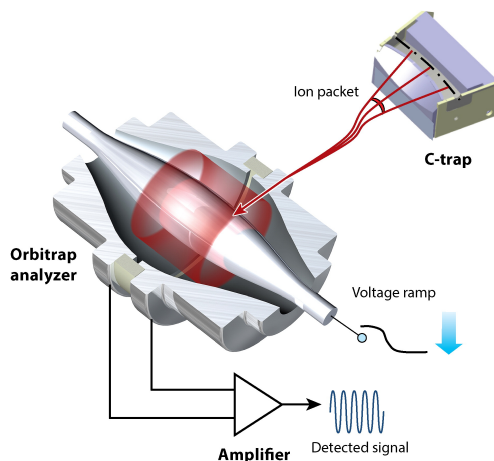


Figure 1.10: Illustration of an Orbitrap and its barrel-like outer electrode and a spindle-like central electrode. The ions are trapped by an orbital motion around the central spindle and oscillate along the length of the electrode [Eliuk and Makarov, 2015].

The most common fragmentation methods are collisional induced dissociation (CID) [Mitchell Wells and McLuckey, 2005, Shukla and Futrell, 2000], higher energy collision induced dissociation (HCD) [Michalski et al., 2012, Olsen et al., 2007] and electron transfer dissociation (ETD) [Coon et al., 2005, Good et al., 2007]. Depending on the applied method a different series of fragment ions is created (Figure 1.11).

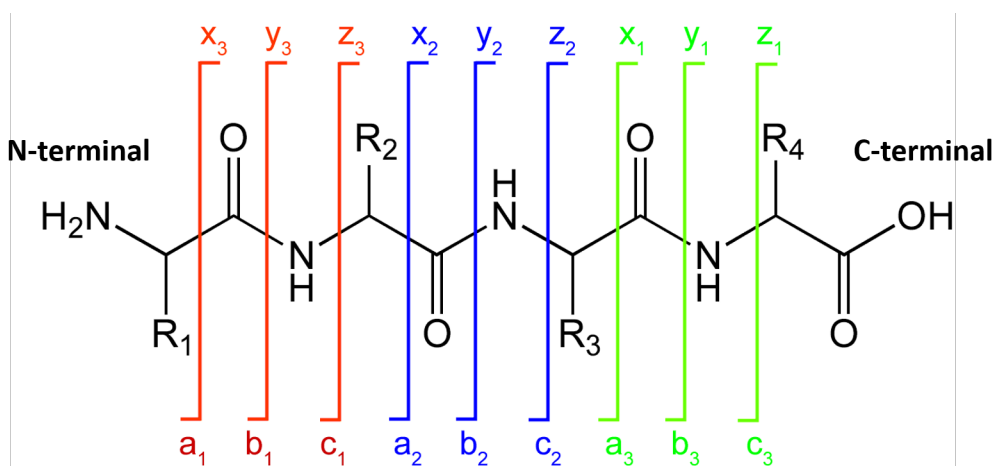


Figure 1.11: Resulting backbone fragment ion series in respect to the different fragmentation methods. CID and HCD fragmentation results in a, b- and y-ion series, while ETD fragmentation produces a, c- and z-ion series. Furthermore, the ion types can be grouped into N-terminal fragment ions (a, b and c) and C-terminal fragment ions (x, y and z). Adapted from (https://commons.wikimedia.org/wiki/File:Peptide_fragmentation.gif, Access date: 2021-08-28)

In CID, peptide fragmentation is induced by collisions of the ion with the atoms of a neutral gas such as helium, argon or nitrogen. The collision leads to a transformation of the ions' kinetic energy into internal vibrational energy that is randomly distributed over all atom bonds, causing breakage to occur on the weakest bond [Boersema et al., 2009, Han et al., 2008, Sobott et al., 2009]. Typically, the weakest bond is the peptide bond unless the peptide is modified by a labile PTM such as phosphorylation. In that case, the covalent bond of the phosphate group is the weakest leading to a so-called neutral loss. The obtained MS/MS spectra of phosphorylated peptides provide a poor fragmentation pattern and have a dominating peak that is indicative of a neutral loss [Boersema et al., 2009].

Both CID and HCD fragmentation provide MS/MS spectra based on N-terminal b- and C-terminal y-ions. However, HCD uses higher energy for the dissociation generating higher quality MS/MS spectra in terms of resolution and mass accuracy. Additionally, HCD produces spectra with highly abundant y-ions and might contain a-ions due to further fragmentation of b-ions [Michalski et al., 2012].

In proteomics, MS/MS analysis is typically carried out in positive ion mode resulting in positively charged peptide or precursor ions [Huang and McLuckey, 2010]. Fragmentation by ETD is fundamentally different than CID and HCD in that it triggers fragmentation by transferring the electrons from radical anions to multiply protonated peptides. Singly charged peptides would lose their charge, therefore at least doubly charged peptides are required. In contrast to CID and HCD, ETD is more suitable for longer and multiple charged peptides and has further the advantage that labile PTMs such as phosphorylation remain on the amino acid side chains of peptides [Penkert et al., 2019, Seidler et al., 2010].

The MS/MS analysis produces as output so-called MS/MS spectra, which contains a list of m/z values and the corresponding intensity values. An MS/MS spectrum represents the fragmentation pattern of a peptide, which is indicative of the amino acid sequence [Steen and Mann, 2004].

Therefore, high quality spectra are of great significance to achieve high identification rates. The quality of an MS/MS spectrum in turn depends on the MS instrument and its incorporated ion source and mass analyser composition. Hence, before conducting an MS experiment, it is important to consider the following instrument related quality parameters [Brenton and Godfrey, 2010]:

- **Mass resolving power** is the capability of separating two peaks of nearly identical m/z and is based on the mathematical method "Full Width at Half Maximum" (FWHM) that calculates the ratio of a peak's m/z value and the peak's width that is measured typically at 50% of the peak height [G. Marshall et al., 2013, Scigelova et al., 2011].
- **Mass accuracy** determines how good the experimental measurement is in comparison to the true value (i.e., theoretical mass calculated based on the known elemental formula, isotopic composition and charge state). Accuracy is the ratio of the m/z measurement error to the true m/z [Brenton and Godfrey, 2010, Scigelova et al., 2011]. The higher the mass accuracy the more likely the elemental composition of a measured ion can be determined.

- *Dynamic range* defines the difference between the most abundant and least abundant proteins [Steen and Mann, 2004].
- *Sensitivity* indicates how much signal intensity can be obtained of the smallest amount of a targeted analyte [Saah and Hoover, 1997].
- *Specificity* measures how well target analytes can be distinguished from other proteins that are present in the sample [Saah and Hoover, 1997].

1.5 Mass spectrometry-based proteomic strategies for peptide and protein identification

There are two complementary proteomic MS strategies that can be pursued for characterising proteins (Figure 1.12): top-down proteomics (TDP) and bottom-up proteomics (BUP). BUP coupled to HPLC is also referred to as “shotgun proteomics” [Zhang et al., 2013]. The objective of TDP is the study of intact proteins, whereas in BUP the proteins are digested by proteases to complex peptide mixtures [Kelleher, 2004]. The main advantage of TDP compared to BUP is the analysis of proteins in their complete form, thus, it can characterise the complete protein including accurate site localisation of complex PTM confirmations. However, TDP remains experimentally and computationally challenging due to the complexity of the acquired data [Brown et al., 2020], which makes BUP the commonly applied approach.

In BUP, the complex peptide mixture can be either analysed by peptide mass fingerprinting (PMF) or MS/MS. In PMF [Thiede et al., 2005], protein identification is based on the acquired mass pattern of its peptides typically after MALDI-MS analysis and does not require identification of the amino acid sequence. PMF is a simple approach, but of great importance due to its capability to distinguish for example collagen proteins of different species, which is essential for human history to gain insights into prehistoric animal husbandry by analysing used skin-based materials [Buckley et al., 2010].

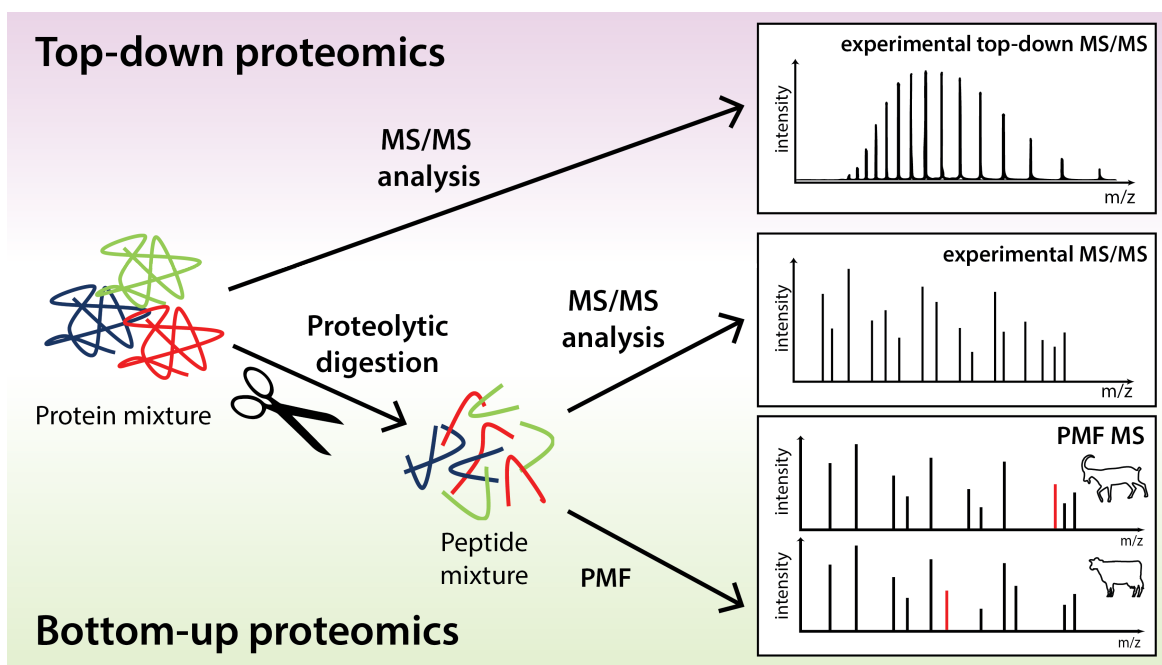


Figure 1.12: The two mass spectrometry-based proteomics analysis strategies: top-down proteomics (TDP) and bottom-up proteomics (BUP). In TDP intact proteins are submitted to MS analysis whereas in BUP the proteins are digested into a complex peptide mixture. BUP can be further divided into peptide mass fingerprinting (PMF) and MS/MS analysis. The MS/MS performs additionally sequence-specific peptide fragmentation.

BUP reduces spectrum complexity by digesting the proteins to peptides enabling simpler MS analysis. However, the resulting peptides lose their association to the protein after breakdown causing the challenge of protein inference in MS/MS analysis [Nesvizhskii and Aebersold, 2005]. Consequently, the protein identification is based on mapping back the identified peptide sequences to known protein sequences, impeded by the fact that peptide sequences can match to multiple protein sequences. Typically, HPLC is directly coupled to MS/MS (LC-MS/MS) which is commonly referred to as typical shotgun proteomics (Figure 1.13) [Zhang et al., 2013].

Shotgun proteomics has become the standard method for large-scale proteome analysis as it comes with a complete and well matured package. The package comprises sample preparation protocols, high-resolution MS instruments and a comprehensive selection of software tools for protein and peptide identification as well as post-processing methods such as the Perseus computational platform [Tyanova et al., 2016b].

For the aim of the studies as part of this thesis – the identification of ancient peptides and proteins – MS/MS coupled to HPLC (LC-MS/MS) was chosen because it enables sequence identification due to the additional peptide fragmentation step. Therefore, the analytes subjected to LC-MS/MS are complex peptide mixtures.

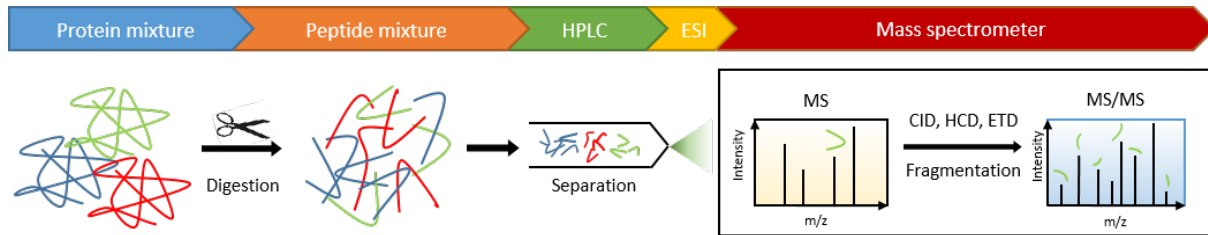


Figure 1.13: LC-MS/MS workflow or commonly referred to as “typical shotgun proteomics workflow”. First, the proteins are extracted from an organic material to be further enzymatically digested into a complex protein mixture. The protein mixture is separated by a reversed high-performance liquid chromatography (HPLC) before it is ionized by an electrospray ioniser (ESI) and subjected to tandem MS (MS/MS) analysis. The obtained peptide fragmentation pattern is used for sequence identification.

1.6 Peptide and protein sequence identification based on shotgun proteomics data

After LC-MS/MS analysis, the acquired MS/MS spectra can be computationally analysed for peptide and protein identification as well as quantification. As this thesis focuses on the identification of ancient protein and peptide sequences, computational methods for quantitative proteomics studies are not covered.

Over the years, several different identification approaches have been developed and can be categorised in spectral library search, sequence database search and de-novo sequencing 1.14. The identification efficiency of all approaches is highly dependent on the mass accuracy and resolution of the MS instrument. The better the measurement, the more informative is the MS/MS spectrum and consequently, leading to a higher rate of confidently identified peptide sequences.

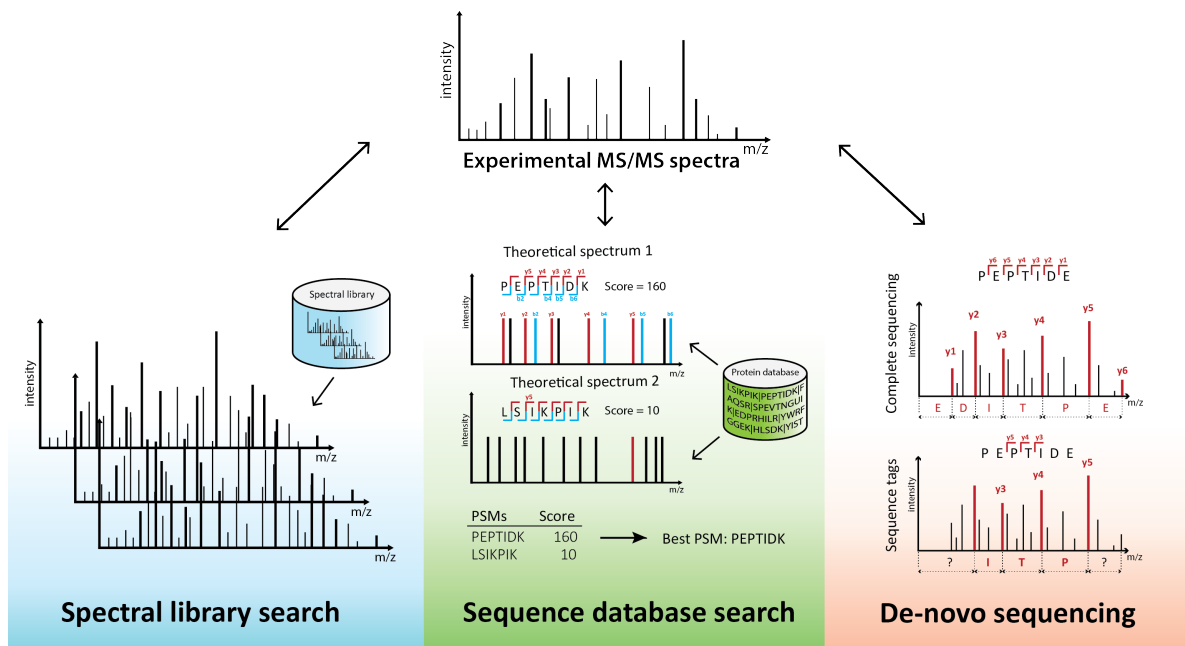


Figure 1.14: Illustration of the three different identification strategies: spectral library search, sequence database search and de-novo sequencing.

1.6.1 Spectral library search

In spectral library search [Craig et al., 2006, Yates et al., 1998, Lam et al., 2008, Stein and Scott, 1994], experimentally acquired MS/MS spectra are compared to a spectral library, which is a collection of already identified experimental MS/MS spectra of previous MS measurements. Only peptides that are contained in the library can be identified, which makes this approach suitable for targeted proteomic studies [Nesvizhskii, 2010], focusing only on a subset of proteins of interest in a sample [Marx, 2013]. Spectral library search approaches are not suitable for ancient protein studies, because of insufficient data availability to generate spectral libraries. More importantly, ancient proteins typically contain sequence

variations, which are not considered by spectral libraries, thus, the measured peptides would remain unidentified. Spectral library searching is not within the scope of this thesis, thus, not further elaborated.

1.6.2 Sequence database search

Sequence database search is the most frequently used method for peptide sequence identification, which is based on the comparison of experimentally acquired MS/MS spectra with theoretical MS/MS spectra. The search engine used for the studies that are part of this thesis is Andromeda, which is incorporated in MaxQuant, a software package for mass spectrometry-based shotgun proteomics [Cox and Mann, 2008, Tyanova et al., 2016a]. Theoretical spectra can be computer assisted (in-silico) generated by considering the following main database search parameters [Sinitcyn et al., 2018a, Tyanova et al., 2016a]:

- a *reference protein sequence database* defines all target proteins of an organism including possible contaminants
- an *enzyme and its cleavage specificity* to determine cleavage sites in a protein sequence to generate a list of peptide sequences
- a *fragmentation method* to in-silico calculate the m/z values of the peptides fragmentation pattern
- a *list of fixed or variable PTMs* to account for the molecular weight of PTMs that introduces a mass shift in the peptide's fragmentation pattern

Reference protein sequence databases are FASTA files, text-based files that comprise reference sequences of which each sequence is defined as single line header followed by the lines of sequence information. FASTA files can be downloaded from online platforms such as UniProt [Bateman et al., 2017] or Ensembl [Zerbino et al., 2018]. The reference protein sequences should be selected with respect to the sample analysed. For example, if a human-derived sample was analysed, the reference database should contain all known human proteins.

Next, the protein sequences provided by a FASTA file are split into a list of peptides based on a certain *enzyme and its cleavage specificity* that must be in accordance with the enzyme used during sample preparation. In shotgun proteomic experiments, trypsin is the most commonly used protease and cleaves proteins on the N-terminal side of arginine (R) and lysine (K) residues [Steen and Mann, 2004]. Subsequently, for each peptide a theoretical spectrum is calculated. To calculate the m/z values of each fragment ion to simulate the peptides fragmentation pattern, the type of ion-series is required which is defined by the applied *fragmentation method* in the MS analysis (Section 1.4.3 Fragmentation methods).

If peptides are expected to be modified, the PTMs need to be defined as a search parameter as a *list of fixed or variable PTMs*. Fixed modifications are always present on certain amino acids, thus, inducing a fixed mass shift, whereas variable modifications may or may not be present. As a result, for fixed modifications only one theoretical spectrum needs to be generated, while for variable modifications multiple theoretical spectra are generated to provide the fragmentation pattern of the unmodified peptide and its modified counterpart.

The experimental MS/MS spectra can be matched to a subset of theoretical spectra, selected by the peptide mass. Such a match is called peptide spectrum match (PSM). The search space of a search engine is highly dependent on parameters, such as the number of variable PTMs or the size of the reference database. The larger the database or the higher the number of variable PTMs the more peptide candidates are generated leading to an increased search space [Nesvizhskii, 2010]. Consequently, more false identifications are introduced, because the probability of obtaining a high scoring PSM by chance increases. When looking at all theoretically possible amino acid sequence combinations, only a small portion of sequences is derived from an organism, thus, short peptide sequence can be considered to be already highly protein-specific [Seidler et al., 2010]. Consequently, peptide candidates with a minimum length of typically seven amino acids are typically selected.

The quality of a PSM is scored via specific scoring functions (see details later in the section) and the peptide sequence of the theoretical spectrum with the best score is the suggested sequence for the experimental MS/MS spectrum. Each of the available database search engines SEQUEST [Washburn, 2015], Mascot [Perkins et al., 1999], MS-GF+ [Kim et al., 2010] or MS Amanda [Dorfer et al., 2014] has its own scoring function that optimises peptide identification by reflecting the similarity between experimental and theoretical spectra.

While database search engines only allow mass shifts introduced by defined PTMs, open search [Chick et al., 2015] and dependent peptide [Savitski et al., 2006] approaches allow unrestricted searches to look for all possible PTMs or chemical modifications [Nesvizhskii, 2010, Sinitcyn et al., 2018a].

1.6.2.1 Andromeda

The search engine Andromeda [Cox et al., 2011] employs a binomial probability-based scoring model to calculate the score of a PSM as shown in the following Equation 1.1:

$$s(q, loss) = -10 \log_{10} \sum_{j=k}^n \left[\binom{n}{j} \left(\frac{q}{100} \right)^j \left(1 - \frac{q}{100} \right)^{n-j} \right] \quad (1.1)$$

where k is the number of matching ions between the total number of n fragment masses of a theoretical spectrum, and the number of selected peaks q of the experimental MS/MS spectrum within a mass interval of 100 Th representing the average mass of amino acids. First, the probability of getting at least k matches out of n theoretical masses by chance is calculated and the higher k is compared to n , the lower the probability that the matches are happening randomly. Next, the negative decadic logarithm of the resulting probability is calculated to obtain the Andromeda score. The final Andromeda score is optimized over a number of different highest intensity peaks taken from the experimental spectrum and the incorporation of modification-specific neutral losses such as water or ammonia losses.

For nearly every experimental spectrum, a PSM can be assigned and many of them have a poor matching quality indicated by low Andromeda scores. Low scoring PSMs are very likely the result of incorrectly assigned peptide sequences due to coincidentally similar

fragmentation patterns and as such are false positive (FP) peptide identifications. However, to distinguish FP and true positive (TP) PSMs with a medium score requires more advanced separation measures. The following section introduces the concept of false discovery rate, which assesses the confidence of peptide and protein identifications globally and is the most accepted validation method for proteomics data.

1.6.2.2 False discovery rate

The false discovery rate (FDR) is an estimation of the error rate of the entire dataset [Elias and Gygi, 2007, Elias and Gygi, 2010, Käll et al., 2008]. A target-decoy search strategy is most commonly used to control and estimate the rate of falsely identified experimental spectra by searching not only against a reference protein database, functioning as a target database, but also against a decoy database. The purpose of the decoy database is to introduce intended FP PSMs that might be otherwise considered correct. While the target database contains real protein sequences, derived from the studied organism, the decoy database contains nonsense protein sequences that do not occur in nature. Decoy sequences are typically generated by reversing the target protein sequences, thus, the number of target and decoy protein sequences are equal. This form of transformation is a simple way of ensuring similarity between target and decoy sequences in terms of amino acid frequencies, protein and peptide length distributions and mass distributions of theoretical peptides.

The FDR is estimated by dividing the number of FP identifications by the total number of PSMs (Figure 1.15) and typically a score threshold is applied to control for a certain FDR, which is usually 1%. An FDR of 1% means that amongst a list of PSMs, 99% of the matches are correctly identified while 1% are not. By increasing the FDR also the number of PSMs will increase, but consequently also the number of FPs [Käll et al., 2007b].

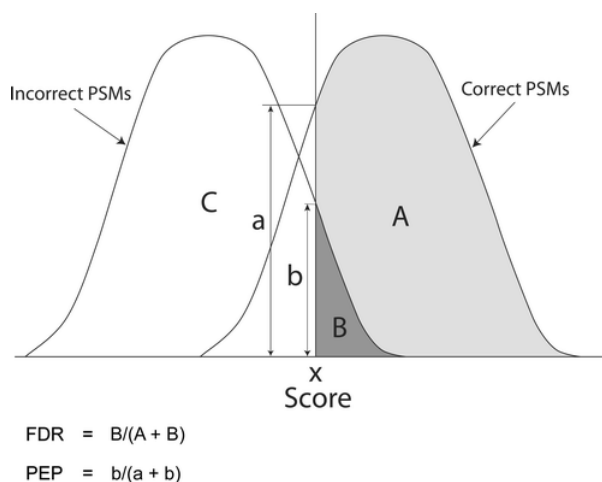


Figure 1.15: Calculation of the false discovery rate (FDR) and posterior error probability (PEP) based on the mixed score distribution of known incorrect and correct peptide spectrum matches (PSMs) [Käll et al., 2007b].

In most cases, a threshold is not defined based on the search engine score, but rather based on the posterior error probability (PEP) or q -value. The PEP and q -value are statistical scores that are used to validate the reliability of a PSM identification.

The PEP is the probability that an individually observed PSM is incorrect. For example, a PEP value of 10% for a certain PSM means that there is a 90% chance that the identified sequence was included in the sample that was measured by the MS analysis [Käll et al., 2007b]. By comparing a mixed score distribution generated by the PSMs that match either to target or decoy sequences, the PEP can be calculated by dividing the number of incorrect PSMs at a certain score value by the number of total PSMs having the same score value (Figure 1.15). For a better separation of true and false positive identifications, other peptide features, such as the peptide length and number of missed cleavages, can be considered besides the score value. The PEP approach implemented in MaxQuant is not only conditional on the score, but also on the aforementioned peptide features including the charge state and the number of PTMs [Tyanova et al., 2016a].

The q -value is defined as the minimal false discovery rate at which the identification is considered to be correct [Käll et al., 2008]. To demonstrate an example: a q -value of 0.01 for a PSM means that 1% is the minimal FDR threshold at which the PSM is amongst a list of identified PSMs. To calculate the q -value of a certain PSM, first all PSMs need to be sorted by their scores, or more commonly by their PEP values, and the top scoring PSMs are listed first. Next, a subset of PSMs is defined by selecting PSMs that have a higher score than the PSM of interest. Based on this subset the q -value is calculated in the same way as the FDR by dividing the number of FPs by the total number of PSMs.

Importantly, the q -value depends on the data set in which the PSM occurs. This means that the q -value for a certain PSM can change when the composition of the subset assigned to the selected PSM changes. A different subset can be the result of changing the search space (e.g., changing the reference protein sequence or the set of variable PTMs) that is used by the database search. Thus, the peptide candidates of a certain experimental spectrum might differ which can subsequently lead to different sequence identification that can either explain the experimental spectrum better or not.

Instead of calculating the PEP values for each PSM to be able to calculate the q -values - as it is implemented in MaxQuant - other tools, such as PeptideProphet [Keller et al., 2002, Choi and Nesvizhskii, 2008] and Percolator [Käll et al., 2007a] use machine learning for the same aim. For example, Percolator uses a semi-supervised machine learning model by training support vector machine models to discriminate between correct and decoy spectrum identifications. Percolator provides as output directly a q -value for each PSM, which is subsequently used to define the FDR threshold. Based on the set of FDR-controlled PSMs, proteins can be identified by protein inference and are typically also FDR controlled. MaxQuant provides the identified proteins in the form of protein groups, since peptide sequences can be shared by different proteins [Tyanova et al., 2016a].

Database search engine analysis controlled for FDR on different levels, such as the PSM and protein level, is currently the most commonly used and reliable approach in the proteomics research. However, besides the quality of the MS measurements, the identification success is strongly dependent on the content of the reference database, restricting the possibility of discovering new peptides. If proteins are not part of the reference database, but are present in the sample there is no chance that the according experimental MS/MS spectrum can be

correctly identified. Computational approaches based on de-novo sequencing tackle this challenge by database independent identification.

1.6.3 De-novo sequencing

By applying de-novo sequencing approaches, the measured peptides in the form of acquired MS/MS spectra of a typical shotgun proteomics experiment are identified uniquely based on the MS/MS spectrum and its fragmentation pattern without the need of any reference database. This independent way of identification offers the great possibility to explore novel peptide sequences such as protein sequence variants as a result of alternative splicing. Furthermore, de-novo sequencing allows protein analysis of organisms with unsequenced genomes. Important other applications include monoclonal antibody sequencing [Tran et al., 2016] and comprehensive sequencing of the HLA peptides [Khodadoust et al., 2017, Ternette et al., 2016].

The peptide sequence can be determined directly by converting the mass difference of consecutive ions in a spectrum to amino acids, which ideally are all from the same ion-series. Complementary ion-pairs are beneficial, since they not only provide additional sequence information, but also give rise to the precursor ion mass based on the sum of their masses [Seidler et al., 2010].

The brute force approach was one of the first approaches used to obtain the peptide sequence based on the MS/MS spectrum content. Based on the precursor mass, all amino acid combinations are generated. For each combination a score is calculated and the highest scoring combination is accepted as the identified sequence [Allmer, 2011]. However, de-novo sequencing based on brute force has limitations, because the number of sequence combinations increases exponentially with increasing precursor masses. This exponential growth leads to a combinatorial explosion, which is heavily restricted by available computing power. Therefore, alternative de-novo sequencing approaches have been investigated, which are based on different methods such as graph theory, dynamic programming, hidden Markov model and machine learning. Amongst others, recent de-novo algorithms are Novor [Ma, 2015], DeepNovo [Tran et al., 2017] and PointNovo [Qiao et al., 2021], which are all based on machine learning approaches.

The success of peptide identification by de-novo sequencing is strongly dependent on the quality of the acquired MS/MS spectrum. MS/MS spectra of peptides with a high fragmentation efficiency, i.e., a high coverage of the fragment ion series with high intensities, are more likely to be correctly sequenced than low abundant peptides providing poor fragmentation efficiency. In addition, poor fragmentation efficiency can cause incomplete ion-series. Complete sequencing is also hindered, when the sequence contains non-standard amino acids (e.g., amino acids that are not encoded in the genetic code such as selenocysteine or pyrrolysine) or uncommon PTMs, because they are not considered in the algorithm [Seidler et al., 2010]. Additionally, the higher the accuracy, the better the information about the elemental composition, which is specifically important to be able to distinguish amino acids with similar masses (e.g., K and Q).

However, complete peptide sequencing in a reference database independent manner is rather the exception. MS/MS spectra are mostly sequenced incompletely and the identified sequence is referred to as 'sequence tags'. Such sequence tags are used in homology-based approaches, which are a combination of de-novo sequencing and database search. This combined identification approach can identify peptides that do not exactly match to a reference sequence database. Implemented algorithms are GutenTag [Tabb et al., 2003], MS-Homology [Kayser et al., 2004] and SPIDER [Han et al., 2005].

1.7 Applications of ancient proteins

The study of ancient proteomes by applying MS-based approaches is termed palaeoproteomics. In 2000, the continuous analytical and computational advancements of MS enabled the first analysis of ancient proteins, when osteocalcin (OC) of an ancient bone was successfully identified based on PMF using MALDI-TOF MS [Ostrom et al., 2000]. Since then, the emerge of palaeoproteomics has led to an increase in ancient protein studies in a diverse range of fields and application: cultural heritage objects, archaeological studies and evolutionary studies [Dallongeville et al., 2016, Hendy, 2021, Welker, 2018b].

Cultural heritage objects are key to understanding the culture of each other's society to foster relations and communication. Many of these objects are based on organic compounds (e.g., paint binders, manuscripts, leather objects) and hence, are rich in proteins. Therefore, palaeoproteomics lends itself as a valuable tool to study these objects.

In art history, paintings have been the subject of several studies [Dallongeville et al., 2016], with the aim to identify the protein composition of paints used in different layers. The protein composition provides valuable information for art historians [Zadrozna et al., 2003], because they gain insights into the artistic technique or history of the object as well as insights into society's life style. Especially important for museum conservators, the identification of used materials and their biological source also sheds light on its degradation process which is essential for the selection of suitable preservation and conservation techniques.

Paint typically consists of a colour pigment and an organic binder [Tokarski et al., 2006]. In paintings based on tempera, the binding material is usually a glutinous material such as egg yolk whereas in oil paintings oil-based binders are applied. In the early 2000s, the proteins of the binder in colour layers of a painting from the 19th century were successfully identified by PMF and the results suggest the use of rabbit glue as protein binder [Hynekl et al., 2004]. With improving methodologies and the application of MS/MS, it became possible to identify egg white and egg yolk proteins [Tokarski et al., 2006]. One of the studied old art paintings was the Benedetto Bonfigli's triptych, *The Virgin and Child, St. John the Baptist, St. Sebastian* (XVth century) that is shown in Figure 1.16, A.

The improved methodology for ancient protein studies and its success opened up new applications of MS-based ancient protein analysis to archaeology [Hendy, 2021]. The first species identification of faunal remains based on Zooarchaeology by Mass Spectrometry (ZooMS) was published in 2010 [Buckley et al., 2010]. ZooMS aims to provide a cheap and fast method to distinguish species (e.g., sheep and goat) based on collagen peptide sequencing typically analysed with PMF using MALDI-MS, which is particularly useful to enable identification of species that are morphologically very similar [Zeder and Pilaar, 2010, Zeder and Lapham, 2010] or where remains are highly fragmented. The species can be distinguished based on the generated mass fingerprints of the analysed collagen peptide sequence by comparing it to reference sequences of known species. A limitation of ZooMS can be insufficient sequence variability based on evolutionary distance, thus, it is typically possible to distinguish between family, but discrimination becomes challenging on the species-level [Hendy, 2021]. ZooMS has been extensively applied for species

identification of faunal assemblages, which provides insights into past subsistence practices and ecologies. For example, MS/MS analysis by MALDI-TOF/TOF have advanced the understanding of domestication by the identification of well-preserved food remains and its ingredients [Hong et al., 2012]. Hong et al. studied the ancient proteins from food residues taken from a pottery bowl (Figure 1.16, B, C) at cemetery sites located in the Turpan Basin, China, and successfully identified milk additives such as bovine milk or curd. Species identification of food remains allows to draw conclusions about animal domestication, providing insights into ancient farming practices.

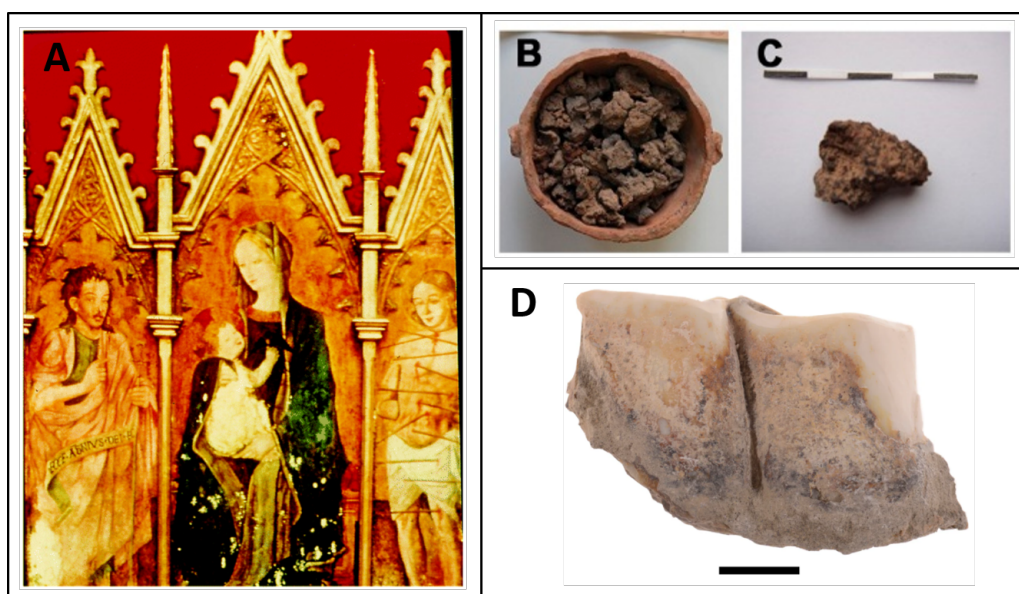


Figure 1.16: A) Benedetto Bonfigli's triptych, The Virgin and Child. St. John the Baptist. St. Sebastian (Italian School, XVth century), Petit Palais museum, Avignon, France. ©CRRMF. [Tokarski et al., 2006] B,C) Ancient proteins were extracted from remaining black food residues taken from a pottery bowl [Hong et al., 2012]. D) Dental enamel and dentine proteins were extracted from the molar assigned to a *Stephanorhinus* [Cappellini et al., 2019].

Ancient protein studies have also shed light on the domestication of cereals in early farming and dietary habits. For example, Colonese et al. identified wheat proteins when studying the remaining proteins of well-preserved Early Bronze Age wooden containers from the Swiss Alps [Colonese et al., 2017]. Another analysis with the aim to gain insights into the dietary habits of humankind was performed on the stomach content of the 5,300-year-old glacier mummy, usually also referred to as "Ötzi the Iceman" [Maixner et al., 2018]. By combining aDNA and ancient proteomics analysis approaches, it was possible to identify traces of meat and cereals in Ötzi's last meal.

Ancient proteins are not only very valuable to extend our knowledge in the field of cultural heritage and archaeology, but also to better understand human evolution [Welker, 2018b]. Since the first time Neanderthal mtDNA was successfully sequenced [Krings et al., 1997] in 1997 and with the introduction and advancements of high-throughput sequencing methods, large-scale research on extinct hominins has become possible [Marciniak and Perry, 2017]. Complete genome sequencing of different Neanderthal

specimen [Castellano et al., 2014, Green et al., 2010, Prüfer et al., 2014] provide important sequence variation information to further understand hominin population structure, size, and movements [Kuhlwilm et al., 2016, Rogers et al., 2017]. Additionally, “Denisovans” have been discovered as a sister-clade to Neanderthals [Krause et al., 2010, Meyer et al., 2012].

However, due to the biomolecular degradation, no ancient hominin DNA older than 0.43 million years has been recovered [Meyer et al., 2016]. In 2015, evolutionary relationships of extinct species were successfully elucidated by obtaining amino acid sequences of ancient bone proteins (mainly of collagen type 1) with shotgun proteomics [Buckley et al., 2015, Cleland et al., 2016, Welker et al., 2015].

Besides the bone tissue, other tissue such as dental enamel and dentine have been explored. A recent study has shown, that it is possible to address the taxonomic placement of the Eurasian Pleistocene Rhinocerotidae by applying shotgun proteomics to a ~1.77 million years old dental enamel and dentine of a *Stephanorhinus* specimen (Figure 1.16, D) excavated from the Dmanisi archaeological site in Georgia [Cappellini et al., 2019].

1.8 Challenges of ancient protein sequence identification

The success of recent ancient protein studies and the application to a variety of organic materials covering different time periods have provided valuable insights advancing our understanding about the past. However, there are important challenges in every step of ancient proteomics experiments that need to be considered.

Proteomics experiments require destructive sampling to extract the proteins from biological materials. Analysis of modern samples such as cell lines can be mostly repeated. For example, cell lines such as the HeLa cells [Masters, 2002], the first continuous cancer cell line - isolated from an aggressive glandular cervical cancer of a woman called Henrietta Lacks, can be reproduced. However, objects from cultural heritage, archaeology and palaeontology are precious, unique and most importantly limited. Therefore, before conducting a proteomics experiment that requires destructive sampling, it is important to ensure the success of the analysis but also that the knowledge gain is worth the permanent loss of the material. The success of obtaining proteins from the material under study is not only dependent on the age, but also on the preservation condition.

Organic compounds are preserved better in environments that provide cool, dry, dark, anaerobic and slightly alkaline conditions (pH of 7.8) [Bollongino et al., 2008]. Even though proteins are more resistant to degradation than DNA, diagenesis still affects protein sequences, so that ancient materials tend to contain shorter and altered peptide fragments [Bada et al., 1973, Barker, 1981, Schroeder and Bada, 1976]. Based on the aforementioned preservation conditions, burial deposition has the advantage of providing a good preservation environment. However, the surrounding microbial can cause high contamination. Thus, the obtained sample for a proteomics analysis usually contains on the one hand low abundant proteins in low concentrations and on the other hand highly abundant contaminant proteins.

In addition, contamination can also be induced during excavation, storage and analysis [Hendy et al., 2018], for example by human tissue or clothes made of proteinaceous material during sample preparation. Contamination of modern human tissue is specifically critical when working with ancient material that is suspected to be of human origin. To not mask the protein sequence obtained from an object of interest by that of contaminants, it is most important to take highest precautions during sampling, measuring as well as in the computational analysis, to be able to distinguish endogenous proteins from contaminants.

In comparison to the modern counterpart, ancient proteins have typically an increased number of PTMs, which can either be biological, such as glycosylation [Ozcan et al., 2014], phosphorylation [Cappellini et al., 2019] and hydroxylation [Ehrlich et al., 2010] but also introduced by diagenesis, ageing or other environmental factors such exposure to light and oxygen. PTMs related to ageing and diagenesis are oxidation [Mackie et al., 2018], glycation [Cleland et al., 2015] and deamidation [Schroeter and Cleland, 2016]. Particularly important is the modification deamidation located on the amino acids asparagine (Asn) and glutamine (Gln). Deamidation can occur enzymatically in living organisms, but also non-enzymatically over time and is therefore often considered as an indicator for protein damage [Robinson and Robinson, 2001]. Consequently, based on deamidation, endogenous proteins obtained from the ancient material can be distinguished from contaminants by comparing their deamidation rates, which is typically higher in ancient samples [Cappellini et al., 2019, Ramsøe et al., 2020].

Despite recent developments and improvements of high-resolution mass spectrometers to detect proteins in complex mixtures, identification of ancient proteins is still challenging due to the low abundance and concentration of ancient proteins in a sample compared to the high abundance of accompanying contaminants. Database search engines (Section 1.6.2 Sequence database search) are the most frequently used identification approach to determine the sequence from shotgun proteomics experiments. The resulting MS/MS spectra of the measured ancient peptide mixture usually provide a poor fragmentation pattern with low intensities. The identification of such spectra usually leads to low scoring PSMs due to insufficient matching ions. Additionally, a set of PTMs needs to be included in the search, which leads to an enormous increase in search space and, as a result, the increase of FP identifications. The overall number of FP is controlled by the FDR (Section 1.6.2.2 False discovery rate). Nevertheless, this leads to a decrease of identified sequences. Moreover, protein sequences of unsequenced organisms or extinct species might not be identified by current database search algorithms, as they may be evolutionarily distinct, even from close extant relatives [Welker, 2018b].

As mentioned previously, an alternative identification approach is de-novo sequencing to be able to identify un-sequenced organism (Section 1.6.3 De-novo sequencing). However, de-novo sequencing even on modern sample provides mainly incomplete tag sequences, due to insufficient obtained fragmentation patterns. The application of hybrid approaches, which are a combination de-novo sequencing and database search, can increase the identification rate by using the tag as lookup to find homologous sequences in the database.

For taxonomic placement, it is important to provide reliable sequence identifications. However, current identification approaches are optimised to identify protein sequences of modern samples. Therefore, they do not cover all aspects that are important for the identification of ancient protein. For example, deamidation converts asparagine (Asn) and glutamine (Gln) to aspartic acid (Asp) and glutamic acid (Glu), respectively, and due to their similar masses, no distinction can be made.

Ancient proteomics is an emerging field that provides valuable insights into the past of humans, animals and plants in several fields, such as cultural heritage, archaeology and evolutionary history. Despite recent successes, it is still challenging to achieve high identification rates, due to advanced diagenesis of ancient proteins. While it is possible to achieve an identification rate of 40% of all acquired MS/MS spectra in modern datasets, such as the cell line HeLa [Masters, 2002] by applying high-resolution shotgun proteomics, only 10% or less PSMs can be identified in ancient datasets. Therefore, many acquired MS/MS spectra remain unidentified stressing the need of improved identification algorithms.

Chapter 2

Purpose and structure of the thesis

To further advance palaeoproteomic methods, the work described in this thesis aims to improve the performance of the database search engine Andromeda, incorporated in the MaxQuant software package for shotgun proteomics data analysis, which is the most frequently used tool in the proteomics community. More specifically, the goal was the development of a novel scoring method that improves the identification by including predicted fragmentation ion intensities of theoretical spectra (Chapter 3 – Publication 1).

Identification by database search engines is the standard approach for protein and peptide identification, facilitated by the availability of large reference protein sequence databases of already sequenced organisms. However, protein sequences of unsequenced or extinct species typically contain sequence variation due to their evolutionary distance to modern species. These variations prohibit exact matching to the sequences of a reference database, which causes many measured peptides to remain unidentified. To tackle this challenge, I was working on MaxNovo, a novel de-novo sequencing algorithm implemented in MaxQuant to provide reliable complete and in-complete sequence identifications independent of any reference knowledge (Chapter 4 – Publication 2).

The two following chapters are ancient studies related to evolutionary history. In the first study (Chapter 5 – Publication 3), the phylogenetic relationships of *Homo antecessor* from the excavation site Atapuerca in Spain and of *Homo erectus* from the excavation site Dmanisi in Georgia were investigated. In the second study (Chapter 6 – Publication 4), we applied an ancient proteomics experiment to analyse the dental enamel proteome of *Gigantopithecus blacki*, a giant hominid from Southeast Asia dated to be ~1.9 million years old.

List of publications that are part of this thesis:

1. Tiwary, S., Levy, R., Gutenbrunner, P. et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nat Methods* 16, 519–525 (2019). <https://doi.org/10.1038/s41592-019-0427-6>
2. Gutenbrunner, P. et al. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra (submitted to bioRxiv, MCP)

3. Welker, F., Ramos-Madrigo, J., Gutenbrunner, P. et al. The dental proteome of *Homo antecessor*. *Nature* 580, 235–238 (2020). <https://doi.org/10.1038/s41586-020-2153-8>
4. Welker, F., Ramos-Madrigo, J., Kuhlwilm, M. et al. Enamel proteome shows that *Gigantopithecus* was an early diverging pongine. *Nature* 576, 262–265 (2019). <https://doi.org/10.1038/s41586-019-1728-8>

A description of my contribution to each publication can be found in a dedicated contribution chapter, which is placed after the discussion.

Chapter 3

High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis

Shivani Tiwary^{1,5}, Roie Levy^{2,5}, Petra Gutenbrunner^{1,5}, Favio Salinas Soto¹, Krishnan K. Palaniappan², Laura Deming³, Marc Berndl³, Arthur Brant², Peter Cimermancic^{2*} & Jürgen Cox^{1,4*}

Nature Methods 16, 519–525 (2019). <https://doi.org/10.1038/s41592-019-0427-6>
Published: 27 May 2019

¹Computational Systems Biochemistry Research Group, Max Planck Institute of Biochemistry, Martinsried, Germany.

²Verily Life Sciences, South San Francisco, CA, USA.

³Google LLC, Mountain View, CA, USA.

⁴Department of Biological and Medical Psychology, University of Bergen, Bergen, Norway.

⁵These authors contributed equally

*email: cpeter@verily.com; cox@biochem.mpg.de

3.1 Abstract

Peptide fragmentation spectra are routinely predicted in the interpretation of mass-spectrometry-based proteomics data. However, the generation of fragment ions has not been understood well enough for scientists to estimate fragment ion intensities accurately. Here, we demonstrate that machine learning can predict peptide fragmentation patterns in mass spectrometers with accuracy within the uncertainty of measurement. Moreover, analysis of our models reveals that peptide fragmentation depends on long-range interactions within a peptide sequence. We illustrate the utility of our models by applying them to the analysis of both data-dependent and data-independent acquisition datasets. In the former case, we observe a q -value-dependent increase in the total number of peptide identifications. In the latter case, we confirm that the use of predicted tandem mass spectrometry spectra is nearly equivalent to the use of spectra from experimental libraries.

3.2 Introduction

Peptide identification by fragmentation is a fundamental part of bottom-up mass-spectrometry-based proteomics [Cottrell, 2011, Sinitcyn et al., 2018a]. Peptide molecules are fragmented with the aid of one of several technologies, including collision-induced dissociation [Mitchell Wells and McLuckey, 2005] (CID), higherenergy collisional dissociation [Olsen et al., 2007] (HCD) and electron transfer dissociation [Coon et al., 2005, Good et al., 2007], producing a pattern of fragments that is indicative of the amino acid sequence [Steen and Mann, 2004]. The frequency with which a peptide backbone bond breaks determines the relative signal intensities in a fragmentation spectrum. Theoretically, the intensities can be calculated from first principles by quantum chemistry. However, for molecules as large as peptides, this is too computationally expensive to be practical. Simpler models, such as the mobile proton hypothesis [Boyd and Somogyi, 2010], exist for qualitative considerations, but they are not precise enough to be beneficial to the peptide identification process. Hence, the intensity information contained in fragmentation spectra remains underused in many peptide identification strategies.

This problem is an ideal situation in which to employ machine learning. It can learn the relationship between sequence and fragment abundances based on a large dataset of training examples, without explicit knowledge of the physical mechanisms behind it. Furthermore, the predictive models do not have to remain black boxes, but can be examined with specialized methods that identify features or combinations thereof that are most relevant for making a prediction. While fragment intensity prediction has been attempted before using a variety of methods [Arnold et al., 2006, Degroev et al., 2013, Dong et al., 2014, Park et al., 2017], they have had limited success. Here, we present a deep learning [LeCun et al., 2015] method whose accuracy is close to the theoretical limitation. Furthermore, we demonstrate its utility by integrating it into data-dependent acquisition [Wolters et al., 2001] (DDA) and data-independent acquisition [Doerr, 2014] (DIA) computational proteomics workflows, and our results suggest that both can benefit from the improved spectrum prediction.

We developed two different regression strategies to model peak intensities. The first one, termed DeepMass:Prism, is a deep learning approach using a bidirectional recurrent neural network (RNN). Its predictive performance reaches the theoretical limit set by the reproducibility of technical replicates. The second approach, termed wiNNer (window-based neural network being easily retrainable), follows a classical sliding sequence window-based machine learning strategy. The latter model is less accurate than DeepMass:Prism. However, it has the advantage of being less computationally expensive to train, which makes ad hoc model creation for a given dataset more feasible. For both strategies, a single model can accommodate peptides of any length, unlike in other approaches [Degroev et al., 2013]. In the deep learning approach, a single model can also accommodate multiple fragmentation types, or other dataset-specific information such as the fragmentation energy

3.3 Methods

Data. For evaluating the DeepMass:Prism model, we obtained 25 raw datasets from the PRIDE MS repository [Vizcaíno et al., 2016]. The data span five organisms (*Homo sapiens*, *Mus musculus*, *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*), and contain 4,624 liquid chromatography–mass spectrometry runs (Supplementary Table 3.1). The datasets were processed using MaxQuant v.1.5.8.7. The datasets comprise ~60 million MS/MS spectra in total (Supplementary Fig. 3.7), with 1.4 million unique sequence/charge state/fragmentation method/mass analyzer combinations. For each unique combination, a single representative MS/MS spectrum was used corresponding to the one with the best Andromeda score. Spectra with Andromeda scores below 100 were discarded. The unique combinations were randomly split into training, validation and testing sets with the ratio 90:5:5. The intensity values were normalized to a 0–10,000 range and were not log-transformed.

For evaluating the wiNNeR model, we used the ProteomeTools dataset [Zolg et al., 2017] (PXD004732). Different models were generated for CID+2, CID+3, HCD+2 and HCD+3, and for each one peptide sequence features were split into training, validation and testing sets with the ratio 90:5:5. All models contain unique peptides, which were selected taking the maximum Andromeda score; if this score was under 100, the peptide was discarded. Intensity values were normalized to 0–1 and then log-transformed using $\log_2(1 + \text{Intensity} \times 10,000)$.

Details of the RNN. Our model takes as an input a peptide amino acid sequence with its associated metadata, and returns intensities of different fragment ion types (that is, y and b ions with and without neutral losses) at each position along the input sequence (Fig. 3.1). The architecture of our neural network comprises two main modules: layers of recurrent cells (encoder) and layers of fully connected neurons (decoder). The encoder contains three bidirectional layers of long short-term memory [Hochreiter and Schmidhuber, 1997] (LSTM) cells and emits a fixed-length representation of the input peptide. The fixed-length representation is concatenated to corresponding metadata (that is, precursor peptide length, charge state, fragmentation method and mass analyzer type) and then input into the decoder. Finally, the decoder outputs intensities of different fragment ion types at each position of the input sequence. Bidirectional LSTM cells and regular perceptron units with the rectifier activation function [Hahnloser et al., 2000] were used to build the RNN and fully connected modules, respectively. The neural network was implemented in Tensorflow v.1.7.0 [Abadi et al., 2016]. The learning and dropout rates, the number of layers, the number of hidden units in each module and the batch size hyper-parameters were optimized using Google Vizier [Golovin et al., 2017] and the validation dataset. The model was trained on GPUs using the Adam optimization method [Kingma and Ba, 2015]. The best model contained 3 bidirectional LSTM layers with 384 hidden units in each layer, and 4 fully connected layers with 768 neurons in each layer. Examples for the best and worst five predictions of the model can be found in Supplementary Fig. 3.16. A tab-separated text file with all spectrum predictions for the tryptic peptides in the human proteome (charge=2, HCD) can be downloaded from the PRIDE dataset PXD010382 (uniprot-filtered-reviewed-human-peptides-ftms-hcd-charge2.tsv).

Details of the sliding-window-based machine learning. For the conventional machine learning approach on sequence windows, we use as feature space the adjacent amino acids around the backbone bond for which the y- and b-ion intensities should be predicted. In addition, the amino acids directly on the peptide N and C termini, as well as the distance of the bond to the termini and the length of the peptides, are used. Each amino acid feature was converted to 21 binary features by one-hot encoding. The 21st state represents cases in which the sliding window extends over a terminus of the peptide. The intensities of a single peptide were normalized by the maximum over y and b ions, and intensities for missing peaks were set to zero. For each fragment series, each backbone segment between two adjacent amino acids corresponds to an instance for the machine learning algorithm. Distances and peptide lengths were normalized so that they ranged from zero to one. In the wiNNer method we trained separate regression models for each combination of precursor charge and fragmentation type, and report results for CID+2, CID+3, HCD+2 and HCD+3.

To evaluate the wiNNer model, we calculated the PCC between true and predicted peak intensities for each peptide in the testing dataset. We used Keras (<https://keras.io>) v.2.0.8, a high-level neural network application programming interface, to train a simple two-layer neural network model. TensorFlow v.1.3.0 was used as backend in Keras. The architecture of the neural network includes two hidden layers. The model was trained for y and b ions in the same neural network model; hence there are two output units. The input layer contains 549 features for a window size of 24. We repeated this analysis for sequence window sizes of 4, 8, 16 and 24 residues, the result of which is shown for CID+2 in Supplementary Fig. 3.13. The PCC increases monotonically with window size for all combinations of precursor charge and fragmentation type. Hence, we selected a window size of 24 for further analysis. Note that most peptides completely fit into the window since they are shorter than 24 residues. The number of hidden layers used for this window size is 312. Batch size, number of epochs, dropout and learning rate were optimized for all of the models separately. To test the window-based model, we used the independent test data taken from the 25 datasets, as shown in Supplementary Fig. 3.7. The PCCs of CID+2, CID+3, HCD+2 and HCD+3 models were 0.898, 0.793, 0.882 and 0.762, respectively. All four wiNNer models performed better than MS2PIP, as shown in Supplementary Fig. 3.17.

In addition to the neural network model, we explored support vector regression (SVR) and random forests from scikit-learn (<http://scikit-learn.org>, v.0.19.1) as machine learning algorithms. Because of practical limitations in training set size, we had to train the SVR on batches of 100,000 instances, the outputs of which were averaged. A radial basis function kernel was used, for which the width parameter was tuned in cross-validation. Then we trained a random forest in which we could use all training instances in one model. As a further option, we put a random forest layer on top of the output of SVR. Among all of these options, the neural network approach showed the best performance. For the comparison of different machine learning approaches, CID+2 was used with a sliding window size of 8, which reduced the computing power needed as compared with the optimal window size of 24. This is why performance is lower here on average compared with the final model. However, we believe that relevant conclusions can be drawn for relative comparisons between machine learning methods.

Benchmarking and comparison with known methodologies. Using the validation and testing datasets, we compared the performance of our models with that of MS2PIP, taking into account only singly charged y- and b-ion series. Although DeepMass:Prism is capable of predicting peak intensities of fragments with neutral losses of water and ammonia, we had to base the comparisons on y- and b-ion peak intensities owing to the limitations of the other predictor. The MS2PIP server (<https://iomics.ugent.be/ms2pip>) with default settings was used for all analyses. We also compared the performances of our models with the best possible theoretical performance. We determined this by calculating the technical variability between spectra in our validation and testing sets against their random replicate observation in the entire dataset.

Plasma sample processing. EDTA plasma samples collected from three healthy patients were pooled together and then clarified by centrifugation at 17,000g for 10min at 4°C. Aliquots were prepared and stored at -80 °C. Immediately before processing, plasma aliquots were thawed at room temperature. Subsequently, four replicate samples were prepared following the Biognosys Sample Preparation Pro Kit. Each sample was transferred into a LoBind tube (Eppendorf), dried by vacuum centrifugation and then stored at -80 °C.

Mass spectrometry for plasma samples. Dried peptide samples were resuspended by the addition of 20µl of 0.1% formic acid in water and water bath sonication for 10min. Samples were subjected to centrifugation at 17,000g for 5min at 4°C. Subsequently, 18µl was transferred into a new LoBind tube (Eppendorf) followed by the addition of 2µl of 10x iRT (indexed retention time) solution (Biognosys). Liquid chromatography–tandem mass spectrometry experiments were performed using 1-µl injections. Samples were subjected to reversed-phase chromatography with an Easy-nLC 1000 HPLC (Thermo Scientific) connected in-line with a Q Exactive Plus (Thermo Scientific) mass spectrometer. External mass calibration was performed before analysis. A binary solvent system consisting of buffer A (0.1% formic acid in water (v/v)) and buffer B (0.1% formic acid in 95% acetonitrile (v/v)) was employed. The mass spectrometer was outfitted with a nanospray ionization source (Thermo Nanoflex source). The liquid chromatography was performed using a PepMap100 3-µm C18 (75µm×2cm) trapping column followed by a PepMap RSLC 2-µm C18 (75µm×25cm) analytical column. For both DDA and DIA experiments, the same 120-min biphasic method was used, consisting of a gradient from 4% to 25% buffer B for 105min followed by 25% to 35% for 15min, at a flow rate of 300nl min⁻¹.

DDA of plasma samples. Each full-scan mass spectra was recorded in positive ion mode over the m/z scan range of 375–1,700 in profile mode at a resolution of 70,000. The automatic gain control (AGC) target was 3×10^6 with a maximum injection time of 50ms. The 12 most intense peaks were selected for HCD fragmentation. Tandem spectra were collected at a resolution of 17,500 with an AGC target of 1×10^5 and maximum injection time of 60ms. Dynamic exclusion and charge state screening were enabled, rejecting ions with an unknown or +1 charge state. An isolation window of 1.5 and normalized collision energy of 28 were used when triggering a fragmentation event.

DIA of plasma samples. Two scan groups were employed. First, using the selected-ion-monitoring scan group, we recorded a full-scan mass spectrum in positive ion mode over the m/z scan range of 400–1,200 in profile mode at a resolution of 70,000. The AGC target

was 3×10^6 with a maximum injection time of 100ms. Next, the DIA scan group was used to acquire 32 DIA segments of 15Da each at a resolution of 35,000. The AGC target was 1×10^6 with a maximum injection time of 120ms. An isolation window of 20 and normalized collision energy of 28 were used when triggering a fragmentation event. A global inclusion list was used to define each DIA segment.

Mass spectrometric data analysis of plasma samples. DDA data were processed with Proteome Discoverer (v.2.1), using Mascot as search algorithm. Fixed modifications included cysteine carbamidomethylation, and variable modifications included methionine oxidation and N-terminal acetylation. The files were searched against the human UniProt proteome database (downloaded 17 February 2016). DIA data were processed with Spectronaut (v.11, Biognosys) using default settings.

DeepMass:Prism model interpretation. To interpret our model, we applied the method of integrated gradients. Integrated gradients attributes the predicted output of a neural network to the set of input features (analogous to inspecting the product of the input features and coefficients in a linear model). For a given peptide sequence and precursor metadata, this method indicates the influence between each amino acid residue in the peptide sequence and the predicted intensity of each spectrum peak. Residues can either positively or negatively influence a peak's predicted intensity. Essentially, this provides a square $N \times N$ attribution matrix denoting the influence of residue i on peak j , where N is the peptide length (Supplementary Fig. 3.11). The diagonal elements of this matrix represent the degree to which the peak intensity is influenced by the identity of the residue at the cleavage site, while off-diagonal elements denote long-range interactions, where a peak's intensity is influenced by the identity of a distant residue. To focus on the influence of peptide identity, we held constant the precursor-level metadata and did not calculate gradients at the context. Specifically, for all peptides analyzed, we assumed a +2 charge state, fragmentation by HCD and Fourier transform mass spectrometry mass analyzer. The sum along a column of the attribution matrix equals the predicted intensity of the represented peak. Columns of this matrix were normalized by their maximum value such that the most positive attribution to peak intensity had value 1.0; other attributions were scored relative to this value. As peak values are non-negative, all columns have at least 1 element equal to 1.0, yet negative attributions can decrease below -1.0 in extreme cases.

To determine distances between interactions (Supplementary Fig. 3.12), we used an attribution threshold of ± 0.7 ; any normalized value more extreme than this threshold was considered a major attribution. The directed distances between this major attribution and cleavage site were determined such that positive distances corresponded to instances where the attributed residue was situated downstream of bond cleavage (toward the C terminus). Finally, distances were normalized to the length of each particular peptide, such that a distance of ± 1.0 corresponded to a full fragment ion length.

To determine the influence of specific amino acids on peptide fragmentation, we repeated the analysis on a per-residue basis (Supplementary Fig. 3.18). Specifically, for each amino acid, we calculated the distribution of distances of major attributions. Except in a few notable exceptions, amino acids did not greatly deviate from the general trend already described (Fig. 3.4). Nonetheless, we saw relevant clustering of per-residue profiles in the positive

attribution of b ions. Broadly, we observed hydrophilic amino acids clustering distinctly from large hydrophobic amino acids. Proline is expected to show distinct behavior because of the well-known proline effect [Hunt et al., 1986]. Indeed, it represents a notable outlier, and had substantially longer-reaching positive influence on predicted intensities up- and downstream (Supplementary Fig. 3.18, upper-right plot). Similarly, among negative attribution profiles, we identified two trends. First, branched-chain amino acids and proline had an influence relatively concentrated at the cleavage site, and second, they had a smoother distribution of influence downstream; asparagine was a notable outlier, with its strongest influence on peaks just upstream of it.

Evaluation of intensity prediction models in Andromeda scoring performance. We analyzed published HeLa whole-cell lysate data, obtained from Kelstrup et al. [Kelstrup et al., 2018] (list of raw files: 20161213_NGHF.DBJ_SA_Exp3A_HeLa_1ug_60min_15000_01.raw, 20161213_NGHF.DBJ_SA_Exp3A_HeLa_1ug_60min_15000_02.raw, 20161213_NGHF.DBJ_SA_Exp3A_HeLa_1ug_60min_15000_03.raw).

To compare the identification of MS/MS spectra using the conventional Andromeda scoring and the intensity-informed scoring, we predicted intensities for all candidate PSMs using DeepMass:Prism, wiNNer or MS2PIP. For the search configuration parameters, Trypsin/P was specified as enzyme, carbamidomethylation of cysteine was specified as a fixed modification and no variable modifications were selected. Protein FDR control was disabled to report results dependent on a q value on the PSM level. The examples in Supplementary Figs. 3.19 and 3.20) were taken from dataset PXD004732 in the PRIDE archive (ProteomeTools).

When comparing the Andromeda scores without and with using intensity prediction on a dataset consisting of synthetic peptides and, hence, with known ground truth [Zolg et al., 2017], we saw several examples for which the correct PSM was not the highest-scoring one for that MS/MS spectrum when the conventional Andromeda score was used but became the highest-scoring PSM when the intensity-informed Andromeda score was used (Supplementary Figs. 3.19 and 3.20). As illustrated in Supplementary Figs. 3.19 and 3.20), examples where the highest-scoring PSM changed after including intensity information included cases where two adjacent amino acids were swapped and cases where a completely new peptide sequence became the top-scoring PSM.

Retention time prediction. We constructed an RNN model with a bidirectional LSTM layer with 40 hidden units, followed by another LSTM layer with 20 hidden units. The last output in the output sequence was then fed into 2 dense layers, the first with 20 hidden neurons and hyperbolic tangent activation function, and the last with 1 neuron and linear activation function. The input to the model is a onehot-encoded sequence of amino acid residues, and the output is a predicted iRT value for the input peptide. We used the Adam optimizer and mean squared error as the loss function, and applied the dropout rate of 0.2 to each layer. The model was implemented using the Keras library and Tensorflow backend. The model was trained on data collected in-house from three different samples: human plasma, HeLa cell lysate and yeast cell lysate. The 69,680 peptide-charge pairs were split into training, validation and test set in a 75:20:5 ratio. It is worth noting that this model is not complete yet—it

was built to assess the maximum possible peptide identification rate when a spectral library is completely generated in silico (that is, predicting both fragment ion intensities and retention times for each peptide sequence). As the model was trained using a limited number of samples that were subjected to liquid chromatography conditions identical to the DDA and DIA data obtained for plasma samples, we have not evaluated how well the model will generalize (that is, different liquid chromatography and column systems). It is even conceivable that the future model will have to be fine-tuned on each liquid chromatography–mass spectrometry setup independently.

To evaluate the full potential for generating spectral libraries completely by computation (that is, peptide fragment ion abundances and peptide retention time information are generated in silico), we predicted both precursor peptide retention times and MS/MS spectra. This combined model we termed DeepMass:Drip. It predicts iRT-calibrated retention times with high accuracy (R^2 of 0.97 and a median error of 4.84 as compared with R^2 of 0.88 and median error of 8.68 for SSRCalc [Krokhin, 2006]). To evaluate this method, we first generated a spectral library using DeepMass:Prism+Drip (that is, predicting fragment ion intensities and retention times for each of the 7,441 peptides from the DDA library). After performing Spectronaut searches, we quantified on average 4,957 peptides, 291 (5.5%) peptides fewer than the DDA library (Fig. 3.4). Here, too, peptides unidentified by DeepMass:Prism+Drip had low-confidence Spectronaut scores during the DDA library searches, with 118 (41%) peptides having q values worse than 10^{-3} (Supplementary Fig. 3.15).

Statistics. In box plots, each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5 multiples past the interquartile range between the low and high quartiles. Data points beyond these ranges are considered outliers and are plotted as individual data points. Numbers of data points used in each box plot are provided in the respective figure legends. The q values for PSM FDRs were estimated in MaxQuant with its standard targetdecoy search strategy.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The mass spectrometry proteomics data including summary tables have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD010382.

Code availability

We offer the trained DeepMass:Prism model for use via the Google Cloud ML platform (<https://github.com/verilylifesciences/deepmass/tree/master/prism>). To obtain the trained DeepMass:Prism model to run locally, please contact the corresponding authors. A user-friendly interface will be made available in the future MaxQuant releases.

3.4 Results

Bidirectional RNN for spectral prediction. The problem of accurately predicting tandem mass spectra has long eluded conventional machine learning approaches for several reasons. First, because peptide sequences vary in length, they are incompatible with many algorithms that take fixed-length representation as an input. Second, different fragmentation and acquisition methods can be used to generate and acquire tandem mass spectra, each producing considerably different results. Moreover, precursor peptides are fragmented into different types of ions, where the abundance of any one ion type can be dependent on another. Training multiple models for each fragmentation method and ion type does not take advantage of such dependencies. Lastly, with large, publicly available mass spectrometry repositories, training of conventional algorithms (for example, support vector machines) becomes difficult, but complex, nonlinear approaches of deep learning become viable. Taking all of this into consideration, we selected RNNs [Graves et al., 2009]. RNNs are a class of artificial neural networks that are designed to work with sequential information, can accept inputs at different levels (for example, amino acid, peptide fragment and machine type) and of different types (for example, amino acid identities or their physicochemical properties), can predict multiple values simultaneously and can support training on datasets with millions of entries (3.3 Methods and Fig. 3.1).

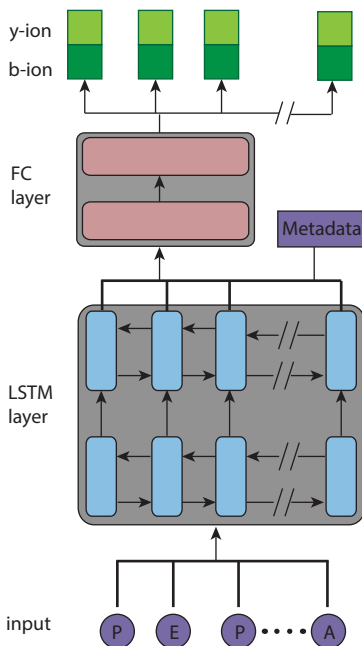


Figure 3.1: Bidirectional RNN architecture for the prediction of fragment intensities. The neural network contains two basic modules: an RNN encoder and a perceptron decoder. The encoder takes a one-hot-encoded peptide sequence as an input and outputs its fixed-length representation vector. The sequence representation vector is then combined with metadata features and input into the decoder. The decoder contains a set of fully connected layers that outputs intensities of different fragment ion types (for example, y and b ions) at each position in the input peptide sequence. FC, fully connected.

Predictive performance of the bidirectional RNN. The datasets that are used for measuring the predictive performance are described in detail in the Methods. In brief, 25 complete datasets containing more than 60 million tandem mass spectrometry (MS/MS) spectra were used for training, testing and validation (Supplementary Fig. 3.7 and Supplementary Table 3.1). We first evaluated the performance of DeepMass:Prism against that of MS2PIP (ref. [Degroeve et al., 2013]) using the Pearson correlation coefficient (PCC) between true and predicted intensities for each peptide. When we compared all peptides in our testing set, we found that the accuracy of our model was markedly better than that of MS2PIP, with a PCC of 0.944 versus 0.871 (Fig. 3.2). We also calculated the PCC of repeatedly collected mass spectra of the same peptides to quantify technical variability in our dataset. Interestingly, our model’s PCC nearly approached the theoretical maximum imposed by this measurement reproducibility of 0.976 (Fig. 3.2 and Supplementary Fig. 3.8).

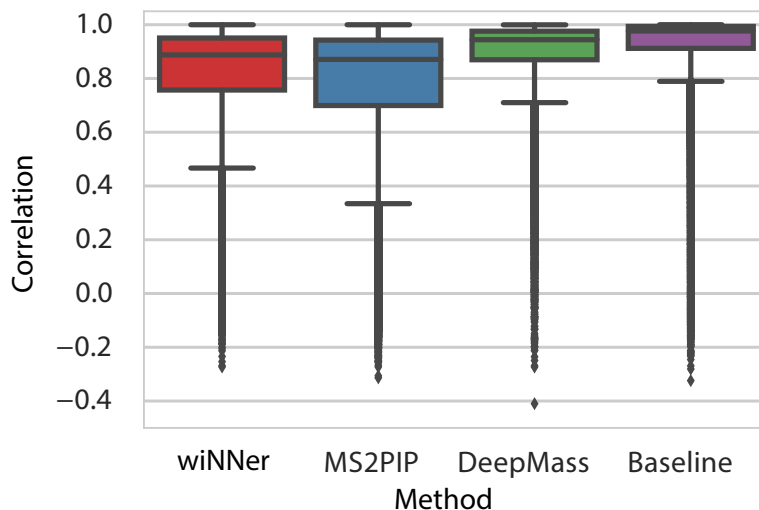


Figure 3.2: Comparing the performance of fragment ion intensity predictions for the different machine learning strategies. The box plots show distributions of PCCs between actual and predicted y- and b-ion peak intensities for each peptide in our testing dataset. The box plots contain 69,888, 69,888, 65,996 and 62,486 unique PSMs from the independent testing datasets for DeepMass:Prism, wiNNer, MS2PIP and technical variability, respectively. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5 multiples past the interquartile range between the low and high quartiles. Data points beyond these ranges are considered outliers and are plotted as diamonds.

DeepMass:Prism was highly accurate across CID and HCD fragmentation methods (median PCCs across all peptides in our testing dataset of 0.958 and 0.925, respectively; Supplementary Fig. 3.8) and for Fourier transform mass spectrometry and ion trap mass spectrometry mass analyzers (median PCCs of 0.924 and 0.949, respectively). The model was also accurate for precursor ion charges from 1 to 3 (median PCCs of 0.931, 0.952 and 0.901, respectively), with performance dropping at higher charges because of the lack of data. The length of the peptide only slightly affected the performance, with PCC for peptides with 5–10 amino acids

being only marginally better than that for peptides with 30–35 residues (0.964 and 0.908, respectively). Similarly, the Andromeda score of a peptide-spectrum match (PSM) minimally affected the performance, with PCC for PSMs with a score of 200 being slightly lower than that for PSMs with a score of 700 (0.937 and 0.954, respectively). Finally, metadata features are crucial for accurate predictions; a model that does not take any metadata as inputs performs poorly (median PCC of 0.810; Supplementary Fig. 3.9).

Sliding-window-based prediction. Another way to construct a regression model that can be applied to peptides of variable sequence length is to use a sliding-window-based approach. Prediction of local protein properties on protein sequence windows has a long tradition and has been applied, for instance, to predict secondary structure, solvent accessibility and trans-membrane regions [Garnier et al., 1996, Rost et al., 1994]. The feature space is constructed from a sequence window centered on the backbone bond that is fragmented, extending k amino acids to the left and to the right from the backbone bond under consideration (Fig. 3.3). Amino acids at the N and C termini, their corresponding distance to the target bond and length of the peptide were also included in the feature space.

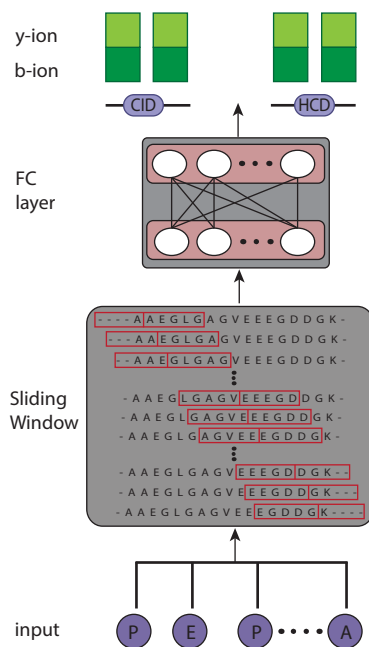


Figure 3.3: Sliding-window-based regression model for prediction of fragment intensities. A symmetrical sliding window is placed around the target peptide bond for which the b- and y-ion intensity should be predicted (red boxes). Amino acids in the window are translated into 0/1 variables by one-hot encoding. Additional features, including the amino acids at the N and C termini, the distances of the bond to the termini and the peptide length, are added to the feature space. This process is repeated for each position in the input peptide sequences. A fully connected two-hidden-layer neural network is then trained and outputs the logarithmic b- and y-ion intensities.

We applied three types of machine learning algorithms to this feature space: support vector machines [Vapnik, 1995, Drucker et al., 1997], random forests [Breiman, 2001] and a two-hidden-layer neural network. Of the three approaches, the neural network strategy wiNner showed the best performance (Supplementary Fig. 3.10). Although wiNner did not reach the prediction accuracy of DeepMass:Prism (Fig. 3.2), it performed better than the best existing conventional machine learning method for all tested combinations of precursor charge and fragmentation type. And, similar to DeepMass:Prism, with wiNner only one model was needed to cover peptides of various lengths. Overall, this approach could be advantageous in situations where fast training is needed, such as when dataset-specific models are needed for the analysis of laboratory-specific proteomics data.

Interpretation of the DeepMass:Prism model. We first explored the agreement between the outputs of our RNN model and observed fragmentation efficiencies between different amino acid pairs [Shao et al., 2014] (Supplementary Fig. 3.11). For a set of A-A-A-[X]-[Z]-A-AA-R peptides, where [X]-[Z] represents all possible combinations of amino acid residue pairs, we predicted the fragmentation efficiencies between the [X]-[Z] residue pair (Supplementary Fig. 3.11). Similar to the previous findings, our model reported notably higher fragmentation efficiency between [X]-Pro residues (where [X] can be any other residues), for both y- and b-ion types. The model also correctly predicted less efficient fragmentation between [X]-[Z] residue pairs where [X] is a hydrophobic residue. Furthermore, the model also correctly identified an increased fragmentation efficiency for b ions between His-[Z] pairs (Supplementary Fig. 3.11).

We next tested whether residues further up- or downstream from the site of fragmentation also contribute to the peak intensity assignment. Such long-range ‘interactions’ in peptide fragmentation have not been extensively studied, even though they can be observed in our dataset (Supplementary Fig. 3.12). For example, analysis of 63 pairs of peptides with a single residue mismatch showed notable differences in b-ion intensities 5–10 residues toward the C terminus from the mismatch site, while y-ion intensities showed less variability. Moreover, the fact that the window-based approaches that use small window sizes (Supplementary Fig. 3.13) performed poorly further supports the existence of these long-range interactions.

Based on these findings, we used our model to study long-range interactions. After randomly selecting 1,000 peptides from the independent testing set, we calculated the integrated gradients [Mukund Sundararajan, Ankur Taly, 2017] over each position of the input peptide sequence and each ion type. Using these gradients, we attributed each peak’s predicted intensity to the summed influence of every amino acid residue in the precursor fragment (Fig. 3.4); residues are able to positively or negatively influence any peak intensity. Finally, for each intensity prediction, we termed the most influential residues ‘major attributions’ (3.3 Methods), which can have both signs.

We found abundant evidence of long-range interactions, that is, major attributions not adjacent to the site of cleavage (Fig. 3.4 and Supplementary Fig. 3.14). Specifically, within b ions, major positive attributions were abundant toward the N terminus from the target residue, up to 75% of the length of the precursor peptide. While major negative attributions were notably less common, we observed many negative attributions a similar distance toward the C terminus from the target residue. We observed a different pattern in y ions: positive

and negative attributions were more even on either side of the target residue. Nonetheless, while positive attributions were broadly spread about the length of the fragment, negative attributions were more tightly concentrated at the cleavage site, with a smaller cluster near the C terminus (an analysis on a per-residue basis is described in the 3.3 Methods section).

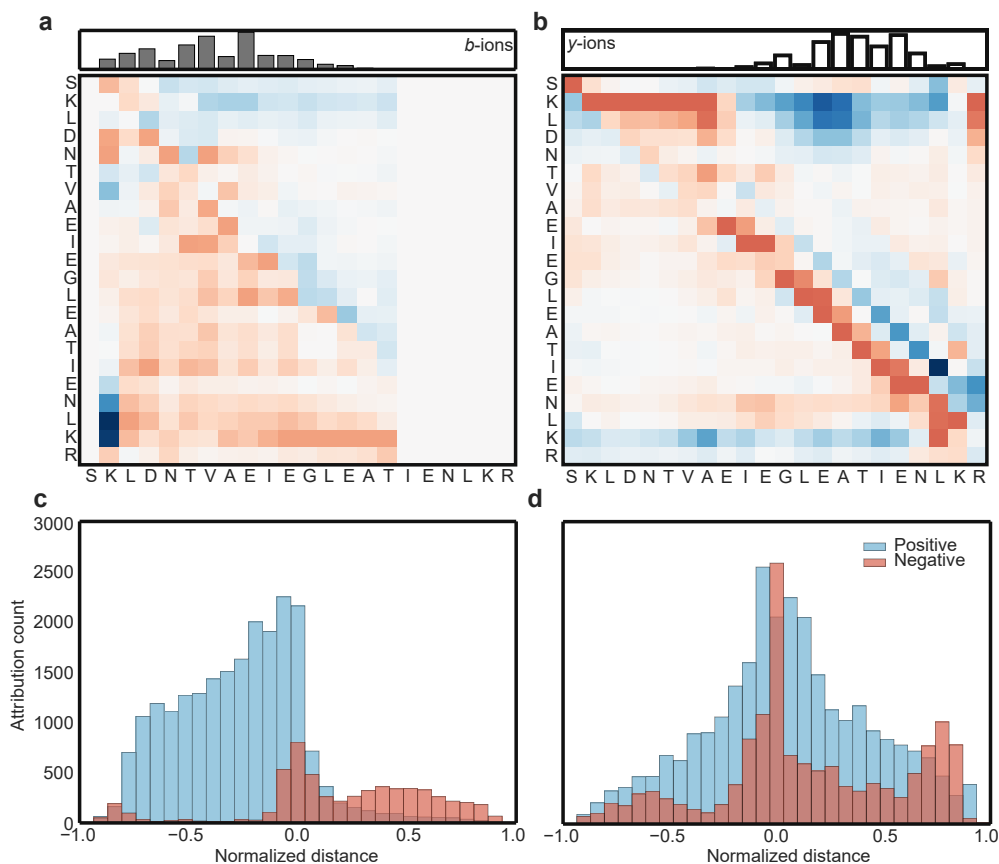


Figure 3.4: Interpretation of the DeepMass:Prism model. Left, y ions; right, b ions. **Top:** to help illustrate how the DeepMass:Prism prediction model works, the output for an example peptide sequence, SKLDNTVAEIEGLEATIENLKR, is plotted in an attribution table, which correlates the relationship between a position along the peptide sequence with the observed fragment ions. At the top are the scaled predicted peak intensities of the fragment ions corresponding to cleavage of the target residue at that position. Each pixel (i, j) in a heat map corresponds to the influence of residue i on peak intensity j. Blue pixels correspond to positive influence, and red to negative; values across each column j are normalized by the maximum value in that column. **Bottom:** histogram plots illustrate the distribution of distance between peak intensity and major attributions. Major attributions are attribution values with absolute values greater than or equal to 0.7 (based on heat map pixels in the top panel). Directional distances are normalized by peptide length; positive distances indicate a residue influencing cleavage toward the C terminus. Similar plots, but with unnormalized distances, are shown in Supplementary Fig. 3.14.

DIA spectrum matching to predicted spectra. An accurate prediction of MS/MS peptide spectra is expected to benefit areas in which reference spectral libraries are utilized (and generated) for data analysis, as is the case for DIA and selected/multiple reaction monitoring [Schubert et al., 2015, Wu et al., 2016]. The most common approach to analyze DIA data requires using a spectral library to determine the peptide identity. Although some library-free methods exist, such as DIA-Umpire [Tsou et al., 2015] and DirectDIA, they are typically less sensitive than library-based methods. Accordingly, this necessitates performing a series of DDA experiments to build a sample-specific reference library. Given the stochastic nature of DDA, a single liquid chromatography–tandem mass spectrometry run is usually incomplete. Instead, replicate runs, and even sample fractionation, are typically required, further increasing experimental costs. Here, we tested whether *in silico*–generated spectral libraries, created using DeepMass:Prism, could replace those that are produced experimentally.

To evaluate this strategy, we processed a pooled human plasma sample into peptides using the Biognosys Sample Preparation Pro Kit, producing four replicate samples. We collected DIA data in triplicate for each sample, as well as DDA data in duplicate for one sample. A sample-specific spectral library was generated from the DDA data with Proteome Discoverer (3.3 Methods), and the DIA data were processed using Spectronaut [Bruderer et al., 2017]. The spectral library contained 7,441 peptides, of which 5,248 (71.0%) were identified and quantified on average (Fig. 3.5 and Supplementary Fig. 3.15). We then used DeepMass:Prism to generate an *in silico* spectral library for the same set of peptides and used it in another Spectronaut search. Specifically, the peak intensities for fragment ions for each of the 7,441 peptides from the DDA library were replaced with values from DeepMass:Prism. However, retention time information was preserved from the DDA-generated spectral library. In an ideal scenario, this approach would identify the same number of peptides as when using the experimental generated library, and we came close—we identified and quantified on average 5,181 peptides, only 103 (1.9%) peptides fewer (Fig. 3.5), and with a high overlap (5,131 peptides or 97.1%). We repeated the analysis with a model that also predicted peptide retention time information *in silico*, with similar results (3.3 Methods).

Interestingly, the 103 peptides unidentified by DeepMass:Prism typically had low-confidence Spectronaut scores, with 46 peptides (45%) having *q* values worse than 10⁻³ in the DDA spectral library searches (Supplementary Fig. 3.15). Importantly, we observed highly correlated peptide abundance quantification in Spectronaut searches between the experimental library and the *in silico* library (PCC of 0.99; Supplementary Fig. 3.15c). As a control, we also generated a spectral library by predicting fragment intensities using MS2PIP (and preserving retention time information from the DDA data). We observed a much lower peptide identification rate compared with that for DDA libraries (on average, 3,976 or 75.8% peptides were identified; PCC of 0.96). Some of the difference between the numbers of peptides identified by MS2PIP- versus DeepMass:Prism-based libraries can be attributed to MS2PIP's lack in predicting spectra for peptides with modifications. However, even after the removal of methionine oxidation, the DeepMass:Prism-based library identified 4,661 peptides (17.2% more than MS2PIP).

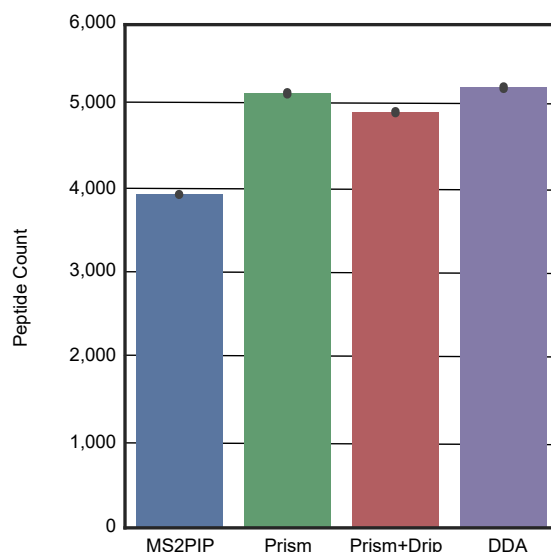


Figure 3.5: Application to spectral library generation for DIA data analysis. The bar plot compares the average number of identified and quantified peptides in plasma samples when using spectral libraries generated with MS2PIP, DeepMass:Prism, DeepMass:Prism including retention time prediction and the standard approach (that is, DDA experiments).

Application to DDA peptide search engine scores. Intensity information in MS/MS spectra is highly informative and is expected to help in finding the correct PSM for a given MS/MS spectrum. Despite this, most search engines make no or little usage of the intensity information. For instance, the Andromeda search engine [Cox et al., 2011] uses peak intensities only to give higher weights for matches to the high-intensity peaks. However, it does not utilize expectations of peak intensities as they relate to sequence context when scoring PSMs. Here, we show that scoring PSMs can benefit from predicted fragment ion intensity information. We integrated intensity predictions into the Andromeda scoring function in the following way: the Andromeda score of a given PSM was replaced by a maximum of several attempts to score the same spectrum. One of these scores was the original Andromeda score. We then calculated the score on subsets of the ions in theoretical spectra, always taking a certain number of top intensity peaks (by default we took the top 3, 5, 7, 10 and 13 peaks) from the theoretical spectrum with intensities predicted by DeepMass:Prism. This strategy was similar to focusing on only the most intense transitions in the analysis of the selected reaction monitoring data. Naively, one may expect that reducing the number of theoretical peaks would reduce the number of matching fragments and hence reduce the score. Despite this, the Andromeda score of a match may still increase, even if the number of matches decreases, if the summed probability of finding this many or more matches by chance, given the number of provided theoretical fragments, decreases [Cox et al., 2011].

We compared the performances for a complex sample, in this case HeLa whole-cell lysate. Overall, we found that including intensity predictions increased both the PSM and peptide identification rates. We then compared the relative performance among DeepMass:Prism, wiNNer and MS2PIP by varying the PSM-level false discovery rate (FDR) in MaxQuant, which corresponds to scanning the q value (Fig. 3.6). Over the whole range of q values tested, there was a gain in PSM identifications when we used DeepMass:Prism and wiNNer, both of which outperformed MS2PIP (Fig. 3.6a). In particular, improvements through intensity prediction were the largest in the high-specificity range (that is, q value < 0.01) (Fig. 3.6b). Note that the improvements in MS/MS identification rates are on top of already high rates of 50% for the conventional Andromeda search (Fig. 3.6c). In terms of gains in unique peptide sequences, DeepMass:Prism also outperformed wiNNer and MS2PIP (Fig. 3.6d). Finally, DeepMass:Prism also outperformed wiNNer and MS2PIP in the number of identified protein groups (at 1% protein-level and PSM-level FDR), with a gain of 3.9% versus 2.3% and 2.1%, for wiNNer and MS2PIP, respectively.

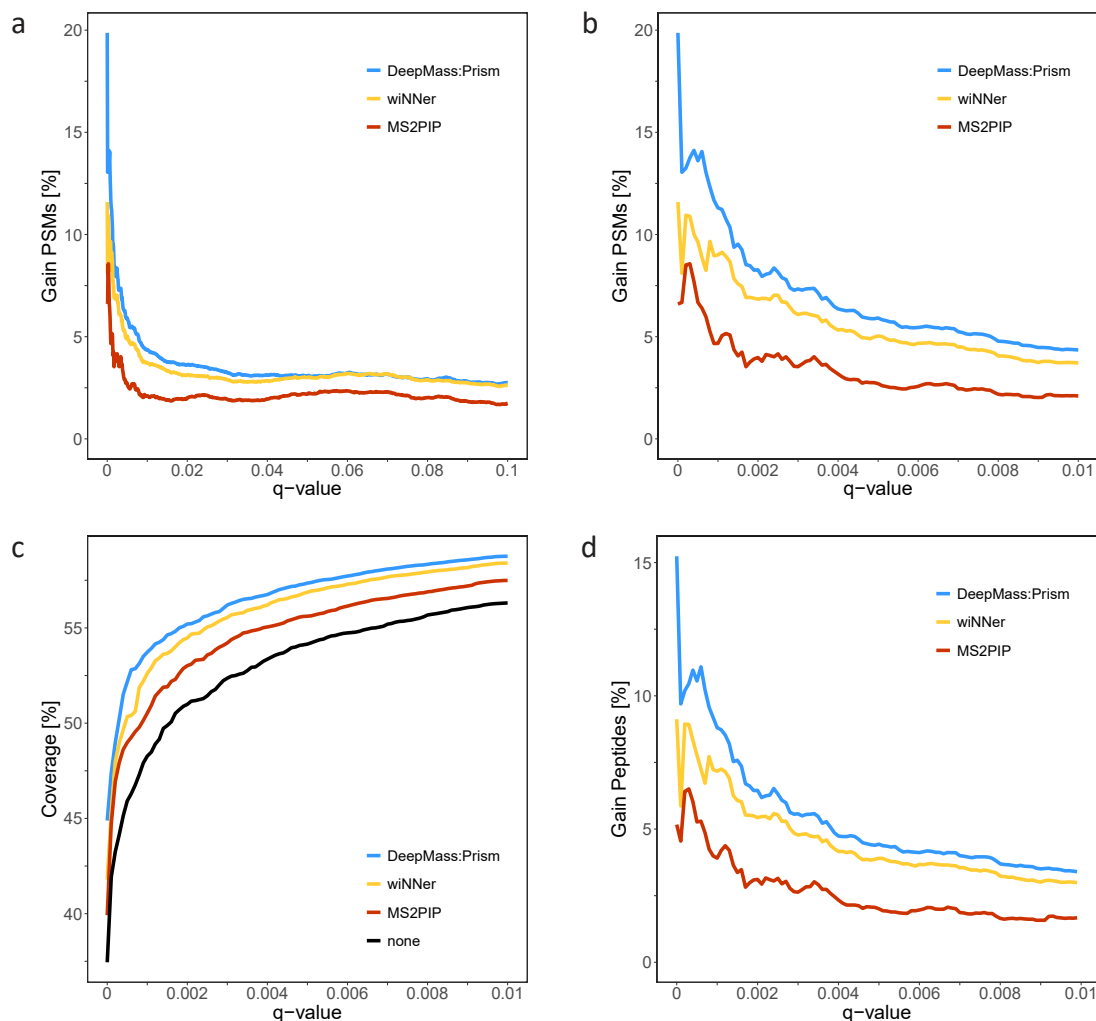


Figure 3.6: Application to PSM scoring for DDA data analysis. Identification-rate improvements on a complex HeLa dataset as a function of q value are shown for MS2PIP, wiNner and DeepMass:Prism. The q values are calculated based on target-decoy search in MaxQuant. **a**, The percentage increase in PSMs compared with Andromeda searches without intensity prediction. The q value ranges from 0% to 10%. The curves are based on 146,606, 146,411 and 145,160 forward PSMs for DeepMass:Prism, wiNner and MS2PIP, respectively. **b**, Same as **a**, but zoomed into the q value range from 0% to 1% and based on 127,806, 127,024 and 125,046 forward PSMs for DeepMass:Prism, wiNner and MS2PIP, respectively. **c**, Coverage (that is, the percentage of identified MS/MS spectra) as a function of q value, without using intensity prediction and when using each of the three prediction models. **d**, The percentage increase in unique peptides compared with Andromeda searches without intensity predictions. The curves are based on 42,505, 42,335 and 41,793 peptides for DeepMass:Prism, wiNner and MS2PIP, respectively

3.5 Discussion

Through the use of machine learning, we demonstrate that MS/MS spectrum prediction can be nearly as accurate as the limits of technical reproducibility. Importantly, this can be taken advantage of in both DDA and DIA computational workflows to improve peptide identification rates and reduce the reliance on spectral libraries. Further, the deep learning regression models described here are highly interpretable, capturing how sequence features, including the interactions across multiple amino acid residues, contribute to peptide fragment ion abundance and the mobile proton hypothesis. Additionally, although the conventional window-based machine learning approach we described has slightly inferior predictive performance, it is less computationally intense to train. For both DDA and DIA application, integration of intensity prediction into the MaxQuant [Cox and Mann, 2008, Tyanova et al., 2016a] environment is currently ongoing.

So far, we have restricted the spectrum predictions to peptides that are not carrying modifications, except for methionine oxidation. However, the generalization to PTMs is straightforward. The modified residues can be encoded as the 21st, 22nd and so on amino acids. Modification-specific neutral losses other than water and ammonia will need to be added for some modifications, as for instance serine and threonine phosphorylation. Non-tryptic peptides can already be accommodated with the current model, and predictions can currently be made for them. However, since the training dataset was from shotgun proteomics data submitted to the PRIDE database [Vizcaíno et al., 2016], it is biased toward tryptic peptides, and DeepMass:Prism performs better for these. As more data for nontryptic peptides and more machine types become available, we will update our models to improve predictions for all peptides.

While these models to predict peptide fragment intensities will benefit peptide search engines, we anticipate that their greatest impact will be through providing in silico-based spectral libraries. For example, if an existing spectral library was generated using a fragmentation mode different from the data that needed to be analyzed, a new library could be generated in silico rather than experimentally. When it comes to the content of a spectral library, approaches are needed to construct a list of peptides based on previous knowledge. For example, if plasma samples are analyzed, proteins and peptides from the Plasma Proteome Database [Nanjappa et al., 2014] could be used to construct the library. Another potential application will be to supplement an existing experimental library with a small number of hypothesis-driven peptides, such as peptides that harbor genetic variants, tumor mutations or post-translational modifications, expanding the range of interesting scientific and clinical questions beyond measuring the levels of proteins in a sample. A further option for analyzing cellular or tissue proteomes will be to construct an in silico library for the entire proteome. Such a spectral library could be supported by both RNA-sequencing data and peptide observability predictions [Mallick et al., 2007, Sanders et al., 2007]. We continue to explore these strategies and envisage a time when predicted spectral libraries will become a necessary and beneficial tool for proteomics and proteogenomics.

3.6 Author's contribution

For accurate peptide and protein identification of MS-based proteomics data using database search engines, not only the quality of the acquisition data is of importance but also the in-silico generated theoretical MS/MS spectra, representing the peptide sequences of a reference database. The intensity values of calculated m/z ratios are typically set to a constant value given the mathematical challenge to determine the frequency with which a peptide backbone breaks.

My main contribution to this publication was the extension of the database search engine Andromeda to replace the constant intensity values with predicted intensity values of ions. The predicted intensities of theoretical add another layer of peptide sequence context to the scoring and, thus, increase the identification rate. To achieve this, I contributed to the modification of the Andromeda score, tested the algorithm and developed required adaptations. For the inclusion of intensity values in the scoring, I extended MaxQuant to provide an interface to read the predicted intensity values from an external file or directly from the intensity prediction model.

Additionally, I validated the extended scoring method based on a synthetic dataset to compare the results to a known ground truth and to a biological dataset. For the validation process, I extended the functionality of MaxQuant to provide intermediate outputs of the top 15 PSMs for each experimental spectrum to be able to predict the intensity values of their sequences. Further, I benchmarked the intensity prediction models—DeepMass:Prism and wiNNeR—to state-of-the-art models based on the extended Andromeda score. For the benchmark, I also provided the MS2PIP intensity predictions. Finally, I was part of the dataset selection and analysis for the development process of our deep learning model (DeepMass:Prism) and our neural network model (wiNNeR).

The work described in this publication successfully applied machine learning techniques to predict the fragment ion intensity values with high accuracy. We demonstrated that the inclusion of predicted intensities into the Andromeda score calculation increases the number of identified spectra at 1% FDR. In particular, improvements through intensity prediction were with nearly 20% gain largest within the high-specificity range (q -value < 0.01).

3.7 Additional information

Acknowledgements

This project has received funding from the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 686547 (S.T.) and from FP7 grant agreement no. GA ERC-2012-SyG_318987-ToPAG (J.C.). J.C. and P.G. are supported by the Marie Skłodowska-Curie European Training Network TEMPERA, a project funded by the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 722606. We thank E. Deutsch, J. Bingham, M. Liu, R. Perrone, P. Kheradpour, B. Brown, M. Edwards, L. Cao, N. Soltero, J. Lehar, T. Snyder, D. Glazer and T. Stanis for their help, support and suggestions.

All author's contributions

S.T., R.L., P.G., F.S.S., P.C. and J.C. designed and developed the code, and performed the analyses. M.B. and L.D. helped with deep learning architecture design, as well as with reviewing the code and analyses. A.B. helped with data ingestion and preprocessing. K.K.P. carried out the wet-laboratory experiments and the DIA data analysis. R.L., P.C. and J.C. wrote the manuscript and directed the project. All authors read and approved the final manuscript.

3.8 Supplementary information

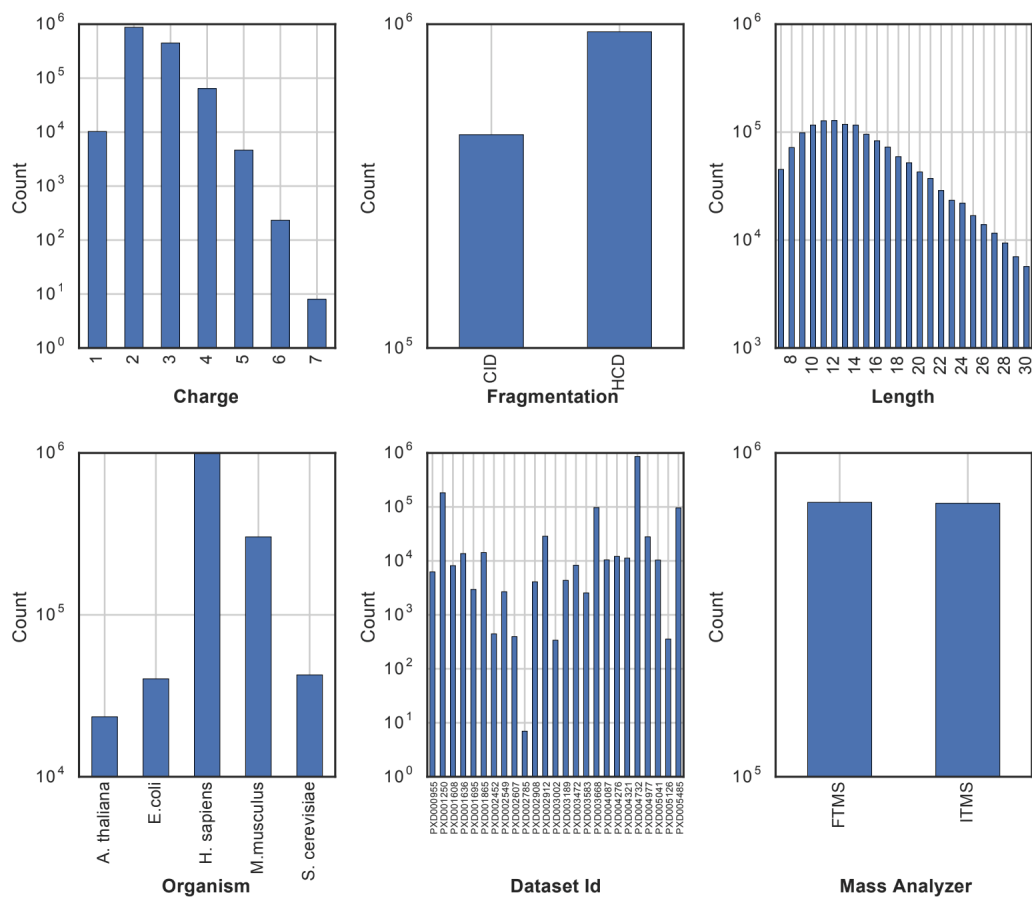


Figure 3.7: Summary of training, validation and testing datasets. The charts show distribution of various peptide and acquisition properties in our training dataset (based on total of 1,263,431 unique sequence/charge/fragmentation/mass analyzer peptide combinations). Note that all counts on the vertical axes are on a logarithmic scale.

3. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis

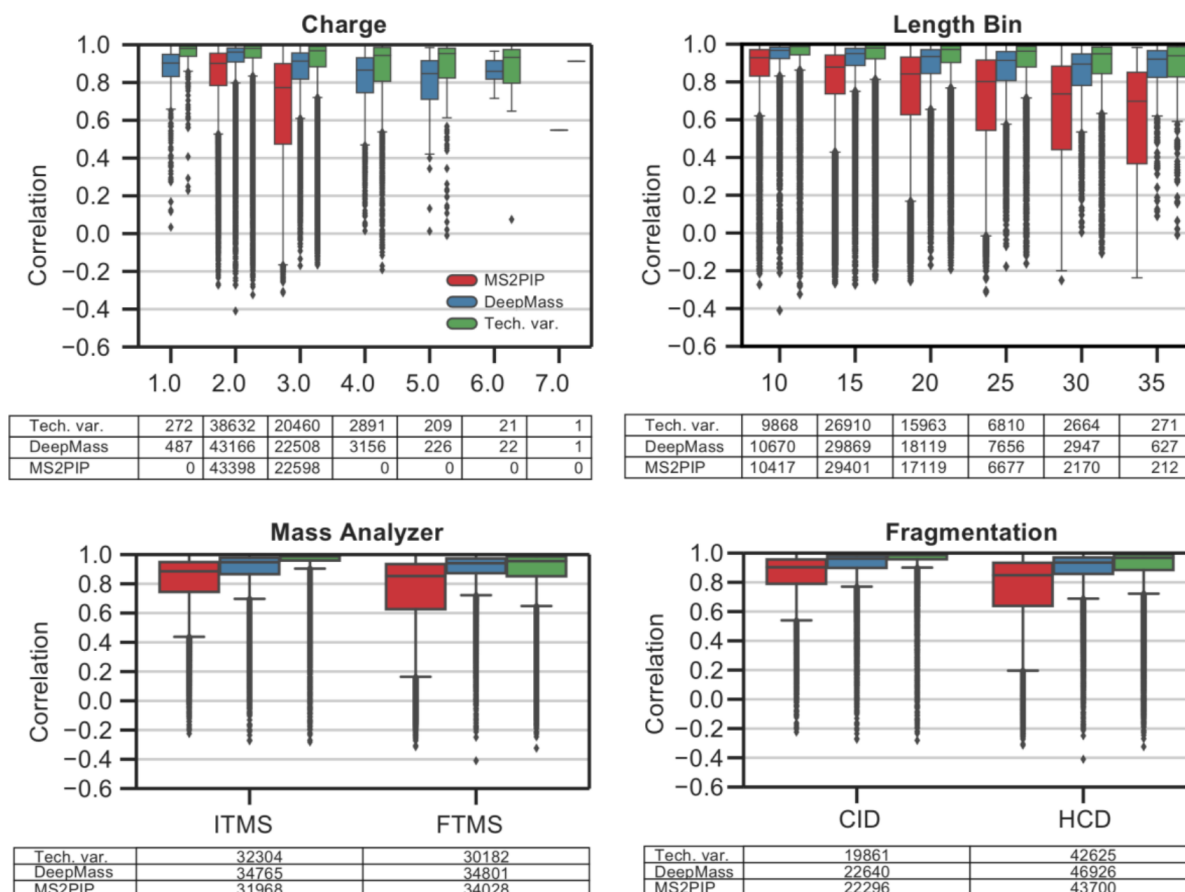


Figure 3.8: Performance of DeepMass:Prism. The box plots show distributions of correlation coefficients (PCC) between actual and predicted y- and b-ion peak intensities for each peptide in our testing dataset, stratified by peptide charge and length, and by mass analyzer and fragmentation type. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Values beyond these ranges are considered outliers, and are plotted as diamonds. Precursor counts represented by each box (N) are listed in tables below each box plot.

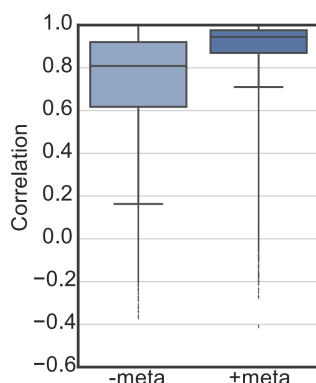


Figure 3.9: Performance of DeepMass:Prism with and without metadata. The box plots show distributions of correlation coefficients (PCC) between actual and predicted y - and b -ion peak intensities for each peptide in our testing dataset, based on our final model with metadata features (precursor length, precursor charge, mass analyzer, fragmentation type; right), and a model with no metadata features (left). Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Values beyond these ranges are considered outliers, and are plotted as diamonds. The number of precursor (N) in both distributions is 69,888.

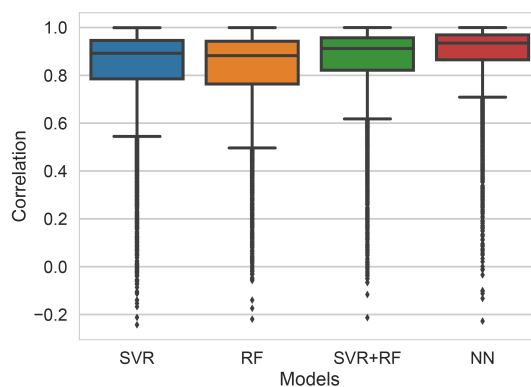


Figure 3.10: Sliding window-based approach using different machine learning algorithms. Distribution of PCCs of the window-based approach using different machine learning algorithms such as support vector regression (SVR), random forest (RF), RF layer on top of the output of SVR (SVR+RF) and two layer neural networks (NN). The comparison was done on CID+2 model and window size 8. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Values beyond these ranges are considered outliers, and are plotted as diamonds. The boxplots contain 9,214 unique PSMs from the ProteomeTools dataset (PXD004732).

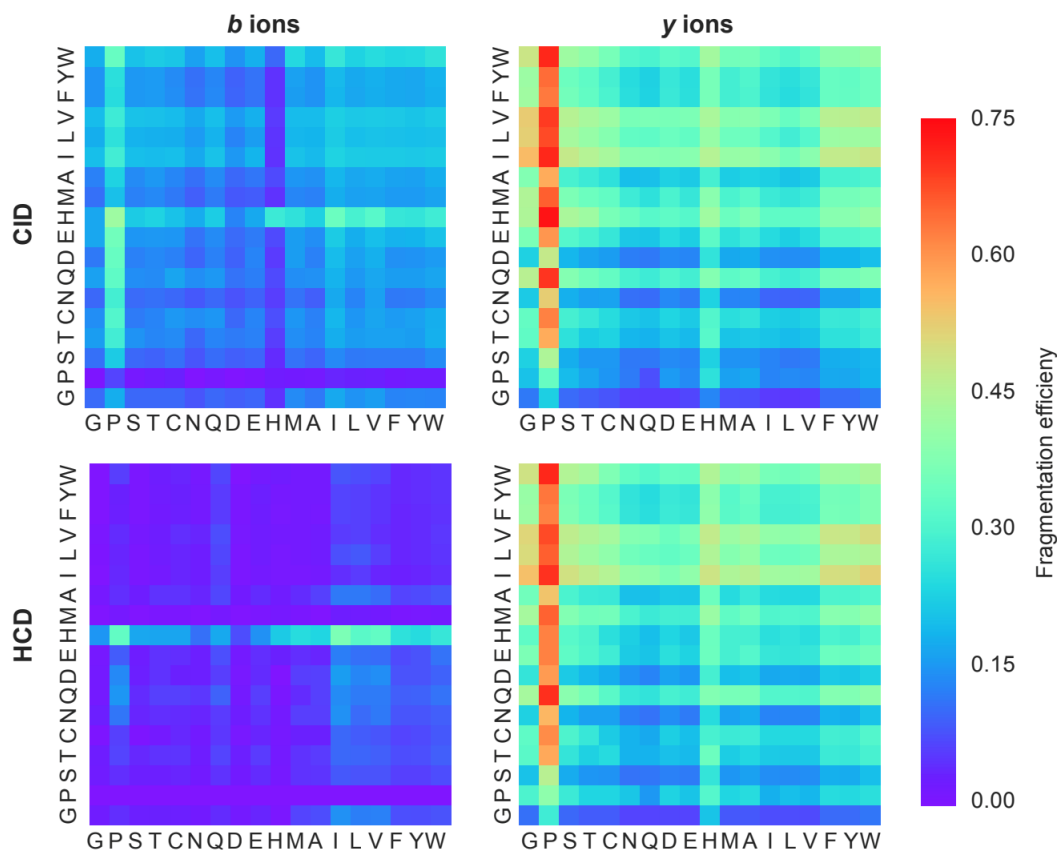


Figure 3.11: Fragmentation efficiency between residue pairs. Fragmentation efficiencies were calculated for b- and y- fragment ions generated between [X] (y-axis) and [Z] (x-axis) residues in A-AA-[X]-[Z]-A-A-A-R peptides, for both CID and HCD fragmentation. The fragmentation efficiency is defined as a predicted peak intensity, normalized by the total sum of peak intensities of the same ion type. Similar to the previous findings [Shao et al., 2014], our model reports significantly higher fragmentation efficiency between [X]-Pro residues (where [X] can be any other residues), for both y- and b-ion types. The model also correctly predicts less efficient fragmentation between [X]-[Z] residue pairs where [X] is a hydrophobic residue. Furthermore, the model also correctly identified an increased fragmentation efficiency for b-ions between His-[Z] pairs.

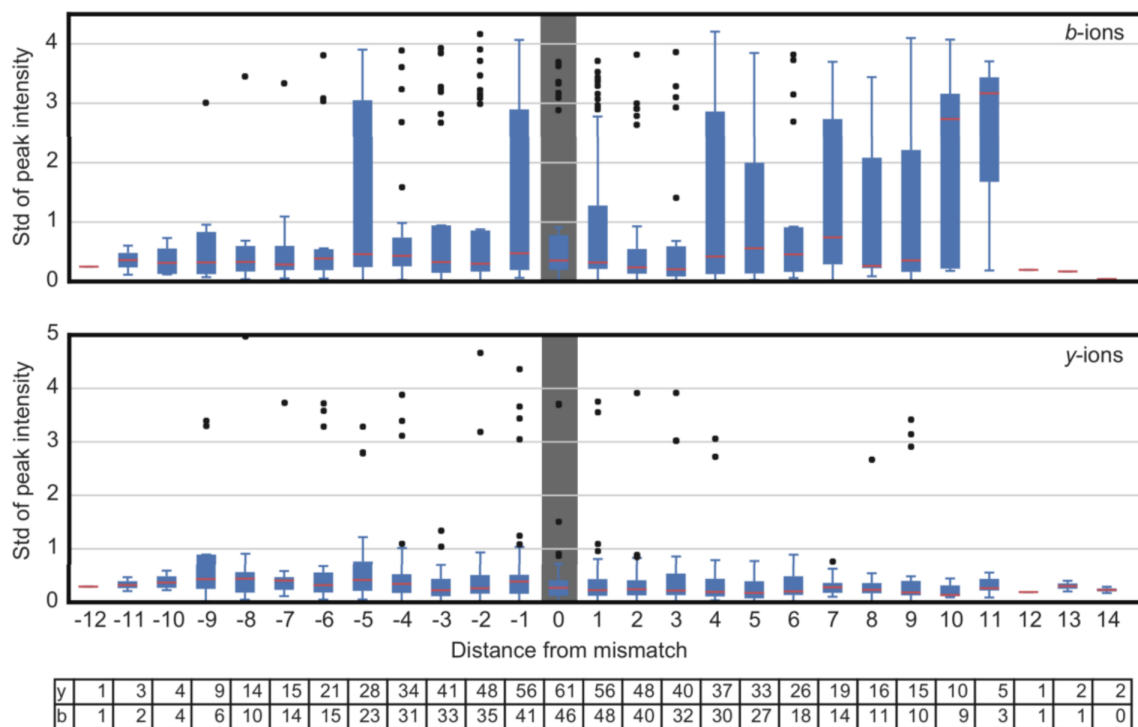


Figure 3.12: Long-range interactions in the training data. In our dataset we found 63 peptide pairs with a single residue mismatch, but the same precursor charge, fragmentation, and mass analyzer types. We evaluated the effect of a position more than one residue away from the site of fragmentation by measuring standard deviations in peak intensities across all 63 peptide pairs (for example, between the y_3 -ion intensity in spectra for both peptides in the pair). The box plots show distributions of peak intensities when looking at positions that are N residues (x -axis) away from the site of the mismatch. Each box extends from the lower to upper quartile values of the data, with a red line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Data points beyond these ranges are considered outliers, and are plotted as diamonds. The numbers of pairs (N) for each box are shown in the table below the box plots.

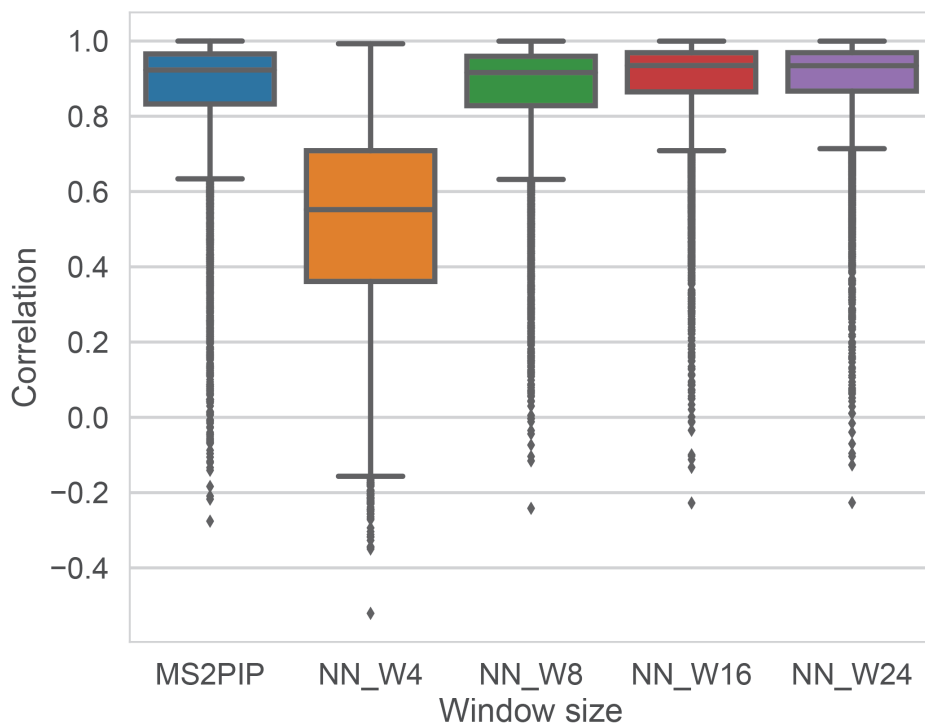


Figure 3.13: Window size dependent performance of the sliding window approach. Distribution of PCCs for MS2PIP and for the window-based neural network model for sliding window sizes of 4, 8, 16 and 24 residues on CID+2 model. The predictive performance increases monotonically with window size. The window-based neural network model outperforms MS2PIP at sufficiently large window sizes. by deleting or overwriting this text; captions may run to a second page if necessary. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Values beyond these ranges are considered outliers, and are plotted as diamonds. The boxplots contain 9214 unique PSMs from the ProteomeTools dataset (PXD004732).

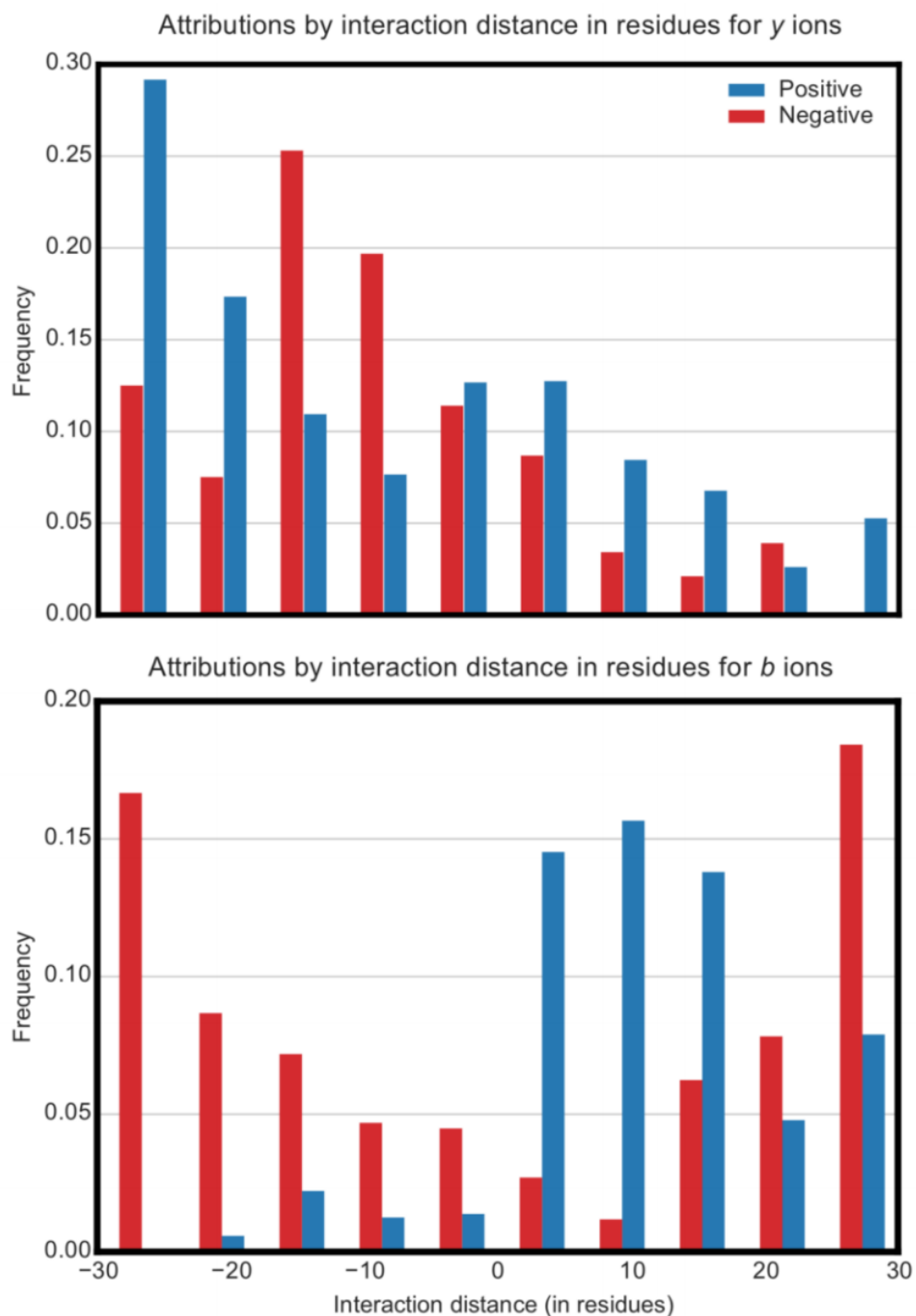


Figure 3.14: Interpretation of the DeepMass:Prism model. Distributions of distance between peak intensity and major attributions are shown for y (top) and b (bottom) ions. Major attributions are attribution values (heat map pixels in top panels of Fig. 3.4) with absolute values greater than or equal to 0.7. Directional distances are in amino-acid residue counts from the cleavage site.

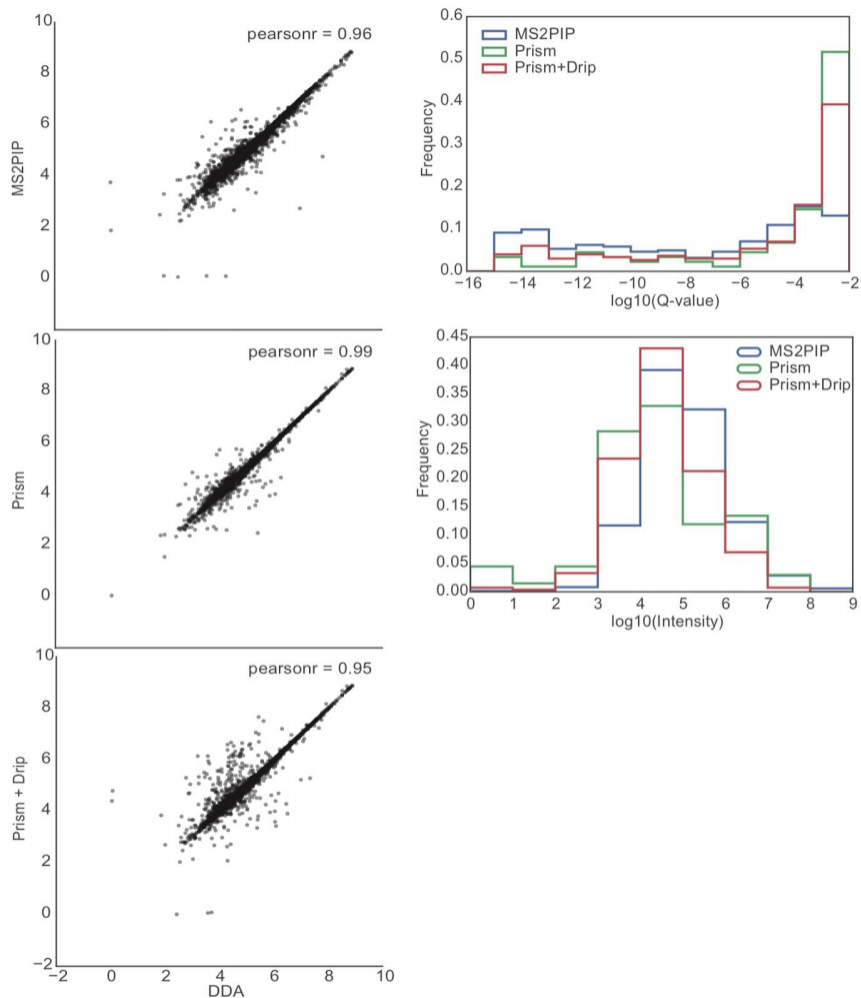


Figure 3.15: Application to spectral library generation for DIA data analysis. Left: The scatter plots show the correlation between log-transformed peptide peak intensities as identified using spectral libraries based on DDA experiments and on MS2PIP (top), DeepMass:Prism (middle), or DeepMass:Prism+Drip (bottom). Pearson correlation coefficients are shown in the upper-left corner of each correlation plot. Right (top): The histograms show the distributions of logtransformed Q-values for peptides detected in DDA-based spectral library searches, but that were missed in searches with spectral libraries based on MS2PIP, DeepMass:Prism, and DeepMass:Prism+Drip. As illustrated, a majority of the missed peptides have a Qvalue worse than 10^{-3} . Right (bottom): The histogram shows the distributions of log-transformed peak intensities for peptides detected with the DDA-based spectral library, but missed in results with spectral libraries based on MS2PIP, DeepMass:Prism, and DeepMass:Prism+Drip. The numbers of data points (N) in these plots are based on the number of identified peptides: 5,248, 5,181, 3,976 for the DeepMass:Prism-, DeepMass:Prism+Drip-, and MS2PIP-based spectral libraries, respectively. The Q-values in the upperright chart are as reported by Spectronaut.

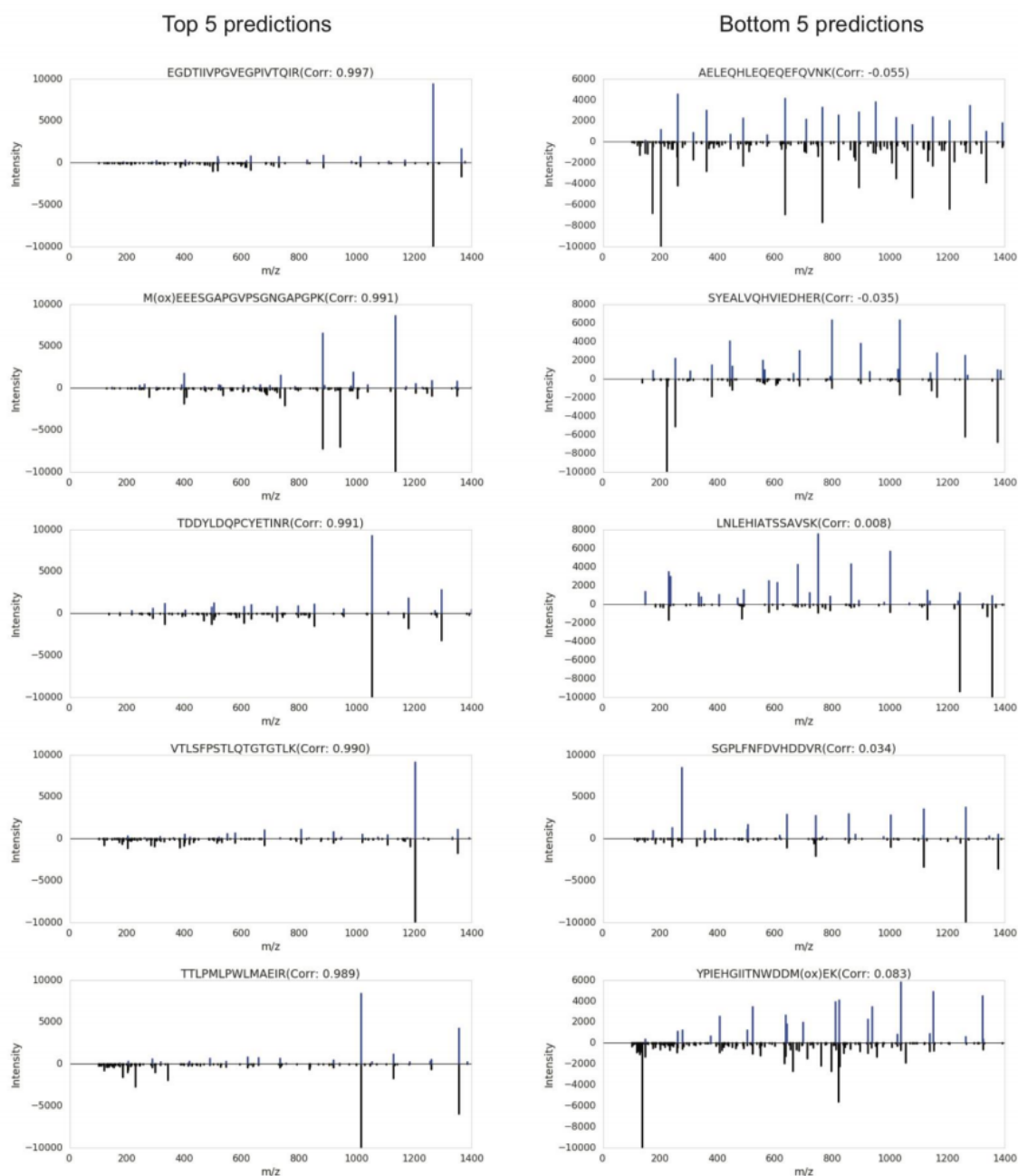


Figure 3.16: Top and bottom 5 predictions by DeepMass:Prism from our testing set. The five best and worst predictions are shown. For each example, the predicted MS/MS spectra and the actual MS/MS spectra are shown above and below the x-axis, respectively. Badly predicted spectra tend to have highly intense fragments at the beginning or the end of a series. The correlation analyses with the peptides from the testing sets were only performed once.

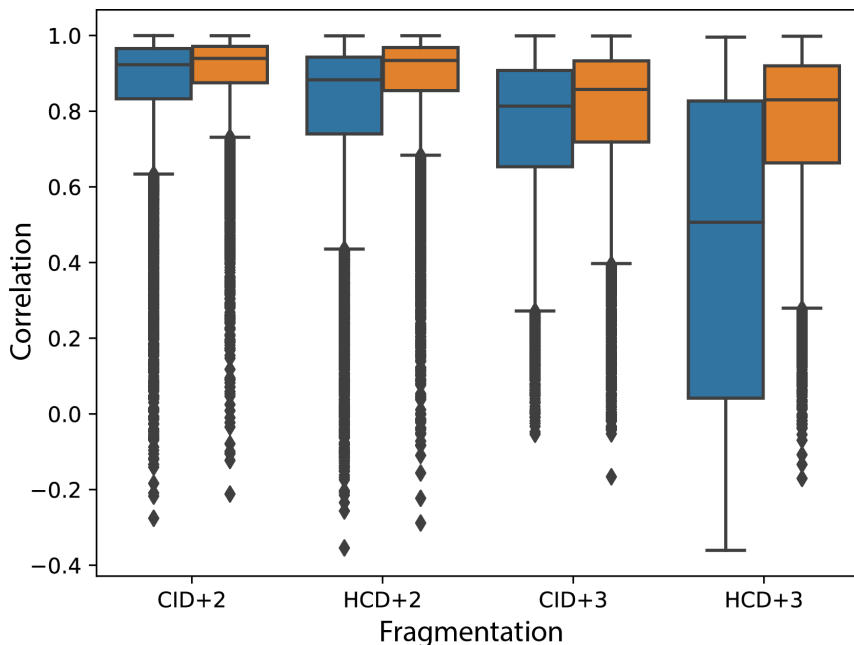


Figure 3.17: Performance of the sliding window approach split by precursor charge and fragmentation type. Distribution of PCCs for the sliding window-based neural network model with a window size of 24 (orange) and MS2PIP (blue) for comparison. The neural network approach results in better performance in all subclasses. Each box extends from the lower to upper quartile values of the data, with a line at the median. The whiskers extend to 1.5-multiple past the interquartile range between the low and high quartiles. Values beyond these ranges are considered outliers, and are plotted as diamonds. The boxplots contain 9,214, 9,317, 6,540 and 6,773 unique PSMs, respectively, for CID+2, HCD+2, CID+3 and HCD+3 test data from the ProteomeTools dataset (PXD004732).

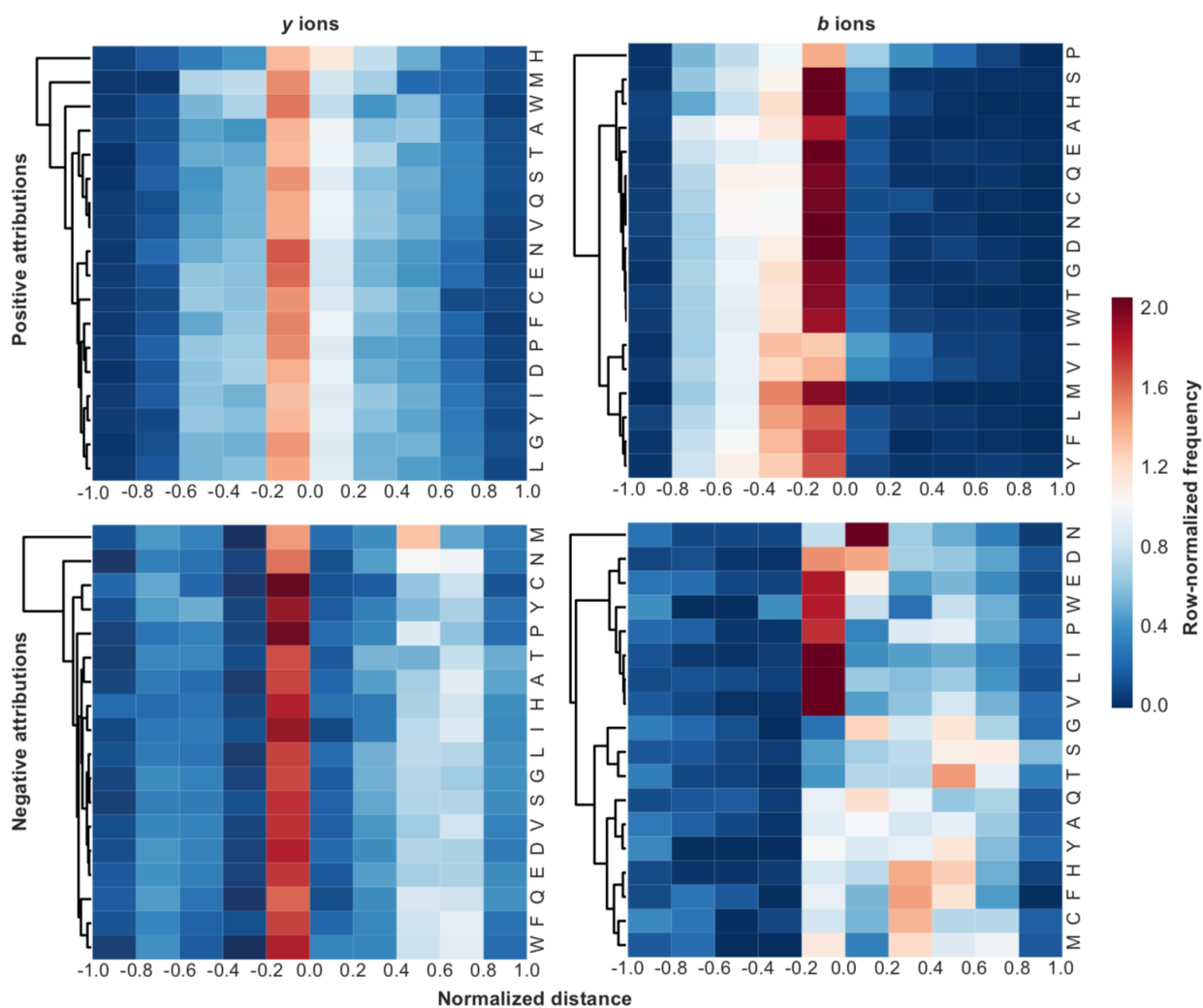


Figure 3.18: Per-residue distances of major attributions. Each row represents a single histogram of the type displayed in Fig. 3.3, on a per-residue basis. Specifically, each row represents the distribution of distances over which a given residue has a strong influence (absolute attribution value ≥ 0.7) on a peak's predicted intensity. Each row is frequency-normalized such that the area under the distribution sums to 1.0. Rows of each set of distributions were clustered via single linkage of PCC. Most rows resemble the overall trend, and notable exceptions are discussed in the text. The distributions in the heatmaps are based on 1000 randomly selected peptide sequences.

3. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis

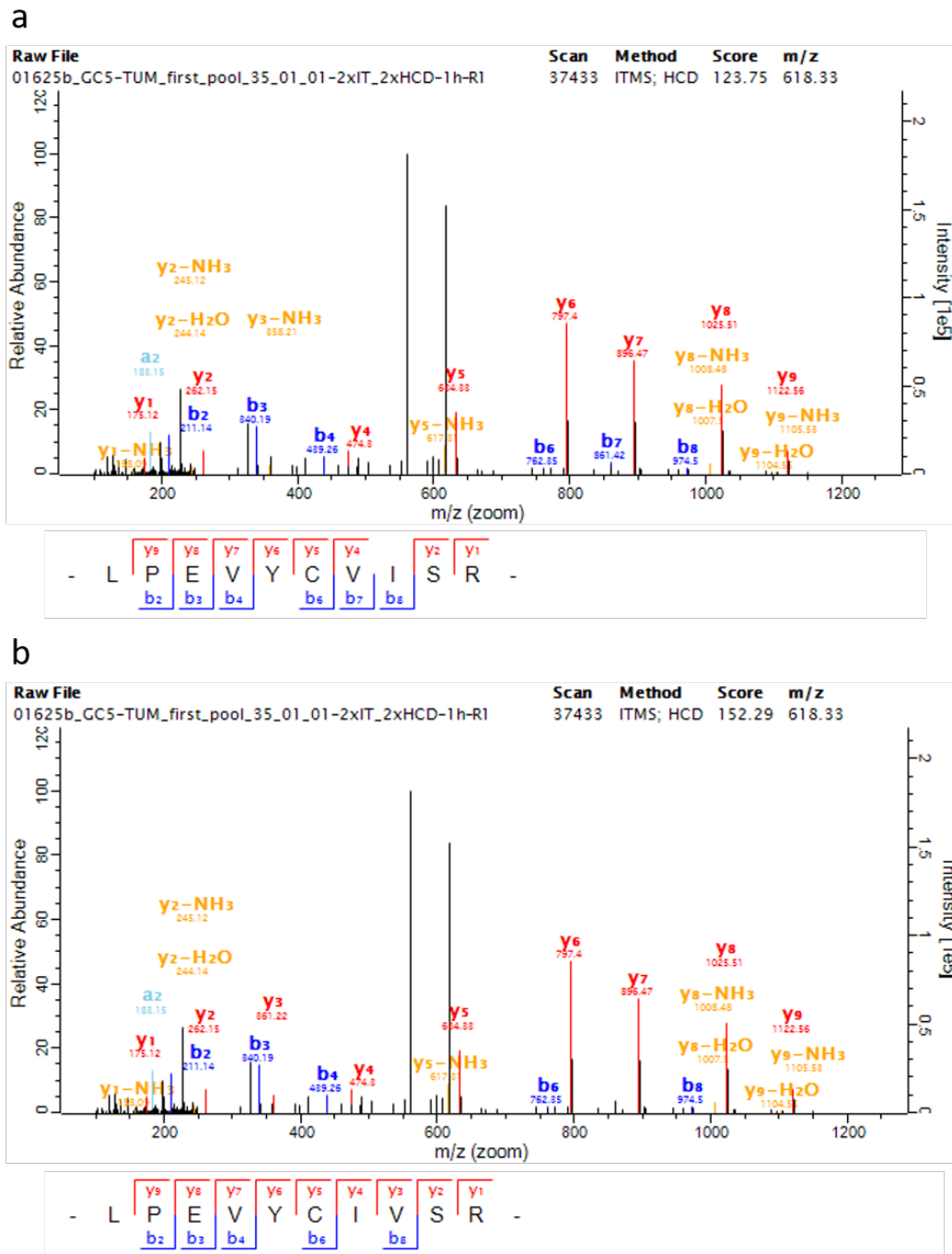


Figure 3.19: Example I for correct PSM recovery by intensity prediction. Including intensity predictions for theoretical spectra in the Andromeda score calculation enables the correct PSM to move from the second best score to the top scoring position. For example, before considering intensity predictions, the sequence of two adjacent amino acids is incorrect (a), but when including intensity information in the score, the sequence with the correct order becomes the highest-scoring PSM (b).

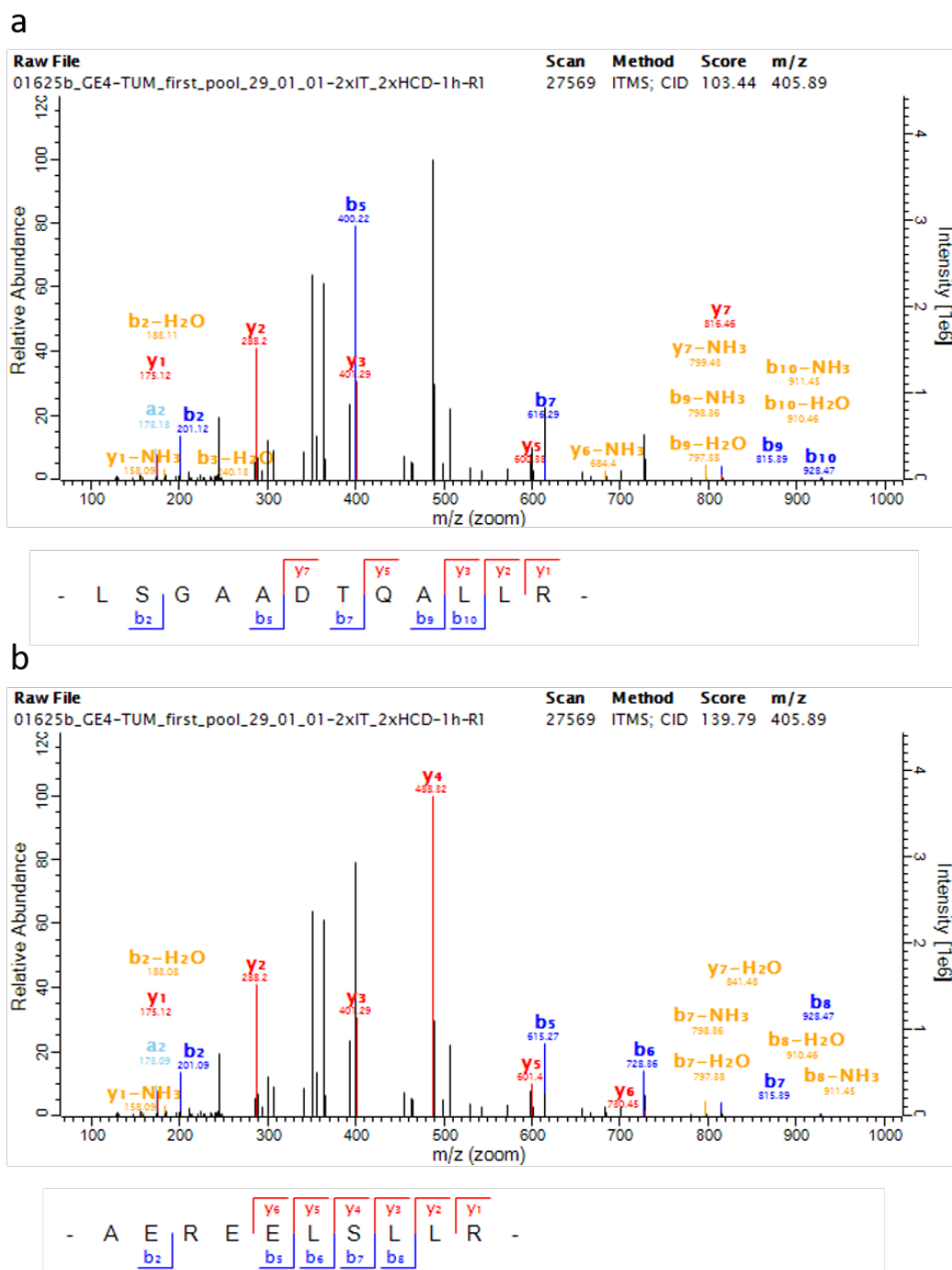


Figure 3.20: Example II for correct PSM recovery by intensity prediction. Including intensity predictions for theoretical spectra in the Andromeda score calculation enables a completely different peptide sequence to become the top-scoring PSM. For example, before considering intensity predictions, the peptide sequence matched to a particular MS/MS spectrum was low but passing (a); however, after including intensity predictions, a new sequence with more matches to higher intense peaks and a more complete ion-series identification was obtained (b).

Species	PRIDE ID	Instrument
<i>Homo sapiens</i>	PXD003668	Q Exactive
	PXD004977	Q Exactive Plus
	PXD001608	Q Exactive
	PXD004732	Orbitrap Fusion ETD
	PXD002549	Q Exactive
<i>Mus musculus</i>	PXD001250	Q Exactive
	PXD002452	LTQ Orbitrap Velos
	PXD005485	LTQ Orbitrap Velos
	PXD001636	LTQ Orbitrap Velos
<i>Saccharomyces cerevisiae</i>	PXD004087	Q Exactive
	PXD001695	Q Exactive, LTQ Orbitrap
	PXD000955	LTQ Orbitrap
	PXD005041	Q Exactive
	PXD003472	Q Exactive, LTQ Orbitrap Velos
<i>Arabidopsis thaliana</i>	PXD001865	LTQ Orbitrap
	PXD002607	LTQ Orbitrap Velos
	PXD002908	LTQ Orbitrap
	PXD003002	LTQ Orbitrap Velos
	PXD004276	LTQ Orbitrap
<i>E.coli</i>	PXD003189	LTQ Orbitrap
	PXD003583	Q Exactive
	PXD004321	Q Exactive
	PXD005126	Q Exactive
<i>E.coli</i>	PXD002912	LTQ Orbitrap Velos
	PXD002785	Q Exactive

Table 3.1: PRIDE datasets used for the DeepMass:Prism model

Chapter 4

Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

Petra Gutenbrunner^{1,2}, Pelagia Kyriakidou¹, Frido Welker & Jürgen Cox^{1,4*}

submitted to bioRxiv, MCP

¹Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.

²Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 Munich, Germany.

³Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Øster Voldgade 5, 1350 Copenhagen, Denmark.

⁴Department of Biological and Medical Psychology, University of Bergen, Jonas Liesvei 91, 5009 Bergen, Norway.

*email: cox@biochem.mpg.de

4.1 Abstract

We describe MaxNovo, a novel spectrum graph-based peptide de-novo sequencing algorithm integrated into the MaxQuant software. It identifies complete sequences of peptides as well as sequence tags that are incomplete at one or both of the peptide termini. MaxNovo searches for the highest-scoring path in a directed acyclic graph representing the MS/MS spectrum with peaks as nodes and edges as potential sequence constituents consisting of single amino acids or pairs. The raw score is a sum of node and edge weights, plus several reward scores, for instance, for complementary ions or protease compatibility. For search-engine identified peptides, it correlates well with the Andromeda search engine score. We use a particular score normalization and the score difference between the first and second-best solution to define a combined score that integrates all available information. To evaluate its performance, we use a human cell line dataset and take as ground truth all Andromeda-identified MS/MS spectra with an Andromeda score of at least 100. MaxNovo outperforms

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

other software in particular in the high-sensitivity range of precision-coverage plots. We also identify incomplete sequence tags and study their statistical properties. Next, we apply MaxNovo to ion mobility-coupled time of flight data. Here we achieve excellent performance as well, except for potential swaps of the two amino acids closest to the C-terminus, which are not well resolved due to the low end of the mass range in MS/MS spectra in this dataset. We demonstrate the applicability of MaxNovo to palaeoproteomics samples with a Late Pleistocene hominin proteome dataset that was generated using three proteases. Interestingly, we did not use any machine learning in the construction of MaxNovo, but implemented expert domain knowledge directly in the definition of the score. Yet, it performs as good as or better than the leading deep learning-based algorithm.

4.2 Introduction

De-novo sequencing [Taylor and Johnson, 1997, Dančák et al., 1999] has the goal of determining the amino acid sequence of a peptide directly from its tandem mass spectrum without making use of a peptide sequence database. Potential fields of applications in proteomics are widespread and include the study of ancient samples [Cappellini et al., 2018], identification of HLA peptides [Ternette et al., 2016, Khodadoust et al., 2017, Laumont et al., 2018], monoclonal antibody sequencing [Bandeira et al., 2008b, Tran et al., 2016], and detection of endogenous, non-ribosomal peptides [Bandeira et al., 2008a]. From early on it was found advantageous to represent the spectrum as a graph [Dančák et al., 1999] in which the fragment peaks correspond to nodes which are connected by edges, whenever the mass difference is interpretable as a mass of one or more amino acids, and hence the connected peaks are adjacent in an ion series. Optimal paths in these graphs are then determined with diverse computational methods including dynamic programming [Chen et al., 2001], hidden Markov models [Fischer et al., 2005] and probabilistic network modeling [Frank and Pevzner, 2005]. More recently, deep learning was applied to the problem [Tran et al., 2017, Karunratanakul et al., 2019, Tran et al., 2019, Yang et al., 2019, Qiao et al., 2021], which has led to the best performing methods to date.

The MaxNovo algorithm described in this manuscript makes use of the spectrum graph representation as well. It performs an exhaustive search for the best path using a cost function that is designed in a way such that the resulting score is similar to the Andromeda18 search engine score, in cases where the database search led to an identification as well. MaxNovo is fully integrated in MaxQuant [Cox et al., 2011, Sinitcyn et al., 2018b] which allows to make use of results obtained from the search engine-based workflow, as for instance the accurate calibrated precursor masses and the three-dimensional MS1 features and isotope patterns. On an Orbitrap HeLa benchmark dataset, we show that MaxNovo performs well in terms of total number of correct de-novo identifications with controlled number of wrong identifications. It outperforms other software, as, for instance PEAKS [Ma et al., 2003], in the high specificity region. Further studies on tims-TOF Pro data, as well as an ancient proteomics application, demonstrate MaxNovo's applicability to diverse mass spectrometry proteomics data sets.

4.3 Methods

4.3.1 HeLa dataset

Mass spectrometric raw data from a HeLa cell line was obtained from the PRIDE repository PXD006932 [Kelstrup et al., 2018]. What we refer to as ‘single-shot’ dataset are the three biological replicates measured on a Q Exactive HF-X using an Orbitrap resolution setting of 15000 (Thermo Fisher Scientific, Bremen, Germany). ‘Fractionated data’ are the 46 fractions obtained by Q Exactive HF-X using an Orbitrap resolution setting of 7500. The MaxQuant runs for both the single shot and the fractionated data were done against the *H. sapiens* proteome (UP000005640), including isoforms, and downloaded from UniProt on 07/04/2021. In total six separate MaxQuant runs were performed for the single shot data. One run with all five scores (complement score, protease score, water-loss score, ammonia-loss score, a₂ score) contributing to the calculation of the raw score and other five runs with each one of the score contributions omitted. For the fractionated data only one MaxQuant run was made with the default calculation of the raw score including all five scores. Default values for all parameters were used in the MaxQuant analysis.

4.3.2 Tims-TOF pro dataset

HeLa DDA data acquired on a tims-TOF Pro instrument was downloaded from the PRIDE repository PXD022582 [Sinitcyn et al., 2021] and was searched with MaxQuant against the *H. sapiens* proteome (UP000005640) including isoforms downloaded from UniProt on 07/04/2021. In the MaxQuant analysis for all parameters default values were used. To switch on tims-TOF analysis the parameter ‘Type’ was set to TIMS-DDA

4.3.3 Ancient dataset

Mass spectrometric raw data from a hominin bone specimen was obtained from the PRIDE repository PXD018264 [Lanigan et al., 2020]. The three biological replicates from the samples digested with the proteases trypsin, chymotrypsin, and Glu-C, were grouped per protease and searched in MaxQuant in three separate runs using different proteome databases. One MaxQuant run was searched against the *H. sapiens* proteome (UP000005640) downloaded from UniProt on 05/08/2021. Two subsequent MaxQuant runs were made, one against the *Gorilla gorilla gorilla* proteome (UP000001519) and one against the *Pan troglodytes* (UP000002277), both downloaded from UniProt on 04/08/2021. All three Uniprot databases contain one protein sequence per gene. In the MaxQuant analysis the default settings were applied except for the following settings: No fixed modification was selected and as variable post-translational modification oxidation of Methionine, deamidation of asparagine and glutamine, hydroxylation of proline, and carbamidomethylation of cysteine. The minimum peptide length was set to eight (default value is seven) and the minimum score for unmodified peptides was set to 40 (default is 0). The Glu-C specificity was configured to also include C-terminal cleavage after glutamine (Q), alongside C-terminal cleavage of glutamic acid (E) and aspartic acid (D) as it was defined in the original publication.

4.3.4 Pre-processing of MS/MS spectra for MaxNovo search

All MS/MS spectra are subject to the standard pre-processing that is also applied before submitting spectra to the Andromeda search. Isotope patterns are found based on correlation to the averagine model [Senko et al., 1995]. In case at least two peaks were put together in an isotope pattern, the corresponding monoisotopic peak replaces them with the summed intensity of the member peaks. In case the charge determined from the isotope pattern is larger than one, the peak is added as a singly charged version. If more than one charge state was found for a fragment (within a user definable tolerance), these are summed up into a single peak.

4.3.5 NOVOR data preparation and analysis

The single-shot and fractionated HeLa raw files were converted to the mgf file format with ProteoWizard [Chambers et al., 2012] version 3.0.11579 for processing with Novor [Ma, 2015] (Version v1.06.0634, Java SDK 16). The protease is trypsin, which is the only supported option. HCD was selected as fragmentation method and FT as mass analyzer. The precursor error tolerance was set to 15 ppm and the error tolerance for fragment ions to 0.02 Da.

4.3.6 PEAKS data analysis

Both HeLa datasets, single-shot and fractionated, were analyzed by the de-novo algorithm in the PEAKS software [Zhang et al., 2012] (version PEAKS X Pro, Peaks Studio 10.6 build 20201221). For the de-novo search, the instrument 'Orbitrap (Orbi-Orbi)' was selected which has set by default the parent mass error tolerance to 15 ppm and the fragment mass error tolerance to 0.02 Da. Trypsin was defined as digestion enzyme and its cleavage specificity was configured that it cleaves after arginine and lysine also if a proline follows. As fixed modification carbamidomethylation of cysteine was included and as variable modifications oxidation of methionine and acetylation of protein N-term. For the maximal number of variable PTMs the default value of three is selected. We increased the number of reported candidates per spectrum from five up to ten. The parameter "Feature association for chimera scans" in the data-refinement section was deactivated, since co-eluting peptides are not within the scope of our benchmark setup. For the top scoring de-novo sequence comparison, the de-novo sequence with the highest de-novo score was selected. In case of multiple de-novo sequences having the same top de-novo score, all sequences are taken as top sequence.

4.3.7 Benchmark based on the HeLa datasets

For both HeLa datasets, single-shot and fractionated, as ground truth the identified MS/MS spectra based on the Andromeda database search engine at 1% PSM and 1% Protein FDR (default settings) were taken which were further filtered by an Andromeda score greater than or equal to 100. All isoleucine in the database sequence as well as de-novo sequence were replaced by leucine. The de-novo identified MS/MS spectra of each of the tools MaxNovo, PEAKS and Novor are joined based on the raw file name and scan number information.

4.3.8 BLAST search

The de-novo sequence identifications, which were not identified by the Andromeda database search engine, were validated by an iterative local BLAST (version 2.11) search. Only completely de-novo sequenced MS/MS spectra were considered and filtered to have a combined score of at least 91.715. Next, for each de-novo sequence all combinations, up to a maximal number of 300, were generated and submitted to four separate BLAST searches each using a different database. For the first three searches blastp was performed by setting the following parameters: `window_size = 40`, `word_size = 2`, `evaluator = 1000`, `max_hsps = 1`, `-threshold 11` and against one of the following databases: 1) Swiss-Prot filtered by only human, 2) Swiss-Prot filtered by only *Bos Taurus* and 3) Swiss-Prot. The fourth search was a tblastn search against a manually generated HeLa nucleotide database based on RNA-seq data [Zarnack et al., 2013, Seo et al., 2016] (ERR127306.1.fastq, ERR127306.2.fastq, ERR127307.1.fastq, ERR127307.2.fastq). All BLAST results were filtered to have for each submitted de-novo sequence identification the best BLAST hit. In case a de-novo sequence identification was validated by multiple database hits the following database priority was applied: 1) Swiss-Prot human, 2) HeLa RNA-seq, 3) Swiss-Prot *Bos taurus* and 4) Swiss-Prot.

4.3.9 Software and data availability

MaxNovo is integrated into the MaxQuant software from version 2.0.3.0 onward and can be downloaded from <https://www.maxquant.org/maxquant/>. A user guide on how to run MaxNovo in MaxQuant is provided as part of the 4.8 Supplementary Information, Supplementary Table 4.2 contains a list of MaxQuant parameters relating to MaxNovo with their explanations. Supplementary Table 4.3 describes all new MaxNovo associated columns in the 'msmsScans.txt' output file. The re-analyzed MS proteomics data have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository.

4.4 Results and discussion

MaxNovo spectrum graph

Input for the MaxNovo algorithm are individual MS/MS spectra in which the fragment peaks are de-isotoped and transformed to charge state one (Fig. 4.1a), in the same way as MS/MS spectra are prepared in MaxQuant for the peptide database search with Andromeda 18 (see 4.3 Methods). Hence, the MaxNovo algorithm assumes that fragment ions are singly charged. The precursor masses are derived from three-dimensional features, spanned by m/z , retention time and signal intensity, after the standard nonlinear mass calibration in MaxQuant has been applied, leading to high accuracy mass estimates for the precursor ions. For each MS/MS spectrum, we construct a graph with the peaks in the spectrum as nodes (Fig. 4.1b). The nodes that correspond to peaks in the MS/MS spectrum we call 'internal nodes'. Four additional nodes are added to the graph, which represent the N-terminus and the C-terminus each twice, once for being reached at low masses, corresponding to the beginning of an ion series, and once for being reached at high masses, corresponding to the end of an ion series.

We then build a directed acyclic graph (DAG) by placing edges between vertices from lower to higher mass. This is based on the 20 common amino acids, and modified versions of some amino acids, which can be specified by the user as fixed and variable modifications. An edge is placed between two internal nodes if their mass difference fits a single amino acid mass or the sum of two amino acid masses (Fig. 4.1c). Mass steps that have equal mass or are so close in mass that they are not discernible based on the data are grouped together and treated as the same amino acid (as, for instance leucine and isoleucine). Similarly, as for internal nodes, for the termini we connect the beginning node with internal nodes that can be reached with a single or a double amino acid step, considering the appropriate terminal masses for the y or b ion series. It is only known to which ion series (e.g. y or b) a path through the network belongs by arriving at one of the terminal nodes. By default, we require a mass difference between two edges to match with a maximal error of 25 ppm. Then weights are assigned to edges and nodes as defined in the next subsection. In this graph, we perform an exhaustive search for the path with the highest raw score (Fig. 4.1d) based on a recursive algorithm. Either the best path represents the y or the b series, depending on which of the two achieves a better score. In case the path with the highest raw score does not reach the termini at one end, a second search for a best path is performed, with the constraint to fill the missing mass. This is necessary for the case that y and b series are not overlapping.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

76

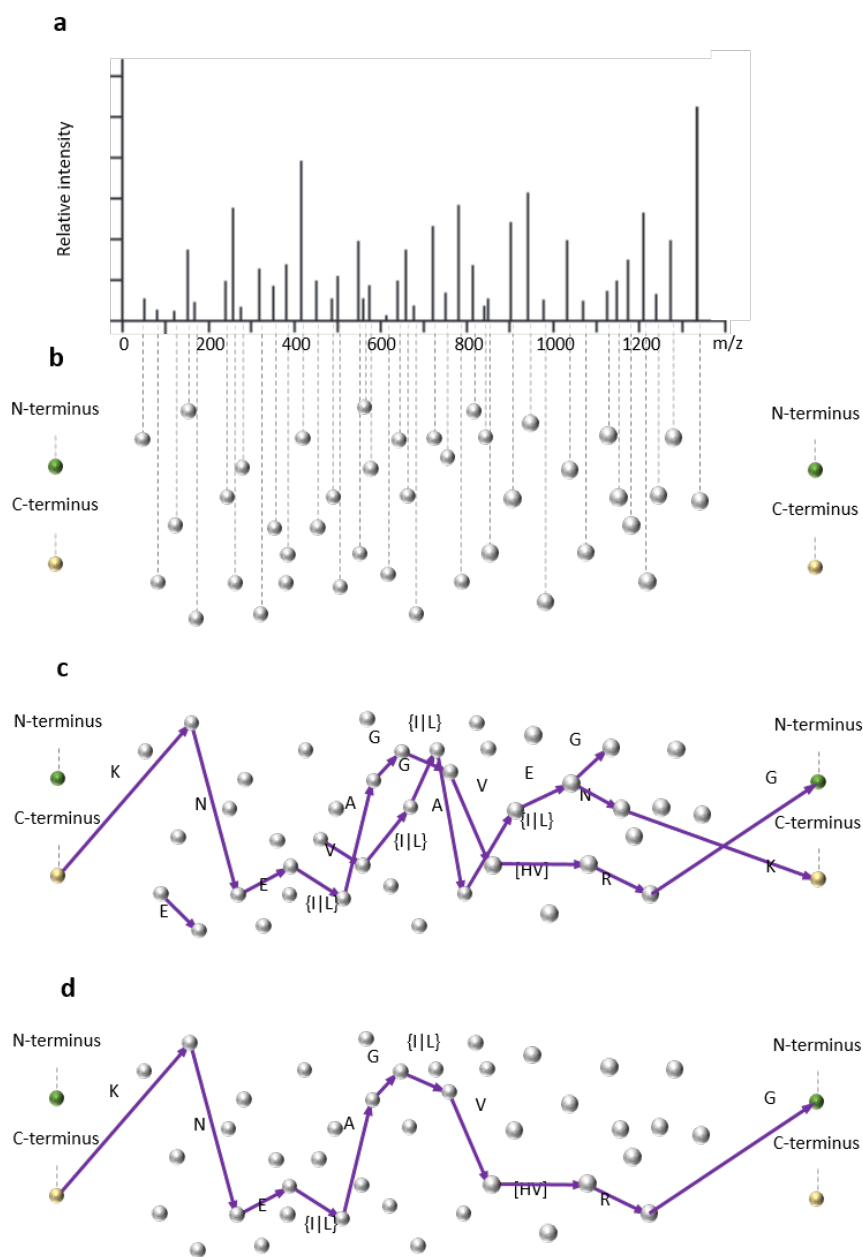


Figure 4.1: Schematics of the MaxNovo spectrum graph. **a.** Starting point for the algorithm is the processed MS/MS spectrum with de-isotoped fragments transferred to charge equals one, and the accurate precursor mass obtained from three-dimensional MS1 features after their nonlinear recalibration. **b.** Each peak in the processed spectrum becomes a potential 'internal node' of the spectrum graph. Four 'terminal nodes' are added, corresponding to N- or C- terminus reached at low or high masses. **c.** A directed edge from low to high mass is put between two nodes, whenever their mass difference fits single or paired amino acids. **d.** The path with the highest raw score is found in an exhaustive search. It may start and end with terminal or external nodes. The raw score is the sum of weights of traversed nodes and edges plus several reward scores as described in the main text.

MaxNovo raw score

Each path through the spectrum DAG gets a raw score assigned. The path in this graph with the highest raw score is defined to be the result of the de-novo search for this MS/MS spectrum and this highest score is the MaxNovo raw score for this spectrum. Either the best path can connect both termini, in which case the complete peptide has been sequenced, or it can be incomplete at one or both of the termini. Hence, the MaxNovo algorithm is a combination of complete de-novo sequencing algorithm and sequence tag finding algorithm within one unified scoring scheme. The raw score assigned to each path consists of six contributions:

$$\text{rawscore} = \text{direct path score} + \text{complement score} + \text{protease score} + \\ a_2 \text{ score} + \text{water loss score} + \text{ammonia loss score}$$

which correspond to specific rewards or penalties. The *direct path score* is the sum of scores defined on the edges and nodes that constitute the path and is scoring one main ion series, which is supposedly the one that contributes most to the identification. Therefore, the path represents either part of, or the complete b-series or part of, or the complete y-series. It is not supposed to mix contributions from N-terminal or C-terminal series, since it receives contributions from a path consisting of steps that correspond to single amino acid or amino acid pair mass differences within one ion series. Two-amino-acid steps are allowed in order to be able to have complete solutions connecting the termini also when one or several peaks in a series are missing. By default, we allow up to two two-steps in a path. Each node visited by the path contributes $-\log_{10}((g + 1) / 100)$ to the score, where g is the number of peaks in a 100 Da interval centered on the current peak which have a higher intensity than the current peak. This contribution assigns a higher reward to peaks with a high intensity relative to other peaks in the surrounding 100 Da interval. A traversed edge contributes $-\log_{10}(s)$, where s is the number of potential steps of the type that was actually taken that could have been taken from the current node. This can have two values: if a single step was taken, then it is the total number of single steps available, and if a two-step was taken, this is the total number of two-steps. Essentially, this is a penalty on how many mass differences were tried in order to reach the next node. For instance, there are more potential two-steps than there are single-steps, which get down-weighted accordingly. Overall, a path length dependent score contribution of $-\log_{10}(n)$ is added to the total direct path score, where n is the total number of steps in the path.

The *complement score* looks for the presence of nodes corresponding to complementary ions to the ions found in the direct path. For instance, if the path describes b-ions, these complementary ions correspond to y-ions that each match with one of the b-ions as a complementary pair. The roles of y and b ions could also be the opposite. For each complementary peak found, a contribution $-\log_{10}((g + 1) / 100) / 4$ is added, similarly as the node weight in the direct path score. Additionally, if a complementary ion is found that resolves the order of the amino acids in a two-step of the direct path a corresponding score contribution is added that equals to the situation as if the resolving peak was found in the direct path. The *protease score* adds a reward in case the path reaches a terminus at which an expectation is met regarding the protease used for digesting proteins to peptides in the sample preparation. For instance, in the case of trypsin, a path reaching the C-terminus

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

78

would be rewarded if the last step were either an arginine or a lysine, in case it is a single amino acid step, or it contains these amino acids, in case it is a two amino acid step. The a_2 score adds a contribution in case the a_2 ion is found. This is only possible in case the sequence path is continued to the N-terminus in order to know that it follows the b-series. The *water loss score* checks for the presence of one of the amino acids D, E, S or T in the path traversed so far. In case any of these is present, it is checked if peaks are present at the expected position(s) for water losses. The *ammonia loss score* similarly checks for peaks at the characteristic masses for ammonia losses based on the presence of K, N, Q or R in the so far identified sequence.

For convenience, we provide brief definitions of these scores and scores defined in subsequent subsections in Table 4.1. In Fig. 4.2 we show a scatter plot of the raw score and the Andromeda score for identified spectra in the single-shot HeLa dataset, which have a Pearson correlation of 0.71, which reduces to 0.60 when no Andromeda score filter is applied. In the construction of the raw score, the main purpose was that it reflects well the total evidence that is present in a spectrum for the peptide sequence that it claims to have found. This is confirmed by the good correlation with the Andromeda score.

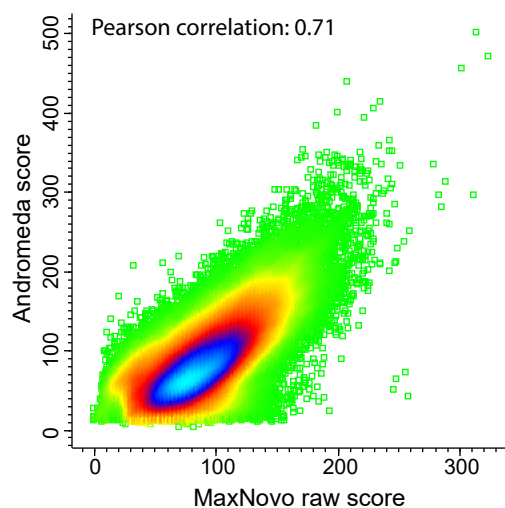


Figure 4.2: Scatter plot of the raw score against the Andromeda score. Data is filtered for Andromeda score >100 . Data points are colored by their density. The Pearson correlation is 0.71.

Normalized score, gap score and combined score

The raw score introduced in the previous section is optimized for finding the best solution, either a sequence tag or a full-length sequence, for a given MS/MS spectrum. However, it is not made for comparing two optimal solutions in two different MS/MS spectra. The ability to compare scores across different spectra is, however, crucial for obtaining confidence in an identification and for estimating the percentage of false positives. To see this in a better way, we define as a dataset with essentially known ground truth all MS/MS spectra in the

HeLa dataset that were identified in the MaxQuant analysis with default protein and PSM level false discovery rates (FDRs) of 1% each, and additionally filter the spectra to have an Andromeda score of at least 100. We consider these spectra as our ground truth and the aim of the de-novo sequencing algorithm is to identify as many peptide sequences of these correctly, if possible the complete sequence for each, and otherwise as many amino acids as possible in a sequence tag. For now, we restrict ourselves to identifying complete peptide sequences.

We will frequently use precision-coverage plots, which are created by ranking the spectra according to a given score. A de-novo identification is counted as correct only if it completely agrees with the Andromeda sequence. If there is only one deviation, the whole peptide counts as incorrect. For a given score threshold we define precision as the number of correctly identified sequences divided by all spectra with a score above the threshold. With coverage, we mean the number of correctly identified spectra above the threshold divided by the total number of spectra in the ground truth. Our definition of coverage measures how many spectra from the whole ground truth have been correctly identified with a complete sequence.

If we calculate such a precision-coverage plot for the raw score defined in the previous subsection (green curve in Fig. 4.3a) we see that the performance is not ideal. In particular, in the high specificity range on the left side, no good precision values are achieved. In order to fix this problem, we define the normalized score, which is the raw score divided by the precursor mass. It is a measure for sequence evidence per length. In particular, in a situation where there is good sequence evidence only in parts of the sequence and lack of it otherwise, the normalized score will not be particularly high. Indeed, the normalized score has better precision-coverage characteristics (purple curve in Fig. 4.3a). As 'complete score', we define it to be equal to the normalized score in case it suggests a full-length peptide sequence and zero otherwise (blue curve in Fig. 4.3a). Another aspect of potential relevance for judging the correctness of a de-novo sequence is how well the second-best solution scores compared to the best. If the score difference between the two highest solutions is small, there is a certain likelihood that the second-best solution is the correct one, while if this gap is large, this strengthens the plausibility of the top-scoring solution. The precision-coverage curve based on this gap score (red curve in Fig. 4.3a) results in a similar area under the curve as the complete score. Since the gap score and the complete score measure different aspects of the identification, it should be beneficial to combine the two scores. Indeed, we define the combined score as the sum of the two ranks of the complete score and the gap score and it achieves the best precision-coverage characteristics of all scores (orange curve in Fig. 4.3a). The combined score is hence the method of choice for finding complete sequences, and it is implicitly meant when referring to the MaxNovo score without further specification.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

80

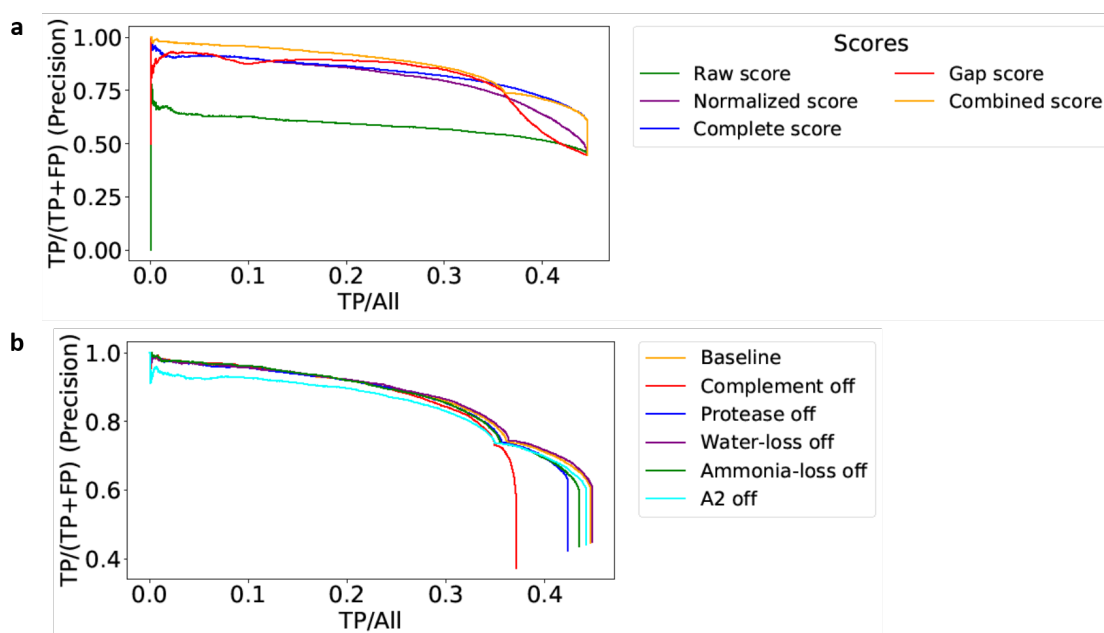


Figure 4.3: Precision-coverage plots of several MaxNovo scores. a. Data was obtained from the single-shot HeLa dataset with all identified MS/MS spectra with an Andromeda score of at least 100. Only full peptide sequences without any amino acid mistake are counted as correct. The curves correspond to raw score (green), normalized score (purple), complete score (blue), gap score (red) and combined score (yellow). **b.** The combined score is shown again as baseline (color). For the other curves, each time one score contribution has been omitted.

Next, we investigated what the benefits are of the individual additive contributions to the raw score is to the overall performance. As the baseline, we take the precision-coverage curve based on the combined score (yellow curve in Fig. 4.3b). Then we remove each of the contributions to the raw score except the direct path score one at a time and record precision-coverage curves for these as well. In the high-specificity region, the strongest effect came from the scoring of the a_2 ion. Conversely, the inclusion of the complementary ion series (the b-series ions in case the main series scored by the direct-path score is the y series) made a big difference in the low specificity region. Water loss, ammonia loss and protease score all contribute in less significant and similar amounts.

Benchmark and comparison to other software

In order to compare the performance of MaxNovo to other software we analyzed the same data with the PEAKS and Novor algorithms (see 4.3 Methods). Fig. 4.4 shows precision-coverage plots for all three programs. While Novor is less performant, the other two are close in performance. Up to coverage 0.35, the curve for MaxNovo is on top. The PEAKS curve reaches farther to the right end, meaning it finds slightly more sequences with low certainty and the respective areas under the curves are very similar for PEAKS and MaxNovo. We conclude that these two software platforms are showing comparable performances with MaxNovo performance being better in the higher specificity region.

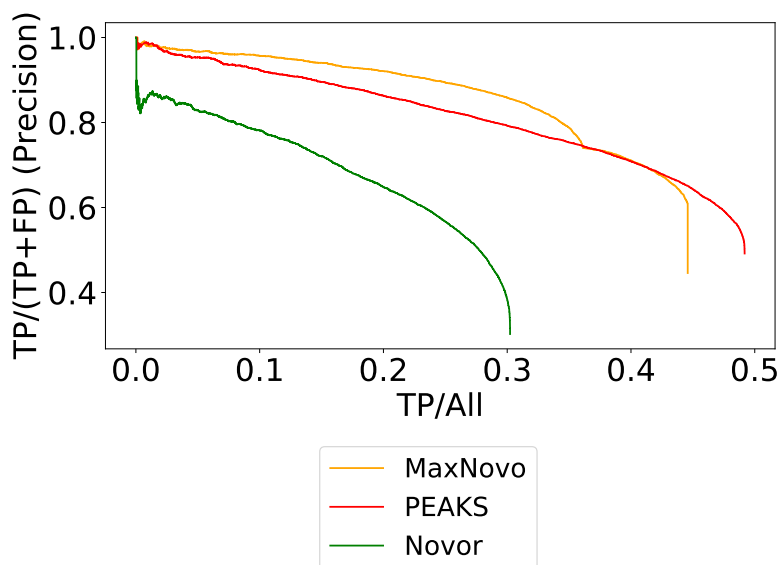


Figure 4.4: Performance comparison of the combined score. Precision-coverage curve comparing the combined score of MaxNovo with PEAKS and Novor.

Incomplete sequences

To measure the performance on incomplete sequences, we recorded precision-coverage plots using the normalized score and we counted as correct, when the predicted sequence agrees completely with the Andromeda sequence or a sub-sequence of it (Fig. 4.5a). This we do for the single shot and the fractionated HeLa datasets separately. We find that we obtain curves that are similar to each other above a precision of 0.8 and are deviating below that. Next, we determine what the expected precision of the results is at a given normalized score threshold (Fig. 4.5b). The curves for fractionated and single-shot data agree well in particular at larger score values, which indicates that the dependency of the precision on the score could be generalizable. For instance, a precision of 0.8 corresponds to a normalized score threshold of approximately ten.

Fig. 4.5c shows histograms of N- and C-terminal missing masses, having medians of 692.3 and 258.1 Da, respectively in the fractionated dataset. The y-axis of the histogram is in logarithmic scale and one can see that higher masses are much less frequent. A density scatter plot of peptide sequence length vs. tag length (Fig. 4.5d) is dominated by the values on the diagonal corresponding to full-length sequences. The distribution of partial sequences has a median length of twelve.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

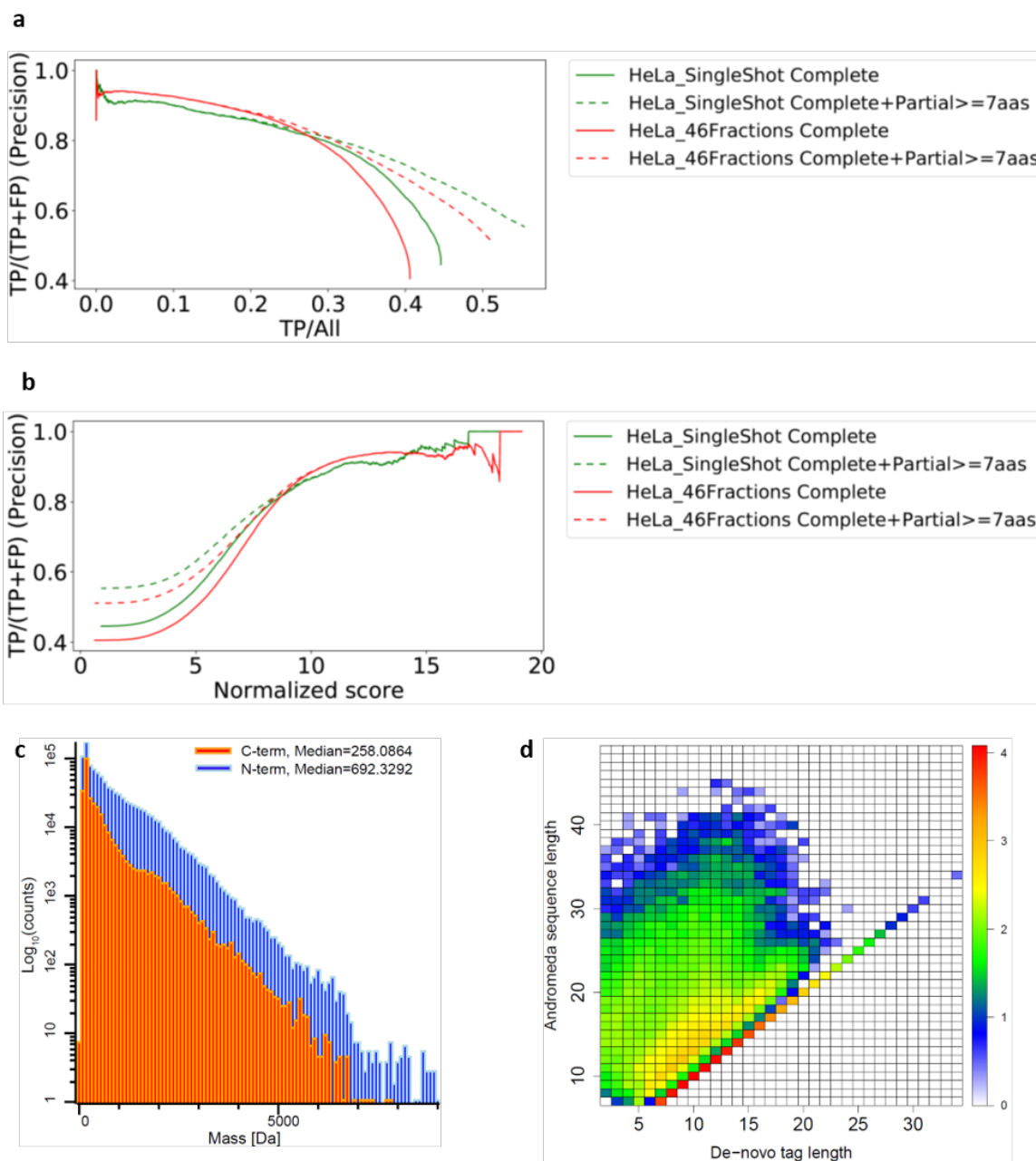


Figure 4.5: Incomplete sequences. **a.** Precision-coverage curve based on the MaxNovo normalized score including complete and incomplete sequences calculated for the single-shot and the fractionated dataset. A minimum tag length of seven was applied. **b.** The precision values from panel a are plotted against the normalized score. **c.** Histograms of N- and C-terminal missing masses for the fractionated HeLa dataset (all MS/MS spectra with missing terminal mass >0). **d.** Density plot of tag length vs. peptide sequence length of all MS/MS spectra that have also been identified by Andromeda.

New identifications

Next, we focus on the sequences in the fractionated HeLa dataset that have not been identified by the Andromeda search engine. 15.2%, or 135,908 in total, of the complete sequences or sequence tags that are at least seven amino acids long are from MS/MS spectra that have not been identified by Andromeda. We perform a multi-tier BLAST search against protein databases containing all human proteins, proteins derived from the HeLa genome, the *Bos taurus* proteome and finally against a SwissProt database containing all species. We further filtered the non-identified MS/MS spectra by a combined score of 91.715 to allow an error rate of 10%, which results in 28,126 sequences that are submitted to BLAST. 28,123 sequences find a match with the same length in at least one of the four tiers. Out of these, 18,492 (65.8%) are matching exactly, while the remaining ones have at least one substitution (Fig. 4.6a). We further classify the sequences without mismatch in Fig. 4.6b.

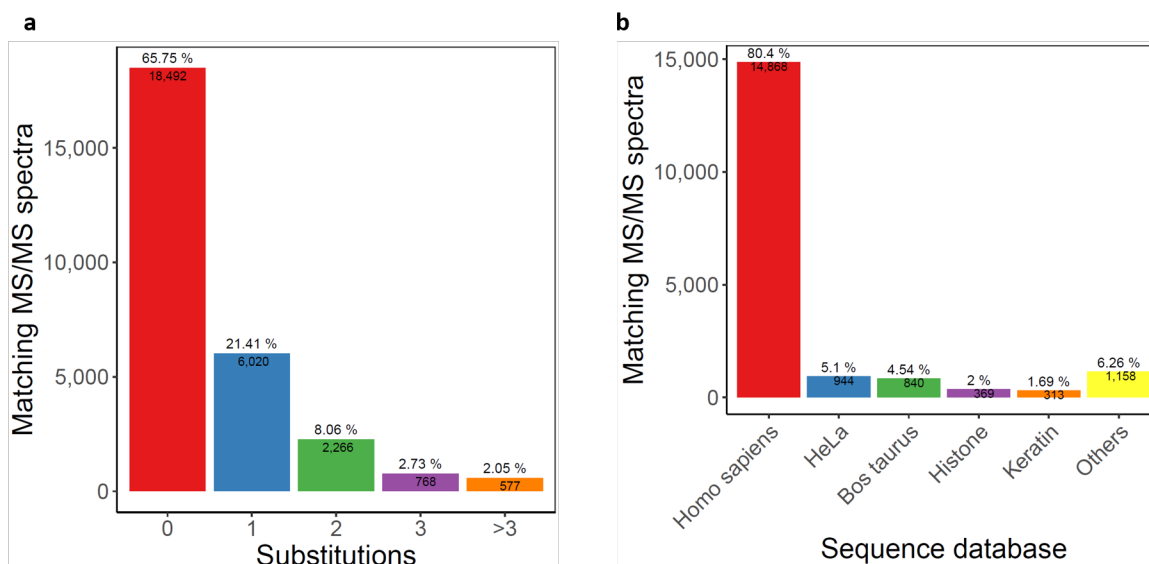


Figure 4.6: New identifications. **a.** Number of MS/MS spectra in the fractionated HeLa dataset whose full-length de-novo prediction was matched by BLAST with the specified number of substitutions. **b.** The de-novo sequences from panel a that were matched without substitutions are classified according to the sequence database that they were matching to.

Tims TOF data

We next applied MaxNovo to a HeLa dataset acquired on a tims-TOF Pro instrument. The precision-coverage plot based on the combined score (Fig. 4.7) shows at first less performance than in Fig. 4.3. However, most of the wrong de-novo sequences have only a swap of the two C-terminal amino acids and are otherwise correct. This is due to the lower mass range limit in the MS/MS spectra of 200 Da in this particular dataset. If we ignore the order of the two C-terminal amino acids, we obtain a much better precision-coverage curve (Fig. 4.7). Hence, it is recommended to use a lower mass range limit for MS/MS spectra or to ignore the order of the last two amino acids in each peptide.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

84

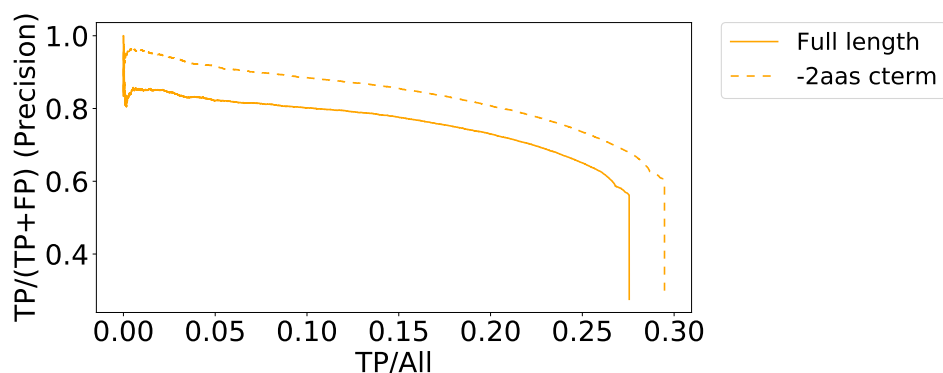


Figure 4.7: timsTOF Pro data. Precision-coverage curve for the timsTOF Pro data based on the combined score once checking complete correctness of the sequence and once ignoring the two C-terminal amino acids.

Application to ancient sequences

Ancient proteomes present a challenging application of de-novo and error-tolerant search approaches, both because comparative genomes of closely-related species are frequently not available and because the determination of sequence variation is essential when phylogenetic analysis is the main purpose of the proteomic study. Ancient proteomes contain a large number of variable PTMs [Cappellini et al., 2019], many of which might only be present at low frequencies, and have increasing rates of peptide bond hydrolysis for older samples, making the occurrence of semi-specific and non-specific peptides in resulting datasets much more likely [Chen et al., 2019]. Finally, dentine and bone proteomes are dominated by collagen type I, which is heavily hydroxylated, increasing search complexity as well as the presence of incomplete fragmentation series, particularly around proline positions. We therefore tested MaxNovo performance against a Late Pleistocene hominin proteome dataset that was previously generated using three proteases in parallel [Lanigan et al., 2020]. We observe that, as for modern data (Fig. 4.5d), de-novo sequence tag lengths are on average shorter when compared to the Andromeda-derived sequence solution (Fig. 4.8a). Likewise, as with PEAKS [Welker, 2018a], the probability that a top-ranking full-length MaxNovo solution exists and is correct for a given spectrum is highly dependent on peptide amino acid length, although the MaxNovo scores provide the possibility to apply stringent selection criteria and partly remove this dependency (Fig. 4.8b). Despite this, MaxNovo correctly resolved the true peptide sequence when presented with protein sequence databases of chimpanzee (*Pan*, Fig. 4.8c) and gorilla (*Gorilla*, Fig. 4.8d), including PSMs containing diagenetic modifications (deamidation) and proline hydroxylation.

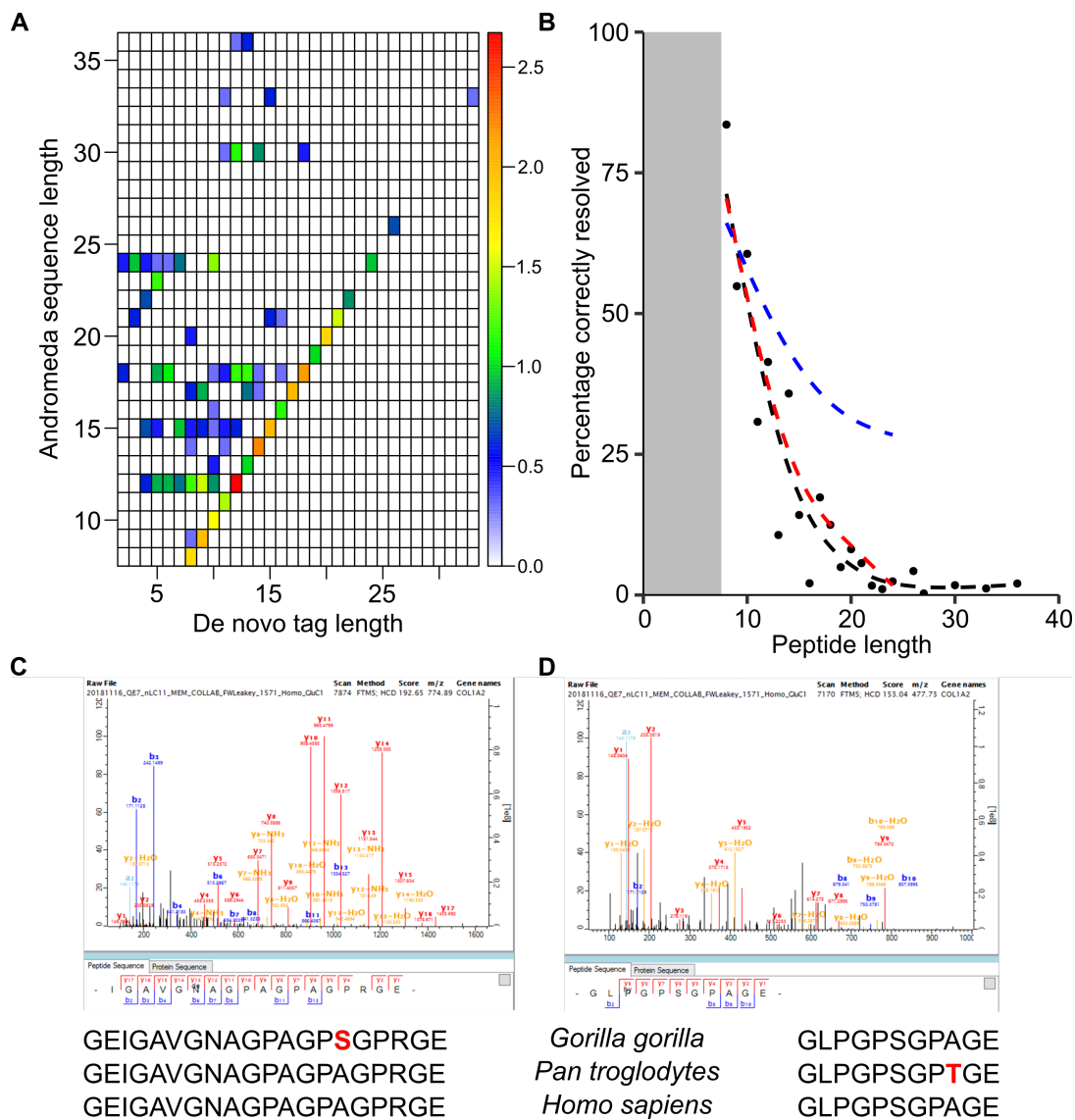


Figure 4.8: Validation on ancient samples. **a.** Density plot of tag length (MaxNovo) vs. peptide sequence length (Andromeda) for PSMs where the de-novo solution contains the database search sequence solution ($n=1,726$). **b.** Correctness of MaxNovo top-ranking sequence solutions are highly dependent on peptide sequence length (black, $n=14,249$). Filtering for combined scores over 75 increases average correctness moderately for longer peptides (red, $n=5,703$), but significantly when only taking into account normalized scores >10 (blue, $n=1,427$). **c.** Example of a successfully resolved MaxNovo PSM containing asparagine deamidation (combined score = 98.9451, GluC digest). **d.** Example of a successfully resolved MaxNovo PSM containing proline hydroxylation (combined score = 98.907, GluC digest). For both **c** and **d**, a protein sequence alignment across COL1A2 is given, indicating relevant positions with SAP in *Gorilla* (**c**, A279S) and *Pan* (**d**, A564T). Coordinates in reference to UniProt accession number P08123.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

Score	Description
raw score	This is the score assigned to each path, which is the sum of the six following score. The optimal path is found maximizing this score in an exhaustive search over all paths.
direct path score	Score summing up weights that are collected along the path with contributions from traversed nodes and edges.
complement score	Score contribution of ions that are complementary to the ions in the direct path.
protease score	Reward for terminal parts of the sequence that are in agreement with the specified protease that was used for generating peptides from proteins.
a ₂ score	Reward for the presence of an a ₂ ion.
water loss score	Reward for the presence of ions resulting from the loss of a water molecule in case the main path contains any of the amino acids D, E, S or T.
ammonia loss score	Reward for the presence of ions resulting from the loss of an ammonia molecule in case the main path contains any of the amino acids K, N, Q or R.
normalized score	This is the raw score divided by the precursor mass. While the raw score adds up the total spectral evidence for a peptide, the normalized score rather corresponds to the spectral evidence per peptide length.
complete score	The complete score equals the normalized score, in case the sequence goes from terminus to terminus, i.e. is completely sequencing the peptide. Otherwise the complete score equals zero.
gap score	The gap score is the difference in raw scores between the best and the second-best scoring solution. If there is no second-best solution, the gap score equals the raw score.
combined score	The combined score is a combination of the complete score and the gap score. Both of these scores are ranked, Then the sum of the two ranks is taken and normalized to lie between 0 and 100.

Table 4.1: Definition of scores used in this publication. The scores with their names in *italic* are additive contributions to the raw score. The combined score is the score of choice for ranking full-length sequences. Incomplete sequence tags are best ranked using the normalized score.

4.5 Conclusion

We introduced MaxNovo, a novel de-novo sequencing algorithm that is integrated into the MaxQuant software. It shows a performance, in terms of sequence identifications, which is as good as or better than software that is currently in use. It is interesting to observe, that no machine learning was used in its construction, but that we rationalized a scoring function based on expert knowledge on peptide fragmentation spectra. Based on its integration into the MaxQuant environment, we can expect MaxNovo to make significant introductions to the recovery of peptide sequences of relevance in clinical, biological, and evolutionary settings.

4.6 Author's contribution

MaxNovo is a novel spectrum graph-based peptide de-novo sequencing algorithm implemented in MaxQuant, allowing for reference database-independent identification. It provides peptide sequences of completely as well incompletely sequenced spectra including mass information for the remaining terminal sequence.

I worked on the implementation and optimisation of the MaxNovo de-novo sequencing algorithm. Furthermore, I individually inspected a range of MS/MS spectra identified by MaxNovo to text and improve the algorithm's sequencing performance. I selected spectra to reflect various characteristics that pose challenges to de-novo sequencing.

The identified de-novo sequence is provided as sequence pattern allowing multiple amino acids or amino acid pairs for a single position due to isobaric masses of certain amino acids or amino acid pairs. I provided a de-novo sequence parser to be able to retrieve all possible sequence combinations of a single identified de-novo sequence to be able to validate the performance of MaxNovo.

For the validation of the MaxNovo algorithm, I selected two HeLa benchmark datasets for which I performed the MaxNovo, Novor and PEAKS analysis. Furthermore, I set up a local BLAST search and a computational post-processing workflow to identify high scoring de-novo sequences, previously unidentified by the Andromeda search engine. Finally, I processed the ancient hominin bone data and provided the results for downstream analysis.

The study demonstrated that MaxNovo performs as well as or better than state-of-the-art deep learning-based de-novo sequencing algorithms and outperforms the benchmarked algorithms in the high-sensitivity range of precision-coverage plots.

4.7 Additional information

Acknowledgements

We thank Christoph Wichmann, Hamid Hamzeiy and Sule Yilmaz for helpful discussions. This project was partially funded by the German Ministry for Science and Education (BMBF) funding action MSCoreSys, reference number FKZ 031L0214D. P.K. is supported by the Marie Skłodowska-Curie European Training Network (ETN) PUSHH, a project funded by the European Union's Framework Program for Research and Innovation Horizon 2020 (grant agreement no. 861389). P.G. is supported by the Marie Skłodowska-Curie European Training Network TEMPERA, a project funded by the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 722606. F.W. is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 948365).

All author's contributions

P.G. and J.C. designed and developed the main code. F.W. provided and analyzed the hominin bone data. All authors analyzed the data and wrote the manuscript.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

4.8 Supplementary information

Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

Petra Gutenbrunner^{1,2}, Pelagia Kyriakidou¹, Frido Welker & Jürgen Cox^{1,4*}

¹Computational Systems Biochemistry Research Group, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.

²Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 Munich, Germany.

³Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Øster Voldgade 5, 1350 Copenhagen, Denmark.

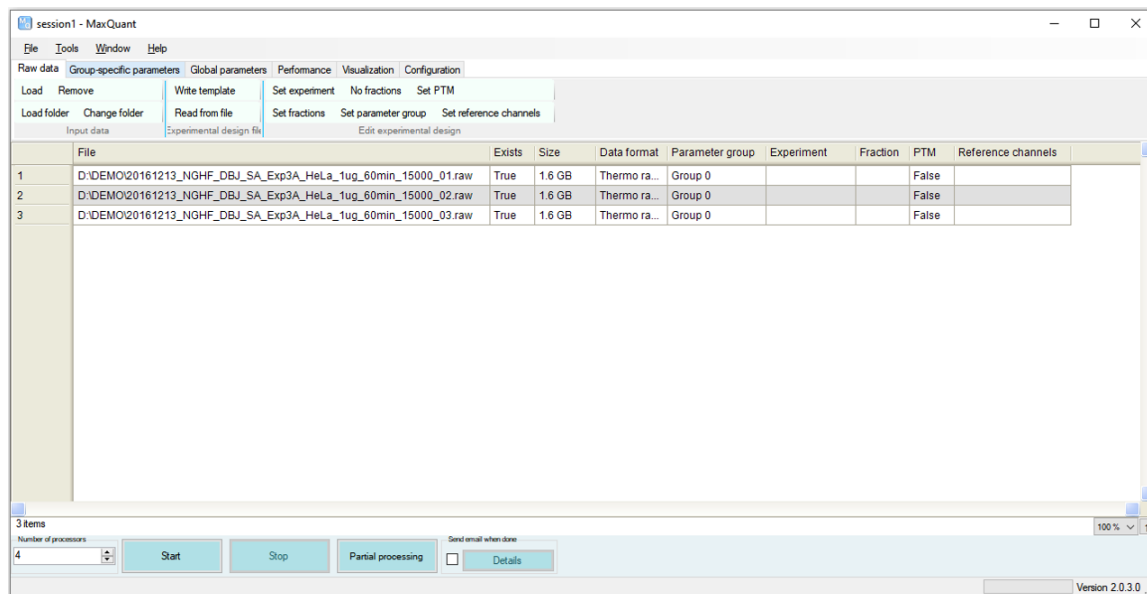
⁴Department of Biological and Medical Psychology, University of Bergen, Jonas Liesvei 91, 5009 Bergen, Norway.

*email: cox@biochem.mpg.de

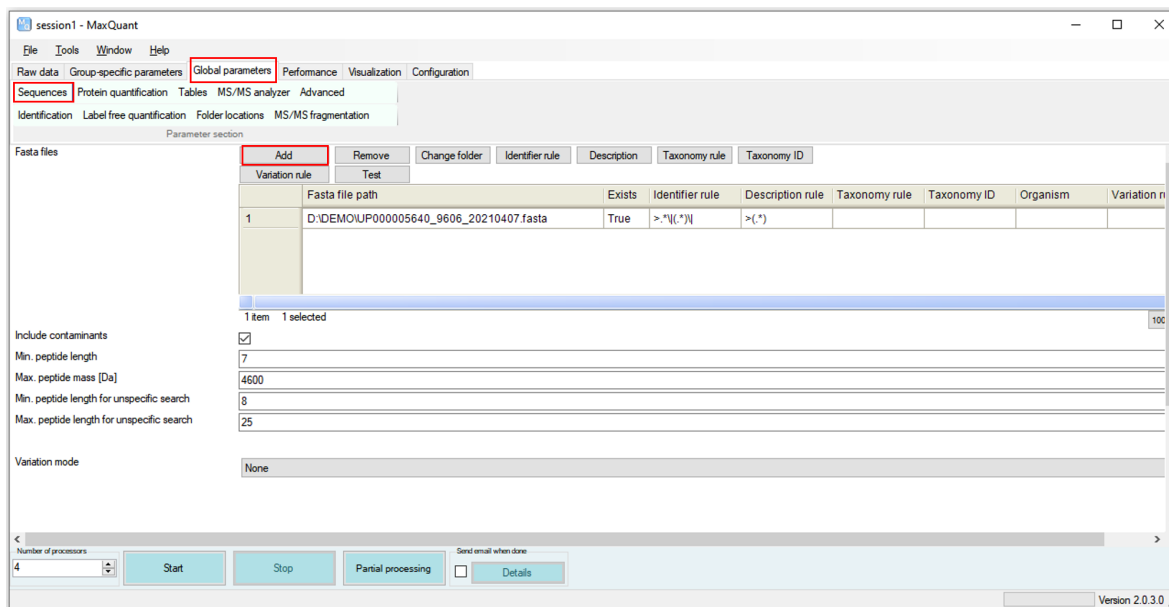
4.8.1 A user guide on how to run MaxNovo in MaxQuant

MaxNovo is integrated into the MaxQuant environment. You can download MaxQuant from <https://maxquant.org/maxquant/>. You must make sure to install .NET Core 3.1 SDK x64 from <https://dotnet.microsoft.com/download/dotnet/3.1>.

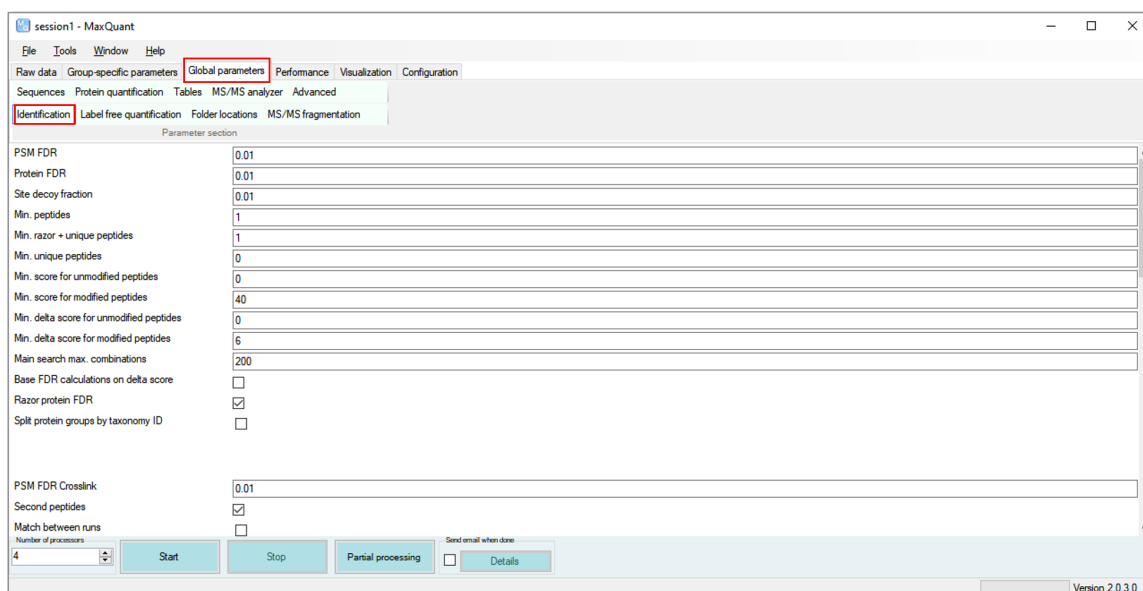
1. Load your raw files and set your experiment design.



2. Go to “Global parameters” tab → “Sequences” tab and load your FASTA file(s).

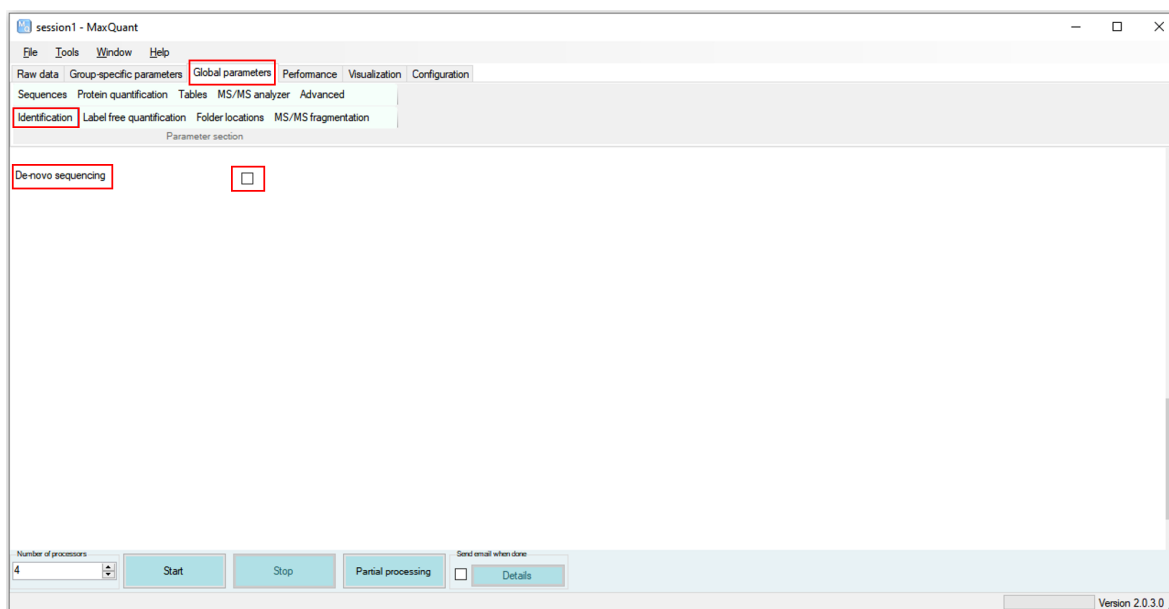


3. Go to “Global parameters” tab → “Identification” tab.

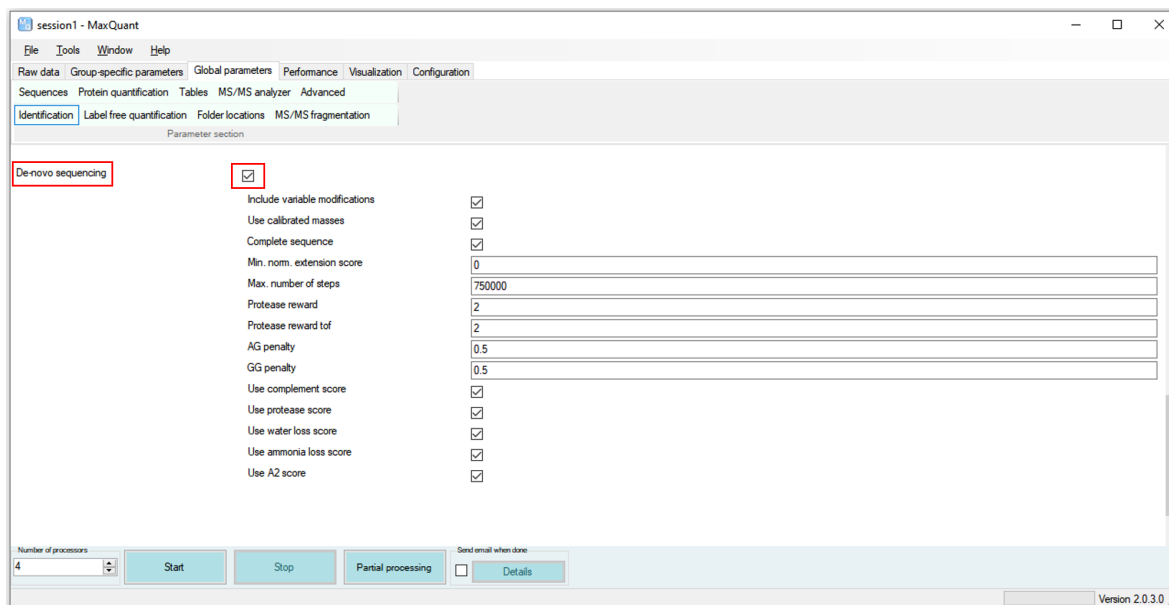


4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

3. a. Scroll down till you find the “De-novo sequencing” checkbox.

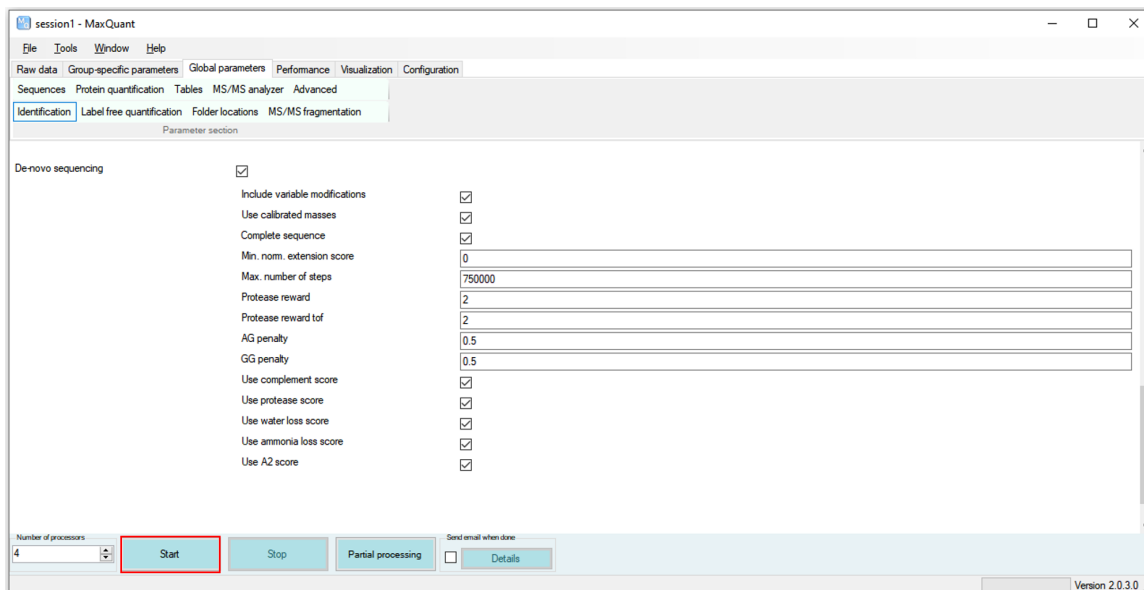


3. b. Enable the De-novo sequencing by clicking on it.

















3.c. Set the MaxNovo parameters that appeared after clicking on the “De-novo sequencing” checkbox.

Please find the description of each one of the parameters at the Supplementary Table 4.2 in the Supporting Information.

4. Press start to start a MaxQuant run with the MaxNovo denovo identification enabled.

After MaxQuant finishes navigate to the “combined” folder that MaxQuant created and then to the “txt” folder. In the “txt” folder there is a file called “msmsScans.txt” where you can find all the additional de-novo information from your experiment. Please find the description of each one of the columns at the Supplementary Table 4.3 in the Supporting Information.

-  allPeptides.txt
-  evidence.txt
-  matchedFeatures.txt
-  modificationSpecificPeptides.txt
-  ms3Scans.txt
-  msms.txt
-  **msmsScans.txt**
-  mzRange.txt
-  Oxidation (M)Sites.txt
-  parameters.txt
-  peptides.txt
-  proteinGroups.txt
-  summary.txt
-  tables.pdf

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

Supplementary Table 4.2

Description of the de-novo parameters in MaxQuant. Find these parameters in MaxQuant when you enable the de-novo algorithm under Global parameters → Identification → De-novo sequencing.

Parameter	Default value	Description
Include variable modifications	true	When checked, the variable modifications (specified under Group-specific parameters → Modifications → Variable modifications) are considered when MaxNovo runs.
Use calibrated masses	true	When checked, the masses are the ones after the re-calibration step.
Complete sequence	true	When checked, if the de-novo sequence is not completely resolved, a complementary ion series is looked up that can extend it.
Min. norm. Extension score	0	the minimum normalized score the extended sequence has to have in order for the de-novo sequence to be extended (in case incomplete).
Max. number of steps	750000	The upper limit of steps in a path that are tested in the exhaustive search along the directed acyclic graph representing each MS/MS spectrum.
Protease reward	2	The weight of the contribution of the "protease score" in the calculation of the "raw score".
Protease reward tof	2	The weight of the contribution of the "protease score" in the calculation of the "raw score" for timsTOF data.
AG penalty	0.5	Score penalty for picking AG instead of Q.
GG penalty	0.5	Score penalty for picking GG instead of N.
Use complement score	true	When checked, complement score contributes to the calculation of the "raw score".
Use protease score	true	When checked, protease score contributes to the calculation of the "raw score".
Use water loss score	true	When checked, water-loss score contributes to the calculation of the "raw score".
Use ammonia loss score	true	When checked, ammonia-loss score contributes to the calculation of the "raw score".
Use a ₂ score	true	When checked, a ₂ score contributes to the calculation of the "raw score".

Table 4.2

Supplementary Table 4.3

Description of the new de-novo associated columns in the “msmsScans.txt” output file from MaxQuant.

Column name	Description
DN sequence	The de-novo AA regex sequence that corresponds to the one with the biggest raw score among all de-novo AA regex sequences for the specific MS/MS scan.
DN length	The length of the sequence that is stored in the column “DN sequence”.
DN min levenshtein distance	The minimum Levenshtein distance between the identified Andromeda sequence(column: “Sequence”) and the de-novo sequences resulting from the de-novo regex (column: “DN sequence”). Levenshtein distance between two peptide sequences is the minimum number of single-amino acid edits (insertions, deletions, or substitutions) required to change one sequence into the other. The MS/MS scans with no identified Andromeda sequence or no de-novo sequence or missing both will have DN min levenshtein distance equal to -1.
DN extension sequence	Sequence that was attached to the incomplete main path sequence in a second search for explaining the missing mass at a terminus.
DN extended	Whether or not a second search for explaining the missing mass at a terminus was successful.
DN complete	When marked with ‘+’ the de-novo regex sequence (column: “DN sequence”) is full length with no mass missing.at the N- or/and C-terminal.
DN raw score	This is the score assigned to the de-novo regex sequence (column: “DN sequence”) and it is the sum of other 6 scores. Raw score = direct path score+complement score+protease score+A ₂ score+water-loss score+ammonia-loss score
DN extension score	Score that was achieved in a second search for explaining the missing mass at a terminus.
DN extension norm. score	Normalized score that was achieved in a second search for explaining the missing mass at a terminus.
DN (normalized) score	This is the raw score divided by the precursor mass.
DN complete score	The complete score equals the normalized score, in case the sequence goes from terminus to terminus, i.e. is completely sequencing the peptide. Otherwise the complete score equals zero.
DN combined score	The combined score is a combination of the complete score and the gap score. Both of these scores are ranked, and then the sum of the two ranks is taken and normalized to lie between 0 and 100.
DN nterm mass	The mass that is missing at the N-terminal of the de-novo sequence tag. The sum of the de-novo sequence tag mass and the DN nterm mass is equal to the precursor mass associated with the specific MS/MS scan.

4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide identification rates in collisional dissociation MS/MS spectra

DN cterm mass	The mass that is missing at the C-terminal of the de-novo sequence tag. The sum of the de-novo sequence tag mass and the DN cterm mass is equal the precursor mass associated with the specific MS/MS scan.
DN missing mass	The sum of DN nterm mass and DN cterm mass.
DN nterm delta score	The score difference between the solution and the best solution that is not allowed to connect to the N-terminus.
DN cterm delta score	The score difference between the solution and the best solution that is not allowed to connect to the C-terminus.
DN term delta score	The maximum of the above two scores.
DN full length delta score	The difference in raw scores between the best and the second-best scoring solution. If there is no second-best solution, the gap score equals the raw score.
DN protease score	The reward for terminal parts of the sequence that are in an agreement with the specified protease that was used for generating peptides from proteins.
DN complement score	The score contribution of ions that are complementary to the ions in the direct path.
DN a ₂ score	The reward for the presence of an a ₂ ion.
DN water loss score	The reward for the presence of ions resulting from the loss of a water molecule in case the main path contains any of the amino acids D, E, S, or T.
DN ammonia loss score	The reward for the presence of ions resulting from the loss of an ammonia molecule in case the main path contains any of the amino acids K, N, Q, or R.
DN agrees with andromeda	When marked with '+' the de-novo regex sequence (column: "DN sequence") matches the sequence from the database search (column: "Sequence") that passed the PSM FDR control (Column: "Identified"=="+"). Here a "match" is considered even if the de-novo regex sequence is a subsequence of the database search one.
DN agrees with andromeda complete	When marked with '+' the de-novo regex sequence (column: "DN sequence") matches the sequence from the database search (column: "Sequence") that passed the PSM FDR control (Column: "Identified"=="+"). Here a "match" is considered only the full-length match.
DN all sequences	All the de-novo AA regex sequences associated with a specific MS/MS scan separated by ";" sorted by their normalized score in descending order.
DN all scores	The normalized score of all the de-novo AA regex sequences associated with a specific MS/MS scan separated by ";" in descending order.

DN all agrees	“+” or “-” separated by “;” for each one of the de-novo regex sequences (column: “DN all sequences”) in the same order. When marked with ‘+’ the de-novo regex sequence matches the sequence from the database search (column: “Sequence”). Here a “match” is considered even if the de-novo regex sequence is a subsequence of the database search one.
DN any agrees	When marked with ‘+’ any of the de-novo regex sequences (column: “DN all sequences”) matches the sequence from the database search (column: “Sequence”). Here a “match” is considered even if the de-novo regex sequence is a subsequence of the database search one.
DN number of steps	The number of different steps that correspond to the path associated with the de-novo regex sequence (column: “DN sequence”).
DN is dominantly y	It is true if the main path corresponds to the y ion series. This is known for sequences that connect to at least one terminus.

Table 4.3

Three different kinds of brackets (round, square, and curly) can be found in a regex de-novo-sequence. The round brackets “()” contain the modification of the amino acid that is located on the left side of the brackets. For example M(Oxidation (M)). The square brackets “[]” contain two amino acids that their order cannot be distinguished by MaxNovo due to the absence of fragment ion peaks in the MS/MS scan. For example [WR] can be either WR or RW. The curly brackets “{}” contain amino acids that are separated by the pipe character “—”. The separated by “—” amino acids are equally possible to be at that position because they are isobaric or almost isobaric (up to mass tolerance error). For example I—L can be either I or L.

**4. Spectrum graph-based de-novo sequencing algorithm MaxNovo achieves high peptide
98 identification rates in collisional dissociation MS/MS spectra**

Chapter 5

The dental proteome of *Homo antecessor*

Frido Welker^{1,22*}, Jazmín Ramos-Madrugal^{2,22}, Petra Gutenbrunner^{1,22}, Meaghan Mackie^{1,3}, Shivani Tiwary², Rosa Rakownikow Jersie-Christensen³, Cristina Chiva^{4,5}, Marc R. Dickinson⁶, Martin Kuhlwilm⁷, Marc de Manuel⁷, Pere Gelabert⁷, María Martín-Torres^{8,9}, Ann Margvelashvili¹⁰, Juan Luis Arsuaga^{11,12}, Eudald Carbonell^{13,14}, Tomas Marques-Bonet^{4,7,15,16}, Kirsty Penkman⁶, Eduard Sabidó^{4,5}, Jürgen Cox², Jesper V. Olsen³, David Lordkipanidze^{10,17}, Fernando Racimo¹⁸, Carles Lalueza-Fox⁷, José María Bermúdez de Castro^{8,9*}, Eske Willerslev^{18,19,20,21*} & Enrico Cappellini^{1*}

Nature 580, 235–238 (2020). <https://doi.org/10.1038/s41586-020-2153-8>

Published: 01 April 2020

¹Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ²Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany. ³The Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark. ⁴Center for Genomic Regulation (CNAG-CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. ⁵Proteomics Unit, University Pompeu Fabra, Barcelona, Spain. ⁶Department of Chemistry, University of York, York, UK. ⁷Institute of Evolutionary Biology (UPF-CSIC), University Pompeu Fabra, Barcelona, Spain. ⁸Centro Nacional de Investigación sobre la Evolución Humana (CENIEH), Burgos, Spain. ⁹Anthropology Department, University College London, London, UK. ¹⁰Georgian National Museum, Tbilisi, Georgia. ¹¹Centro Mixto UCM-ISCIH de Evolución y Comportamiento Humanos, Madrid, Spain. ¹²Departamento de Paleontología, Facultad de Ciencias Geológicas, Universidad Complutense de Madrid, Madrid, Spain. ¹³Departamento d'Història i Història de l'Art, Universitat Rovira i Virgili, Tarragona, Spain. ¹⁴Institut Català de Paleoecologia Humana i Evolució Social (IPHES), Tarragona, Spain. ¹⁵Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain. ¹⁶Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain. ¹⁷Tbilisi State University, Tbilisi, Georgia. ¹⁸Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ¹⁹Department of Zoology, University of Cambridge, Cambridge, UK. ²⁰Wellcome Sanger Institute, Hinxton, UK. ²¹Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark.

²²These authors contributed equally: Frido Welker, Jazmín Ramos-Madrugal, Petra Gutenbrunner.

*e-mail: frido.welker@sund.ku.dk;josemaria.bermudezdecastro@cenieh.es;ewillerslev@sund.ku.dk;ecappellini@sund.ku.dk

5.1 Abstract

The phylogenetic relationships between hominins of the Early Pleistocene epoch in Eurasia, such as *Homo antecessor*, and hominins that appear later in the fossil record during the Middle Pleistocene epoch, such as *Homo sapiens*, are highly debated [Gabunia et al., 2000, Zhu et al., 2018, Stringer, 2016, Hublin, 2009, Rightmire, 1998]. For the oldest remains, the molecular study of these relationships is hindered by the degradation of ancient DNA. However, recent research has demonstrated that the analysis of ancient proteins can address this challenge [Cappellini et al., 2019, Chen et al., 2019, Welker et al., 2019]. Here we present the dental enamel proteomes of *H. antecessor* from Atapuerca (Spain) [Bermúdez de Castro et al., 1997, Carbonell et al., 1995] and *Homo erectus* from Dmanisi (Georgia) [Gabunia et al., 2000], two key fossil assemblages that have a central role in mod-

els of Pleistocene hominin morphology, dispersal and divergence. We provide evidence that *H. antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins, including modern humans, Neanderthals and Denisovans. This placement implies that the modern-like face of *H. antecessor*—that is, similar to that of modern humans—may have a considerably deep ancestry in the genus *Homo*, and that the cranial morphology of Neanderthals represents a derived form. By recovering AMELY-specific peptide sequences, we also conclude that the *H. antecessor* molar fragment from Atapuerca that we analysed belonged to a male individual. Finally, these *H. antecessor* and *H. erectus* fossils preserve evidence of enamel proteome phosphorylation and proteolytic digestion that occurred in vivo during tooth formation. Our results provide important insights into the evolutionary relationships between *H. antecessor* and other hominin groups, and pave the way for future studies using enamel proteomes to investigate hominin biology across the existence of the genus *Homo*.

5.2 Introduction

Since 1994, over 170 human fossil remains have been recovered from level TD6 of the Gran Dolina site of the Sierra de Atapuerca [Carbonell et al., 1995] (Burgos, Spain) (Extended Data Fig. 5.4, Supplementary Information). These fossils have been dated to the late Early Pleistocene epoch and exhibit a unique combination of cranial, mandibular and dental features [Bermúdez de Castro et al., 1997, Duval et al., 2018]. To accommodate the variation observed in the human fossils from TD6, a new species of the genus *Homo* – *H. antecessor* – was proposed in 1997 [Bermúdez de Castro et al., 1997]. The relationship of this species to earlier or later hominins in Eurasia – such as the *H. erectus* specimens from Dmanisi or Neanderthals, Denisovans and modern humans, respectively – have been the subject of considerable debate [Stringer, 2016, Hublin, 2009, Freidline et al., 2013, Lacruz et al., 2013]. These issues remain unresolved owing to the fragmentary nature of hominin fossils at other sites, and the failure to recover ancient DNA in Eurasia that dates to the Early, and most of the Middle, Pleistocene epoch.

By contrast, recent developments in the extraction and tandem mass spectrometric analysis of ancient proteins have made it possible to retrieve phylogenetically informative protein sequences from Early Pleistocene contexts [Cappellini et al., 2019, Welker et al., 2019]. We therefore applied ancient protein analysis to a *H. antecessor* molar from sublevel TD6.2 of the Gran Dolina site of the Sierra de Atapuerca (specimen ATD6-92) (Extended Data Fig. 5.5a). This specimen, identified as an enamel fragment of a permanent lower left first or second molar, has been directly dated to 772–949 thousand years ago (ka) using a combination of electron spin resonance and U-series dating [Duval et al., 2018]. In addition, we sampled dentine and enamel from an isolated *H. erectus* upper first molar (specimen D4163) (Extended Data Fig. 5.5b) from Dmanisi (Georgia) that has been dated to 1.77 million years ago (Ma) [Gabunia et al., 2000, Ferring et al., 2011, Lordkipanidze et al., 2013], as amino acid racemization analysis of this specimen indicated the presence of an endogenous protein component in the intracrystalline enamel fraction of the tooth (Extended Data Fig. 5.6, Supplementary Information). On both specimens, we performed digestion-free peptide extraction optimized for the recovery of short, degraded protein remains [Cappellini et al., 2019]. Nanoscale liquid chromatography–tandem mass spectrometry (nanoLC–MS/MS) acquisition was replicated in two independent proteomic laboratories (Extended Data Table 5.1), implementing common precautions and analytical workflows to minimize protein contamination (5.3 Methods). We compared the proteomic datasets retrieved from the Pleistocene hominin tooth specimens with those generated from a positive control, a recent human premolar (Ø1952; which is from a male individual and is approximately three centuries old), as well as previously published Holocene teeth [Stewart et al., 2017] (5.3 Methods, 5.10 Supplementary Information). Finally, to validate our enamel peptide spectrum matches, we performed machine-learning-based MS/MS spectrum intensity prediction using the wiNNeR algorithm [Tiwary et al., 2019].

5.3 Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Site location and specimen selection

Recent human control specimens

We analysed Ø1952, a human premolar recovered in an archaeological excavation in Copenhagen (Almindeligt Hospital Kirkegård, excavated in 1952, from kisse '2'). The tooth is approximately three centuries old, as the cemetery was in use from approximately ad 1600 to approximately ad 1800, and originates from a male individual. We also re-analysed previously published data [Stewart et al., 2017] related to specimens that are dated to between approximately 5,700 and 200 years ago; of these specimens, we took SK339 as a recent example in our comparative figures (a male individual from Fewston (UK) dated to the nineteenth century ad).

Atapuerca

One fragmentary permanent lower left first or second molar (ATD6-92; field number and museum accession number at CENIEH) was used for ancient protein analysis (Extended Data Fig. 5.5a, 5.10 Supplementary Information). ATD6-92 originates from sublevel TD6.2 of the Gran Dolina cave site. Sublevel TD6.2 contains a large number of faunal remains, about 170 hominin fossils and about 830 archaeological artefacts. All hominin specimens from sublevel TD6.2, including ATD6-92, are attributed to *H. antecessor* [Bermúdez de Castro et al., 1997]. ATD6-92 has recently been directly dated through electron spin resonance, laser-ablation inductively coupled plasma mass spectrometry U-series and bulk U-series dating [Duval et al., 2018]. Together with previous chronological research at the site, these analyses constrain the age of ATD6-92 to 772–949 thousand years old [Duval et al., 2018].

Dmanisi

One fragmentary permanent upper first molar (D4163; field number and museum accession number at the Georgian National Museum) was used for ancient protein analysis (Extended Data Fig. 5.5b, 5.10 Supplementary Information). D4163 derives from layer B1 in excavation block M6 (Dmanisi). Layer B1 at Dmanisi contains one of the richest palaeontological assemblages attributed to the Eurasian Early Pleistocene epoch, including several hominin crania. Below, we refer to these specimens as *H. erectus* (Dmanisi). They represent the earliest hominin fossils outside Africa, and are dated to 1.77 Ma [Ferring et al., 2011]. Faunal material from the site previously demonstrated ancient protein survival for most specimens, but a total absence of ancient DNA [Cappellini et al., 2019] (Fig. 5.3).

Amino acid racemization

Chiral amino acid analysis was undertaken on one Pleistocene sample from the hominin tooth (D4163) to test the endogeneity of the enamel protein through its degradation patterns. The tooth chip was separated into the enamel and dentine portions, and each was powdered with an agate pestle and mortar. All samples were prepared using previously published procedures [Penkman et al., 2008], modified to be optimized for enamel, using a bleach time of 72 h to isolate the intracrystalline protein, demineralization in HCl, KOH neutralization and formation of a biphasic solution through centrifugation [Dickinson et al., 2019]. Two sub-samples were analysed from each portion: one fraction was directly demineralized and the free amino acids analysed, and the second was treated to release the peptide-bound amino acids, thus yielding the total hydrolysable amino acid fraction. Samples were analysed in duplicate by reversed-phase high-performance liquid chromatography, with standards and blanks analysed alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamidation to aspartic acid (Asp) and glutamic acid (Glu), respectively [Hill, 1965]. It is therefore not possible to distinguish between the acidic amino acids and their derivatives, and they are reported together as Asx and Glx, respectively. Additional descriptions of the methods, as well as additional results, are given in the Supplementary Information.

Proteomic extraction and nanoLC–MS/MS

Protein extraction

Protein extraction was conducted on enamel samples (from the Atapuerca *H. antecessor*, Dmanisi *H. erectus* and Ø1952) and a dentine sample (Dmanisi), using one of three protocols. In brief, the first extraction method used HCl for demineralization, but included no subsequent reduction, alkylation or digestion. The second extraction method used a more standard approach, in which the pellet left from the demineralization in extraction one was reduced, alkylated and digested with LysC and trypsin. The third extraction method used TFA for demineralization, and had no subsequent reduction, alkylation or digestion. The first and third extraction approaches provided more extensive peptide recovery in ancient enamel proteomes [Cappellini et al., 2019] compared to the second extraction approach [Mackie et al., 2018]. Further details can be found in the 5.10 Supplementary Information and a previous publication [Cappellini et al., 2019]. Ø1952 was processed using extraction methods one and three. No proteinase and phosphatase inhibitors were used during extraction, as we assumed that catalytically active enzymes were not present in our specimens and the high acidic conditions during our extraction would have irreversibly denatured any proteases possibly present as contaminants in our reagents. Extended Data Table 5.1 provides a breakdown of the use of specific extraction methods, hominin samples and hominin tissues.

NanoLC–MS/MS analysis.

Shotgun proteomic data were obtained on peptide extracts of both hominins at separate facilities at the Novo Nordisk Centre for Protein Research (University of Copenhagen) and the Proteomics Unit (Centre for Genomic Regulation, Barcelona Institute of Science and Technology). Full peptide elutions were injected, in some cases across replicate runs in both Copenhagen and Barcelona. In brief, samples processed in Copenhagen were suspended in 0.1%

trifluoroacetic acid, 5% acetonitrile, and analysed on a Q-Exactive HF or HF-X mass spectrometer (Thermo Fisher Scientific) coupled to an EASYnLC 1200 (Thermo Fisher Scientific). The HF or HF-X mass spectrometer was operated in positive ion mode with a nanospray voltage of 2 kV and a source temperature of 275 °C. Data-dependent acquisition mode was used for all mass spectrometric measurements. Full mass spectrometry scans were done at a resolution of 120,000 with a mass range of m/z 300–1,750 and 350–1,400 for the HF and HF-X mass spectrometers, respectively, with detection in the Orbitrap mass analyser. Fragment ion spectra were produced at a resolution of 60,000 via high-energy collision dissociation (HCD) at a normalized collision energy of 28% and acquired in the Orbitrap mass analyser. In addition, test runs for the Dmanisi sample were performed at a shorter gradient (Supplementary Information). In Barcelona, samples were dissolved in 0.1% formic acid and analysed on a LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific) coupled to an EASY-nLC 1000. The mass spectrometer was operated similarly to the parameters stated for the HF and HF-X mass spectrometers in Copenhagen, except the nanospray voltage was 2.4 kV and full mass spectrometry scans with 1 micro scan were used over a mass range of m/z 350–1,500. Further details of the LC-MS/MS analysis can be found in the Supplementary Information.

Proteomic data analysis

Protein sequence database construction

We constructed an initial Hominidae sequence database containing protein sequences of all major and minor enamel proteins derived from all extant great apes, a hylobatid (*Nomascus leucogenys*) and a macaque (*Macaca mulatta*). Additionally, we added protein sequences translated from extinct Late Pleistocene hominins [Meyer et al., 2016, Castellano et al., 2014], and sequences from *Gorilla beringei*, *Pongo pygmaeus* and *Pongo tapanuliensis* [De Manuel et al., 2016, Nater et al., 2017, Prado-Martinez et al., 2013]. For each protein, we reconstructed the protein sequence of ancestral nodes in the Hominidae family through PhyloBot [Hanson-Smith and Johnson, 2016] to minimize cross-species proteomic effects [Welker, 2018a], and added missing isoform variation on the basis of the isoforms present for each protein in the human proteome as given by UniProt (5.10 Supplementary Information). Furthermore, we downloaded the entire human reference proteome from UniProt (4 September 2018) for a single separate search to allow matches to proteins previously not encountered in enamel proteomes. To each constructed database, we added a set of known or possible laboratory contaminants to allow for the identification of possible protein contaminants [Hendy et al., 2018].

Proteomic software, settings and false-discovery rate

Raw mass spectrometry data were searched for each specimen and tissue separately in either PEAKS [Zhang et al., 2012] (v.7.5) or MaxQuant [Cox and Mann, 2008] (v.1.5.3.30). No fixed modifications were specified in any search. For PEAKS, variable posttranslational modifications were set to include proline hydroxylation, glutamine and asparagine deamidation, oxidation (M), phosphorylation (STY), carbamidomethylation (C) and pyroglutamic acid (from Q and E). For MaxQuant, the following variable post-translation modifications were additionally included: ornithine formation (R), oxidation (W), dioxidation (MW), histidine to

aspartic acid (H > D), and histidine to hydroxyglutamate. Searches were conducted with unspecific digestion. For PEAKS, precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.05 Da, and the false-discovery rate of peptide spectrum matches was set to equal $\leq 1.0\%$. For MaxQuant, default settings of 20 ppm for the first search and 4.5 ppm for the final search were used, a fragment mass tolerance of 20 ppm, and peptide spectrum match (PSM) and protein false-discovery rate was set to 1.0%, with a minimum required Andromeda score of 40 for all peptides. Protein matches were accepted with a minimum of two unique peptide matches in either the PEAKS or MaxQuant search. Proteins that conform to these criteria are detailed in Extended Data Table 5.2. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (single amino acid polymorphisms) are included as Supplementary Data 1.

Data search iterations

For both the proteomes of Dmanisi and Atapuerca specimens, we conducted two separate initial searches. First, we conducted a search in PEAKS against the entire human proteome. Only standard enamel proteins were identified in these searches, allowing us to continue with more specific searches. For the Dmanisi dentine sample, this first search resulted in a small number of peptides matching to collagen type I only. On the basis of the limited amount of sequence data, no further analysis of the Dmanisi dentine data was therefore conducted. Second, for the enamel data, we conducted a search in PEAKS and MaxQuant against the entire enamel proteome database of all extant and extinct Hominidae. This search was used to observe single amino acid polymorphisms outside the known sequence variation in PEAKS and MaxQuant through the de novo, error-tolerant and/or dependent peptide approaches implemented in each of these search engines. These initial searches indicate overall good protein preservation in both samples and the presence of peptide matches to *Pan*- and *Homo*-derived proteins only.

On the basis of these two initial searches, a novel protein sequence database was used that only includes sequences from the genus *Pan*, the genus *Homo*, their predicted ancestral sequences and novel protein sequences observed for both the Dmanisi or Atapuerca samples. Final searches and subsequent data analysis were conducted against this database using the above search and post-translational modification settings. Positions supported by insufficient spectral data were replaced by 'X', in resulting peptide alignments before phylogenetic analysis.

Data analysis of Ø1952 and the previously published¹⁶ dataset was conducted only in MaxQuant against a database restricted to *H. sapiens*. All other search settings and database restrictions were similar between these two recent human controls and the ancient hominin proteomes.

Peptide sequence and single amino acid polymorphism validation

To validate the PSMs covering single amino acid polymorphisms of interest, we performed peptide spectrum intensity prediction and validation on our dataset using wiNNeR [Tiwary et al., 2019]. Data from the ancient specimens (Dmanisi *H. erectus* and Atapuerca *H. antecessor*) were divided into a subset that contained phylogenetically informative peptide

sequences and a larger subset that did not contain these peptides. A training dataset was prepared by taking a subset of the latter peptides, and adding a previously published dataset of enamel proteomes from Dmanisi fauna [Cappellini et al., 2019]. We built two models, one for HCD +2 spectra and one for HCD +3 spectra. We took into account the large number of variable modifications observed in our ancient enamel proteomes, and split the retained data for each model into subsets for training, validation and testing (80:10:10). We then obtained Pearson correlation coefficients for the predicted and true fragment intensities in the test dataset and the phylogenetically informative spectra. The architecture of wiNner was built using Keras (version 2.0.8; <https://keras.io>) and Tensorflow (version 1.3.0). The wiNner analysis indicated close correspondence between predicted and true fragment ion intensities (Pearson correlation coefficient medians between 0.85 and 0.76 for different subsets of the data), indicating adequate peptide sequence identification for all our peptides, including phylogenetically informative positions and the variable post-translational modifications. The wiNner model can be accessed on GitHub (<https://github.com/cox-labs/wiNner.git>). Additional methodological details of the wiNner architecture are given in the Supplementary Information.

Protein damage analysis

Ancient proteins can be modified diagenetically in a variety of ways compared to their modern counterparts. We quantify glutamine and asparagine deamidation following a previously published publication [Mackie et al., 2018] for MaxQuant output, based on MS1 spectral intensities and protein-based bootstrapping (1,000 bootstraps). Further details can be found in the previous publication [Mackie et al., 2018]. We observe that both glutamines and asparagines are almost all deamidated to glutamic acid and aspartic acid, respectively (Extended Data Fig. 5.9a–c). In addition, peptide length distributions were obtained for datasets presented here and elsewhere [Cappellini et al., 2019, Welker et al., 2019], demonstrating a shortening of average peptide length and overall peptide length distributions for older samples (Extended Data Fig. 5.9d).

Protein *in vivo* modification analysis

The existing literature on enamel and enamel proteome biomineralization describes three processes that are key to the maturation of the enamel proteome: protein hydrolysis by MMP20 and KLK4 [Chun et al., 2010, Yamakoshi et al., 2006, Iwata et al., 2007, Nagano et al., 2009], *in vivo* phosphorylation of serine residues [Cappellini et al., 2019, Welker et al., 2019, Tagliabracci et al., 2012] and expression of different isoforms of AMELX, AMBN and AMTN [Chun et al., 2010, Nagano et al., 2009, Fukae et al., 1996]. We sought to explore the presence of both *in vivo* protein hydrolysis and serine phosphorylation modifications in our Pleistocene hominin proteomes.

For protein hydrolysis by MMP20 and KLK4, we made use of the Atapuerca digestion-free dataset and the described locations of AMBN, AMELX and AMELY, and ENAM cleavage by MMP20 and KLK4 [Chun et al., 2010, Yamakoshi et al., 2006, Iwata et al., 2007, Nagano et al., 2009]. We compared the experimentally observed cleavage sites to a random cleavage model of each protein separately and tested whether the cleavage sites are present in a larger portion of PSMs in the ancient sample. Here we can indeed show an increased

presence of PSMs with termini at, or close to, known MMP20 and KLK4 cleavage locations (Extended Data Fig. 5.10). This corresponds with our observation that protein regions with continuous sequence coverage correspond to known proteolytic fragments after MMP20 and KLK4 activity (Extended Data Fig. 5.7).

Phosphorylation of serines (S), threonines (T) and tyrosines (Y) was assessed using Icelogo [Colaert et al., 2009] sequence motif analysis. This analysis was based on the MaxQuant results, from which only identified phosphorylation sites with a localization probability of ≥ 0.95 were selected. STY sites with no phosphorylation or localization probabilities ≤ 0.95 were taken as the non-phosphorylated background, and a sequence motif window of 7 amino acids on either side of the STY was selected.

Sequence motif analysis indicates a strong preference for the phosphorylation of S with a glutamic acid (E) on the +2 position (S-X-E motif) (Fig. 5.1a, b) in both hominin enamel proteomes. This substrate motif and the S-X-phosphorylated S motif are recognized by the kinase FAM20C, which is known to be active in vivo on extracellular proteins involved in biomineralization [Tagliabracci et al., 2012], and has previously been reported for ancient, non-hominin enamel proteomes as well [Cappellini et al., 2019, Welker et al., 2019].

To compare phosphorylation occupancy between the Dmanisi and Atapuerca enamel proteomes, we performed a separate MaxQuant database search (5.10 Supplementary Information) and restricted our analyses to amino acid positions covered by phosphorylated and nonphosphorylated peptides, observed in both hominins and quantified through label-free quantification.

Phylogenetic analysis

Comparison between the ancient protein sequences and modern reference proteins

We compared the reconstructed ancient protein sequences from the Dmanisi *H. erectus* and Atapuerca *H. antecessor* with protein sequences from great apes [De Manuel et al., 2016, Prado-Martinez et al., 2013], three Neanderthals [Prüfer et al., 2014, Castellano et al., 2014, Prüfer et al., 2017], a Denisovan [Meyer et al., 2012] and a panel of present-day humans, including 256 samples from the Simons Genome Diversity Panel [Mallick et al., 2016] and 41 high-coverage individuals from the 1000 Genomes Project [The 1000 Genomes Project, 2015]. Altogether, our reference data represent worldwide human and great ape variation (Supplementary Tables 5.9, 5.10). Additionally, we included protein sequences from macaque (*M. mulatta*) and gibbon (*N. leucogenys*) to root phylogenetic trees. The protein sequences were retrieved from the UniProt database or reconstructed from the reference whole-genome sequences as described in Supplementary Methods.

The ancient and reference protein sequences were aligned using mafft [Katoh, 2002]. We aligned the sequences of each protein separately and obtained an alignment for each of the ancient individuals independently (Supplementary Table 5.11). The isobaric amino acids leucine (L) and isoleucine (I) cannot be distinguished with the experimental procedure used for this study. Therefore, we have to take the following precautions to avoid unintentional sequence differences. If either I or L was present at a specific amino acid position in the

reference protein sequences, we replaced all corresponding amino acids in the ancient protein sequences with the amino acid that is present. Alternatively, if both amino acids are present in the reference protein sequence, we replace all I to L for all sequences. We used sequence information for seven proteins (ALB, AMBN, AMELX, AMELY, COL17 α 1, ENAM and MMP20) for the *H. antecessor* individual and six proteins for the *H. erectus* individual (ALB, AMBN, AMELX, COL17 α 1, ENAM and MMP20) with a total of 22.08% and 22.14% non-missing sites, respectively (Supplementary Table 5.11). We were able to recover a unique single amino acid polymorphism for *H. antecessor*; however, for *H. erectus* no unique single amino acid polymorphism was detected (Supplementary Tables 5.11-5.13, Supplementary Figs. 5.21-5.23).

Phylogenetic reconstruction

We built phylogenetic trees using our protein sequence alignments following three approaches: a maximum likelihood approach using PhyML v.3 [Guindon et al., 2010], and two Bayesian approaches using mrBayes [Ronquist et al., 2012] and BEAST [Bouckaert et al., 2019].

For the maximum likelihood approach, we built maximum likelihood trees for each protein independently and for a concatenated alignment consisting of all of the available protein sequences for each of the ancient samples (Supplementary Figs. 5.24, 5.25). We used PhyML v.3 and the parameters described in the Supplementary Information section 5.10.2.3.5a to build and optimize the tree topologies, branch length and substitutions rates for each of the alignments. Support for each bipartition was obtained based on 100 non-parametric bootstrap replicates. We evaluated the effect of significant missingness in the ancient samples on the inferred topology. Finally, we looked at the effect of varying which of the subset of present-day human samples was included in the tree (Supplementary Information section 5.10.2.3.5b, c).

For the Bayesian approach using mrBayes, to assess the robustness of the maximum likelihood inference results, we performed Bayesian phylogenetic inference on the basis of the concatenated alignments using mrBayes 3.2 and the parameters described in Supplementary Information section 5.10.2.3.5d (Extended Data Fig. 5.11, Supplementary Fig. 5.27). Bayesian inference was performed using the CIPRES Science Gateway [Miller et al., 2010].

For the Bayesian approach using BEAST, we used BEAST 2.5 to obtain a time calibrated tree for the seven proteins used for *H. antecessor*. For this analysis, we used concatenated alignments including the Neanderthals, the Denisovan, seven randomly chosen *H. sapiens* individuals and a single individual per great ape species. The alignment was partitioned by gene and a coalescent constant population model was used for the tree prior. The dates of the ancient samples included in the analysis (Vindija Neanderthal, 52 ka [Prüfer et al., 2017]; Altai Neanderthal, 112 ka [Prüfer et al., 2014]; Denisovan, 72 ka [Meyer et al., 2012] and *H. antecessor*, 860.5 ka [Duval et al., 2018]) were used as tip dates for calibration. For each partition, we used the Jones–Taylor–Thornton substitution model with four categories for the gamma parameter, for which we allowed the Markov chain Monte Carlo chain to sample the shape of the gamma distribution (with an exponentially distributed prior) and assigned independent clock models. Additionally, we set a prior for the divergence time of great

apes to 23.85 ± 2.5 Ma (normally distributed) [Besenbacher et al., 2019], and rooted the tree using the macaque (*M. mulatta*). The overall topology of the tree was estimated for the seven partitions jointly. The convergence of the algorithm was assessed using Tracer v.1.7.0 [Rambaut et al., 2018]. Finally, we repeated this analysis with 100 alignments, each of them consisting of 7 present-day humans chosen randomly. Although the topology within the clade consisting of present-day humans, Neanderthals and Denisovan was not consistent across the replicates, 99 of the replicates consistently place the *H. antecessor* sequence as an outgroup to this clade (Fig. 5.2a).

Further details on phylogenetic analysis and results can be found in the 5.10 Supplementary Information. Example MS/MS spectra from the MaxQuant search and overlapping sites of phylogenetic interest (single amino acid polymorphisms) for both hominins are included as Supplementary Data 1.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD014342. Generated ancient protein consensus sequences used for phylogenetic analysis for *H. antecessor* (Atapuerca) and *H. erectus* (Dmanisi) hominins can be found in the Supplementary Data 2, which is formatted as a .fasta file. Full protein sequence alignments used during phylogenetic analysis can be accessed via Figshare (<https://doi.org/10.6084/m9.figshare.9927074>). Amino acid racemization data are available online through the NOAA database. The wiNNer model can be accessed on GitHub (<https://github.com/cox-labs/wiNNer.git>).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2153-8>.

5.4 Results

The results show that the wiNner model retrained for randomly cleaved and heavily modified peptides provides a predictive performance similar to that of the wiNner model trained on modern, trypsin-digested samples, assuring accurate sequence identification for the phylogenetically informative peptides (median Pearson correlation coefficients of ≥ 0.76) (5.3 Methods, Supplementary Fig. 5.17, 5.10 Supplementary Information).

Protein recovery from the Dmanisi dentine sample was limited to sporadic collagen type I fragments, and therefore in-depth analysis of this material was not further pursued. By contrast, we recovered ancient proteomes from both hominin enamel samples. We found that the composition of these proteomes is similar to that of the recent human specimen that we processed as a positive control, as well as to previously published proteomes from ancient enamel [Cappellini et al., 2019, Stewart et al., 2017, Castiblanco et al., 2015, Asaka et al., 2009] (Extended Data Table 5.2, Supplementary Table 5.8). The enamel-specific proteins include amelogenin (both AMELX and AMELY isoforms), enamelin (ENAM), ameloblastin (AMBN), amelotin (AMTN) and the enamel-specific protease matrix metalloproteinase 20 (MMP20). Serum albumin (ALB) and collagens (COL1 α 1, COL1 α 2 and COL17 α 1) are also present. For the enamel-specific proteins, the peptide sequences that we retrieved cover approximately the same protein regions in all of the specimens that we analysed (Extended Data Fig. 5.7). Although destructive, our sampling of Pleistocene hominin teeth resulted in higher protein sequence coverage than acid-etching of Holocene enamel surfaces [Stewart et al., 2017, Porto et al., 2011] (Supplementary Fig. 5.18). The AMTN-specific peptides largely derive from a single sequence region involved in hydroxyapatite precipitation through the presence of phosphorylated serines [Gasse et al., 2015]. Finally, the observation of the AMELY-specific peptides (which is coded on the non-recombinant portion of the Y chromosome) demonstrates that the *H. antecessor* molar that we studied belonged to a male individual [Stewart et al., 2017] (Extended Data Fig. 5.8).

Besides proteome composition and sequence coverage, several further lines of evidence independently support the endogenous origin of the hominin enamel proteomes. Unlike exogenous trypsin, keratins and other human-skin contaminants that we identified, the enamel proteins have high deamidation rates (Extended Data Fig. 5.9) – above the rate observed for the recent human specimens (Supplementary Fig. 5.19). Both Pleistocene hominins have average peptide lengths that are shorter than those observed for our recent human controls (Extended Data Fig. 5.9d). The average peptide length is shorter in the Dmanisi hominin, but longer in the younger Atapuerca hominin (Extended Data Fig. 5.9d). By contrast, we observe that the peptide lengths in enamel from the Dmanisi hominin are indistinguishable from those of the faunal remains from the same site. Together, our protein data are therefore consistent with theoretical and experimental [Cappellini et al., 2019, Demarchi et al., 2016] expectations for samples of their relative age.

In addition to diagenetic modifications, we observe two kinds of in vivo modification in our recent and ancient enamel proteomes. First, we detect serine (S) phosphorylation within the S-X-E motif (Fig. 5.1a, b). This motif, as well as the S-X-phosphorylated S motif, is recognized by the FAM20C secreted kinase, which is active in the phosphorylation of extracellular proteins [Tagliabracci et al., 2012, Hu et al., 2005]. The presence of phosphoserine in

fossil enamel and its location in the S-X-E and/or S-X-phosphorylated S motifs has also previously been observed in other Pleistocene enamel proteomes [Cappellini et al., 2019, Glimcher et al., 1990]. Phosphorylation occupancy can be computed successfully for ancient and recent samples, and reveals differences in the ratios of phosphorylated peptides between samples (Fig. 5.1c, Supplementary Table 5.7). Second, the peptide populations that we retrieve primarily cover the ameloblastin, enamelin and amelogenin sequence regions, representing cleavage products deriving from in vivo activity of the proteases MMP20 and – subsequently – kallikrein 4 (KLK4) (Extended Data Fig. 5.7, 5.3 Methods). The peptide populations are also enriched in N and C termini that correspond to known MMP20 and KLK4 cleavage sites (Extended Data Fig. 5.10, Supplementary Fig. 5.20). FAM20C phosphorylation and MMP20 and KLK4 proteolysis are the two main processes that occur in vivo during enamel biomineralization. Our observation of products deriving from both processes opens up the possibility of studying in vivo processes of hominin tooth formation across the Pleistocene epoch.

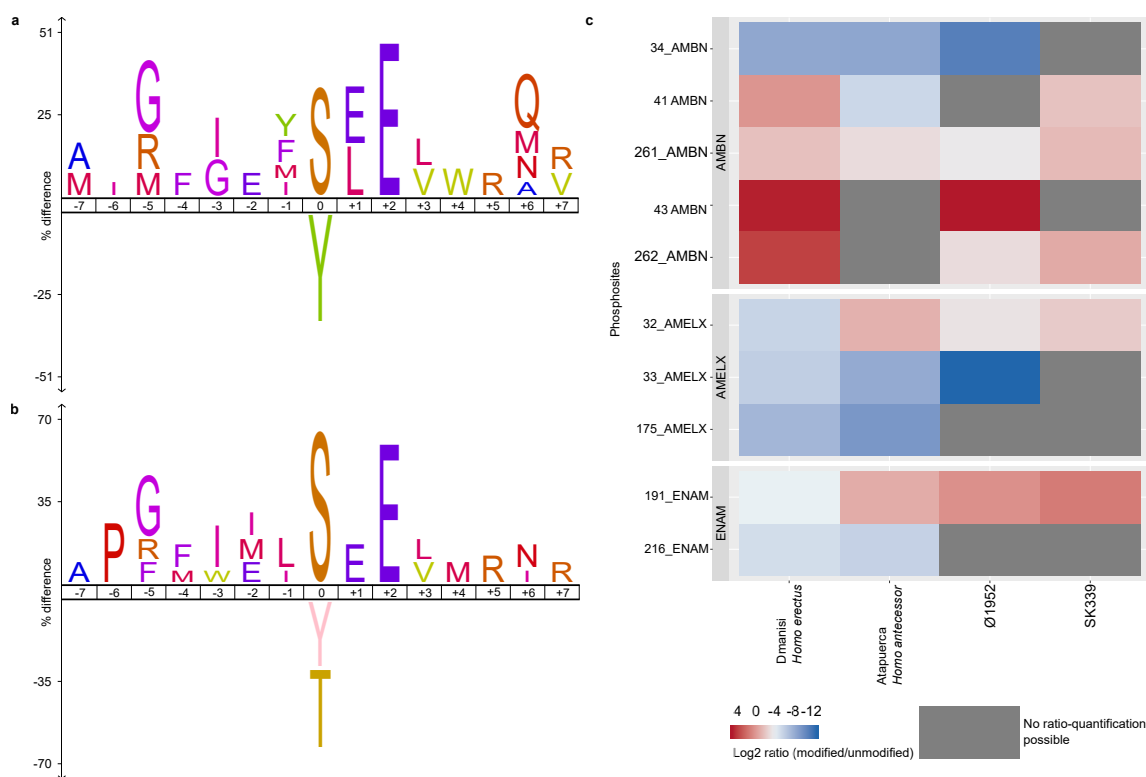


Figure 5.1: Phosphorylation of hominin enamel proteomes. **a**, Phosphorylation sequence motif analysis of *H. antecessor* specimen ATD6-92. **b**, Phosphorylation sequence motif analysis of *H. erectus* specimen D4163. **c**, Phosphorylation occupancy comparison, expressed as log₂-transformed summed intensity ratio of modified and unmodified peptides, for amino acid sites for which data are available for at least two specimens. y axis labels indicate the position of the phosphorylated amino acids for each protein (UniProt accession numbers Q9NP70 (AMBN), Q99217 (AMELX) and Q9NRM1 (ENAM)). SK339 denotes an archaeological specimen from a modern human, which is approximately three centuries old (see ‘Recent human control specimens’ in the 5.3 Methods for details).

Homo antecessor is known only from the Gran Dolina TD6 assemblage in Atapuerca [Bermúdez de Castro et al., 1997]. Its relationship with other European Middle Pleistocene fossils is heavily debated [Stringer, 2016, Hublin, 2009, Rightmire, 1998, Wagner et al., 2010, Martínón-Torres et al., 2007]. It remains contentious as to whether *H. antecessor* represents the last common ancestor of *H. sapiens*, Neanderthals and Denisovans [Bermúdez de Castro et al., 1997], or whether it represents a sister lineage to the last common ancestor of these species [Bermúdez de Castro et al., 2017, Gómez-Robles et al., 2013]. We address this issue by conducting phylogenetic analyses on the basis of our ancient protein sequences from *H. antecessor* (ATD6-92), a panel of present-day great ape genomes and protein sequences translated from archaic hominin genomes (5.3 Methods).

We built several phylogenetic trees using maximum likelihood and Bayesian methods (Fig. 5.2a, Supplementary Figs. 5.24 - 5.27). In these trees, the *H. antecessor* sequence represents a sister taxon that is closely related to, but not part of, the group composed of Late Pleistocene hominins for which molecular data are available (Fig. 5.2a, Supplementary Figs. 5.24, 5.26, 5.27). The enamel protein sequences do not resolve the relationships between *H. sapiens*, Neanderthals and Denisovans owing to the low number of informative single amino acid polymorphisms.

However, pairwise divergence of the amino acid sequences between *H. antecessor* and the clade containing *H. sapiens*, Neanderthals and the Denisovan is larger than the divergence between the members of this clade (Fig. 5.2b, Supplementary Fig. 5.23, 5.10 Supplementary Information). The concatenated gene tree may be subjected to incomplete lineage sorting, and we have too little sequence data to discard this possibility at the moment. However, if we use the concatenation of available gene trees as a best guess for the population tree, and assume that such a population tree is a good descriptor of the relationships among ancient hominins, then our results support the placement of *H. antecessor* as a closely related sister taxon of the last common ancestor of *H. sapiens*, Neanderthals and Denisovans. The phylogenetic position of *H. antecessor* agrees with a divergence of the *H. sapiens* and Neanderthal + Denisovan lineages between 550 and 765 ka [Meyer et al., 2016, Prüfer et al., 2014], as ATD6-92 has been dated to 772–949 ka [Duval et al., 2018]. This is further supported by recent reconsiderations of the morphology of *H. antecessor* in relation to Middle and Late Pleistocene hominins [Gómez-Robles et al., 2013].

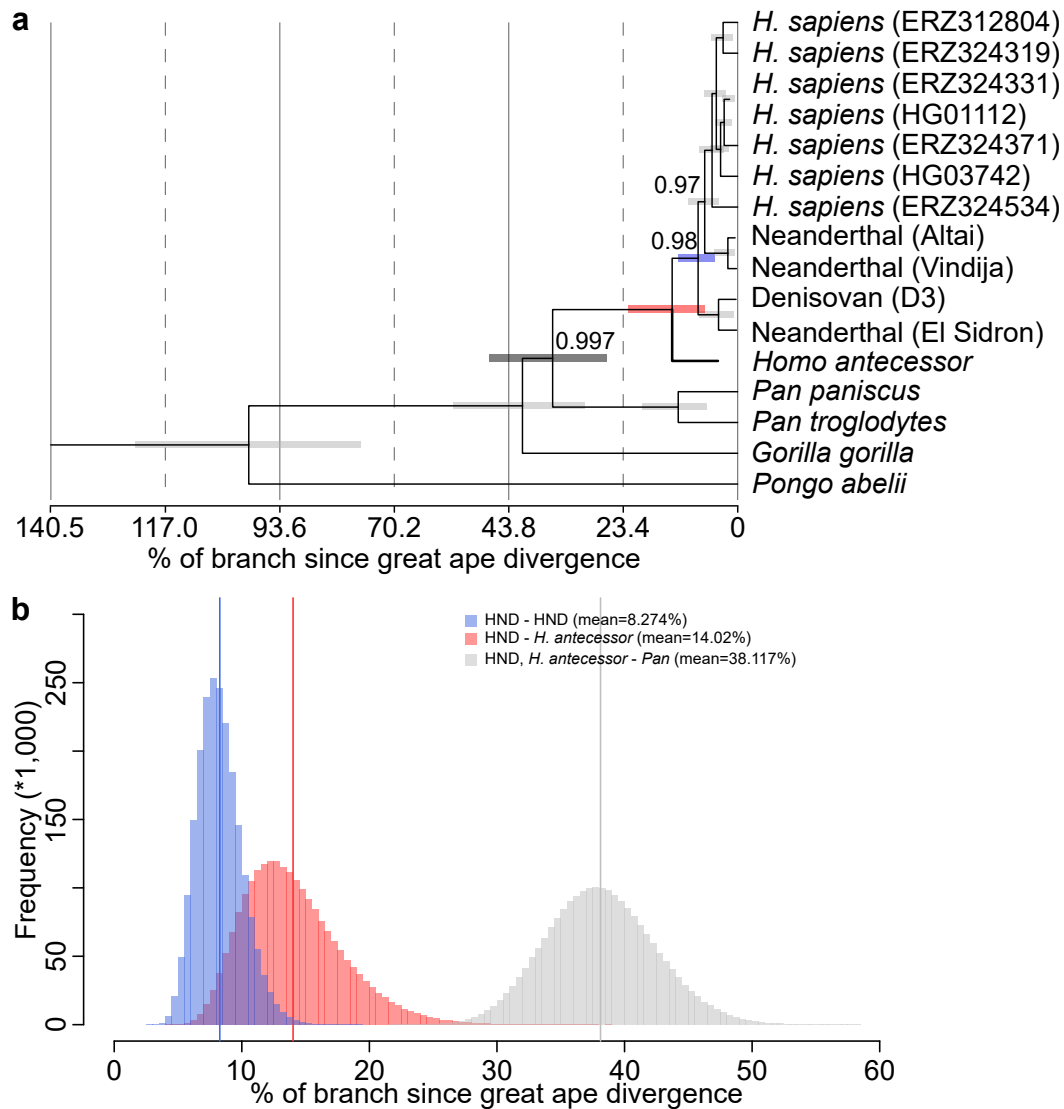


Figure 5.2: Phylogenetic analysis of *H. antecessor* ATD6-92. **a**, Maximum credibility tree estimated using BEAST and a concatenated alignment of seven protein sequences recovered for the ancient sample. Posterior Bayesian probabilities are indicated at nodes with a probability of ≤ 1 . Horizontal error bars at each node indicate the 95% highest posterior density intervals for the split time estimates. The position of *H. antecessor* is consistent with that obtained via maximum likelihood (Supplementary Fig. 5.24) and Bayesian (Supplementary Fig. 5.27) analyses. ERZ and HG codes in parentheses after *H. sapiens* refer to identifiers for data from the Simons Genome Diversity Panel [Mallick et al., 2016] and 1000 Genomes Project [The 1000 Genomes Project, 2015], respectively (see ‘Comparison between the ancient protein sequences and modern reference proteins’ in the 5.3 Methods for details). **b**, Histograms of the divergence times obtained for the split between *H. antecessor* and the *H. sapiens*, Neanderthal and Denisovan clade (HND) (red), the HND–HND split (blue), and the *Pan*–(HND + *H. antecessor*) split (grey). Divergence times in **a** and **b** are shown as a percentage of the time since the divergence of all great apes.

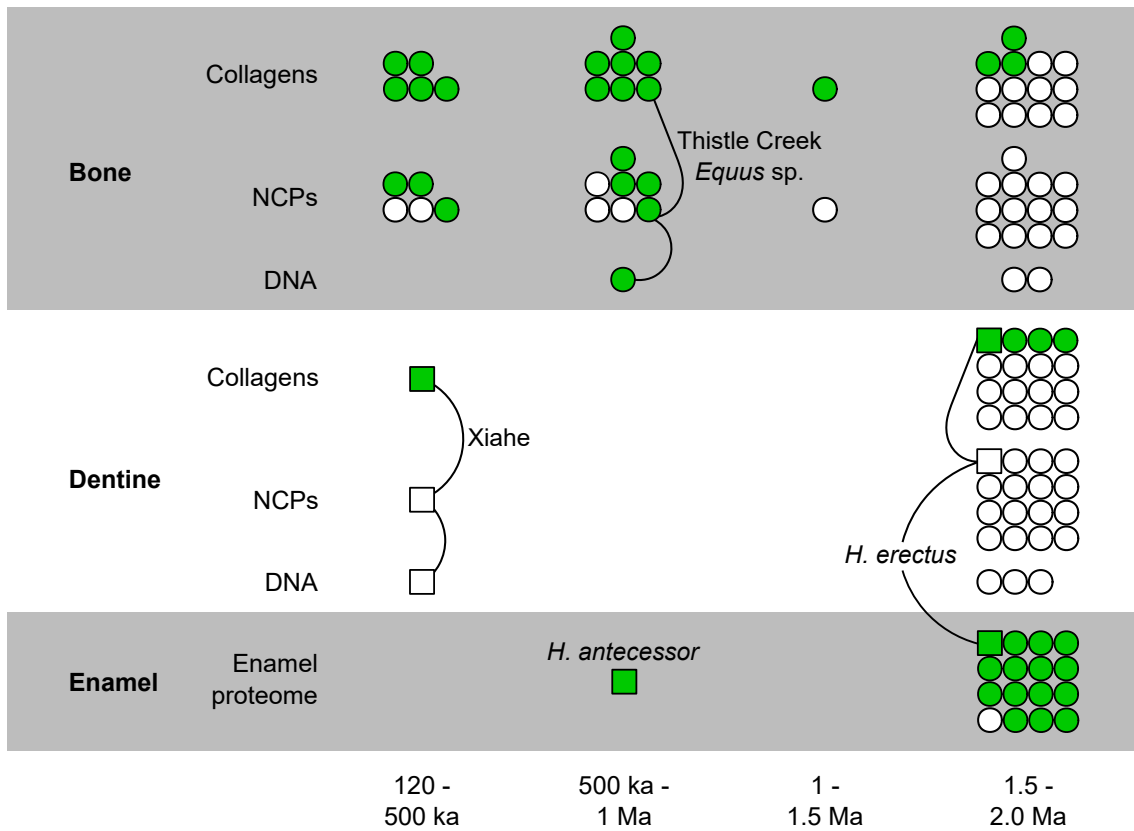


Figure 5.3: Skeletal proteome preservation in the Middle and Early Pleistocene epoch (0.12–2.6 Ma) For each sample, the presence (green) or absence (blank) of endogenous DNA, collagens, non-collagenous proteins (NCPs) or an enamel proteome is given. Only samples for which mammalian proteomes are published are considered [Cappellini et al., 2019, Chen et al., 2019, Welker et al., 2019, Welker et al., 2017, Hill et al., 2015, Wadsworth and Buckley, 2014, Orlando et al., 2013]. Hominin samples are indicated with squares, other mammalian samples are indicated with circles. Selected specimens have their separate molecular components joined, and are named. Xiahe refers to the Xiahe mandible [Chen et al., 2019]; the Thistle Creek *Equus* refers to a horse metapodial from the Canadian permafrost [Orlando et al., 2013].

5.5 Discussion

Homo antecessor has tentatively been proposed as the last common ancestor of Neanderthals and modern humans [Bermúdez de Castro et al., 1997]. The similarities observed between the modern-like mid-facial topography of *H. antecessor* and *H. sapiens* – including a modern pattern of coronal orientation of the infraorbital surface, the sloping and directionality of this plane, as well as the anterior flexion of the maxillary surface and arching of the zygomatic-alveolar crest – were key in this proposal [Bermúdez de Castro et al., 1997, Lacruz et al., 2019]. Additional studies of the face of ATD6-69 have confirmed that *H. antecessor* exhibits the oldest known modern-like face in the fossil record [Freidline et al., 2013, Lacruz et al., 2013]. The phylogenetic placement of *H. antecessor* implies that this modern-like face – as represented by *H. antecessor* – must have a considerably deep ancestry in the genus *Homo*. Findings made between 2003 and 2005 have shown that the *H. antecessor* hypodigm includes some features that were previously considered Neanderthal autapomorphies [Bermúdez de Castro et al., 2017]. Our results suggest that these features appeared in Early Pleistocene hominins, and were retained by Neanderthals and lost by modern humans.

By contrast, the phylogenetic tree built with the *H. erectus* specimen from Dmanisi has only moderate resolution (Extended Data Fig. 5.11, Supplementary Fig. 5.22), despite deeper shotgun protein sequencing for this specimen (Extended Data Table 5.1). This partly inconclusive result might be due to the shorter average peptide lengths compared to the Atapuerca *H. antecessor* specimen (Extended Data Fig. 5.9d, 5.3 Methods) and an absence of uniquely segregating single amino acid polymorphisms (Supplementary Table 5.11). Although our *H. erectus* data from Dmanisi demonstrate that ancient hominin proteins can be reliably obtained from the Early Pleistocene epoch, they also highlight the current limits of ancient protein analysis when applied to the phylogenetic placement of Early Pleistocene hominin remains.

Our dataset provides a unique molecular resource of hominin biomolecular sequences from Early and Middle Pleistocene hominins, and represents – to our knowledge – the oldest ancient hominin proteomes presented to date. Comparison of hominin and fauna proteomes from different skeletal tissues reveals that the dental enamel proteome outlasts dentine and bone proteome preservation (Fig. 5.3). Here the prolonged survival of hominin enamel proteomes is exploited to show that *H. antecessor* represents a hominin taxon closely related to the last common ancestor of *H. sapiens*, Neanderthals and Denisovans. In addition, our datasets demonstrate that *in vivo* proteome modifications, such as serine phosphorylation, survive over time scales of hundreds of thousands of years. Current research therefore suggests that dental enamel, the hardest tissue in the mammalian skeleton, is the material of choice for the analysis of hominin evolution in deep time.

5.6 Author's contribution

The phylogenetic relationships of early hominins, specifically of the Early Pleistocene epoch in Eurasia and the Middle Pleistocene epoch, are highly debated due to the lack of morphological features and fully degraded aDNA of fossil remains. Therefore, we performed an ancient protein analysis to an enamel fragment of *Homo antecessor* and to dentin and enamel of a *Homo erectus* molar.

My contribution to this publication was strongly focussed on the performed data analysis obtained from a LC-MS/MS experiment. First, I performed a MaxQuant data analysis, which included identification not only by the database search engine Andromeda, but also from the MaxNovo de-novo sequencing algorithm (Chapter 4 - Publication 2) and the dependent peptide algorithm. This analysis enabled the identification of peptides and protein sequences and their sequence variations.

Furthermore, I developed a post-processing pipeline that uses the obtained MaxQuant results for protein sequence reconstruction. The pipeline facilitates phylogenetic analysis by extracting sequence variations. For the development of sequence reconstruction, I aligned all identified sequences to a reference protein sequence. Based on this alignment, I was able to reconstruct a protein sequence that incorporates the identified and validated sequence variations, which were further validated and subsequently used for phylogenetic analysis.

Additionally, I performed the phosphorylation occupancy analysis to investigate the quantity of phosphorylation that occurred in vivo during tooth formation. Finally, for the validation of identified peptide sequences with our machine learning (Chapter 3 – Publication 1; wiNNer), I performed MaxQuant analysis of several ancient datasets to provide sufficient training data.

For *Homo antecessor*, we successfully obtained sufficient sequence coverage of the enamel-specific proteins to provide evidence that *Homo antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins, including modern humans, Neanderthals and Denisovans.

5.7 Additional information

Acknowledgement

F.W. is supported by a Marie Skłodowska Curie Individual Fellowship (no. 795569). E. Cappellini was supported by VILLUM FONDEN (no. 17649). E.W. is supported by the Lundbeck Foundation, the Danish National Research Foundation, the Novo Nordisk Foundation, the Carlsberg Foundation, KU2016 and the Wellcome Trust. Without the effort of the members of the Atapuerca research team during fieldwork, this work would have not been possible; we make a special mention of J. Rosell, who supervises the excavation of the TD6 level. The research of the Atapuerca project has been supported by the Dirección General de Investigación of the Ministerio de Ciencia, Innovación y Universidades (grant numbers PGC2018-093925-B-C31, C32, and C33); field seasons are supported by the Consejería de Cultura y Turismo of the Junta de Castilla y León and the Fundación Atapuerca. We acknowledge The Leakey Foundation through the personal support of G. Getty (2013) and D. Crook (2014–2016, 2018, and 2019) to M.M.-T., as well as F.W. (2017). Restoration and conservation work on the material have been carried out by P. Fernández-Colón and E. Lacasa from the Conservation and Restoration Area of CENIEH-ICTS and L. López-Polín from IPHES. The picture of the specimen ATD6-92 was made by M. Modesto-Mata. E. Cappellini, J.C., J.V.O. and P. Gutenbrunner are supported by the Marie Skłodowska-Curie European Training Network (ETN) TEMPERA, a project funded by the European Union's Framework Program for Research and Innovation Horizon 2020 (grant agreement no. 722606). Amino acid analyses were undertaken thanks to the Leverhulme Trust (PLP-2012-116) and NERC (NE/K500987/1). T.M.-B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social 'La Caixa' and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). C.L.-F. is supported by a FEDER-MINECO grant (PGC2018- 095931-B-100). M.K. was supported by the Postdoctoral Junior Leader Fellowship Programme from 'la Caixa' Banking Foundation (LCF/BQ/PR19/11700002). M.M. is supported by the Danish National Research Foundation award PROTEIOS (DNRF128). Work at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation from the Novo Nordisk Foundation (grant number NNF14CC0001). The CRG/UPF Proteomics Unit is part of the Spanish Infrastructure for Omics Technologies (ICTS OmicsTech) and it is a member of the ProteoRed PRB3 consortium, which is supported by grant PT17/0019 of the PE I+D+i 2013-2016 from the Instituto de Salud Carlos III (ISCIII) and ERDF. We acknowledge support from the Spanish Ministry of Science, Innovation and Universities, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208, and 'Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya' (2017SGR595). D.L. and A.M. are supported by the John Templeton Foundation (no. 52935) and by the Shota Rustaveli National Science Foundation of Georgia (no. FR-18-27262). We thank M. L. Schjellerup Jørkov for providing specimen Ø1952.

All author's contributions

E. Cappellini, E.W., J.M.B.d.C., D.L., C.L.-F. and F.W. designed the study. E. Cappellini, M.M., F.W., J.R.-M., R.R.J.-C., M.R.D., C.C. and M.d.M. performed experiments. E. Cap-

pellini, A.M., J.L.A., E. Carbonell, P. Gelabert, E.S., J.C., J.V.O., T.M.-B. and D.L. provided material, reagents or research infrastructure. F.W., J.R.-M., P. Gutenbrunner, S.T., E. Cappellini, F.R., M.M.-T., J.M.B.d.C., M.K., M.R.D., C.L.-F. and K.P. analysed data. F.W., E. Cappellini and J.M.B.d.C. wrote the manuscript with input from all other authors.

5.8 Extended data figures

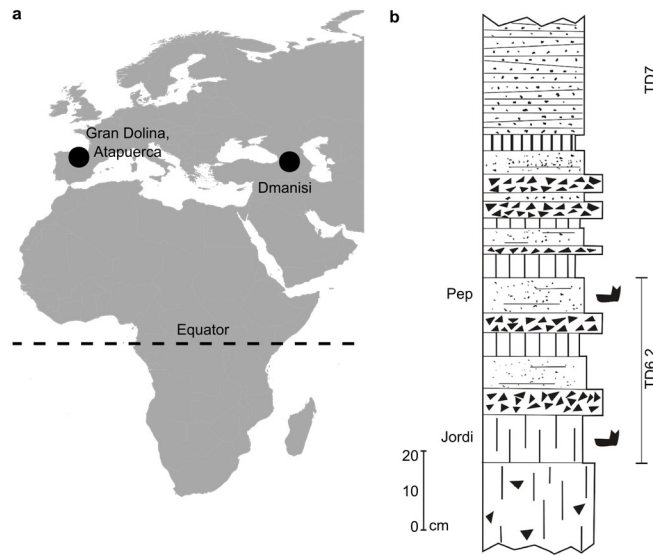


Figure 5.4: Location and stratigraphy of the hominin fossils studied. **a**, Geographic location of Gran Dolina and Dmanisi. Base map was generated using public domain data from www.naturalearthdata.com. **b**, Summarized stratigraphic profile of Gran Dolina, including the location of hominin fossils in layers ‘Pep’ and ‘Jordi’ of sublevel TD6.2.

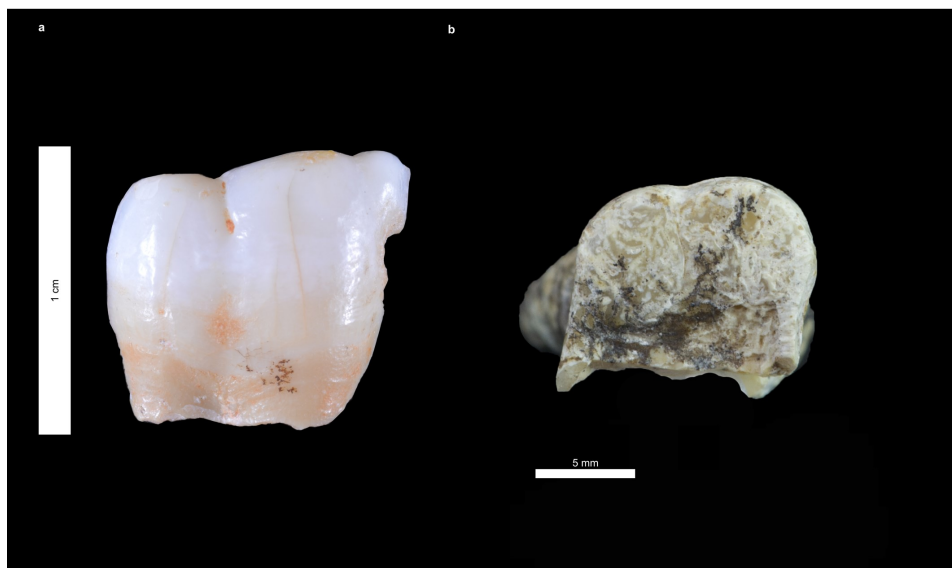


Figure 5.5: Hominin specimens studied. **a**, ATD6-92 in buccal view. The fragment represents a portion of a permanent lower left first or second molar. **b**, D4163 in occlusal view. The specimen is a fragmented right upper first molar. Scale bar differs between **a** and **b**.

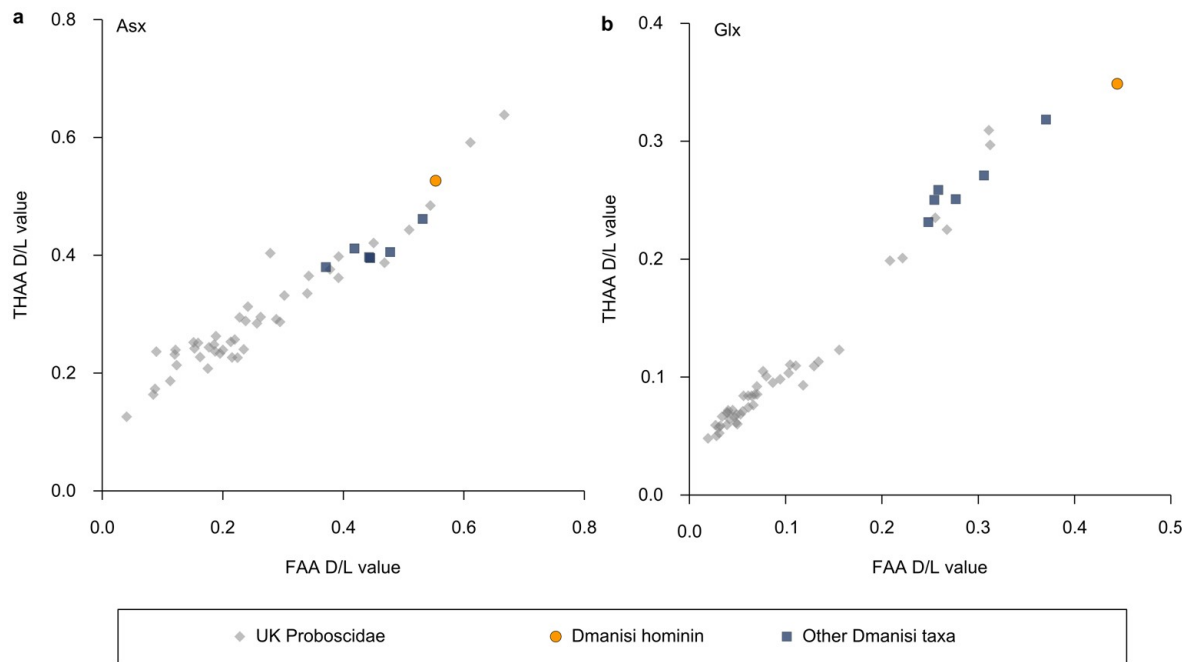


Figure 5.6: Amino acid racemization of D4163. **a, b,** The extent of intracrystalline racemization in enamel for the free amino acid (FAA) (x axis) fraction and the total hydrolysable amino acids (THAA) (y axis) fraction for aspartic acid plus asparagine (here denoted Asx) (**a**), and glutamic acid plus glutamine (here denoted Glx) (**b**), demonstrates endogenous amino acids breaking down within a closed system. The hominin value is displayed in relation to values for enamel samples from other fauna from Dmanisi [Cappellini et al., 2019] (blue squares) and a range of previously obtained Pleistocene and Pliocene Proboscidea from the UK [Dickinson et al., 2019] (grey diamonds). Fauna are shown for comparison, but different rates in their protein breakdown mean that they will show different extents of racemization. The x and y axis are on different scales.

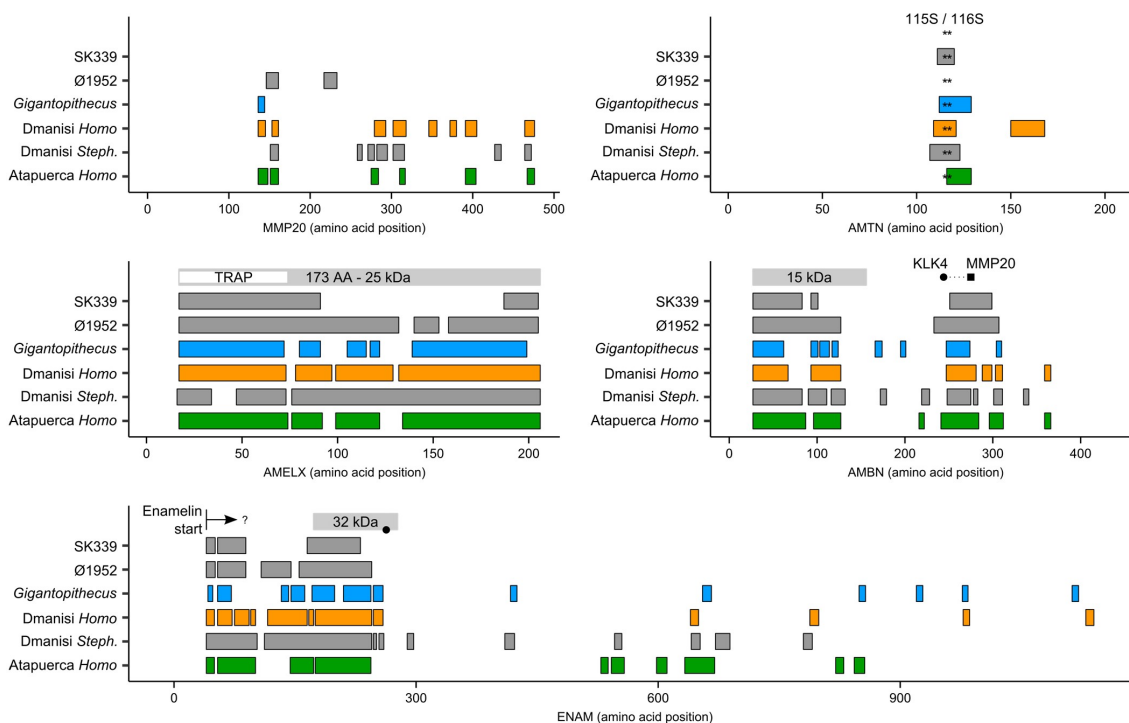


Figure 5.7: Sequence coverage for five enamel-specific proteins across Pleistocene samples and recent human controls. For each protein, the bars span protein positions covered, with positions remapped to the human reference proteome. The top row indicates the position of a selection of known MMP20 and KLK4 cleavage products of the enamel-specific proteins AMELX [Nagano et al., 2009], AMBN [Chun et al., 2010] and ENAM [Fukae et al., 1996]. Several in vivo proteolytic degradation fragments of ENAM share the same N terminus, but have unknown C termini [Yamakoshi et al., 2006]. Dotted line for AMBN indicates a putative cleavage product based on known MMP20 (squares) and KLK4 (circles) in vivo cleavage positions. For AMTN, serines (S) at positions 115 and 116 (indicated by asterisks) are conserved among vertebrates and involved in mineral-binding [Gasse et al., 2015]. Additional cleavage products as well as MMP20 and KLK4 cleavage sites are known in all enamel-specific proteins. SK339 [Stewart et al., 2017] and Ø1952 are two recent human control samples (5.3 Methods). AA, amino acids; *Steph.*, *Stephanorhinus* [Cappellini et al., 2019]; TRAP, tyrosine-rich amelogenin polypeptide.

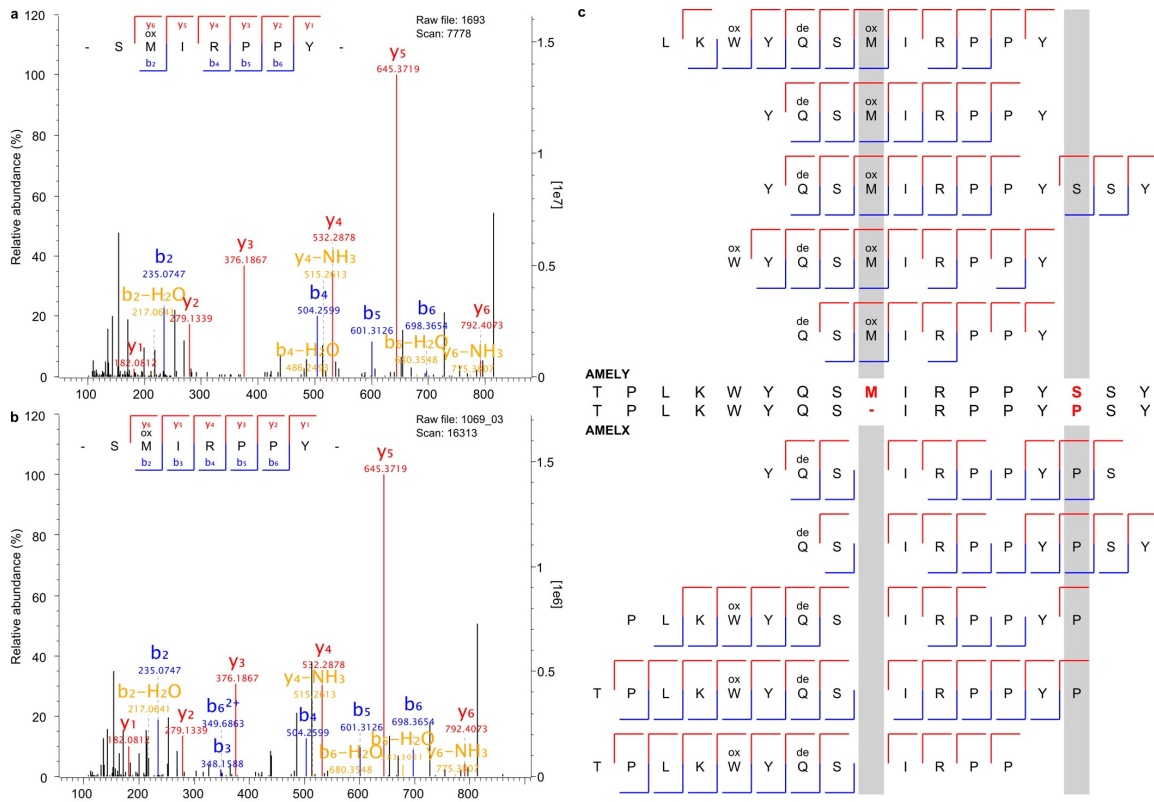


Figure 5.8: *Homo antecessor* specimen ATD6-92 represents a male hominin. **a**, Mass spectrum of an AMELY-specific peptide from the recent human control Ø1952. **b**, Mass spectrum of the same AMELY-specific peptide from *H. antecessor*. **c**, Alignment of a selection of AMELY- and AMELX-specific peptide fragment ion series deriving from *H. antecessor*. The alignment stretches along human AMELX isoform 1, positions 37 to 52 only (Uniprot accession numbers Q99217 (AMELX), Q99218 (AMELY)). See Supplementary Fig. 5.16 for another example of an AMELY-specific MS2 spectrum.

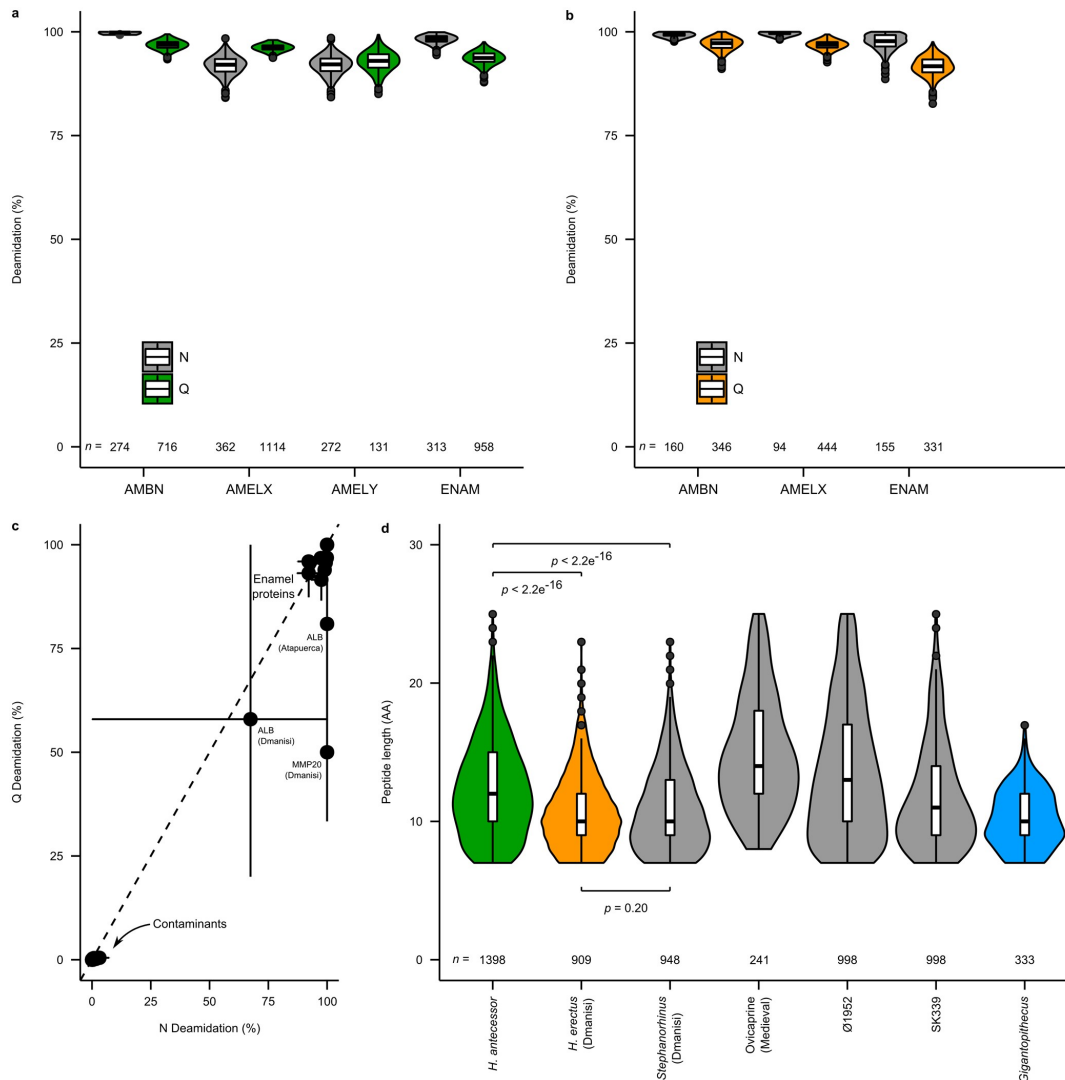


Figure 5.9: Enamel proteome damage. **a, b**, Glutamine (Q) and asparagine (N) deamidation of enamel-specific proteins from *H. antecessor* (Atapuerca) (**a**), and *H. erectus* (Dmanisi) (**b**). Values are based on 1,000 bootstrap replications of protein deamidation. **c**, Relationship between mean asparagine (N) and glutamine (Q) deamidation for all proteins in both the Atapuerca and Dmanisi hominin datasets. Error bars represent 95% confidence interval window of 1,000 bootstrap replications of protein deamidation. Dashed line is $x = y$. **d**, Peptide length distribution of *H. antecessor* (Atapuerca), *H. erectus* (Dmanisi), four previously published enamel proteomes [Cappellini et al., 2019, Welker et al., 2019, Stewart et al., 2017] and one additional human Medieval control sample (Ø1952). For **a, b** and **d**, the number of peptides (n) is given for each violin plot. The box plots within the violin plots define the range of the data (whiskers extend to $1.5\times$ the interquartile range), outliers (black dots, beyond $1.5\times$ the interquartile range), 25th and 75th percentiles (boxes), and medians (centre lines). P values of two-sided t -tests conducted between sample pairs are indicated. No independent replication of these experiments was performed.

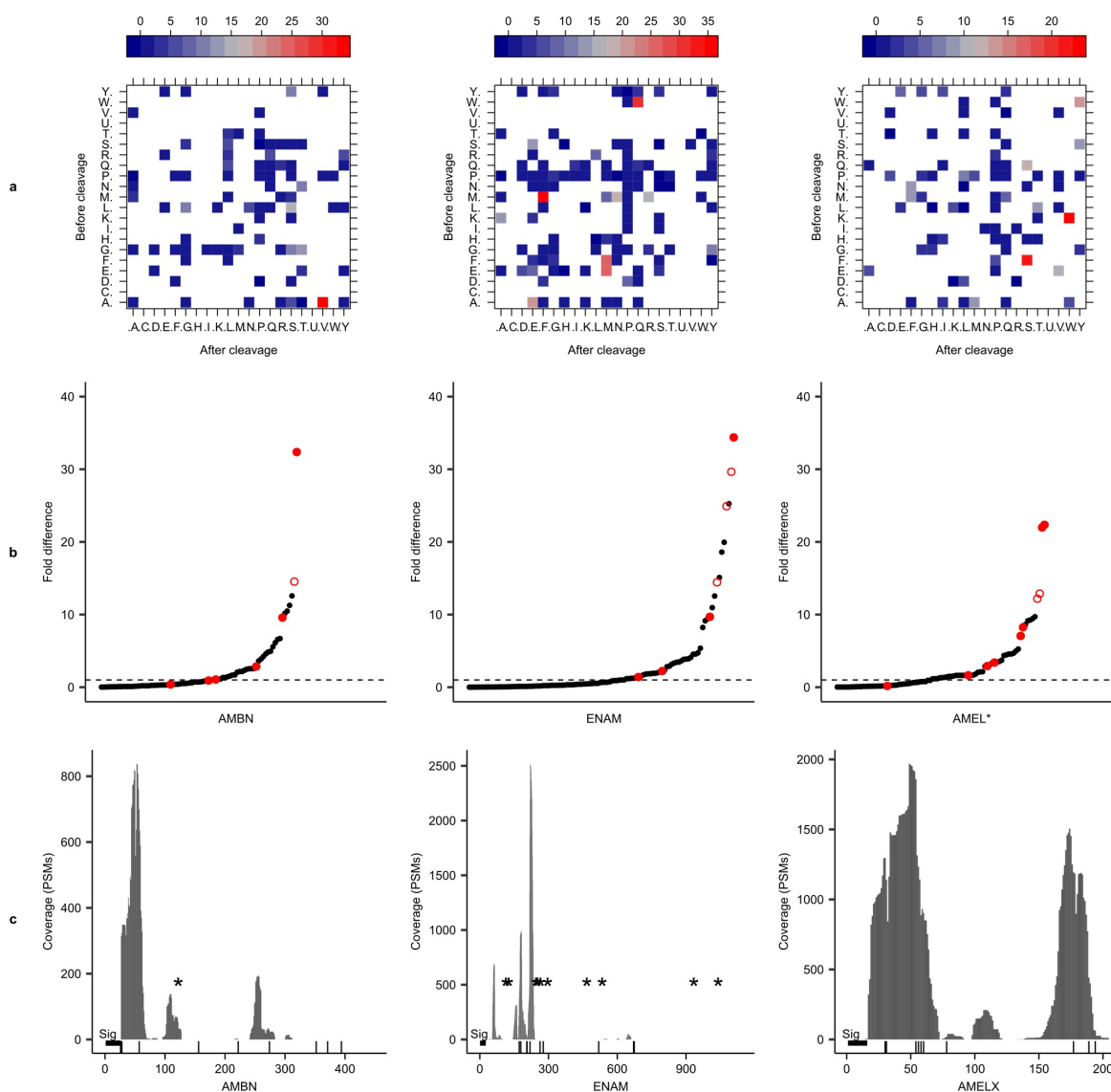


Figure 5.10: Survival of in vivo MMP20 and KLK4 cleavage sites in the Atapuerca enamel proteome. **a**, Experimentally observed cleavage matrices for ameloblastin (AMBN), enamelin (ENAM) and amelogenin (AMELX and AMELY) (Methods). Fold differences are colour-coded by comparing observed PSM cleavage frequencies to a random cleavage matrix for each protein separately [Chen et al., 2019]. **b**, Fold differences for all observed cleavage pairs per protein. Red filled circles represent MMP20, KLK4 and signal peptide cleavage sites mentioned in the literature [Yamakoshi et al., 2006, Iwata et al., 2007, Nagano et al., 2009, Fukae et al., 1996]. Red open circles indicate cleavage sites located up to two amino acid positions away from such sites. **c**, PSM coverage for each protein. The signal peptide (thick horizontal bar labelled 'sig'), known MMP20 and KLK4 cleavage sites (vertical bars), and O- and N-linked glycosylation sites (asterisks) are also indicated. For AMELX, peptide positions for all three known isoforms were remapped to the coordinates of isoform 3, which represents the longest isoform (UniProt accession Q99217-3). The x and y axes differ between the three panels of **c**.

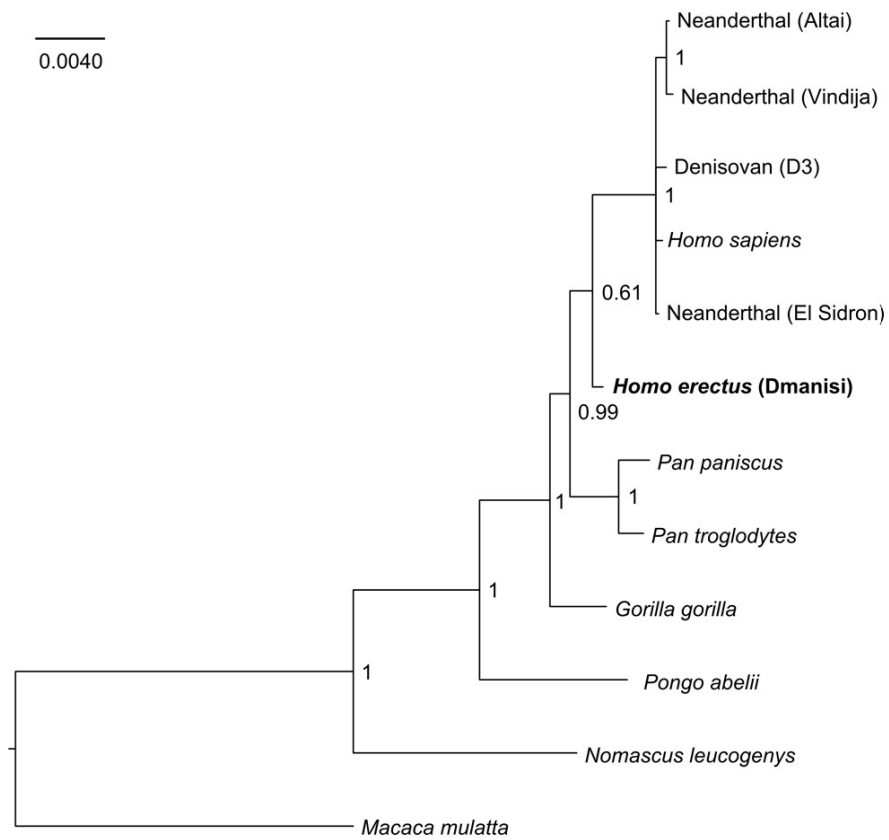


Figure 5.11: Phylogenetic position of D4163 through Bayesian analysis. *Nomascus leucogenys* and *M. mulatta* were used as outgroups.

5.9 Extended data tables

Stage Tip number	Tissue	Protein extraction method*	Mass Spectrometer	Mass Spectrometer location	Replicates
<i>Homo antecessor</i> , specimen ATD6-92, Atapuerca					
1069	Enamel	1	QE-HF	Copenhagen	4
1069	Enamel	1	Fusion Lumos	Barcelona	1
<i>Homo erectus</i> , specimen D4163, Dmanisi					
1138	Enamel	1	QE-HF	Copenhagen	2
1141	Enamel	2	QE-HF	Copenhagen	2
1138	Enamel	1	Fusion Lumos	Barcelona	1
1141	Enamel	2	Fusion Lumos	Barcelona	1
1139	Dentine	1	QE-HF	Copenhagen	2
1142	Dentine	2	QE-HF	Copenhagen	2
1139	Dentine	1	Fusion Lumos	Barcelona	1
1142	Dentine	2	Fusion Lumos	Barcelona	1
1386	Enamel	1	QE-HF	Copenhagen	1
1387	Enamel	3	QE-HF	Copenhagen	1
1388	Enamel	1	QE-HF	Copenhagen	1

Table 5.1: Extraction and mass spectrometry details of analyses conducted on both ancient hominin specimens. QE-HF, Q Exactive HF (or HF-X) hybrid quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). Fusion Lumos, LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific). *Extraction method 1: demineralization in HCl, with no subsequent proteolytic digestion. Extraction method 2: demineralization in HCl, reduction, alkylation and digestion with LysC and trypsin. Extraction method 3: demineralization in TFA, with no subsequent proteolytic digestion. See 5.10 Supplementary Information for further details.

Protein	Primary accession	MaxQuant				PEAKS				Combined Coverage (%)
		Peptides	Unique peptides	Coverage (AA)	Coverage (%)	Peptides	Unique peptides	Coverage (AA)	Coverage (%)	
<i>Homo antecessor</i> , specimen ATD6-92, Atapuerca										
AMELX	Q99217*	527	527	170 (0)	82.9	737	12	171 (1)	83.4	83.4
AMELY	Q99218*	220	86	131 (0)	63.6	341	6	141 (10)	68.4	68.4
AMBN	Q9NP70*	289	289	160 (3)	35.8	351	350	166 (9)	37.1	37.8
AMTN	Q6UX39	4	4	14	6.7	5	5	14	6.7	6.7
ENAM	Q9NRM1	424	424	233 (18)	20.4	586	586	245 (32)	21.5	23.0
MMP20	O60882	12	12	65 (0)	13.5	14	14	66 (1)	13.7	13.7
ALB	P02768	11	11	69 (17)	11.3	12	7	76 (24)	12.5	15.3
COL1a1	P02452	17	17	34 (21)	2.3	15	15	29 (16)	2.0	3.4
COL1a2	P08123	1	1	23	1.7	2	2	23	1.7	1.7
COL17a1	Q9UMD9	27	27	96 (24)	6.4	42	42	88 (16)	5.9	7.5
<i>Homo erectus</i> , specimen D4163, Dmanisi										
AMELX	Q99217*	357	357	182 (9)	88.8	297	297	173 (0)	84.4	88.8
AMBN	Q9NP70*	219	219	123 (1)	27.5	182	182	139 (17)	31.1	31.3
AMTN	Q6UX39	6	6	31 (13)	15.3	1	1	18 (0)	9.1	14.8
ENAM	Q9NRM1	306	306	224 (78)	19.6	293	293	160 (14)	14.0	20.8
MMP20	O60882	13	13	90 (15)	18.6	16	16	84 (9)	17.4	20.5
ALB	P02768	33	33	216 (12)	35.5	41	28	233 (29)	38.3	40.2
COL1a1	P02452	10	10	202 (44)	13.8	17	17	414 (256)	28.3	31.3
COL1a2	P08123	9	9	130 (3)	9.5	11	11	197 (66)	14.6	14.6
COL17a1	Q9UMD9	10	10	67 (45)	4.5	1	1	22 (0)	1.5	4.5

Table 5.2: Ancient hominin enamel proteome composition and coverage. Proteins are included only if two or more unique peptides were observed in either the PEAKS or MaxQuant searches. Primary accession refers to the *H. sapiens* entry in UniProt. Protein sequence coverage in the final column indicates the coverage obtained after combining PEAKS and MaxQuant peptide recovery. For ‘coverage (AA)’ columns, numbers in parentheses refer to the number of amino acid (AA) positions uniquely identified in PEAKS or MaxQuant searches. For AMELX and AMELY, coverage statistics combine counts for all isoforms present, whereas peptide counts refer only to the highest-ranking isoform or database entry. Direct comparisons between PEAKS and MaxQuant are uninformative owing to fundamental differences in spectral identification, protein and/or peptide assignment, and peptide counting approaches. *Combined coverage calculated against the longest isoforms for each protein.

5.10 Supplementary information

5.10.1 Anthropological background

5.10.1.1 Atapuerca

María Martín-Torres, Juan Luis Arsuaga, Eudald Carbonell, José María Bermúdez de Castro

The Sierra de Atapuerca is placed about 15 km east of the city of Burgos in Northern Spain ($3^{\circ} 41' 59.22''$ W, $42^{\circ} 20' 27''$ N). It is a small hill with an area of about 25 km² and a maximum altitude of 1082 meters, placed between the basins of the river Duero to the southwest and the river Ebro to the northeast. The Sierra de Atapuerca has a strategic position at the end of the so-called Bureba corridor (Fig. 5.12), a relatively narrow passage between the Cordillera Cantábrica range (to the north) and the Sierra de la Demanda Iberian range (located towards the south). The Bureba corridor was probably an important passage for the migrations of different species between Mediterranean areas and the interior of the Iberian Peninsula during the Pleistocene.

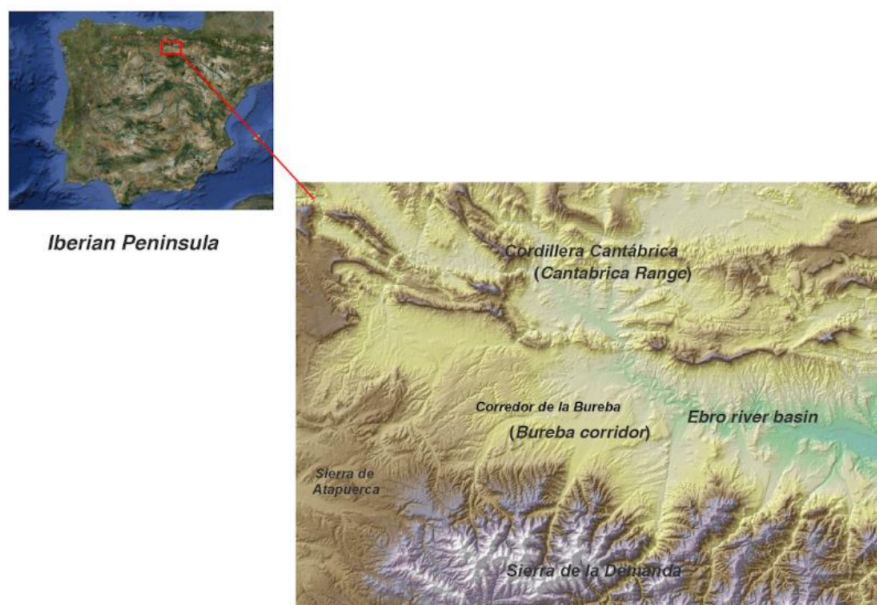


Figure 5.12: Location of the Sierra de Atapuerca in the Iberian Peninsula. Note that the Sierra de Atapuerca is located at the end of the Bureba corridor, which connects the basin of the Ebro River with the Castellana Plateau and the Duero river basin.

The Sierra de Atapuerca is made of Mesozoic limestone and from a structural point of view belongs to the Iberian range [Olivé et al., 1990, Pineda and Arce, 1997]. The study of the geomorphology revealed that the Sierra de Atapuerca is an NNW-SSE anticline verging on the NE. This structure forms an anticline ridge made up of Upper Cretaceous limestone, dolomites, and calcarenites [Pérez-González et al., 1995]. The core of the anticline is made up of Keuper (Upper Triassic) shales and Triassic-Jurassic dolomites

[Parés and Pérez-González, 1999]. The karst system, where the archaeological and paleontological sites are located, developed in the Upper Cretaceous. Unconformably overlying the Cretaceous limestone are calcareous conglomerates and red sandstone of Oligocene in age [Olivé et al., 1990, Pineda and Arce, 1997]. Additional information on the geomorphology of the Sierra de Atapuerca can be found in Zazo et al. [Zazo et al., 1983], and Parés and Pérez-González [Parés and Pérez-González, 1995]. Ortega-Martínez [Ortega-Martínez, 2009] carried out the main description of the multilevel karst system. The construction of a railway trench at the southern slope of the Sierra de Atapuerca at the end of the XIX century, as well as the opening of a quarry during the mid-XX century in the abandoned trench, revealed the presence of some karstic fillings containing archaeological and paleontological information. From 1978 onwards, a systematic excavation of these fillings has produced rich archaeological and fossil records. One of the most important cavities is Gran Dolina. This cavity is 28 meters deep and is fully infilled with Early and Middle Pleistocene sediments. The first description of the Gran Dolina infilling was carried out by Gil et al. [Gil et al., 1987]. These authors describe eleven main lithostratigraphic levels, TD1 to TD11 (from bottom to top; Fig. 5.13). Later, Parés and Pérez-González [Parés and Pérez-González, 1999] and Parés and Pérez-González [Parés and Pérez-González, 1995] published a more detailed description of the stratigraphy of Gran Dolina. The finding of Early Pleistocene human fossil remains in level TD6 was very important for the scientific Atapuerca project. This finding, made in 1994 during the excavation of an archaeological pit of less than six square meters [Carbonell et al., 1999a], initiated the systematic excavation and study of the fossiliferous filling of the Gran Dolina. An exhaustive stratigraphic study of the TD6 level has been carried out by Campaña et al. [Campaña et al., 2016].

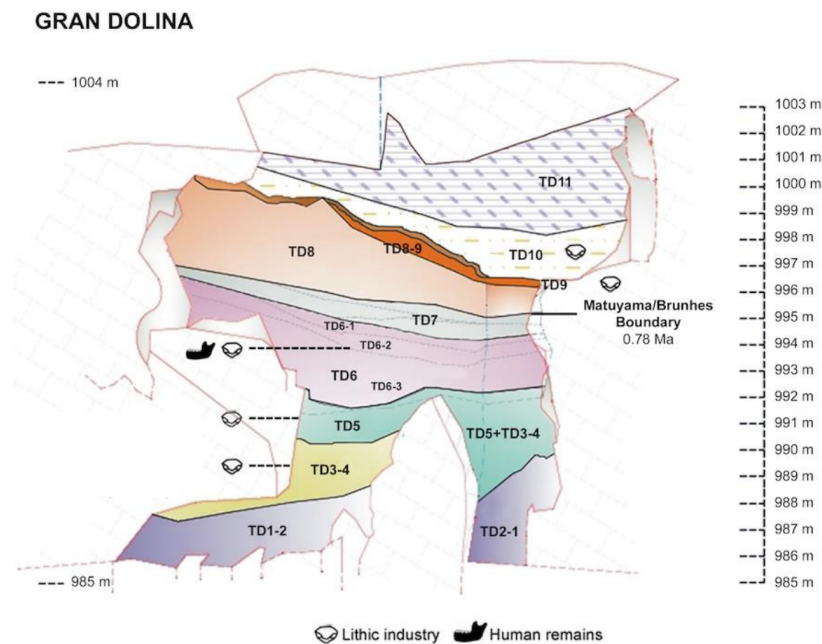


Figure 5.13: Stratigraphic profile of deposits and levels at the Gran Dolina cave site. A detailed view of the sublevels of TD6 can be found in 5.1b. This figure has been made by Jordi Rosell.

5.10.1.1.1 Atapuerca TD6.2 sublevel

The TD6 level has been divided in three sublevels: TD6.1, TD6.2, and TD6.3 [Campaña et al., 2016, Bermúdez de Castro et al., 2012] (Fig. 5.4). The human fossils, as well as more than 830 artefacts and several thousands of micro- and macromammal fossil remains [Cuenca-Bescós et al., 1999, van der Made, 1999, García and Arsuaga, 1999, Carbonell et al., 1999b, Bermúdez de Castro et al., 1999], come from the sublevel TD6.2 [Parés and Pérez-González, 1999, Parés and Pérez-González, 1995]. Parés and Pérez-González [Parés and Pérez-González, 1999, Parés and Pérez-González, 1995] and more recently Álvarez Posada [Álvarez-Posada et al., 2018] observed a polarity reversal between TD7 and TD8, interpreted as the Matuyama/Brunhes boundary, meaning that levels TD8 to TD11 were deposited during the Middle Pleistocene, whereas levels TD7 to TD1 were deposited during the Early Pleistocene. The combination of the paleomagnetic data and ESR/U-series ages suggests an age range between 0.78 and 0.86 million years ago (Ma) for the TD6 level [Falguères et al., 1999]. Thermoluminescence (TL) dates on samples taken at the TD7 level, one meter below the Brunhes/Matuyama boundary, give a weighted mean age of 0.96 ± 0.12 million years ago (Ma) for TD7 [Berger et al., 2008]. The ESR dating applied to optically bleached quartz grains from TD6 gives dates between 0.60 ± 0.09 Ma and 0.95 ± 0.09 Ma [Moreno et al., 2015]. These authors also obtained dates of 0.73 ± 0.13 Ma and 0.85 ± 0.14 Ma for the TD7 level, from samples taken under the Matuyama/Brunhes boundary. Using thermally transferred OSL (TT-OSL) dating of individual quartz grains, Arnold et al. [Arnold and Demuro, 2015] obtained a weighted mean age of 0.84 ± 0.06 Ma for the TD6 level. Arnold and Demuro [Arnold et al., 2015] have undertaken a series of TT-OSL suitability assessments on known-age samples from TD6. Using this method, they obtained a weighted average age of 0.85 ± 0.04 Ma for TD6.3. The first direct Electron Spin Resonance (ESR) dating study of *Homo antecessor* using one hominin tooth has provided an age estimate ranging from 0.72 Ma to 0.95 Ma [Duval et al., 2018]. Finally, a recent paleomagnetic study of the interior facies of TD1 places the TD6.2 hominins between the Matuyama/Brunhes boundary and the Jaramillo subchron [Parés et al., 2018]. Summarizing, and taking into account the biostratigraphic information from TD6 [Cuenca-Bescós et al., 1999, Cuenca-Bescós et al., 2015], we consider that the TD6 hominins can be assigned to the Marine Isotope Stage 21 (MIS 21).

5.10.1.1.2 ATD6-92

We sampled an isolated lower left first or second molar (ATD6-92). The specimen was excavated from square F14, sublevel TD6.2, in 2004. The fragment contains both enamel and dentine, although the root and dentine interior is largely absent. ATD6-92 derives from a stratigraphic unit containing over one hundred and seventy hominin fossils, all of which have been attributed to *Homo antecessor*. The specimen has been μ CT scanned previously, as well as directly dated using a combination of Electron Spin Resonance (ESR), LA-ICP-MS U-series analysis, and bulk U-series dating [Duval et al., 2018]. The dating analysis constrains the age of the tooth to between 624-949 thousand years ago (ka). As sublevel TD6.2 has reversed polarity, as indicated by paleomagnetic data, this age range can be shortened to 772-949 ka, and is in full agreement with other lines of chronological evidence for Gran Dolina TD6 [Duval et al., 2018, Falguères et al., 1999, Arnold and Demuro, 2015] (see above).

5.10.1.2 Dmanisi

Ann Margvelashvili, David Lordkipanidze

The archaeological and paleontological site of Dmanisi, Georgia, is located in the South Caucasus, 55 km southwest of Tbilisi (41°20'10"N and 44°20'38"E). The site is situated on the promontory that overlooks the confluence of the Mashavera and Pinesauri rivers. Lower Pleistocene deposits are set between the Medieval ruins and above the 1.85-Ma Mashavera Basalt which originated from the Javakheti volcanic highland. Hominin fossils attributed to *Homo erectus ergaster georgicus* (Dmanisi) derive from the layer B1, which is dated to between 1.78 and 1.76 Ma, based on ⁴⁰Ar/³⁹Ar dates, paleomagnetism, and paleontologic constraints [Ferring et al., 2011]. Besides the uniquely well-preserved and numerous hominin fossils, the site has produced a rich paleontological assemblage, from some of which the recovery of ancient enamel proteins was possible. These specimens were discovered in different excavation blocks and/or have different preservation conditions [Cappellini et al., 2019].

5.10.1.2.1 D4163

The isolated molar D4163 was discovered in M6, an excavation area located about 50 m northwest from Block II, where the majority of the hominin bones have been recovered. The excavations of M6 ceased after 2008, when the section reached the Mashavera basalt. The tooth was found in layer B1. This layer is dated to 1.77 Ma and correlates with the earliest Upper Matuyama chron, as it displays reversed polarity [Ferring et al., 2011]. B1 ashes of all excavation blocks in Dmanisi have distinctive laminated calcretes – retarded or arrested water percolations - whereas these calcretes are absent in M6 [Ferring et al., 2011].

D4163 is a right upper first molar. It is poorly preserved, with half of the crown missing and the third root broken off (Fig. 5.5b). The tooth is worn down (grade 4-5 according to Molnar 1971 [Molnar, 1971]). There is also the clear presence of taphonomic modifications of the dental surfaces. Together, these have resulted in the obliteration of the cusp shapes of the occlusal surface. Therefore, the fissure morphology of the occlusal surface is not observable, as the grooves are absent due to wear, taphonomy, and missing parts of the crown. The occlusal outline is asymmetrical, as far as can be assessed from the preservation state. The tooth is rather rhomboidal, instead of square. The remnants of a flat wear patch (contact point) are present on the distal surface of the crown. A more concaved wear patch can be traced on the mesial surface as well. Only two major cusps are present (paracone and metacone) with a faint buccal groove running between them. A shallow central fossa is present as well. The paracone is larger than the metacone, which is different from the other Dmanisi first molars. However, even though half of the crown is missing, it can be assumed that the bucco-lingual dimensions are larger than the mesio-distal dimensions. This pattern resembles the pattern observed for the first and second molars of the D4500 cranium [Lordkipanidze et al., 2013]. Similarity between the D4500 molars and D4163 also include the massive roots present on all these specimens. The roots slightly angle distally, and trifurcation is in the cervical third of the root. The roots are set widely apart from each other. The largest root (lingual/palatinal) is broken off. The bucco-distal root is the longest, whereas the bucco-mesial root is shorter, more triangular, and compressed mesio-distally. The molar does not fit to any current Dmanisi individuals or their associated mandibles. It is therefore likely that D4163 belongs to another *Homo erectus* (Dmanisi) individual.

5.10.2 Supplementary methods and results

5.10.2.1 Amino Acid Racemization

Marc R. Dickinson, Kirsty Penkman

5.10.2.1.1 Amino acid racemization methods

Chiral amino acid analysis was undertaken on one Pleistocene sample from the hominin tooth from Dmanisi to test the endogeneity of the enamel protein through its degradation patterns. The current technique of amino acid analysis developed for geochronological purposes [Penkman et al., 2008] combines a reverse-phase high-pressure liquid chromatography (RP-HPLC) method of analysis [Kaufman and Manley, 1998] with the isolation of an 'intracrystalline' fraction of amino acids by bleach treatment [Sykes et al., 1995]. This combination of techniques results in the analysis of D/L values of multiple amino acids from the chemically protected (closed system) protein fraction within the biomineral, thereby enabling both decreased sample mass and increased reliability of the analysis.

The tooth chip was separated into the enamel and dentine portions and each powdered with an agate pestle and mortar. All samples were prepared using modified procedures of Penkman et al. [Penkman et al., 2008], but optimized for enamel, using a bleach time of 72 hours to isolate the intra-crystalline protein [Dickinson et al., 2019]. Two subsamples were analyzed from each portion: one fraction was directly demineralized and the free amino acids analyzed (referred to as the 'free' amino acids, FAA, F), and the second was treated to release the peptide-bound amino acids, thus yielding the 'total hydrolysable' amino acid fraction (THAA, H*). After demineralization of the enamel, the pH of the solution was raised with KOH and then centrifuged for 5 min at 13,000 rpm, whereupon a biphasic solution formed [Dickinson et al., 2019]. The supernatant was extracted and dried via centrifugal evaporation. Samples were analyzed in duplicate by RP-HPLC, with standards and blanks run alongside samples. During preparative hydrolysis, both asparagine (Asn) and glutamine (Gln) undergo rapid irreversible deamination to aspartic acid (Asp) and glutamic acid (Glu), respectively [Hill, 1965]. It is therefore not possible to distinguish between the acidic amino acids and their derivatives and they are reported together as Asx and Glx respectively. The secondary amino acids proline and hydroxyproline are not detectable in the chiral amino acid analysis used here, and are therefore not reported below. The DL ratios of aspartic acid/asparagine, glutamic acid/glutamine, phenylalanine and alanine (D/L Asx, Glx, Phe and Ala) are assessed to provide an overall estimate of intra-crystalline protein decomposition (IcPD). In a closed system, the amino acid ratios of the FAA and the THAA subsamples should be highly correlated, enabling the recognition of compromised samples (e.g. Preece & Penkman, 2005 [Preece and Penkman, 2005]; Dickinson et al., 2019 [Dickinson et al., 2019]). The D/L of all the amino acids will increase with time, but each amino acid racemises at a different rate, therefore providing different resolution over different timescales.

5.10.2.1.2 Amino acid racemization results

The amino acid composition of the hominin enamel differs from dentine, but closely resembles that of enamel from other taxa from Dmanisi [Cappellini et al., 2019] (e.g. *Megacerini*; Fig. 5.14), indicating the amino acids studied are derived from the same proteins entrapped within the intra-crystalline fraction of enamel. In a closed system, all of the products of protein degradation should be retained, so the racemization of the FAA and THAA should be highly correlated over geological time. More labile amino acids, such as those that are unbound (FAA) are more likely to leach out of an open system, so open system behavior tends to result in data points falling away from the general closed system trend. Unfortunately, there is currently no directly comparative enamel data from the same species, but we have compared the hominin enamel both to other species from Dmanisi, and to Proboscidea data from the UK, for which there is a larger dataset (Extended Data Fig. 5.6).

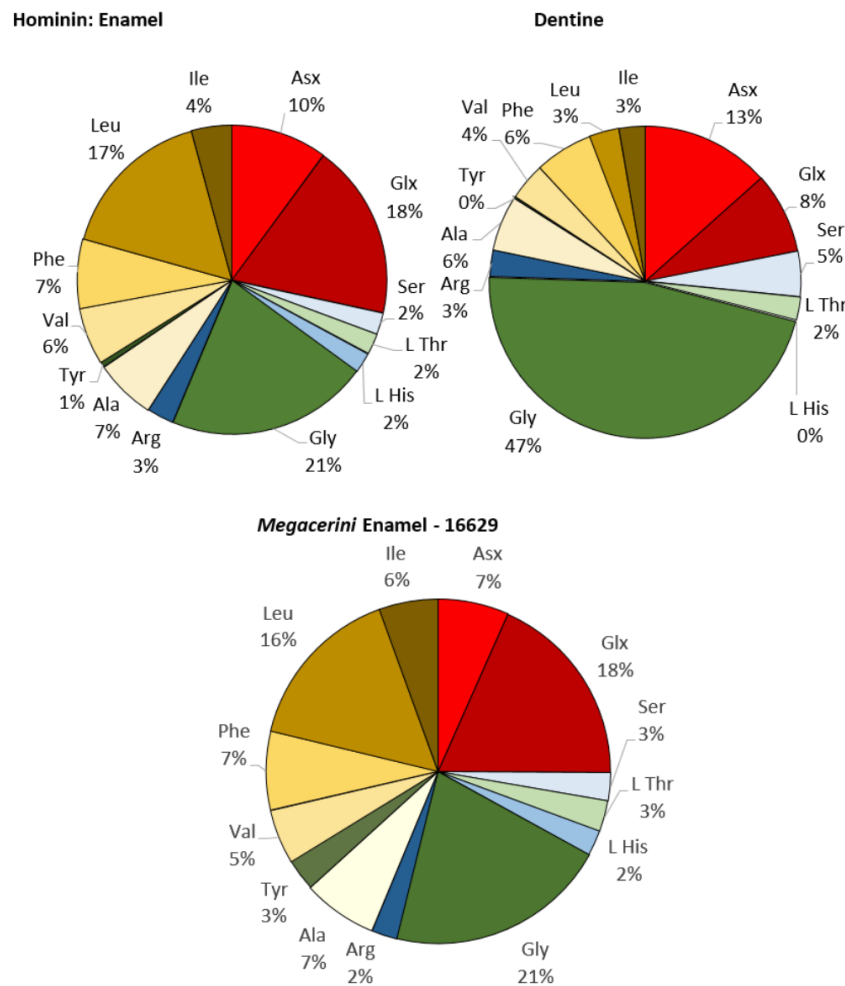


Figure 5.14: THAA amino acid compositions of enamel and dentine from the hominin tooth D4163. A profile of a *Megacerini* enamel sample from Dmanisi is shown for comparison (taken from Cappellini et al., 2019 [Cappellini et al., 2019]). The amino acid composition of the hominin enamel closely resembles that of the *Megacerini* enamel.

Whilst taxonomic effects influence the rates of racemization (so different taxa will have a different extent of racemization, despite being the same age), it is likely that the relative relationship between the racemization of FAA and THAA is similar enough between taxa to identify if significant leaching/contamination of the original protein has occurred. The FAA and THAA racemization values from the hominin enamel plot along the same trajectory as those from Proboscidea enamel, indicating that the intra-crystalline enamel amino acids from the hominin tooth are endogenous and show closed system behavior. The extent of racemization in all four amino acids studied is comparable to other taxa from the Dmanisi deposits of the same age (Extended Data Fig. 5.6), but slightly higher, indicating that hominin enamel amino acids racemise at a faster rate. This fast rate of degradation is also consistent with the high percentage of FAA, implying higher rates of hydrolysis (Supplementary Fig. 5.15). The extent of racemization and hydrolysis in the hominin enamel indicates a great antiquity, which is consistent with 1.8 Ma amino acids from this region. In summary, the protein composition, level of peptide bond hydrolysis, and extent of racemization in the hominin sample is consistent with the isolation of a closed system of endogenous protein from the intracrystalline fraction of the enamel.

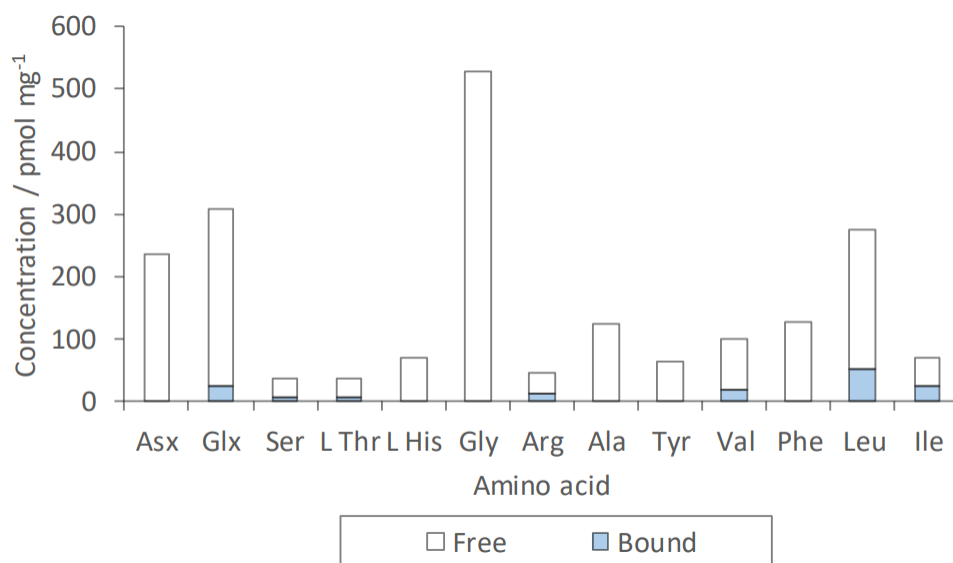


Figure 5.15: Absolute concentrations of the free and bound amino acids. Bound amino acid concentrations have been calculated by subtracting the FAA from the THAA concentrations. If the calculation results in a negative number, the bound concentration is assumed to be zero; this approach is likely to slightly overestimate the %FAA due to amino acid breakdown during the hydrolysis step for the THAA.

5.10.2.2 Proteomic Analysis

Frido Welker, Petra Gutenbrunner, Shivani Tiwary, Meaghan Mackie, Rosa Rakownikov, Jessie Christensen, Cristina Chiva, Pere Gelabert, Eduard Sabido, Jurgen Cox, Jesper V. Olsen, Enrico Cappellini

5.10.2.2.1 Protein extraction

We attempted protein analysis on enamel fragments from both hominins and our recent human control (Ø1952), and in addition also conducted protein extractions on a dentine sample from the Dmanisi hominin. Enamel powders might contain traces of dentine, as no dedicated chemical or physical separation was attempted after drilling. We did not perform sub-superficial acid-etching of exposed enamel surfaces [Stewart et al., 2017, Porto et al., 2011, Stewart et al., 2017]. Three separate protocols were utilized. See Extended Data Table 5.1 for further details on the division of extraction approaches across samples.

Extraction method 1. Approximately 250 mg of enamel was demineralized in 1.2 M HCl at 4°C, changing and saving the HCl supernatant every 24 hours. The peptides suspended in the HCl demineralization fractions were then directly cleaned, concentrated and immobilized, skipping reduction, alkylation and digestion steps, using in house assembled C18 solid-phase extraction Stage-Tips, using previously published methods [Cappellini et al., 2019].

Extraction method 2. After the demineralization of the Dmanisi sample from Extraction method 1, there were some remnants of the sample that did not appear to be mineral, and therefore digestion on this fraction was performed using a guanidine hydrochloride (GuHCl) solution, in case some protein could be extracted, following Mackie et al. [Mackie et al., 2018]. In short, the sample was reduced and alkylated using 2-Chloroacetamide (CAA) and Tris (2-carboxyethyl) phosphine (TCEP) within a GuHCl solution and incubated at 80°C. Subsequently, the sample was digested with the enzymes rLysC and Trypsin (Promega) overnight. Samples were then acidified to under pH 2 using 10% TFA in ultrapure water. Peptides were collected and cleaned on C18 Stage-Tips, as in method 1.

Extraction method 3. A single sample of the Dmanisi *Homo erectus* was demineralized in 10% TFA Peptide clean-up on C18 StageTips followed the details outlined for Extraction method 1.

5.10.2.2.2 LC-MS/MS analysis

Protein extractions were analyzed in two independent laboratories. First, extracts were all analyzed at the Novo Nordisk Centre for Protein Research, University of Copenhagen (Denmark). Second, a significant subset of extracts were also analyzed at the Proteomics Unit of the Centre for Genomic Regulation, Pompeu Fabra University (Barcelona, Spain). Several of the extracts were also analyzed multiple times on the same instrument at the same facility, providing further depth to our proteomic data (Extended Data Table 5.1). For all analyses, injection blanks preceded and followed sample injections to minimize the risk of sample carry-over between runs. Furthermore, whenever possible, ancient enamel proteomes were only injected after extensive cleaning of the MS.

Copenhagen (Denmark). Samples were analysed using in an EASY-nLC 1200 (Thermo Fisher Scientific (Proxeon), Odense, Denmark) coupled to either a Q-Exactive HF or HF-X (Thermo Fisher Scientific, Bremen, Germany) orbitrap mass spectrometer. The peptides were eluted from the StageTips using 20 µL 40% acetonitrile (ACN) and subsequently 10 µL

60% ACN into a 96 well plate. For the Atapuerca sample, a quarter of the eluted sample was dried and sent to Barcelona. In Copenhagen, after elution, all samples were vacuum centrifuged at 40°C until approximately 3 µL was left. The remainder of the Atapuerca samples were then rehydrated with 10 µL of 0.1% TFA, 5% ACN. The Dmanisi samples were rehydrated with 9 µL of 0.1% TFA, 5% ACN, 4 µL of which being saved and sent to Spain. Different LC and MS methods were used in Copenhagen based on the instrument and column length. In addition, for the Dmanisi runs, a short test run of each sample was also performed before the main analysis. Columns were either 15 or 50 cm PicoFrit® columns (75 µm inner diameter) in-house packed with 1.9 µm C18 beads (Reprosil-AQ Pur, Dr. Maisch). Buffer B consisted of 80% acetonitrile and buffer A of milliQ water, both containing 0.1% TFA. Tables 5.3-5.5 summarise the different LC and MS methods used in Copenhagen.

Sample set	Instrument	Column length (cm)	Injection volume (µl)	Flow rate (nl/min)	Time (mm:ss)	Duration (min, from time)	% Buffer B
Atapuerca	Q-Exactive HF	50	5	200	00:00	0	2
					110:00	110	25
					135:00	25	40
					140:00	5	60
					145:00	5	60
					150:00	5	2
					165:00	15	2
Dmanisi	Q-Exactive HF test run	15	1	250	00:00	0	20
					11:00	11	35
					13:00	2	80
					14:00	1	80
					16:00	2	5
					26:00	10	5
					Q-Exactive HF	50	4
	110:00	110	25				
	135:00	25	40				
	140:00	5	60				
	145:00	5	60				
	150:00	5	2				
	165:00	15	2				
	Q-Exactive HF-X	15	5	250	00:00	0	5
	50:00				50	30	
	60:00				10	45	
	62:00				2	80	
	67:00				5	80	
	72:00				5	5	
77:00	5				5		

Table 5.3: nLC operation parameters for samples when coupled to a Q-Exactive HF or Q-Exactive HF-X mass spectrometer (Copenhagen).

Sample set	Instrument	TopN	Resolution	Mass range	ACG	Max IT
Atapuerca	Q-Exactive HF	10	120,000	300-1750 m/z	3e6	20 ms
Dmanisi	Q-Exactive HF test run	10	60,000	350-1400 m/z	3e6	20 ms
	Q-Exactive HF	10	120,000	300-1750 m/z	3e6	20 ms
	Q-Exactive HF-X	10	120,000	350-1400 m/z	3e6	25 ms

Table 5.4: Mass spectrometer full scan operation parameters for samples on the Q-Exactive HF and QExactive HF-X (Copenhagen). Resolution is at m/z 200. IT: injection time, ACG: auto gain control target.

Sample set	Instrument	Resolution	Fixed first mass	ACG	Max IT	NCE	Isolation window	DE
Atapuerca	Q Exactive HF	60,000	100 m/z	2e5	108 ms	28	1.3 m/z	30 s
Dmanisi	Q Exactive HF test run	30,000	100 m/z	1e5	45 ms	28	1.3 m/z	10 s
	Q Exactive HF	60,000	—	2e5	108 ms	25	1.3 m/z	30 s
	Q Exactive HF-X	60,000	100 m/z	2e5	108 ms	28	1.2 m/z	20 s

Table 5.5: Mass spectrometer fragment scan operation parameters for samples on the Q-Exactive HF and Q-Exactive HF-X (Copenhagen). Resolution is at m/z 200. IT: injection time, NCE: normalized collision energy, ACG: auto gain control target, DE: dynamic exclusion.

Barcelona (Spain). Samples were dissolved in 0.1% formic acid and analyzed by LC-MSMS using a LTQ-Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, San Jose, CA, USA) coupled to an EASY-nLC 1000 (Thermo Fisher Scientific (Proxeon), Odense, Denmark). Peptides were loaded directly onto the analytical column and separated by reversed-phase chromatography using a 50-cm column with an inner diameter of 75 μm , packed with 2 μm C18 particles (Thermo Scientific, San Jose, CA, USA). Chromatographic gradients started at 95% buffer A and 5% buffer B with a flow rate of 300 nl/min for 5 minutes and gradually increased to 22% buffer B and 78% A in 105 min and then to 35% buffer B and 65% A in 15 min. After each analysis, the column was washed for 10 min with 10% buffer A and 90% buffer B. Buffer A: 0.1% formic acid in water. Buffer B: 0.1% formic acid in acetonitrile.

The mass spectrometer was operated in positive ionization mode with nanospray voltage set at 2.4 kV and source temperature at 275°C. Ultramark 1621 was used for external calibration of the FT mass analyzer prior the analyses, and an internal calibration was performed using the background polysiloxane ion signal at m/z 445.1200. The acquisition was performed in data-dependent acquisition (DDA) mode and full MS scans with 1 micro scans at resolution of 120,000 were used over a mass range of m/z 350-1500 with detection in the Orbitrap mass analyzer. Auto gain control (AGC) was set to 4E5 and charge state filtering disqualifying singly charged peptides was activated. In each cycle of data-dependent acquisition analysis, following each survey scan, the most intense ions above a threshold ion count of 10000 were selected for fragmentation. The number of selected precursor ions for fragmentation was determined by the "Top Speed" acquisition algorithm and a dynamic exclusion of 60 seconds. Fragment ion spectra were produced via high-energy collision dissociation (HCD)

at normalized collision energy of 28% and acquired in the Orbitrap mass analyzer. AGC was set to 3E4, and an isolation window of 1.6 m/z and maximum injection time of 100 ms were used. All data were acquired with Xcalibur software v4.1.31.9. Several blank samples were injected before and after each sample to avoid sample carryover.

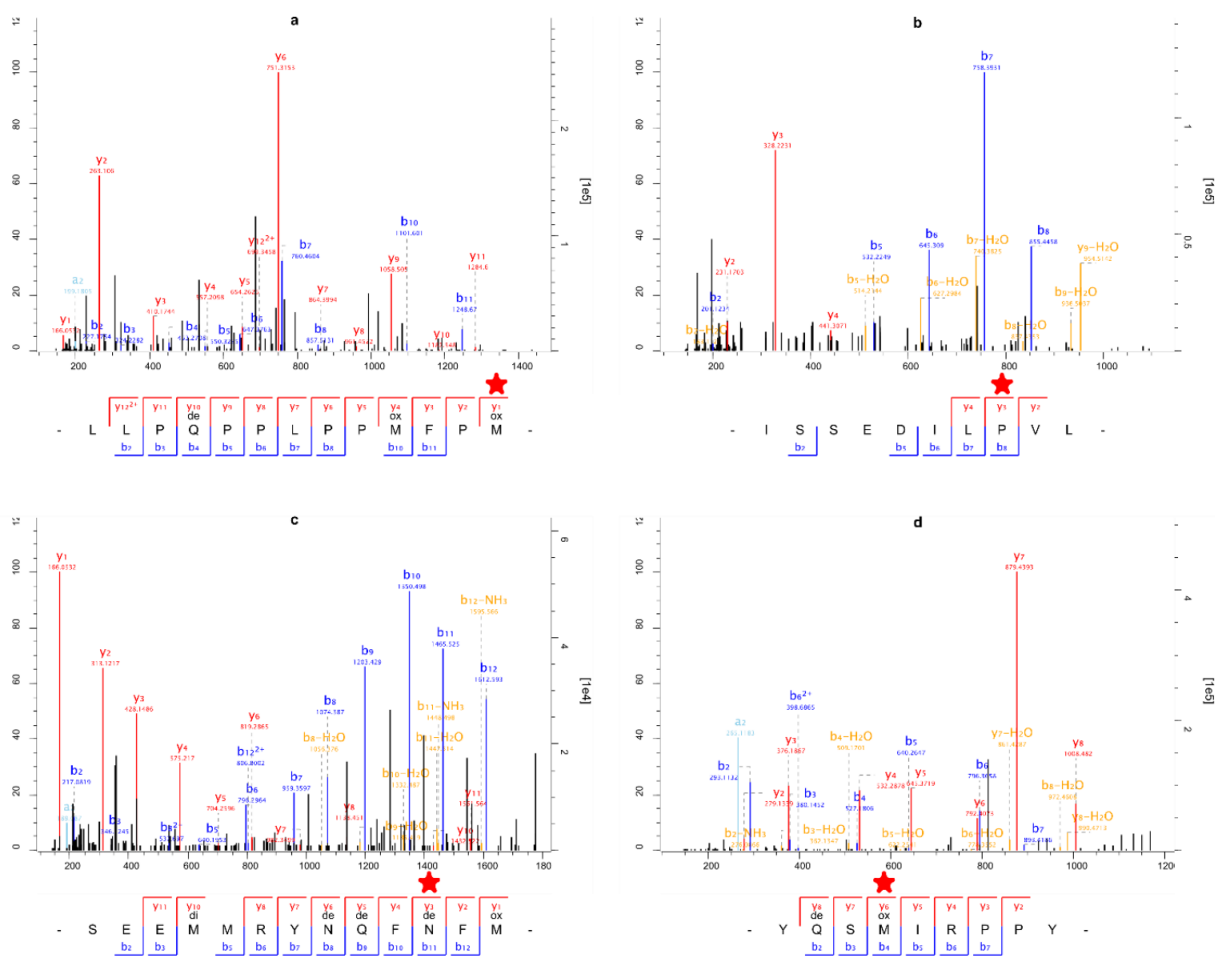


Figure 5.16: Examples of LC-MS/MS spectra of *Homo antecessor*. **a**, Ancestral SAP (L179M) in AMELY present in *Homo antecessor*. Stagetip number: 1069. Scan number: 31841. **b**, *Homo antecessor*-derived SAP (L1089P) in COL17 α 1. Stagetip number: 1069. Scan number: 31158. **c**, Derived SAP (T64N) in ENAM shared with *Homo sapiens*, Neanderthals, and Denisovans. Stagetip number: 1069. Scan number: 18166. **d**, AMELY-specific peptide. Stagetip number: 1069. Scan number: 9847. Note the presence of deamidation (de) and oxidation (ox, di) modifications in **a**, **c**, and **d**. Informative positions are highlighted (red star). Spectral acquisition was replicated across multiple extractions and injections for each informative position.

5.10.2.2.3 Protein sequence database

We constructed an initial Hominidae sequence database containing protein sequences of all major and minor enamel proteins derived from all extant great apes, a hylobatid (*Nomascus leucogenys*), and a macaque (*Macaca mulatta*). Additionally, we added protein sequences translated from extinct Late Pleistocene hominins [Meyer et al., 2016, Castellano et al., 2014], and sequences from *Gorilla beringei*, *Pongo pygmaeus*, and *Pongo tapanuliensis* [De Manuel et al., 2016, Nater et al., 2017, Prado-Martinez et al., 2013]. See 5.10.2.3.1 and 5.10.2.3.2 for details on the prediction of protein sequences for published genomes of great apes and extinct hominins. Accession numbers for UniProt or Genbank accessions of extant great apes and humans can be found in Supplementary Table 5.6. For each protein, we reconstructed the protein sequence of ancestral nodes in the Hominidae family through PhyloBot [Hanson-Smith and Johnson, 2016] to minimize cross-species proteomic effects [Welker, 2018a], and added missing isoform variation based on the isoforms present for each protein in the human proteome as given by UniProt. Ancestral Sequence Reconstruction (ASR) was conducted across the entire Hominoidea phylogeny using PhyloBot [Hanson-Smith and Johnson, 2016]. Input sequences were constrained phylogenetically to (*Macaca*,(*Nomascus*,((*Pongo abelii*, *Pongo pygmaeus*),*Gorilla*,(*Homo*,((*Pan paniscus*, *Pan troglodytes*)))))). After obtaining protein sequences of extant, extinct, and ancestral nodes across Hominoidea, we imputed isoform variation known to exist for AMELX, AMELY, AMBN, AMTN, KLK4 and TUFT1 as we realized that isoforms for most of these proteins are not present in all great ape reference sequences available in UniProt. We assumed that alternative splicing would be placed identically for all great apes, took those protein positions from the human reference sequences (which does have isoform variation for each of these proteins), and placed them on the non-human great ape and ASR sequences. These manually created great ape and ASR isoforms were added to the protein sequence database, with sequence names appended with “ManIso2” and/or “_ManIso3”. Furthermore, we downloaded the entire human reference proteome from UniProt (downloaded 04.09.2018) for a single separate search to allow matches to proteins previously not encountered in enamel proteomes. To each constructed database we added a set of known or possible laboratory contaminants, to allow for the identification of possible protein contaminants [Hendy et al., 2018].

Protein	<i>Homo sapiens</i>	<i>Pan paniscus</i>	<i>Pan troglodytes</i>	<i>Gorilla gorilla</i>	<i>Pongo abelii</i>	<i>Nomascus leucogenys</i>	<i>Macaca Mulatta</i>
COL1 α 1	P02452	XP_003817507	H2QDE6	G3RBN8	H2NVM9	XP_012358721	H9Z595
COL1 α 2	P08123	XP_003809763	H2QUY2	G3QT97	H2PMW7	G1RZZ2	H9Z2D1
COL17 α 1	Q9UMD9	XP_008949124.1	H2Q2J4	G3QE20	H2NBI5	G1RZC4	(absent)
ALB	P02768	XP_003832390	H2RBT1	G35791	Q5NVH5	G1R8T8	Q28522
AMBN	Q9NPF70	XP_003809040	H2R148	G3RCU1	H2PDI5	G1R841	F7HLX4
AMELY	Q99218	(absent)	Q861X8	C3UJP7	(absent)	(absent)	A0A1D5RDA1
AMELX	Q99217	XP_003805726	A5JJS6	G3SDK0	H2PUX0	G1RCS3	A5JJS8
ENAM	Q9NRM1	B2L7U5	H2QPM0	B2L7U8	H2PDI6	G1R843	F7H832
TUFT1	Q9NNX1	XP_003817293	K7CQG4	G3QY68	H2N5V2	G1RGY4	G7MDK9
KLK4	Q9Y5K2	XP_003813692	XP_009434410	G3QU55	A0A2J8U913	G1R1C5	G7NMD1
MMP20	O60882	XP_003828430	H2Q4M8	G3QLA8	H2NF32	G1R6B1	F7GQW6
AMTN	Q6UX39	XP_003809041	H2QPL9	G3RJV5	H2PDI4	G1R825	F6VN65
ODAM	A1E959	XP_003809049	A1YQ94	G3QY18	H2PDH6	G1R7Z0	A1YQ92
AHSG	P02765	XP_008953975	Q9N2D0	E1U7Q5	H2PC98	G1R4B1	F6VZ47

Table 5.6: Accession numbers of publicly available protein sequences utilized in database construction and phylogenetic analysis.

5.10.2.2.4 wiNNer peptide sequence and SAP validation

Dataset. The wiNNer model was trained for the prediction of the phylogenetically informative peptide sequences in the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*). Ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*) were divided into phylogenetically informative peptide sequences and the subset not containing such phylogenetically informative peptides. A training dataset was prepared by taking a subset of the latter peptides, and adding a previously published dataset of enamel proteomes from Dmanisi fauna [Cappellini et al., 2019]. The dataset only has HCD fragmentation, so we build two models. HCD+2 contains 5,555 unique modification-specific peptides, and HCD+3 contains 692 unique modification-specific peptides. For each unique modification-specific peptide, the spectrum with the highest Andromeda score was taken. Spectra with an Andromeda score below 50 were discarded. The retained data for each model was split into a division of 80:10:10 for training, validation and test sets, respectively. We kept test data for evaluating the wiNNer model by calculating the Pearson correlation coefficient (PCC) between true and predicted intensities for each peptide.

The training data has non-tryptic peptides. In addition, variable modifications such as Oxidation (M), Deamidation (NQ), Gln → pyro-Glu, Glu → pyro-Glu, Oxidation (P), Carbamidomethyl (C), Dioxidation (MW), Oxidation (W), His → GluOH (H), His → Asp (H), Arg Ornithine, Phospho (STY) and Phospho (S) were added when processed with MaxQuant, and peptide sequences containing these variable modifications were taken into account.

wiNNer model. The conventional neural network model wiNNer [Tiwary et al., 2019] (window-based neural network being easily retrainable) uses sequence windows to compute feature space around the backbone bond for which y- and b-ion intensities need to be predicted. Amino acids on the N- and C- terminus, the distance of the bond from the peptide termini, and the length of peptide, were also used as features. Each amino acid residue and modified amino acid residues were converted to a 38 binary feature by one hot encoding to include residue specific modification and sliding window extending termini. For example, to include deamidation (NQ), the modification of two extra residues N(de) and Q(de) were added for one hot encoding. The missing intensities were set to zero and the intensities of a single peptide was normalized by the maximum value among y- and b-ions. Then, they were log transformed: $\log_2(1 + \text{Intensity} * 10000)$. We trained two regression models, one for HCD+2 and one for HCD+3.

The architecture of the wiNNer model is built using Keras (version 2.0.8; <https://keras.io/>), a high-level neural network, to train the neural network model. Tensorflow version 1.3.0 was used as backend in Keras. The architecture of neural network includes 5 dense layers. The input layers contain 991 features for a window size of 24. The hidden unit is reduced from 600, 400, 200 to 50 in subsequent dense layers, and the output layer contains 2 units for y- and b-ion peak intensities. Hyper-parameters such as batch size, dropout, learning rate and number of epochs were optimized separately for different models. The wiNNer model can be accessed on GitHub (<https://github.com/cox-labs/wiNNer.git>).

Results. The supplementary Figure 5.17a shows the PCC of HCD+2 and HCD+3 models were 0.85 and 0.81 respectively, between true and predicted intensities for each peptide in the test sets. These results are close to the wiNner model for unmodified sequences, where the PCC is 0.88 for HCD+2 and 0.76 for HCD+317. Figure S6b shows the PCC between true and predicted spectra intensity of all the phylogenetically informative peptide sequences in the ancient samples. The PCC of All peptides, Oxidation (M), Deamidation (NQ), Oxidation (P), Dioxidation (MW) and Phospho (S) were 0.77, 0.79, 0.76, 0.68, 0.76 and 0.76 respectively. Again, these distributions are similar to the overall test performance of our wiNner model, and indicate accurate peptide sequence identification for our phylogenetically informative spectra.

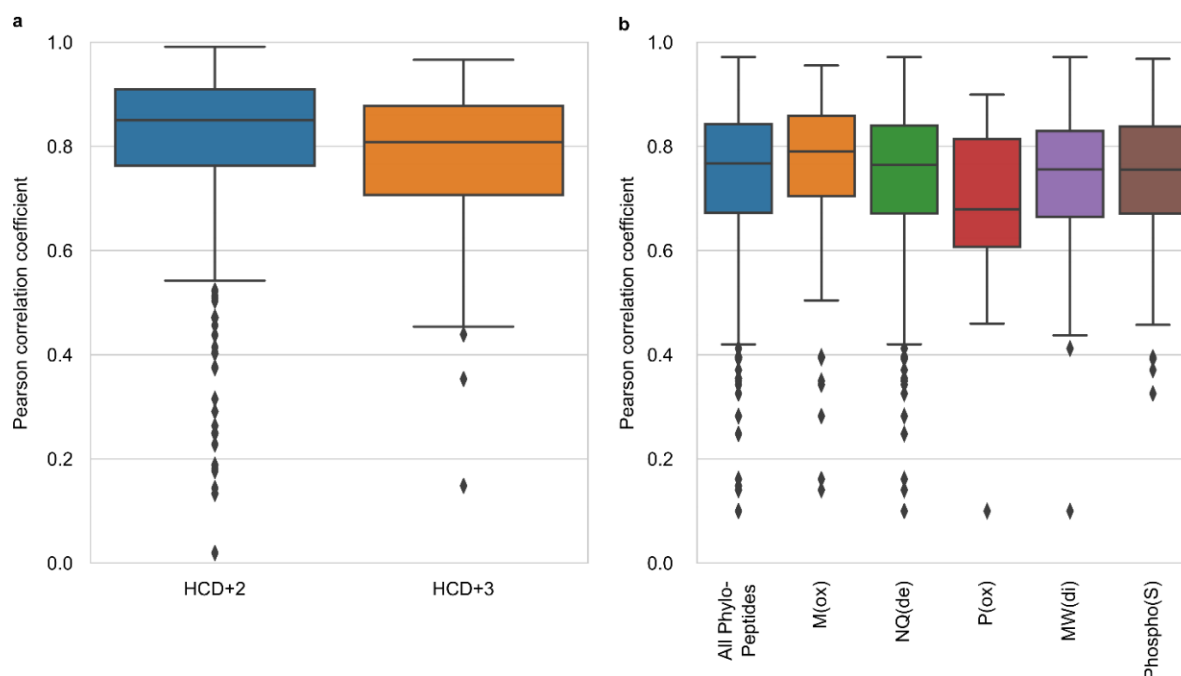


Figure 5.17: Results of wiNner on ancient enamel proteomes. **a**, Performance of the wiNner model for HCD+2 and HCD+3. The boxplots shows the PCC distribution between true and predicted spectrum intensity for each peptide sequence in the test data. The boxplots contains 556 unique PSMs for HCD+2 and 70 unique PSMs for HCD+3, respectively. **b**, PCC distribution of true and predicted intensities in phylogenetically informative peptides, and for specific classes of variable modifications. Distribution of PCCs of true and predicted *y*- and *b*-ion peak intensities in all the phylogenetically informative peptides (“All Phylo-Peptides”), and peptides including variable modifications, such as Oxidation (M), Deamidation (NQ), Oxidation (P), Dioxidation (MW) and Phospho (S). The boxplots contain 359, 125, 278, 22, 47, and 114 unique peptides, respectively. Lower mean PCC correlation for classes of variable modifications are generally related to smaller number of unique peptides included in wiNner test and validation. For **a** and **b**, each box extends from the lower to upper quartile values of the data. The line indicates the median. The whiskers extends up to 1.5 times the interquartile range. Values above and below the interquartile ranges are plotted in diamonds as outliers.

5.10.2.2.5 Proteomic data analysis

Proteomic data analysis focused on diagenetic degradation (deamidation and peptide bond hydrolysis), and the survival of in vivo modifications in the two Pleistocene hominin proteomes.

Deamidation analysis. Glutamine and asparagine deamidation has recently been investigated repeatedly in ancient proteome studies in an effort to understand patterns of protein damage between proteins and/or across time. We quantified glutamine and asparagine deamidation separately by following Mackie et al. [Mackie et al., 2018] for MaxQuant. In short, protein-based deamidation rates are based on spectral intensities of matching PSMs, subsequently re-assessed through 1000 bootstraps to derive a 95% confidence interval. See the original paper for further details.

Peptide length distributions. Our digestion-free extraction approach allows us to explore peptide length distributions across and between samples. Here, peptide length distributions are based on MaxQuant peptide matches only. Comparisons are made with enamel proteomes extracted and analyzed in similar manners presented earlier [Cappellini et al., 2019, Welker et al., 2019].

MMP20 and KLK4 cleavage patterns. The existing literature on enamel proteomes and enamel proteome biomineralization describes the processes by which the enamel proteome is shaped through targeted protein hydrolysis by the proteolytic enzymes MMP20 and KLK4 [Chun et al., 2010, Yamakoshi et al., 2006, Iwata et al., 2007, Nagano et al., 2009]. These two enzymes digest the major enamel proteome components AMBN, AMELX, and ENAM into shorter amino acid chains. The in vivo cleavage locations of MMP20 and KLK4 on AMBN, AMEL(X/Y), and ENAM have been described in the literature [Chun et al., 2010, Yamakoshi et al., 2006, Iwata et al., 2007, Nagano et al., 2009]. However, although various studies have previously obtained proteomic data on fully mineralized enamel samples, including ancient proteomes, no studies have explored whether in vivo MMP20 and KLK4 cleavage survives in (ancient) enamel proteomes.

Our digestion-free extraction approach allows us to explore whether our ancient proteome datasets preserve such in vivo patterns. As the Dmanisi sample contains several extracts either digested or alkylated, we explore MMP20 and KLK4 cleavage patterns solely using the Atapuerca dataset. In addition, we only focus on the proteins AMBN, AMEL(X/Y), and ENAM, since for these proteins we have large numbers of identified peptide-spectrum matches (PSMs), as well as sufficient literature on known MMP20 or KLK4 cleavage sites.

As a background, we calculated for each protein a theoretical matrix (21 × 21) containing the percentage that each possible amino acid (P1) – amino acid (P1') pair occurred in the UniProt sequence of that protein (AMBN_HUMAN, AMELX_HUMAN, AMELY_HUMAN, ENAM_HUMAN). For AMELX, this calculation was based on the isoform 3 variant, which is the longest of the three AMELX variants present in UniProt. Similarly, for AMELY, we selected isoform 1, because it is the longest isoform. For AMEL(X/Y), we had to construct a joined theoretical matrix, as AMELX and AMELY significantly overlap in peptide populations, while the sampled *Homo antecessor* individual is a male. We obtained this combined theoretical matrix by first joining the AMELX and AMELY matrices in a ratio of

9:1, which is close to the experimental proteomic and tRNA ratios observed in developing enamel [Stewart et al., 2016, Fincham et al., 1991, Parker et al., 2019]. The protein-derived ratios were then recalculated to a percentage for each amino acid pair. This theoretical background thereby is a simple model of protein fragmentation in which each amino acid pair (P1-P1') has an equal chance of being cleaved and being observed, regardless of the amino acids involved, their potential PTM modification, or their placement within the sequence or structure of the protein.

As for the observed peptide cleavage pairs, we took the C- and N- termini of the peptide-spectrum matches to each protein and counted the total occurrence of each amino acid pair. This matrix was then turned into percentages, and divided by the theoretical matrix to obtain a value indicating the fold difference between the observed frequency of each hydrolyzed peptide bond (P1-P1') and random peptide bond cleavage (P1-P1') for each possible combination of amino acids in both positions.

The results of this exercise show that, against a general background of peptide bond hydrolysis, there are amino acid pairs that stand out by very large positive fold-differences in their occurrence (Extended Data Fig. 5.10a, b). When mapped to known MMP20 and KLK4 cleavage sites, bonds with high rates of fragmentation are without exception known cleavage sites of either of those enzymes, or are within 2 amino acids in either C- or N-terminal direction of such a known cleavage site (Extended Data Fig. 5.10b, c). This finding provides experimental evidence that *in vivo* MMP20 and KLK4 protein cleavage survives in Middle Pleistocene proteomes. In addition, for ENAM we observe a set of cleavage pairs that have highly enriched fold cleavages located closely together (positions 56-57, 58-59, 65-66, 66-67; accession ENAM_HUMAN in UniProt), suggesting that extensive hydrolysis, either *in vivo* or diagenetically, occurs at or around these positions (Extended Data Fig. 5.10b).

Furthermore, our data shows an absence of PSMs to known locations of N-linked glycosylation in ENAM (Extended Data Fig. 5.10c), as revealed in comparison with such modifications for ENAM included in UniProt. This might either be due to enhanced molecular diagenesis at such positions or aspects of our extraction and/or spectral identification approach preventing the recovery of PSMs from such sequence regions.

Phosphorylation occupancy. Phosphorylation occupancy was calculated using MaxQuant (v1.5.3.30) by measuring the intensity ratio of phosphorylated and non-phosphorylated peptides. The raw mass spectrometry data of both hominins, SK339 and Ø1952 was analyzed in a combined search in MaxQuant for occupancy comparison between the samples. For the database search, the same settings were applied as described in the Methods section. We modified the used sequence database in such a way that it only contains one protein sequence entry per species. The removal of AMELX and AMBN isoforms ensures that occupancy is calculated only once for each site. Additionally, we excluded AMELY protein sequences, since we did not confidently identify AMELY in the initial database search in the Dmanisi specimen.

In total, we could identify 18 phosphorylation sites. For three sites, ratio calculation was possible, since in all four specimen the phosphorylated peptides and its non-phosphorylated counterparts were present and quantified through label-free quantification (Supplementary Table 5.7, Group 1). For nine sites, it was not possible to quantify the phosphorylated peptide

in at least one out of all 4 specimens (Supplementary Table 5.7, Group 2). For six sites, only the phosphorylated peptides were present but not the non-phosphorylated counterpart at least one out of all four specimens (Supplementary Table 5.7, Group 3).

Group 1: identification of both, phosphorylated and unmodified peptides							
	AMELX		AMBN		ENAM		
Phosphosite	32_AMELX	43_AMBN	191_ENAM				
<i>Homo erectus</i>	0.017859	0.23628	0.053114				
<i>Homo antecessor</i>	0.36654	0.10561	0.42624				
Ø1952	0.088134	0.073352	0.98228				
SK339	0.17361	0.28135	1.7769				
Group 2: 0 – quantification of only unmodified peptides							
	AMELX					AMBN	
Phosphosite	33_AMELX	42_AMELX	175_AMELX	181_AMELX	182_AMELX	34_AMBN	41_AMBN
<i>Homo erectus</i>	0.013984	0.001532	0.0053254	0	0	0.0028724	0.84794
<i>Homo antecessor</i>	0.0033069	0	0.0014622	0	0	0.0027449	0.020293
Ø1952	0.00016585	0	0	0.010832	0	0.00051682	0
SK339	0	0	0	0	0	0	0.22245
	ENAM						
Phosphosite	54_ENAM	216_ENAM					
<i>Homo erectus</i>	0	0.025636					
<i>Homo antecessor</i>	0.20649	0.01626					
Ø1952	0	0					
SK339	0	0					
Group 3: ∞ – identification of only phosphorylated peptides							
	AMTN		AMBN				
Phosphosite	115_AMTN	116_AMTN	101_AMBN	261_AMBN	262_AMBN	303_AMBN	
<i>Homo erectus</i>	∞	∞	∞	17.033	8.7485	∞	
<i>Homo antecessor</i>	∞	∞	∞	∞	∞	0.025266	
Ø1952	∞	∞	0	0.19077	0.10544	∞	
SK339	∞	∞	∞	∞	0.44953	∞	

Table 5.7: Description of all identified phosphorylation sites in both ancient hominins and both modern human control samples, including their summed intensity ratios of phosphorylated and unmodified peptides. The sites are grouped, depending on the presence of both phosphorylated and nonphosphorylated peptides (group 1), the sole quantification of unmodified peptides in at least one of the samples (0, group 2), or the presence of phosphorylated peptides only (∞ , group 3). Ø1952 represents our recent human control. SK339 is an example of a recent control taken from Stewart et al. [Stewart et al., 2017].

5.10.2.2.6 Recent modern human control samples

To contrast our ancient hominin enamel proteomes with modern human data from less diagenetically altered environments, we also processed a single Medieval human tooth from Copenhagen (Ø1952) using extraction methods 1 and 3 (two injections in total, one based on HCl and one based on TFA; see Section 5.10.2.2.1). LC-MS/MS set-up was identical as described above (see Section 5.10.2.2.2). Ø1952 represents a male individual, derives from the former cemetery of the Almindeligt Hospital which was in use from approximately 1600-1800 A.D., and which was excavated in 1952. The specimen is stored and owned by the Laboratory of Biological Anthropology, Department of Forensic Medicine, University of Copenhagen. Additionally, we processed published human enamel proteome data from Stewart et al. [Stewart et al., 2017]. The latter utilizes a 5% HCl acid-etching procedure that is minimally invasive, but also only recovers proteins from the outer enamel surface. Stewart et al. used an extraction method without enzymatic digestion, and is therefore partly comparable to the workflow used for Ø1952 and the ancient samples processed using extraction 1 and 3. Their samples range in chronological age between approximately 5700 and 200 years old. Proteomic data from Stewart et al. and Ø1952 was processed using the same MaxQuant as the ancient samples against a protein sequence database restricted to *Homo sapiens*. All other search settings were as specified in the Methods, and are therefore identical between the modern human control samples and the ancient hominins.

Our analysis of Ø1952 and the Stewart et al. samples indicate that all these enamel proteome extracts are composed of the core enamel proteome (ENAM, AMBN, AMELX/Y, MMP20, AMTN) with the addition of additional collagens (COL1 α 1, COL1 α 2, COL17 α 1) and plasma proteins (AHSG, ALB) in some cases (Supplementary table 5.8). It is unclear whether the collagens and plasma proteins derive from residual dentine in the Ø1952 sample. It is likely these proteome components are endogenous to the enamel, however, as the acid-etch surface extractions performed by Stewart et al. also contains these proteins at low frequency. We confirm the sex assignment of the Stewart et al. samples by observing an absence of AMELY-unique peptides in female samples but the presence of AMELY-specific peptides in all male samples. We note that MMP20 is absent in acid-etching of enamel surfaces. Conversely, ODAM and AMTN are only present in the acid-etches and not in the destructively-sampled Ø1952.

Analysis of proteome degradation performed on the *Homo antecessor* and *Homo erectus* datasets were contrasted with Ø1952 and SK339, a male individual of the Stewart et al. dataset. SK339 was chosen as it represents an approximation of the average protein sequence recovery across the Stewart et al. samples and is comparatively young in chronological age (19th century AD).

Protein sequence coverage obtained for our human control (Ø1952) and the ancient hominins is comparable, but on average lower compared to the recent controls generated through subsuperficial acid-etching of exposed enamel surfaces (Supplementary Fig. 5.18). This is likely due to the deep-sequencing of the ancient samples across multiple LC-MS/MS injections, while data for SK339 is limited to a single LC-MS/MS injection and data for Ø1952 is limited to two LC-MS/MS injections. The decision to not employ subsuperficial acid etching of enamel surfaces, but opt for destructive sampling, is supported by this observation of lower protein sequence recovery for the acid-etched datasets (Supplementary Fig. 5.18).

Protein	Guthlac_F	Seaham_F	SK363	SK378	Whitwell_F	Guthlac_M	Seaham_M	SK119	SK130	SK339	SK351	SK366	Whitwell_M	Ø1952
Sex	F	F	F	F	F	M	M	M	M	M	M	M	M	M
ALB		1.3 [1]			1.3 [1]	1.1 [1]		1.3 [1]	2.5 [2]	1.3 [1]		1.3 [1]	3.8 [1]	7.1 [4]
AHSG			2.5 [1]			2.2 [1]								12.8 [12]
COL1α1						6.3 [10]								67.3 [591]
COL1α2						1.8 [8]								44.8 [206]
COL17α1			1.1 [3]		1.7 [2]	1.6 [2]	0.9 [1]			0.8 [1]	0.8 [1]			17.2 [22]
ODAM		3.2 [1]	3.2 [1]		3.2 [1]		3.2 [1]	3.2 [1]		3.2 [1]		3.2 [1]	6.5 [2]	
AMTN				5.3 [1]				5.3 [2]		5.3 [2]	4.8 [1]		3.4 [1]	
AMELX	49.8 [124]	37.6 [89]	45.9 [131]	56.6 [120]	37.6 [112]	41 [77]	44.9 [113]	37.1 [121]	35.1 [93]	45.9 [147]	43.4 [147]	40.5 [132]	23.4 [36]	86.8 [413]
AMELY						22.4 [2]	22.9 [8]	31.8 [12]	27.6 [8]	33.3 [16]	32.8 [14]	29.2 [17]	17.7 [1]	64.6 [54]
AMBN	21.9 [50]	22.1 [62]	24.6 [53]	25.5 [68]	26 [80]	21.3 [48]	23.7 [69]	22.8 [58]	19.7 [50]	25.5 [67]	22.4 [55]	26.2 [62]	21 [54]	39.4 [312]
ENAM	10 [79]	9.1 [61]	9.9 [60]	7.6 [40]	10.7 [96]	9.1 [46]	9.6 [101]	9.8 [68]	8.9 [46]	10.6 [75]	10.7 [70]	10.3 [81]	6.7 [40]	15.5 [214]
MMP20														6.8 [5]

Table 5.8: Sequence coverage (%) of proteins identified in the modern human enamel proteomes from Stewart et al. and Ø1952. For AMTN, AMELX, and AMELY, sequence coverage is calculated for the longest isoform present. Numbers in brackets indicate unique+razor counts per protein. Unique+razor peptide counts for AMELX are summed for the three isoforms present in the database – note that a large number of shared peptides between these three isoforms are therefore not taken into consideration here.

Still, major sequence regions for the enamel-specific proteins recovered in the ancient hominins are also present in the modern human controls, regardless of extraction approach (Extended Data Fig. 5.7).

We contrasted protein deamidation and peptide lengths of the recent human controls with those of the ancient hominin samples. We observe that protein deamidation is high for both recent controls, but less advanced compared with *Homo antecessor* and *Homo erectus* (Supplementary Fig. 5.19). Interestingly, although the ancient hominins and Ø1952 have an expected pattern of more advanced asparagine (N) deamidation compared to glutamine (Q) deamidation, the SK339 proteome displays the opposite pattern of with more advanced glutamine deamidation. It is unclear why this would be the case. Enamel-specific proteins are fragmented in vivo through the combined action of MMP20 and KLK4 [Chun et al., 2010, Nagano et al., 2009, Fukae et al., 1996]. Nevertheless, we observe that the peptides recovered from the recent control samples are, on average, longer than those observed in the *Homo antecessor* and *Homo erectus* samples (Extended Data Fig. 5.9d; Ø1952 – *Homo antecessor*: t-test(6.87), df=1692, p=9.0e⁻¹²; Ø1952 – *Homo erectus*: t-test(15.93), df=1573, p<2.2e⁻¹⁶). We also observe that the peptide lengths for SK339 are significantly shorter than

those obtained for Ø1952 (t-test(-6.67), df=570, $p=5.9e^{-11}$) or a Medieval ovicaprine enamel specimen published previously⁶ (t-test(-8.90), df=487, $p<2.2e^{-16}$). These observations are consistent with advanced enamel protein fragmentation in vivo, but also highlight the potential of significant variation in protein fragmentation between samples or analytical procedures during protein extraction and LC-MS/MS analysis.

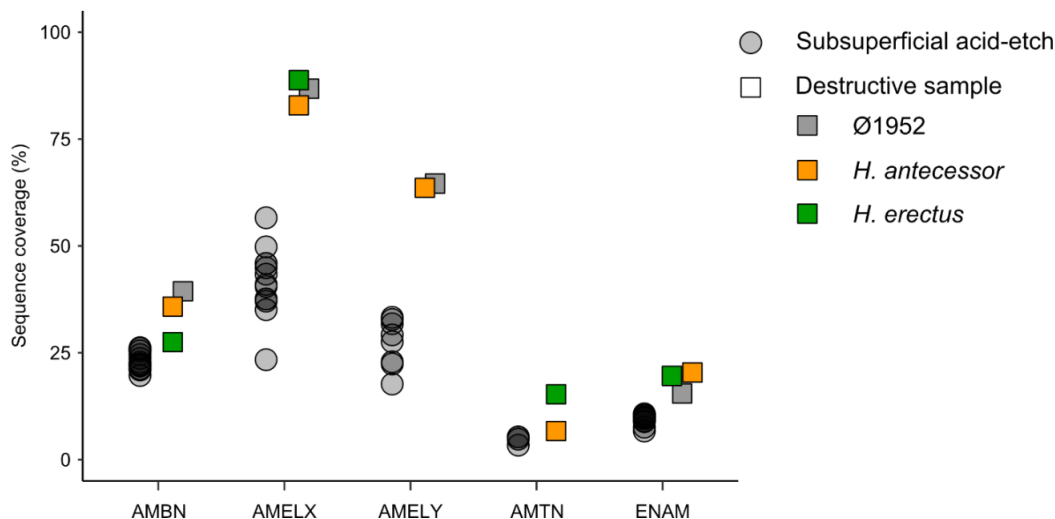


Figure 5.18: Comparison of protein sequence coverage between the subsuperficial acid-etches ($n=13$, Stewart et al.16) and destructive sampling ($n=3$, this study).

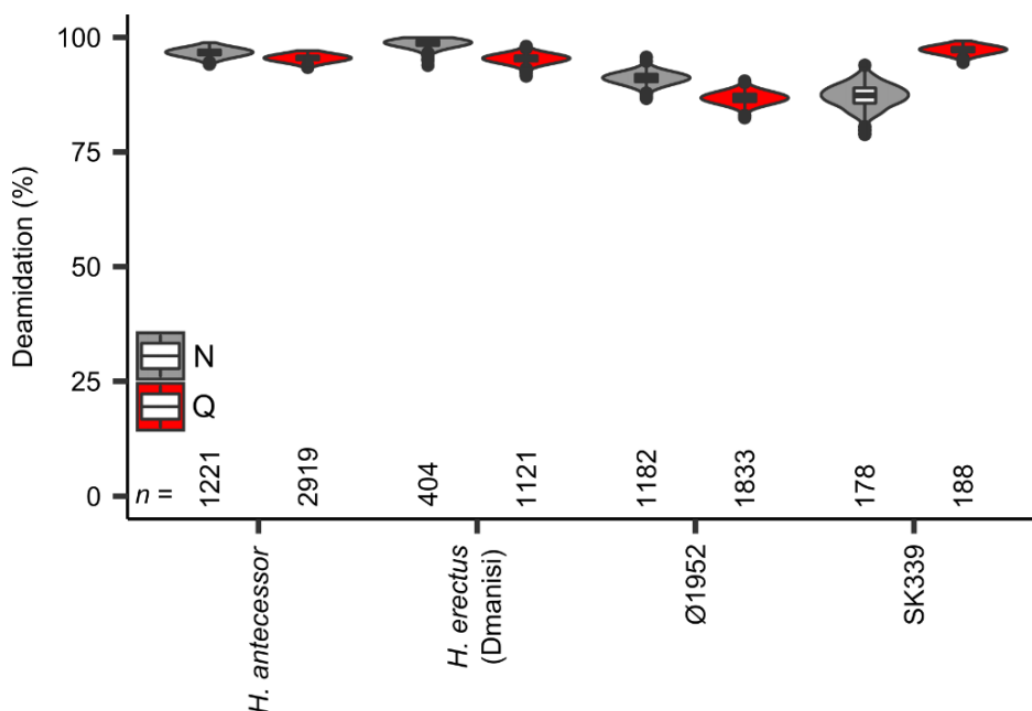


Figure 5.19: Enamel proteome deamidation of human controls and ancient hominin proteomes. Only peptides matching to AMELX, AMELY, AMBN, and ENAM are taken into account. Deamidation calculated following Mackie et al. [Mackie et al., 2018] for 1,000 bootstrap replicates. The number of peptides (n) is given for each violin plot. The boxplots within define the range of the data (whiskers extending to 1.5x the interquartile range), outliers (black dots, beyond 1.5x the interquartile range), 25th and 75th percentiles (boxes), and medians (centre lines). The boxplots define the range of the data (whiskers extending to 1.5 the interquartile range), 25th and 75th percentiles (boxes), and medians (dots).

Finally, we compare the amino acid cleavages between our recent controls and the *Homo antecessor* dataset. We observe that SK339 and Ø1952 display highly similar fold differences in cleavage site occurrence compared to a random cleavage model for each protein (Supplementary Fig. 5.20). Large fold-differences in P1-P1' occurrence correspond to known MMP20 and/or KLK4 cleavage sites in these proteins. The *Homo antecessor* dataset contains a wider range of cleavage sites, compatible with diagenetic hydrolysis of peptide bonds. This has been observed previously for an ancient proteome as well [Chen et al., 2019]. Still, also for the *Homo antecessor* dataset, the observed P1-P1' pairs with a high fold difference compared to a random fragmentation model correspond to sites of known MMP20 and/or KLK4 activity, or peptide bond locations located in close proximity to such sites (see Extended Data Fig. 5.10).

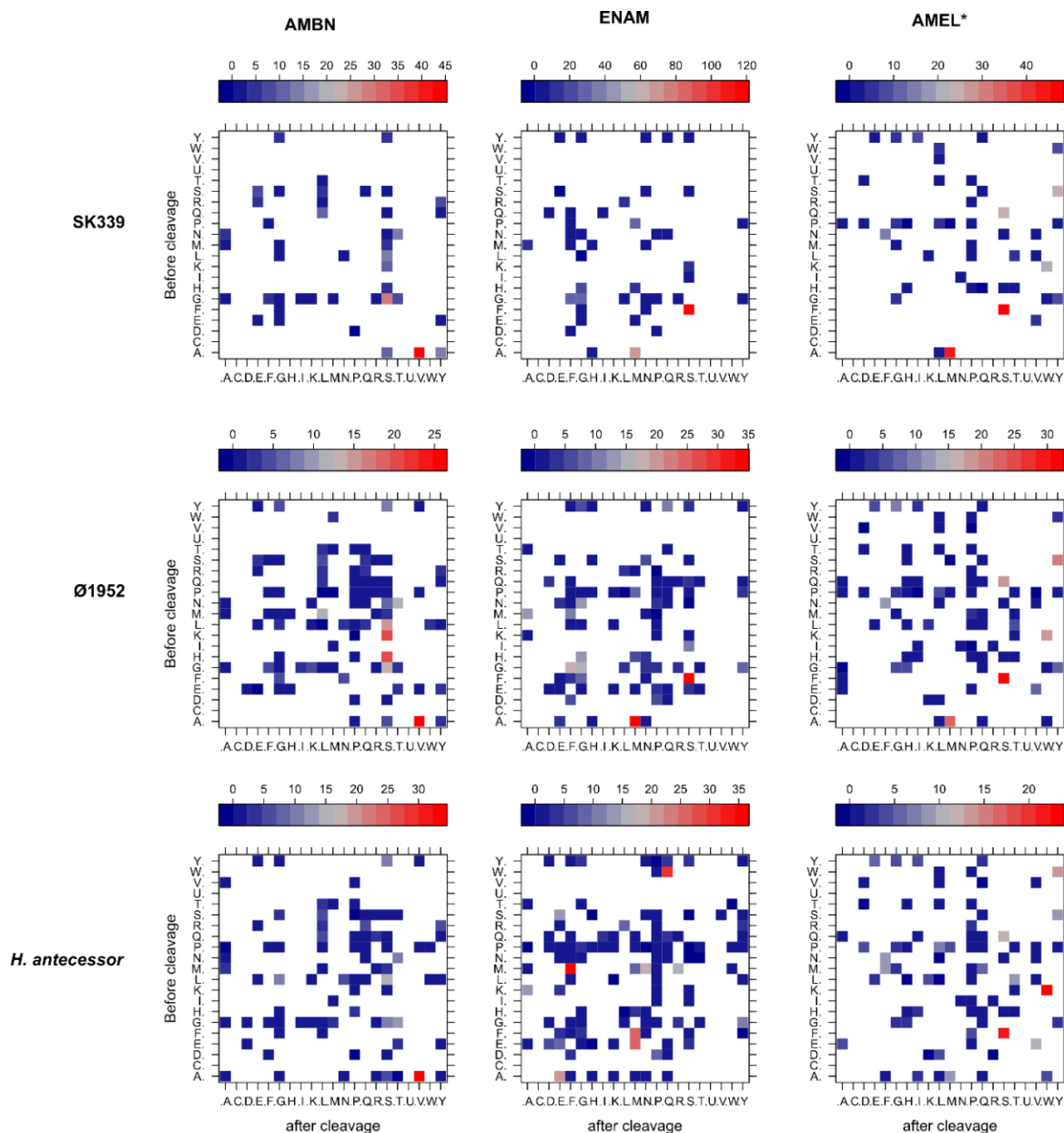


Figure 5.20: Cleavage matrices of SK339, Ø1952, and the *Homo antecessor* proteome for AMBN, ENAM, and AMELX/Y. Values indicate the fold difference between the observed cleavage frequency of each amino acid pair P1-P1' and its theoretical expectancy based on a random protein cleavage (PSMs). Counts based on peptide-spectrum-matches for each protein separately. The bottom row is identical to the top row in Extended Data Fig. 5.10. AMEL* contains PSMs matching to both AMELX and AMELY, with the theoretical matrix of both proteins joined in the ratio 9:1, respectively.

5.10.2.3 Phylogenetic analysis

Jazmín Ramos-Madrigal, Martin Kuhlwilm, Marc de Manuel, Tomas Marques-Bonet, Carles Lalueza-Fox, Eske Willerslev, Fernando Racimo

5.10.2.3.1 Reference data

We combined the new ancient protein sequences from *Homo erectus* and *Homo antecessor* with protein sequences from great apes [De Manuel et al., 2016, Nater et al., 2017, Prado-Martinez et al., 2013, Xue et al., 2015], three Neanderthals [Prüfer et al., 2014, Castellano et al., 2014, Prüfer et al., 2017], a Denisovan [Meyer et al., 2012], and a panel of present-day humans, including 256 samples from the Simons Genome Diversity Panel (SGDP, IDs: ERZ312767-ERZ312784, ERZ312788-ERZ312789, ERZ312791-ERZ312797, ERZ312799-ERZ312848, ERZ324259-ERZ324404, ERZ324529-ERZ324546, ERZ324867-ERZ324876, ERZ325059, ERZ325062, ERZ329670-ERZ329671) [Mallick et al., 2016] and 41 high-coverage individuals from the 1000 Genomes Project [The 1000 Genomes Project, 2015] (IDs: HG00096, HG00119, HG00183, HG00268, HG00419, HG00436, HG00640, HG00759, HG01051, HG01112, HG01136, HG01500, HG01565, HG01583, HG01595, HG01879, HG02568, HG02922, HG03006, HG03052, HG03642, HG03742, NA12413, NA12878, NA12891, NA12892, NA18525, NA18562, NA18939, NA18985, NA19017, NA19238, NA19239, NA19240, NA19625, NA19648, NA19685, NA19700, NA20502, NA20581, NA20845), representing worldwide populations (Supplementary Table. 5.9). Additionally, we included protein sequences from macaque (*Macaca mulatta*) and gibbon (*Nomascus leucogenys*) to root phylogenetic trees. When available, for some great apes, we obtained the protein sequences of interest from the UniProt database (Supplementary Table 5.10), or, in the case of the Neanderthals and Denisovan, from published data [Castellano et al., 2014]. For the rest of the samples, including those from the SGDP and 1000 Genomes, we reconstructed the protein sequences from publicly available read alignments or genotype calls, as described below.

5.10.2.3.2 Reconstructing protein sequences from whole-genome sequencing data

Read alignments were obtained for 41 individuals from Phase3 of the 1000 Genomes Project (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3>) and 256 individuals from the SGDP (ENA accession PRJEB9586). For each individual, we downloaded the reads overlapping the regions of interest using *samtools* [Li et al., 2009] (v0.1.18) and built a majority-count-based consensus sequence using ANGSD [Korneliussen et al., 2014] (v0.913); reads with mapping quality lower than 30 and bases with base quality lower than 20 were discarded. For great apes, VCF files with called diploid genotypes were provided by the authors of the corresponding original publications [De Manuel et al., 2016, Nater et al., 2017, Prado-Martinez et al., 2013]. In this case, we built FASTA files for each of the five genes of interest, for each sample. Indels were not considered and a random allele was chosen at heterozygous positions. For each of these consensus sequences, we identified and removed introns (in silico splicing) from each gene using the annotation of the reference human genome (h19), as provided by ENSEMBL [Zerbin et al., 2018]. We then performed a *tblastn* search using the human reference protein as the query, and each of the previously 'spliced genes' as

a subject, separately. Finally, protein sequences were obtained from the resulting translated alignments.

Species	Subspecies	Inds.	Common name	Reference(s)
<i>Gorilla beringei</i>	<i>Gorilla_beringei_beringei</i>	7	Mountain gorilla	Xue <i>et al.</i> ⁹⁷
	<i>Gorilla_beringei_grauer</i>	8	Eastern lowland gorilla	Prado-Martinez, <i>et al.</i> ⁴³ & Xue <i>et al.</i> ⁹⁷
<i>Gorilla gorilla</i>	<i>Gorilla_gorilla_diehli</i>	1	Cross River gorilla	Prado-Martinez, <i>et al.</i> ⁴³
	<i>Gorilla_gorilla_gorilla</i>	27	Western lowland gorilla	
<i>Homo sapiens</i>	<i>Homo_sapiens</i>	41	Present-day human	1000 Genomes Project ⁵⁸
		256		SGDP ⁵⁷
<i>Pan paniscus</i>	<i>Pan_paniscus</i>	10	Bonobo	Prado-Martinez, <i>et al.</i> ⁴³
<i>Pan troglodytes</i>	<i>Pan_troglodytes_elliotti</i>	10	Nigerian chimpanzee	Prado-Martinez, <i>et al.</i> ⁴³ , de Manuel, <i>et al.</i> ⁴¹ and Auton, <i>et al.</i> ¹⁰¹
	<i>Pan_troglodytes_schweinfurthii</i>	19	Eastern chimpanzee	
	<i>Pan_troglodytes_troglodytes</i>	18	Central chimpanzee	
	<i>Pan_troglodytes_verus</i>	12	Western chimpanzee	
	<i>Pan_troglodytes_verus/troglodytes</i>	1	Chimpanzee hybrid	
<i>Pongo pygmaeus</i>	<i>Pongo_pygmaeus</i>	15	Bornean orangutan	Nater <i>et al.</i> ⁴² and Prado-Martinez, <i>et al.</i> ⁴³
<i>Pongo abelii</i>	<i>Pongo_abelii</i>	11	Sumatran orangutan	
<i>Pongo tapanuliensis</i>	<i>Pongo_tapanuliensis</i>	1	Tapanuli orangutan	
Ancient samples		Inds		Reference(s)
Denisovan		1		Castellano S, <i>et al.</i> ⁴⁰
Neanderthals		3		Prüfer K, <i>et al.</i> ^{31,55}

Table 5.9: Reference samples for which sequencing data was used to reconstruct the protein sequences. Inds. = Individuals.

Protein	ALB	AMB	ENAM	COL17a1	MMP20	AMELY	AMELX
Chromosome (HG37)	4	4	4	10	11	Y	X
<i>Homo sapiens</i>	P02768	Q9NP70	Q9NRM1	Q9UMD9	O60882	Q99218	Q99217
<i>Pan paniscus</i>	XM_00383234 2.2	XM_00380899 2.1	B2L7U5	XM_00895087 6.1	XM_00382838 2.2		XM_00380567 8.1
<i>Pan troglodytes</i>	H2RBT1	H2R148	H2QPM0	H2Q2J4	H2Q4M8	Q861X8	A5JJS6
<i>Gorilla gorilla gorilla</i>	G3S791	G3RCU1	B2L7U8	G3QE20	G3QLA8	C3UJP7	G3SDK0
<i>Pongo abelii</i>	Q5NVH5	H2PDI5	H2PDI6	H2NBIS	H2NF32	A0A2J8W4N8	H2PUX0
<i>Nomascus leucogenys</i>	G1R8T8	G1R841	G1R843	G1RZC4	G1R6B1		G1RCS3
<i>Macaca mulatta</i>	Q28522	F7HLX4	F7H832		F7GQW6	A0A1D5RDA	A5JJS8

Table 5.10: Accession numbers of reference sequences obtained from the UniProt or Genbank databases.

5.10.2.3.3 Protein alignments

We compared the protein sequences retrieved from the ancient samples and samples in the reference dataset. COL1 α 1 and COL1 α 2 were excluded from phylogenetic analysis, as their deamidation values could not exclude a contaminating origin of these proteins. AMTN was excluded since the retrieved sequences overlap a conserved sequence region with no phylogenetic SAPs within Hominidae. For the proteins considered endogenous and informative (AMBN, AMELX, AMELY, ENAM, MMP20, COL17 α 1, ALB), we present aligned fragment ion series and MS/MS spectra in the Supplementary Data 1 file. For each of these proteins, and for each of the ancient samples (Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*), we generated two multiple sequence alignments using *mafft* [Katoh, 2002]: one including all individuals in Supplementary Table 5.7 and 5.8 and the corresponding ancient sample (Full alignments), and a second one including a single individual from each species/group (Single species alignments; Supplementary Table 5.11). To account for isobaric amino acids (leucine and isoleucine), which cannot be distinguished with the methods used to obtain the protein sequences for ancient specimens, we did the following: 1) for positions where the ancient sample carried either a leucine or an isoleucine, and the reference samples carried only one of those amino acids, we set the ancient sample to the amino acid found in the reference samples, and 2) for positions where the ancient sample carried either a leucine or an isoleucine, and both amino acids were present among the reference samples, we set all individuals to carry a leucine.

Protein	Total sites	Non-missing sites in <i>Homo antecessor</i>	Unique amino acid substitutions in <i>Homo antecessor</i>	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Substitutions uniquely shared between <i>H. antecessor</i> and any HND individual.
ALB	609	93	0	53 (10)	16 (2)	67 (13)	22 (4)	0
AMBN	447	169	0	64 (21)	16 (5)	77 (27)	35 (12)	0
AMELX	206	171	0	7 (4)	1 (0)	17 (12)	3 (2)	0
AMELY	206	141	0	28 (21)	17 (12)	36 (28)	26 (21)	0
COL17 α 1	1,498	112	1	62 (4)	21 (1)	124 (8)	62 (2)	0
ENAM	1,142	262	0	159 (31)	38 (8)	201 (39)	91 (16)	2
MMP20	483	66	0	32 (10)	8 (3)	50 (14)	21 (5)	1
Protein	Total sites	Non-missing sites in <i>Homo erectus</i>	Unique amino acid substitutions in <i>Homo erectus</i>	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Polymorphic sites ³	Polymorphic sites (not singletons) ³	Substitutions uniquely shared between <i>H. erectus</i> and any HND individual.
ALB	609	245	0	53 (24)	16 (7)	67 (30)	22 (10)	0
AMBN	447	140	0	64 (18)	16 (7)	77 (23)	35 (12)	0
AMELX	206	182	0	7 (5)	1 (1)	17 (14)	3 (3)	0
COL17 α 1	1,498	67	0	61 (2)	21 (1)	123 (3)	62 (1)	0
ENAM	1,142	238	0	159 (26)	38 (4)	200 (37)	90 (12)	0
MMP20	483	99	0	32 (12)	8 (3)	50 (17)	21 (5)	0

Table 5.11: Description of the protein alignments used to compare the two ancient samples sequenced and the reference dataset. ¹Single species alignment: built using a single individual per species/group. ²Full alignment: built using all individuals in the reference dataset (Supplementary Table 5.9), Supplementary Table 5.10)). ³The numbers of sites where the ancient samples are non-missing are indicated in brackets. HND refers to any individual contained within the clade composed of *H. sapiens*, Neanderthals, and Denisovans.

a. Inspecting the protein alignments for informative amino acid substitutions in the ancient samples. We found one unique amino acid substitution in *Homo antecessor* and three substitutions that are uniquely shared among *Homo antecessor* and at least one other hominin (Neanderthal, Denisovan or present-day human; Supplementary Table 5.11). In contrast, we found no substitutions that are unique to Dmanisi or are uniquely shared between Dmanisi and at least one other hominin. The unique amino acid substitution present in *Homo antecessor* requires one nucleotide change (Alanine (A) to Proline (P); codon GCT, most likely substituted to CCT; Supplementary Table 5.12)). To evaluate the plausibility of the unique amino acid substitution in *Homo antecessor*, we assessed whether the corresponding nucleotide position is segregating in any of the groups included in the full alignment and in the 1000 Genomes diversity panel. For the 1000 Genomes diversity panel, we used the Phase3 VCFs. We found that the three nucleotides coding for the amino acid present in great apes, including humans, are invariant; the A at position 1,089 in the protein COL17 α 1 is coded by the GCT codon present in all samples in the alignment (Supplementary Table 5.12). We further examined the frequency of this type of amino acid substitution in the set of proteins recovered from *Homo antecessor* and present in our protein alignments. We found 12 instances of an A-to-P substitution among the reference proteins analyzed. Alanine and proline are chemically compatible (they are both non-polar and hydrophobic), so this makes a plausible change to have occurred in the *Homo antecessor* lineage. Finally, to get a sense of how likely it

was to find one unique substitution in the *Homo antecessor* lineage given the sequences recovered, we estimated the average of unique substitutions in the modern and archaic hominin lineages. To do so, we counted the number of unique substitutions in each hominin sample when compared to *Homo antecessor*, *Gorilla*, *Pan*, and *Pongo* in the alignment comprising the seven concatenated proteins. In this case, we define an amino acid substitution as 'unique' in a given sample when present in that sample but not in *Homo antecessor*, *Gorilla*, *Pan*, or *Pongo*. To account for missing data in *Homo antecessor*, we only included the sites where the ancient sample is non-missing. We observed an average of 1.605 ± 0.68 unique substitutions for present-day humans, 2 unique substitutions for the Altai and Sidron Neanderthals and the Denisovan, and one unique substitution for the Vindija Neandethal (Supplementary Figure 5.21).

b. Informative amino acid substitutions recovered in both Atapuerca *Homo antecessor* and Dmanisi *Homo erectus* samples.

Among the proteins recovered for Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*, we obtained amino acid positions covered in both samples (Extended Data Fig. 5.7). Most of such positions correspond to non-segregating amino acids in hominids, or are singletons. These are not particularly informative for differentiating between groups. There are, however, three positions recovered for both hominins that are segregating in our alignment: position 281 in MMP20, 648 in ENAM, and 255 in AMBN (Supplementary Table 5.13). The amino acid present at position 281 in the MMP20 protein is fixed to a threonine in *Pan* ($n=72$), *Gorilla* ($n=44$), and *Pongo* ($n=28$), Neanderthals ($n=7$) and the Denisovan, but it is segregating in present-day humans (both threonine (T) and glutamine (N) are present; $n=302$) (Supplementary Table 5.13; Supplementary Figure 5.22a). In this position, both Dmanisi and *Homo antecessor* carry a threonine. Similarly, the amino acid present at position 648 in ENAM is fixed to a T (threonine) in *Pan* ($n=72$; with the exception of a single sample), *Gorilla* ($n=44$), and *Pongo* ($n=28$), but it is segregating in present-day humans (both threonine/isoleucine are present; $n=302$) and Neanderthals ($n=7$) and the Denisovan carry the derived amino acid (I). The sequence support of the single *Pan* individual carrying an isoleucine is restricted to just three sequence reads, and hence not well supported. Here, *Homo antecessor* carries an isoleucine - which is almost exclusively present in presentday humans, Neanderthals, while the Denisovan and Dmanisi carry a threonine (Supplementary Table 5.13; Supplementary Figure 5.22b). Finally, the amino acid present at position 255 in the AMBN protein is fixed to an alanine in *Pan* ($n=72$), *Gorilla* ($n=44$), and *Pongo* ($n=28$), Neanderthals ($n=6$) and the Denisovan, but it is segregating in present-day humans (both alanine (A) and valine (V) are present; $n=302$) (Supplementary Table 5.13; Supplementary Figure 5.22c). In this case, both the Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* carry an alanine.

	COL17a1- (chr10:105796813-105796815) A -> P ³		
Nucleotide	A	G	C
SGDP (n=256)	1.0	1.0	1.0
1000 genomes (n=41/2,504) ¹	1.0	1.0	1.0
Neanderthal (n=6) ²	1.0	1.0	1.0
Denisovan (n=1)	1.0	1.0	1.0
<i>Pan</i> ssp. (n=43)	1.0	1.0	1.0
<i>Gorilla</i> ssp. (n=70)	1.0	1.0	1.0
<i>Pongo</i> ssp. (n=27)	1.0	1.0	1.0

Table 5.12: Nucleotide allele frequencies at the codons that translate to the amino acid where *Homo antecessor* carries a unique amino acid substitution.¹ None of these genomic sites are present in the 1000 genomes diversity panel. The frequencies of the alleles were obtained from the read alignments in the 41 high coverage genomes.² Allele frequencies for the Neanderthal and Denisovan genomes were obtained from the read alignments in Castellano, et al. [Castellano et al., 2014], Meyer, et al. [Meyer et al., 2012], and Hajdinjak, et al. [Hajdinjak et al., 2018]. For samples in Hajdinjak, et al., only the following samples were non-missing at the sites of interest: Mezmaiskaya_2, Les_Cottes_Z4-1514 and Goyet_Q56-1.³ This amino acid substitution is only found in *Homo antecessor*. While all groups presented in the table carry the codon GCT (AGC, since the protein is coded in the reverse strand) which translates into Alanine (A), *Homo antecessor* most likely carries triplet CCT (AGG) that translates into Proline (P).

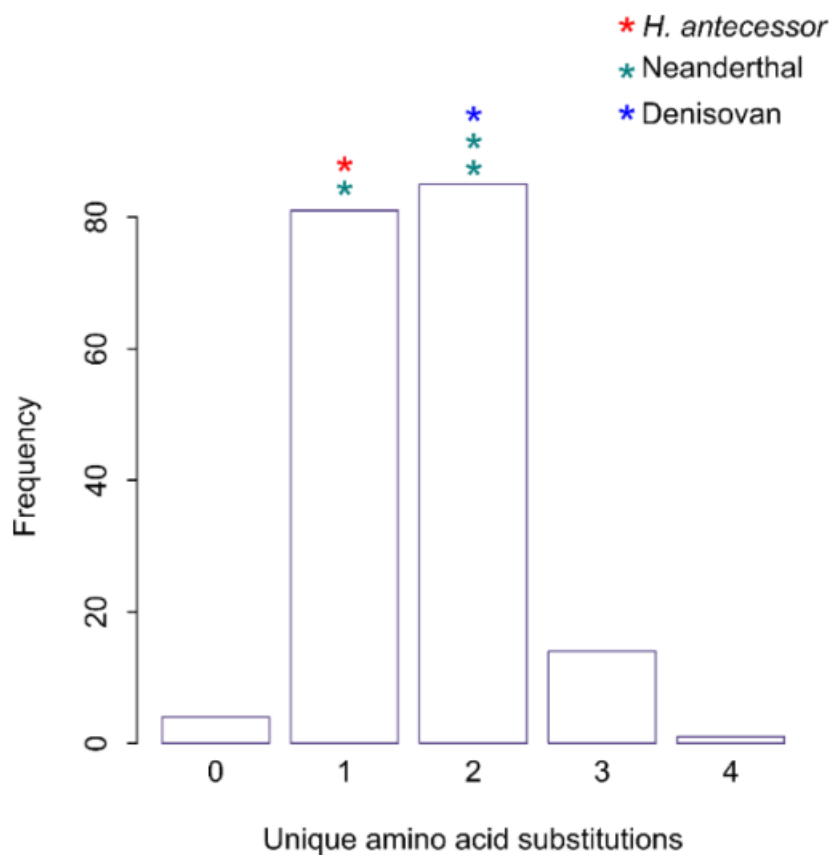


Figure 5.21: Frequency of unique amino acid substitutions among present-day humans. For each present-day human sample, we counted how many unique amino acid substitutions it carries when compared to *Homo antecessor*, *Gorilla*, *Pan*, and *Pongo*. Each bar corresponds to the number of present-day human samples that carry 0, 1, 2, 3 or 4 unique amino acid substitutions. The number of unique substitutions present in the archaic samples is indicated as stars in the top of the corresponding bar. This comparison only includes the sites where *Homo antecessor* is non-missing and the seven proteins recovered the ancient sample.

	ENAM+ (chr4:71509085-71509087) T->I			MMP20- (chr11:102477376-102477378) T->N			AMBN+ (chr4:71469603-71469605) A->V		
Nucleotide	A	C/T (rs7671281)	A	A	G/T (rs1784424)	T	G	C/T (rs7439186)	C
SGDP (n=256)	1.0	0.138/0.862	1.0	1.0	0.494/0.506	1.0	1.0	0.896/0.104	1.0
1000 genomes (n=41/2,504) ¹	1.0	0.182/0.818	1.0	1.0	0.580/0.420	1.0	1.0	0.881/0.119	1.0
Neanderthal (n=6-7) ²	1.0	0.0/1.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
Denisovan (n=1)	1.0	0.0/1.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Pan</i> ssp. (n=72)	1.0	0.977/0.023	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Gorilla</i> ssp. (n=44)	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0
<i>Pongo</i> ssp. (n=28)	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0	1.0	1.0/0.0	1.0

Table 5.13: Allele frequencies of the codons that translate to informative amino acid substitutions recovered for both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor*. ¹

The allele frequencies at positions chr4:71509086 and chr11:102477377 were obtained from the 1000 Genomes diversity panel that comprise ~2,500 individuals. The frequencies of the invariant sites, that were not present in the 1000 Genomes VCFs, were obtained from the read alignments in the 41 high coverage genomes also from the 1000 genomes. ²Allele frequencies for the Neanderthal and Denisovan genomes were obtained from the read alignments in Castellano et al. [Castellano et al., 2014], Meyer et al. [Meyer et al., 2012], and Hajdinjak et al. [Hajdinjak et al., 2018]. For samples in Hajdinjak, et al., only the following samples were non-missing at the sites of interest: Spy_94a (only for ENAM and MMP20), Mezmaiskaya_2 (only for MMP20 and AMBN), Les_Cottes_Z4-1514 and Goyet_Q56-1 (only for ENAM and MMP20).

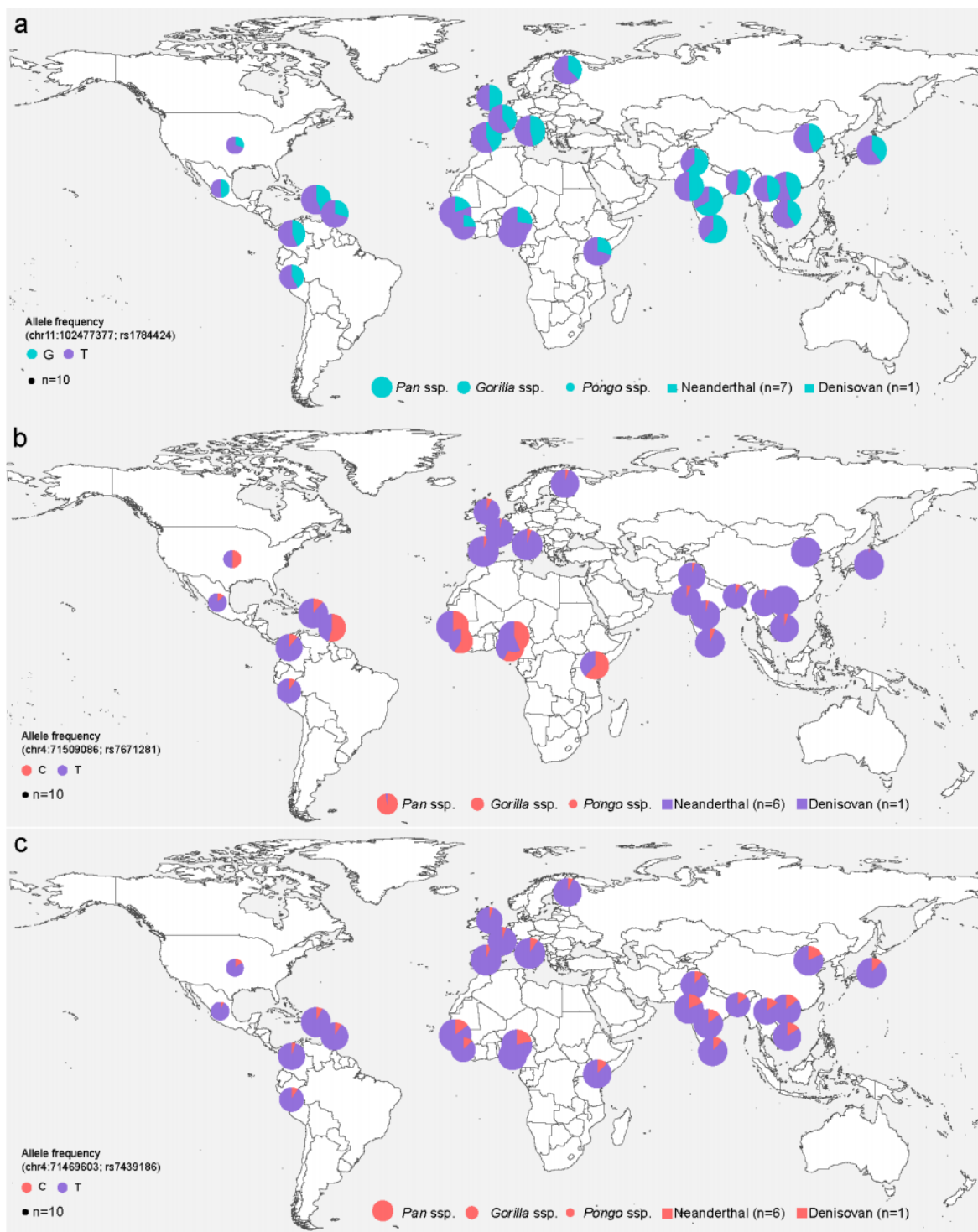


Figure 5.22

Figure 5.22: (Previous page.) **Allele frequencies in the 1000 Genomes diversity panel, among other great apes and in archaic hominins, at SNPs that code for three informative amino acids recovered for both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples.** **a**, Genomic site corresponding to the second position of the codon that translates to a threonine (if G) or an asparagine (if T). This codon corresponds to position 281 in the MMP20 protein. Both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples carry a threonine at this position. **b**, Genomic site corresponding to the second nucleotide in the codon that translates into a threonine (if C) or an isoleucine (if T). This codon corresponds to position 648 of the ENAM protein. Dmanisi *Homo erectus* carries a threonine. While we cannot experimentally distinguish between isobaric amino acids (L/I) for *Homo antecessor*, it is most likely that the Atapuerca *Homo antecessor* carries an isoleucine, since only isoleucine is present among present-day human variation. **c**, Genomic site corresponding to the second position of the codon that translate to an alanine (if C) or a valine (if T). This codon corresponds to position 255 in the AMBN protein. Both Dmanisi *Homo erectus* and Atapuerca *Homo antecessor* samples carry an alanine at this position. The base map was generated using public domain data from <http://www.natureearthdata.com/>.

5.10.2.3.4 Pairwise-distance between the ancient and modern samples

We used a distance-based approach to compare the *Homo antecessor* and *Homo erectus* protein sequences with present-day humans, Neanderthals and the Denisovan. The latter three hominin groups are hereafter collectively referred to as the HND group. We estimated the pairwise distance between the Atapuerca *Homo antecessor* or Dmanisi *Homo erectus* and individuals of the HND group, while considering all the proteins retrieved for each ancient sample and excluding X and Y chromosome-located proteins (AMELX and AMELY). Pairwise distances between samples were estimated using the phangorn R package [Schliep et al., 2017], considering the LG model [Le and Gascuel, 2008] and pairwise-deletions. In this case, we used the full alignment comprising the complete set of samples in the dataset (multiple individuals per species). The average distance between *Homo antecessor* and individuals from the HND group was 0.0036228 (0.0035712 when excluding AMELX and AMELY), whereas the average distance between members of the HND group is approximately three times smaller (0.0010827, or 0.0011429 when excluding AMELX and AMELY; Supplementary Figure 5.20a, b). The distance between *Homo antecessor* and the HND group is significantly different to that between pairs of samples from the HND group (p -value < 0.001; Mann-Whitney U Test). This suggests that the sequence is distantly related to the corresponding sequences in Neanderthals, Denisovans, and present-day humans.

In the case of Dmanisi *Homo erectus* individual, we obtained an average pairwise-distance between the ancient sample and individuals in the HND group of 0.0017148 (0.0020855, when excluding AMELX) and of 0.0011 (0.0011455 when excluding AMELX) between members of the HND group (Supplementary Figure 5.23c, d). In this case, we cannot exclude that Dmanisi belongs to the HND group, possibly due to the limited amount of informative positions retrieved for this sample, and the absence of a Dmanisi-unique SAP.

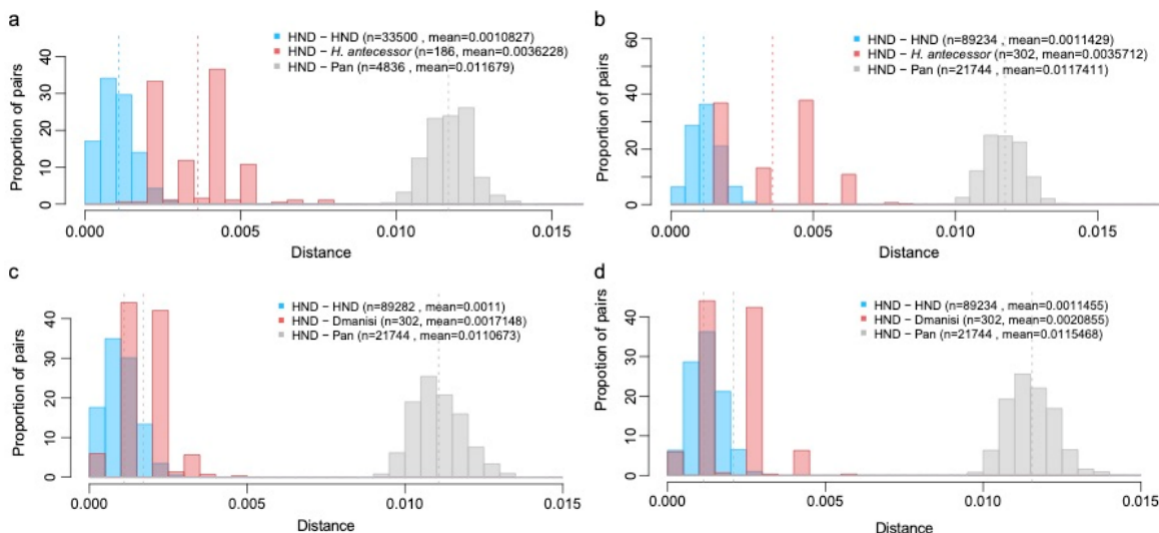


Figure 5.23: Pairwise-distances. **a**, Distribution of pairwise distances between individuals of the HND group (Humans, Neanderthals and Denisovan), *Homo antecessor*, and *Pan* for the seven proteins (ALB, AMBN, ENAM, MMP20, COL17 α 1, AMELX and AMELY), and **b**, excluding X- and Y-chromosome located proteins (AMELX and AMELY). **c**, Distribution of pairwise distances between individuals of the HND group, Dmanisi *Homo erectus*, and *Pan* for the six proteins (ALB, AMBN, ENAM, MMP20, COL17 α 1 and AMELX), and **d**, excluding the X-chromosome located protein (AMELX). The distance between pairs of samples is estimated based on the concatenated alignments and using the phangorn R package via the LG model and accounting for pairwise-deletions. HND refers to any individual contained within the clade composed of *H. sapiens*, Neanderthals, and Denisovans.

5.10.2.3.5 Phylogenetic placement of the ancient sequences

We sought to build phylogenetic trees using the aforementioned sequences, by applying two different approaches: a maximum likelihood and a Bayesian approach.

Maximum-likelihood (ML) approach. Using the single species alignments, we built ML trees for each protein and for a concatenated alignment of all seven of the available protein sequences using PhyML v3, for each of the ancient samples [Guindon et al., 2010] (Supplementary Figure 5.21, 5.22). For each alignment we optimized the tree topologies, branch lengths and substitution rates (-o tlr) under the JTT model (-m JTT). Additionally, we obtained maximum likelihood estimates for the gamma distribution shape parameter (-a e) and the proportion of invariable sites (-v e). For each alignment, we started from three random trees and for each optimization step, we kept the best tree between those generated through the 'nearest neighbor interchange' and 'subtree prune and regraft' routines (-s BEST -rand_start -n_rand_starts 3). Support for each bipartition was obtained based on 100 non-parametric bootstrap replicates.

In the case of *Homo antecessor* (Supplementary Fig. 5.24), each of the individual gene trees produced using the individual proteins provided poor resolution for resolving known species relationships. Four out of the seven gene trees place *Homo antecessor* within a hominin-only clade or as the closest outgroup to a hominin-only clade. The tree based on the

concatenated alignment helps to resolve most of the known species relationships, and places the *Homo antecessor* sequence as the closest outgroup to Neanderthals, present-day humans and the Denisovan with 100% bootstrap support.

For the *Homo erectus* sample (Supplementary Figure 5.25), none of the individual gene trees produced a reliable topology for great ape relationships, and most inferred nodes have poor bootstrap support. In the tree based on the concatenated alignment, the Dmanisi *Homo erectus* is placed as the closest outgroup to present-day humans, Neanderthals and the Denisovan. However, the bootstrap support is, again, very poor (54%).

Maximum-likelihood approach including all present-day human samples.

In addition to the single-species tree, we estimated a maximum likelihood tree including all present-day human samples and other great apes present in the dataset for the Atapuerca *Homo antecessor* individual (Supplementary Table 5.9). Note that, to include the AMELY protein recovered for *Homo antecessor*, we only included male individuals in the tree. We used the seven-protein concatenated alignment, PhyML v3 and the same parameters described above. Additionally, when more than one individual was identical across the seven proteins considered, we kept only one of them. Consistent with the single-species tree, the resulting phylogeny places *Homo antecessor* as an outgroup to all individuals in the HND clade (Supplementary Figure 5.26). However, the support for this bipartition is smaller: 41% of the trees place *Homo antecessor* as an outgroup to the HND clade, while in the remaining 59% *Homo antecessor* is placed within the HND clade. We then tested, whether using different combinations of present-day human samples in the phylogenetic inference affects the position of *Homo antecessor*. We created 1,000 different alignments, each containing 7 randomly sampled present-day humans from the 305 in our dataset. For each alignment, we used PhyML v3 and the same parameters described above, and evaluated the position obtained for *Homo antecessor* with the respect to the HND clade. In 98.7% of the trees, *Homo antecessor* is placed as an outgroup to the HND clade, while in the remaining 1.3% it is placed within Neanderthals, forming a clade with the Vindija Neanderthal.

Effect of missing data on the phylogenetic inference.

To assess the effect of missing data in *Homo antecessor*, and whether missing data could artificially yield the phylogenetic result we observe in the ancient sample, we performed a downsampling experiment by adding missing data to one present-day human sample (ERZ324268) and the Altai Neanderthal. For each of those samples, we created 100 independent replicates, where each contains a similar amount of missing data to *Homo antecessor*, distributed randomly in blocks of similar size as observed in the ancient sample. For each sample and replicate, we created a concatenated protein alignment consisting of the given replicate and all samples present in the single-species dataset, excluding the ancient sample. We then used PhyML (with the parameters described in the section above) to estimate a ML tree for each replicate and sample, and compared the resulting topologies to the topology obtained using the same alignment without missing data. For sample ERZ324268, we recover a topology similar to the one obtained without missing data in 99% of the replicates, while in 1% of the replicates the sample forms a polytomy with the Denisovan and the Sidron Neanderthal. For the Altai Neanderthal, we recover the topology obtained without missing data in 43% of the cases, while 57% of the replicates result in a polytomy involving all hominin samples.

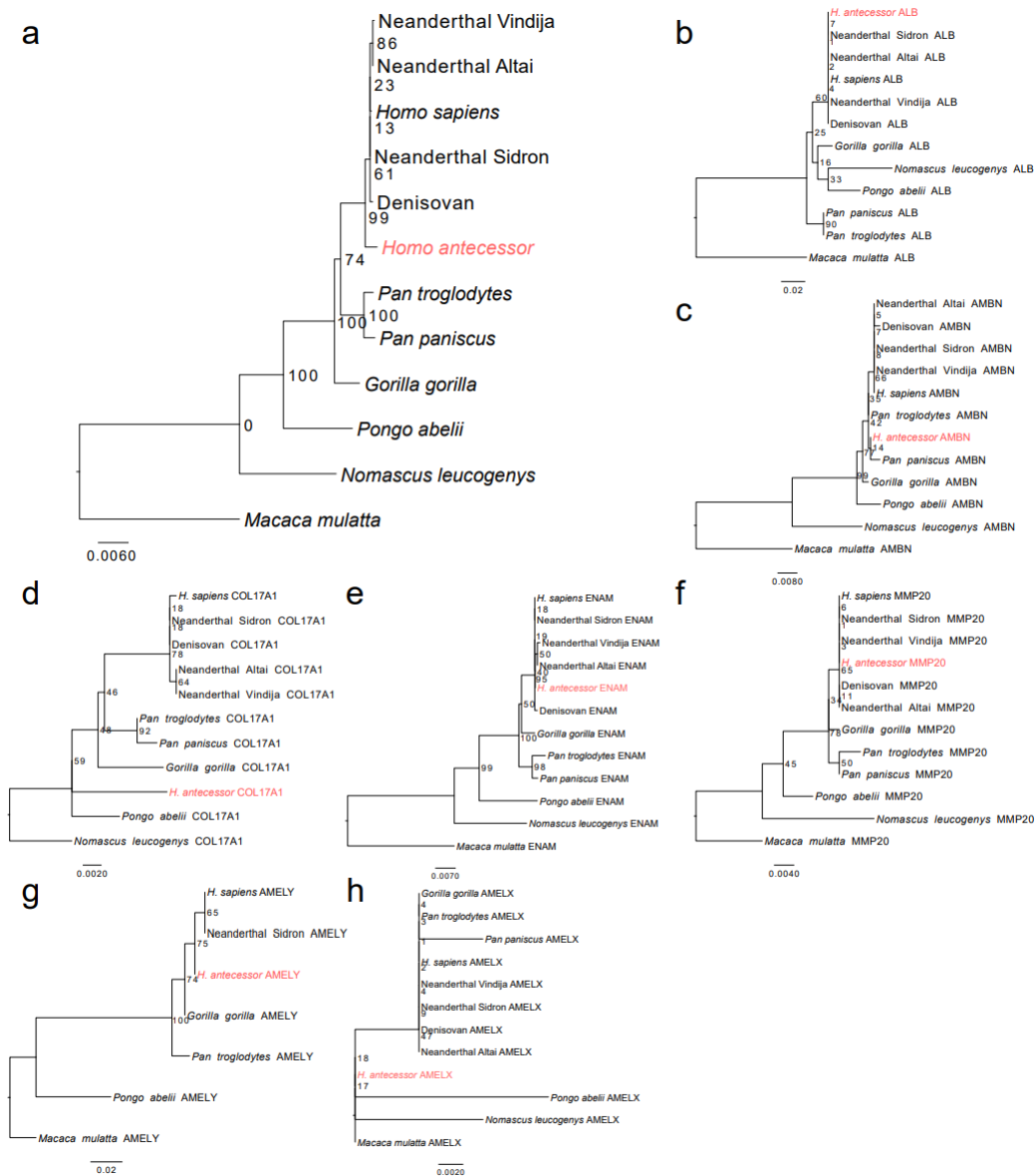


Figure 5.24: Atapuerca *Homo antecessor* maximum likelihood trees. a, Maximum-likelihood tree based on a concatenated alignment of the seven reconstructed *Homo antecessor* proteins. b-h, Individual protein trees. Support values for each bipartition were estimated through 100 non-parametric bootstrap replicates.

Bayesian phylogenetic inference.

To assess the robustness of the ML inference results, we also performed Bayesian phylogenetic inference based on the concatenated alignments using mrBayes [Ronquist et al., 2012] (v3.2). In this case, we partitioned the alignments by gene and for each partition we estimated the substitution rates, the shape parameter of a gamma distribution to model across-site rate variation (exponentially distributed prior), and the proportion of invariable sites (uniformly distributed prior; unlink Statefreq=(all) Ratemultiplier=(all) Aamodel=(all) Shape=(all) Pinvar=(all)). The tree topology and branch lengths were inferred for all the par-

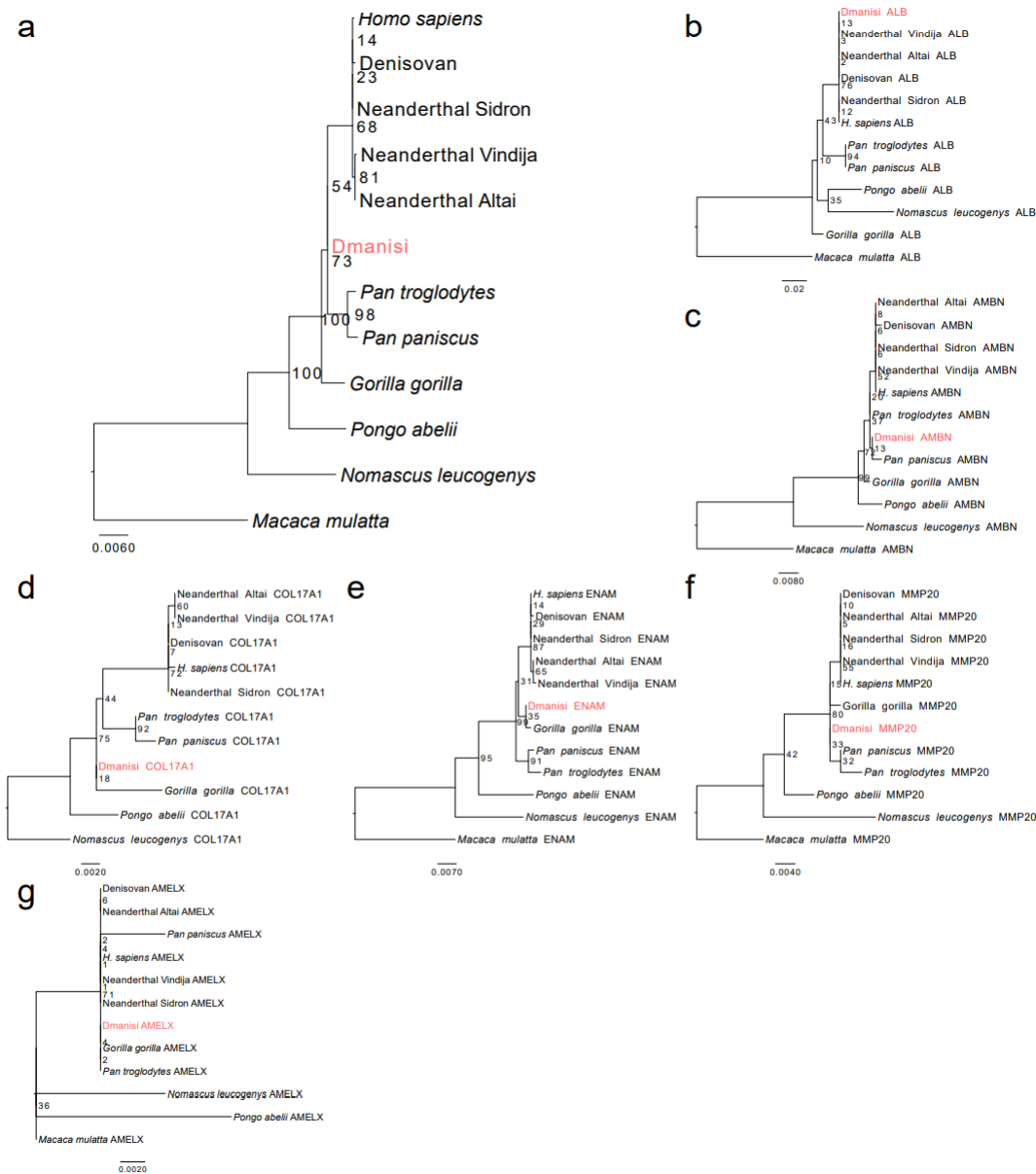


Figure 5.25: Dmanisi *Homo erectus* maximum likelihood trees. a, Maximum-likelihood tree obtained using PhyML and a concatenated alignment of the six reconstructed proteins for Dmanisi individual. b-h, Individual protein trees. Support values for each bipartition were estimated through 100 non-parametric bootstrap replicates.

titions jointly. To obtain posterior probabilities of unknown parameters, we ran a Metropolis-coupled Markov-chain Monte Carlo (MCMCMC) algorithm with four chains and a temperature parameter of 0.2 for 5,000,000 cycles sampling every 500 steps, after which we discarded the first 1,250,000 runs as burn-in. Convergence of the algorithm was evaluated using Tracer [Rambaut et al., 2018] (v.1.7.0). In particular, we required that the effective sample sizes for each estimated parameter were greater than 200, and that the overall log-likelihood of the run did not fluctuate substantially. Bayesian inference performed via mrBayes was

performed using the CIPRES Science Gateway [Miller et al., 2010]. In agreement with the maximum-likelihood approach, the Atapuerca *Homo antecessor* and Dmanisi *Homo erectus* individuals are placed as an outgroup to the HND clade with a posterior probability of 1 and 0.613, respectively (Supplementary Figure 5.27; Extended Data Figure 5.11).

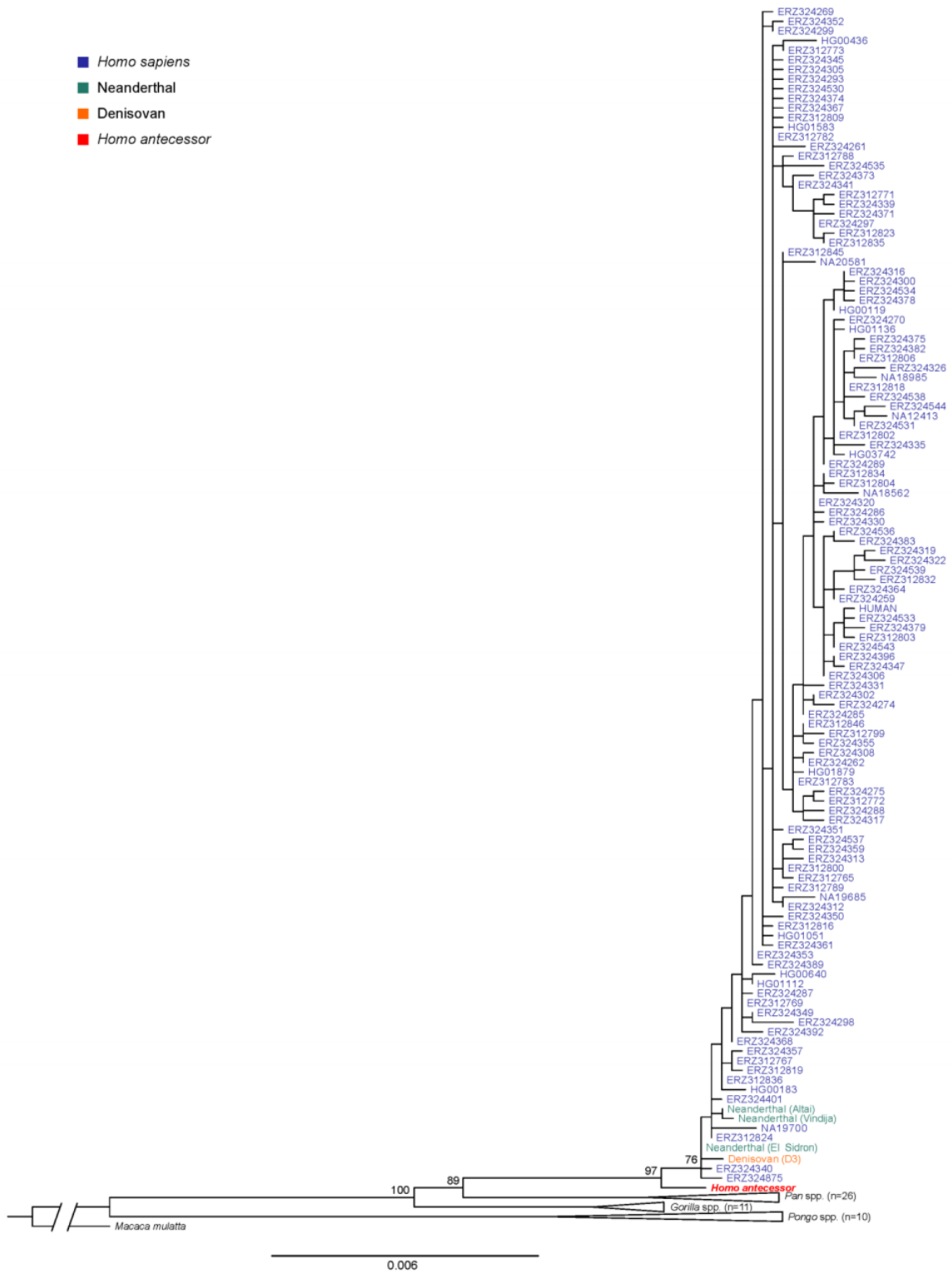


Figure 5.26: Maximum-likelihood tree based on a concatenated alignment of the seven protein sequences recovered for the Atapuerca *Homo antecessor* individual. Support values were estimated through 100 non-parametric bootstrap replicates.

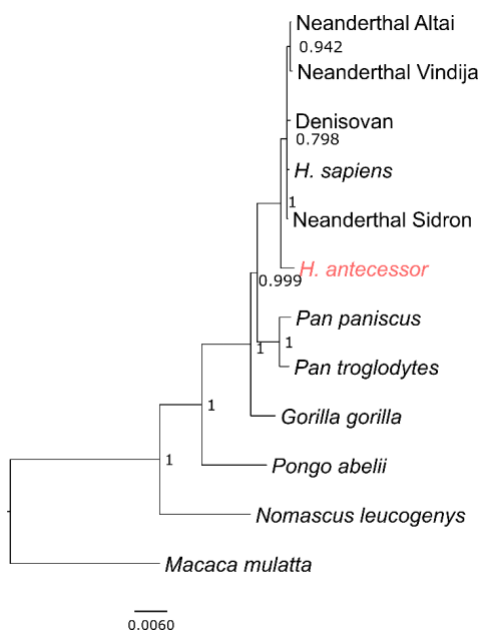


Figure 5.27: *Homo antecessor* Bayesian phylogenetic tree. Tree obtained using the seven-protein concatenated alignment and MrBayes. Posterior probabilities are indicated for each bipartition. *Macaca mulatta* was used as outgroup.

5.10.2.3.6 BEAST analysis

We used the Bayesian approach implemented in BEAST 2.5 [Bouckaert et al., 2019] to infer sequence divergence times for the proteins sequenced and between the different groups in the dataset and *Homo antecessor*. For this analysis, we used the seven-protein concatenated alignment including the Neanderthals, the Denisovan, seven randomly chosen *Homo sapiens* sequences and a single individual per great ape species. The alignment was partitioned by gene and a coalescent constant population model was used for the tree prior. The ages of the ancient samples included in the analysis (Vindija Neanderthal: 52 ka [Prüfer et al., 2017], Altai Neanderthal: 112 ka [Prüfer et al., 2017], Denisovan: 72 ka [Prüfer et al., 2017] and *Homo antecessor* 860.5 ka [Duval et al., 2018]) were used as tip dates for calibration. For each of the seven partitions we used the JTT substitution model with four categories for the gamma parameter, for which we allowed the MCMC chain to sample the shape of the gamma distribution (with an exponentially distributed prior) and assigned independent clock models. Additionally, we set a prior for the divergence time of great apes to 23.85 ± 2.5 Ma (normally distributed) [Besenbacher et al., 2019], and rooted the tree using the macaque (*Macaca mulatta*). Note that the overall topology of the tree was estimated for the seven partitions jointly. We assessed the convergence of the algorithm using Tracer [Rambaut et al., 2018] v1.7.0. Finally, since including all present-day humans in the analyses led to convergence problems, we instead repeated the analyses for 100 alignments, each of them consisting of seven different present-day humans chosen randomly. While the topology within the clade comprising present-day humans, Neanderthals and Denisovan (HND) was not consistent in all of the 100 alignments, 99 of them consistently placed *Homo antecessor* as an outgroup

to the HND clade. The one alignment that did not place *Homo antecessor* as an outgroup, it placed it as an outgroup to Neanderthals and the Denisovan.

The inferred tree for one of the 99 replicates is shown in Figure 5.3a. We note that the estimated tree in Figure 5.3a does not place Neanderthals as a monophyletic clade, which is likely due to incomplete lineage sorting, and the poor phylogenetic resolution afforded by sampling only a few gene trees. We assessed the proportion of sampled trees from the MCMC chain in which the inferred HND-*Homo antecessor* divergence was more ancient than that estimated for the HND clade for each of the 100 replicates. We found that in 95% of the trees, the divergence of *Homo antecessor* predated the first divergence within the HND clade (Fig. 5.2b).

These estimates rely on gene trees reconstructed from only seven protein sequences for which it was possible to obtain data for *Homo antecessor*. We therefore caution that these divergence times only reflect an approximation to the average genetic divergence time. They are likely an upper boundary for the population split time between these groups, which is expected to be more recent than the average genomic divergence. Nevertheless, assuming incomplete lineage sorting is not prevalent in these gene trees, our results support the placement of *Homo antecessor* as an outgroup to the HND group.

Chapter 6

Enamel proteome shows that *Gigantopithecus* was an early diverging pongine

Frido Welker^{1*}, Jazmín Ramos-Madrigal¹, Martin Kuhlwilm², Wei Liao^{3,4}, Petra Gutenbrunner⁵, Marc de Manuel², Diana Samodova⁶, Meaghan Mackie^{1,6}, Morten E. Allentoft⁷, Anne-Marie Bacon⁸, Matthew J. Collins^{1,9}, Jürgen Cox⁵, Carles Lalueza-Fox², Jesper V. Olsen⁶, Fabrice Demeter^{7,10}, Wei Wang^{11*}, Tomas Marques-Bonet^{2,12,13,14*} & Enrico Cappellini^{1*}

Nature 576, 262–265 (2019). <https://doi.org/10.1038/s41586-019-1728-8>
Published: 13 November 2019

¹Evolutionary Genomics Section, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

²Institute of Evolutionary Biology (UPF-CSIC), University Pompeu Fabra, Barcelona, Spain.

³School of Earth Sciences, China University of Geosciences, Wuhan, China.

⁴Anthropology Museum of Guangxi, Nanning, China.

⁵Computational Systems Biochemistry, Max Planck Institute of Biochemistry, Martinsried, Germany.

⁶Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark.

⁷Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark.

⁸CNRS FRE 2029 BABEL, Université Paris Descartes, Faculté de Chirurgie Dentaire, Paris, France.

⁹Department of Archaeology, University of Cambridge, Cambridge, UK.

¹⁰UMR7206 Eco-anthropologie, Muséum national d'Histoire naturelle, Musée de l'Homme, Paris, France.

¹¹Institute of Cultural Heritage, Shandong University, Qingdao, China.

¹²Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain.

¹³Centre for Genomic Regulation (CNAG-CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.

¹⁴Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain.

*e-mail: frido.welker@bio.ku.dk; wangw@sdu.edu.cn; tomas.marques@upf.edu; ecappellini@bio.ku.dk

6.1 Abstract

Gigantopithecus blacki was a giant hominid that inhabited densely forested environments of Southeast Asia during the Pleistocene epoch [Zhang and Harrison, 2017]. Its evolutionary relationships to other great ape species, and the divergence of these species during the Middle and Late Miocene epoch (16–5.3 million years ago), remain unclear [Harrison, 2010, Begun, 2007]. Hypotheses regarding the relationships between *Gigantopithecus* and extinct and extant hominids are wide ranging but difficult to substantiate because of its highly derived dentognathic morphology, the absence of cranial and postcranial remains [Zhang and Harrison, 2017, Begun, 2007, Kelley, 2002, Miller et al., 2008,

Grehan and Schwartz, 2009], and the lack of independent molecular validation. We retrieved dental enamel proteome sequences from a 1.9-million-year-old *G. blacki* molar found in Chuifeng Cave, China [Shao et al., 2014, Wang, 2009]. The thermal age of these protein sequences is approximately five times greater than that of any previously published mammalian proteome or genome. We demonstrate that *Gigantopithecus* is a sister clade to orangutans (genus *Pongo*) with a common ancestor about 12–10 million years ago, implying that the divergence of *Gigantopithecus* from *Pongo* forms part of the Miocene radiation of great apes. In addition, we hypothesize that the expression of alpha-2-HS-glycoprotein, which has not been previously observed in enamel proteomes, had a role in the biomineralization of the thick enamel crowns that characterize the large molars in *Gigantopithecus* [Bartlett et al., 2006, Dean and Schrenk, 2003]. The survival of an Early Pleistocene dental enamel proteome in the subtropics further expands the scope of palaeoproteomic analysis into geographical areas and time periods previously considered incompatible with the preservation of substantial amounts of genetic information.

6.2 Introduction

G. blacki is an extinct, potentially giant hominid species that once inhabited Asia. It was first discovered and identified by von Koenigswald in 1935, when he described an isolated tooth that he found in a Hong Kong drugstore [von Koenigswald, 1935]. The entire *G. blacki* fossil record, dated between the Early Pleistocene (about 2 million years ago (Ma)) and the late Middle Pleistocene (about 0.3 Ma) [Zhang et al., 2014], includes thousands of teeth and four partial mandibles from subtropical Southeast Asia [Zhang and Harrison, 2017, Zhao and Zhang, 2013, Pei, 1965]. All the locations at which *G. blacki* remains have been found are in or near southern China, stretching from Longgupo Cave, just south of the Yangtze River, to the Xinchong Cave on Hainan Island and, possibly, into northern Vietnam and Thailand [Bocherens et al., 2017, Ciochon et al., 1996].

To address the evolutionary relationships between *Gigantopithecus* and extant hominoids, we performed protein extractions on dentine and enamel samples of a single molar (CF-B-16) found in Chui Feng Cave, China (Extended Data Figs. 6.4, 6.5), that has been morphologically assigned to *G. blacki* [Shao et al., 2014, Wang, 2009]. The site has been dated, using multiple approaches, to 1.9 ± 0.2 Ma. We processed enamel and dentine samples using recently established digestion-free protocols that were optimized for extremely degraded ancient proteomes [Cappellini et al., 2019] (see 6.3 Methods). Enamel demineralization was replicated using two acids, trifluoroacetic acid (TFA) and hydrochloric acid (HCl).

6.3 Methods

Chuifeng Cave

The Chuifeng Cave (23°34'27" N, 107°0'22" E) is one of the most representative sites for the Early Pleistocene *G. blacki* fauna [Wang, 2009]. The site is located in the Bubing Basin in the northwestern part of the Guangxi Zhuang Autonomous Region, south China (Extended Data Fig. 6.4). The cave is 19 m long, 0.5–2 m wide and 1.5–5 m high, penetrating the limestone from southeast to northwest at a height of ~77 m above the local valley floor. A fossiliferous sandy clay with a few limestone breccias fills most of the cave, with an average depth of 1.3 m (Extended Data Fig. 6.5). Four excavation areas (A, B, C and D) were excavated down to limestone bedrock at 10-cm intervals. Twenty-four large mammalian species, including 92 *G. blacki* teeth, were unearthed from the cave [Wang, 2009]. The Chuifeng Cave mammalian fauna is characterized by the occurrence of typical Early Pleistocene species, such as *Hystrix magna*, *Sinomastodon sp.*, *Stegodon preorientalis*, *Ailuropoda microta*, *Pachycrocuta licenti*, *Tapirus sanyuanensis* and *Sus peii* [Wang, 2009]. This mammalian fauna is comparable with other *Gigantopithecus*-containing faunas of the Early Pleistocene in southern China, such as Baikong [Jin et al., 2014], Longgupo [Huang R; Gu, Y; Larick, R; Fang, Q; Yonge, C; de Vos, J; Schwarcz, H P; Rink, W J, 1995] and Liucheng [Pei, 1957]. The mammalian fauna composition is consistent with the results of combined electron spin resonance (ESR)/U-series dating and sediment palaeomagnetic studies (~ 1.9 Ma) [Sun et al., 2014]. In the present study, we collected one well-preserved *G. blacki* tooth (excavation number CF-B-16) for palaeoproteomic analysis. This tooth was excavated from area B at a depth of 90 cm from the sediment surface and, based on its stratigraphic position, dated to ~1.9 Ma. No other specimens were tested before CF-B-16, and no specific selection was made as to which *Gigantopithecus* tooth would be analysed.

Thermal age

Thermal age was calculated to allow comparison with previously published ancient genomes, ancient proteomes and collagen peptide mass fingerprinting studies, from other temporal and geographical localities. Temperature estimates for the hominin occupation of Dmanisi based on herpetological fauna suggest a temperature about 3.1 °C above the current MAT [Blain et al., 2014], while the sea surface temperature record used [Demarchi et al., 2016] predicts a negative ΔT at the time of hominin occupation. Given this discrepancy and the widely different temperature estimates for the Last Glacial Maximum (LGM) in the Caucasus, we conservatively use a scale factor of 0, correlating with a ΔT of approximately -0.2 °C and a current MAT of 11.2 °C. Our thermal age prediction for Dmanisi (2.2 Myr@10 °C) should therefore be seen as conservative. The thermal age for Chuifeng Cave was calculated with a general lapse rate between mean annual temperature (MAT) and altitude of 5.0 °C/km, a scale factor of 0.7 and a ΔT at LGM of -3 °C. The actual ΔT at LGM might have been more pronounced, again leading to a conservative estimate for the thermal age of Chuifeng Cave also. MAT was estimated based on the ten closest weather stations listed in publicly accessible World Meteorological Organization (WMO) data (Extended Data Table 6.2). Thermal age calculations are, among other factors, altitude dependent, but only five out of these ten weather stations have altitudes directly associated with them. We therefore estimated the altitude of the other five

weather stations through an online resource (<https://www.advancedconverter.com/map-tools/find-altitude-by-coordinates>). The correlation between WMO altitude and estimated altitude was $R^2 = 0.99$, providing sufficient validity to our estimated altitudes. The MATs for all weather stations were then averaged to obtain an approximate MAT for Chuifeng Cave. Next, thermal age was calculated for chronological ages of 1.7 Myr, 1.9 Myr and 2.1 Myr, giving estimates of the minimum (9.2 Myr@10 °C), maximum (15.0 Myr@10 °C) and mean (11.8 Myr@10 °C) thermal ages associated with the Chuifeng Cave fauna within a 95% confidence interval (Fig. 6.1). The Chuifeng Cave proteome is therefore, to our knowledge, substantially older than the oldest collagen peptide mass fingerprint (Ellesmere Island, 0.003 Myr@10 °C), oldest mammalian genome (Thistle Creek, 0.03 Myr@10 °C), oldest hominin genome (Sima de los Huesos, 0.25 Myr@10 °C) and oldest enamel proteome (Dmanisi, 2.2 Myr@10 °C) published to date [Demarchi et al., 2016]. Full thermal age calculations can be found in Supplementary Data 2.

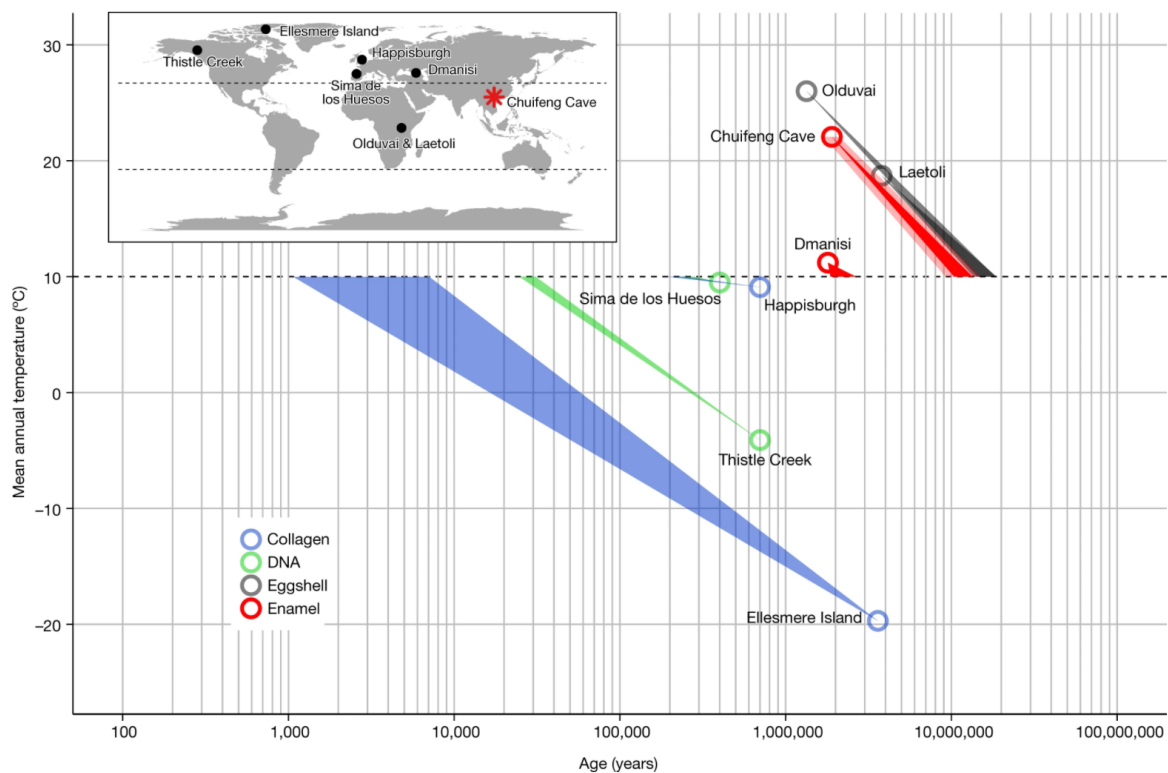


Figure 6.1: Chronological (circles) and thermal (coloured wedges, @10 °C) ages of other Cenozoic ancient genomes and proteomes are given for comparison. For Chuifeng Cave, dark red shading indicates thermal age calculated at a chronological age of 1.9 Ma and lighter shading indicates 95% confidence intervals based on a chronological age of 1.7 and 2.1 Ma, respectively. Inset, geographical location of Chuifeng Cave in the subtropics of Southeast Asia (red asterisk). Base map was generated using public domain data from <http://www.natureearthdata.com/>. The red asterisk also encloses the entire known geographical range of *G. blacki* fossils.

Protein extraction

Extraction of ancient proteins took place in facilities at the Natural History Museum of Denmark dedicated to extracting ancient DNA and ancient proteins. These laboratories include clean rooms fitted with filtered ventilation and positive air pressure [Hendy et al., 2018]. A negative extraction blank was processed alongside the ancient extractions, with the additional inclusion of injection blanks during tandem mass spectrometry (MS/MS) analysis to monitor potential protein contamination during all stages of analysis.

Two samples of enamel (185 and 118 mg) and one of dentine (192 mg) were removed from the same molar (CF-B-16) using a sterilized drill and crushed to a rough powder. One enamel sample and the dentine sample were demineralized in 1.2 M HCl at 3 °C for 24 h, and the other enamel sample was demineralized at the same temperature and duration using 10% TFA. Subsequently, solubilized protein residues were cleaned, concentrated and immobilized on C18 Stage-Tips using previously published methods [Cappellini et al., 2019]. No other samples from Chuifeng Cave were analysed before or during the analysis of CF-B-16.

LC-MS/MS analysis

The extracts were analysed by nanoflow liquid chromatography–tandem mass spectrometry (nanoLC-MS/MS) using a 15-cm capillary column (75 µm inner diameter, packed with 1.9 µm C18 beads (Reprosil-AQ Pur, Dr. Maisch)) on an EASY-nLC 1200 system (Proxeon) connected to a Q-Exactive HF-X mass spectrometer (Thermo Scientific). The nanoLC gradient and MS parameters followed a previously published Q-Exactive HF-X method [Mackie et al., 2018]. System wash blanks were performed before and after every sample to hinder cross-contamination.

Database construction

We constructed a protein sequence database for Hominoidea proteins known to be present in enamel proteomes (Supplementary Table 6.4), to which we added the homologous sequences from one Cercopithecoid (*Macaca mulatta*) as an outgroup for phylogenetic analysis. As few protein sequences are publicly available for *Pongo pygmaeus*, we predicted those sequences from publicly available genomic sequence data using the known gene coordinates of *Pongo abelii* homologues. Similarly, we generated de novo AMELY sequences for *P. abelii* and *P. pygmaeus*. Finally, we added common laboratory contaminants to allow spectra from such proteins to be confidently identified [Hendy et al., 2018].

Ancestral sequence reconstruction

Previous research indicates that cross-species proteomic effects, observed during spectral identification, substantially reduce the identification of phylogenetically informative amino acid positions at large evolutionary distances [Welker, 2018a]. We reasoned that this was likely to occur in the case of *Gigantopithecus* proteins [Welker, 2018b], and therefore reconstructed the ancestral protein sequences of enamel-specific proteins. Ancestral sequence reconstruction (ASR) was conducted across the entire Hominoidea phylogeny using PhyloBot [Hanson-Smith and Johnson, 2016]. Input sequences were constrained phylogenetically to (*Macaca*,(*Nomascus*,((*Pongo abelii*, *Pongo pygmaeus*),*Gorilla*),(*Homo*,((*Pan paniscus*, *Pan*

troglydytes)))))). We added those sequences to the reference protein database to account for them in the database search of PEAKS and MaxQuant.

Isoform variation

After obtaining complete protein sequences for all extant hominids, we added isoforms not present in UniProt or GenBank for the proteins AMELX, AMELY, AMBN, AMTN, KLK4 and TUFT1, including the reconstructed ASR sequences of these proteins, to the database. We assumed that the isoforms for these non-human hominids would result from identically placed alternative splicing across species and ancestral nodes (as also supported by all UniProt isoforms present for the studied proteins). Thus, we copied these alternative splicing sites onto the available reference sequences to create the missing isoforms. Database sequence names for these proteins were appended with ‘_ManIso2’ or ‘_ManIso3’.

Proteomic data analysis

Raw mass spectrometry data were searched per sample type (enamel, dentine, extraction blank and injection blanks) against a sequence database containing all common enamel proteins for all extant hominids (see above). We used PEAKS [Zhang et al., 2012] (v.7.5) and MaxQuant [Cox and Mann, 2008] (v.1.6.2.6) software. The de novo and error-tolerant implementations of PEAKS, and the dependent peptide algorithm implemented in MaxQuant, were used to generate possible, additional, SAP variation in enamel protein sequences. Such novel SAPs could represent unique amino acid substitutions on the *Gigantopithecus* lineage, which are not relevant to its phylogenetic placement but are relevant for dating the *Pongo–Gigantopithecus* divergence. Next, these potential sequence variants were added to a newly constructed sequence database and verified in separate searches in PEAKS and MaxQuant. We defined as variable modifications methionine oxidation, proline hydroxylation, glutamine and asparagine deamidation, pyro-glutamic acid from glutamic acid, pyro-glutamic acid from glutamine and phosphorylation (STY). No fixed modifications were selected. We did not use an enzymatic protease during sample preparation, therefore the digestion mode was set to ‘unspecific’. For PEAKS, peptide spectrum matches were only accepted with a false-discovery rate (FDR) $\leq 1.0\%$, and precursor mass tolerance was set to 10 ppm and fragment mass tolerance to 0.05 Da. For MaxQuant, peptide spectrum matches and protein FDR were set at $\leq 1.0\%$, with a minimum Andromeda score of 40 for all peptides. Protein matches were accepted with a minimum of two unique peptide sequences in at least one of the MaxQuant or PEAKS searches, including the removal of non-specific peptides after BLASTp searches of peptides matching non-enamel proteins against UniProt and GenBank databases. Proteins that are retained after applying these criteria are listed in Extended Data Table 6.1. Examples of annotated MS/MS spectra after MaxQuant analysis can be found in Supplementary Figs. 6.14 - 6.23.

Assessment of protein damage and degradation followed protocols explained elsewhere [Cappellini et al., 2019, Mackie et al., 2018, Welker et al., 2016] and included rates of deamidation and a comparison of observed peptide lengths. GRAVY index scores of peptide hydrophobicity were calculated using the R package Peptides, with the scale set to ‘KyteDoolittle’.

Phylogenetic and divergence analysis

Comparative reference dataset

We assembled a reference dataset with five protein sequences retrieved from the ancient sample (AMBN, AMELX, AMTN, ENAM and MMP20) and relevant extant species (Supplementary Table 6.4). Protein sequences for human (*Homo sapiens*), common chimpanzee (*P. troglodytes*), bonobo (*P. paniscus*), Sumatran orangutan (*P. abelii*), Western gorilla (*Gorilla gorilla*), rhesus macaque (*M. mulatta*) and the white-cheeked gibbon (*Nomascus leucogenys*) were obtained from the UniProt database. Additionally, we expanded our dataset with protein sequences from publicly available whole-genome sequence data from present-day great apes (in total 27 orangutans, 42 gorillas, 11 bonobos and 61 chimpanzees [Prado-Martinez et al., 2013, Nater et al., 2017, De Manuel et al., 2016]), as well as 19 human individuals from the Simons Genome Diversity Project (SGDP) [Mallick et al., 2016]. See 6.9 Supplementary Information for the human sample numbers taken from the SGDP dataset.

Reconstruction of protein sequences from whole-genome sequencing data

DNA sequence reads for reference samples used were mapped to the human genome (version hg19) using BWA-MEM v.0.7.5a-r405 (<http://bio-bwa.sourceforge.net/bwa.shtml>) with default parameters. PCR and optical duplicates were identified and removed using PICARD v.1.91 (<https://sourceforge.net/projects/picard/files/picard-tools/1.91/>). Single-nucleotide polymorphisms were called on the read alignments using the GATK UnifiedGenotyper (https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php).

To reconstruct the protein sequences from the genotype calls, we first created a consensus sequence for each of the five genes of interest and for each sample. Indels were not considered and a random allele was chosen at heterozygous positions. Next, we removed the intron sequences from each gene using the annotation of the reference human genome (hg19) available in the ENSEMBL database. For each of the in silico spliced genes, we performed a tblastn search [Altschul et al., 2002] using the human reference protein as the query. Finally, we obtained the translated protein sequences from the resulting alignments.

Assessing the phylogenetic position of *G. blacki*

We compared the *G. blacki* protein sequences with the corresponding homologues of the species in the reference panel. For each gene, we built two multiple sequence alignments using mafft [Katoh and Frith, 2012]. The first incorporated all samples in the reference panel ($n = 164$). The second incorporated only a single sample per species (Supplementary Table 6.5). To account for isobaric amino acids (leucines and isoleucines), which cannot be distinguished in the ancient protein data, we changed all isoleucines to leucines at positions where the ancient sample carried either of those amino acids. To assess the phylogenetic position of the ancient sample, two inference approaches were used: a maximum-likelihood and a Bayesian inference.

Maximum-likelihood approach. PhyML v.3.1 [Guindon et al., 2010] was used to infer a maximum-likelihood tree, branch lengths and substitutions rates for each individual protein alignment (Supplementary Fig. 6.13) and for the concatenated alignment. For each alignment, we started from three random trees (`-n_rand_starts 3 -s BEST -rand_start`), used the JTT model (`-m JTT -f m`), and obtained maximum likelihood estimates for the gamma distribution shape parameter (`-a e`) and the proportion of invariable sites (`-v e`). Support values were obtained for each bipartition based on 100 non-parametric bootstrap replicates. The bootstrap results per branch split are shown in Extended Data Fig. 6.9b.

Bayesian approach. As a complementary approach, we used MrBayes [Ronquist et al., 2012] and the concatenated alignment to infer the phylogenetic position of the ancient sample (Fig. 6.3), Extended Data Fig. 6.9b). We set an independent bipartition for each gene and estimated: substitution rates, across-site rate variation and the proportion of invariable sites (`unlink Statefreq = (all) Ratemultiplier = (all) Aamodel = (all) Shape = (all) Pinvar = (all)`). MrBayes was executed using the CIPRES portal [Miller et al., 2010]. The MCMC algorithm was set to 5,000,000 cycles with 4 chains and a temperature parameter of 0.2. The convergence of the algorithm was assessed using Tracer v.1.6.0 after discarding 25% of the iterations as burn-in. MrBayes was run against the reference sequence for each species (Extended Data Fig. 6.9c) or against 162 great ape individuals, one hylobatid and one cercopithecoid (Extended Data Fig. 6.10). Both of these analyses, as well as the PHyML maximum likelihood approach, resulted in the same topology. The analysis using a large number of individuals shows, however, that resolution within the genus *Pongo* is limited (Extended Data Fig. 6.10). Nevertheless, the placement of *Gigantopithecus* is fully supported.

Divergence time of *Gigantopithecus*

We estimated the divergence time between *Gigantopithecus* and the *Pongo* branch first by using a distance-based approach. We used the alignment of the amino acid sequences of reference genome sequences for each species as well as diversity data (see above). A distance matrix was created from the concatenated protein sequences of all individuals using the function `dist.ml` from the R package `phangorn` [Schliep, 2011] under the LG amino acid substitution model [Le and Gascuel, 2008]. We used pairwise exclusion to increase the amount of data for the present-day branches. We then calculated the mean difference of all orangutan sequences to all sequences from *Homo*, *Pan* and *Gorilla*, and the mean difference of all orangutan sequences to *Gigantopithecus* (Extended Data Fig. 6.9a). We used the average distance between orangutan and the other extant great apes as a scaling factor, assuming a divergence time between these branches [Besenbacher et al., 2019] of 23.8 Ma. Under this assumption, the molecular divergence of *Gigantopithecus* from the *Pongo* branch is 9.98 Ma. However, because Chuifeng Cave is dated to 1.9 Ma, this branch is likely to be underestimated and its age needs to be corrected to 11.88 Ma. We combine the 95% confidence interval of the distance matrix with the 95% confidence interval of the mutation rate estimate [Besenbacher et al., 2019], and add the upper and lower values of the 95% confidence interval for the Chuifeng Cave dating (1.7–2.1 Ma), and therefore suggest conservative upper and lower boundaries for the divergence of 8.91 and 15.65 Ma, respectively.

If mutation rates did not substantially differ between extant *Pongo* and *Gigantopithecus*, this estimate should reflect the molecular evolution of their common branch. We calculated the divergence between the other great apes, taking into account the mutation rate differences on these lineages as scaling factors [Besenbacher et al., 2019]. The resulting divergence time between *Gorilla* and the *Homo/Pan* branch is estimated at 10.27 Ma (7.9–13.25 Ma, 95% confidence interval), and the divergence between *Homo* and *Pan* at 8.72 Ma (8.06–13.81 Ma, 95% confidence interval). These values are in strong agreement with previous estimates [Besenbacher et al., 2019], suggesting that these protein sequences represent well the known phylogeny of the great apes. Clearly, all divergence time estimates scale with assumptions of mutation rates. We also caution that the small number of mutations in the peptide fragments in *Gigantopithecus* constitutes a severe limitation on the precision of these estimates on this branch. However, the phylogenetic position of *Gigantopithecus* as a sister clade to orangutans is also well supported in this analysis: a phylogenetic tree from a distance matrix of the reference sequences for these species (neighbour joining tree in phangorn; maximum likelihood computed with the pml function; 1,000 bootstrap replicates) separates *Gigantopithecus* from orangutans with 100% bootstrap support.

We used the program MrBayes [Ronquist et al., 2012] to estimate divergence time estimates in a Bayesian framework using the reference genome sequences. We defined *M. mulatta* as outgroup, grouped *Pan*, *Homo* and *Gorilla* together as well as *Pongo* and *Gigantopithecus*, and set the divergence time of the two groups with a uniform distribution of 17.739–26.061 Ma, using a previously published estimate [Besenbacher et al., 2019]. Furthermore, we set the divergence time of the macaques and apes at 26.061–39.9 Ma (from the maximum divergence time of the hominids to a very high divergence time of the apes). We used a variable mutation rate and the VT amino acid substitution model [Müller and Vingron, 2001] in five million iterations. This results in a divergence time for *Gigantopithecus*–*Pongo* of 10.14 Ma (4.76–15.79 Ma, 95% HPD interval). The divergence of *Gorilla* from the *Homo/Pan* branch is estimated at 8.59 Ma (4.62–13.56 Ma, 95% HPD interval), and the divergence of *Homo* and *Pan* at 5.78 Ma (2.64–9.53 Ma, 95% HPD interval). These are largely consistent with, but slightly younger than, previous estimates [Besenbacher et al., 2019, Langergraber et al., 2012], possibly owing to a mutation slowdown in these lineages compared to the *Pongo* lineage, which is not taken into account here. However, they seem in agreement with the fossil record indicating the origin of hominins around 6–8 Ma and the dating of a possible early Gorillini (*Chororapithecus*) around 7–9 Ma [Langergraber et al., 2012, Katoh et al., 2016, Senut et al., 2001, Brunet et al., 2002, Haile-Selassie, 2001]. Therefore, we conclude that the relative branch lengths of the tree (Fig. 5.3b) are concordant with the overall phylogeny and the estimates presented above.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All the mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD013838. Generated ancient protein consensus sequences for *G. blacki* can be found in Supplementary Data 1. Full thermal age calculations can be found in Supplementary Data 2.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2153-8>.

6.4 Results

We identified no endogenous proteins from the dentine, but instead recovered an ancient enamel proteome composed of 409 unique peptides matching 6 endogenous proteins: amelogenin X (AMELX), ameloblastin (AMBN), amelotin (AMTN), enamelin (ENAM), matrix metalloproteinase-20 (MMP20) and alpha-2-HS-glycoprotein (AHSG, also known as FETUA; Extended Data Table 6.1). This observation extends the survival of ancient mammalian proteins to a thermal age, obtained by normalizing the chronological age to a constant temperature of 10 °C, of approximately 11.8 million years (Myr)_{10 °C} (Extended Data Table 6.2). Such a thermal age is well beyond the thermally oldest DNA (0.25 Myr_{10 °C}, Sima de los Huesos, Spain [Meyer et al., 2016]), collagen (0.22 Myr_{10 °C}, Happisburgh, UK [Wadsworth and Buckley, 2014]) and enamel proteome (2.2 Myr_{10 °C}, Dmanisi, Georgia [Cappellini et al., 2019]) reported to date. The Chuifeng Cave enamel proteome is thus, to our knowledge, the oldest Cenozoic skeletal proteome currently reported (Fig. 6.1). The survival of a subtropical proteome from approximately 2 Ma suggests that chronologically older specimens from higher latitudes are likely to contain preserved ancient proteomes as well.

The content of the recovered enamel proteome is consistent with previously reported ancient enamel proteomes [Cappellini et al., 2019, Stewart et al., 2016, Castiblanco et al., 2015], with the addition of several peptides deriving from a single region of AHSG. Peptide matches to these proteins covered a minimum of 43 informative single amino acid polymorphisms (SAPs; Supplementary Table 6.5). In addition, the retrieved protein regions fell largely within areas previously recovered from an Early Pleistocene *Stephanorhinus* enamel proteome from Dmanisi [Cappellini et al., 2019] (Supplementary Fig. 6.12). The absence of peptides specific to amelogenin Y (AMELY) suggests that the sampled molar might have belonged to a female *Gigantopithecus*. Alternatively, male-diagnostic AMELY-specific peptides were not observed due to their degradation beyond the limit of detection of the instrument. The endogenous enamel proteome sequence coverage of 456 amino acids is lower than the previously recovered sequence coverage for a Dmanisi *Stephanorhinus* specimen (875 amino acids [Cappellini et al., 2019]; Supplementary Table 6.3). This observation is consistent with the older thermal age for Chuifeng Cave, compared to Dmanisi [Cappellini et al., 2019].

We replicated enamel demineralization using two acids (TFA and HCl). The chromatograms of these two extracts showed that different peptide populations were released (Extended Data Fig. 6.6). Owing to the partial acidic hydrolysis [Cristobal et al., 2017], which potentially occurs alongside demineralization, peptide populations with a wider range of acidity (Extended Data Fig. 6.7a) and hydrophobicity (Extended Data Fig. 6.7c) are generated using TFA. Our TFA-based demineralization returned 127 more unique non-overlapping peptide sequences than did the HCl-based demineralization (Extended Data Fig. 6.7e). The TFA extract, therefore, outperformed the HCl-based extraction, despite being carried out on a smaller amount of starting material [Cappellini et al., 2019]. Ultimately, the extended coverage of TFA-based demineralization increased the identification rate of informative SAPs, enhancing the phylogenetic information obtained (Extended Data Fig. 6.7d). Finally, the HCl- and TFA-demineralized samples had similar deamidation rates and average peptide lengths (Extended Data Fig. 6.8), which indicates that the two acids released peptide populations that were modified to the same extent.

The *Gigantopithecus* enamel proteome is characterized by extensive diagenetic modifications, such as high rates of deamidation (Fig. 6.2a) and a high degree of degradation, as indicated by relatively short peptide lengths (Fig. 6.2b); this is expected for an ancient proteome preserved in subtropical conditions. When quantifying peptide intensities using label-free quantification (LFQ) implemented in MaxQuant [Cox and Mann, 2008], we observed that summed and normalized MS1 spectral intensities were higher for shorter peptides than for longer peptides (Extended Data Fig. 6.7b). Finally, the peptide lengths of the Chui Feng Cave enamel proteome were shorter than those identified in thermally younger enamel proteomes (Fig. 6.2b).

Enamel-specific proteins are modified *in vivo* through protein phosphorylation, alternative splicing of AMELX, and MMP20- and KLK4-mediated proteolysis. Such modifications could survive in ancient proteomes. We detected evidence of surviving *in vivo* post-translational modifications, such as serine phosphorylation in the S-x-E/phS motif, which is recognized by the secreted kinase FAM20C (Fig. 6.2c). The FAM20C kinase regulates the phosphorylation of extracellular proteins involved in biomineralization [Tagliabracci et al., 2012]. We also observed two alternative-splicing-derived isoforms of AMELX (Fig. 6.2d). These observations are similar to those from other Early Pleistocene enamel proteomes [Cappellini et al., 2019]. The *Gigantopithecus* enamel proteome therefore demonstrates that such *in vivo* modifications can be recovered from hominid samples across the Pleistocene.

To achieve a protein-based phylogenetic placement of *Gigantopithecus*, we compared the enamel proteome sequences we retrieved with those of extant apes (Hominoidea). We used publicly available whole-genome sequence data to predict enamel protein sequences from relevant species [Prado-Martinez et al., 2013, Nater et al., 2017] (Supplementary Table 6.4, Supplementary Figs. 6.13–6.23). Our results show that *Gigantopithecus* represents a sister taxon to all extant orangutans (*Pongo spp.*) and forms a monophyletic group with extant pongines (Fig. 6.3a, Extended Data Figs. 6.9, 6.10). We then attempted to estimate the divergence time between *Gigantopithecus* and *Pongo* species using two approaches: (1) a pairwise distance approach and (2) a Bayesian approach using MrBayes (see 6.3 Methods). Although confidence intervals obtained for the divergence estimates of the *Pongo*–*Gigantopithecus* split are large, our results indicate that *Gigantopithecus* diverged from the extant *Pongo* species in the Middle or Late Miocene (approximately 10 Ma or 12 Ma using the Bayesian or pairwise distance approaches, respectively; Fig. 6.3b). This suggests that, despite an exclusively Pleistocene fossil record, *Gigantopithecus* is a member of an early radiation of pongines, whose diversity peaked during the Middle and Late Miocene (Fig. 6.3b). Our results thereby resolve not only the phylogenetic position of *Gigantopithecus*, but also renew the debate on the evolutionary relationships between extant hominids and early hominids present in the fossil record [Harrison, 2010].

The presence of AHSG in the *Gigantopithecus* proteome is intriguing, as this protein is not commonly observed in (modern) hominid enamel proteomes. All retrieved peptides derive from a single, highly conserved region that is bordered by disulfide cysteine bonds on either side (Extended Data Fig. 6.11). AHSG is highly glycosylated *in vivo*, but we observed no glycosylation during our bioinformatics analysis. The observed sequence contains regularly spaced aspartic acid residues that provide a suitable motif for binding to basic cal-

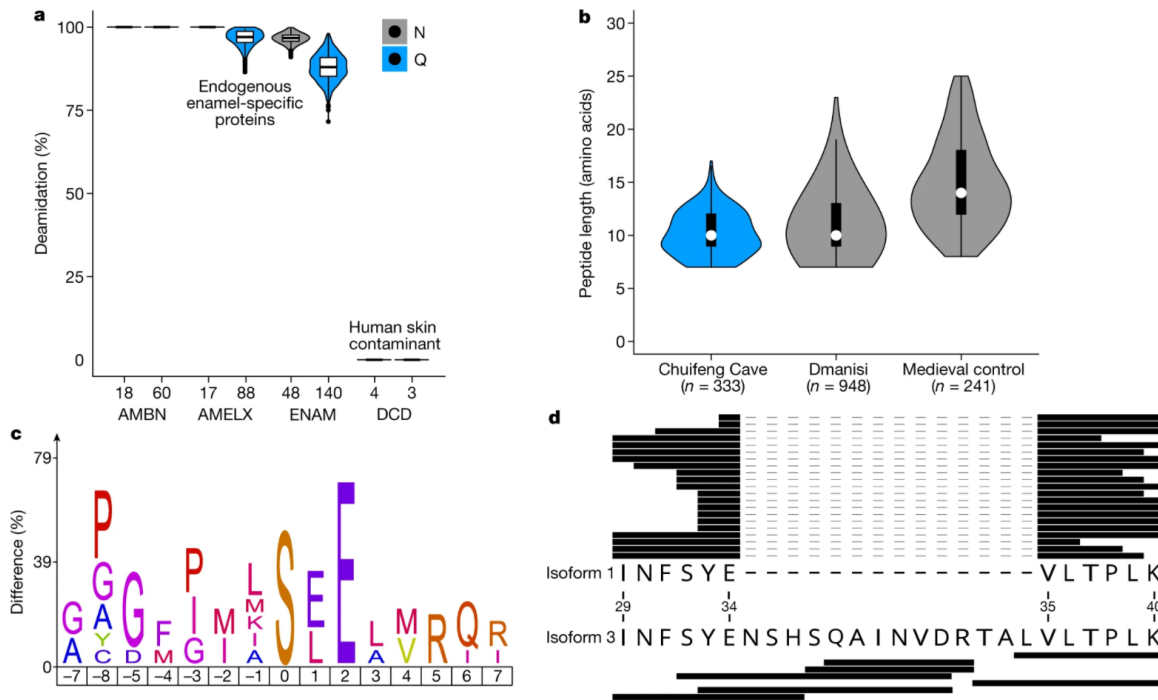


Figure 6.2: **a**, Violin plots of asparagine (N) and glutamine (Q) deamidation for selected proteins ($n = 1,000$ bootstrap replicates of intensity-based peptide deamidation [Mackie et al., 2018]). Human dermcidin (DCD) is included as an example of a non-deamidated contaminant. For AMBN, all observed asparagines and glutamines are deamidated. For AMELX, all asparagines are deamidated. For DCD, no observed asparagines and glutamines are deamidated. The number of peptides used for the calculation are shown at the bottom. **b**, Violin plots of peptide lengths for *Gigantopithecus*, an Early Pleistocene rhinoceros from Dmanisi and a Medieval control sample [Cappellini et al., 2019]. **c**, Sequence-motif analysis of the over-representation of specific amino acids around the phosphorylated amino acid (position 0; $n = 14$). **d**, Peptide coverage of AMELX protein isoforms. Matching peptides are indicated by black bars for isoform 1 ($n = 21$) and isoform 3 ($n = 7$). The latter includes an insertion due to alternative splicing between isoform 1, amino acid positions 34 and 35 (coordinates in reference to UniProt Accession number: Q99217-1 [AMELX_HUMAN]). **b** includes data on AMELX, AMBN, ENAM, AMTN and MMP20 only. For **a** and **b**, box plots define the range of the data (whiskers extending to $1.5\times$ the interquartile range), 25^{th} and 75^{th} percentiles (boxes) and medians (dots/lines). For **a**, outliers are indicated as black dots (beyond $1.5\times$ the interquartile range).

cium phosphate lattices [Tang and Skibsted, 2016]. The notion that this specific peptide sequence is involved in biomineral binding is supported by three observations: (1) this region is presented on the external surface of AHSG [Heiss et al., 2003]; (2) such surfaces have been demonstrated to bind biominerals in other systems as well [Demarchi et al., 2016]; and (3) this type of binding enhances peptide preservation [Demarchi et al., 2016]. AHSG acts as a key component of bone and dentine mineralization by inhibiting the extrafibrillar mineralization of collagen type I helices [Price et al., 2009] and has previously been hypothesized

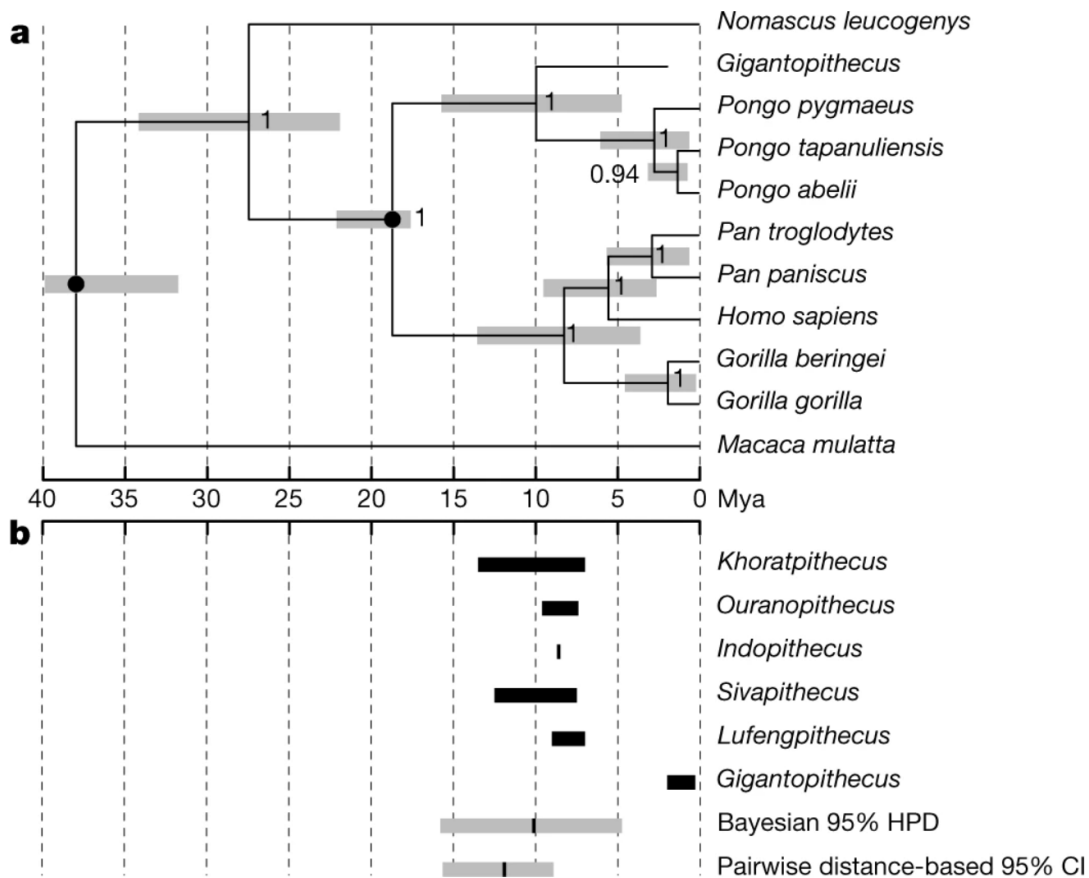


Figure 6.3: **a**, Time-calibrated Bayesian phylogeny of *Gigantopithecus*. Circled nodes were fixed for topology (see section 6.3 Methods). Grey error bars represent the 95% highest posterior density (HPD) intervals for the divergence estimates. **b**, Distribution of probable and possible extinct pongines known from the fossil record, including *Gigantopithecus* (black bars). Grey bars represent the 95% HPD interval obtained from the Bayesian approach and the 95% confidence interval obtained from the pairwise-distance-based approach, of the *Gigantopithecus*–*Pongo* divergence.

to have a role in amelogenesis [Bartlett et al., 2006]. Our extracts contained no endogenous plasma proteins, such as human serum albumin, or other common dentine proteins, such as collagen type I. We also did not identify any AHSG peptides in our dentine sample. We therefore exclude the possibility that the AHSG peptides derived from dentine. *Gigantopithecus* is known to have relatively long enamel formation times and thick enamel compared to several extant and extinct hominids, including its phylogenetically closest relatives [Dean and Schrenk, 2003, Kono et al., 2014]. We therefore hypothesize that *Gigantopithecus* recruited AHSG as an additional molecular component to favour enamel biomineralization during prolonged amelogenesis, ultimately playing a part comparable to the one it has in bone and dentine mineralization [Bartlett et al., 2006].

6.5 Discussion

Our results reveal the long-debated phylogenetic position of *Gigantopithecus* as an early diverging pongine. We demonstrate that ancient enamel proteomes can be retrieved from Early Pleistocene samples preserved in subtropical conditions, well beyond the current limitations of biomolecular research in hominid and hominin evolution. In addition, the survival of an Early Pleistocene *Gigantopithecus* enamel proteome enables us to assess the presence of multiple forms of in vivo modification. Finally, palaeoproteomic analysis allowed us to identify a hitherto unknown biological component of tooth formation in an extinct hominid. These findings suggest that the palaeoproteomic analysis of hominid enamel has the potential to provide a molecular perspective on human and great ape evolution.

6.6 Author's contribution

Prior to this study, the evolutionary relationship of *Gigantopithecus blacki* to other great ape species had remained unclear. Therefore, we performed a proteomic analysis of its dental enamel molar.

For this study, I performed the MaxQuant analysis of the data provided by the LC-MS/MS experiment. In the analysis, I used the available toolset provided by MaxQuant to identify the peptide sequences and its sequence variations. Additionally, I applied the post-processing pipeline that I previously implemented for protein sequence reconstruction (Chapter 5 - Publication 3). The reconstructed sequences comprise the identified and validated sequence variations, which are necessary for phylogenetic analysis.

Despite the subtropical climate in Southeast Asia, we retrieved protein sequences of a dental enamel from *Gigantopithecus blacki*, which is dated to be the 1.9 million years old. Phylogenetic analysis revealed that *Gigantopithecus blacki* is a sister clade to orangutans with a common ancestor about 12-10 million years ago.

6.7 Additional information

Acknowledgements

E.C. and F.W. are supported by VILLUM FONDEN (17649) and by the European Commission through a Marie Skłodowska-Curie (MSCA) Individual Fellowship (795569). T.M.-B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), NIH grant U01 MH106874, Howard Hughes International Early Career grant, Obra Social 'La Caixa' and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). E.C., J.C., J.V.O., D.S. and P.G. are supported by the Marie Skłodowska-Curie European Training Network (ETN) TEMPERA, a project funded by the European Union's EU Framework Program for Research and Innovation Horizon 2020 under grant agreement no. 722606. M.J.C. and M.M. are supported by the Danish National Research Foundation award PROTEIOS (DNRF128). Work at the Novo Nordisk Foundation Center for Protein Research is funded in part by a donation from the Novo Nordisk Foundation (NNF14CC0001). Research at Chuifeng Cave is made possible by support from the National Natural Science Foundation of China (41572023) and by a grant from the Bagui Scholar of Guangxi. M.K. was supported by a Deutsche Forschungsgemeinschaft (DFG) fellowship (KU 3467/1-1) and the Postdoctoral Junior Leader Fellowship Programme from 'la Caixa' Banking Foundation (LCF/BQ/PR19/11700002). M.E.A. is supported by the Independent Research Fund Denmark (7027-00147B). We thank E. Willerslev for critical reading of the manuscript, scientific support and guidance.

All author's contributions

F.W., E.C., F.D. and T.M.-B. designed the study. W.W. conducted excavation of Chuifeng Cave. W.W. and W.L. carried out faunal analysis of Chuifeng Cave. W.W., W.L. and M.E.A. provided ancient samples. F.W., J.R.-M., M.K., P.G., D.S., M.M., J.V.O. and M.d.M. performed data generation and analysed data with support from A.-M.B., M.J.C., J.C., C.L.-F., F.D. and T.M.-B. F.W., E.C. and T.M.-B. wrote the manuscript with contributions from all authors.

6.8 Extended data figures



Figure 6.4: a, Landscape outside Chuifeng Cave. b, Elevated altitude of Chuifeng Cave (arrow points to the entrance). Photo credit: W.W.

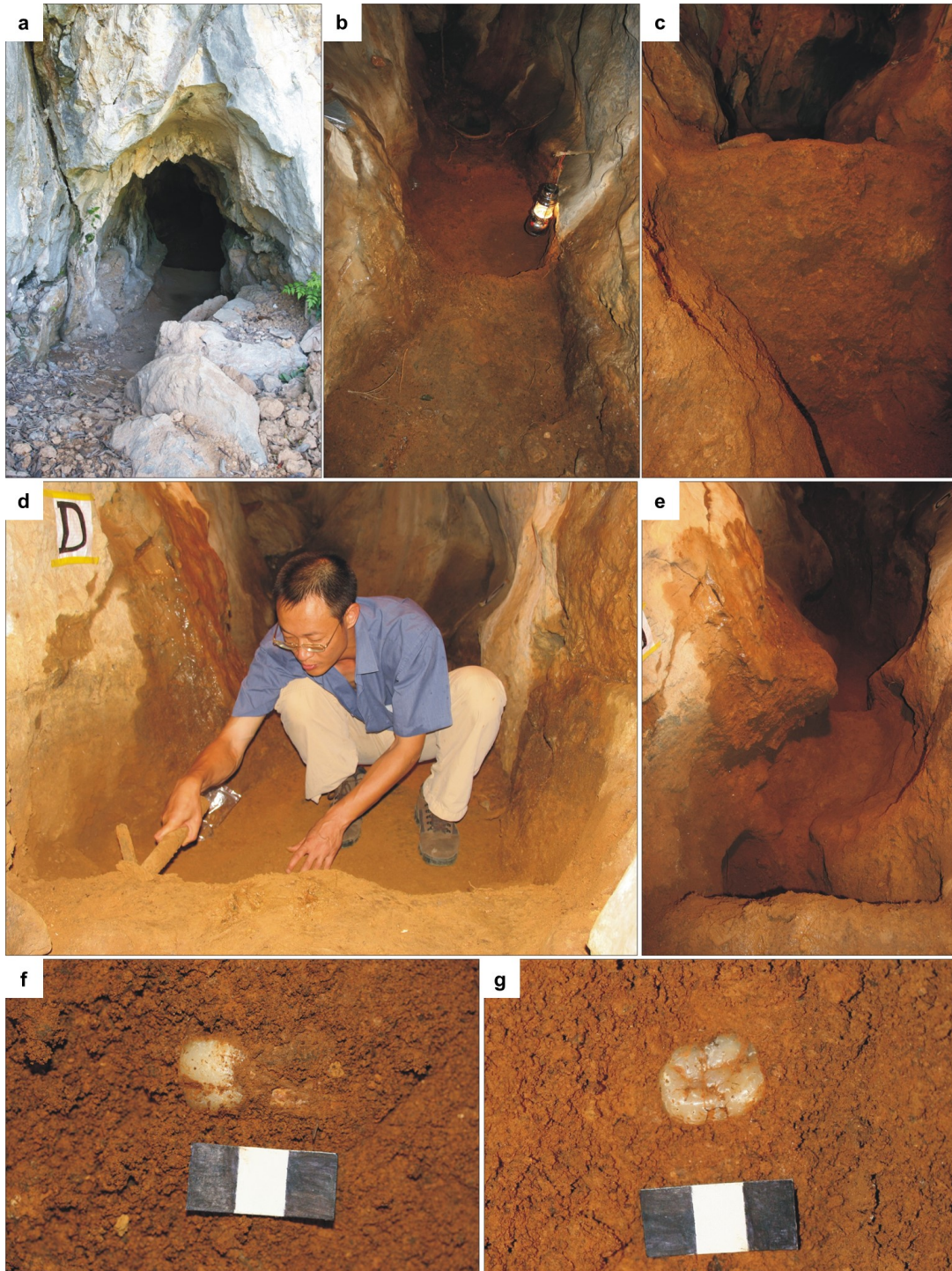


Figure 6.5: **a**, Main entrance of Chuifeng Cave. **b**, Well-preserved deposits before excavation. **c**, The stratigraphic profile (1.3 m in height) of area D. **d**, W.W. excavating in area D. **e**, Excavated channel. **f,g**, In situ *G. blacki* teeth (scale bars, 3 cm). Photo credit: W.W.

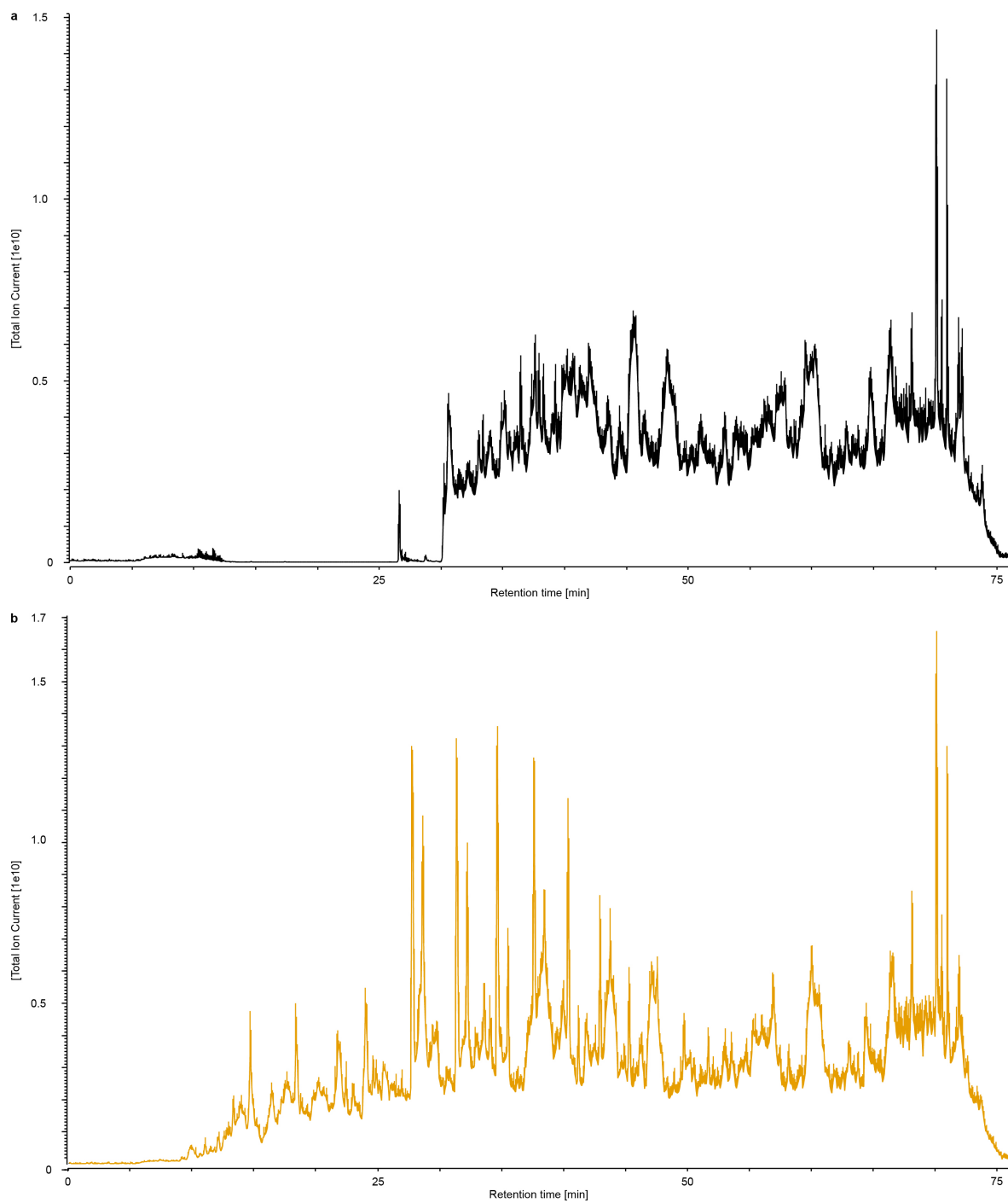


Figure 6.6: **a**, HCl extract. **b**, TFA extract. Note differences in maximum total ion currents on the y axes. Each extract was analysed only once.

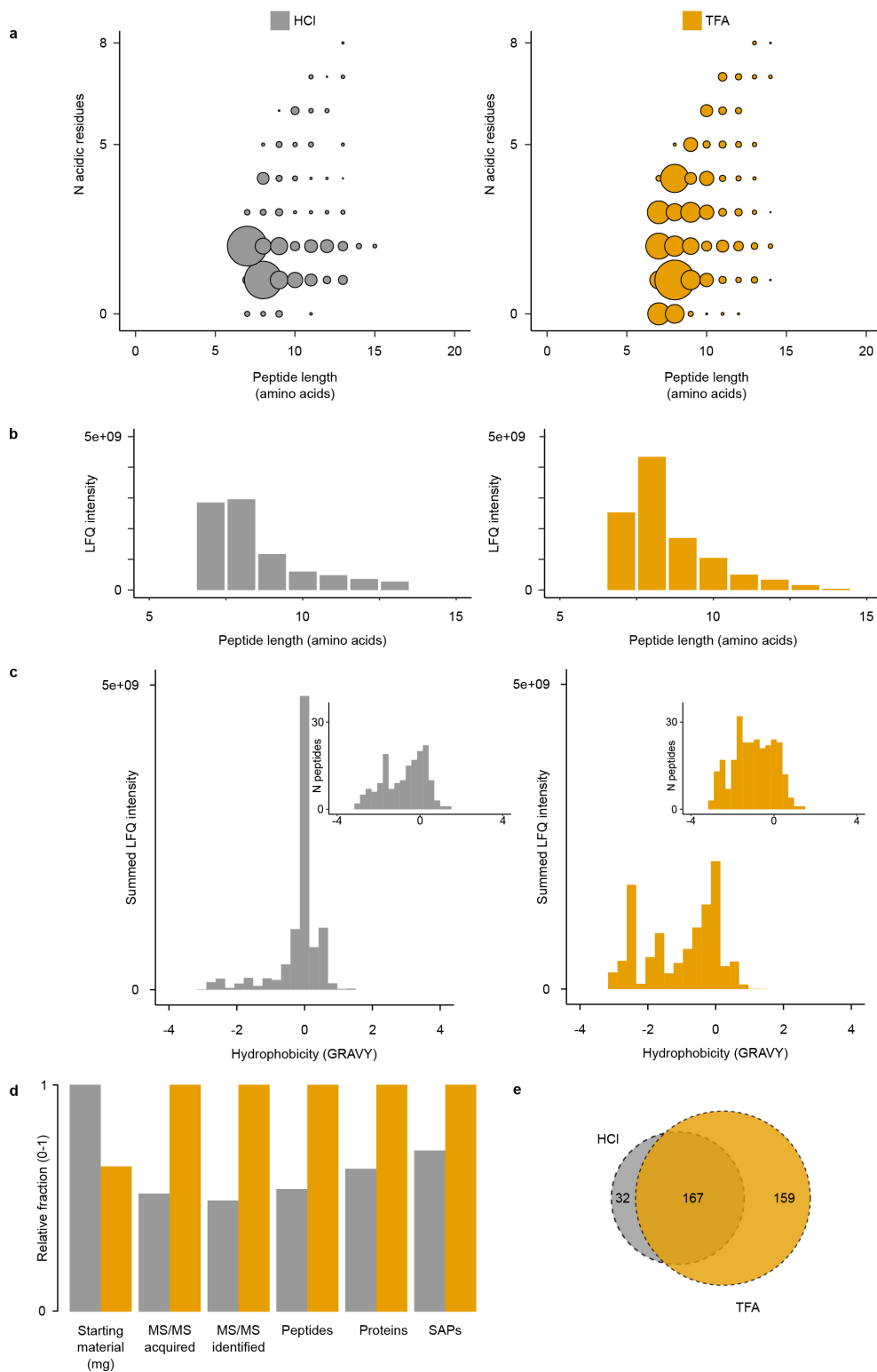


Figure 6.7

Figure 6.7: **a**, Summed and normalized peptide intensities for each combination of peptide length and number of acidic residues (aspartic acid, glutamic acid, deamidated asparagine, deamidated glutamine). Circle sizes are proportional to the percentage of the total intensity, for each combination of peptide length and number of acidic residues. **b**, Summed and normalized intensities by peptide length. **c**, Summed and normalized peptide intensities across peptide hydrophobicity (GRAVY index values calculated using the R package *Peptides*, scale 'KyteDoolittle'). Insets show peptide count distribution across peptide hydrophobicity. **d**, Extraction performance for various data categories. Values scaled to one and compared to the best-performing extraction method for each category independently. SAPs refer to those SAPs informative within Hominoidea. **e**, Proportional Venn diagram of unique peptide sequences identified in the two demineralization methods. All comparisons based on MaxQuant LFQ data only. N, number of peptides.

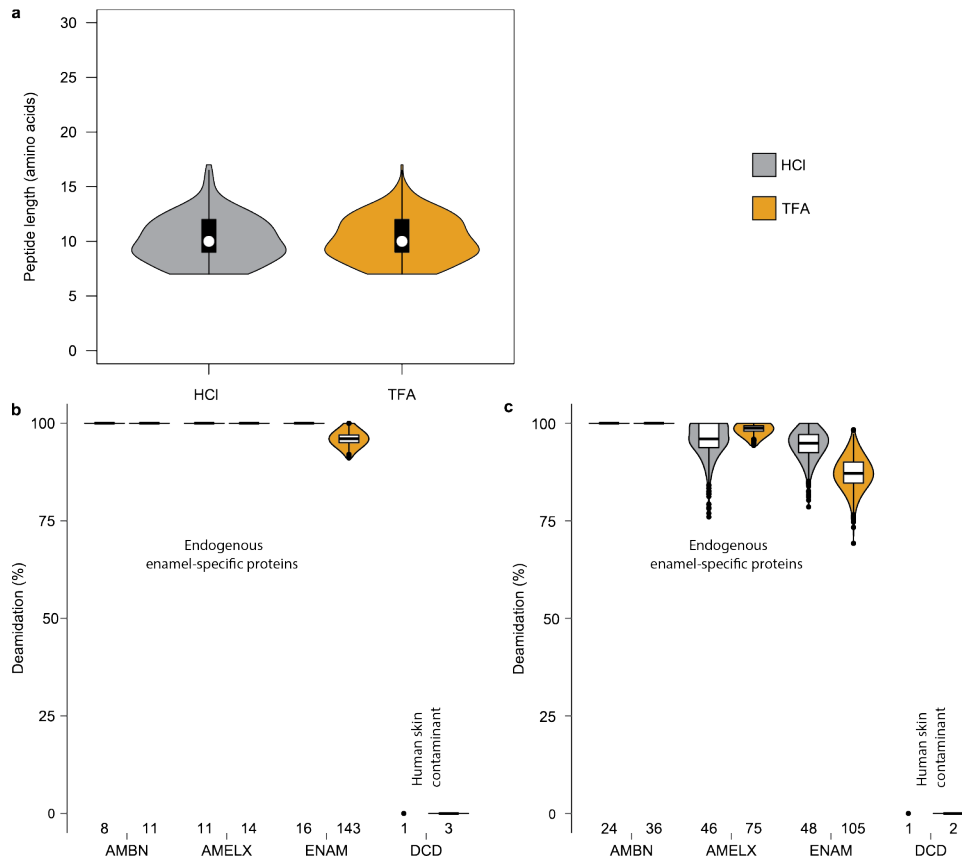


Figure 6.8: **a**, Comparison of peptide lengths, showing an identical distribution for the TFA ($n = 305$) and HCl ($n = 191$) extractions (two-sided t -test, $t_{394} = -0.599$, $P = 0.5495$). **b**, Comparison of asparagine deamidation. **c**, Comparison of glutamine deamidation. **b**, **c**, Violin plots describe the distribution of bootstrap replicates ($n = 1,000$) of intensity-based peptide deamidation³². The number of peptides used for the calculation are shown at the bottom. For some proteins, only deamidated asparagines or glutamines were observed (for example, AMBN), while DCD is included as an example of a non-deamidated contaminant. All comparisons based on MaxQuant data only. For **a–c**, box plots define the range of the data (whiskers extending to $1.5\times$ the interquartile range), outliers (beyond $1.5\times$ the interquartile range), 25^{th} and 75^{th} percentiles (boxes) and medians (dots).

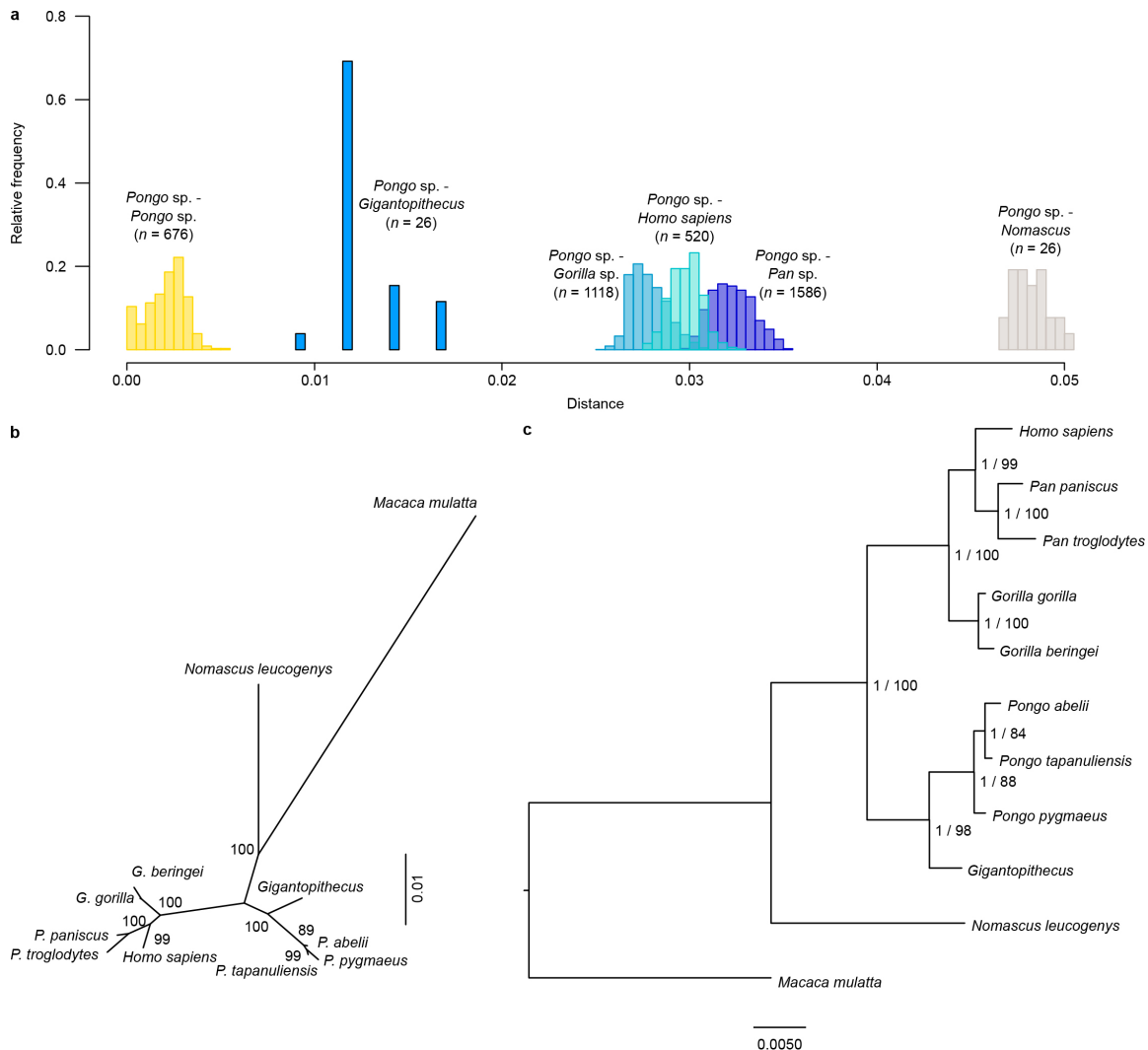


Figure 6.9: **a**, Pairwise distances between groups of selected Hominoids and *Pongo* estimated using the concatenated protein alignments and the phangorn R package. n shows number of pairwise comparisons. **b**, Maximum-likelihood tree computed on a distance matrix using pml R function. Support values were obtained from 1,000 bootstrap replicates. **c**, Rooted phylogenetic tree obtained using MrBayes. For each bipartition, we show the posterior probability (0–1) obtained from the Bayesian approach and the support values obtained from 100 non-parametric bootstrap replicates in a PHyML maximum-likelihood (0–100) tree. PHyML and MrBayes recover the same topology. **b** and **c** are based on the same concatenated alignment of the five proteins retrieved from *Gigantopithecus*, and resulted in the same tree topology.

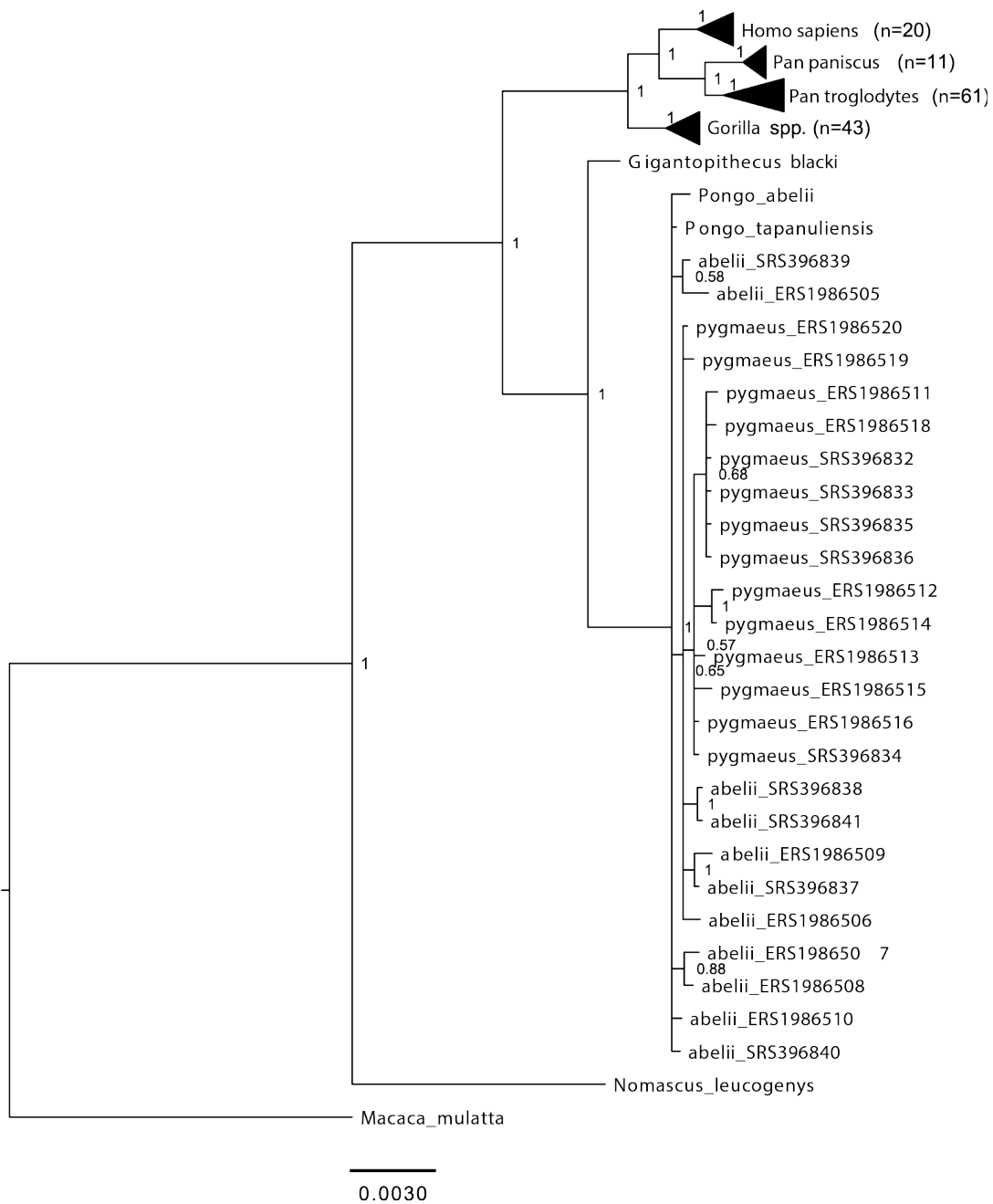


Figure 6.10: Tree obtained from the concatenated alignment using MrBayes. *M. mulatta* was used as an outgroup. The internal nodes corresponding to the *Gorilla*, *Pan* and *Homo* clades are collapsed for visualization purposes. The number of individuals in each of these nodes is indicated in parentheses.

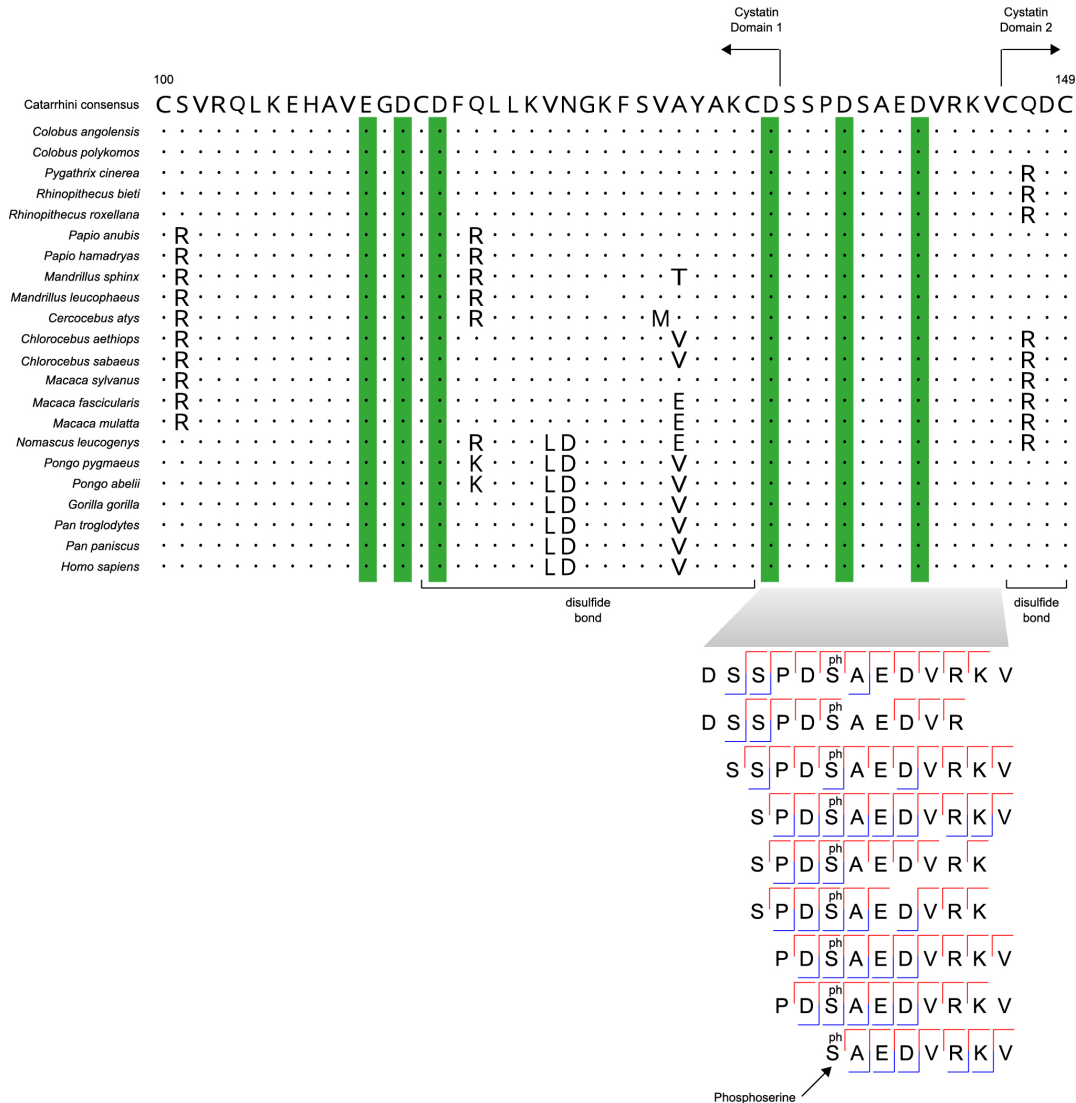


Figure 6.11: All AHSG-specific peptides, identified by PEAKS and MaxQuant, derive from a single sequence region bridging cystatin domains 1 and 2. The surviving sequence region is evolutionarily conserved across Catarrhini. It contains a regular repeat of acidic amino acid residues (aspartic acid, D, on positions 133, 137 and 141) that enable binding of basic calcium phosphate (residues highlighted in green), similarly to a conserved region just N-terminal (glutamic acid, E, on position 111, and aspartic acid, D, on position 113 and 115). At the bottom, a fragment ion alignment is given of MaxQuant-identified AHSG peptides. The serine is phosphorylated in all matching spectra. The majority-based amino acid consensus sequence for the alignment of Catarrhini is shown at the top for amino acid positions 100–149 (amino acid coordinates following UniProt accession P02765 [FETUA_HUMAN]).

Protein	Primary entry	Protein accession	MaxQuant peptides (all unique)	MaxQuant amino acids	PEAKS peptides (all unique)	PEAKS amino acids	Combined sequence coverage (%)
AMELX	H2PUX0_PONAB	H2PUX0	149	135 (4)	270	141 (10)	70.7
AMBN	H2PDI5_PONAB	H2PDI5	55	105 (15)	79	107 (11)	27.5
AMTN	H2PDI4_PONAB	H2PDI4	2	18 (0)	2	18 (0)	8.6
ENAM	H2PDI6_PONAB	H2PDI6	125	129 (5)	189	181 (57)	16.3
MMP20	H2NF32_PONAB	H2NF32	2	9 (0)	1	9 (0)	1.9
AHSG	H2PC98_PONAB	H2PC98	7	13 (0)	12	13 (0)	3.5
ALB	ALBU_Bovin	P02769	2				
DCD	DCD_human	P81605	3		8		
K1C9	K1C9_human	P35527			3		

Table 6.1: Only proteins with two unique peptides in at least either MaxQuant or PEAKS searches were accepted. No protein matches in the dentine or blank extractions fulfilled this criterion. Primary entries for proteins used for phylogenetic reconstruction refer to the *P. abelii* accessions in UniProt for reference purposes. Protein sequence coverage in the final column indicates the coverage obtained after combining PEAKS and MaxQuant peptide recovery and is only reported for proteins considered endogenous. ALB, DCD and K1C9 are considered to be contaminants. For amino acid columns, numbers in brackets refer to amino acid positions uniquely identified in PEAKS or MaxQuant searches.

Protein	Primary entry	Protein accession	MaxQuant peptides (all unique)	MaxQuant amino acids	PEAKS peptides (all unique)	PEAKS amino acids	Combined sequence coverage (%)
AMELX	H2PUX0_PONAB	H2PUX0	149	135 (4)	270	141 (10)	70.7
AMBN	H2PDI5_PONAB	H2PDI5	55	105 (15)	79	107 (11)	27.5
AMTN	H2PDI4_PONAB	H2PDI4	2	18 (0)	2	18 (0)	8.6
ENAM	H2PDI6_PONAB	H2PDI6	125	129 (5)	189	181 (57)	16.3
MMP20	H2NF32_PONAB	H2NF32	2	9 (0)	1	9 (0)	1.9
AHSG	H2PC98_PONAB	H2PC98	7	13 (0)	12	13 (0)	3.5
ALB	ALBU_Bovin	P02769	2				
DCD	DCD_human	P81605	3		8		
B2MG	B2MG_human	P61769	2				
K1C9	K1C9_human	P35527			3		

Table 6.2: The ten geographically closest meteorological weather stations included in publicly available WMO data were used to estimate current MAT at Chuifeng Cave. Correlation between online altitude estimation and WMO provided altitude is $R^2 = 0.99$ (Pearson correlation).

6.9 Supplementary information

Humans from the Simons Genome Diversity Project

The following 19 humans sequenced as part of the Simons Genome Diversity Panel were used to construct the human reference panel used for phylogenetic alignment (“Full alignments”):

Homo_sapiens-Africa_20ERR1419150	Homo_sapiens-Africa_21ERR1419161
Homo_sapiens-Africa_22ERR1419175	Homo_sapiens-Africa_23ERR1419183
Homo_sapiens-Africa_24ERR1347655	Homo_sapiens-Africa_25ERR1419136
Homo_sapiens-Africa_26ERR1419145	Homo_sapiens-Africa_27ERR1419179
Homo_sapiens-Africa_28ERR1419171	Homo_sapiens-WestEurasia_0ERR1395604
Homo_sapiens-WestEurasia_1ERR1419103	Homo_sapiens-WestEurasia_2ERR1419088
Homo_sapiens-WestEurasia_3ERR1419098	Homo_sapiens-WestEurasia_4ERR1419111
Homo_sapiens-WestEurasia_5ERR1419121	Homo_sapiens-WestEurasia_6ERR1419141
Homo_sapiens-WestEurasia_7ERR1419091	Homo_sapiens-WestEurasia_8ERR1419102
Homo_sapiens-WestEurasia_9ERR1419117	

Supplementary Table 6.3. Proteome characteristics of Dm.5/157–16635 (*Stephanorhinus* sp., Dmanisi, Georgia, main text reference 17) and CF-B-16 (*Gigantopithecus blacki*, Chuifeng Cave, China, this study).

	<i>Stephanorhinus</i>	<i>Gigantopithecus</i>
Chronological age (million years ago)	1.77	1.9
Total number of spectra acquired	117,516	38,082
Total number of spectra matched	6,393	1,398
% Spectra matched	5.44	3.67
Number of proteins identified	6	6
Number of peptides identified	987	409
Number of amino acids covered	875	456
Average peptide length	10.89	10.20
Average deamidation rate N*	96.34 ± 1.08	92.85 ± 3.29
Average deamidation rate Q*	96.57 ± 0.57	92.30 ± 2.19

Table 6.3: *mean ± 1 standard deviation obtained after 1,000 bootstrap replicates of MaxQuant proteomic data analysis (endogenous proteins only, see Methods)

Supplementary Table 6.4. Entry names of publicly available protein sequences utilized in database construction and phylogenetic analysis.

Protein	<i>Homo sapiens</i>	<i>Pan paniscus</i>	<i>Pan troglodytes</i>	<i>Gorilla gorilla</i>	<i>Pongo abelii</i>	<i>Nomascus leucogenys</i>	<i>Macaca Mulatta</i>
COL1 α 1	COL1A1 _HUMAN	XP _003817507	H2QDE6 _PANTR	G3RBN8 _GORGO	H2NVM9 _PONAB	XP _012358721	H9Z595 _MACMU
COL1 α 2	CO1A2 _HUMAN	XP _003809763	H2QUY2 _PANTR	G3QT97 _GORGO	H2PMW7 _PONAB	G1RZZ2 _NOMLE	H9Z2D1 _MACMU
ALB	ALBU _HUMAN	XP _003832390	H2RBT1 _PANTR	G3S791 _GORGO	ALBU _PONAB	G1R8T8 _NOMLE	ALBU _MACMU
AMBN	AMBN _HUMAN	XP _003809040	H2R148 _PANTR	G3RCU1 _GORGO	H2PDI5 _PONAB	G1R841 _NOMLE	F7HLX4 _MACMU
AMELY	AMELY _HUMAN	(absent)	AMELY _PANTR	C3UJP7 _9PRIM	(absent)	(absent)	A0A1D5 RDA1 _MACMU
AMELX	AMELX _HUMAN	XP _003805726	A5JJS6 _PANTR	G3SDK0 _GORGO	H2PUX0 _PONAB	G1RCS3 _NOMLE	A5JJS8 _MACMU
ENAM	ENAM _HUMAN	B2L7U5 _PANPA	H2QPM0 _PANTR	B2L7U8 _9PRIM	H2PDI6 _PONAB	G1R843 _NOMLE	F7H832 _MACMU
TUFT1	TUFT1 _HUMAN	XP _003817293	K7CQG4 _PANTR	G3QY68 _GORGO	H2N5V2 _PONAB	G1RGY4 _NOMLE	G7MDK9 _MACMU
KLK4	KLK4 _HUMAN	XP _003813692	XP _009434410	G3QU55 _GORGO	A0A2J8 U913 _PONAB	G1R1C5 _NOMLE	G7NMD1 _MACMU
MMP20	MMP20 _HUMAN	XP _003828430	H2Q4M8 _PANTR	G3QLA8 _GORGO	H2NF32 _PONAB	G1R6B1 _NOMLE	F7GQW6 _MACMU
AMTN	AMTN _HUMAN	XP _003809041	H2QPL9 _PANTR	G3RJV5 _GORGO	H2PDI4 _PONAB	G1R825 _NOMLE	F6VN65 _MACMU
ODAM	ODAM _HUMAN	XP _003809049	ODAM _PANTR	G3QY18 _GORGO	H2PDH6 _PONAB	G1R7Z0 _NOMLE	ODAM _MACMU
AHSG	FETUA _HUMAN	XP _008953975	FETUA _PANTR	E1U7Q5 _9PRIM	H2PC98 _PONAB	G1R4B1 _NOMLE	F6VZ47 _MACMU

Table 6.4: Entry names either refer to UniProt or GenBank.

Supplementary Table 6.5. Description of the protein alignments used in the study.

Protein	Total sites	Non-missing sites	Unique Sites	Single-individual alignment		Full alignment	
				Poly-morphic sites	Polymorphic sites & non singletons	Poly-morphic sites	Polymorphic sites & non singletons
AMBN	447	116	0	64 (7)	19 (3)	74 (11)	28 (5)
AMELX	206	145	2	9 (4)	2 (0)	17 (11)	3 (1)
AMTN	209	18	0	26 (0)	13 (0)	39 (1)	24 (1)
ENAM	1,142	186	1	161 (29)	54 (12)	189 (31)	87 (16)
MMP20	483	9	0	35 (3)	11 (0)	41 (3)	20 (0)

Table 6.5: Single individual alignment: alignment built using a single individual per species. This alignment was used to estimate divergence times for each node. **Full alignment:** alignment built using all individuals in the reference dataset. 1: Private mutations/polymorphism found in the ancient sample. 2: The number of sites where the ancient sample is non-missing are indicated in brackets.

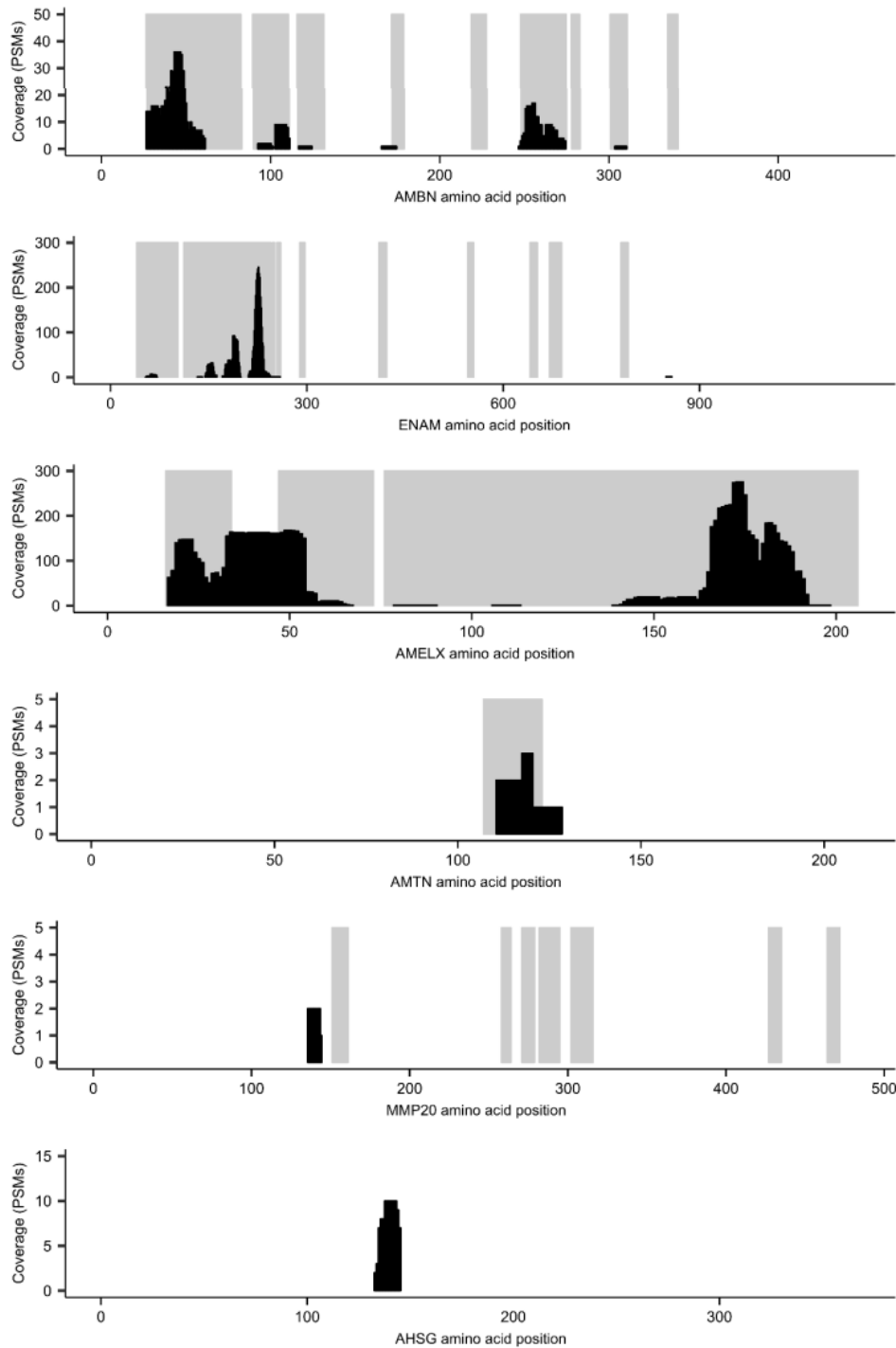


Figure 6.12: MaxQuant PSM coverage plot for the six endogenous enamel proteins recovered from the *Gigantopithecus* enamel. Grey background shading indicates regions with peptide sequence coverage in a previously published Early Pleistocene enamel proteome from Dmanisi (*Stephanorhinus sp.*; 19). Although differences are expected, due to differences in primary sequence, recovered peptide sequences from *Gigantopithecus* generally fall within sequence regions identified in the thermally younger *Stephanorhinus sp.* specimen from Dmanisi.

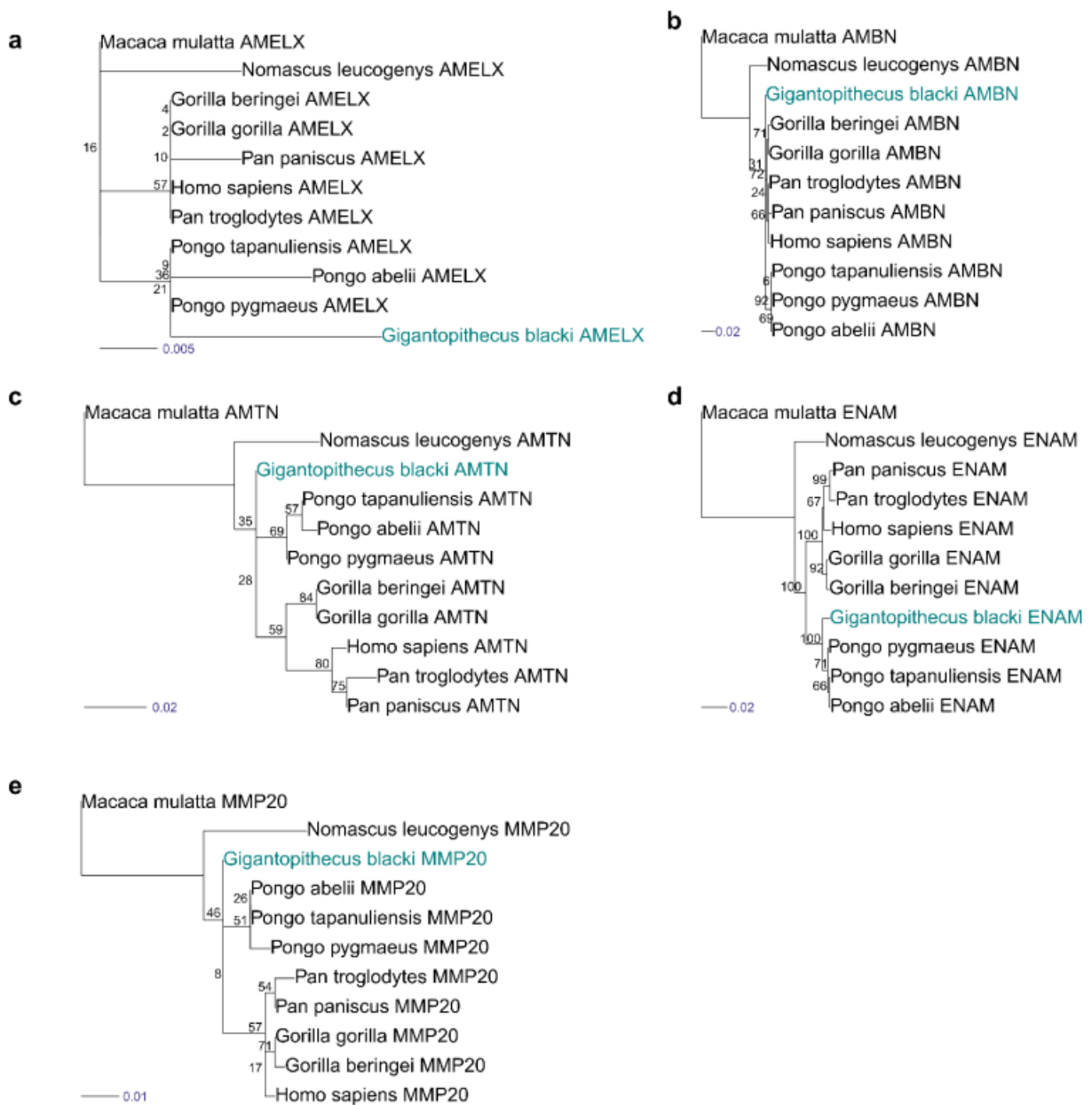


Figure 6.13: Supplementary Figure S2. PHyML individual protein trees for the five endogenous proteins recovered from the *Gigantopithecus blacki* enamel proteome (AHSG excluded). **a**, AMELX. **b**, AMBN. **c**, AMTN. **d**, ENAM. **e**, MMP20. Nodal support values are after 100 bootstraps.

Note on Supplementary Figures 6.14 to 6.23

Supplementary figures 6.14 to 6.23 contain MaxQuant-derived examples of annotated MS/MS spectra. Fragment ions are labelled in the “Standard” annotation mode provided in MaxQuant. For each panel, both the annotated MS/MS spectrum and an amino acid sequence, covered by the relevant b- and y-ions, are shown. For the latter, additional PTM annotations on relevant amino acid positions include “ph” (phosphorylation), “de” (deamidation), and “ox” (oxidation). In some of these cases, additional PSMs were also identified by MaxQuant (Supplementary Figure 6.12) and PEAKS.

Supplementary figures 6.14 to 6.19 include unique spectra matching to peptide sequences representing one of the six endogenous proteins identified in the *Gigantopithecus* proteome. Supplementary Figures 6.20 to 6.23 include several examples of MS/MS spectra overlapping SAPs of phylogenetic interest.

All the mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the data set identifier PXD013838.

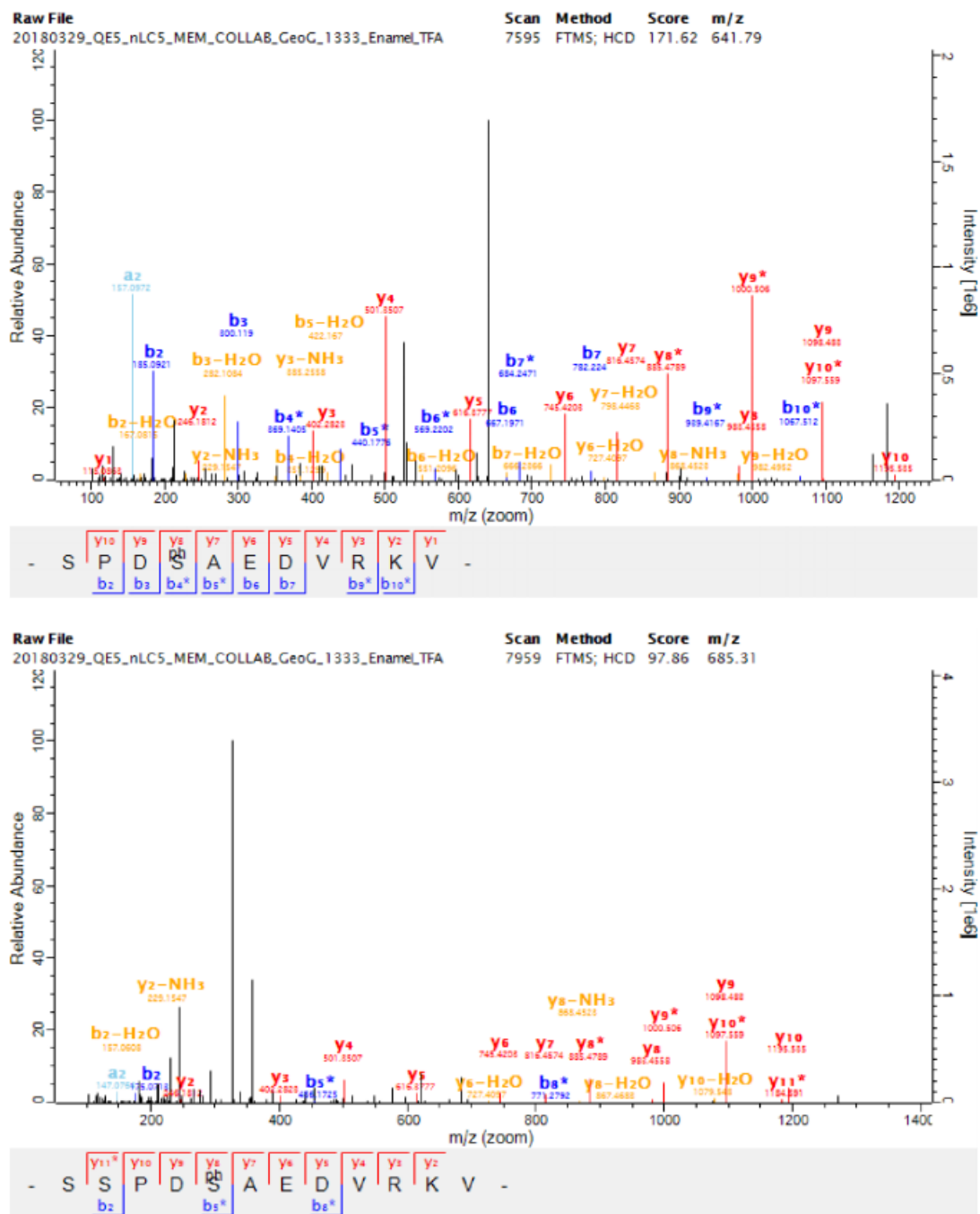


Figure 6.14: Examples of annotated MS/MS spectra unique to AHSG. Top, amino acid positions 135-145. Bottom, amino acid positions 134-145. Coordinates refer to UniProt entry H2PC98_PONAB.

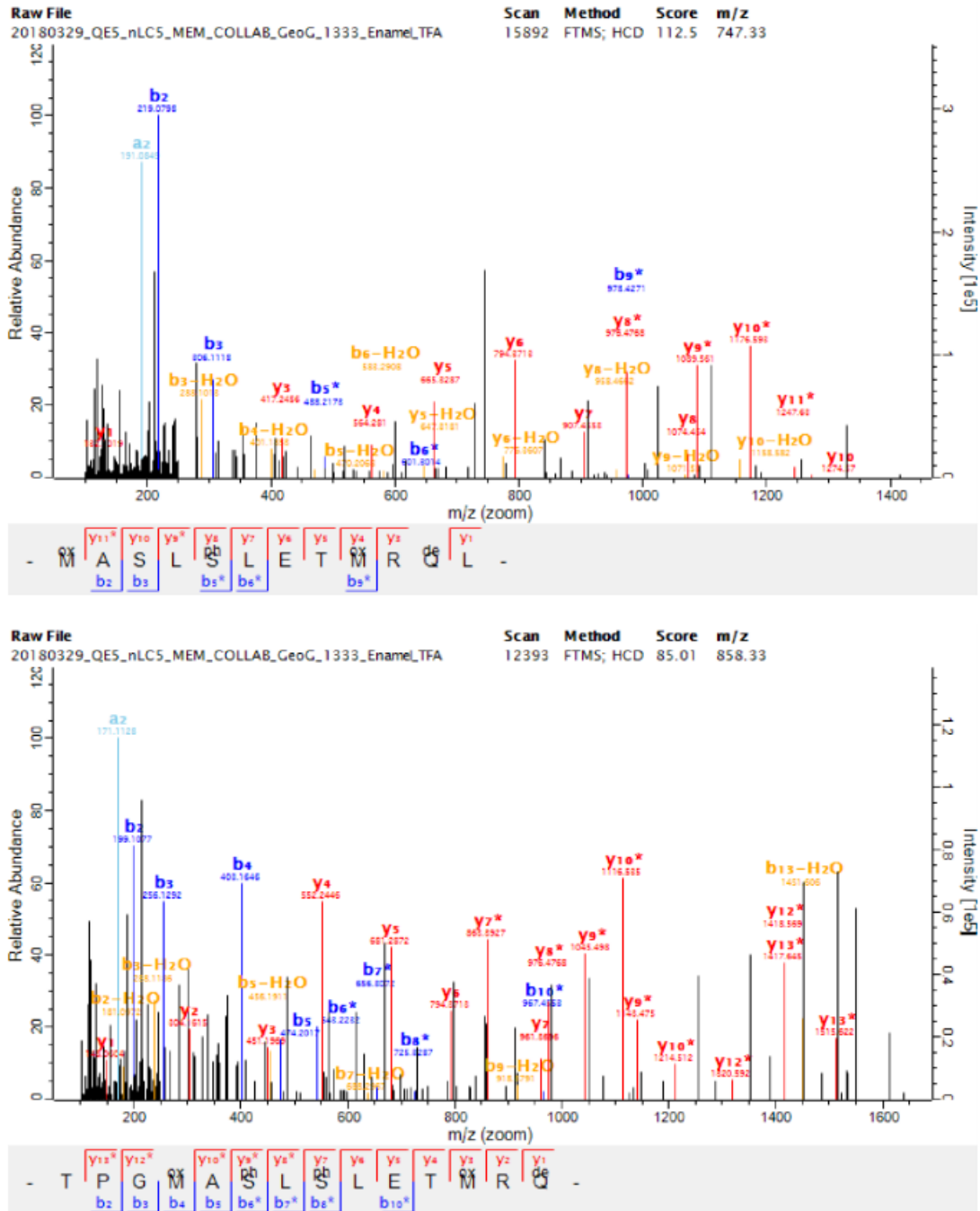


Figure 6.15: Examples of annotated MS/MS spectra unique to AMBN. Top, amino acid positions 39-50. Bottom, amino acid positions 36-49. Coordinates refer to UniProt entry H2PDI5_PONAB.

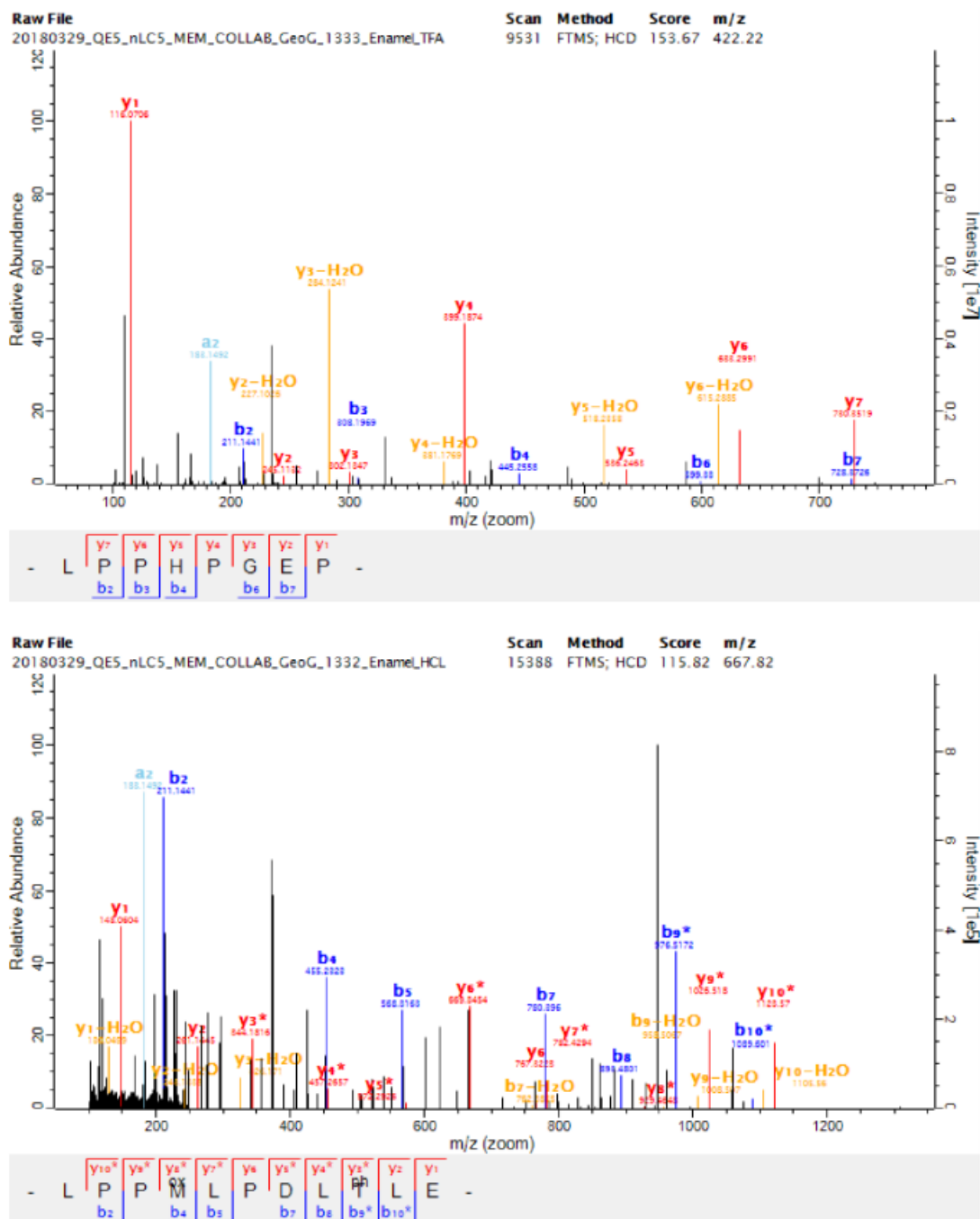


Figure 6.16: Examples of annotated MS/MS spectra unique to AMELX. Top, amino acid positions 20-27. Bottom, amino acid positions 182-192. Coordinates refer to UniProt entry H2PUX0_PONAB.

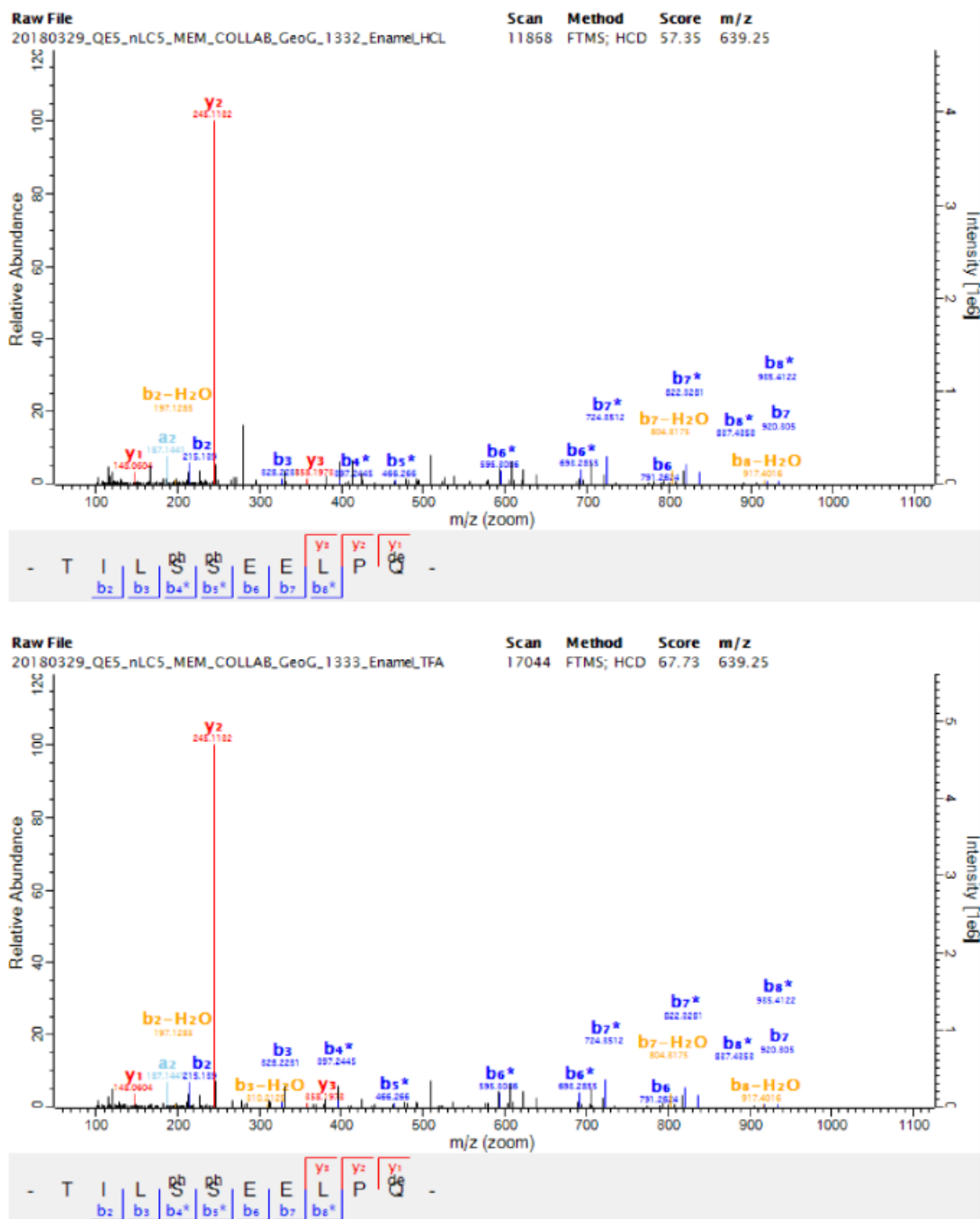


Figure 6.17: Examples of annotated MS/MS spectra unique to AMTN. Top, amino acid positions 112-121. Bottom, amino acid positions 112-121. Coordinates refer to UniProt entry H2PDI4_PONAB.

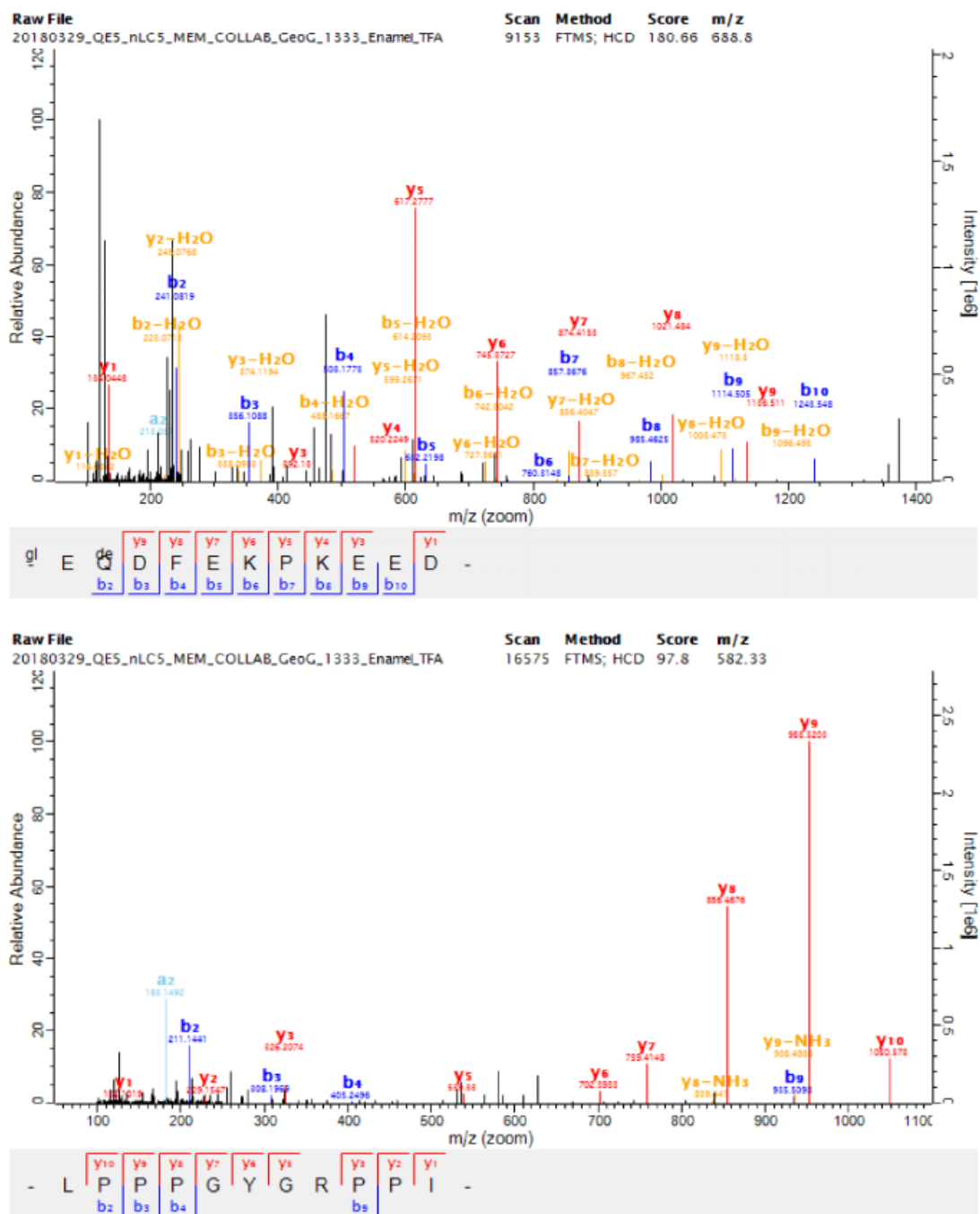


Figure 6.18: Examples of annotated MS/MS spectra unique to ENAM. Top, amino acid positions 221-231. Bottom, amino acid positions 180-190. Coordinates refer to UniProt entry H2PDI6_PONAB.

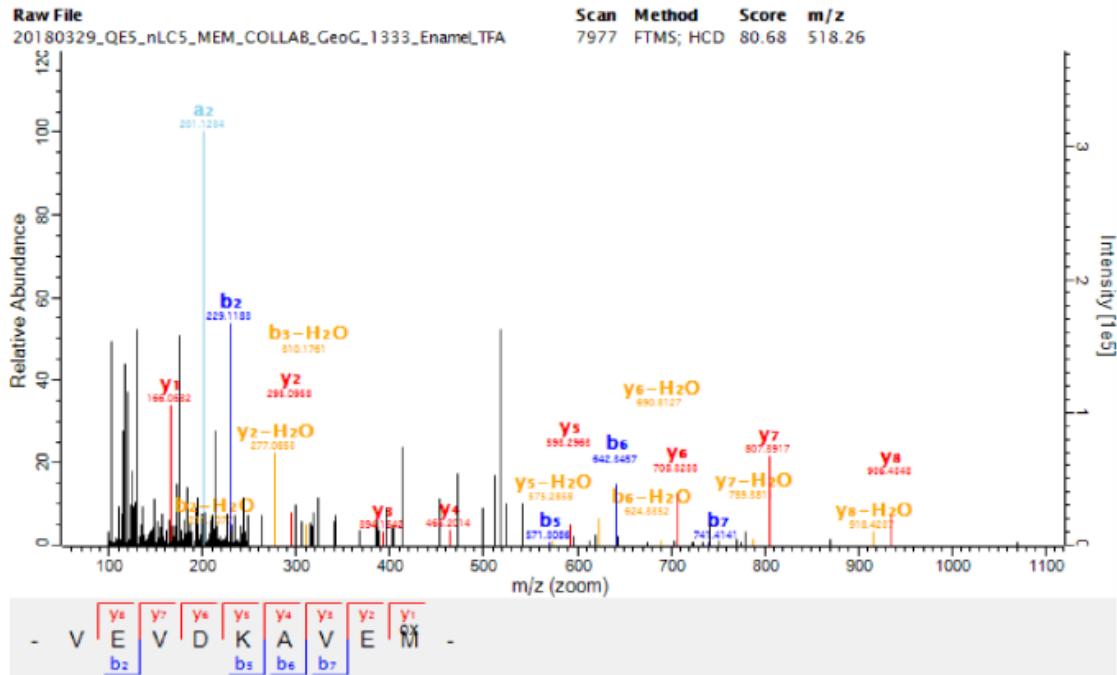


Figure 6.19: Example of an annotated MS/MS spectrum unique to MMP20. Amino acid positions 136-144. Coordinates refer to UniProt entry H2NF32_PONAB.

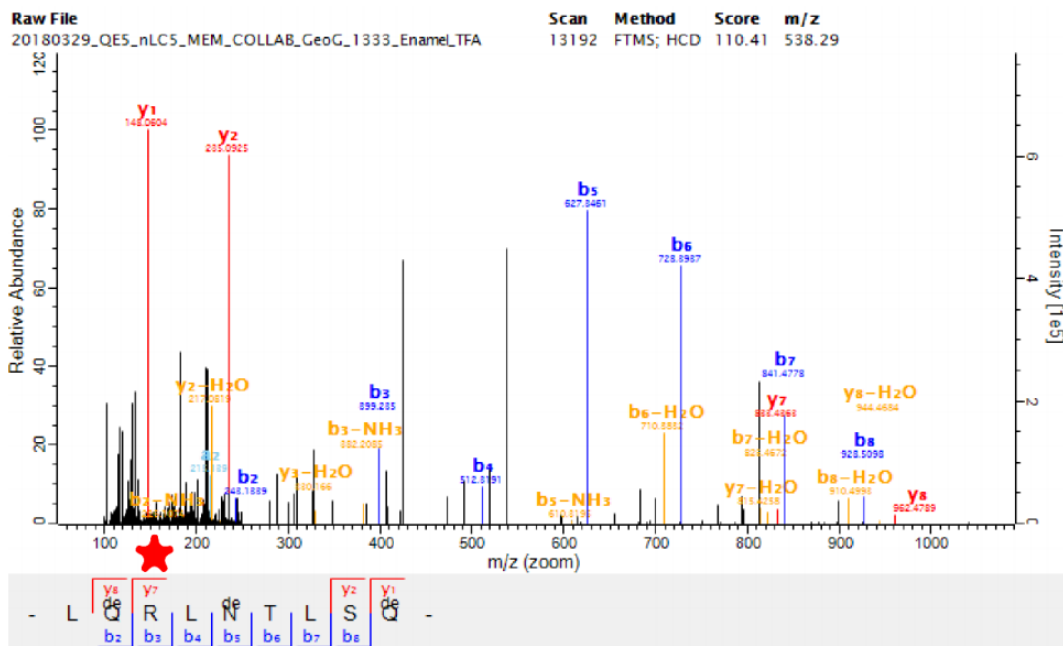


Figure 6.20: Example of an annotated MS/MS spectrum of a phylogenetically informative SAP in AMBN. Amino acid positions 53-61, with a Hominidae-derived “R” on amino acid position 55. Coordinates refer to UniProt entry H2PDI5_PONAB.

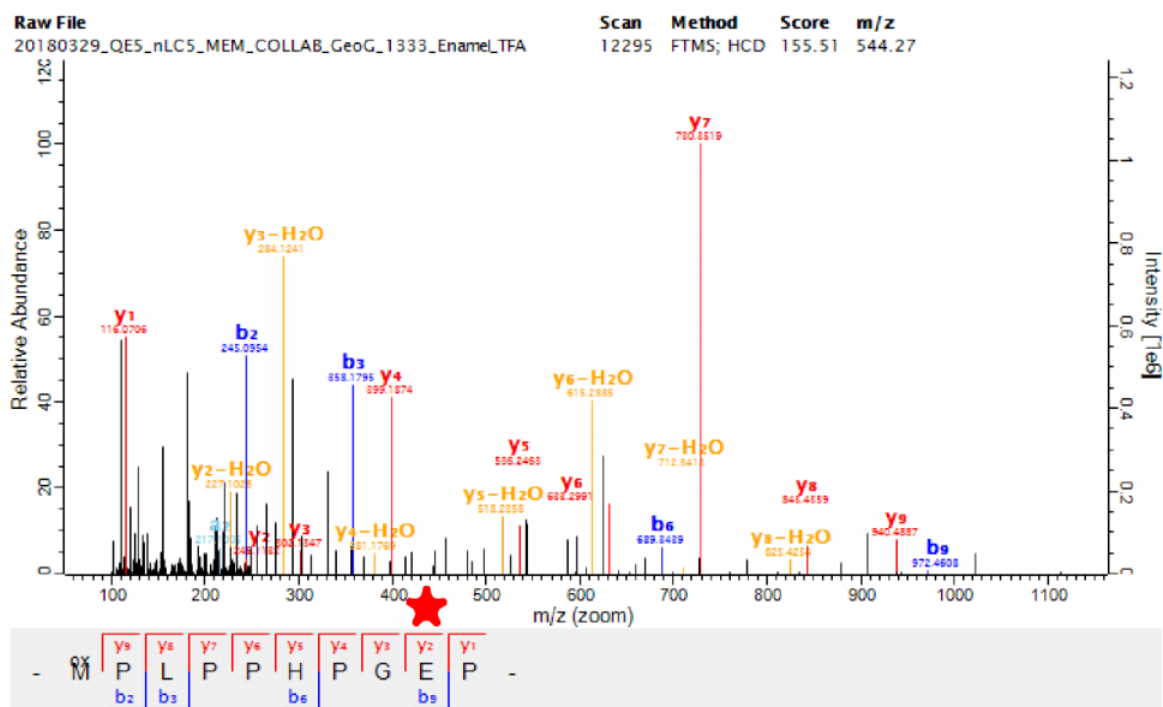
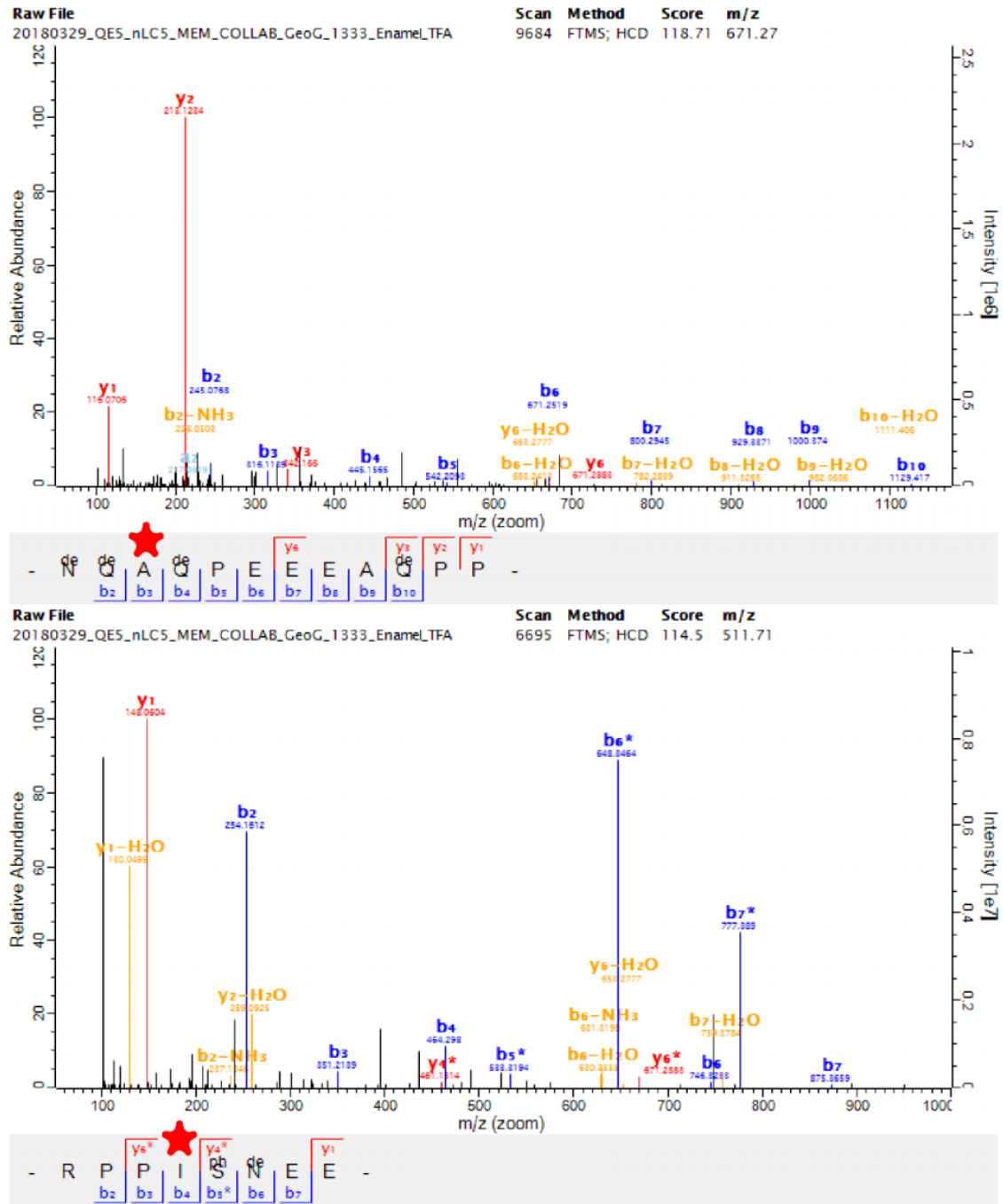


Figure 6.21: Examples of annotated MS/MS spectra of phylogenetically informative SAPs in AMELX. Amino acid positions 18-27, with a *Gigantopithecus*-derived “E” on amino acid position 26. Coordinates refer to UniProt entry H2PUX0_PONAB.



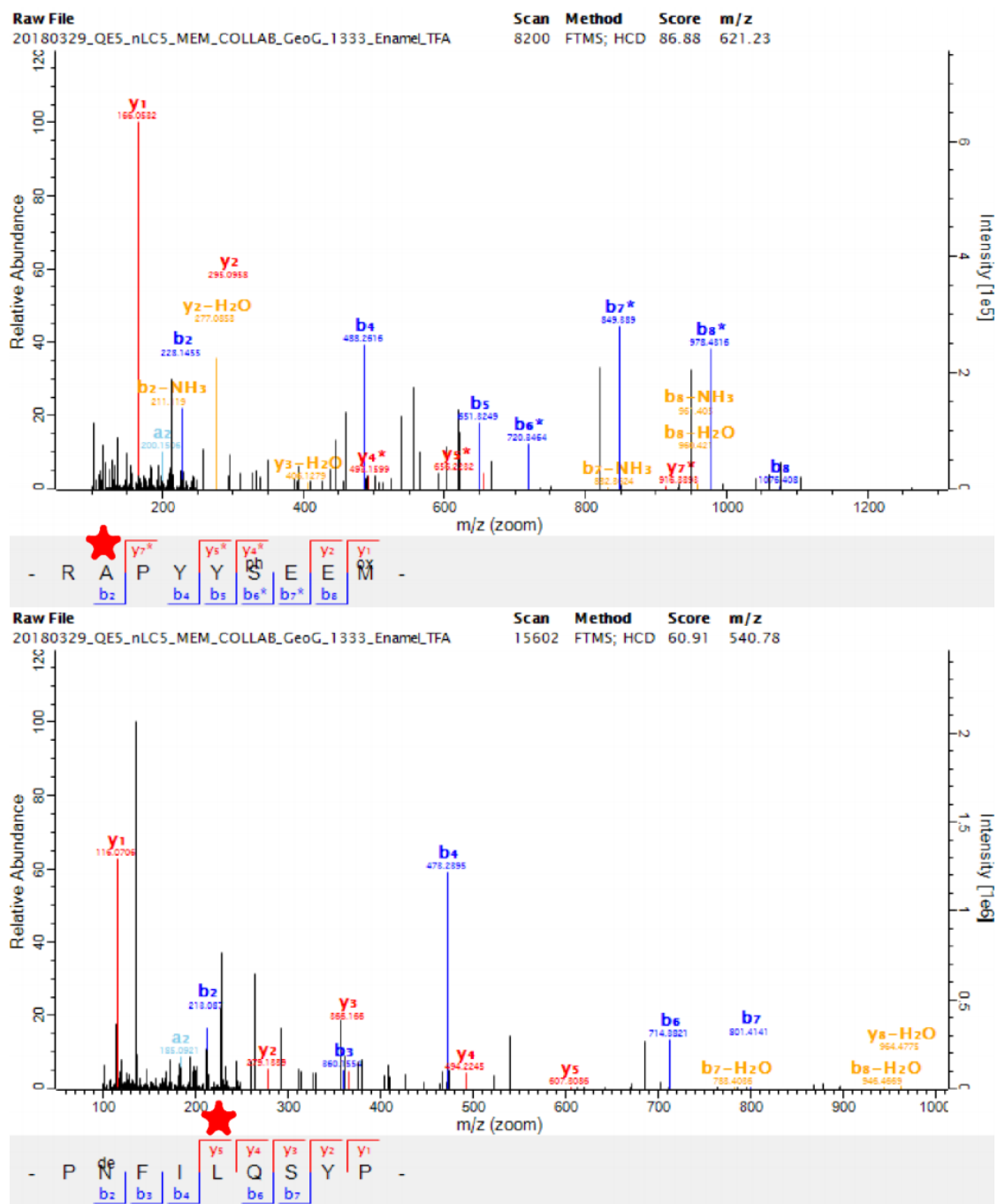


Figure 6.23: Examples of annotated MS/MS spectra of phylogenetically informative SAPs in ENAM. Top, amino acid positions 211-219, with a *Pongo+Gigantopithecus* derived “A” on amino acid position 212 (“A” in other Hominoids). Bottom, amino acid positions 849-857, with a *Gigantopithecus*-derived “L” on amino acid position 853. Coordinates refer to UniProt entry H2PDI6.PONAB.

Chapter 7

Conclusion and outlook

Ancient biomolecules, such as aDNA and ancient proteins are essential resources to unravel the past. Due to the lack of availability of morphological features, these biomolecules are the sole option for species identification and phylogenetic placement of highly fragmented fossil remains.

aDNA is known to degrade faster than ancient proteins, which makes ancient proteins vital when investigating fossil remains that date back more than a million years or when deposition conditions are unfavorable for the preservation of DNA. Ancient protein studies have highly contributed to deepen our knowledge in the fields of cultural heritage, archaeology and palaeontology by studying surviving proteins of organic materials or harbouring on the surface of archaeological objects (Section 1.7 Application of ancient proteins). PMF and LC-MS/MS are the most used MS approaches for ancient peptide and protein identification. PMF is mainly performed for species identification by comparing the measured peptide mass pattern of collagenous protein sequences to known peptide mass patterns of other species. Species identification of faunal assemblages have provided insights into past subsistence practices and ecologies. Unlike PMF, LC-MS/MS approaches generate a detailed peptide fragmentation pattern, which are indicative of the amino acid sequences and are commonly identified by database search engines.

Typically, ancient proteins are obtained from unsequenced or extinct species, which are evolutionarily distant to their extant relatives, thus, their protein sequences comprise unknown sequence variations. Database search engines, however, are only capable of identifying sequences that exactly match to a reference sequence database. Consequently, identification by sequence database search is an obstacle for ancient protein studies causing many peptides to remain unidentified (Section 1.6.2 Sequence database search, 1.8 Challenges of ancient protein sequence identification). De-novo sequencing has the potential to overcome this challenge by directly inferring the sequence based on the peptides fragmentation pattern. Yet, de-novo sequencing is hampered by the generally poor fragmentation quality due to the low abundant and highly fragmented peptides caused by advanced ancient protein degradation (Section 1.6.3 De-novo sequencing). Furthermore, proteomics analysis requires destructive sampling of ancient objects, which are primarily very precious and most importantly limited. Thus, ancient protein analysis is only performed when there is a guaranteed successful outcome to avoid an irrecoverable information loss.

To this end, the aim of this thesis was to computationally increase the identified peptide rate on two levels: sequence database searching (Chapter 3 – Publication 1) and de-novo sequencing (Chapter 4 – Publication 2). To improve the identification rate of database searching, Andromeda was extended to include predicted fragmentation intensities. By adding the intensity prediction, the sequence context is not only included in form of fragment mass but also in form of fragmentation efficiencies. We successfully demonstrated an increase in peptide identifications by using the intensity informed Andromeda score in comparison to the conventional Andromeda score (Chapter 3 - Publication 1). To provide highly accurate peptide fragment intensities, we developed in collaboration with the research organisation Verily Life Sciences two machine learning models: DeepMass:Prism and wiNNer. DeepMass:Prism is based on deep learning and was trained on large datasets covering a variety of different species and MS settings, therefore it can be used as standard prediction model for tryptic datasets obtained by LC-MS/MS experiments. In contrast, wiNNer is based on neural networks and was trained on a smaller dataset allowing fast re-training. The possibility to re-train prediction models fast and easily with standard computing power is crucial for ancient datasets. These datasets can be different to common datasets in terms of enzymatic digestion, instrumental setup or because of non-standard protein characteristics. We demonstrated that wiNNer can be re-trained based on ancient datasets that are characterized by non-tryptic digestion, short and highly modified peptides (Chapter 5 - Publication 3). Despite the limited available training data, wiNNer provides high quality predictions for ancient proteomics data.

The current status of the extended Andromeda score is a proof-of-principle and requires further development to make it accessible for public usage for which the following steps need to be developed: 1) integration of the DeepMass:Prism and wiNNer model directly in the MaxQuant workflow to enable real-time prediction during the identification step; 2) extension of the current intensity prediction models to additionally allow prediction of modified peptides, as currently only prediction of unmodified peptides is possible.

Next, with the development of MaxNovo, integrated in the MaxQuant software package, we provide a novel spectrum graph-based de-novo sequencing algorithm that achieves high peptide identification in collisional dissociated MS/MS spectra (Chapter 4 - Publication 2). MaxNovo is implemented using expert domain knowledge, whereas state-of-the-art de-novo algorithms are based on deep learning algorithms. While deep learning models are often referred to as “black box”, the development of the MaxNovo algorithm is transparent and can be fully explained. Furthermore, MaxNovo performs as well as or better than leading deep learning-based algorithms, which rely on sufficient training data and might require re-training to provide reliable performance for uncommon datasets. Lastly, we have shown that MaxNovo is able to provide additional peptide identifications in ancient datasets previously unidentified by the database search engine Andromeda.

Additionally, as part of this thesis, I contributed to two evolutionary ancient proteomics studies, where we provided evidence that *H. antecessor* is a close sister lineage to subsequent Middle and Late Pleistocene hominins, including modern humans, Neanderthals and Denisovans (Chapter 5 – Publication 3). With proteomic analysis of dental enamel from a 1.9 million year old molar, we revealed that *Gigantopithecus blacki* is a sister clade to orangutans with a common ancestor about 12-10 million years ago (Chapter 6 – Publication 4).

Despite the challenges due to the characteristics of ancient protein, previous ancient studies as well as the two ancient studies included in this thesis demonstrate the success of retrieving ancient protein sequences of fossils dated to Late and Middle Pleistocene. Furthermore, based on the ancient analysis of the molar of *G. blacki* we can show that the limits of ancient protein studies have not yet been reached. Despite archaeological deposition in subtropical climate, we were able to retrieve an enamel proteome in sufficient quality to allow phylogenetic placement from an Early Pleistocene sample, which is currently the oldest Cenozoic skeletal proteome reported [Welker et al., 2019]. The successful palaeoproteomic analysis on a sample older than previous ones and preserved in subtropical conditions enhances the application range of palaeoproteomic analysis on samples from climatically problematic areas and earlier time periods. This extends the scope of samples that hold the potential to gain new insights into great ape and human evolution.

For both ancient studies that are part of this thesis, I developed a post-processing computational workflow based on the MaxQuant outputs to reconstruct ancient protein sequences to determine the amino acid sequence variations. For the reconstruction, all identified peptide sequences of different identification approaches are aligned to a reference protein sequence that is representative for a gene, but independent of the species information, since the investigated species is mainly of unknown origin. To date, there is no software tool available that specifically addresses the challenges of ancient proteins. Therefore, a possible future development would be to transform this ancient protein sequence reconstruction workflow into a novel standalone software tool. Moreover, this software can be extended by the integration of the intensity-informed Andromeda search engine including a dedicated ancient wiNNer model for intensity prediction and MaxNovo to enable the identification of novel peptides containing sequence variations. Such a software would provide a complete package for ancient protein sequence identification and reconstruction.

The current scope of this thesis was focused on the identification of ancient protein and peptide sequences. However, the deamidation rate of peptides is crucial to validate the age of the fossils because the deamidation rate is expected to increase with the age of the sample. Therefore, the possibility to not only identify ancient proteins but also quantify their deamidation rate would offer a complete package that covers all important aspects for ancient protein studies. Thus, a comprehensive software tool covering the aforementioned functionalities would be highly beneficial for the ancient proteomics community.

Bibliography

- [Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., and Brain, G. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, pages 265–284.
- [Abelson P.H., 1954] Abelson P.H. (1954). Amino acids in fossils. *Science*, 119(3096):576–576.
- [Aebersold and Mann, 2003] Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *422(6928):198–207*.
- [Allentoft et al., 2012] Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., and Bunce, M. (2012). The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748):4724–4733.
- [Allmer, 2011] Allmer, J. (2011). Algorithms for the de novo sequencing of peptides from tandem mass spectra.
- [Altschul et al., 2002] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (2002). Papers to appear in Ecotoxicology and Environmental Safety Environmental Research, Section B. *Environmental Research*, 90(1):67–68.
- [Álvarez-Posada et al., 2018] Álvarez-Posada, C., Parés, J. M., Cuenca-Bescós, G., Van der Made, J., Rosell, J., Bermúdez de Castro, J. M., and Carbonell, E. (2018). A post-Jaramillo age for the artefact-bearing layer TD4 (Gran Dolina, Atapuerca): New paleomagnetic evidence. *Quaternary Geochronology*, 45:1–8.
- [Andrews et al., 1985] Andrews, J. T., Miller, G. H., Davies, D. C., and Davies, K. H. (1985). Generic identification of fragmentary Quaternary molluscs by amino acid chromatography: A tool for Quaternary and palaeontological research. *Geological Journal*, 20(1):1–20.
- [Arenas, 2015] Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 0(OCT):319.
- [Arnold and Demuro, 2015] Arnold, L. J. and Demuro, M. (2015). Insights into TT-OSL signal stability from single-grain analyses of known-age deposits at Atapuerca, Spain. *Quaternary Geochronology*, 30:472–478.

- [Arnold et al., 2015] Arnold, L. J., Demuro, M., Parés, J. M., Pérez-González, A., Arsuaga, J. L., Bermúdez de Castro, J. M., and Carbonell, E. (2015). Evaluating the suitability of extended-range luminescence dating techniques over early and Middle Pleistocene timescales: Published datasets and case studies from Atapuerca, Spain. *Quaternary International*, 389:167–190.
- [Arnold et al., 2006] Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H., and Radivojac, P. (2006). A machine learning approach to predicting peptide fragmentation spectra. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 230:219–230.
- [Asaka et al., 2009] Asaka, T., Akiyama, M., Domon, T., Nishie, W., Natsuga, K., Fujita, Y., Abe, R., Kitagawa, Y., and Shimizu, H. (2009). Type XVII Collagen is a Key Player in Tooth Enamel Formation. *The American Journal of Pathology*, 174(1):91–100.
- [Bada et al., 1973] Bada, J. L., Kvenvolden, K. A., and Peterson, E. (1973). Racemization of amino acids in bones. *Nature*, 245(5424):308–310.
- [Bandeira et al., 2008a] Bandeira, N., Ng, J., Meluzzi, D., Lington, R. G., Dorrestein, P., and Pevzner, P. A. (2008a). De novo sequencing of nonribosomal peptides. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4955 LNBI, pages 181–195. Springer, Berlin, Heidelberg.
- [Bandeira et al., 2008b] Bandeira, N., Pham, V., Pevzner, P., Arnott, D., and Lill, J. R. (2008b). Automated de novo protein sequencing of monoclonal antibodies.
- [Banerjee and Mazumdar, 2012] Banerjee, S. and Mazumdar, S. (2012). Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte. *International Journal of Analytical Chemistry*, 2012:1–40.
- [Barker, 1981] Barker, C. (1981). Biogeochemistry of amino acids. *Geochimica et Cosmochimica Acta*, 45(10):1965–1966.
- [Bartlett et al., 2006] Bartlett, J. D., Ganss, B., Goldberg, M., Moradian-Oldak, J., Paine, M. L., Snead, M. L., Wen, X., White, S. N., and Zhou, Y. L. (2006). Protein-Protein Interactions of the Developing Enamel Matrix. *Current Topics in Developmental Biology*, 74:57–115.
- [Bateman et al., 2017] Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley, A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cucho, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Noupikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire,

- C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169.
- [Begun, 2007] Begun, D. R. (2007). How to identify (as opposed to define) a homoplasy: Examples from fossil and living great apes. *Journal of Human Evolution*, 52(5):559–572.
- [Behjati and Tarpey, 2013] Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, 98(6):236–238.
- [Berger et al., 2008] Berger, G. W., Pérez-González, A., Carbonell, E., Arsuaga, J. L., Bermúdez de Castro, J. M., and Ku, T. L. (2008). Luminescence chronology of cave sediments at the Atapuerca paleoanthropological site, Spain. *Journal of Human Evolution*, 55(2):300–311.
- [Bermúdez de Castro et al., 1997] Bermúdez de Castro, J. M., Arsuaga, J. L., Carbonell, E., Rosas, A., Martínez, I., and Mosquera, M. (1997). A Hominid from the Lower Pleistocene of Atapuerca, Spain: Possible Ancestor to Neandertals and Modern Humans. *Science*, 276(5317):1392–1395.
- [Bermúdez de Castro et al., 1999] Bermúdez de Castro, J. M., Carbonell, E., Cáceres, I., Díez, J. C., Fernández-Jalvo, Y., Mosquera, M., Ollé, A., Rodríguez, J., Rodríguez, X. P., Rosas, A., Rosell, J., Sala, R., Vergés, J. M., and van der Made, J. (1999). The TD6 (Aurora stratum) hominid site. Final remarks and new questions. *Journal of Human Evolution*, 37(3):695–700.
- [Bermúdez de Castro et al., 2012] Bermúdez de Castro, J. M., Carretero, J. M., García-González, R., Rodríguez-García, L., Martín-Torres, M., Rosell, J., Blasco, R., Martín-Francés, L., Modesto, M., and Carbonell, E. (2012). Early pleistocene human humeri from the gran dolina-TD6 site (sierra de atapuerca, spain). *American Journal of Physical Anthropology*, 147(4):604–617.
- [Bermúdez de Castro et al., 2017] Bermúdez de Castro, J. M., Martín-Torres, M., Arsuaga, J. L., and Carbonell, E. (2017). Twentieth anniversary of Homo antecessor (1997-2017): a review. *Evolutionary Anthropology: Issues, News, and Reviews*, 26(4):157–171.
- [Besenbacher et al., 2019] Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., and Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology and Evolution*, 3(2):286–292.
- [Blain et al., 2014] Blain, H. A., Agustí, J., Lordkipanidze, D., Rook, L., and Delfino, M. (2014). Paleoclimatic and paleoenvironmental context of the Early Pleistocene hominins from Dmanisi (Georgia, Lesser Caucasus) inferred from the herpetofaunal assemblage. *Quaternary Science Reviews*, 105:136–150.
- [Bludau and Aebersold, 2020] Bludau, I. and Aebersold, R. (2020). Proteomic and interactomic insights into the molecular basis of cell functional diversity.
- [Bocherens et al., 2017] Bocherens, H., Schrenk, F., Chaimanee, Y., Kullmer, O., Mörike, D., Pushkina, D., and Jaeger, J. J. (2017). Flexibility of diet and habitat in Pleistocene South

- Asian mammals: Implications for the fate of the giant fossil ape *Gigantopithecus*. *Quaternary International*, 434:148–155.
- [Boersema et al., 2009] Boersema, P. J., Mohammed, S., and Heck, A. J. (2009). Phosphopeptide fragmentation and analysis by mass spectrometry.
- [Bollongino et al., 2008] Bollongino, R., Tresset, A., and Vigne, J. D. (2008). Environment and excavation: Pre-lab impacts on ancient DNA analyses. *Comptes Rendus - Palevol*, 7(2-3):91–98.
- [Bouckaert et al., 2019] Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Poppinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650.
- [Boyd and Somogyi, 2010] Boyd, R. and Somogyi, Á. (2010). The mobile proton hypothesis in fragmentation of protonated peptides: A perspective.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [Brenton and Godfrey, 2010] Brenton, A. G. and Godfrey, A. R. (2010). Accurate Mass Measurement: Terminology and Treatment of Data. *Journal of the American Society for Mass Spectrometry*, 21(11):1821–1835.
- [Brown et al., 2020] Brown, K. A., Melby, J. A., Roberts, D. S., and Ge, Y. (2020). Top-down proteomics: challenges, innovations, and applications in basic and clinical research.
- [Bruderer et al., 2017] Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D., and Reiter, L. (2017). Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. *Molecular & Cellular Proteomics*, page mcp.RA117.000314.
- [Brunet et al., 2002] Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J. R., De Bonis, L., Coppens, Y., Dejax, J., Denys, C., Douring, P., Eisenmann, V., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau, F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Campomanes, P. P., De Leon, M. P., Rage, J. C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A., and Zollikofer, C. (2002). A new hominid from the upper Miocene of Chad, Central Africa. *Nature*, 418(6894):145–151.
- [Buckley et al., 2015] Buckley, M., Farina, R. A., Lawless, C., Tambusso, P. S., Varela, L., Carlini, A. A., Powell, J. E., and Martinez, J. G. (2015). Collagen sequence analysis of the extinct giant ground sloths *lestodon* and *megatherium*. *PLoS ONE*, 10(11):e0139611.
- [Buckley et al., 2010] Buckley, M., Whitcher Kansa, S., Howard, S., Campbell, S., Thomas-Oates, J., and Collins, M. (2010). Distinguishing between archaeological sheep and goat bones using a single collagen peptide. *Journal of Archaeological Science*, 37(1):13–20.
- [Cain, 2020] Cain, A. (2020). taxonomy — Definition, Examples, Levels, & Classification — Britannica.

- [Calisher, 2007] Calisher, C. H. (2007). Taxonomy: What's in a name? Doesn't a rose by any other name smell as sweet? *Croatian medical journal*, 48(2):268.
- [Calvete et al., 2014] Calvete, J. J., Bini, L., Hochstrasser, D., Sanchez, J. C., and Turck, N. (2014). The magic of words.
- [Campaña et al., 2016] Campaña, I., Pérez-González, A., Benito-Calvo, A., Rosell, J., Blasco, R., de Castro, J. M. B., Carbonell, E., and Arsuaga, J. L. (2016). New interpretation of the Gran Dolina-TD6 bearing Homo antecessor deposits through sedimentological analysis. *Scientific Reports*, 6:34799.
- [Cann et al., 1987] Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* 1987 325:6099, 325(6099):31–36.
- [Cano et al., 1993] Cano, R. J., Poinar, H. N., Pieniazek, N. J., Acra, A., and Poinar, G. O. (1993). Amplification and sequencing of DNA from a 120-135-million-year-old weevil. *Nature*, 363(6429):536–538.
- [Cantino and Queiroz, 2003] Cantino, P. and Queiroz, K. D. (2003). PhyloCode: a phylogenetic code of biological nomenclature. www.ohiou.edu/phylocode, (October 2014):1–62.
- [Cappellini et al., 2018] Cappellini, E., Prohaska, A., Racimo, F., Welker, F., Pedersen, M. W., Allentoft, M. E., De Barros Damgaard, P., Gutenbrunner, P., Dunne, J., Hammann, S., Roffet-Salque, M., Ilardo, M., Moreno-Mayar, J. V., Wang, Y., Sikora, M., Vinner, L., Cox, J., Evershed, R. P., and Willerslev, E. (2018). Ancient Biomolecules and Evolutionary Inference.
- [Cappellini et al., 2019] Cappellini, E., Welker, F., Pandolfi, L., Ramos-Madrigal, J., Samodova, D., Rütther, P. L., Fotakis, A. K., Lyon, D., Moreno-Mayar, J. V., Bukhsianidze, M., Rakownikow Jersie-Christensen, R., Mackie, M., Ginolhac, A., Ferring, R., Tappen, M., Palkopoulou, E., Dickinson, M. R., Stafford, T. W., Chan, Y. L., Götherström, A., Nathan, S. K., Heintzman, P. D., Kapp, J. D., Kirillova, I., Moodley, Y., Agusti, J., Kahlke, R. D., Kildadze, G., Martínez-Navarro, B., Liu, S., Sandoval Velasco, M., Sinding, M. H. S., Kelstrup, C. D., Allentoft, M. E., Orlando, L., Penkman, K., Shapiro, B., Rook, L., Dalén, L., Gilbert, M. T. P., Olsen, J. V., Lordkipanidze, D., and Willerslev, E. (2019). Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny. *Nature*, 574(7776):103–107.
- [Carbonell et al., 1995] Carbonell, E., Bermudez de Castro, J. M., Arsuaga, J. L., Diez, J. C., Rosas, A., Cuenca-Bescos, G., Sala, R., Mosquera, M., and Rodriguez, X. P. (1995). Lower Pleistocene hominids and artifacts from Atapuerca-TD6 (Spain). *Science*, 269(5225):826–830.
- [Carbonell et al., 1999a] Carbonell, E., Esteban, M., Nájera, A. M., Mosquera, M., Rodríguez, X. P., Ollé, A., Sala, R., Vergès, J. M., Bermúdez de Castro, J. M., and Ortega, A. I. (1999a). The Pleistocene site of Gran Dolina, Sierra de Atapuerca, Spain: a history of the archaeological investigations. *Journal of Human Evolution*, 37(3):313–324.
- [Carbonell et al., 1999b] Carbonell, E., García-Antón, M., Mallol, C., Mosquera, M., Ollé, A., Rodríguez, X. P., Sahnouni, M., Sala, R., and Vergès, J. M. (1999b). The TD6 level lithic industry from Gran Dolina, Atapuerca (Burgos, Spain): production and use. *Journal of Human Evolution*, 37(3):653–693.

- [Castellano et al., 2014] Castellano, S., Parra, G., Sánchez-Quinto, F. A., Racimo, F., Kuhlwilm, M., Kircher, M., Sawyer, S., Fu, Q., Heinze, A., Nickel, B., Dabney, J., Siebauer, M., White, L., Burbano, H. A., Renaud, G., Stenzel, U., Lalueza-Fox, C., De La Rasilla, M., Rosas, A., Rudan, P., Brajkoviæ, D., Kucan, Ž., Gušic, I., Shunkov, M. V., Derevianko, A. P., Viola, B., Meyer, M., Kelso, J., Andrés, A. M., and Pääbo, S. (2014). Patterns of coding variation in the complete exomes of three Neandertals. *Proceedings of the National Academy of Sciences of the United States of America*, 111(18):6666–6671.
- [Castiblanco et al., 2015] Castiblanco, G. A., Rutishauser, D., Ilag, L. L., Martignon, S., Castellanos, J. E., and Mejía, W. (2015). Identification of proteins from human permanent erupted enamel. *European Journal of Oral Sciences*, 123(6):390–395.
- [Chambers et al., 2012] Chambers, M. C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics.
- [Chen et al., 2019] Chen, F., Welker, F., Shen, C. C., Bailey, S. E., Bergmann, I., Davis, S., Xia, H., Wang, H., Fischer, R., Freidline, S. E., Yu, T. L., Skinner, M. M., Stelzer, S., Dong, G., Fu, Q., Dong, G., Wang, J., Zhang, D., and Hublin, J. J. (2019). A late Middle Pleistocene Denisovan mandible from the Tibetan Plateau. *Nature*, 569(7756):409–412.
- [Chen and Liu, 2014] Chen, H. and Liu, Y. (2014). Teeth. In *Advanced Ceramics for Dentistry*, pages 5–21. Butterworth-Heinemann.
- [Chen et al., 2001] Chen, T., Tepel, M., Rush, J., Church, G. M., and Kao, M. Y. (2001). A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8(3):325–337.
- [Chick et al., 2015] Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature Biotechnology*, 33(7):743–749.
- [Choi and Nesvizhskii, 2008] Choi, H. and Nesvizhskii, A. I. (2008). Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *Journal of Proteome Research*, 7(1):254–265.
- [Choudhuri, 2014] Choudhuri, S. (2014). Phylogenetic Analysis. *Bioinformatics for Beginners*, pages 209–218.
- [Chun et al., 2010] Chun, Y. H. P., Yamakoshi, Y., Yamakoshi, F., Fukae, M., Hu, J. C. C., Bartlett, J. D., and Simmer, J. P. (2010). Cleavage Site Specificity of MMP-20 for Secretory-stage Ameloblastin. *Journal of Dental Research*, 89(8):785–790.

- [Ciochon et al., 1996] Ciochon, R., Long, V. T., Laricki, R., González, L., Grüni, R., De Vos, J., Yonge, C., Taylor, L., Yoshida, H., and Reagan, M. (1996). Dated co-occurrence of *Homo erectus* and *Gigantopithecus* from Tham Khuyen Cave, Vietnam. *Proceedings of the National Academy of Sciences of the United States of America*, 93(7):3016–3020.
- [Cleland et al., 2016] Cleland, T. P., Schroeter, E. R., Feranec, R. S., and Vashishth, D. (2016). Peptide sequences from the first *Castoroides ohioensis* skull and the utility of old museum collections for palaeoproteomics. *Proceedings of the Royal Society B: Biological Sciences*, 283(1832).
- [Cleland et al., 2015] Cleland, T. P., Schroeter, E. R., and Schweitzer, M. H. (2015). Biologically and diagenetically derived peptide modifications in moa collagens. *Proceedings of the Royal Society B: Biological Sciences*, 282(1808).
- [Colaert et al., 2009] Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009). Improved visualization of protein consensus sequences by iceLogo. *Nature Methods*, 6(11):786–787.
- [Collins and Jukes, 1994] Collins, D. W. and Jukes, T. H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics*, 20(3):386–396.
- [Colonese et al., 2017] Colonese, A. C., Hendy, J., Lucquin, A., Speller, C. F., Collins, M. J., Carrer, F., Gubler, R., Kühn, M., Fischer, R., and Craig, O. E. (2017). New criteria for the molecular identification of cereal grains associated with archaeological artefacts. *Scientific Reports*, 7(1):1–7.
- [Coon et al., 2005] Coon, J. J., Syka, J., Shabanowitz, J., Hunt, D. F., and Others (2005). Tandem mass spectrometry for peptide and protein sequence analysis. *Biotechniques*, 38(4):519–521.
- [Cottrell, 2011] Cottrell, J. S. (2011). Protein identification using MS/MS data.
- [Cox and Mann, 2008] Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26(12):1367–1372.
- [Cox et al., 2011] Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10(4):1794–1805.
- [Craig et al., 2006] Craig, R., Cortens, J. C., Fenyo, D., and Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research*, 5(8):1843–1849.
- [Cristobal et al., 2017] Cristobal, A., Marino, F., Post, H., Van Den Toorn, H. W., Mohammed, S., and Heck, A. J. (2017). Toward an Optimized Workflow for Middle-Down Proteomics. *Analytical Chemistry*, 89(6):3318–3325.
- [Cuenca-Bescós et al., 2015] Cuenca-Bescós, G., Blain, H.-A., Rofes, J., Lozano-Fernández, I., López-García, J. M., Duval, M., Galán, J., and Núñez-Lahuerta, C. (2015). Comparing two

- different Early Pleistocene microfaunal sequences from the caves of Atapuerca, Sima del Elefante and Gran Dolina (Spain): Biochronological implications and significance of the Jaramillo subchron. *Quaternary International*, 389:148–158.
- [Cuenca-Bescós et al., 1999] Cuenca-Bescós, G., Laplana, C., and Canudo, J. I. (1999). Biochronological implications of the Arvicolidae (Rodentia, Mammalia) from the Lower Pleistocene hominid-bearing level of Trincheras Dolina 6 (TD6, Atapuerca, Spain). *Journal of Human Evolution*, 37(3):353–373.
- [Dabney et al., 2013] Dabney, J., Meyer, M., and Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, 5(7):a012567.
- [Dallongeville et al., 2016] Dallongeville, S., Garnier, N., Rolando, C., and Tokarski, C. (2016). Proteins in Art, Archaeology, and Paleontology: From Detection to Identification.
- [Dančák et al., 1999] Dančák, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999). De novo peptide sequencing via tandem mass spectrometry. In *Journal of Computational Biology*, volume 6, pages 327–342. Mary Ann Liebert, Inc. Pp.
- [Dayhoff et al., 1978] Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. (1978). A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, pages 345–352.
- [De Manuel et al., 2016] De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F., Casals, F., Angedakin, S., Arandjelovic, M., Boesch, C., Kühl, H., Vigilant, L., Langergraber, K., Novembre, J., Gut, M., Gut, I., Navarro, A., Carlsen, F., Andrés, A. M., Siegmund, H. R., Scally, A., Excoffier, L., Tyler-Smith, C., Castellano, S., Xue, Y., Hvilsum, C., and Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354(6311):477–481.
- [Dean and Schrenk, 2003] Dean, M. C. and Schrenk, F. (2003). Enamel thickness and development in a third permanent molar of *Gigantopithecus blacki*. *Journal of Human Evolution*, 45(5):381–388.
- [Degroeve et al., 2013] Degroeve, S., Martens, L., and Jurisica, I. (2013). MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–3203.
- [Demarchi et al., 2016] Demarchi, B., Hall, S., Roncal-Herrero, T., Freeman, C. L., Woolley, J., Crisp, M. K., Wilson, J., Fotakis, A., Fischer, R., Kessler, B. M., Jersie-Christensen, R. R., Olsen, J. V., Haile, J., Thomas, J., Marean, C. W., Parkington, J., Presslee, S., Lee-Thorp, J., Ditchfield, P., Hamilton, J. F., Ward, M. W., Wang, C. M., Shaw, M. D., Harrison, T., Domínguez-Rodrigo, M., Macphee, R. D., Kwekason, A., Ecker, M., Horwitz, L. K., Chazan, M., Kroger, R., Thomas-Oates, J., Harding, J. H., Cappellini, E., Penkman, K., and Collins, M. J. (2016). Protein sequences bound to mineral surfaces persist into deep time. *eLife*, 5(September).
- [Dent et al., 2004] Dent, B. B., Forbes, S. L., and Stuart, B. H. (2004). Review of human decomposition processes in soil.

- [DeSalle et al., 1992] DeSalle, R., Gatesy, J., Wheeler, W., and Grimaldi, D. (1992). DNA sequences from a fossil termite in oligo-miocene amber and their phylogenetic implications. *Science*, 257(5078):1933–1936.
- [Dickinson et al., 2019] Dickinson, M., Lister, A. M., and Penkman, K. E. H. (2019). A new method for enamel amino acid racemization dating: a closed system approach. *Quaternary Geochronology*, 50:29–46.
- [Doerr, 2014] Doerr, A. (2014). DIA mass spectrometry. *Nature Methods*, 12(1):35.
- [Dong et al., 2014] Dong, N. P., Liang, Y. Z., Xu, Q. S., Mok, D. K. W., Yi, L. Z., Lu, H. M., He, M., and Fan, W. (2014). Prediction of Peptide Fragment Ion Mass Spectra by Data Mining Techniques. *Analytical Chemistry*, 86(15):7446–7454.
- [Dorfer et al., 2014] Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., and Mechtler, K. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, 13(8):3679–3684.
- [Dorsey and Dill, 1989] Dorsey, J. G. and Dill, K. A. (1989). The Molecular Mechanism of Retention in Reversed-Phase Liquid Chromatography. *Chemical Reviews*, 89(2):331–346.
- [Drucker et al., 1997] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems*, 9(x):155–161.
- [Duval et al., 2018] Duval, M., Grün, R., Parés, J. M., Martín-Francés, L., Campaña, I., Rosell, J., Shao, Q., Arsuaga, J. L., Carbonell, E., and Bermúdez de Castro, J. M. (2018). The first direct ESR dating of a hominin tooth from Atapuerca Gran Dolina TD-6 (Spain) supports the antiquity of Homo antecessor. *Quaternary Geochronology*, 47:120–137.
- [Edman and Begg, 1967] Edman, P. and Begg, G. (1967). A Protein Sequenator. In *European Journal of Biochemistry*, volume 1, pages 80–91. Eur J Biochem.
- [Efron, 2000] Efron, B. (2000). The Bootstrap and Modern Statistics. *Journal of the American Statistical Association*, 95(452):1293–1296.
- [Efron and Tibshirani, 1986] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75.
- [Ehrlich et al., 2010] Ehrlich, H., Deutzmann, R., Brunner, E., Cappellini, E., Koon, H., Solazzo, C., Yang, Y., Ashford, D., Thomas-Oates, J., Lubeck, M., Baessmann, C., Langrock, T., Hoffmann, R., Wörheide, G., Reitner, J., Simon, P., Tsurkan, M., Ereskovsky, A. V., Kurek, D., Bazhenov, V. V., Hunoldt, S., Mertig, M., Vyalikh, D. V., Molodtsov, S. L., Kummer, K., Worch, H., Smetacek, V., and Collins, M. J. (2010). Mineralization of the metre-long biosilica structures of glass sponges is templated on hydroxylated collagen. *Nature Chemistry*, 2(12):1084–1088.
- [Elias and Gygi, 2007] Elias, J. E. and Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214.

- [Elias and Gygi, 2010] Elias, J. E. and Gygi, S. P. (2010). Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology (Clifton, N.J.)*, 604:55–71.
- [Eliuk and Makarov, 2015] Eliuk, S. and Makarov, A. (2015). Evolution of Orbitrap Mass Spectrometry Instrumentation.
- [Falguères et al., 1999] Falguères, C., Bahain, J.-J., Yokoyama, Y., Arsuaga, J. L., Bermúdez de Castro, J. M., Carbonell, E., Bischoff, J. L., and Dolo, J.-M. (1999). Earliest humans in Europe: the age of TD6 Gran Dolina, Atapuerca, Spain. *Journal of Human Evolution*, 37(3):343–352.
- [Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.
- [Feng, 2009] Feng, X. (2009). Chemical and Biochemical Basis of Cell-Bone Matrix Interaction in Health and Disease. *Current Chemical Biology*, 3(2):189–196.
- [Ferring et al., 2011] Ferring, R., Oms, O., Agustí, J., Berna, F., Nioradze, M., Shelia, T., Tappen, M., Vekua, A., Zhvania, D., and Lordkipanidze, D. (2011). Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85-1.78 Ma. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10432–10436.
- [Fincham et al., 1991] Fincham, A. G., Bessem, C. C., Lau, E. C., Pavlova, Z., Shuler, C., Slavkin, H. C., and Snead, M. L. (1991). Human developing enamel proteins exhibit a sex-linked dimorphism. *Calcified Tissue International*, 48(4):288–290.
- [Fischer et al., 2005] Fischer, B., Roth, V., Roos, F., Grossmann, J., Baginsky, S., Widmayer, P., Gruissem, W., and Buhmann, J. M. (2005). NovoHMM: A hidden Markov model for de novo peptide sequencing. *Analytical Chemistry*, 77(22):7265–7273.
- [Francalacci, 1995] Francalacci, P. (1995). DNA recovery from ancient tissues: problems and perspectives. *Human Evolution*, 10(1):81–91.
- [Frank and Pevzner, 2005] Frank, A. and Pevzner, P. (2005). PepNovo: De novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77(4):964–973.
- [Freidline et al., 2013] Freidline, S. E., Gunz, P., Harvati, K., and Hublin, J.-J. (2013). Evaluating developmental shape changes in Homo antecessor subadult facial morphology. *Journal of Human Evolution*, 65(4):404–423.
- [Fukae et al., 1996] Fukae, M., Tanabe, T., Murakami, C., Dohi, N., Uchida, T., and Shimizu, M. (1996). Primary Structure of the Porcine 89-kDa Enamelin. *Advances in Dental Research*, 10(2):111–118.
- [G. Marshall et al., 2013] G. Marshall, A., T. Blakney, G., Chen, T., K. Kaiser, N., M. McKenna, A., P. Rodgers, R., M. Ruddy, B., and Xian, F. (2013). Mass Resolution and Mass Accuracy: How Much Is Enough? *Mass Spectrometry*, 2(Special Issue):S0009–S0009.
- [Gabunia et al., 2000] Gabunia, L., Vekua, A., Lordkipanidze, D., Swisher, C. C., Ferring, R., Justus, A., Nioradze, M., Tvalchrelidze, M., Antón, S. C., Bosinski, G., Jöris, O., Lumley,

- M. A., Majsuradze, G., and Mouskhelishvili, A. (2000). Earliest Pleistocene hominid cranial remains from Dmanisi, Republic of Georgia: taxonomy, geological setting, and age. *Science*, 288(5468):1019–1025.
- [García and Arsuaga, 1999] García, N. and Arsuaga, J. L. (1999). Carnivores from the Early Pleistocene hominid-bearing Trinchera Dolina 6 (Sierra de Atapuerca, Spain). *Journal of Human Evolution*, 37(3):415–430.
- [Garnier et al., 1996] Garnier, J., Gibrat, J.-F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *methods in enzymology*, Volume 266(1995):540–553.
- [Gasse et al., 2015] Gasse, B., Chiari, Y., Silvent, J., Davit-Béal, T., and Sire, J.-Y. (2015). Amelotin: an enamel matrix protein that experienced distinct evolutionary histories in amphibians, sauropsids and mammals. *BMC Evolutionary Biology*, 15(1):47.
- [Gil et al., 1987] Gil, E., Aguirre, E., and Hoyos, M. (1987). Contexto estratigráfico. In Aguirre Carbonell, E., & Bermúdez de Castro, J.M., E., editor, *El hombre fósil de Ibeas y el Pleistoceno de la Sierra de Atapuerca*. Junta de Castilla y León, Consejería de Cultura y Turismo, Valladolid, Spain.
- [Glimcher et al., 1990] Glimcher, M. J., Cohen-Solal, L., Kossiva, D., and de Ricqlès, A. (1990). Biochemical analyses of fossil enamel and dentin. *Paleobiology*, 16(2):219–232.
- [Golovin et al., 2017] Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., and Sculley, D. (2017). Google Vizier : A Service for Black-Box Optimization. *Kdd '17*, page 10.
- [Gómez-Robles et al., 2013] Gómez-Robles, A., Bermúdez de Castro, J. M., Arsuaga, J.-L., Carbonell, E., and Polly, P. D. (2013). No known hominin species matches the expected dental morphology of the last common ancestor of Neanderthals and modern humans. *Proceedings of the National Academy of Sciences*, 110(45):18196–18201.
- [Good et al., 2007] Good, D. M., Wirtala, M., McAlister, G. C., and Coon, J. J. (2007). Performance characteristics of electron transfer dissociation mass spectrometry. *Molecular & cellular proteomics : MCP*, 6(11):1942–1951.
- [Graves et al., 2009] Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868.
- [Graves and Haystead, 2002] Graves, P. R. and Haystead, T. A. J. (2002). Molecular Biologist's Guide to Proteomics. *Microbiology and Molecular Biology Reviews*, 66(1):39–63.
- [Green et al., 2010] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H. Y., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušić, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., De La Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso,

- J., Lachmann, M., Reich, D., and Pääbo, S. (2010). A draft sequence of the neandertal genome. *Science*, 328(5979):710–722.
- [Grehan and Schwartz, 2009] Grehan, J. R. and Schwartz, J. H. (2009). Evolution of the second orangutan: Phylogeny and biogeography of hominid origins. *Journal of Biogeography*, 36(10):1823–1844.
- [Guindon et al., 2010] Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.
- [Hahnloser et al., 2000] Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.
- [Haile-Selassie, 2001] Haile-Selassie, Y. (2001). Late Miocene hominids from the Middle Awash, Ethiopia. *Nature*, 412(6843):178–181.
- [Hajdinjak et al., 2018] Hajdinjak, M., Fu, Q., Hübner, A., Petr, M., Mafessoni, F., Grote, S., Skoglund, P., Narasimham, V., Rougier, H., Crevecoeur, I., Semal, P., Soressi, M., Talamo, S., Hublin, J.-J., Gušić, I., Kučan, Ž., Rudan, P., Golovanova, L. V., Doronichev, V. B., Posth, C., Krause, J., Korlević, P., Nagel, S., Nickel, B., Slatkin, M., Patterson, N., Reich, D., Prüfer, K., Meyer, M., Pääbo, S., and Kelso, J. (2018). Reconstructing the genetic history of late Neanderthals. *Nature*, 555(7698):652–656.
- [Han et al., 2008] Han, X., Aslanian, A., and Yates, J. R. (2008). Mass spectrometry for proteomics.
- [Han et al., 2005] Han, Y., Ma, B., and Zhang, K. (2005). Spider: Software for protein identification from sequence tags with de novo sequencing error. *Journal of Bioinformatics and Computational Biology*, 3(3):697–716.
- [Hanson-Smith and Johnson, 2016] Hanson-Smith, V. and Johnson, A. (2016). PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. *PLoS Computational Biology*, 12(7):e1004976.
- [Harrison, 2010] Harrison, T. (2010). Apes among the tangled branches of human origins.
- [Heiss et al., 2003] Heiss, A., DuChesne, A., Denecke, B., Grötzinger, J., Yamamoto, K., Renneé, T., and Jahnen-Dechent, W. (2003). Structural basis of calcification inhibition by α 2-HS glycoprotein/fetuin-A: Formation of colloidal calciprotein particles. *Journal of Biological Chemistry*, 278(15):13333–13341.
- [Hendy, 2021] Hendy, J. (2021). Ancient protein analysis in archaeology.
- [Hendy et al., 2018] Hendy, J., Welker, F., Demarchi, B., Speller, C., Warinner, C., and Collins, M. J. (2018). A guide to ancient protein studies.
- [Higuchi et al., 1984] Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., and Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282–284.

- [Hill et al., 2015] Hill, R. C., Wither, M. J., Nemkov, T., Barrett, A., Alessandro, A., Dzieciatkowska, M., and Hansen, K. C. (2015). Preserved Proteins from Extinct Bison latifrons Identified by Tandem Mass Spectrometry; Hydroxylysine Glycosides are a Common Feature of Ancient Collagen. *Molecular & Cellular Proteomics*, 14(7):1946–1958.
- [Hill, 1965] Hill, R. L. (1965). Hydrolysis of proteins. *Advances in Protein Chemistry*, 20:37–107.
- [Ho et al., 2003] Ho, C. S., Lam, C. W. K., Chan, M. H. M., Cheung, R. C. K., Law, L. K., Lit, L. C. W., Ng, K. F., Suen, M. W. M., and Tai, H. L. (2003). Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical biochemist. Reviews*, 24(1):3–12.
- [Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. J. (1997). Long short-term memory. *Neural Computation*, 9(8):1–32.
- [Hong et al., 2012] Hong, C., Jiang, H., Lü, E., Wu, Y., Guo, L., Xie, Y., Wang, C., and Yang, Y. (2012). Identification of milk component in ancient food residue by proteomics. *PLoS ONE*, 7(5).
- [Hu et al., 2005] Hu, J. C. C., Yamakoshi, Y., Yamakoshi, F., Krebsbach, P. H., and Simmer, J. P. (2005). Proteomics and Genetics of Dental Enamel. *Cells Tissues Organs*, 181(3-4):219–231.
- [Huang and McLuckey, 2010] Huang, T. Y. and McLuckey, S. A. (2010). Gas-phase chemistry of multiply charged bioions in analytical mass spectrometry. *Annual Review of Analytical Chemistry*, 3(1):365–385.
- [Huang R; Gu, Y; Larick, R; Fang, Q; Yonge, C; de Vos, J; Schwarcz, H P; Rink, W J, 1995] Huang R; Gu, Y; Larick, R; Fang, Q; Yonge, C; de Vos, J; Schwarcz, H P; Rink, W J, W. C. (1995). Earliest hominids and artifacts from Asia: Longgupo Cave, Central China. *Nature*, 378:275–278.
- [Hublin, 2009] Hublin, J. J. (2009). The origin of Neandertals. *Proceedings of the National Academy of Sciences*, 106(38):16022.
- [Hunt et al., 1986] Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*.
- [Hynekl et al., 2004] Hynekl, R., Kuckova, S., Hradilova, J., and Kodicek, M. (2004). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry as a tool for fast identification of protein binders in color layers of paintings. *Rapid Communications in Mass Spectrometry*, 18(17):1896–1900.
- [Iwata et al., 2007] Iwata, T., Yamakoshi, Y., Hu, J. C. C., Ishikawa, I., Bartlett, J. D., Krebsbach, P. H., and Simmer, J. P. (2007). Processing of Ameloblastin by MMP-20. *Journal of Dental Research*, 86(2):153–157.
- [Jágr et al., 2012] Jágr, M., Eckhardt, A., Pataridis, S., and Mikšík, I. (2012). Comprehensive proteomic analysis of human dentin. *European Journal of Oral Sciences*, 120(4):259–268.

- [Jin et al., 2014] Jin, C., Wang, Y., Deng, C., Harrison, T., Qin, D., Pan, W., Zhang, Y., Zhu, M., and Yan, Y. (2014). Chronological sequence of the early Pleistocene Gigantopithecus faunas from cave sites in the Chongzuo, Zuojiang River area, South China. *Quaternary International*, 354:4–14.
- [Jukes, T.H. and Cantor, 1969] Jukes, T.H. and Cantor, C. (1969). Evolution of protein molecules. *Munro, H.N., Ed., Mammalian Protein Metabolism, Academic Press, New York, 21-132*, pages 21–132.
- [Käll et al., 2007a] Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007a). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925.
- [Käll et al., 2007b] Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2007b). Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research*, 7(1):40–44.
- [Käll et al., 2008] Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases.
- [Kannan and Wheeler, 2012] Kannan, L. and Wheeler, W. C. (2012). Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology*, 7(1):1–10.
- [Kapli et al., 2020] Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444.
- [Karas and Hillenkamp, 1988] Karas, M. and Hillenkamp, F. (1988). Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons.
- [Karunratanakul et al., 2019] Karunratanakul, K., Tang, H. Y., Speicher, D. W., Chuangsuwanich, E., and Sriswasdi, S. (2019). Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular and Cellular Proteomics*, 18(12):2478–2491.
- [Kato, 2002] Kato, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.
- [Kato and Frith, 2012] Kato, K. and Frith, M. C. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics*, 28(23):3144–3146.
- [Kato et al., 2016] Kato, S., Beyene, Y., Itaya, T., Hyodo, H., Hyodo, M., Yagi, K., Gouzu, C., WoldeGabriel, G., Hart, W. K., Ambrose, S. H., Nakaya, H., Bernor, R. L., Boissarie, J. R., Bibi, F., Saegusa, H., Sasaki, T., Sano, K., Asfaw, B., and Suwa, G. (2016). New geological and palaeontological age constraint for the gorilla-human lineage split. *Nature*, 530(7589):215–218.
- [Kaufman and Manley, 1998] Kaufman, D. S. and Manley, W. F. (1998). A new procedure for determining DL amino acid ratios in fossils using reverse phase liquid chromatography. *Quaternary Science Reviews*, 17(11):987–1000.
- [Kayser et al., 2004] Kayser, J. P., Vallet, J. L., and Cerny, R. L. (2004). Defining parameters for homology-tolerant database searching. *Journal of Biomolecular Techniques*, 15(4):285–295.

- [Kelleher, 2004] Kelleher, N. L. (2004). Peer Reviewed: Top-Down Proteomics. *Analytical Chemistry*, 76(11):196 A–203 A.
- [Keller et al., 2002] Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20):5383–5392.
- [Kelley, 2002] Kelley, J. (2002). The hominoid radiation in Asia. In Hartwig, W. C., editor, *The Primate Fossil Record*, pages 369–384. Cambridge University Press, Cambridge.
- [Kelstrup et al., 2018] Kelstrup, C. D., Bekker-Jensen, D. B., Arrey, T. N., Hoglebe, A., Harder, A., and Olsen, J. V. (2018). Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *Journal of Proteome Research*, 17(1):727–738.
- [Khodadoust et al., 2017] Khodadoust, M., Olsson, N., Wagar, L., Haabeth, O., Chen, B., Swaminathan, K., Rawson, K., Liu, C., Steiner, D., Lund, P., Rao, S., Zhang, L., Marceau, C., Stehr, H., Newman, A., Czerwinski, D. K., Carlton, V., Moorhead, M., Faham, M., Kohrt, H., Carette, J., Green, M., Davis, M., Levy, R., Elias, J. E., and Alizadeh, A. (2017). Antigen Presentation Profiling Reveals Recognition of Lymphoma Immunoglobulin Neoantigens. *Nature*, 543(7647):723.
- [Kim et al., 2010] Kim, S., Mischerikow, N., Bandeira, N., Navarro, J. D., Wich, L., Mohammed, S., Heck, A. J., and Pevzner, P. A. (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: Applications to database search. *Molecular and Cellular Proteomics*, 9(12):2840–2852.
- [Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.
- [Kingma and Ba, 2015] Kingma, D. P. and Ba, J. L. (2015). Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations 2015*, pages 1–15.
- [Kono et al., 2014] Kono, R. T., Zhang, Y., Jin, C., Takai, M., and Suwa, G. (2014). A 3-dimensional assessment of molar enamel thickness and distribution pattern in *Gigantopithecus blacki*. *Quaternary International*, 354:46–51.
- [Korneliussen et al., 2014] Korneliussen, T., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):356.
- [Krause et al., 2010] Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., and Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464(7290):894–897.
- [Krings et al., 1997] Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M., and Pääbo, S. (1997). Neandertal DNA sequences and the origin of modern humans. *Cell*, 90(1):19–30.
- [Krokhin, 2006] Krokhin, O. V. (2006). Sequence-specific retention calculator. Algorithm for peptide retention prediction in ion-pair RP-HPLC: Application to 300- and 100-Åpore size C18 sorbents. *Analytical Chemistry*.

- [Kuhlwilm et al., 2016] Kuhlwilm, M., Gronau, I., Hubisz, M. J., De Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., De La Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Ž., Gušić, I., Marques-Bonet, T., Andrés, A. M., Viola, B., Pääbo, S., Meyer, M., Siepel, A., and Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, 530(7591):429–433.
- [Lacruz et al., 2013] Lacruz, R. S., de Castro, J. M. B., Martínón-Torres, M., O’Higgins, P., Paine, M. L., Carbonell, E., Arsuaga, J. L., and Bromage, T. G. (2013). Facial Morphogenesis of the Earliest Europeans. *PLOS ONE*, 8(6):e65199.
- [Lacruz et al., 2019] Lacruz, R. S., Stringer, C. B., Kimbel, W. H., Wood, B., Harvati, K., O’Higgins, P., Bromage, T. G., and Arsuaga, J.-L. (2019). The evolutionary history of the human face. *Nature Ecology & Evolution*, 3(5):726–736.
- [Lam et al., 2008] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., Stein, S. E., and Aebersold, R. (2008). Building consensus spectral libraries for peptide identification in proteomics. *Nature Methods*, 5(10):873–875.
- [Lane, 2005] Lane, C. S. (2005). Mass spectrometry-based proteomics in the life sciences.
- [Langergraber et al., 2012] Langergraber, K. E., Prüfer, K., Rowney, C., Boesch, C., Crockett, C., Fawcett, K., Inoue, E., Inoue-Muruyama, M., Mitani, J. C., Muller, M. N., Robbins, M. M., Schubert, G., Stoinski, T. S., Viola, B., Watts, D., Wittig, R. M., Wrangham, R. W., Zuberbuler, K., Pääbo, S., and Vigilant, L. (2012). Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(39):15716–15721.
- [Lanigan et al., 2020] Lanigan, L. T., Mackie, M., Feine, S., Hublin, J. J., Schmitz, R. W., Wilcke, A., Collins, M. J., Cappellini, E., Olsen, J. V., Taurozzi, A. J., and Welker, F. (2020). Multi-protease analysis of Pleistocene bone proteomes. *Journal of Proteomics*, 228.
- [Laumont et al., 2018] Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J. P., Gendron, P., Courcelles, M., Hardy, M. P., Côté, C., Durette, C., St-Pierre, C., Benhammadi, M., Lanoix, J., Vobecky, S., Haddad, E., Lemieux, S., Thibault, P., and Perreault, C. (2018). Noncoding regions are the main source of targetable tumor-specific antigens. *Science Translational Medicine*, 10(470).
- [Le and Gascuel, 2008] Le, S. Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7):1307–1320.
- [LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- [Li et al., 2009] Li, H., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Handsaker, R. E. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, 25(16).
- [Liu et al., 2013] Liu, H., Ponniah, G., Neill, A., Patel, R., and Andrien, B. (2013). Accurate determination of protein methionine oxidation by stable isotope labeling and LC-MS analysis. *Analytical Chemistry*, 85(24):11705–11709.

- [Lordkipanidze et al., 2013] Lordkipanidze, D., de Leon, M. S. P., Margvelashvili, A., Rak, Y., Rightmire, G. P., Vekua, A., and Zollikofer, C. P. E. (2013). A complete skull from Dmanisi, Georgia, and the evolutionary biology of early Homo. *Science*, 342(6156):326–331.
- [Ma, 2015] Ma, B. (2015). Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry*, 26(11):1885–1894.
- [Ma et al., 2003] Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(20):2337–2342.
- [Mackie et al., 2018] Mackie, M., R  ther, P., Samodova, D., Di Gianvincenzo, F., Granzotto, C., Lyon, D., Peggie, D. A., Howard, H., Harrison, L., Jensen, L. J., Olsen, J. V., and Cappellini, E. (2018). Palaeoproteomic Profiling of Conservation Layers on a 14th Century Italian Wall Painting. *Angewandte Chemie - International Edition*, 57(25):7369–7374.
- [Maixner et al., 2018] Maixner, F., Turaev, D., Cazenave-Gassiot, A., Janko, M., Krause-Kyora, B., Hoopmann, M. R., Kusebauch, U., Sartain, M., Guerriero, G., O’Sullivan, N., Teasdale, M., Cipollini, G., Paladin, A., Mattiangeli, V., Samadelli, M., Tecchiati, U., Putzer, A., Palazoglu, M., Meissen, J., L  sch, S., Rausch, P., Baines, J. F., Kim, B. J., An, H. J., Gostner, P., Egarter-Vigl, E., Malfertheiner, P., Keller, A., Stark, R. W., Wenk, M., Bishop, D., Bradley, D. G., Fiehn, O., Engstrand, L., Moritz, R. L., Doble, P., Franke, A., Nebel, A., Oeggl, K., Rattei, T., Grimm, R., and Zink, A. (2018). The Iceman’s Last Meal Consisted of Fat, Wild Meat, and Cereals. *Current Biology*, 28(14):2348–2355.e9.
- [Makarov, 2000] Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, 72(6):1156–1162.
- [Mallick et al., 2007] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., and Aebersold, R. (2007). Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*.
- [Mallick et al., 2016] Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J. P., Song, Y. S., Poletti, G., Balloux, F., Van Driem, G., De Knijff, P., Romero, I. G., Jha, A. R., Behar, D. M., Bravi, C. M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O. L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M. S., Ruiz-Linares, A., Beall, C. M., Di Rienzo, A., Jeong, C., Starikovskaya, E. B., Metspalu, E., Parik, J., Villems, R., Henn, B. M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatoyannopoulos, G., Wee, J. T., Khusainova, R., Khusnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M. F., Kivisild, T., Klitz, W., Winkler, C. A., Labuda, D., Bamshad, M., Jorde, L. B., Tishkoff, S. A., Watkins, W. S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Pa  bo, S., Kelso, J., Patterson, N., and Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.
- [Marciniak and Perry, 2017] Marciniak, S. and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation.

- [Marks, 1988] Marks, J. (1988). Molecular evolutionary genetics. By M. Nei. New York: Columbia University Press. 1987. x + 512 pp., tables, figures, indexes. \$50.00 (cloth). *American Journal of Physical Anthropology*, 75(3):428–429.
- [Martínón-Torres et al., 2007] Martínón-Torres, M., Castro, J. M. B. D., Gómez-Robles, A., Arsuaga, J. L., Carbonell, E., Lordkipanidze, D., Manzi, G., and Margvelashvili, A. (2007). Dental Evidence on the Hominin Dispersals during the Pleistocene. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33):13279–13282.
- [Marx, 2013] Marx, V. (2013). Targeted proteomics. *Nature Methods*, 10(1):19–22.
- [Masters, 2002] Masters, J. R. (2002). HeLa cells 50 years on: The good, the bad and the ugly.
- [McWilliams and Suomalainen, 2019] McWilliams, T. G. and Suomalainen, A. (2019). Mitochondrial DNA can be inherited from fathers, not just mothers.
- [Meyer et al., 2016] Meyer, M., Arsuaga, J. L., De Filippo, C., Nagel, S., Aximu-Petri, A., Nickel, B., Martínez, I., Gracia, A., De Castro, J. M. B., Carbonell, E., Viola, B., Kelso, J., Prüfer, K., and Pääbo, S. (2016). Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature*, 531(7595):504–507.
- [Meyer et al., 2012] Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., De Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J., Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226.
- [Michalski et al., 2012] Michalski, A., Neuhauser, N., Cox, J., and Mann, M. (2012). A systematic investigation into the nature of tryptic HCD spectra. *Journal of Proteome Research*, 11(11):5479–5491.
- [Miller et al., 2010] Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees.
- [Miller et al., 2008] Miller, S. F., White, J. L., and Ciochon, R. L. (2008). Assessing mandibular shape variation within *Gigantopithecus* using a geometric morphometric approach. *American Journal of Physical Anthropology*, 137(2):201–212.
- [Mitchell Wells and McLuckey, 2005] Mitchell Wells, J. and McLuckey, S. A. (2005). Collision-induced dissociation (CID) of peptides and proteins.
- [Molnar, 1971] Molnar, S. (1971). Human tooth wear, tooth function and cultural variability. *American Journal of Physical Anthropology*, 34(2):175–189.
- [Moreno et al., 2015] Moreno, D., Falguères, C., Pérez-González, A., Voinchet, P., Ghaleb, B., Despriée, J., Bahain, J.-J., Sala, R., Carbonell, E., Bermúdez de Castro, J. M., and Arsuaga, J. L. (2015). New radiometric dates on the lowest stratigraphical section (TD1 to TD6) of Gran Dolina site (Atapuerca, Spain). *Quaternary Geochronology*, 30:535–540.

- [Mukund Sundararajan, Ankur Taly, 2017] Mukund Sundararajan, Ankur Taly, Q. Y. (2017). Axiomatic Attribution for Deep Networks. *arxiv.org*.
- [Müller and Vingron, 2001] Müller, T. and Vingron, M. (2001). Modeling amino acid replacement. *Journal of Computational Biology*, 7(6):761–776.
- [Muñoz and Heck, 2014] Muñoz, J. and Heck, A. J. (2014). From the human genome to the human proteome. *Angewandte Chemie - International Edition*, 53(41):10864–10866.
- [Nagano et al., 2009] Nagano, T., Kakegawa, A., Yamakoshi, Y., Tsuchiya, S., Hu, J. C. C., Gomi, K., Arai, T., Bartlett, J. D., and Simmer, J. P. (2009). Mmp-20 and Klk4 Cleavage Site Preferences for Amelogenin Sequences. *Journal of Dental Research*, 88(9):823–828.
- [Nanjappa et al., 2014] Nanjappa, V., Thomas, J. K., Marimuthu, A., Muthusamy, B., Radhakrishnan, A., Sharma, R., Ahmad Khan, A., Balakrishnan, L., Sahasrabudde, N. A., Kumar, S., Jhaveri, B. N., Sheth, K. V., Kumar Khatana, R., Shaw, P. G., Srikanth, S. M., Mathur, P. P., Shankar, S., Nagaraja, D., Christopher, R., Mathivanan, S., Raju, R., Sirdeshmukh, R., Chatterjee, A., Simpson, R. J., Harsha, H. C., Pandey, A., and Prasad, T. S. (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Research*.
- [Nascimento et al., 2017] Nascimento, F. F., dos Reis, M., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis.
- [Nater et al., 2017] Nater, A., Mattle-Greminger, M. P., Nurcahyo, A., Nowak, M. G., de Manuel, M., Desai, T., Groves, C., Pybus, M., Sonay, T. B., Roos, C., Lameira, A. R., Wich, S. A., Askew, J., Davila-Ross, M., Fredriksson, G., de Valles, G., Casals, F., Prado-Martinez, J., Goossens, B., Verschoor, E. J., Warren, K. S., Singleton, I., Marques, D. A., Pamungkas, J., Perwitasari-Farajallah, D., Rianti, P., Tuuga, A., Gut, I. G., Gut, M., Orozco-terWengel, P., van Schaik, C. P., Bertranpetit, J., Anisimova, M., Scally, A., Marques-Bonet, T., Meijaard, E., and Krützen, M. (2017). Erratum: Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species (*Current Biology* (2017) 27(22) (3487–3498.e10) (S0960982217312459)(10.1016/j.cub.2017.09.047)). *Current Biology*, 27(22):3576–3577.
- [Nesvizhskii, 2010] Nesvizhskii, A. I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.
- [Nesvizhskii and Aebersold, 2005] Nesvizhskii, A. I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data: The protein inference problem.
- [Olivé et al., 1990] Olivé, A., Ramirez-Merino, J. L., and Ortega, L. I. (1990). Mapa Geológico de España a Escala 1:50,000 (Belorado, 201).
- [Olsen et al., 2007] Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, 4(9):709–712.
- [Orlando et al., 2013] Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliusson, T., Malaspinas, A. S., Vogt, J., Szklarczyk,

- D., Kelstrup, C. D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A. M., Cahill, J., Rasmussen, M., Wang, X., Min, J., Zazula, G. D., Seguin-Orlando, A., Mortensen, C., Magnussen, K., Thompson, J. F., Weinstock, J., Gregersen, K., Røed, K. H., Eisenmann, V., Rubin, C. J., Miller, D. C., Antczak, D. F., Bertelsen, M. F., Brunak, S., Al-Rasheid, K. A., Ryder, O., Andersson, L., Mundy, J., Krogh, A., Gilbert, M. T. P., Kjær, K., Sicheritz-Ponten, T., Jensen, L. J., Olsen, J. V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J., and Willerslev, E. (2013). Recalibrating equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74–78.
- [Ortega-Martínez, 2009] Ortega-Martínez, A. I. (2009). *La evolución geomorfológica del karst de la sierra de Atapuerca (Burgos) y su relación con los yacimientos pleistocenos que contiene*. PhD thesis, Burgos, Spain.
- [Ostrom et al., 2000] Ostrom, P. H., Schall, M., Gandhi, H., Shen, T. L., Hauschka, P. V., Strahler, J. R., and Gage, D. A. (2000). New strategies for characterizing ancient proteins using matrix-assisted laser desorption ionization mass spectrometry. *Geochimica et Cosmochimica Acta*, 64(6):1043–1050.
- [Ozcan et al., 2014] Ozcan, S., Kim, B. J., Ro, G., Kim, J. H., Bereuter, T. L., Reiter, C., Dimapasoc, L., Garrido, D., Mills, D. A., Grimm, R., Lebrilla, C. B., and An, H. J. (2014). Glycosylated proteins preserved over millennia: N-glycan analysis of Tyrolean Iceman, Scythian Princess and Warrior. *Scientific Reports*, 4(1):1–8.
- [Pääbo et al., 1989] Pääbo, S., Higuchi, R. G., and Wilson, A. C. (1989). Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology.
- [Pääbo et al., 2004] Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., and Hofreiter, M. (2004). Genetic analyses from ancient DNA.
- [Panawala, 2017] Panawala, L. (2017). Difference Between Taxonomy and Systematics.
- [Parés et al., 2018] Parés, J. M., Álvarez, C., Sier, M., Moreno, D., Duval, M., Woodhead, J. D., Ortega, A. I., Campaña, I., Rosell, J., Bermúdez de Castro, J. M., and Carbonell, E. (2018). Chronology of the cave interior sediments at Gran Dolina archaeological site, Atapuerca (Spain). *Quaternary Science Reviews*, 186:1–16.
- [Parés and Pérez-González, 1995] Parés, J. M. and Pérez-González, A. (1995). Paleomagnetic age for hominid fossils at Atapuerca archaeological site, Spain. *Science*, 269(5225):830–832.
- [Parés and Pérez-González, 1999] Parés, J. M. and Pérez-González, A. (1999). Magnetostratigraphy and stratigraphy at Gran Dolina section, Atapuerca (Burgos, Spain). *Journal of Human Evolution*, 37(3):325–342.
- [Park et al., 2017] Park, J., Piehowski, P. D., Wilkins, C., Zhou, M., Mendoza, J., Fujimoto, G. M., Gibbons, B. C., Shaw, J. B., Shen, Y., Shukla, A. K., Moore, R. J., Liu, T., Petyuk, V. A., Tolić, N., Paša-Tolić, L., Smith, R. D., Payne, S. H., and Kim, S. (2017). Informed-Proteomics: Open-source software package for top-down proteomics. *Nature Methods*, 14(9):909–914.

- [Parker et al., 2019] Parker, G. J., Yip, J. M., Eerkens, J. W., Salemi, M., Durbin-Johnson, B., Kiesow, C., Haas, R., Buikstra, J. E., Klaus, H., Regan, L. A., Rocke, D. M., and Phinney, B. S. (2019). Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel. *Journal of Archaeological Science*, 101:169–180.
- [Pedersen et al., 2015] Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., Wales, N. A., Carøe, C., Campos, P. F., Schmidt, A. M. Z., Gilbert, M. T. P., Hansen, A. J., Orlando, L., and Willerslev, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130383.
- [Pei, 1965] Pei, W. (1965). Excavation of Liucheng Gigantopithecus cave and exploration of other caves in Kwangsi. *Memoir of the Institute of Vertebrate Palaeontology and Palaeoanthropology, Academia Sinica*, 7:1–54.
- [Pei, 1957] Pei, W.-c. (1957). Discovery of Gigantopithecus mandibles and other material in Liu-cheng district of central Kwangsi in south China. *Vertebrata Palasiatica*, 1(2):65–71, plates I–III.
- [Penkert et al., 2019] Penkert, M., Hauser, A., Harmel, R., Fiedler, D., Hackenberger, C. P., and Krause, E. (2019). Electron Transfer/Higher Energy Collisional Dissociation of Doubly Charged Peptide Ions: Identification of Labile Protein Phosphorylations. *Journal of the American Society for Mass Spectrometry*, 30(9):1578–1585.
- [Penkman et al., 2008] Penkman, K. E. H., Kaufman, D. S., Maddy, D., and Collins, M. J. (2008). Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quaternary Geochronology*, 3(1-2):2–25.
- [Pérez-González et al., 1995] Pérez-González, A., Aleixandre, T., Pinilla, A., Gallardo, J., Benayas, J., Martínez, M. J., and Ortega, A. I. (1995). An Approach to the Galeria Stratigraphy in the Sierra de Atapuerca Trench (Burgos). In Bermúdez de Castro, J. M., Arsuaga, J. L., and Carbonell, E., editors, *Human Evolution in Europe and the Atapuerca evidence*, pages 99–122. Junta de Castilla y León, Consejería de Cultura y Turismo, Valladolid, Spain.
- [Peris et al., 2020] Peris, D., Janssen, K., Barthel, H. J., Bierbaum, G., Delclòs, X., Peñalver, E., Solórzano-Kraemer, M. M., Jordal, B. H., and Rust, J. (2020). DNA from resin-embedded organisms: Past, present and future. *PLoS ONE*, 15(9 September):e0239521.
- [Perkins et al., 1999] Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. In *Electrophoresis*, volume 20, pages 3551–3567. Electrophoresis.
- [Pineda and Arce, 1997] Pineda, A. and Arce, J. M. (1997). Mapa Geológico de España a escala 1:50,000 (Burgos, 200).
- [Porto et al., 2011] Porto, I. M., Laure, H. J., de Sousa, F. B., Rosa, J. C., and Gerlach, R. F. (2011). New techniques for the recovery of small amounts of mature enamel proteins. *Journal of Archaeological Science*, 38(12):3596–3604.
- [Prado-Martinez et al., 2013] Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere,

- G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M. L., Stevison, L., Camprub, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R. E., Pusey, A., Lankester, F., Kiyang, J. A., Bergl, R. A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S. A., Mullikin, J. C., Wilson, R. K., Gut, I. G., Gonder, M. K., Ryder, O. A., Hahn, B. H., Navarro, A., Akey, J. M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M. H., Hvilsom, C., Andrés, A. M., Wall, J. D., Bustamante, C. D., Hammer, M. F., Eichler, E. E., and Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459):471–475.
- [Preece and Penkman, 2005] Preece, R. C. and Penkman, K. E. H. (2005). New faunal analyses and amino acid dating of the Lower Palaeolithic site at East Farm, Barnham, Suffolk. *Proceedings of the Geologists' Association*, 116(3-4):363–377.
- [Price et al., 2009] Price, P. A., Toroian, D., and Lim, J. E. (2009). Mineralization by inhibitor exclusion. The calcification of collagen with fetuin. *Journal of Biological Chemistry*, 284(25):17092–17101.
- [Prüfer et al., 2017] Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., Dannemann, M., Nelson, B. J., Key, F. M., Rudan, P., Kućan, Ž., Gušić, I., Golovanova, L. V., Doronichev, V. B., Patterson, N., Reich, D., Eichler, E. E., Slatkin, M., Schierup, M. H., Andrés, A. M., Kelso, J., Meyer, M., and Pääbo, S. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, 358(6363):655–658.
- [Prüfer et al., 2014] Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.
- [Qiao et al., 2021] Qiao, R., Tran, N. H., Xin, L., Chen, X., Li, M., Shan, B., and Ghodsi, A. (2021). Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3(5):420–425.
- [Queiroz, 2012] Queiroz, K. (2012). Biological Nomenclature from Linnaeus to the PhyloCode. *Bibliotheca Herpetologica*, 9(1-2):135–145.
- [Rambaut et al., 2018] Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

- [Ramsøe et al., 2020] Ramsøe, A., van Heekeren, V., Ponce, P., Fischer, R., Barnes, I., Speller, C., and Collins, M. J. (2020). DeamiDATE 1.0: Site-specific deamidation as a tool to assess authenticity of members of ancient proteomes. *Journal of Archaeological Science*, 115:105080.
- [Rannala and Yang, 1996] Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, 43(3):304–311.
- [Rightmire, 1998] Rightmire, G. (1998). Human evolution in the Middle Pleistocene: the role of *Homo heidelbergensis*. *Evolutionary anthropology*, 6(6):218–227.
- [Robinson and Robinson, 2001] Robinson, N. E. and Robinson, A. B. (2001). Molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(3):944–949.
- [Rogers et al., 2017] Rogers, A. R., Bohlender, R. J., and Huff, C. D. (2017). Early history of Neanderthals and Denisovans. *Proceedings of the National Academy of Sciences of the United States of America*, 114(37):9859–9863.
- [Ronquist et al., 2012] Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.
- [Rost et al., 1994] Rost, B., Sander, C., and Schneider, R. (1994). PHD-an automatic mail server for protein secondary structure prediction. *Bioinformatics*, 10(1):53–60.
- [Saah and Hoover, 1997] Saah, A. J. and Hoover, D. R. (1997). ‘Sensitivity’ and ‘specificity’ reconsidered: The meaning of these terms in analytical and diagnostic settings.
- [Salmon et al., 2013] Salmon, C. R., Tomazela, D. M., Ruiz, K. G. S., Foster, B. L., Paes Leme, A. F., Sallum, E. A., Somerman, M. J., and Nociti, F. H. (2013). Proteomic analysis of human dental cementum and alveolar bone. *Journal of Proteomics*, 91:544–555.
- [Salter, 2004] Salter, L. A. (2004). The phylogenetic handbook: A practical approach to DNA and protein phylogeny. *American Journal of Human Biology*, 16(3):354–355.
- [Sanders et al., 2007] Sanders, W. S., Bridges, S. M., McCarthy, F. M., Nanduri, B., and Burgess, S. C. (2007). Prediction of peptides observable by mass spectrometry applied at the experimental set level. In *BMC Bioinformatics*.
- [Savitski et al., 2006] Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006). Modifi-Comb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Molecular and Cellular Proteomics*, 5(5):935–948.
- [Schliep et al., 2017] Schliep, K., Potts, A. J., Morrison, D. A., and Grimm, G. W. (2017). Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution*, 8(10):1212–1220.
- [Schliep, 2011] Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

- [Schroeder and Bada, 1976] Schroeder, R. A. and Bada, J. L. (1976). A review of the geochemical applications of the amino acid racemization reaction. *Earth Science Reviews*, 12(4):347–391.
- [Schroeter and Cleland, 2016] Schroeter, E. R. and Cleland, T. P. (2016). Glutamine deamidation: An indicator of antiquity, or preservational quality? *Rapid Communications in Mass Spectrometry*, 30(2):251–255.
- [Schubert et al., 2015] Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., Lam, H., Amodei, D., Mallick, P., Maclean, B., and Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, 10(3):426–441.
- [Schuster, 2008] Schuster, S. C. (2008). Next-generation sequencing transforms today's biology.
- [Scigelova et al., 2011] Scigelova, M., Hornshaw, M., Giannakopoulos, A., and Makarov, A. (2011). Fourier transform mass spectrometry.
- [Scott and Gras, 2012] Scott, R. and Gras, R. (2012). Comparing distance-based phylogenetic tree construction methods using an individual-based ecosystem simulation, ecosim. In *Artificial Life 13: Proceedings of the 13th International Conference on the Simulation and Synthesis of Living Systems, ALIFE 2012*, pages 105–110. MIT Press Journals.
- [Seidler et al., 2010] Seidler, J., Zinn, N., Boehm, M. E., and Lehmann, W. D. (2010). De novo sequencing of peptides by MS/MS.
- [Senko et al., 1995] Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233.
- [Senut et al., 2001] Senut, B., Pickford, M., Gommery, D., Mein, P., Cheboi, K., and Coppens, Y. (2001). Premier hominidé du Miocène (formation de Lukeino, Kenya). *Comptes Rendus de l'Académie de Sciences - Serie IIA: Sciences de la Terre et des Planètes*, 332(2):137–144.
- [Seo et al., 2016] Seo, J., Singh, N. N., Ottesen, E. W., Lee, B. M., and Singh, R. N. (2016). A novel human-specific splice isoform alters the critical C-terminus of Survival Motor Neuron protein. *Scientific Reports*, 6.
- [Shao et al., 2014] Shao, Q., Wang, W., Deng, C., Voinchet, P., Lin, M., Zazzo, A., Douville, E., Dolo, J. M., Falguères, C., and Bahain, J. J. (2014). ESR, U-series and paleomagnetic dating of Gigantopithecus fauna from Chuifeng Cave, Guangxi, southern China. *Quaternary Research (United States)*, 82(1):270–280.
- [Shoulders and Raines, 2009] Shoulders, M. D. and Raines, R. T. (2009). Collagen structure and stability.
- [Shukla and Futrell, 2000] Shukla, A. K. and Futrell, J. H. (2000). Tandem mass spectrometry: Dissociation of ions by collisional activation.

- [Sinitcyn et al., 2021] Sinitcyn, P., Hamzeiy, H., Salinas Soto, F., Itzhak, D., McCarthy, F., Wichmann, C., Steger, M., Ohmayer, U., Distler, U., Kaspar-Schoenefeld, S., Prianichnikov, N., Yilmaz, , Rudolph, J. D., Tenzer, S., Perez-Riverol, Y., Nagaraj, N., Humphrey, S. J., and Cox, J. (2021). MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology*, pages 1–11.
- [Sinitcyn et al., 2018a] Sinitcyn, P., Rudolph, J. D., and Cox, J. (2018a). Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science*, 1:207–234.
- [Sinitcyn et al., 2018b] Sinitcyn, P., Tiwary, S., Rudolph, J., Gutenbrunner, P., Wichmann, C., Yllmaz, , Hamzeiy, H., Salinas, F., and Cox, J. (2018b). MaxQuant goes Linux.
- [Sobott et al., 2009] Sobott, F., Watt, S. J., Smith, J., Edelmann, M. J., Kramer, H. B., and Kessler, B. M. (2009). Comparison of CID Versus ETD Based MS/MS Fragmentation for the Analysis of Protein Ubiquitination. *Journal of the American Society for Mass Spectrometry*, 20(9):1652–1659.
- [Steen and Mann, 2004] Steen, H. and Mann, M. (2004). The ABC’s (and XYZ’s) of peptide sequencing. *Nat Rev Mol Cell Biol*, 5(9):699–711.
- [Stein and Scott, 1994] Stein, S. E. and Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866.
- [Stewart et al., 2017] Stewart, N. A., Gerlach, R. F., Gowland, R. L., Gron, K. J., and Montgomery, J. (2017). Sex determination of human remains from peptides in tooth enamel. *Proceedings of the National Academy of Sciences of the United States of America*, 114(52):13649–13654.
- [Stewart et al., 2016] Stewart, N. A., Molina, G. F., Mardegan Issa, J. P., Yates, N. A., Sosovicka, M., Vieira, A. R., Line, S. R. P., Montgomery, J., and Gerlach, R. F. (2016). The identification of peptides by nanoLC-MS/MS from human surface tooth enamel following a simple acid etch extraction. *RSC Advances*, 6(66):61673–61679.
- [Stringer, 2016] Stringer, C. (2016). The origin and evolution of Homo sapiens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698).
- [Sun et al., 2014] Sun, L., Wang, Y., Liu, C., Zuo, T., Ge, J., Zhu, M., Jin, C., Deng, C., and Zhu, R. (2014). Magnetochronological sequence of the early pleistocene gigantopithecus faunas in Chongzuo, Guangxi, southern China. *Quaternary International*, 354:15–23.
- [Sykes et al., 1995] Sykes, G. A., Collins, M. J., and Walton, D. I. (1995). The significance of a geochemically isolated intracrystalline organic fraction within biominerals. *Organic Geochemistry*, 23(11-12):1059–1065.
- [Taanman, 1999] Taanman, J. W. (1999). The mitochondrial genome: Structure, transcription, translation and replication.
- [Tabb et al., 2003] Tabb, D. L., Saraf, A., and Yates, J. R. (2003). GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Analytical Chemistry*, 75(23):6415–6421.

- [Tagliabracci et al., 2012] Tagliabracci, V. S., Engel, J. L., Wen, J., Wiley, S. E., Worby, C. A., Kinch, L. N., Xiao, J., Grishin, N. V., and Dixon, J. E. (2012). Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science*, 336(6085):1150–1153.
- [Tajima and Nei, 1984] Tajima, F. and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, 1(3):269–285.
- [Tang and Skibsted, 2016] Tang, N. and Skibsted, L. H. (2016). Calcium binding to amino acids and small glycine peptides in aqueous solution: Toward peptide design for better calcium bioavailability. *Journal of Agricultural and Food Chemistry*, 64(21):4376–4389.
- [Taylor and Johnson, 1997] Taylor, J. A. and Johnson, R. S. (1997). Sequence database searches via de Novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11(9):1067–1075.
- [Ternette et al., 2016] Ternette, N., Yang, H., Partridge, T., Llano, A., Cedeño, S., Fischer, R., Charles, P. D., Dudek, N. L., Mothe, B., Crespo, M., Fischer, W. M., Korber, B. T. M., Nielsen, M., Borrow, P., Purcell, A. W., Brander, C., Dorrell, L., Kessler, B. M., and Hanke, T. (2016). Defining the HLA class I-associated viral antigen repertoire from HIV-1-infected human cells. *European Journal of Immunology*, 46(1):60–69.
- [Thammana, 2016] Thammana, M. (2016). A Review on High Performance Liquid Chromatography (HPLC). *Journal of Pharmaceutical Analysis*, 5(2).
- [The 1000 Genomes Project, 2015] The 1000 Genomes Project, C. (2015). A global reference for human genetic variation. *Nature*, 526:68–74.
- [Thiede et al., 2005] Thiede, B., Höhenwarter, W., Krah, A., Mattow, J., Schmid, M., Schmidt, F., and Jungblut, P. R. (2005). Peptide mass fingerprinting. *Methods*, 35(3 SPEC.ISS.):237–247.
- [Thompson et al., 2003] Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry*, 75(8):1895–1904.
- [Tiwary et al., 2019] Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K. K., Deming, L., Berndl, M., Brant, A., Cimermancic, P., and Cox, J. (2019). High quality MS/MS spectrum prediction for data-dependent and -independent acquisition data analysis. *Nat Methods*.
- [Tokarski et al., 2006] Tokarski, C., Martin, E., Rolando, C., and Cren-Olivé, C. (2006). Identification of proteins in Renaissance paintings by proteomics. *Analytical Chemistry*, 78(5):1494–1502.
- [Tran et al., 2019] Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., and Li, M. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, 16(1):63–66.
- [Tran et al., 2016] Tran, N. H., Rahman, M. Z., He, L., Xin, L., Shan, B., and Li, M. (2016). Complete de Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports*, 6.

- [Tran et al., 2017] Tran, N. H., Zhang, X., Xin, L., Shan, B., and Li, M. (2017). De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 114(31):8247–8252.
- [Tsou et al., 2015] Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A. I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, 12(3):258–264.
- [Tyanova et al., 2016a] Tyanova, S., Temu, T., and Cox, J. (2016a). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11(12):2301–2319.
- [Tyanova et al., 2016b] Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016b). The Perseus computational platform for comprehensive analysis of (prote)omics data.
- [van der Made, 1999] van der Made, J. (1999). Ungulates from Atapuerca TD6. *Journal of Human Evolution*, 37(3):389–413.
- [van der Valk et al., 2021] van der Valk, T., Pečnerová, P., Díez-del Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., Lister, A. M., Götherström, A., and Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, 591(7849):265–269.
- [Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York.
- [Vizcaíno et al., 2016] Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R., and Hermjakob, H. (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, 44(D1):D447—D456.
- [von Koenigswald, 1935] von Koenigswald, G. H. R. (1935). Eine fossile Säugetiernefauna mit simia aus Sudchina. *Proceedings, Koninklijke Akademie van Wetenschappen, Amsterdam*, 38:872–879.
- [Wadsworth and Buckley, 2014] Wadsworth, C. and Buckley, M. (2014). Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. *Rapid Communications in Mass Spectrometry*, 28(6):605–615.
- [Wagner et al., 2010] Wagner, G. A., Krbetschek, M., Degering, D., Bahain, J.-J., Shao, Q., Falguères, C., Voinchet, P., Dolo, J.-M., Garcia, T., and Rightmire, G. P. (2010). Radiometric dating of the type-site for *Homo heidelbergensis* at Mauer, Germany. *Proceedings of the National Academy of Sciences*, 107(46):19726–19730.
- [Walsh et al., 2005] Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J. (2005). Protein post-translational modifications: The chemistry of proteome diversifications.

- [Wang, 2009] Wang, W. (2009). New discoveries of *Gigantopithecus blacki* teeth from Chuifeng Cave in the Bubing Basin, Guangxi, south China. *Journal of Human Evolution*, 57(3):229–240.
- [Washburn, 2015] Washburn, M. P. (2015). The H-Index of 'An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database'.
- [Welker, 2018a] Welker, F. (2018a). Elucidation of cross-species proteomic effects in human and hominin bone proteome identification through a bioinformatics experiment. *BMC Evolutionary Biology*, 18(1):1–11.
- [Welker, 2018b] Welker, F. (2018b). Palaeoproteomics for human evolution studies. *Quaternary Science Reviews*, 190:137–147.
- [Welker et al., 2015] Welker, F., Collins, M. J., Thomas, J. A., Wadsley, M., Brace, S., Cappellini, E., Turvey, S. T., Reguero, M., Gelfo, J. N., Kramarz, A., Burger, J., Thomas-Oates, J., Ashford, D. A., Ashton, P. D., Rowsell, K., Porter, D. M., Kessler, B., Fischer, R., Baessmann, C., Kaspar, S., Olsen, J. V., Kiley, P., Elliott, J. A., Kelstrup, C. D., Mullin, V., Hofreiter, M., Willerslev, E., Hublin, J. J., Orlando, L., Barnes, I., and Macphee, R. D. (2015). Ancient proteins resolve the evolutionary history of Darwin's South American ungulates. *Nature*, 522(7554):81–84.
- [Welker et al., 2016] Welker, F., Hajdinjak, M., Talamo, S., Jaouen, K., Dannemann, M., David, F., Julien, M., Meyer, M., Kelso, J., Barnes, I., Brace, S., Kamminga, P., Fischer, R., Kessler, B. M., Stewart, J. R., Pääbo, S., Collins, M. J., and Hublin, J. J. (2016). Palaeoproteomic evidence identifies archaic hominins associated with the Châtelperronian at the Grotte du Renne. *Proceedings of the National Academy of Sciences of the United States of America*, 113(40):11162–11167.
- [Welker et al., 2019] Welker, F., Ramos-Madrigal, J., Kuhlwilm, M., Liao, W., Gutenbrunner, P., de Manuel, M., Samodova, D., Mackie, M., Allentoft, M. E., Bacon, A. M., Collins, M. J., Cox, J., Lalueza-Fox, C., Olsen, J. V., Demeter, F., Wang, W., Marques-Bonet, T., and Cappellini, E. (2019). Enamel proteome shows that *Gigantopithecus* was an early diverging pongine. *Nature*, 576(7786):262–265.
- [Welker et al., 2017] Welker, F., Smith, G. M., Hutson, J. M., Kindler, L., Garcia-Moreno, A., Villaluenga, A., Turner, E., and Gaudzinski-Windheuser, S. (2017). Middle Pleistocene protein sequences from the rhinoceros genus *Stephanorhinus* and the phylogeny of extant and extinct Middle/Late Pleistocene Rhinocerotidae. *PeerJ*, 5:e3033.
- [Willerslev and Cooper, 2005] Willerslev, E. and Cooper, A. (2005). Ancient DNA.
- [Wolters et al., 2001] Wolters, D. A., Washburn, M. P., and Yates, J. R. (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, 73(23):5683–5690.
- [Wu et al., 2016] Wu, J. X., Song, X., Pascovici, D., Zaw, T., Care, N., Krisp, C., and Molloy, M. P. (2016). SWATH Mass Spectrometry Performance Using Extended Peptide MS/MS Assay Libraries. *Molecular & Cellular Proteomics*, 15(7):2501–2514.

- [Xue et al., 2015] Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsum, C., Mudakikwa, A., Eichler, E. E., Cranfield, M. R., Marques-Bonet, T., Tyler-Smith, C., and Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231):242–245.
- [Yamakoshi et al., 2006] Yamakoshi, Y., Hu, J. C. C., Fukae, M., Yamakoshi, F., and Simmer, J. P. (2006). How do enamelysin and kallikrein 4 process the 32-kDa enamelin? *European Journal of Oral Sciences*, 114(s1):45–51.
- [Yamashita and Fenn, 1984] Yamashita, M. and Fenn, J. B. (1984). Electrospray ion source. Another variation on the free-jet theme. *Journal of Physical Chemistry*, 88(20):4451–4459.
- [Yang et al., 2019] Yang, H., Chi, H., Zeng, W. F., Zhou, W. J., and He, S. M. (2019). PNovo 3: Precise de novo peptide sequencing using a learning-to-rank framework. In *Bioinformatics*, volume 35, pages i183–i190. Bioinformatics.
- [Yates et al., 1998] Yates, J. R., Morgan, S. F., Gatlin, C. L., Griffin, P. R., and Eng, J. K. (1998). Method to Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis. *Analytical Chemistry*, 70(17):3557–3565.
- [Zadrozna et al., 2003] Zadrozna, I., Połec-Pawlak, K., Głuch, I., Ackacha, M. A., Mojski, M., Witowska-Jarosz, J., and Jarosz, M. (2003). Old master paintings - A fruitful field of activity for analysts: Targets, methods, outlook.
- [Zarnack et al., 2013] Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3):453–466.
- [Zazo et al., 1983] Zazo, C., Goy, J. L., and Hoyos, M. (1983). Estudio geomorfológico de los alrededores de la Sierra de Atapuerca (Burgos). *Estudios Geológicos*, 39:179–185.
- [Zeder and Lapham, 2010] Zeder, M. A. and Lapham, H. A. (2010). Assessing the reliability of criteria used to identify postcranial bones in sheep, Ovis, and goats, Capra. *Journal of Archaeological Science*, 37(11):2887–2905.
- [Zeder and Pilaar, 2010] Zeder, M. A. and Pilaar, S. E. (2010). Assessing the reliability of criteria used to identify mandibles and mandibular teeth in sheep, Ovis, and goats, Capra.
- [Zerbino et al., 2018] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2018). Ensembl 2018. *Nucleic acids research*, 46(D1):D754–D761.

- [Zhang et al., 2012] Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G. A., and Ma, B. (2012). PEAKS DB: De novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular and Cellular Proteomics*, 11(4):M111.010587.
- [Zhang et al., 2013] Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C., and Yates, J. R. (2013). Protein analysis by shotgun/bottom-up proteomics.
- [Zhang and Harrison, 2017] Zhang, Y. and Harrison, T. (2017). Gigantopithecus blacki: a giant ape from the Pleistocene of Asia revisited. *American Journal of Physical Anthropology*, 162:153–177.
- [Zhang et al., 2014] Zhang, Y., Jin, C., Cai, Y., Kono, R., Wang, W., Wang, Y., Zhu, M., and Yan, Y. (2014). New 400-320ka Gigantopithecus blacki remains from Hejiang Cave, Chongzuo City, Guangxi, South China. *Quaternary International*, 354:35–45.
- [Zhao and Zhang, 2013] Zhao, L. X. and Zhang, L. Z. (2013). New fossil evidence and diet analysis of Gigantopithecus blacki and its distribution and extinction in South China. *Quaternary International*, 286:69–74.
- [Zhu et al., 2018] Zhu, Z., Dennell, R., Huang, W., Wu, Y., Qiu, S., Yang, S., Rao, Z., Hou, Y., Xie, J., Han, J., and Ouyang, T. (2018). Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago. *Nature*, 559(7715):608–612.
- [Zolg et al., 2017] Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., DeLanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H. C., Weininger, M., Yu, P., Schlegl, J., Kramer, K., Schmidt, T., Kusebauch, U., Deutsch, E. W., Aebersold, R., Moritz, R. L., Wenschuh, H., Moehring, T., Aiche, S., Huhmer, A., Reimer, U., and Kuster, B. (2017). Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, 14(3):259–262.

Acknowledgement

This dissertation would not have been possible without the extensive support and assistance of my supervisors, colleagues, friends and especially my family and partner.

I would like to thank my supervisor, PD Dr. Gertrud E. Rößner, whose professional support and guidance was invaluable to the success of this dissertation. A big thank you goes to my supervisor Dr. Jürgen Cox for giving me the opportunity to pursue my PhD in his research group and for his brilliant insights and analytical skills that brought this work to a higher level.

Many thanks to the entire TEMPERA ETN, especially to Prof. Enrico Cappellini who did an amazing job in guiding and organising the network, but also shared his extensive expertise on ancient proteins. I immensely benefited from and enjoyed our collaborations and also I am very grateful for all the international and fun experience. I especially would like to thank my co-authors for the amazing collaboration on such impactful research and the open as well as detailed discussions during my stay in Copenhagen. Frido, your attention to detail for ancient studies and guidance was indispensable. Also thank you to Fabiana, for the detailed proofreading of this dissertation, you've also become a dear friend to me.

I also would like to acknowledge my colleagues Christoph, Hamid and Şule: I would not have been able to successfully complete this dissertation without you. Thanks also to my entire group for all the fruitful discussions and laughter. You were always available when I needed you. Thanks for being such great colleagues and friends.

The biggest thanks goes to Lena, I am very grateful for the amazing friendship, the great laughter and thank you for being always there for me. Also for taking the time to give me the best feedback ever when proofreading this dissertation. Angelo, I would like to thank you for a great friendship, your guidance and input on future perspectives and opportunities.

The greatest thanks go to my amazing family who provided immense support during my entire PhD and being understanding for the little time I was available. Finally, I would like to thank my partner and best friend James who was always there for me. Your unwavering belief in me kept me motivated throughout my journey.