Molecular signatures of sperm-level selection in humans and their evolutionary conservation

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

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

[2] Department of Ecology and Evolution, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

[3] John Innes Centre, Norwich Research Park, Norwich, UK

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

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

[#] These authors share first authorship.

*** Corresponding authors:** S. Immler (s.immler@uea.ac.uk)

Keywords: within-ejaculate selection, sperm selection, fertility, reproduction, sperm proteomics, sperm phenotype, polygenic scores

SUMMARY PARAGRAPH

Of the millions of sperm in a human ejaculate, only one fertilises an egg. Selection on sperm is strong, offering an opportunity to proofread novel germline genotypes generated by segregation, recombination, and de novo mutations. Here, we combine within-ejaculate selection on phenotypic sperm quality with deep-coverage whole-genome sequencing and tandem mass tag proteomics to identify molecular biomarkers of human sperm quality. High-quality sperm are characterised by increased longevity, motility, morphological normality, and DNA condensation. We show that high-quality sperm are genetically distinct from low-quality sperm at >1,000 loci mapping to 511 genetic elements, 64% of which fall within protein-coding portions including intronic regions. Genes diverging between high- and low-quality sperm pools are involved in cell signalling, cell proliferation, and neurodevelopmental processes. Polygenic scores differ between high- and low-quality sperm and reveal variants linked to overall health and healthy lifestyle choices, linked by putative negative correlation. TMT proteomics show significant differences in the abundance of 1,178 proteins, with known markers of inflammation, immune function, and cancer-related cell proliferation showing lower abundance in high-quality sperm. The genomic signature of sperm selection in humans and zebrafish shows significant overlap, indicating evolutionarily conserved processes determining sperm function across taxa. Our finding of a direct link between phenotypic variation and genomic variation in human sperm highlights the importance of within-ejaculate selection and its key contribution to determining offspring fitness and health.

MAIN TEXT

A human ejaculate contains between 20 and 150 million sperm, of which one fertilises an egg. Sperm within the ejaculate of a man vary phenotypically and genetically, offering a strong opportunity for natural selection to act upon[1–3]. Although the importance of selection among germ cells is increasingly recognised [4], the role and magnitude of selection among mature sperm within an ejaculate remain largely overlooked [1,3]. Recent findings in zebrafish (*Danio rerio*) show that longer-lived sperm differ from shorter-lived sperm at many loci distributed across the genome [5]. Large-scale studies of gene expression during spermatogenesis provide clear evidence for expression in post-meiotic haploid spermatids [6], and incomplete sharing of transcripts between meiotic products [7] supports the idea that genetic differences among sperm engender phenotypic differences within an ejaculate. This notion is further supported by accelerated evolution of genes expressed at post-meiotic stages [6], as well as elevated purifying selection on these genes [8].

However, the genes and associated molecular markers under haploid selection in sperm are not known. Identifying molecular biomarkers of high-quality sperm is central to understanding the fundamental biology of human reproduction and fertility and will also inform improved sperm selection for use in Assisted Reproductive Technology [9,10].

Here, we designed two in vitro assays (swim up and methyl cellulose; see Methods) to select for high-quality sperm within the ejaculate of human donors and combined this selection with deep-coverage whole-genome sequencing (WGS) and tandem mass tag (TMT) proteomics to resolve the molecular architecture of human sperm quality. Complementary analyses using polygenic scores from the PAN-UK Biobank further enabled assessment of putative implications of selection on variants underlying sperm quality. By comparing the genomic signature of haploid selection in human sperm with that in zebrafish, we confirmed the biological relevance and generality of our findings across taxa.

Selection on phenotypic sperm quality

We selected distinct sperm-quality pools from semen samples of 35 donors using the swim-up assay and compared fertilisation-relevant traits (sperm kinematics, morphology, chromatin, DNA integrity) across raw and washed samples and selected sperm incubated for 0, 4 and 24 hours (Suppl. Table S1). The fraction of motile sperm decreased from an average of 60% in fresh and washed samples to 50% after 4 hours of incubation and 25% after 24 hours of incubation prior to swim-up (Suppl. Fig. S1 and Suppl. Table S2). Sperm kinematics improved continuously with increasing incubation time (Fig. 1A; Suppl. Tables S3–S10; Suppl. Fig. S2),

as did chromatin condensation (Fig. 1B; Suppl. Table S11) and gross morphology (Fig. 1C; Suppl. Table S12), whereas DNA integrity was slightly higher in swim-up samples compared to all other samples (Fig. 1D; Suppl. Table S13).

Genomic divergence among selected sperm pools

We linked phenotypic differences between selected high- and low-quality sperm with genomic differences by performing high-coverage WGS on selected sperm pools using both assays. For the methyl cellulose assays, we used ejaculates from four donors and collected sperm from the Centre (low-quality) and the Edge (high-quality) of the incubation dish; for the swim-up assays, we used ejaculates from five donors and collected sperm at 0 hours (T0; low-quality) and at 4 and 24 hours (T4 and T24; high-quality). Each donor represented an experimental replicate and samples were analysed separately. For DNA extraction and sequencing, we standardised the number of cells to 2 million per sample and performed WGS at ~100× per sample.

To assess genetic variation across donors, we conducted a comprehensive analysis of allele frequencies and identified significant loci for interpretation. We compared allele frequencies at an average of 2,280,804 loci per donor (min = 2,022,290; max = 2,754,360; Suppl. Table S14) and retained >99.9th-quantile significant loci for interpretation. We identified heterozygous sites for each donor, filtered for a minimum total read number of 32 per locus, and performed likelihood-ratio tests on allele frequencies to test for significant deviations from expected Mendelian frequencies using the Centre/T0 samples as reference. We identified significant variants at an average of 1,139 loci per donor (min = 783; max = 2,578; Fig. 2A,B; Suppl. Table S15 & S16; Suppl. Fig. S10; coverage was not correlated with significance; Suppl. Table S17; Suppl. Fig. S4)).

We mapped the significantly diverging loci to an average of 511 genetic elements (min = 340; max = 697) per comparison, of which 64% were intron and exon regions of protein-coding genes (min = 58.4%; max = 69.4%) and 31.6% were lncRNAs (min = 26.4%; max = 37.5%; Fig. 2E,F), with no significant differences in element representation across donors and time points ($\chi^2 = 103.42$, df = 228, p = 1.0). Significantly diverging protein-coding sequences were longer than the average protein-coding gene in the reference genome (Centre: $t = 8.62$, $p < 0.001$; T4: $t = 5.66$, $p < 0.001$; T24: $t = 7.42$, $p < 0.001$; T48: $t = 4.42$, $p < 0.001$; Suppl. Fig. S5). Gene functions and Gene Ontology (GO) terms overlapped across time points (Suppl. Fig. S6–S8); we focus here on methyl cellulose and swim up at T4. Overall, significantly diverging

genes were enriched for processes of cell assembly and neuronal signalling. In methyl cellulose, enrichment included transmembrane receptor protein tyrosine kinase signalling, cell junction assembly and neuron development (Fig. 2C), and molecular functions associated with cell communication and membrane transport such as protein serine kinase activity, G-protein-coupled neurotransmitter receptor activity, calcium ion binding and glutamate receptor activity (Suppl. Fig. S7). In swim up at T4, most enrichment was associated with nervous system development, including synapse assembly, neuron differentiation and chemical synaptic transmission (Fig. 2D), and molecular processes involving cellular transport and cell adhesion such as calcium ion binding, glutamate receptor activity and calcium-activated potassium channel activity (Suppl. Fig. S8).

Across assays and time points, we identified 19 genes that diverged in at least six of nine donors (Suppl. Table S18); the top five were *DLG2, PTPRD, CSMD1, CDH18* and *NRXN1*. *DLG2, PTPRD, CSMD1* and *CDH18* are well-described tumour suppressor genes [11–13], and *NRXN1* has a role in lung cancer development. Certain alleles within these five genes have been associated with altered risk for schizophrenia, autism spectrum disorder (ASD), and neurodevelopmental disorders, as well as involvement in inflammasome formation and cellular senescence [14–18]. *CSMD1* is associated with male and female infertility [19], *DLG2* is downregulated in human sperm with compromised performance [20], *CDH18* is implicated in cell adhesion in testes [21] and deletions in *NRXN1* have been observed in azoospermic patients [22]. Analysing assays separately, we detected *RBFOX1, EGLN3, CDH13, LRP1B* and *LINC00960* across all four donors in methyl cellulose (Suppl. Fig. S9A) and *CDH18* and *ZFPM2* across all donors in swim up at T4 (Suppl. Fig. S9B). *RBFOX1* and *CDH13* contribute to neuronal development [23,24], whereas *CDH18, EGLN3, LRP1B* and *LINC00960* participate in cancer cell migration and tumour growth or suppression [25–27].

Polygenic scores of variants diverging between sperm pools

We connected the genomic signature of selection for sperm quality with traits by generating polygenic scores (see Methods) using the Pan-UK Biobank database [43]. Using raw PGS that included all heterozygous loci for each sperm pool (T0/T4 and Centre/Edge), we calculated the score difference between the two pools for each donor and normalised the difference (Fig. 1A,B). Based on normalised *Z*-scores in the 99th percentile, we identified the highest-scoring traits as total, direct and indirect bilirubin, supervision of normal pregnancy, lipoprotein(a) and urate and some lifestyle and diet traits were also among the top traits (Fig. 2G,H). The emerging terms were remarkably consistent across donors (Suppl. Fig. S10), and changes in polygenetic

151 scores for bilirubin-related traits showed a negative correlation to changes in scores for lifestyle
152 traits, indicating genetic correlation between traits. The direction of effects was not always
153 consistent between donors and experiments. This supports the idea that selection on sperm
154 performance is likely polygenic, rather than driven by a small number of variants. In addition,
155 the variability across genomic backgrounds may indicate potential epistatic interactions,
156 although polygenicity alone does not always imply epistasis. We tested for genetic correlations
157 between trait-associated polygenic scores, which yielded similar patterns across both assays.
158 (Suppl. Fig. S4A; Suppl. Fig. S11). None of the traits that showed the strongest changes in their
159 polygenic scores (Fig. 2G,H) displayed significant genetic correlations. Overall, these results
160 support a highly polygenic response to sperm selection, and that many organism-level traits
161 have genetic underpinnings that are visible to selection at the sperm stage.
162

163 **Evolutionarily conserved genomic signature of haploid selection between humans and**
164 **zebrafish**

165 We re-analysed zebrafish sperm-pool data from Alavioon *et al.* (2017) [5] using the same rigorous
166 approach as for the human data and compared genes diverging between sperm pools in humans
167 and zebrafish using human gene annotations [5] (Suppl. Tables S19–S21; Suppl. Fig. S12–S13).
168 We identified on average 5,233 significantly diverging loci per male (range: 4,545–5,722; Fig.
169 3A), with distributions of genomic elements like those observed in human samples (Fig. 3C).
170 Enrichment analyses in zebrafish revealed biological processes including cell-to-cell
171 interactions and neurodevelopmental processes (differentiation and generation of neurons, cell
172 signalling, developmental cell differentiation and organ formation; Fig. 3B), and molecular
173 functions related to ion transport and channel activity (calcium-activated cation channel
174 activity, sodium channel activity and ion transmembrane transporter activity; Suppl. Fig. S15).
175 We found 153 genes shared among human methyl cellulose, swim up and the human homologs
176 of zebrafish genes (Fig. 3D). The top 20 significant biological GO terms for these 153 genes
177 indicated functions centred on neural development and cell proliferation (Fig. 3E). Semantic
178 similarity analysis between GO terms identified in humans and the human homologs of
179 zebrafish genes revealed high similarity (scores of 0.85 for methyl cellulose and 0.7 for swim
180 up at T4; Suppl. Fig. S14). Finally, three of the five top candidate genes in humans were also
181 significantly diverging among zebrafish sperm pools—*DLG2/dlg2, CSMD1/csmd1a* and
182 *CDH18/cdh18*. Altogether, despite ~450 million years of evolutionary divergence between
183 zebrafish and humans [28], the biological processes and gene functions under haploid selection
184 in sperm appear highly conserved across taxa.

185

186 **Proteomic biomarkers of selected sperm**

187 To identify molecular biomarkers of sperm quality, we performed methyl cellulose and swim
188 up assays in an additional four human donor samples (two donors per assay) and applied TMT
189 proteomics to the selected sperm samples (see *Methods*), including a technical replicate for one
190 donor to confirm robustness. Of 8,191 peptide matches, 4,616 proteins were identified and
191 retained after rigorous filtering, and 2,232 overlapped across all donors (Fig. 4A). In total,
192 1,178 proteins showed significant differences in abundance between sperm pools in at least one
193 donor, with 634 proteins showing significantly ($\log_2$-fold change > 2; $p_{adj} < 0.05$) lower
194 abundance and 544 showing significantly higher abundance in high-quality sperm (Fig. 4B).
195 Analysing assays separately, we identified 52 and 425 proteins with higher and lower
196 abundance, respectively, in methyl cellulose and 33 and 594 proteins with higher and lower
197 abundance, respectively, in swim up (range: 43-113 higher and 86-431 lower per donor; Suppl.
198 Table S22). In swim up, proteins less abundant in high-quality sperm were strongly associated
199 with RNA processing and protein translation (Suppl. Fig. S15A), whereas proteins more
200 abundant in high-quality sperm were enriched for motility and membrane organisation. In
201 methyl cellulose, proteins were associated with lipid transport and cholesterol efflux pathways
202 characteristic of capacitation (Suppl. Fig. S15B). KEGG analysis showed that proteins less
203 abundant in selected sperm were enriched for RNA- and ribosome-binding activities, including
204 "ribonucleoprotein complex binding" and "structural constituent of ribosome" (Suppl. Fig.
205 S15A). Adhesion-related terms such as "collagen binding" and "cell–substrate junction" were
206 also observed. Proteins more abundant in selected sperm were enriched for functions related to
207 enzyme inhibition and structural support, including endopeptidase inhibitor activity, peptidase
208 regulator activity and antioxidant activity (Suppl. Fig. S16B). Functional enrichment analyses
209 revealed two broad, consistent patterns. First, proteins with low abundance in high-quality
210 sperm are linked to translational and splicing activity—features of immature sperm absent from
211 mature, high-quality sperm [29]. Their depletion in selected fractions supports the interpretation
212 that sperm surviving the selection assays represent a more functionally mature population.
213 Second, proteins with increased abundance in high-quality sperm were related to motility,
214 consistent with evidence that flagellar structure and activity are central to sperm motility and
215 fertilisation competence [30–32]. Other enriched functions included lipid metabolism and
216 membrane-associated protein complexes, processes tightly linked to capacitation and
217 regulation of membrane fluidity [33]. Enrichment for antioxidant and enzyme inhibitor activities

218 is also consistent with requirements to withstand oxidative and proteolytic stress during transit
219 through the female tract [34,35]. Applying an additional filtering step based on significance
220 ($p_{adj} < 0.05$) and the median ratio of protein abundance among sperm pools across all four
221 donors, we identified six key proteins that were downregulated and located on the membrane
222 in high-quality sperm: OLFM4, CTNG, PLA2, DEFB105A, midkine and protein K, all
223 associated with human disease phenotypes including various cancers, inflammation and cell
224 growth [36–40]. We validated OLFM4 and CTNG as possible biomarkers based on their abundance
225 and putative cell-surface localisation using fluorophore-conjugated antibodies and FACS in
226 five additional donors (see Methods). Both proteins showed consistently higher abundance in
227 low-quality sperm across all five donors (Suppl. Table S23; Suppl. Fig. S17, S18), supporting
228 the TMT results. Genes underlying the significantly diverging proteins showed some overlap
229 with genomic signatures in the WGS data (Suppl. Fig. 4E).
230
231 **Discussion**
232 Within-ejaculate selection for sperm fitness, based on a combination of longevity and
233 penetration ability, resulted in significant improvement of overall phenotypic sperm quality.
234 These phenotypic differences have a genomic basis, as indicated by genome-wide signals of
235 selection involving genes with fundamental roles in organismal health and fitness across
236 generations. The genomic markers of haploid selection are remarkably conserved and show
237 strong parallels between species as distant as humans and zebrafish. Beyond coding regions
238 exhibiting divergence in both species, non-coding regions and lncRNAs in human sperm and
239 lincRNAs in zebrafish sperm were particularly frequent. lnc/linc RNAs exert widespread
240 regulatory functions on downstream gene expression, suggesting substantial regulatory activity
241 in genes determining sperm function. Differences in lncRNA expression have also been
242 reported in sperm of men with compromised fertility [42] where they regulate genes involved in
243 cytoplasmic transport, biosynthetic processes and spermatogenesis. Notably, we found little
244 overlap with geno-informative markers previously identified in the mammals: mouse,
245 macacque and cattle [7].
246 Our genomic and proteomic data indicate that this gametic stage may play a key role in
247 favouring high-functioning genomes, with potential to reduce transmission of heritable disease
248 and to improve overall health and fitness. While the genomic data highlight clear divergence
249 between sperm pools, reduced abundance of proteins related to neurodevelopmental processes
250 and cancer in high-quality sperm suggests a direct link between phenotypic sperm quality and
251 genetic quality. A link between sperm fitness and offspring fitness has been demonstrated in

zebrafish [5,41], and we provide the genomic architecture underlying these differences. Neurodevelopmental disorders point to early developmental differences, which may also explain the observed 5–7% difference in survival during early zebrafish embryo development [5]. In contrast, cancer is an ageing-related heritable disease, and the reduction of proteomic cancer biomarkers in high-quality sperm aligns with our previous observation of higher reproductive fitness and increased lifespan in zebrafish sired by longer-lived sperm [41]. The signals among variants contributing to Polygenic Scores further support a strong link between overall genetic sperm quality and human health.

Overall, we detected stable core patterns in the molecular biomarkers of human sperm quality across donors, assays and species. Nevertheless, we observed notable differences in the identity of diverging genes and proteins, consistent with sperm fitness - like organismal fitness - being shaped by a large network of genes. Allelic divergence in these genes varies among individuals, and these regulatory networks appear to arise from genetic correlations, but this idea needs further testing. Our study provides molecular signatures of gametic selection and its consequences for organismal fitness and health. These findings open opportunities to use sperm-quality biomarkers as indicators of overall genomic health in applied contexts such as Assisted Reproductive Technologies, particularly IVF and ICSI [10], and to improve outcomes in human reproduction, 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. Haploid selection in “diploid” organisms. *Annu. Rev. Ecol. Evol. Syst.* **50**, 219–236 (2019).
3. Immler, S. Haploid selection in “diploid” organisms. *Annu. Rev. Ecol. Evol. Syst.* **50**, 219–236 (2019).
4. Neville, M. D. C. *et al.* Sperm sequencing reveals extensive positive selection in the male germline. *Nature* **630**, 123–132 (2025). doi:10.1038/s41586-025-09448-3.
5. Alavioon, G. *et al.* Haploid selection within a single ejaculate increases offspring fitness. *Proc. Natl Acad. Sci. USA* **114**, 8053–8058 (2017).
6. Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *Nature* **613**, 308–316 (2023).
7. Bhutani, K. *et al.* Widespread haploid-biased gene expression enables sperm-level natural selection. *Science* **371**, eabb1723 (2021).

8. Otto, S. P., Scott, M. F. & Immler, S. Evolution of haploid selection in predominantly diploid organisms. *Proc. Natl Acad. Sci. USA* **112**, 15952–15957 (2015).
9. Alavioon, G., Marcu, D. & Immler, S. Within-ejaculate sperm selection and its implications for assisted reproduction technologies. In *XIIIth International Symposium on Spermatology* 127–133 (Springer, 2021).
10. van Oosterhout, C., Marcu, D. & Immler, S. Accounting for the genetic load in assisted reproductive technology. *Clin. Transl. Med.* **12**, e864 (2022).
11. Griesius, S. *et al.* Reduced expression of the psychiatric risk gene *DLG2* (PSD93) impairs hippocampal synaptic integration and plasticity. *Neuropsychopharmacology* **47**, 1367–1378 (2022).
12. Yu, X. *et al.* Protein tyrosine phosphatase receptor-type δ acts as a negative regulator suppressing breast cancer. *Oncotarget* **8**, 98798–98811 (2017).
13. Veeriah, S. *et al.* The tyrosine phosphatase *PTPRD* is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proc. Natl Acad. Sci. USA* **106**, 9435–9440 (2009).
14. 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).
15. 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).
16. 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).
17. Ishizuka, K. *et al.* Functional characterization of rare *NRXN1* variants identified in autism spectrum disorders and schizophrenia. *J. Neurodev. Disord.* **12**, 25 (2020).
18. Rujescu, D. *et al.* Disruption of the *neurexin 1* gene is associated with schizophrenia. *Hum. Mol. Genet.* **18**, 988–996 (2009).
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. Li, X. *et al.* Cadherin-18 loss in prospermatogonia and spermatogonial stem cells enhances cell adhesion through a compensatory mechanism. *Zool. Res.* **45**, e003 (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. Nixon, B. *et al.* New horizons in human sperm selection for assisted reproduction. *Front. Endocrinol.* **14**, 1198265 (2023).
30. Lindemann, C. B. & Lesich, K. A. Functional anatomy of the mammalian sperm flagellum. *Cytoskeleton* **73**, 652–669 (2016).
31. Chen, Z. A. *et al.* Native structure of the mammalian sperm axoneme revealed by cryo–electron tomography. *Nat. Struct. Mol. Biol.* **30**, 222–234 (2023).
32. Leung, M. R. *et al.* Architecture of the mammalian sperm flagellum. *Nature* **636**, 455–463 (2025).
33. Gadella, B. M. & Luna, C. Cell biology and functional dynamics of the mammalian sperm surface. *Theriogenology* **81**, 74–84 (2014).
34. Bisht, S., Faiq, M., Tolahunase, M. & Dada, R. Oxidative stress and male infertility. *Nat. Rev. Urol.* **14**, 470–485 (2017).
35. Cesari, A., Monclus, M. de los A., Tejón, G. P., Clementi, M. & Fornes, M. W. Regulated serine proteinase lytic system on mammalian sperm surface: there must be a role. *Theriogenology* **74**, 699–711.e5 (2010).

36. Vecchi, L. *et al.* Phospholipase A2 drives tumorigenesis and cancer aggressiveness through its interaction with annexin A1. *Cells* **10**, 1472 (2021).
37. Uhlitz, F. *et al.* Mitogen-activated protein kinase activity drives cell trajectories in colorectal cancer. *EMBO Mol. Med.* **13**, e14123 (2021).
38. Ma, H. W. *et al.* Olfactomedin 4 produces dysplasia but suppresses metastasis of colon cancer. *Cancer Gene Ther.* **30**, 694–703 (2023).
39. Neumaier, E. E., Rothhammer, V. & Linnerbauer, M. The role of midkine in health and disease. *Front. Immunol.* **14**, 1157024 (2023).
40. Sudhakaran, M. & Doseff, A. I. Role of heterogeneous nuclear ribonucleoproteins in the cancer-immune landscape. *Int. J. Mol. Sci.* **24**, 5086 (2023).
41. 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).
42. Zhang, X. *et al.* Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *Int. J. Mol. Sci.* **20**, 5573 (2019).
43. Karczewski, K.J., Gupta, R., Kanai, M. *et al.* Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. *Nat. Genet.* **57**, 2408–2417 (2025).

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.

For TMT 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)). For the biomarker validation of OFLM4 and CTNG, we recruited an additional five donors from the University of East Anglia (UEA), Norwich, UK.

384 The samples from University of East Anglia and Cape Town were collected under ethics permit
385 no. ETH2324-0997 and the samples collected at the Rigshospitalet Copenhagen were collected
386 under ethics permit no. H-17012149.
387
388 **Sperm selection assays**
389 All samples were collected near the laboratory and allowed to liquefy for 30 minutes at 37°C
390 prior to the selection assay. We used two different assays to select for sperm within ejaculates:
391 *Methyl Cellulose Assay*: A 10 cm-diameter watch glass was filled with 1% (w/v)
392 methylcellulose (Sigma, Denmark) up to a 4cm radius and 10ml total volume. Then, 300µl of
393 raw ejaculate was placed in the centre of the glass (the ejaculate was allowed to sink and filled
394 a circle of 0.5cm in the centre of the glass). For each donor, we used five 300 µl aliquots
395 allocated into five watch glasses per donor in parallel to ensure sufficient cells in the edge
396 population for subsequent whole genome sequencing. The samples were incubated in the watch
397 glasses for two hours at 37°C under 5% CO2 supplementation. 4ml aliquots were then collected
398 from the outer circle of each watch glass at a radius of 3-3.5 cm, 2.5-3cm or 2-2.5cm
399 (depending on the distance travelled by the sperm starting at the radius from the further circle
400 with >10 cells under the microscope for each male) and placed in a 15ml tube. All five edge
401 samples collected from the outer circle were pooled for each donor to ensure high enough cell
402 numbers for sequencing. We also collected a sample (5 ml) from the centre of each watch glass
403 at least 0.5 cm above the bottom of the glass to avoid contamination with dead sperm and
404 somatic cells. To extract the sperm from the methylcellulose, 1µl of cellulase was added for
405 every 1ml of collected sample and the sample was mixed thoroughly on a shaking table for two
406 hours at 50°C. The samples were then centrifuged, the supernatant removed, and the sperm
407 pellets were stored at -80° C for subsequent DNA extraction.
408
409 *Swim Up Assays*: For the phenotypic analyses, we collected a raw (R T0), washed (W T0), and
410 sub-samples from swim ups following incubation (T0, T4, T8, T24, T48) and for the genomic
411 analyses we collected a raw sample (T0) and sub-samples from swim ups following incubation
412 (T4, T24 for all five donors and T8 (n = 1 donor), T24 and T48 (n = 3 donors).
413 For the swim ups sub-samples, aliquots of 1ml raw ejaculate were placed in 15ml centrifuge
414 tubes at a 1:1 ratio of human tubal fluid (HTF) medium containing (in mM): 72.8 NaCl, 4.69
415 KCl, 0.2 $MgSO_4$, 0.37 KH_2PO_4, 2.04 $CaCl_2$, 0.33 sodium pyruvate, 21.4 sodium lactate, 2.78
416 glucose, 21 HEPES, 25 $NaHCO_3$, adjusted to pH 7.4 with NaOH (Fisher, UK) and incubated
417 for 2, 4, 8, 12, 24 and 48 hours at 37°C with 5% CO2 in the air. Motile sperm, at each time

418 point, were recovered by direct swim up as follows: after flicking each tube to homogenise the
419 sample, 1ml was aliquoted and placed under 4ml of HTF medium in a new 15ml centrifuge
420 tube at 37°C for 60 min in 5% CO2 as described above. The top layer of the swim up (3.5ml)
421 was collected and centrifuged at 700xg for 10-20 min at room temperature (RT) prior to
422 resuspension in 2ml HTF. After assessing sperm concentrations of samples using image
423 cytometry, the cells were pelleted, the supernatant removed, and the pellet stored at -80°C until
424 further use.
425
426 **Phenotypic analyses**
427 *Sample prep and selection assay*
428 We used samples from 35 donors for sperm selection by Swim Up assay and subsequent
429 phenotypic analyses. All raw samples were split into three aliquots: one third of the raw sample
430 was used directly for a swim up and the remaining 2/3 were washed and split into two aliquots
431 and incubated for 4h and 24h, respectively. In the washing step, the samples were washed twice
432 in HTF lacking HAS. The sample was centrifuged at 500g for 10 minutes and after removal of
433 the supernatant, 3ml of HFT+ was added and the sample was split again into two labelled tubes
434 (T4 and T24). Following incubation, the samples were centrifuged at 500g for 10 minutes to
435 supernatant and perform a direct swim up to isolate motile cells as described above.
436 The swim ups were performed in an Eppendorf tube as follows: 1 ml of human tubal fluid
437 (HTF+) medium containing (in mM): 72.8 NaCl, 4.69 KCl, 0.2 $MgSO_4$, 0.37 KH_2PO_4, 2.04
438 $CaCl_2$, 0.33 sodium pyruvate, 21.4 sodium lactate, 2.78 glucose, 21 HEPES, 25 $NaHCO_3$ and
439 3.5% human serum albumin (HSA) (Fisher, UK) adjusted to pH 7.4 with HCL was added to
440 the semen/sperm pellet. During the swim up assay, cells were allowed to swim up for 1h at
441 37°C with 5% CO_2 supplementation after which we collected 700µl of the sample from the top
442 of the Eppendorf tube.
443 We then assessed phenotypic traits including sperm concentration, motility and velocity, sperm
444 viability, sperm morphology, chromatin and DNA integrity (see *Supplementary Methods* for
445 details on each of these).
446
447 **Genomic analyses of sperm human sperm pools**
448 *Whole genome sequencing*
449 DNA was extracted using a standard phenol:chloroform:isoamyl alcohol protocol (see
450 Supplementary Methods for details on DNA extraction) from the sperm pools and prepared
451 using a PCR-free library preparation kit (KAPA HyperPrep kit; Roche) for a standard insert

size (400-500 bp), (see Supplementary Methods for details). Sequencing was conducted by the Earlham Institute, UK, on an Illumina NovaSeq6000 to generate 150 bp paired-end reads.

Bioinformatic analyses of genome data

All scripts and step-by-step instructions are available on: https://github.com/alicegodden/spermpool.

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. Alignment of the reads from T0 raw ejaculate and selected sperm pools were used to generate base count profiles using pileup2pro2 (https://github.com/douglasgscofield/gameteUtils). SAMtools v1.9 option -q30, -Q30 and d100000 was used to exclude sites with low quality, allowing for sites with deep coverage.

For the likelihood ration tests, we only retained sites with a total coverage of > 31x or more than 2*total mean coverage were excluded (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). The heterozygous sites and base count profiles of the T0 raw and high performance sperm pools were used for likelihood ratio tests using the script hetPoolLikelihoods.pl available at https://github.com/douglasgscofield/gameteUtils following 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. 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, identified loci were mapped onto an annotated *Homo sapiens* reference genome GRCh38. 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].

Polygenic Scores

VCF files were generated from each sperm pool BAM file with a bcftools v1.15.1 [3] pipeline as follows:

bcftools mpileup -a AD,ADF,ADR -r "${ncbi_chr}" -q 30 -Q 30 -f GRCh38_latest_genomic.fna "${bam_file}" | bcftools call -mv -f GQ -Oz -o "$output_file" [1].

Chromosome files were concatenated, and chromosome identifiers converted to Ensembl format. To allow integration with the Pan-UK Biobank dataset, the VCF files were lifted over to the GRCh37 reference genome coordinate system using GATK v4.6.0.0, with Picard LiftoverVCF. The GRCh37 reference genome was obtained from Ensembl, Homo_sapiens.GRCh37.dna_sm.primary_assembly.fa.gz , version 27/11/2015. The chain file GRCh37_to_GRCh38.chain.gz, version 25/07/2014, https://ftp.ensembl.org/pub/assembly_mapping/homo_sapiens/ was downloaded from Ensembl.

Polygenic Score (PGS) data were obtained from the Pan-UK Biobank phenotype manifest: https://docs.google.com/spreadsheets/d/1AeeADtT0U1AukliiNyiVzVRdLYPkTbruQSk38DeutU8/edit?gid=1450719288#gid=1450719288 in May 2025. The data were filtered to retain CHR, POS, REF, ALT and beta_EUR columns, with the latter representing the effect size of the alternative allele as estimated in UK Biobank participants of inferred European ancestry. We extracted information on allele frequency, chrom, pos, ref and alt columns from the sperm pool VCF files and retained all positions that overlapped with variants in the Pan-UK Biobank dataset. The sperm pool allele frequencies and beta_EUR of all overlapping loci for a phenotype were multiplied and the product for each phenotype was added up, to obtain a score for each donor.

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) [4]. 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

520 potential sequencing bias in the zebrafish data, the cut-off point for genome-wide identification
521 of significant loci in the zebrafish data was set to the 99.9th quantile; critical values were 13.99,
522 26.51, 10.39 for male ID: Male_31, Male_32 and Male_34, respectively. We then used Venn
523 diagrams on significant genes by converting zebrafish genes into human gene homologs to
524 identify genes shared between humans and zebrafish and performed GO analyses using
525 *ShinyGO v0.75* [2] on the 153 shared genes.
526
527 **TMT proteomics**
528 *Sperm selection assays for proteomics*
529 Samples from four donors in total, two for each assay were used. The samples were washed in
530 human tubal fluid (HTF) and resuspended in 200 µL HTF. We collected two samples
531 representing a less fit (Centre/T0) and a fitter (Edge/T4) sperm pool from each donor from
532 Methyl Cellulose and Swim Up (assays as described above; see Supplementary Methods for
533 exact details) and stored them at -80°C.
534
535 *Protein extraction and proteomics*
536 For protein extraction, samples were thawed, and 450 ml of acetone was added to each of the
537 thawed samples. Samples were then vortexed for 15 minutes, then centrifuged at 10,000 rpm
538 for 10 minutes. Supernatant was then removed, and 500 ml of acetone added to each sample.
539 Each sample was then spun gently for two minutes manually then centrifuged for 15 mins at
540 10000 rpm. Supernatant was removed and samples were air dried for 1 and a half hours.
541
542 Protein pellets were resuspended in 100 µl of 2.5% sodium deoxycholate (SDC; Merck) in 0.2
543 M EPPS-buffer (Merck), pH 8.5, and the samples vortexed under heating. The samples were
544 treated with dithiothreitol and iodoacetamide to alkylate cysteine residues and digested with
545 trypsin in the SDC buffer according to standard procedures. After the digest, the SDC was
546 precipitated by adjusting to 0.2% TFA, and the clear supernatant subjected to C18 SPE
547 (Reprosil-Pur 120 C18-AQ, 5 um, Dr. Maisch GmbH, Germany). Peptide concentration was
548 estimated by running an aliquot of the digests on LCMS (see below). TMT labelling was
549 performed using 8 channels from a TMT™10plex kit (TMT lot: VH306773, ThermoFisher
550 Scientific, Hemel Hempstead, UK) according to the manufacturer's instructions with slight
551 modifications; samples were dissolved in 90 µl of 0.2 M EPPS buffer (MERCK)/10%
552 acetonitrile, and 200 µg TMT reagent dissolved in 22 µl of acetonitrile was added. Samples

553 were assigned to the TMT channels in two separate experiments, according to the table below.
554 After 2 h incubation, aliquots of 1.5 µl from each sample from each individual TMT experiment
555 were combined in 250 µl 0.2% TFA, desalted, and analysed on the mass spectrometer to check
556 the labelling efficiency and estimate total sample abundances. The main sample aliquots were
557 quenched by adding 8 µl of 5% hydroxylamine and combined to roughly level abundances.
558 The peptides were desalted using a C18 Sep-Pak cartridge (200 mg, Waters, Wilmslow, UK).
559 The eluted peptides were dissolved in 500 µl of 25 mM NH_4HCO_3 and fractionated by high pH
560 reversed phase HPLC. The samples were loaded to an XBridge® 3.5 µm C18 column (150 x
561 3.0 mm, Waters). Fractionation was performed on an ACQUITY Arc Bio System (Waters) with
562 the following gradient of solvents A (water), B (acetonitrile), and C (25 mM NH_4HCO_3 in
563 water) at a flow rate of 0.5 ml min^{-1}: solvent C was kept at 10% throughout the gradient; solvent
564 B: 0-5 min: 5%, 5-10 min: 5-10%, 10-60 min: 10-40%, 60-75 min: 40-80%, followed by 5 min
565 at 80% B and re-equilibration to 5% for 24 min. Fractions were collected every 1 min and
566 concatenated by combining fractions of similar peptide concentration to produce 20 final
567 fractions for MS analysis. A replicate of one donor sample (2 TMT channels) was fractionated
568 using the Pierce™ High pH Reversed-Phase Peptide Fractionation Kit (Thermo) to produce 7
569 fractions.
570
571 *TMT analyses*
572 Aliquots of all fractions were analysed by nanoLC-MS/MS on an Orbitrap Eclipse™ Tribrid™
573 mass spectrometer equipped with a FAIMS Pro Duo interphase coupled to an UltiMate® 3000
574 RSLCnano LC system (Thermo Fisher Scientific, Hemel Hempstead, UK). The samples were
575 loaded onto a trap cartridge (PepMap™ Neo 5 µm C18 300 µm X 5 mm Trap Cartridge,
576 Thermo) with 0.1% TFA at 15 µl min^{-1} for 3 min. The trap column was then switched in-line
577 with the analytical column (Aurora Frontier TS, 60 cm nanoflow UHPLC column, ID 75 µm,
578 reversed phase C18, 1.7 µm, 120 Å; IonOpticks, Fitzroy, Australia) for separation using the
579 following gradient of solvents A (water, 0.1% formic acid) and B (80% acetonitrile, 0.1%
580 formic acid) at a flow rate of 0.23 µl min^{-1} : 0-3 min 1% B (parallel to trapping); 3-10 min
581 linear increase B to 9 %; 10-105 min increase B to 50%; followed by a ramp to 99% B and re-
582 equilibration to 1% B.
583 Data were acquired with the following parameters in positive ion mode: MS1/OT: resolution
584 120K, profile mode, mass range *m/z* 400-1600, AGC target $4e^5$, max inject time 50 ms, FAIMS
585 device set to three compensation voltages (-35V, -50V, -65V) for 1 s each; MS2/IT: for each
586 CV, data dependent analysis with the following parameters: 1 s cycle time Rapid mode, centroid

587 mode, quadrupole isolation window 0.7 Da, charge states 2-5, threshold $1.9e^4$, CID CE=33,
588 AGC target 1e4, max. inject time 50 ms, dynamic exclusion 1 count for 15 s with mass
589 tolerance of 7 ppm; MS3 synchronous precursor selection (SPS): 10 SPS precursors, MS2
590 isolation window 0.7 Da, HCD fragmentation with CE=65, Orbitrap Turbo TMT and TMTpro
591 resolution 30k, AGC target 200%, max inject time 100 ms; Real Time Search (RTS): protein
592 database UP000005640_human_2024_9606.fasta (Uniprot, 20,609 entries), enzyme trypsin,
593 1 missed cleavage, oxidation (M) as variable, carbamidomethyl (C) and TMT6plex as fixed
594 modifications, precursor tolerance 10 ppm, Xcorr = 1.4, dCn = 0.1.
595
596 *Proteome data analyses*
597 Scripts are on: https://github.com/alicegodden/spermpool.
598 The acquired raw data were processed and quantified in Proteome Discoverer 3.1 (Thermo
599 Fisher Scientific); all mentioned tools of the following workflow are nodes of the proprietary
600 Proteome Discoverer (PD) software.
601 The UP000005640_human_2024_9606.fasta protein fasta database (Uniprot, 20,609 entries)
602 was imported into PD adding a reversed sequence database for decoy searches. The database
603 search was performed using the incorporated search engines CHIMERYS (MSAID, Munich,
604 Germany) and Comet.
605 The processing workflow for both search engines included recalibration of MS1 spectra (RC),
606 reporter ion quantification by most confident centroid (20 ppm) and a search on the human
607 protein database (as imported above). For CHIMERYS the Top N Peak Filter was applied with
608 20 peaks per 100 Da. Then the inferys_3.0.0. fragmentation prediction model was used with
609 fragment tolerance of 0.3 Da, enzyme trypsin with 1 missed cleavage, variable modification
610 oxidation (M), fixed modifications carbamidomethyl (C) and TMT6plex on N-terminus and K.
611 For Comet the version 2019.01 rev.0 parameter file was used with default settings except
612 precursor tolerance set to 6 ppm and trypsin missed cleavages set to 1. Modifications were the
613 same as for CHIMERYS.
614 The consensus workflow included the following parameters: intensity-based abundance,
615 normalisation on total peptide abundances, protein abundance-based ratio calculation, only
616 unique peptides (protein groups) for quantification, TMT channel correction values applied
617 (Lot VH306773), co-isolation/SPS matches/CHIMERYS Coefficient thresholds 50%/70%,
618 0.8, missing values imputation by low abundance resampling, hypothesis testing by *t*-test
619 (background based), adjusted p-value calculation by BH-method. The results were exported
620 into a Microsoft Excel table including data for normalised and un-normalised abundances,

621 ratios for the specified conditions, the corresponding p-values and adjusted p-values, number
622 of unique peptides, q-values, PEP-values, identification scores from both search engines; FDR
623 confidence filtered for high confidence (strict FDR 0.01) only. We used *AlphaFold* [5-7] to
624 generate molecule views of the candidate proteins.
625
626 **Validation of OLFM4 and CTNG**
627 ***Sample collection and preparation***
628 Following sample collection, liquefaction and CASA analysis to assess sperm motility in each
629 sample, samples were aliquoted to approximately 1.5 million cells in 1 mL of human tubal fluid
630 (HTF) medium. Aliquots were incubated for 4 hours at 37 °C with 5% CO_2 prior to antibody
631 staining. Following incubation, sperm samples were reassessed for motility and cell number
632 using CASA, then immediately stained with eFluor™ 520 fixable viability dye (eBioscience,
633 Invitrogen) according to the manufacturer's protocol. After viability staining, cells were fixed
634 in 4% paraformaldehyde (PFA) at 4 °C.
635 Fixed cells were then blocked using Human Fc Receptor Blocking Solution (eBioscience,
636 Invitrogen) to minimise non-specific antibody binding. Fluorophore-conjugated primary
637 antibodies were subsequently added and incubated under standard conditions. The following
638 antibodies were used: anti-OLFM4 clone 12, conjugated to R-phycoerythrin (PE), at a 1:10
639 dilution (Novus Biologicals, NBP3-06415PE); anti-CTNG (CTNG/2155R), conjugated to
640 mFluor Violet 450 SE, at 1:10 dilution (Novus Biologicals, NBP3-08485MFV450); and anti-
641 AOC3 (VAP1/AOC3 - TK8-14), conjugated to Alexa Fluor 594, at 1:10 dilution (Bio-Techne,
642 NBP2-81043AF594.). Throughout staining, cells underwent minimal washing steps to preserve
643 cell numbers and were stored overnight at 4 °C prior to flow cytometry analysis.
644
645 *Flow cytometry*
646 Samples were analysed using the BD FACSDiscover™ S8 Cell Sorter (BD Biosciences),
647 operated in fully spectral mode. No cell sorting was performed. Imaging and fluorescence data
648 were acquired simultaneously for each sample with fixable viability dye (eFluor™ 520)
649 detected using imaging laser B1 to allow for imaging. Events were gated to isolate the sperm
650 population (P1), based on forward and side scatter properties. All subsequent data acquisition
651 and analysis were restricted to the P1 population. For each antibody-stained sample, between
652 5,000 and 50,000 individual events within the P1 gate were recorded, depending on sample
653 quality and cell availability.

Spectral unmixing was performed using BD FACSChorus™ software (BD Biosciences), based on single-stained controls for each fluorophore-conjugated antibody and the fixable viability dye. An unstained sperm sample was included in experiment set to control for cellular autofluorescence during unmixing. Compensation and unmixing matrices were applied prior to further analysis.

Flow cytometry image files were merged with the primary fluorescence data using BD CellView™ Image Extractor. The resulting datasets were processed in R v4.4.3 [8]. Data were normalised using the CytoNorm package [9], and quality control was performed using PeacoQC [10] to identify and remove anomalous events. Gated and cleaned data were subsequently analysed in FlowJo v 10.10.0 (BD Life Sciences – FlowJo, LLC, Ashland, OR, USA. https://www.flowjo.com), with all events restricted to the P1 sperm population (see Suppl. Fig. S5-2 for gating strategy).

Statistical modelling of marker expression in live and dead cell populations was conducted in R v4.4.3 [8], with residual diagnostics simulated using the DHARMa package [11] to evaluate model fit and distribution assumptions. Model summaries were formatted using the Stargazer package for publication [12].

ACKNOWLEDGEMENTS

We thank Carlo Martins for his help with the proteomics work and 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: Sample collection, phenotypic data collection, data analyses, WGS data analysis, data interpretation, writing. JCK: Proteomics sample preparation, proteomics data analysis, data interpretation, writing. AG: WGS data analyses, data interpretation, writing. GA: sperm selection assay development, sample collection for WGS. GS: Proteomics sample preparation, data analysis, data interpretation. AB: Polygenic scores analyses, data interpretation, writing. 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, 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 in the European Nucleotide Archive 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. Danecek, P. *et al.* BCFtools v1.15.1. *GigaScience* **10**, giab008 (2021); available at https://samtools.github.io/bcftools/
4. Alavioon, G. *et al.* Haploid selection within a single ejaculate increases offspring fitness. *Proc. Natl Acad. Sci. USA* **114**, 8053–8058 (2017).
5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
6. 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).

7. 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).
8. R Core Team. *R: A Language and Environment for Statistical Computing* (v4.4.3). R Foundation for Statistical Computing, Vienna (2025).
9. Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y. & Aghaeepour, N. CytoNorm: a normalization algorithm for cytometry data. *Cytometry A* **97**, 268–278 (2020).
10. Emmaneel, A. *et al.* PeacoQC: peak-based selection of high-quality cytometry data. *Cytometry A* **101**, 325–338 (2022).
11. Hartig, F., Lohse, L. & de Souza Leite, M. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.4.6 (2024).
12. Hlavac, M. Stargazer: well-formatted regression and summary statistics tables from R. R package version 5.2.3 (2022).

FIGURE LEGENDS

Figure 1: Phenotypic signature of within-ejaculate selection on sperm quality. Phenotypic traits 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). **(A)** Curvilinear velocity, **(B)** % sperm with mature chromatin (using aniline), **(C)** % sperm with normal morphology (using SpermBlue), and **(D)** % sperm without single- or double-strand breaks (using toluidine blue). Data for all donors are shown as individual dots and overall mean ± standard error are shown in black for each time point (sample sizes: 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 human sperm quality. Loci with significantly diverging alleles between sperm pools collected **(A)** from Centre (low-quality) and Edge (high-quality) in the Methyl Cellulose assay (n = 4) and **(B)** at T0 (low quality) and T4 (high quality) from Swim Up assays (n = 5; likelihood ratio tests, > 99.9th quantile). Top 20 significant Biological GO terms for the genes showing significant divergence in **(C)** Methyl cellulose and **(D)** Swim Up. Distribution of genomic elements among significantly diverging loci for **(E)** Methyl Cellulose and **(F)** Swim Up. The heatmap shows the Z-score normalised average difference in polygenic scores for the top traits selected based on the 99th percentile of absolute score differences across donors. Positive values (red) indicate

754 traits with higher-than-average score differences, while negative values (blue) indicate lower-
755 than-average differences for **(G)** Methyl Cellulose and **(H)** Swim Up.
756
757 **Figure 3: Comparison of genomic signature of within-ejaculate selection in zebrafish and**
758 **humans. (A)** Loci with significantly diverging alleles between sperm pools collected from
759 centre (low quality) and edge of a water droplet (n = 3; data from Alavioon et al 2017[5]) and (B)
760 the associated top 20 significant biological GO terms. **(C)** Distribution of genomic elements
761 among significantly diverging loci. **(D)** Venn diagram showing overlap of significantly
762 diverging genes between human Methyl Cellulose, human Swim Up and zebrafish. **(E)** Top 20
763 significant biological GO terms for the 153 genes share between zebrafish, human Methyl
764 Cellulose and human Swim Up assays.
765
766 **Figure 4: Proteomic signature of within-ejaculate selection on sperm performance. (A)**
767 Venn diagram showing overlap of differentially expressed proteins between donors,
768 experiments and technical runs. **(B)** Volcano plots for each donor displaying significantly
769 differentially expressed proteins (pink; (significance cut-off points are $p_{adj} < 0.05$; $\log_2$ fold
770 change > |2|) between Centre/T0 and Edge/T4 sperm pools assays with $\log_2$ fold changes > 0
771 indicating increased expression in Centre/T0 sperm pools and $\log_2$ fold changes < 0 indicating
772 increased expression in Edge/T0 sperm pools. Two technical runs were performed for Donor A
773 in the Swim Up assay. The top five significant proteins shared among all samples are
774 highlighted by black circle and protein ID. **(C)** Validation of OLFM and CTNG as markers of
775 human sperm quality using FACS on cells stained with a combination of antibody labelling and
776 live-dead staining. 'Positive' indicates staining for marker protein above gating threshold and
777 'negative' indicates sub-threshold labelling for antibodies (see Supplementary Material for
778 gating details).
779

Fig. 1

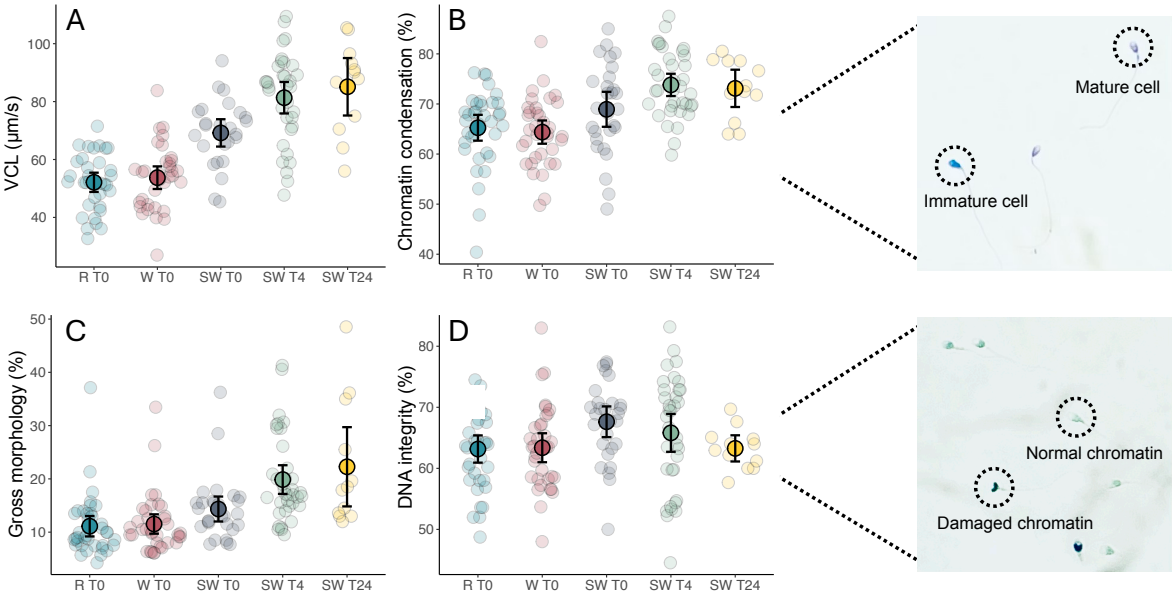

A
VCL (µm/s)
100
80
60
40
R T0
W T0
SW T0
SW T4
SW T24
B
Chromatin condensation (%)
80
70
60
50
40
R T0
W T0
SW T0
SW T4
SW T24
Mature cell
Immature cell
C
Gross morphology (%)
50
40
30
20
10
R T0
W T0
SW T0
SW T4
SW T24
D
DNA integrity (%)
80
70
60
50
R T0
W T0
SW T0
SW T4
SW T24
Normal chromatin
Damaged chromatin

Fig. 2

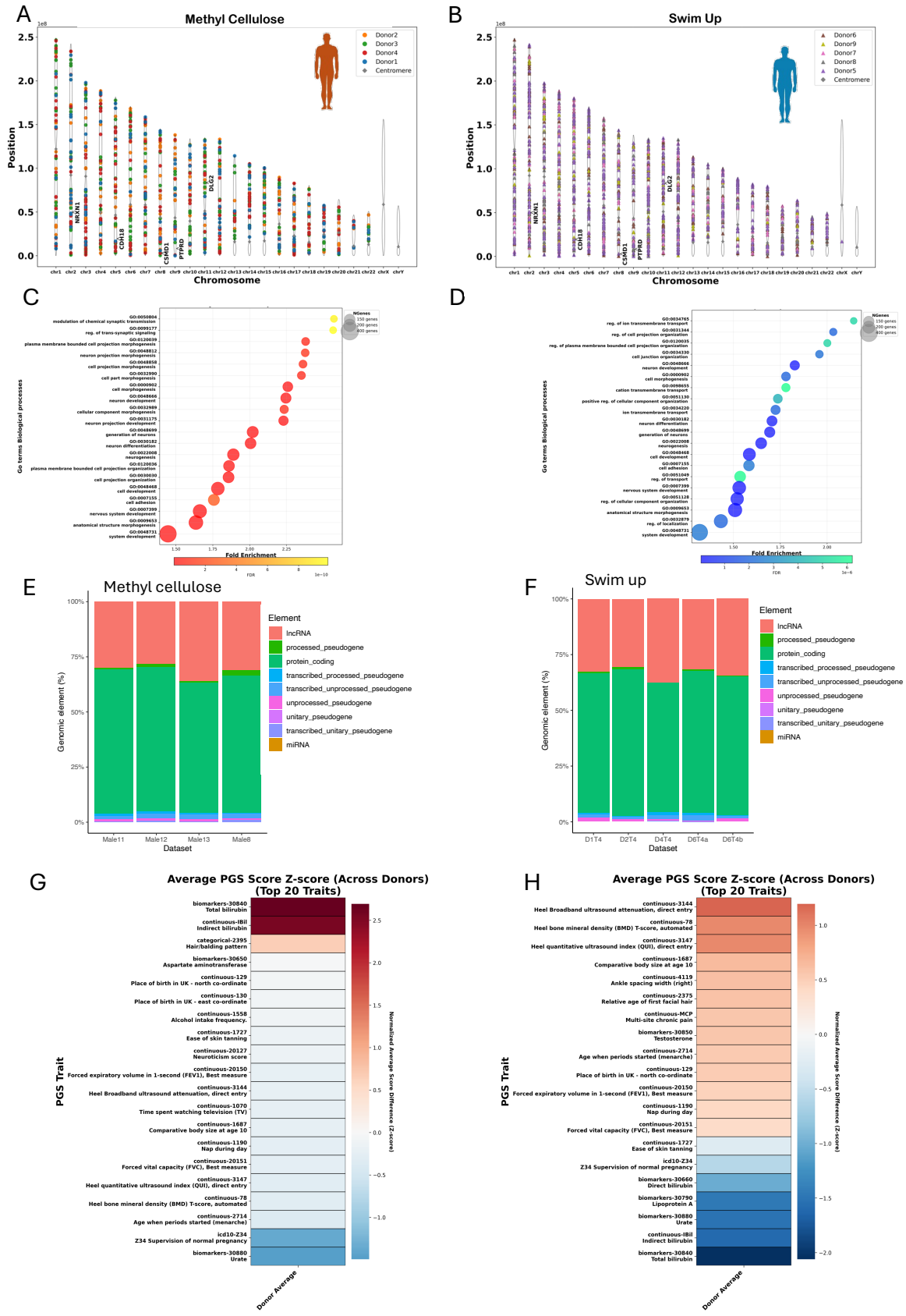

A
Methyl Cellulose
Position
Chromosome
Donor2
Donor3
Donor4
Donor1
Centromere
NRXN1
CDH18
CSMD1
PTPRD
DLG2
B
Swim Up
Donor6
Donor9
Donor7
Donor8
Donor5
C
Go terms Biological processes
Fold Enrichment
NGenes
D
E
Methyl cellulose
Genomic element (%)
Dataset
Element
lncRNA
processed_pseudogene
protein_coding
transcribed_processed_pseudogene
transcribed_unprocessed_pseudogene
unprocessed_pseudogene
unitary_pseudogene
transcribed_unitary_pseudogene
miRNA
Male11
Male12
Male13
Male8
F
Swim up
D1T4
D2T4
D4T4
D6T4a
D6T4b
G
Average PGS Score Z-score (Across Donors)
(Top 20 Traits)
PGS Trait
Donor Average
Normalized Average Score Difference (Z-score)
H

Fig. 3

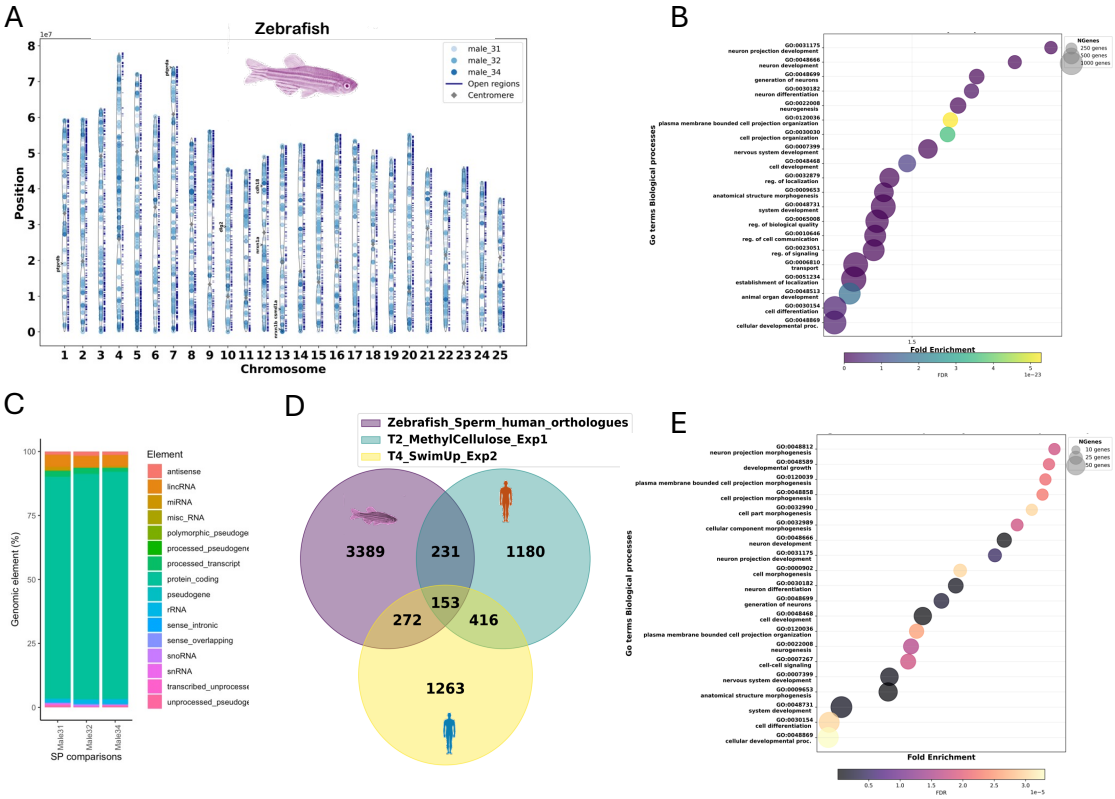
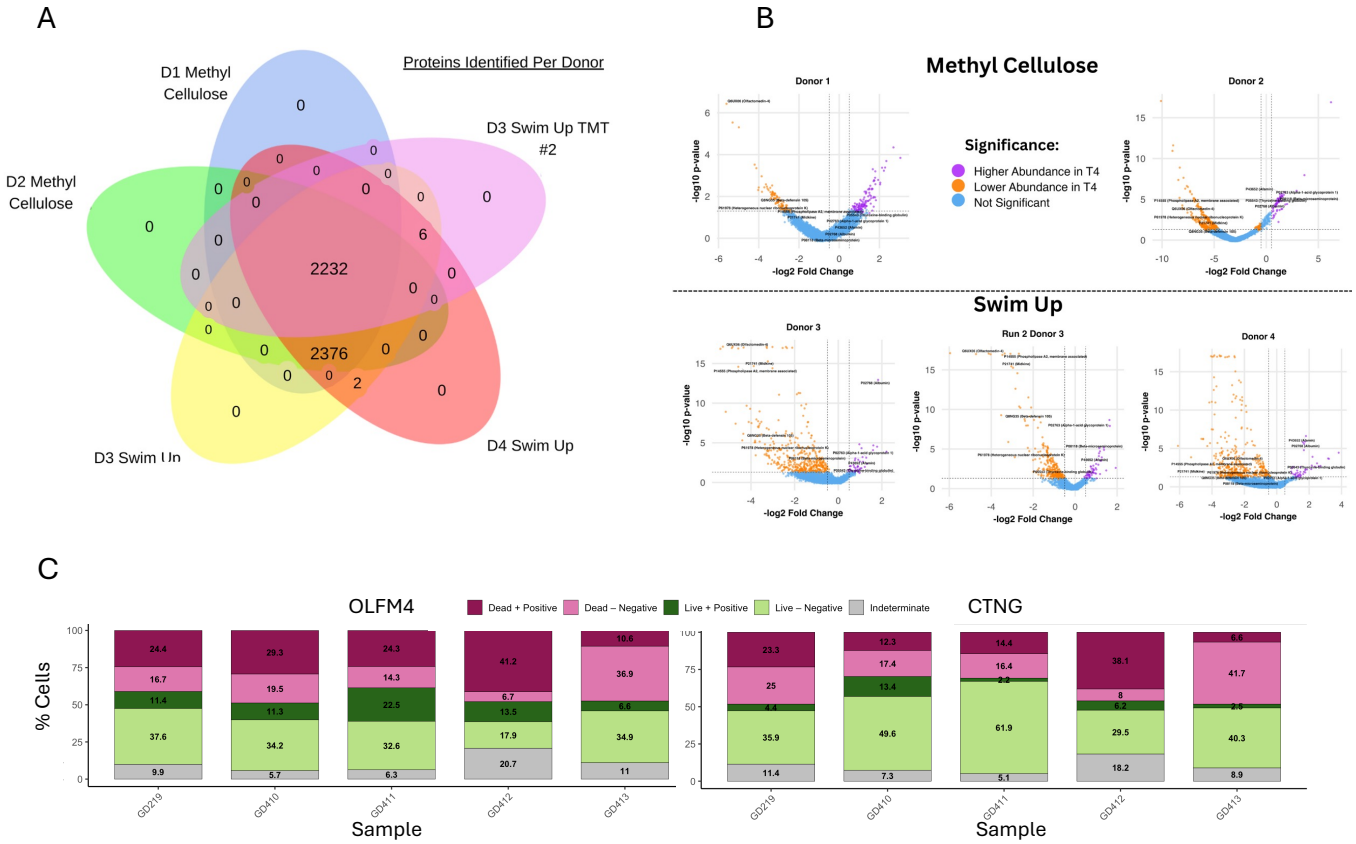

A
Zebrafish
male_31
male_32
male_34
Open regions
Centromere
Position
Chromosome
B
Go terms Biological processes
Fold Enrichment
FDR
NGenes
C
Genomic element (%)
SP comparisons
Element
antisense
lincRNA
miRNA
misc_RNA
polymorphic_pseudoge
processed_pseudogene
processed_transcript
protein_coding
pseudogene
rRNA
sense_intronic
sense_overlapping
snoRNA
snRNA
transcribed_unprocesse
unprocessed_pseudoge
Male31
Male32
Male34
D
Zebrafish_Sperm_human_orthologues
T2_MethylCellulose_Exp1
T4_SwimUp_Exp2
3389
231
1180
153
272
416
1263
E
Go terms Biological processes
Fold Enrichment
FDR
NGenes

Fig. 4


A
Proteins Identified Per Donor
D1 Methyl Cellulose
D2 Methyl Cellulose
D3 Swim Up TMT #2
D3 Swim Up
D4 Swim Up
2232
2376
B
Methyl Cellulose
Donor 1
Donor 2
Significance:
Higher Abundance in T4
Lower Abundance in T4
Not Significant
-log10 p-value
-log2 Fold Change
Swim Up
Donor 3
Run 2 Donor 3
Donor 4
C
OLFM4
CTNG
Dead + Positive
Dead – Negative
Live + Positive
Live – Negative
Indeterminate
% Cells
Sample
GD219
GD410
GD411
GD412
GD413