The conserved transcriptional program of metazoan male germ cells uncovers ancient origins of human infertility

Rion Brattig Correia[1,2,*], Joana M. Almeida[1,3,*], Margot J. Wyrwoll[4,*], Irene Julca[5], Daniel Sobral[6,7], Chandra Shekhar Misra[1,8], Leonardo G. Guilgur[1], Hans-Christian Schuppe[9], Neide Silva[1], Pedro Prudêncio[10], Ana Nóvoa[1], Ana S. Leocádio[1], Joana Bom[1], Moisés Mallo[1], Sabine Kliesch[11], Marek Mutwil[5], Luis M. Rocha[1,2,**], Frank Tüttelmann[4,**], Jörg D. Becker[1,8,**,#] & Paulo Navarro-Costa[1,3,**,#]

1- Instituto Gulbenkian de Ciência. Oeiras, Portugal.

2- Department of Systems Science and Industrial Engineering, Binghamton University. New York, USA.

3- Instituto de Saúde Ambiental, Faculdade de Medicina, Universidade de Lisboa. Portugal.

4- Institute of Reproductive Genetics, University of Münster. Germany.

5- School of Biological Sciences, Nanyang Technological University. Singapore.

6- Associate Laboratory i4HB - Institute for Health and Bioeconomy, NOVA School of Science and Technology, NOVA University Lisbon. Caparica, Portugal.

7- UCIBIO - Applied Molecular Biosciences Unit, Department of Life Sciences, NOVA School of Science and Technology, NOVA University Lisbon. Caparica, Portugal.

8- Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa. Oeiras, Portugal.

9- Clinic of Urology, Pediatric Urology and Andrology, Justus-Liebig-University. Giessen, Germany.

10- Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa. Portugal.

11- Centre of Reproductive Medicine and Andrology, University Hospital Münster. Germany.

* and ** Equal contribution

e-mail: JDB (jbecker@itqb.unl.pt); PNC (navarro-costa@medicina.ulisboa.pt)

Germ cells provide the cellular basis for sexual reproduction in multicellular animals. In males, germ cells differentiate into sperm, one of the most morphologically diverse eukaryotic cell types. Owing both to this remarkable diversity and to the rapid evolution of reproduction-related genes, the transcriptional program of male germ cells is widely regarded as divergent across species[1,2]. However, the possibility that these cells retain a distinctive evolutionarily-conserved genetic basis remains largely unexplored. Here we show, using phylostratigraphy, that the complex male germ cell transcriptome has an old evolutionary origin shared between vertebrate and invertebrate species. Through multilayer network analysis of the human, mouse and fruit fly meiotic transcriptome, we establish that old genes serve as a genetic scaffold from which complexity has evolved, and identify a core set of 79 ancient functional interactions at the heart of male germ cell identity. By silencing a cohort of 920 candidate genes likely to affect the acquisition and maintenance of this identity, we uncover 164 previously unknown spermatogenesis genes. Integrating this information with whole-exome sequencing data from azoospermic men reveals three novel genetic causes of human male infertility shared between species that have diverged for more than 600 million years. Our results redefine the importance of evolutionary history on human reproductive disease and illustrate the translational potential of a new synergy between comparative network biology and medical genetics. We anticipate that our open-access and easily-adaptable interdisciplinary research platform will be harnessed in the context of other cell types and diseases.

Understanding what defines the uniqueness of a given cell type out of the 843 predicted cellular fates in the human body is a complex and fascinating problem[3]. Through Conrad Waddington's foundational work, we have come to appreciate that developmental trajectories ultimately dictate cell type identity via the establishment of specific transcriptional programs[4]. The fact that transcriptomes tend to cluster by tissue type rather than by species[5] clearly indicates that gene expression identity can be maintained across many million years of evolutionary divergence. This echoes the modular nature of eukaryotic biological processes, whose intervening macromolecular complexes are typically built by the addition of younger components to a core block of ancient subunits[6].

The emergence of germ cells is considered one of the first cell type specializations in metazoan history[7]. Since the capability to undergo both sexual reproduction and gametogenesis were already present in the unicellular ancestor of all metazoa[8], the split between germ line and soma presumably provided early multicellular organisms increased robustness against mutations while minimizing genetic conflict between different cell lineages[9]. Accordingly, the role of germ cells as conduits for the life history of species exposes them to a unique array of powerful evolutionary forces at the level of postcopulatory sexual selection. These are considered driving factors for the rapid evolution observed in many reproduction-related genes[1], and for the well-documented tendency for new genes to originate in male germ cells[10]. Therefore, it is not surprising that when it comes to the genetics of gamete differentiation, the emphasis is often placed on the new rather than on the old.

However, our recent data argue for a critical reappraisal of the role of old genes in germ cells. More specifically, we observed, in a wide range of plant species, a substantial contribution of old genes to the pollen transcriptome, suggestive of an ancient transcriptional program common to plant male gametes[11]. Here, we bridge this concept to multicellular animals to determine if the identity of metazoan male germ cells has an old, evolutionarily-conserved genetic basis.

To test this hypothesis, we devised an interdisciplinary research platform based on four combined approaches (**Fig. 1a**). First, we determined, through phylostratigraphy, the age of the gene expression program of male germ cells from three evolutionarily-distant metazoan species: humans (*Homo sapiens*), mice (*Mus musculus*) and fruit flies (*Drosophila melanogaster*). Then we used network science to infer the significance of old genes within the context of the complex male germ line transcriptome. Subsequently, through developmental biology (*in vivo* RNAi in fruit fly testes), we defined the role of a key subset of the conserved germ cell transcriptome in male reproductive fitness. Finally, we merged this information with clinical genetics to identify new causes of human infertility. Overall, we show that old genes play a prominent role in male germ cell regulation and that the disruption of this ancient genetic program leads to human reproductive disease.

Male germ cell transcription is intrinsically complex

Male germ cell development is divided into three conserved stages[12]. The first is the pre-meiotic stage and corresponds to the mitotic expansion of committed precursors (spermatogonia). Meiosis defines the second stage, with the newly-differentiated spermatocytes undergoing reductive division. In the third stage, the post-meiotic cells (spermatids) embark on a cytodifferentiation program that culminates with the formation of mature male gametes.

To understand to what extent male germ cell transcription quantitatively differs from that in somatic lineages, we collected previously published high-quality RNA-Seq datasets from pre-meiotic, meiotic and post-meiotic germ cell populations, and compared them to representative somatic cell types of the primary embryonic layers: neurons (ectoderm), muscle (mesoderm) and enterocytes (endoderm; **Fig. 1b** and **Sup. Table 1**). We centered our analysis on three evolutionarily-distant gonochoric species with excellent genome annotations - humans, mice and fruit flies - and observed that both vertebrate and invertebrate male germ cells have a generally more complex transcriptome than their somatic counterparts (as measured by the percentage of the entire genome each cell type expresses). At the more permissive minimum average expression cut-off of >0.01 transcripts per million (TPM), meiotic and pre-meiotic germ cells expressed from 76.7 to 92.3% of the protein-coding genome, depending on the species. These values were 14.3 to 19.4% higher, on average, than in corresponding somatic cells. The conservation of such increased complexity across species with different genome sizes (13,947 protein-coding genes in fruit flies *vs.* 22,802 and 22,287 in humans and mice, respectively) and sperm competition levels (particularly high in fruit flies[13]) is suggestive of an intrinsic demand for a comprehensive transcriptional program across metazoan male germ cells, preceding the extensive epigenetic remodeling that occurs at the post-meiotic stage[14].

The male germ cell transcriptome has an old evolutionary origin

We next assessed the contribution of old genes to the complexity of the germ cell transcriptome. For this we used phylostratigraphy, a technique that determines the evolutionary age of groups of homologous genes (defined as orthogroups in our analysis) by mapping their last common ancestor in a species tree[15]. We assembled our tree based on the proteomes of 25 phylogenetically-representative eukaryotic species, and assigned each orthogroup to the oldest phylogenetic node (phylostratum) it could be traced back to (**Fig. 1c** and **Sup. Fig. 1**). This way,

phylostrata ranked 1 to 5 were the oldest and contained orthogroups common to all metazoa, while phylostratum 16 contained the youngest, species-specific orthogroups. A total of 113,757 orthogroups were identified in the 25 representative eukaryotes, 85.5% of which were species-specific (97,270). In light of the rapid divergence typically associated with reproduction-related genes, it was rather unexpected to find that the majority of genes expressed in male germ cells (65.2 to 70.3% depending on the species, minimum average TPM cut-off >1) mapped to the oldest-ranking phylostrata containing orthogroups common to all metazoa (**Fig. 1d**). Indeed, in all three tested species this percentage was not significantly different from that recorded in the somatic cell types. The fraction of old male germ cell orthogroups that were ubiquitously expressed (proxy to their involvement in cellular housekeeping processes) varied from 49.7 to 63.4%, depending on the species (**Sup. Fig. 2a-b**). This strongly suggests that a quite significant number of old genes expressed by male germ cells are involved in more specific roles than just the maintenance of basal cellular functions.

By summing the products of the age of all expressed genes and their expression levels at a given developmental stage (a metric known as the transcriptome age index - TAI[16]) we determined the transcriptome age of the different male germ cell stages (**Fig. 1e**). Meiotic and pre-meiotic cells across all three tested species had lower TAIs than post-meiotic cells, indicative of older transcriptomes. This trend was less obvious in the fruit fly as in this species post-meiotic transcription is largely residual[17]. Collectively, we observed that, both in vertebrates and invertebrates, the complexity of the male germ cell transcriptome has an old evolutionary origin that is tempered by the increased expression of younger genes at later developmental stages.

Conserved genes have a prominent role in the male meiotic transcriptome

We next addressed the possible significance of the abundant expression of old genes in the male germ line. We reasoned that if a conserved gene expression program were to define male germ cell identity, this program would preferentially build on the ancient origins of meiosis - a process dating back to the last eukaryotic common ancestor[18]. Indeed, male prophase I chromatin organization is highly conserved[19], and such organization activates the expression of the spermatogenic program across distant metazoan species[20]. Thus, we assembled the transcriptome networks of human, mouse, and fruit fly spermatocytes, where nodes represent all

genes expressed at a minimum average TPM cut-off >1 and edge weights indicate the probability of the connected genes contributing to a specific function according to the STRING protein-protein interaction database[21] (**Fig. 1f**). The structure of these meiotic networks reflects the multiple genetic interdependencies responsible for cellular function, as illustrated by the characteristic clustering of functionally-related genes into topologically-defined modules[22]. Network edges were filtered to only include high confidence interactions (score ≥0.5, **Sup. Fig. 3**). The resulting meiotic networks contained between 7,961 and 11,322 genes, depending on the species, and an average of 290,000 edges.

Consistent with phylostratigraphy, the meiotic networks included a substantial fraction of genes conserved across all metazoa (45.1 to 58.1% of all genes; **Fig. 1g**). Of note, these conserved genes had two important properties: i- they were significantly more connected than their non-conserved counterparts (higher degree centrality, in line with previous observations on essential genes[23]); and ii- their interactors were themselves more connected than those of non-conserved genes (higher page rank, Kolmogorov-Smirnov test for both analyses; **Fig. 1h-i**). To address a possible ascertainment bias associated with more available information on conserved genes, we tested to what extent degree centrality and page rank were affected by network rewiring. In this approach, a variable percentage (20 to 100%) of all edges are randomly shuffled across the network, thus diluting any latent biases in the datasets. We observed that both network metrics remained higher in conserved genes even when 80% of all network edges were rewired (**Sup. Fig. 4**). Such results strongly suggest that the increased centrality and connectivity of conserved genes are mostly driven by their intrinsic properties, rather than by differences stemming from the amount of source data available. Indeed, both properties were so salient that it was possible for machine learning classifiers to reliably predict if a gene was conserved just based on network features. More specifically, using a Random Forest classifier, the AUC (area under the receiver operating characteristic curve) score was 0.74, 0.75, and 0.82 in the human, mouse, and fruit fly datasets, respectively (**Fig. 1j**). Classification performance was equally high when using a different supervised learning algorithm (linear Support Vector Machine) and an alternative performance metric (precision and recall; **Sup. Fig. 5**), demonstrating that the evolutionary conservation of a gene can be robustly predicted from the more central network connectivity pattern of its interactions. Based on these analyses, we conclude that conserved genes are

central components in the male meiotic transcriptome of evolutionarily-distant species. Through this prominent role, they likely represent an ancient genetic scaffold for the progressive acquisition of transcriptional complexity in male germ cells.

An ancient gene regulatory module of male germ cell identity

We have previously shown that it is possible to simplify the complexity of biological networks by removing redundancy and only retaining the most relevant interactions that sustain the dynamics of the system[24]. To do so, one can extract a network backbone consisting of the edges that are sufficient to compute all shortest paths in the original network[25]. In the context of transcriptome networks, this distance backbone uncovers the subgraph of gene interactions more likely to convey the key functional characteristics of the gene expression program[26].

By extending orthology to the backbone, we have developed a new type of multilayer network: the orthoBackbone. It expands on the concept of interologs[27,28], as it consists of only the edges that connect the same pair of orthologs in the transcriptome backbone of different species (**Fig. 2a**). Since the orthoBackbone contains key interaction subnetworks that have been maintained across evolution, it offers unprecedented access to what can be considered the core gene regulatory module of a given cellular state.

The orthoBackbone drastically reduced the redundancy of the meiotic networks, retaining only 1.7 to 2.7% of all edges, depending on the species (**Fig. 2b**). Despite this extreme reduction, it remarkably preserved approximately 70% of all conserved genes (**Sup. Table 2**), thus validating our assertion that old genes serve as a genetic scaffold for the meiotic transcriptome (**Fig. 2c**). Gene Ontology (GO) term enrichment analysis for biological processes revealed that orthoBackbone genes were preferentially involved in gene expression and protein regulation, in contrast with the other conserved genes that were mainly associated with cell signaling pathways (**Fig. 2d** and **Sup. Fig. 6**). The preferential contribution of the orthoBackbone to the gene expression machinery further emphasized the association of this subnetwork with the acquisition/maintenance of cellular identity. Using GO annotations, we further selected orthoBackbone interactions involving gene expression regulators previously linked to spermatogenesis. By doing so we obtained a subset of 79 functional interactions between 104 human genes (out of the total of 3,596 in the orthoBackbone; **Fig. 2e**). These were ascribed to a

wide gamut of regulatory processes and involved noteworthy spermatogenesis genes such as *RFX2* (transcription), *CDYL* (chromatin remodeling) and *BOLL* (translation), among others[29,30,31]. Based on its conservation in the meiotic orthoBackbone of evolutionarily-distant species and on the significant spermatogenic role of the intervening genes, this subnetwork of 79 functional interactions can be regarded as an ancient gene regulatory module at the heart of male germ cell identity.

Male germ cell identity has a broad functional basis

We next set out to determine the functional consequences of disrupting the orthoBackbone. Since the latter represents a still sizeable interaction subnetwork consisting of more than 3,000 genes, we decided to select candidate genes likely to affect the acquisition and maintenance of male germ cell identity. For this, we took into consideration the profound changes that meiosis introduces on germ cell transcription - an effect that has been likened to that of cellular reprogramming[32,33]. We reasoned that genes that are upregulated at meiotic entry and/or downregulated at meiotic exit likely represent instructive elements for the meiotic cell state, hence preferential routes to tamper with male germ cell identity.

Through differential gene expression analysis, we identified, in human and mouse spermatocytes, all expressed genes that shared a similar upregulation at meiotic entry and/or downregulation at meiotic exit. Of these 970 mammalian meiotic differentially expressed genes (DEGs), we selected as candidates for functional assessment only those that were also expressed in fruit fly germ cells (**Fig. 2f**). We did not take into account differential expression in the latter species, since the largely residual levels of post-meiotic transcription in fruit fly spermatogenesis thwarts direct comparisons with the mammalian system[17]. The resulting 920 fruit fly genes (homologous to 797 and 850 in humans and mice, respectively) were silenced specifically prior to meiotic entry by Drosophila *in vivo* RNAi, using the well-established *bam*-GAL4 driver[34]. Of these, a total of 250 genes (27.2%) were essential for male fertility, as their silencing resulted, upon mating with wild-type females, in egg hatching rates below the cut-off of 75% (>2 standard deviations of the mean observed in negative controls: 91.6 ± 8.5%; **Fig. 2g**). Significantly, 190 out of these 250 hits (76.0%) were part of the orthoBackbone. Cytological analysis of all 250 genes by testicular phase-contrast microscopy in the RNAi males revealed diverse origins for the infertility phenotype, with the

earliest manifestation of cellular defects ranging from the pre-meiotic to the mature gamete stage (**Fig. 2h**). By exploring publicly-available information (see **Methods**), we determined that 164 (65.6%) of all hits had never been previously reported as functionally required for male reproduction in any species (**Fig. 2i**). Accordingly, these 164 new conserved spermatogenesis genes (homologous to 183 and 190 in humans and mice, respectively) represent both a substantial advance in our understanding of the genetic basis of male germ cell development and a valuable resource to explore from a precision medicine perspective. To facilitate open access to this information, we made all data available in the form of a user-friendly gene browser (Meiotic Navigator; **Fig. 2j**). Overall, by merging our results with previously published data we conclude that at least 42.9% of a developmentally-regulated subset of the conserved meiotic transcriptome (our 250 hits + 145 previously reported male fertility genes that gave a negative result in our assay) is required for a surprisingly diverse range of germ cell functions from the pre-meiotic to the mature gamete stage. This highlights the pervasive influence of meiosis on the spermatogenic program and argues for a broad functional basis underlying male germ cell identity.

Ancient origins of human infertility

One of our 164 newly-identified spermatogenesis genes - the Drosophila RING finger protein 113 (*dRNF113*, a spliceosomal component also known as *mdlc*[35]) - emerged as a particularly interesting meiotic regulator. By analyzing our in-house whole exome database containing sequencing data from 74 cases of human male meiotic arrest, we identified an infertile man harboring a homozygous loss-of-function (LoF) variant in a testis-specific paralog (*RNF113B*) belonging to the same orthogroup[36]. This frameshift variant c.556_565del;p.(Thr186GlyfsTer119) leads to the abrogation of the protein's two functional domains (a C3H1-type zinc finger and a RING finger) and to its truncation (**Fig. 3a**). The identified man (M1911) is part of a consanguineous family of Middle-Eastern ancestry and shared the homozygous state of the *RNF113B* LoF variant with his equally infertile brother, but not with his fertile heterozygous brother (**Sup. Fig. 7** and **Sup. Information**). Remarkably, the results from the testicular biopsy of M1911 revealed an equivalent meiotic arrest phenotype to that observed in *dRNF113*-silenced fruit flies. Indeed, spermatocytes were, by far, the predominant cell type in the male gonads of both species: they were the most advanced cell stage observed in 89.0% of all assessed human seminiferous

tubules (*vs.* 9% in controls) and occupied an average of 64.1% of the entire testis area in fruit flies (*vs.* 12.3% in controls; **Fig. 3b-c**). Early (round) spermatids were practically absent in both species, with cellular debris accumulating in the post-meiotic region of the fruit fly gonad. By analyzing our previously published single-cell RNA-Seq dataset of normal human spermatogenesis[37], we observed that *RNF113B* was predominantly expressed at meiotic entry, peaking at the diplotene stage of prophase I (**Fig. 3d**). This mirrored the protein localization pattern of fruit fly dRNF113, characterized by a substantial nuclear accumulation in primary spermatocytes (**Fig. 3e**).

Strengthening our interest in the *RNF113* genes was their prevalent effect on meiotic transcription. Indeed, by analyzing M1911's testicular transcriptome via RNA-Seq, we observed the deregulation of 21.7% of all expressed genes (**Fig. 3f**). Notably, such deregulation had a significant impact on the meiotic orthoBackbone, with 30.0% of its network edges being disrupted (i.e., containing at least one deregulated gene). A similar effect was patent in the dRNF113 RNAi, with 20.3% of the fruit fly testicular transcriptome being deregulated, resulting in the disruption of 26.3% of the orthoBackbone (**Fig. 3g**). For comparison, the silencing of *Prp19*, another meiotic DEG and spliceosomal component that RNF113 proteins associate to[38], had a lower effect on the orthoBackbone (15.2% of disrupted edges), despite a similar meiotic arrest phenotype (**Sup. Fig. 8**). Based on the above, we conclude that the *RNF113* genes have retained a key regulatory role over meiotic transcription and cell cycle progression for more than 600 million years of evolution, thus serving as one of the potential gatekeepers of male germ cell identity. Furthermore, they provide us the means to interfere with the meiotic orthoBackbone in humans and fruit flies - an advantage we next explored from a clinical perspective.

The orthoBackbone as an ancillary tool in clinical genetics

Male infertility is a complex human disease with a poorly-defined genetic component. This contributes to a low causative diagnostic yield (typically just 30%), and to a paucity of clinically-validated genes (currently just 104, in contrast with the more than 700 already associated with other clinical disorders such as intellectual disability)[39,40]. Since male infertility affects up to 12% of all men[41], addressing such knowledge gap is an issue of clear medical importance. To attempt to narrow the gap, we explored the possibility that spermatogenesis is particularly sensitive to

disturbances in the meiotic orthoBackbone. Thus, we harnessed the sizeable effect of RNF113B on this network as means of identifying additional genetic causes of human infertility. Indeed, we noted that 40.5% (32 out of 79 functional interactions) of the gene regulatory module of male germ cell identity was deregulated in M1911 (i.e., contained at least one differently expressed gene). Therefore, we posited that the similar spermatogenic impairment recorded in the human *RNF113B* LoF variant and on the fruit fly dRNF113 RNAi ultimately reflected the downregulation, in both species, of a common set of orthoBackbone genes.

By defining the overlap between differently expressed orthoBackbone genes in the testicular transcriptome of M1911 and of the dRNF113 RNAi, we identified 61 conserved human genes that were similarly downregulated in both species. These formed a connected network (based on STRING data) suggestive of their functional co-involvement in related biological processes (**Fig. 3h**). Of the 61 genes, 27 had already been linked to male germ cell defects in different species: four were clinically-validated male infertility genes (*CDC14A*, *CFAP91*, *DNAI1* and *DNAI2*), and the other 23 had been previously reported in animal models (see **Methods**). The fact that among the latter was *BOLL*, one of the oldest known metazoan gametogenesis genes[31], was a particularly noteworthy observation. We next tested if these 27 genes could be used to identify new genetic causes of human infertility. For this we analyzed whole-exome sequencing data of 1,021 azoospermic men from the MERGE cohort[42]. Filtering these exomes for LoF variants in the 27 selected orthoBackbone genes revealed two new human male infertility genes: *HSPA2* and *KPNA2*.

HSPA2 - heat shock protein family A member 2 - is a molecular chaperone of the highly conserved 70-kDa heat shock protein (HSP70) gene family. HSP70 members are involved in cellular proteostasis from bacteria to human, with the mouse and fruit fly *HSPA2* homologs (*Hspa2* and *Hsc70-1* to *Hsc70-5*, respectively) being required for meiotic progression past the primary spermatocyte stage[43,44]. Indeed, we observed that the silencing of *Hsc70-1* in the fruit fly testis, also with the *bam*-GAL4 driver, resulted in an equivalent meiotic arrest to that of *dRNF113*, strongly suggesting that the misexpression of this chaperone is a central element of the latter phenotype (**Sup. Fig. 9a-b**). Two different heterozygous LoF variants were detected in our male infertility cohort, with this gene having a predicted autosomal-dominant inheritance (**Sup. Fig. 9c-f, Sup. Information** and **Methods**). The first, detected in individual M1678, is the early stop-gain

variant c.175C>T;p.(Gln59Ter) that truncates more than 90% of the protein, likely leading to nonsense-mediated decay. The second, identified in individual M2190, is the frameshift variant c.1806dup;p.(Glu603ArgfsTer81) which affects the distal end of the protein's nucleotide-binding domain. Histopathological analysis of M2190's testicular tissue revealed a complete loss of germ cells (SCO: Sertoli cell-only phenotype) in 259 assessed seminiferous tubules, a likely aggravation of the extensive cell death accompanied by tubule vacuolization reported in *Hspa2* mutant mice[45] (**Fig. 3i**).

KPNA2 - karyopherin subunit alpha 2 - is a nuclear importin that functions as an adapter protein in nucleocytoplasmic transport. Its mouse ortholog (*Kpna2*) is required for the nuclear accumulation of Hop2[46], a conserved regulator of meiotic progression from yeast to mammals[47,48]. The main function of Hop2 is to repair meiotic DNA double-strand breaks (DSBs), with the corresponding mouse mutant being characterized by a primary spermatocyte arrest coupled to extensive cell death. Since male fruit flies dispense with meiotic DSBs[49], the inclusion of *KPNA2* in the orthoBackbone seems counter-intuitive. Yet, Drosophila importin alpha 2 (*dKPNA2*, also known as *Pendulin*) is required for post-meiotic development[50], with its silencing specifically at meiotic entry being associated with aborted spermiogenesis (**Sup. Fig. 10a-b**). This post-meiotic function might also be present in its mammalian orthologs, as suggested by the nuclear localization of Kpna2 in elongating mouse spermatids[46]. We detected two heterozygous *KPNA2* LoF variants in our cohort of infertile men, with this gene also having a predicted autosomal-dominant inheritance (**Sup. Fig. 10c-g**, **Sup. Information** and **Methods**). The splice-site variant (c.667-2A>G;p.?) in individual M1645 is predicted to disrupt the correct splicing of intron 6 due to the loss of an acceptor site. Segregation analysis revealed the *de novo* occurrence of this variant, strongly supporting both its pathogenicity and our recent results pointing to the important contribution of new mutations to male infertility[51]. The other *KPNA2* LoF variant, in individual M2098, is the frameshift variant c.1158_1161del;p.(Ser387ArgfsTer14) which affects 27% of the protein, including its minor nuclear localization signal binding site. The available testicular histopathology report for the latter individual lists a SCO phenotype as the cause of the azoospermia, again suggestive of the possible deterioration of an initially meiotic phenotype (**Fig. 3i**).

In summary, by experimentally disrupting the meiotic orthoBackbone across evolutionarily-distant species, we were able to uncover two new genetic causes of human infertility (*HSPA2* and *KPNA2*) in addition to *RNF113B*, affecting a combined total of 5 individuals. This successful merger between basic and clinical research highlights the advantages of interdisciplinarity when tackling complex medical conditions.

Discussion

Our study indicates that the male germ cell transcriptome has an old evolutionary origin that serves as a genetic scaffold from which complexity has evolved. In particular, we propose that the transcriptional identity of metazoan male germ cells is built around a relatively small network of ancient gene interactions with an overarching functional impact. These observations provide a unifying genetic basis for the recently reported deep conservation of fundamental germ cell biological processes such as sperm motility and gamete fusion[52,53]. The existence of a shared genetic identity in metazoan spermatocytes can be regarded as a ramification of an ancestral multipotency program already present in germ line precursor cells[54]. Indeed, even in species with divergent germ line segregation strategies, conserved functional interactions at the post-translational level are required for primordial germ cell specification[55,56]. The observation that metazoan chromosomes have deeply conserved gene linkages, some stretching as far back as unicellular organisms, provides a plausible mechanistic basis for the persistence of ancient transcriptional programs[57]. Accordingly, one can speculate that chromatin interactions have been preserved throughout metazoan evolution to ensure both gene regulation and chromosome pairing during meiosis.

It is likely that meiosis emerged as the preferential stage for the acquisition of male germ cell identity due to its underlying connection with sexual reproduction, and its transcriptionally-permissive state resulting from the attenuated compartmentalization of prophase I chromatin[33]. In this regard, it is important to note the conceptual distinction between cell cycle genes involved in the basic control of cellular division (caretakers) and those regulating the process from a developmental perspective (gatekeepers)[58]. Although typically applied in the context of cancer, this distinction can also offer insight into the effects of the cell cycle on cellular identity. More specifically, gatekeepers, by influencing the cooperation between cells and their balance in the

context of multicellular systems, represent a suitable home for cell fate-determining genes. The *RNF113* genes are particularly illustrative examples of gatekeeping functions, as they are involved in cell cycle progression and in the regulation of developmental programs both in reproductive and neuronal tissue[35].

The actual benefit of comparative biology for the identification of new genetic causes of human disease is often a contentious topic[59]. Central to this debate is the moderate success in translating animal data to the clinical setting, coupled with the fact that the vast majority of human genetic variants are not shared with other species[60]. By focusing on the deep evolutionary past of human spermatogenesis, we have identified 79 core functional interactions for male germ cell identity and 183 novel functionally-validated candidate genes. The latter have so far uncovered three new genetic causes of human infertility (*RNF113B*, *HSPA2* and *KPNA2*). Taken together, our results emphasize the often-overlooked contribution of evolutionary history to human disease and illustrate how interdisciplinary research can significantly expand our knowledge of fundamental cellular processes. Accordingly, all the code required for repurposing our analytical pipeline to other cell types and pathologies is available as an open-access resource at https://github.com/rionbr/meionav. These resources will likely contribute to a renewed appreciation of comparative biology in the medical field.

Data and code availability

An interactive online application compiling experimental results on 920 evolutionarily-conserved spermatocyte genes and associated microscopy images (Meiotic Navigator) is available at https://pages.igc.pt/meionav. Computational data and custom R and Python scripts used in the analysis are available from https://github.com/rionbr/meionav. All testicular RNA-Seq data generated in this study (**Sup. Table 1**) were deposited in the European Genome-Phenome Archive (human data) and in the Sequence Read Archive (fruit fly data).

Acknowledgments

The authors wish to thank Gabriel Martins and José Marques from the Imaging Unit of Instituto Gulbenkian de Ciência (IGC, Portugal) for assistance in acquiring and processing microscopy data. We acknowledge the IGC's Genomics Unit for their involvement in preparing the fruit fly

RNA-Seq libraries, Nicole Terwort (University of Münster) for preparing the RNA extractions from human testes, and the Core Facility Genomics of the Medical Faculty of the University of Münster for conducting library preparation and RNA sequencing in samples from individual M1911. Renate Renkawitz-Pohl (Philipps-Universität Marburg, Germany) and Chris Doe (University of Oregon, USA) kindly provided us the *bam*-GAL4 line and an aliquot of the anti-dRNF113 antibody, respectively. We thank Élio Sucena, Raquel Oliveira and Mónica Bettencourt-Dias (all from the IGC), Patrícia Beldade (Faculty of Sciences, University of Lisbon), and Peter Ellis (University of Kent, UK) for insightful discussions. We gratefully acknowledge all men that gave their informed consent to be included in the Male Reproductive Genomics (MERGE) study cohort.

Both the Transgenics and the Vivaria Units of the IGC were supported by Congento (LISBOA-01-0145-FEDER-022170), and co-financed by Fundação para a Ciência e a Tecnologia (FCT) under the Portugal 2020 Agreement (European Regional Development Fund). The Genomic Facility of the IGC was partially supported by projects LISBOA-01-0246-FEDER-000037 Single cell HUB and LISBOA-01-0145-FEDER-016417 (ONEIDA), co-funded by Programa Operacional Regional Lisboa 2020. The MERGE cohort and associated human genetics analyses were funded in the frame of the German Research Foundation Clinical Research Unit ‘Male Germ Cells’ (DFG, CRU326; grants to Frank Tüttelmann). Rion Brattig Correia was supported by a contract within the scope of FCT project PTDC/MEC-AND/30221/2017. Irene Julca was supported by the Singaporean Ministry of Education grant MOE2018-T2-2-053. Chandra Shekhar Misra acknowledges a doctoral fellowship from FCT (PD/BD/114362/2016) under the Plants for Life PhD Program. Luis M. Rocha was partially funded by NIH, National Library of Medicine Grant 1R01LM012832; by a Fulbright Commission fellowship; and by National Science Foundation Research Traineeship “Interdisciplinary Training in Complex Networks and Systems” Grant 1735095. Jörg D. Becker received salary support from FCT (grant CEECIND/03345/2018). Paulo Navarro-Costa was supported by Portuguese national funding through the following FCT project grants: PTDC/MEC-AND/30221/2017 and EXPL/MEC-AND/0676/2021.

Author contributions

PNC, JDB, FT and LMR designed experiments. RBC, JMA, MJW, IJ, DS, CSM, LGG, HS, NS, PP, AN, ASL, JB, MM and PNC performed the experiments. PNC, LMR, FT, MM, RBC, JMA,

MJW, IJ, DS, CSM and SK analyzed the data. PNC conceived the study and wrote the manuscript with input from all authors.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Swanson, W. J. & Vacquier, V. D. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3**, 137–144 (2002).

2. Pitnick, S., Hosken, D. J. & Birkhead, T. R. Sperm morphological diversity. in *Sperm Biology* 69–149 (Elsevier, 2009).

3. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).

4. Waddington, C. H. *The Strategy of the Genes*. (Routledge, 1957).

5. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**, 1593–1599 (2012).

6. Wan, C. *et al.* Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015).

7. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).

8. Sebé-Pedrós, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).

9. Extavour, C. G. M. Evolution of the bilaterian germ line: lineage origin and modulation of specification mechanisms. *Integr. Comp. Biol.* **47**, 770–785 (2007).

10. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20**, 1313–1326 (2010).

11. Julca, I. *et al.* Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. *Nat. Plants* **7**, 1143–1159 (2021).

12. Navarro-Costa, P., Molaro, A., Misra, C. S., Meiklejohn, C. D. & Ellis, P. J. Sex and

suicide: The curious case of Toll-like receptors. *PLOS Biol.* **18**, e3000663 (2020).

13. Clark, A. G. Sperm competition and the maintenance of polymorphism. *Heredity* **88**, 148–153 (2002).

14. Xia, B. *et al.* Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates. *Cell* **180**, 248-262.e21 (2020).

15. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

16. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).

17. Lim, C., Tarayrah, L. & Chen, X. Transcriptional regulation during Drosophila spermatogenesis. *Spermatogenesis* **2**, 158–166 (2012).

18. Goodenough, U. & Heitman, J. Origins of Eukaryotic Sexual Reproduction. *Cold Spring Harb. Perspect. Biol.* **6**, a016154 (2014).

19. Patel, L. *et al.* Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat. Struct. Mol. Biol.* **26**, 164–174 (2019).

20. Maezawa, S. *et al.* Super-enhancer switching drives a burst in gene expression at the mitosis-to-meiosis transition. *Nat. Struct. Mol. Biol.* **27**, 978–988 (2020).

21. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

22. The DREAM Module Identification Challenge Consortium *et al.* Assessment of network module identification across complex diseases. *Nat. Methods* **16**, 843–852 (2019).

23. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).

24. Gates, A. J., Correia, R. B., Wang, X. & Rocha, L. M. The effective graph reveals redundancy, canalization, and control pathways in biochemical regulation and signaling. *Proc. Natl. Acad. Sci.* **118**, e2022598118 (2021).

25. Simas, T., Correia, R. B. & Rocha, L. M. The distance backbone of complex networks. *J. Complex Netw.* **9**, cnab021 (2021).

26. Ren, Y., Ay, A. & Kahveci, T. Shortest path counting in probabilistic biological networks.

BMC Bioinformatics **19**, 465 (2018).

27. Walhout, A. J. M. *et al.* Protein Interaction Mapping in *C. elegans* Using Proteins Involved in Vulval Development. *Science* **287**, 116–122 (2000).

28. Sun, M. G. & Kim, P. M. Evolution of biological interaction networks: from models to real data. *Genome Biol.* **12**, 235 (2011).

29. Wu, Y. *et al.* Transcription Factor RFX2 Is a Key Regulator of Mouse Spermiogenesis. *Sci. Rep.* **6**, 20435 (2016).

30. Liu, S. *et al.* Chromodomain Protein CDYL Acts as a Crotonyl-CoA Hydratase to Regulate Histone Crotonylation and Spermatogenesis. *Mol. Cell* **67**, 853-866.e5 (2017).

31. Shah, C. *et al.* Widespread Presence of Human BOULE Homologs among Animals and Conservation of Their Ancient Reproductive Function. *PLoS Genet.* **6**, e1001022 (2010).

32. Guo, J. *et al.* The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141–1157 (2018).

33. Alavattam, K. G. *et al.* Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development. *Nat. Struct. Mol. Biol.* **26**, 175–184 (2019).

34. White-Cooper, H. Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the Drosophila testis. *Spermatogenesis* **2**, 11–22 (2012).

35. Carney, T. D., Struck, A. J. & Doe, C. Q. *midlife crisis* encodes a conserved zinc-finger protein required to maintain neuronal differentiation in *Drosophila*. *Development* **140**, 4155–4164 (2013).

36. Szcześniak, M. W., Ciomborowska, J., Nowak, W., Rogozin, I. B. & Makałowska, I. Primate and Rodent Specific Intron Gains and the Origin of Retrogenes with Splice Variants. *Mol. Biol. Evol.* **28**, 33–37 (2011).

37. Di Persio, S. *et al.* Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis. *Cell Rep. Med.* **2**, 100395 (2021).

38. Gatti da Silva, G. H., Jurica, M. S., Chagas da Cunha, J. P., Oliveira, C. C. & Coltri, P. P. Human RNF113A participates of pre-mRNA splicing in vitro. *J. Cell. Biochem.* **120**, 8764–8774 (2019).

39. Houston, B. J. *et al.* A systematic review of the validated monogenic causes of human

male infertility: 2020 update and a discussion of emerging gene–disease relationships. *Hum. Reprod. Update* **28**, 15–29 (2021).

40. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17**, 9–18 (2016).

41. Agarwal, A., Mulgund, A., Hamada, A. & Chyatte, M. R. A unique view on male infertility around the globe. *Reprod. Biol. Endocrinol.* **13**, 37 (2015).

42. Wyrwoll, M. J. *et al.* Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. *Am. J. Hum. Genet.* **107**, 342–351 (2020).

43. Eddy, E. M. Role of heat shock protein HSP70-2 in spermatogenesis. *Rev. Reprod.* **4**, 23–30 (1999).

44. Azuma, M., Ogata, T., Yamazoe, K., Tanaka, Y. & Inoue, Y. H. Heat shock cognate 70 genes contribute to Drosophila spermatocyte growth progression possibly through the insulin signaling pathway. *Dev. Growth Differ.* **63**, 231–248 (2021).

45. Dix, D. J. & Poorman-Allenii, P. Targeted gene disruption of Hsp70-2 results in failed meiosis, germ cell apoptosis, and male infertility. *Dev. Biol.* **93**, 3264–3268 (1996).

46. Ly-Huynh, J. D. *et al.* Importin Alpha2-Interacting Proteins with Nuclear Roles During Mammalian Spermatogenesis. *Biol. Reprod.* **85**, 1191–1202 (2011).

47. Leu, J.-Y., Chua, P. R. & Roeder, G. S. The Meiosis-Specific Hop2 Protein of S. cerevisiae Ensures Synapsis between Homologous Chromosomes. *Cell* **94**, 375–386 (1998).

48. Petukhova, G. V., Romanienko, P. J. & Camerini-Otero, R. D. The Hop2 Protein Has a Direct Role in Promoting Interhomolog Interactions during Mouse Meiosis. *Dev. Cell* **5**, 927–936 (2003).

49. John, A., Vinayan, K. & Varghese, J. Achiasmy: Male Fruit Flies Are Not Ready to Mix. *Front. Cell Dev. Biol.* **4**, Article 75 (2016).

50. Mason, D. A., Fleming, R. J. & Goldfarb, D. S. Drosophila melanogaster Importin α1 and α3 Can Replace Importin α2 During Spermatogenesis but Not Oogenesis. *Genetics* **161**, 157–170 (2001).

51. Oud, M. S. *et al.* A de novo paradigm for male infertility. *Nat. Commun.* **13**, 154 (2022).

52. Speer, K. F., Allen-Waller, L., Novikov, D. R. & Barott, K. L. Molecular mechanisms of

sperm motility are conserved in an early-branching metazoan. *Proc. Natl. Acad. Sci.* **118**, e2109993118 (2021).

53. Moi, D. *et al.* Archaeal origins of gamete fusion. *bioRxiv* (2021).

54. Fierro-Constaín, L. *et al.* The Conservation of the Germline Multipotency Program, from Sponges to Vertebrates: A Stepping Stone to Understanding the Somatic and Germline Origins. *Genome Biol. Evol.* evw289 (2017).

55. Perillo, M., Swartz, S. Z. & Wessel, G. M. A conserved node in the regulation of Vasa between an induced and an inherited program of primordial germ cell specification. *Dev. Biol.* **482**, 28–33 (2022).

56. Colonnetta, M. M. *et al.* Preformation and epigenesis converge to specify primordial germ cell fate in the early Drosophila embryo. *PLOS Genet.* **18**, e1010002 (2022).

57. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**, eabi5884 (2022).

58. Domazet-Lošo, T. & Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* **8**, 66 (2010).

59. Pound, P. & Ritskes-Hoitinga, M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *J. Transl. Med.* **16**, 304 (2018).

60. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).

Figure legends

Figure 1. The complexity of the male germ cell transcriptome has an old evolutionary origin. a- Overview of the experimental strategy. **b-** Male germ cells have a higher transcriptome complexity than representative somatic cell types. Human post-meiotic cells are a noteworthy exception. TPM: transcripts per million; see **Sup. Table 1** for information on RNA-Seq datasets. **c-** Clade tree for mapping the time of origin of genes in the three selected species: human (Primata), mouse (Rodentia), and fruit fly (Diptera). Genes assigned to phylostrata 1-5 are

common to all metazoan species. Mya: million years ago; see **Sup. Fig.1** for the list of representative species of each clade and number of genes in each phylostratum. **d-** The majority of genes expressed by male germ cells are common to all Metazoa (phylostrata 1-5, green outline). This fraction is similar to that found in representative somatic cell types of each selected species. Minimum average expression cut-off: TPM >1. ns- no significant difference (p >0.3472; Mann–Whitney U test). Sg.: Spermatogonia, Sc.: Spermatocytes, St.: Spermatids, En.: Enterocytes, Ne.: Neurons and Ms.: Muscle. **e-** Post-meiotic male germ cells have younger transcriptomes than meiotic and pre-meiotic cells. Transcriptome age indices (TAIs) are split between the different phylostrata. **f-** Male germ cells have large, structured meiotic networks. Graphs represent the largest connected component of the spermatocyte transcriptome (minimum average expression cut-off: TPM >1) according to STRING functional association scores. Gene conservation (across all Metazoa) was defined based on eggNOG orthogroups. Networks were filtered to only include edges with combined scores ≥0.5 (see **Sup. Fig.3**). **g-** Meiotic networks contain a substantial number of conserved genes. **h-** Conserved genes (red) are more connected than non-conserved genes (blue) in the meiotic networks. ****p <0.0001 (Kolmogorov-Smirnov test). **i-** Conserved genes (red) have more connected interactors than non-conserved genes (blue). ****p <0.0001 (Kolmogorov-Smirnov test). **j-** Machine-learning algorithms reliably predict evolutionary conservation based solely on meiotic network features. Values correspond to AUC (area under the curve) scores. “Coin toss” corresponds to a random classification. Four-fold cross-validation results are shown. ROC: receiver operating characteristic curves; SVM: support-vector machine; see **Sup. Fig.5** for precision and recall curves.

Figure 2. The orthoBackbone uncovers core features of male germ cell identity. a- The orthoBackbone methodology. First, the most relevant associations are determined by defining the network metric backbone (based on shortest paths). Of the backbone edges (in green), those connecting the same orthologous genes in different species are selected as part of the evolutionarily-conserved orthoBackbone (in red, with asterisks). In case of a one-to-many conserved edge relationship (due to the presence of paralogs), inclusion depends on at least one of the multiple edges being part of the backbone. Letters depict different genes, B, B’ and B’’ correspond to paralogs, and numbers indicate distances between genes. **b-** The meiotic

orthoBackbone represents less than 3% of all functional interactions (edges) in the meiotic networks. **c-** The orthoBackbone connects >70% of all conserved genes expressed in spermatocytes. Gene conservation (across all Metazoa) was defined based on eggNOG orthogroups. **d-** orthoBackbone genes are preferentially involved in gene expression regulation. Charts represent the top 5 terms of an unfiltered gene ontology (GO) enrichment analysis for biological processes of the human meiotic orthoBackbone. False discovery rate ≤0.05; see **Sup. Fig.6** for the expanded GO analyses. **e-** The orthoBackbone reveals an ancient gene regulatory module of male germ cell identity consisting of 79 functional interactions. Post-transc. reg.: Post-transcriptional regulation; RNA mod.: RNA modification. **f-** Conserved meiotic differentially expressed genes (DEGs) were defined based on their upregulation at mammalian meiotic entry and/or downregulation at meiotic exit. In both cases, genes also had to be expressed in insect spermatogenesis. Green lines link orthologs (920 in fruit flies, 797 in humans and 850 in mice) based on eggNOG orthogroups. Expression level in normalized absolute log(FPKM+1). **g-** An *in vivo* RNAi screen in fruit fly testes uncovers the functional requirement of 250 meiotic DEGs (27.2%) for male reproductive fitness. Silencing of the 920 conserved meiotic DEGs was induced prior to meiotic entry using the *bam*-GAL4 driver. Color-code for the recorded testicular phenotype as in “h”. Results reflect a total of four independent experiments. Threshold for impaired reproductive fitness (red horizontal line) corresponds to a 75% fertility rate (>2 standard deviations of the mean observed in negative controls). **h-** Conserved meiotic DEGs are required for diverse spermatogenic processes. Testicular phenotypes of the 250 hits were defined by phase-contrast microscopy and assigned to five classes based on the earliest manifestation of the phenotype. **i-** Meiotic DEGs reveal 164 new, evolutionarily-conserved spermatogenesis genes (65.6% of all hits, homologous to 183 and 190 in humans and mice, respectively). Phenotype novelty was defined by lack of previously published evidence of a role in male fertility / spermatogenesis in humans, mice or fruit flies. **j-** All data acquired in this screen are freely available in the form of an open-access gene browser (Meiotic Navigator).

Figure 3. Ancient origins of human infertility. a- Similar domain structure of the RNF113 proteins in humans (RNF113B) and fruit flies (dRNF113). Both contain a C3H1-type zinc finger and a RING finger domain. The effect of the *RNF113B* loss of function (LoF) variant identified in

individual M1911 is represented (frameshift). Numbers indicate amino acid residue position. **b-** *RNF113B* is required for meiotic progression past the primary spermatocyte stage. Testicular histology of M1911 (*RNF113B* LoF variant) and of a control sample with normal spermatogenesis. Arrowheads: primary spermatocytes; arrows: spermatids. Scale bars: 100 μm (overview), 50 μm (insets), and 10 μm (meiotic region). For quantification, testicular tubules were analyzed for the most advanced germ cell stage present: spermatogonia (yellow bars), spermatocytes (green), spermatids (blue), and Sertoli cells / tubular shadows (grey). C: control; V: *RNF113B* LoF variant. l and r indicate left and right testis, respectively. **c-** The silencing of fruit fly *dRNF113* also results in meiotic arrest. Phase-contrast microscopy. Arrowheads: primary spermatocytes; asterisks: early (round) spermatids; arrows: late (elongating) spermatids; sv- seminal vesicle. Scale bars: 50 μm (whole testis) and 20 μm (meiotic region). Meiotic area indicates the fraction of the entire testis occupied by primary spermatocytes. Two independent RNAi reagents were used. ****p <0.0001 (unpaired t-test). **d-** *RNF113B* is predominantly expressed at meiotic entry. Data analyzed from our previously published single cell RNA-Seq atlas of normal spermatogenesis[37]. Usg.: undifferentiated spermatogonia; Dsg.: differentiated spermatogonia / preleptotene; Lp.: leptotene; Zg.: zygotene; Pc.: pachytene; Dp.: diplotene; M: meiotic divisions; Rsp.: round spermatids; Esp.: elongating spermatids. **e-** The nuclear levels of the dRNF113 protein increase at meiotic entry. Images are maximum projections of the entire nuclear volume. Spermatocytes correspond to late prophase I cells. Dotted lines delimit the nuclear envelope (as assessed by fluorescent wheat germ agglutinin). Scale bar: 5 μm. a.u- arbitrary units. ****p <0.0001 (unpaired t-test). **f-** *RNF113B* is required for normal gene expression during spermatogenesis. Differential gene expression (DGE) analysis of RNA-Seq data obtained from testicular biopsies of M1911 (*RNF113B* LoF variant, left and right testis) and of three controls with normal spermatogenesis. Down and upregulated genes in blue and red, respectively. orthoBackbone differentially expressed genes (DEGs) are outlined. FC: fold change. FDR: false discovery rate. Edge disruption corresponds to the number of orthoBackbone edges containing at least one DEG. **g-** *dRNF113* also regulates gene expression in fruit fly gonads. Whole testes samples (in triplicate) in both experimental conditions. **h-** Network of functional associations between orthoBackbone genes downregulated both in the *RNF113B* LoF and dRNF113 RNAi. Node size indicates result of the page rank metric in the meiotic network (measure of the connectivity of

interacting genes), and color specifies if the gene has a known role in spermatogenesis (in any species). Testicular phenotype of men affected by variants in *HSPA2* and *KPNA2* (red nodes) are depicted in **i**. Edge thickness indicates STRING functional association scores and color specifies main source of data for the associations. **i-** LoF variants in the orthoBackbone genes *HSPA2* and *KPNA2* associate with human male infertility. Testicular histology of individuals M2190 and M2098 (*HSPA2* and *KPNA2* variants, respectively) reveals a complete loss of germ cells (Sertoli cell-only phenotype). Arrowheads: Sertoli cells. Scale bars: 100 μm (overview), 50 μm (insets), and 10 μm (intratubular region).

Methods

Phylostratigraphic analysis

For humans, mice and fruit flies, published RNA-Seq data of different stages of spermatogenesis and of representative cell types of the three primary embryonic layers (**Sup. Table 1**) were downloaded from the Sequence Read Archive. Data were checked for quality control and preprocessed by trimming adaptor sequences (Trim galore!; v0.5.0; Babraham Bioinformatics). Gene expression was quantified as transcripts per million (TPM) using Salmon v0.14.1[1]. Orthogroups (set of homologous genes derived from the last common ancestor) were defined based on the proteomes of 25 species representing key phylogenetic positions (**Sup. Fig. 1**), and were assembled using OrthoFinder v2.4.0[2], with DIAMOND v0.9.24.125 as aligner (default settings)[3]. A species tree reflecting the current consensus of the eukaryotic phylogeny[4,5,6] served as basis for the phylostratigraphic analysis (**Sup. Fig. 1**). Each orthogroup was assigned to a phylostratum (node) by identifying the oldest clade found in the orthogroup[7] using ETE v3.0 with the “get_common_ancestor” option. Phylostrata were assigned a node number, ranging from 1 (oldest) to 16 (youngest, species-specific). Genes with a minimum average TPM >1 were considered expressed. After mapping all expressed genes to the phylostratum containing their corresponding orthogroup, the distribution of the transcriptome allocated to phylostrata 1 to 5 (orthogroups common to all metazoa) was compared, in a pairwise manner, between germ cells and soma using the Mann–Whitney U test as implemented in the SciPy python package[8]. The transcriptome age index (TAI) of the germ cell samples was calculated by dividing the product of each gene’s TPM value and node number by the sum of all TPM values[9]. Higher TAI values represent younger transcriptomes.

Network construction

To assemble the meiotic networks, previously published RNA-Seq datasets of purified human and mouse spermatocytes and of the spermatocyte-enriched median region of the fruit fly testes were used to identify all expressed genes (see **Sup. Table 1** for the list of all selected datasets). Salmon v0.14.1 was used to quantify gene expression levels, and genes were considered expressed based on a minimum average expression level of TPM >1. All possible interactions

between expressed genes were retrieved from the STRING v11 protein-protein interactions (PPI) database[10]. To select for high-confidence interactions, only PPI with a STRING combined score ≥0.5 were included (**Sup. Fig. 3**). The transcriptome networks of human, mouse, and fruit fly spermatocytes were then constructed from the set of expressed genes and the corresponding significant STRING PPI (**Fig. 1f**). These networks were represented as weighted graphs, where edge weights (given by the STRING PPI scores) denote the probability of the connected genes interacting and thus jointly contributing to a specific function. A multilayer network representation was employed to define orthologous relationships, with each layer corresponding to a particular species and nodes belonging to the same orthogroup being connected across layers. Orthologous connections were established using the EggNOG v5 database at the metazoan level (meNOG)[11] and a one-to-many relationship was employed when identifying orthologs between network layers.

OrthoBackbone computation

The initial step to define the meiotic OrthoBackbone was the extraction of each species' spermatocyte network metric backbone (i.e., each layer of the multilayer network). The metric backbone is the subgraph that is sufficient to compute all shortest paths in the network, thus removing edges that break the triangle inequality (and are therefore redundant in regards to the shortest paths). This network retains all metric edges and preserves all the nodes in the original network[12,13]. Considering that a set of the PPI in each layer are evolutionarily-conserved, we developed the orthoBackbone: a single shared subnetwork obtained by collapsing the metric backbone of all layers according to the orthologous relationships. More precisely, the orthoBackbone corresponds to a subgraph of the metric backbone of every layer, where each edge has an analogous edge connecting the same orthologous genes in all other species' layers. For cases where there was not a one-to-one conserved edge relationship (i.e., an edge of one layer's backbone mapped to several edges in other species' layers due to the existence of paralogs), the inclusion criterion was for at least one of these homologous edges to be part of the backbone (**Fig. 2a**). Notice that because only edges between ortholog genes can be part of the orthoBackbone, nodes (including ortholog nodes) that are left with no edges are removed. Thus, the orthoBackbone typically has fewer nodes than its original network layer and it is not

necessarily itself a metric backbone. Next, annotations retrieved from the Gene Ontology resource database (release 2020-09-10) were used to select all human orthoBackbone edges connecting at least one spermatogenesis-related gene linked to gene expression regulation. This led to the identification of the 79 functional interactions in spermatocytes that form an ancient gene regulatory module of male germ cell identity (**Fig. 2e**).

Differential gene expression analysis of the meiotic transitions

Reads were aligned to their respective genomes (GRCh38, GRCm38 and BDGP6.22) using HISAT2 v2.1.0 under default parameters[14]. Uniquely mapped read counts were generated using FeatureCounts v1.5.0-p1 and ENSEMBL GTF annotations[15]. Differential gene expression analysis was performed using a likelihood test[16], as implemented in the edgeR package[17]. Genes that were significantly upregulated at meiotic entry (spermatocyte *vs.* spermatogonia) and/or downregulated at meiotic exit (spermatid *vs.* spermatocyte) were selected for further analysis. Three criteria were employed to define differentially expressed genes (DEGs): false discovery rate ≤ 0.05; abs[$\log_2$(fold change)] ≥ 1; and average $\log_2$(normalized counts per million) ≥ 1. Subsequently, EggNOG orthogroups (defined at the metazoan level)[11] were used to establish which genes had a similar up/downregulation behavior in both species. This list was further trimmed by only retaining those whose fruit fly orthologs were also expressed at average $\log_2$(normalized counts per million) >1 in the corresponding testes, leading to 920 fruit fly genes (homologous to 797 and 850 in humans and mice, respectively).

***Drosophila in vivo* RNAi**

An *in vivo* UAS-GAL4 system was used to silence the 920 conserved meiotic DEGs specifically as male germ cells prepare to enter meiosis[18]. Silencing was induced using the *bam*-GAL4 driver (kindly provided by Renate Renkawitz-Pohl, Philipps Universität Marburg, Germany), which promotes high levels of shRNA expression in late spermatogonia and early primary spermatocytes[19]. UAS-hairpin lines targeting the selected genes were purchased from the Bloomington Drosophila Stock Centre (BDSC) and the Vienna Drosophila Resource Centre (VDRC). Lines previously associated with a phenotype in the literature, regardless of the tissue, were preferentially chosen for this experiment. In the ten cases where no lines were available,

these were generated in-house following standard procedures[20]. Briefly, shRNA were designed using DSIR[21], with the corresponding sequences being cloned into the pWalium 20 vector. Constructs were injected into fruit flies carrying an attp site on the third chromosome (BDSC stock #24749) and transgene-carrying progeny were selected to establish the UAS-hairpin lines. A similar strategy was employed to generate a second independent RNAi reagent for dRNF113 (dRNF113 #2). The antisense sequences selected for each locus are listed in **Sup. Table 3**, and information on all tested lines is available at the Meiotic Navigator gene browser.

For assessing fertility, gene-silenced males were mated with wild-type Oregon-R virgin females (2 males:4 females) for 12 hours at 25ºC. Laid eggs were left to develop for 24 hours at 25ºC before the percentage of egg hatching was determined. This percentage (fertility rate) served as measure of the male reproductive fitness associated with each tested gene. All fruit flies were 3 to 7 days post-eclosion and fertility rates correspond to an average of four independent experiments, with a minimum of 25 eggs scored per replicate. Every batch of experiments included a negative control [RNAi against the mCherry fluorophore (BDSC stock #35785) - a sequence absent in the fruit fly genome], and a positive control [RNAi against Ribosomal protein L3 (BDSC stock #36596) - an essential unit of the ribosome]. A cut-off of <75% fertility rate was established to define impaired reproductive fitness based on the rate observed in the negative controls (>2 standard deviations of 91.6 $\pm$ 8.5%). For dRNF113, Hsc70-1 (VDRC stock #106510) and dKPNA2 (VDRC stock #34265) extended fertility tests were conducted, with ~100 eggs scored per replicate. To this end, egg-laying cages with apple juice agar plates as substrate were set-up with a 1:2 male to female ratio. All *Drosophila* lines were maintained at 25ºC in polypropylene bottles containing enriched medium (cornmeal, molasses, yeast, beet syrup and soy flour).

The novelty of the hits was determined based on the following exclusion criteria (applied to all 250 hits and their orthologs): i- clinically-validated human infertility genes[22]; ii- genes associated with a mouse male infertility phenotype (MP:0001925) in the Mouse Genome Informatics database[23]; and iii- fruit fly genes with an associated male germ cell or male reproductive phenotype in the FlyBase database[24]. For the latter two categories, information reflects data available in both systems as of February 2022.

Human testicular imaging

Testicular biopsies of the azoospermic individuals M1911 (*RNF113B* variant), M2098 (*KPNA2* variant) and M2190 (*HSPA2* variant) were collected at the Centre for Reproductive Medicine and Andrology (CeRA) in Münster, Germany or at the University Hospital Giessen, Germany. As a control for testicular imaging, tissue of individual M2951 (who underwent a vasectomy reversal at the CeRA) was used. All testicular tissues were fixed in Bouin's solution and embedded in paraffin using an automatic ethanol and paraffin row (Bavimed Laborgeräte GmbH). Serial sections (5µm thick) were stained with periodic acid-Schiff according to standard procedures. Testicular histopathology was performed as part of the routine clinical work up. For quantification, at least 100 testicular tubules per testis were screened for the most advanced germ cell type present (categories: spermatogonia, spermatocytes, round spermatids and elongating spermatids), or in their absence, for Sertoli cells or tubular shadows.

***Drosophila* testicular imaging**

Squash preparations of freshly-dissected Drosophila testes were performed as previously described[25] and examined using a phase contrast microscope (Nikon Eclipse E400). Phenotypes of all silenced genes associated with decreased reproductive fitness (250 in total) were assigned to one of five classes, based on the earliest stage in which the cytological defects were detected: 1- pre-meiotic (abnormal late spermatogonia); 2- meiotic (failure to successfully progress through meiosis); 3- post-meiotic (abnormal spermiogenesis); 4- gamete (low mature sperm numbers and/or motility); 5- undetectable (no cytological defects). Images from at least 2 pairs of testes were acquired per genotype and were corrected for background illumination as previously described[26]. These are all available in the Meiotic Navigator gene browser.

For the quantification of the meiotic area, phase-contrast images of different testicular regions were acquired at 40x and stitched together, using the MosaicJ tool[27] in the ImageJ software (v1.8, National Institutes of Health), to reconstruct a high-resolution image of the entire testis. On average, 8 individual images were acquired to assemble each complete testis. The meiotic area corresponds to the ratio between the area (in pixels) occupied by primary spermatocytes and that of the entire gonad. A total of 15 different testes were quantified per genotype across 3 independent experiments, and groups were compared using unpaired t-tests.

For the dRNF113 immunofluorescence assay, a modified whole-mount protocol specifically designed for the analysis of *Drosophila* spermatocytes was employed[28]. Briefly, testes were dissected in testis buffer (183mM KCl, 47mM NaCl, 10mM Tris-HCl, 1mM EDTA and 1mM PMSF), transferred to a pre-fix solution containing 4% formaldehyde (PolySciences) in PBS, and then fixed for 20min using a heptane-fixative mix at 3:1. The fixative consisted of 4% formaldehyde in a PBS + 0.5% NP-40 (Merck) solution. After washing the fixative, samples were incubated for 1h in PBS supplemented with 0.3% Triton X-100 (Sigma-Aldrich), 1% (w/v) bovine serum albumin (BLIRT) and 1% (w/v) donkey serum (Thomas Scientific). Primary antibody incubation (anti-dRNF113 at 1:250; kindly provided by Chris Doe, University of Oregon, USA) was performed overnight at 4°C in PBS supplemented with 1% (w/v) bovine serum albumin and 1% (w/v) donkey serum. After washing the primary antibody solution, samples were incubated for 1h with a goat anti-guinea pig secondary antibody (Alexa Fluor 488 conjugate; 1:1000; Invitrogen) and wheat germ agglutinin to label the nuclear envelope (Alexa Fluor 647 conjugate; 1:500; Invitrogen). Testes were mounted in Vectashield mounting medium with DAPI (Vector Laboratories) and the entire nuclear volume of individual cells was acquired as 1µm-thick slices using a 63x oil immersion objective and 10x digital zoom in a Leica SP5 confocal microscope. Slices were stacked into maximum intensity Z-projections and the relative intensity of the dRNF133 signal was measured as previously described[29]. A total of 30 late spermatogonia and 30 mature primary spermatocytes (late prophase I) were quantified across 3 independent experiments, and the two groups were compared using an unpaired t-test.

RNA-Seq in human and *Drosophila* testes

Total RNA from two snap-frozen testicular tissue samples of the azoospermic individual M1911 (*RNF113B* variant) as well as from one sample each from three unrelated individuals with normal spermatogenesis [M1544 (obstructive azoospermia), M2224 (anorgasmia) and M2234 (previous vasectomy)[30]] were extracted using the Direct-zol RNA Microprep kit (Zymo Research), following the manufacturer's instructions. RNA quality was estimated by electrophoresis (Agilent Technologies), with all samples having a RNA integrity number (RIN) >4.5 (range: 4.5 - 5.6), except one replicate of M1911 (with RIN= 2.0). rRNA depletion was performed with the NEBNext rRNA Depletion Kit v2 (New England Biolabs), and total RNA libraries were prepared with the

NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs), using 700ng of RNA per sample. Paired-end sequencing with 150bp per read was performed using the NextSeq2000 system (Illumina) at the University of Münster, Germany. An average of 44 million reads were generated per sample. The expression pattern of *RNF113B* was determined in a control testis by retrieving the corresponding data from our recently published single-cell RNA-Seq dataset of normal human spermatogenesis[31].

For fruit flies, RNA was extracted from 40 pairs of adult testes per sample per condition (3 to 7 days post-eclosion) using the PureLink RNA Mini Kit (Thermo Fisher Scientific) following the manufacturer's instructions. Three conditions were analyzed: dRNF113 RNAi (meiotic arrest), Prp19 RNAi (meiotic arrest) and the mCherry RNAi negative control (normal spermatogenesis), with three independent biological replicates per condition. Extracted RNA was treated with DNAse (Thermo Fisher Scientific), with a RNA quality number of 10 for all samples, as assessed by electrophoresis (Advanced Analytical Technologies). Total RNA libraries were prepared using the Zymo-Seq RiboFree Total RNA-Seq Library Kit (Zymo Research), using 1000ng of RNA per sample. Paired-end sequencing with 150bp per read was performed using DNBSEQ technology (BGI Group), with an average of 46 million reads per sample. Confirmation of knockdown efficiency via quantitative RT-PCR was performed for both the dRNF113 and Prp19 RNAi. For this, cDNA was synthetized using the Transcriptor First Strand cDNA Synthesis Kit (Roche), according to the manufacturer's instructions. Reactions were performed in a QuantStudio 6 Real-Time PCR System (Thermo Fisher Scientific), using SYBR Green chemistry (Applied Biosystems). All reactions were set up in triplicates and Act57B and RpL32 were used as internal controls. Primers are listed in **Sup. Table 3**.

RNA-Seq reads were aligned against the human and fruit fly genomes (GRCh38 and BDGP6.32, respectively) using the STAR aligner[32]. Gene-level counts were obtained using feature-counts, taking into account the strandedness. Counts were normalized with the TMM method[33] and differential gene expression analysis was performed using a quasi-likelihood F-test[34], as implemented in the edgeR R package. DEGs were defined using the previously listed criteria (see "Differential gene expression analysis of the meiotic transitions").

Whole exome sequencing and orthoBackbone analysis

Whole exome sequencing (WES) in the 1,021 azoospermic men included in the MERGE study cohort was performed as previously described[35]. All men provided written informed consent, in agreement with local requirements. The study protocol was approved by the Münster Ethics Committees/Institutional Review Boards (Ref. No. Münster: 2010-578-f-S) in accordance with the Helsinki Declaration of 1975.

Exome sequencing data were filtered for high-confidence loss-of-function (LoF) variants in the 27 orthoBackbone genes that were similarly downregulated in the testicular transcriptome of M1911 and of the dRNF113 RNAi, and for which there were available functional data supporting a role in spermatogenesis. These genes were: *BAZ1A*, *BOLL*, *CDC14A*, *CFAP61*, *CFAP91*, *CNOT6*, *CNOT7*, *CUL3*, *DHX36*, *DNAH7*, *DNAH8*, *DNAI1*, *DNAI2*, *DRC7*, *EIF2A*, *HSPA2*, *KPNA2*, *LIPE*, *MTMR7*, *PPP1R2*, *PSMF1*, *SKP1*, *SRPK1*, *TOB1*, *TPGS2*, *TUBA4A* and *UBE2K*. From this list, all those previously associated with known causes of human infertility or with congenital syndromic conditions were excluded (*CFAP61*, *CFAP91*, *DNAH7*, *DNAH8*, *DNAI1*, *DNAI2* and *LIPE*). The inheritance mode for each candidate gene was predicted using the DOMINO algorithm[36] and the gnomAD observed/expected (o/e) constraint score for LoF variants[37]. Autosomal-dominant inheritance was only assumed when consistently predicted by DOMINO and by an o/e score <0.35. Filtering criteria were: stop-gain, frameshift and splice site variants with a minor allele frequency (MAF) <0.01 in gnomAD, a maximum occurrence of 10x in our in-house database and a read depth >10. Variants were only considered when detected in accordance with the predicted mode of inheritance.

References

1. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

2. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

3. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

4. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree

of life. *Nature* **452**, 745–749 (2008).

5. Fairclough, S. R. *et al.* Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome Biol.* **14**, R15 (2013).

6. Laumer, C. E. *et al.* Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B Biol. Sci.* **286**, 20190831 (2019).

7. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

8. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

9. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815–818 (2010).

10. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

11. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

12. Simas, T., Correia, R. B. & Rocha, L. M. The distance backbone of complex networks. *J. Complex Netw.* **9**, cnab021 (2021).

13. Correia, R. B., Barrat, A. & Rocha, L. M. The metric backbone preserves community structure and is a primary transmission subgraph in contact networks. *bioRxiv* (2022).

14. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

15. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma. Oxf. Engl.* **30**, 923–930 (2014).

16. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

17. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* **26**, 139–140 (2010).

18. Brand, A. H. & Perrimon, N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Dev. Camb. Engl.* **118**, 401–415 (1993).

19. White-Cooper, H. Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the Drosophila testis. *Spermatogenesis* **2**, 11–22 (2012).

20. Perkins, L. A. *et al.* The Transgenic RNAi Project at Harvard Medical School: Resources and Validation. *Genetics* **201**, 843–852 (2015).

21. Vert, J.-P., Foveau, N., Lajaunie, C. & Vandenbrouck, Y. An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* **7**, 520 (2006).

22. Houston, B. J. *et al.* A systematic review of the validated monogenic causes of human male infertility: 2020 update and a discussion of emerging gene–disease relationships. *Hum. Reprod. Update* **28**, 15–29 (2021).

23. Baldarelli, R. M. *et al.* The mouse Gene Expression Database (GXD): 2021 update. *Nucleic Acids Res.* **49**, D924–D931 (2021).

24. Larkin, A. *et al.* FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.* **49**, D899–D907 (2021).

25. Bonaccorsi, S., Giansanti, M. G., Cenci, G. & Gatti, M. Preparation of live testis squashes in Drosophila. *Cold Spring Harb. Protoc.* **2011**, prot5577 (2011).

26. Landini, G. Background illumination correction – Novel context-based segmentation algorithms for intelligent microscopy. https://blog.bham.ac.uk/intellimic/g-landini-software/background-illumination-correction/ (2020).

27. Thévenaz, P. & Unser, M. User-friendly semiautomated assembly of accurate image mosaics in microscopy. *Microsc. Res. Tech.* **70**, 135–146 (2007).

28. Raich, N., Contremoulins, V. & Karess, R. E. Immunostaining of Whole-Mount Drosophila Testes for 3D Confocal Analysis of Large Spermatocytes. *J. Vis. Exp. JoVE* (2020).

29. Prudêncio, P. *et al.* The Trithorax group protein dMLL3/4 instructs the assembly of the zygotic genome at fertilization. *EMBO Rep.* **19**, e45728 (2018).

30. Siebert-Kuss, L. M. *et al.* Stage-specific gene and transcript dynamics in human male germ cells. *bioRxiv* (2022).

31. Di Persio, S. *et al.* Single-cell RNA-seq unravels alterations of the human spermatogonial stem cell compartment in patients with impaired spermatogenesis. *Cell Rep. Med.* **2**, 100395 (2021).

32. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

33. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

34. Lun, A. T. L., Chen, Y. & Smyth, G. K. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. *Methods Mol. Biol. Clifton NJ* **1418**, 391–416 (2016).

35. Wyrwoll, M. J. *et al.* Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. *Am. J. Hum. Genet.* **107**, 342–351 (2020).

36. Quinodoz, M. *et al.* DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. *Am. J. Hum. Genet.* **101**, 623–629 (2017).

37. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Supplementary Information and Figures

Clinical genetics

The homozygous frameshift variant c.556_565del;p.(Thr186GlyfsTer119) in *RNF113B* in individual M1911 was previously identified by filtering, for rare LoF variants (see **Methods** for criteria), whole exome sequencing data of 74 men with complete bilateral meiotic arrest. The variants in the *HSPA2* and *KPNA2* genes (2 cases for each, for a total of four unrelated men) were identified by analyzing data from 1,021 azoospermic men from the MERGE cohort. These were predicted to follow an autosomal-dominant inheritance according to DOMINO and to o/e scores of 0.32 and 0.21, respectively. For *HSPA2*, the heterozygous stop-gain variant c.175C>T;p.(Gln59Ter) was identified in individual M1678, and the heterozygous frameshift variant c.1806dup;p.(Glu603ArgfsTer81) in M2190. For *KPNA2*, the heterozygous splice acceptor variant c.667-2A>G;p.? was detected in individual M1645, and the heterozygous frameshift variant c.1158_1161del;p.(Ser387ArgfsTer14) in M2098.

To rule out possible alternative monogenic causes for the infertility phenotype in all 5 selected individuals, the sequences of 230 genes previously identified with an at least limited level of evidence of being associated with male infertility[1] were screened in the exomes of M1911 (*RNF113B* variant), M1678 (*HSPA2* variant), M2190 (*HSPA2* variant), M1645 (*KPNA2* variant) and M2098 (*KPNA2* variant; see **Methods** for detailed criteria). Of these variants, only those affecting genes with a quantitative impact on spermatogenesis, that were consistent with the reported mode of inheritance of the respective gene, and occurring <10x in our in-house database were considered for further analysis. These criteria identified the heterozygous missense variant c.2377G>T;p.(Ala793Ser) in *DNMT1* as a possible alternative cause for male infertility in M1911, a possibility that was disproven by subsequent familial studies. Briefly, DNA from M1911's mother, a fertile brother and an azoospermic brother were available for segregation analysis, which was performed by Sanger sequencing. The azoospermic brother of M1911 was also homozygous for the frameshift variant in *RNF113B*, while the mother and the fertile brother were heterozygous for the variant. Of note, the missense variant c.2377G>T p.(Ala793Ser) in *DNMT1* was not detected neither in the mother nor in the infertile brother of M1911, but only in the fertile brother in a heterozygous state (primers are listed in **Sup. Table 3**). The latter result strongly suggests that

the *DNMT1* variant has no pathogenic effect on male fertility. Segregation analysis was also performed in M1654's parents, revealing that the splice-site variant c.667-2A>G;p.? was not present neither in his father nor in his mother (i.e., indicating that it was a *de novo* variant). Paternity was confirmed by short-tandem-repeat analysis. No chromosomal aberrations or Y-chromosomal microdeletions in the AZF regions were identified in any of the selected individuals.

Clinical data of individuals M1678 and M2190 (*HSPA2* variants)

The testicular phenotype of M1678 is unknown, as this individual did not undergo a testicular biopsy. However, FSH levels of 26.7 U/L (normal range 1-7 U/L) and a bi-testicular volume of 10 mL (normal value >24 mL) are clearly indicative of non-obstructive azoospermia. M2190 underwent a testicular biopsy at the Urology and Andrology department of the University Hospital of Giessen, Germany. He was diagnosed with complete bilateral SCO and sperm retrieval was not successful.

Clinical data of individuals M1645 and M2098 (*KPNA2* variants)

Individual M1645 underwent a testicular biopsy in a different clinical unit, with the corresponding pathology report stating complete bilateral SCO without successful sperm retrieval. Histology data was not available for this man. M2098 was also diagnosed with complete bilateral SCO and sperm retrieval was equally unsuccessful.

References

1. Houston, B. J. *et al.* A systematic review of the validated monogenic causes of human male infertility: 2020 update and a discussion of emerging gene–disease relationships. *Hum. Reprod. Update* **28**, 15–29 (2021).

Supplementary Figure Legends

Supplementary Figure 1. List of representative species of each clade for the phylostratigraphic analysis. Phylostrata are ranked from 1 (older, in red) to 16 (younger, in blue). The number of genes assigned to each phylostratum in the human (H), mouse (M) and fruit fly (insect, I) genomes is indicated above phylostratum rank. Divergence time (in grey) is indicated in million years ago (Mya).

Supplementary Figure 2. Ubiquitously expressed genes in male germ cells. a- Number of potentially ubiquitously-expressed old genes (mapping to the oldest-ranking phylostrata 1-5, see **Fig.1d**) in the male germ cell transcriptome. These were identified, in each species, by being also expressed in all three representative somatic cell types of the primary embryonic layers: neurons (ectoderm), muscle (mesoderm) and enterocytes (endoderm). Germ cells corresponds to spermatogonia + spermatocytes + spermatids. “Other” corresponds to germ cell-enriched and variably-expressed genes. Minimum expression cut-off: transcripts per million (TPM) >1. **b-** Potentially ubiquitously-expressed genes tend to be evolutionarily older than cell type-enriched genes, as determined by the transcriptome age index. Muscle is a noteworthy exception. Gene sets were determined irrespectively of their age. Germ cell-enriched corresponds to spermatogonia + spermatocytes + spermatids. Dots represent replicates, n= number of genes in each group. Minimum expression cut-off: TPM >1.

Supplementary Figure 3. Filtering the meiotic transcriptome networks. a- The majority of STRING functional association scores in the unfiltered network are weak (i.e., have a low combined score). Networks were filtered to only include functional associations (edges) with a STRING combined score of >0.5 (cut-off represented by the red lines). **b-** Filtering significantly reduced the total number of edges in the meiotic networks. Solid colors represent the number of remaining associations (edges) after filtering. **c-** Filtering maintained the majority of expressed genes in the meiotic networks. Solid colors represent the number of remaining genes after filtering. **d-** Meiotic networks have non-normal structure. Degree distribution (the likelihood of a gene being connected to other genes) follows a long tail distribution, with most genes having a

small number of connections and only a few being highly connected. For comparison, the pink line shows a normal distribution. **e-** Overall contribution of each of the six types of STRING data sources for the filtered meiotic networks. Contribution is measured by the percentage of all edges in the network containing information from each source. **f** to **k-** Numerical contribution of each individual STRING data source to the overall combined scores in the filtered meiotic networks.

Supplementary Figure 4. Rewiring the meiotic transcriptome networks. a- Effect on degree centrality metrics of randomly shuffling a variable percentage of all network edges. The difference between conserved and non-conserved genes is progressively attenuated but not erased. Gene conservation (across all Metazoa) was defined based on eggNOG orthogroups. **b-** Same as in "a", but for page rank metrics.

Supplementary Figure 5. Precision and recall curves confirm the reliability of machine-learning algorithms to predict evolutionary conservation. These curves plot positive predictive value (precision) against sensitivity (recall). Values correspond to AUC (area under the curve) scores. "Coin toss" corresponds to a random classification. Note that unbalanced datasets can offset the baseline of the coin toss results. Four-fold cross-validation results are shown. SVM: support-vector machine.

Supplementary Figure 6. Top 10 terms of an unfiltered gene ontology (GO) enrichment analysis of the human meiotic orthoBackbone. Tested category: biological processes. **a-** Meiotic orthoBackbone genes. **b-** Other evolutionarily-conserved meiotic network genes not part of the orthoBackbone. Gene conservation (across all Metazoa) was defined based on eggNOG orthogroups. False discovery rate $\leq$0.05.

Supplementary Figure 7. The *RNF113B* loss of function variant in individual M1911. a- Pedigree of M1911's family. Black indicates a diagnosis of infertility. DNA was available for individuals IV.9, V.3 (M1911, arrow), V.4 and V.5. A simplified genotype is indicated. LoF (loss-of-function): c.556_565del;p.(Thr186GlyfsTer119); WT ("wild-type"): reference allele. **b-**

Validation of the *RNF113B* LoF variant by Sanger sequencing. In all panels, arrows indicate the location of the variant / reference sequence. **c-** Clinical presentation of M1911.

Supplementary Figure 8. Silencing the spliceosome component *Prp19* in the fruit fly testis. Prp19 RNAi results in an comparable meiotic arrest phenotype to that of *dRNF113*, despite a less severe disruption of the orthoBackbone. **a-** *Prp19* is essential for male fertility. ****p <0.0001 (unpaired t-test). Male germ line driver: *bam*-GAL4. **b-** Prp19 RNAi is highly effective in the testis and does not downregulate dRNF113 levels. For comparison, the efficiency of the dRNF113 RNAi is also represented (right). RT-qPCR: quantitative reverse transcription PCR. **c-** Silencing *Prp19* leads to a primary spermatocyte arrest comparable to that of the dRNF113 RNAi. Phase-contrast microscopy. Arrowheads: primary spermatocytes; asterisks: early (round) spermatids; arrows: late (elongating) spermatids; sv- seminal vesicle. Scale bars: 50 μm (whole testis) and 20 μm (meiotic region). Meiotic area indicates the fraction of the entire testis occupied by primary spermatocytes. ****p <0.0001 and **p =0.002 (unpaired t-tests). **d-** Prp19 RNAi impacts the testicular transcriptome. Differential gene expression (DGE) analysis of RNA-Seq data obtained from whole testes samples (in triplicate) in both conditions. Down and upregulated genes in blue and red, respectively. orthoBackbone differentially expressed genes (DEGs) are outlined. FC: fold change. FDR: false discovery rate. Edge disruption corresponds to the number of orthoBackbone edges containing at least one DEG. orthoBackbone disruption is lower than in the dRNF113 RNAi (**Fig. 3g**).

Supplementary Figure 9. Association of *HSPA2* loss of function variants and the fruit fly Hsc70-1 RNAi with male infertility. a- *Hsc70-1*, one of the fruit fly homologs of *HSPA2*, is essential for male fertility. ****p <0.0001 (unpaired t-test). Male germ line driver: *bam*-GAL4. **b-** Silencing *Hsc70-1* phenocopies the dRNF113 RNAi meiotic arrest. Phase-contrast microscopy. Arrowheads: primary spermatocytes; sv- seminal vesicle. Scale bars: 50 μm (whole testis) and 20 μm (meiotic region). Meiotic area indicates the fraction of the entire testis occupied by primary spermatocytes. ****p <0.0001 and ns: no significant difference (unpaired t-tests). **c-** Validation of the *HSPA2* heterozygous stop-gain variant in individual M1678 by Sanger sequencing. Arrows indicate the location of the variant / reference sequence. **d-** Validation of the *HSPA2* LoF variant

in individual M2190 (heterozygous frameshift) by Sanger sequencing. **e-** Clinical presentation of M1678. **f-** Clinical presentation of M2190.

Supplementary Figure 10. Association of *KPNA2* loss of function variants and the fruit fly dKPNA2 RNAi with male infertility. a- *dKPNA2*, the fruit fly ortholog of *KPNA2*, is essential for male fertility. ****p <0.0001 (unpaired t-test). Male germ line driver: *bam*-GAL4. **b-** Silencing *dKPNA2* at meiotic entry aborts spermatogenesis at the late post-meiotic stage. Phase-contrast microscopy. Yellow dashed lines map the seminal vesicle (sv) insets. Note the lack of male gametes inside the seminal vesicles of the dKPNA2 RNAi, despite the presence of multiple elongating spermatids in the testis. These vesicles were either filled with cellular debris or empty. Arrows: late (elongating) spermatids; s- mature sperm; Ctr.- Control RNAi; RNAi- dKPNA2 RNAi. 20 seminal vesicles were scored in each condition for the presence of mature sperm. Percentages refer to the release of male gametes from the scored vesicles. Scale bars: 50 μm (whole testis) and 20 μm (seminal vesicles). **c-** Pedigree of individual M1645's family. DNA was available for individuals II.2, II.3, and III.2 (M1645, arrow). Black indicates a diagnosis of infertility. Individual III.5 was diagnosed with infertility for unknown reasons. A simplified genotype is represented. LoF (loss-of-function): c.667-2A>G;p.?; WT ("wild-type"): reference allele. **d-** Validation of the *KPNA2* LoF variant in M1645 by Sanger sequencing. Arrows indicate the location of the variant / reference sequence. **e-** Validation of the *KPNA2* LoF variant in individual M2098 (heterozygous frameshift) by Sanger sequencing. **f-** Clinical presentation of M1645. **c-** Clinical presentation of M2098.

Figure 1

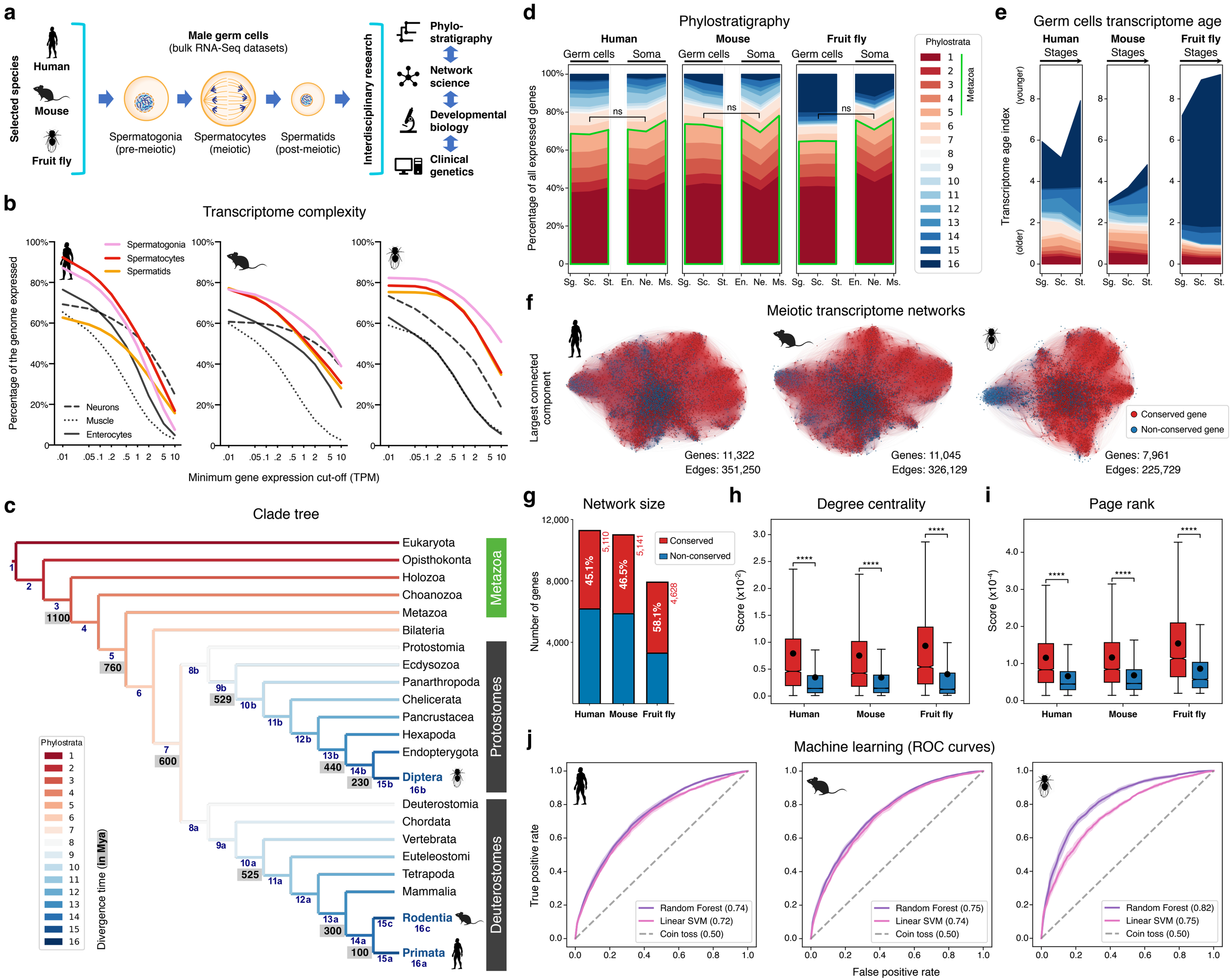

a
Selected species
Human
Mouse
Fruit fly
Male germ cells
(bulk RNA-Seq datasets)
Spermatogonia (pre-meiotic)
Spermatocytes (meiotic)
Spermatids (post-meiotic)
Interdisciplinary research
Phylo-stratigraphy
Network science
Developmental biology
Clinical genetics
b
Transcriptome complexity
Percentage of the genome expressed
Spermatogonia
Spermatocytes
Spermatids
Neurons
Muscle
Enterocytes
Minimum gene expression cut-off (TPM)
c
Clade tree
Eukaryota
Opisthokonta
Holozoa
Choanozoa
Metazoa
Bilateria
Protostomia
Ecdysozoa
Panarthropoda
Chelicerata
Pancrustacea
Hexapoda
Endopterygota
Diptera
Deuterostomia
Chordata
Vertebrata
Euteleostomi
Tetrapoda
Mammalia
Rodentia
Primata
Metazoa
Protostomes
Deuterostomes
Phylostrata
Divergence time (in Mya)
d
Phylostratigraphy
Human
Mouse
Fruit fly
Germ cells
Soma
Percentage of all expressed genes
ns
Phylostrata
Metazoa
e
Germ cells transcriptome age
Stages
Transcriptome age index
(younger)
(older)
f
Meiotic transcriptome networks
Largest connected component
Genes: 11,322
Edges: 351,250
Genes: 11,045
Edges: 326,129
Genes: 7,961
Edges: 225,729
Conserved gene
Non-conserved gene
g
Network size
Number of genes
Conserved
Non-conserved
45.1%
5,110
46.5%
5,141
58.1%
4,628
h
Degree centrality
Score (x10⁻²)
i
Page rank
Score (x10⁻⁴)
j
Machine learning (ROC curves)
True positive rate
False positive rate
Random Forest (0.74)
Linear SVM (0.72)
Coin toss (0.50)
Random Forest (0.75)
Linear SVM (0.74)
Coin toss (0.50)
Random Forest (0.82)
Linear SVM (0.75)
Coin toss (0.50)

Figure 2

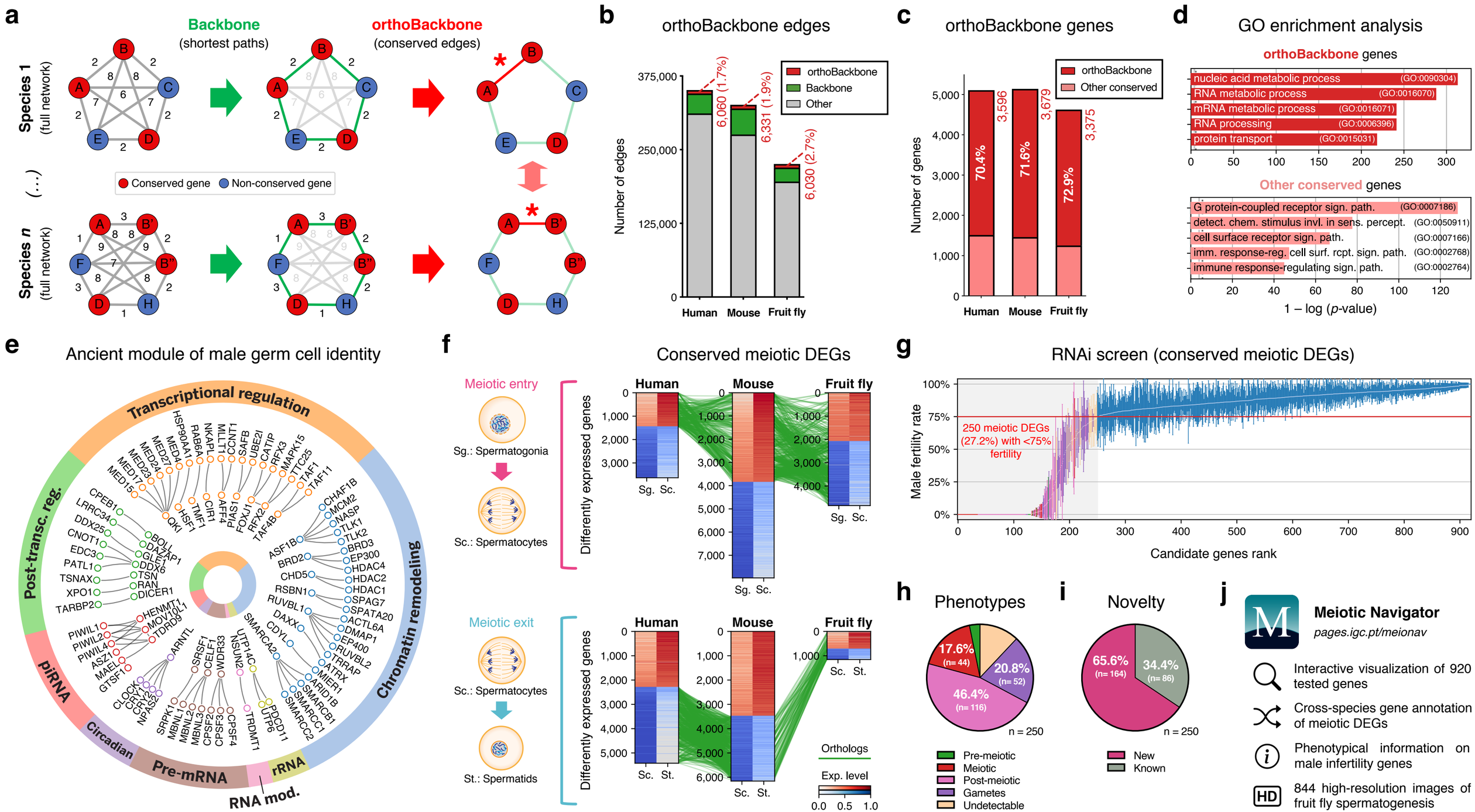
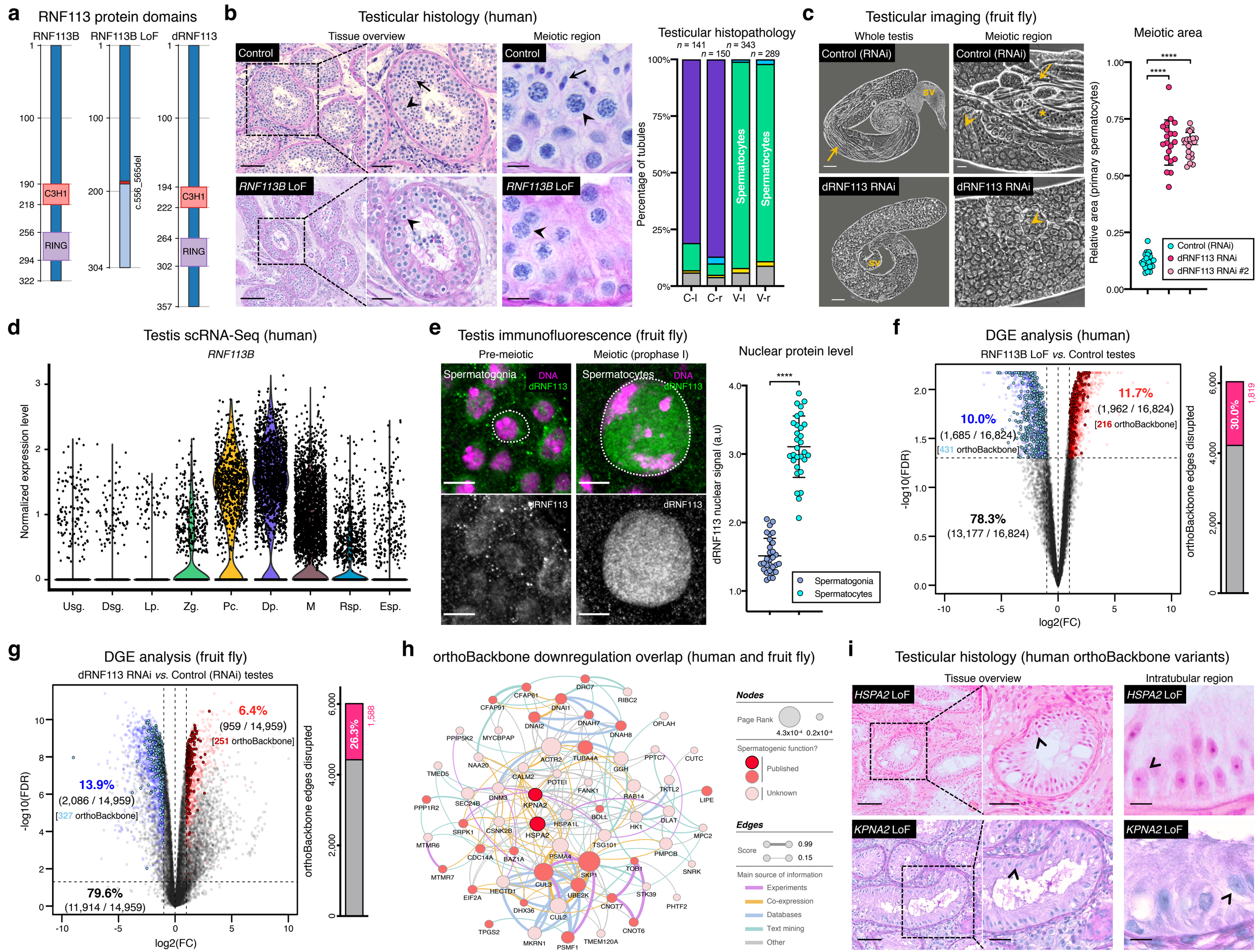

a
Backbone
(shortest paths)
orthoBackbone
(conserved edges)
Species 1
(full network)
(...)
Species n
(full network)
Conserved gene
Non-conserved gene
b
orthoBackbone edges
orthoBackbone
Backbone
Other
Number of edges
375,000
250,000
125,000
0
6,060 (1.7%)
6,331 (1.9%)
6,030 (2.7%)
Human
Mouse
Fruit fly
c
orthoBackbone genes
orthoBackbone
Other conserved
Number of genes
5,000
4,000
3,000
2,000
1,000
0
70.4%
71.6%
72.9%
3,596
3,679
3,375
Human
Mouse
Fruit fly
d
GO enrichment analysis
orthoBackbone genes
nucleic acid metabolic process (GO:0090304)
RNA metabolic process (GO:0016070)
mRNA metabolic process (GO:0016071)
RNA processing (GO:0006396)
protein transport (GO:0015031)
0 50 100 150 200 250 300
Other conserved genes
G protein-coupled receptor sign. path. (GO:0007186)
detect. chem. stimulus invl. in sens. percept. (GO:0050911)
cell surface receptor sign. path. (GO:0007166)
imm. response-reg. cell surf. rcpt. sign. path. (GO:0002768)
immune response-regulating sign. path. (GO:0002764)
0 20 40 60 80 100 120
1 – log (p-value)
e
Ancient module of male germ cell identity
Transcriptional regulation
Post-transc. reg.
Chromatin remodeling
piRNA
Circadian
Pre-mRNA
rRNA
RNA mod.
f
Conserved meiotic DEGs
Meiotic entry
Sg.: Spermatogonia
Sc.: Spermatocytes
Meiotic exit
Sc.: Spermatocytes
St.: Spermatids
Differently expressed genes
Human
Mouse
Fruit fly
Sg. Sc.
Sc. St.
Orthologs
Exp. level
0.0 0.5 1.0
g
RNAi screen (conserved meiotic DEGs)
Male fertility rate
100%
75%
50%
25%
0%
250 meiotic DEGs
(27.2%) with <75%
fertility
0 100 200 300 400 500 600 700 800 900
Candidate genes rank
h
Phenotypes
17.6%
(n= 44)
20.8%
(n= 52)
46.4%
(n= 116)
n = 250
Pre-meiotic
Meiotic
Post-meiotic
Gametes
Undetectable
i
Novelty
65.6%
(n= 164)
34.4%
(n= 86)
n = 250
New
Known
j
Meiotic Navigator
pages.igc.pt/meionav
Interactive visualization of 920 tested genes
Cross-species gene annotation of meiotic DEGs
Phenotypical information on male infertility genes
844 high-resolution images of fruit fly spermatogenesis

Figure 3


a
RNF113 protein domains
RNF113B
RNF113B LoF
dRNF113
C3H1
RING
c.556_565del
b
Testicular histology (human)
Tissue overview
Meiotic region
Control
RNF113B LoF
Testicular histopathology
Percentage of tubules
Spermatocytes
C-l
C-r
V-l
V-r
c
Testicular imaging (fruit fly)
Whole testis
Meiotic region
Control (RNAi)
dRNF113 RNAi
SV
Meiotic area
Relative area (primary spermatocytes)
Control (RNAi)
dRNF113 RNAi
dRNF113 RNAi #2
d
Testis scRNA-Seq (human)
RNF113B
Normalized expression level
Usg.
Dsg.
Lp.
Zg.
Pc.
Dp.
M
Rsp.
Esp.
e
Testis immunofluorescence (fruit fly)
Pre-meiotic
Meiotic (prophase I)
Spermatogonia
Spermatocytes
DNA
dRNF113
Nuclear protein level
dRNF113 nuclear signal (a.u)
Spermatogonia
Spermatocytes
f
DGE analysis (human)
RNF113B LoF vs. Control testes
11.7%
(1,962 / 16,824)
[216 orthoBackbone]
10.0%
(1,685 / 16,824)
[431 orthoBackbone]
78.3%
(13,177 / 16,824)
-log10(FDR)
log2(FC)
orthoBackbone edges disrupted
30.0%
1,819
g
DGE analysis (fruit fly)
dRNF113 RNAi vs. Control (RNAi) testes
6.4%
(959 / 14,959)
[251 orthoBackbone]
13.9%
(2,086 / 14,959)
[327 orthoBackbone]
79.6%
(11,914 / 14,959)
-log10(FDR)
log2(FC)
orthoBackbone edges disrupted
26.3%
1,588
h
orthoBackbone downregulation overlap (human and fruit fly)
Nodes
Page Rank
Spermatogenic function?
Published
Unknown
Edges
Score
Main source of information
Experiments
Co-expression
Databases
Text mining
Other
i
Testicular histology (human orthoBackbone variants)
Tissue overview
Intratubular region
HSPA2 LoF
KPNA2 LoF