SPOCD1 is an essential executor of piRNA-directed 1 de novo DNA methylation

SPOCD1 is an essential executor of piRNA-directed *de novo* DNA methylation

Ansgar Zoch[1,2], [#]Tania Auchynnikava[2], [#]Rebecca V. Berrens[3], [#]Yuka Kabayama[1,2], Theresa Schöpp[1,2], Madeleine Heep[1,2], Lina Vasiliauskaitė[1], Yuvia A. Pérez-Rico[4], Atlanta G. Cook[2], Alena Shkumatava[4], Juri Rappsilber[2,5], Robin C. Allshire[2] and *Dónal O'Carroll[1,2]

1. MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, 5 Little France Drive, Edinburgh, EH16 4UU, UK.
2. Wellcome Centre for Cell Biology, University of Edinburgh, Edinburgh, EH9 3BF, UK.
3. Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK.
4. Institut Curie, PSL Research University, CNRS UMR 3215, INSERM U934, 26 rue d'Ulm, 75248 Paris, France.
5. Institute of Biotechnology, Technische Universität Berlin, Berlin, Germany.

[#]These authors contributed equally to this work.

*Corresponding author: Dónal O'Carroll (donal.ocarroll@ed.ac.uk)

In mammals, the acquisition of the germline from the soma provides the germline with an essential challenge, the necessity to erase and reset genomic methylation[1]. In the male germline RNA-directed DNA methylation silences young active transposable elements (TEs)[2-4]. The PIWI protein MIWI2 (PIWIL4) and its associated PIWI-interacting RNAs (piRNAs) instruct TE DNA methylation[3,5]. PiRNAs are proposed to tether MIWI2 to nascent TE transcripts, however the mechanism by which MIWI2 directs *de novo* TE methylation is poorly understood but central to the immortality of the germline. Here, we define the interactome of MIWI2 in foetal gonocytes that are undergoing *de novo* genome methylation and identify a novel MIWI2-associated factor, SPOCD1, that is essential for young TE methylation and silencing. The loss of *Spocd1* in mice results in male-specific infertility but impacts neither piRNA biogenesis nor localization of MIWI2 to the nucleus. SPOCD1 is a nuclear protein and its expression is restricted to the period of *de novo* genome methylation. We found SPOCD1 co-purified *in vivo* with DNMT3L and DNMT3A, components of the *de novo* methylation machinery as well as constituents of the NURD and BAF chromatin remodelling complexes. We propose a model whereby tethering of MIWI2 to a nascent TE transcript recruits repressive chromatin remodelling activities and the *de novo* methylation apparatus through SPOCD1. In summary, we have identified a novel and essential executor of mammalian piRNA-directed DNA methylation.

The germline gives rise to the sperm and egg cells that are the basis of reproduction and heredity. One of the biggest threats to the integrity of the germline are TEs that have the ability to cause mutation through transposition. In mammals, DNA methylation is an important determinant of TEs silencing[6]. The mammalian germline is derived from somatic cells early during development[7] and this acquisition from the soma necessitates the process of germline

47 reprogramming and *de novo* genome methylation to reset genomic DNA methylation patterns[1].
48 In the mouse male germline, the process of *de novo* DNA methylation occurs in gonocytes
49 mostly during embryonic development[4,6]. DNMT3L is the key mediator of genome
50 methylation that interacts with and stimulates the DNMT3 (DNMT3A, DNMT3B and
51 DNMT3C) *de novo* DNA methyltransferases[8-13]. DNMT3A and DNMT3B may act
52 redundantly on TEs during *de novo* genome methylation whereas the rodent-specific DNMT3C
53 has a specialized function in TE methylation[12-15]. The first wave of *de novo* methylation is
54 indiscriminate leading to the bulk genomic methylation[4]. Many active long interspersed nuclear
55 element-1 (LINE1) and intracisternal A-particle (IAP) copies escape the first round of
56 methylation and remain expressed, threatening the genomic integrity of the germline[4]. The
57 PIWI proteins and their associated small non-coding PIWI-interacting RNA (piRNAs)
58 eliminate this threat through post-transcriptional and transcriptional silencing mechanisms[16].
59 The PIWI protein MILI (PIWIL2) destroys cytoplasmic TE transcripts by piRNA-guided
60 endonucleolytic cleavage that leads to the initiation of effector piRNA production[17]. The
61 resulting effector piRNAs are proposed to guide the nuclear PIWI protein MIWI2 to active TE
62 loci by tethering the ribonucleoprotein particle to the nascent transcript and instructing DNA
63 methylation by an unknown mechanism[16].
64 To explore the mechanism of piRNA-instructed *de novo* DNA methylation, we employed a
65 proteomics approach using our *Miwi2*^{HA} allele[18] that encodes an endogenously expressed fully
66 functional N-terminal epitope-tagged HA-MIWI2 and performed anti-HA
67 immunoprecipitation coupled with quantitative mass spectrometry (IP-MS) from extracts of
68 *Miwi2*$^{+/+}$ (negative control) and *Miwi2*$^{HA/HA}$ embryonic day 16.5 (E16.5) foetal testes (Fig. 1a,
69 Supplementary Data Table 1). This approach identified 28 MIWI2-associated proteins
70 (enrichment >4-fold, $P<0.05$). Encouragingly, 12 of these have been implicated in piRNA
71 biogenesis[16] with 5 being novel MIWI2 interactors (Fig. 1b). We also identified 16 additional

(14 novel) interacting proteins (Fig. 1b) that could either participate in piRNA biogenesis (cytoplasmic) or nuclear MIWI2 functions. To identify nuclear factors required for the execution of MIWI2 function, we applied the following criteria. First, the expression of the gene should be restricted to the period of *de novo* genome methylation as is the case for *Miwi2* and *Dnmt3l* (Fig. 1c and Extended Data Fig. 1a). This criterion would likely exclude novel piRNA biogenesis factors as they would be expected to be expressed also in adult spermatogenic populations, as exemplified by *Mili* (*Piwil2*) and *Vasa* (*Ddx4*) (Fig. 1c and Extended Data Fig. 1a). Second, the gene should encode a protein with a nuclear localization signal (NLS). Applying these criteria, we found a single gene *Spocd1* (Fig. 1c and Extended Data Fig. 1b) of unknown function that encodes for a 1015 amino acid protein with a TFIIS-M domain and a SPOC domain (Fig. 1d). Intriguingly, a SPOC domain previously described in SHARP (SPEN, MINT) has been shown to recruit the transcriptional co-repressor NCoR/SMRT[19,20]. The SPOC domain of SPOCD1 is closely related to the one found in PHF3 and DIDO1; and both of these proteins also contain a TFIIS-M domain (Extended Data Fig. 2a-b). Indeed, phylogenetic analysis supports that *Spocd1* originated from a duplication of *Phf3* in the common ancestor of lobe-finned fishes and tetrapods (Extended Data Fig. 3). In summary, we have identified SPOCD1 as a MIWI2 interactor that is a strong candidate for a facilitator of nuclear MIWI2 function.

The loss of piRNA-pathway factors or the *de novo* DNA methylation machinery results in male infertility, arrested spermatogenesis in meiosis and deregulation of LINE1 and IAP elements in mice[2,3,5,9,12,13,17,21]. To explore a potential role for SPOCD1 in the piRNA pathway, we generated a mutant allele *(Spocd1^{-})* in the mouse (Extended Data Fig. 4a-c). *Spocd1*-deficiency resulted in male-specific infertility with the complete absence of spermatozoa in the epididymis (Fig. 2a, b and Extended Data Fig. 4d). The testes of *Spocd1$^{-/-}$* mice were atrophic (Fig. 2c); histological analyses of *Spocd1$^{-/-}$* testis revealed aberrant seminiferous tubules that lacked

spermatids and presented a meiotic arrest at the early pachytene stage (Fig. 2d and Extended Data Fig. 4e). In addition, chromosome pairing is defective in *Spocd1*$^{-/-}$ meiotic cells (Extended Data Fig. 4f), a hallmark of mutations that result in LINE1 derepression[9]. The expression of LINE1 and IAP were both detected in the seminiferous tubules of adult *Spocd1*$^{-/-}$ testis (Fig. 2e, f). To explore the full repertoire of deregulated TEs, we performed RNA-seq that revealed that the same families of TEs are deregulated in post-natal day 20 (P20) testis of *Miwi2*$^{-/-}$ or *Spocd1*$^{-/-}$ mice (Fig. 2g and Extended Data Fig. 4g-i). Phosphorylation of the histone variant H2AX (γH2AX) is a marker of double stranded breaks[22]; staining of testis sections revealed the characteristic foci observed in meiotic cells in *Spocd1*$^{+/+}$ mice whereas a strong γH2AX stain indicative of extensive DNA damage was observed in *Spocd1*$^{-/-}$ meiocytes (Fig. 2h). Indeed, widespread apoptosis of meiotic cells was observed in *Spocd1*$^{-/-}$ testes (Fig. 2i). In summary, SPOCD1 is essential for spermatogenesis and is required for transposon repression. The piRNA pathway is required for *de novo* DNA methylation of IAP elements and several sub-families of LINE1[2-5]. We next sought to determine if SPOCD1 is required for piRNA-directed *de novo* DNA methylation. We thus isolated genomic DNA from wildtype, *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ P14 spermatogonia and performed whole genome methylation sequencing (Methyl-seq). We choose this time point as it is after completion of *de novo* genome methylation, prior to the onset of *Spocd1*$^{-/-}$ phenotypic defects with SPOCD1. Indeed, no major changes in methylation in *Spocd1*$^{-/-}$ spermatogonia were observed in genic, intergenic, CpG island or promoter regions (Fig. 3a and Extended Data Fig. 5a,b). Globally, *Spocd1*-deficiency did not affect collective transposon (all TEs grouped) methylation levels (Fig. 3a and Extended Data Fig. 5a,b). Consistent with *Miwi2*-deficiency, IAPEy and MMERVK10C as well as the young LINE1 families L1Md_A, L1Md_T and L1Md_Gf failed to be fully methylated in *Spocd1*$^{-/-}$ spermatogonia (Fig. 3b and Extended Data Fig. 5c,d). Metaplot analysis demonstrated defective *de novo* methylation specifically at TE promotor elements in *Spocd1*$^{-/-}$ spermatogonia

(Fig. 3c and Extended Data Fig. 6a), which is a hallmark of piRNA- and DNMT3C- directed methylation (Extended Data Fig. 6b)[12,23]. The loss of methylation was particularly evident in young LINE1 families and elements such as L1Md_T and L1Md_Gf compared to the older L1Md_F (Fig. 3c-d and Extended Data Fig. 6c). MIWI2 is required specifically for the methylation of one imprinted locus, *Rasgrf1*[24]. Consistently, among imprinted loci, only *Rasgrf1* methylation is dependent upon SPOCD1 function (Fig. 3e). In summary, SPOCD1 is required for *de novo* DNA methylation of the TEs that are regulated by the piRNA pathway.

The dependency of piRNA-mediated silencing of TEs on SPOCD1 may indicate a role for SPOCD1 as a downstream effector of nuclear MIWI2 function or alternatively in piRNA biogenesis, amplification or loading; as mutations that disrupt these processes will lead to the same phenotypic outcome. We therefore sequenced small RNA from *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ E16.5 foetal testes to analyse piRNA biogenesis. We found no major impact of *Spocd1*-deficiency on length distribution (Fig. 3f), annotation of mapped piRNAs (Fig. 3g and Extended Data Fig. 7a), relative piRNA counts (Extended Data Fig. 7b), piRNA amplification (Fig. 3h Extended Data Fig. 7c-e) or piRNAs mapping to TEs (Extended Data Fig. 7f). piRNA binding to MIWI2 licences its entry to the nucleus, thus disruption of piRNA biogenesis, amplification or loading results in the dramatic reduction of MIWI2's nuclear localization[17,18,25]. The fact that MIWI2 exhibits normal localization in the absence of *Spocd1* (Fig. 3i) confirms that SPOCD1 is not required for piRNA processing but suggests its involvement in the execution of MIWI2's nuclear function. A possible alternative nuclear function could be that SPOCD1 acts as transcription factor required for either transposon or gene expression. However, RNA-seq from E16.5 foetal gonocytes revealed *Spocd1*-deficiency had a minimal impact on gene expression and normal expression for the majority of TEs but, importantly, the piRNA-regulated TEs were mostly expressed at or above normal levels in *Spocd1*$^{-/-}$ E16.5 foetal gonocytes (Extended Data Fig. 8a-b, Supplementary Data Table 2 and

147 3). Collectively, these data are not supportive of a role for SPOCD1 as a transcription or a
148 piRNA biogenesis factor.
149 We next sought to explore how SPOCD1 contributes to *de novo* TE methylation. We thus
150 engineered the *Spocd1*^{HA} allele where the sequence encoding the HA epitope tag has been
151 inserted into the *Spocd1* locus to generate an endogenously expressed, fully functional C-
152 terminal epitope-tagged SPOCD1-HA (Fig. 4a and Extended Data Fig. 9a-d). Confocal
153 immunolocalization of SPOCD1-HA on E16.5 foetal testis sections revealed that SPOCD1 is
154 restricted to foetal gonocytes and predominantly nuclear (Fig. 4b). Furthermore, SPOCD1
155 expression is restricted to the period of *de novo* DNA methylation (Fig. 4c and Extended Data
156 Fig. 9e-f). Expression of SPOCD1 commenced at E14.5 preceding MIWI2 expression by a day
157 with expression of both proteins extinguished by P5 (Fig. 4c, d and Extended Data Fig. 9e-g).
158 To explore how SPOCD1 might mediate *de novo* DNA methylation we performed anti-HA IP-
159 MS from extracts of *Spocd1*$^{+/+}$ (negative control) and *Spocd1*$^{HA/+}$ E16.5 foetal testes (Fig 4e, f,
160 Supplementary Data Table 4). We identified 72 proteins (enrichment >4-fold, $P<0.05$) that
161 associate with SPOCD1-HA, amongst which were DNMT3L and DNMT3A, components of
162 the *de novo* methylation machinery (Fig 4e, f, Supplementary Data Table 4). A few peptides
163 corresponding to DNMT3C were detected in the SPOCD1 precipitates, but their abundance
164 was insufficient to meet our stringent co-purification criteria. We confirmed the co-
165 precipitation of SPOCD1 with components of the *de novo* methylation machinery using HEK
166 cells as an orthologous system (Extended Data Fig. 10). Several components of the repressive
167 chromatin remodelling NURD and BAF complexes co-purified with SPOCD1 (Fig 4e, f,
168 Supplementary Data Table 4). At least one paralogue of all components of the core NURD
169 complex bar one[26] and several of the BAF complex[27] were enriched in SPOCD1 IPs (Fig 4e, f,
170 Supplementary Data Table 4). We also found MIWI2 significantly enriched in the SPOCD1 IP
171 but less than the stringent 4-fold cut off ($P<0.011$, 1.9-fold enriched). We noted a poor overlap

between the factors co-precipitated by MIWI2 and SPOCD1, which could arise from different extraction procedures between the respective IPs. We re-performed MIWI2 IP-MS but included Benzonase to aid chromatin solubilization, as was done in the SPOCD1 IP-MS experiment. This revealed a major overlap in co-precipitated proteins between MIWI2 and SPOCD1 (Fig 4g, h, Supplementary Data Table 5). Importantly, the interaction with SPOCD1 was confirmed and we now found NURD (MTA3) and BAF (ARID1A and SMARCA5) components in the MIWI2 IP using the same stringent association criteria (enrichment >4-fold, $P<0.05$). Moreover, we also found DNMT3L as well as additional BAF and NURD components significantly enriched (<4-fold) in the MIWI2 IP (Supplementary Data Table 6). In summary, we have shown that SPOCD1 is a nuclear protein, specifically expressed during the period of *de novo* DNA methylation and co-precipitates the *de novo* DNA methylation machinery as well as several chromatin remodelling complexes.

Here we have defined MIWI2-associated factors in E16.5 foetal gonocytes; among these we have identified SPOCD1 and shown its requirement for piRNA-directed TE methylation. While SPOCD1 robustly co-purified with MIWI2, we did not observe MIWI2 in SPOCD1 immunoprecipitates within the stringent high enrichment and confidence interactors. We interpret this observation as indicating that only a fraction of SPOCD1 is bound either directly or indirectly to MIWI2. This fraction likely merits the portion of MIWI2 that has identified an active TE and engaged in silencing. Indeed, it may be important to uncouple MIWI2 from the effector machinery until a *bona fide* target has been identified to avoid precocious aberrant methylation and possible epimutations that would be transmitted to the next generation. We propose a tentative model of MIWI2-piRNA directed DNA methylation whereby high complementarity base pairing of the piRNA to a nascent TE transcript licences MIWI2 to engage SPOCD1 and the associated chromatin remodelling and DNA methylation machinery.

In conclusion, we have identified SPOCD1 as an essential nuclear effector of MIWI2 function and provide the first mechanistic insights into mammalian piRNA-directed methylation.

Methods summary

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

References

1 Tang, W. W., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M. A. Specification and epigenetic programming of the human germ line. *Nat Rev Genet* **17**, 585-600, doi:10.1038/nrg.2016.88 (2016).

2 Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744-747, doi:10.1126/science.1142612 (2007).

3 Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**, 503-514, doi:10.1016/j.devcel.2007.03.001 (2007).

4 Molaro, A. *et al.* Two waves of de novo methylation during mouse germ cell development. *Genes Dev* **28**, 1544-1549, doi:10.1101/gad.244350.114 (2014).

5 Kuramochi-Miyagawa, S. *et al.* DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22**, 908-917, doi:10.1101/gad.1640708 (2008).

6 Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**, 116-117, doi:10.1038/2413 (1998).

7 Ohinata, Y. *et al.* Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207-213, doi:10.1038/nature03813 (2005).

8 Chedin, F., Lieber, M. R. & Hsieh, C. L. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc Natl Acad Sci U S A* **99**, 16916-16921, doi:10.1073/pnas.262443999 (2002).

9 Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96-99, doi:10.1038/nature02886 (2004).

10 Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H. & Tajima, S. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *J Biol Chem* **279**, 27816-27823, doi:10.1074/jbc.M400181200 (2004).

11 Webster, K. E. *et al.* Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proc Natl Acad Sci U S A* **102**, 4068-4073, doi:10.1073/pnas.0500702102 (2005).

12 Barau, J. *et al.* The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**, 909-912, doi:10.1126/science.aah5143 (2016).

13 Jain, D. *et al.* rahu is a mutant allele of Dnmt3c, encoding a DNA methyltransferase homolog required for meiosis and transposon repression in the mouse male germline. *PLoS Genet* **13**, e1006964, doi:10.1371/journal.pgen.1006964 (2017).

14 Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429**, 900-903, doi:10.1038/nature02633 (2004).

15 Kato, Y. *et al.* Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Hum Mol Genet* **16**, 2272-2280, doi:10.1093/hmg/ddm179 (2007).

16 Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* **20**, 89-108, doi:10.1038/s41576-018-0073-3 (2019).

17 De Fazio, S. *et al.* The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* **480**, 259-263, doi:10.1038/nature10547 (2011).

18 Vasiliauskaitė, L. *et al.* A MILI-independent piRNA biogenesis pathway empowers partial germline reprogramming. *Nat Struct Mol Biol* **24**, 604-606, doi:10.1038/nsmb.3413 (2017).

19 Ariyoshi, M. & Schwabe, J. W. A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling. *Genes Dev* **17**, 1909-1920, doi:10.1101/gad.266203 (2003).

20 Mikami, S. *et al.* Structural insights into the recruitment of SMRT by the corepressor SHARP under phosphorylative regulation. *Structure* **22**, 35-46, doi:10.1016/j.str.2013.10.007 (2014).

21 Hata, K., Kusumi, M., Yokomine, T., Li, E. & Sasaki, H. Meiotic and epigenetic aberrations in Dnmt3L-deficient male germ cells. *Mol Reprod Dev* **73**, 116-122, doi:10.1002/mrd.20387 (2006).

22 Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem* **273**, 5858-5868 (1998).

23 Manakov, S. A. *et al.* MIWI2 and MILI Have Differential Effects on piRNA Biogenesis and DNA Methylation. *Cell Rep* **12**, 1234-1243, doi:10.1016/j.celrep.2015.07.036 (2015).

24 Watanabe, T. *et al.* Role for piRNAs and noncoding RNA in de novo DNA methylation of the imprinted mouse Rasgrf1 locus. *Science* **332**, 848-852, doi:10.1126/science.1203919 (2011).

25 Aravin, A. A. *et al.* A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**, 785-799, doi:10.1016/j.molcel.2008.09.003 (2008).

26 Kloet, S. L. *et al.* Towards elucidating the stability, dynamics and architecture of the nucleosome remodeling and deacetylase complex by using quantitative interaction proteomics. *FEBS J* **282**, 1774-1785, doi:10.1111/febs.12972 (2015).

27 Mashtalir, N. *et al.* Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. *Cell* **175**, 1272-1288.e1220, doi:10.1016/j.cell.2018.09.032 (2018).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

This research was supported by the Welcome Trust funding to DOC (106144), R.B. (213612), JR (103139), RCA (095021, 200885), the Wellcome Centre for Cell Biology (203149) and a multi-user equipment grant (108504). DOC's laboratory is also supported by the European Union H2020 program grant GermAge. A.Z. was funded by a German Research Foundation fellowship (DFG, award ZO 376/1-1). We acknowledge the EMBL GeneCore facility in Heidelberg, Germany for preparing the microarray dataset, RNA-seq dataset of E16.5

gonocytes and sequencing all next-generation sequencing libraries and namely Ferris Jung at GeneCore for preparation of the Methyl-seq libraries.

Author contributions

A.Z. contributed to the design, execution and analysis of most experiments. T.A. helped established the IP conditions and performed the mass-spectrometry analysis under the guidance of J.R. and R.C.A. R.B. and Y.K. performed the bioinformatic analysis of the Methyl-seq data or sRNA-seq as well as RNA-seq data, respectively. T.S. prepared the sRNA-seq libraries and together with Y.K. the RNA-seq libraries of P20 testes. M.H and A.C. performed the homology alignment of the SPOC and TFIIS-M domains. L.V. performed the IF staining of HA-MIWI2 and generated the gonocytes microarray dataset. Y.R.P. performed the phylogenetic analysis under guidance of A.S. D.O'C. conceived and supervised this study. D.O'C. and A.Z wrote the final version of the manuscript.

Author information

The authors declare no competing financial interests.

Correspondence and requests for materials should be addressed to Dónal O'Carroll (donal.ocarroll@ed.ac.uk).

Figures legends

Figure 1 | Definition of the MIWI2 interactome and identification of SPOCD1 from gonocytes undergoing *de novo* genome methylation.

a, Volcano plot showing enrichment (log_2(mean LFQ ratio of HA-MIWI2 IP/control IP from *Miwi2*$^{+/+}$ foetal testis) and confidence (-log_{10}(P-value of two-sided Student's t-test)) of proteins

co-purifying with HA-MIWI2 from E16.5 testis lysates (n=3). Dotted line indicates factors with enrichment >4-fold and significance P<0.05. Red: Known piRNA pathway members, blue: SPOCD1. **b**, List of known piRNA biogenesis factors and non-piRNA pathway-associated proteins co-purifying with HA-MIWI2. Novel identified MIWI2 interactors are underlined. **c**, Relative expression of indicated transcripts as measured by Affymetrix microarray in E16.5 gonocytes (n=2), adult spermatogonia (n=3) and spermatocytes (n=3). Data are mean and s.e.m., normalized to peak expression of each transcript. **d**, Schematic representation of SPOCD1 domain structure.

Figure 2 | SPOCD1 is required for spermatogenesis and LINE1/IAP silencing.

a, Number of E16.5 embryos per plug of studs with the indicated *Spocd1* genotype mated to wildtype females are presented. Data are mean and s.e.m. from n=3 *Spocd1*$^{+/+}$ studs (9 plugs total) and n=4 *Spocd1*$^{-/-}$ studs (10 plugs total). **P~0.001, two-sided Student's t-test. **b**, Representative images of PAS & Haematoxylin stained epididymis sections from (n=3) adult mice with the indicated genotype are shown. Scale bars, 20 µm. **c**, Average testicular weight in mg from adult mice with the indicated *Spocd1* genotype is plotted. Insert shows a representative image of wildtype (left) and *Spocd1*$^{-/-}$ testes. Data are mean and s.e.m. from n=3 wildtype and n=5 *Spocd1*$^{-/-}$ mice. **P~0.01, two-sided Student's t-test. **d**, Representative PAS & Haematoxylin stained testis sections of (n=3) adult mice of the indicated *Spocd1* genotype is shown. Scale bars, 50 µm. **e**, **f**, Representative images of testis sections from (n=3) adult wildtype and *Spocd1*$^{-/-}$ mice stained for LINE1 ORF1p (e) or IAP-GAG protein (f) (red) are shown. DNA was stained with DAPI (blue). Scale bars, 50 µm. **g,** RNA-seq derived heat maps depicting fold-change of expression relative to wildtype for the 10 most up-regulated LINE and ERVK TEs in (n=3) *Miwi2*$^{-/-}$ and *Spocd1*$^{-/-}$ P20 testis. **h**, **i**, Representative images of testis sections of (n=3) adult wildtype and *Spocd1*$^{-/-}$ mice stained for the DNA damage response

marker γH2AX (h) and TUNEL staining revealing apoptotic cells (i) (red). DNA was stained with DAPI (blue). Scale bars, 50 μm.

Figure 3 | SPOCD1 is required for *de novo* TE DNA methylation loci but not piRNA expression.

a-e, Analyses of genomic CpG methylation of undifferentiated P14 spermatogonia from (n=3) wildtype, *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ mice are presented. **a**, **b**, Percentages of CpG methylation levels of the indicated genomic features (with genic, promoter and CpG islands non-overlapping TEs and intergenic non-overlapping TEs or genes) or TEs (non-overlapping genes) for (n=3) biological replicates per genotype is shown as box plots. Boxes represent interquantile range from 25th to 75th percentile, the horizontal line the median, whiskers denote the data range of median ± 2x interquantile range and dots datapoints outside of this data range. **c**, Metaplots of mean CpG methylation over LINE1 elements and adjacent 2 kb are shown. **d**, Correlation analysis of mean CpG methylation loss relative to wildtype for individual TEs of the indicated LINE1 family in relation to their divergence from the consensus sequence is shown for *Spocd1*$^{-/-}$ spermatogonia. **e**, Heatmap of mean CpG methylation level of indicated maternal and paternal imprinted regions is shown. *Rasgrf1* imprinted control region is shown in detail. **f-h**, piRNA analysis of small RNAs sequenced from E16.5 testes from (n=3) *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ mice is presented. **f**, Nucleotide (nt) length distribution of small RNAs is shown. Data represent the mean and s.e.m. No significant differences were observed (P=1.0, Bonferroni adjusted two-tailed Student's t-tests). **g**, Annotation of piRNAs from merged replicates. **h**, Ping-pong analysis of piRNAs: Relative frequency of the distance between 5' ends of complementary piRNAs mapping to the LINE1 L1Md_T family is shown. **i**, Representative images (of n=3 wildtype and *Spocd1*$^{-/-}$ mice) of MIWI2 localization in E16.5

Spocd1$^{+/-}$ and *Spocd1*$^{-/-}$ gonocytes. Scale bars, 15 µm. Insert shows a zoom in of the indicated cell. Scale bars, 2 µm.

Figure 4 | SPOCD1 is a nuclear protein that associates with the *de novo* DNA methylation machinery and repressive chromatin remodelling complexes.

a, Schematic representation of the *Spocd1*^{HA} allele and the C-terminal HA-tagged SPOCD1 protein. **b**, Representative image of SPOCD1-HA localization in gonocytes at E16.5 from (n=3) *Spocd1*$^{HA/+}$ mice. Scale bar, 20 µm. Insert shows a zoom in of the indicated foetal gonocyte. Scale bar, 2 µm. **c**, **d**, Representative images of expression of SPOCD1-HA (c) and HA-MIWI2 (d) in gonocytes at the indicated time points from (n=3) *Spocd1*$^{HA/+}$ and *Miwi2* $^{HA/+}$ mice, respectively, are shown. Scale bars, 2 µm. **e**, Volcano plot showing enrichment ($\log_2$(mean LFQ ratio of SPOCD1-HA IP/control IP from *Spocd1*$^{+/+}$ foetal testis) and statistical confidence (-$\log_{10}$(P-value of two-sided Student's t-test)) of proteins co-purifying with SPOCD1-HA from E16.5 testis lysates (n=4). Dotted line indicates enrichment >4-fold and significance P<0.05. DNMT3L and DNMT3A (green), members of the NURD (violet) and BAF (blue) complexes are highlighted. **f**, Schematic representation of selected proteins co-purifying with SPOCD1. **g**, Volcano plot as presented in panel e of proteins co-purifying with MIWI2-HA from E16.5 testis Benzonase-solubilized extracts (n=4). **h,** Schematic representation of overlap of proteins co-purifying with both SPOCD1 and MIWI2.

Methods

Mouse strains and experimentation

The *Miwi2*^{HA} and *Miwi2*^{tdTomato} (*Miwi2*^{tdTom}) mouse alleles have been previously produced in the O'Carroll laboratory[18,28]. These lines were kept on a C57BL/6N genetic background. The

Miwi2^{tdTom} mouse allele generates a null allele and was used as a *Miwi2*^{null} allele in this study[28]. The *Spocd1*^{null} and *Spocd1*^{HA} alleles were created using CRISPR-Cas9 gene editing technology using B6CBAF1/Crl genetic background fertilized 1-cell zygotes as previously described[29,30]. For *Spocd1*^{null}, we injected a single sgRNA (GCAGGTTGAAGAGCAGGCTG) together with CAS9 mRNA and F_0 offspring screened by PCR and Sanger sequencing for frame-shift mutations. The *Spocd1*^{HA} allele was generated by injection of a single sgRNA (CCCCTCCTCAGATTCAGCAT) together with CAS9 mRNA and a single stranded DNA oligo containing a GGGGS linker, HA-epitope tag and a PAM site mutation flanked by 72 nucleotides of homology arms (AAACAGACTGCAGAACAGATACAAACTAGGCAGGTGTGGGAGAGCTCACTCGC CCCTCCTCAGATTCAGCATCtGTAAAGGAATCAAGCGTAATCTGGAACATCGTAT GGGTAGGATCCTCCGCCTCCACACTCATGTTCTGGTGGCTCTAAAGGGTCTGACC CCTCTGGTGGGGGACAGTTAGAGCCACCTCCATCCA). F_0 offspring were then screened by PCR and Sanger sequencing for the correct allele. Both lines were established from one founder animal and back-crossed several times to a C57BL/6N genetic background. Thus, the mice analysed were on a mixed B6CBAF1/Crl; C57BL/6N genetic background. Mice were genotyped by PCR using the following primer pairs, *Spocd1*^{null} (F: GAAGATGAGGTAGAGGCCATCG, R: TGAGCCACTTTGAGAAACAGGT) and *Spocd1*^{HA} (F: CCCCATCCACTGTAGTATCTGC, R: ATACAAACTAGGCAGGTGTGGG). For foetal testes collection for IP-MS, the *Miwi2*^{HA} line was additionally back-crossed twice to an Hsd:ICR (CD1) outbred genetic background, which shows a characteristic large litter size (*Miwi2*^{HA}.CD1). Mice were mated for 4 days and females checked for plugs daily. Plugged females were separated from studs and the day of the plug counted as E0.5. Foetal testes for the immuno-precipitation and mass-spectrometry experiments were collected from matings of *Miwi2*$^{HA/HA}$ studs to *Miwi2*$^{HA/HA}$ females or *Spocd1*$^{HA/HA}$ studs to Hsd:ICR (CD1) wildtype

females. Male fertility was assessed by mating studs to Hsd:ICR (CD1) wildtype females counting the number of embryos at E16.5 for each plugged female. Female fertility was assessed by mating *Spocd1*$^{-/-}$ females to *Spocd1*$^{+/-}$ studs and comparing number of embryos at E16.5 for each plugged female to matings of C57BL/6N wildtype mice.

Animals were maintained at the University of Edinburgh, UK in accordance with the regulation of the UK Home Office. Ethical approval for the mouse experimentation has been given by the University of Edinburgh's Animal Welfare and Ethical Review Body and the work done under licence from the United Kingdom's Home Office.

Immuno-precipitation and mass-spectrometry (IP-MS)

Foetal testes were isolated from E16.5 embryos and snap frozen in liquid nitrogen. 50 testes per replicate were lysed and homogenized in 1 ml of hypotonic lysis buffer (10 mM Tris-HCl pH 8, 10 mM KCl, 5 mM $MgCl_2$, 0.1 % IGEPAL CA-630, complete protease inhibitor EDTA-free (Roche)) with 20 strokes in a glass douncer. For the IP-MS experiments presented in figure 4, lysates of *Spocd1*$^{HA/+}$ or *Miwi2*$^{HA/+}$ testes and corresponding wildtype controls were further incubated for 30 min at 4 ˚C after addition of 50 U/ml Benzonase (Millipore). Lysates were cleared by centrifugation for 10 min at 21,000 x g. 50 µl anti-HA magnetic beads (Pierce) (additionally cross-linked with 20 mM dimethyl-pimelidate in borate buffer pH 9) were resuspended in 600 µl hypotonic lysis buffer. 900 µl of cleared lysate was then added to the resuspended beads and incubated for 30 min at 4 ˚C. Beads were washed four times with wash buffer (50 mM Tris-HCl pH 8, 100 mM KCl, 5 mM $MgCl_2$, 0.1 % IGEPAL CA-630) and bound proteins were eluted with 0.1 % Rapigest (Waters) in 50 mM Tris-HCl pH 8 for 15 min at 50 ˚C.

Eluted proteins were trypsin digested as described[31], desalted using STAGE tips[32], resuspended in 0.1 % trifluoroacetic acid (v/v) and subjected to LC-MS. Peptides were separated on an ultra-

high resolution nano-flow liquid chromatography nanoLC Ultimate 3000 unit fitted with an Easyspray (50 cm, 2 µm particles) column coupled to the high resolution/accurate-mass mass-spectrometer Orbitrap Fusion Lumos operated in DDA(data-dependent-acquisition)-mode (Thermo Fisher Scientific). Samples we separated using a 2 % - 40 % - 95 % 190 min gradient (Mobile phase A - 0.1 % aqueous formic acid, B – 80 % acetonitrile in 0.1 % formic acid). The MS acquisition parameters were as follows – cycle time was set to 3 s, the MS1 scan Orbitrap resolution was set to 120,000, RF lens to 30 %, AGC target to 4.0e5, and maximum injection time to 50 ms, detected intensity threshold was 5.0e3. The MS2 scan was performed with the Ion Trap using rapid scan setting. The AGC target was set to 2.0e4, and maximum injection time was 50ms. This set-up achieves a detection limit in the low attomole (10^{-18})-range and has been used in large proteome and interactome screens[33,34]. Raw data were processed using MaxQuant version 1.6.1.0. Label-free quantitation (LFQ) was performed using the MaxQuant LFQ algorithm[35]. Peptides were searched against the mouse UniProt database (date 21.07.2017) with commonly observed contaminants (e.g. trypsin, keratins, etc.) removed during Perseus analysis[35-37]. For visualization, LFQ intensities were imported into Perseus version 1.6.0.2[37] and processed as described[38].

Nuclear localization signal (NLS) prediction

The presence of an NLS was predicted by cNLS mapper[39], searching the entire protein for bipartite NLSs and a cut-off score of 5.0.

Affymetrix microarray datasets

The Affymetrix microarray datasets of spermatogonia, spermatocytes, mouse embryonic fibroblasts (MEFs) and bone marrow (ArrayExpress: E-MTAB-4828, E-MTAB-7067, E-MTAB-5056) have been previously described[28,40,41]. The new microarray data for gonocytes

was generated as previously described[40] from gonocytes purified by FACS from E16.5 foetal testes using the *Miwi2*^{tdTom} reporter allele[28].

Domain alignment

Alignments of SPOCD1 domains to homologous proteins were generated using ClustalW[42]. Alignment of the SPOC domain was adjusted based on SPOCD1 models generated by PHYRE2[43] superimposed on human SHARP (PDBid 1OW1[19]) and *A. thaliana* FPA SPOC (PDBid 5KXF[44]) domains. Alignments are presented using Jalview[45] with secondary structure elements from human SHARP (PDBid 1OW1[19]), PHF3 (PDBid 2DME) and human TFIIS (PDBid 3NDQ). The following sequence identifiers of homologues were used in the alignment of the SPOCD1 SPOC domain: Q6ZMY3, H9GUJ8, F7FFW6, B2RQG2, Q92576, H9GF02, B8A483, XP_028916420.1, Q8C9B9, Q9BTC0, G1KE55, F1QQA3, F7DIQ2, Q62504, Q96T58, H9GKA7, F1QMN6, XP_028921280.1. The SPOCD1 TFIIS-M domain was aligned based on primary sequence to the following sequences: Q6ZMY3, H9GUJ8, F7FFW6, B2RQG2, Q92576, H9GF02, B8A483, XP_028916420.1, Q8C9B9, Q9BTC0, G1KE55, F1QQA3, F7DIQ2, P10711, P23193, H9GPX5, Q7T3C1, F7BX76.

Phylogenetic analyses

Spocd1, *Phf3* and *Dido1* sequences were searched via tblastn[46] using the mouse and alligator protein sequences as queries and the non-redundant nucleotide collection as database. Transcript sequences with significant alignments were downloaded and processed to keep only cDNA sequences. Additional cDNA sequences were included based on described orthologous relationships from Ensembl 90[47] and Ensembl 91[48]. Axolotl and western clawed frog sequences were extracted from the UCSC genome browser[49]. *Dido1* orthologous sequences from *Drosophila* and *C. elegans* were identified by the ortholog annotations of Flybase[50] (release FB2017_05). For phylogeny reconstruction including only *Spocd1* orthologs, cDNA sequences

were aligned using the RevTrans 2.0b webserver[51] with T-COFFEE (v.11.0) as alignment method. For the phylogeny including *Spocd1* and its paralogs, cDNA sequences were first virtually translated with the Virtual Ribosome tool version 2.0[52] and aligned using Clustal Omega[42,53] with default parameters. Protein alignments were saved and used as scaffolds to align cDNA sequences using the RevTrans 2.0b webserver. Nucleotide alignments generated by RevTrans were formatted into NEXUS interleave files that were used for phylogenetic reconstruction with MrBayes v3.2.6[54,55] through the CIPRES Science Gateway V. 3.3[56]. The selected evolutionary model for MrBayes was *GTR* + I + Γ (nst=6 rates=invgamma) and priors on state frequencies were left with default values. Two analyses (nruns=2) were run with four MCMC chains for each one (nchains=4) and a heating parameter of 0.2. Sample and diagnostic frequencies were set to 1,000 and 5,000, respectively. Analyses were stopped after 2 x 10^6 generations and the consensus trees were obtained using a burnin fraction of 0.25. Tree appearance was edited with FigTree v1.4.2 (http//tree.bio.ed.ac.uk/software/figtree).

Histology

Isolated testes and epididymis were fixed overnight in Bouin's fluid, washed three times in 70 % ethanol and embedded in paraffin. 6 µm sections were cut on a microtome (Leica) and deparaffinised in a graded alcohol series according to standard laboratory procedures. The rehydrated sections were then stained with the periodic-acid-Schiff (PAS) staining kit (TCS Biosciences) according to the manufacturer's recommendations. The stained sections were subsequently de-hydrated in a reverse alcohol series and mounted on coverslips with Pertex mounting media (Pioneer Research Chemicals) according to standard laboratory procedures. Slides were imaged on a Zeiss AxioScan scanning microscope using the 40x objective. Cropped images of the scan were exported using the Zeiss Zen software and further processed in ImageJ.

Immuno-fluorescence

Immuno-fluorescence was performed on freshly cut 6 μm sections of OCT embedded testes as previously described[18] Primary antibodies were incubated overnight in blocking buffer (dilutions: anti-HA (C29F4, Cell Signaling Technologies) 1:200 (for SPOCD1-HA) or 1:500 (for HA-MIWI2); anti-HA (6E2, Cell Signaling Technologies) 1:200; anti-LINE1-ORF1P[57] 1:500; anti-IAP-GAG (a kind gift from B. Cullen, Duke University, Durham, NC, USA) 1:500; anti-γH2AX (IHC-00059, Bethyl Laboratories) 1:500), anti-SCP1 (ab15090, abcam) 1:300 and anti-SCP3 (D1, sc74569, Santa Cruz Biotechnology) 1:300. Sections were then stained with DAPI and the appropriate donkey anti-rabbit or donkey anti-mouse labelled with an Alexa Fluor (488, 568 or 647) dye and mounted on coverslips with Prolong Gold (Invitrogen). Images were acquired on a Zeiss Observer, Leica SP8 confocal microscope or Zeiss LSM880 with Airyscan module. Images acquired using the Airy scan module were deconvoluted with the Zeiss Zen software "Airyscan processing" on settings "3D" and a strength of 6. Images were processed and analysed with ImageJ and Zeiss Zen software.

Terminal deoxynucleotidyl transferase dUTP nick end labelling (TUNEL assay)

Paraffin embedded testes were sectioned and re-hydrated as described above. Sections were pre-treated with proteinase K (10 μg/ml in 10 mM Tris pH 8; Thermo Scientific) and labelled using the Click-iT TUNEL assay, Alexa Fluor 647 dye (Invitrogen) according to the manufacturer's instructions. Sections were counter-stained with DAPI (1 μg/ml), embedded with Prolong Gold (Invitrogen) and imaged on a Zeiss Observer microscope. Images were processed as above.

RNA sequencing and analysis

Total RNA was extracted from sorted *Miwi2*^{tdTOM}-positive E16.5 gonocytes using QIAzol reagent (Qiagen). Libraries for low input RNA-seq were then prepared with RiboGone and the SMARTer Stranded RNA-seq kit from Clontech and sequenced on an Illumina HiSeq 4000 in 75 bp paired-end mode. For RNA-seq of P20 testis, total RNA was extracted from 1 testis using QIAGEN RNeasy Mini kit, following the manufacturer's protocol including on-column DNase treatment. Total RNA was used for library preparation with NEBNext Ultra II Directional RNA Library Prep Kit for Illumina and 8 cycles of PCR was performed. These libraries were sequenced on an Illumina NextSeq 500 in 150 bp single-end read mode.

For analysis of differentially expressed genes, reads were mapped to GRCm38 genome_tran (release 84) with HISAT2(2.1.0)[58] using following options: --no-mixed --no-discordant --qc-filter --trim5 3. Mapped reads per gene were counted with htseq-count (HTSeq 0.11.1)[59] providing a GTF file and differentially expressed genes were analysed using DESeq2(1.26.0)[60]. For analysis of differentially expressed retrotransposons, adapter sequences were removed from reads using cutadapt(1.8.1)[61] with default settings. Processed reads were mapped to consensus sequence of rodent retrotransposons retrieved from Repbase(24.01)[62] using bowtie2[63] (2.3.4.3) with default settings. Mapped reads per retrotransposon were counted and significantly de-regulated species were analysed using DESeq2.

Fluorescence activated cell sorting (FACS)

CD9^{+} spermatogonia were sorted from P14 testes as previously described[18] with minor alterations: 100 µg/ml DNase I (Sigma-Aldrich) was added together with foetal calf serum to stop the trypsin digest and the cell suspension was further incubated for 3 min at 32 ˚C to facilitate complete degradation of released DNA. Cells were blocked with Fc block (anti-CD16/32, clone 93, eBioscience, 1:50), followed by labelling with anti-CD45 (clone 30-F11, eBioscience, 1:200) and anti-CD51 (clone RMV-7, Biolegend, 1:50) biotin conjugated

antibodies. Cells were then stained with anti-CD9^{APC} (clone eBioKMC8, eBioscience, 1:200), anti-cKit$^{PE\text{-}Cy7}$ (clone 2B8, eBioscience, 1:1600), streptavidin^{V450} (BD bioscience, 1:250) and 1 µg/ml DAPI and sorted into medium on a BD Aria II sorter (gating strategy shown in Supplementary Figure 2a). Sorted cells were pelleted for 5 min at 500 g and snap frozen in liquid nitrogen.

E16.5 gonocytes were FACS purified from *Miwi2*$^{tdTom/+}$ foetal testes by dissecting the testes in a drop of goni-mem (DMEM (Life Technologies) supplemented with penicillin-streptomycin (Life Technologies), NEAA (Life Technologies), sodium pyruvate (Life Technologies) and sodium lactate (Sigma-Aldrich) and digestion in 0.25 % Trypsin-EDTA (Gibco) at 37 ˚C for 10 minutes. Digestion was stopped by addition of 20 % foetal calf serum (FCS) and cells pelleted for 5 min at 100 x g. The pellet was treated with 10 µl 5 mg/ml DNase I (Sigma-Aldrich) for 2 min and cells rigorously resuspended in PBS containing 2 % FCS by pipetting 50 times. tdTomato-positive cells were sorted on a BD Aria Fusion sorter into PBS, lysed in Qiazol (Qiagen) and snap frozen in liquid nitrogen (gating strategy shown in Supplementary Figure 2b).

Whole genome methylation sequencing (Methyl-seq) and analysis

DNA from FACS-isolated P14 spermatogonial stem cells was isolated by proteinase K digest (10 mM Tris-HCl pH 8 , 5 mM EDTA , 1 % SDS, 0.3 M Na-acetate, 0.2 mg/ml proteinase K) overnight, followed by two rounds of phenol/chloroform/isoamylalcohol (25:24:1, Sigma-Aldrich) extraction and one round of chloroform extraction. The DNA was precipitated at -20 ˚C after addition of 1/10 volume 3 M Na-acetate, 10 µg linear acrylamide (Invitrogen) and 1 volume of isopropanol, washed two times and solubilized in 5 mM Tris-HCl pH 8. Methyl-seq libraries were prepared using the NEBnext Enzymatic Methyl-seq kit (NEB) according to the manufacturer's instructions and sequenced by Illumina NextSeq and HiSeq sequencing in 150

bp paired-end read mode. Whole genome bisulfite sequencing (WGBS) datasets of adult *Mili*$^{-/-}$ spermatocytes[4] and P10 *Dnmt3c*$^{+/-}$, *Dnmt3c*$^{-/-}$, *Dnmt3l*$^{+/-}$, *Dnmt3l*$^{-/-}$ germ cells[12] were obtained from public repositories (accession numbers SRP037785 and GSE84140, respectively).

Raw sequence reads were trimmed to remove both poor-quality calls and adapters using Trim Galore (v0.4.1,www.bioinformatics.babraham.ac.uk/projects/trim_galore/, Cutadapt[61] version 1.8.1, parameters: --paired --length 25 --trim-n --clip_R2 5). Trimmed reads were aligned to the mouse genome in paired-end mode to be able to use overlapping parts of the reads only once. Alignments were carried out with Bismark v0.22.1[64] with the following set of parameters: bismark --score_min L,0,-0.4 --paired. CpG methylation calls were extracted from the deduplicated mapping output using the Bismark methylation extractor (v0.22.1). The mapping statistics were calculated using the SeqMonk (www.bioinformatics.babraham.ac.uk/projects/seqmonk/) datastore summary report of aligned deduplicated bam files. The methylation conversion rate was calculated by mapping all reads to the spiked-in CpG unmethylated lambda and CpG methylated pUC19 DNA using the Bismark pipeline as outlined above (Supplementary Table 7).

50 adjacent CpG running window probes were generated for probes containing at least 10 reads and mean percentage methylation of the 3 replicates was calculated for each probe. For analysis of specific genome features these were defined as follows: Genic regions were defined as probes overlapping genes and promoter as probes overlapping 2000 bp upstream of annotated transcripts, as annotated by Ensembl (GRCm38.p6). CpG islands (CGIs) probes overlapping the Ensembl (GRCm38.p6) CGI annotation. For genic, promoters and CGIs genome features reads overlapping transposons were filtered out. For transposons, UCSC repeat masker annotations were downloaded from the table browser (https://genome.ucsc.edu/cgi-bin/hgTables, 02/2019). The transposon annotation, which includes retro- and DNA transposons, was sorted to exclude simple repeats as well as any small non-coding RNA

annotations. Analysis of TEs in our data was performed by unique mapping in the genome and excluding any repeats overlapping gene bodies. Transposon families were assessed by mapping only to full length elements defined as > 5 kb for LINE1 elements, > 6 kb for IAP families and > 4.5 kb for MMERVK10C. Intergenic regions were defined as regions non-overlapping genes or transposons. The methylation level was expressed as the mean percentage of individual CG sites. The metaplots, scatterplots and correlation analysis were performed by extracting the reads overlapping the respective genomic regions from SeqMonk and plotting in RStudio. The methylation difference analysis was performed using the divergence (milliDiv) from the consensus sequences. We extracted the imprinted control regions (ICR) from (https://atlas.genetics.kcl.ac.uk/). For CpG methylation of the *Rasgrf1* imprinted region we quantified individual CpGs with a minimum of 1 read mapping. Graphing and statistics were performed using SeqMonk and RStudio.

Small RNA sequencing (sRNA-seq) and analysis

For each replicate, 6 foetal testes were pooled and RNA was isolated using the QIAzol reagent following the manufacturer's instructions. Total RNA was size selected for 15–40 nucleotides (nt) using a 15 % TBE-Urea gel (Invitrogen) and a small RNA marker (Abnova) with 2x Gel loading Buffer II (Ambion). RNA was purified from the gel by addition of nuclease-free water and two successive 1 hour incubation steps at 37 ˚C, 1000 rpm with a freeze/thaw step in between. Samples were then transferred onto spin columns (Corning) plugged with filter paper (Whatman) and centrifuged at max speed for 1 min. RNA was precipitated overnight at -20 ˚C in 2.5 volumes ethanol 100 % and 1 µl GlycoBlue (Life Technologies), washed with 80 % ethanol and dissolved in 10 µl nuclease-free water. For generation of the library the NEBnext Multiplex Small RNA Library Prep Set for Illumina (NEB) was used following the manufacturer's instructions with 4 µl size-selected RNA per reaction, adaptors diluted 1:2 and

16 cycles of PCR amplification. Concentration was measured with the Qubit high sensitivity dsDNA kit on a Qubit fluorometer (Life Technologies) and quality of the library was checked using a HSD1000 tape on a Tapestation 2200 instrument (Agilent). 4 ng of each sample were used for the final library pool and sequenced on a HiSeq2500 sequencer (Illumina) in 50 bases single-end read mode.

Adapter sequences were removed from 3' end of the raw fastq file using cutadapt[61] with default settings. Annotation of processed reads of 18–32 nt for each sample were retrieved as described[65] using bowtie 1.2.1.1[66]. Up to 3 mismatches were allowed when reads were mapped to genomic TE sequences retrieved from RepeatMasker(mm10, October 2015). Mapped piRNA reads (25–30 nt) were categorized according to annotations with reads not mapping to any recorded genomic element included in 'other'. To compare expression of individual piRNAs, only mapped reads of 25–30 nt with more than 10 counts were considered and visualised as scatter plot. The piRNA amplification analysis and mapping of LINE1 and IAP was performed as described[17]. The consensus sequence of L1MdTfI, L1MdGfI, IAPEYI and IAPEZI were retrieved from Repbase[62].

Cell culture, transfection and IP-Western blot

HEK 293T cells (sourced from the O'Carroll laboratory stock, University of Edinburgh; not additionally authenticated and regularly tested for mycoplasma contamination) were cultured at 37 ˚C, 5 % CO_2 in Glasgow minimum essential medium (Sigma Aldrich) supplemented with 10 % foetal calf serum (Gibco), 2 mM L-glutamine and 1 mM Na-pyruvate (Invitrogen). For transfection $4*10^5$ cells were seeded per well on a 6-well plate on day 0 followed by transfection of 1 µg of each plasmid (pcDNA3.1-DNMT3A-FLAG: GenScript clone ID OMu22132D; pcDNA3.1-DNMT3L-FLAG: GenScript clone ID OMu18257D; pcDNA3.1-DNMT3C-FLAG: encoding the long isoform as defined by Barau et al. 2016[12], synthesised by

GenScript; pcDNA3.1-SPOCD1-HA: encoding XP_017175994.1, synthesised by GenScript) by Jetprime transfection (Polyplus) according to the manufacturer's instructions on day 1. 65 hours later cells were washed twice with ice-cold PBS, scrapped of the plate in 1 ml lysis buffer (IP buffer: 150 mM KCl, 2.5 mM $MgCl_2$, 0.5 % Triton X-100, 50 mM Tris pH 8, supplemented with 1x protease inhibitors (cOmplete ULTRA EDTA-free, Roche) and 37 U/ml Benzonase (Millipore)) and lysed for 30 minutes at 4 ˚C on a rotating wheel. Lysates were cleared for 5 minutes at 21,000 g and 400 µl each was incubated for 2 hours at 4 ˚C with 20 µl of anti-HA beads (Pierce) and 20 µl control Protein G Dynabeads (Life Technologies), which had been washed twice in PBS, 0.5 % Triton X-100 and resuspended in 500 µl lysis buffer. Immuno-precipitates were eluted for 10 minutes at 50 ˚C in 35 µl 0.1 % SDS (sodium dodecyl sulfate), 50 mM Tris pH 8. Lysates and eluates were separated on a 4-12 % bis-tris acrylamide gel (Invitrogen) and blotted onto nitrocellulose membrane (Amersham Protran 0.45 NC) according to standard laboratory procedures. The membrane was stained for protein with 0.1 % (w/v) Ponceau S in 5 % (v/v) acetic acid solution for 5 minutes, blocked with blocking buffer (4 % (w/v) skimmed milk powder (Sigma-Aldrich) in TBS-T (tris buffered saline, 0.1 % Tween-20)), incubated with primary antibodies for 1 hour (anti-HA (6E2, Cell Signaling Technologies) 1:1000; anti-FLAG (M2, Sigma-Aldrich) 1:1000) in blocking buffer, washed 4 times for 5 minutes in TBS-T, incubated with secondary antibodies (IRDye 680RD donkey anti-rabbit & IRDye 800CW donkey anti-mouse, LI-COR, 1:10,000) in Immobilon® Block – PO blocking solution (Millipore), washed 4 times for 5 minutes in TBS-T and imaged on a LI-COR Odyssey Fc system. Exposure of the entire images was adjusted in Image Studio Lite (LI-COR) and regions of interest cropped for presentation.

Statistical information

Statistical testing was performed with R(3.3.1) using the R Studio software and with Perseus for the mass-spectrometry data. Unpaired, two-tailed Student's t-tests were used to compare differences between groups and adjusted for multiple testing using Bonferroni correction where indicated, except for RNA-seq data analysis where Wald's tests and Benjamini-Hochberg correction were used. Averaged data are presented as mean ± s.e.m. (standard error of the mean), unless otherwise indicated. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Data availability

All mRNA expression data that support the findings of this study have been deposited at Array Express under accession numbers E-MTAB-7985. The Methyl-seq data generated in this study have been deposited at ArrayExpress under the accession number E-MTAB-7997. The sRNA-seq and RNA-seq data generated in this study have been deposited at Gene Expression Omnibus under the accession number GSE131377. Data for the IP-MS experiments were deposited at ProteomeXchange under the accession number PXD016701.

Code availability

Scripts used for the Methyl-seq, RNA-seq and sRNA-seq analysis are available on github (https://github.com/rberrens/SPOCD1-piRNA_directed_DNA_met).

Additional References

28 Carrieri, C. *et al.* A transit-amplifying population underpins the efficient regenerative capacity of the testis. *J Exp Med* **214**, 1631-1641, doi:10.1084/jem.20161371 (2017).

29 Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910-918, doi:10.1016/j.cell.2013.04.025 (2013).

30 Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154**, 1370-1379, doi:10.1016/j.cell.2013.08.022 (2013).

31 Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359-362, doi:10.1038/nmeth.1322 (2009).

32 Rappsilber, J., Ishihama, Y. & Mann, M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* **75**, 663-670 (2003).

33 Hein, M. Y. *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712-723, doi:10.1016/j.cell.2015.09.053 (2015).

34 Richards, A. L. *et al.* One-hour proteome analysis in yeast. *Nat Protoc* **10**, 701-714, doi:10.1038/nprot.2015.040 (2015).

35 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).

36 The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

37 Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731-740, doi:10.1038/nmeth.3901 (2016).

740 38 Hubner, N. C. & Mann, M. Extracting gene function from protein-protein interactions
741 using Quantitative BAC InteraCtomics (QUBIC). *Methods* **53**, 453-459,
742 doi:10.1016/j.ymeth.2010.12.016 (2011).
743 39 Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. Systematic identification of cell
744 cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite
745 motifs. *Proc Natl Acad Sci U S A* **106**, 10171-10176, doi:10.1073/pnas.0900604106
746 (2009).
747 40 Morgan, M. *et al.* mRNA 3' uridylation and poly(A) tail length sculpt the mammalian
748 maternal transcriptome. *Nature* **548**, 347-351, doi:10.1038/nature23318 (2017).
749 41 Morgan, M. *et al.* A programmed wave of uridylation-primed mRNA degradation is
750 essential for meiotic progression and mammalian spermatogenesis. *Cell Res* **29**, 221-
751 232, doi:10.1038/s41422-018-0128-1 (2019).
752 42 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence
753 alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75
754 (2011).
755 43 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2
756 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858,
757 doi:10.1038/nprot.2015.053 (2015).
758 44 Zhang, Y., Rataj, K., Simpson, G. G. & Tong, L. Crystal Structure of the SPOC Domain
759 of the Arabidopsis Flowering Regulator FPA. *PLoS One* **11**, e0160694,
760 doi:10.1371/journal.pone.0160694 (2016).
761 45 Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M. & Barton, G. J. Jalview
762 Version 2--a multiple sequence alignment editor and analysis workbench.
763 *Bioinformatics* **25**, 1189-1191, doi:10.1093/bioinformatics/btp033 (2009).

764 46 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
765 database search programs. *Nucleic Acids Res* **25**, 3389-3402,
766 doi:10.1093/nar/25.17.3389 (1997).
767 47 Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res* **45**, D635-D642,
768 doi:10.1093/nar/gkw1104 (2017).
769 48 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754-D761,
770 doi:10.1093/nar/gkx1098 (2018).
771 49 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**,
772 D493-496, doi:10.1093/nar/gkh103 (2004).
773 50 Gramates, L. S. *et al.* FlyBase at 25: looking to the future. *Nucleic Acids Res* **45**, D663-
774 D671, doi:10.1093/nar/gkw1016 (2017).
775 51 Wernersson, R. & Pedersen, A. G. RevTrans: Multiple alignment of coding DNA from
776 aligned amino acid sequences. *Nucleic Acids Res* **31**, 3537-3539,
777 doi:10.1093/nar/gkg609 (2003).
778 52 Wernersson, R. Virtual Ribosome--a comprehensive DNA translation tool with support
779 for integration of sequence feature annotation. *Nucleic Acids Res* **34**, W385-388,
780 doi:10.1093/nar/gkl252 (2006).
781 53 Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI.
782 *Nucleic Acids Res* **38**, W695-699, doi:10.1093/nar/gkq313 (2010).
783 54 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under
784 mixed models. *Bioinformatics* **19**, 1572-1574, doi:10.1093/bioinformatics/btg180
785 (2003).
786 55 Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis
787 coupled Markov chain Monte Carlo for Bayesian phylogenetic inference.
788 *Bioinformatics* **20**, 407-415, doi:10.1093/bioinformatics/btg427 (2004).

789 56 Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for
790 inference of large phylogenetic trees. *2010 Gateway Computing Environments*
791 *Workshop (GCE)* (2010).
792 57 Di Giacomo, M., Comazzetto, S., Sampath, S. C. & O'Carroll, D. G9a co-suppresses
793 LINE1 elements in spermatogonia. *Epigenetics Chromatin* **7**, 24, doi:10.1186/1756-
794 8935-7-24 (2014).
795 58 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low
796 memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
797 59 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-
798 throughput sequencing data. *Bioinformatics* **31**, 166-169,
799 doi:10.1093/bioinformatics/btu638 (2015).
800 60 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and
801 dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-
802 014-0550-8 (2014).
803 61 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
804 reads. *2011* **17**, 3, doi:10.14806/ej.17.1.200 (2011).
805 62 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive
806 elements in eukaryotic genomes. *Mob DNA* **6**, 11, doi:10.1186/s13100-015-0041-9
807 (2015).
808 63 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*
809 *Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
810 64 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for
811 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572,
812 doi:10.1093/bioinformatics/btr167 (2011).

65 Kabayama, Y. *et al.* Roles of MIWI, MILI and PLD6 in small RNA regulation in mouse growing oocytes. *Nucleic Acids Res* **45**, 5387-5398, doi:10.1093/nar/gkx027 (2017).

66 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

Extended Data figure legends

Extended Data Figure 1 | Expression pattern and presence of nuclear localization signals for novel MIWI2 interactors.

a, **b**, Relative expression of indicated transcripts as measured by Affymetrix microarray in E16.5 gonocytes (n=2), adult spermatogonia (n=3), spermatocytes (n=3), MEFs (n=3) and bone marrow (n=2). Data are mean and s.e.m. NLS indicates presence of a nuclear localization signal as predicted by cNLS mapper.

Extended Data Figure 2 | Homology alignment of SPOCD1 SPOC and TFIIS-M domains.

a, Multiple sequence alignment of the SPOC domain from SPOCD1 with representative vertebrate sequences from PHF3, DIDO1 and SPEN orthologues. The numbering for mouse SPOCD1 is shown above. Secondary structure elements for the human SHARP SPOC domain (PDBid 1OW1, SHARP is the human SPEN orthologue) are shown below the sequence, with dark rectangles for alpha helices and lighter arrows for beta strands. **b,** Multiple sequence alignment of TFIIS-M domain of SPOCD1 with equivalent sequences from TFIIS, PHF3 and DIDO1. Secondary structure elements from human PHF3 (PDBid 2DME) and human TFIIS (PDBid 3NDQ) are shown below, using the same annotation as in (a). Sequences are coloured according to sequence identity.

838

839 **Extended Data Figure 3 | Phylogeny of SPOCD1.**

840 Bayesian phylogeny of Spocd1 (blue) and its vertebrate paralogs Phf3 (red) and Dido1 (green)
841 inferred from cDNA sequences. Posterior probabilities of splits are shown as node labels.
842 Branch lengths measure the expected substitutions per site as indicated in the scale bar.

843

844 **Extended Data Figure 4 | Generation and characterisation of the *Spocd1*^{null} mouse allele.**
845 **a**, Schematic representation of the *Spocd1* locus and the encoded 1015 amino acids (aa) protein
846 (transcript XM_017320505.1) as well as design of the sgRNA targeting *Spocd1* exon 7, which
847 harbours part of the TFIIS-M domain. **b**, Schematic representation and sequencing trace
848 (lower) of the part of *Spocd1*^{null} exon 7 harbouring the mutation site. The mutated site,
849 highlighted in red, contains 2 premature stop codons and causes a frame-shift. Sequencing was
850 repeated with identical results on n=3 animals. **c**, Representative image of genotyping results
851 for *Spocd1*$^{+/+}$, *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ animals. Similar results were obtained for all animals of
852 the *Spocd1*$^{-}$ line. **d,** Number of E16.5 embryos per plug from matings of mice with the indicated
853 *Spocd1* genotypes are presented. Mean and s.e.m. from n=7 *Spocd1*$^{+/+}$ dams mated to n=3
854 *Spocd1*$^{+/+}$ studs and n=12 *Spocd1*$^{-/-}$ dams mated to n=5 *Spocd1*$^{+/-}$ studs is plotted. NS, non-
855 significant difference (P~0.98), two-tailed Student's t-test. **e,** Representative PAS and
856 haematoxylin stained histological testis sections of different stages of the seminiferous cycle
857 are shown of (n=3) *Spocd1*$^{+/+}$ and *Spocd1*$^{-/-}$ animals, indicating a germ cell differentiation arrest
858 at the early pachytene stage. Scale bar, 5 µm. eP, early pachytene; RS, round spermatids, eS(13)
859 elongating spermatids (step 13); PL, pre-leptotene; P, pachytene; L, leptotene; Z, zygotene;
860 m2, secondary meiocytes. **f,** Representative images of zygotene spermatocytes in wildtype and
861 *Spocd1*$^{-/-}$ adult testis sections stained for the synaptonemal complex proteins SCP1 (red) and
862 SCP3 (green). DNA stained with DAPI (blue). Scale bar, 1 µm. The representative images

presented in panels e and f are from n=3 mice per genotype. **g, h** Analysis of TE expression in P20 testes from n=3 wildtype, *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ mice by RNA-seq. **g,** Comparison of TE expression in *Miwi2*$^{-/-}$ and wildtype testes is shown. TEs with a significantly different (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test) change in expression (>2-fold) are highlighted in red and the top 12 most up-regulated TEs in *Miwi2*$^{-/-}$ testes are labelled. **h,** Comparison of TE expression in *Spocd1*$^{-/-}$ and wildtype testes is shown. TEs with a significantly different (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test) change in expression (>2-fold) are highlighted in red and same TEs as in (a) are labelled. **i,** Comparison of TE expression in *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ testes is shown. TEs with a significantly different (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test) change in expression (>2-fold) are highlighted in red. TEs which are significantly up-regulated in *Miwi2*$^{-/-}$ relative to wildtype are highlighted in black.

Extended Data Figure 5 | CpG Methylation analysis of different genomic features and TE families.

Analysis of genomic CpG methylation of undifferentiated P14 spermatogonia from (n=3) wildtype, *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ mice is presented. **a, b,** Scatter plots comparing CpG methylation levels for the respective genomic features between wildtype and *Spocd1*$^{-/-}$ or *Miwi2*$^{-/-}$ (**a**) and between *Spocd1*$^{-/-}$ or *Miwi2*$^{-/-}$ spermatogonia (**b**) are shown. **c, d,** Scatter plots comparing CpG methylation levels for the respective TE families between wildtype and *Spocd1*$^{-/-}$ or *Miwi2*$^{-/-}$ (**c**) and between *Spocd1*$^{-/-}$ or *Miwi2*$^{-/-}$ spermatogonia (**d**) are shown. Data is mean from n=3 biological replicates per genotype and shown as individual data points (grey) overlayed by a density map.

Extended Data Figure 6 | Methylation analysis of TE families.

Analysis of genomic CpG methylation of undifferentiated P14 spermatogonia from (n=3) wildtype, *Spocd1*$^{-/-}$ and *Miwi2*$^{-/-}$ mice is presented. **a**, Metaplots of CpG methylation over L1Md_A, IAPEy and MMERVK10C elements and adjacent 2 kb are shown. Schematic representation of the element is shown below. **b**, Metaplots of mean CpG methylation over LINE1 elements and adjacent 1 kb are shown. The Methyl-seq datasets of P14 wildtype, *Miwi2*$^{-/-}$ and *Spocd1*$^{-/-}$ spermatogonia are compared to WGBS datasets of adult *Mili*$^{-/-}$ spermatocytes (Molaro et al. 2014[4]) and P10 *Dnmt3c*$^{+/-}$, *Dnmt3c*$^{-/-}$, *Dnmt3l*$^{+/-}$, *Dnmt3l*$^{-/-}$ germ cells (Barau et al. 2016[12]). Schematic representation of LINE1 is shown below. **c**, Correlation analysis of mean CpG methylation loss relative to wildtype over individual elements of the indicated TE family in relation to their divergence from the consensus sequence is shown for *Miwi2*$^{-/-}$ and *Spocd1*$^{-/-}$ spermatogonia.

Extended Data Figure 7 | piRNA analysis.

piRNA analyses of small RNAs sequenced from E16.5 testis from (n=3) *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ mice are presented. **a**, Relative frequency of piRNAs mapping to LINE1 and IAP families from *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ E16.5 testes. Plots are shown for all piRNA or anti-sense piRNAs. Data are mean and s.e.m. Adjusted P-values are listed, P=1.0 values are denoted as NS (Bonferroni adjusted two-sided Student's t-test). **b**, Scatter plots showing mean expression of all (n=124411) piRNAs. The identity line is shown in red. r, Pearson's correlation coefficient. **c**, Nucleotide features of piRNA from *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ E16.5 testes. Frequency of mapped piRNAs with a U at position 1 (1U) and with an A at position 10 (10A) are shown for L1Md_T elements. Data represent the mean and s.e.m. Adjusted P-values are shown (Bonferroni adjusted two-sided Student's t-test) **d**, Ping-pong analysis of piRNAs from *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ E16.5 testis. Relative frequencies of the distances between 5' ends of complementary piRNAs are shown for the indicated LINE1 and IAP families. **e**, Nucleotide

features of piRNA from *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ E16.5 testis. Relative frequencies of piRNAs with a U at position 1 (1U) and with an A at position 10 (10A) are shown for respective elements shown in (d). Data are mean and s.e.m. Adjusted P-values are listed, P=1.0 values denoted as NS (Bonferroni adjusted two-sided Student's t-test) **f**, Positions of piRNAs mapped to the consensus sequence of L1Md_T. Positive and negative values indicate sense and antisense piRNAs, respectively. Schematic representation of L1Md_T is shown above.

Extended Data Figure 8 | TE and gene expression in *Spocd1*$^{-/-}$ gonocytes.

Analysis of TE and gene expression in E16.5 *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ gonocytes by RNA-seq from n=3 mice per genotype. **a,** Comparison of TE expression in *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ gonocytes is shown. TEs up-regulated in *Miwi2*$^{-/-}$ testes at P20 are highlighted in black. **b,** Comparison of gene expression in *Spocd1*$^{+/-}$ and *Spocd1*$^{-/-}$ gonocytes is shown. Significantly expressed genes (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test, >2-fold change) are highlighted in red.

Extended Data Figure 9 | Generation of the *Spocd1*^{HA} mouse allele.

a, Schematic representation of the SPOCD1 protein and *Spocd1* locus as well as design of the sgRNA targeting the 3' UTR near the translation termination site on *Spocd1* exon 15. The *Spocd1*^{HA} allele encodes for a carboxy-terminal GGGGS linker followed by the HA epitope tag. The protospacer adjacent motif (PAM) site was mutated to inhibit re-targeting of the *Spocd1*^{HA} allele by the sgRNA-CAS9 complex. All inserted nucleotides and corresponding encoded amino acids are highlighted in red. The SPOCD1-HA protein is shown as a schematic representation. **b**, Schematic representation of the targeting strategy to generate the *Spocd1*^{HA} allele with a short single stranded DNA oligo donor (ssODN) of 200 nucleotides containing 5' and 3' homology arms (5'HA and 3'HA) of 72 nucleotides. **c**, Representative image of

genotyping results for *Spocd1*$^{+/+}$, *Spocd1*$^{HA/+}$ and *Spocd1*$^{HA/HA}$ animals. Similar results were obtained for all animals of the *Spocd1*^{HA} line. **d**, Sequencing trace of part of a PCR amplicon of the HA epitope tag insertion site from a *Spocd1*$^{HA/HA}$ animal. The experiment was repeated with identical results on n=2 animals. **e**, **f**, **g**, Representative images of wildtype (e), *Spocd1*$^{HA/+}$ (f) and *Miwi2*$^{HA/+}$ (g) testis sections at the indicated developmental time point probed with anti-HA antibody in green. DNA stained with DAPI in blue. Scale bars, 10 µm. The representative images presented in panels e to g are from experiments done n=3 mice as biological replicates with similar results.

Extended Data Figure 10 | Co-immunoprecipitation experiments of SPOCD1 and DNMT3A/L/C in HEK cells.

Western blot analysis of co-immunoprecipitation of SPOCD1-HA with DNMT3L-FLAG, DNMT3A-FLAG, DNMT3C-FLAG or GFP in HEK cells. Shown are lysate sample (L), control IP (protein G beads) (B) and anti-HA IP (IP) for 4 experiments. For uncropped source data, see Supplementary Figure 1.

Figure 1

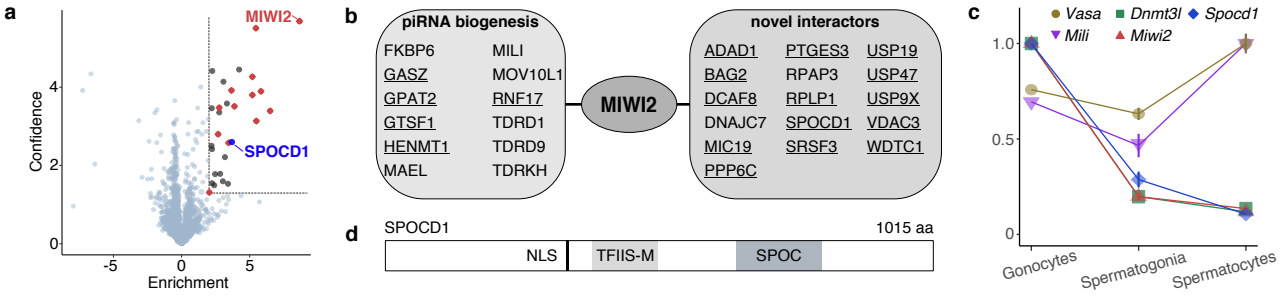

a
MIWI2
SPOCD1
Confidence
Enrichment
b
piRNA biogenesis
FKBP6
GASZ
GPAT2
GTSF1
HENMT1
MAEL
MILI
MOV10L1
RNF17
TDRD1
TDRD9
TDRKH
MIWI2
novel interactors
ADAD1
BAG2
DCAF8
DNAJC7
MIC19
PPP6C
PTGES3
RPAP3
RPLP1
SPOCD1
SRSF3
USP19
USP47
USP9X
VDAC3
WDTC1
d
SPOCD1
1015 aa
NLS
TFIIS-M
SPOC
c
Vasa
Dnmt3l
Spocd1
Mili
Miwi2
Gonocytes
Spermatogonia
Spermatocytes

Figure 2

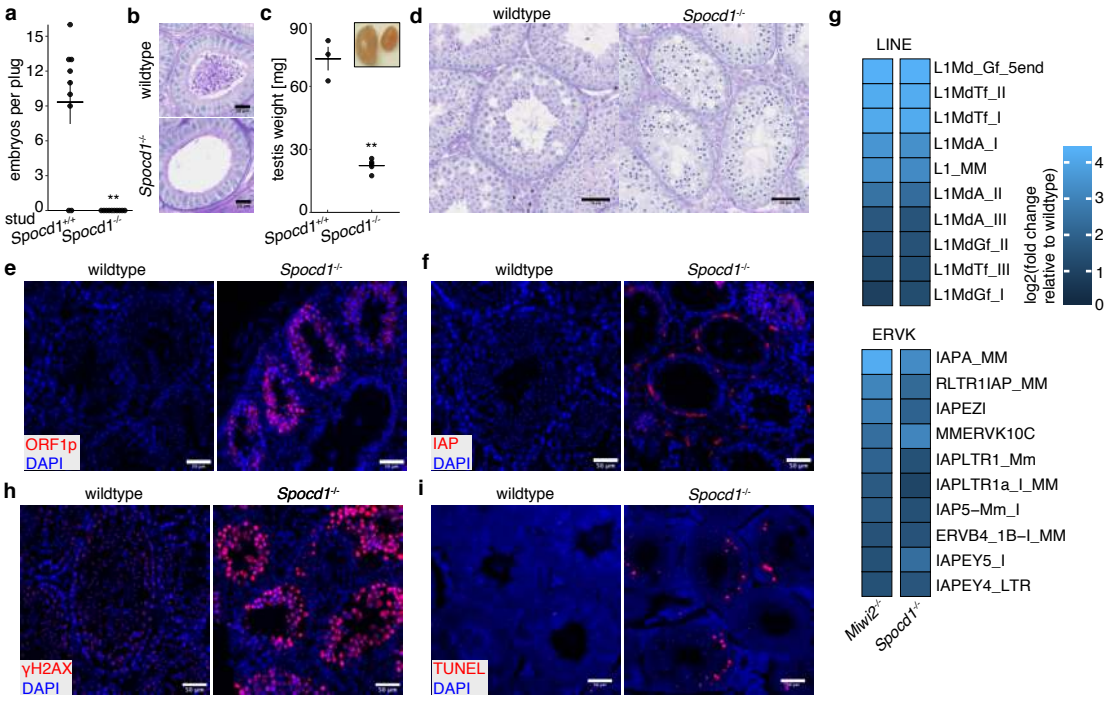

a
embryos per plug
15
12
9
6
3
0
**
stud
Spocd1+/+
Spocd1-/-
b
wildtype
Spocd1-/-
c
testis weight [mg]
90
60
30
0
**
Spocd1+/+
Spocd1-/-
d
wildtype
Spocd1-/-
e
wildtype
Spocd1-/-
ORF1p
DAPI
f
wildtype
Spocd1-/-
IAP
DAPI
g
LINE
L1Md_Gf_5end
L1MdTf_II
L1MdTf_I
L1MdA_I
L1_MM
L1MdA_II
L1MdA_III
L1MdGf_II
L1MdTf_III
L1MdGf_I
log2(fold change relative to wildtype)
4
3
2
1
0
ERVK
IAPA_MM
RLTR1IAP_MM
IAPEZI
MMERVK10C
IAPLTR1_Mm
IAPLTR1a_I_MM
IAP5–Mm_I
ERVB4_1B–I_MM
IAPEY5_I
IAPEY4_LTR
Miwi2-/-
Spocd1-/-
h
wildtype
Spocd1-/-
γH2AX
DAPI
i
wildtype
Spocd1-/-
TUNEL
DAPI

Figure 3

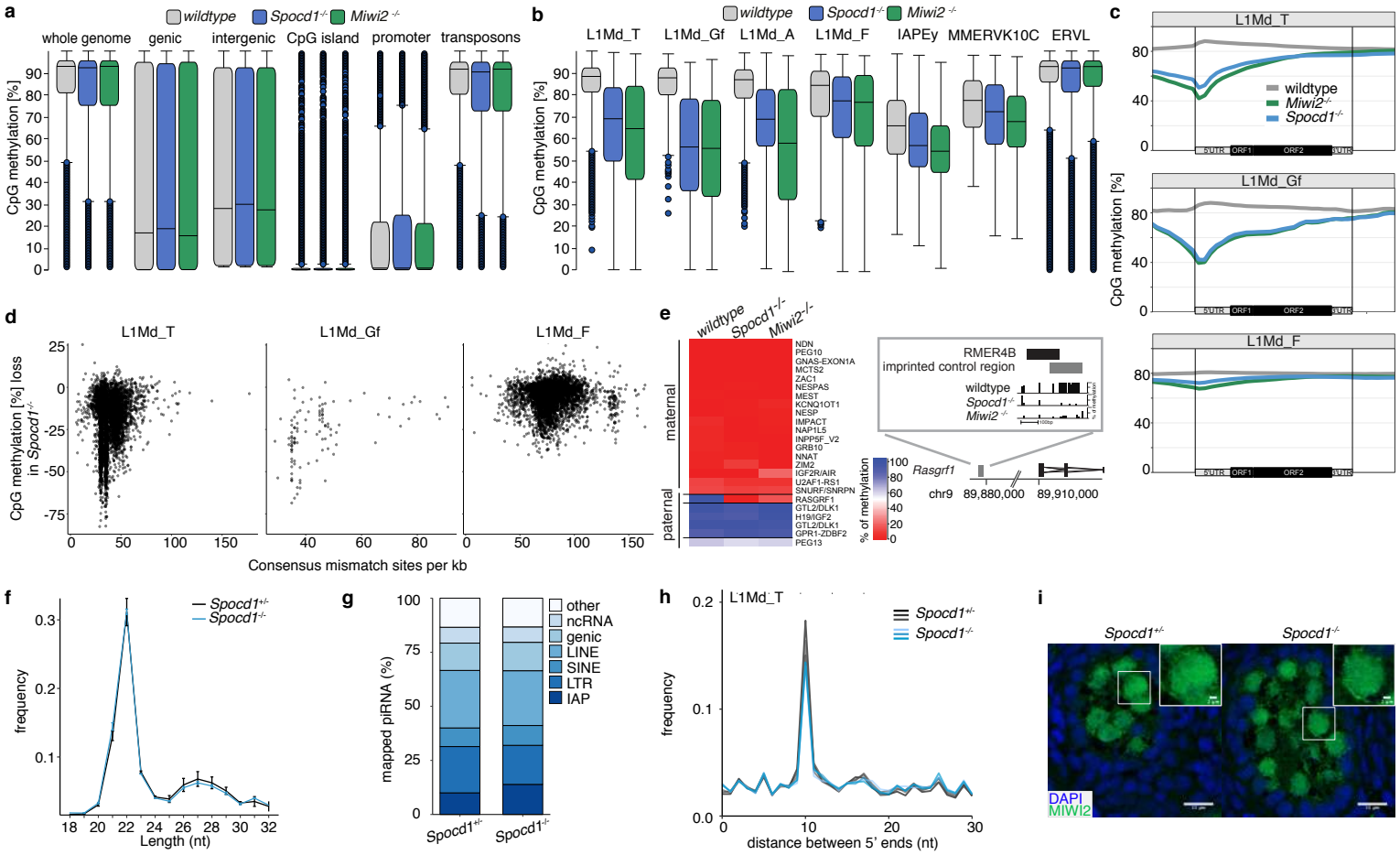
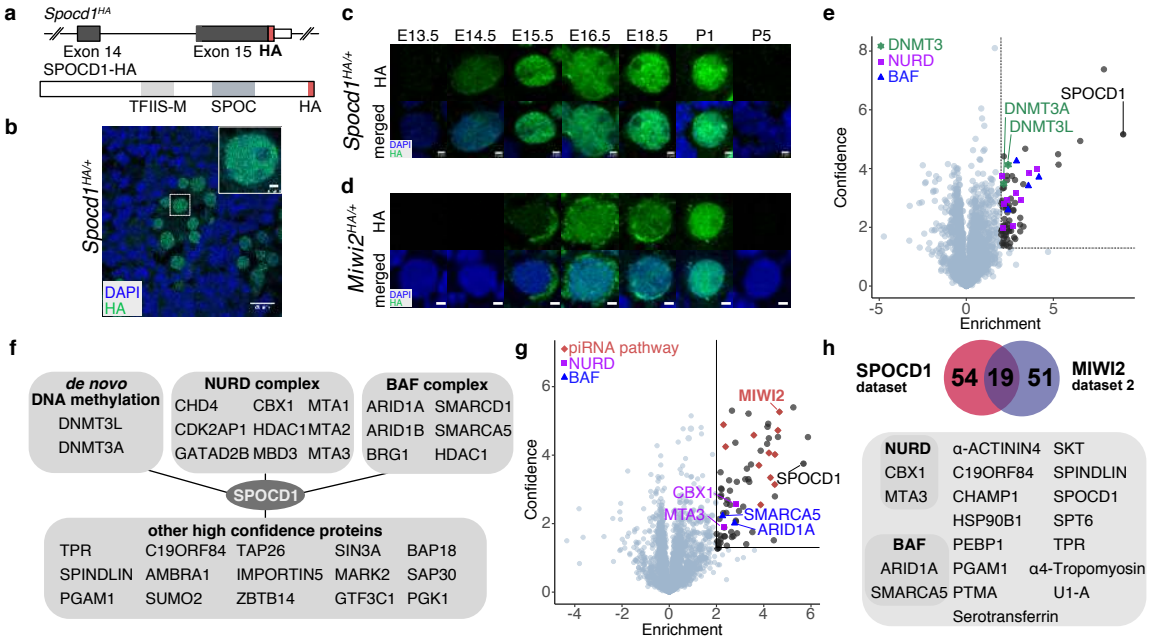
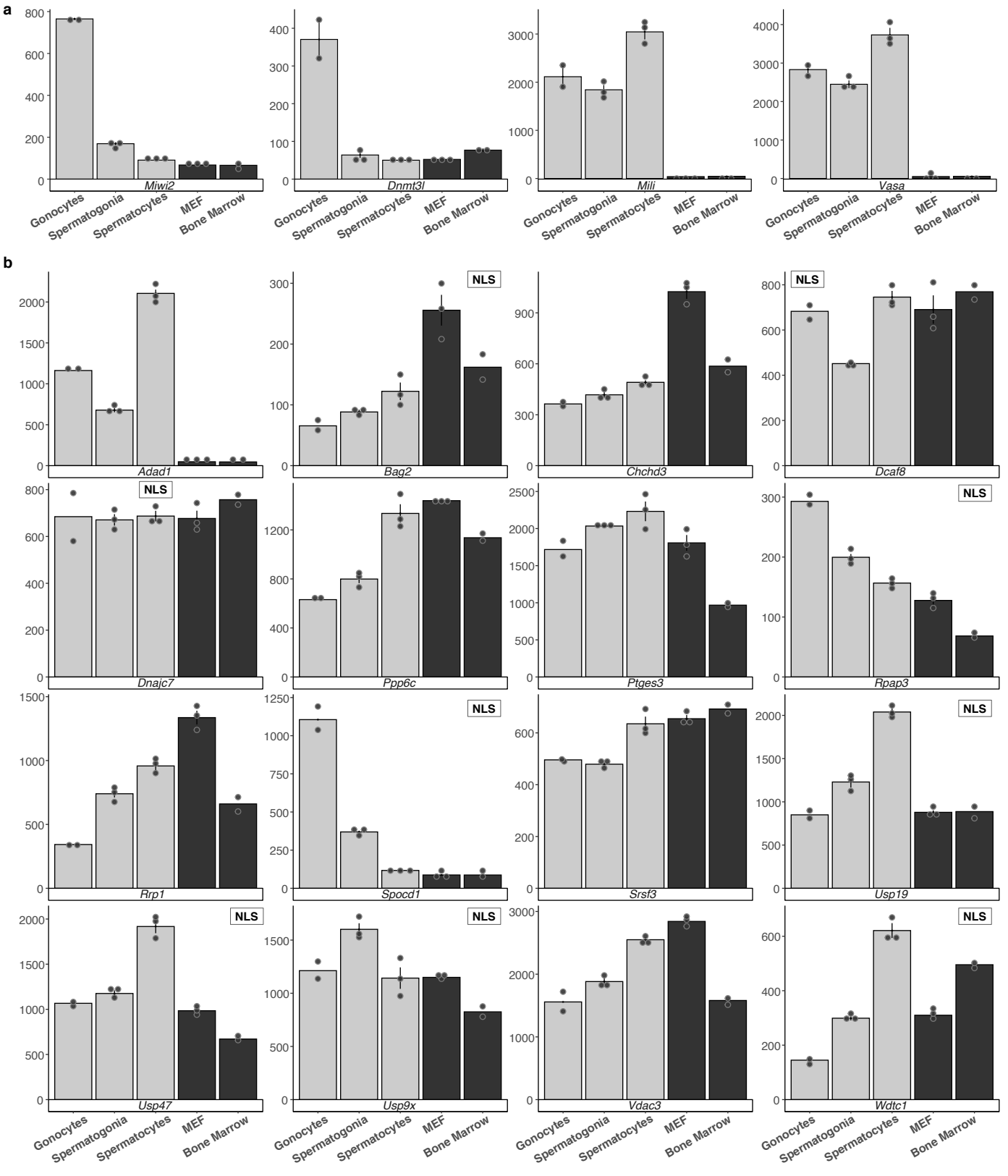
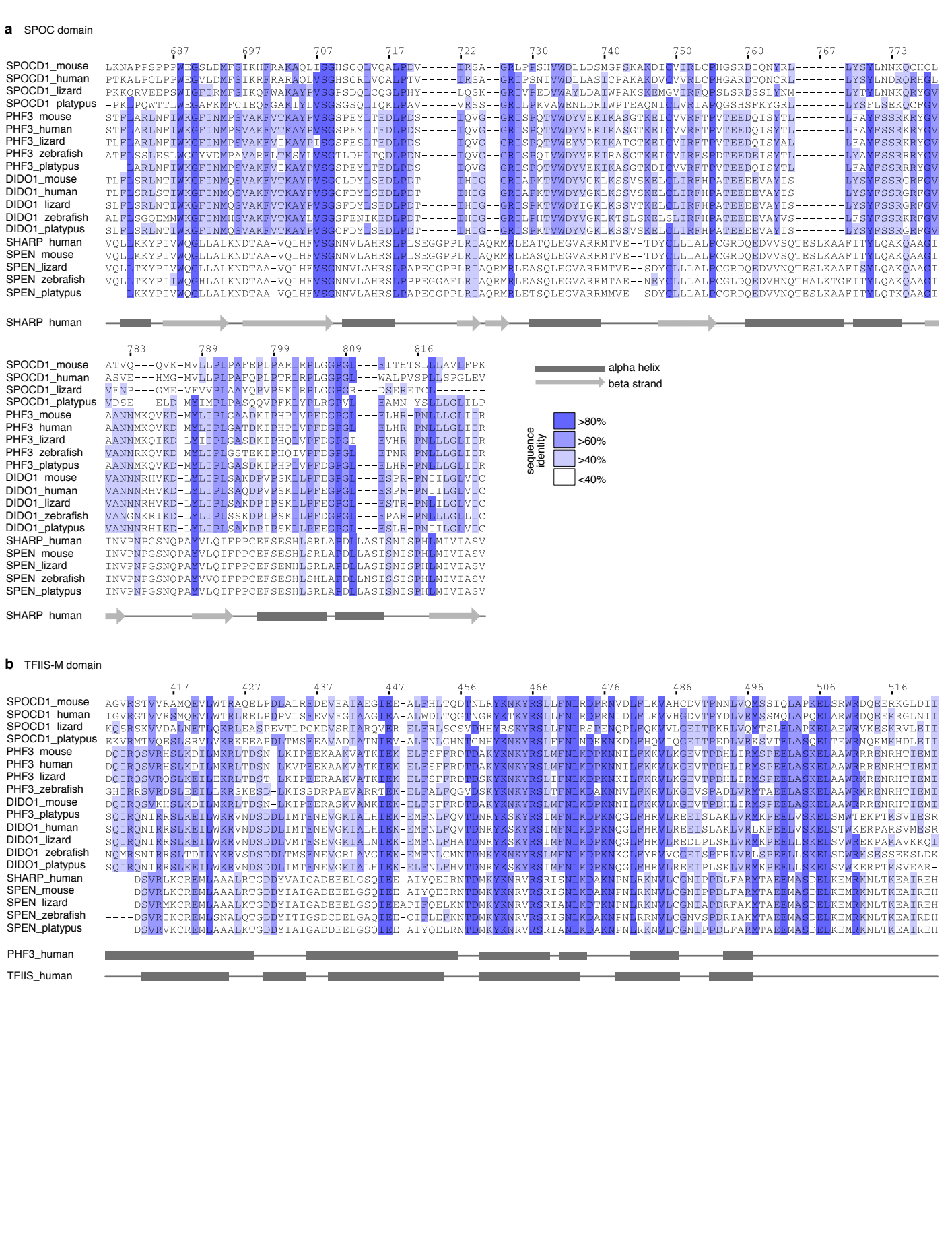

a
wildtype
Spocd1-/-
Miwi2 -/-
whole genome
genic
intergenic
CpG island
promoter
transposons
CpG methylation [%]
b
L1Md_T
L1Md_Gf
L1Md_A
L1Md_F
IAPEy
MMERVK10C
ERVL
c
L1Md_T
L1Md_Gf
L1Md_F
wildtype
Miwi2-/-
Spocd1-/-
5'UTR
ORF1
ORF2
3'UTR
d
L1Md_T
L1Md_Gf
L1Md_F
CpG methylation [%] loss in Spocd1-/-
Consensus mismatch sites per kb
e
wildtype
Spocd1-/-
Miwi2-/-
maternal
paternal
NDN
PEG10
GNAS-EXON1A
MCTS2
ZAC1
NESPAS
MEST
KCNQ1OT1
NESP
IMPACT
NAP1L5
INPP5F_V2
GRB10
NNAT
ZIM2
IGF2R/AIR
U2AF1-RS1
SNURF/SNRPN
RASGRF1
GTL2/DLK1
H19/IGF2
GTL2/DLK1
GPR1-ZDBF2
PEG13
% of methylation
RMER4B
imprinted control region
wildtype
Spocd1-/-
Miwi2-/-
Rasgrf1
chr9
89,880,000
89,910,000
f
Spocd1+/-
Spocd1-/-
frequency
Length (nt)
g
mapped piRNA (%)
other
ncRNA
genic
LINE
SINE
LTR
IAP
Spocd1+/-
Spocd1-/-
h
L1Md_T
Spocd1+/-
Spocd1-/-
frequency
distance between 5' ends (nt)
i
Spocd1+/-
Spocd1-/-
DAPI
MIWI2

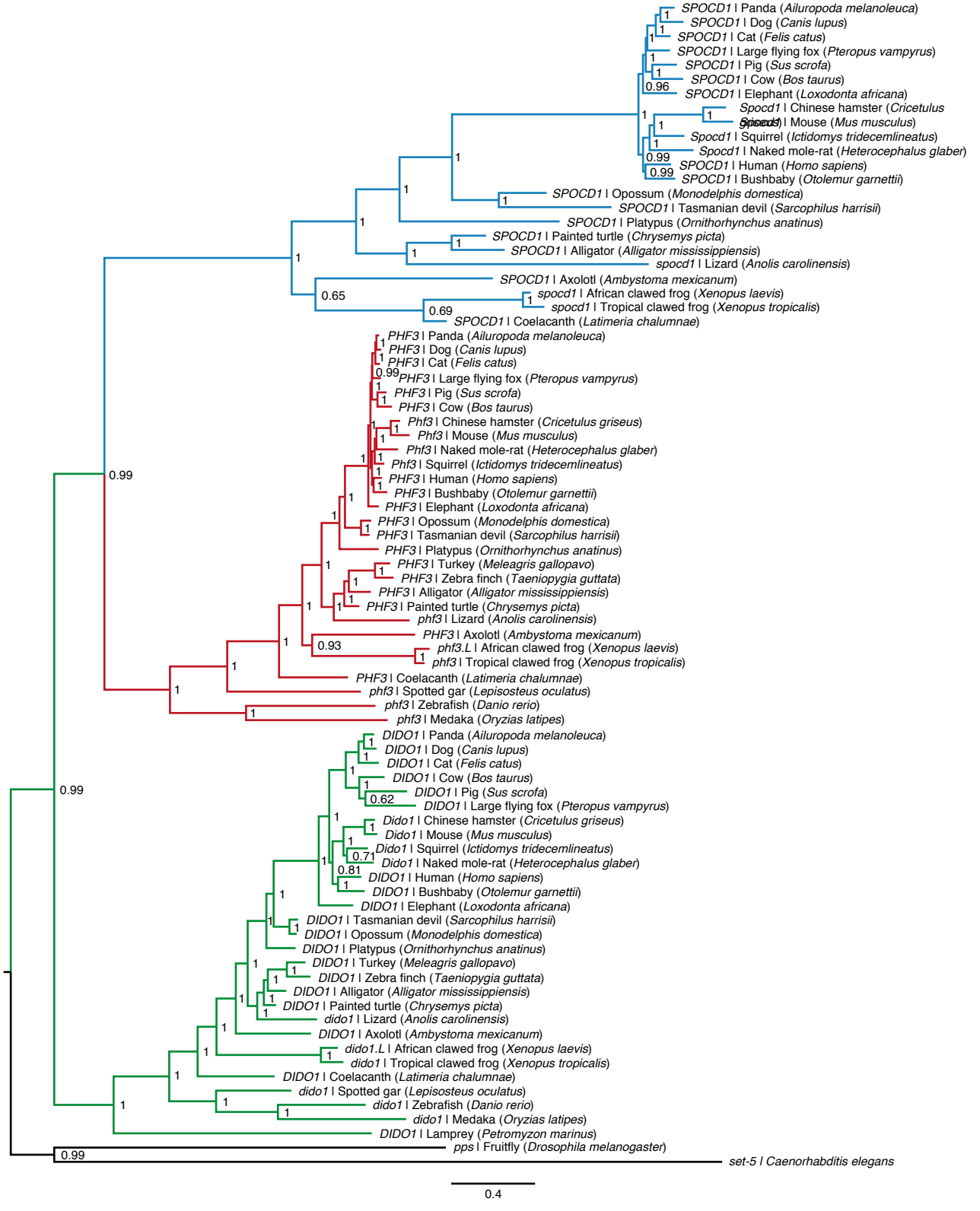
Figure 4

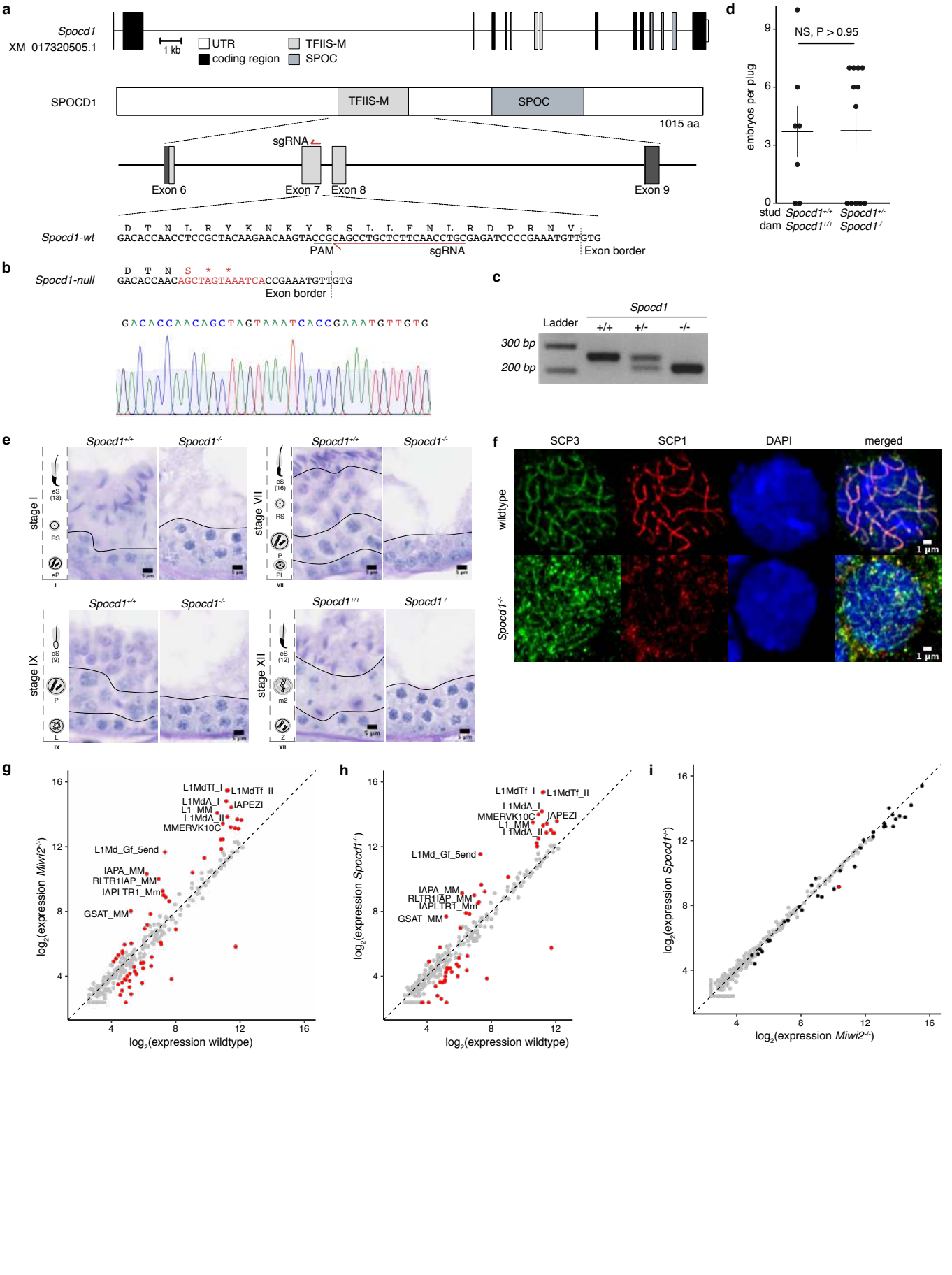

a
Spocd1HA
Exon 14
Exon 15 HA
SPOCD1-HA
TFIIS-M
SPOC
HA
b
Spocd1HA/+
DAPI
HA
c
E13.5 E14.5 E15.5 E16.5 E18.5 P1 P5
Spocd1HA/+
HA
merged
DAPI
HA
d
Miwi2HA/+
HA
merged
DAPI
HA
e
DNMT3
NURD
BAF
SPOCD1
DNMT3A
DNMT3L
Confidence
Enrichment
f
de novo DNA methylation
DNMT3L
DNMT3A
NURD complex
CHD4 CBX1 MTA1
CDK2AP1 HDAC1 MTA2
GATAD2B MBD3 MTA3
BAF complex
ARID1A SMARCD1
ARID1B SMARCA5
BRG1 HDAC1
SPOCD1
other high confidence proteins
TPR C19ORF84 TAP26 SIN3A BAP18
SPINDLIN AMBRA1 IMPORTIN5 MARK2 SAP30
PGAM1 SUMO2 ZBTB14 GTF3C1 PGK1
g
piRNA pathway
NURD
BAF
MIWI2
SPOCD1
CBX1
MTA3
SMARCA5
ARID1A
Confidence
Enrichment
h
SPOCD1 dataset
54 19 51
MIWI2 dataset 2
NURD α-ACTININ4 SKT
CBX1 C19ORF84 SPINDLIN
MTA3 CHAMP1 SPOCD1
HSP90B1 SPT6
BAF PEBP1 TPR
ARID1A PGAM1 α4-Tropomyosin
SMARCA5 PTMA U1-A
Serotransferrin

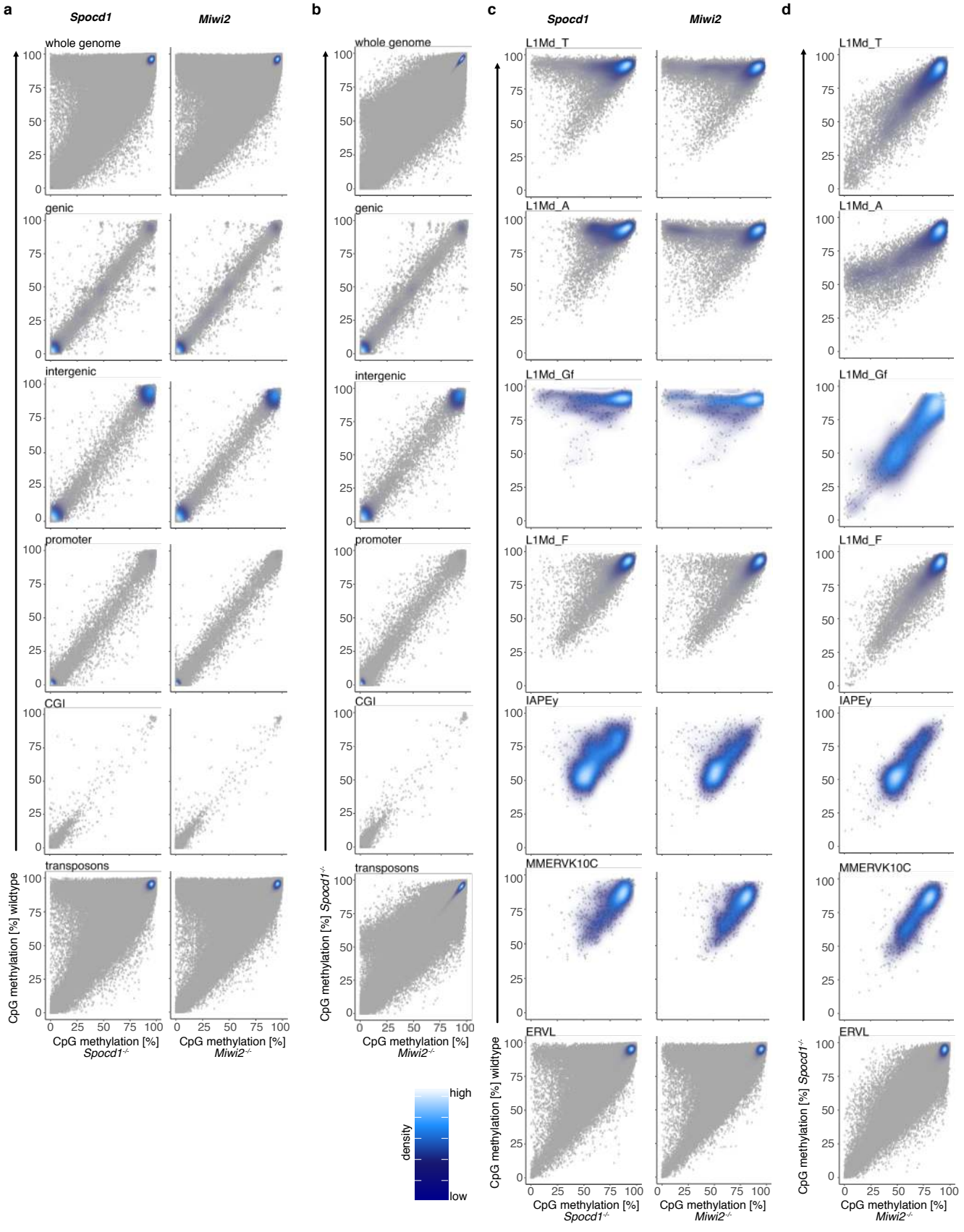
Extendend Data

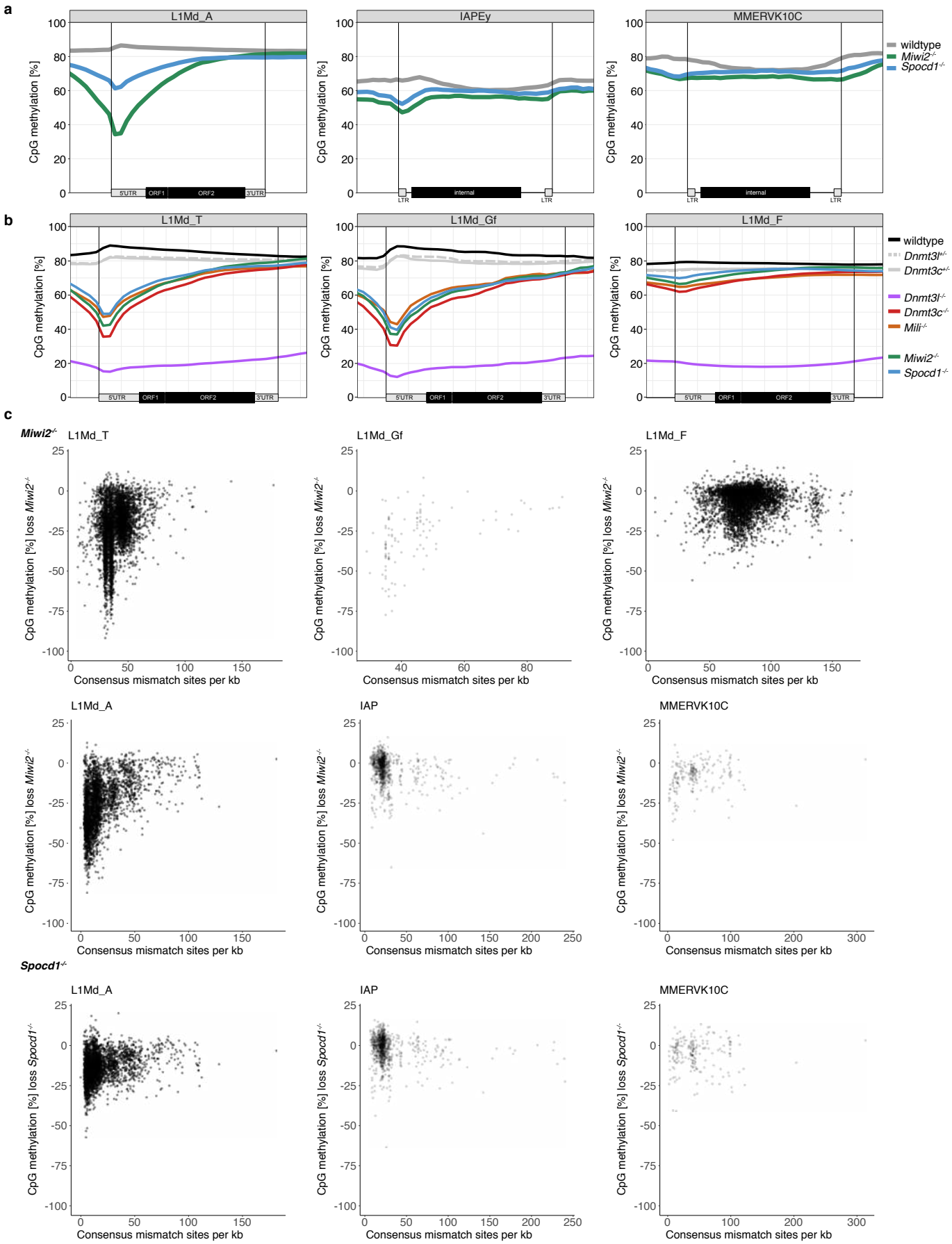

a
Miwi2
Dnmt3l
Mili
Vasa
b
Adad1
Bag2
Chchd3
Dcaf8
Dnajc7
Ppp6c
Ptges3
Rpap3
Rrp1
Spocd1
Srsf3
Usp19
Usp47
Usp9x
Vdac3
Wdtc1
NLS
Gonocytes
Spermatogonia
Spermatocytes
MEF
Bone Marrow

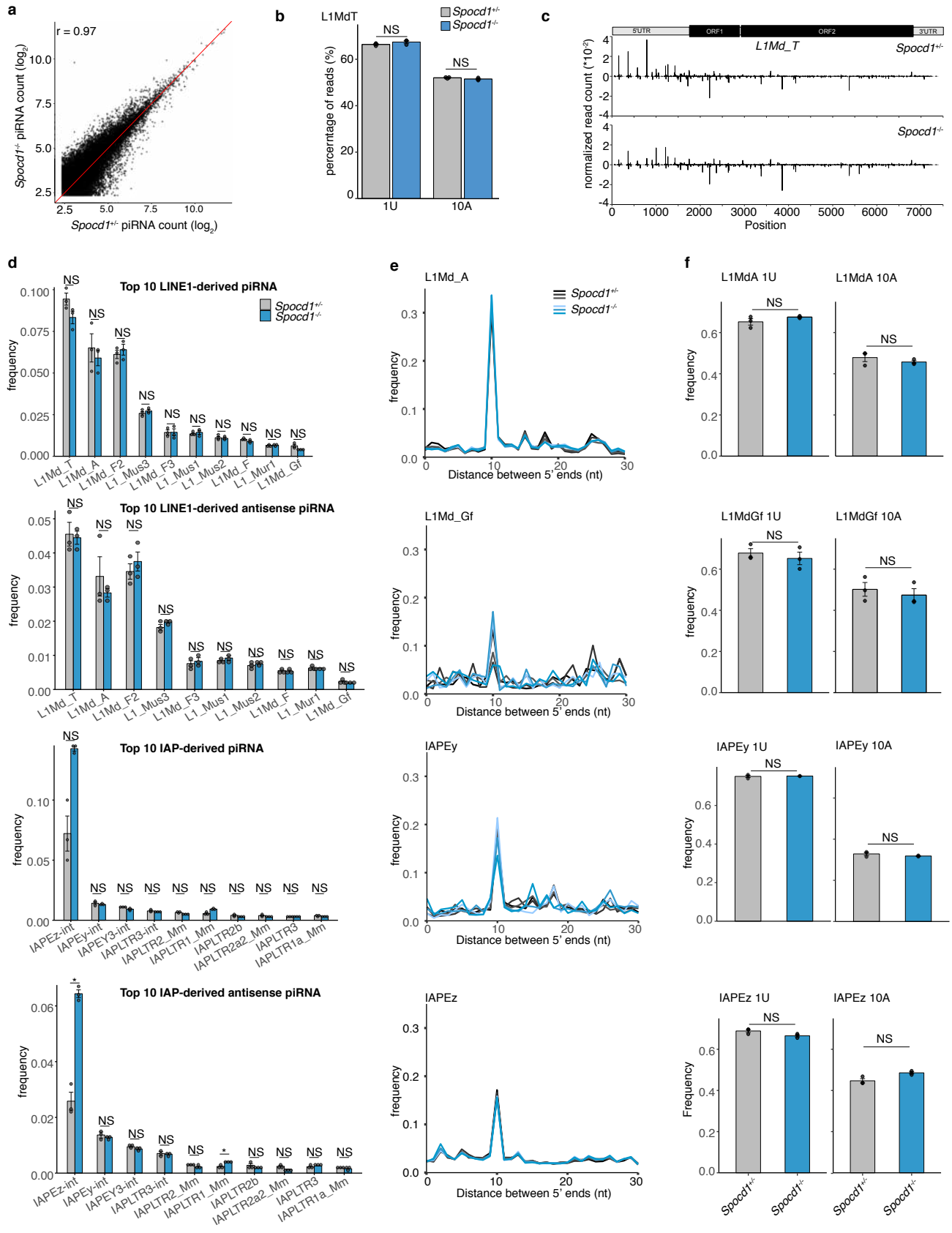

a SPOC domain
SPOCD1_mouse
SPOCD1_human
SPOCD1_lizard
SPOCD1_platypus
PHF3_mouse
PHF3_human
PHF3_lizard
PHF3_zebrafish
PHF3_platypus
DIDO1_mouse
DIDO1_human
DIDO1_lizard
DIDO1_zebrafish
DIDO1_platypus
SHARP_human
SPEN_mouse
SPEN_lizard
SPEN_zebrafish
SPEN_platypus
SHARP_human
alpha helix
beta strand
sequence identity
>80%
>60%
>40%
<40%
b TFIIS-M domain
PHF3_human
TFIIS_human

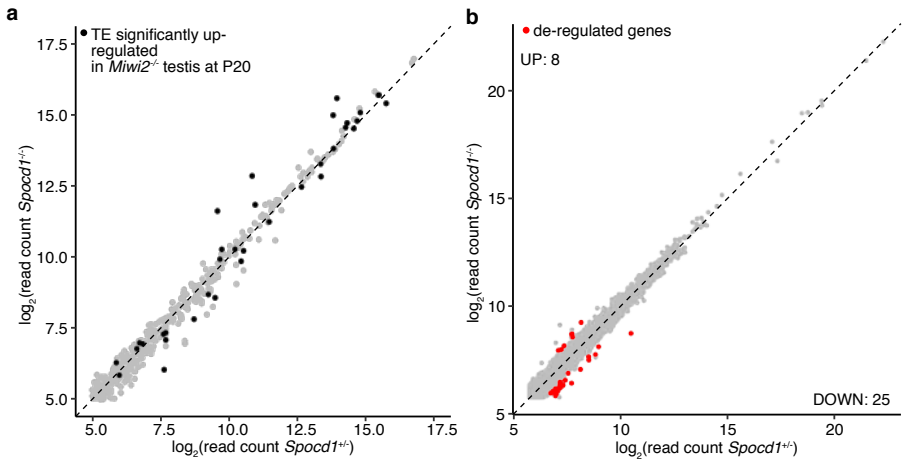

SPOCD1 | Panda (Ailuropoda melanoleuca)
SPOCD1 | Dog (Canis lupus)
SPOCD1 | Cat (Felis catus)
SPOCD1 | Large flying fox (Pteropus vampyrus)
SPOCD1 | Pig (Sus scrofa)
SPOCD1 | Cow (Bos taurus)
SPOCD1 | Elephant (Loxodonta africana)
Spocd1 | Chinese hamster (Cricetulus griseus)
Spocd1 | Mouse (Mus musculus)
Spocd1 | Squirrel (Ictidomys tridecemlineatus)
Spocd1 | Naked mole-rat (Heterocephalus glaber)
SPOCD1 | Human (Homo sapiens)
SPOCD1 | Bushbaby (Otolemur garnettii)
SPOCD1 | Opossum (Monodelphis domestica)
SPOCD1 | Tasmanian devil (Sarcophilus harrisii)
SPOCD1 | Platypus (Ornithorhynchus anatinus)
SPOCD1 | Painted turtle (Chrysemys picta)
SPOCD1 | Alligator (Alligator mississippiensis)
spocd1 | Lizard (Anolis carolinensis)
SPOCD1 | Axolotl (Ambystoma mexicanum)
spocd1 | African clawed frog (Xenopus laevis)
spocd1 | Tropical clawed frog (Xenopus tropicalis)
SPOCD1 | Coelacanth (Latimeria chalumnae)
PHF3 | Panda (Ailuropoda melanoleuca)
PHF3 | Dog (Canis lupus)
PHF3 | Cat (Felis catus)
PHF3 | Large flying fox (Pteropus vampyrus)
PHF3 | Pig (Sus scrofa)
PHF3 | Cow (Bos taurus)
Phf3 | Chinese hamster (Cricetulus griseus)
Phf3 | Mouse (Mus musculus)
Phf3 | Naked mole-rat (Heterocephalus glaber)
Phf3 | Squirrel (Ictidomys tridecemlineatus)
PHF3 | Human (Homo sapiens)
PHF3 | Bushbaby (Otolemur garnettii)
PHF3 | Elephant (Loxodonta africana)
PHF3 | Opossum (Monodelphis domestica)
PHF3 | Tasmanian devil (Sarcophilus harrisii)
PHF3 | Platypus (Ornithorhynchus anatinus)
PHF3 | Turkey (Meleagris gallopavo)
PHF3 | Zebra finch (Taeniopygia guttata)
PHF3 | Alligator (Alligator mississippiensis)
PHF3 | Painted turtle (Chrysemys picta)
phf3 | Lizard (Anolis carolinensis)
PHF3 | Axolotl (Ambystoma mexicanum)
phf3.L | African clawed frog (Xenopus laevis)
phf3 | Tropical clawed frog (Xenopus tropicalis)
PHF3 | Coelacanth (Latimeria chalumnae)
phf3 | Spotted gar (Lepisosteus oculatus)
phf3 | Zebrafish (Danio rerio)
phf3 | Medaka (Oryzias latipes)
DIDO1 | Panda (Ailuropoda melanoleuca)
DIDO1 | Dog (Canis lupus)
DIDO1 | Cat (Felis catus)
DIDO1 | Cow (Bos taurus)
DIDO1 | Pig (Sus scrofa)
DIDO1 | Large flying fox (Pteropus vampyrus)
Dido1 | Chinese hamster (Cricetulus griseus)
Dido1 | Mouse (Mus musculus)
Dido1 | Squirrel (Ictidomys tridecemlineatus)
Dido1 | Naked mole-rat (Heterocephalus glaber)
DIDO1 | Human (Homo sapiens)
DIDO1 | Bushbaby (Otolemur garnettii)
DIDO1 | Elephant (Loxodonta africana)
DIDO1 | Tasmanian devil (Sarcophilus harrisii)
DIDO1 | Opossum (Monodelphis domestica)
DIDO1 | Platypus (Ornithorhynchus anatinus)
DIDO1 | Turkey (Meleagris gallopavo)
DIDO1 | Zebra finch (Taeniopygia guttata)
DIDO1 | Alligator (Alligator mississippiensis)
DIDO1 | Painted turtle (Chrysemys picta)
dido1 | Lizard (Anolis carolinensis)
DIDO1 | Axolotl (Ambystoma mexicanum)
dido1.L | African clawed frog (Xenopus laevis)
dido1 | Tropical clawed frog (Xenopus tropicalis)
DIDO1 | Coelacanth (Latimeria chalumnae)
dido1 | Spotted gar (Lepisosteus oculatus)
dido1 | Zebrafish (Danio rerio)
dido1 | Medaka (Oryzias latipes)
DIDO1 | Lamprey (Petromyzon marinus)
pps | Fruitfly (Drosophila melanogaster)
set-5 | Caenorhabditis elegans
0.4

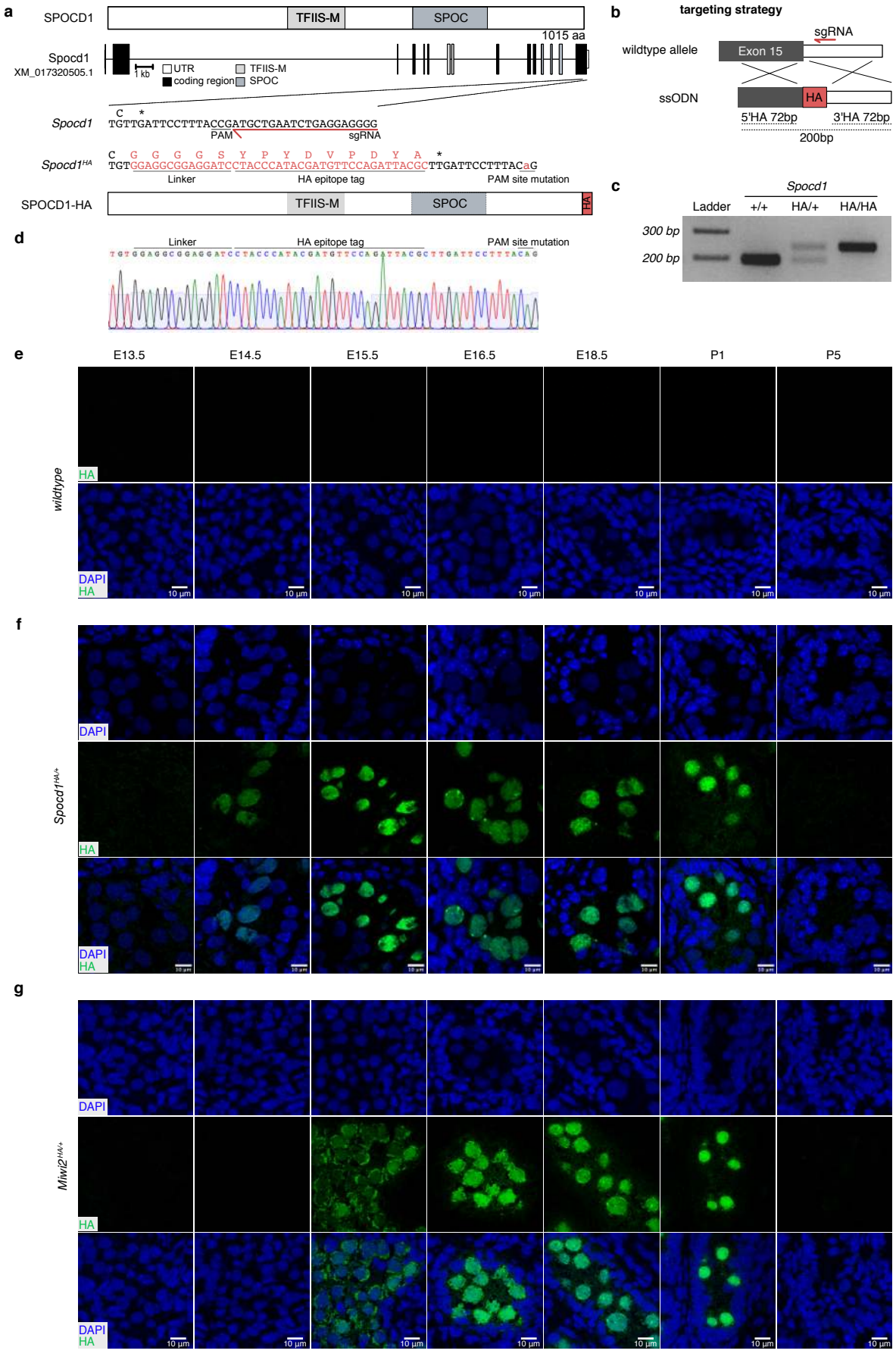

a
Spocd1
XM_017320505.1
1 kb
UTR
coding region
TFIIS-M
SPOC
SPOCD1
1015 aa
sgRNA
Exon 6
Exon 7
Exon 8
Exon 9
Spocd1-wt
PAM
sgRNA
Exon border
b
Spocd1-null
Exon border
c
Spocd1
Ladder
+/+
+/-
-/-
300 bp
200 bp
d
NS, P > 0.95
embryos per plug
stud Spocd1+/+ Spocd1+/-
dam Spocd1+/+ Spocd1-/-
e
Spocd1+/+
Spocd1-/-
stage I
stage VII
stage IX
stage XII
f
SCP3
SCP1
DAPI
merged
wildtype
Spocd1-/-
1 µm
g
log2(expression Miwi2-/-)
log2(expression wildtype)
L1MdTf_I
L1MdTf_II
L1MdA_I
IAPEZI
L1_MM
L1MdA_II
MMERVK10C
L1Md_Gf_5end
IAPA_MM
RLTR1IAP_MM
IAPLTR1_Mm
GSAT_MM
h
log2(expression Spocd1-/-)
log2(expression wildtype)
i
log2(expression Spocd1-/-)
log2(expression Miwi2-/-)


a
Spocd1
Miwi2
whole genome
genic
intergenic
promoter
CGI
transposons
CpG methylation [%] wildtype
CpG methylation [%]
Spocd1-/-
CpG methylation [%]
Miwi2-/-
b
whole genome
genic
intergenic
promoter
CGI
transposons
CpG methylation [%] Spocd1-/-
CpG methylation [%]
Miwi2-/-
high
density
low
c
Spocd1
Miwi2
L1Md_T
L1Md_A
L1Md_Gf
L1Md_F
IAPEy
MMERVK10C
ERVL
CpG methylation [%] wildtype
CpG methylation [%]
Spocd1-/-
CpG methylation [%]
Miwi2-/-
d
L1Md_T
L1Md_A
L1Md_Gf
L1Md_F
IAPEy
MMERVK10C
ERVL
CpG methylation [%] Spocd1-/-
CpG methylation [%]
Miwi2-/-

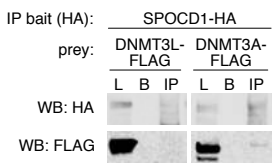

a
L1Md_A
IAPEy
MMERVK10C
CpG methylation [%]
5'UTR
ORF1
ORF2
3'UTR
internal
LTR
wildtype
Miwi2-/-
Spocd1-/-
b
L1Md_T
L1Md_Gf
L1Md_F
wildtype
Dnmt3l+/-
Dnmt3c+/-
Dnmt3l-/-
Dnmt3c-/-
Mili-/-
Miwi2-/-
Spocd1-/-
c
Miwi2-/-
L1Md_T
L1Md_Gf
L1Md_F
CpG methylation [%] loss Miwi2-/-
Consensus mismatch sites per kb
L1Md_A
IAP
MMERVK10C
Spocd1-/-
CpG methylation [%] loss Spocd1-/-


a
r = 0.97
Spocd1-/- piRNA count (log2)
Spocd1+/- piRNA count (log2)
b
L1MdT
Spocd1+/-
Spocd1-/-
percentage of reads (%)
NS
1U
10A
c
5'UTR
ORF1
ORF2
3'UTR
L1Md_T
Spocd1+/-
Spocd1-/-
normalized read count (*10-2)
Position
d
Top 10 LINE1-derived piRNA
frequency
L1Md_T
L1Md_A
L1Md_F2
L1_Mus3
L1Md_F3
L1_Mus1
L1_Mus2
L1Md_F
L1_Mur1
L1Md_Gf
Top 10 LINE1-derived antisense piRNA
Top 10 IAP-derived piRNA
IAPEz-int
IAPEy-int
IAPEY3-int
IAPLTR3-int
IAPLTR2_Mm
IAPLTR1_Mm
IAPLTR2b
IAPLTR2a2_Mm
IAPLTR3
IAPLTR1a_Mm
Top 10 IAP-derived antisense piRNA
e
L1Md_A
L1Md_Gf
IAPEy
IAPEz
Distance between 5' ends (nt)
f
L1MdA 1U
L1MdA 10A
L1MdGf 1U
L1MdGf 10A
IAPEy 1U
IAPEy 10A
IAPEz 1U
IAPEz 10A
Frequency
Spocd1+/-
Spocd1-/-

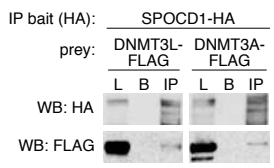

a
TE significantly up-regulated in Miwi2-/- testis at P20
log2(read count Spocd1-/-)
log2(read count Spocd1+/-)
17.5
15.0
12.5
10.0
7.5
5.0
b
de-regulated genes
UP: 8
DOWN: 25
20
15
10
5


a
SPOCD1
TFIIS-M
SPOC
1015 aa
Spocd1
XM_017320505.1
1 kb
UTR
coding region
TFIIS-M
SPOC
Spocd1
C *
TGTTGATTCCTTTACCGATGCTGAATCTGAGGAGGGG
PAM
sgRNA
Spocd1HA
C G G G G S Y P Y D V P D Y A *
TGTGGAGGCGGAGGATCCTACCCATACGATGTTCCAGATTACGCTTGATTCCTTTACaG
Linker
HA epitope tag
PAM site mutation
SPOCD1-HA
TFIIS-M
SPOC
HA
b
targeting strategy
sgRNA
wildtype allele
Exon 15
ssODN
HA
5'HA 72bp
3'HA 72bp
200bp
c
Spocd1
Ladder
+/+
HA/+
HA/HA
300 bp
200 bp
d
Linker
HA epitope tag
PAM site mutation
e
E13.5
E14.5
E15.5
E16.5
E18.5
P1
P5
wildtype
HA
DAPI
HA
10 µm
f
Spocd1HA/+
DAPI
HA
DAPI
HA
10 µm
g
Miwi2HA/+
DAPI
HA
DAPI
HA
10 µm

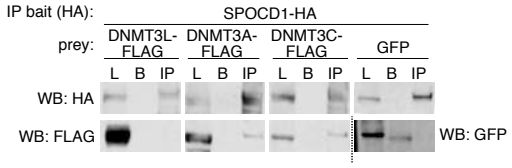
Experiment 1



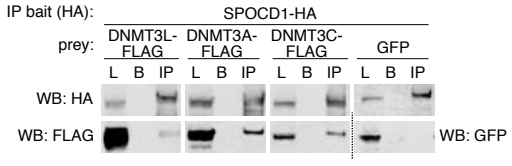
Experiment 2



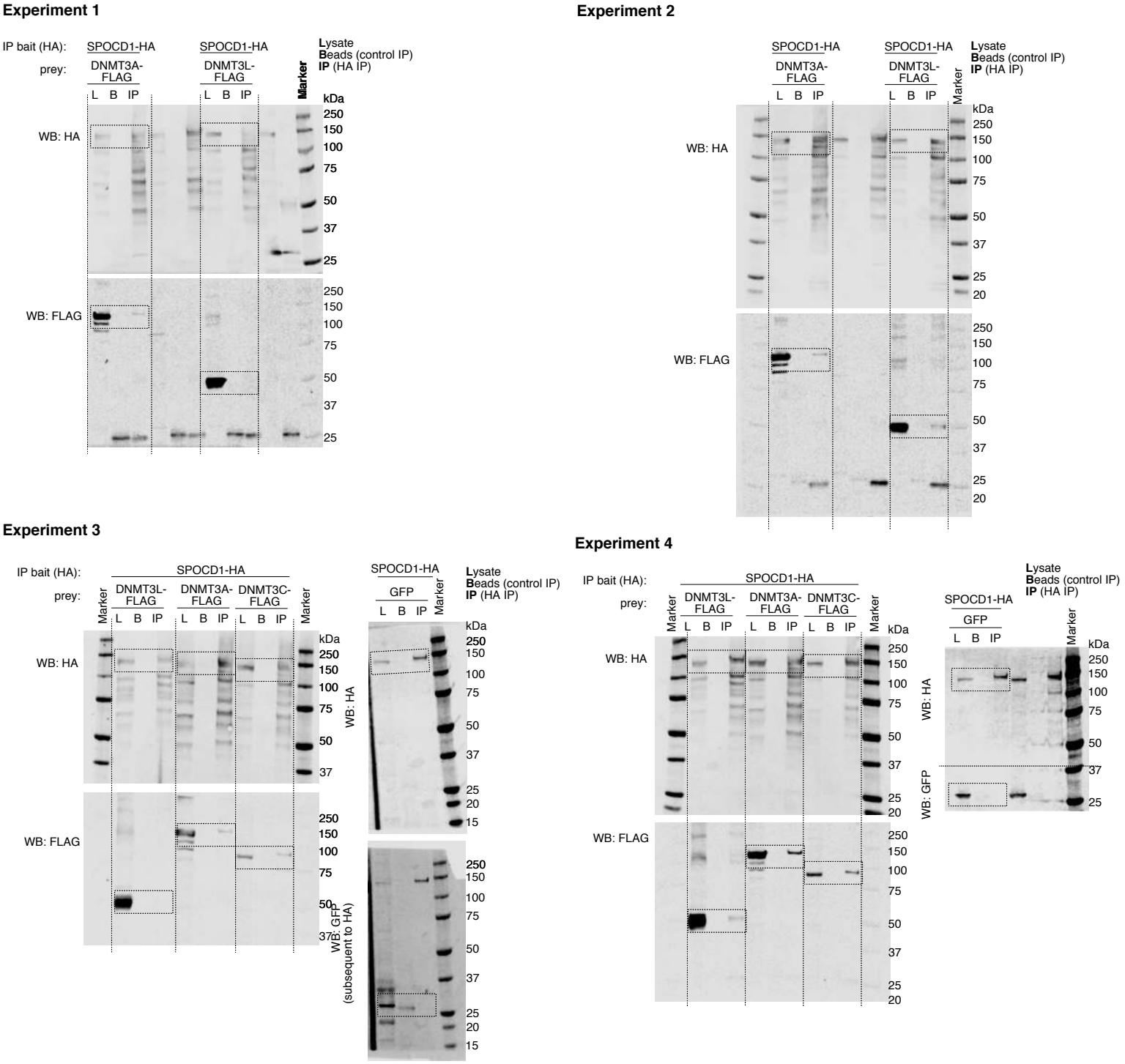
Experiment 3



Experiment 4



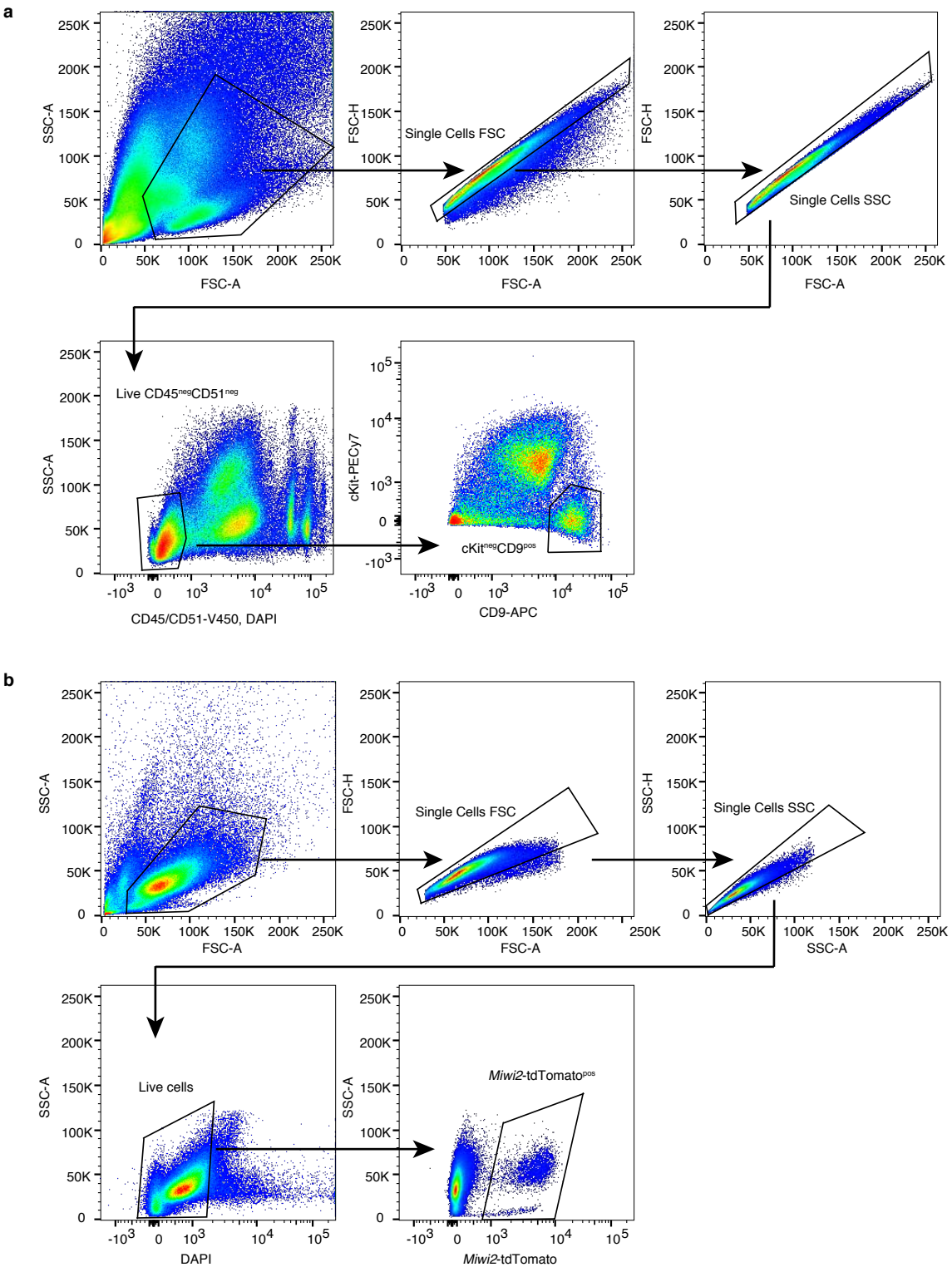
Supplementary Figure 1



Supplementary Figure 1 | Uncropped Western blot images

Uncropped source data of experiments shown in Extended Data Figure 10: Western blot analysis of co-immunoprecipitation of SPOCD1-HA with DNMT3L-FLAG, DNMT3A-FLAG, DNMT3C-FLAG or GFP in HEK cells. Shown are lysate sample (L), control IP (protein G beads) (B) and anti-HA IP (IP) for 4 experiments.

Supplementary Figure 2



Supplementary Figure 2 | FACS Gating strategy

a, FACS isolation of post-natal day 14 (P14) undifferentiated spermatogonia. Representative example of a gating strategy for sorting live CD45^{neg}CD51^{neg}c-Kit^{pos}CD9^{pos} undifferentiated spermatogonia from P14 mice, where a single-cell suspension of testicular cells was stained with anti-CD45^{biotin}, anti-CD51^{biotin}, anti-c-Kit$^{PE\text{-}Cy7}$ and anti-CD9^{APC} antibodies as well as streptavidin^{V450} and DAPI. A representative experiment of n=3 mice is shown. **b,** FACS isolation of E16.5 foetal gonocytes using the *Miwi2*^{tdTOM} reporter allele. Representative example of a gating strategy for sorting live *Miwi2*-tdTomato^{pos} gonocytes from E16.5 mice, where a single-cell suspension of testicular cells was stained with DAPI. A representative experiment of n=3 mice is shown.

Supplementary Table 1

protein.ID	protein	gene	P-value	fold-enrichment
Q8CGT6	MIWI2	*Piwil4*	2.1E-06	404.5
Q91XW8	FKBP6	*Fkbp6*	3.1E-06	44.0
E9Q9M5	USP19	*Usp19*	3.5E-05	18.7
D3Z7C6	PTGES3	*Ptges3*	3.8E-05	4.8
Q14BI7	TDRD9	*Tdrd9*	5.4E-05	36.8
Q8N7N5	DCAF8	*Dcaf8*	7.2E-05	8.4
Q14DK4	GPAT2	*Gpat2*	1.1E-04	9.0
Q99MV7	RNF17	*Rnf17*	1.2E-04	12.7
A0A0G2JFB2	TDRKH	*Tdrkh*	1.3E-04	56.8
F2YMC1	MOV10L1	*Mov10l1*	1.6E-04	36.7
Q9D706	RPAP3	*Rpap3*	2.6E-04	10.1
Q8VD46	GASZ	*Asz1*	3.1E-04	14.8
D3YUE6	GTSF1	*Gtsf1*	3.3E-04	6.8
Q91YN9	BAG2	*Bag2*	3.4E-04	4.7
Q80ZK9	WDTC1	*Wdtc1*	4.4E-04	6.8
Q8CAE2	HENMT1	*Henmt1*	7.3E-04	45.0
Q99MV1	TDRD1	*Tdrd1*	1.6E-03	6.4
B1ASB6	SPOCD1	*Spocd1*	2.4E-03	12.6
Q8CDG1	MILI	*Piwil2*	2.6E-03	11.0
A0A0N4SVL9	PPP6C	*Ppp6c*	3.1E-03	4.6
Q58E35	RPLP1	*Rplp1*	3.8E-03	4.7
Q9CRB9	MIC19	*Chchd3*	6.1E-03	9.1
Q3TX38	VDAC3	*Vdac3*	1.6E-02	7.1
Q5SUE7	ADAD1	*Adad1*	1.7E-02	5.5
Q3U781	SRSF3	*Srsf3*	2.5E-02	8.2
Q3UL32	DNAJC7	*Dnajc7*	2.8E-02	4.7
Q8BY87	USP47	*Usp47*	2.9E-02	10.7
Q4FE56	USP9X	*Usp9x*	3.2E-02	5.3
Q8BVN9	MAEL	*Mael*	4.9E-02	4.1

Supplementary Data Table 1 | Proteins identified as MIWI2 interactors.

Table listing all statistically significant (P<0.05, two-sided Student's t-test, n=3) proteins that are at least 4-fold enriched in the HA-MIWI2 immuno-precipitation.

Supplementary Table 2

TE	fold-change	P-value
IAPA_MM	5.09	8.72E-05
RLTR1IAP_MM	4.67	1.25E-07
IAPEZI	3.52	1.85E-05
IAPLTR1a_I_MM	2.46	2.65E-04
ERVB4_1B-I_MM	0.48	5.69E-11
IAPEY5_LTR	0.48	9.05E-04
RLTR6I_MM	0.43	7.46E-12
ERVB4_2-I_MM	0.39	9.74E-06
RLTR10B2	0.39	1.07E-11
MURVY-int	0.38	3.74E-04
ERVB2_1A-I_MM	0.30	6.34E-09
ZP3AR	0.26	3.19E-11
ERVB4_3-I_MM	0.19	4.55E-06

Supplementary Data Table 2 | Deregulated TEs in E16.5 *Spocd1*$^{-/-}$ gonocytes.

Table of all statistically significant (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test), deregulated (>2-fold) TEs identified in E16.5 *Spocd1*$^{-/-}$ gonocytes, listing fold-change of expression (relative to wildtype) and P-values as determined from RNA-seq data. n=3 for both genotypes.

Supplementary Table 3

ensemble ID	gene_name	fold-change	P-value
ENSMUSG00000058267	*Mrps14*	3.02	1.23E-03
ENSMUSG00000034544	*Rsrc1*	2.69	2.11E-05
ENSMUSG00000041923	*Nol4*	2.66	1.09E-05
ENSMUSG00000031099	*Smarca1*	2.63	9.75E-07
ENSMUSG00000026748	*Plxdc2*	2.50	2.07E-03
ENSMUSG00000010721	*Lmbr1*	2.48	4.96E-04
ENSMUSG00000072774	*Zfp951*	2.39	2.30E-04
ENSMUSG00000023087	*Noct*	2.29	7.60E-03
ENSMUSG00000004668	*Abca13*	0.44	2.61E-05
ENSMUSG00000037747	*Phyhipl*	0.40	1.11E-05
ENSMUSG00000025014	*Dntt*	0.35	6.42E-03
ENSMUSG00000079173	*Zan*	0.33	3.55E-06
ENSMUSG00000036526	*Card11*	0.33	9.75E-07
ENSMUSG00000030259	*Rassf8*	0.28	8.21E-04
ENSMUSG00000028784	*Spocd1*	0.25	1.40E-18
ENSMUSG00000030577	*Cd22*	0.24	9.26E-03
ENSMUSG00000031022	*BC051019*	0.23	8.10E-03
ENSMUSG00000074277	*Phldb3*	0.22	4.09E-04
ENSMUSG00000036634	*Mag*	0.18	8.08E-03
ENSMUSG00000078234	*Klhdc7a*	0.13	3.49E-05
ENSMUSG00000002980	*Bcam*	0.11	1.69E-06
ENSMUSG00000114818	*Gm35164*	0.11	4.96E-04
ENSMUSG00000083767	*Gm11405*	0.10	1.59E-03
ENSMUSG00000109613	*Gm45405*	0.10	1.77E-13
ENSMUSG00000068758	*Il3ra*	0.09	1.85E-06
ENSMUSG00000066361	*Serpina3c*	0.07	1.45E-04
ENSMUSG00000109326	*Gm9165*	0.06	8.95E-04
ENSMUSG00000090126	*4930519F09Rik*	0.06	1.42E-04
ENSMUSG00000040525	*Cblc*	0.06	8.21E-04
ENSMUSG00000085658	*Gm15704*	0.06	5.15E-06
ENSMUSG00000082398	*Gm11927*	0.04	3.90E-03
ENSMUSG00000108865	*Gm44617*	0.02	2.52E-05
ENSMUSG00000109136	Gm45114	0.01	1.42E-04

Supplementary Data Table 3 | Deregulated genes in E16.5 *Spocd1*$^{-/-}$ foetal gonocytes.

Table of all statistically significant (P<0.01, Benjamini-Hochberg adjusted two-sided Wald's test), de-regulated (>2-fold) genes identified in E16.5 *Spocd1*$^{-/-}$ gonocytes, listing fold-change of expression (relative to wildtype) and P-values as determined from RNA-seq data. n=3 for both genotypes.

Supplementary Table 4

protein.ID	protein	gene	P-value	fold-enrichment
Q7M739	TPR	*Tpr*	4.2E-08	239.9
B1ASB6	SPOCD1	*Spocd1*	6.8E-06	516.5
Q61142	SPINDLIN	*Spin1*	1.2E-05	93.6
Q3U7Z6	PGAM1	*Pgam1*	2.1E-05	10.5
H3BKT1	C19ORF84H	*Gm38999*	3.3E-05	38.7
A2AH22	AMBRA1	*Ambra1*	3.8E-05	4.5
A2BH40	ARID1A	*Arid1a*	5.2E-05	7.4
A3EWM2	DNMT3L	*Dnmt3l*	7.2E-05	5.3
H7BWX9	SUMO2	*Sumo2*	7.3E-05	39.4
P83917	CBX1	*Cbx1*	1.0E-04	16.6
Q3U3A7	MTA3	*Mta3*	1.4E-04	12.3
Q8R2N0	TAP26	*Ccdc59*	1.8E-04	5.1
Q8BKC5	IMPORTIN5	*Ipo5*	1.8E-04	6.2
A1L3S7	GATAD2B	*Gatad2b*	1.8E-04	4.2
Q91ZW3	SMARCA5	*Smarca5*	1.8E-04	18.0
A0A0R4J2B6	RBBP5	*Rbbp5*	2.5E-04	7.0
Q544H8	ZBTB14	*Zbtb14*	2.5E-04	5.1
O88508	DNMT3A	*Dnmt3a*	3.2E-04	4.4
Q60520	SIN3A	*Sin3a*	3.4E-04	9.7
E9Q6R4	ARID1B	*Arid1b*	3.6E-04	11.9
E9Q9N6	MARK2	*Mark2*	4.5E-04	4.5
Q8K284	GTF3C1	*Gtf3c1*	4.8E-04	4.3
Q9R190	MTA2	*Mta2*	6.9E-04	7.1
Q9DCT6	BAP18	*Bap18*	1.1E-03	5.9
E9QAS4	CHD4	*Chd4*	1.2E-03	5.1
F8WHY8	MTA1	*Mta1*	1.2E-03	8.7
O88574	SAP30	*Sap30*	1.2E-03	4.2
Q3TKT4	BRG1	*Smarca1*	1.5E-03	7.0
P09411	PGK1	*Pgk1*	1.5E-03	4.1
Q9Z2D8	MBD3	*Mbd3*	1.6E-03	4.4
P24549	RALDH1	*Aldh1a1*	1.7E-03	5.5
Q99MX1	USP26	*Usp26*	1.7E-03	4.7
Q9CU65	ZMYM2	*Zmym2*	1.9E-03	5.2
A0A1W2P7G2	SARNP	*Sarnp*	2.1E-03	4.3
Q923G2	RPABC3	*Polr2h*	2.1E-03	5.3
Q6DID5	MUM1	*Mum1*	2.3E-03	4.9
Q99JR8	SMARCD1	*Smarcd2*	2.4E-03	5.3
Q5EBQ2	PEBP1	*Pebp1*	2.5E-03	6.5
Q921I1	Serotransferrin	*Tf*	2.7E-03	6.5
Q6ZWM8	PPP1CC	*Ppp1cc*	3.1E-03	5.1
Q62383	SPT6	*Supt6h*	4.7E-03	8.6
Q6URW6	Myosin14	*Myh14*	5.2E-03	4.6
A0A1D5RLL4	TRRAP	*Trrap*	5.9E-03	5.6
G3X9B1	HEAT1R1	*Heatr1*	7.8E-03	5.2
A8DUK2	beta-globin	*Hbbt1*	8.1E-03	8.5
A2AQ25	SKT	*Skt*	8.3E-03	4.7
Q8CFG0	SULF2	*Sulf2*	8.7E-03	4.4
Q58E49	HDAC1	*Hdac1*	9.2E-03	6.5
Q3ULT2	α-ACTININ4	*Actn4*	9.4E-03	9.8
P10639	Thioredoxin	*Txn*	1.0E-02	5.6
O35207	Cdk2ap1	*Cdk2ap1*	1.1E-02	4.3
P48036	AnnexinA5	*Anxa5*	1.2E-02	4.2
A3KMF2	MKL2	*Mkl2*	1.2E-02	4.2
Q61103	DPF2	*Dpf2*	1.2E-02	5.2
A0A140T8S5	CHAMP1	*Champ1*	1.3E-02	6.0
A0A0N4SVC2	Transformer2A	Tra2a	1.3E-02	4.0
F8VPR5	CREB-BP	*Crebp*	1.3E-02	4.3
Q9DBJ3	BAIAP2L1	*Baiap2l1*	1.4E-02	5.5
P15626	GSTM2	*Gstm2*	1.4E-02	8.6
Q6IRU2	α4-Tropomyosin	*Tpm4*	1.5E-02	4.9
Q3UEK9	AHSG	*Ahsg*	1.8E-02	4.0
Q9JJV2	Profilin2	*Pfn2*	1.9E-02	4.8
Q61391	Neprilysin	*Mme*	1.9E-02	4.6
Q91V38	HSP90B1	*Hsp90b1*	2.2E-02	6.7
F6VXG5	WT1	*Wt1*	2.4E-02	4.5
Q3U6E4	PTMA	*Ptma*	2.5E-02	7.0
Q8C7S2	LIMA1	*Lima1*	2.6E-02	5.1
Q8CIG3	LSD2	*Kdm1b*	3.5E-02	4.2
Q9Z103	ADNP	*Adnp*	3.5E-02	5.7
Q8CGZ0	CHERP	*Cherp*	3.5E-02	4.5
Q9JI95	CPN10-like	*Cpn10-rs1*	4.0E-02	4.9
Q3U5S6	Syndecan	*Sdc4*	4.1E-02	4.5
Q62189	U1-A	*Snrpa*	4.6E-02	5.6

Supplementary Data Table 4 | Proteins identified as SPOCD1 interactors.

Table listing all statistically significant (P<0.05, two-sided Student's t-test) proteins that are at least 4-fold enriched in the SPOCD1-HA immuno-precipitation.

Supplementary Table 5

protein.ID	protein	gene	P-value	fold-enrichment
Q7M739	TPR	*Tpr*	4.03E-06	38.15
Q71LX8	HSP90AB1	*Hsp90ab1*	5.03E-06	6.30
Q8CGT6	MIWI2	*Piwil4*	5.46E-06	25.37
A0A0R4J086	OLFML3	*Olfml3*	6.13E-06	10.24
Q80Y52	HSP90AA1	*Hsp90aa1*	1.21E-05	18.25
Q91XW8	FKBP6	*Fkbp6*	1.29E-05	4.91
A0A0A6YWX1	USP19	*Usp19*	1.49E-05	17.88
D3YWG8	MOV10L1	*Mov10l1*	1.89E-05	24.23
Q4FJN2	FKBP5	*Fkbp5*	2.61E-05	11.96
P30416	FKBP4	*Fkbp4*	2.95E-05	24.28
Q3THQ5	STIP1	*Stip1*	2.98E-05	58.52
B2RQL4	ADAD2	*Adad2*	3.22E-05	18.59
A0A0N4SVL9	PPP6C	*Ppp6c*	4.33E-05	6.96
Q8BY87	USP47	*Usp47*	4.68E-05	17.67
Q99MV1	TDRD1	*Tdrd1*	5.66E-05	5.22
Q8CI32	BAG5	*Bag5*	7.43E-05	14.63
D3YUE6	GTSF1	*Gtsf1*	8.63E-05	18.64
Q99MV7	RNF17	*Rnf17*	9.59E-05	22.21
B1ASB6	SPOCD1	*Spocd1*	1.77E-04	51.36
F8VPN2	TEX15	*Tex15*	1.87E-04	11.01
Q14DK4	GPAT2	*Gpat2*	1.98E-04	13.89
Q3U7Z6	PGAM1	*Pgam1*	2.48E-04	9.76
P47810	WEE1	*Wee1*	3.17E-04	4.50
Q8CAE2	HENMT1	*Henmt1*	4.53E-04	19.48
A6H5Y3	Methione Synthase	*Mtr*	5.02E-04	10.18
Q99PT1	Rho-GDI1	*Arhgdia*	5.06E-04	9.31
H3BKT1	C19ORF84H	*Gm38999*	5.45E-04	5.79
Q8N7N5	DCAF8	*Dcaf8*	6.01E-04	7.26
Q61142	SPINDLIN	*Spin1*	6.03E-04	10.48
A0A0G2JFB2	TDRKH	*Tdrkh*	7.22E-04	22.08
F7BX26	PPP5C	*Ppp5c*	9.41E-04	4.58
P62311	LSM3	*Lsm3*	9.97E-04	4.65
M0QWY0	EMC8	*Emc8*	1.09E-03	5.94
Q8VDN2	ATP1A1	*Atp1a1*	1.48E-03	4.58
D3Z7C6	PTGES3	*Ptges3*	1.60E-03	22.20
Q569Z6	THRAP3	*Thrap3*	1.65E-03	4.53
Q6P5H2	NESTIN	*Nes*	2.03E-03	5.52
Q9R0P9	UCHL1	*Uchl1*	2.64E-03	5.37
P83917	CBX1	*Cbx1*	2.71E-03	7.13
Q8VD46	GASZ	*Asz1*	2.82E-03	14.61
P02104	HemoglobinY2	*Hbby2*	2.83E-03	5.97
Q4KL76	HSPE1	*Hspe1*	3.19E-03	8.76
Q3ULT2	α-ACTININ4	*Actn4*	4.27E-03	7.74
Q62383	SPT6	*Supth6*	4.85E-03	5.34
Q80TU6	ACIN1	*Acin4*	4.91E-03	4.88
Q91ZW3	SMARCA5	*Smarca5*	5.57E-03	4.79
Q3UI98	GTF3C5	*Gtf3c5*	7.30E-03	4.36
Q921I1	Serotranferrin	*Tf*	8.15E-03	7.05
Q3TYJ0	STUB1	*Stub1*	8.75E-03	7.08
A2BH40	ARID1A	*Arid1a*	9.09E-03	6.79
Q3U3A7	MTA3	*Mta3*	1.27E-02	5.05
A0A140T8S5	CHAMP1	*Champ1*	1.31E-02	5.02
Q6IRU2	α4-Tropomyosin	*Tpm4*	1.75E-02	6.17
P27661	H2AX	*H2afx*	1.77E-02	11.48
Q5EBQ2	PEBP1	*Pebp1*	1.92E-02	7.31
Q14AI7	COPS4	*Cops4*	1.99E-02	4.02
B2RSN3	β-TUBULIN	*Tubb2b*	2.24E-02	4.47
Q5SUE7	ADAD1	*Adad1*	2.33E-02	6.57
A2AQ25	SKT	*Skt*	2.34E-02	5.02
P10649	GSTM1	*Gstm1*	2.36E-02	5.12
Q80ZK9	WDTC1	*Wdtc1*	2.47E-02	6.27
Q3UF95	BAG6	*Bag6*	2.87E-02	4.52
P14733	LAMIN-B1	*Lmnb1*	3.07E-02	21.48
Q80X90	FilaminB	*Flnb*	3.27E-02	4.10
Q8BT07	CEP55	*Cep55*	3.50E-02	4.22
A0A1B0GSX7	NUP98	*Nup98*	3.61E-02	4.50
Q91V38	HSP90B1	*Hsp90b1*	3.98E-02	9.58
Q3U6E4	PTMA	*Ptma*	4.21E-02	4.24
P40142	Transketolase	*Tkt*	4.39E-02	5.42
Q62189	U1-A	*Snrpa*	4.98E-02	4.29

Supplementary Data Table 5 l Proteins identified as MIWI2 interactors using benzonase for protein extraction.

Table listing all statistically significant (P<0.05, two-sided Student's t-test) proteins that are at least 4-fold enriched in the HA-MIWI2 immuno-precipitation.

Supplementary Table 6

protein.ID	protein	gene	complex	P-value	fold-enrichment
Q9Z2D8	MBD3	*Mbd3*	NURD	0.003	2.55
Q9R190	MTA2	*Mta2*	NURD	0.006	3.51
E9Q6R4	ARID1B	*Arid1b*	BAF	0.008	3.10
E9QAS4	CHD4	*Chd4*	NURD	0.009	2.50
F8WHY8	MTA1	*Mta1*	NURD	0.023	3.19
A3EWM2	DNMT3L	*Dnmt3l*	DNMT3	0.032	1.94

Supplementary Data Table 6 | Additional NURD, BAF and *de novo* methylation components identified in the MIWI2 IP-MS that used benzonase for protein extraction.

Table listing statistically significant (P<0.05, two-sided Student's t-test) but less than 4-fold enriched proteins from the NURD, BAF and *de novo* methylation machinery complexes co-purified in the HA-MIWI2 IPMS.

Supplementary Table 7

sample.ID	total read count	duplicated reads (removed)	duplication rate [%]	mean read length	total read length	fold coverage	conversion rate [%]
WT_1	74,950,362	5,623,257	6.98	111	24,913,208,414	9.14	99.06
WT_2	57,568,270	3,845,611	6.26	110	18,978,827,940	6.96	98.93
WT_3	43,149,146	3,102,389	6.71	110	14,222,296,134	5.22	99.10
Spocd1_1	67,538,386	4,600,180	6.38	112	22,682,079,257	8.32	98.96
Spocd1_2	52,572,930	3,622,534	6.45	108	17,040,648,058	6.25	98.95
Spocd1_3	47,609,543	3,204,297	6.31	108	15,478,149,570	5.68	99.05
Miwi2_1	68,500,420	5,295,444	7.18	111	22,785,333,541	8.36	99.01
Miwi2_2	61,952,244	4,576,972	6.88	111	20,702,606,662	7.60	99.14
Miwi2_3	66,368,499	4,948,138	6.94	112	22,339,543,890	8.20	99.09

Supplementary Data Table 7 | Sequencing statistics of Methyl-seq datasets

Table listing sequencing and alignment statistics for the Methyl-seq libraries.