Within-ejaculate haploid selection reduces disease biomarkers in human sperm

Daniel Marcu[1], Jayme Cohen-Krais[1], Alice Godden[1], Ghazal Alavioon[2], Carlo Martins[3], Kristian Almstrup[4,5], Gerhard Saalbach[3], Simone Immler[1,2]*

[1] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK

[2] Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden

[3] John Innes Centre

[4] Department of Growth and Reproduction, Copenhagen University Hospital – Rigshospitalet, Denmark

[5] Department of Cellular and Molecular Medicine, University of Copenhagen, Denmark

*** Corresponding author:** S. Immler (s.immler@uea.ac.uk); D. Marcu (d.marcu@uea.ac.uk)

Keywords: haploid selection, sperm selection, fertility, reproduction, sperm proteomics, sperm phenotype

SUMMARY PARAGRAPH

Males produce millions of sperm in a single ejaculate, but only a single sperm fertilises the egg. Fertilisation is the perfect opportunity for natural selection where sperm of poor quality are eliminated, and the fittest sperm sire the offspring. However, the biomarkers of the fittest sperm are currently elusive. Here, we selected human sperm pools for fitness within ejaculates of healthy donors using *in vitro* assays and identify phenotypic, genomic and proteomic biomarkers of fitter sperm. Fitter sperm showed prolonged motility and incrsed morphological normality and DNA condensation, and their genomes and proteomes differed from less fit sperm. The genes diverging between sperm pools are involved in neurodevelopmental processes, cell signalling and cell proliferation and are linked to heritable diseases including neurodevelopmental disorders and cancers. The genomic signatures of haploid selection in human sperm strongly overlapped in function with the genomic signature in zebrafish sperm suggesting that the sperm functions under haploid selection are highly conserved. The five top peptides downregulated in fitter sperm across all donors are known markers for inflammation, immune function and cancer-related cell proliferation. Selecting for high-performance long-lived sperm prior to fertilisation may therefore help prevent ageing-related heritable diseases in the offspring later in life.

MAIN TEXT

A human ejaculate contains between 20 and 150 million sperm, but only one fertilises the egg. Sperm within the ejaculate of a male vary phenotypically and genetically, offering a great opportunity for selection to act upon [1–3]. The genetic and phenotypic traits of healthy human sperm that successfully reach the site of fertilisation, fuse with the oocyte and sire healthy offspring are still elusive. Evidence for the potential benefits of within ejaculate selection mainly comes from externally fertilising species where selection for longer lived sperm improves offspring fitness [4–7]and show genetic divergence from shorter lived sperm[6]. Evidence of post-meiotic gene expression and incomplete sharing among haploid mouse spermatids further supports the idea that the haploid genotype of sperm may directly influence the sperm phenotype, with a list of putative Geno-Informative Markers (GIMs) identified in mammals [8]. However, the genes and associated molecular markers under haploid selection in sperm are not known. Identifying molecular biomarkers of high-quality sperm is not only a key to understanding the fundamental aspects of human reproduction and fertility, but it will also help improving selecting sperm for the use of Assisted Reproductive Technology [9,10]. Here, we provide a list of genes linked to phenotypic traits in human sperm and show that selection for sperm fitness based on longevity and performance results in a reduction of biomarkers for cancer and other heritable diseases.

The phenotypic traits of sperm selected for fitness

We used the ejaculates from 35 apparently healthy volunteers (Suppl. Table S1) and selected sperm within the ejaculate of each donor based on their ability to survive for an incubation period of four (T4) and 24 hours (T24) and successfully perform a swim-up assay following the incubation period (Figure 1A). We compared the sperm collected from swim up assays at T4 and T24 with raw, washed and fresh sperm after sim up (T0) samples for sperm kinematics, morphology, chromatin and DNA integrity. As expected, the fraction of viable (motile or non-motile) sperm dropped from around 60% in fresh and washed samples to 50% after 4 hours incubation and 25% after 24 hours of incubation prior to swim-up (Suppl. Fig. S1 & Suppl. Table S2). Sperm kinematics continuously improved from T0 to T24 (Figure 1B; Suppl. Table S3 – S10; Suppl. Fig S2) as did gross morphology (Figure 1D; Supplementary Table S11) and chromatin condensation (Figure 1C, Suppl. Table S12), whereas DNA integrity improved in swim up samples at T0 but continuously decreased at T4 and T24 (Figure 1E; Suppl. Table S13).

The genomic signature of haploid selection in human sperm

To link the phenotypic differences between fitter and less fit sperm described above with genotypic differences, we performed two selection assays, one Methyl Cellulose and one using Swim Up as described above. For the Methyl Cellulose assays (Figure 1A; see Methods for details), we used ejaculates from four donors and selected for fitter sperm by placing 300µl of raw ejaculate into the centre of a glass staining dish containing methylcellulose with a radius of 4 cm and incubated for two hours at 37°C with 5% CO2 supplementation. Following the incubation period, we collected 4 ml aliquots from the centre (less fit sperm) of the dish and from the edge (fitter sperm) where at least >10 cells were found in a field of vision (minimum radius ranged from 2-3 cm distance from centre). For the Swim Up assays (Figure 1A, see Methods for details), we used ejaculates from five donors, for which we collected a raw sub-sample (T0, less fit sperm) and a sub-sample from the top of swim up assays at time points T4 and T24 (fitter sperm). For each time point and sperm pool, we standardised the number of cells to 2 million/ml and performed DNA extraction and library preparation without PCR steps to avoid amplification bias in the data. We performed whole-genome sequencing on the less fit (Centre/T0) and fitter (Edge/T4/T8/T24/T48 etc.) sperm pools from all nine donors to an average of 100x per sample (Suppl.Tables S16&17; Suppl. Fig S4). We identified all heterozygous sites (loci with at least two alleles) in each donor (coverage was not associated with significance; Suppl. Table S17; Suppl. Fig. S4), filtered them for a minimum total read number (sum of alleles) of 32 per locus and performed likelihood ratio tests on allele frequencies at these loci to test for significant deviations from Mendelian frequencies using the Centre/T0 pools as reference. We compared allele frequencies at $>2x10^6$ loci per donor (mean = 2,280,804; min = 2,022,290, max = 2,754,360; Suppl. Table S14) between less fit and fitter sperm pools using maximum-likelihood ratio tests with a highly rigorous significance cutoff point of the top > 99th quantile loci. We observed statistically significant divergence in 0.05% of loci among pools in all donors across different assays and time points (mean=1139; min=783, max=2578; Figure 2 A&B; Suppl. Table S15) with most genetic divergence occurring on autosomes. We used these data to look for indications of differential longevity of X- and Y-chromosomes [11,12] by comparing the coverage of X- and Y-linked loci, but we found no evidence of differential performance of X- and Y-sperm at any time point (Suppl. Table S17; Suppl. Fig S10).

Based on the significantly diverging loci, we mapped around 500 genetic elements (mean=511; min=340, max=697) per sample across all time points and found no significant differences in representation of genetic elements across donors and time points (χ^2 = 103.42, *df = 228*, *p* = 1), and a consistent distribution of protein-coding genes and non-regulatory elements (Figure 1G). Identified elements included protein coding genes (mean = 64.0%; min= 58.4%, max= 69.4%) and lncRNAs (mean = 31.6%; min = 26.4%, max = 37.5%) as predominant categories. Gene coding sequences diverging between the sperm pools were longer than the average coding gene length in the reference genome (T2 pools: *t* = 8.62, *p* <0.001; T4 pools: *t* = 5.66, *p* <0.001; T24 pools; *t* = 7.42, *p* <0.001; T48 pools: *t* = 4.42, *p* <0.001; Suppl. Fig. S5).

For discussion of gene functions and Gene Ontology (GO), we here focus on results from Methyl Cellulose and Swim Up at T4, because the phenotypic analyses suggested a strong effect of selection on phenotypes at T4 and because the terms overlapped across time points (for T8/T24/T48 see Suppl. Fig S6, S8 & S9). Gene Ontology (GO) analyses revealed a significant enrichment for genes related to cell assembly and neuronal. In Methyl Cellulose, diverging genes were enriched for biological processes associated with cell signalling and structure (transmembrane receptor protein tyrosine kinase signalling pathway, cell junction assembly, neuron development, Figure 2B) and molecular functions associated with cell communication and membrane transport (protein serine kininase activity, G protein-coupled neurotransmitter receptor activity, calcium ion binding and glutamate receptor; Suppl. Fig. S8). In Swim Up T4, the top most enriched biological processes were primarily associated with nervous system development (synapse assembly, cell morphogenesis involved in neuron differentiation and chemical synaptic transmission; Figure 2D) and the molecular processes were related to cellular transport and cell adhesion (calcium ion binding, glutamate receptor activity and calcium-activated potassium channel activity Suppl. Fig. S9) across all selection time points and donors. Overall, our results show that most of the genetic divergence observed between sperm pools primarily occurs in genomic regions relevant to neuronal and synaptic development.

The top five significantly diverging genes of 19 genes found in at least six of the nine donors, and in at least half of all assays and time points (20 in total, Supplementary Table S18) were: Discs Large MAGUK Scaffold Protein 2 (*DLG2*), Receptor-type Tyrosine-Protein Phosphatase Delta (*PTPRD*), CUB and Sushi Multiple Domains 1 (*CSMD1*),

Cadherin 18 (*CDH18*) and Neurexin 1 (*NRXN1*). *DLG2*, *PTPRD*, *CSMD1*, *CDH18* are well described tumour suppressor genes [13,14] and *NRXN1* plays a role in lung cancer development. All five genes have also been associated with an increased risk for schizophrenia, autism spectrum disorder (ASD) and (neuro)developmental disorders as well as the formation of an inflammasome and senescence [15–18]. *CSMD1* is associated with male and female infertility [19], and *DLG2* is downregulated in human sperm with compromised performance [20], *CDH18* plays a role in cell adhesion in testes [21] and deletions in *NRXN1* were observed in azoospermic patients [22].

We also looked for statistical overlap in significantly diverging genes among donors in Methyl Cellulose and Swim UP T4 assays separately. We found five genes shared among all four donors in Methyl Cellulose (Suppl. Fig. S7A): RNA binding protein, fox-1 (*RBFOX1)*, egl-9 family hypoxia inducible factor 3 (*EGLN3*), cadherin13 (*CDH13*), LDL receptor related protein 1B (*LRP1B*) and long intergenic non-protein coding RNA 960 (*LINC00960*). We found two genes shared across all five donors in Swim Up T4: cadherin 18 (*CDH18*) and Zinc finger protein, multitype 2 protein (*ZFPM2*) (Suppl. Fig. S7B). *ROBFOX1* and *CDH13* play a role in neuronal development [23,24] whereas *CDH18*, *EGLN3*, *LRP1B* and *LINC00960* are involved in cancer cell migration and tumour growth and suppression [25–27].

Evolutionary conserved genomic signature of haploid selection between humans and zebrafish

We compared the genes diverging between sperm pools in humans with genes diverging between sperm pools in zebrafish using a published dataset from our lab and applying human gene annotations [6]; see Suppl. Tables S19S21; Suppl. Fig. S11-S12). Enrichment analyses on the zebrafish revealed biological processes such as cell-to-cell processes and neurodevelopmental processes (differentiation and generation of neurons, cell signalling, developmental cell differentiation and organ formation; Fig 2F), while molecular functions involved ion transport and channel activity (calcium activated cation channel activity, sodium channel activity and ion transmembrane transporter activity; Suppl. Fig. S13).

We found that a total of 153 genes shared among human Methylcellulose and Swim up T4 and the human homologs in zebrafish. The top 20 significant biological GO terms for these 153 genes indicated functions around neural development and cell proliferation (Figure 2I&J). We also ran a semantic similarity analysis between GO terms identified in the human

and the human homologs of zebrafish genes. Our analysis revealed a high degree of semantic similarity with scores of 0.85 (Methyl Cellulose) and 0.7 (Swim UP T4) (Suppl. Fig S14). Finally, three of the five top candidate genes in humans were also significantly diverging among sperm pools in zebrafish, namely *DLG2/ dlg2*, *CSMD1/ csmd1a* and *CDH18/ cdh18*. Altogether, our results indicate that despite the evolutionary divergence of over 450 million years between zebrafish and humans [28]), the biological processes and gene functions under haploid selection in sperm are highly conserved across taxa.

Proteomic biomarkers of selected sperm

We performed Methylcellulose and Swim Up T4 assays in an additional four human donor samples (two per assay) to compare the proteome of fitter and less fit sperm using Tandem Mass Tag (TMT) proteomics analyses (see Methods for details). For one donor, we performed a second TMT run as a technical replicate (Donor A in Swim Up, only sample containing enough sperm for two TMT runs), and we added this run into our statistical and filtering steps for improved rigour. We carefully filtered the raw data following a data processing and consensus workflow (see Methods for details). The overall number of proteins identified across the sperm samples of all four donors was 6864, of which 4556 overlapped in all donors (Suppl. Fig. S15) and 1719 were significantly differentially expressed in at least one donor ($\log_2$-fold change > 2 and adjusted $p < 0.05$; Figure 3B). The significant KEGG GO terms of all significantly differentially expressed proteins across all donors yielded a list of processes involved in basic cellular processes and metabolic pathways, and proteins involved associated with neurodegenerative disorders and metabolic disorders (Figure 3D).

We found four significantly differentially down-regulated proteins in selected sperm (Edge/T4) shared across all donors and runs (Figure 3B), Protein-Glutamine Gamma-Glutamyltransferase 4 (TGM4), Testis-specific Serine Kinase Substrate (TSKS), Phospholipase A2 (PLA2) and Olfactomedin-4 (OLFM4). Using an additional rigorous filtering step based on significance levels ($p_{adj} > 0.05$) and the median ratio of protein abundance among sperm pools across all four donors, we identified five key proteins that were down-regulated in selected sperm of all four donors and potential cell-surface biomarkers for sperm selection assays, two of which were also in the statistically significant top list: OLFM4, PLA2, Beta Defensin 105 (DEFB105A), Heterogeneous Nuclear Ribonucleoprotein K (protein K) and Midkine. Six proteins (all except TSKS) are associated

with human disease phenotypes including various cancers, more general inflammation and cell growth [29–33] whereas TSKS is associated with sperm motility and a potential target for male contraception [34].

Discussion

Within-ejaculate selection for sperm fitness based on a combination of longevity and penetration ability resulted in a significant improvement of overall phenotypic sperm quality. These phenotypic differences clearly have a genomic bases as indicated by the finding of genome-wide signs of selection involving genes playing fundamental roles in organismal health and fitness across several generations [7]. The idea that selection on sperm fitness may improve not only sperm quality but also affect offspring fitness is further supported by our findings of a significant decline in proteins that are established biomarkers of heritable diseases including cancer and neurodevelopmental disorders.

The genomic markers of haploid selection are remarkably evolutionarily conserved and show strong parallels between species as distant as humans and zebrafish. Besides a list of coding regions showing genetic divergence in both species we also find that non-coding regions and lncRNAs in human sperm and lincRNAs in zebrafish sperm are particularly frequent in our data. lnc/linc RNAS have very general widespread regulatory functions for downstream gene expression [35], and our finding suggests significant regulatory activity in genes determining in sperm function. In fact, differences in lncRNA expression were also found in sperm of men with compromised fertility [36], where they regulate genes involved in cytoplasmic transport, biosynthetic processes and spermatogenesis. Interestingly, we found little overlap with GIMs identified in mammals [8].

The genomic and proteomic signatures of selected sperm strongly suggest that this gametic stage may play a key role in selecting for high-functioning genomes and that this stage has the potential to reduce the passing down of heritable diseases to the next generations and improve overall health and fitness. The link between sperm fitness and offspring fitness has been shown in externally fertilising species before, including the zebrafish [4–7], and here we provide the genomic architecture underlying these differences. Neurodevelopmental disorders indicate early developmental differences, which may also explain the 5-7% difference in survival during early zebrafish embryo development [6]. In contrast, cancer is a well-known heritable ageing-related disease, and the finding that proteomic cancer

biomarkers are reduced on fitter sperm is in line with our previous finding of overall higher reproductive fitness and lifespan in zebrafish [7]. Moreover, alleles exposed to haploid sperm selection are directly linked to fertility as indicated by the enrichment for cell adhesion and calcium channel activity in both human and zebrafish sperm. In mammals, these pathways respond to chemo-attractants, pH and oviductal ligands [37] and in humans, the CatSper ion channel is activated in the fallopian tube, which leads to Ca^{2+} influx into the cell and leads to hyperpolarisation and hyperactivation [38]. Whilst zebrafish sperm lack CatSper, they have cyclic nucleotide-gated channels located on the sperm head that react to the low K^{+} in freshwater and hyperpolarise the cells to allow Ca^{2+} efflux resulting in activation and guidance into the egg's micropyle [39].

While we observed major overlaps in the general patterns across individuals, approaches (experimental assays, proteomics and genomics) and species, we also observed notable differences in the exact identity of diverging genes and proteins, which supports the idea that sperm fitness - just as organismal fitness - is governed by a large network of genes. Allelic divergence in these genes varies across individuals and these gene networks are likely a result of additive and epistatic effects, which require an in-depth population genomic approach with large sample sizes to fully disentangle the genomic architecture of haploid selection. Nevertheless, our study sets a new path into our understanding of the key role of gametic selection for organismal health and fitness and opens new doors for its use in applied fields such as Assisted Reproductive Technologies, and IVF and ICSI in particular [10]. Understanding the biology of the fittest sperm is a key step towards improving the success rates and outcome of these technologies in human reproduction and fertility as well as animal stock breeding and aquaculture.

REFERENCES

1. Joseph, S. B. & Kirkpatrick, M. Haploid selection in animals. *Trends Ecol. Evol.* **19**, 592–597 (2004).
2. Immler, S. & Otto, S. P. Evolution of haploid selection in predominantly diploid organisms - PubMed. https://pubmed.ncbi.nlm.nih.gov/26669442/.

3. Immler, S. Haploid Selection in “Diploid” Organisms | Annual Reviews. https://www.annualreviews.org/content/journals/10.1146/annurev-ecolsys-110218-024709.

4. Crean, A. J., Dwyer, J. M. & Marshall, D. J. Fertilization Is Not a New Beginning: The Relationship between Sperm Longevity and Offspring Performance. *PLOS ONE* **7**, e49167 (2012).

5. Immler, S., Hotzy, C., Alavioon, G., Petersson, E. & Arnqvist, G. Sperm variation within a single ejaculate affects offspring development in Atlantic salmon. *Biol. Lett.* **10**, 20131040 (2014).

6. Alavioon, G. *et al.* Haploid selection within a single ejaculate increases offspring fitness. *Proc. Natl. Acad. Sci.* **114**, 8053–8058 (2017).

7. Alavioon, G., Cabrera Garcia, A., LeChatelier, M., Maklakov, A. A. & Immler, S. Selection for longer lived sperm within ejaculate reduces reproductive ageing in offspring. *Evol. Lett.* **3**, 198–206 (2019).

8. Bhutani, K. *et al.* Widespread haploid-biased gene expression enables sperm-level natural selection. *Science* **371**, eabb1723 (2021).

9. Alavioon, G., Marcu, D. & Immler, S. Within-Ejaculate Sperm Selection and Its Implications for Assisted Reproduction Technologies. in *XIIIth International Symposium on Spermatology* (eds. Björndahl, L., Flanagan, J., Holmberg, R. & Kvist, U.) 127–133 (Springer International Publishing, Cham, 2021). doi:10.1007/978-3-030-66292-9_20.

10. van Oosterhout, C., Marcu, D. & Immler, S. Accounting for the genetic load in assisted reproductive technology. *Clin. Transl. Med.* **12**, e864 (2022).

11. You, Y.-A. *et al.* Sex chromosome-dependent differential viability of human spermatozoa during prolonged incubation. *Hum. Reprod.* **32**, 1183–1191 (2017).

12. Song, W.-H. *et al.* Effect of endocrine disruptors on the ratio of X and Y chromosome-bearing live spermatozoa. *Reprod. Toxicol.* **82**, 10–17 (2018).

13. Griesius, S. *et al.* Reduced expression of the psychiatric risk gene DLG2 (PSD93) impairs hippocampal synaptic integration and plasticity. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **47**, 1367–1378 (2022).

14. Yu, X. *et al.* Protein tyrosine phosphatase receptor-type δ acts as a negative regulator suppressing breast cancer. *Oncotarget* **8**, 98798–98811 (2017).

15. Cukier, H. N. *et al.* Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol. Autism* **5**, 1 (2014).

16. Kushima, I. *et al.* Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights. *Cell Rep.* **24**, 2838–2856 (2018).

17. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).

18. Ishizuka, K. *et al.* Functional characterization of rare NRXN1 variants identified in autism spectrum disorders and schizophrenia. *J. Neurodev. Disord.* **12**, 25 (2020).

19. Lee, A. S. *et al.* Rare mutations in the complement regulatory gene CSMD1 are associated with male and female infertility. *Nat. Commun.* **10**, 4626 (2019).

20. Joshi, M. *et al.* Coding and regulatory transcriptome comparisons between fertile and infertile spermatozoa identify RNA signatures of male infertility. *Andrologia* **54**, e14437 (2022).

21. Xx, L. *et al.* Cadherin-18 loss in prospermatogonia and spermatogonial stem cells enhances cell adhesion through a compensatory mechanism. *Zool. Res.* **45**, (2024).

22. Liu, Y., Shi, D., Zhang, F., Dai, L. & Ma, W. Analysis of copy number variations of the autosomal genome in 156 patients with azoospermia. *Transl. Androl. Urol.* **11**, 675–682 (2022).

23. Fogel, B. L. *et al.* RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum. Mol. Genet.* **21**, 4171–4186 (2012).

24. Mossink, B. *et al.* Cadherin-13 is a critical regulator of GABAergic modulation in human stem-cell-derived neuronal networks. *Mol. Psychiatry* **27**, 1–18 (2022).

25. Jin, Y. *et al.* Inactivation of EGLN3 hydroxylase facilitates Erk3 degradation via autophagy and impedes lung cancer growth. *Oncogene* **41**, 1752–1766 (2022).

26. Príncipe, C., Dionísio de Sousa, I. J., Prazeres, H., Soares, P. & Lima, R. T. LRP1B: A Giant Lost in Cancer Translation. *Pharmaceuticals* **14**, 836 (2021).

27. Elango, R. *et al.* Long noncoding RNA profiling unveils LINC00960 as unfavorable prognostic biomarker promoting triple negative breast cancer progression. *Cell Death Discov.* **10**, 1–12 (2024).

28. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).

29. Vecchi, L. *et al.* Phospholipase A2 Drives Tumorigenesis and Cancer Aggressiveness through Its Interaction with Annexin A1. *Cells* **10**, 1472 (2021).

30. Uhlitz, F. *et al.* Mitogen activated protein kinase activity drives cell trajectories in colorectal cancer. *EMBO Mol. Med.* **13**, e14123 (2021).

31. Ma, H. W. *et al.* Olfactomedin 4 produces dysplasia but suppresses metastasis of colon cancer. *Cancer Gene Ther.* **30**, 694–703 (2023).

32. Neumaier, E. E., Rothhammer, V. & Linnerbauer, M. The role of midkine in health and disease. *Front. Immunol.* **14**, (2023).

33. Sudhakaran, M. & Doseff, A. I. Role of Heterogeneous Nuclear Ribonucleoproteins in the Cancer-Immune Landscape. *Int. J. Mol. Sci.* **24**, 5086 (2023).

34. Salicioni, A. M. *et al.* Testis-specific serine kinase protein family in male fertility and as targets for non-hormonal male contraception†. *Biol. Reprod.* **103**, 264–274 (2020).

35. Dai, Y.-J., Hu, F., He, S.-Y. & Wang, Y.-Y. Epigenetic landscape analysis of lncRNAs in acute myeloid leukemia with DNMT3A mutations. *Ann. Transl. Med.* **8**, 318–318 (2020).

36. Zhang, X. *et al.* Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *Int. J. Mol. Sci.* **20**, 5573 (2019).

37. Brenker, C. *et al.* Synergistic activation of CatSper Ca2+ channels in human sperm by oviductal ligands and endocrine disrupting chemicals. *Hum. Reprod.* **33**, 1915–1923 (2018).

38. Navarro, B., Kirichok, Y., Chung, J.-J. & Clapham, D. E. Ion channels that control fertility in mammalian spermatozoa. *Int. J. Dev. Biol.* **52**, 607–613 (2008).

39. Fechner, S. *et al.* A K+-selective CNG channel orchestrates Ca2+ signalling in zebrafish sperm. *eLife* **4**, e07624 (2015).

METHODS

Sample collection from human donors

For the assessing the effects of sperm selection on sperm phenotypic quality, we used the samples from 35 self-reported healthy donors collected at the University of East Anglia, UK.

For the genomics analyses, we included ejaculates from nine self-reported healthy volunteers with sperm concentration >15 M/ml: nine samples used for genomics analyses were collected at the Rigshospitalet Copenhagen, Denmark.

The samples were collected near the laboratory and allowed to liquefy for 30 minutes at 37°C. Prior to the selection assays, a portion of the raw ejaculates (200µl) from each donor was stored at -80°C to be used as the reference "low performance sperm pool" for later allele frequency estimation.

For proteomics analyses we used four human sperm ejaculates: two from the University of East Anglia (UEA), Norwich, UK; two from Cape Town, South Africa (ZA)).

The samples from University of East Anglia and Cape Town were collected under ethics permit no. ETH2324-0997 and the samples collected at the Rigshospitalet Copenhagen were collected under ethics permit no. H-17012149.

Sperm selection assays

We used two different assays to select for sperm within ejaculates:

Methyl Cellulose Assay: A 10 cm-diameter watch glass was filled with 1% (w/v) methylcellulose (Sigma, Denmark) up to a 4cm radius and 10ml total volume. Then, 300µl of raw ejaculate was placed in the centre of the glass (the ejaculate was allowed to sink and filled a circle of 0.5cm in the centre of the glass). For each donor, we used five 300 µl aliquots allocated into five watch glasses per donor in parallel to ensure sufficient cells in the edge population for subsequent whole genome sequencing. The samples were incubated in the watch glasses for two hours at 37°C under 5% CO2 supplementation. 4ml aliquots were then collected from the outer circle of each watch glass at a radius of 3-3.5 cm, 2.5-3cm or 2-2.5cm (depending on the distance travelled by the sperm starting at the radius from the further circle with >10 cells under the microscope for each male) and placed in a 15ml tube. All five edge samples collected from the outer circle were pooled for each donor to ensure high enough cell numbers for sequencing. We also collected a sample (5 ml) from

the centre of each watch glass at least 0.5 cm above the bottom of the glass to avoid contamination with dead sperm and somatic cells. To extract the sperm from the methylcellulose, 1µl of cellulase was added for every 1ml of collected sample and the sample was mixed thoroughly on a shaking table for two hours at 50°C. The samples were then centrifuged, the supernatant removed, and the sperm pellets were stored at -80° C for subsequent DNA extraction.

Swim Up Assays: For the phenotypic analyses, we collected a raw (R T0), washed (W T0), and sub-samples from swim ups following incubation (T0, T4, T8, T24, T48) and for the genomic analyses we collected a raw sample (T0) and sub-samples from swim ups following incubation (T4, T24 for all five donors and T8 (n = 1 donor), T24 and T48 (n = 3 donors).

For the swim ups sub-samples, aliquots of 1ml raw ejaculate were placed in 15ml centrifuge tubes at a 1:1 ratio of human tubal fluid (HTF) medium containing (in mM): 72.8 NaCl, 4.69 KCl, 0.2 $MgSO_4$, 0.37 KH_2PO_4, 2.04 $CaCl_2$, 0.33 sodium pyruvate, 21.4 sodium lactate, 2.78 glucose, 21 HEPES, 25 $NaHCO_3$, adjusted to pH 7.4 with NaOH (Fisher, UK) and incubated for 2, 4, 8, 12, 24 and 48 hours at 37°C with 5% CO2 in the air. Motile sperm, at each time point, were recovered by direct swim up as follows: after flicking each tube to homogenise the sample, 1ml was aliquoted and placed under 4ml of HTF medium in a new 15ml centrifuge tube at 37°C for 60 min in 5% CO2 as described above (*Figure 3.1*). The top layer of the swim up (3.5ml) was collected and centrifuged at 700xg for 10-20 min at room temperature (RT) prior to resuspension in 2ml HTF. After assessing sperm concentrations of samples using image cytometry as described in Egeberg et al. (2013), the cells were pelleted, the supernatant removed, and the pellet stored at -80°C until further use.

Phenotypic analyses

Sample prep and selection assay

We used samples from 35 donors for sperm selection by Swim Up assay and subsequent phenotypic analyses. All raw samples were split into three aliquots: one third of the raw sample was used directly for a swim up and the remaining 2/3 were washed and split into two aliquots and incubated for 4h and 24h, respectively. In the washing step, the samples were washed twice in HTF lacking HAS. The sample was centrifuged at 500g for 10 minutes and after removal of the supernatant, 3ml of HFT+ was added and the sample was split again into two labelled tubes (T4 and T24). Following incubation, the samples were centrifuged at

500g for 10 minutes to supernatant and perform a direct swim up to isolate motile cells as described above.

The swim ups were performed in an Eppendorf tube as follows: 1 ml of human tubal fluid (HTF+) medium containing (in mM): 72.8 NaCl, 4.69 KCl, 0.2 $MgSO_4$, 0.37 KH_2PO_4, 2.04 $CaCl_2$, 0.33 sodium pyruvate, 21.4 sodium lactate, 2.78 glucose, 21 HEPES, 25 $NaHCO_3$ and 3.5% human serum albumin (HSA) (Fisher, UK) adjusted to pH 7.4 with HCL was added to the semen/sperm pellet. During the swim up assay, cells were allowed to swim up for 1h at 37°C with 5% CO_2 supplementation after which we collected 700µl of the sample from the top of the Eppendorf tube.

We then assessed phenotypic traits including sperm concentration, motility and velocity, sperm viability, sperm morphology, chromatin and DNA integrity (see Supplementary Methods for details on each of these).

Genomic analyses of sperm human sperm pools

Whole genome sequencing

DNA was extracted using a standard phenol:chloroform:isoamyl alcohol protocol (see Supplementary Methods for details on DNA extraction) from the sperm pools and prepared using a PCR-free library preparation kit (KAPA HyperPrep kit; Roche) for a standard insert size (400-500 bp)(see Supplementary Methods for details).

Bioinformatic analyses of genome data

Allele frequency estimations

Following quality control (see *Supplementary Methods* for details), variant detection of the alignments of reads from the T0 raw ejaculate followed the best practice hard-filtering workflow recommended by GATK. In brief, GATK 4.9.1 Haplotype Caller was used to call the variants and filter the heterozygous paternal sites. Single nucleotide polymorphisms (SNPs) were filtered using SelectVariant and the SNP filter expression was set as QD < 2.0 && FS > 60.0 && MQ < 40.0 && SOR > 3.0 && MQRankSum < -12.5 && ReadPosRankSum < - 8.0 && QUAL < 30. Alignment of the reads from T0 raw ejaculate and selected sperm pools were used to generate base count profiles using pileup2pro2. SAMtools 1.9 option -q30, -Q30 and d100000 was used to exclude sites with low quality, allowing for sites with deep coverage.

The heterozygous sites and base count profiles of the T0 raw and high performance sperm pools were used for likelihood ratio tests using the scrip hetPoolLikelihoods.pl available at[1]https://github.com/douglasgscofield/gameteUtils based on Lynch et al. (2014)[1]. The most significant allelic differences between the T0 raw and T4/24 sperm pools were identified by the largest likelihood ratios. Briefly, the major and minor alleles were assigned according to counts within the T0 raw sperm pools. Sites with a minimum of 32x and a maximum of 2*total mean coverage (mean coverage was 206.69, 200.68, 176.31, 174.9, 261.17, 221.03, 271.5, 239.1023 and 211.443 for 9 donors: Donor 1, Donor 2, Donor 3, Donor 4, Donor 5, Donor 6, Donor 7, Donor 8 and Donor 9, respectively) were excluded. The error rate was estimated based on the number of putative error reads/number of total reads (Lynch et al., Eq 3a). Then, the probability that a random read was recoded as major (ΦM) and minor allele (Φm), respectively, was calculated (Lynch et al. (2014) [1], Eq 1a and 1b). Next, we estimated the ratio of the major allele frequency in each pool (p□) accounting for the putatively non-erroneous and erroneous reads that represent the major allele type (Lynch et al. (2014) [1], Eq 3b). The log likelihood of the data following our model as well as data under the assumption of monomorphism was then computed following Lynch et al., (2014) Eq 4a and 4b, respectively. Finally, the likelihood ratio test statistic was computed following a Chi-squared distribution with 1 degree of freedom (LRT_Maf in hetPoolLikelihoods.pl). To identify statistically significant loci between T0 raw and T4/T8/T24/T48 sperm pools, the threshold for genome-wide significance was set to > 99.9th quantile (critical values Donor1: 17.08, Donor 2: 11.91, Donor 3: 15.36, Donor 4:11.60, Donor 5: 11.37, Donor 6: 11.30, Donor 7: 11.67, Donor 8: 11.37, Donor 9:11.6).

Predictive role of genetically divergent regions

To identify genetic elements residing at statistically significant loci, the statistically significant loci identified using likelihood ratio tests were overlapped with the annotated *Homo sapiens* reference genome using a customised R script. To uncover the biological significance of the genetic differences resulting from the comparisons between T0 raw and selected fit sperm pools, the resulting gene overlaps were then used for gene enrichment and gene length analysis using *ShinyGO v0.75* [2].

Comparison between human and zebrafish data

To identify conserved genetic pathways between zebrafish and human sperm pools selected for fitness, we reanalysed the zebrafish sperm pool data from Alavioon et al. (2017)[3]. The

full experimental design and selection assay for longevity in zebrafish is described in Alavioon et al. (2017) [3]. To process the genomic data, we used the latest zebrafish GRCz11 *Danio rerio* reference genome to align the genomic reads and employed the same analysis pipeline and hard filtering as for human samples (see *Supplementary Methods* for details). To further minimise potential sequencing bias in the zebrafish data, the cut-off point for genome-wide identification of significant loci in the zebrafish data was set to the 99.9th quantile; critical values were 13.99, 26.51, 10.39 for male ID: Male_31, Male_32 and Male_34, respectively. We then used Venn diagrams on significant genes by converting zebrafish genes into human gene homologs to identify genes shared between humans and zebrafish and performed GO analyses using *ShinyGO v0.75*[2] on the 153 shared genes.

Proteomics analyses

Sperm selection assays for proteomics

Samples from four donors in total, two for each assay were used. The samples were washed in human tubal fluid (HTF) and resuspended in 200 µL HTF. We collected two samples representing a less fit (Centre/T0) and a fitter (Edge/T4) sperm pool from each donor from Methyl Cellulose and Swim Up (assays as described above; see Supplementary Methods for exact details) and stored them at -80°C.

Protein extraction and proteomics

For protein extraction, samples were thawed, and 450 ml of acetone was added to each of the thawed samples. Samples were then vortexed for 15 minutes, then centrifuged at 10,000 rpm for 10 minutes. Supernatant was then removed, and 500 ml of acetone added to each sample. Each sample was then spun gently for two minutes manually then centrifuged for 15 mins at 10000 rpm. Supernatant was removed and samples were air dried for 1 and a half hours.

Protein pellets were resuspended in 100 µl of 2.5% sodium deoxycholate (SDC; Merck) in 0.2 M EPPS-buffer (Merck), pH 8.5, and the samples vortexed under heating. The samples were treated with dithiothreitol and iodoacetamide to alkylate cysteine residues and digested with trypsin in the SDC buffer according to standard procedures. After the digest, the SDC was precipitated by adjusting to 0.2% TFA, and the clear supernatant subjected to C18 SPE (Reprosil-Pur 120 C18-AQ, 5 um, Dr. Maisch GmbH, Germany). Peptide concentration was estimated by running an aliquot of the digests on LCMS (see below). TMT labelling was

performed using 8 channels from a TMT™10plex kit (TMT lot: VH306773, ThermoFisher Scientific, Hemel Hempstead, UK) according to the manufacturer's instructions with slight modifications; samples were dissolved in 90 µl of 0.2 M EPPS buffer (MERCK)/10% acetonitrile, and 200 µg TMT reagent dissolved in 22 µl of acetonitrile was added. Samples were assigned to the TMT channels in two separate experiments, according to the table below. After 2 h incubation, aliquots of 1.5 µl from each sample from each individual TMT experiment were combined in 250 µl 0.2% TFA, desalted, and analysed on the mass spectrometer to check the labelling efficiency and estimate total sample abundances. The main sample aliquots were quenched by adding 8 µl of 5% hydroxylamine and combined to roughly level abundances. The peptides were desalted using a C18 Sep-Pak cartridge (200 mg, Waters, Wilmslow, UK). The eluted peptides were dissolved in 500 µl of 25 mM NH_4HCO_3 and fractionated by high pH reversed phase HPLC. The samples were loaded to an XBridge® 3.5 µm C18 column (150 x 3.0 mm, Waters). Fractionation was performed on an ACQUITY Arc Bio System (Waters) with the following gradient of solvents A (water), B (acetonitrile), and C (25 mM NH_4HCO_3 in water) at a flow rate of 0.5 ml min^{-1}: solvent C was kept at 10% throughout the gradient; solvent B: 0-5 min: 5%, 5-10 min: 5-10%, 10-60 min: 10-40%, 60-75 min: 40-80%, followed by 5 min at 80% B and re-equilibration to 5% for 24 min. Fractions were collected every 1 min and concatenated by combining fractions of similar peptide concentration to produce 20 final fractions for MS analysis. A replicate of one donor sample (2 TMT channels) was fractionated using the Pierce™ High pH Reversed-Phase Peptide Fractionation Kit (Thermo) to produce 7 fractions.

TMT analyses

Aliquots of all fractions were analysed by nanoLC-MS/MS on an Orbitrap Eclipse™ Tribrid™ mass spectrometer equipped with a FAIMS Pro Duo interphase coupled to an UltiMate® 3000 RSLCnano LC system (Thermo Fisher Scientific, Hemel Hempstead, UK). The samples were loaded onto a trap cartridge (PepMap™ Neo 5 µm C18 300 µm X 5 mm Trap Cartridge, Thermo) with 0.1% TFA at 15 µl min^{-1} for 3 min. The trap column was then switched in-line with the analytical column (Aurora Frontier TS, 60 cm nanoflow UHPLC column, ID 75 µm, reversed phase C18, 1.7 µm, 120 Å; IonOpticks, Fitzroy, Australia) for separation using the following gradient of solvents A (water, 0.1% formic acid) and B (80% acetonitrile, 0.1% formic acid) at a flow rate of 0.23 µl min^{-1} : 0-3 min 1% B (parallel to trapping); 3-10 min linear increase B to 9 %; 10-105 min increase B to 50%; followed by a ramp to 99% B and re-equilibration to 1% B.

Data were acquired with the following parameters in positive ion mode: MS1/OT: resolution 120K, profile mode, mass range *m/z* 400-1600, AGC target $4e^5$, max inject time 50 ms, FAIMS device set to three compensation voltages (-35V, -50V, -65V) for 1 s each; MS2/IT: for each CV, data dependent analysis with the following parameters: 1 s cycle time Rapid mode, centroid mode, quadrupole isolation window 0.7 Da, charge states 2-5, threshold $1.9e^4$, CID CE=33, AGC target 1e4, max. inject time 50 ms, dynamic exclusion 1 count for 15 s with mass tolerance of 7 ppm; MS3 synchronous precursor selection (SPS): 10 SPS precursors, MS2 isolation window 0.7 Da, HCD fragmentation with CE=65, Orbitrap Turbo TMT and TMTpro resolution 30k, AGC target 200%, max inject time 100 ms; Real Time Search (RTS): protein database UP000005640_human_2024_9606.fasta (Uniprot, 20,609 entries), enzyme trypsin, 1 missed cleavage, oxidation (M) as variable, carbamidomethyl (C) and TMT6plex as fixed modifications, precursor tolerance 10 ppm, Xcorr = 1.4, dCn = 0.1.

Proteome data analyses

The acquired raw data were processed and quantified in Proteome Discoverer 3.1 (Thermo Fisher Scientific); all mentioned tools of the following workflow are nodes of the proprietary Proteome Discoverer (PD) software.

The UP000005640_human_2024_9606.fasta protein fasta database (Uniprot, 20,609 entries) was imported into PD adding a reversed sequence database for decoy searches. The database search was performed using the incorporated search engines CHIMERYS (MSAID, Munich, Germany) and Comet.

The processing workflow for both search engines included recalibration of MS1 spectra (RC), reporter ion quantification by most confident centroid (20 ppm) and a search on the human protein database (as imported above). For CHIMERYS the Top N Peak Filter was applied with 20 peaks per 100 Da. Then the inferys_3.0.0._fragmentation prediction model was used with fragment tolerance of 0.3 Da, enzyme trypsin with 1 missed cleavage, variable modification oxidation (M), fixed modifications carbamidomethyl (C) and TMT6plex on N-terminus and K. For Comet the version 2019.01 rev.0 parameter file was used with default settings except precursor tolerance set to 6 ppm and trypsin missed cleavages set to 1. Modifications were the same as for CHIMERYS.

The consensus workflow included the following parameters: intensity-based abundance, normalisation on total peptide abundances, protein abundance-based ratio calculation, only unique peptides (protein groups) for quantification, TMT channel correction values applied (Lot VH306773), co-isolation/SPS matches/CHIMERYS Coefficient thresholds 50%/70%, 0.8, missing values imputation by low abundance resampling, hypothesis testing by *t*-test (background based), adjusted p-value calculation by BH-method. The results were exported into a Microsoft Excel table including data for normalised and un-normalised abundances, ratios for the specified conditions, the corresponding p-values and adjusted p-values, number of unique peptides, q-values, PEP-values, identification scores from both search engines; FDR confidence filtered for high confidence (strict FDR 0.01) only. We used *AlphaFold* [4,5,6]) to generate molecule views of the candidate proteins.

ACKNOWLEDGEMENTS

We thank Sally Otto and Cock van Oosterhout for feedback on earlier drafts of this manuscript. This research was supported by grants from the European Research Council (ERC, *HapSelA-336633* and *SELECTHAPLOID-101001341*) and Natural and Environmental Research Council (NERC, *NE/S011188/1*) to SI and a BBSRC PhD studentship to DM.

AUTHOR CONTRIBUTIONS

DM: phenotypic data collection, phenotypic data analyses, bioinformatic data analyses of genome sequencing data, sperm selection assays for proteomics, MS writing. JCK: Proteomics sample collection and preparation, proteomics data interpretation, MS writing. AG: Data analyses and interpretation. CM: Proteomics sample preparation and data analysis. GA: sperm selection assay development, sample collection for genomics. GS: Proteomics sample preparation and data analysis. KA: Donor recruitment, sample collection for genomics, ethics permit holder DK. SI: Conception and design of study, supervision of project, ethics permit holder UK, data interpretation, MS writing.

DATA AVAILABILITY

All raw datasets have been uploaded to respective platforms. The sperm phenotypic data, R scripts associated with phenotype statistical analyses, the list of significantly differentially expressed genes form the human genomic analyses and the list of significantly differentially expressed proteins can be found on dryad (doi:10.5061/dryad.4qrfj6qkn). The data files

containing the raw whole-genome sequencing data for all human sperm pools can be found on EBI under accession number PRJEB80525. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD056057 and 10.6019/PXD056057; PXD056228 and 10.6019/PXD056228; and PXD056234 and 10.6019/PXD056234"".

ADDITIONAL INFORMATION

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to SI (s.immler@uea.ac.uk). The authors declare no competing interests.

METHODS REEFERENCES

1. Lynch, M., Bost, D., Wilson, S., Maruki, T. & Harrison, S. Population-Genetic Inference from Pooled-Sequencing Data. Genome Biol. Evol. 6, 1210–1218 (2014).
2. Ge, S. X., Jung, D. & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36, 2628–2629 (2020).
3. Alavioon, G. et al. Haploid selection within a single ejaculate increases offspring fitness. Proc. Natl. Acad. Sci. 114, 8053–8058 (2017).
4. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
5. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439–D444 (2022).
6. Varadi, M. et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 52, D368–D375 (2024).

FIGURE LEGENDS

Figure 1: Experimental design and phenotypic signature of within-ejaculate selection on sperm performance. (**A**) Experimental design for selecting sperm for performance phenotypes using a combination of incubation and Methyl Cellulose or incubation and Swim Up assays. Sperm pools collected in both assays were further used for phenomic (n = 35), genomic (n = 9) and proteomic (n=4) analyses. **(B-E)** Change in phenotypic of sperm in raw (R T0), washed (W T0), swim up after 0 hours incubation (SW T0), after four hours incubation (SW T4) and 24 hours incubation (SW T24). **(B)** curvilinear velocity, **(C)** % sperm with mature chromatin, **(D)** gross morphology, and **(E)** % sperm without double-strand breaks. Data for all donors are shown as individual dots and overall mean ± standard error are shown in black for each time point. Donor number for the different time points were: Raw: n=35; W: n=35; T0: n=25; T4: n=35 and T24: n=12.

Figure 2: Genomic signature of within-ejaculate selection on sperm performance. (A) Loci with significantly diverging alleles between sperm pools collected from the Centre and the Edge in the Methyl Cellulose assay (likelihood ratio tests, > 99.9th quantile) for four donors and **(B)** top 20 significant Biological GO terms for the genes showing significant divergence. **(C)** Loci with significantly diverging alleles between sperm pools collected from Swim Up assays at T0 and T4 for five donors and **(D)** top 20 significant Biological GO terms for the genes showing significant divergence. **(E)** Loci with significantly diverging alleles between sperm pools collected from the Centre and the Edge in zebrafish, and **(F)** top 20 significant Biological GO terms for the human gene homologs translated from significantly diverging genes in zebrafish. **(G)** Distribution of genomic elements among significantly diverging loci for each human donor and time point and **(H)** for three zebrafish males from Alavioon et al 2017[6]. **(I)** Venn diagram showing overlap of significantly diverging genes between human Methyl Cellulose, human Swim Up and zebrafish and **(J)** top 20 significant biological GO terms for the 153 genes share between all three datasets.

Figure 3: Proteomic signature of within-ejaculate selection on sperm performance. (A) Volcano plots for each donor displaying significantly differentially expressed proteins (pink; (significance cut-off points are $p_{adj} < 0.05$; $\log_2$ fold change > |2|) between Centre/T0 and Edge/T4 sperm pools assays with $\log_2$ fold changes > 0 indicating increased expression in Centre/T0 sperm pools and $\log_2$ fold changes < 0 indicating increased expression in

Edge/T0 sperm pools. Two technical runs were performed for Donor A in the Swim Up assay. The top five significant proteins shared among all samples are highlighted by black circle and protein ID. **(B)** Venn diagram showing overlap of differentially expressed proteins between donors, experiments and technical runs. **(C)** Top five significantly differentially expressed proteins following stringent filtering steps shared among all donors. **(D)** Top 20 significant KEGG pathways for the differentially expressed proteins identified in both experiments. **(E)** Venn diagram showing overlap between differentially expressed genes from proteomic analyses (translated from protein IDs) and from genomic analyses for both experiments.

Figure 1

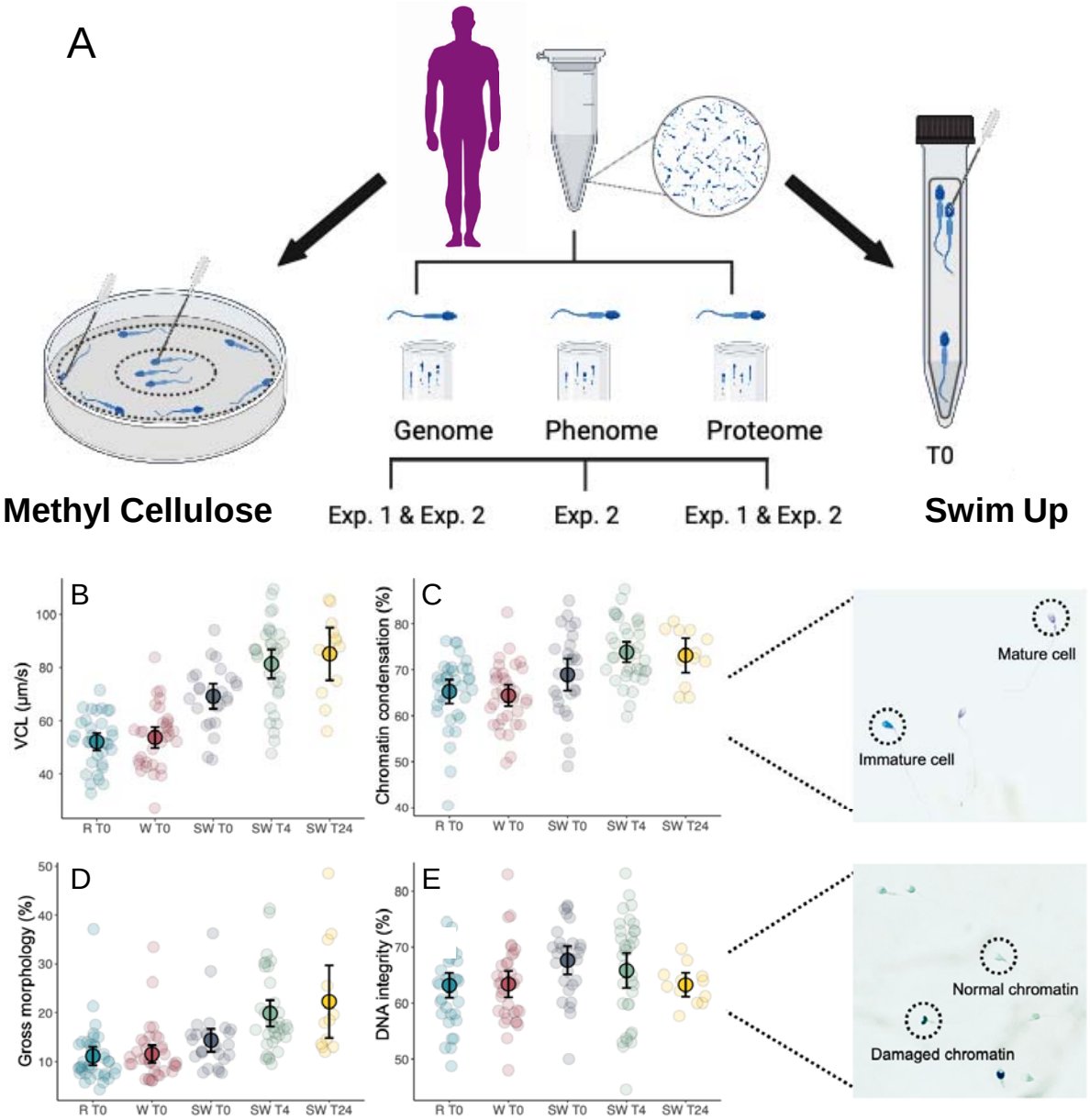

A
Genome
Phenome
Proteome
T0
Methyl Cellulose
Exp. 1 & Exp. 2
Exp. 2
Exp. 1 & Exp. 2
Swim Up
B
VCL (µm/s)
C
Chromatin condensation (%)
Mature cell
Immature cell
D
Gross morphology (%)
E
DNA integrity (%)
Normal chromatin
Damaged chromatin
R T0
W T0
SW T0
SW T4
SW T24

Figure 2

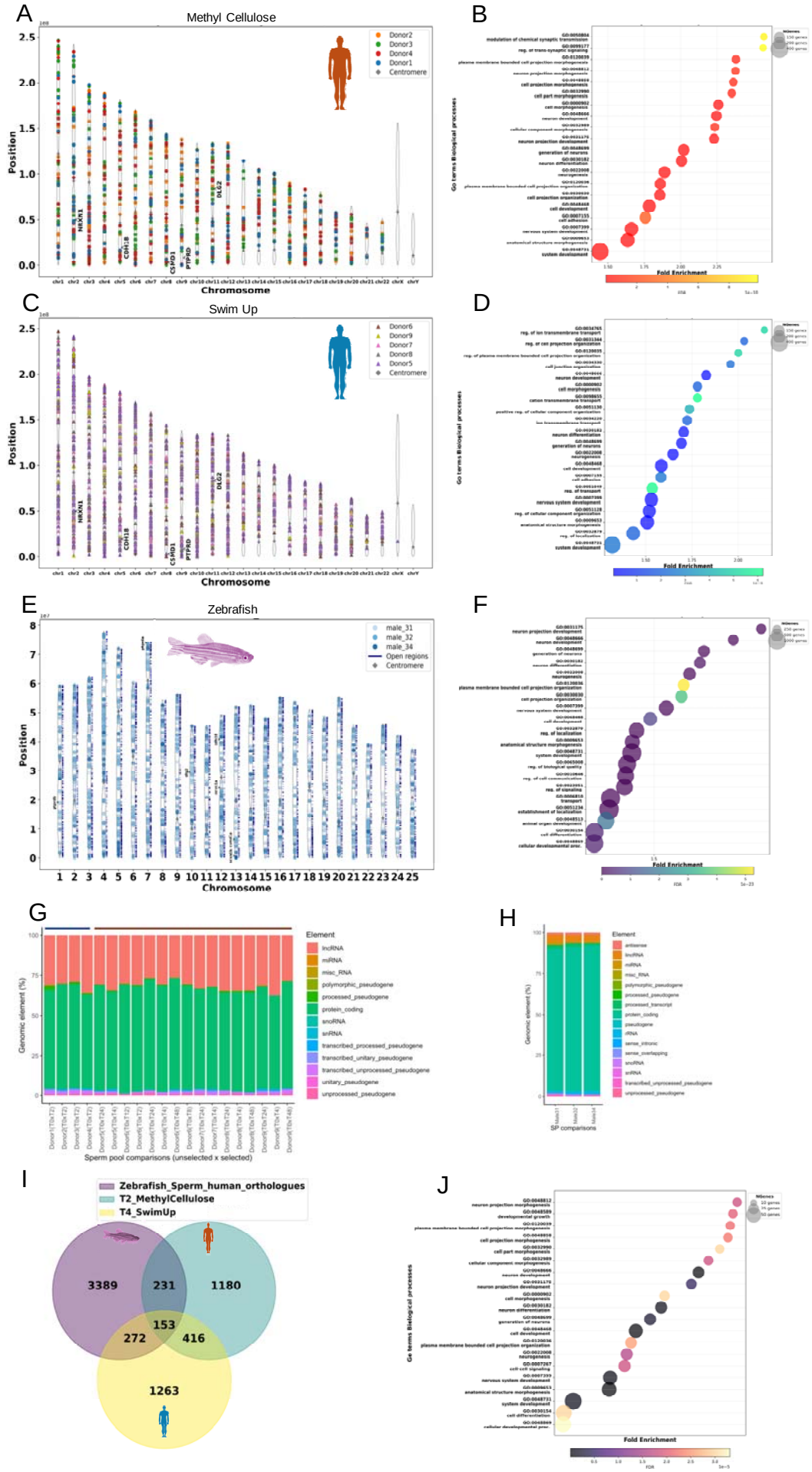
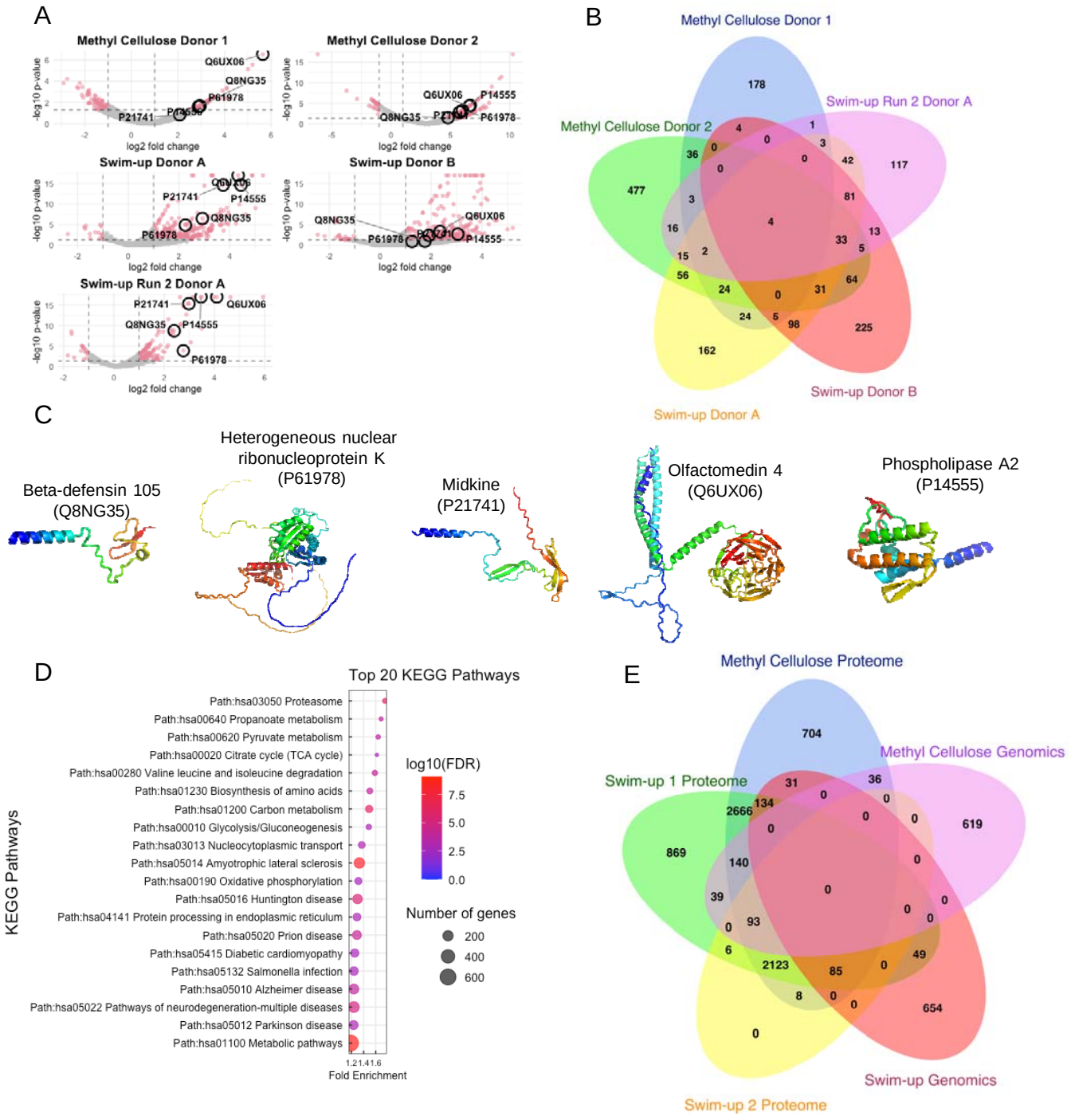

A
Methyl Cellulose
Position
Chromosome
Donor2
Donor3
Donor4
Donor1
Centromere
B
Go terms Biological processes
Fold Enrichment
C
Swim Up
Donor6
Donor9
Donor7
Donor8
Donor5
D
E
Zebrafish
male_31
male_32
male_34
Open regions
F
G
Genomic element (%)
Sperm pool comparisons (unselected x selected)
Element
lncRNA
miRNA
misc_RNA
polymorphic_pseudogene
processed_pseudogene
protein_coding
snoRNA
snRNA
transcribed_processed_pseudogene
transcribed_unitary_pseudogene
transcribed_unprocessed_pseudogene
unitary_pseudogene
unprocessed_pseudogene
H
SP comparisons
I
Zebrafish_Sperm_human_orthologues
T2_MethylCellulose
T4_SwimUp
3389
231
1180
153
272
416
1263
J

Figure 3


A
Methyl Cellulose Donor 1
Methyl Cellulose Donor 2
Swim-up Donor A
Swim-up Donor B
Swim-up Run 2 Donor A
-log10 p-value
log2 fold change
Q6UX06
Q8NG35
P61978
P21741
P14555
B
Methyl Cellulose Donor 1
Swim-up Run 2 Donor A
Methyl Cellulose Donor 2
Swim-up Donor B
Swim-up Donor A
C
Beta-defensin 105 (Q8NG35)
Heterogeneous nuclear ribonucleoprotein K (P61978)
Midkine (P21741)
Olfactomedin 4 (Q6UX06)
Phospholipase A2 (P14555)
D
Top 20 KEGG Pathways
KEGG Pathways
Path:hsa03050 Proteasome
Path:hsa00640 Propanoate metabolism
Path:hsa00620 Pyruvate metabolism
Path:hsa00020 Citrate cycle (TCA cycle)
Path:hsa00280 Valine leucine and isoleucine degradation
Path:hsa01230 Biosynthesis of amino acids
Path:hsa01200 Carbon metabolism
Path:hsa00010 Glycolysis/Gluconeogenesis
Path:hsa03013 Nucleocytoplasmic transport
Path:hsa05014 Amyotrophic lateral sclerosis
Path:hsa00190 Oxidative phosphorylation
Path:hsa05016 Huntington disease
Path:hsa04141 Protein processing in endoplasmic reticulum
Path:hsa05020 Prion disease
Path:hsa05415 Diabetic cardiomyopathy
Path:hsa05132 Salmonella infection
Path:hsa05010 Alzheimer disease
Path:hsa05022 Pathways of neurodegeneration-multiple diseases
Path:hsa05012 Parkinson disease
Path:hsa01100 Metabolic pathways
Fold Enrichment
log10(FDR)
Number of genes
E
Methyl Cellulose Proteome
Methyl Cellulose Genomics
Swim-up 1 Proteome
Swim-up Genomics
Swim-up 2 Proteome