Proteomics
www.proteomics-journal.com

Analysis of the Preserved Amino Acid Bias in Peptide Profiles of Iron Age Teeth from a Tropical Environment Enable Sexing of Individuals Using Amelogenin MRM

Valerie C. Wasinger, Darren Curnoe,* Sonia Bustamante, Raynold Mendoza, Rasmi Shoocongdej, Lewis Adler, Andy Baker, Kanoknart Chintakanon, Ceridwen Boel, and Paul S.C. Tacon*

The first dental proteomic profile of Iron Age individuals (ca. 2000–1000 years B.P.), collected from the site of Long Long Rak rock shelter in northwest Thailand is described. A bias toward the preservation of the positively charged aromatic, and polar amino acids is observed. It is evident that the 212 proteins identified (2 peptide, FDR <1%) comprise a palimpsest of alterations that occurred both ante-mortem and post-mortem. Conservation of amino acids within the taphonomically resistant crystalline matrix enabled the identification of both X and Y chromosome linked amelogenin peptides. A novel multiple reaction monitoring method using the sex specific amelogenin protein isoforms is described and indicate the teeth are of male origin. Functional analysis shows an enrichment of pathways associated with metabolic disorders and shows a capacity for harboring these conditions prior to death. Stable isotope analysis using carbon isotopes highlights the strongly C_3 based (≈80%) diet of the Long Long Rak cemetery people, which probably comprised rice combined with protein from freshwater fish among other food items. The combination of proteomics and stable isotope analysis provides a complementary strategy for assessing the demography, diet, lifestyle, and possible diseases experienced by ancient populations.

1. Introduction

The investigation of ancient biomolecules has begun to revolutionize our understanding of human prehistory. The most dramatic findings over the last decade or so have resulted from the analysis of ancient DNA and include major advances on issues like the geographic source and timing of ancient human migrations,[1] diversity, and phylogeny of archaic hominins[2] and discovery and characterization of admixture between modern humans and archaic taxa.[3] Yet the preservation of DNA in ancient skeletal samples is a rare event, making the achievements of this nascent field all the more remarkable. In contrast, human tissue contains a high abundance of proteins which are characterized by at least a three-fold longer preservation time than DNA,[4] under similar preservation conditions. While

Dr. V. C. Wasinger, S. Bustamante, Dr. L. Adler
Bioanalytical Mass Spectrometry Facility
Mark Wainwright Analytical Centre
University of New South Wales Sydney
NSW, 2052, Australia
E-mail: v.wasinger@unsw.edu.au

Dr. V. C. Wasinger, Prof. D. Curnoe, R. Mendoza, Prof. A. Baker
Dr. C. Boel
Palaeontology, Geobiology and Earth Archives Research Centre
University of New South Wales Sydney
NSW, 2052, Australia
E-mail: d.curnoe@unsw.edu.au

Prof. D. Curnoe, R. Mendoza, Dr. C. Boel
ARC Centre of Excellence for Australian Biodiversity and Heritage
University of New South Wales Sydney
NSW, 2052, Australia

Dr. R. Shoocongdej
Department of Archaeology
Silpakorn University
Bangkok, 10200, Thailand

Dr. R. Shoocongdej
Interaction between Prehistoric Population and Environments in Highland Pang Mapha Project
Bangkok, 10170, Thailand

Dr. K. Chintakanon
Advanced Dental Technology Center
Thailand Science Park
Amphoe Khlong Luang, Chang Wat Pathum
Thani, 12120, Thailand

Prof. P. S. C. Tacon
PERAHU
Griffith Centre for Social and Cultural Research & School of Humanities Languages and Social Science
Griffith University
Gold Coast campus
QLD, 4222, Australia

DOI: 10.1002/pmic.201800341

the analysis of ancient proteins has significant potential for addressing a wide range of bioarchaeological and evolutionary questions it has lagged considerably behind ancient DNA research in its application and methodological development.

While the human genome contains around 19 000 genes, there are up to one million proteins and their isoforms in the human body.[5] Proteins are crucial to the normal functions as well as pathological processes that occur within cells,[6] reflecting a specific set of conditions within a given cell, tissue, or organ at a particular point in time.[6] Protein modifications occur as a result of altered gene expression in addition to the interactions that occur among various proteins and between different types of cells. The description and characterization of ancient protein peptides through mass spectrometry (palaeoproteomics) offers the opportunity to assess normal physiological and pathological processes that were at play in an individual prior to their death, and by extension, to understand group, population or even species-level environmental conditions, evolutionary pressures, and genetic adaptations. As some skeletal proteins are specific to developmental phases, or are expressed preferentially within the sexes, palaeoproteomics also offers the chance to assess the identity of individuals or to compare the life history characteristics among populations or across taxa.

Establishing the sex of anthropological remains is a fundamental bioarchaeological quest that has relied on the presence of specific DNA markers or physiological differences in skeletal structure.[7] Reliability is influenced by the degree of degradation, as well as conflicting changes in skeletal material resulting from hormone levels, age, and growth rates.[8] It also precludes that these structures are available for assessment. The application of the amelogenin protein dimorphism has recently been shown to be a reliable approach that complements or replaces these techniques in the absence of quality DNA or other osteological features.[9]

Relatively simple procedures in palaeopreotemics have been deployed over several decades and used to establish protein preservation in deep time[10] and to establish the taxonomic identity of faunal[11] including dinosaur, and hominin remains from isolated bones and fossils.[10,12,14] However, recent developments in high-throughput, high-resolution, mass spectrometry have led to greatly improved sensitivity and accuracy, which have now begun to be applied in the field of palaeopathology. For example, shotgun mass spectrometry has been used to identify highly antigenic proteins associated with bacteria such as gingipains in medieval human dental calculus,[15,16] periodontic disease-associated bacteria in Roman period dental calculus,[16] establish the possible presence of *Mycobacterium* through buccal swabs of a 500-year-old Inca mummy,[17] the possibility of infection in 360-year-old bone samples from Tokyo using a label-free approach,[14] and identification of proteins associated with tissue inflammation in skin samples of ≈4200-year-old mummies from Egypt.[18] In the only other exploration of the endogenous dental paleoproteome to our knowledge, peptides associated with plague bacterium *Yersinia pestis* were able to be identified in 300-year-old dental pulp from a funeral site in France.[19]

The human teeth we examine here were all collected during archaeological investigations undertaken at Long Long Rak (LLR) rock shelter. This limestone cave is located within Mae Hong Son Province, Northwest Thailand, at approximately 735 m above sea level (≈10 m above present ground surface). LLR was found to contain 19 teak coffins with human skeletal remains, earthenware sherds, animal bones, iron tools, a wooden loom, textiles, a wooden basket and large numbers of beads. Radiocarbon dating shows these log coffins to date within the range of 1960 ± 30 cal. yr BP to 1636 ± 44 cal. yr BP. The Log Coffin "Culture" present at LLR represents a distinctive burial practice of northwest Thailand. LLR is the only such locality that has been subjected to systematic archaeological excavation. In comparison to the Iron age mortuary practices of the lowland areas in central and northeast Thailand, the log coffin culture shows similar practices to those documented in highland central Vietnam and also the hanging coffins seen in South China.[20,21]

Here we apply for the first time proteomic analysis to six archaeological human teeth from LLR. We analyze the properties of the peptides identified including modifications and physicochemical properties of the amino acids preserved in this subset of proteins, and take advantage of these properties to develop an MRM technique capable of determining the sex of four individuals. We combine these methods with carbon and oxygen stable isotope analysis to broadly identify the diet of this community, and to test the proposition that broad dietary information may offer insights into the interpretation of proteins identified through palaeoproteomic investigation.

Significance Statement

Recovery of proteins preserved over time is challenging. This study presents the first recovery of the tooth proteome from a tropical humid environment (ca. 2000–1000 years B.P.). We highlight a novel application of co-detection within the ProgenesisQI environment, using ion alignment prior to identification, to detect common peptide features across all samples. This allowed the identification of LLR peptides based on strong MS/MS spectra in common to control peptides. We have termed this approach "ion-scaffolding." Analysis of common peptides has also shown the conservation of amino acids which can be classified as "small," "hydrophobic," and "acidic" in nature; this has likely contributed to their preservation over time and may be unique to the crystalline structure of the tooth matrix. Two important amelogenin protein isoforms have also been preserved, which can allow the determination of sex. Here we describe a novel MRM method using the isoforms of both X-related (a number of isotypes have been observed) and Y-related forms of the amelogenin protein. We also describe for the first time the combination and potential utility of stable isotope and protein pathway enrichment analysis at a group level to explore the association of diets and lifestyle consequences.

2. Experimental Section

2.1. Sample Collection and Study Population

Six Teeth were collected from two log coffins (C2 and C6) from Chamber 1 of a three-chambered cave at LLR. Archaeological excavations were conducted over three field seasons as a part of

multidisciplinary research on interaction between "Prehistoric Populations and Their Environments in Highland Pang Mapha Project," which aimed to examine mortuary variability and human use of the cave.[20] A total of 18 radiometric dates from samples of log coffins and lacquer suggests they represent a single culture with at least three occupational periods in which LLR functioned as a family cemetery.[21] Controls included deciduous teeth from two healthy individuals ($n = 2$) and extracted adult teeth from two healthy individuals (one male, one female) recovered 30 years ago from modern day humans. Control samples were run in duplicate and sufficient protein was obtained from sample 2B to be run in duplicate. LLR samples 2B, 2D, 6A, 6B, 6D where used in shotgun experiments and were treated as biological replicates representing the LLR population; LLR samples 2C, 2D, 6C, 6D were used for MRM experiments.

2.2. Treatment of Teeth

Methods for the extraction of dentin and enamel proteins has been adapted from the groups of Jàgr,[22] Hill,[23]and Capellini.[24] Briefly, teeth were cut lengthwise with a diamond blade. One portion of half of the tooth was used for proteomic analysis. All preparations for protein extractions were carried out in a biological safety cabinet in a PC1 lab environment. Appropriate protective equipment including gloves were worn when handling specimens. All care was taken to avoid cross-contamination by wiping down all surfaces between sampling. A dental burr was used to abrade the tooth surface and remove any surface contamination prior to ultrasonic cleaning for 30 min. Other potential contamination was handled with further wash steps as follows. Teeth were washed overnight in 0.5 M HCl at 4 °C followed by an overnight wash in 500 mM NaCl in 50 mM NH_4CO_3, pH8 with added protease inhibitors (Roche, EDTA free) at 10 RPM on a shaker at 4 °C. The tooth samples were then rinsed for 5 mins in ddH_20 twice and air-dried in a biological safety cabinet. The tooth enamel/dentin was crushed in a mortar and pestle under liquid nitrogen. An equal amount of powder was weighed out (100 mg) to continue the analysis. The samples were treated with 500 µL of 0.5 M HCl at 4 °C for 18 h, followed by 5 min centrifugation at 14 000 g. The pellet was retained and resuspended in 50 µL of 50 mM NH_4CO_3in 10 mM DTT, 2 M Urea, pH8. Protein concentration was measured with 2-D Quant Kit (GE Healthcare-Life sciences) following the manufacturer's instructions and 10 µg was taken for digestion with Trypsin overnight at 25 °C in a 1:100 ratio. Digestion was halted by acidification, and peptides dried down.

2.3. Mass Spectrometry of Samples

Digested peptides were reconstituted in 5 µL 0.1% formic acid and separated by nano-LC using an Ultimate 3000 HPLC and autosampler (Dionex, Amsterdam, Netherlands). The sample, 1.6 µg (0.8 µL), was loaded onto a virgin micro C18 pre-column (300 µm × 5 mm, Dionex) with H_2O:CH_3CN (98:2, 0.1% TFA) at 10 µL min^{-1}. After washing, the pre-column was switched (Valco 10 port valve, Dionex) into line with a virgin fritless nano column (75 µm i.d × 12 cm) containing reverse phase C18 media (1.9 µm, 120 Å, Dr. Maisch GmbH HPLC). Peptides were eluted using a linear gradient of H_2O:CH_3CN (98:2, 0.1% formic acid) to H_2O:CH_3CN (64:36, 0.1% formic acid) at 250 nl min^{-1} over 120 min. The OrbitrapVelos (Thermo Electron, Bremen, Germany) mass spectrometer was run in DDA mode as previously described.[25] LLR teeth samples were run prior to modern teeth samples to prevent cross contamination. The column and trap cartridge were tested for any contaminating proteins within the LC system by running buffer blanks (no sample) prior to analysis. The column was also washed with buffer injections (no sample) run in-between each sample to limit carry-over effects. An extraction blank was also prepared to confirm proteins identified were endogenous to the samples. In addition to the methods already described here, all measures taken to confirm protein authenticity, the significance of the tests and description of methods and results are given in Table S1, Supporting Information. This includes determining aspartic acid racemization, amino acid deamidation, and PCA analysis.

2.4. Determination of Protein Modifications

Although it may be expected that ancient proteins will have modifications due to diagenesis; the extent and variety of them from human tooth enamel from humid and tropical environments is not well understood. The search engine Peaks (v8.0) was used to profile post-translational modifications (PTM) present in the LLR samples. This preliminary search allowed us to refine the number of modifications unique to these samples. PTM profiling revealed that oxidation Δ M = 15.99 (M, K, P), deamidation Δ M = 0.98 (N,Q), and Carbamidomethyl Δ M = 57.02 (N-terminus) contributed to the modifications present in the LLR samples. These modifications obtained *Ascores* of 1000. These parameters were set as "variable" modifications in subsequent database searches.

2.5. Ion Feature Mapping and Identification of Peptides

Ion feature mapping was achieved by alignment of retention time (RT) using ProgenesisQI LC-MS data analysis software v4 (Nonlinear Dynamics, Newcastle upon Tyne, UK). The peptide intensities were normalized against total intensity (sample specific log-scale abundance ratio scaling factor) and compared between groups by one-way analysis of variance (ANOVA, $p \leq 0.05$ for statistical significance) and post-hoc multiple comparison procedures.[26] Application of ion alignment prior to identification restricted the analysis of peptides to those common across all sample types. This allowed the identification of LLR peptides based on strong MS/MS spectra in common to control peptides and eliminated the majority of environmental contaminants associated with the LLR samples.

Protein dataset-peak lists were generated using ProgenesisQI to create a mascot generic file and submitted to Mascot 2.6 (Matrix Science, London, UK, www.matrixscience.com). All MS/MS spectra were searched against the Uniprot and the CRAPome database (downloaded Jan 2018) using the Mascot search

program for protein identification with the following criteria: 1) species, *Homo sapien*; 2) allowed three missed cleavage; 3) variable modifications, Oxidation (M, K, P), Deamidation (N, Q), Carbamidomethyl (N-terminus), 4) peptide tolerance, ± 4 ppm; 5) MS/MS tolerance, ± 0.4 Da; 6) peptide charge +2 and +3; and 7) enzyme specificity, semi-tryptic. The peptides were considered to be confidently identified when matches had an ion score >34 and peptides were assigned to a protein. Only proteins identified from the Uniprot database with an adjusted FDR of 1%, with two unique peptides were used in further analysis. Comparison of amino acids (AA) length was done using NCSS 9 software and significance was calculated using a nonparametric, 2-sample Mann-Whitney *U* test.

2.6. Pathway Enrichment Analysis

Canonical Pathways Analysis tool (Ingenuity Systems http://www.ingenuity.com) was used to identify any enrichment of signaling and metabolic pathways associated with the preservation of the proteins in the LLR samples. The significance of the association between the dataset and the canonical pathway was measured by the fold-enrichment, *z*-score, and the significance (*p*-value). Gene ontologies were also explored using AmiGO 2 (http://amigo.geneontology.org/amigo), using the common list of proteins.[27]

Proteomic data are available via ProteomeXchange via the PRIDE partner repository[28] with the dataset identifier PXD009418.

2.7. Targeted Amelogenin Analysis Using Synthetic Surrogate Peptides for Confirmation

Amelogenin peptides identified during shotgun proteomics were used as a target for qualitative MRM. The identified peptides were common to all 3 X-isoforms described in the UniProt database. The WYQSIRPPYPSYG peptide target monitored 11 transitions (parent ion and product ion combinations to create a specific and unique signature), while the peptide WYQSIRPPYPSY, a shorter version of the same peptide also observed in some samples, monitored 15 transitions to identify the peptide. A specific Y-related isoform consisting of the sequence IALVLTPLK reflecting one of the two Y-specific isoforms (monitoring six transitions) present in the database, was assessed in four teeth. These were the only sex related amelogenin peptides able to be deciphered from the LLR dataset. Transitions are available in Supporting Information. Relative amounts of each peptide based on the retention times and ion specific abundances were analyzed using Skyline software v3.6.[29] Mean abundance, standard error was assessed using NCSS 9 software (NCSS, Utah, USA)[30]across three replicates were calculated. The presence or absence of peptides was gauged against a control female and control male tooth as well as synthetic versions of two peptides (Sigma-Aldrich, TX, USA).

2.8. Stable Isotope Analysis

Tooth samples were individually prepared following an optimized preparation procedure.[31] All further details followed the work of Ji et.al.[32] Comparative data for prehistoric samples from Sri Lanka, Bornean Sarawak, and Thailand were obtained from studies.[33–36]

3. Results

3.1. Authenticity of Data

Mitigation of contamination through sample collection, handling, and analysis are an essential element of paleoproteomic workflows because the conserved protein homology across species boundaries, as well as the nature and impact of diagenesis contrasting with the biological changes that precede mortem are difficult to establish. Workflows that incorporate controls, use multiple complementary techniques and check-stops through the project are vital in ascertaining authenticity.[37] Our summarized approach is detailed in **Figure 1**, with details and results of all approaches taken to confirm authenticity and control for contamination available in Table S1 and Figures S1–S3, Supporting Information.

3.2. Palaeoproteomics

Protein modification and degradation, with a reduced diversity of peptide ions were apparent in the MS profiles of the Iron-Age samples compared to modern samples (**Figure 2**). A frequent modification in peptides is the conversion of the naturally occurring L- to the D-enantiomer of aspartic acid following deimidation of asparagines.[38] Isomerization of amino acids does not alter the mass of a peptide and is therefore silent in the m/z plane, however, it does change the structural property and therefore it can be distinguished by a shift in retention time (RT).[39] This is also observable in Figure 2.

Correlating the RT and the mass to charge (m/z) properties of peptides using the *ProgenesisQI* alignment algorithm enabled the use of the control ions as an alignment "scaffold" to map common LLR peptides. We have termed this "ion-scaffolding." It works on the principle of using the MS1 mass of ions and the vector pattern of neighboring ions to align and co-detect common ions across samples. The utility of this "ion-scaffolding" approach was observed by Principal Component Analysis (PCA). PCA was performed on all ions of significance ($p < 0.05$), FDR <1%, power>0.8, for peptides generating MS/MS data; and showed that 80% of all differences could be mapped to the appropriate sample type; ie clustering of ions was observed for control and LLR samples (Figure S1, Supporting Information). Peptides common to both control and LLR samples were evaluated for differences in modification, amino acid length, and pathway enrichment. The number of proteins identified common to LLR and control groups equates to 212; with an adjusted FDR of 1%, with two unique peptides, and ion scores >34. Only these proteins were evaluated in further studies.

Gene ontology (GO) was mapped in Figure 2C and show just over 24% of identified proteins relate to the cellular processes (GO: 0009987) and 16% of proteins mapped to metabolic pro-

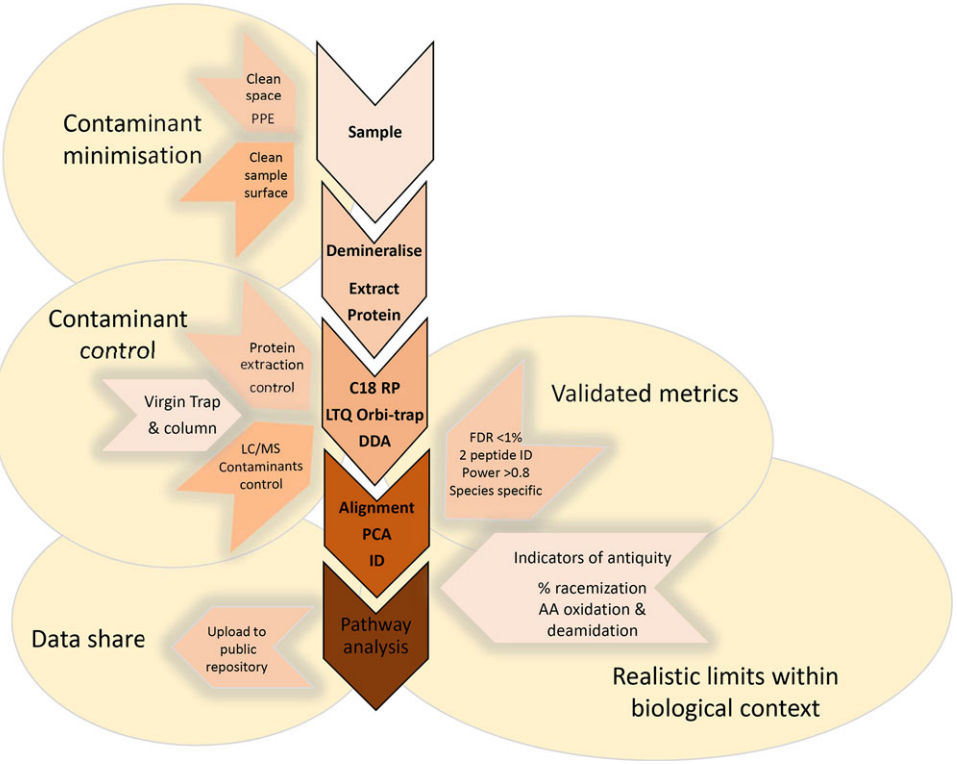


Figure 1. Sample management, protein extraction and data analysis workflow inclusive of contaminant minimization and control procedures.

cesses (GO:0008152). Pathway enrichment analysis showed one of the main networks found to be enriched in the LLR samples is related to metabolic disease (**Figure 3**). Cancer, immunological, and gastrointestinal diseases also were of significance in LLR samples (Figure 3A,B). Comparing *z*-score enriched biofunctions in the LLR and control samples highlighted an increase in proteins associated with cell death, mortality, and a concomitant decrease in cellular movement, signaling, and transport as well as transcription (Figure 3C). These are indications of a system in bio-shutdown at the onset of necrosis. Thus, the proteomic snapshot is one indicating the antemortem proteome with significant contributions of post-mortem activity. While we do not have the study power to examine in further detail disease at an individual level—this approach shows the potential for illness prior to death. An association with lifestyle and diet of the LLR individuals can be observed with a significant increase toward metabolic disorders and effects on carbohydrate metabolism. This association was explored further by assessing the results of stable isotopes (below).

A factor influencing proteome complexity can be related to the length of the identified peptides. The number of AA in the identified tryptic fragments were counted using only peptides that were present in both datasets with scores >34 and $p < 0.05$. The peptides were grouped into highest mean relative difference between LLR and control, and AA length was determined for all peptides in that grouping. **Figure 4**A shows the mean AA length of the unmodified peptides (Seen as the dark dashed line). When collagen peptides were analyzed separately, significant differences between LLR and control peptides were reached with a reduction in LLR peptide length with multiple modifications present ($p = 0.04$). Differences in deamidated peptide length were also observed between collagen and non-collagen peptides ($p = 0.005$), proline oxidation ($p = 0.03$) and multiple modifications ($p = 0.004$). Very few non-collagen peptides displayed proline oxidation and there was no significant diagenetic difference between LLR and control samples for collagen peptides ($p = 0.8$). In comparison, the human UniProt database contains the in silico tryptic digests of all known proteins, with the average length of peptides calculated to be ten amino acids. This in silico length is nearly equivalent to the 11 AA mean length of control teeth peptides, but significantly different when compared to modified peptides from LLR teeth (Calculated from 1 066 058 peptides in release 15 human Swiss-Prot database). These modifications should also be assessed in the context of the preserved amino acids profiles. The fidelity of amino acids within the tooth matrix is affected over time. We assessed the representation of each amino acid within the identified peptides. Peptides identified only in both datasets were plotted as percentage amino acid composition and normalized to control samples (see Figure S3, Supporting Information). The data shows an overrepresentation of the residues (N,E,H,P,F) in LLR samples (collagen removed). Figure 4B compares the properties of these preserved residues to the entire human database to observe preservation of particular physicochemical properties of amino acids in the ancient proteome. The results show that charged and polar amino acids have been retained in the LLR dataset.

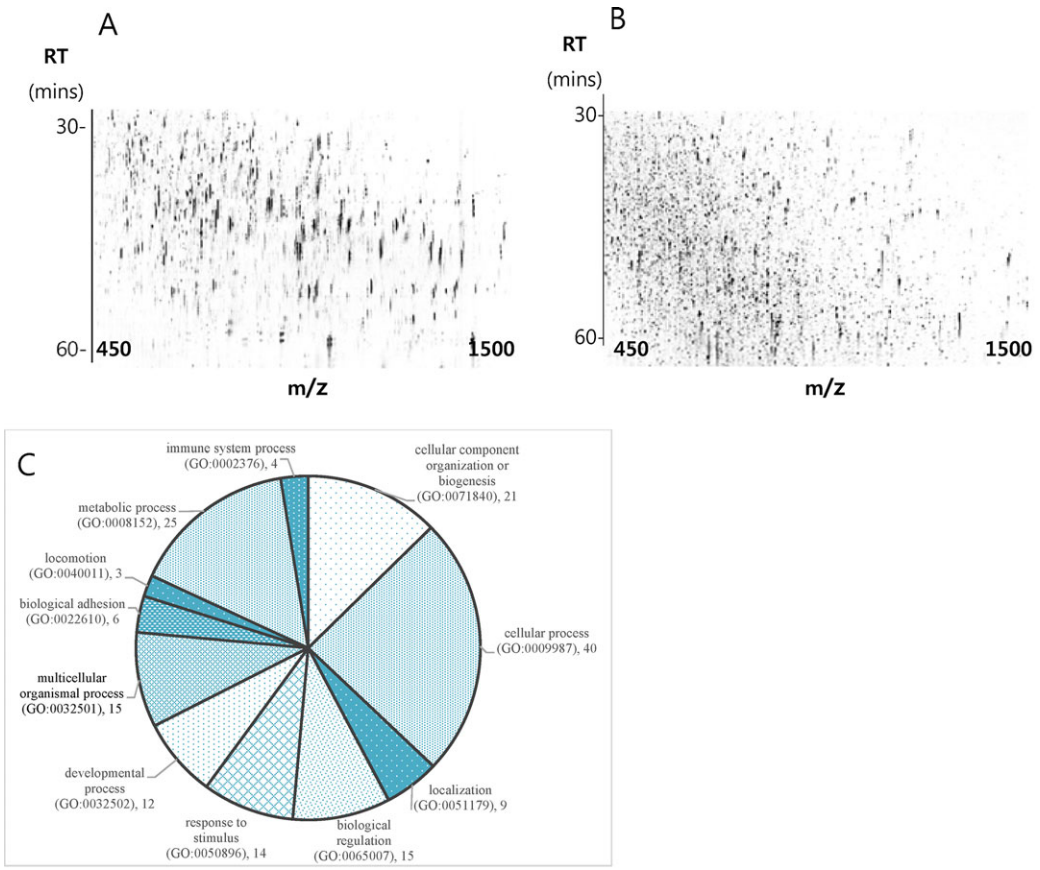


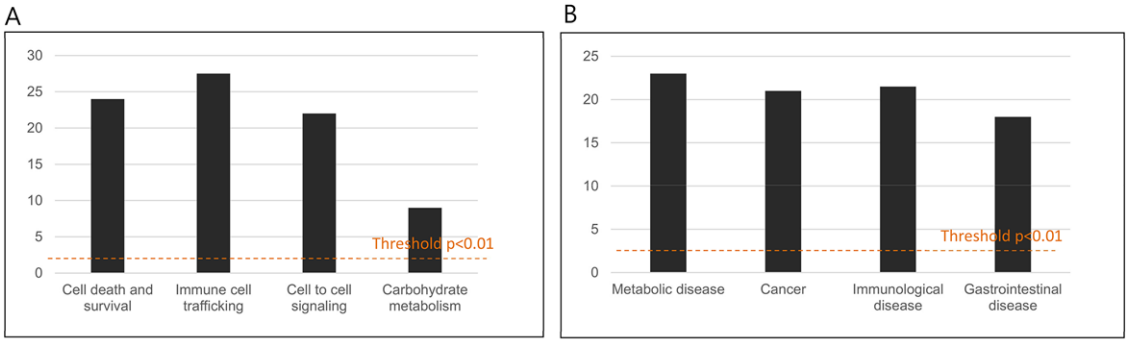
Figure 2. Representative ion-map regions showing equivalent protein amounts separated using reverse-phase ESI-MS/MS. A) LLR specimen, B) control specimen. This reveals a reduction in the overall diversity and number of peptide ions originating from LLR samples. Significant spreading of ions across retention time also indicate a greater degree of racemization. C) LLR gene-ontologies mapped according to biological processes.

3.3. Amelogenin Sexing

Amelogenin protein is expressed from both X and Y chromosomes with approximately 90% of transcripts dominated by the X-version with additional splicing and proteolytic processing creating a number of isoform variants.[40] There are three main X-related forms and two Y-related forms listed in the UniProt database. The ClustalW sequence alignment (Clustal.org) is given in Figure S4, Supporting Information along with the shotgun sequenced X related isoform and the synthetically generated sequence of both X and Y-related isoforms used as a standard for the amelogenin protein (**Figure 5**). This difference in sequence can be exploited for the determination of the sex of the LLR samples. Shotgun proteomics revealed variants from isoform 1 and 3 of the X-translated copy (Samples 2D, 2B, 6B, and 6D). Amelogenins constitute ≈90% of total enamel matrix proteins during the secretory stage of enamel formation and together they form the supramolecular framework supporting enamel crystal growth. Amelogenin is a 20 kDa hydrophobic protein, and along with enamelin (65 kDa acidic protein) and ameloblastin, they constitute the principal enamel matrix proteins (all three proteins were identified in the dataset). Comparison of the amelogenin isoforms to the average amino acid preservation in LLR (non-collagen) peptides shows consistent preservation of unmodified P and no significant alteration in the preservation of V, I, L, or S residues (Figure S3, Supporting Information) enabling its utility in distinguishing sex in these samples from a tropical environment. Targeted MRM on the two shot-gun identified peptide sequences and additionally a Y-specific peptide sequence was assessed on male and female control teeth as well as the LLR teeth seen in Figure 5. This qualitative analysis suggests all four ancient samples tested have evidence of the X-related peptides (TALVLTPLK and WYQSIRPPYPSY(G)), while the Y-related peptide (IALVLTPLK) was also present suggesting the four samples tested are likely to be male (Figure 5).

3.4. Stable Isotope Analysis

Stable carbon isotope ($\delta^{13}C$) values from carbonate provide an indication of an individual's whole diet during their lifetime.[41] Data could only be obtained for four of the six teeth investigated. Carbon stable isotope values vary between −14.41‰ and −13.19‰, with a small range of 1.22‰. These values indicate that the people buried at LLR rock shelter subsisted predominantly on C_3 food sources but with a reasonable proportion (≈20%) of their diet deriving from C_4 food sources. Further results of the stable isotope analysis is detailed in the Supporting Information.



C

Biofunction	Z-score
Cell movement	-4.84
Molecular transport	-2.36
Cell signaling	-2.17
Expression of RNA	-1.26
Transcription	-0.98
Death	5.74
Mortality	5.77

Figure 3. A) Canonical pathway analysis of LLR samples show one of the most enriched pathways are associated with cell death, with B) enrichment of proteins associating with biofunctions and disease of a metabolic nature. The Threshold represents a significance of $p = 0.01$. C) Effects analysis of functions with z-score ≥ 2 or ≤ -2 show the greatest effect was observed for pathways and biofunctions related to death and mortality for LLR teeth compared to control teeth with decreases observed in cell signaling and movement and reduction in transcription at the time of this proteomic "snap-shot."

4. Discussion

Despite their tropical burial and geological age of 1960 ± 30 cal. yr BP to 1636 ± 44 cal. yr BP the LLR rock shelter teeth contained 212 proteins identifiable through high-resolution, shot-gun mass spectrometry. From them we have been able to construct a proteome from LLR humans inhabiting a tropical environment. The observable proteome of the LLR individuals is a palimpsest of proteins resulting from contributions from molecular pathways active just prior to death and networks operating during post-mortem bio-shutdown over immediate time-scales. These were followed by chemical degradation and hydrolysis events, modifications and peptide loss which occurred over thousands of years, processes associated with diagenesis. Despite the sequential nature of these events, the contribution of each of them to the proteomes recovered from the LLR rock shelter teeth is difficult to establish. Inferring conclusions across geological time-scales empirically remains, therefore, a major challenge for palaeoproteomic analyses.

Collagen, the most abundant protein present in teeth, begins to breakdown and decay in vivo by increasing the number of cross-links as individuals age ontogenetically.[42] Collagen and its breakdown products occur with death, while various physical and chemical alterations of osteological samples occur as tissues interact with various elements of their burial environment (diagenesis). Autolysis is the first such postmortem process and involves the body's own enzymes causing cellular destruction. Putrefaction begins soon thereafter. Teeth are less susceptible to structural decay post-mortem but are affected by diagenesis.[42] Physicochemical alterations of proteins follow decay, with dissolution, especially in acidic conditions, causing cortical bone to breakdown and exposing the organic content of teeth and bones to destruction of peptide fidelity. A bacterial component to this process is also present which is normally accelerated in warmer climates and results in rapid decline in peptide fidelity of susceptible proteins and a concomitant reduction in proteomic complexity.

The preservation of sequence and retardation of diagenesis, is seen for peptides with a greater number of N, E, H, P, and F amino acids. Our data show that protein preservation over longer time-scales may be advantaged by the presence of aromatic (F), positively charged (H) and negatively charged (E) polar amino acid residues, while polar residues (D, K, T) are less likely to persist in non-collagenous proteins. This is consistent with published studies which demonstrate a lack of C and Y amino acids following the amino acid asparagine (N); a residue particularly prone to deamination events[39,43]; while nonenzymatic glycation of K residues and reduced solubility were observed in teeth[44] and have been hypothesized to affect preservation of ancient proteins[45] . The presence of the small, nonpolar, amino acid proline results in increased proline oxidation to hydroxyproline in collagen peptides, however its presence in non-collagenous peptides is typically in its unmodified form. The hydroxylation of proline increases the peptides conformational stability, as does the presence of the small hydrophobic amino acid glycine.[46] It is likely that the preservation of these residues is contributed to by the preservation of protein structure in collagenous proteins.

The utility of correlating the RT and m/z properties of LLR peptides to that of a modern control sample in a method we have termed "ion-scaffolding" is demonstrated. This approach has allowed for the identification of 212 proteins which were further

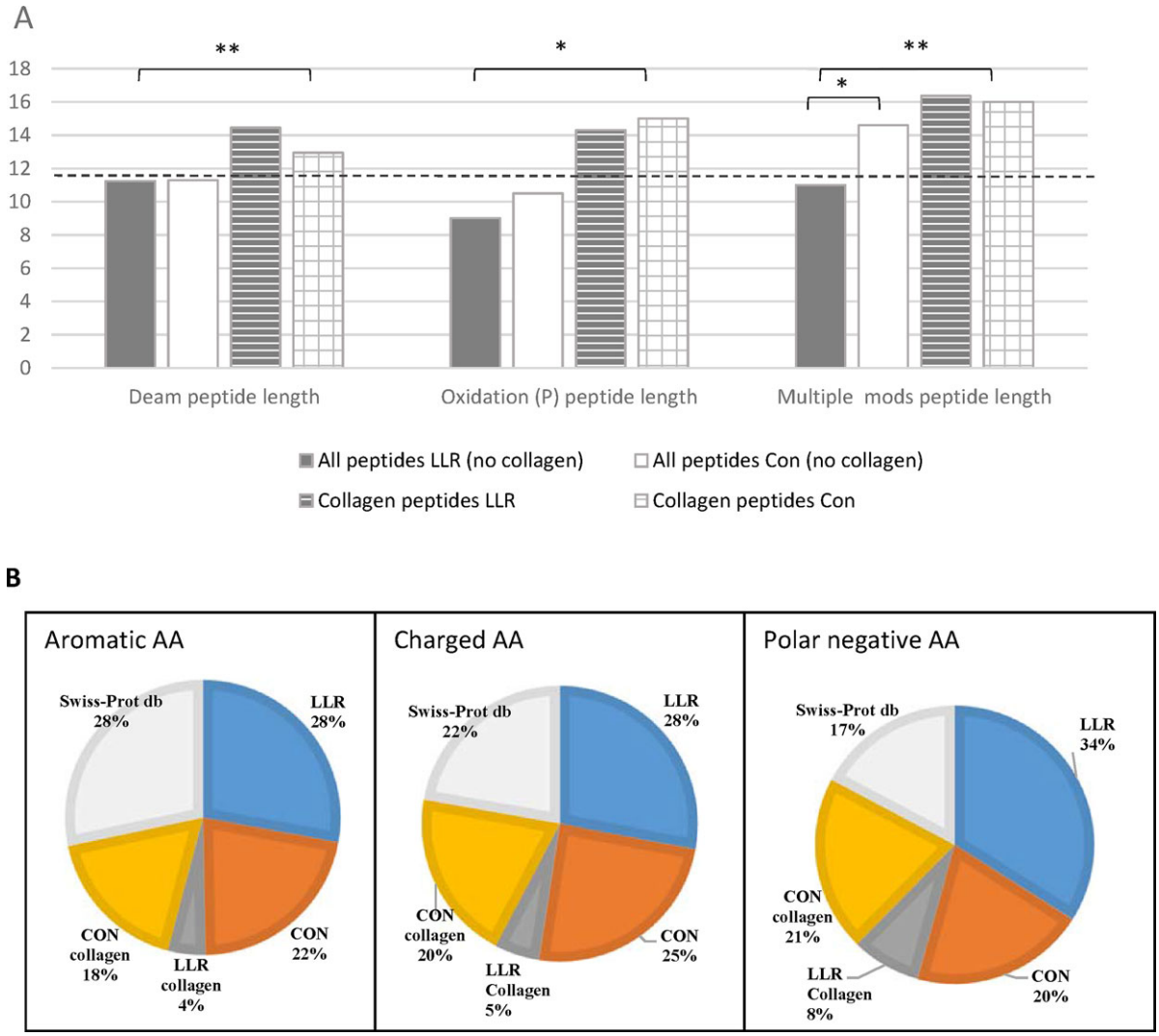


Figure 4. A) The mean amino acid length of (confidence score > 34, $p < 0.05$, 2 peptide identification of proteins, FDR < 1%) is shown for both collagen and non-collagen peptides. The graph shows the mean amino acid length decreases in LLR samples for peptides containing multiple modifications. * denotes significance <0.05; ** significance <0.005. B) pie shows the over represented amino acid classes as percentage of all detected amino acids for the LLR sample set in comparison to control and the human Swiss-Prot dataset.

evaluated for changes in the proteomes. Peptides common to both data sets were evaluated for differences in modifications, length, and pathway enrichment. The premise that a reduction in proteome complexity has resulted from peptide bond cleavage resulting in shorter length peptides is not supported by the LLR data with no significant difference reached between unmodified peptides. However, multiple modifications on peptides was correlated with a mean reduction in peptide length ($p = 0.04$). We have shown an association between the presence of particular amino acids and the preservation potential of these peptides within the protein sequence, while other peptide sequence stretches have been lost over time, as only peptide sequences present in both datasets were accounted for using this ion scaffolding method. It is interesting to note that the relative differences in protein concentrations should reflect the difference in protein regulation and expression between the LLR and modern samples and this likely will also have a functional impact on enriched pathways. Differences in the proteins seen here to determine antemortem pathologies are of a much more transient nature when compared to the autolysis and degradation and diagenetic processes that are occurring over a much longer timescale. So, while the pathologies and evolution of ancient diseases are of immense importance to understanding modern day diseases, any pathways determined here to indicate modern day pathologies are only suggestive of the potential to harbor disease (due to the low power of this study). However, if the LLR proteomes have the potential to harbor disease, then we can extend this to the potential that some in the population may have succumb to these diseases. There is a likelihood that metabolic diseases were present in this population as supported by secondary information using stable isotope analysis of diets which reflected a diet based broadly (80%) on plants such as rice.

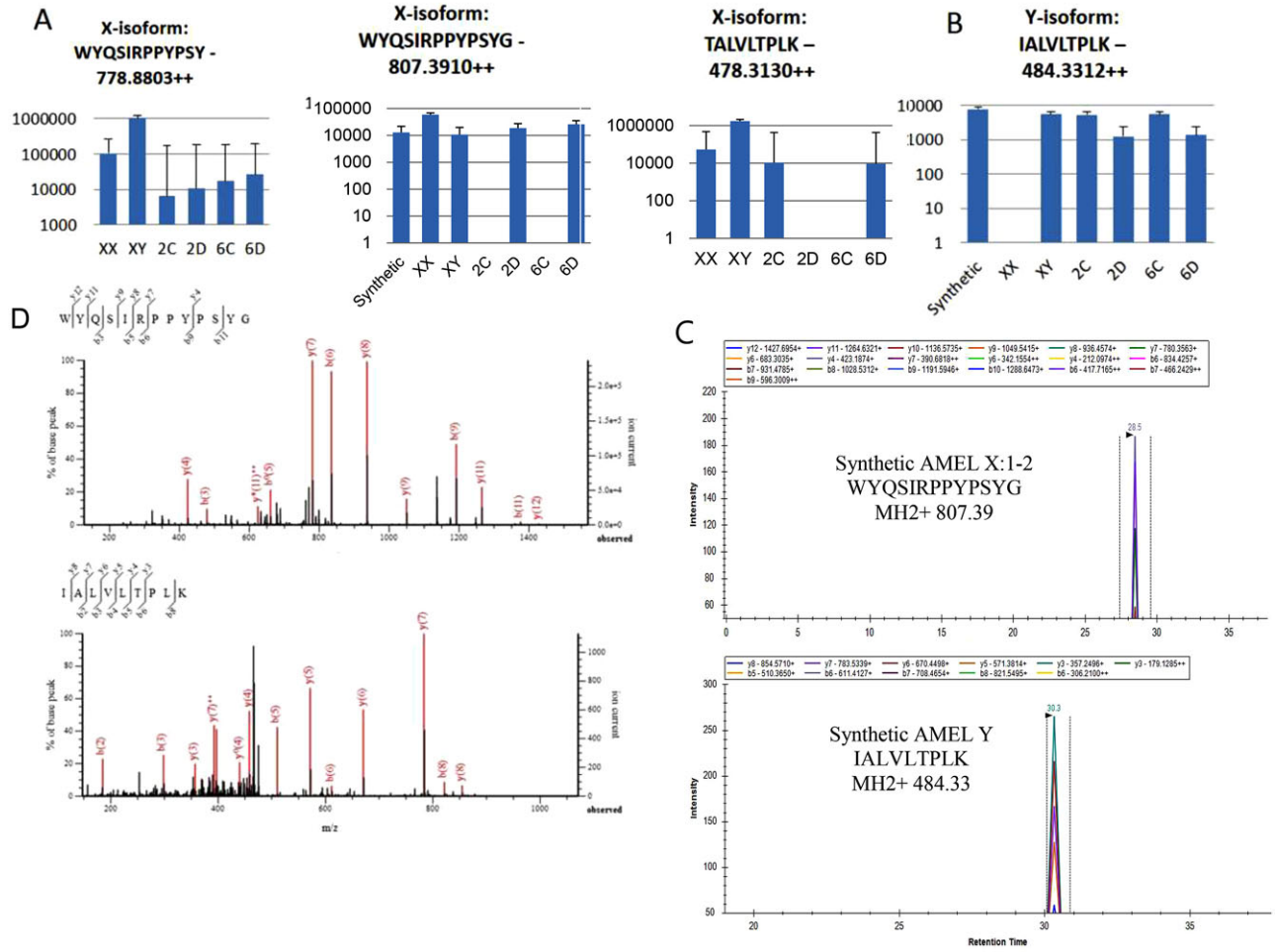


Figure 5. Amelogenin MRM analysis showing three peptides from A) X-specific isoform and B) one peptide from the Y-specific isoform; C) synthetic peptide targeted MRM's showing daughter ion signatures. D) The MS profile is also provided for both peptides.

We note the presence of proteins in the LLR proteomes that directly interact in pathways of inflammation, metabolic, and cancer related pathologies (Figure 3). In this present context, the consequence is likely to also be influenced by apoptosis of aberrant cells and concomitant attempts to stimulate cell proliferation and development while proper immune recognition of pathogens nor immune response is fully operational. It demonstrates a snap-shot of a proteome with significant contributions post-mortem possibly as a result of LLR burial practices. Pozhitkov and colleagues[47] have recently shown that some genes are amplified well after death, and this should beneficially spur inflammation, activate the immune system, counteract cellular stress and hypoxia and additionally, re-awakens developmental genes. A number of cancer related genes are also amplified.[47] In the samples from LLR we have shown the same pathways to be enriched. While Pozhitkov et al. could demonstrate gene activity through DNA microarrays and mRNA over a 96 h period post-mortem, our data demonstrate that: 1) effector molecules, not just upstream messengers (mRNA) are transcribed and translated despite a decrease in the transcriptional processes (Figure 3C), 2) many of these proteins reflect pathways of immune shutdown and carcinogenic up regulation (Figure 3A), 3) the LLR proteome is a snap-shot of protein activity of both ante- and post-mortem protein activity, and 4) this activity can be measured in geological time-scales even for samples from tropical environments.

Of interest was the identification of peptides from the protein amelogenin. The LLR rockshelter samples contained peptides representative of isoforms 1 and 3 from X-linked gene, and isoform 2 from the Y-linked gene. MRM was used to confirm the findings, with specimen labeled 2C, 2D, 6C, and 6D supporting evidence of male origin. Samples labeled 2B and 6B had insufficient MRM evidence to determine sex conclusively. As the amelogenin protein is made up of contributions from both parents; there can be a number of variations of X-isoforms (from the X-gene of either parent for female offspring). Male offspring will have X-related isoforms from the mothers X-related gene, while only inheriting the Y-isoform from the Y-related sex chromosome of the male parent). A recent study has shown the applicability of amelogenin sexing accuracy in historic samples.[9] Here, sex determination of archaeological samples was compared to presence or absence of transitions observed relative to control samples and synthetic peptide standards. All samples

Table 1. Comparison of Long Long Rak results with samples from prehistoric Thailand.

Sample [*n*][a)]	$\delta^{13}C_{VPDB}$[b)]	*p*-value of LLR versus prehistoric Thailand
LLR-IA (8)	−13.86 ± 0.46 (−14.4, −13.19)	—
BNW-NL (23)	−11.94 ± 1.74 (−13.63, −8.18)	**2.57e-5**
BNW-BA (76)	−13.20 ± 0.35 (−14.19, −12.13)	0.17
BNW-IA (18)	−12.39 ± 1.21 (−13.32, −8.19)	**2.63e-5**
BLK-BA (26)	−14.01 ± 0.43 (−15.10, −13.10)	0.99
NUL-IA (34)	−13.78 ± 0.38 (−14.53, −12.91)	0.99
KPD-EP (66)	−12.29 ± 1.09 (−14.10, −6.10)	**2.57e-5**

[a)] LLR-IA = Long Long Rak-Iron Age; BNW-NL = Ban Non Wat-Neolithic; BNW-BA = Ban Non Wat-Bronze Age; BNW-IA = Ban Non Wat-Iron Age; BLK-BA = Ban Lum Khao-Iron Age; NUL = Noen U-Loke-Iron Age; KPD-EP = Kok Phanom Di-Epipalaeolithic.; [b)] Mean ± sd (min.,max.)

displayed presence of various X-related isoforms, however female control samples did not display any signature transitions of the Y-related isoform. The hydrophobic nature of these peptides also supports their persistence over geological timescales and indicates their significant potential use for sexing ancient samples in future studies.

The results of ANOVA with Tukey's post hoc tests comparing the LLR mean $\delta^{13}C$ values with means for various prehistoric samples from Thailand are provided in **Table 1**. Mean differences are significant when compared with the Ban Non Wat Neolithic and Iron Age samples from the Mun River Valley of northeastern Thailand and the Kok Phnom Di Epipalaeolithic sample from the coastal plain of eastern Thailand. However, the mean $\delta^{13}C$ value for LLR is essentially indistinguishable from samples of Bronze Age skeletons at Ban Non Wat and Ban Lum Khao Bronze Age and Iron Age individuals Noen U-Loke, all three sites also belonging within the Mun River Valley (Table 1). This suggests that the people buried at LLR rock shelter had a diet that focused heavily on cultivated rice with some protein probably from freshwater fish, as is the case with these other three samples,[34,35] whereas the Neolithic and Iron Age inhabitants of Ban Non Wat and the Epipalaeolithic people living at Kok Phnom Di had much more diverse diets that were depleted in C_3 food sources.[34,36] This is reflective of not only cultural differences but also environmental and geographical constraints.

As diet is a leading modifiable risk factor for disease[48] complementing stable isotope data with proteomic profiles of a population provides a unique framework to asses diet over natural lifespans. The combination of proteomics and stable isotope analysis provides a new strategy for assessing various aspects of the biology, diet, and lifestyle of ancient populations including how they might differ between the sexes. It presents a novel way to establish the broad nature of dietary capacity and to infer environment and lifestyle trends in populations. Heritable diseases at a population level also have the potential to be identified. In the present study, we have found a possible association between lifestyle and diet (excessive consumption of carbohydrates) of the LLR individuals observed as a significant increase toward metabolic disorders and effects on lipid and carbohydrate metabolism. Several challenges remain, however, including distinguishing protein enrichment of molecular pathways associated antemortem versus post-mortem processes and events, more broadly allowing us to link them with lifestyle and dietary behaviors identified through archaeological and stable isotope evidence with greater certainty.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

R.S. and K.C. excavated and curated the samples. V.W., D.C., L.A., S.B., R.M., A.B., and C.B. performed the analyses. D.C, A.B., and V.W. designed the study. V.W. and D.C. wrote the paper with inputs from L.A, S.B, R.M, A.B., C.B., R.S., K.C., and P.S.C.T. This research was funded by Australian Research Council grants LP120200144 (awarded to DC and PT), FT120100168 (awarded to D.C.) and a PANGEA Research Centre grant (awarded to V.W. and D.C.). Archaeological research at Long Long Rak was funded by a Thailand Research Fund (TRF) grants RTA6080001 and RDG55H0006 (awarded to R.S. and K.C.). Amelogenin analysis was funded by Mark Wainwright Analytical Centre Grant (awarded to V.W., D.C., R.M.). The authors thank Mark Wilkins and Simone Li for assistance in calculating in silico human proteome in the Swiss-Prot database. The authors acknowledge the tooth fairy for the kind donation of control teeth.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

amino acids, diet, MRM, sex diagnostic, tooth proteome

Received: September 13, 2018
Revised: January 18, 2019
Published online:

[1] a) M. Slatkin, F. Racimo, Proc. Natl. Acad. Sci. **2016**, *113*, 6380; b) J. K. Pickrell, D. Reich, *Trends Genet.* **2014**, *30*, 377.
[2] a) S. Sawyer, G. Renaud, B. Viola, J. J. Hublin, M. T. Gansauge, M. V. Shunkov, A. P. Derevianko, K. Prüfer, J. Kelso, S. Pääbo, *Proc. Natl. Acad. Sci. U S A* **2015**, *112*, 15696; b) K. Prüfer, F. Racimo, N. Patter-

son, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. De Filippo, H. Li, *Nature* **2014**, *505*, 43.
[3] a) M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, H. A. Burbano, C. Lalueza-Fox, M. de La Rasilla, A. Rosas, *Nature* **2016**, *530*, 429; b) Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, *Nature* **2015**, *524*, 216.
[4] a) B. Demarchi, S. Hall, T. Roncal-Herrero, C. L. Freeman, J. Woolley, M. K. Crisp, J. Wilson, A. Fotakis, R. Fischer, B. M. Kessler, R. R. Jersie-Christensen, *ELife* **2016**, *5*; b) E. Cappellini, M. J. Collins, M. T. P. Gilbert, *Science* **2014**, *343*, 1320.
[5] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, M. L. Tress, *Hum. Mol. Genet.* **2014**, *23*, 5866.
[6] V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, I. Humphery-Smith, *Electrophoresis* **1995**, *16*, 1090.
[7] C. M. Bauer, H. Niederstaetter, G. McGlynn, H. Stadler, W. Parson, *Forensic Sci. Int.* **2013**, *7*, 581.
[8] K. Krishan, P. M. Chatterjee, T. Kanchan, S. Kaur, N. Baryah, R. K. Singh, *Forensic Sci. Int.* **2016**, *261*, 165.e1.
[9] N. A. Stewart, R. F. Gerlach, R. L. Gowland, K. J. Gron, J. Montgomery, *PNAS* **2017**, *114*, 13649.
[10] M. H. Schweitzer, W. Zheng, T. P. Cleland, M. Bern, *Bone* **2013**, *52*, 414.
[11] M. Buckley, M. Collins, J. Thomas-Oates, *Anal. Biochem.* **2007**, *374*, 325.
[12] F. Welker, M. Hajdinjak, S. Talamo, K. Jaouen, M. Dannemann, F. David, M. Julien, M. Meyer, J. Kelso, I. Barnes, S. Brace, *Proc. Natl. Acad. Sci. U S A* **2016**, *113*, 201605834.
[13] S. Brown, T. Higham, V. Slon, S. Pääbo, M. Meyer, K. Douka, F. Brock, D. Comeskey, N. Procopio, M. Shunkov, A. Derevianko, M. Buckley, *Sci. Rep.* **2016**, *6*, 23559.
[14] R. Sawafuji, E. Cappellini, T. Nagaoka, A. K. Fotakis, R. R. Jersie-Christensen, J. V. Olsen, K. Hirata, S. Ueda, *R. Soc. Open Sci.* **2017**, *4*, 161004.
[15] C. Warinner, J. F. M. Rodrigues, R. Vyas, C. Trachsel, N. Shved, J. Grossmann, A. Radini, Y. Hancock, R. Y. Tito, S. Fiddyment, C. Speller, *Nature Genet.* **2014**, *46*, 336.
[16] M. Mackie, J. R. Hendy, A. D. Lowe, A. Sperduti, M. R. Holst, M. J. Collins, C. F. Speller, *STAR* **2017**, *1*.
[17] A. Corthals, A. Koller, D. W. Martin, R. Rieger, E. I. Chen, M. Bernaski, G. Recagno, L. M. Dávalos, *PLoS One* **2012**, *7*, e41244.
[18] J. Jones, M. Mirzaei, P. Ravishankar, D. Xavier, D. H. Shin, R. Bianucci, P. A. Haynes, *Phil. Trans. R. Soc. A* **2016**, *374*, 20150373.
[19] R. Barbieri, R. Mekni, A. Levasseur, E. Chabrierè, M. Signoli, S. Tzortzis, G. Aboudharam, M. Drancourt, *PLoS One* **2017**, *12*, e0180552.
[20] R. Shoocongdej, Bangkok: Charansanitwong Printing. (In Thai) **2016**.
[21] N. Pumijumnong, S. Wannasri, *Teak Log Coffins in Northwest Thailand: Dated by Dendrochronology and 14C wiggle Matching.*, Vol. *37*, Somerset, UK **2017**.
[22] M. Jágr, A. Eckhardt, S. Pataridis, I. Mikšík, *Eur. J. Oral Sci.* **2012**, 120, *259*.
[23] R. C. Hill, M. J. Wither, T. Nemkov, A. Barrett, A. D'Alessandro, M. Dzieciatkowska, K. C. Hansen, *Mol. Cell Biol.* **2015**, *14*, 1946.
[24] E. Cappellini, L. J. Jensen, D. Szklarczyk, A. Ginolhac, R. A. da Fonseca, T. W. Stafford, S. R. Holen, M. J. Collins, L. Orlando, E. Willerslev, M. T. Gilbert, J. V. Olsen, *J. Proteome Res.* **2012**, *11*, 917.
[25] J. Beretov, V. C. Wasinger, E. K. A., Miller, P. Schwartz, P. H. Graham, Y. Li, *PLoS One* **2015**, *10*, e0141876.
[26] a) J. D. Storey, R. Tibshirani, *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 9440; b) N. A. Karp, P. S. McCormick, M. R. Russell, K. S. Lilley, *Mol. Cell. Proteomics* **2007**, *6*, 1354.
[27] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, P. D. Thomas, *NAR* **2017**, *45*, 183.
[28] E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D. S. Campbell, M. Bernal-Llinares, S. Okuda, S. Kawano, R. L. Moritz, J. J. Carver, M. Wang, Y. Ishihama, N. Bandeira, H. Hermjakob, J. A. Vizcaíno, *NAR* **2017**, *45*, D1100.
[29] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, M. J. MacCoss, *Bioinformatics* **2010**, *26*, 966.
[30] J. Hintze, in https://doi.org/www.ncss.com, **2013**.
[31] a) P. L. Koch, N. Tuross, M. L. Fogel, *J. Archaeol. Sci.* **1997**, *24*, 417; b) S. J. Garvie-Lok, T. L. Varney, M. A. Katzenberg, *J. Archaeol. Sci.* **2004**, *31*, 763; c) B. E. Crowley, P. V. Wheatley, *Chem. Geol.* **2014**, *381*, 234.
[32] X. Ji, D. Curnoe, P. S. C. Taçon, B. Zhende, L. Ren, R. Mendoza, H. Tong, J. Ge, C. Deng, L. Adler, A. Baker, B. Du, *J. Archeol. Sc. Reports.* **2016**, *8*, 277.
[33] P. Roberts, N. Perera, O. Wedage, S. Deraniyagala, J. Perera, S. Eregama, A. Gledhill, M. D. Petraglia, J. A. Lee-Thorp, *Science* **2015**, *347*, 1246.
[34] a) C. L. King, R. A. Bentley, N. Tayles, U. S. Viðarsdóttir, G. Nowell, C. G. Macpherson, *J. Archaeological Sci.* **2013**, *40*, 1681; b) R. A. Bentley, K. Cox, N. Tayles, C. Higham, C. Macpherson, G. Nowell, M. Cooper, T. E. Hayes, *Asian Perspect.* **2009**, *48*, 79.
[35] K. J. Cox, R. A. Bentley, N. Tayles, H. R. Buckley, C. G. Macpherson, M. J. Cooper, *J. Archaeological Sci.* **2011**, *38*, 665.
[36] R. A. Bentley, N. Tayles, C. Higham, C. Macpherson, T. C. Atkinson, *Curr. Anthropol.* **2007**, *48*, 301.
[37] J. Hendy, F. Welker, B. Demarchi, C. Speller, C. Warinner, M. J. Collins, *Nat. Ecol. Evol.* **2018**, *2*, 791.
[38] a) P. Helfman, J. Bada, *Proc. Natl. Acad. Sci. U S A* **1975**, *72*, 2819; b) J. L. Bada, *Ann. Rev. Earth Planet. Sci.* **1985**, *13*, 241.
[39] H. Yang, R. A. Zubarev, *Electrophoresis* **2010**, *31*, 1764.
[40] E. C. Salido, P. H. Yen, K. Koprivnikar, L. C. Yu, L. J. Shapiro, *Am. J. Hum. Genet.* **1992**, *50*, 303.
[41] S. Jim, S. H. Ambrose, R. P. Evershed, *Geochimica et Cosmochimica Acta* **2004**, *68*, 61.
[42] M. Buckley, C. Wadsworth, *Aleogeography Paleoclimatol. Paleoecol.* **2014**, *416*, 69.
[43] B. Demarchi, M. J. Collins, P. J. Tomiak, B. J. Davies, K. E. H. Penkmana, *Quat. Geochronol.* **2013**, *16*, 158.
[44] Y. Matsuda, J. Miura, M. Shimizu, T. Aoki, M. Kubo, S. Fukushima, M. Hashimoto, F. Takeshige, T. Araki, *J. Dent. Res.* **2016**, *95*, 1528.
[45] T. P. Cleland, E. R. Schroeter, M. H. Schweitzer, *Proc. Natl. Acad. Sci. U S A* **2015**, *282*, 20150015.
[46] S. R. Trevino, S. Schaefer, J. M. Scholtz, C. N. Pace, *J. Mol. Biol.* **2007**, *373*, 211.
[47] A. E. Pozhitkov, R. Neme, T. Domazet-Lošo, B. G. Leroux, S. Soni, D. Tautz, P. A. Noble, *Open Biol.* **2017**, *7*.
[48] D. O'Brien, *Annu. Rev. Nutr.* **2015**, *35*, 565.