

UC Irvine

UC Irvine Previously Published Works

Title

Early origin of the human seminal plasma proteome during vertebrate evolution

Permalink

<https://escholarship.org/uc/item/15d4t654>

Journal

Genome Biology and Evolution, 18(9)

ISSN

1759-6653

Authors

Ranz, Jose M

Civetta, Alberto

Publication Date

2026-09-18

DOI

10.1093/gbe/evag236

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

**Early origin of the human seminal plasma proteome during
vertebrate evolution**

Jose M. Ranz^{1,2*} and Alberto Civetta^{3*}

¹ Department of Ecology and Evolutionary Biology, University of California Irvine, CA 92697, USA
² Department of Systems Biology, University of California Irvine, CA 92697, USA
³ Department of Biology, University of Winnipeg, Winnipeg, MB R3B 2E9, Canada

* To whom correspondence may be addressed:
Jose Ranz (jranz@uci.edu) and Alberto Civetta (a.civetta@uwinnipeg.ca)

Abstract

Seminal Fluid Proteins (SFPs)-encoding genes are rapidly evolving at the sequence level and often considered prone to frequent phylogenetic turnover. One mechanism underlying such rapid turnover is the lineage-restricted origin of new SFP genes, which results in a complement of phylogenetically young genes. Recent advances in genome resources and comparative genomics now enable more comprehensive cross-taxa analyses, providing improved estimates of the timing of gene origin during evolution. Using these resources alongside proteomic data from human seminal plasma, we traced the evolutionary origins of 646 putative human SFP genes. We find that most SFPs trace back to the bony vertebrate (Eutelostomi) radiation. These ancient SFP genes are more pleiotropic than younger ones, which are enriched in immunity-related functions and regulate proteolytic activity, particularly peptidase inhibitors. Together, these results underscore that the human proteome components of the seminal plasma were originated early during vertebrate evolution, providing insights into the history of the reproductive biology in deuterostomes.

Keywords: Seminal plasma proteome, vertebrate evolution, gene age, humans

Significance Statement

Seminal fluid proteins (SFPs) influence a wide range of traits affecting male and female post-mating fitness, contributing as well to the reproductive isolation between populations. SFP-encoding genes typically exhibit rapid sequence evolution, and their origination patterns through the phylogeny drive ejaculate protein composition divergence across species. Here, we present a comprehensive analysis of human seminal plasma proteins, tracing their evolutionary origins across vertebrates. We find that the human SFP gene complement is overall evolutionary ancient, with many genes present already in the genome of the common ancestor of bony fishes and humans. Notably, the evolutionary age of SFP genes is tightly associated with their functional breadth: ancient SFP genes tend to display a wider range of functions, while younger SFP genes are enriched primarily for immune and proteolytic functions. Our results highlight the contribution of an

evolutionarily ancient gene repertoire to the contemporary reproductive functions of human SFPs.

Introduction

Seminal Fluid Proteins (SFPs) are transferred to females during copulation where they exert a number of important post-mating physiological and behavioral effects. Although these responses have been primarily studied in insects (Avila, et al. 2011), mammalian SFPs have also been shown to facilitate fertilization by protecting the sperm against immune challenges, mediate the formation of a sperm-storage reservoir, and trigger female responses that can induce ovulation, among others (McGraw, et al. 2015; Schjenken and Robertson 2020). Notably, despite their central role in fertilization, SFP genes show, on average, rapid sequence evolution that is phylogenetically widespread (Clark and Swanson 2005; Ramm, et al. 2008; Swanson and Vacquier 2002). In addition to this overall pattern of rapid sequence evolution, there is evidence of a major overhaul of the secretory function of tissues and anatomy of the glands that produce the seminal fluid (McGraw, et al. 2015). However, our understanding of the evolution of SFP genes is largely incomplete, particularly in vertebrates, taxa in which we lack a systematic analysis of how the protein complement of the seminal fluid arose across key evolutionary transitions. Recently, *D. melanogaster* SFPs had been shown to have 62% of its encoding genes originating before the radiation of the genus *Drosophila* (Ranz, et al. 2025). Whether this pattern of ancient origin holds beyond protostomes remains untested.

Unlike previous work, which has largely emphasized the divergence and turnover of these genes, our focus here is to identify potentially phylogenetically old members of the SFP gene complement and to evaluate their prevalence. Specifically, by examining human SFPs phylogenetic origin, we address whether the previously reported pattern of a phylogenetically old SFP gene complement found for *D. melanogaster* is an oddity restricted to protostomes or a more universal pattern that extends to deuterostomes. First, we capitalize on proteomic characterizations of human seminal plasma from semen samples from two different studies (Batruch, et al. 2011; Pilch and Mann 2006) to delineate an updated list of 646 human SFPs common to both studies. Then, using upgraded genomic resources from 23 vertebrate species (Chen, et al. 2025; Shao, et al. 2019), we

establish the branch of origin of human SFP genes. We find that the majority of human SFPs are phylogenetically ancient and that the proportion of ancient genes among SFPs is overrepresented relative to the rest of the genome. The potential repurposing of the molecular machinery underlying male reproductive functions has important implications for the evolution of reproductive systems across the animal kingdom.

Results and Discussion

We retrieved information from two different proteomic studies of seminal plasma. The first identified 879 non-redundant seminal plasma proteins obtained from three ejaculates of a single healthy donor, while the second documented 1,923 non-redundant human SFPs obtained from pooling ejaculates from different healthy individuals (Batruch, et al. 2011; Pilch and Mann 2006). Among the 2,156 non-redundant list of SFPs derived from combining both initial lists, we found a common set of 646 proteins to both studies (Fig. 1a). Due to methodological differences in the mass spectrometry experiments performed in the two studies, and the potential presence of broadly expressed housekeeping proteins that are not necessarily SFPs, for example carry-over proteins shared among a variety of secretory pathways, we examined both the non-redundant list and the common set of proteins for characteristics expected among reputed human SFPs. Specifically, we found that the common set of 646 proteins exhibits a significant increase in the proportion of proteins associated with the functional terms extracellular region (GO:0005576) and extracellular space (GO:0005615) compared to the total 2,156 proteins, consistent with the possibility of an enrichment in the common set of soluble proteins (746/2156 or 34.6% vs 353/646 or 54.7%; two-sample test for equality of proportions with continuity correction (2TEP): $\chi^2 = 83.52$, d.f. = 1, $P < 2.2 \times 10^{-16}$; Fig. 1b). In addition, we delineated a reinforced subset by considering peptide sequence coverage and number of peptides identified for each protein. Specifically, we identified a list of 340 SFPs that had at least 2 peptides per protein in both studies (reinforced list; Fig. 1a). This reinforced list of SFPs featured an even tighter association with the mentioned cellular compartment GO terms (353/645 or 54.7% vs 219/340 or 64.4%; %; 2TEP: $\chi^2 = 8.2$, d.f. = 1, $P = 4.2 \times 10^{-03}$) (Fig. 1b).

We reasoned that any properly curated list of human SFPs should account for most of the protein content of the samples analyzed. Using the protein level estimates across the three

seminal samples assayed by Batruch et al. 2011 (Batruch, et al. 2011), we calculated the average relative abundance of each putative SFP in our common set. The proteins in the common set accounted for 86.6% of the total protein content of the samples analyzed by Batruch et al. 2011 (Batruch, et al. 2011). Examining the interplay between protein abundance and the common list of SFPs revealed the presence of 367 of the 506 proteins (i.e. 56.8%) in the top quartile most abundant proteins. The top 10% of the most abundant proteins present in our common list, i.e. 64, accounting for 68.4% of the total protein content (Fig. 1c), did not differ from the content represented by the top 10% in the reinforced list (34, accounting for 63.4%) (two-proportion z-test, $z=1.58$, $P=0.113$). The top 10 most abundant putative human SFPs were identical across both the common and the reinforced lists. These included proteins known to be relevant for semen coagulation after ejaculation (SEMG1 and SEMG2) (Anamthathmakula and Winuthayanon 2020), structural components of the seminal coagulum (FN1) (Lilja, et al. 1987), bacterial protection (LTF and PIP) (Gopalkrishnan, et al. 1994; Nowicka-Bauer, et al. 2023), seminal coagulum liquefaction (KLK3) (Anamthathmakula and Winuthayanon 2020), and sperm capacitation roles (ALB and NPC2) (Ozkocer and Konac 2021; Zhao, et al. 2021) (Fig. 1d). Importantly, neither the common set nor the reinforced set excluded other much less abundant, but well-known, SFPs, such as prostasin (PRSS8) and the prostate stem cell antigen (PSCA) (Reiter, et al. 1998; Takahashi, et al. 2003). Collectively, we concluded that our filtering strategy effectively controlled for false positives while retaining reputed SFPs.

Subsequently, to date the branch of origin during the vertebrate phylogeny of our list of SFPs, we leveraged a comprehensive human gene age dating, which combines microsynteny information across 23 vertebrates within a framework of maximum parsimony, to date the origin of 19,665 protein-coding human genes across six evolutionary transitions following the Euteleostomi: Tetrapoda, Amniota, Mammalia, Theria, Eutheria, and Primate (Chen, et al. 2025; Shao, et al. 2019). The combination of the proteomics and gene dating datasets resulted in a catalog of 615 human SFP-encoding genes dated across seven age classes (Fig. 2a). For example, relative to the examples of known SFPs mentioned above, their respective encoding genes were originated during the evolutionary transition of the Euteleostomi (FN1, NPC2, and PRSS8), during that of the Tetrapoda (ALB), during that of the Mammalia (PIP), or during that of the Eutheria (KLK3, LTF, PSCA, SEMG1 and SEMG2). To determine whether this diversity in time of origin is representative of SFP genes more broadly, we evaluated

whether they are proportionally represented across evolutionary transitions relative to the abundance of the different gene age classes in the genome, finding that this was not the case (chi-square goodness-of-fit, $\chi^2 = 48.23$, d.f. = 6, $P = 1.06 \times 10^{-08}$). Importantly, post hoc tests further uncovered an excess of human SFP genes that were present already in the ancestral genome of the Euteloestomi, i.e. branch 0, prior to 429 mya (Kumar, et al. 2022), and a deficit in the three youngest evolutionary transitions (Fig. 2a). This pattern of branch of origin can be further simplified as a contrast between the branch corresponding to the Euteloestomi and older vs the rest (i.e. *br0* vs *br1-6*; 2-tailed exact binomial test, observed proportion of success = $410/615 = 0.667 > \text{expected} = 0.566$, with a 95% CI of 0.628-0.704, $P=4.16 \times 10^{-7}$) (Fig. 2b).

Further, we focused on the reinforced subset of 328 human SFPs. We also found a departure from the random expectation (chi-square goodness-of-fit, $\chi^2 = 168.52$, d.f. = 6, $P < 2.2 \times 10^{-16}$) (Fig. 2c). Subsequent post hoc tests confirmed the enrichment for human SFPs originated in branch 0, which was further supported by the two-categories test (*br0* vs *br1-6*; 2-tailed exact binomial test, observed proportion of success = $216/328 = 0.659 > \text{expected} = 0.566$, with a 95% CI of 0.604-0.71, $P=6.7 \times 10^{-4}$) (Fig. 2d). Post hoc tests also confirmed an underrepresentation of SFPs originated in recent evolutionary transitions, but unlike for the common list, this signal was only significant for the Primate radiation, possibly as a result of the most limited statistical power associated with the reinforced list of SFPs. Overall, our findings suggest that many of the human SFPs were already present in the ancestor to all vertebrate lineages.

One potential confounding factor to consider is a nonrandom association between SFP abundance, and therefore probability of detection in proteomic analysis, and gene age. If the most ancient SFPs were characterized by higher levels of expression compared to younger SFPs, then ancient SFP would be more likely to be detected, thus explaining the statistical enrichment detected. To ascertain this possibility, we tested for differences in relative abundance (Batruch, et al. 2011) among SFPs across different age classes. The analysis of the common set did not reveal any significant difference in relative abundance across the seven age classes (Kruskal–Wallis rank-sum test, $\chi^2 = 5.92$, d.f. = 6, $P = 0.433$) nor when considering only two broadly defined categories, i.e. *br0* vs *br1-6*, (Kruskal–Wallis rank-sum test, $\chi^2 = 1.33$, d.f. = 1, $P = 0.248$). The equivalent analyses with the reinforced set did not offer a substantially different picture, with the comparison across seven age classes not entailing any significant difference either (Kruskal–Wallis rank-sum

test, $\chi^2 = 10.67$, d.f. = 6, $P = 0.099$) and the test between the two broadly defined categories revealing significant differences in abundance but not in the direction of the proposed confounding scenario as genes in the age category *br1-6* were more highly expressed than those in the *br0* category (Kruskal–Wallis rank-sum test, $\chi^2 = 4.12$, d.f. = 1, $P = 0.042$). Collectively, these results do not support that the SFPs originated in the ancestor to all vertebrates are more abundant, being therefore more easily detectable, thus explaining their excess in relation to SFPs from younger age classes.

As gene age is associated with gene properties that reflect different functional constraints, we hypothesized that older human SFP genes should be more pleiotropic than younger SFP genes (Hao, et al. 2010). We examined two different proxies of pleiotropy, breadth of tissue expression and gene ontology, thus reducing dependence on any single measure of pleiotropy. We found that ancient genes originating during the Eutelostomi radiation have a lower tissue-specificity (τ —tau index; Fig. 3) and a more diverse array of gene ontologies (Fig. 4a; Table S1), particularly in relation to biological processes and molecular functions, compared to those originated subsequently within other evolutionary transitions. These results were consistent across the list of SFPs considered. Importantly, as older genes are known to have a more diverse array of gene ontologies (Hao et al. 2010), we replicated our analyses by searching SFP enrichment patterns relative to the rest of genes within the same age category instead of using all the genes in the genome as a background reference set (Fig. 4b; Table S2). In this second set of analyses, we found the same result, i.e. a broader collection of ontologies associated with ancient SFPs compared to younger SFPs (Fig. 4b). The biological process, molecular function, and cellular component enrichment patterns of the evolutionarily younger SFP gene complement are unique, showing a distinctive enrichment of proteolytic activity (e.g. proteasome complex and peptidase activity) and immune terms such as antibacterial humoral response (Fig. 4a and 4b). Moreover, the specific association of 'specific granule lumen', a compartment characteristic of neutrophils containing antimicrobial proteins (Okuda, et al. 1981), with younger SFPs further supports an immune-related profile. The enrichment for immune-related and protease terms is likely related to inflammatory responses in mammals that may aid in fertilization (McGraw, et al. 2015). Protease inhibitors, in particular, may protect sperm from proteolysis, thereby promoting immune tolerance and supporting pregnancy (Laflamme and Wolfner 2013). The younger components of the human seminal plasma proteome reflect a more recent evolutionary

207 acquisition of genes specialized in post-mating functions, which may bias reproductive
208 outcomes in contexts such as male-male competition.

209 **Conclusions**

210 Our results in humans, and previously in the protostome *D. melanogaster* (Ranz, et al.
211 2025), suggest that a substantial fraction of SFP genes are not evolutionary young.
212 Instead, our results reveal a deep evolutionary origin that spans major transitions during
213 the evolution of the vertebrates. These findings open new avenues for exploring how
214 ancient molecular frameworks have been used across the diversity of reproductive and
215 mating systems seen across the vertebrate phylogeny. Particularly, it remains to be
216 clarified whether the molecular mechanisms whereby SFPs impact reproductive functions
217 are ancient, and therefore potentially widespread across multiple contemporary species,
218 or have been more recently coopted, therefore being associated to specific lineages.
219 Distinguishing between both scenarios will require extensive proteomic characterization
220 of seminal plasma across the vertebrate phylogeny.

Material and Methods

Sfp gene data.

We retrieved data from two independent studies that characterized the human seminal plasma proteome from one (Pilch and Mann 2006) and three healthy individuals (Batruch, et al. 2011), respectively. Both studies employ mass spectrometry and rigorous filtering strategies to ensure high-confidence protein identification, primarily relying on score-based validation derived from database search engines such as Mascot (Perkins, et al. 1999). Additionally, both workflows incorporate layers of validation, such as replicate analyses and manual spectrum inspection, to improve the reliability of their results. Further, a reinforced list was obtained by excluding proteins with less than two peptide counts in either study. Additionally, this reinforced list was filtered by using coverage information provided by Batruch et al. 2011 (Batruch, et al. 2011) to require a minimum sequence coverage of 10%, thus retaining more robust protein identifications for downstream analyses. Average protein abundance was calculated using the data from Batruch et al. 2011 (Batruch, et al. 2011).

Tau values for human genes were taken as calculated (Palmer, et al. 2021), which were based on the GTEx (version 8, <https://www.gtexportal.org/home/>) expression values for genes and transcripts. Tau values range from 0 to 1, from broader to more restricted expression, respectively (Yanai, et al. 2005). Gene Ontology (GO) enrichment was assessed using DAVID (Sherman, et al. 2022) and redundant terms were clustered under parental terms using the rrvgo R package (Sayols 2023), using a similarity threshold of 0.7 to reduce redundancy among GO terms. As the choice of the background gene set can substantially influence the enrichment results and their interpretation, the analysis was conducted using all protein-coding genes in the genome, and only the protein-coding genes present in the analyzed age category, as the background reference gene set. *P*-values were corrected for multiple tests following Benjamini-Hochberg (Benjamini and Hochberg 1995). Only functional terms with an associated adjusted *p*-value below 0.01 were considered. For each parental term, the enrichment score ($-\log_{10}(P_{adj})$) and the fold enrichment ($\log_2(FE)$) were calculated.

Phylogenetic gene age data.

The gene age information for 19,665 protein-coding human genes across seven evolutionary transitions or age classes along the vertebrate phylogeny was taken from

Chen et al. 2025 (Chen, et al. 2025). The authors used the Ensembl hg38_ver95 release, and an upgraded version of GenTree (<http://gentree.ioz.ac.cn/>) (Shao, et al. 2019), which relies on multiple pairwise sequence alignments between humans and 22 additional species: chimp, orangutan, gibbon, rhesus monkey, marmoset, rabbit, guinea pig, rat, mouse, horse, dog, cow, elephant, opossum, platypus, chicken, zebra finch, lizard, western clawed frog, stickleback, pufferfish, and zebrafish. Briefly, GenTree uses microsynteny information and a parsimony-based framework assigning gene origin at the deepest phylogenetic node. This approach is robust across genome releases and insensitive to the quality of gene annotation (Shao, et al. 2019).

Statistical analyses.

Test for chi-square goodness-of-fit, analysis of residuals, Benjamini–Hochberg correction for multiple testing, two-sample test for equality of proportions with continuity correction, two-proportion z-test, exact binomial test, and Kruskal–Wallis rank-sum test were conducted using built-in functions in R (R Development Core Team 2016).

Declaration of Generative AI and AI-Assisted Technologies.

During the preparation of this manuscript, the authors used ZotGPT (Anthropic Claude 4.6 Sonnet), UCI's AI assistant, to improve the readability and language of the text. Upon using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

All data and materials used are included or cited in the article, supporting Tables S1-S2, and in a dataset stored at Dryad repository (http://datadryad.org/share/LINK_NOT_FOR_PUBLICATION/kWHUa5lOia2rHKMRpz4MADlwYfR6aEEs2l3y-RHSQo).

Figure Legends

Figure 1. Delineation of a catalog of human SFPs.

(a) Relationship of the two proteomic datasets used to delineate a catalog of human SFP-encoding genes (Batruch, et al. 2011; Pilch and Mann 2006). SFPs detected in both datasets, i.e. the common dataset, were considered in downstream analyses. Additionally, genes encoding proteins supported by two peptides defined a reinforced set of SFP genes. (b) Percentage of genes in different sets that have associated the functional terms extracellular region (GO:0005576) and extracellular space (GO:0005615), proxies for soluble properties that characterize SFPs. (c) Cumulative fraction of the averaged protein content across three samples of seminal plasma (Batruch, et al. 2011) associated with the 200 most abundant proteins in the common set of putative human SFPs. The 10% most abundant putative SFPs among the 646 that define the common set, which account for 68.4% of all protein content (Batruch, et al. 2011), are indicated. (d) Relative averaged fraction of the protein content across seminal samples (Batruch, et al. 2011) represented by the 200 most abundant proteins in the common set of putative human SFPs delineated here. The top 10 most abundant human SFPs are indicated: ACPP, Isoform 1 of Prostatic acid phosphatase; ALB, Isoform 1 of Serum albumin; AZGP1, alpha-2-glycoprotein 1, zinc; FN1, Isoform 15 of Fibronectin; KLK3, Prostate-specific antigen; LTF, Truncated lactoferrin; MUC6, mucin 6, gastric; PIP, Prolactin-inducible protein; SEMG1, Semenogelin-1; and SEMG2, Semenogelin-2.

Figure 2. Origin of human SFP encoding genes across the phylogeny of the vertebrates.

Phylogeny showing the seven phylogenetic transitions considered in order to determine the time of origin of human Sfp genes (Chen, et al. 2025) for the common set of proteins to two proteomic studies (Batruch, et al. 2011; Pilch and Mann 2006) (a) and for a reinforced set that includes only those proteins represented by at least two peptides in both studies (c). The seven evolutionary transitions or age classes are associated with seven numbered branches (from *br0* to *br6*) following Chen et al. 2025 (Chen, et al. 2025). For each branch, it is indicated the number of genes originated, the standardized residual value from the chi-square goodness-of-fit, and the trend when significant according to post hoc tests at a 5% FDR. ↑ and ↓ denote enrichment and depletion, respectively, relative to

the share of the different age classes across the whole genome. **(b and d)** Observed number of human SFP genes originated in the branch of the Euteleostomi and older, i.e. *br0*, and the expected number according to the share of this age class vs the rest of age classes (i.e. *br1-br6*) in the entire genome following Chen et al. 2025 (Chen, et al. 2025). Error bar, upper and lower boundaries of the 95% CI for the observed counts according to the boundaries obtained with an exact binomial test.

Figure 3. The expression breadth of human SFP genes increases with their evolutionary age.

Violin (foreground) and box (background) plots showing the distribution of the tau expression-specificity index of SFP-encoding genes of different age classes. **(a)** results for SFP genes that are part of the common dataset from two independent studies (Batruch, et al. 2011; Pilch and Mann 2006); **(b)** results after filtering by peptide support. Boxes represent the interquartile range (IQR) around the median, shown as black dots and with values displayed above the violin. The higher tau values indicate an increase in tissue-specific expression, a proxy of reduced pleiotropy. The *p*-value associated with the Kruskal–Wallis rank-sum test for differences across age classes is provided on top.

Figure 4. The breadth of functional enrichment patterns is larger for older than for younger human SFP genes.

(a) Biological process, molecular function, and cellular compartment Gene Ontology terms enrichment for SFP genes of the common dataset across two independent studies (Batruch, et al. 2011; Pilch and Mann 2006) (left panel) and after filtering by peptide support (right panel), using all the genes in the genome as a background reference gene set **(b)**. Like **(a)** but using the genes within the same gene age category as a background reference gene set. The plots show a larger number of terms' enrichment for genes that originated at the time of the bony vertebrate (Euteleostomi) radiation (*br0*) relative to those with later origin (*br1-6*). For each parental term, the enrichment score ($-\log_{10}(P_{adj})$) and the fold enrichment ($\log_2(FE)$) are shown.

References

- Anamthathmakula P, Winuthayanon W 2020. Mechanism of semen liquefaction and its potential for a novel non-hormonal contraception dagger. *Biol Reprod* 103: 411-426. doi: 10.1093/biolre/ioaa075
- Avila FW, Sirot LK, LaFlamme BA, Rubinstein CD, Wolfner MF 2011. Insect seminal fluid proteins: identification and function. *Annu Rev Entomol* 56: 21-40. doi: 10.1146/annurev-ento-120709-144823
- Batruch I, et al. 2011. Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system. *J Proteome Res* 10: 941-953. doi: 10.1021/pr100745u
- Benjamini Y, Hochberg Y 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* 57: 289-300.
- Chen JH, et al. 2025. Evolutionarily new genes in humans with disease phenotypes reveal functional enrichment patterns shaped by adaptive innovation and sexual selection. *Genome Res* 35: 379-392. doi: 10.1101/gr.279498.124
- Clark NL, Swanson WJ 2005. Pervasive adaptive evolution in primate seminal proteins. *Plos Genetics* 1: e35. doi: 10.1371/journal.pgen.0010035
- Gopalkrishnan K, Joseph R, Sheth AR 1994. Alteration of semen characteristics and regulatory factors in human semen with bacterial infection. *Arch Androl* 32: 213-218. doi: 10.3109/01485019408987788
- Hao L, et al. 2010. Human functional genetic studies are biased against the medically most relevant primate-specific genes. *BMC Evol Biol* 10: 316. doi: 10.1186/1471-2148-10-316
- Kumar S, et al. 2022. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* 39. doi: 10.1093/molbev/msac174
- Laflamme BA, Wolfner MF 2013. Identification and function of proteolysis regulators in seminal fluid. *Mol Reprod Dev* 80: 80-101. doi: 10.1002/mrd.22130
- Lilja H, Oldbring J, Rannevik G, Laurell CB 1987. Seminal vesicle-secreted proteins and their reactions during gelation and liquefaction of human semen. *J Clin Invest* 80: 281-285. doi: 10.1172/JCI113070
- McGraw LA, Suarez SS, Wolfner MF 2015. On a matter of seminal importance. *Bioessays* 37: 142-147. doi: 10.1002/bies.201400117
- Nowicka-Bauer K, Kamieniczna M, Olszewska M, Kurpisz MK 2023. Proteomic approach towards identification of seminal fluid biomarkers from individuals with severe oligozoospermia, cryptozoospermia and non-obstructive azoospermia: a pilot study. *Transl Androl Urol* 12: 1497-1510. doi: 10.21037/tau-23-130

382 Okuda M, Takenaka T, Kawabori S, Ogami Y 1981. Ultrastructural study of the specific
383 granule of the human eosinophil. *J Submicrosc Cytol* 13: 465-471.

384 Ozkocer SE, Konac E 2021. The current perspective on genetic and epigenetic factors in
385 sperm maturation in the epididymis. *Andrologia* 53: e13989. doi: 10.1111/and.13989

386 Palmer D, Fabris F, Doherty A, Freitas AA, de Magalhaes JP 2021. Ageing
387 transcriptome meta-analysis reveals similarities and differences between key
388 mammalian tissues. *Aging (Albany NY)* 13: 3313-3341. doi: 10.18632/aging.202648

389 Perkins DN, Pappin DJ, Creasy DM, Cottrell JS 1999. Probability-based protein
390 identification by searching sequence databases using mass spectrometry data.
391 *Electrophoresis* 20: 3551-3567. doi: 10.1002/(SICI)1522-
392 2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

393 Pilch B, Mann M 2006. Large-scale and high-confidence proteomic analysis of human
394 seminal plasma. *Genome Biol* 7: R40. doi: 10.1186/gb-2006-7-5-r40

395 R Development Core Team. 2016. R: A language and environment for statistical
396 computing. Vienna, Austria: R Foundation for Statistical Computing.

397 Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD 2008. Sexual selection and the
398 adaptive evolution of mammalian ejaculate proteins. *Mol Biol Evol* 25: 207-219. doi:
399 10.1093/molbev/msm242

400 Ranz JM, Flacchi C, Hariyani IE, Civetta A 2025. Gene age shapes functional and
401 evolutionary properties of the *Drosophila* seminal fluid proteome. *Proc Natl Acad Sci U S*
402 *A* 122: e2505490122. doi: 10.1073/pnas.2505490122

403 Reiter RE, et al. 1998. Prostate stem cell antigen: A cell surface marker overexpressed
404 in prostate cancer. *Proceedings of the National Academy of Sciences of the United*
405 *States of America* 95: 1735-1740. doi: DOI 10.1073/pnas.95.4.1735

406 Sayols S 2023. rrvgo: a Bioconductor package for interpreting lists of Gene Ontology
407 terms. *MicroPubl Biol* 2023. doi: 10.17912/micropub.biology.000811

408 Schjenken JE, Robertson SA 2020. The Female Response to Seminal Fluid. *Physiol*
409 *Rev* 100: 1077-1117. doi: 10.1152/physrev.00013.2018

410 Shao Y, et al. 2019. GenTree, an integrated resource for analyzing the evolution and
411 function of primate-specific coding genes. *Genome Res* 29: 682-696. doi:
412 10.1101/gr.238733.118

413 Sherman BT, et al. 2022. DAVID: a web server for functional enrichment analysis and
414 functional annotation of gene lists (2021 update). *Nucleic Acids Res* 50: W216-W221.
415 doi: 10.1093/nar/gkac194

416 Swanson WJ, Vacquier VD 2002. The rapid evolution of reproductive proteins. *Nat Rev*
417 *Genet* 3: 137-144.

418 Takahashi S, et al. 2003. Down-regulated expression of prostaticin in high-grade or
419 hormone-refractory human prostate cancers. Prostate 54: 187-193. doi:
420 10.1002/pros.10178

421 Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level
422 relationships in human tissue specification. Bioinformatics 21: 650-659. doi:
423 10.1093/bioinformatics/bti042

424 Zhao R, et al. 2021. Direct activation of the proton channel by albumin leads to human
425 sperm capacitation and sustained release of inflammatory mediators by neutrophils. Nat
426 Commun 12: 3855. doi: 10.1038/s41467-021-24145-1

427

Acknowledgments

We thank Ben Hopkins and Rama Singh for comments on early versions of the manuscript.

Funding

This work was supported by a National Science Foundation Grant to J.M.R (MCB-2129845) and a Natural Sciences and Engineering Research Council Discovery Grant (RGPIN-2025-04591) to A.C.

Authors contributions

Jose M. Ranz, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing; **Alberto Civetta**, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing.

Additional information

Supplemental information includes two tables.