Powell *et al. BMC Genomics* (2020) 21:277
https://doi.org/10.1186/s12864-020-6681-2

BMC Genomics

RESEARCH ARTICLE Open Access

Entropy based analysis of vertebrate sperm protamines sequences: evidence of potential dityrosine and cysteine-tyrosine cross-linking in sperm protamines

Christian D. Powell[1,2], Daniel C. Kirchoff[1], Jason E. DeRouchey[1] and Hunter N. B. Moseley[2,3,4*]

Abstract

Background: Spermatogenesis is the process by which germ cells develop into spermatozoa in the testis. Sperm protamines are small, arginine-rich nuclear proteins which replace somatic histones during spermatogenesis, allowing a hypercondensed DNA state that leads to a smaller nucleus and facilitating sperm head formation. In eutherian mammals, the protamine-DNA complex is achieved through a combination of intra- and intermolecular cysteine cross-linking and possibly histidine-cysteine zinc ion binding. Most metatherian sperm protamines lack cysteine but perform the same function. This lack of dicysteine cross-linking has made the mechanism behind metatherian protamines folding unclear.

Results: Protamine sequences from UniProt's databases were pulled down and sorted into homologous groups. Multiple sequence alignments were then generated and a gap weighted relative entropy score calculated for each position. For the eutherian alignments, the cysteine containing positions were the most highly conserved. For the metatherian alignment, the tyrosine containing positions were the most highly conserved and corresponded to the cysteine positions in the eutherian alignment.

Conclusions: High conservation indicates likely functionally/structurally important residues at these positions in the metatherian protamines and the correspondence with cysteine positions within the eutherian alignment implies a similarity in function. One possible explanation is that the metatherian protamine structure relies upon dityrosine cross-linking between these highly conserved tyrosines. Also, the human protamine P1 sequence has a tyrosine substitution in a position expecting eutherian dicysteine cross-linking. Similarly, some members of the metatherian Planigales genus contain cysteine substitutions in positions expecting plausible metatherian dityrosine cross-linking. Rare cysteine-tyrosine cross-linking could explain both observations.

Keywords: Protamine, Sperm Chromatin, Relative Entropy, Cross-Linking

*Correspondence: hunter.moseley@uky.edu
[2]Markey Cancer Center, University of Kentucky, 800 Rose Street, Pavilion CC 40536, Lexington, USA
[3]Department of Molecular & Cellular Biochemistry, University of Kentucky, 40508, Lexington, USA
Full list of author information is available at the end of the article

Background

The process in which male germ cells develop into sperm cells is called spermatogenesis. During spermatogenesis, DNA undergoes hypercondensation in order to form a smaller nucleus. This is accomplished through the final replacement of a vast majority of somatic DNA histones (>90%) with one of three nuclear proteins; sperm-specific histones, protamine-like proteins, or protamines [1]. In mammals, sperm protamines are small (<60 amino acids), arginine-rich nuclear proteins. After hypercondensation of DNA mediated by protamines, a haploid male germ cell nucleus is formed, which is genetically inactive but just 1/20th the size of a somatic cell nucleus [2]. This reorganization of the spermatozoa DNA is also thought to protect the paternal genome against oxidative damage [1, 3–5].

Genetically the family of sperm protamines is highly diverse, being observed across the tree of life. For example, a single species of fish can contain multiple genes for protamine and protamine-like proteins, whereas birds tend to have two identical copies of a single protamine gene [1]. While all protamines perform the task of binding and condensing DNA, the sizes and structural components of the protamines can vary greatly from species to species. Despite these differences most sperm protamines include large arginine-rich DNA binding regions and phosphorylation sites [1, 6, 7]. In addition, arginine content in mammalian sperm protamine P1s appear to be under sexual selection pressure [8]. The positively charged arginine residues in these binding regions are able to engage in an electrostatic interaction with the negatively charged DNA phosphate backbone [6]. These interactions form a toroid shaped protamine-DNA complex conforming to an internal hexagonal lattice [9]. The various phosphorylation sites in the protamine sequences are involved in a number of post-translational modifications and are thought to regulate the interactions with DNA.

Some of the simplest protamines are those of fish. Like most sperm protamines, the sequences of fish protamines are populated with a large percentage of arginine residues. However, fish protamines tend to be under 35 amino acids in length and contain increased frequencies of arginine (approximately 70%) in comparison to their mammalian analogs. The secondary structure of the protamines consist of multiple beta turns, with limited CD, NMR, and fluorescence data, indicating the formation of a possible globular structure [10, 11].

Mammals have a relatively conserved set of protamines, with metatherian mammals having only one protamine gene, while eutherian mammals have two to three varieties. These mammalian sequences tend to start with MARYR at the N-terminus, typically followed by a region containing one or more phosphorylation sites , then a DNA binding region comprised of multiple blocks of arginine residues, and ending with a varied *C*-terminal region [7, 12].

The eutherian protamine P1 is encoded by the *PRM1* gene. Alignment of the sequences of eutherian mammal sperm protamine P1 have shown the sequences to be relatively conserved. In eutherian sperm protamine P1 sequences, the arginine-rich DNA binding regions are broken up by cysteine residues, which are involved in both inter- and intramolecular disulfide cross-linkings [7, 13, 14]. In bull protamine P1, the intra-protamine disulfide bonds were shown to create a hairpin-like structure, with disulfide crosslinks formed between the cysteines in positions 7 and 15 as well as the cysteines at positions 40 and 48. The remaining cysteine positions in bull protamine P1 are involved in inter-protamine bonding. More recently, we have shown that this disulfide mediated secondary structure of the bull protamine is required for proper chromatin remodeling [15] (Kirchoff et al.: Disulfide-mediated secondary structure in protamine is critical for dna condensation in mammalian sperm chromatin, in preparation).

The other two eutherian sperm protamine types are encoded in the *PRM2* gene. These other protamine proteins are longer than the eutherian sperm protamine P1 type protamine and include a number of post-translational truncation sites in the N-terminal tail [1]. Unlike the eutherian P1 type sperm protamines, the P2 protamines engage in zinc ion binding that is stoichiometrically 1:1 for many eutherian mammals [16]. This zinc ion binding is achieved with highly conserved cysteine and histidine residues in the P2 protamine sequence. Both eutherian P1 and P2 type protamines engage in intermolecular disulfide cross-linking with one another when forming the DNA protamine complex [14]. For all eutherian sperm protamine types, it is thought that the cysteine cross-linkages are important for protecting the spermatozoa from oxidative damage [3–5].

Also in eutherian mammals, a testis-specific variant of glutathione peroxidase (GPx4) is involved in the formation of the thiol cross-linking between and within the protamines and with protecting the sperm cells from oxidative stress due to reactive oxygen species [17, 18]. In particular, Conrad et al. showed that without GPx4, sperm develop abnormal heads likely due to a lack of stabilizing disulfide cross-linking [18].

In contrast to eutherian sperm protamines, little is known about metatherian sperm protamines, except that metatherian sperm protamines tend to lack cysteine residues with the only exception to this tendency involving species of the Planigale genus [19]. To our knowledge, there is no consensus on the structure of metatherian sperm protamines, nor is there prior evidence to suggest that inter- and intramolecular cross-linking occurs

```
                              10          20          30
PRTA_ACIST          -ARRRRRHAS TKLKRRR--- -----RRRRH GKKSHK
PRT_ORYLA           ----MRRQAS LPARRRRRVR RTRVVRRRRR VGRRRH
PRT_PERFV           --PRRRRHAA RPVRRRRRTR RSSRVHRRRR AVRRRR
PRTB_MUGCE          --PRRRRETS RPIRRRRRAR RAPI-RRRRR VVRRRR
PRT1_SCOSC          -MPRRRRRAS RPVRRRRRAR RSTAVRRRRR VVRRRR
PRT2_SCOSC          -MPRRRRRAS RPIRRRRRAR RSTAVRRRRR VVRRRR
PRT_DICLA           --PRRRRQAS RPVRRRRRTR RSTAERRRRR VVRRRR
PRTY_THUTH          --PRRRRQAS RPVRRRRRYR RSTAARRRRR VVRRRR
PRTZ_THUTH          --PRRRRRSS RPVRRRRRYR RSTVARRRRR VVRRRR
PRTZ1_SAROR         --PRRRRRSS RPVRRRRRYR RSTAARRRRR VVRRRR
...
```

Fig. 1 Fish Protamine Alignment. Alignment showing a selection of 10 fish protamine amino acid sequences from the alignment of fish protamine sequences. The full multiple sequence alignment used 34 sequences from the 2019_05 release of the UniProt knowledgebase and was aligned with MUSCLE 3.8.31. For full alignment see Additional file 1: Supplemental Figure 1

in metatherian sperm protamines, with the exception of species of the Planigale genus where it was suggested [19]. Additionally, it is unclear if GPx4 is required for the proper function of metatherian sperm protamines, although it is known that metatherian mammals do express glutathione peroxidase for defense against oxidative stress [20]. Sequence data also exists for the testes specific version of glutathione peroxidase in Tasmanian Devils (G3WAH0_SARHA) [21]. Metatherian spermatozoa are more susceptible to oxidative damage, likely due to a lack of stabilizing disulfide cross-linkages [3–5]. These current gaps in knowledge prompted the following analyses of multiple sequence alignments (MSAs) of eutherian P1, eutherian P2, metatherian P1, and fish sperm protamine sequences, which provide some insight into the structures of protamines and mechanism behind protamine mediated DNA condensation.

Results

MSAs were generated for 145 eutherian sperm protamine P1, 16 eutherian sperm protamine P2, 95 metatherian sperm protamine, and 34 fish protamine sequences, all retrieved from the UniProt knowledgebase and aligned using MUSCLE 3.8.31 [21, 22]. The MSAs were then analyzed using the relative entropy method described.

Fish protamine

From the fish protamine MSA (see Fig. 1 or Additional file 1: Supplemental Figure 1 for the full fish protamine MSA), the relative entropy-based analysis showed that no position in the alignment had conservation scores above the relative entropy threshold. The most highly conserved positions were those containing only arginine residues. There was in fact a four-way tie for the most highly conserved position with positions 15, 16, 17, and 27 all having a conservation score equal to the conservation threshold of 4.135.

Eutherian sperm protamine P1

For the eutherian sperm protamine P1 MSA (see Fig. 2 or Additional file 1: Supplemental Figure 2 for the full eutherian sperm protamine P1 MSA), a total of nine positions were determined to be highly conserved, based on relative entropy scores and conservation threshold described

```
                        10         20         30         40         50         60
HSP1_CAPHI      MARYRCCLTH --SRSRCR-R ---RRRRRCR -RRRRRFGRR --RRR-RVCC RRY--TVVRC TRQ-
HSP1_SHEEP      MARYRCCLTH --SRSRCR-R ---RRRRRCR -RRRRRFGRR --RRR-RVCC RRY--TVVRC TRQ-
HSP1_BOVIN      MARYRCCLTH --SGSRCR-R ---RRRRRCR -RRRRRFGRR --RRR-RVCC RRY--TVIRC TRQ-
HSP1_PIG        MARYRCCRSH --SRSRCR-P ---R-RRRCR -RRRRRCCPR --RRR-AVCC RRY--TVIRC RRC-
HSP1_HORSE      MARYRCCRSQ --SQSRCR-R ---RRRRRCR -RRRRRSVRQ --RR---VCC RRY--TVLRC RRRR
HSP1_ORCOR      MARNR-CRSP --SQSRCR-R ---P-RRRCR --RRIRCCRR --QR--RVCC RRY--TTTRC ARQ-
HSP1_MOUSE      MARYRCCRSK --SRSRCR-R ---R-RRRCR -RRRRRCCRR --RR--RRCC RRRRSYTIRC KKY-
HSP1_RAT        MARYRCCRSK --SRSRCR-R ---R-RRRCR -RRRRRCCRR --RR--RRCC RRRRSYTFRC KRY-
HSP1_HUMAN      MARYRCCRSQ --SRSRYY-R ---Q-RQRSR -RRRRRSCQT --RRRAMRCC RPR--YRPRC RRH-
HSP1_GORGO      MARYRCCRSQ --SRSRCY-R ---Q-RQTSR -RRRRRSCQT --QRRAMRCC RRR--NRLRR RKH-
...
```

Fig. 2 Eutherian P1 Sperm Protamine Alignment. Alignment showing a selection of sperm protamine P1 amino acid sequences from 10 common eutherian mammals. The full multiple sequence alignment used 145 sequences from the 2019_05 release of the UniProt knowledgebase and was aligned with MUSCLE 3.8.31. Positions with relative entropy score greater than the conservation threshold of 4.135 are highlighted (see Additional file 1: Supplemental Table 1). For full alignment see Supplemental Figure 2

above. All nine highly conserved positions within the alignments were positions comprised primarily of cysteine residues. In descending conservation score, the most highly conserved positions were positions 7, 49, 50, 60, 38, 17, 29, 6, and 37. For a listing of all the positions in the eutherian protamine P1 MSA which were above the conservation score see Supplemental Table 1. All highly conserved positions within the alignment were composed of over 69% cysteine residues (excluding gaps) and the most highly conserved positions tended to all be within one residue of a known intramolecular cross-linking region (positions 7, 49, 50, 60, 17, and 6) [13].

Eutherian sperm protamine P2

For the processed eutherian sperm protamine P2 MSA (see Fig. 3 or Additional file 1: Supplemental Figure 3 for the full unprocessed eutherian sperm protamine P2 MSA), a total of eleven positions were determined to be highly conserved, based on relative entropy scores and conservation threshold described above. While intramolecular cross-linking in sperm protamine P2 proteins have yet to be determined [14], half of the positions identified as highly conserved in the P2 alignment consisted primarily of cysteine residues (positions 59, 75, 83, 93, and 107). The remaining positions are either primarily composed of histidine (positions 68, 89, 53, 85, and 110) or tyrosine (position 54) residues. For a listing of all the positions in the processed eutherian protamine P2 MSA which were above the conservation score see Supplemental Table 2. For a listing of all the positions in the full, unprocessed eutherian protamine P2 MSA which were above the conservation score see Supplemental Table 3.

Metatherian sperm protamine P1

When the relative entropy method was applied to the metatherian sperm protamine alignment (see Fig. 4 or Additional file 1: Supplemental Figure 4 for the full metatherian sperm protamine MSA), similar results are found in the eutherian sperm protamine P1 alignment. Instead of nine positions determined to be conserved, like in the eutherian P1 alignment, the metatherian alignment only has seven highly conserved positions. Of these seven positions, six were found to primarily contain tyrosine (positions 4, 57, 16, 62, 75, and 34). The remaining highly conserved position in the alignment primarily consisted of histidine residues (position 7). For a listing of all the positions in the metatherian protamine MSA which were above the conservation score see Supplemental Table 4.

Whole sequence arginine-Lysine density analysis

The arginine-lysine frequencies for each protamine in each homologous protamine group are shown in Fig. 5 and in Table 1. It is clear from the differences in these arginine-lysine frequency distributions that the relative proportion of the DNA binding region to the whole protamine sequence is quite different for the protamine groups, especially the eutherian P2 protamines.

DNA binding region arginine-Lysine density analysis

Figure 6 and Table 2 show the arginine-lysine frequencies in the hypothesized DNA binding regions for each protamine in the eutherian P1 and metatherian MSAs. Fish protamines were included in Fig. 6 for comparison. The distributions were analyzed using a Welch's t-test which showed that each protamine group's arginine-lysine frequency distribution in their hypothesized DNA binding region is statistically discrete from any other groups. Comparing the DNA binding region of the eutherian P1 sperm protamine group to that of the metatherian sperm protamine group yielded a p-value of 2.184e-2. Comparing the DNA binding region of the eutherian P1 sperm

```
                          58        68        78        88        98        108
PRM2_RATTU       RG--HHRHRR CSRKRLHRIH KRR-RSCRRR RRHSCCHRRR HRRGCRRSRR RRRCRCRKCR RQCH
PRM2_MOUSE       RGHHHHRHRR CSRKRLHRIH KRR-RSCRRR RRHSCRHRRR HRRGCRRSRR RRRCRCRKCR RHHH
PRM2_RATFU       RG--HHRHRR CSRKRLHRIH KRR-RSCRRR RRHSCCHRRR HRRGCRRSRR RRRCKCRKCR RHCH
PRM2_ALOSE       QGCYGYRRRL CSRRRLYRVH RRQRRSCRRR C---CRYRRR NRRGCRT-RR RT------CR RH--
PRM2_CALJA       QGYSSYRRRR CSRRRRYRIH RRRSRSCRRR RRRSCRYRRR PRRGCRSRRR RR------CR RY--
PRM2_SEMEN       QGYSHHRRRR CSRRRLYRIH RRRHRSCRRR RRRSCRHRRR HRRGCRT-RR RR------CR RY--
PRM2_ERYPA       QGHSHHRRRR CSQRRLHRIH RRRHRSCRRR RRRSCRHRRR HRRGCRT-RR RR------CR RY--
PRM2_MACNE       RGHSHHRRRR CSRRRLHRIH RRRHRSCRRR RRRSCRHRRR HRRGCRT-RR RR------CR RH--
PRM2_MACFU       RGHSHHRRRR CSRRRLHRIH RRRHRSCRRR RRRSCRHRRR HRRGCRT-RR RR------CR RH--
PRM2_MACMU       -GHSHYRRRH CSRRRLHRIH RRRHRSCRRR RRRSCRHRRR HRRGCRT-RR RR------CR RH--
PRM2_GORGO       -GHSHYRRRH CSRRRLRRIH RQQHRSCRRR KRRSCRHRRR HRKGCRT-RR RT------CR RH--
PRM2_PANPA       -GHSHYRRRH CSRRRLRRIH RQQHRSCRRR KRRSCRHRRR HRRGCRT-RR RT------CR KH--
PRM2_PANTR       -GHSHYRRRH CSRRRLRRIH RQQHRSCRRR KRRSCRHRRK HRRGCRT-RR RT------CR RH--
PRM2_HUMAN       -GQSHYRRRH CSRRRLHRIH RRQHRSCRRR KRRSCRHRRR HRRGCRT-RK RT------CR RH--
PRM2_PONPY       -GHSHYRRRH CSRRRLHRIH RQQHRSCKRR RRHSCRHRRK HRRGCRT-RR RT------CR RH--
PRM2_HYLLA       -GHSHYRRRH CSRRRLHRIH RQQHRSCGRR RRRSCRQRRR HRRGCRT-RR RR------CR RH--
```

Fig. 3 Eutherian P2 Sperm Protamine Alignment. Alignment showing processed sperm protamine P2 amino acid sequences from the alignment of 19 eutherian mammals. Sequences were pulled from the 2019_05 release of the UniProt knowledgebase and was aligned with MUSCLE 3.8.31. Positions with relative entropy score greater than the conservation threshold of 4.135 are highlighted (see Supplemental Table 2 for the truncated alignment and see Supplemental Table 3 for the untruncated alignment). For the untruncated alignment see Supplemental Figure 3

```
                        10         20         30         40         50         60
HSP1_MONDO      MARYRRRSRS RSRSRYGRRR RRSRS----R RRRSRRRRR- ---RRGRR-- ---GRGYHRR
HSP1_DIDVI      MARYRRRSRS RSRSRYGRRR RRSRS----R RRRSRRRRR- ---RRGRR-- ---GRGYHRR
HSP1_PHACI      MARYR-HSRS RSRSRY-QRR RRRRSRYRSQ RRRYRRRRGS RRRRRRGR-- ---RRGY-RR
HSP1_PSECU      MARYR-HSRS RSRSRYRRRR RRRRSRYRGR RRRYRRSRR- RRRRGRRRGN CLGRRGYRRR
HSP1_MACRU      MARYR-HSRS RSRSRY-RRR RRRRSRYRSQ RRRYRGRRR- -RRSRRGR-- ---RRGYSRR
HSP1_MACGI      MARYR-HSRS RSRSRY-RRR RRRRSRYRSR RRRYRGRRR- RRSRRGRR-- ---RRGYSRR
HSP1_WALBI      MARYR-HSRS RSRSRY-RRR RRRRSRYRSR RRRYRGRRR- RRSRRGRR-- ---RRGYSRR
HSP1_THYCY      MARYRRHSRS RSRSRY-RRR RRRRSRHHNR RRTYRRSRR- HSRRRRGR-- ---RRGYSRR
HSP1_PLAMS      MARCRRHSRS RSRSRN-QCQ RRRRRRY-NR RRTYRRSRR- HSRRRRGR-- ---RRGCSRR
HSP1_PLAIN      MARSRRHSRS RSRSR--NQC QRRRRRTYNR RRTMREKPR- HSRRRRVR-- ---RRGCSCR
...
                        70         80
HSP1_MONDO      SPHRR---RR RRRR------ -
HSP1_DIDVI      SPHRR---RR RRRR------ -
HSP1_PHACI      RYS-----RR --RRY----- -
HSP1_PSECU      RYS-----RR RRRRYY---- -
HSP1_MACRU      RYS------R RRRRY----- -
HSP1_MACGI      RYS------R RRRRY----- -
HSP1_WALBI      RYS-----RR RRRRY----- -
HSP1_THYCY      RYS-----RR GRRRY----- -
HSP1_PLAMS      RYS-----RR GRRRY----- -
HSP1_PLAIN      RCS-----RR RRRRC----- -
...
```

Fig. 4 Metatherian Sperm Protamine Alignment. Alignment showing a selection of sperm protamine P1 amino acid sequences from 10 metatherian mammals. The full multiple sequence alignment used 95 sequences from the 2019_05 release of the UniProt knowledgebase and was aligned with MUSCLE 3.8.31. Positions with relative entropy score greater than the conservation threshold of 4.135 are highlighted (see Additional file 1: Supplemental Table 4). For full alignment see Supplemental Figure 4

protamine group to the whole fish sequence sperm protamine group yielded a p-value of 4.762e-11. Comparing the DNA binding region of the metatherian sperm protamine group to the whole fish sequence sperm protamine group yielded a p-value of 4.987e-8. However, the differences between the medians of these distributions is less than 0.017 (see Table 2) and the possible functional consequences of these relatively small differences in arginine-lysine frequencies are unclear given the relatively high variance of each group. It is possible that the arginine-lysine frequency of the DNA binding region across all of the protamines is just a species- and protein-specific optimization of the DNA binding function.

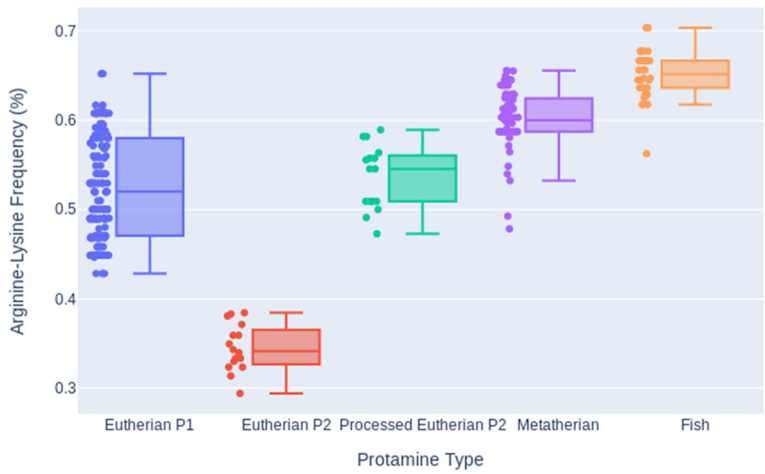

Fig. 5 Arginine-Lysine Frequencies for Protamine Groups. Box plots of the arginine-lysine frequency for eutherian protamine P1, processed eutherian P2, metatherian P1, and fish protamines. Individual arginine-lysine frequencies are plotted next to the associated box plot. For min, quartile, and max data see Table 1

Table 1 Arginine-lysine frequency low, quartile, and max information for each sperm protamine group

Alignment	Min	Q1	Median	Q3	Max
Eutherian P1	0.429	0.471	0.52	0.58	0.652
Eutherian P2	0.294	0.328	0.341	0.362	0.385
Processed Eutherian P2	0.473	0.509	0.545	0.559	0.589
Metatherian	0.478	0.587	0.6	0.624	0.656
Fish	0.562	0.636	0.652	0.667	0.704

Table 2 Arginine-lysine frequency low, quartile, and max information for the hypothesised protamine DNA binding regions

Alignment	Min	Q1	Median	Q3	Max
Eutherian P1 DNA	0.478	0.636	0.714	0.773	0.857
Metatherian DNA	0.529	0.667	0.697	0.719	0.781

Discussion

The results from the MSA analysis of fish protamines showed no highly conserved residues above arginine. This likely indicates that the only functionally/structurally important residues in the fish protamines are arginine. The arginine-lysine density analysis showed that fish have a greater charge density across their entire protamines sequences than any of the other protamine groups.

The results from the MSA analysis of the eutherian protamine P1 sequences showed that the most highly conserved positions tend to be cysteine containing. The high evolutionary sequence conservation indicates that the positions are of great functional/structural importance. When these highly conserved positions are overlaid onto a proposed schematic structure for bull sperm protamine P1 [13, 14], it is clear that the conserved positions align with the cysteines involved in intra- and intermolecular bonding in bull sperm protamine P1. It is also notable that the cysteines involved in the intramolecular cross-linkings were shown to be more highly conserved than those involved in the intermolecular cross-linkings. This likely supports the hypothesis that the hairpin-like secondary structure of eutherian sperm protamine P1s is required for proper DNA hypercondensation [15].

Comparing the metatherian P1 MSA to the eutherian P1 MSA, we find a number of commonalities. Both the N-terminal regions contain phosphorylation sites followed by blocks of arginine residues broken up by residues which can engage in cross-linking (cysteine in eutherians and tyrosine in metatherians) [7]. A proposed schematic structure for metatherian sperm protamine P1 with conserved tyrosine positions is shown in Fig. 7. The conserved tyrosine positions are visualized interacting in a similar cross-linking pattern as is observed in the cysteine containing eutherian mammal sperm protamines. Due to the similar conserved nature and similar spacing between the tyrosine residues in the metatherian protamine MSA and the cysteine residues in the eutherian protamine MSA, we hypothesize that metatherian protamines undergo intramolecular and potentially intermolecular crosslinking, enabling an analogous structure and folding mechanism as their eutherian counterparts. Folding of the metatherian protamines could possibly be facilitated by an orthologous enzyme to the glutathione peroxidase found in eutheria or an analogous peroxidase. There are specific peroxidases (e.g. certain myeloperoxidases) that are capable of catalysing dityrosine cross-linking in proteins [23–25]. However pi-pi stacking of two nearby tyrosines can represent another possible structural motif hypothesis [26].

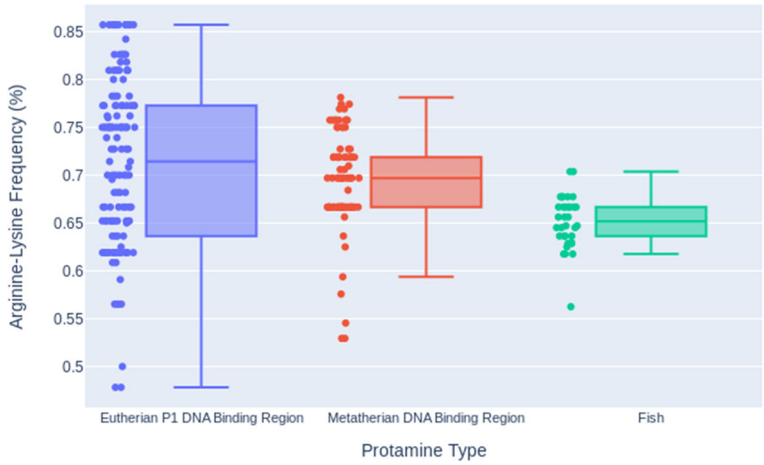

Fig. 6 Arginine-Lysine Frequencies for Protamine DNA Binding Regions. Box plots of the arginine-lysine frequency for the DNA binding region of eutherian P1 protamines, the hypothesised DNA binding region of metatherian sperm protamines, and the whole sequence of fish protamines for comparison. Individual arginine-lysine frequencies are plotted next to the associated box plot. For min, quartile, and max data see Table 2

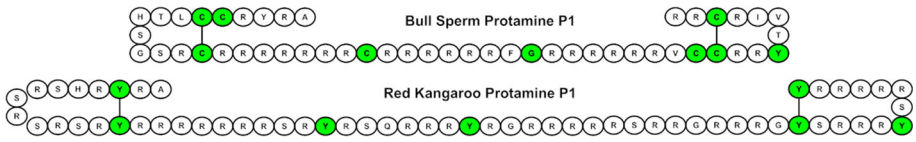

Fig. 7 Proposed Schematic Structure for Eutherian and Metatherian P1 Type Protamines. Highly conserved positions highlighted and overlaid onto proposed schematic structures for Bos taurus (Bull) and Macropus rufus (Red Kangaroo) sperm protamine P1s. Schematic structure of Bull P1 sperm protamines based off of Balhorn 1991 [13]. Red Kangaroo schematic structure assumed from similarity

The comparison of arginine-lysine frequencies of the differing protamine groups showed that although metatherian sperm protamines and fish protamines have similarly high charged residue frequencies across their entire sequences, metatherian mammals have an increased density of charged residues towards the middle of the sequences. This trend is also found in eutherian sperm protamine P1s where there is a central DNA binding region. The charged residue frequencies of the hypothesized DNA binding regions of eutherian P1 and metatherian sperm protamines were the most similar distributions by a significant factor.

Additional deviations are also apparent in the MSAs. One example is the tyrosine substitutions (tyr15cys and tyr16cys) of two positions in human sperm protamine P1, which are expected to be involved in intramolecular dicysteine cross-linking in the N-terminal staple fold. It is possible that cysteine-tyrosine cross-linking preserves the cross-linking function in human sperm protamines with these substitutions. A proposed schematic structure for human sperm protamine P1 with conserved cysteine and tyrosine positions is shown in Fig. 8. Cysteine-tyrosine cross-linking, while not novel, is extremely rare in nature and are known to be mediated by copper metalloprotein enzymes [27]. Also, prior research has shown that human spermatozoa are more susceptible to severe oxidative stress (>= 5 mM H2O2) in comparison to other eutherian mammals [3]. These substitutions could be an explanation for these observations.

Likewise, a counter deviation found in the metatherian alignment is with species of the Planigale genus, for which all species but P. maculata contain cysteine substitutions in a number of highly conserved positions where tyrosine residues are expected (Additional file 1: Supplemental Figure 4). It is notable that the change from tyrosine to cysteine is a single nucleotide substitution and that the arginine-lysine heavy regions are still found in the sequences of the Planigale species. This could suggest that some metatherian protamine sequences are convergently evolving towards similar motifs as seen in the protamine sequences of more highly evolved mammals [28].

Conclusions

In summary, the common patterns of sequence conservation between eutherian and metatherian protamine P1 sequence families support hypotheses for dityrosine cross-linking in the metatherian P1 protamines and a rare cysteine-tyrosine cross-linking in human sperm protamine P1. The presence of cysteine cross-linking in a number of species of the Planigale genus also indicates that metatherian sperm protamines are likely capable of taking on a hairpin-like structure via intramolecular cross-linking analogous to eutherian sperm protamine P1. This was additionally supported by the finding that metatherian sperm protamines also contain an increased density of arginine and lysine residues in the center of the sequence, which likely represents a large DNA binding region analogous to the DNA region found in eutherian sperm protamine P1 sequences. In addition to directly testing these hypotheses with wet lab and analytical experiments, the next logical steps involve searching for an analogous peroxidase enzyme with expression localized to the testis in metatherian mammals to provide further evidence of dityrosine cross-linking and possibly an analogous peroxidase with a copper binding cofactor or new mechanism to support the formation of a cysteine-tyrosine cross-linking in human sperm protamines. Moreover, these proposed mechanisms and structures may play a role in fertility, particularly human fertility.

Methods

We used an entropy-based method to determine the functionally important residues in the MSAs of various protamine groups. Additionally, the charged residues

Fig. 8 Proposed Schematic Structure for Human Sperm Protamine P1. Highly conserved positions highlighted and overlaid onto proposed schematic structures for Homo sapien sperm protamine P1. Tyrosine containing positions highlighted in red and cysteine and other residue containing positions highlighted in green. Human schematic structure assumed from similarity [13]

densities of the protamine sequences in each group were determined. By comparing the conserved residues and the charged residue densities, we made predictions about structural features for related protamine groups and mechanisms behind their ability to bind DNA.

Creation of homologous protamine groups

All entries containing the keyword "protamine" were downloaded from the May 2019 release (release 2019_05) of the UniProt KnowledgeBase (SwissProt/TrEMBL) [21] to create the initial dataset of protamine and protamine-like proteins.

The protamine dataset was then broken down into four homologous groups based on existing UniProt gene name and organism classification annotations. The four groups were eutherian sperm protamine P1, eutherian sperm protamine P2, metatherian sperm protamine P1, and fish protamine. The eutherian sperm protamine P1 group was parsed by collecting all sequences which contained 'Eutheria' in their organism classification and the gene name of either 'PRM1' or 'Prm1'. The same approach was performed for the eutherian sperm protamine P2 group, but using the gene name of either 'PRM2' or 'Prm2'. The metatherian sperm protamine P1 group was parsed by collecting all sequences which contained 'Metatheria' in their organism classification and with the gene name of either 'PRM1' or 'Prm1'. The fish protamine group was parsed by collecting any sequence which contained 'Actinopterygii' in their organism classification and that did not contain the word 'like' in their description. For the eutherian and metatherian groups only a single entry was allowed per organism per group. If multiple sequences existed in a single group, preference was given to the Swiss-Prot entry since these are reviewed entries. Therefore, each group is composed of orthologous genes, with the exception of the fish protamine group where some species have more than one protamine gene in the group.

Additional filtering and truncation of protamine p2 sequences

Sperm protamine P2's contain multiple post-translational cleavage sites, which lead to the removal of 40% of the amino terminus of these proteins [1]. After processing, the protein sequence is slightly longer than that of protamine P1 and the processed protein of P2 also has a higher arginine frequency than that of the unprocessed sequence [1]. The unprocessed P2 protamine binds to DNA and is truncated over several days leaving only the processed protamine [29–31]. As the processed protamine P2 sequences has higher similarity to the P1 protamine sequences the P2 alignment is truncated down to include only the processed regions. This is achieved by determining the closest post-translational processing site to each protein's DNA binding region in the eutherian sperm protamine P2 alignment. The post-translating sites were found by using mouse sperm protamine P2 (MOUSE_PRM2) as a reference [1]. MOUSE_PRM2 residue 44 is the closest post-translational processing site to the protein's DNA binding region in mice. MOUSE_PRM2 residue 44 can be found at position 48 in the eutherian P2 sperm protamine alignment. Position 48 of the alignment was therefore used as the truncation site.

Aberrant sperm protamine P2 sequences (Rat, Boar, Bovin) caused gapping in an initial MSA and were found to lack significant translational expression in prior literature [32, 33]. Also, the sequence for Chinese hamster PRM2 is a pseudogene [34]. Therefore these sequences were removed before final alignment.

Multiple sequence alignment and conservation analysis

A fasta file was generated for each homologous protamine group (i.e., eutherian P1, eutherian P2, and metatherian P1) and then MUSCLE 3.8.31 [22] was used to create an MSA using default settings.

Relative entropy (Kullback-Leibler divergence) was used to determine residue conservation scores for each position (column of residues) in the alignment. Relative entropy incorporates background frequencies of amino acids to measure the distance between the amino acid frequency in a position of the alignment versus the background frequencies [35–37].

$$D_{KL}(P||Q) = \sum_{a \epsilon AA} P(a) log \frac{P(a)}{Q(a)} \quad (1)$$

D(P‖Q) is calculated for each position in the alignment and uses all 20 of the standard amino acids plus Asx (B) for Asp or Asn, Glx (Z) for Glu or Gln, and Xaa (X) for unknown. P(a) is the frequency of the amino acid in the position. Q(a) is the background frequency of an amino acid. For this analysis, the natural abundance of amino acids determined by the UniProt knowledgebase was used [21]. Relative entropy has been shown to be one of the most effective algorithms for determining functionally/structurally important residues from alignments and tied for the most effective method for determining positions playing a role in protein-protein interactions [35, 37].

Additionally, a weighting was used to deal with the presence of gaps in the alignment. The gap weighting is incorporated by multiplying the calculated relative entropy measure by the percent of non-gap residues in the position.

$$G_W = D_{KL}(P||Q) * \%nongapresidues \quad (2)$$

To determine which positions are conserved in the alignment, a conservation score threshold equal to a position entirely composed of arginine residues (~4.1354) was used. If the gap weighted conservation score was greater

than the threshold, the position was determined to be conserved. Methionine residues at the beginning of protein sequences are ignored.

Arginine-Lysine density analysis

The arginine-lysine density of each protamine group was calculated by counting the number of arginine and lysine residues in each protamine sequence in the group. Histidine was left out of the analysis as it is mostly deprotonated at physiological pH. Lysine was included above a simple arginine frequency due to known DNA interactions for lysine residues in a variety of DNA binding proteins, but more importantly, an observed reduction in the severity of lower bound outliers for all protamine groups analyzed. This improvement in the lower bound outliers is greater for lysine than for the inclusion of any other amino acid (see Supplemental Tables 5-9). It is also important to note that there is an evolutionary trend for the replacement of lysine with arginine suggesting arginine is favored over lysine for the condensation of DNA during spermatogenesis [38]. The quartile ranges for each group was then calculated and graphed using Plotly [39]. For eutherian sperm protamine P2, the processed sequence was used, as determined by the method mentioned above. Additionally, for each protein in the eutherian P1 and metatherian sperm protamine alignments the charged residue density within the hypothesized DNA binding region was calculated. The hypothesized DNA binding region for the eutherian P1 alignment begins at position 17 and ends at position 46. The hypothesized DNA binding region for the metatherian alignment begins at position 16 and ends at position 56.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6681-2.

Additional file 1: Additional figures and tables referenced within this article. Supplemental information is available as a PDF file.
The additional figures and tables referenced within this article are available in the figshare repository, https://doi.org/10.6084/m9.figshare.10292573.v1 [40]. The supplemental information file (.pdf) can be found in the .zip archive under ProtamineAnalysisFigshare/Docs/supplemmental.pdf.

Abbreviations

GPx4: Glutathione peroxidase 4 protein; MSA: Multiple sequence alignment; PRM1: Protamine 1 gene; PRM2: Protamine 2 gene; P1: Protamine 1 protein; P2: Protamine 2 protein

Acknowledgements

The authors would like to acknowledge the efforts of the UniProt Consortium in their mission to provide high-quality and freely accessible protein data [21].

Authors' contributions

CDP performed the analyses. DCK, JED, and HNBM were all involved with the study's conception and design. All authors read, edited, and approved the final manuscript.

Funding

This work was supported by funding from NSF 1419282 (Moseley), NSF 1453168 (DeRouchey), and NIH UL1TR001998-01 (Kern). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets, figures, scripts, and supplemental information supporting the conclusions of this article are available in the figshare repository, https://doi.org/10.6084/m9.figshare.10292573.v1 [40]. All scripts used in this paper are available under the BSD 3-Clause Clear License.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Chemistry, University of Kentucky, 161 Jacobs Science Building, 40506, Lexington, USA. [2]Markey Cancer Center, University of Kentucky, 800 Rose Street, Pavilion CC 40536, Lexington, USA. [3]Department of Molecular & Cellular Biochemistry, University of Kentucky, 40508, Lexington, USA. [4]Institute for Biomedical Informatics, University of Kentucky, 40536, Lexington, USA.

Received: 3 December 2019 Accepted: 17 March 2020
Published online: 03 April 2020

References

1. Balhorn R. The protamine family of sperm nuclear proteins. Genome Biol. 2007;8(9):227.
2. Steger K, Balhorn R. Sperm nuclear protamines: A checkpoint to control sperm chromatin quality. Anat Histol Embryol. 2018;47(4):273–9.
3. Bennetts LE, Aitken RJ. A comparative study of oxidative dna damage in mammalian spermatozoa. Mol Reprod Dev Inc Gamete Res. 2005;71(1):77–87.
4. Villani P, Eleuteri P, Grollino MG, Rescia M, Altavista P, Spanò M, Pacchierotti F, Cordelli E. Sperm dna fragmentation induced by dnase i and hydrogen peroxide: an in vitro comparative study among different mammalian species. Reproduction. 2010;140(3):445.
5. Enciso M, Johnston SD, Gosálvez J. Differential resistance of mammalian sperm chromatin to oxidative stress as assessed by a two-tailed comet assay. Reprod Fertil Dev. 2011;23(5):633–7.
6. Biegeleisen K. The probable structure of the protamine–dna complex. J Theor Biol. 2006;241(3):533–40.
7. Queralt R, Adroer R, Oliva R, Winkfein R, Retief J, Dixon G. Evolution of protamine p1 genes in mammals. J Mol Evol. 1995;40(6):601–7.
8. Lüke L, Tourmente M, Roldan ER. Sexual selection of protamine 1 in mammals. Mol Biol Evol. 2016;33(1):174–84.
9. Brewer LR. Deciphering the structure of dna toroids. Integr Biol. 2011;3(5):540–7.
10. Arellano A, Canales M, Jullian C, Brunet JE. Fluorescence studies on clupein protamines: evidence for globular conformation. Biochem Biophys Res Commun. 1988;150(2):633–9.
11. Cid H, Arellano A. Secondary structure prediction of protamines. Int J Biol Macromol. 1982;4(1):3–8.
12. Soler-Ventura A, Gay M, Jodar M, Vilanova M, Castillo J, Arauz-Garofalo G, Villarreal L, Ballescà JL, Vilaseca M, Oliva R. Characterization of human sperm protamine proteoforms through a combination of top-down and bottom-up mass spectrometry approaches. J Proteome Res. 2019;19(1):221–237. https://doi.org/10.1021/acs.jproteome.9b00499.
13. Balhorn R, Corzett M, Mazrimas J, Watkins B. Identification of bull protamine disulfides. Biochemistry. 1991;30(1):175–81.
14. Vilfan ID, Conwell CC, Hud NV. Formation of native-like mammalian sperm cell chromatin with folded bull protamine. J Biol Chem. 2004;279(19):20088–95.

15. Hutchison JM, Rau DC, DeRouchey JE. Role of disulfide bonds on dna packaging forces in bull sperm chromatin. Biophys J. 2017;113(9):1925–33.
16. Bench G, Corzett M, Kramer C, Grant P, Balhorn R. Zinc is sufficiently abundant within mammalian sperm nuclei to bind stoichiometrically with protamine 2. Mol Reprod Dev Inc Gamete Res. 2000;56(4):512–9.
17. Pfeifer H, Conrad M, Roethlein D, Kyriakopoulos A, Brielmeier M, Bornkamm GW, Behne D. Identification of a specific sperm nuclei selenoenzyme necessary for protamine thiol cross-linking during sperm maturation. FASEB J. 2001;15(7):1236–8.
18. Conrad M, Moreno S, Sinowatz F, Ursini F, Kölle S, Roveri A, Brielmeier M, Wurst W, Maiorino M, Bornkamm G. The nuclear form of phospholipid hydroperoxide glutathione peroxidase is a protein thiol peroxidase contributing to sperm chromatin stability. Mol Cell Biol. 2005;25(17):7637–44.
19. Retief JD, Rees JS, Westerman M, Dixon GH. Convergent evolution of cysteine residues in sperm protamines of one genus of marsupials, the planigales. Mol Biol Evol. 1995;12(4):708–12.
20. Whittington A, Parkinson A, Spencer P, Grigg G, Hinds L, Gallagher C, Kuchel P, Agar N. Comparative study of the antioxidant defence systems in the erythrocytes of australian marsupials and monotremes. Comp Biochem Physiol Part C Pharmacol Toxicol Endocrinol. 1995;110(3):267–72.
21. Consortium U. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47(D1):506–15.
22. Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
23. Bayse GS, Michaels AW, Morrison M. The peroxidase-catalyzed oxidation of tyrosine. Biochim Biophys Acta BBA Enzymol. 1972;284(1):34–42.
24. Heinecke JW. Tyrosyl radical production by myeloperoxidase: a phagocyte pathway for lipid peroxidation and dityrosine cross-linking of proteins. Toxicology. 2002;177(1):11–22.
25. Mai K, Smith NC, Feng Z-P, Katrib M, Šlapeta J, Šlapetova I, Wallach MG, Luxford C, Davies MJ, Zhang X, et al. Peroxidase catalysed cross-linking of an intrinsically unstructured protein via dityrosine bonds in the oocyst wall of the apicomplexan parasite, eimeria maxima. Int J Parasitol. 2011;41(11):1157–64.
26. Lee J, Ju M, Cho OH, Kim Y, Nam KT. Tyrosine-rich peptides as a platform for assembly and material synthesis. Adv Sci. 2019;6(4):.
27. Martinie RJ, Godakumbura PI, Porter EG, Divakaran A, Burkhart BJ, Wertz JT, Benson DE. Identifying proteins that can form tyrosine-cysteine crosslinks. Metallomics. 2012;4(10):1037–42.
28. E Kasinsky H, Maria Eirin-Lopez J, Ausió J. Protamines: structural complexity, evolution and chromatin patterning. Protein Pept Lett. 2011;18(8):755–71.
29. Green G, Balhorn R, Poccia D, Hecht N. Synthesis and processing of mammalian protamines and transition proteins. Mol Reprod Dev. 1994;37(3):255–63.
30. de Mateo S, Ramos L, de Boer P, Meistrich M, Oliva R. Protamine 2 precursors and processing. Protein Pept Lett. 2011;18(8):778–85.
31. Lüke L, Tourmente M, Dopazo H, Serra F, Roldan ER. Selective constraints on protamine 2 in primates and rodents. BMC Evol Biol. 2016;16(1):21.
32. Bunick D, Balhorn R, Stanker LH, Hecht NB. Expression of the rat protamine 2 gene is suppressed at the level of transcription and translation. Exp Cell Res. 1990;188(1):147–52.
33. Maier W-M, Nussbaum G, Domenjoud L, Klemm U, Engel W. The lack of protamine 2 (p2) in boar and bull spermatozoa is due to mutations within the p2 gene. Nucleic Acids Res. 1990;18(5):1249–54.
34. Lüke L, Vicens A, Serra F, Luque-Larena JJ, Dopazo H, Roldan ER, Gomendio M. Sexual selection halts the relaxation of protamine 2 among rodents. PLoS One. 2011;6(12):.
35. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23(15):1875–82.
36. Cover TM, Thomas JA. Elements of information theory. 2012. Wiley.
37. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol. 2000;303(1):61–76.
38. Eirín-López JM, Ausio J. Origin and evolution of chromosomal sperm proteins. Bioessays. 2009;31(10):1062–70.
39. Inc PT. Collaborative data science. Montreal: Plotly Technologies Inc Montral; 2015.
40. Powell CD, Moseley HN. Entropy Based Analysis of Vertebrate Sperm Protamines Sequences: Evidence of Dityrosine and Cysteine-Tyrosine Cross-Linking in Sperm Protamines. 2019. https://doi.org/10.6084/m9.figshare.10292573.v1.

Publisher's Note