Sex estimation using sexually dimorphic amelogenin protein fragments in human enamel

Glendon J. Parker, Julia M. Yip, Jelmer W. Eerkens, Michelle Salemi, Blythe Durbin-Johnson, Caleb Kiesow, Randall Haas, Jane E. Buikstra, Haagen Klaus, Laura A. Regan, David M. Rocke, Brett S. Phinney

Keywords: Sex estimation, Teeth, Enamel, Amelogenin, Proteomics, Mass spectrometry, Paleoproteomics

ABSTRACT

Amelogenin genes are located on both X and Y sex chromosomes in humans and are a major focus of DNA-based sex estimation methods. Amelogenin proteins, AMELX_HUMAN and AMELY_HUMAN, are expressed in the tooth organ and play a major role in mineralization of enamel, the most taphonomically resistant, archaeologically persistent human tissue. We describe shotgun liquid chromatography mass spectrometry analysis of 40 enamel samples representing 25 individuals, including modern third molars and archaeological teeth from open-air contexts including permanent adult (400 to 7300 BP) and deciduous teeth (100 to 1000 BP). Peptides specific to the X-chromosome isoform of amelogenin were detected in all samples. Peptides specific to the sexually dimorphic Y-chromosome isoform were also detected in 26 samples from 13 individuals, across all time periods, including previously unsexed deciduous teeth from archaeological contexts. While the signal of each gene product can vary by more than an order of magnitude, we show close agreement between osteological and amelogenin-based sex estimation and thus demonstrate that the protein-based signal can be obtained reliably from open-air archaeological contexts dating to at least 7300 years ago. While samples with AMELY_HUMAN peptides are unambiguously male, samples with no AMELY_HUMAN signal may either be low signal male false negatives or female samples. In order to estimate sex in these samples we developed a probability curve of female sex as a function of the logarithm of AMELX_HUMAN signal (p < 0.0001) using logistic regression. This is also the first demonstration using proteomics to estimate sex in deciduous teeth and pushes back the application of the method to teeth that are at least 7300 years old.

1. Introduction

Sex estimation represents one of the essential osteological variables documented in virtually any study of human remains. Sex estimation is fundamental in the bioarchaeological assessment of ancient demographic variation, paleoepidemiological reconstructions of ancient disease patterns, and differential diagnosis in paleopathology (Buikstra and Ubelaker, 1994; White et al., 2012). Currently, sex estimation relies on two methods: analysis of sexually dimorphic osteological features and detection of DNA markers that are specific to the X- and Y-chromosome (Krishan et al., 2016; McFadden and Oxenham, 2016; Mittnik et al., 2016; Phenic, 1969; Skoglund et al., 2013; Spradley and Jantz, 2011). Though a range of skeletal elements have sexually dimorphic sex estimation and thus demonstrate that the protein-based signal can be obtained reliably from open-air archaeological contexts dating to at least 7300 years ago. While samples with AMELY_HUMAN peptides are unambiguously male, samples with no AMELY_HUMAN signal may either be low signal male false negatives or female samples. In order to estimate sex in these samples we developed a probability curve of female sex as a function of the logarithm of AMELX_HUMAN signal (p < 0.0001) using logistic regression. This is also the first demonstration using proteomics to estimate sex in deciduous teeth and pushes back the application of the method to teeth that are at least 7300 years old.
2. Methods

2.1. Sample procurement and processing

Three enamel sample types were used in this study. First, modern third molars from individuals self-identified as male or female were collected as part of a study that sought, to document carbon, oxygen, strontium and lead isotope signatures across the continental United States in order to prove provenance human remains (Regan, 2006). The subjects were all born between 1964 and 1987 and were procured between 2005 and 2006. This initial study was determined to be exempt by the overseeing institutional review board (HQ USAFA IRB FAC2005026H) and found to be research, not human subject research, by the University of California - Davis institutional review board (IRB 79755-2). Second, archaeological teeth, primarily second and third molars, were collected from individuals with osteological sex estimations. These samples were from open-air contexts in central California and highland Peru, and range in age from 400 to at least 7300 BP (Eerkens et al., 2014; Eerkens et al., 2016a; Eerkens et al., 2016b; Eerkens et al., 2017; Greenwald et al., 2016; Haas et al., 2017; Haas and Viviano Llave, 2015). Third, decedent teeth completely lacking in contextual sex information were also analyzed. These include three teeth from Mississippian sites (~1000 BP; provided by the Center of American Archeology), and one deciduous tooth from an historic Euroamerican burial ca. AD 1850 (Table 1). Enamel was dissected in 20 - 35 mg blocks using a diamond coated dental wheel (Brasseler USA, 9188.11.180 DS DIAM DISC) with an NSK Ultimate XL Micromotor. Where possible the dissected dental blocks included the dentin enamel junction (Mitsiadis et al., 2014).
hydrochloric acid was added to each 2 mL sample vial with seven 2.8 mM ceramic beads (Omini-International Inc.), milled for 3 min at 7000 rpm in a MagnaLyzer (Roche Inc.), then centrifuged for 5 min at 16000 g. To reduce soluble proteins, 6 μL of 0.5M dithioerythritol (DTE) was added to each sample vial and incubated in the dark at 25 °C for 60 min. The carbamidomethylation was quenched by adding 12 μL of 0.5M iodoacetamide to the sample vials and incubated at room temperature for 30 min. Alkylation was performed by adding 12 μL of 0.5M iodoacetamide to the sample vials and incubated in the dark at 25 °C for 60 min. The carboxymethylation reaction was quenched by adding 12 μL of 0.5M DTE to the sample vials and incubated in room temperature for 5 min. After incubation, 0.01% ProteaseMAX (6 μL of 0.5% w/v, Promega Inc.) was added to the sample vials along with mass spectrometry grade trypsin (1 μL of 0.5 μg/μL, Thermo Pierce Inc.). Each sample was incubated at room temperature for 20 h at 300 rpm. After incubation, the sample vials were centrifuged for 5 min and 200 μL of the supernatant transferred to 0.22 μm centrifugal filters and centrifuged for 30 min. The filtrate was then transferred to clean Eppendorf® Protein LoBind tubes for sample clean up. Organic contaminants in aqueous stocks and solutions were removed by prior passage over solid phase extraction (SepPak, tC18, Waters Inc.).

2.3. Data acquisition

Digested peptides were desalted and concentrated using ZipTip C18 pipette tips (Millipore Inc.) with the eluted material lyophilized and resuspended in 2% (v/v) acetonitrile and 0.1% (v/v) TFA. The peptide concentration was measured using the Pierce™ Quantitative Fluorometric Peptide Assay (Thermo Pierce™) and 1 μg of peptide, or 40% of the total sample if the sample was too diluted, was applied to mass spectrometry. The sample was applied to LC-MS/MS on a Thermo Scientific Q Exactive Plus Orbitrap mass spectrometer in conjunction with a Proxeon Easy-nLC II HPLC (Thermo Scientific) and Proxeon nanospray source. The digested peptides were loaded on a 100 μm x 25 mm Magic C18 100 Å 5U reverse phase trap where they were desalted online before being separated using a 75 μm x 150 mm Magic C18 200 Å 3U reverse phase column. Peptides were eluted using a 65 min gradient with a flow rate of 300 nL/min. An MS survey scan was obtained for the m/z range 300 - 1600, MS/MS spectra were acquired using a method that incorporated an inclusion list of 28 ions (Table A.1) that were subjected to HCD (High Energy Collisional Dissociation). When inclusion list ions were not found, MS/MS was done on other ions in the MS survey scan. An isolation mass window of 1.6 m/z was used for precursor ion selection, and normalized collision energy of 27% used for fragmentation. A 5 s duration was used for the dynamic exclusion. Washes were applied between each sample. After 10 samples a blank run of bovine serum albumin standards was applied to test for sample-to-sample contamination. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009781 (http://www.proteomexchange.org) (Deutsch et al., 2017).

2.4. Bioinformatic processing

Mass spectrometry datasets (.RAW format) were processed using PEAKS™ (version 8.5) peptide matching software (Bioinformatics Solutions Inc., Waterloo, ON) (Zhang et al., 2012). The FASTA formatted UNIPROT Homo sapiens reference protein database (http://www.uniprot.org/proteomes/UP00005640) was modified to include additional FASTA protein entries of splice variants associated with AMELY_HUMAN (Q99217-1, −2, −3) and AMELY_HUMAN (Q99218-1, −2) gene products (Salido et al., 1992; Simmer, 1995). The reference database was further modified to incorporate a decoy database and was validated in PEAKs™ Software (Zhang et al., 2012). Peptide matching spectral assignment was conducted using default conditions with the following exceptions: error tolerance, precursor mass = 15 ppm, fragment ion = 0.5 Da, cleavage with trypsin with up to 2 missed cleavages, and up to two non-specific cleavages. The algorithm assumed all cytoines were carbamidomethylated, and the peptide was partially modified by deamidation (NQ), oxidation (MHW), pyroglutamate conversion from glutamate and glutamine, and methionine dioxidation. All peptide assignments were filtered by a 1% false discovery rate. Each peptide was quantified by summing the ion peak intensity of each primary precursor mass (MS1).
AMELX_HUMAN and AMELY_HUMAN protein were quantified using label free quantitation by combining the peak ion intensity measurements (PEAKS™ software, version 8.5) of all peptides (> 1% FDR) specific to either protein and normalized for enamel mass (CI/mg) (Zhang et al., 2012). A value of 1 was added to all combined measurements prior to normalization to avoid a null value. All values were log transformed prior to analysis. Sex was modeled as a function of log AMELX_HUMAN and log AMELY_HUMAN using Firth’s bias-reduced logistic regression (Agresti, 2013; Firth, 1993). Receiver operating characteristic (ROC) curves from models using log AMELX_HUMAN only, log AMELY_HUMAN only, and both log AMELX_HUMAN and log AMELY_HUMAN were compared using a resampling-based cross validation procedure in which, for each of 1000 resampling steps, 5 observations were randomly selected as a test set on which to evaluate the model as fitted on the remaining observations. Analyses were conducted using R, version 3.4.4 (R Core Team, 2018). Firth’s bias-reduced logistic regression was implemented using the R package bgcm, version 0.6.1 (Kosmidis, 2017). ROC curves were plotted using the R package ROCR, version 1.0–7, and confidence intervals for the cross-validated area under the ROC curve (cvAUC) were calculated using the R package cvaUC, version 1.1.0 (LeDell et al., 2014; Sing et al., 2005). The resulting logistic regression was used to determine the probability of female sex (Pr(F)) as a function of the logarithmic values of AMELX_HUMAN and AMELY_HUMAN signals. Probability values above 0.5 were estimated as female sex, and values below 0.5 estimated as male sex.

3. Results

3.1. The Y-chromosome isoform of amelogenin protein contains sexually dimorphic peptides

A bioinformatic analysis of the X- and Y-chromosome forms of amelogenin identifies numerous amino acid substitutions in the primary protein sequences (Fig. 1, blue squares). These differences increase in frequency after the inserted methionine at M59 in the Y-chromosomal form (AMELY_HUMAN, Q99217-3). The X- and Y-chromosomal protein forms also contain two and one additional sequences respectively as the result of splice variation (Fig. 1). The range of AMELY_HUMAN and AMELX_HUMAN signal varies over 3.2 and 3.7 orders of magnitude respectively.

3.2. Proteomic analysis of enamel

Overall, 2337 ± 785 peptide spectrum matches (PSM), 178 ± 56 proteins, and 929 ± 256 unique peptides (peptide sequences) were observed in modern teeth; 2005 ± 1367 PSM, 102 ± 25 proteins, and 575 ± 329 unique peptides in archaeological adult teeth; and 788 ± 882 PSM, 100 ± 32 proteins, and 281 ± 274 unique peptides in archaeological deciduous teeth. The combined ion peak intensity (CI) attributable to amelogenin proteins was 20 ± 14% of the combined ion peak intensity values of all peptides in the dataset. When AMELY_HUMAN was detected, the ratio of AMELX_HUMAN to AMELY_HUMAN was 8.54 ± 6.35%, which is consistent with the level of relative expression of 10% reported in the literature and 13% coefficient of linear least-squares regression reported below (Fincham et al., 1991).

3.3. Construction of amelogenin calibration curve

Using PEAKS™ software, the ion peak intensity values from precursor peptides specific for both AMELX_HUMAN and AMELY_HUMAN proteins were combined and normalized for enamel mass (CI/mg), and plotted in Cartesian space (Fig. 3). Samples identified as male (based on DNA, osteologically, or self-identification) are shown as blue squares and cluster away from female samples (plotted as red circles). The exceptions are two low-signal male enamel samples with no detectable AMELY_HUMAN signal and one sample with a high AMELX_HUMAN signal but no detectable AMELX_HUMAN signal (Soro Mikaya Patjca, Burial 12). Eight of nine of the unsexed deciduous molars (black diamonds) plot with the main group of male samples. The four originating juvenile skeletons can therefore be estimated as male sex.

A linear line of best fit was calculated using all samples with a detectable AMELX_HUMAN signal (y = 0.130x - 4.4 × 10^6, r = 0.939, DF = 24, p < 0.0001). The slope of the line (a = 0.13) indicates that the Y-chromosomal form was expressed at around one eighth of the level of the AMELX gene in male enamel, consistent with measured transcription levels (Fincham et al., 1991). The range of AMELX_HUMAN and AMELY_HUMAN signal varies over 3.2 and 3.7 orders of magnitude respectively.
3.4. Use of a logistic regression model to estimate female sex

Detection of peptides from the AMELY_HUMAN gene product is an unambiguous indicator of male sex. The lack of AMELY_HUMAN signal however may be due to either female sex or low signal male samples with no detectable AMELY_HUMAN peptides, either the result of endogenous biological variation or exogenous taphonomic or environmental processes. The probability of a male false negative sex estimate will decrease, and female true positive estimate increase, as the corresponding AMELX_HUMAN signal increases.

To establish a statistical framework for female sex estimation, a probability curve was constructed using data from modern samples and samples with sex estimates based on DNA methodology (n = 11, Fig. 4A). Logarithmic values were calculated for both the AMELY_HUMAN and AMELX_HUMAN combined ion peak intensity signal (CI/mg). The logarithmic values of the AMELX_HUMAN cumulative ion peak intensity signal (CI/mg) from samples with no AMELY_HUMAN signal from juvenile skeletons or samples with sex estimated from osteological markers were used to calculate the probability of female sex (Pr(F)) based on the logistic regression (Fig. 4B). The logistic regression of the sex estimation curve using logarithmic values, resulted in the following formula for Pr(F) given AMELY_HUMAN = 0:

\[
\ln(Pr(F)/(1 - Pr(F))) = -12.8622 + 0.7496 \times \ln(AMELX_HUMAN + 1)
\]

Samples with no detectable AMELY_HUMAN peptides, and an AMELX_HUMAN cumulative peak ion intensity above 5.31 \times 10^8 CI/mg have a greater than 90% probability of having female sex (Pr(F)) (Fig. 4B).

In order to evaluate the predictive capabilities of the above model, a receiver operating curve was constructed after randomly selecting 5 values as a test set, and then comparing the predicted values with the actual values, and repeating this procedure 1000 times. Resulting receiver operating (ROC) curves are included in the appendix (Figs. A.2 and A.3). The cross-validated area under the ROC curve (cvAUC) when using signal from both proteins was calculated as 0.9380 (Appendix Material: A.1 Results).

Estimation of sex using the above protein-based analytical framework was directly compared to sex estimation using osteological methodology (Table 2). Male samples, defined as having detectable levels of AMELY_HUMAN specific peptides had a null probability of female sex of (Pr(F) = 0). Unsexed samples and samples with osteological sex estimation with no detectable AMELY_HUMAN peptides were applied to the logistic regression curve and probability of female sex calculated (Pr(F)). Samples with a Pr(F) value greater than 0.5 were assigned female sex, samples with a probability of less than 0.5 were assigned male sex. There were three examples of the latter in this sample cohort, two samples (JY2 and JY15) had replicate samples with AMELY_HUMAN peptides and unambiguous male assignment. The remaining sample JY5 from Burial 3 at site SCL-928 (California) had a Pr(F) value of 0.16 and corresponding probability of male sex (1-Pr(F)) of 0.84. There was one point of difference with osteological sex estimation, sample JY63 from Burial 12 at Soro Mik'aya Patjxa (Peru, 8000 - 6500 BP) (Table 2) (Haas et al., 2017). The osteological estimation of male sex in this case is based on secondary sex characteristics of the robust mastoid process and supraorbital margin that by itself has a higher false positive rate (Haas et al., 2017). The probability of female sex in this sample was calculated at Pr(F) = 0.96. Each unsexed deciduous tooth from a juvenile skeleton (JY12 - 19, 65, 66) was unambiguously assigned as male sex (Pr(F) = 0.00, Table 2).

3.5. Persistence of amelogenin peptides over an archaeological time frame

Quantitative data for peptides specific to AMELY_HUMAN (Y) and AMELX_HUMAN protein (X) were plotted for each sample as a function of archaeological time (Fig. 5). Male (blue square) samples contained peptides specific to the sexually dimorphic AMELY_HUMAN gene product (Y) in 18 of 21 measurements from 11 individuals. Samples without detectable AMELY_HUMAN peptides included samples with estimated female sex (red circles) and low signal male sex or unsexed deciduous samples (black diamond). These samples are included in the analysis due to the addition of a single unit to prevent a null result. All samples measured contained at least some peptides specific to the AMELY_HUMAN gene product. While the signal of each gene product can vary by an order of magnitude or greater at each time point, the degree signal is persistent over archaeological time. Indeed, the oldest samples (ca. 7300 BP) are among those with the greatest signal.
3.6. Protein deamidation over archaeological time

The side chains of the amino acids glutamine and asparagine can change over time to form glutamate and aspartate (Clarke, 1987). This process, called deamidation, is accelerated in environments of extreme pH, higher temperatures, and the level of movement permitted at each amino acid residue (Clarke, 1987; Geiser and Clarke, 1987; Robinson, 2002; Tyler-Cross and Schirch, 1991; Wilson et al., 2012). Protein movement is highly restricted when proteins bind to mineralized surfaces (Demarchi et al., 2016; van Doorn et al., 2012; Wilson et al., 2012). Analysis of AMELX_HUMAN peptides in older enamel samples showed almost complete deamidation at glutamate 57 (Q57) compared to modern samples (Fig. 6A). The rate of deamidation at Q57 was quantified by combining the peak intensity of ions currents (CI) for each peptide containing either glutamine 57 (Q57) or glutamate 57 (E57), calculating the proportion of glutamate 57 bearing peptides, and then plotting as a function of time (Fig. 6B). Based on a linear least squares curve fit, deamidation occurred at rate of 1% every 305 years (R = 0.53, p < 0.0001).

4. Discussion

This study demonstrates the use of amelogenin peptide measurement from enamel to estimate sex. The presence of peptides specific to the Y-chromosome form of amelogenin protein (AMELY_HUMAN) is an unambiguous indicator of male sex. Since the absence of AMELY_HUMAN may be due to female sex or a low signal male sample, female sex should be expressed as a probability (Pr(F)), which will increase with the quantity of peptides specific to the X-chromosome form of amelogenin protein (AMELY_HUMAN). We demonstrate that this framework can be used to estimate sex of subadult teeth. The detection of amelogenin peptides is robust with signal being detected over a 7300 year time frame. This is in spite of a wide range of amelogenin signal. Based on this approach we identified one sample with potential false sex estimation based on osteological markers. Proteomic sex estimation concludes that an individual is likely a female instead of a male (Pr(F) = 0.96, Table 2).

The most characterized sex-specific genes, the amelogenin gene family, are expressed as protein in the most robust and taphonomically
persistent tissue in the human body. It is logical that recent studies have raised and then demonstrated the potential for using the Y-chromosome form of amelogenin as a means of sexing enamel samples (Porto et al., 2011a; Porto et al., 2011b; Stewart et al., 2016). Three recent studies have successfully identified peptides from AMELX_HUMAN in archaeological material in samples up to 5700 BP (Stewart et al., 2016, 2017). These innovative studies focus on detection of single masses consistent with characterized peptides from each amelogenin isoform (Stewart et al., 2017; Stewart et al., 2016). Detection of both peptides is taken as an indicator of male sex, the detection of AMELX_HUMAN peptide alone is used to determine female sex. Another innovative aspect of these studies is the minimal impact to the enamel by use of surface demineralization, with brief acid-etching, without engaging in more extensive destruction of the tooth or loss of gross anatomical features.

This project differs from the recently published work by Stewart et al., in four aspects. Firstly, the extraction of peptides from each amelogenin protein is maximized through destructive analysis and more extensive chemical extraction through milling and longer treatment with acid demineralization. Secondly, the resulting detected signals, maximal peak intensities, from multiple protein specific peptides were combined bioinformatically. Both of these approaches increase the proteomic signal from an enamel sample. Because the individual assignment of a peptide sequences to a fragmentation spectrum was screened for false positive assignment, the specificity of each signal was also increased (Zhang et al., 2012). Thirdly, the approach used in this study was quantitative, whereas the Stewart study was binary and did not take into account the possibility of false negative assignment. The quantitative approach is important for two reasons. The signal of AMELY_HUMAN was one sixth to one eighth of the magnitude of AMELX_HUMAN protein, therefore in low signal samples signals, maximal peak intensities, from multiple protein specific peptides were combined bioinformatically. Both of these approaches increase the proteomic signal from an enamel sample. Because the individual assignment of a peptide sequences to a fragmentation spectrum was screened for false positive assignment, the specificity of each signal was also increased (Zhang et al., 2012). Thirdly, the approach used in this study was quantitative, whereas the Stewart study was binary and did not take into account the possibility of false negative assignment. The quantitative approach is important for two reasons. The signal of AMELY_HUMAN was one sixth to one eighth of the magnitude of AMELX_HUMAN protein, therefore in low signal samples
it would drop out first resulting in a false positive estimate (Fincham et al., 1991). We observed variation in endogenous signal that ranged over two orders of magnitude. Both of these factors will result in some low signal male samples having no detectable AMELY_HUMAN (Fig. 3). Finally, because we always detected the AMELY_HUMAN protein, a subset of low signal male samples would be falsely assigned as female. The probability of a false negative sample however will decrease with an increase in amelogenin signal. Female sex should always be provided as a probability. Such a framework is not considered in the Stewart study. Nevertheless, the different approach in this study should not be interpreted as an advantage over the method described in the Stewart paper. Anthropologists requiring maintenance of the gross anatomical features of a tooth could employ the innovative method as outlined by Stewart.

The biogenesis of enamel is highly regulated and mediated through a group of enamel specific proteins that include both of the amelogenin proteins, ameloblastin, enamelin, and enamelysin, all of which were detected in this analysis (data are available via ProteomeXchange with identifier PXD009781, http://www.proteomexchange.org). The physical organization, calcium binding, and regulated removal of these proteins is necessary for the precise deposition of calcium apatite onto the growing faces of the microcrystal tufts to provide the right physical combination of hardness and flexibility in the enamel tissue. The mechanism is still the subject of research, but the organized and targeted degradation of these proteins by coordinated proteolysis is a necessary step (Carneiro et al., 2016; Kwak et al., 2016; Mazumder et al., 2014; Prajapati et al., 2016; Tarasevich et al., 2007). The result of this process is a range of overlapping non-tryptic peptides remaining in the tissue. By distributing the peptide signal through a wide variety of peptide masses, the signal from any one peptide is reduced. This is a concern when depending on one peptide as a source of sex information (Stewart et al., 2017; Stewart et al., 2016). The signal dilution is exacerbated by the high level of post-translational modifications particularly methionine oxidation and deamidation, which further increase the number of possible amelogenin peptide masses (Figs. 2 and 6). With time, other environmental chemistries, such as advanced glycation end-products, Maillard reactions, carbonyl oxidation and hydrolysis, will further diversify available peptide structures and reduce the signal from any one peptide species (Arena et al., 2014; Cabisco et al., 2014). The use of trypsin in the proteomic sex assay has the potential to mitigate this slightly by reducing the diversity of sexually dimorphic peptides. These environmental processes are unlikely to account for the wide range of signal (combined peak intensities) associated with each gene product, which across the study is over 3 orders of magnitude, although less than 2 order of magnitude at any time point (Figs. 3 and 5).

Our results (Fig. 5) suggest that the quantified signal was stable over 7300 years, although this may be in part due to a reduction in the complexity of the enamel proteome, with the mineral-associated amelogenin comprising a greater proportion of the remaining peptides. All proteins and peptides within the sample will degrade over time, however the amelogenin peptides, associated with the mineral apatite surfaces will have less freedom of movement and therefore less degradative chemistry (Demarchi et al., 2016; Moradian-Oldak, 1998; Zhu et al., 2014). Over time the proportion of amelogenin in the proteome will increase and lower abundance peptides will contribute to the quantification of the protein.

A consequence of the high variation in combined peak ion intensity is that male false negative samples should be expected at some level requiring female sex estimation to be expressed as a probability (Fig. 4, Table 2). There are limitations to our approach however. We made an assumption that the probability of female sex as a function of total amelogenin signal would follow a symmetrical sigmoidal curve, following the premise that high signal samples would be less likely to be a low signal male false negative sample. The precise configuration of the sigmoidal probability curve (Fig. 4B) will change as more data from samples with known sex becomes available. One weakness of the current curve is that it depends too much on one low signal data point, a male sample with no detectable AMELY_HUMAN peptides. In theory, the lowest probability of female sex at the extreme end of low signal samples may well be higher than observed in Fig. 4B. Low signal female samples will also occur. The problem of insufficient low signal samples becomes greater if replicates from the same skeleton are averaged; a duplicate of the low signal sample contained a high level of AMELY_-HUMAN peptides. A logistic regression is a more appropriate model for a discrete outcome, such as sex estimation, than regression models that have continuous outcomes (Agresti, 2013). A higher number of samples encompassing a range of signal levels will refine the parameters of the logistic regression curve and increase its accuracy. Alternatively if lower signal samples are not found, application of less material could be used to mimic low signal samples.

Protein deamidation is slower when the peptide backbone is less flexible, either because of protein structure or association with bio-mineral surfaces (Solazzo et al., 2014; Tyler-Cross and Schirch, 1991; van Doorn et al., 2012). Factors that increase peptide movement, such as temperature, will increase deamidation. The slower rate of deamidation at glutamine 57, as compared to glutamines 79, 86 and 87, could be due to increased binding to the calcium apatite surface or secondary protein structure, there is an adjacent tyrosine and tryptophan. The variation in the proportion of glutamine 57 deamidation at each time point however precludes this phenomenon from being used to estimate the archaeological age of the sample. Furthermore, the rate of deamidation increases at low pH and higher temperature, and the sample...
protocol includes treatment in 1.2 M HCl at 56 °C for one hour. A high baseline is therefore expected and changes in the protocol to reduce deamidation during sample processing would be required to evaluate the effects of sample preparation on deamidation. While the degree of deamidation at glutamine 57 is not precise enough for age estimation, like DNA depurination it can be used to demonstrate consistency with what is expected from an ancient sample.

As with any new forensic and bioarchaeological method, sex estimation using proteomic identification of the Y-chromosome form of amelogenin (AMELY_HUMAN) will need to meet several criteria before it can be routinely employed. The chemical processing of samples needs to be optimized. The extent of peptide survivability and composition of the peptide signal over archaeological time needs to be demonstrated and the lower limits of detection and quantification determined. Different strategies can be used to further increase the sensitivity of mass spectrometry. An inclusion list is used in this study to focus on and increase detection of specific peptides. Peptides on this list are analyzed first before scanning for other peptides, as this increases sensitivity and ensures that relevant peptides proceed to fragmentation (Liebler and Zimmerman, 2013). With characterization of additional amelogenin peptides this list can be expanded, increasing sensitivity of proteomic sex estimation. This targeted approach can be expanded further using specialized techniques such as parallel reaction monitoring and data-independent analysis (Gallien et al., 2015; Li et al., 2016; Searle et al., 2015; Shi et al., 2016). In parallel reaction monitoring the instrumentation focuses exclusively on relevant peptides, increasing the sensitivity of detection, sometimes by orders of magnitude (Keshishian et al., 2007). In data independent acquisition, this approach is applied broadly across all peptides (Searle et al., 2015). The routine use of a targeted approach also reduces the cost and time of the protocol and allows for potential automation (Legg et al., 2017).

Proteomic detection of sex-chromosome specific amelogenins (AMELY_HUMAN and AMELX_HUMAN) in enamel from archaeological and forensic contexts forms the basis of protein-based sex estimation. As with DNA-based methods, the protein-based sex estimation assay could be applied just as easily to deciduous or permanent teeth in subadult individuals, and to skeletons with missing, degraded, or ambiguous osteological sex markers. Additionally, the use of proteomic methods in sex estimation complements other proteomic applications in archaeological contexts, such as species identification, phylogenetics, and paleopathology. Further development of these methods, particularly if using a targeted data acquisition strategy, could be applicable over a range of mass spectrometry platforms, provide good levels of sensitivity, and be time and cost efficient. It provides a statistical and proteomic framework that can complement existing osteological and DNA-based sex estimation methods and extend the potential of sex estimation to additional archaeological assemblages, including subadult skeletons.

**Disclaimers**

A patent based on the concept and some data presented in this study has been awarded (US 8,877,455 B2, Australian Patent 2011229918, Canadian Patent CA 2794248, and European Patent EP11759843.3, GJP inventor). The patent is owned by Parker Proteomics LLC. Protein-Based Identification Technologies LLC (PBIT) has an exclusive license to develop the intellectual property and is co-owned by Utah Valley University and GJP. This ownership of PBIT and associated intellectual property does not alter policies on sharing data and materials. These financial conflicts of interest are administered by the Research Integrity and Compliance Office, Office of Research at the University of California, Davis to ensure compliance with University of California Policy.

**Acknowledgements**

The authors wish to thank Dr. Bethany Livermore-Turner (Georgia- State University) for samples and advice on study design and Drs. Robert H Rice and Tammy Bounasera for advice on editing. For the California samples, we thank Most Likely Descendants Ramona Garibay and Katherine Perez for supporting the work, and William Self- Associates, Far Western Anthropological Research Group, and the UC Davis Anthropology Museum for providing access to the samples. GJP acknowledges the support of the National Institutes of Justice/ Department of Justice (DN-BX-K065). JWE acknowledges the support of the National Science Foundation (#BCS-1318532).

**Appendix A. Supplementary data**

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jas.2018.08.011.

**Table A. 1**

| Inclusion List for Mass Spectrometry Data Acquisition. The above mass over charge values (Z = 2 and 3) with corresponding peptide sequences were used as the basis of an inclusion list in acquiring mass spectrometry data. |
|-----------------------------------------------|----------------|
| | m/z | M+2 | M+3 |
| **AMELY** | | |
| 46 - 54 | IALVTLTPKL | 484.33 | 323.22 |
| | WYQSMIRPPPY | 670.83 | 477.55 |
| | WYESMRPPPY | 671.32 | 447.88 |
| | WYQ6mIRPPPY | 678.82 | 452.89 |
| | WYESMoxIRPPPY | 679.32 | 453.21 |
| 55 - 65 | WYQSMIRPPPSYS | 714.83 | 476.89 |
| | WYQ6moxIRPPPSYS | 722.34 | 481.90 |
| | WYESMoxIRPPPSYS | 722.83 | 482.22 |
| 55 - 66 | WYQSMIRPPPSYS | 757.86 | 505.57 |
| | WYESMRPPPSYS | 758.35 | 505.90 |
| | WYQSMIRPPPSYS | 765.86 | 510.91 |
| | WYESMRPPPSYS | 766.35 | 511.23 |
| **AMELX** | | |
| 46 - 54 | TALVTLTPLK | 478.31 | 319.21 |

(continued on next page)
**Fig. A.1.** Amino Acid Variation in Amelogenin Sex-Specific Isoforms. Amelogenin genes are transcribed in 3 (AMELX: isoform 1:NM_001142.2, isoform 2:NM_182681.1, isoform 3:NM_182680.1) and 2 (AMELY: isoform 2:XP_016885531.1, isoform 1:NP_001134.1) transcripts respectively (Salido et al., 1992; Simmer, 1995). Amino acid variation between the two proteins are indicated in blue squares. The putative signal peptide is indicated in green. UNIPROT protein accession numbers are indicated.

**Fig. A.2.** Receiver Operator Curve (ROC) for models with log AMELY plus log AMELX, log AMELY only, and log AMELX only. X-axis shows the false negative rate (1 – specificity) and the y-axis shows the true positive rate (sensitivity). Values for the model with log AMELX and log AMELY are shown in red, values for the model with log AMELY only are shown in blue, and values for the model with log AMELX only are shown in green.
References


