Prediction of meiosis-essential genes based upon the dynamic proteomes responsive to spermatogenesis

Kailun Fang[1,2,3,8], Qidan Li[3,4,5,8], Yu Wei[2,3,8], Jiaqi Shen[6,8], Wenhui Guo[3,4,5,7,8], Changyang Zhou[2,3,8], Ruoxi Wu[1], Wenqin Ying[2], Lu Yu[1,2], Jin Zi[5], Yuxing Zhang[3,4,5], Hui Yang[2,3,9*], Siqi Liu[3,4,5,9*], Charlie Degui Chen[1,3,9*]

1. State Key Laboratory of Molecular Biology, Shanghai Key Laboratory of Molecular Andrology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China.

2. State Key Laboratory of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Institute of Neuroscience, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.

3. University of the Chinese Academy of Sciences, Beijing 100049, China.

4. CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

5. BGI Genomics, BGI-Shenzhen, Shenzhen 518083, China.

6. United World College Changshu China, Jiangsu 215500, China.

7. Malvern College Qingdao, Shandong, 266109, China

8. These authors contributed equally: Kailun Fang, Qidan Li, Yu Wei, Jiaqi Shen, Wenhui Guo, Changyang Zhou.

9. These authors jointly supervised this work: Charlie Degui Chen, Siqi Liu, Hui Yang

*email: siqiliu@genomics.cn; cdchen@sibcb.ac.cn; huiyang@ion.ac.cn

ABSTRACT

Mammalian meiosis is a cell division process specific to sexual reproduction, whereas a comprehensive proteome related to different meiotic stages has not been systematically investigated. Here, we isolated different types of germ cells from the testes of spermatogenesis-synchronized mice and quantified the corresponding proteomes with high-resolution mass spectrometry. A total of 8,002 proteins were identified in nine types of germ cells, while the protein signatures of spermatogenesis were characterized by the dynamic proteomes. A supervised machine learning package, FuncProFinder, was developed to predict meiosis-essential candidates based on the proteomic dataset. Of the candidates with unannotated functions, four of the ten genes at the top prediction scores, Zcwpw1, Tesmin, 1700102P08Rik and Kctd19, were validated as meiosis-essential genes by knockout mouse models. The proteomic analysis towards spermatogenic cells indeed setups a solid evidence to study the mechanism of mammalian meiosis. The proteome data are available via ProteomeXchange with identifier PXD017284.

INTRODUCTION

Meiosis is a cell division process specific to germ cells, in which DNA replicates once and divides twice to generate four gametes. It is well accepted that mammalian meiosis is a complex process including homologous recombination, synapsis and so on, while the molecular mechanisms involved in such process are still to be explored yet[1, 2]. Since the genes participating in yeast meiosis are well studied, homology comparison to yeast is a common strategy to scrutinize the meiotic mechanism of mammalian (e.g., such as Spo11, Dmc1, Psmc3ip, and Rnf212)[3-7]. On the other hand, the regulatory mechanism of mammalian meiosis is more complicated than that of yeast, and the genes specifically participated in mammalian meiosis could not be found by this strategy. Knock-out of genes with testis or oocyte-specific-expression pattern could be another approach to identify meiosis-essential genes in mammalian. For instance, a total of 54 testis-specific gene were knocked out by Miyata's group, however, none of the knockout mice exhibited a meiosis-essential phenotype[8]. It is thus clear that an efficient approach to find mammalian meiosis-essential genes is badly required in the frontier.

The status of gene expression is a fundamental characteristic tightly associated with physiological functions, while a dynamic atlas of gene expression throughout spermatogenesis would be extremely useful for exploration of meiosis-essential genes.

Up to now, transcriptional gene expression in thousands of germ cells covering various developmental stages of spermatogenesis were quantified at single cell level[9-15], resulting in very detail transcriptome landscape throughout spermatogenesis, yet few studies explored the data for further functional excavating. As gene expression at protein level are downstream of transcription, proteomic abundance change of genes could be more directly associated with phenotype or functional change. Importantly, multiple studies clarified a poor correlation between mRNA and protein abundance in testes[16, 17], therefore, a global proteomic profiling of gene expression in spermatogenesis is of great meaningful to unravel functional molecules of meiosis. However, the report regarding systematic profiling of proteomics during meiosis was limited. Only one type of meiotic cells-pachytene spermatocytes was quantified in previous proteomics studies[7, 17], leaving protein expression remained unknown in most of the stages in meiosis. Therefore, in contrast to meiotic dependence of transcriptomes in details, quantified profiling of meiotic proteome has remained a large room to be improved, as well as digging for functional molecules from a big omics dataset.

In this work, to understand the molecular basis of mouse meiosis and predict meiosis-essential proteins, 7 consecutive types of meiotic cells plus pre-meiotic spermatogonia and post-meiotic round spermatids were isolated and the proteins in each cell-type were identified and quantified by high-resolution mass spectrometry with a label-free mode. The meiosis-dependent signatures were characterized by protein abundance changes. Furthermore, a supervised ensemble machine learning package, FuncProFinder, was developed to predict the meiosis-essential proteins. The meiosis-related phenotypes for the five proteins at the top scores of the prediction, Pdha2, Zcwpw1, Tesmin, Kctd19 and 1700102P08Rik were verified by knockout mice. Therefore, comprehensive proteomics data paves a path to efficiently discover meiosis-essential proteins and to figure out their functions in meiosis.

RESULTS

Isolation for the mouse spermatogenic cells around meiosis

To quantify protein expression change and closely monitor the molecular events in response to mouse meiosis, we isolated the spermatogenic cells around meiosis in C57BL/6 mouse testes, including pre-meiotic TypeA undifferentiated spermatogonia, consecutive types of meiotic cells, and post-meiotic round spermatids.

The isolation workflow of spermatogenic cells is illustrated in Fig.1a. TypeA undifferentiated THY1+ c-KIT- spermatogonia (Aundiff) were isolated from the testes of postnatal day 7 (P7) mice using magnetic activated cell sorting according to an established method[18] (Fig.S1a, S1c). The immuno-fluorescence staining of PLZF, a well-known Aundiff marker, revealed that the percentage of PLZF+ cells increased from 10% to 70% after purification (Fig. S1b, S1d, Table S1), implying the Aundiff cells greatly enriched. The haploid round spermatids (RS) were purified by DNA-content based cell sorting from the testes of P28 mice (Fig.S1e, S1g). DAPI-staining images indicated that the purity of isolated RS reached almost 100% (Fig.S2f-S2i, Table S1).

In the seminiferous tubules of mouse testes, consecutive types of meiotic cells are mixed and difficult to be separated from each other. To simplify the types of spermatocytes in testes, we applied a spermatogenesis synchronization method described before[9, 19, 20]: mouse spermatogonia differentiation was inhibited by WIN18,446 for 7 days, and re-activated synchronously by retinoic acid (RA) injection on P9 (Fig.1b). Four weeks after RA treatment, testes of P37 to P46 mice exhibited only one or two types of meiotic spermatocytes at a given point (Fig.1c-1g), greatly facilitating DNA-content based cell sorting for purification (Fig.1h-1l). To assess the purity of the isolated meiotic cells, we performed immuno-fluorescence staining with antibodies against the synaptonemal complex marker SYCP3 and the DNA damage marker γH2AX (Fig.1m-1t) and recognized spermatocyte cell-types with the criteria described previously[21]. Based on the quantitative evaluation upon fluorescence, most of isolated meiotic cells were of high purity around 90% (Table 1). Considering the protein amount of isolated early Leptotene and Leptotene were less than 120 μg, we mixed these two adjacent cell types together as an earlyL/L group for the following proteomic analysis. Thus, a total of seven types of meiotic cells, early leptotene and leptotene (earlyL/L), zygotene (Z), early pachytene (earlyP), middle pachytene (midP), late pachytene (lateP), early diplotene (earlyD), and late diplotene (lateD) were prepared for further protein study.

A quantitative proteomic atlas of mouse meiosis during spermatogenesis

In the nine types of spermatogenic cells isolated above, a total of 8,002 proteins were identified (unique peptides≥2), with between 6,000–7,000 proteins in each cell type (Table S2). In all the identified proteins, 7,742 proteins were only detected in seven sub-stages of meiosis, and 5,108 proteins were globally identified through all the nine cell types. To obtain high quality of quantification data, each sample was triplicated in LC-MS/MS. The Pearson correlation coefficients for all the triplicates in the same sample reached around 0.99 (Fig.2a), indicating the proteomic quantification was highly replicated. Additionally, the comparison of protein expression correlation among 9 different cell-types revealed that the protein abundance changed dramatically between earlyP and midP, strongly implied that spermatocytes may undergo a cell state transition after passing of the midPachytene checkpoint.

Next, to evaluate the consistency of our protein quantification with previous knowledge, we tracked the protein abundance change of 15 well-known cell-type-specific biomarkers throughout spermatogenesis (e.g., Lin28a, Stra8, Spo11, Tnp2 etc.). As depicted in Fig.2b, protein abundances of all those biomarkers appeared typical phase-dependent, which was basically in agreement with previous researches[22-25]. In addition, proteins in several meiosis representative processes are generally recognized as meiotic-phase dependent, while the proteomic evidence in this study further implied their functions around meiosis (Fig.2c). For instance, synaptonemal complex (SC) mainly forms from Z to lateP and decreased after lateP, and most SC components in proteomics data showed consistent with previous knowledge. However, Syce1, a key SC component, remained stable protein abundance from D to RS. A similar Syce1 transcriptional expression pattern could be observed in a single-cell RNA sequencing study[9], implying that Syce1 might perform additional functions except synapsis after meiosis Prophase I.

To further explore the phase-dependent dynamic processes around meiosis, the abundance of all the 8,002 identified proteins in nine cell types were input to a statistical software, Perseus, for differential expression proteins (DEPs) analysis. A DEP was defined as its abundance with significant changes between any two sub-stages when Q-value less than 0.001. A total of 6,020 proteins were determined as DEPs, and these DEPs were divided to 4 groups by K-means analysis, C1 matched with Aundiff, C2 with earlyL/L, C3 with Z-earlyP, and C4 matched with midP-RS (Fig.2d, Table S3). Gene ontology (GO) analysis towards DEPs in each cluster led to uncover the biological processes enriched in different phases around meiosis. As illustrated in Fig.2e, cell-cell adhesion and actin cytoskeleton organization related proteins were enriched in Aundiff cells. Nucleic acid related processes such as rRNA processing, DNA replication were enriched in earlyL/L cells. Meiotic cell cycle related proteins were enriched in Z-earlyP cells, and piRNA metabolism, sperm function-related proteins were enriched in midP-RS cells. The KEGG pathways analysis towards 4 DEP clusters were also illustrated in Fig.S2a, and proteins in 4 representative pathways, DNA replication (enriched in C2), spliceosome, proteasome and oxidative phosphorylation (C4) were typically differentially expressed during meiosis (Fig.S2b-e). Taking all the information above, the proteomic information both qualitative and quantitative was not only highly agreed with prior knowledge, but also offered new clues to understand meiotic molecules, approaching the additional functions of meiotic proteins, functionally categorizing the meiotic DEPs and uncovering previously uncharacterized molecular signatures and dynamic processes from the protein abundance change during meiosis.

Supervised machine learning analysis of proteomic data can predict meiosis-essential candidates during spermatogenesis

Although 8,002 proteins were identified in the spermatogenesis and were further divided to four groups with relevant biological processes, the question was not well clarified which protein was essential to meiosis. Recently, supervised machine-learning approaches were applied to systemically predict functional genes[26]. Here, we established a supervised ensemble machine learning Matlab package called FuncProFinder to predict and discover meiosis-essential candidates (Methods and Supplementary methods). According to MGI phenotype annotation, a protein is called meiosis-essential because knockout of the protein leads to meiosis arrest, while a non-essential protein is termed that the protein knockout mice does not have lethal or meiosis-arrest phenotype. From the 8,002 identified proteins in this study, 159 proteins were essential and 2,151 were non-essential (Fig.3a, Table S3). With protein abundance of these proteins as train sets, three

methods of the FuncProFinder, radial basis function (RBF)[27], naive Bayesian model (NBM) and support vector machine (SVM)[28], were used to construct classifiers to predict whether or not a given protein was meiosis-essential. Based on the FuncProFinder package, the prediction precision reached to 47.70% (RBF), 30.97% (NBM) and 20.71% (SVM) with the recall setting at 0.2 tested by Monte-Carlo cross validation[29]. While AUCs for the receiver operation characteristic (ROC) curve of the prediction were 0.7364 (RBF), 0.7150 (NBM) and 0.6711 (SVM), respectively (Fig.3b). As the prediction performance of RBF in both precision and AUC was over the other two algorithms in this dataset, FuncProFinder-RBF was accepted to predict meiosis-essential possibility for a protein.

Next, the identified proteins were scored through the FuncProFinder-RBF algorithm (Table S4), the higher the meiotic confidence scores, the more possible to be meiosis-essential. 500 proteins on the top of the scores were filtrated and their functional information in MGI were shown in Fig. 3c. Without the filtration, the ratio of meiosis-essential proteins against unlethal proteins was 6.54% (159:2310), whereas after filtration the ratio changed to 49.30% (35:71), indicating well-known meiosis-essential proteins were greatly enriched. The top 500 candidates exhibited dynamic protein abundance change during spermatogenesis (Fig.3d), containing more DEPs (94.00%) compared to total 8,002 proteins (75.23% DEPs), and 83.60% proteins of top-500 were highly abundant in 3 meiotic sub-phases (C2-C4 group), compared to only 57.01% before RBF filtration (Fig.3e), consistent with the hypothesis that a meiosis-essential candidate could be a DEP expressed highly in sub-phases of meiosis. Taking all above, the RBF algorithm could be a potential method to select the meiosis-dependent candidates from the large pool of identified proteins.

Pyruvate dehydrogenase alpha 2 (PDHA2) is essential for meiosis

Of the top 500 candidates, there were 176 proteins whose phenotype were not validated by KO mouse models according to the MGI database (Fig.4a). Understanding of the functional pathways of those predicted meiosis-essential candidates could apply new views of molecular basis of meiosis. Enrichr, a pathway enrichment tool, was implemented to find out enriched functions of these 176 candidates based on KEGG 2019 Mouse database. Fig.4b unraveled the top 6 functional categories after enrichment analysis, and only one common metabolism pathway, pyruvate metabolism, was highly enriched in these meiosis-essential candidates. It has been reported that pyruvate metabolism was required in the isolated Pachytene spermatocytes cultured in vitro[30]. However, whether pyruvate-related proteins were essential in meiosis is not verified in vivo. PDHA2, among the top-500 meiosis-essential candidates, is a catalytic subunit of the pyruvate dehydrogenase complex (PDC), associated with other four proteins, PDHB, DLAT, DLD and PDHX[31]. Lack of any component in the complex could lead to activity loss. In previous study, the transcription status of Pdha2 gene was dynamically changed during spermatogenesis, increased in Pachytene and gradually decreased in spermatids[32]. With our proteomic data, the abundance changes of all the proteins in the PDC during spermatogenesis were further illustrated in Fig.4c. The protein abundance of two catalytic subunits of PDC, PDHA1 and PDHA2, appeared changes in the opposite directions, X-chromatin-linked protein PDHA1 decreasing during meiosis due to meiotic sex chromosome inactivation (MSCI), whereas PDHA2 increasing from earlyP to lateD. As regards the other four components of PDC, the change patterns of their abundance were similar to PDHA2, implying that the PDC had an integrity structure of catalytic functions during meiotic development.

To further verify physiological roles of Pdha2 during meiosis, a Pdha2 knockout mouse model was generated by the CRISPR/Cas-mediated genome engineering. A 13-basepair deletion was induced into the Pdha2 exon, which led to the reading-frame shift of Pdha2 and early termination (Fig.4d), and the knockout result was examined by genotyping (Fig.4e). The testes weight of 8-weeked adult Pdha2$^{-/-}$ mouse were obviously smaller than Pdha2$^{+/-}$ mouse (Fig.4f). The H&E staining of cross sections to the mouse testes and epididymis were depicted on Fig.4g and Fig.4h. In Pdha2$^{+/-}$ mice, the testes were comparable to WT mice, in which all types of germ cells were observed, from spermatogonia to spermatozoa, while mature sperms were fully filled in their epididymis. In contrast to their heterozygous littermates, the Pdha2$^{-/-}$ mice showed that the post-meiotic cells were totally absent, whereas Pachytene-like spermatocytes were accumulated in their testes. Furthermore, no spermatozoa was observed in their epididymis. Thus, with the help of RBF prediction, this study provides an evidence that the PDHA2 is a meiotic-regulation factor, knockout of which is likely to stop the meiosis at Pachytene. As PDC catalyzes pyruvate to acetyl-CoA and decides the energy level in a cell, it is a reasonable deduction that PDHA2, as a key component

of PDC, could regulate ATP generation in spermatocytes and affect the meiotic process.

Phenotype verification of the top 10 male meiosis-essential candidates without function annotation

Among the 176 proteins mentioned above, 41 proteins appeared without KEGG pathway annotation (Fig.5a). Whether they are essential for meiosis need to be further verified by experiments. The top 10 candidates upon meiotic confidence score were selected and knocked-out in mice by the CRISPR/Cas-mediated genome engineering (Fig.S3a-j). Abundance changes of the selected 10 proteins were exhibited in Fig.5b. After knock-out treatment, we obtained survival pups from 8 of the 10 genes, as the deficiency of the Gapvd1 and 1700037H04Rik in mice might lead to lethal. The reproductive anatomy and fertility of these non-lethal mice were carefully examined by testis weight and histological analysis of testes. In the Txnl1$^{-/-}$, AA467197$^{-/-}$, Lrrc40$^{-/-}$ and Naxe$^{-/-}$mice, no significant change was observed in their testes weight and H&E staining images as depicted in Fig.S3a-h. However, knockout of the other 4 genes indeed affected the testes morphology of the homozygous deficient mice. Generally, the testis weights of the Zcwpw1$^{-/-}$, Tesmin$^{-/-}$, Kctd19$^{-/-}$ or 1700102P08Rik$^{-/-}$ mice were significantly lighter as compared with the heterozygous littermates (Fig.5c-f). Specifically, H&E staining images of the testes derived from Zcwpw1$^{-/-}$, Tesmin$^{-/-}$, and 1700102P08Rik$^{-/-}$ mice appeared Pachytene arrested phenotype, Pachytene spermatocytes with condensed nuclei, lack of post-Pachytene cells, and tubules highly vacuolized (Fig.5g-i). In the Kctd19$^{-/-}$ mice, the mice exhibited typical Metaphase I arrested phenotype, containing spermatocytes from Leptotene to Metaphase I, but with no post-meiotic cells (Fig.5j).

A molecular mechanism of Pachytene arrest is hypothesized to be resulted from failure of DNA repair or synapsis[33, 34]. As knock out of Zcwpw1$^{-/-}$, Tesmin$^{-/-}$, or 1700102P08Rik$^{-/-}$ led to Pachytene-arrest, the molecular mechanisms underlying the Pachytene arrest phenotypes in these three knockout mice lines need to be verified. To address the question, the spermatocytes in Zcwpw1$^{-/-}$, Zcwpw1$^{+/-}$, Tesmin$^{-/-}$, Tesmin$^{+/-}$, 1700102P08Rik$^{-/-}$ or 1700102P08Rik$^{+/-}$mice were chromosome-spread and immune-stained with the antibodies against DNA recombination and synapsis events, including SYCP3 and SYCP1 as the components of synaptonemal complex, γH2AX as an indicator of DNA damage and MLH1 as a marker of crossover formation. The immunostaining of the 4 different antibodies against the spermatocytes from Tesmin and 1700102P08Rik in both heterozygous and homozygous knock out mice exhibited no difference as shown in Fig.S4f-n, implying either DNA repair or synapsis were not affected by gene knockout. However, the immunostaining of these antibodies against the Zcwpw1$^{-/-}$ spermatocytes was quite different with Zcwpw1$^{+/-}$ (Fig.6a-e). The staining signal distribution of γH2AX in the spermatocytes at Leptotene of the Zcwpw1$^{-/-}$ mice was comparable with that of Zcwpw1$^{+/-}$, suggesting that the formation of double strand breaks (DSBs) was not affected by the absence of ZCWPW1 (Fig.6a). However, in the spermatocytes at Pachytene, the γH2AX staining was only seen in the sex body region in Zcwpw1$^{+/-}$, whereas it still spread out in the autosome regions in Zcwpw1$^{-/-}$, indicating that the DSB repair was not finished on the autosome once ZCWPW1 not expressed (Fig.6b). Co-immuno-staining of SYCP3 and SYCP1 in the heterozygous mice were highly merged in the chromosomes, whereas the locations of the two structure proteins were not fully overlapped in Zcwpw1$^{-/-}$ mice, suggesting that the synaptonemal complex were not fully formed due to the lack of ZCWPW1 (Fig.6c). Furthermore, the co-immuno-staining of SYCP3 and MLH1 in the Pachytene spermatocytes appeared nearly no MLH1 loci in the chromosome of Zcwpw1$^{-/-}$ mice, whereas the staining signals were perceived in Zcwpw1$^{+/-}$ mice and were counted within normal range, implicating that the crossover formation was inhibited because of ZCWPW1 knock out (Fig.6d, 6e). On one hand, the ZCWPW1 plays a key role in both DNA repair and synapsis, on the other hands, the functions of 1700102P08Rik and Tesmin are not clearly clarified even though the two proteins indeed participate in the regulation of meiosis.

To summarize the knockout experiments for the meiosis-essential proteins predicted, the RBF offered a set of satisfactory candidates. Of the top 10 candidates, 40% at least were verified as meiosis-essential. As their functions are not annotated yet, their involvement of meiosis would be an interesting direction for functional exploration. For example, the Pachytene arrest in Zcwpw1$^{-/-}$ mice were found to be resulted from failed DSB repair and incomplete synapsis.

DISSUCSSION

In this study, one of the fundamental goals is to acquire global and quantitative information of proteomics during different stages of mouse meiosis. How to obtain such information is a long-lasting question. The proteomic investigation towards the entire mouse testis and one type of meiotic cells-Pachytene spermatocytes have been accomplished in several labs[7, 17, 35]. However, without isolation of different types of meiotic cells, these studies could not generate a precise picture into the different stages of mouse meiosis.

Here, first, we designed a project that enabled a comprehensive proteomic profiling around meiosis. To reach the goal, 7 consecutive types of meiotic cells plus pre-meiotic spermatogonia and post-meiotic round spermatids were well-isolated and proteins in each cell-type were identified and quantified by high-resolution mass spectrometry. A total of 8,002 proteins were identified, including 6,020 differentially expressed proteins, which was the largest dataset of proteomics related to meiosis in the relevance research area, offering global information of protein quantities in different stages of the entire meiosis process.

Second, this comprehensive proteomics is likely to provide new views for understanding of meiosis. For instance, it is generally accepted that the spermatocytes could be categorized to several sub-stages judged by the status of chromosome morphology[21, 36]. Based on this criterium, earlyP and midP are categorized to the similar type of Pachytene cells, nevertheless, the protein abundance correlation revealed these two types of spermatocytes with the similar appearance were totally different (Fig.2a and Fig.2d). And for another example, components of the synaptonemal complex were assumed to mainly exist from Z to lateP and decreased after lateP. However, Syce1, a key SC component, was consistently detected after lateP to RS with relatively higher abundance (Fig.2c), implying that Syce1 might perform additional functions except synapsis after meiosis Prophase I. Therefore, the refine profile of quantitative proteome appeared a new assessment of molecular events related to meiosis.

Third, informatics analysis towards proteomic data related with spermatogenesis is likely to pave a path for functional exploration of mouse meiotic genes. Today, genes involved in critical meiotic events, such as homologous recombination and synapsis, remain poorly understood. Miyata et al. selected 54 testis-specific genes and made the correspondent knockout mice, unfortunately, they did not find any gene related to meiosis[8]. In this study, an ensemble strategy of machine learning was taken to predict meiosis-essential proteins based on the abundance change of meiosis-essential or non-essential proteins. With such strategy, enrichment of meiosis-essential proteins was raised from 6.54% to 49.30% after prediction on the test set (Fig.3c). Moreover, at least 40% of top 10 candidates without KEGG pathway annotation were confirmed as meiosis-essential proteins by gene-knockout mouse (Fig.4c-j). Hence, functional exploration upon proteomics seems a significant improvement on prediction efficiency.

Besides, with the machine learning prediction and gene knockout validation, meiosis development of 3 genes, Zcwpw1, Tesmin and 1700102P08Rik, were found to be arrested at Pachytene in homologous gene knockout mice. Immuno-fluorescence images in this study divulged that Zcwpw1 played a role of DNA repair and synapsis. Recently, Zcwpw1 was identified as a histone H3K4me3 reader required for synapsis and repair of PRDM9-dependent DSBs by other three research groups[37-40], which was consistent with our observations. Whereas, knockout of the other two genes, Tesmin and 1700102P08Rik, were found have no effect on DNA repair and synapsis, indicating they could be involved in presently unknown molecular events inspected by midPachytene checkpoint. These meiosis-essential biological events need further exploration in the future. Additionally, a list of meiosis-essential candidates without KEGG annotation were presented here (Table.S4). Based on this list, more meiosis-essential genes could be verified and new biological events could be uncovered to the molecular basis of mouse meiosis in the future.

METHODS

Mice for experiments

Wild-type mice, C57BL/6Slac, were purchased from SLAC China. All the animal treatment were under the guidelines in the Animal Care and Use Committee at Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Science with approved Ethical review (SIBCB-NAF-14-003-s213-002).

Spermatogenesis synchronization

Spermatogenesis was synchronized as previously described with modifications[9, 19, 20]. Briefly, C57BL/6Slac mice, from P2 to P8, were fed on WIN 18,446 suspended in 1% gum tragacanth at 100 μg/g body weight, which could block spermatogonia differentiation and synchronize the spermatogenesis process. These mice were re-initiated spermatogonia differentiation at P9 through intraperitoneal injection of retinoic acid in dimethyl sulfoxide at 35 μg/g body weight. The testes of P37 to P46 mice were collected and evaluated for synchronous efficiency with histological analysis and cell sorting.

Isolation of mouse spermatocytes

Spermatocytes were isolated by fluorescent-activated cell sorting (FACS) as previously described with modifications[21]. Briefly, testes of an individual spermatogenesis-synchronized mouse were collected in GBSS. After removal of the tunica albuginea, the testes were incubated in 5 ml of DMEM containing 120 U/ml of collagenase type I at 32 °C with gentle agitation for 15 min. The dispersed seminiferous tubules were further digested with 5 ml of 0.15% trypsin and DNase I (10 μg/ml) at 32 °C for 30 min, and the digestion was terminated by adding 0.5 ml of fetal bovine serum (FBS). The suspension of dissociated testicular cells was filtered through a DMEM-prewetted cellular filter, followed by centrifugation at 500 × g for 5 min at 4 °C. The cells were resuspended in DMEM with Hoechst 33,342 (5 mg/ml) to their concentrations at 1×10^6 cells/ml, and were treated with Propidium Iodide (2mg/ml) and DNase I (10 μg/mL) in a rotator for 30 min at 32 °C at 10 rpm/min. The treated cells were centrifuged at 500× g for 5 min at 4 °C and resuspended in 1 ml DMEM for sorting based on their Hoechst 33,342 staining by FACSAria II cell sorter (BD Biosciences).

Isolation of mouse THY1+ c-KIT- spermatogonia

Undifferentiated spermatogonia (THY1+ c-KIT- spermatogonia) were isolated using magnetic activated cell sorting (MACS) with magnetic microbeads conjugated to anti-THY1 (130-049-101, Miltenyi Biotech) and anti-c-KIT (130-091-224, Miltenyi Biotech) as described previously[18]. Briefly, Testes of P7 mice were digested with collagenase type I and trypsin. After digestion, the testis cells were suspended in PBS, then layered on 2 ml 30% percoll, followed by centrifugation at 600 × g for 8 min. The cell pellets were resuspended with PBS containing anti-c-KIT magnetic microbeads. With 20 min incubation, the mixtures were loaded on magnetic device to collect c-KIT- cells. These c-KIT- cells were incubated with anti-THY1 magnetic microbeads for 20 minutes, and the THY1+ c-KIT- cells were enriched by MS columns (130-042-201, Miltenyi Biotech) and MiniMACS separator (130-042-102, Miltenyi Biotech). The purity of THY1+ c-KIT- cells were estimated by anti-PLZF (SC-22839, Santa Cruz Biotechnology) and DAPI staining.

Histological analysis

Testes were fixed in Bouin's solution, embedded in paraffin and sectioned. The sections were dewaxed with xylene and were re-hydrated by a series concentration of ethanol. The treated sections were stained with hematoxylin and eosin (H&E) and were sealed with nail polish. Spermatogenesis stages in seminiferous tubule cross-sections were recognized as previously described[36].

Meiotic chromosome spreading and immunofluorescence

Meiotic spreads were made following the protocol previously described with modifications[41]. Briefly, the cells were re-suspended in hypotonic extraction buffer, then in 100 mM sucrose. The cell suspension was pipetted onto glass slides that were coated in thin layer of 1% PFA and 0.15% Triton X-100. The slides were dried slowly in a humid chamber at room temperature. For immunofluorescence staining, the slides were washed with PBS and blocked with Tris-HCl buffer saline

containing 0.5% Tween-20 and 3% BSA for 30 min, and were incubated with different antibodies, as anti-SCP1 (ab15090, Abcam), anti-SCP3 (sc-74569, Santa Cruz), anti-SCP3 (ab15093, Abcam), anti-γH2AX (05-636, Millipore), anti-γH2AX (#9718, CST) and anti-MLH1 (550838, BD Pharmingen). Finally, the slides were incubated with Alexa Fluor 488- or 594-conjugated secondary antibodies (711545152 and 711585152, Jackson ImmunoResearch Laboratories) to detect meiotic-relevance signals, and were treated with DAPI for defining nucleus. The spread cells were monitored under confocal laser scanning microscope FV3000 (Olympus).

Protein extraction and digestion.

The mouse cells were homogenized by pipetting in lysis buffer containing 7 M urea, 2 M thiourea, 0.2% SDS, 100 mM Tris-HCl, 10 mM DTT and 1xcocktail-free protease inhibitor (Promega), pH 7.4. The proteins in lysates were reduced with 5 mM DTT and were alkylated with 55 mM IAM, then were further extracted by cold acetone precipitation. The precipitated proteins were resolved in 7 M urea lysis buffer, and the protein concentrations were estimated by Bradford protein assay (Bio-Rad). For each sample, 200 μg protein was loaded into a 10 kD spin filters (Millipore) and was centrifuged at 12,000g for 20 min. The filter was washed in order with urea lysis buffer and 1 M TEAB, and was treated with the trypsin digestion buffer (Promega) at 37°C for 16 hours, shaking at 600 rpm in a thermo mixture (Eppendorf). Tryptic peptides were collected by centrifugation and were quantified using Quantitative Colorimetric Peptide Assay (Thermo Scientific).

Peptide fractionation on RP-HPLC

For each sample, approximate 100 μg peptides were dissolved in elution buffer A containing 20 mM ammonium bicarbonate and 5% acetonitrile, pH 9.8. The dissolved peptides were loaded on a Phenomenex C18 column (5 μm particle, 110 Å pore and 250 mm × 4.6 mm) that was mounted on a Shimadzu liquid chromatography system and was pre-equilibrated with elution buffer B containing 20 mM ammonium bicarbonate and 90% acetonitrile, pH 9.8. The peptides were eluted through a stepped gradient program as follows: 0–3 min, 5% B; 3–7 min, 9% B; 7–11 min, 13% B; 11–15 min, 19% B; 15–19 min, 80% B; 19-21 min, 5% B; 21-21.5 min, 5%-80% B; 21.5-22.5 min, 80% B; 22.5-23 min, 80%-5% B and 23-29 min, 5% B at a flow of 1 ml/min. Twenty four fractions from 3-26 min were collected and these fractions were further combined to five fractions according to the absorption peaks at 260 nm during chromatography, fractions 1-5 as F1, 6-10 as F2, 11-14 as F3, 15-18 as F4 and 19-24 as F5, respectively.

Peptide detection by LC-MS/MS

Identification of peptide was conducted on a quadrupole Orbitrap mass spectrometer (Q Exactive HF, Thermo Fisher Scientific) coupled to an U3000 HPLC system (Thermo Fisher Scientific) via a nano-electrospray ion source. About 1 μg of peptides were loaded on an C18 trap column (75 μm I.D. × 1.5 cm; in house packed using Welch C18 3 μm silica beads) and were directly gone to an C18 analysis column (75 μm I.D. × 20 cm; in house packed using Welch C18 3 μm silica beads). The peptides getting into the mass spectrometer were eluted at 300 nl/min and 40 °C with two elution buffers, buffer A: 0.1% formic acid and 2% acetonitrile, and buffer B: 0.1% formic acid and 98% acetonitrile, following a gradient program, 0–5 min, 5% B, 5–7 min, 5–7% B, 7–67 min, 7-28% B, 67-80min, 28-43% B, 80-82 min, 43%-98% B, 82-84 min, 98%B, 84-85min, 5% B and 85-90 min, 5% B. The mass spectrometer was operated in "top-30" data-dependent mode, collecting MS spectra in the Orbitrap mass analyzer (120,000 resolution at 350-1500 *m/z* range) with an automatic gain control (AGC) target of 3E6 and a maximum ion injection time of 50 ms. The ions with higher intensities were isolated with an isolation width of 1.6 *m/z* and were fragmented through higher-energy collisional dissociation (HCD) with a normalized collision energy (NCE) of 28%. The MS/MS spectra were collected at 15,000 resolution with an AGC target of 1E5 and a maximum ion injection time of 45 ms. Precursor dynamic exclusion was enabled with a duration of 60 s.

Peptide search and protein quantification by Maxquant

Tandem mass spectra were searched against the 2018 Swissprot mouse databases (downloaded 11-19-2018) using MaxQuant (v 1.5.3.30)[42] with a 1% FDR at peptide and protein level. The search parameters for a peptide were set as, trypsin

digestion only, maximum of two missed cleavages of trypsin, minimum length of six amino acids, cysteine carbamidomethylation as fixed modification, N-terminal acetylation and methionine oxidations as variable modifications. The 'Match Between Runs' option was used. Label-free quantification (LFQ) was estimated with MaxLFQ algorithm, using a minimum ratio count of 1, and the specifically relative LFQ for a protein was defined by the ratios of the protein LFQ at certain sub-stage being divided by the protein maximum LFQ.

Proteomic informatics analysis

Bioinformatics analysis towards the identified and quantified proteins was performed with the Perseus software (version 1.6.1.3)[43] and R statistical software. Differentially expressed proteins (DEPs) among sub-stages were defined by two-way ANOVA analysis in Perseus filtered with adjusted q value < 0.001. To look for the DEP groups whose protein abundance changes during spermatogenesis share the similar patterns, the DEPs were first evaluated by NbClust package in R (version 3.5.1)[44] aiming at the optimum K value, and the relative LFQ of DEPs were clustered using K-means analysis. Gene Ontology and KEGG pathway analysis were performed using David GOBP (version 6.8)[45] and Enrichr[46], in which an enriched function was accepted upon p values less than 10e-8 and only the GO or pathway terms with <250 gene number in that gene sets were included. Heatmap data was visualized by Seaborn 0.9.0[47].

Machine learning for prediction of meiosis-essential proteins

Three sub-classifiers, regularized RBF, NBM or SVM, were employed to gain a better prediction for meiosis-essential protein candidates, respectively. Protein abundance in 9 types of germ cells were inputted to the classifiers to train the weights. The meiosis-essential proteins defined as the proteins overlapped MGI database and proteomics in this study were labeled as positive and non-essential as negative. The ensemble learning process was broadly conducted in the two steps as described previously with modifications[48], the details of regularized RBF, NBM and SVM were presented in Supplementary methods:

Step1. All the labeled proteins were randomly divided into two sets, train (80%) and test (20%). One single sub-classifier was constructed by the train set and the corresponding output function were generated for prediction. The predicted score for the proteins in test set were estimated by the output function. This process was repeated for 1,000 times to construct the ensemble algorithm.

Step2. To test the performance of the algorithm, a Monte-Carlo cross validation was applied[29]. The final predicted score of each protein, called meiotic confidence score, is defined the mean value of the individual scores in step1. Precision-recall curve and receiver operating characteristic (ROC) curve were used to evaluate the prediction performance of the three algorithms based on regularized RBF, NBM and SVM.

The FuncProFinder Matlab package was uploaded on: https://github.com/sjq111/FuncProFinder.

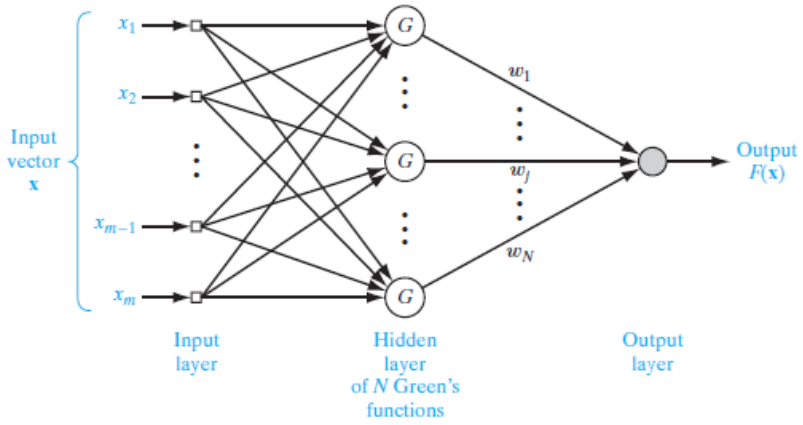
Generation of Gene Knockout Mice with CRISPR/Cas9 System.

For generation of knockout mice corresponding to the genes interested by this study, sgRNAs were designed upon their genome structures that were listed in Table S5. T7-Cas9 PCR product was gel purified and used as the template for in vitro transcription (IVT) using mMESSAGE mMACHINE™ T7 ULTRA transcription kit (AM1345, ThermoFisher Scientific). The T7-sgRNA PCR product was gel purified and used as the template for IVT using MEGAshortscript™ T7 transcription kit (AM1354, ThermoFisher Scientific). Both the Cas9 mRNA and the sgRNAs were purified using MEGAclear™ transcription clean-Up kit (AM1908, ThermoFisher Scientific) and eluted in RNase-free water (10977015, ThermoFisher Scientific). The Cas9 mRNA and sgRNAs were injected to one-cell embryos as described previously[49-51]. The injected embryos were cultured in vito to develop to 2-cell embryos and transplanted to oviducts to generate knockout pups. After pups were born, genotyping was performed by direct sequencing following PCR to validate the knockout consequences. Genotyping primer sequences that were used are listed in Table S5.

Supplementary methods

Regularized radial basis function network (Regularized RBF Network)

Regularized radial basis function network[27] was applied as one type of sub-classifiers for meiosis-essential protein prediction. The structure of this network is shown below,



In this paper, m=9 and $\boldsymbol{x_1, x_2, \cdots, x_m}$ represented protein abundance of 9 types of germ cells. And N, the number of Green's functions in the hidden layer, is equal to the number of proteins in training set. Several regularization methods could be applied to avoid the over fitting problem caused by the noise and uncertainty of data. In this research, we applied Tikhonov's regularization method. Based on the method, the loss function of the network contains two terms shown below.

$$\mathcal{E}(F) = \mathcal{E}_s(F) + \lambda \mathcal{E}_c(F)$$

$$\mathcal{E}_s(F) = \frac{1}{2}\sum_{i=1}^{N}\left(d_i - F(\boldsymbol{x_i})\right)^2 \text{ and } \mathcal{E}_c(F) = \frac{1}{2}\|\boldsymbol{D}F\|^2$$

Where $\mathcal{E}_s(F)$ is the standard error term and $\mathcal{E}_c(F)$ is the regularizing term, λ is the regularization parameter and $\boldsymbol{D}$ is a linear differential operator.

In this research,

$$G(\boldsymbol{x}, \boldsymbol{x_i}) = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{x_i}\|^2} \text{ is the Green functions}$$

$$\boldsymbol{D} = \sum_n {\alpha_n}^{\frac{1}{2}}\left(\frac{\partial}{\partial x_1} + \frac{\partial}{\partial x_2} + \cdots + \frac{\partial}{\partial x_m}\right)^n \text{ where } \alpha_n = \frac{\sigma^{2n}}{n!\,2^n}$$

In this research $\sigma = 1$ and we estimated the optimal choice of λ was about 3×10^{-4}. To avoid affection of extreme values of RBF-predicted scores, we applied $F^* = tanh(F)$ to constrained the output scores into$[-1,1]$.

Naive Bayesian Model (NBM)

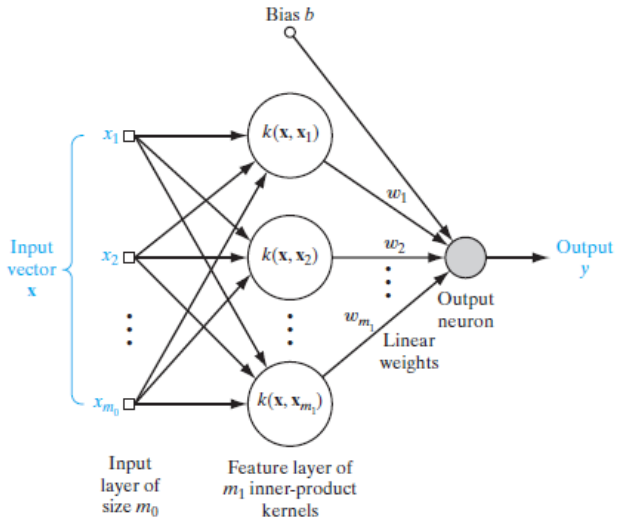
Naive Bayesian Model was applied as the second type of sub-classifiers for meiosis-essential protein prediction. $\boldsymbol{x} = (x_1, x_2, \cdots, x_9)$ represented expression abundance of each protein in 9 types of cells. Two different classes of data, meiosis-essential and non-essential, were denoted to be C_1 and C_2 respectively. This sub-classifier is supposed to estimate

$P(\mathcal{C}_1 |\boldsymbol{x})$ and $P(\mathcal{C}_2 |\boldsymbol{x})$ based on Bayes formula.

Bayes formula claims that $P(\mathcal{C}_k |\boldsymbol{x}) = \frac{\prod_{i=1}^{9} P(x_i|\mathcal{C}_k)P(\mathcal{C}_k)}{P(\boldsymbol{x})}$ where k=1 or 2. We use $P(x_i|\mathcal{C}_k) = \frac{s_{ki}}{s_k}$ and $P(\mathcal{C}_k) = \frac{s_k}{s}$ to estimate the corresponding probabilities, where s_k is the number of $\mathcal{C}_k$ data in the training set, s is the number of data in the training set and s_{ki} is the number of $\mathcal{C}_k$ data in the training set which has component x_i. As the data in this research is continuous, we divided each component into 12 classes. $P(\boldsymbol{x})$ as the common denominator is not considered, since the $P(\mathcal{C}_k |\boldsymbol{x})$ can be calculated by normalization.

Support Vector Machine (SVM)

Support Vector Machine (SVM) was applied as the third type of sub-classifiers for meiosis-essential protein prediction[28]. The structure of the sub-classifier is shown below.



In this research, m_0=9 and $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{m_0}}$ represented expression abundance of each protein in 9 types of cells, $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_{m_1}}$ were expression abundance of all the proteins in the training set. We choose bias $b = 0$ and the kernel function $k(\boldsymbol{x}, \boldsymbol{x_i}) = e^{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{x_i}\|^2}$ where $\sigma = 1$. In this research, we trained weights by solving the soft margin SVM problem shown below.

$$\min_{\boldsymbol{\omega}} \left(\frac{\lambda}{2}\|\boldsymbol{\omega}\|^2 + \frac{1}{m_1}\sum_{i=1}^{m_1} max\{0,\ 1 - d_i\langle \boldsymbol{\omega}, \boldsymbol{\psi}(\boldsymbol{x_i})\rangle\} \right)$$

Where $d_i \in \{-1,1\}$ is the expected output, and $\langle \boldsymbol{\psi}(\boldsymbol{x}), \boldsymbol{\psi}(\boldsymbol{x_i})\rangle = k(\boldsymbol{x}, \boldsymbol{x_i})$.

We applied Stochastic Gradient Descent (SGD) method[52] to solve the problem and let $\lambda = 1$.

To avoid affection of extreme values of SVM-predicted scores, we applied $y^* = tanh(y)$ to constrained the output scores into$[-1,1]$.

Data availability.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[53] partner

repository with the dataset identifier PXD017284

ACKNOWLEDGEMENTS

We thank Dr. Dangsheng Li (Shanghai Institute of Biochemistry and Cell Biology) for helpful discussion of the manuscript, Degui Chen was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB19000000), and the National Natural Science Foundation of China (91753128, 81772472). Liu Siqi was supported by the National Key R&D Program of China (2017YFC0908400 and 2017YFC0906703), the National Natural Science Foundation of China (NO.31700728) and the Shenzhen Engineering Laboratory for Proteomics (DRC-SZ[2016]749). Yang Hui was supported by R&D Program of China (2018YFC2000100 and 2017YFC1001302), CAS Strategic Priority Research Program (XDB32060000), National Natural Science Foundation of China (31871502, 31522037), Shanghai Municipal Science and Technology Major Project (2018SHZDZX05), Shanghai City Committee of science and technology project (18411953700, 18JC1410100). We thank the Histology, Flow Cytometry, and Animal services at SIBCB.

AUTHOR CONTRIBUTIONS

D.C. and K.F conceived the project. K.F performed the synchronization of mouse spermatogenesis, isolated different types of germ cells, validated the phenotype of knockout mice and contributed to the manuscript; W.G performed the experiments to produce the proteomic data and contributed to proteomic data analysis; Q.L performed the proteomic bioinformatic data analysis and wrote the manuscript; J.S developed the machine learning package FuncProFinder; C.Z, Y.W, W.Y and R.W constructed all the knockout mouse models. L.Y and Y.Z helped to the visualization of proteomic data. H.Y supervised the knockout mouse production; S.L supervised the proteomic data production, data analysis and refined the manuscript; D.C supervised the project. All authors contributed to the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCE

1. Cahoon, C.K. & Hawley, R.S. Regulating the construction and demolition of the synaptonemal complex. *Nat Struct Mol Biol* **23**, 369-377 (2016).

2. Keeney, S., Lange, J. & Mohibullah, N. Self-organization of meiotic recombination initiation: general principles and molecular pathways. *Annu Rev Genet* **48**, 187-214 (2014).

3. Romanienko, P.J. & Camerini-Otero, R.D. The mouse Spo11 gene is required for meiotic chromosome synapsis. *Mol Cell* **6**, 975-987 (2000).

4. Baudat, F., Manova, K., Yuen, J.P., Jasin, M. & Keeney, S. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol Cell* **6**, 989-998 (2000).

5. Bannister, L.A. *et al.* A dominant, recombination-defective allele of Dmc1 causing male-specific sterility. *PLoS Biol* **5**, e105 (2007).

6. 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).

7. Wang, L. *et al.* Proteomic Analysis of Pachytene Spermatocytes of Sterile Hybrid Male Mice. *Biol Reprod* **95**, 52 (2016).

8. Miyata, H. *et al.* Genome engineering uncovers 54 evolutionarily conserved and testis-enriched genes that are not required for male fertility in mice. *Proc Natl Acad Sci U S A* **113**, 7704-7710 (2016).

9. Chen, Y. *et al.* Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res* **28**, 879-896 (2018).

10. Green, C.D. *et al.* A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Dev Cell* **46**, 651-667 e610 (2018).

11. Lukassen, S., Bosch, E., Ekici, A.B. & Winterpacht, A. Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific reports* **8**, 6521 (2018).

12. Jung, M. *et al.* Unified single-cell analysis of testis gene regulation and pathology in five mouse strains. *Elife* **8** (2019).

13. Ernst, C., Eling, N., Martinez-Jimenez, C.P., Marioni, J.C. & Odom, D.T. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat Commun* **10**, 1251 (2019).

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

15. Hermann, B.P. *et al.* The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. *Cell Rep* **25**, 1650-1667 e1658 (2018).

16. Cagney, G. *et al.* Human tissue profiling with multidimensional protein identification technology. *J Proteome Res* **4**, 1757-1767 (2005).

17. Gan, H. *et al.* Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. *Mol Cell Proteomics* **12**, 1144-1157 (2013).

18. Kubota, H., Avarbock, M.R. & Brinster, R.L. Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells. *Proc Natl Acad Sci U S A* **101**, 16489-16494 (2004).

19. Hogarth, C.A. *et al.* Turning a spermatogenic wave into a tsunami: synchronizing murine spermatogenesis using WIN 18,446. *Biol Reprod* **88**, 40 (2013).

20. Romer, K.A., de Rooij, D.G., Kojima, M.L. & Page, D.C. Isolating mitotic and meiotic germ cells from male mice by developmental synchronization, staging, and sorting. *Dev Biol* **443**, 19-34 (2018).

21. Gaysinskaya, V., Soh, I.Y., van der Heijden, G.W. & Bortvin, A. Optimized flow cytometry isolation of murine spermatocytes. *Cytometry A* **85**, 556-565 (2014).

22. Bellani, M.A., Boateng, K.A., McLeod, D. & Camerini-Otero, R.D. The expression profile of the major mouse SPO11 isoforms indicates that SPO11beta introduces double strand breaks and suggests that SPO11alpha has an additional role in prophase in both spermatocytes and oocytes. *Mol Cell Biol* **30**, 4391-4403 (2010).

23. Chakraborty, P. *et al.* LIN28A marks the spermatogonial progenitor population and regulates its cyclic expansion. *Stem Cells* **32**, 860-873 (2014).

24. Kleene, K.C. & Flynn, J.F. Characterization of a cDNA clone encoding a basic protein, TP2, involved in chromatin condensation during spermiogenesis in the mouse. *J Biol Chem* **262**, 17272-17277 (1987).

25. Zhou, Q. *et al.* Expression of stimulated by retinoic acid gene 8 (Stra8) in spermatogenic cells induced by retinoic acid:

an in vivo study in vitamin A-sufficient postnatal murine testes. *Biol Reprod* **79**, 35-42 (2008).

26. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat Neurosci* **19**, 1454-1462 (2016).

27. Poggio, T. & Girosi, F. Networks for approximation and learning. *Proceedings of the IEEE* **78**, 1481-1497 (1990).

28. Cortes, C. & Vapnik, V. Support-Vector Networks. *Machine Learning* **20**, 273-297 (1995).

29. Picard, R.R. & Cook, R.D. Cross-validation of regression models. *Journal of the American Statistical Association* **79**, 575-583 (1984).

30. Grootegoed, J.A., Jansen, R. & Van der Molen, H.J. The role of glucose, pyruvate and lactate in ATP production by rat spermatocytes and spermatids. *Biochim Biophys Acta* **767**, 248-256 (1984).

31. Patel, M.S., Nemeria, N.S., Furey, W. & Jordan, F. The pyruvate dehydrogenase complexes: structure-based function and regulation. *J Biol Chem* **289**, 16615-16623 (2014).

32. Takakubo, F. & Dahl, H.H. The expression pattern of the pyruvate dehydrogenase E1 alpha subunit genes during spermatogenesis in adult mouse. *Exp Cell Res* **199**, 39-49 (1992).

33. Roeder, G.S. & Bailis, J.M. The pachytene checkpoint. *Trends Genet* **16**, 395-403 (2000).

34. Ashley, T., Gaeth, A.P., Creemers, L.B., Hack, A.M. & de Rooij, D.G. Correlation of meiotic events in testis sections and microspreads of mouse spermatocytes relative to the mid-pachytene checkpoint. *Chromosoma* **113**, 126-136 (2004).

35. Shao, B. *et al.* Unraveling the proteomic profile of mice testis during the initiation of meiosis. *J Proteomics* **120**, 35-43 (2015).

36. Ahmed, E.A. & de Rooij, D.G. Staging of mouse seminiferous tubule cross-sections. *Methods Mol Biol* **558**, 263-277 (2009).

37. Huang, T. *et al.* The histone modification reader ZCWPW1 links histone methylation to repair of PRDM9-induced meiotic double stand breaks. *bioRxiv*, 836023 (2019).

38. Li, M. *et al.* The histone modification reader ZCWPW1 is required for meiosis prophase I in male but not in female mice. *Science advances* **5**, eaax1101 (2019).

39. Mahgoub, M. *et al.* Dual Histone Methyl Reader ZCWPW1 Facilitates Repair of Meiotic Double Strand Breaks. *bioRxiv*, 821603 (2019).

40. Wells, D. *et al.* ZCWPW1 is recruited to recombination hotspots by PRDM9, and is essential for meiotic double strand break repair. *bioRxiv*, 821678 (2019).

41. Peters, A.H., Plug, A.W., van Vugt, M.J. & de Boer, P. A drying-down technique for the spreading of mammalian meiocytes from the male and female germline. *Chromosome Res* **5**, 66-68 (1997).

42. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372 (2008).

43. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*

13, 731-740 (2016).

44. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *2014* **61**, 36 %J Journal of Statistical Software (2014).

45. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57 (2009).

46. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-W97 (2016).

47. Waskom, M. *et al.* mwaskom/seaborn: v0. 9.0 (July 2018). Zenodo. (2018).

48. Breiman, L. Bagging predictors. *Machine learning* **24**, 123-140 (1996).

49. Wang, H. *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* **153**, 910-918 (2013-05).

50. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370-1379 (2013).

51. Yang, H., Wang, H. & Jaenisch, R. Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. *Nat Protoc* **9**, 1956-1968 (2014).

52. Bottou, L. Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010* 177-186 (Springer, 2010).

53. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaíno JA (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47(D1):D442-D450 (PubMed ID: 30395289).

Fig. 1

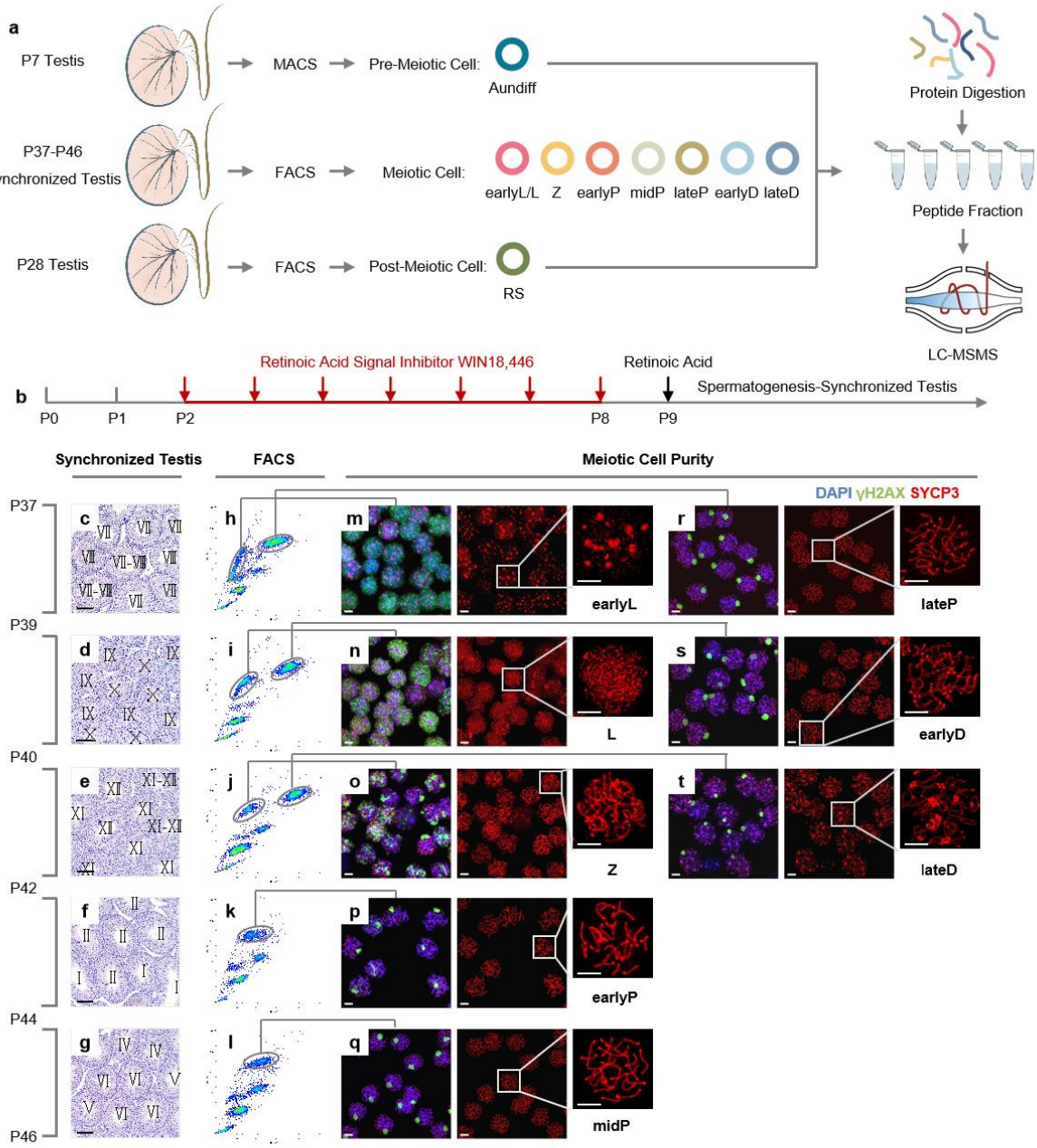


Table 1. The purity of the prepared and separated spermatocytes at different stages of meiosis

Cell type	Mouse Num.	Cell Num. (×10⁶)	Cell Purity		Cell type	Mouse Num.	Cell Num. (×10⁶)	Cell Purity	
earlyL	9	4.99	earlyL	95.36%	midP	6	5.78	midP	97.99%
			L	4.64%				earlyP	2.01%
L	5	4.34	L	78.57%	lateP	9	21.16	lateP	96.62%
			earlyL	21.43%				earlyD	3.38%
Z	7	7.04	Z	88.40%	earlyD	3	6.63	earlyD	90.59%
			earlyP	11.60%				lateP	9.41%
earlyP	8	6.06	earlyP	93.64%	lateD	7	7.27	lateD	96.38%
			Z	6.36%				Diak.	3.62%

Fig.1 Isolation of spermatocytes during meiotic prophase I. a Workflow from the germ cell collection to mass spectrometry analysis. **b** The strategy to obtain mice with synchronized spermatogenesis. **c-g** Cross sections of Hematoxylin and Eosin (H&E) stained testes from spermatogenesis-synchronized mice on Postnatal Day 37 (P37) to P46. Roman numerals in each tubule designate the stage represented by the cellular constitution with the criteria described previously[36]. Scale bar = 50 μm. h-l FACS plots of Hoechst 33,342 stained testes cells from the mice whose spermatogenesis were synchronized to the corresponding stages in c-g. **m-t** Chromatin spreading of FACS-sorted spermatocytes, that were co-immuno-stained with DAPI (blue), anti-γH2AX (green) and anti-SYCP3 (red). Scale bar = 5 μm.

Fig. 2

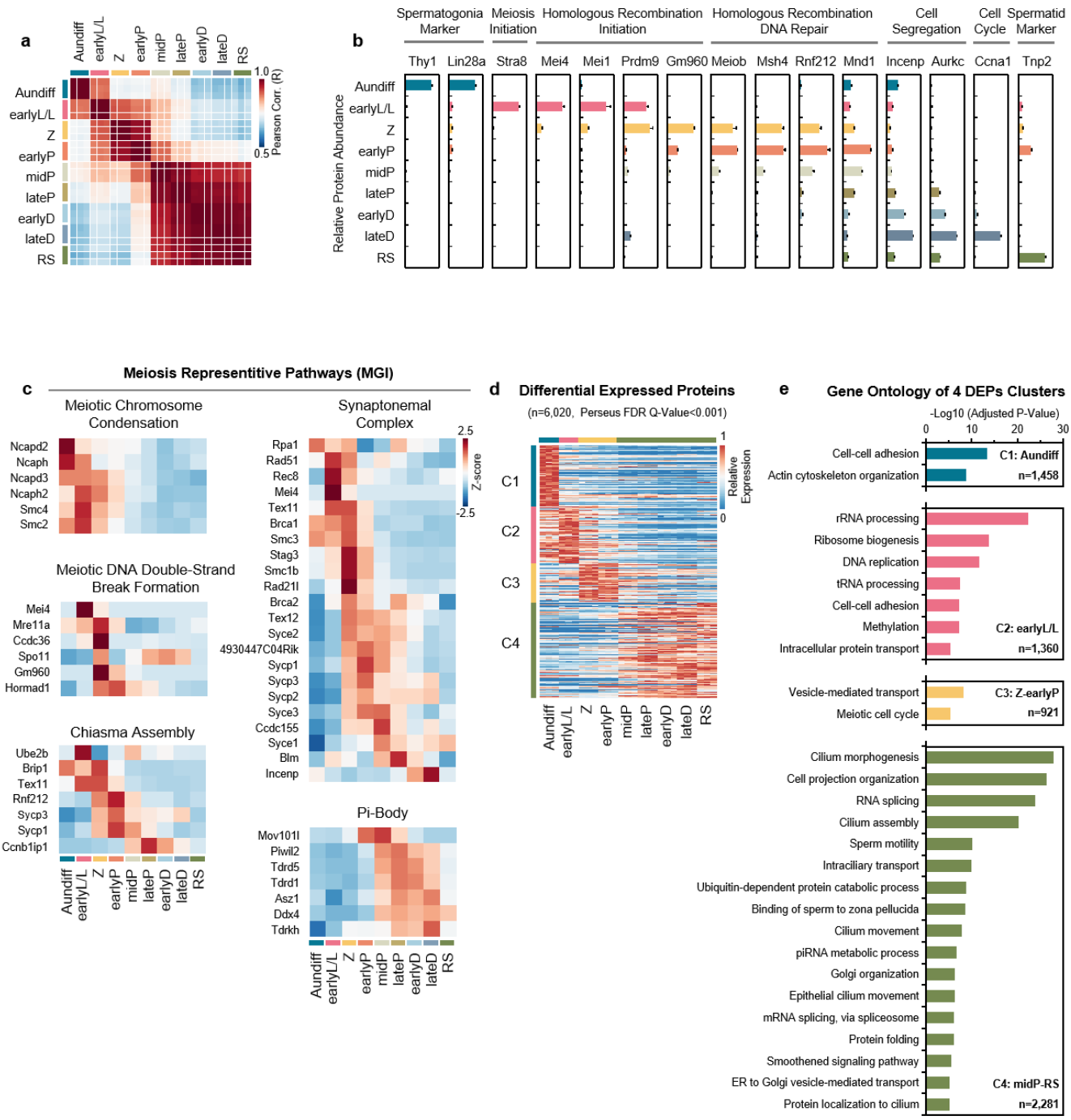


Fig.2 Proteomic informatics of mouse germ cells around meiotic prophase I. a Pearson correlation coefficients of the protein LFQ intensity of all quantified proteins among nine stages of germ cells. Color key represents the value of Pearson correlation coefficient. **b** The relative abundance of the typical biomarkers of germ cells in response to spermatogenesis. Error bars represent SEM in triplicates. **c** Heatmap of dynamic abundance of the proteins involved in five representative meiotic pathways (MGI database). **d** K-means clusters of relative protein abundance elicited from the DEPs. Color key represents the relative protein abundance. **e** Gene ontology (GO) analysis of the enriched biological processes for four DEPs clusters.

Fig. 3

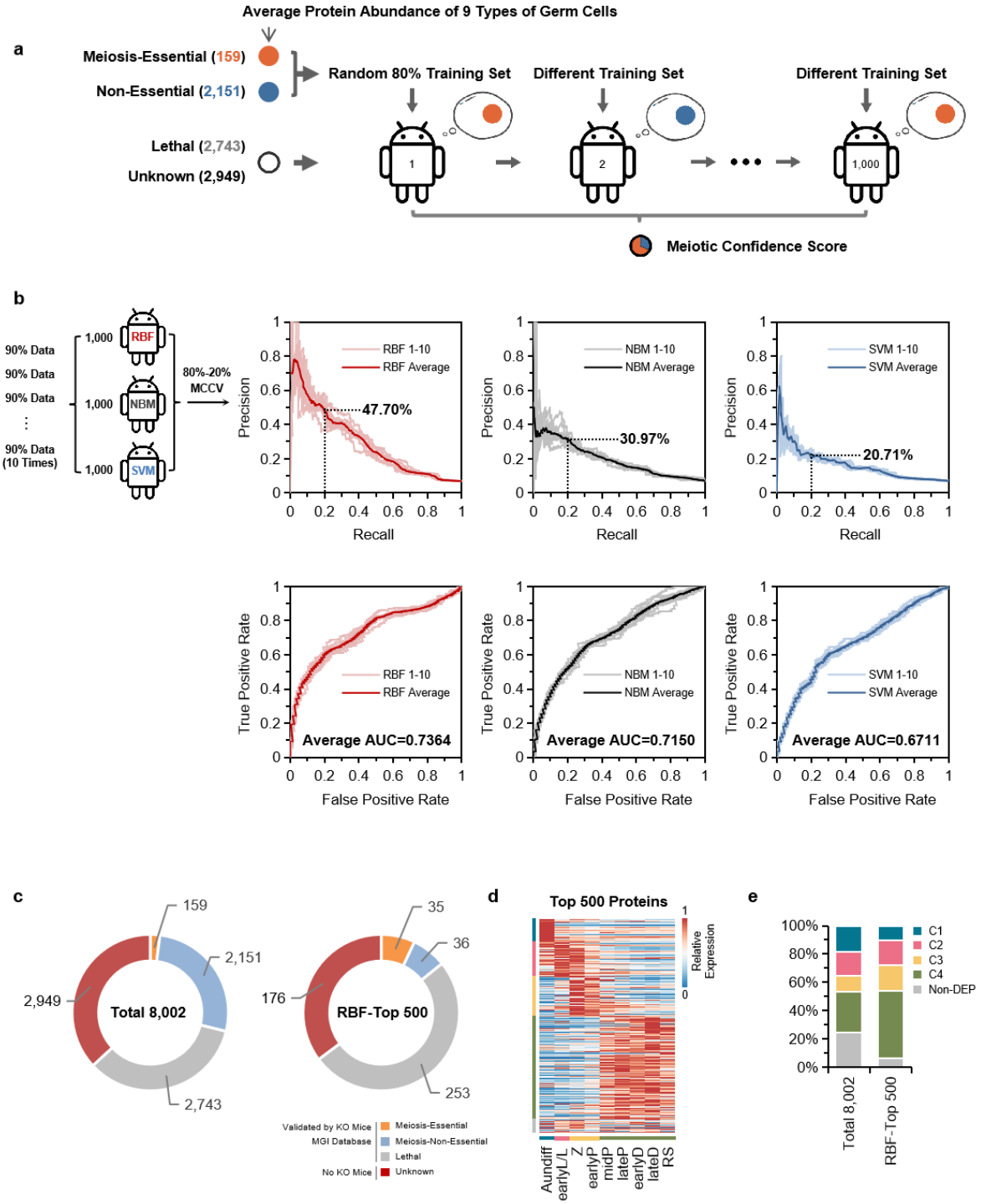


Fig.3 Prediction of meiotic-essential cadidates using machine learning. a Workflow to build machine learning algorithms based on proteome data. **b** Comparison of the prediction results, Precision-Recall curve (upper panels) and ROC curve (lower panels) generated from three sub-classifiers, regularized RBF, NBM and SVM. **c** On the basis of annotation of MGI database, the distribution of meiosis-essential, meiosis-non-essential, lethal and unknown proteins in 8,002 quantified proteins (left panel) and the top 500 meiosis candidates derived from RBF prediction (right panel). **d** Heatmap of dynamic abundance of the top 500 meiosis candidates derived from RBF prediction. **e** Comparison of the relative distribution of the DEP clusters treated with/without RBF filtration.

Fig. 4

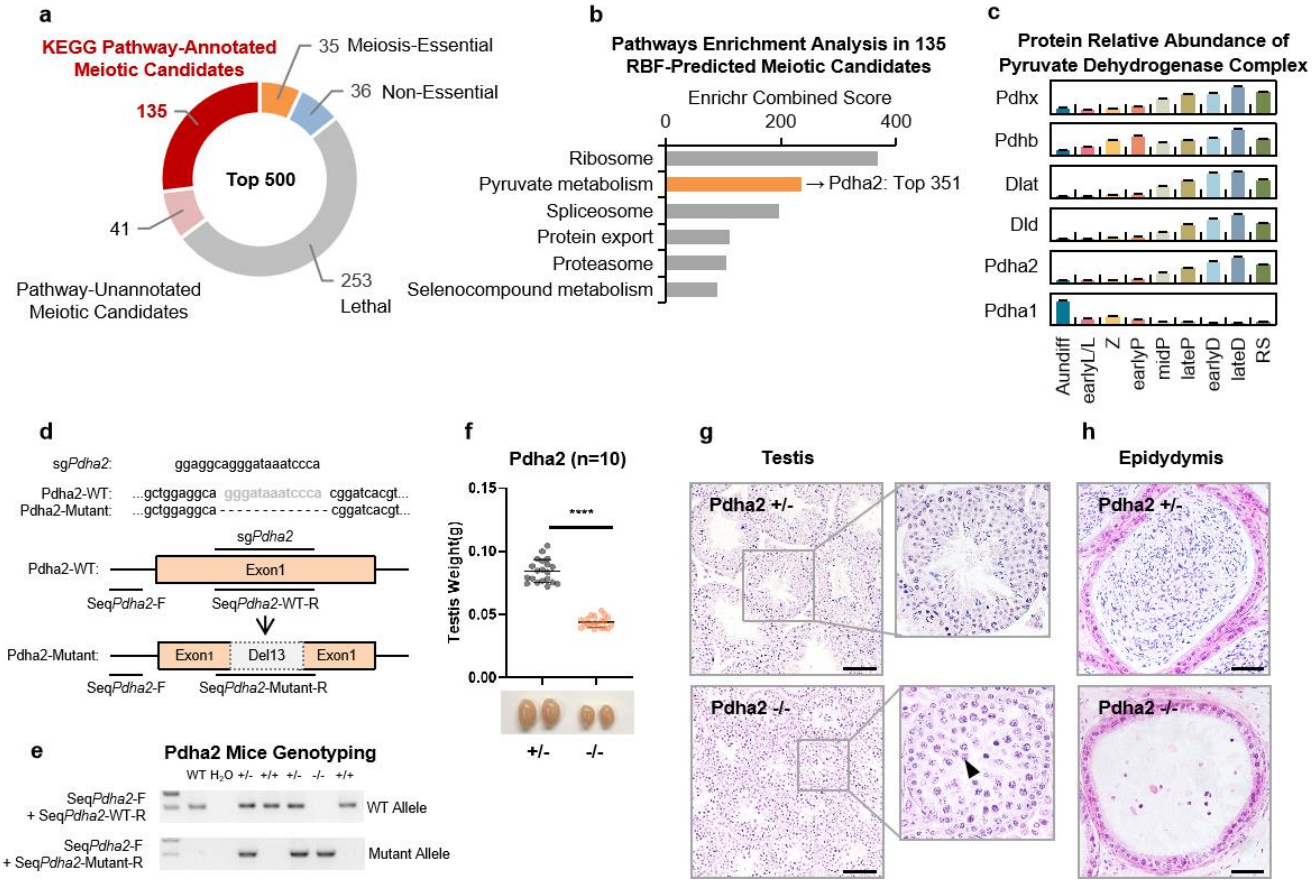


Fig.4 Phenotypic validation of the PDHA2 during meiosis. a The proteins with clearly functional annotation in the 176 proteins without KO evidence in MGI database. **b** The KEGG pathways enriched in the proteins indicated in Fig.4a. **c** The dynamically relative abundance of the 6 proteins of the PDC complex. Error bar represents SEM in triplicates. **d** The strategy to construct Pdha2 knockout mouse. **e** The genotyping of Pdha2 knock out mice verified by PCR. **f** Comparison of the testes weight between Pdha2$^{+/-}$ and Pdha2$^{-/-}$ mice (n=10, unpaired two-tailed t test, p< 0.0001). **g** Cross-sections of H&E stained seminiferous tubules from 8-weeked Pdha2$^{+/-}$ (upper panel) and Pdha2$^{-/-}$ mice (lower panel), insets denote the specific one seminiferous tubule under higher magnification. Arrow points to a Pachytene-like cell. **h** Cross-sections of epididymis from 8-weeked Pdha2$^{+/-}$(upper panel) and Pdha2$^{-/-}$ (lower panel) mice stained with H&E. Scale bar = 50 μm.

Fig. 5

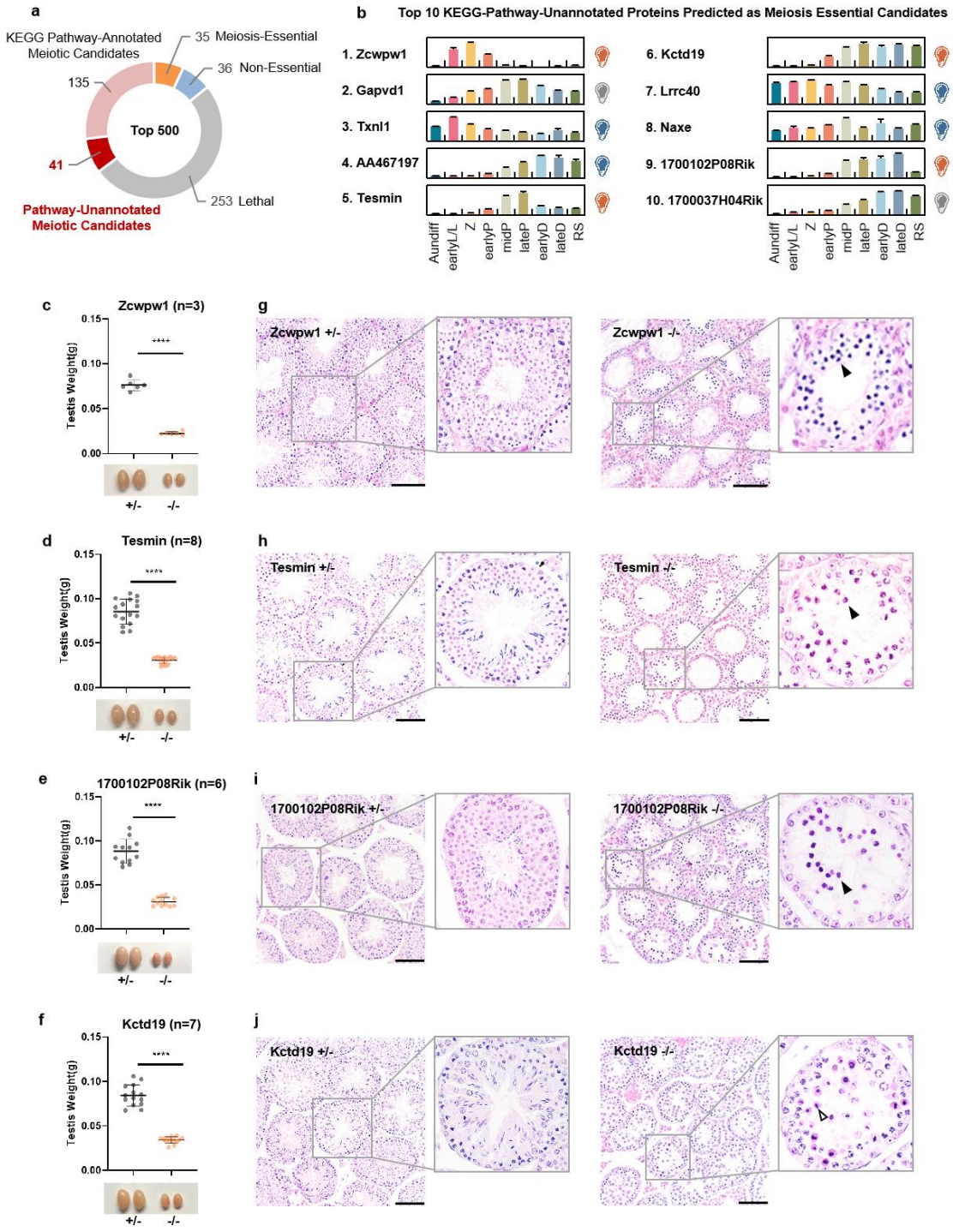


Fig.5 Phenotypic validation of the meiotic-essential proteins predicted by RBF. a The proteins without functional annotation in the 176 proteins without KO evidence in MGI database. **b** The dynamical relative abundance of the top 10 RBF-ranked candidates during spermatogenesis. Error bar represents SEM in triplicates. The bulbs on the figure right indicate the phenotype of KO mice, orange as meiosis-essential, blue as meiosis non-essential and grey as lethal. **c-f** Comparison of testis weights derived from 8-weeked Zcwpw1+/- and Zcwpw1-/- mice (**c**), Tesmin+/- and Tesmin -/- mice (**d**), Kctd19+/- and Kctd19-/- mice (**e**), and 1700102P08Rik+/- and 1700102P08Rik -/- mice (**f**). **** represents $p< 0.0001$ in unpaired two-tailed t test. **g-j**. Cross-sections of H&E stained seminiferous tubules from the heterozygous and homozygous knockout mice of the four genes above, insets denote the specific one seminiferous tubule under higher magnification. Filled arrow points to a Pachytene-like cell. Hollow arrow points to a Metaphase I-like cell. Scale bar = 50 μm.

Fig. 6

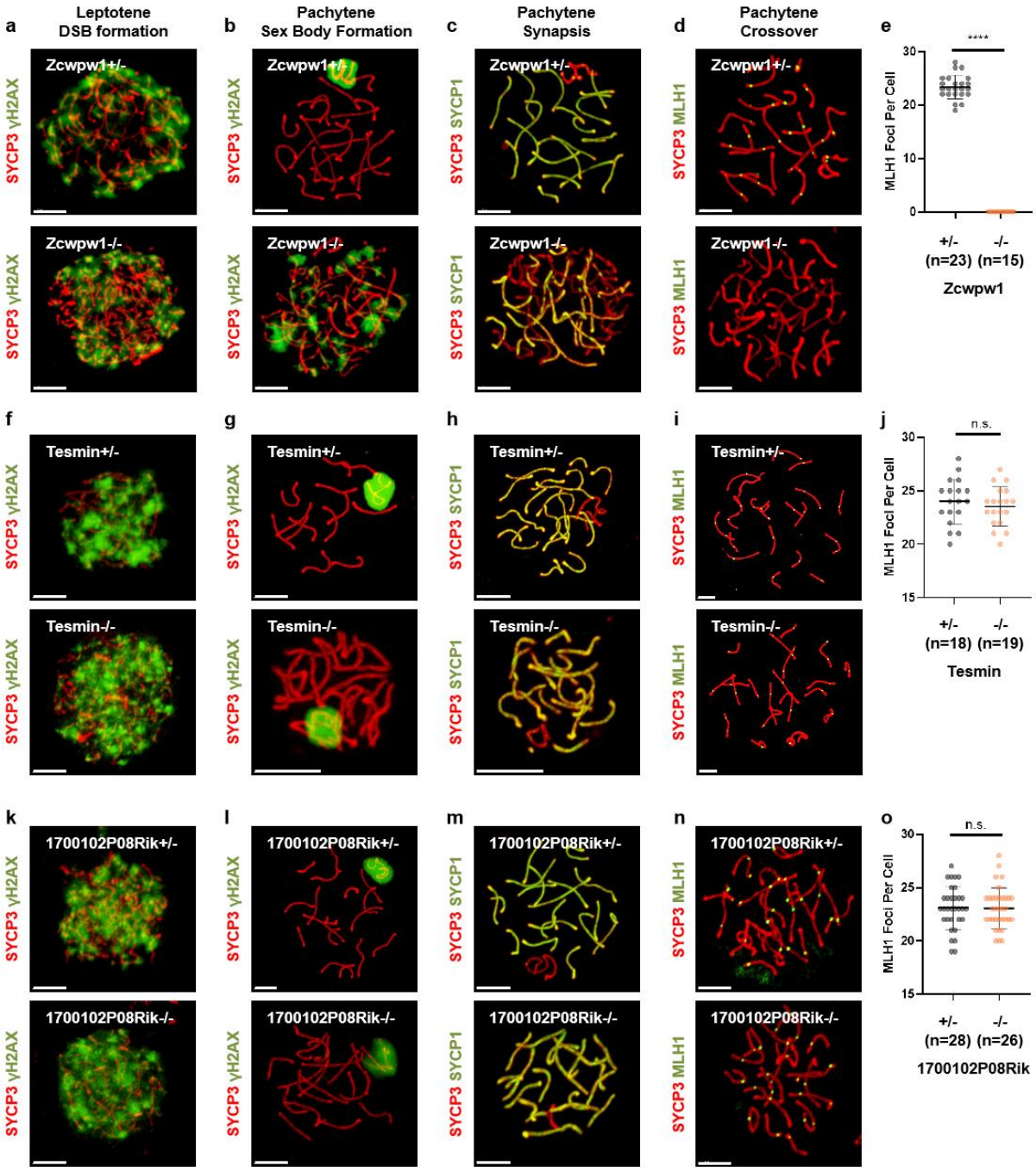


Fig.6 Analysis of the homologous recombination and synapsis states by confocal images in the heterozygous and homozygous gene knockout mice, Zcwpw1 (a-d), Tesmin (f-i) and 1700102P08Rik (k-n). The confocal images in two columns of left side, the cells were co-immuno-stained with anti-SYCP3 (red) and anti-γH2AX (green), in the right column with anti-SYCP3 (red) and anti-SYCP1 (green), and in the most right column with anti-SYCP3 (red) and anti-MLH1 (green). Scale bar = 50 μm. m-o Comparison of MLH1 foci number in spermatocytes derived from Zcwpw1$^{+/+}$ and Zcwpw1$^{-/-}$ (**e**), Tesmin $^{-/-}$ and Tesmin $^{+/-}$(**j**), 1700102P08Rik $^{+/-}$ and 1700102P08Rik $^{-/-}$ (**o**) mice. **** represents p< 0.0001 in unpaired two-tailed t test and n.s. means no significance.

Fig. S1

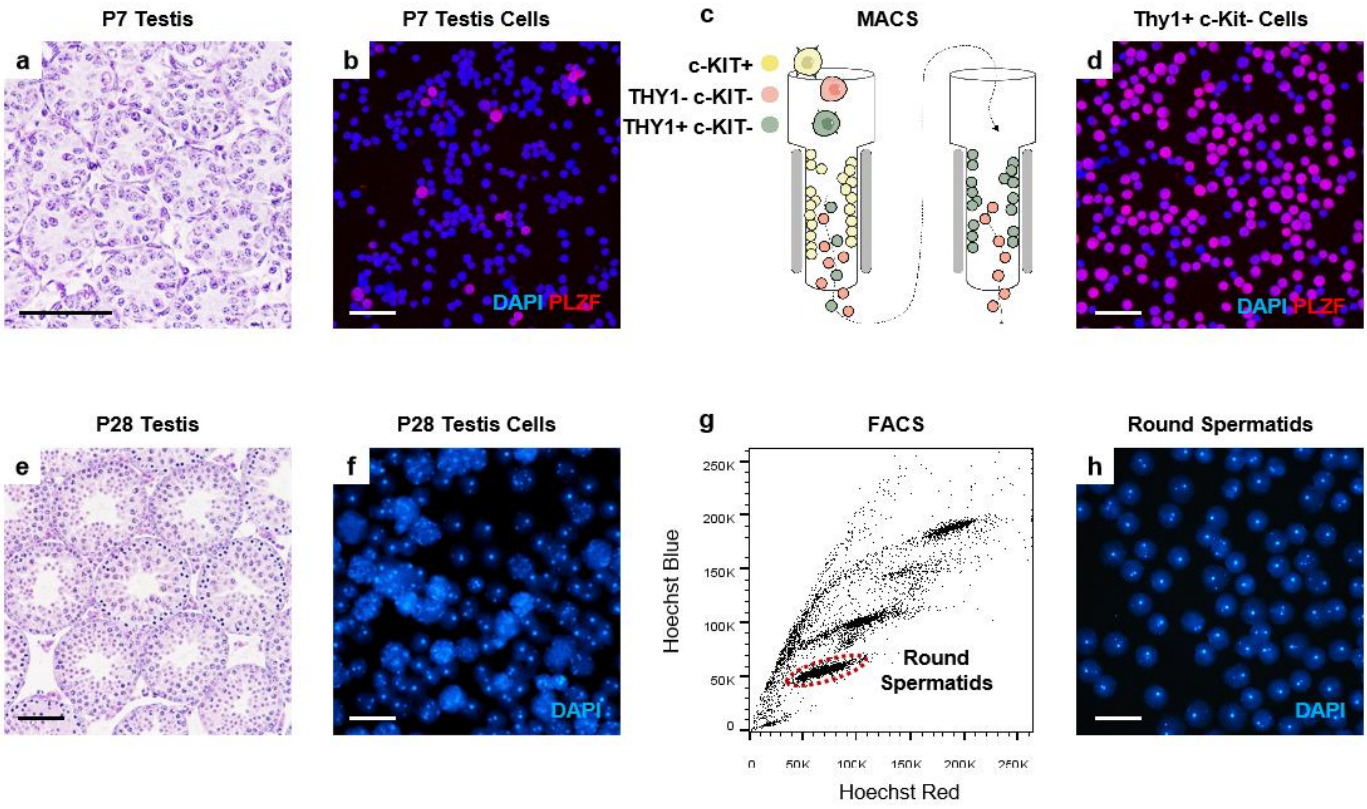


Table S1. The purity of the prepared and separated Aundiff and RS

Cell type	Mouse Num.	Cell Num. (×10⁶)	Cell Purity	
Aundiff	96	2.97	PLZF+	79.38%
			PLZF-	20.62%
RS	10	20.85	RS	100.00%

Fig.S1 Isolation of undifferentiated spermatogonia (Aundiff) and round spermatids (RS). **a** Cross section of H&E stained testes from P7 mouse. **b** DAPI (blue) and anti-PLZF (red) immuno-staining of digested P7 testes cells before MACS purification. **c** Illustration of magnetic activated cell sorting (MACS) strategy to isolate THY1+ c-KIT- Aundiff. Differentiated spermatogonia and other testes somatic cells expressed c-KIT (c-KIT+ cells, showed as yellow cells) were depleted first by binding to the anti-c-KIT antibody and magnetic columns. The unbinded c-KIT- cells were subsequently separated by MACS to enrich the THY1+ c-KIT- cells (showed as green cells). **d** DAPI (blue) and anti-PLZF (red) immuno-staining of THY1+ c-KIT- cells after MACS purification. **e** Cross section of H&E stained testes of P28 mouse. **f** DAPI staining of digested P28 testes cells before FACS purification. **g** FACS plot of digested testes cells stained with Hoechst 33,342. Gate for sorting of haploid round spermatids was circled. **h** DAPI staining of round spermatids after FACS purification. All scale bars = 50 μm.

Fig. S2

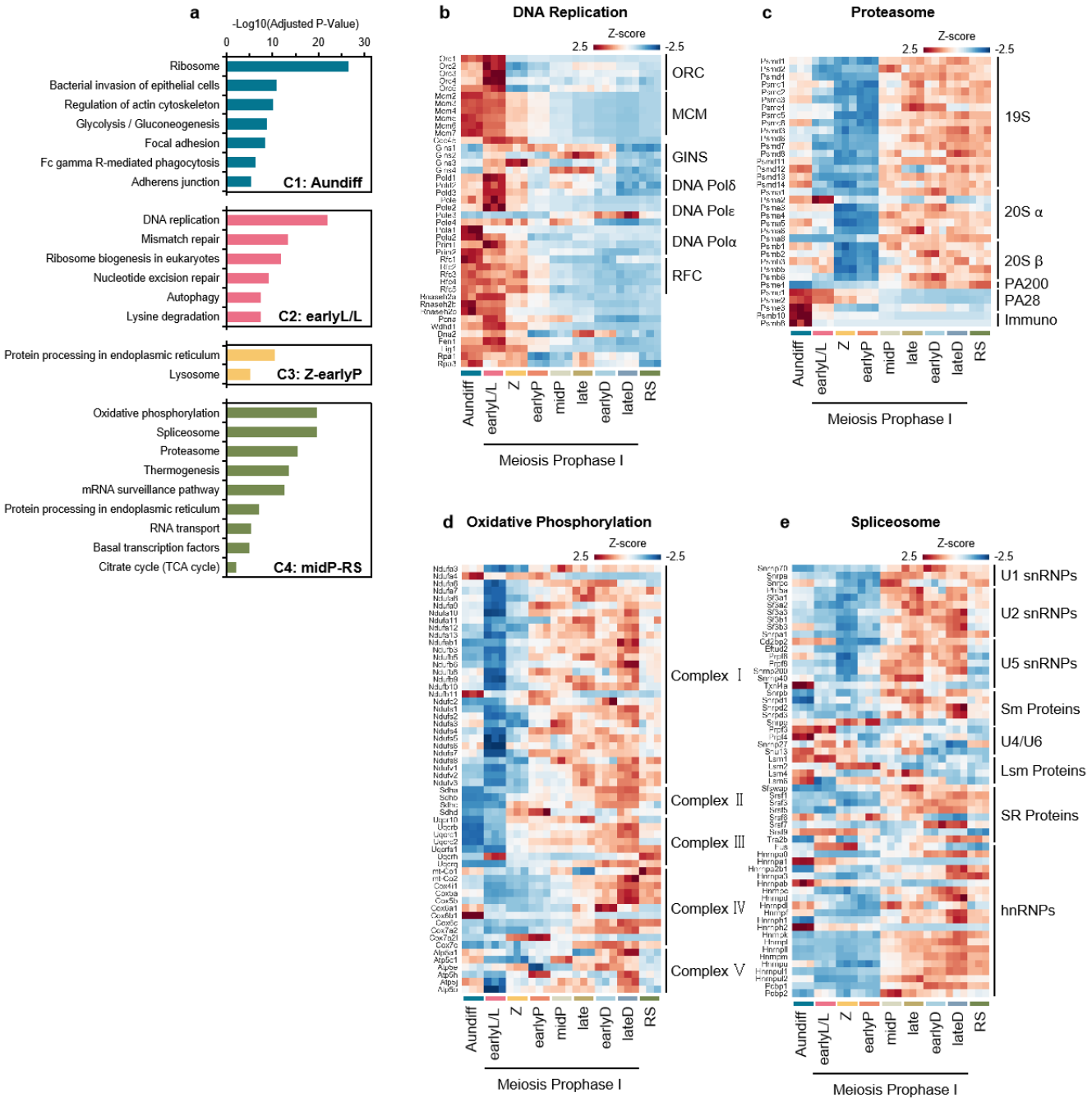


Fig.S2 Enriched KEGG pathways for four DEPs clusters. a KEGG pathway analysis of the enriched functions for four DEPs clusters. **b-e** Heatmap of dynamic abundance of the proteins involved in DNA replication (**b**), proteasome (**c**), oxidative phosphorylation (**d**) and spliceosome (**e**). Color key represents the Z-score of relative protein abundance.

Fig. S3

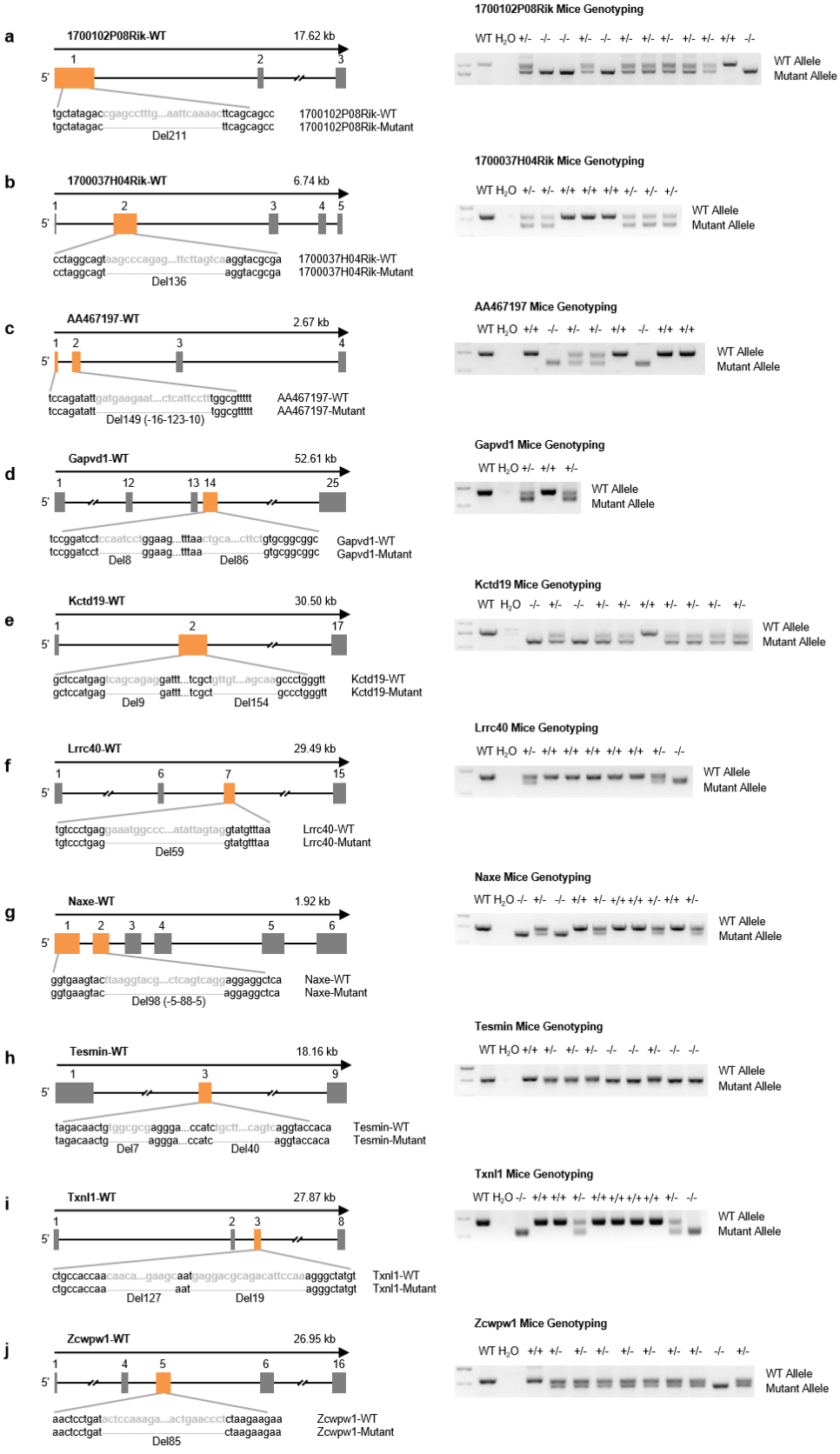


Fig.S3 Construction and genotyping of 10 knock out mice of the top 10 RBF-ranked candidates. a-j The strategy to construct knockout mouse of the top 10 RBF-ranked candidates (left panel) and the genotyping of knock out mice verified by PCR (right panel).

Fig. S4

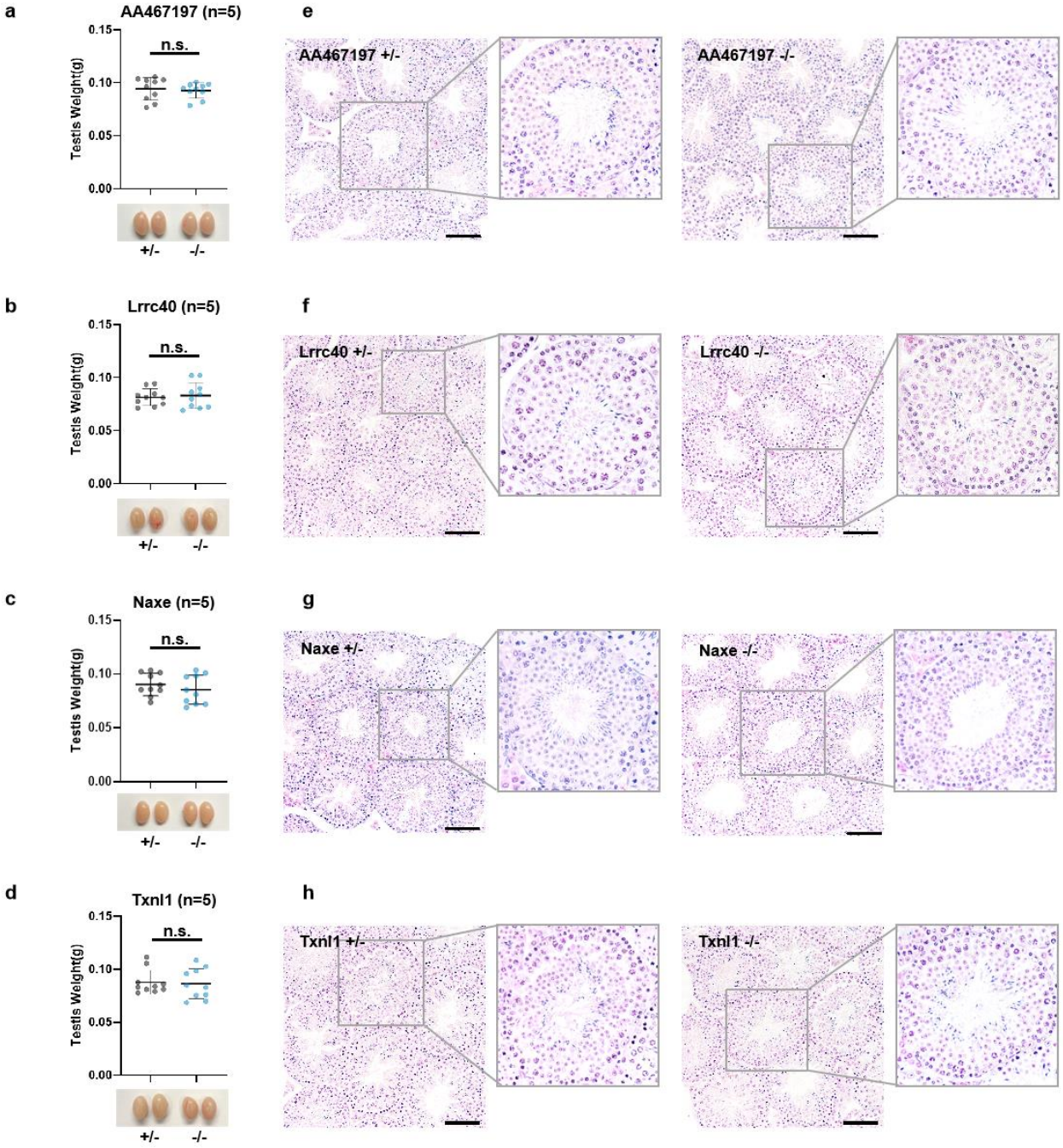


Fig.S4 Phenotypic validation of the meiosis non-essential proteins predicted by RBF. **a-d** Comparison of testis weights derived from 8-weeked Txnl1$^{+/-}$ and Txnl1$^{-/-}$ mice (**a**), AA467197$^{+/-}$ and AA467197 $^{-/-}$ mice (**b**), Lrrc40$^{+/-}$ and Lrrc40$^{-/-}$ mice (**c**), and Naxe$^{+/-}$ and Naxe $^{-/-}$ mice (**d**). n.s. means no significance in unpaired two-tailed t test. **e-h** Cross-sections of H&E stained seminiferous tubules from the heterozygous and homozygous knockout mice of the four genes above, insets denote the specific one seminiferous tubule under higher magnification.