Functional diversity and evolution of the *Drosophila* sperm proteome

Martin D. Garlovsky[a,✉], **Jessica Sandler**[b], **and Timothy L. Karr**[b,c,✉]

[a]Department of Applied Zoology, Faculty of Biology, Technische Universität Dresden, Dresden 01069, Germany
[b]Biosciences Mass Spectrometry Core Research Facility, Knowledge Enterprise, Arizona State University, USA
[c]Neurodegenerative Disease Research Center, The Biodesign Institute, Arizona State University, USA

Given the central role fertilization plays in the health and fitness of sexually reproducing organisms and the well-known evolutionary consequences of sexual selection and sperm competition, knowledge gained by a deeper understanding of sperm (and associated reproductive tissues) proteomes has proven critical to the field's advancement. Due to their extraordinary complexity, proteome depth-of-coverage is dependent on advancements in technology and related bioinformatics, both of which have made significant advancements in the decade since the last *Drosophila* sperm proteome was published. Here we provide an updated version of the *Drosophila melanogaster* sperm proteome (DmSP3) using improved separation and detection methods and an updated genome annotation. We identified 2563 proteins, with label-free quantitation (LFQ) for 2125 proteins. Combined with previous versions of the sperm proteome, the DmSP3 contains a total of 3176 proteins. The top 20 most abundant proteins contained the structural elements α- and β-tubulins and sperm leucyl-aminopeptidases (S-Laps). Both gene content and protein abundance were significantly reduced on the X chromosome, a finding consistent with prior genomic studies of the X chromosome gene content and evolution. We identified 9 of the 16 Y-linked proteins, including known testis-specific male fertility factors. LFQ measured significant levels for 75/83 ribosomal proteins (RPs) we identified, including a number of core constituents. The role of this unique subset of RPs in sperm is unknown. Surprisingly, our expanded sperm proteome also identified 122 seminal fluid proteins (Sfps), proteins found predominantly in the accessory glands. The possibility of tissue contamination from seminal vesicle or other reproductive tissues was addressed using concentrated salt and detergent treatments. Salt treatment had little effect on sperm proteome composition suggesting only minor contamination during sperm isolation while a significant fraction of Sfps remained associated with sperm following detergent treatment suggesting Sfps may arise within, and have additional functions, in sperm *per se*.

Keywords: Spermatozoa | seminal fluid proteins | ribosomes | meiotic sex chromosome inactivation | fertility | evolution | discovery proteomics | human disease | OMIM | *Drosophila*

Abbreviations: AUC, area-under-the-curve; BP, biological processes; CC, cellular components; CDS, coding sequences; CID, collision-induced fragmentation; DAVID, database for visualisation and integrated discovery; DmSP, *Drosophila melanogaster* sperm proteome; FBgn, FlyBase gene number; FDR, false discovery rate; FPKM, fragments per kilobase of transcript per million mapped reads; GO, Gene ontology; LFQ, Label-free quantitation; MF, molecular functions; MIPS, monoisotopic peak determination; OMIM, Online Mendelian Inheritance in Man; PAML, phylogenetic analysis by maximum likelihood; RPs, ribosomal proteins; S-Laps, sperm leucyl-aminopeptidases; Sfp, Seminal fluid protein; TEAB, triethylammonium bicarbonate

Correspondence: ***martin.garlovsky@tu-dresden.de; tkarr@asu.edu***

Introduction

Spermatozoa form, function and evolution is determined in large measure by its proteome (1). High throughput proteomics using liquid-chromatography tandem mass-spectrometry (LC-MS) has now been used to characterize the composition of the sperm proteome in a wide range of animals (1–3). These studies have revealed several common features of sperm as expected for an ancient cell type with a highly conserved function despite exhibiting exceptional morphological diversity across the tree of life (4–6). For instance, across taxa, sperm show enrichment of metabolic processes, mitochondria, axoneme, microtubules and cytoskeletal components (2, 3, 7, 8).

Recent advances in LC-MS technology, particularly in data acquisition time and improved liquid chromatographic systems, provide enhanced proteome coverage of complex cell and tissue types (9, 10). These advances accordingly allow routine and accurate quantitation of both label- and label-free methodologies, an essential element for comparative studies of sperm composition and function (11, 12). Additionally, these advances permit direct injection of sample peptides without the need for pre-fractionation using polyacrylamide gel electrophoresis thus avoiding sample loss. Accordingly, in the current study we re-interrogated the *Drosophila melanogaster* sperm proteome using direct solubilization of sperm followed by on-line fractionation of tryptic peptides and report on a significant increase in both proteome size and content.

D. melanogaster with its excellent genome annotation provides a powerful genetic and functional genomics model system to understand reproduction and fertility (e.g., (13)). Our previous efforts identified over 1,000 *D. melanogaster* sperm proteins with prior versions designated DmSP1 (7) and DmSP2 (14). The DmSP3 described in this study significantly increases coverage and refinement of the *D. melanogaster* sperm proteome, from the 1108 sperm proteins identified in the combined DmSP1 and DmSP2 (14) to more than 3000 proteins in the DmSP3 (Table 1). Table 1 highlights our extended knowledge base not only in terms of absolute numbers of sperm proteins, but also discovery of new protein groups including the surprising findings of substantial numbers of seminal fluid proteins and ribosomal proteins in the DmSP3. We use the increased proteome coverage and quantitative LFQ information in the DmSP3 to provide a detailed analysis of relative abundance of sperm proteins for the

first time, and re-examine the evolutionary dynamics, gene age, and chromosomal distribution of proteins in the DmSP3. The analyses provide stronger support for previous claims and in particular cements the subjective prior findings supporting the meiotic sex-chromosome inactivation model for male-specific gene and X chromosome evolution (15–18).

Methods

A. Fly stocks and sample preparation. We used laboratory wild-type strain Oregon-R *D. melanogaster* males, aged 5-7 days. All dissections tissue removal and sperm isolation were performed at room temperature in freshly prepared phosphate buffered saline (PBS) with or without protease inhibitors (HALT, Thermo Fisher). We anaesthetized flies and removed reproductive tracts with forceps under a stereo dissecting microscope as previously described (14). Briefly, each biological replicate from 10 males (20 paired seminal vesicles) were prepared separately over the course of no more than one hour by first removing the seminal vesicles from each male reproductive tract (containing testes, seminal vesicles, and accessory glands) into a fresh drop of PBS. Sperm were then carefully removed using fine needles to a 1.5ml microcentrifuge tube containing PBS (on ice). Sperm were then pelleted at 15,000 rpm for 15 minutes at 4ºC, washed 3X with PBS and immediately solubilized in 25 microliters of 5% SDS/50mM TEAB containing 50mM dithiothreitol. Solubilized samples were then incubated for 10-15 minutes at 95ºC and spun again at 15,000 rpm for 15 minutes at 20ºC. No visible pellets were observed and the supernatants removed and stored at -20ºC or immediately processed as described below.

Solubilized sperm proteins were quantified using EZQ Protein Quantitation Kit (Thermo Fisher) and 2.25ug alkylated (Pierce) using 40mM final concentration freshly prepared iodoacetamide for 30 minutes in the dark at room temperature. Samples were processed using the Protifi S-trap Micro Columns and instructions were given via the S-trap Ultra High Recovery Protocol (Protifi). Briefly, samples were acidified by addition of 12% phosphoric acid to a final concentration of$\sim$1.2% phosphoric acid. Proteins were digested by addition of 2.0 µg of porcine trypsin (MS grade, Pierce) and incubated at 30ºC for 2 hours. S-trap buffer (90% methanol, 100 mM TEAB final) was also added in volumes 7X our total sample volume. Acidified sample and the S-trap buffer was filtered through columns. Columns were washed 3X with S-trap buffer. An additional 0.5 µg of trypsin and 25 µL of 50 mM TEAB was added to the top of each column and incubated for 1 hour at 47ºC. Samples were eluted off the S-trap columns using three elution buffers: 50 mM TEAB, 0.2% formic acid in water, and 50% acetonitrile/50% water + 0.2% formic acid. Samples were dried down via speed vac and resuspended in 20-30 µL of 0.1% formic acid.

B. Liquid-chromatography tandem mass-spectrometry. All LC-MS analyses were performed at the Biosciences Mass Spectrometry Core Facility (https://cores.research.asu.edu/mass-spec/) at Arizona State University. All data-dependent mass spectra were collected in positive mode using an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific) coupled with an UltiMate 3000 UHPLC (Thermo Scientific). One µL of peptides were fractionated using an Easy-Spray LC column (50 cm × 75 µm ID, PepMap C18, 2 µm particles, 100 Å pore size, Thermo Scientific) equipped with an upstream 300um x 5mm trap column. Electrospray potential was set to 1.6 kV and the ion transfer tube temperature to 300ºC. The mass spectra were collected using the "Universal" method optimized for peptide analysis provided by Thermo Scientific. Full MS scans (375–1500 *m/z* range) were acquired in profile mode with the Orbitrap set to a resolution of 120,000 (at 200 *m/z*), cycle time set to 3 seconds and mass range set to "Normal". The RF lens was set to 30% and the AGC set to "Standard". Maximum ion accumulation time was set to "Auto". Monoisotopic peak determination (MIPS) was set to "peptide" and included charge states 2-7. Dynamic exclusion was set to 60s with a mass tolerance of 10ppm and the intensity threshold set to 5.0e3. MS/MS spectra were acquired in a centroid mode using quadrupole isolation window set to 1.6 (*m/z*). Collision-induced fragmentation (CID) energy was set to 35% with an activation time of 10 milliseconds. Peptides were eluted during a 240-minute gradient at a flow rate of 0.250 uL/min containing 2-80% acetonitrile/water as follows: 0-3 minutes at 2%, 3-75 minutes 2-15%, 75-180 minutes at 15-30%, 180-220 minutes at 30-35%, 220-225 minutes at 35-80% 225-230 at 80% and 230-240 at 80-5%.

C. Label-free quantification (LFQ). We analysed raw files searched against the Uniprot (www.uniprot.org) *D. melanogaster* database (Dmel_UP000000803.fasta) using Proteome Discover 2.4 (Thermo Scientific). Raw files were searched using SequestHT that included Trypsin as enzyme, maximum missed cleavage site 3, min/max peptide length 6/144, and precursor ion (MS1) mass tolerance set to 20 ppm and fragment mass tolerance set to 0.5 Da and a minimum of 1 peptide identified. Carbamidomethyl (C) was specified as fixed modification, and dynamic modifications set to Aceytl and Met-loss at the N-terminus, and oxidation of Met. A concatenated target/decoy strategy and a false-discovery rate (FDR) set to 1.0% was calculated using Percolator (19). The data was imported into Proteome Discoverer 2.4, and accurate mass and retention time of detected ions (features) using Minora Feature Detector algorithm. The identified Minora features were then used to determine area-under-the-curve (AUC) of the selected ion chromatograms of the aligned features across all runs and relative abundances calculated.

D. Gene ontology enrichment. We performed gene ontology (GO) enrichment network analyses using the website version of DAVID (v6.8) (20) and Cytoscape (v3.9.0) (21). We used the ClueGO plugin v2.5.8 (22) for Cytoscape to generate enriched GO categories using a right-sided hypergeometric test and *p*-values, adjusted using Benjamini-Hochberg for multiple testing correction, reported. We performed network comparisons between the DmSP2 and DmSP3 using ClueGO. Gene lists were uploaded to DAVID

Table 1. History of the *Drosophila melanogaster* sperm proteome (DmSP). DmSP1: (7); DmSP2: (14); DmSP3: this study. The DmSP2 combined the 341 proteins identified in the DmSP1 with the 956 proteins identified in the DmSP2. Likewise, the DmSP3 reported here represents the combined total of all proteins identified in the DmSP2 (n = 1108) with the 2563 proteins identified across all experiments in the current study. Numbers in parentheses denote number of newly identified proteins. *Under = significant gene underrepresentation compared to expected value (see Methods); ns = not significant.

	DmSP1	DmSP2	DmSP3
Methods/Technology	LC-MS^2/Maldi	SDS-PAGE	Cell digest
Machine	Thermo LCQ	LTQ Orbitrap	Orbitrap Fusion Lumos
Proteins identified	341	1108 (+767)	3176 (+2068)
X-linked*	Under	ns	Under
Y-linked	-	4	9 (+5)
Sfps	CG2918	11 (+10)	122 (+111)
Ribosomal proteins	-	9	83 (+74)

(https://david.ncifcrf.gov/tools.jsp) and functional outputs for all three GO categories (BP, CC, MF) and associated statistical values were saved in Excel spreadsheets. Enriched GO categories with FDR values below 1% are reported. Specific parameters details are found in the figure legends.

E. Evolutionary rates. We calculated the rate of non-synonymous (dN) to synonymous (dS) nucleotide substitutions (dN/dS) for *D. melanogaster* genes using an existing pipeline (23). We downloaded amino acid sequences and coding sequences (CDS) for *D. melanogaster* (BDGP6.32), and CDS files for *D. sechellia* (dsec_r1.3), *D. simulans* (ASM75419v3), and *D. yakuba* (dyak_caf1) from Ensembl (24). For each species, we identified the longest isoform of each gene and identified orthologs using reciprocal BLASTn (25), with a minimum 30% identity and 1x10-10 E-value cut-off. We identified reciprocal 1:1 orthologs between all four species by the highest BLAST score and identified open reading frames using BLASTx. We then aligned orthologs using PRANK (26) and masked poorly aligned reads with SWAMP (27) using a minimum sequence length = 150, non-synonymous substitution threshold = 7, and window size = 15. We retained 11715 orthologs for analysis after filtering poorly aligned orthologs and those with sequence length < 30bp. We calculated one-ratio estimates (model 0) with an unrooted phylogeny: ((*D. simulans*, *D. sechellia*), *D. melanogaster*, *D. yakuba*), using the CODEML package in PAML (28), and filtered orthologs with a branch specific dS $\geq$ 2 or where S*dS $\leq$ 1 to avoid mutational saturation. In total we retained dN/dS estimates for 11417 genes after filtering, including 2571 (80.9%) proteins in the DmSP3. We tested for differences in evolutionary rates between independent sets of genes using Mann-Whitney U tests.

F. Experimental design and statistical rationale. We designed experiments to (i) maximize proteome coverage, (ii) measure using label-free quantitation the relative abundance of individual proteins in the proteome and (iii) examine sample purity by measuring the magnitude of adventitious protein binding and contamination in our samples. We performed three independent experiments using three treatments of purified sperm samples as described in Methods. In experiment one, we collected 3 biological replicates of sperm in PBS only. In experiment two sperm were collected in either PBS and Halt protease inhibitor ("Halt" treatment), PBS only ("NoHalt" treatment) or PBS containing 0.1% Triton X100 without protease inhibitor ("PBST" treatment). In experiment three we collected 4 biological replicates of sperm prepared using either PBS ("PBS" treatment) or 2.5M NaCl ("Salt" treatment).

We applied strict thresholds for peptide and protein identification by setting a false-discovery rate (FDR) threshold at 1.0%, calculated using a reverse-concatenated target/decoy strategy in Percolator. To test for differences in abundance between treatments, LFQ ion intensities, calculated using the Minora feature detector in Proteome Discoverer to determine area-under-the-curve (AUC) and summed technical replicates prior to analysis precursor intensities, were fit to protein-wise negative binomial generalized linear models. For experiment two, investigating the effect of detergent treatment, during preliminary analysis we performed pairwise analysis between all three treatments which revealed 16 proteins that showed differential abundance between controls (Halt vs. NoHalt) (Fig. S1). We subsequently performed differential abundance analysis comparing the PBST treatment to the average of both controls (Halt and NoHalt), excluding these 16 proteins (see supplementary analysis). To rank order protein abundances, we calculated a grand mean for each protein, excluding the PBST treatment samples which, as expected, showed substantial differences compared to other samples (see Results).

G. Statistical analysis. We performed all statistical analysis in R v4.03 (29). All code and analyses are available via GitHub.

To test for non-random distribution of sperm proteins across the polytene chromosomes we downloaded the chromosomal location for all genes in the genome from FlyBase.org (30) and calculated the total numbers of genes on each chromosome. We then summed the observed number of genes found in the sperm proteome on each chromosome, and calculated the expected number based on the total number of sperm proteins identified. We calculated χ^2 statistics for each chromosome and the associated *p*-values with one degree of freedom and used the Benjamini-Hochberg procedure to correct for multiple testing. We excluded analysis of the Y chromosome due to the small number of protein coding genes. To test for non-random distribution of sperm genes across ages classes, we downloaded gene age information from http://gentree.ioz.ac.cn/download.php (31) and

grouped as; ancestral (class 0; common to the *Drosophila* genus; n = 12013), subgenus Sophophora (classes 1 + 2; n = 416), melanogaster group (class 3; n = 200), melanogaster subgroup (class 4; n = 334), or recent (classes 5 + 6; n = 120). We tested if sperm genes were randomly distributed across age classes compared to the rest of the genome as above, calculating the observed number of genes in each age class across the genome and among sperm proteins and calculating χ^2 statistics comparing the observed vs. expected number of genes in each age class, using the Benjamini-Hochberg procedure to correct for multiple testing.

To test for differences in abundance between ribosomal proteins compared to the DmSP3 average, independent groups of X-linked- Y-linked- or autosomal- proteins, or between 'high confidence' seminal fluid proteins (Sfps), 'low confidence/transferred' Sfps, or remaining sperm proteins, we calculated the grand mean abundance across all three experiments excluding the PBST treatment. We filtered proteins identified by two or more unique peptides and found in at least 3 biological replicates in at least 1 treatment group (where applicable). We performed Kruskal-Wallace rank-sum tests followed by pairwise Wilcoxon rank-sum tests corrected for multiple testing using the Benjamini-Hochberg procedure. For experiments two and three we performed differential abundance analyses using edgeR (32). For experiment two we filtered proteins with values in 7 out of 9 biological replicates. For experiment three we filtered to include proteins identified in at least 5 replicates (i.e., in at least 3 out of 4 biological replicates of one treatment).

Results

H. Overview of the DmSP3. In the current study we identified 2563 proteins across our three experiments (Fig. S2), of which 1965 (76.7%) proteins were identified by two or more unique peptides in a single experiment (n = 1412) or in two or more replicates across any experiment (n = 1867). Relative protein abundances of 2125 proteins (81.2%) were measured by LFQ. As expected from our previous study (33), α- and β- tubulins, and Sperm-Leucylaminopeptidases (S-Laps) were among the most abundant (Table 2). Also present were proteins of unexpected sperm prevalence including ocnus and janus B, a pair of duplicated gene products encoding a testis-specific phosphohistidine phosphatase (34), numerous seminal fluid proteins (Sfps) and over 80 ribosomal proteins (RPs). Overall, we found highly consistent estimates of protein abundances between experiments. Protein abundances were strongly correlated between experiments (Pearson's correlation = 0.86 – 0.89, all $p < 0.001$; Fig. S3) and median coefficients of variation for each experiment ranged from 0.018-0.054. We performed analyses using the entire DmSP3 (n = 3176; Table S1), combining the 2563 proteins identified in the current study with the 1108 proteins identified in the DmSP2 (7, 14) (Fig. 1a).

I. Gene Ontology and network analyses. The DmSP3 is considerably larger than the DmSP2 (Fig. 1a) and GO analysis identified 24 significantly enriched BP categories (Fig. 2a; Table S2). As expected, major categories included processes involved in energy transduction (e.g., oxidation-reduction, glycolysis, TCA cycle) and reproduction. Other sperm-specific functions included terms related to microtubule and cilium movement. Surprisingly, the GO term "translation" was a prominent member in this analysis containing 78 cytosolic and mitochondrial RPs. To further explore the GO category representation in the DmSP2 and DmSP3, we generated a heat map between the two proteomes in Cytoscape using ClueGO (Fig. 2b). Similar to our previous analysis of the DmSP1 and DmSP2 (14), most of the categories were equal or nearly equal in their shared properties with the one obvious exception being the aforementioned translation BP category as discussed further below.

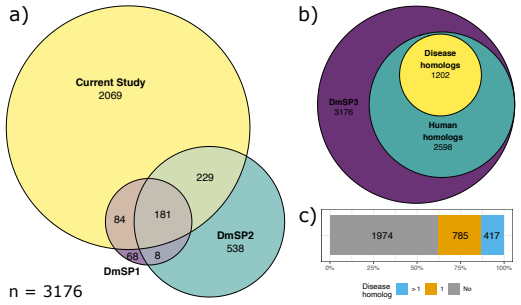

Fig. 1. Proteins identified in the *D. melanogaster* sperm proteome (DmSP). a) Overlap between DmSP1, DmSP2 and the current study, together making up the DmSP3 (n = 3176). b) Number of *D. melanogaster* genes found in the DmSP3 with human homologs and disease associated phenotypes from the Online Mendilian Inheritance in Man database (OMIM.org). c) Number of *D. melanogaster* sperm protein genes with none (grey), one (orange), or more than one (blue) associated disease phenotype.

J. Human disease homologs. Genes in the DmSP3 are highly conserved, with 81.8% (2598/3176) of genes having

Table 2. Most abundant proteins in the DmSP3. Top 20 most abundant proteins in the DmSP3 by LFQ (rank ordered).

FBgn	Name	Chrom.
FBgn0003884	α-Tubulin at 84B	3R
FBgn0003889	β-Tubulin at 85D	3R
FBgn0259795	loopin-1	2R
FBgn0003885	α-Tubulin at 84D	3R
FBgn0033868	Sperm-Leucylaminopeptidase 7	2R
FBgn0035915	Sperm-Leucylaminopeptidase 1	3L
FBgn0052064	Sperm-Leucylaminopeptidase 4	3L
FBgn0045770	Sperm-Leucylaminopeptidase 3	3L
FBgn0039071	big bubble 8	3R
FBgn0034132	Sperm-Leucylaminopeptidase 8	2R
FBgn0041102	ocnus	3R
FBgn0031545	CG3213	2L
FBgn0037862	Mitochondrial aconitase 2	3R
FBgn0038373	CG4546	3R
FBgn0002865	Male-specific RNA 98Ca	3R
FBgn0035240	CG33791	3L
FBgn0025111	Adenine nucleotide translocase 2	X
FBgn0069354	Porin2	2L
FBgn0052351	Sperm-Leucylaminopeptidase 2	3L
FBgn0012036	Aldehyde dehydrogenase	2L

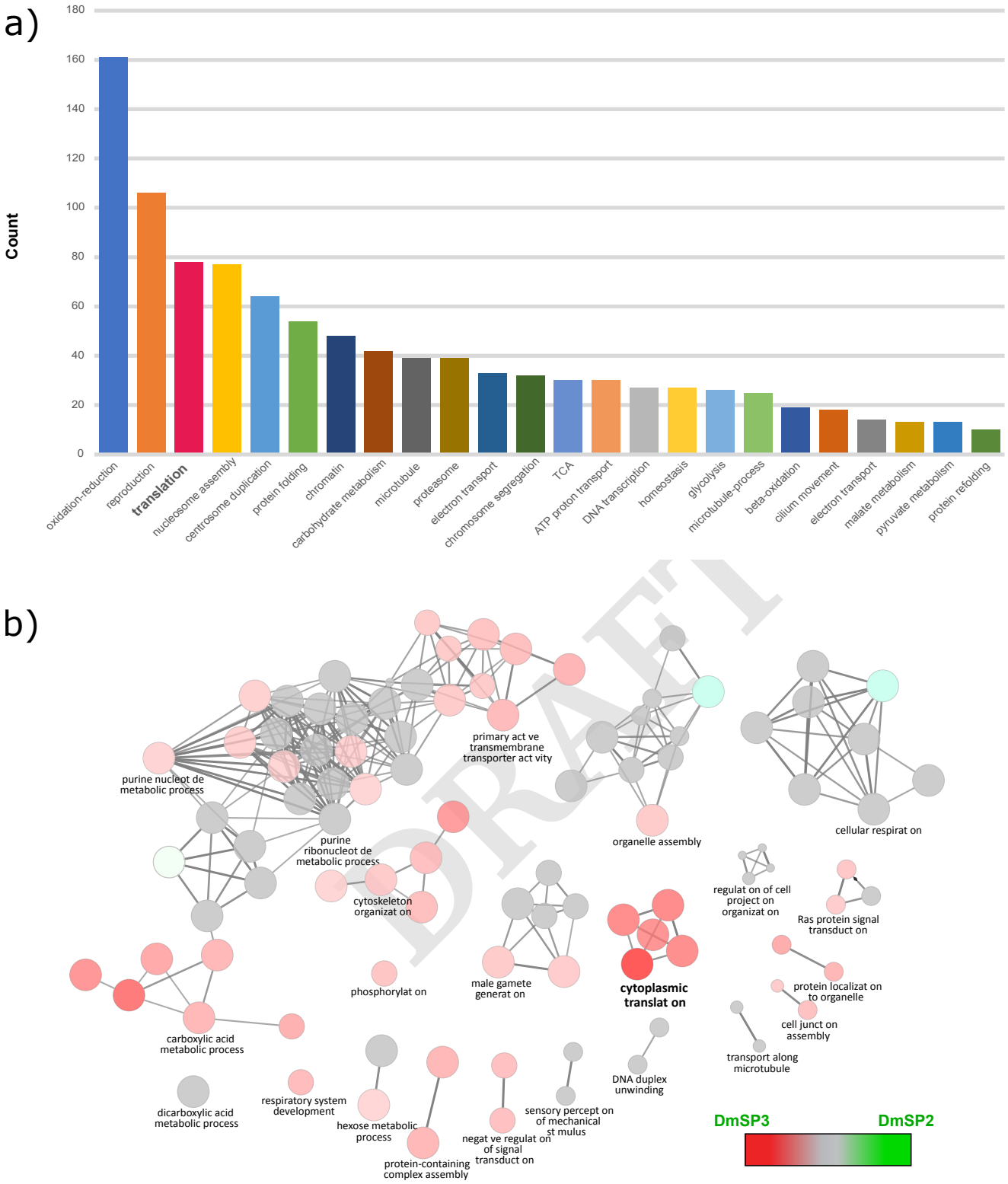

Fig. 2. GO functional network enrichment analysis and comparison of the DmSP2 and DmSP3. (a) Bar graph of the 24 GO Biological Process categories identified in the DmSP3 by DAVID (20). Only functional enrichment groups with Benjamini-Hochberg corrected *p*-values < 0.01 and passing a 1% FDR threshold are shown. Note: some GO terms have been combined for clarity; see Table S2 for complete list of GO terms. (b) GO Biological Process network comparison between the DmSP3 (3176 proteins) and DmSP2 (1108 proteins) using the ClueGO plugin for Cytoscape. Color-coded nodes within the network depict the degree of relative compositional enrichment of each dataset. The network is comprised of 22 groups (each comprised of at least 30 genes associated with a common GO functional term) containing a total of 1431 proteins. Node compositional enrichment for proteins identified in the current study (highlighted in red) when node composition bias exceeds 60% while grey nodes indicate equal representation. **Bold** letters indicate one highly enriched category of proteins involved in cytoplasmic translation.

human homologs, compared to 48% of all *Drosophila* genes (35). Fully 37.8% (1202/3176) of DmSP3 genes have a ho-

molog in humans associated with a known disease or syn-drome in a search against the Online Mendelian Inheritance

in Man database (OMIM.org; Fig. 1b). Over one third (34.7%; 417/1202) of disease associated DmSP3 genes have more than one human disease homolog (Fig. 1c). Among the most prevalent disease phenotypes found were susceptibility to autism, primary ciliary dyskinesia, spermatogenic failure, and myofibrillar- and congenital- myopathy (Table 3).

K. Ribosomal proteins in the DmSP3. Almost one-half of all *D. melanogaster* RPs listed in FlyBase.org (83/169; 49.1%, including paralogs) were identified in the DmSP3 (Table S3). We identified the majority of cytoplasmic RPs (76/93; 81.7%) but only 9.2% (7/76) of mitochondrial RPs. There was no significant difference in RP abundance compared to the DmSP3 average (Kruskal-Wallis rank sum test, $\chi^2 = 0.063$, df = 1, $p = 0.803$; Fig. 3a), suggesting that RPs identified in the DmSP3 are integral to the sperm proteome and not artefactual.

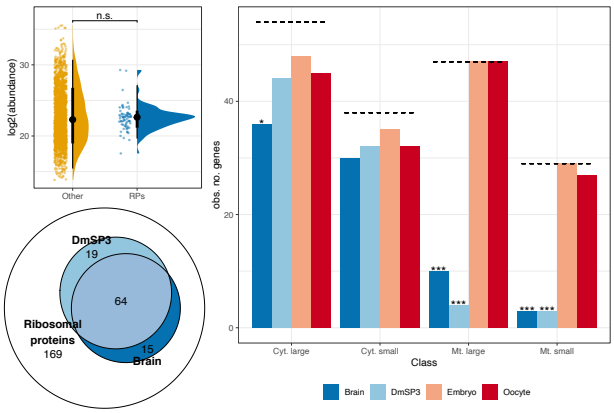

Fig. 3. Ribosomal proteins in the DmSP3. a) Abundance of ribosomal proteins (RPs) identified in the DmSP3 compared to the remaining sperm proteome ('other'). Colored points represent the abundance of individual proteins. Black points show the mean and thick and thin bars represent the 33%, and 66% confidence intervals, respectively. We compared abundances using a Kruskal-Wallace rank sum test. b) Representation of large and small cytoplasmic and mitochondrial RPs in the brain, DmSP3, embryo, or oocyte. The dashed line represents the total number of RPs in each class and asterisks represent results from comparing the observed to expected number of proteins identified using the χ^2 distribution after multiple testing correction. c) Overlap between the total number of RPs identified in the DmSP3 and brain tissue. n.s.; non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The canonical ribosome contains 80 RPs including 13 paralog pairs in *Drosophila* (FlyBase.org). Although the significance of paralog heterogeneity for ribosome function is currently unknown, paralog switching of RPs has been observed in gonads and other tissues (36). We therefore compared RP paralogs in the DmSP3 to those previously described in four tissue types including the testis (36). Significant differences were found between all four tissues as only three RP paralogs were observed in the DmSP3 (RpL22-like, RpS14b and RpS28b) whereas both RPs and paralog RPs were found in all tissues with the exception of two (Rp10Aa and RpS14a; Table S4). Notably, RpL22-like is more abundant in the testis (36), whereas RpL22 was more abundant in the DmSP3. For the remaining paralogs we identified only one member of each pair: the most abundant paralog found in the testis for seven, and the less abundant paralog for two (Table S4). For RpL10Ab and RpS14b only one paralog was identified in both the current study and by Hopes et al. (36), and we did not identify either paralog of RpS10 (RpS10a or RpS10b). Together, these results suggest a complex landscape of paralog switching in the gonad during spermatogenesis and highlight distinct differences between sperm-RP and testis-RP populations.

We next compared the representation of RPs found in the DmSP3 to three other recent proteomic studies in *D. melanogaster* which used Lumos Fusion Orbitrap mass-spectrometers; embryo (37), unfertilised oocyte (38), and brain (39). All four tissue/cell types identified most cytoplasmic RPs, with a slight underrepresentation of large cytoplasmic subunits in the brain (Fig. 3b). The DmSP3 and brain both showed significant underrepresentation of large and small mitochondrial RPs, whereas oocyte and embryo showed almost complete representation of all ribosomal subunits (Fig. 3b). Significantly more RPs identified in the brain or sperm were shared between tissues (64/98; 65.3%) than expected by chance (Fisher's exact test, $p < 0.001$; Fig. 3c).

L. Chromosomal distribution of sperm proteins. Sperm proteins were underrepresented on the X- ($\chi^2 = 12.6$, df = 1, $p = 0.002$) and 3L- ($\chi^2 = 11.8$, df = 1, $p = 0.002$) chromosomes (Fig. 4a); a pattern that was previously reported for X-linked genes in the DmSP1 (7) but not replicated in the DmSP2 (14). Protein abundance of X-linked proteins was significantly lower than those on autosomes (Wilcoxon rank-sum test, $p = 0.041$) or the Y chromosome (Wilcoxon rank-sum test, $p < 0.001$; Fig. 4b). We identified 9 of the 16 known proteins encoded on the Y chromosome (Table 4). The average abundance of Y-linked sperm proteins was higher than autosomal sperm proteins (Wilcoxon rank-sum test, $p < 0.001$); 6 within the top 20% most highly abundant proteins, and all within the top 50% (Fig. 4b; Table 4).

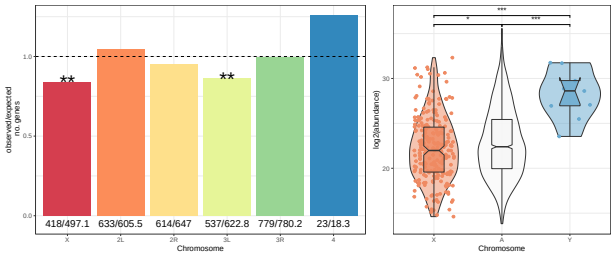

Fig. 4. Chromosomal distribution of DmSP3 proteins. a) Chromosomal distribution of sperm proteins. Numbers below bars are the observed and expected number of genes on each chromosome, respectively, and the dashed line indicates the null expectation. Asterisks represent results from comparing the observed to expected number of genes using the χ^2 distribution after multiple testing correction. b) Abundance of sperm proteins found on autosomes ('A') and sex chromosomes ('X' or 'Y'). Points, representing individual proteins, are omitted from autosomes for clarity. Asterisks represent results from pairwise Wilcoxen rank-sum test corrected for multiple testing using the Benjamini-Hochberg procedure. n.s., non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

M. Seminal fluid proteins identified in the DmSP3. Sfps have been extensively studied in *Drosophila* with over 600 putative Sfps identified to date including 292 that are considered 'high confidence' (42). A surprisingly high number of Sfps were identified in the DmSP3 (122 'high confidence' Sfps; 156 'low confidence/transferred' Sfps; Table S1) (42).

Table 3. Human disease homologs in the DmSP3. Most common human disease phenotypes from the Online Mendelian Inheritance in Man database (OMIM.org) associated with *D. melanogaster* genes found in the DmSP3. N = number of *D. melanogaster* genes associated with each phenotype. Similar disease phenotypes (marked with an asterisk) have been grouped. Complete list of disease associations can be found in Table S15.

OMIM phenotype	N
Autism, Susceptibility To; AUTS20, AUTSX1, AUTSX2	27*
Ciliary Dyskinesia, Primary; CILD40, CILD3, CILD7	25*
Spermatogenic Failure; SPGF39, SPGF45, SPGF46	24*
Myopathy; CFTD, MFM2, Fatal Infantile Hypertonic, Alpha-B Crystallin-Related	24*
Hypertension, Essential	23
Type 2 Diabetes Mellitus; T2D	21
Asperger Syndrome, X-Linked, Susceptibility To; ASPGX1, ASPGX2	18*
Cataract, Multiple Types; CTRCT16, CTRCT9	16*
Ichthyosis, Congenital, Autosomal Recessive; ARCI4A, ARCI4B	16*
46, XY Sex Reversal 8; SRXY8	12
Colorectal Cancer; CRC	11
Encephalopathy, Familial, With Neuroserpin Inclusion Bodies; FENIB	11
Ghosal Hematodiaphyseal Dysplasia; GHDD	10
Plasminogen Activator Inhibitor-1 Deficiency	10
Vitamin D-Dependent Rickets, Type 3; VDDR3	10
Deafness, Autosomal Recessive 91; DFNB91	9
Leukemia, Acute Myeloid; AML	9
Maturity-Onset Diabetes of The Young, Type 8, With Exocrine Dysfunction; MODY8	9
Pseudoxanthoma Elasticum; PXE	9
Cardiomyopathy, Dilated, 1II; CMD1II	8
Charcot-Marie-Tooth Disease, Axonal, Type 2F; CMT2F	8
Neuronopathy, Distal Hereditary Motor, Type IIB; HMN2B	8
Surfactant Metabolism Dysfunction, Pulmonary, 3; SMDP3	8

Table 4. Y-linked sperm proteins in the DmSP3. Genes are rank ordered by mean abundance and link to male fertility from gene knockout/knockdown experiments (40, 41) are shown.

FBgn	Name	Ranked abundance (%)	Sterile
FBgn0267433	male fertility factor kl5	98.8	Yes
FBgn0267432	male fertility factor kl3	98.8	Yes
FBgn0058064	Aldehyde reductase Y	95.5	No
FBgn0001313	male fertility factor kl2	93.6	Yes
FBgn0046323	Occludin-Related Y	92.9	No
FBgn0267449	WD40 Y	86.4	Yes
FBgn0267592	Coiled-Coils Y	86	Not studied
FBgn0046697	Ppr-Y	78.3	No
FBgn0046698	Protein phosphatase 1, Y-linked 2	65.2	No

We found no significant difference in abundance between Sfps and the remaining DmSP3 (Kruskal-Wallace rank-sum test, $\chi^2 = 4.28$, df = 1, $p = 0.118$; Fig. 5a) and 44 'high confidence' Sfps were at, or above, the median abundance of the DmSP3 (Table S5).

We therefore examined the binding characteristics of the Sfps by washing purified sperm with a strong anionic detergent (Triton X-100) known to disrupt plasma membranes. Following detergent treatment 1600 proteins were identified, the majority (1063/1600; 66%) identified by 2 or more unique peptides. We identified 198 proteins that were lower abundance in PBST compared to controls (Table S6) and three proteins more abundant in PBST samples (Fig. 5b).

Of the 60 'high confidence' Sfps identified by two or more unique peptides in experiment two, 17 (28.4%) were filtered out prior to analysis (including 14 which were not detected in PBST samples in any replicate), and 29 (48.3%) were found at significantly lower abundance in PBST samples, together suggesting these proteins are weakly bound or found on the sperm plasma membrane. The remaining 14 (23.3%) Sfps showed no significant difference in abundance, suggesting tight association with sperm (Table 5). Additionally, 13 out of 53 (24.5%) RPs detected in experiment two were significantly lower in abundance after PBST treatment. Proteins lower in abundance after PBST treatment showed GO enrichment of multicellular organism reproduction, mitochondrial transport, transmembrane transport, cytoplasmic translation, and sarcomere organisation (BP). Thus, as expected, PBST treatment stripped lipids and membrane- and membrane- bound proteins (including Sfps) from sperm (Table S7).

In experiment three, we washed sperm samples with high molar salt expected to weaken ionic bonds and eliminate non-specific protein binding to sperm (including Sfps). We

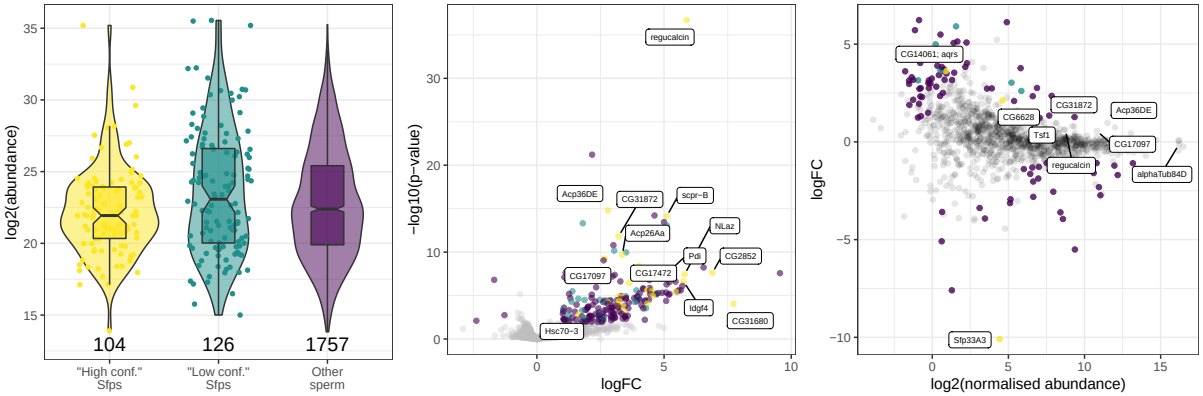

Fig. 5. Seminal fluid proteins in the DmSP3. a) log2 abundance of proteins found in the DmSP3 classified as 'high confidence' Sfps, 'low confidence or transferred' Sfps by Wigby et al. (42), or remaining sperm proteins. Points, representing individual proteins, are omitted from 'other sperm' for clarity. b) Volcano plot for difference between PBST treatment vs. the average of both controls (Halt and NoHalt) in experiment two. Positive values indicate higher abundance in controls. c) MA plot for difference between NaCl treatment vs. PBS control. Positive values indicate higher abundance in NaCl treatment. For b) and c) points are coloured as in a) denoting 'high confidence' (yellow) and 'low confidence/transferred' (turquoise) Sfps or remaining sperm proteins (purple) that showed significant differences in abundance based on a |logFC| > 1 and false discovery rate corrected *p*-value < 0.05. Several Sfps are labelled in b) that showed differential abundance between treatments. Sfps are labelled in c) that were among the top 10% most abundant proteins, and the three Sfps that showed significant differences in abundance between treatments (Sfp33A3, aquarius [CG14061], and CG6628).

identified 1890 proteins, of which 1273 (65%) were identified by two or more unique peptides. After filtering (see Methods) we performed differential abundance analysis for 1202 proteins and identified 92 differentially abundant proteins, including 3 Sfps (Sfp33A3, aquarius [CG14061], and CG6628) (Fig. 5c). The remaining 48 'high confidence' Sfps we identified in this experiment did not show significant differential abundance between treatments, with 6 Sfps in the top 20% most abundance proteins (regucalcin, Acp36DE, CG31872, Transferrin 1, CG17097, and α-Tubulin at 84D).

N. Gene – protein abundance concordance. To explore the relationship between protein abundance and gene expression for the 68 'high confidence' Sfps tightly binding to sperm following detergent or salt treatment ('sperm associated Sfps'; Table S8), we compared gene expression (FPKM; fragments per kilobase of transcript per million mapped reads) for all proteins identified in the DmSP3 between the accessory glands, carcass, ovary, and testis using data retrieved from FlyAtlas2 (43). The average expression of both sperm associated Sfps and the remaining Sfps identified in the DmSP3 was highest in the accessory glands, while the remaining DmSP3 proteins were most highly expressed in testis (Fig. S4a). However, 7 sperm associated Sfps showed higher expression in the testis than accessory glands (Table S9).

The abundance of proteins in the DmSP3 had the strongest correlation (β) and best fit (R^2) in the testis (β = 0.460, R^2 = 0.133, $p < 0.001$, n = 1498) (Fig. S4b). Protein abundance of sperm associated Sfps was positively correlated with gene expression in the testis (β = 0.399, R^2 = 0.152, p = 0.006, n = 49), but not the accessory glands (p = 0.246), carcass (p = 0.052), or ovary (p = 0.271). The abundance of remaining Sfps identified in the DmSP3 was positively correlated with gene expression in the accessory glands (β = 0.274, R^2 = 0.197, p = 0.004, n = 41) and testis (β = 0.281, R^2 = 0.147, p = 0.040, n = 29), but not the carcass (p = 0.109) or ovary (p = 0.677) (Fig. S4c). Therefore, our results suggest sperm associated Sfps show tighter regulation with gene expression in the testis than accessory glands.

Table 5. Seminal fluid proteins remaining in the sperm proteome after PBST treatment.

FBgn	Name	Chrom.
FBgn0011694	Ejaculatory bulb protein II	2R
FBgn0261055	Seminal fluid protein 26Ad	2L
FBgn0004181	Ejaculatory bulb protein	2R
FBgn0003885	α-Tubulin at 84D	3R
FBgn0260745	midline fasciclin	3R
FBgn0036970	Serpin 77Bc	3L
FBgn0036969	Serpin 77Bb	3L
FBgn0259975	Seminal fluid protein 87B	3R
FBgn0034709	Secreted Wg-interacting molecule	2R
FBgn0264815	Phosphodiesterase 1c	2L
FBgn0020414	Imaginal disc growth factor 3	2L
FBgn0050104	Ecto-5'-nucleotidase 2	2R
FBgn0052203	Serpin 75F	3L
FBgn0003748	Trehalase	2R

O. Gene age. A variety of mechanisms drive genomic and protein diversity including gene duplication and retroposition (6, 31) resulting in unique, lineage-specific patterns of gene age (44–46). Newly evolved genes frequently acquire testis-biased gene expression (47) and it was therefore of interest to query the gene age landscape of the DmSP3. There were fewer 'recent' (χ^2 = 6.58, df = 1, p = 0.026), 'melanogaster subgroup' (χ^2 = 9.69, df = 1, p = 0.009), and 'Sophophora-group' (χ^2 = 5.51, df = 1, p = 0.032) age genes than expected by chance, indicating genes encoding sperm proteins are underrepresented in more recent evolutionary time (Fig. 6a). We identified 13 genes of recent origin, of which five were located on the X chromosome (Table S10).

P. Sperm evolutionary rates. Genes in the DmSP3 evolve more slowly than the genome average (Mann-Whitney U test, $p < 0.001$). This pattern remains when considering X-linked

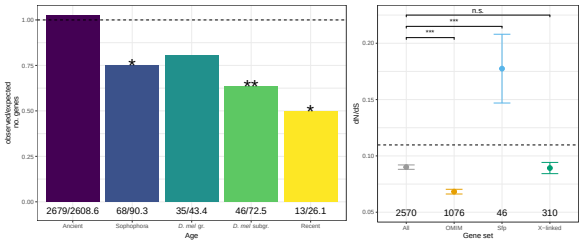

Fig. 6. Sperm evolutionary dynamics. a) Gene age distribution of sperm proteins. Numbers below bars are the observed and expected number of genes in each age class, respectively, and the dashed line at 1 indicates the null expectation. Asterisks represent results from comparing the observed to expected number of genes using the χ^2 distribution after multiple testing correction. b) Mean (± standard error) nonsynonymous (dN) to synonymous (dS) nucleotide substitution rate (dN/dS) estimates for sperm proteins. Asterisks represent results from Mann-Whitney U tests comparing each gene set (OMIM, Sfp, X-linked) to the genome average ("All"), excluding proteins in that set. Dashed line represents the genome average (mean dN/dS = 0.110, standard error = 0.001, n = 11417). Numbers below points indicate numbers of genes in each category. Note: groups are not necessarily mutually exclusive, i.e., 'OMIM' proteins may also be 'X-linked', etc. n.s., non-significant; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

sperm proteins compared to the genome average ($p < 0.001$), which evolve at a similar rate to the DmSP3 average ($p = 0.958$; Fig. 6b). Sfps in the DmSP3 evolve faster than the DmSP3 average ($p < 0.001$), at a similar rate to other Sfps ($p = 0.232$; Fig. S5), whereas genes with a human disease homolog (OMIM.org) evolve more slowly than the DmSP3 average ($p < 0.001$; Fig. 6b).

The top 10%, fastest evolving genes in the DmSP3 (dN/dS [mean ± s.e.] = 0.313 ± 0.009, n = 257, Table S11) showed GO enrichment for multicellular organism reproduction (BP) and extracellular space (CC) (Table S12). The bottom 10%, slowest evolving genes in the DmSP3 (dN/dS = 0.004 ± 0.0002, n = 258, Table S13) showed GO BP enrichment for cytoplasmic translation, centrosome duplication, regulation of cell shape, ribosomal large subunit assembly, tricarboxylic acid cycle, ATP hydrolysis coupled proton transport, cell adhesion, oocyte microtubule cytoskeleton polarisation, and endocytosis (Table S14).

Discussion

In summary, our reanalysis of the *D. melanogaster* sperm proteome (DmSP3) more than doubled the number of identified proteins, dramatically increased representation of RPs, and highlighted several human neurological disease homologs. LFQ identified highly abundant tubulins, Sperm-Leucylaminopeptidases (S-Laps), Y-linked sperm proteins and ocnus, a testis-specific protein. LFQ also provided direct evidence for lowered abundances of X-linked sperm proteins. Sperm genes evolve relatively slowly and are underrepresented in recent age classes, consistent with evolutionary constraint acting on the sperm proteome. Finally, we identified a number of Sfps in the DmSP3 which were resistant to detergent or high molar salt treatment, suggesting some are integral to the sperm proteome.

The increased (> 2-fold) depth of proteome coverage is likely due to improved protein extraction, efficiency of trypsination/peptide recovery and direct injection methods employed in this study. Traditionally SDS-PAGE off-line pre-fractionation has been the method of choice for the analysis of complex proteomes. However, these off-line methods come at a cost: sample loss due to the extra steps involved and the well-known issues of peptide recovery from polyacrylamide gels (48, 49). Although work to alleviate this limitation continues to improve this approach, our results suggest that a combination of high SDS concentrations in the initial solubilization and use of immobilized enzymatic digestion using S-Trap technologies greatly enhanced the yield of usable peptides for bottom-up proteomics. The DmSP3 also contained Yolk protein 2 (Yp2), a protein previously found in sperm (50) but undetected in the DmSP1 or DmSP2. As noted by the authors of this study, detection of Yp2 in sperm required large amounts of input protein for detection on immunoblots, suggesting Yp2 was present at very low levels in the testis and sperm (50). Therefore, detection of Yp2 in our study provides additional confidence in the efficacy or our approach.

The sperm proteome is expected to exhibit dynamic gene movement and expression evolution due to its sex-specific expression and essential role for male fertility (6). We found X-linked genes are underrepresented in the sperm proteome, as reported in the DmSP1 (7). Additionally, we show that X-linked sperm proteins were found in significantly lower abundances, consistent with the downstream effects of meiotic sex chromosome inactivation (16–18), and/or resolution of intralocus sexual conflict (51–53). In contrast, more than half of Y-linked proteins (9/16) including known fertility factors were present in the DmSP3 (40, 41). Our LFQ analysis revealed all 9 Y-linked protein abundances were above the DmSP average, with 7/9 in the top 10%. This is the first quantitative assessment of this important class of sperm proteins in sperm and adds direct empirical evidence in support of the long-standing hypothesized structural role in the assembly of the sperm axoneme (54).

We found sperm proteins evolve more slowly than the genome average. Slow rates of adaptive evolution could be due to purifying selection or weak selection acting on sperm genes as they are shielded from selection in females (7, 55, 56). Sperm proteins were also underrepresented in recent evolutionary age classes and over 80% had human homologs, supporting the idea that sperm genes are under evolutionary constraint. A recent study found Sfps are overrepresented in recent age classes (57), indicating different evolutionary forces acting on sperm vs. non-sperm components of the ejaculate. Sfps in the DmSP3 evolve at a similar rate to Sfps found elsewhere in the genome, and more quickly than the DmSP3 average, suggesting similar evolutionary pressures affecting rates of Sfp evolution across tissue types.

The abundance of RPs in the DmSP3 was unexpected given that sperm are stripped of most cellular machinery prior to maturation. However, sperm may undergo post ejaculatory modifications, perform secondary sexual functions, or provision the developing zygote after fertilization, requiring protein synthesis (58–61). Sperm function beyond delivering a haploid compliment of nuclear material for fertilization still remains relatively underexplored (59, 62, 63). The presence

of a large repertoire of core RPs delivered to the egg during fertilization raises the intriguing possibility that paternally-derived ribosomes are active during zygote formation and perhaps beyond.

Another intriguing finding that sperm had higher abundance of RpL22 versus the paralog RpL22-like, opposite from levels found in the testis (36) suggests a complex pattern of paralog switching and selectivity during spermatogenesis. While the functional significance of this selectivity is unknown, they are interesting to consider in the context of the known mRNA repertoire in *Drosophila* sperm delivered to the egg at fertilization (61). Fully 33% of the total sperm mRNA repertoire encoded ribosomal proteins (47/142; ref. (61)), a striking coincidence that warrants further study. We also found similarity in the underrepresentation of mitochondrial RPs in both the DmSP3 and brain, providing yet another example of the molecular similarities between these two tissue types (64). Finally, we note that the DmSP3 contains as many as 300 entries with GO annotation terms related to neuronal structure and function, lending additional support to the similarities drawn between the brain and testis.

Q. Possible testis origin of seminal fluid proteins. Although some Sfps were previously identified, but not quantified, in the DmSP2 (14), the unexpectedly high numbers (and in some cases, relative abundances) of Sfps in the DmSP3 adds to an expanding landscape of seminal fluid protein biology. As Sfps are thought to be primarily secreted from the paired accessory glands and the ejaculatory bulb in *Drosophila* (65), our results raise the possibility that some Sfps are integral to the sperm proteome, are secreted from the testes or seminal vesicles, or bind to sperm prior to mixing in the ejaculatory duct. We identified 122 'high confidence' Sfps (42) in the DmSP3 which is unlikely artefactual given that many Sfps were found in multiple biological replicates and in independent experiments. Denaturing the sperm plasma membrane using detergent stripped most (75%) Sfps from the sperm proteome, suggesting these Sfps are integral to the sperm plasma membrane or bound to sperm advantageously in the seminal vesicles prior to mixing in the ejaculatory duct. High molar salt had little effect on the composition of the sperm proteome, indicating some Sfps are bound strongly to sperm.

We identified 68 'sperm associated Sfps' that were not depleted by detergent or salt treatment. We suggest several of the 'high-confidence' Sfps identified in the DmSP3 that are highly expressed in the testes (Table S9) should be classified as sperm proteins. In addition, α-Tubulin at 84D (FBgn0003885) is a major constituent of microtubules and involved in sperm axoneme assembly, and therefore likely a sperm protein. Notably, Acp36DE was consistently among the most abundant proteins in our experiments. Acp36DE tightly binds sperm and is essential for efficient sperm storage in the female sperm storage organs (66, 67). The possibility that Sfps bind to sperm in the seminal vesicles prior to mixing in the ejaculatory duct should be investigated further. Moreover, the potential for the testes, seminal vesicles, or perhaps even sperm cells, to secrete proteins, including Sfps, requires further investigation.

Finally, the DmSP3 contains over 1200 human disease homologs. The prominence of several neurological diseases (e.g., Primary Ciliary Dyskinesia, susceptibility to autism, encephalopathy, and neuropathy) may be related to the shared functional designs of sperm and neurons, cells of extraordinary axial ratios transmitting biological information over large distances. It will be of great interest to tease out the significance of this subset of neural-related DmSP3 proteins in the context of sperm function and its related reproductive activities and possible relevance for study of human diseases.

R. Conclusion. Our reanalysis of the *D. melanogaster* sperm proteome using improved separation and detection methods and an updated genome annotation highlights several key features of sperm function and evolution, including the prominence of proteins integral to sperm development (tubulins and S-Laps), the dynamic nature of sex-linked sperm genes, and constraints on sperm proteome evolution. We also show the prevalence of many RPs, despite the expectation that sperm are transcriptionally silent. The parallels in ribosomal protein composition and occurrence of several human neurological disease homologs also lends further support to the functional similarities between sperm and neurons. Finally, we demonstrate that a significant number of seminal fluid proteins are found in the sperm proteome raising the possibility that Sfps mix with sperm in the seminal vesicles, or Sfps may be secreted from the testes, seminal vesicles, or even sperm cells.

S. Data availability. Proteomic data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (68) with the identified PDXXXXXXXX. All code and analyses are available on GitHub.

T. Author contributions. TLK and MDG conceived the study. TLK and JS performed dissections and LC-MS experiments. MDG and TLK performed analyses and wrote the manuscript. All authors agreed on the final version of the manuscript.

ACKNOWLEDGEMENTS

We would like to thank Caitlin McDonough-Goldstein and Maria Vibranovski for helpful discussion, Alison Wright, Daniela Palmer, and Leeban Yusef for advice analysing evolutionary rates, and Eric Sedore and Larne Pekowsky from the Syracuse University HTC Campus Grid and NSF award ACI-1341006 for providing computing services. We are also grateful to the authors whose data was used in this study for making data publicly available and the curators of FlyBase.org for continued maintenance of this essential resource. This work was funded in part by the Biodesign Institute and ASU Knowledge Enterprise Core Research Facilities.

References

1. Steve Dorus and Timothy L. Karr. Sperm proteomics and genomics. In *Sperm Biology: An Evolutionary Perspective*, pages 435–469. Academic press, Burlington, MA, first edition, 2009.
2. Helen L. Bayram, Amy J. Claydon, Philip J. Brownridge, Jane L. Hurst, Alan Mileham, Paula Stockley, Robert J. Beynon, and Dean E. Hammond. Cross-species proteomics in analysis of mammalian sperm proteins. *Journal of Proteomics*, 135:38–50, March 2016. ISSN 1874-3919. doi: 10.1016/j.jprot.2015.12.027.
3. Melissah Rowe, Sheri Skerget, Matthew A. Rosenow, and Timothy L. Karr. Identification and characterisation of the zebra finch (Taeniopygia guttata) sperm proteome. *Journal of Proteomics*, October 2018. ISSN 1874-3919. doi: 10.1016/j.jprot.2018.10.009.
4. Scott Pitnick, David J. Hosken, and Timothy R. Birkhead. Sperm morphological diversity. In *Sperm Biology: An Evolutionary Perspective*, pages 69–149. Academic press, Burlington, MA, first edition, 2009.

5. Ariel F. Kahrl, Rhonda R. Snook, and John L. Fitzpatrick. Fertilization mode drives sperm length evolution across the animal tree of life. *Nat Ecol Evol*, pages 1–12, June 2021. ISSN 2397-334X. doi: 10.1038/s41559-021-01488-y.
6. Elaine C. Rettie and Steve Dorus. Drosophila sperm proteome evolution. *Spermatogenesis*, 2(3):213–223, July 2012. ISSN 2156-5554. doi: 10.4161/spmg.21748.
7. Steve Dorus, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L. Karr. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nature Genetics*, 38(12):1440–1445, December 2006. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng1915.
8. Timothy Karr, L. Fruit flies and the sperm proteome. *Human Molecular Genetics*, 16(R2): R124–R133, July 2007. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddm252.
9. Jürgen Cox and Matthias Mann. Quantitative, High-Resolution Proteomics for Data-Driven Systems Biology. *Annual Review of Biochemistry*, 80(1):273–299, 2011. doi: 10.1146/annurev-biochem-061308-093216.
10. Ruedi Aebersold and Matthias Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620):347–355, September 2016. ISSN 1476-4687. doi: 10.1038/nature19949.
11. Lei Zhao, Xiaoji Cong, Linhui Zhai, Hao Hu, Jun-Yu Xu, Wensi Zhao, Mengdi Zhu, Minjia Tan, and Bang-Ce Ye. Comparative evaluation of label-free quantification strategies. *Journal of Proteomics*, 215:103669, March 2020. ISSN 1874-3919. doi: 10.1016/j.jprot.2020.103669.
12. Fengchao Yu, Sarah E. Haynes, and Alexey I. Nesvizhskii. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics*, 20, January 2021. ISSN 1535-9476, 1535-9484. doi: 10.1016/j.mcpro.2021.100077.
13. Timothy L. Karr. Reproductive proteomics comes of age. *Molecular & Cellular Proteomics*, 18(Supplement 1):S1–S5, March 2019. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.E119.001418.
14. Elizabeth R. Wasbrough, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L. Karr. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *Journal of Proteomics*, 73(11):2171–2185, October 2010. ISSN 1874-3919. doi: 10.1016/j.jprot.2010.09.002.
15. Carine Barreau, Elizabeth Benson, Elin Gudmannsdottir, Fay Newton, and Helen White-Cooper. Post-meiotic transcription in Drosophila testes. *Development*, 135(11):1897–1902, June 2008. ISSN 0950-1991. doi: 10.1242/dev.021949.
16. Maria D. Vibranovski, Hedibert F. Lopes, Timothy L. Karr, and Manyuan Long. Stage-Specific Expression Profiling of Drosophila Spermatogenesis Suggests that Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-Expressed Genes. *PLOS Genetics*, 5(11):e1000731, November 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000731.
17. Maria D. Vibranovski, Domitille S. Chalopin, Hedibert F. Lopes, Manyuan Long, and Timothy L. Karr. Direct Evidence for Postmeiotic Transcription During Drosophila melanogaster Spermatogenesis. *Genetics*, 186(1):431–433, September 2010. doi: 10.1534/genetics.110.118919.
18. Sharvani Mahadevaraju, Justin M. Fear, Miriam Akeju, Brian J. Galletta, Mara M. L. S. Pinheiro, Camila C. Avelino, Diogo C. Cabral-de-Mello, Katie Conlon, Stafania Dell'Orso, Zelalem Demere, Kush Mansuria, Carolina A. Mendonça, Octavio M. Palacios-Gimenez, Eli Ross, Max Savery, Kevin Yu, Harold E. Smith, Vittorio Sartorelli, Haiwang Yang, Nasser M. Rusan, Maria D. Vibranovski, Erika Matunis, and Brian Oliver. Dynamic sex chromosome expression in *Drosophila* male germ cells. *Nature Communications*, 12(1):892, February 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-20897-y.
19. Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*, 4(11):923–925, November 2007. ISSN 1548-7105. doi: 10.1038/nmeth1113.
20. Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, January 2009. ISSN 1750-2799. doi: 10.1038/nprot.2008.211.
21. Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, 13(11):2498–2504, January 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1239303.
22. Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, April 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp101.
23. Alison E. Wright, Peter W. Harrison, Fabian Zimmer, Stephen H. Montgomery, Marie A. Pointer, and Judith E. Mank. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecology*, 24(6):1218–1235, 2015. ISSN 1365-294X. doi: 10.1111/mec.13113.
24. Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, Andrea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E Loveland, Fergal J Martin, Joannella Morales, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J Trevanion, Fiona Cunningham, Kevin L Howe, Daniel R Zerbino, and Paul Flicek. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz966.
25. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.
26. Ari Löytynoja and Nick Goldman. webPRANK: A phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, 11(1):579, November 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-579.
27. Peter W Harrison, Gregory E Jordan, and Stephen H Montgomery. SWAMP: Sliding Window Alignment Masker for PAML. *Evol Bioinform Online*, 10:197–204, December 2014. ISSN 1176-9343. doi: 10.4137/EBO.S18193.
28. Ziheng Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol*, 24(8):1586–1591, August 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm088.
29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2020.
30. Aoife Larkin, Steven J Marygold, Giulia Antonazzo, Helen Attrill, Gilberto dos Santos, Phani V Garapati, Joshua L Goodman, L Sian Gramates, Gillian Millburn, Victor B Strelets, Christopher J Tabone, Jim Thurmond, and FlyBase Consortium. FlyBase: Updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research*, 49(D1):D899–D907, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1026.
31. Yong E. Zhang, Maria D. Vibranovski, Benjamin H. Krinsky, and Manyuan Long. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Res*, 20(11):1526–1533, November 2010. ISSN 1549-5469. doi: 10.1101/gr.107334.110.
32. Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp616.
33. Steve Dorus, Elaine C Wilkin, and Timothy L. Karr. Expansion and functional diversification of a leucyl aminopeptidase family that encodes the major protein constituents of Drosophila sperm. *BMC Genomics*, 12:177, April 2011. ISSN 1471-2164. doi: 10.1186/1471-2164-12-177.
34. John Parsch, Colin D. Meiklejohn, Elisabeth Hauschteck-Jungen, Peter Hunziker, and Daniel L. Hartl. Molecular Evolution of the ocnus and janus Genes in the Drosophila melanogaster Species Subgroup. *Molecular Biology and Evolution*, 18(5):801–811, May 2001. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003862.
35. Shinya Yamamoto, Manish Jaiswal, Wu-Lin Charng, Tomasz Gambin, Ender Karaca, Ghayda Mirzaa, Wojciech Wiszniewski, Hector Sandoval, Nele A. Haelterman, Bo Xiong, Ke Zhang, Vafa Bayat, Gabriela David, Tongchao Li, Kuchuan Chen, Upasana Gala, Tamar Harel, Davut Pehlivan, Samantha Penney, Lisenka E.L.M. Vissers, Joep de Ligt, Shalini N. Jhangiani, Yajing Xie, Stephen H. Tsang, Yesim Parman, Merve Sivaci, Esra Battaloglu, Donna Muzny, Ying-Wooi Wan, Zhandong Liu, Alexander T. Lin-Moore, Robin D. Clark, Cynthia J. Curry, Nichole Link, Karen L. Schulze, Eric Boerwinkle, William B. Dobyns, Rando Allikmets, Richard A. Gibbs, Rui Chen, James R. Lupski, Michael F. Wangler, and Hugo J. Bellen. A Drosophila genetic resource of mutants to study mechanisms underlying human genetic diseases. *Cell*, 159(1):200–214, September 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.09.002.
36. Tayah Hopes, Karl Norris, Michaela Agapiou, Charley G P McCarthy, Philip A Lewis, Mary J O'Connell, Juan Fontana, and Julie L Aspden. Ribosome heterogeneity in Drosophila melanogaster gonads through paralog-switching. *Nucleic Acids Research*, (gkab606), July 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab606.
37. Wen Xi Cao, Sarah Kabelitz, Meera Gupta, Eyan Yeung, Sichun Lin, Christiane Rammelt, Christian Ihling, Filip Pekovic, Timothy C. H. Low, Najeeb U. Siddiqui, Matthew H. K. Cheng, Stephane Angers, Craig A. Smibert, Martin Wühr, Elmar Wahle, and Howard D. Lipshitz. Precise Temporal Regulation of Post-transcriptional Repressors Is Required for an Orderly Drosophila Maternal-to-Zygotic Transition. *Cell Reports*, 31(12), June 2020. ISSN 2211-1247. doi: 10.1016/j.celrep.2020.107783.
38. Caitlin E. McDonough-Goldstein, Scott Pitnick, and Steve Dorus. *Drosophila* oocyte proteome composition covaries with female mating status. *Scientific Reports*, 11(1):3142, February 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82801-4.
39. Jiefu Li, Shuo Han, Hongjie Li, Namrata D. Udeshi, Tanya Svinkina, D. R. Mani, Chuanyun Xu, Ricardo Guajardo, Qijing Xie, Tongchao Li, David J. Luginbuhl, Bing Wu, Colleen N. McLaughlin, Anthony Xie, Pornchai Kaewsapsak, Stephen R. Quake, Steven A. Carr, Alice Y. Ting, and Liqun Luo. Cell-Surface Proteomic Profiling in the Fly Brain Uncovers Wiring Regulators. *Cell*, 180(2):373–386.e15, January 2020. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2019.12.029.
40. Yassi Hafezi, Samantha R. Sruba, Steven R. Tarrash, Mariana F. Wolfner, and Andrew G. Clark. Dissecting Fertility Functions of Drosophila Y Chromosome Genes with CRISPR. *Genetics*, 214(4):977–990, April 2020. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.120.302672.
41. Jiaying Zhang, Junjie Luo, Jieyan Chen, Junbiao Dai, and Craig Montell. The role of Y chromosome genes in male fertility in *Drosophila melanogaster*. *Genetics*, 215(3):623–633, July 2020. ISSN 1943-2631. doi: 10.1534/genetics.120.303324.
42. Stuart Wigby, Nora C. Brown, Sarah E. Allen, Snigdha Misra, Jessica L. Sitnik, Irem Sepil, Andrew G. Clark, and Mariana F. Wolfner. The *Drosophila* seminal proteome and its role in postcopulatory sexual selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1813):20200072, December 2020. doi: 10.1098/rstb.2020.0072.
43. David P Leader, Sue A Krause, Aniruddha Pandit, Shireen A Davies, and Julian A T Dow. FlyAtlas 2: A new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Research*, 46(D1):D809–D815, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx976.
44. Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat Rev Genet*, 10(1):19–31, January 2009. ISSN 1471-0064. doi: 10.1038/nrg2487.
45. Qianwei Su, Huangyi He, and Qi Zhou. On the Origin and Evolution of Drosophila New Genes during Spermatogenesis. *Genes*, 12(11):1796, November 2021. ISSN 2073-4425. doi: 10.3390/genes12111796.
46. Manyuan Long, Nicholas W. VanKuren, Sidi Chen, and Maria D. Vibranovski. New Gene Evolution: Little Did We Know. *Annu. Rev. Genet.*, 47(1):307–333, November 2013. ISSN

0066-4197, 1545-2948. doi: 10.1146/annurev-genet-111212-133301.

47. Esther Betrán, Kevin Thornton, and Manyuan Long. Retroposed New Genes Out of the X in Drosophila. *Genome Res.*, 12(12):1854–1859, January 2002. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.604902.
48. Ayako Takemori, David S. Butcher, Victoria M. Harman, Philip Brownridge, Keisuke Shima, Daisuke Higo, Jun Ishizaki, Hitoshi Hasegawa, Junpei Suzuki, Masakatsu Yamashita, Joseph A. Loo, Rachel R. Ogorzalek Loo, Robert J. Beynon, Lissa C. Anderson, and Nobuaki Takemori. PEPPI-MS: Polyacrylamide-Gel-Based Prefractionation for Analysis of Intact Proteoforms and Protein Complexes by Mass Spectrometry. *J. Proteome Res.*, 19(9): 3779–3791, September 2020. ISSN 1535-3893. doi: 10.1021/acs.jproteome.0c00303.
49. Nobuaki Takemori, Ayako Takemori, Piriya Wongkongkathep, Michael Nshanian, Rachel R. Ogorzalek Loo, Frederik Lermyte, and Joseph A. Loo. Top-down/Bottom-up Mass Spectrometry Workflow Using Dissolvable Polyacrylamide Gels. *Anal. Chem.*, 89(16):8244–8250, August 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.7b00357.
50. Magdalena M. Majewska, Agnieszka Suszczynska, Joanna Kotwica-Rolinska, Tomasz Czerwik, Bohdan Paterczyk, Marta A. Polanska, Piotr Bernatowicz, and Piotr Bebas. Yolk proteins in the male reproductive system of the fruit fly Drosophila melanogaster: Spatial and temporal patterns of expression. *Insect Biochemistry and Molecular Biology*, 47:23–35, April 2014. ISSN 09651748. doi: 10.1016/j.ibmb.2014.02.001.
51. Michael Parisi, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. Paucity of genes on the drosophila X chromosome showing male-biased expression. *Science*, 299(5607):697–701, January 2003. ISSN 00368075.
52. Richard P. Meisel, John H. Malone, and Andrew G. Clark. Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Res.*, 22(7):1255–1265, January 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.132100.111.
53. Qi Zhou and Doris Bachtrog. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science*, 337(6092):341–345, July 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1225385.
54. C Pisano, S Bonaccorsi, and M Gatti. The kl-3 loop of the Y chromosome of Drosophila melanogaster binds a tektin-like protein. *Genetics*, 133(3):569–579, March 1993. ISSN 1943-2631. doi: 10.1093/genetics/133.3.569.
55. Amy L. Dapper and Michael J. Wade. Relaxed selection and the rapid evolution of reproductive genes. *Trends in Genetics*, 0(0), July 2020. ISSN 0168-9525. doi: 10.1016/j.tig.2020.06.014.
56. Helen M. Southern, Mitchell A. Berger, Philippe G. Young, and Rhonda R. Snook. Sperm morphology and the evolution of intracellular sperm–egg interactions. *Ecology and Evolution*, 8(10):5047–5058, May 2018. ISSN 2045-7758. doi: 10.1002/ece3.4027.
57. Bahar Patlar, Vivek Jayaswal, José M. Ranz, and Alberto Civetta. Non-adaptive molecular evolution of seminal fluid proteins in Drosophila. *Evolution*, n/a(n/a), 2021. ISSN 1558-5646. doi: 10.1111/evo.14297.
58. Yael Gur and Haim Breitbart. Mammalian sperm translate nuclear-encoded proteins by mitochondrial-type ribosomes. *Genes Dev*, 20(4):411–416, February 2006. ISSN 0890-9369. doi: 10.1101/gad.367606.
59. Timothy L. Karr, William J. Swanson, and Rhonda R. Snook. The evolutionary significance of variation in sperm–egg interactions. In Tim R. Birkhead, David J. Hosken, and Scott Pitnick, editors, *Sperm Biology*, pages 305–365. Academic Press, London, January 2009. ISBN 978-0-12-372568-4. doi: 10.1016/B978-0-12-372568-4.00008-2.
60. Scott Pitnick, Mariana F. Wolfner, and Steve Dorus. Post-ejaculatory modifications to sperm (PEMS). *Biological Reviews*, 95(2):365–392, 2020. ISSN 1469-185X. doi: 10.1111/brv.12569.
61. Bettina E. Fischer, Elizabeth Wasbrough, Lisa A. Meadows, Owen Randlet, Steve Dorus, Timothy L. Karr, and Steven Russell. Conserved properties of *Drosophila* and human spermatozoal mRNA repertoires. *Proc. R. Soc. B.*, 279(1738):2636–2644, July 2012. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2012.0153.
62. Simone Immler. The sperm factor: Paternal impact beyond genes. *Heredity*, 121(3):239–247, September 2018. ISSN 1365-2540. doi: 10.1038/s41437-018-0111-0.
63. Stephen A. Krawetz. Paternal contribution: New insights and future challenges. *Nat Rev Genet*, 6(8):633–642, August 2005. ISSN 1471-0064. doi: 10.1038/nrg1654.
64. Bárbara Matos, Stephen J. Publicover, Luis Filipe C. Castro, Pedro J. Esteves, and Margarida Fardilha. Brain and testis: More alike than previously thought? *Open Biology*, 11(6): 200322, 2021. doi: 10.1098/rsob.200322.
65. Frank W. Avila, Laura K. Sirot, Brooke A. LaFlamme, C. Dustin Rubinstein, and Mariana F. Wolfner. Insect seminal fluid proteins: Identification and function. *Annu. Rev. Entomol.*, 56 (1):21–40, 2011. ISSN 0066-4170. doi: 10.1146/annurev-ento-120709-144823.
66. D M Neubaum and M F Wolfner. Mated Drosophila melanogaster females require a seminal fluid protein, Acp36DE, to store sperm efficiently. *Genetics*, 153(2):845–857, October 1999. ISSN 0016-6731.
67. Frank W. Avila and Mariana F. Wolfner. Cleavage of the Drosophila seminal protein Acp36DE in mated females enhances its sperm storage activity. *Journal of Insect Physiology*, 101:66–72, August 2017. ISSN 0022-1910. doi: 10.1016/j.jinsphys.2017.06.015.
68. Yasset Perez-Riverol, Jingwen Bai, Chakradhar Bandla, David García-Seisdedos, Suresh Hewapathirana, Selvakumar Kamatchinathan, Deepti J Kundu, Ananth Prakash, Anika Frericks-Zipper, Martin Eisenacher, Mathias Walzer, Shengbo Wang, Alvis Brazma, and Juan Antonio Vizcaíno. The PRIDE database resources in 2022: A hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, 50(D1):D543–D552, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1038.

Supplementary Note 1: Supplementary information

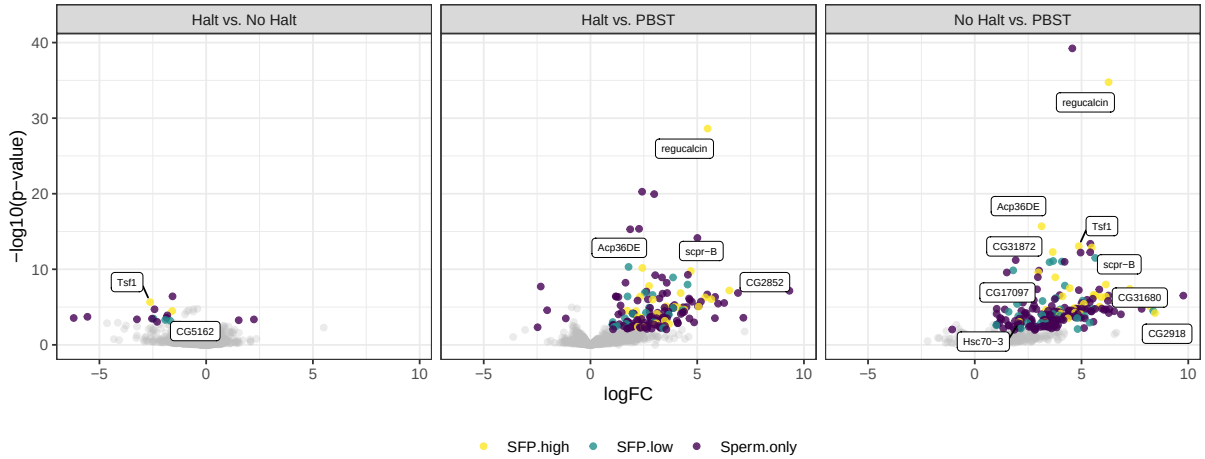

Fig. S1. Volcano plots from pairwise analyses between treatments in experiment two, denoting 'high confidence' (yellow) and 'low confidence/transferred' (turquoise) Sfps or remaining sperm proteins (purple) that showed significant differences in abundance based on a |logFC| > 1 and false discovery rate corrected *p*-value < 0.05. Several Sfps are labelled that showed differential abundance between treatments.

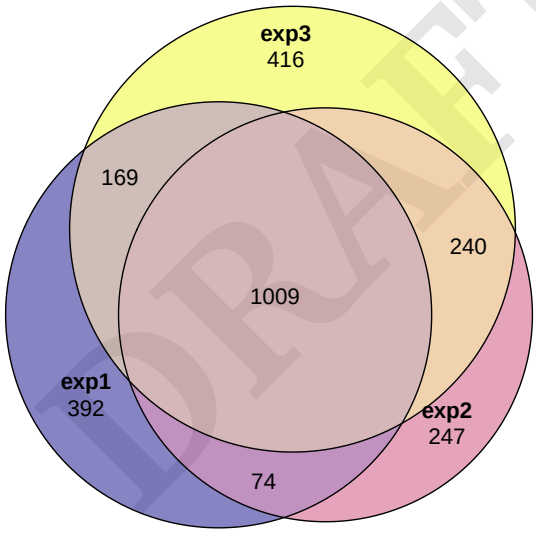

Fig. S2. Overlap between proteins identified in each experiment in the current study.

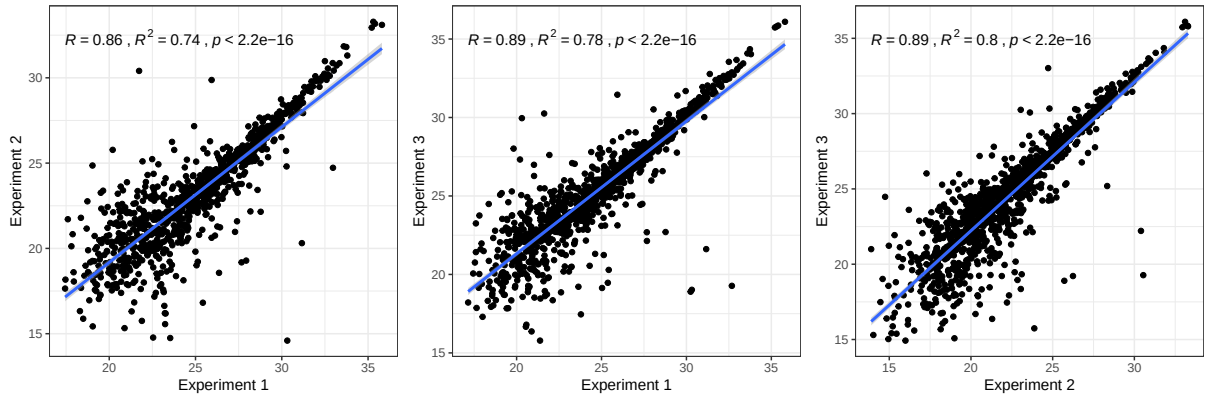

Fig. S3. Correlations in average protein abundance between each experiment in the current study. Mean protein abundance was calculated across all replicates for each experiment, except experiment two which excluded the PBST treatment. Shown are Pearson's correlations and line of best fit.

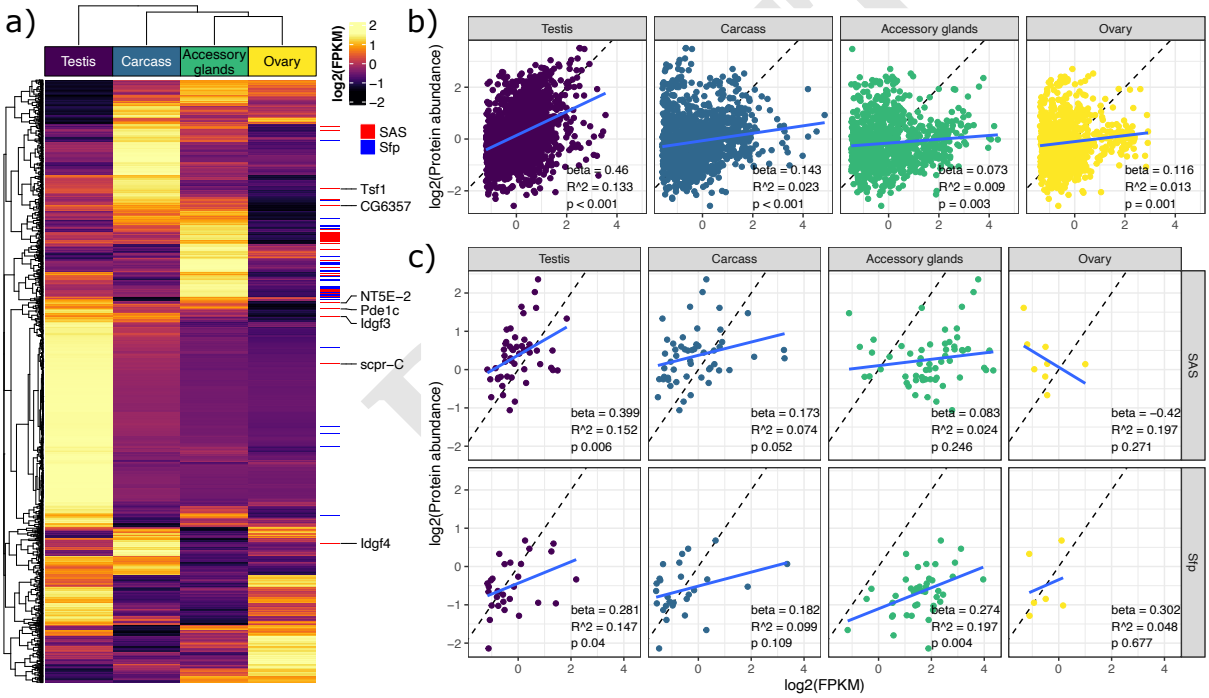

Fig. S4. Gene – protein abundance concordance in the DmSP3. a) Heatmap of mRNA expression of DmSP3 genes (n = 2673) in the accessory glands, carcass, ovary, and testis. Data retrieved from from FlyAtlas2 (43) are log2(FPKM) scaled per gene. The 7 'high confidence' Sfps with higher expression in the testis than accessory glands are highlighted on the right. Labels on the right also show 'sperm associated Sfps' (red) and other 'high confidence' Sfps (blue) identified in the DmSP3. b) Linear regressions of gene expression on protein abundance in the testis (n = 1498), carcass (n = 1165), accessory glands (n = 1001), and ovary (n = 825). c) Linear regressions of gene expression on protein abundance for 'sperm associated Sfps' (SAS) and remaining Sfps identified in the DmSP in each tissue. b) and c) are linear regressions using z-score log2-transformed values after filtering genes with log2-FPKM < 2. Blue lines are model fits from a linear regression, dashed lines indicate a perfect correlation between gene expression and protein abundance.

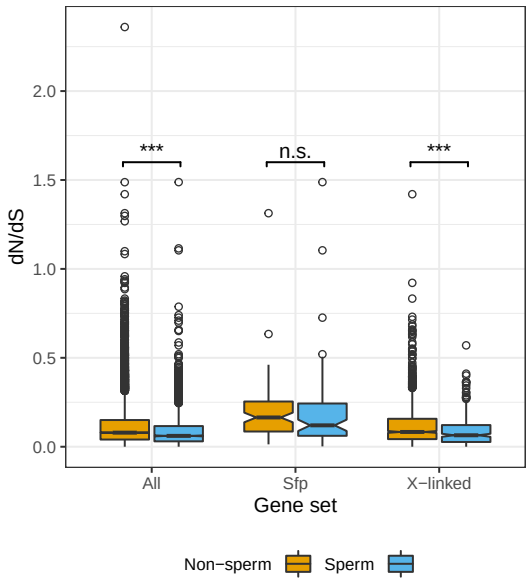

Fig. S5. Nonsynonymous (dN) to synonymous (dS) nucleotide substitution rate (dN/dS) estimates for proteins in the DmSP3 or elsewhere. Asterisks represent results from Mann-Whitney U tests; n.s., non-significant; ***, $p < 0.001$.