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

Rion B. Correia[1,2,3,*], Joana M. Almeida[1,4,*], Margot J. Wyrwoll[5,*], Irene Julca[6], Daniel Sobral[7,8], Chandra Shekhar Misra[1,9], Leonardo G. Guilgur[1], Hans-Christian Schuppe[10], Neide Silva[1], Pedro Prudêncio[11], Ana Nóvoa[1], Ana S. Leocádio[1], Joana Bom[1], Moisés Mallo[1], Sabine Kliesch[12], Marek Mutwil[6], Luis M. Rocha[1,3,**], Frank Tüttelmann[5,**], Jörg D. Becker[1,9,**,#] & Paulo Navarro-Costa[1,4,**,#]

1- Instituto Gulbenkian de Ciência. Oeiras, Portugal.

2- CAPES Foundation, Ministry of Education of Brazil. Brasília, Brazil.

3- Department of Systems Science and Industrial Engineering, Binghamton University. USA.

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

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

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

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

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

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

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

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

12- 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 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 associated with germ cell defects shared across more than 600 million years of evolution. Our results demonstrate the central role of old genes in germ cell function and illustrate how comparative biology can be an important tool in 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 sex, 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 extend this concept to multicellular animals and posit that 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 combined this information with clinical genetics to identify new causes of human infertility. Collectively, 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 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 old evolutionary origin of the male germ cell transcriptome

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, 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 network of human, mouse and fruit fly spermatocytes, with nodes representing all genes expressed at TPM >1 and edges indicating the probability of the connected genes contributing to a specific function according to the STRING protein-protein interactions database[21] (**Fig. 1f**). The structure of these meiotic networks ultimately reflects the complex 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, through thresholding, 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 harbored 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 for conserved genes, we tested to what extent the degree centrality and page rank computations 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 confirm 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 a machine learning algorithm to reliably predict if a gene was conserved just based on network features (**Fig. 1j** and **Sup. Fig. 5**). 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. These values represent a robust predictive power of the evolutionary conservation of a gene, and were largely corroborated using a linear Support Vector Machine classifier, as well as precision and recall metrics. Based on these data, 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.

A set of 79 old functional interactions at the core of male germ cell identity

We have previously shown that it is possible to simplify the complexity of biological networks by removing network redundancy as to only retain the key edges that sustain the system[24]. These edges form the network backbone and are calculated based on the minimum number of shortest paths (proxy to the most relevant interactions) required to maintain network integrity[25]. In the context of transcriptome networks, the backbone uncovers the essential gene interactions that define core pathways in the gene expression program[26]. By extending the concept of orthology to the transcriptome backbone, we have developed a new metric - the orthoBackbone - consisting of all edges that connect the same pair of orthologs in the backbone of different species (**Fig. 2a**). Since the orthoBackbone contains key functional associations that have been maintained across evolution, it offers unprecedented access to what can be considered as defining genetic interactions at the heart of cell type identity.

The orthoBackbone drastically reduced the complexity 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 central network with the acquisition/maintenance of cellular identity. We explored the human orthoBackbone interactions to uncover, using GO annotations, all those involving gene expression regulators previously linked to spermatogenesis. By doing so we defined a set of 79 ancient gene interactions between 104 genes (out of the total of 3,596 in the orthoBackbone) ascribed to a wide gamut of regulatory processes and involving noteworthy spermatogenesis genes such as *RFX2*

(transcription), *CDYL* (chromatin remodeling) and *BOLL* (translation), among others[27,28,29] (**Fig. 2e**). Based on the conservation of these ancient functional associations across evolutionarily-distant species, their unique network properties, and the significant spermatogenic role of the intervening genes, they can be regarded as core constituent elements 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 genetic program 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[30,31]. 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 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[32]. 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 the data generated in the screen 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 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*[33]) - emerged as a particularly interesting meiotic regulator. By analyzing our in-house whole exome database containing sequencing data from 74 cases of human 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[34]. 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. 8** 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[35], we observed that *RNF113B* was predominantly expressed at meiotic entry, peaking at the diplotene stage (**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[36], had a lower effect on the orthoBackbone (15.2% of disrupted edges), despite a similar meiotic arrest phenotype (**Sup. Fig. 7**). 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 potential gatekeepers of male germ cell identity. Furthermore, they provide us the means to interfere with a sizable fraction of the meiotic orthoBackbone network 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)[37,38]. Since male infertility affects up to 12% of all men[39], addressing such knowledge gap is an issue of clear clinical 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, out of the 79 core functional associations we propose as essential constituents of male germ cell identity, 32 (40.5%) were 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 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[29], 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[40]. 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[41,42]. 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**). Two different heterozygous LoF variants were detected in our male infertility cohort, with this gene having a predicted autosomal-dominant inheritance

(**Sup. Fig. 9b, 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[43] (**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[44], a conserved regulator of meiotic progression from yeast to mammals[45,46]. 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[47], 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[48], with its silencing being associated with abnormal spermiogenesis due to spermatid individualization defects. This post-meiotic function might also be present in its mammalian orthologs, as suggested by the nuclear localization of Kpna2 in elongating mouse spermatids[44]. 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. 10**, **Sup. Information** and **Methods**). One, in individual M1645, is the splice-site variant c.667-2A>G;p.? predicted to disrupt the correct splicing of intron 6 due to the loss of an acceptor site. The other, 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 identifying disturbances in 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 number 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[49,50]. 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[51]. 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[52,53]. 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[54]. 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[31]. 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)[55]. 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[33].

The actual benefit of comparative biology for the identification of new genetic causes of human disease is often a contentious topic[56]. 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[57]. By focusing on the deep evolutionary past of human spermatogenesis, we have identified 183 novel functionally-validated candidate genes that have so far uncovered three new genetic causes of human infertility (*RNF113B*, *HSPA2* and *KPNA2*). These 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 the experimental results and 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), and Patrícia Beldade (Faculty of Sciences, University of Lisbon) 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 are 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 is 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 B. Correia was supported by a contract within the scope of FCT project PTDC/MEC-AND/30221/2017. Chandra Shekhar Misra acknowledges a doctoral fellowship from FCT (PD/BD/114362/2016) under the Plants for Life PhD Program. Jörg D. Becker received salary support from FCT (grant CEECIND/03345/2018). Paulo Navarro-Costa is 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). doi:10.1016/B978-0-12-372568-4.00003-3.

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). doi:10.4324/9781315765471.

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. 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) doi:10.1101/2022.02.02.478784.

26. Ren, Y., Ay, A. & Kahveci, T. Shortest path counting in probabilistic biological networks. *BMC Bioinformatics* **19**, 465 (2018).

27. Wu, Y. *et al.* Transcription Factor RFX2 Is a Key Regulator of Mouse Spermiogenesis.

Sci. Rep. **6**, 20435 (2016).

28. 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).

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

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

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

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

33. 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).

34. 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).

35. 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).

36. 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).

37. 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).

38. 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).

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

40. 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).

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

42. 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).

43. 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).

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

45. 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).

46. 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).

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

48. Mason, D. A., Fleming, R. J. & Goldfarb, D. S. Drosophila melanogaster Importin □1 and □3 Can Replace Importin □2 During Spermatogenesis but Not Oogenesis. *Genetics* **161**, 157–170 (2001).

49. 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).

50. Moi, D. *et al.* Archaeal origins of gamete fusion. *bioRxiv* (2021) doi:10.1101/2021.10.13.464100.

51. 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).

52. 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).

53. 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).

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

55. 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).

56. 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).

57. 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 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 meiotic cells have large, structured transcriptome networks. Graphs represent the largest connected component of the spermatocyte transcriptome (minimum expression cut-off: TPM >1) according to STRING functional association scores. Gene conservation (across all Metazoa) was defined based on eggNOG orthogroups. Networks were thresholded 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 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). Letters depict different genes, numbers indicate distances between genes. **b-** The meiotic orthoBackbone represents less than 3% of all functional associations (edges) in the transcriptome 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 79 ancient functional interactions between known spermatogenesis gene expression regulators. 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 the *RNF113B* LoF (from M1911) 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[35]. 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, 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 thresholded 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 human variants in *HSPA2* and *KPNA2* (red nodes) are depicted in **i**. Edge size 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* LoFs, 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).

Figure 1

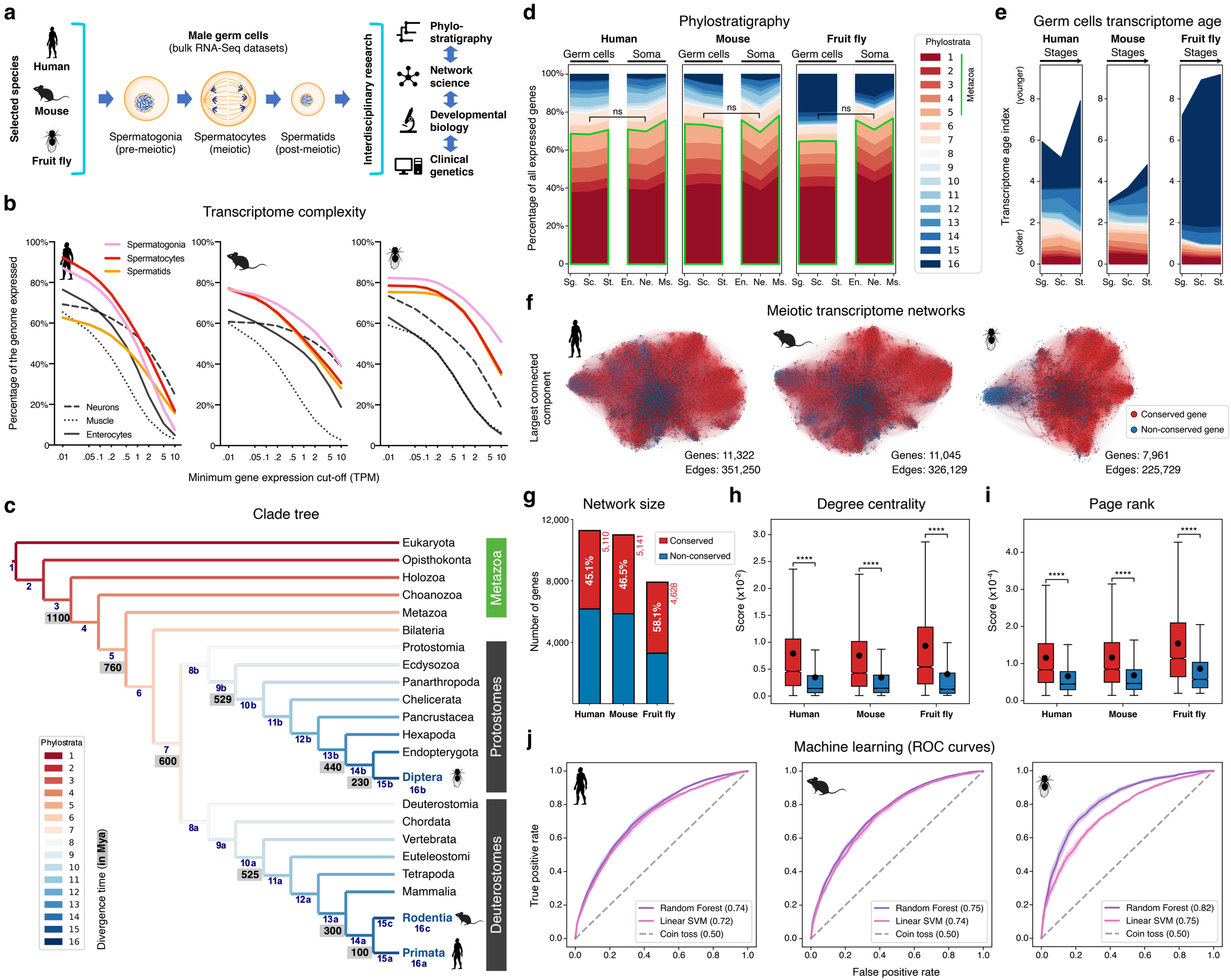

a
Selected species
Human
Mouse
Fruit fly
Male germ cells
(bulk RNA-Seq datasets)
Spermatogonia (pre-meiotic)
Spermatocytes (meiotic)
Spermatids (post-meiotic)
Interdisciplinary research
Phylo-stratigraphy
Network science
Developmental biology
Clinical genetics
b
Transcriptome complexity
Percentage of 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

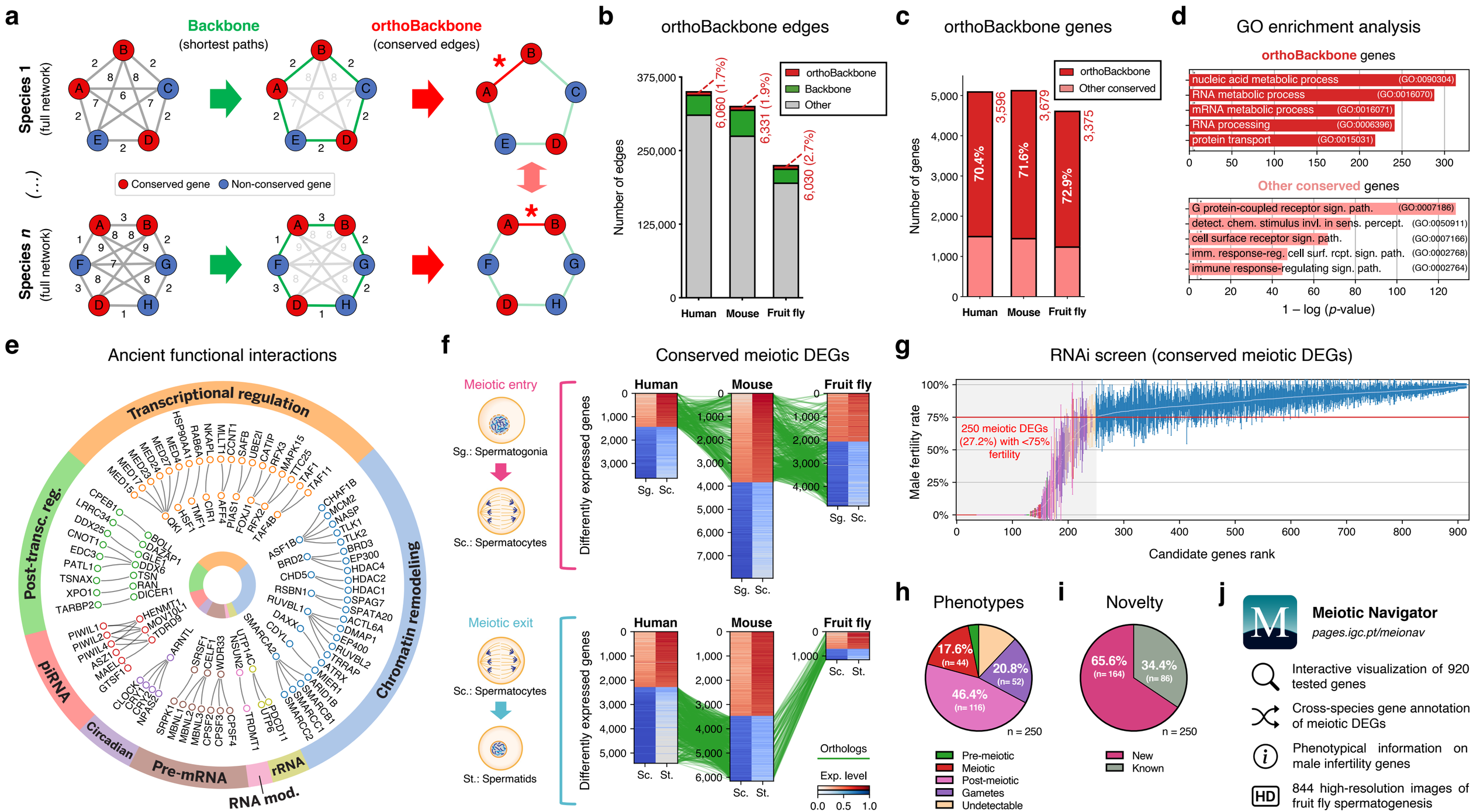
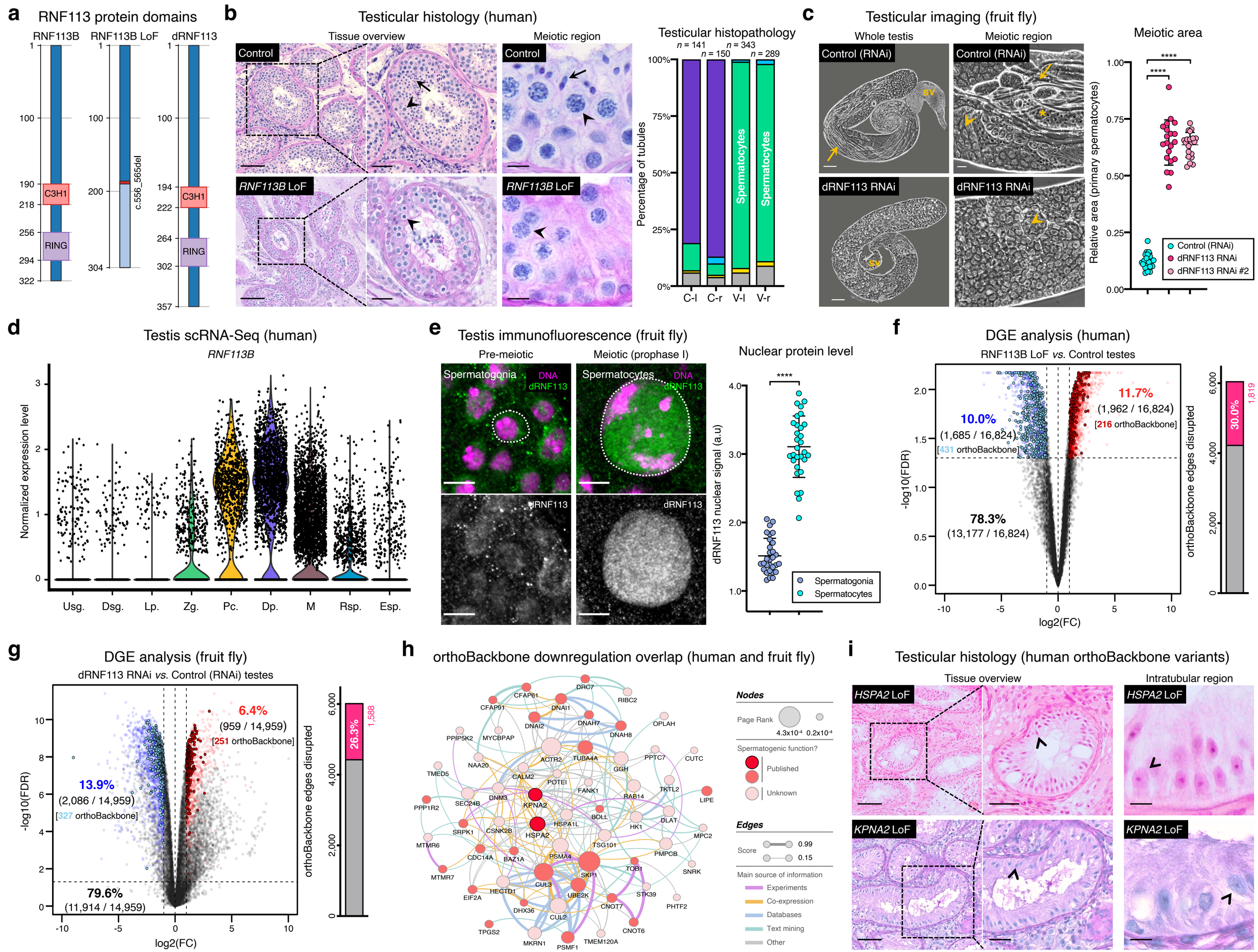

a
Backbone
(shortest paths)
orthoBackbone
(conserved edges)
Species 1
(full network)
(...)
Species n
(full network)
Conserved gene
Non-conserved gene
b
orthoBackbone edges
orthoBackbone
Backbone
Other
Number of edges
375,000
250,000
125,000
0
6,060 (1.7%)
6,331 (1.9%)
6,030 (2.7%)
Human
Mouse
Fruit fly
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 functional interactions
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
n = 141
n = 150
n = 343
n = 289
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