In vivo versus *in silico* assessment of potentially pathogenic missense variants in human reproductive genes

Xinbao Ding[1,4], Priti Singh[1,4,5], Tina N. Tran[1], Robert Fragoza[2,3], Haiyuan Yu[2,3] and John C. Schimenti[1]*

[1] Cornell University, College of Veterinary Medicine, Department of Biomedical Sciences, Ithaca, NY 14853, USA

[2] Cornell University, Department of Computational Biology, Ithaca, NY 14853, USA

[3] Cornell University, Weill Institute for Cell and Molecular Biology, Ithaca, NY 14853, USA

[4] These authors contributed equally to this work.

[5] Current address: Fred Hutchinson Cancer Center, Seattle, WA 98109, USA

* Corresponding author: jcs92@cornell.edu

Abstract

Infertility is a highly heterogeneous condition, with genetic causes estimated to be involved in approximately half of the cases. High-throughput sequencing (HTS) approaches are becoming an increasingly important tool for genetic diagnosis of diseases, including idiopathic infertility. Nevertheless, most rare or minor alleles revealed by HTS are classified as variants of uncertain significance (VUS). Interpreting the functional impacts of VUS is challenging but profoundly important for clinical management and genetic counseling. To determine the consequences of segregating polymorphisms in key fertility genes, we functionally evaluated 8 missense variants in the genes *ANKRD31*, *BRDT*, *DMC1*, *EXOI*, *FKBP6*, *MSH4* and *SEPT12* by generating genome-edited mouse models. Six variants were classified as deleterious by most functional prediction algorithms, and two disrupted a protein-protein interaction in the yeast 2 hybrid assay. Even though these genes are known to be essential for normal meiosis or spermiogenesis in mice, none of the tested human variants compromised fertility or gametogenesis in the mouse models. These results should be useful for genetic diagnoses of infertile patients, but they also underscore the need for more effective VUS categorization. To this end, we evaluated the performance of 10 widely used pathogenicity prediction algorithms in classifying missense variants within fertility-related genes from two sources: 1) the ClinVar database, and 2) those functionally tested in mice. We found that all the algorithms performed poorly in terms of predicting phenotypes of mouse-modeled variants. These studies highlight the importance of functional validation of potential infertility-causing genetic variants.

Introduction

A major challenge in human genetics is to elucidate VUS residing within disease-associated genes. In the absence of experimental data that test functional consequences of such variants (namely single nucleotide polymorphisms, or SNPs), or phenotypic data on individuals and control groups bearing VUS, functional prediction algorithms are commonly used. However, the accuracy of these predictors is not sufficiently reliable for basing clinical decisions without corroborating information [1]. Compounding our relatively weak knowledge of human infertility genetics is the large number of genes that are required for normal fertility [2]. Thus, even for infertile individuals who have undergone whole genome or exome sequencing (WGS and WES, respectively), an actual causal SNP or private mutation will exist within a background of VUSs in candidate genes, making it difficult to conclusively implicate any single variant as being responsible for infertility. Alternatively, for all but the rarest SNPs, insights into their potential effects can be gleaned from genome-wide association studies (GWAS) designed to dissect genetic contributions to disease or phenotypic states. Unfortunately, there has been very limited success using GWAS for mapping infertility alleles, an outcome exacerbated by the likelihood that common variants are not responsible in most cases [3].

CRISPR/Cas9-mediated genome editing provides a means for evaluating potential human disease variants in an appropriate *in vivo* system. We have developed an integrated computational and experimental approach to functionally assess potentially deleterious missense variants in essential reproductive genes by modeling them in mice [4]. Our original approach taken for selecting candidate infertility missense variants for

modeling was based on the following criteria: a) They reside in genes that are essential for fertility in mice; b) they alter an amino acid conserved between mice and humans; and c) the amino acid change is predicted to be deleterious by various bioinformatic tools. Using this approach, we tested several missense variants in genes important for meiosis, finding that a substantial fraction had no clear impact on fertility [5–8]. To increase the likelihood that missense variants selected for mouse modeling impact fertility, we added two *in vitro* pre-screens to our selection pipeline [9–11]. The first pre-screen was to prioritize VUS that disrupt a known protein-protein interaction (PPI), because human disease mutations are overrepresented for such variants [12], and the second was to assess protein stability defects in cultured cells [10].

Recommendations by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) have been widely adopted by clinical laboratories around the world to guide clinical interpretation of sequence variants [13]. These guidelines provided criteria for classifying variants as pathogenic (P), likely pathogenic (LP), VUS, likely benign (LB), or benign (B). These classifications are based on distinct evidence types, each of which is assigned a level of strength. The Clinical Genome Resource (ClinGen) approves self-organized Variant Curation Expert Panels (VCEPs) in specific disease areas, e.g., hearing loss and cardiomyopathy, to make gene-centric specifications to the ACMG/AMP guidelines. The VCEPs provide disease-specific rules for accurate variant interpretation and classify variants specified for the genes and/or diseases within their scope. As a ClinGen partner, the ClinVar database archives genetic variants and assigns predicted or known impacts to diseases and other conditions [14]. The collaboration of these two entities provides complementary resources

to support genomic interpretation. Unfortunately, ClinGen does not have a reproductive- or infertility-related VCEP that would evaluate variants/mutations causing infertility, which impinges upon clinical action. Consequently, functional evidence from the literature becomes crucial. Given the lack of appropriate or *ex vivo* systems to evaluate VUS in genes involved in certain aspects reproduction such as spermatogenesis, and in the absence of well-characterized families segregating a suspect variant, *in vivo* evidence from knock-in/humanized animal models is currently one of the most rigorous ways to test the roles of VUS in fertility.

Here, we report experiments to test 8 missense variants in 7 genes crucial for various processes in meiosis or spermiogenesis. Six of these are known to play essential roles in various steps of meiotic recombination or meiotic chromosome organization (*Ankrd31, Brdt, Dmc1, Exo1, Fkbp6* and *Msh4*), and the other (*Sept12*) is important for sperm morphogenesis and motility. We performed functional interpretation by constructing and analyzing mouse models containing the orthologous amino acid substitutions. Despite *in silico* predictions or Y2H screening indicating that these variants are harmful, we found that mice bearing these variants were fertile. Our assessment of these and other functionally interpreted alleles in mouse models, as well as infertility missense variants in the ClinVar database, reveals that *in vivo* assays are crucial even when informed by computational predictions and *in vitro* assays.

Materials and Methods

Variant prioritization

The possible pathogenic functional effects of missense variants in human *DMC1, FKBP6, EXO1, MSH4, BRDT* and *ANKRD31* were analyzed using SIFT [15] and PolyPhen-2 [16]. The frequency of variants is as reported in Genome Aggregation Database (gnomAD v2.1.1 for *FKBP6, EXO1, MSH4, BRDT* and *ANKRD31*, and v3.1.1 for *DMC1*) (gnomeAD.broadinstitute.org) as of April 2021. The domain structures of human proteins were drawn using Domain Graph [17].

Y2H screening

Y2H screening for *SEPT12* variants was performed as previously described [9]. Disrupted PPIs were identified by the following criteria: a) the mutated protein reduced growth by at least 50% relative to WT as benchmarked by twofold serial dilution experiments; b) neither WT or mutant DB-ORFs were autoactivators; and c) the reduced growth phenotype was reproduced across 3 replicates. A mutation was scored as disruptive if one or more corresponding PPIs were affected.

Generation of mouse models by CRISPR/Cas9

All animal usage was approved by Cornell University's Institutional Animal Care and Use Committee, under protocol 2004-0038 to J.C.S.

Mouse models were generated by the CRISPR/Cas9 genome editing technology, as described previously [5]. sgRNAs and ssODNs are listed in Table S1. In addition to amino acid substitutions, several silent mutations were introduced to prevent the mutated region from being recognized and recut by the ribonucleoprotein. Briefly, the sgRNA, ssODN, and Cas9 mRNA (25 ng/μL, TriLink) were co-injected into zygotes (F1 hybrids between strains FVB/NJ and B6(Cg)-Tyr^{c-2J}/J) then transferred into the oviducts of pseudopregnant females. *Dmc1*, *Fkbp6*, *Exo1*, *Msh4*, *Brdt* and *Ankrd31* founders

carrying at least one copy of the desired alteration were identified and backcrossed into C57BL/6J (or FvB/NJ in the case of *Sept12* mutants) for at least three generations. For genotyping, 4 µL of crude DNA lysate was created as described previously [18] from ear punch biopsy specimens or toes of 7- to 14-day-old mice. The primers used for genotyping the mutant mice are listed in Table S1. To distinguish between WT and SNP alleles, PCR amplicons were analyzed via Sanger sequencing or restriction enzyme digestion.

Fertility testing

Two-month-old heterozygous and homozygous mice were mated with age-matched C57BL/6J (for *Dmc1*, *Fkbp6*, *Exo1*, *Msh4*, *Brdt* and *Ankrd31* mice) or FVB/NJ (for *Sept12* mice) until 7–15 months of age. The number of pups from each mating set and sex was recorded. After at least 3 months without offspring, the analyzed individuals were considered infertile, then sacrificed for more detailed reproductive evaluations as described in the following sections.

Sperm counts

Cauda epididymis sperm counting was performed using a standard method [19]. One cauda epididymis per male was used for each data point.

Histology

For the preparation of paraffin blocks, tissues were fixed overnight in Bouin's solution at room temperature, washed in 70% ethanol and then dehydrated and embedded in paraffin as previously described [5]. Six-micron sections were deparaffinized and stained by hematoxylin and eosin (H&E).

Terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) staining

Testes were fixed in 4% paraformaldehyde and embedded in paraffin. Sections were deparaffinized and performed TUNEL staining using the DeadEnd™ Fluorometric TUNEL System (Promega, G3250) following manufacturer's instructions.

Follicle counts

Ovaries were collected from 3- and 12-wk-old *Ankrd31* mutant females and fixed in Bouin's solution, embedded in paraffin, serially sectioned at 6 µm, and stained by H&E. Follicle quantification was performed as described previously [20]. Every fifth section was examined for the presence of the following classes of oocytes/follicles: primordial, primary, secondary, preantral and antral. For quantitative analyses, they were grouped into primordial and growing follicles (from primary to antral).

Meiotic spread preparation and immunostaining

Meiotic prophase I surface spreads were prepared as described previously [21]. Briefly, testes were removed and decapsulated in hypotonic sucrose extraction buffer and left on ice for 1 h. Tubules were chopped on glass depression slides in a bubble of 0.03% sucrose and added to slides coated in 1% paraformaldehyde. Slides were slow dried and washed in PBS with Photo-Flo (Kodak).

For staining, the slides were incubated for 1 h at RT with sterile filtered antibody dilution buffer (ADB) containing 10 mL normal goat serum, 3 g BSA, 50 µL Triton X-100 and 990 mL PBS, and subsequently incubated overnight with ADB diluted rabbit polyclonal H3K9me3 (Millipore Sigma; 07-442, 1:100) and mouse monoclonal anti-Sycp3 (Abcam; ab12452, 1:200) in a humidified chamber at 4 °C. After washing in ADB, the slides were incubated at 37 °C for 30 min with a 1:1000 dilution of

secondary antibody goat anti-rabbit IgG 594 (Molecular Probes, A11012) and goat anti-mouse IgG 488 (Molecular Probes, A11001), then incubated at 37 °C for 5 min with 500 ng/mL of DAPI (4′,6-diamidino-2-phenylindole) and mounted using Vectashield antifade mounting medium (Vector, H-1000).

Diakinesis chromosome spreading

This procedure was described previously [22]. Briefly, seminiferous tubules were treated by hypotonic solution (1% sodium citrate) for 30 min and minced carefully for 15 min. The cell suspension was centrifuged, and the supernatant was removed. Cells were then fixed three times in methanol: glacial acetic acid (3:1) and dropped onto slides. After drying, the cells were stained with Giemsa staining solution.

Data preprocessing

The infertility-related variants were downloaded using the ClinVar Miner web-based tool (version: 2021-5-29) [23]. Variants/mutants under terms “female infertility”, “genetic infertility”, “infertility disorder” and “male infertility” with “one-star criteria provided” were selected. The variants were organized into five catalogs (pathogenic, likely pathogenic, VUS, likely benign and benign) and only missense variants with a SNP rsID were selected. Sets of deleterious and benign missense variants that have been functionally interpreted in mouse models were collected from our experiments (http://www.infertilitygenetics.org/) and from searching the literature using PubMed, Google Scholar, bioRxiv and medRxiv (“Mouse_all” dataset, n=102 in total, Table S2), in which the infertility variants were grouped to form the “Mouse_infertility” dataset (n=34). The analysis was restricted to SNPs with rsIDs reported in the gnomAD database. The ClinVar annotation and star rating were also collected in Table S2.

Predictor calls of missense variants from the two datasets were derived using dbNSFP v4.1a [24], which is integrated into the Ensembl Variant Effect Predictor (VEP). Only the "pathogenic" and "benign" variants in ClinVar were grouped into the "ClinVar_infertility" dataset (n=697). The cutoff scores for predictors are listed in Table S3.

Evaluation of pathogenicity predictor performances

To compare the performance of the prediction tools, we applied statistical metrics derived from a confusion matrix. We identified a correctly classified variant as a true positive (TP) if, and only if, the variant corresponded to the positive class (deleterious) and as a true negative (TN) if and only if the variant corresponded to the negative class (benign). Accordingly, a false positive (FP) is a negative variant (benign) that is classified as positive (deleterious) and a false negative (FN) is a positive variant (deleterious) classified as a negative one (benign).

The Matthews Correlation Coefficient (MCC) was used to measure the correlation between the actual class of variants and the predictions made by the classifiers:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The value range of MCC is -1 to 1. A coefficient of 1 represents a perfect prediction, 0 is no better than random prediction, and −1 indicates total disagreement between prediction and observation.

Prediction precision, also known as positive predictive value (PPV), of the impact of missense variants were calculated by:

$$PPV = \frac{TP}{TP + FP}$$

The prediction accuracy (ACC) of the impact of missense variants were evaluated using the following statistical measures:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

The range of values of ACC lies between 0 and 1. A perfect tool has an ACC of 1.

We used Receiver Operating Characteristic (ROC) curves to visualize the tradeoffs between sensitivity and specificity in the binary classifiers. The ROC curve is the fraction of the TP over all positives TP+FN (sensitivity) against the fraction of the FP over all negatives TN+FP (1-specificity).

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{TN + FP - TN}{TN + FP} = \frac{FP}{TN + FP}$$

The area under the ROC curve, known as the AUC, was used to measure the performances of predictors in correctly identifying the TPs, i.e. deleterious variants, among the TPs and FPs. The AUC can take values between 0 and 1. A perfect tool has an AUC of 1 and the AUC random tool is 0.5. The prediction scores derived from dbNSFP v4.1a were used to calculate ROC in Statistical Package for the Social Sciences (SPSS) software.

Image collections and statistics

Fluorescent images were captured by an Olympus XM10 camera. Bright-field images were captured by the Olympus SC30 camera. Cropping, color, and contrast adjustments were made with Adobe Photoshop CC 2020, using identical background

adjustments for all images. All data were expressed as mean ± SEM. All statistical calculations were carried out using an unpaired Student's *t-test* or one-way analysis of variance (ANOVA) followed by Tukey's *post hoc* test or two-sided Fisher's exact test with R. Graphs were generated using R. ROC curve was generated by SPSS software (version 20.0; IBM).

Results

Selection of reproduction genes and variants evaluated in this study

The process of spermatogenesis comprises multiple tightly regulated steps of differentiation (Fig. S1). It begins with several mitotic divisions of cohorts of spermatogonial stem cells that ultimately enter meiotic prophase I. During the leptotene substage of prophase I, programmed double strand breaks (DSBs) are introduced at preferred sites throughout the genome. These DSBs are recognized by a number of DNA damage response proteins that ultimately repair the DSBs via homologous recombination, and this process drives pairing and synapsis of homologous chromosomes [25]. Prophase I is followed by two cell divisions without intervening DNA replication, leading to haploid, round spermatids that subsequently undergo a dramatic morphological transformation into spermatozoa during the differentiation process of spermiogenesis.

We evaluated variants in the following genes that are already known to be essential for fertility: *ANKRD31* (MIM: 618423), *BRDT* (MIM: 602144), *EXO1* (MIM: 606063), *FKBP6* (MIM: 604839), *DMC1* (MIM: 602721), *SEPT12* (MIM: 611562), and *MSH4*

(MIM: 602105). Fig. S1 illustrates the stages of mouse spermatogenesis during which they function, and below, a brief description of the key functions of these genes in spermatogenesis is provided. To select potentially deleterious missense variants for functional evaluation in mice, we considered allele frequency (AF), and either *in silico* prediction of pathogenicity or a genetic assay for protein interaction (Fig. 1A). We prioritized missense SNPs (present in the gnomAD database) segregating in human populations at an AF below 1%. Except for the *SEPT12* alleles, SIFT and PolyPhen2 pathogenicity prediction algorithms were used to identify candidate deleterious variants. To further narrow candidates to a manageable number, we limited variants to those that also had an AF above 0.01%, yielding 19 putatively deleterious missense SNPs in 5 genes (*FKBP6*, *EXO1*, *MSH4*, *BRDT* and *ANKRD31*) (Fig.1B).

Mutations that disrupt protein-protein interactions (PPIs) are enriched as a molecular cause of human genetic diseases [12]. To explore if PPI disruption might be an effective molecular screen for potential pathogenic infertility variants, yeast 2 hybrid (Y2H) assays were performed on *SEPT12* variants to determine if the altered amino acids disrupted known interactions with other Septins, SEPT1 and SEPT5 [26–28]. We screened 5 variants, 4 of which disrupted a PPI (Fig. 1C). One variant, p.Gly169Glu encoded by rs138628476, was reported by our groups to cause male subfertility in a mouse model [9]. The other 3 PPI-disrupting variants were added to our candidate list.

Phenotypic analysis of mouse models

Based on our variant prioritization, Y2H screening, and resource capacity, we chose 9 missense variants for functional testing by mouse modeling (Table 1). This includes 1

variant in *DMC1* that met criteria described above, except for having a lower AF than 0.01% cutoff. Using CRISPR/Cas9-mediated genome editing in zygotes, we successfully generated 8 mouse lines modeling their corresponding human amino acid variants (Fig. S2). To identify any potential reproduction defects, the animals were bred to homozygosity and tested for several fertility parameters as described previously [5–10], including testis weight, gonad histology, and sperm counts for males. For genes with known roles in meiosis, we also analyzed spermatocyte surface spread preparations of prophase I chromosomes via immunolabeling for key proteins. We also performed fertility trials of males and females in all cases. A brief description of the genes studied, and the phenotypes of mouse models, are presented below.

DMC1 (DNA Meiotic Recombinase 1) is a meiosis-specific homolog of *E. coli* RecA and is expressed in leptotene-to-zygotene spermatocytes, stages corresponding to initiation of homologous chromosome pairing. Following generation and exonucleolytic processing of meiotic DSBs to expose 3' overhangs bound by single stranded DNA binding protein RPA, DMC1 and a related recombinase RAD51 catalyze strand exchange, promoting homolog recognition and pairing. Male and female mouse mutants null for *Dmc1* are sterile as a consequence of failed DSB repair and interhomolog synapsis, both of which trigger prophase I arrest and checkpoint-mediated elimination of meiocytes [29,30]. We generated mice modeling the missense SNP rs189722264 (g.8500G>T; p.Gly50Cys) (Fig. S2A), which was predicted to be highly deleterious by both SIFT and PolyPhen2 (Table 1). In contrast to mice bearing a frameshift (presumably null) allele generated in the same editing experiment, which were azoospermic as expected, *Dmc1*$^{G50C/G50C}$ female and male mice had normal fecundity

(Fig. 2A). Mutant males had testis sizes and sperm counts similar to WT sibs (Fig. 2B and C), and histology revealed normal spermatogenesis, unlike the maturation arrest observed in *Dmc1*$^{-/-}$ mice (Fig. 2D).

FKBP6 (FK506 binding protein 6) encodes a cis-trans peptidyl-prolyl isomerase, in which resides the amino acid altered by rs3750075 (g.6245C>A; p.Phe72Leu) (Fig. S3). This gene functions in immunoregulation and basic cellular processes involving protein folding and trafficking. Expressed in several tissues [31], FKBP6 localizes to meiotic chromosome cores and regions of homologous chromosome synapsis [32]. Despite its expression in both testis and ovary, its deficiency only causes male-specific infertility [32,33]. *FKBP6* variants have been associated with non-obstructive azoospermia and idiopathic infertility [34,35]. Fertility tests of *Fkbp6*^{F72L} homozygotes revealed normal fecundity (Fig. 2A), testis weights, sperm counts, and testis histology (Fig. 2B-D).

EXO1 (Exonuclease 1) is involved in repair of DNA mismatches and DSBs. Loss of Exo1 causes cancer predisposition and infertility in mice [36], the latter related to premature meiotic homolog separation before alignment at the metaphase plate [37]. The SNP rs4149964 changes an AA residue (g.28941C>A; p.Ser465Tyr) in the MLH1 interaction domain (Fig. S3). *Exo1*$^{S465Y/S465Y}$ males and females showed normal fecundity (Fig. 2A), and all male parameters were normal (Fig. 2B-D).

MSH4 (MutS protein homolog 4) is a mismatch repair protein essential for meiotic recombination. Male and female mice lacking *Msh4* are infertile due to meiotic failure [38]. We modeled rs141042002 (g.56405C>T; p.Thr410Met), which is located in the DNA-binding domain (Fig. S3). The *Msh4*$^{T410M/T410M}$ mutants did not reveal any reproductive phenotypes in males or females (Fig. 2A-D).

BRDT (bromodomain testis associated) belongs to the bromodomain-extra terminal (BET) family of proteins. It is expressed in pachytene spermatocytes (meiocytes in which chromosomes are fully synapsed) and round spermatids. It contains two conserved bromodomains (BD1 and 2) involved in the recognition of acetylated histones, and an extra terminal (ET) motif region involved in PPIs [39–41]. Mice bearing a deletion of BD1 have oligoasthenoteratozoospermia [42]. A null allele is more severe, causing complete meiotic arrest [43]. As an epigenetic regulator, BRDT is required for proper meiotic chromatin organization, meiotic sex chromosome inactivation (MSCI), and normal crossover (CO) formation [44].

We modeled the human *BRDT* p.Asp550His allele, encoded by SNP rs141699970, which resides in the ET domain (Figs. S2E and S3). A mouse line with a 62 nt deletion was also recovered and expanded to serve as a knockout allele (*Brdt*$^{-}$) for comparison (Fig. S4). As previously reported [43], *Brdt*$^{-/-}$ males were infertile, however, *Brdt*$^{D550H/D550H}$ males sired litters of sizes comparable to those of heterozygous littermates when mated to WT females (Fig. 3A), and also had normal testis weights, sperm numbers and testis histology, in contrast to the null (Fig. 3B-D). Both *Brdt*$^{-/-}$ and *Brdt*$^{D550H/D550H}$ females were fertile (Fig. 3A), which was expected since *Brdt* is expressed exclusively in the testis [45]. TUNEL staining revealed more apoptotic cells in the seminiferous tubules of *Brdt*$^{D550H/D550H}$ mice than controls but much less than in *Brdt*$^{-/-}$ mice (Fig. 3E and F). Depletion of BRDT results in weaker H3K9me3 signal in the pseudoautosomal region (PAR) of the sex chromosomes in diplonema [44]. We confirmed this in *Brdt*$^{-/-}$ spermatocytes but did not observe a difference between *Brdt*$^{D550H/D550H}$ and control spermatocytes (Fig. 3G). However, *Brdt*$^{D550H/-}$ mice exhibited slightly reduced testis

weights, presence of metaphase I arrested spermatocytes, and elevated apoptotic spermatogenic cells despite having comparable sperm counts and normal H3K9me3 modification in the PAR (Fig. 3B-G). Therefore, we classify *BRDT*^{D550H} as benign for fertility, although it may be a weak hypomorph.

ANKRD31 (Ankyrin repeat domain 31) contains two separated triplets of Ankyrin repeats, and three conserved domains: a predicted coiled-coil domain and two regions without functional predictions [46]. Male mice lacking *Ankrd31* are sterile due to delayed recombination initiation, altered DSB distribution, and failed recombination in the PAR of sex chromosomes, where it is needed for DSB formation [46–48]. *Ankrd31* null female mice are fertile but exhibit premature ovarian insufficiency (POI) [46,47]. We modeled human rs150791065 (g.72881A>G; p.Thr557Ala; Fig. S2F), which is located within a ANK1 domain (Figs. S4). Fertility test demonstrated that *Ankrd31*$^{T557A/T557A}$ mice of both sexes had normal fecundity (Fig. 4A), and homozygous males also had normal testes and sperm counts (Fig. 4B-D). Nevertheless, since *Ankrd31* is essential for recombination between sex chromosomes [46], we next examined metaphase chromosomes for this possible defect in the mutants. Chiasmata were present between each pair of homologs and between sex chromosomes in the *Ankrd31*$^{T557A/T557A}$ spermatocytes (Fig. 4E). Finally, because null females are fertile despite having a reduced primordial follicle pool [46,47], we quantified follicle numbers in 3- and 12-wk-old mutant ovaries but found no difference compared to controls (Fig. 4F and G).

SEPT12 belongs to the Septin GTP-binding protein family and is exclusively expressed in testis. It localizes around the manchette and the neck region of elongated spermatids, as well as the annulus of mature sperm. Missense mutations in *SEPT12*

have been implicated as causing sperm defects and infertility in men [49]. The phenotype of mice lacking *Sept12* is not clear. The first reported attempt to create mouse mutants did not obtain germline transmission, but most chimeras bearing a null allele proved to be infertile, showing a variety of testicular phenotypes [50]. It is uncertain whether these phenotypes were related to haploinsufficiency for *Sept12*, or unrelated defects in the embryonic stem cells in which the targeting was performed. However, mice bearing a phosphomimetic allele caused subfertility associated with decreased sperm count and sperm motility [50].

The humanized alleles described thus far were selected through algorithmic predictions classifying them as likely pathogenic with respect to protein function. Because none had any effect on fertility, we attempted to increase the likelihood that selected alleles would be consequential by first determining if they disrupted a known PPI. Accordingly, 4 alleles of *SEPT12* were selected with AF < 1% and that also disrupted at least 1 PPI. We previously performed mouse modeling of one such allele ($SEPT12^{G169E/G169E}$) that indeed caused subfertility and poor motility [9]. We failed to generate a mouse model for another (Table 1), but did generate and characterize two mouse alleles, $Sept12^{V160M}$ and $Sept12^{V220I}$, corresponding to human rs144420035 (g.9563G>A) and rs142721632 (g.9998G>A), respectively. Both mutations are located within a GTPase domain (Fig. S3). Males homozygous for both alleles were fertile with normal fecundity (Fig. 5A), testis weights, sperm counts, and testis histology (Fig. 5B-D). Thus, only 1 of 3 with a PPI disruption proved to be pathogenic.

Performance of pathogenicity prediction algorithms in identifying disease-causing variants

To explore why such a high fraction of predicted deleterious variants failed to cause the reproductive phenotypes in mice, we evaluated the effectiveness of 10 different pathogenicity prediction algorithms in classifying 26 human alleles (including those presented here) that our group has modeled in mice, corresponding to 15 essential fertility genes [5–10]. We use the terms 'deleterious' and 'benign' to refer to variants that are or are not associated with disease, respectively.

Unsurprisingly, the predictors performed differently, as they use different underlying criteria. 24 of the 26 alleles were classified as deleterious by both SIFT and PolyPhen, but only 12 caused a phenotype in mice. The REVEL (Rare Exome Variant Ensemble Learner) [51] and CADD (Combined Annotation Dependent Depletion; a cutoff score of >25 was used) [52] algorithms performed best (Fig. 6A). In contrast to the low accuracy of computational predictions alone in predicting phenotypes, *in vitro* assays of variants that revealed protein defects (instability, altered function, or PPI disruptions) correctly predicted mouse phenotypes in 10/14 cases. Of the 12 alleles classified as deleterious by both PolyPhen and SIFT and for which no *in vitro* data was available, only 2 caused a mouse phenotype (Figs. 6A and S5).

The disparity between the predicted and observed effects of these missense variants led us to further explore the better-performing algorithms. We built a dataset (“Mouse_all”) comprising 102 human missense variants that have been modeled in mice, and which reside in genes involved in reproduction or other processes (Materials and Methods; Fig. 6B; Table S2). This group was subdivided in two datasets

(Mouse_all_D and Mouse_all_B, where D=deleterious and B=benign) based on whether the mouse models did or did not show a pathogenic phenotype, respectively (Fig. S6A). For a “control” dataset, we retrieved infertility-related missense variants from the ClinVar database that had been pre-categorized into five categories: ClinVar_B (benign), ClinVar_LB (likely benign), ClinVar_VUS, ClinVar_LP (likely pathogenic), and ClinVar_P (pathogenic) (Fig. S6A). For each of these datasets, we generated prediction scores from 10 algorithms computed by dbNSFP (database for nonsynonymous SNP functional predictions) [24]. The prediction score distribution patterns of SIFT, DANN, GERP++ RS and MutationTaster were similar for all categories (Fig. S6B), indicating that these tools failed to discriminate the variants functionally. Next, we focused on the ClinVar_B and ClinVar_P datasets (n=697 combined, collectively referred to as “ClinVar_infertility”; Fig. S6A), representing the extremes of ClinVar classifications. Whereas the predictions of all the tools corresponded well with the pathogenic (ClinVar_P) dataset, only REVEL and CADD had relatively low FP (type I error) rates (7.95% and 8.3%, respectively) for the benign (ClinVar_B) variants. MutationTaster performed the worst (FP rate of 94.14%; Fig. S6C). Surprisingly, all of the prediction tools had higher FP rates (39.29% to 89.29%) for the Mouse_all_B group than those in ClinVar_B (infertility) group (Fig. S6C), consistent with a previous study showing that prediction tools have high FP values [53].

We next compared the performance of pathogenicity predictors for the Clinvar_infertility and Mouse_all datasets in terms of prediction accuracy (ACC), positive predictive value (PPV), Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) of receiver operating characteristic (ROC). Consistent with previous comparisons using ClinVar variants [54], REVEL was superior to other predictors

for the ClinVar_infertility dataset (ACC=0.92, PPV=0.88, MCC=0.83 and AUC=0.979; Fig. 6C-F). However, all the predictors performed worse for the Mouse_all vs the ClinVar_infertility dataset (Fig. 6C-F). The MCC scores for the Mouse_all dataset approached 0, indicating that the mouse phenotype and *in silico* interpretations are uncorrelated (Fig. 6E). We next focused on a "Mouse_infertility" data subset containing 34 functionally tested variants from the Mouse_all dataset (Fig. 6B). All the predictors performed similarly between Mouse_all and Mouse_infertility datasets (Fig. 6C-F).

To study why the predictors' performances differ between the ClinVar_infertility (mostly *in silico* predicted variants) and Mouse_all datasets (*in vivo* tested variants), we analyzed correlations between variant classifications. We retrieved ClinVar annotations of the Mouse_all dataset and found that 65/102 variants (63.7%) were reported in ClinVar but only 30.4% (31/102) were classified concordantly (Fig. 7). This included 5 of 28 Mouse_all benign variants being classified as "Pathogenic/Likely Pathogenic" in ClinVar, and 12 deleterious variants being classified as "Benign/Likely Benign" (Fig. 7). With the caveat that variants may have different effects in human vs mouse, these results raise serious questions about clinical usage of ClinVar or *in silico* based classifications alone, and argues that *in vivo* functionally-interpreted variants are important for accurate classifications of the phenotypic consequences of genetic variants.

Discussion

WGS and WES are quickly becoming mainstream diagnostic tools for identifying genomic causes of rare diseases and cancers [55,56]. Nevertheless, wider adoption of sequencing as a diagnostic tool for other [potentially genetic] diseases is being limited by our ability to accurately classify and interpret the consequences of genetic variants. Many *in silico* prediction tools were developed to filter and prioritize variants as being potentially causative for disease on the basis of predicted impact on a protein's function, but actual phenotype correlations are difficult to predict. Complicating factors include heterogeneity of molecular mechanisms underlying diseases, lack of a deep mechanistic understanding of most genes and proteins, and complexity of physiological systems. Here, we found that *in silico* prediction tools had high FP rates when used to predict whether a human missense variant causes a phenotypic outcome in mouse models. This shortcoming underscores the need to combine computational predictions with functional evidence from appropriate cellular or animal models in order to be useful for genetic diagnosis of disease origins.

In clinical settings, *in silico* predictions represent one evidentiary criterion for variant interpretation by the ACMG/AMP. However, exclusive dependence upon *in silico* tools can cause erroneous interpretation of variants. For example, SPATA16 p.Arg283Gln [57] and MEIOB p.Asn64Ile [58] were reported as causative variants of human globozoospermia and azoospermia, respectively, but were shown not to cause reproductive phenotypes in mice, even though the latter variant caused some meiotic defects [59,60]. Besides *in silico* predictors, biochemical and *in vitro* approaches are commonly used to evaluate the impact of variants on protein function. While these

approaches help to prioritize or classify variants with greater confidence when used in conjunction with *in silico* predictions, the impact of such functional variants on reproductive organs or phenotype is lacking and unacceptable for use in the clinical setting as a sole diagnostic. The *DMC1* homozygous variant p.Met200Val was reported to cause POI in an African woman [61], and a subsequent study reported that this allele impaired biochemical function [62]. However, we found that the orthologous allele of this highly conserved gene did not impair mouse fertility [7]. Here, we report that two alleles of *SEPT12* did not markedly affect mouse reproduction despite disrupting interaction with other Septins in the Y2H assay. Our data emphasizes the importance of animal models in the functional validation of genetic variants involved in human infertility, and highlights challenges behind truly ascertaining the physiological effects of VUS.

We concluded that the *Brdt*^{D550H} variant is non-pathogenic since mouse homozygotes had normal testis weight, sperm number and fertility despite the presence of increased apoptotic cells in the seminiferous tubules. Interestingly, we found that whereas mice bearing this allele in *trans* to a null (*Brdt*$^{D550H/-}$) had sperm counts comparable to controls, they exhibited a subtle increase in apoptotic and meiotic metaphase I arrested spermatocytes. The mice we used were not entirely inbred, so it is possible that heterosis suppressed the phenotype. Nevertheless, there was no inconsistency amongst replicates in terms of fertility. These results suggest that functional prediction algorithms may correctly identify missense alleles that impact the protein in some subtle way that does not result in a pathogenic phenotype.

As part of a larger effort to identify true infertility variants segregating in people, we have used pathogenicity prediction algorithms as a screen for prioritizing missense

variants. However, the work summarized here reveals the shortcomings of current predictors despite being claimed to have an accuracy ranging from ~65% to 90% [63]. Given that a variety of *in silico* algorithms/pipelines are employed by clinical labs for prioritizing potential disease-causing variants [64], it is important to identify high-confidence classifiers to minimize both false-positive and false-negative prediction rates. Here, we attempted to address this issue by evaluating the performance of predictors in classifying infertility-related missense variants deposited in ClinVar. The tools varied with respect to agreement with ClinVar annotations, presumably reflecting differences in feature sets and scoring methods employed by these tools. Usually, ensemble prediction tools or metapredictors (e.g. CADD and REVEL), which generate their predictions based on the output (scores) of other tools, are purported to have higher classification accuracies than individual tools (e.g. SIFT and PolyPhen2).

To estimate the performance of a predictor, commonly used parameters include ACC, Matthews Correlation Coefficient (MCC), sensitivity, specificity, negative predictive value, PPV, and AUC. Given that all have flaws, we took an alternative approach of considering multiple success rate descriptors simultaneously. Here, we compared variant classification performance using four metrics, MCC, ACC, PPV and AUC. Consistent with a previous study [54], the consensus results from each parameter demonstrated that REVEL outperformed other classifiers in the ClinVar_infertility dataset. However, it is important to recognize that ClinVar variants are sometimes used to train algorithms either directly (such as MutationTaster) or indirectly (e.g. REVEL, which was trained by HGMD and shares greater than half of ClinVar variants/mutants [65]), and conversely, pathogenicity of some variants entered into ClinVar are classified in

part by various prediction algorithms. Such circularities in testing and training data may compromise calls [66]. Hence, *in vivo* tested missense variants, which are classified according to whether the animal model has a disease-causing phenotype or not, most likely constitute a more clinically accurate and relevant (albeit small) testing set.

Apart from these issues, there are other potential explanations for the poor performance of predictors in classifying alleles in the Mouse_all dataset compared to those in the ClinVar_infertility dataset. One is that a significant proportion of ClinVar classifications, which are often used for algorithm training, are inaccurate. Based on data from chemical mutagenesis studies in mice, Miosge *et al*. demonstrated that for *de novo* or rare missense mutants that were algorithmically predicted to be deleterious, nearly half were in fact neutral or nearly neutral by both phenotypic and *in vitro* assays [53]. Similarly, we found that all the predictors yielded high FP values (type I error). A study of human variants ClinVar database provided evidence for extensive misclassifications, suggested that substantial numbers of misclassifications could be corrected by considering allele frequency, and that input from orthogonal functional and genetic studies are crucial for improving variant classification accuracy [67]. A second potential explanation is that there is a substantial disconnect between biochemical and pathogenic effects of variants. That is, living systems have robustness or redundancies that largely mask minor biochemical or structural defects of proteins, although subtle molecular defects can ultimately lead to allele frequency decline from purifying selection [53]. A third possible explanation is that the power of predictors was not sufficiently evaluated by the modest sample size of the Mouse_all dataset. With the development of genome editing approaches, knock-in/humanized animal models can be applied to

interpret more missense variants and the results can be used to evaluate and improve the prediction tools. Finally, despite the evolutionary conservation of all the amino acids that were investigated here, it is possible that mice are more robust to alterations than humans.

Different levels of manual curation are applied in databases. ClinVar uses a three-star rating system to represent the “Review Status” of each submission. These are: "single submitter - criteria provided" (one star); "expert panel" (three stars); and "practice guidelines" (four stars). To assist users in discerning the exploring the data and evidence for variant interpretations, several ClinVar derivative platforms have been developed. We used ClinVar Miner to identify variants of interest for evaluating the performance of pathogenicity predictors, and selected one-star rating variants because there are no infertility-related variants with three- or four-star ratings (no Variant Curation Expert Panels exists for infertility, precluding any variants from having these ratings). The formation of an infertility Variant Curation Expert Panel for ClinGen (https://www.clinicalgenome.org) is sorely needed to critically address the accuracy of pathogenicity calls currently present in ClinVar.

In summary, we provide *in vivo* evidence of 8 human variants in 7 essential reproduction genes that proved to be benign in mice, despite being predicted to have negative impacts upon protein function by *in silico* pathogenicity prediction algorithms or the Y2H assay. We conclude that these variants are likely to be clinically benign, with the caveats that mice may be more tolerant to the protein alteration, and that these alleles may contribute to phenotypic defecst when combined with variations in other genes. We compared 10 predictors and found that REVEL outperformed the others for

evaluating infertility-related variants in ClinVar. However, all the 10 predictors performed worse in classifying variants that were functionally tested in mouse models compared to those classified by ClinVar, possible reflecting a circularity in the classification process. The results underscore the importance of animal models in the functional validation of genetic variants involved in human infertility, and the need for more rigorous and evaluations of VUS.

Supplemental information description

Supplemental data include 6 figures and 3 tables.

Acknowledgments. The authors would like to thank R. Munroe and C. Abratte of Cornell's transgenic facility for generating the mice, as well as D. Conrad and S. Wierbowski for discussions. This work is supported by grants from the National Institutes of Health (R01 HD082568 and P50 HD096723) and contract CO29155 from the NY State Stem Cell Program (NYSTEM). X.D. was supported by a postdoctoral fellowship from the Empire State Stem Cell Fund through New York State Department of Health contract C30293GG.

Declaration of interests

The authors declare no competing interests.

Web resources

Clinvar Miner: https://clinvarminer.genetics.utah.edu

Ensembl VEP: https://useast.ensembl.org/Homo_sapiens/Tools/VEP

gnomAD: https://gnomad.broadinstitute.org

References

1. Mahmood, K., Jung, C.-H., Philip, G., Georgeson, P., Chung, J., Pope, B.J., and Park, D.J. (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum. Genomics *11*, 10.
2. Schimenti, J.C., and Handel, M.A. (2018). Unpackaging the genetics of mammalian fertility: strategies to identify the “reproductive genome”. Biol. Reprod. *99*, 1119–1128.
3. Aston, K.I. (2014). Genetic susceptibility to male infertility: news from genome-wide association studies. Andrology *2*, 315–321.
4. Ding, X., and Schimenti, J.C. (2021). Strategies to identify genetic variants causing infertility. Trends Mol. Med. *27*, 792–806.
5. Singh, P., and Schimenti, J.C. (2015). The genetics of human infertility by functional interrogation of SNPs in mice. Proc. Natl. Acad. Sci. USA *112*, 10431–10436.
6. Tran, T.N., and Schimenti, J.C. (2019). A segregating human allele of SPO11 modeled in mice disrupts timing and amounts of meiotic recombination, causing oligospermia and a decreased ovarian reserve†. Biol. Reprod. *101*, 347–359.
7. Tran, T.N., and Schimenti, J.C. (2018). A putative human infertility allele of the meiotic recombinase DMC1 does not affect fertility in mice. Hum. Mol. Genet. *27*, 3911–3918.
8. Tran, T.N., Martinez, J., and Schimenti, J.C. (2019). A predicted deleterious allele of the essential meiosis gene MND1, present in ~3% of East Asians, does not disrupt reproduction in mice. Mol. Hum. Reprod. *25*, 668–673.

9. Fragoza, R., Das, J., Wierbowski, S.D., Liang, J., Tran, T.N., Liang, S., Beltran, J.F., Rivera-Erick, C.A., Ye, K., Wang, T.-Y., et al. (2019). Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat. Commun. *10*, 4141.

10. Ding, X., Fragoza, R., Singh, P., Zhang, S., Yu, H., and Schimenti, J.C. (2020). Variants in RABL2A causing male infertility and ciliopathy. Hum. Mol. Genet. *29*, 3402–3411.

11. Singh, P., Fragoza, R., Blengini, C.S., Tran, T.N., Pannafino, G., Al-Sweel, N., Schimenti, K.J., Schindler, K., Alani, E.A., Yu, H., et al. (2021). Human MLH1/3 variants causing aneuploidy, pregnancy loss, and premature reproductive aging. Nat. Commun. *12*, 5005.

12. Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F.M., Lee, H.R., Wang, X., Mort, M., Stenson, P.D., Cooper, D.N., et al. (2014). A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet. *10*, e1004819.

13. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

14. Landrum, M.J., Chitipiralla, S., Brown, G.R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. Nucleic Acids Res. *48*, D835–D844.

15. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

16. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

17. Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., and Yao, X. (2009). DOG 1.0: illustrator of protein domain structures. Cell Res. *19*, 271–273.

18. Truett, G.E., Heeger, P., Mynatt, R.L., Truett, A.A., Walker, J.A., and Warman, M.L. (2000). Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). BioTechniques *29*, 52–54.

19. Wang, Y. (2003). Epididymal sperm count. Curr. Protoc. Toxicol. *Chapter 16*, Unit16.6.

20. Bolcun-Filas, E., Rinaldi, V.D., White, M.E., and Schimenti, J.C. (2014). Reversal of female infertility by Chk2 ablation reveals the oocyte DNA damage checkpoint pathway. Science *343*, 533–536.

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

22. Uroz, L., Rajmil, O., and Templado, C. (2008). Premature separation of sister chromatids in human male meiosis. Hum. Reprod. *23*, 982–987.

23. Henrie, A., Hemphill, S.E., Ruiz-Schultz, N., Cushman, B., DiStefano, M.T., Azzariti, D., Harrison, S.M., Rehm, H.L., and Eilbeck, K. (2018). ClinVar Miner: Demonstrating

utility of a Web-based tool for viewing and filtering ClinVar data. Hum. Mutat. *39*, 1051–1060.

24. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. *12*, 103.

25. Gray, S., and Cohen, P.E. (2016). Control of meiotic crossovers: from double-strand break formation to designation. Annu. Rev. Genet. *50*, 175–210.

26. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

27. Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., et al. (2011). Next-generation sequencing to generate interactome datasets. Nat. Methods *8*, 478–480.

28. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The bioplex network: A systematic exploration of the human interactome. Cell *162*, 425–440.

29. Pittman, D.L., Cobb, J., Schimenti, K.J., Wilson, L.A., Cooper, D.M., Brignull, E., Handel, M.A., and Schimenti, J.C. (1998). Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for Dmc1, a germline-specific RecA homolog. Mol. Cell *1*, 697–705.

30. Yoshida, K., Kondoh, G., Matsuda, Y., Habu, T., Nishimune, Y., and Morita, T. (1998). The mouse RecA-like gene Dmc1 is required for homologous chromosome synapsis during meiosis. Mol. Cell *1*, 707–718.

31. Meng, X., Lu, X., Morris, C.A., and Keating, M.T. (1998). A novel human gene FKBP6 is deleted in Williams syndrome. Genomics *52*, 130–137.

32. Crackower, M.A., Kolas, N.K., Noguchi, J., Sarao, R., Kikuchi, K., Kaneko, H., Kobayashi, E., Kawai, Y., Kozieradzki, I., Landers, R., et al. (2003). Essential role of Fkbp6 in male fertility and homologous chromosome pairing in meiosis. Science *300*, 1291–1295.

33. Shah, K., Sivapalan, G., Gibbons, N., Tempest, H., and Griffin, D.K. (2003). The genetic basis of infertility. Reproduction *126*, 13–25.

34. Zhang, W., Zhang, S., Xiao, C., Yang, Y., and Zhoucun, A. (2007). Mutation screening of the FKBP6 gene and its association study with spermatogenic impairment in idiopathic infertile men. Reproduction *133*, 511–516.

35. Miyamato, T., Sato, H., Yogev, L., Kleiman, S., Namiki, M., Koh, E., Sakugawa, N., Hayashi, H., Ishikawa, M., Lamb, D.J., et al. (2006). Is a genetic defect in Fkbp6 a common cause of azoospermia in humans? Cell Mol Biol Lett *11*, 557–569.

36. Wei, K., Clark, A.B., Wong, E., Kane, M.F., Mazur, D.J., Parris, T., Kolas, N.K., Russell, R., Hou, H., Kneitz, B., et al. (2003). Inactivation of Exonuclease 1 in mice results in DNA mismatch repair defects, increased cancer susceptibility, and male and female sterility. Genes Dev. *17*, 603–614.

37. Schaetzlein, S., Chahwan, R., Avdievich, E., Roa, S., Wei, K., Eoff, R.L., Sellers, R.S., Clark, A.B., Kunkel, T.A., Scharff, M.D., et al. (2013). Mammalian Exo1 encodes both structural and catalytic functions that play distinct roles in essential biological processes. Proc. Natl. Acad. Sci. USA *110*, E2470-9.

38. Kneitz, B., Cohen, P.E., Avdievich, E., Zhu, L., Kane, M.F., Hou, H., Kolodner, R.D.,

Kucherlapati, R., Pollard, J.W., and Edelmann, W. (2000). MutS homolog 4 localization to meiotic chromosomes is required for chromosome pairing during meiosis in male and female mice. Genes Dev. *14*, 1085–1097.

39. Pivot-Pajot, C., Caron, C., Govin, J., Vion, A., Rousseaux, S., and Khochbin, S. (2003). Acetylation-dependent chromatin reorganization by BRDT, a testis-specific bromodomain-containing protein. Mol. Cell. Biol. *23*, 5354–5365.

40. Berkovits, B.D., and Wolgemuth, D.J. (2013). The role of the double bromodomain-containing BET genes during mammalian spermatogenesis. Curr Top Dev Biol *102*, 293–326.

41. Shi, J., and Vakoc, C.R. (2014). The mechanisms behind the therapeutic activity of BET bromodomain inhibition. Mol. Cell *54*, 728–736.

42. Shang, E., Nickerson, H.D., Wen, D., Wang, X., and Wolgemuth, D.J. (2007). The first bromodomain of Brdt, a testis-specific member of the BET sub-family of double-bromodomain-containing proteins, is essential for male germ cell differentiation. Development *134*, 3507–3515.

43. Gaucher, J., Boussouar, F., Montellier, E., Curtet, S., Buchou, T., Bertrand, S., Hery, P., Jounier, S., Depaux, A., Vitte, A.-L., et al. (2012). Bromodomain-dependent stage-specific male genome programming by Brdt. EMBO J. *31*, 3809–3820.

44. Manterola, M., Brown, T.M., Oh, M.Y., Garyn, C., Gonzalez, B.J., and Wolgemuth, D.J. (2018). BRDT is an essential epigenetic regulator for proper chromatin organization, silencing of sex chromosomes and crossover formation in male meiosis. PLoS Genet. *14*, e1007209.

45. Jones, M.H., Numata, M., and Shimane, M. (1997). Identification and

characterization of BRDT: A testis-specific gene related to the bromodomain genes RING3 and Drosophila fsh. Genomics *45*, 529–534.

46. Papanikos, F., Clément, J.A.J., Testa, E., Ravindranathan, R., Grey, C., Dereli, I., Bondarieva, A., Valerio-Cabrera, S., Stanzione, M., Schleiffer, A., et al. (2019). Mouse ANKRD31 Regulates Spatiotemporal Patterning of Meiotic Recombination Initiation and Ensures Recombination between X and Y Sex Chromosomes. Mol. Cell *74*, 1069–1085.e11.

47. Boekhout, M., Karasu, M.E., Wang, J., Acquaviva, L., Pratto, F., Brick, K., Eng, D.Y., Xu, J., Camerini-Otero, R.D., Patel, D.J., et al. (2019). REC114 partner ANKRD31 controls number, timing, and location of meiotic DNA breaks. Mol. Cell *74*, 1053–1068.e8.

48. Acquaviva, L., Boekhout, M., Karasu, M.E., Brick, K., Pratto, F., Li, T., van Overbeek, M., Kauppi, L., Camerini-Otero, R.D., Jasin, M., et al. (2020). Ensuring meiotic DNA break formation in the mouse pseudoautosomal region. Nature *582*, 426–431.

49. Kuo, Y.-C., Lin, Y.-H., Chen, H.-I., Wang, Y.-Y., Chiou, Y.-W., Lin, H.-H., Pan, H.-A., Wu, C.-M., Su, S.-M., Hsu, C.-C., et al. (2012). SEPT12 mutations cause male infertility with defective sperm annulus. Hum. Mutat. *33*, 710–719.

50. Lin, Y.-H., Lin, Y.-M., Wang, Y.-Y., Yu, I.-S., Lin, Y.-W., Wang, Y.-H., Wu, C.-M., Pan, H.-A., Chao, S.-C., Yen, P.H., et al. (2009). The expression level of septin12 is critical for spermiogenesis. Am. J. Pathol. *174*, 1857–1868.

51. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am. J. Hum. Genet.

99, 877–885.

52. Rentzsch, P., Schubach, M., Shendure, J., and Kircher, M. (2021). CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. *13*, 31.

53. Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S., et al. (2015). Comparison of predicted and actual consequences of missense mutations. Proc. Natl. Acad. Sci. USA *112*, E5189-5198.

54. Ghosh, R., Oak, N., and Plon, S.E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. *18*, 225.

55. Liu, Z., Zhu, L., Roberts, R., and Tong, W. (2019). Toward Clinical Implementation of Next-Generation Sequencing-Based Genetic Testing in Rare Diseases: Where Are We? Trends Genet. *35*, 852–867.

56. Nakagawa, H., and Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. Cancer Sci. *109*, 513–522.

57. Dam, A.H.D.M., Koscinski, I., Kremer, J.A.M., Moutou, C., Jaeger, A.-S., Oudakker, A.R., Tournaye, H., Charlet, N., Lagier-Tourenne, C., van Bokhoven, H., et al. (2007). Homozygous mutation in SPATA16 is associated with male infertility in human globozoospermia. Am. J. Hum. Genet. *81*, 813–820.

58. Gershoni, M., Hauser, R., Yogev, L., Lehavi, O., Azem, F., Yavetz, H., Pietrokovski, S., and Kleiman, S.E. (2017). A familial study of azoospermic men identifies three novel causative mutations in three new human azoospermia genes. Genet. Med. *19*, 998–1006.

59. Fujihara, Y., Oji, A., Larasati, T., Kojima-Kita, K., and Ikawa, M. (2017). Human Globozoospermia-Related Gene Spata16 Is Required for Sperm Formation Revealed by CRISPR/Cas9-Mediated Mouse Models. Int. J. Mol. Sci. *18*,.

60. Guo, R., Xu, Y., Leu, N.A., Zhang, L., Fuchs, S.Y., Ye, L., and Wang, P.J. (2020). The ssDNA-binding protein MEIOB acts as a dosage-sensitive regulator of meiotic recombination. Nucleic Acids Res.

61. Mandon-Pépin, B., Touraine, P., Kuttenn, F., Derbois, C., Rouxel, A., Matsuda, F., Nicolas, A., Cotinot, C., and Fellous, M. (2008). Genetic investigation of four meiotic genes in women with premature ovarian failure. Eur. J. Endocrinol. *158*, 107–115.

62. Hikiba, J., Hirota, K., Kagawa, W., Ikawa, S., Kinebuchi, T., Sakane, I., Takizawa, Y., Yokoyama, S., Mandon-Pépin, B., Nicolas, A., et al. (2008). Structural and functional analyses of the DMC1-M200V polymorphism found in the human population. Nucleic Acids Res. *36*, 4181–4190.

63. Özkan, S., Padilla, N., Moles-Fernández, A., Diez, O., Gutiérrez-Enríquez, S., and Cruz, X. de la (2021). The computational approach to variant interpretation. In Clinical DNA Variant Interpretation, (Elsevier), pp. 89–119.

64. Kobren, S.N., Baldridge, D., Velinder, M., Krier, J.B., LeBlanc, K., Esteves, C., Pusey, B.N., Züchner, S., Blue, E., Lee, H., et al. (2021). Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. Genet. Med. *23*, 1075–1085.

65. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research,

genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.

66. Grimm, D.G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D.G., Samocha, K.E., Cooper, D.N., Stenson, P.D., Daly, M.J., Smoller, J.W., et al. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum. Mutat. *36*, 513–523.

67. Shah, N., Hou, Y.-C.C., Yu, H.-C., Sainger, R., Caskey, C.T., Venter, J.C., and Telenti, A. (2018). Identification of misclassified clinvar variants via disease population prevalence. Am. J. Hum. Genet. *102*, 609–619.

Figure legends

Figure 1. Selection of causal variants from 7 reproduction genes. A) Overview of the pipeline used for filtering human SNP databases. For *DMC1*, rare variants with MAF (minor allele frequency) less than 0.01% were applied for *in silico* predictions, and rs189722264 was selected for mouse modeling. For *SEPT12*, rare variants (0.01%<MAF<1%) were applied for Y2H screening before mouse modeling. B) Distribution of benign and deleterious nsSNPs (0.01%<MAF<1%) predicted by SIFT and PolyPhen2 in indicated reproduction genes. C) Y2H screening of nsSNPs in SEPT12 disrupting its interaction with the SEPT1 and SEPT5 proteins, respectively. N.A., not analyzed.

Figure 2. Phenotypic characterization of homozygous mice harboring predicted deleterious nsSNPs in *DMC1*, *MSH4*, *EXO1* and *FKBP6*. A) Litter sizes from mating of $Dmc1^{G50C/G50C}$ (GC/GC), $Fkbp6^{F72L/F72L}$ (FL/FL), $Exo1^{S465Y/S465Y}$ (SY/SY) and $Msh4^{T410M/T410M}$ (TM/TM) females and males to WT partners. N=2 for GC/GC, FL/FL and SY/SY male and female mice, N=3 for TM/TM male and female mice. B) Testis weights of 2-month old mice. C) Sperm counts in mutant homozygotes and littermate controls. D) Cross-sections of testes showing comparable histology to the littermate controls. Scale bars = 100 μm. Data in A, B and C are represented as the mean ± SEM and were analyzed using two-tailed unpaired *t test*.

Figure 3. Phenotypic analysis of *Brdt*^{D550H} mice. A) Litter sizes from mating of *Brdt*$^{D550H/D550H}$ (DH/DH), *Brdt*$^{D550H/+}$ (DH/+), *Brdt*$^{+/-}$ (+/-) and *Brdt*$^{-/-}$ (-/-) animals to WT partners. N=2 for DH/DH male and female mice, N=3 for -/- male and female mice. B) Testis weights of 2-month-old males. C) Epididymal sperm counts (Methods). D) Histological sections of 2-month-old testes and cauda epididymides. Round germ cells were present in -/- epididymides. Scale bars = 100 µm. Black arrowheads in inset indicate meiotic metaphase I arrested cells. E) Quantification of TUNEL+ cells in each tubule. F) Number of tubules containing > 4 TUNEL+ cells. No control (DH/+) tubules section had >4. N=3 mice. G) Immunolocalization of H3K9me3 and SYCP3 on meiotic prophase I chromosomes in indicated spermatocytes stages. Dashed boxes highlight XY bodies. Data in A, B and C are represented as the mean ± SEM and were analyzed using one-way analysis of variance (ANOVA) with Tukey's *post hoc* test.

Figure 4. Phenotypic analysis of *Ankrd31* T557A mice. A) Litter sizes from matings of *Ankrd31*$^{T557A/T557A}$ (TA/TA) and *Ankrd31*$^{T557A/+}$ (TA/+) animals to WT partners. N=3 for TA/TA male and female mice. B) Testis weights of 2-month-old mice. C) Sperm counts. D) Histological analyses of 2-mon old testes and cauda epididymides. Scale bars = 100 µm. E) Meiotic metaphase I spermatocyte chromosome preparations. Arrows indicate X and Y chromosomes. F) Histological analyses of 2-month-old ovaries. Scale bars = 100 µm. G) Follicle counts summed across every fifth serial section. Data in A, B, C and G are represented as the mean ± SEM and were analyzed using two-tailed unpaired *t test*.

Figure 5. Phenotypic analysis of *Sept12* ^{V162M} and *Sept12*^{V222I} mice. A) Litter sizes from mating of WT (+/+), *Sept12*$^{V162M/V162M}$ (VM/VM) and *Sept12*$^{V222I/V222I}$ (VI/VI) males to WT partners. N=2 for VM/VM and VI/VI mice. B) Testis weight of 2-month-old mice. C) Sperm counts. D) Histological analyses of 2-month-old testes. Scale bars = 100 μm. Data in A, B and C are represented as the mean ± SEM and were analyzed using one-way ANOVA with Tukey's *post hoc* test.

Figure 6. Comparison of predicted and actual deleterious variants analyzed with various pathogenicity predictors. A) *In silico* prediction outcomes of 10 commonly used algorithms for 26 functionally interpreted variants in mouse models. The *in vitro* experiments of DMC1_M200V variants were performed by Hikiba et.al., 2008. The variants below the dotted line all were tested by at least 1 *in vitro* pre-screen experiment. N.A., not available. B) Overview of how the Mouse_all and Mouse_infertility datasets were derived. For missense variants modeled in mice, only those with SNP rsIDs were considered. C) Prediction Accuracy (ACC), D) Positive Predictive Value (PPV) and E) Matthews correlation coefficient (MCC) of predictors using ClinVar_infertility (n=697, only pathogenic and benign variants were used), Mouse_all (n=102) and Mouse_infertility (n=34) datasets. F) ROC curves (Receiver Operating Characteristic curve) of 10 predictors in three datasets and the AUC (Area Under the Curve) values were labeled in brackets.

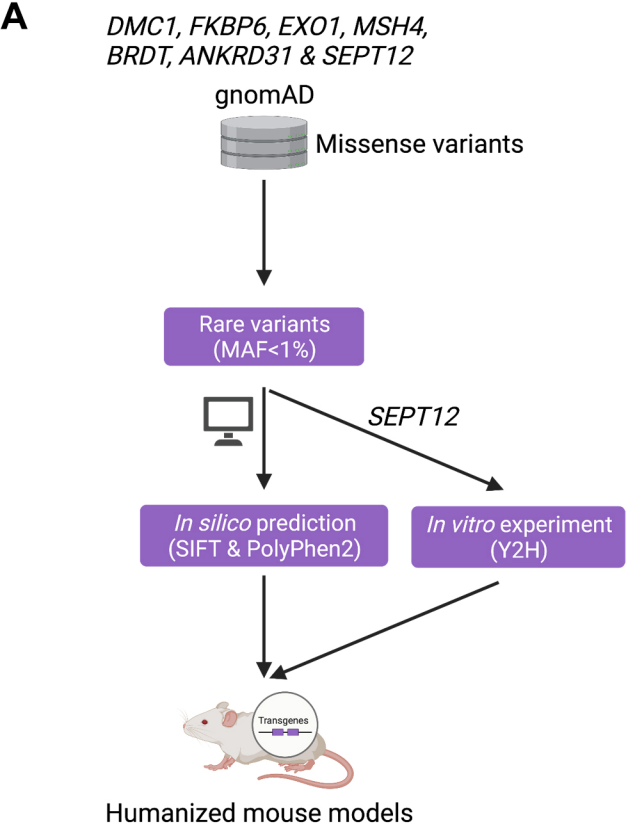
Figure 7. Classification of 102 functionally-interpreted missense variants in ClinVar. Variants modeled in mice (n=102) were classified as either Benign or

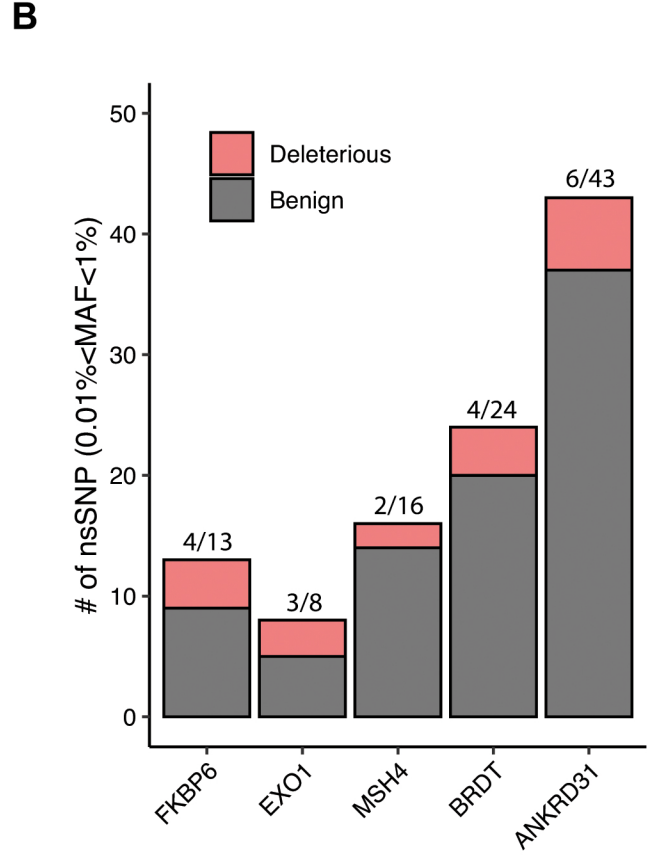
Deleterious according to the reported phenotype description. These variants were correlated with ClinVar classifications (Benign/ Likely Benign, Conflicting interpretations, VUS and Pathogenic/ Likely Pathogenic). 37 of 102 variants (36.3%) were not reported in ClinVar (N.A., not available).

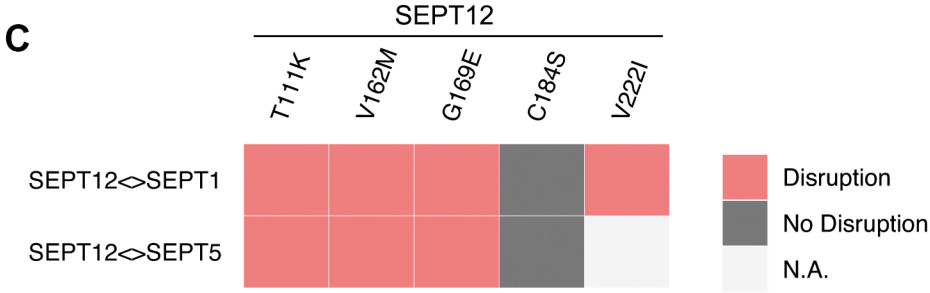
Table 1. Variants tested in this study

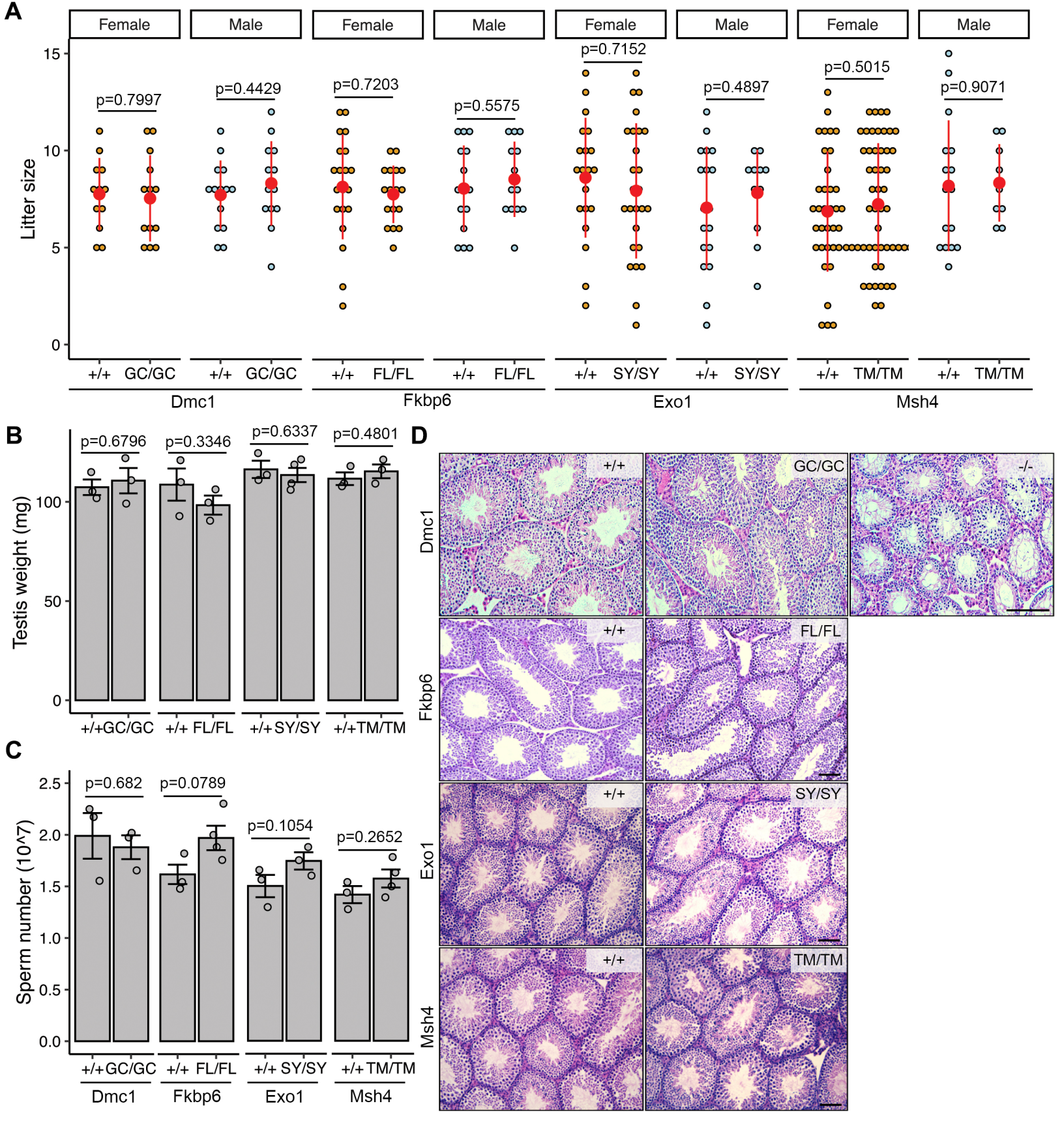
Gene	SNP (rsID)	Hs AAS	Mm AAS	AF (%)	SIFT	PolyPhen2
ANKRD31	rs150791065	T557A	T544A	0.1625	Deleterious (0.01)	Prob._damaging (1)
BRDT	rs141699970	D550H	D546H	0.01488	Deleterious (0.02)	Prob._damaging (0.972)
DMC1	rs189722264	G50C	G50C	0.000657	Deleterious (0)	Prob._damaging (0.99)
EXOI	rs4149964	S456Y	S454Y	0.3472	Deleterious (0)	Prob._damaging (0.986)
FKBP6	rs3750075	F72L	F72L	0.1131	Deleterious (0)	Prob._damaging (0.998)
MSH4	rs141042002	T410M	T432M	0.1172	Deleterious (0.01)	Prob._damaging (0.978)
SEPT12	rs144420035	V162M	V160M	0.04864	Deleterious (0)	Poss._damaging (0.884)
SEPT12	rs142721632	V222I	V220I	0.1893	Tolerated (0.31)	Benign (0.292)
SEPT12	rs145805283[a]	T111K	T109K	0.04176	Deleterious (0)	Prob._damaging (0.958)

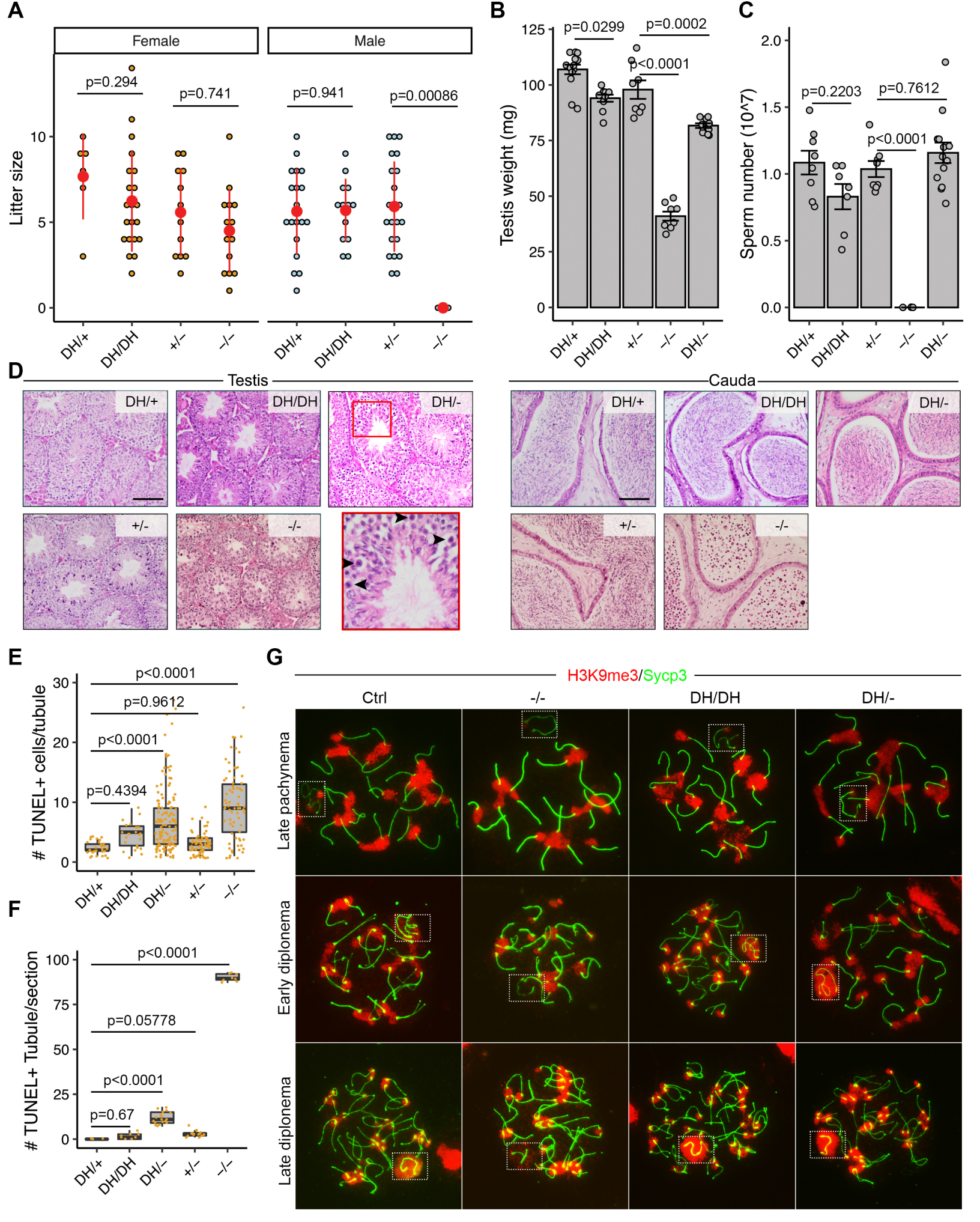
AAS, amino acid substitution. AF (Allele Frequency) based on gnomAD. [a] No mouse model was made for this allele. Variants in *SEPT12* were evaluated by Y2H only.

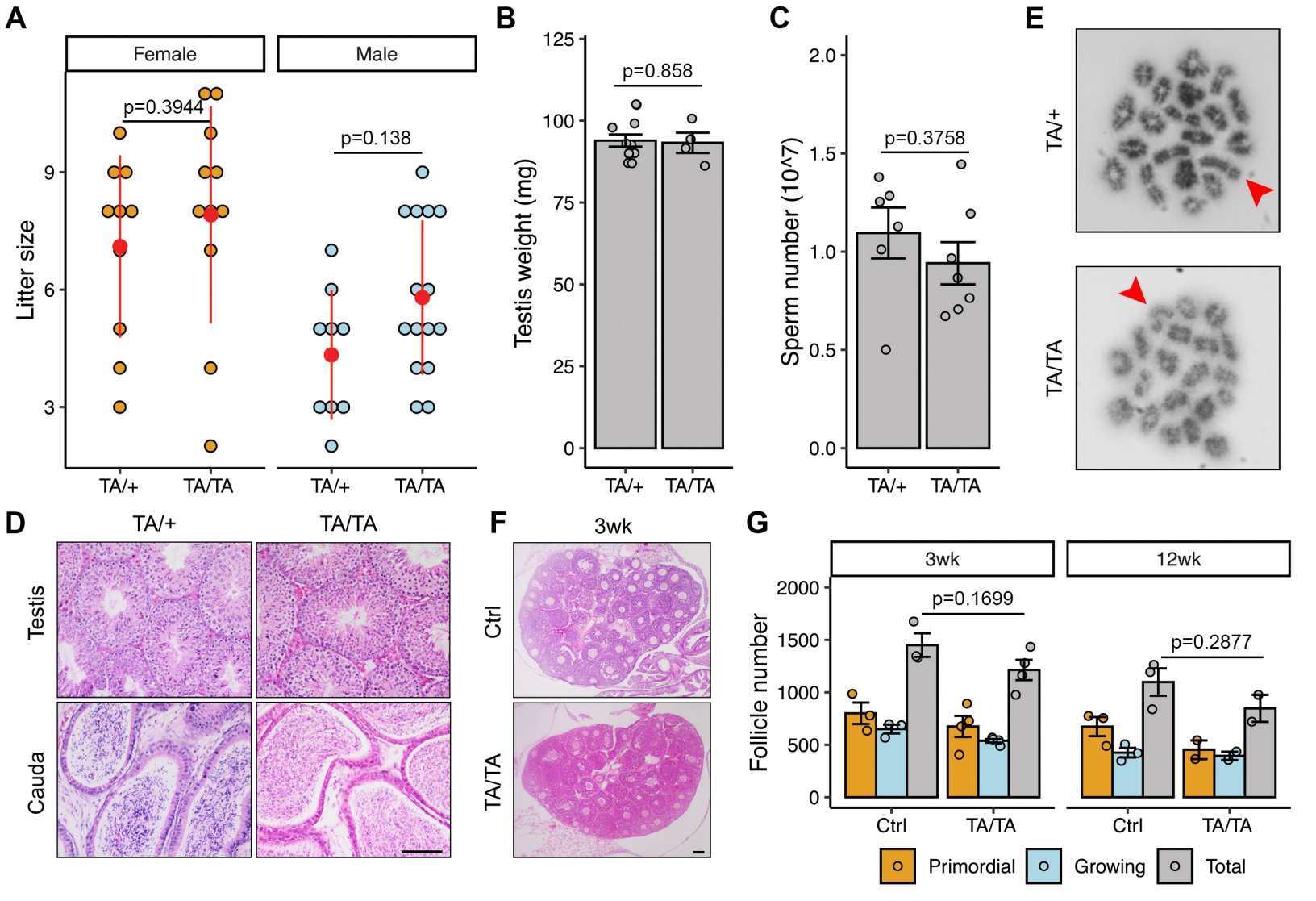

A
DMC1, FKBP6, EXO1, MSH4,
BRDT, ANKRD31 & SEPT12
gnomAD
Missense variants
Rare variants
(MAF<1%)
SEPT12
In silico prediction
(SIFT & PolyPhen2)
In vitro experiment
(Y2H)
Transgenes
Humanized mouse models

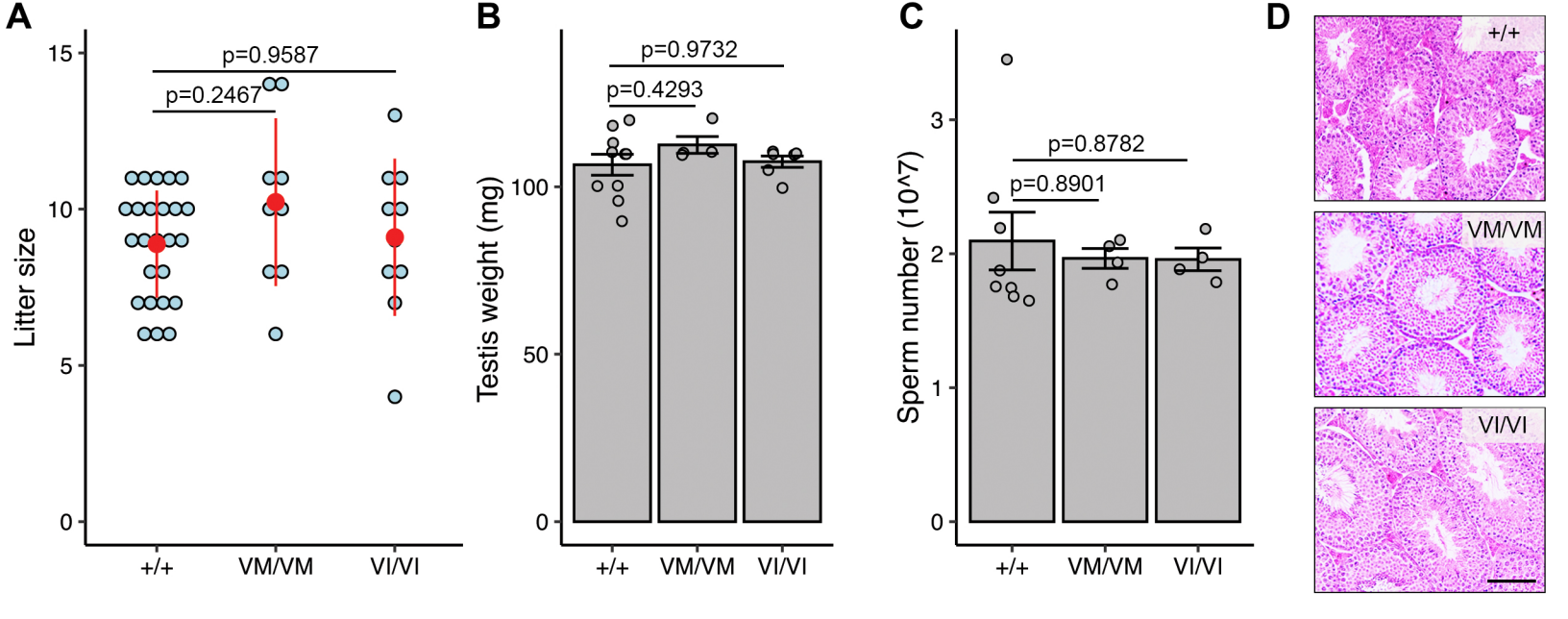

B
Deleterious
Benign
# of nsSNP (0.01%<MAF<1%)
0
10
20
30
40
50
4/13
3/8
2/16
4/24
6/43
FKBP6
EXO1
MSH4
BRDT
ANKRD31

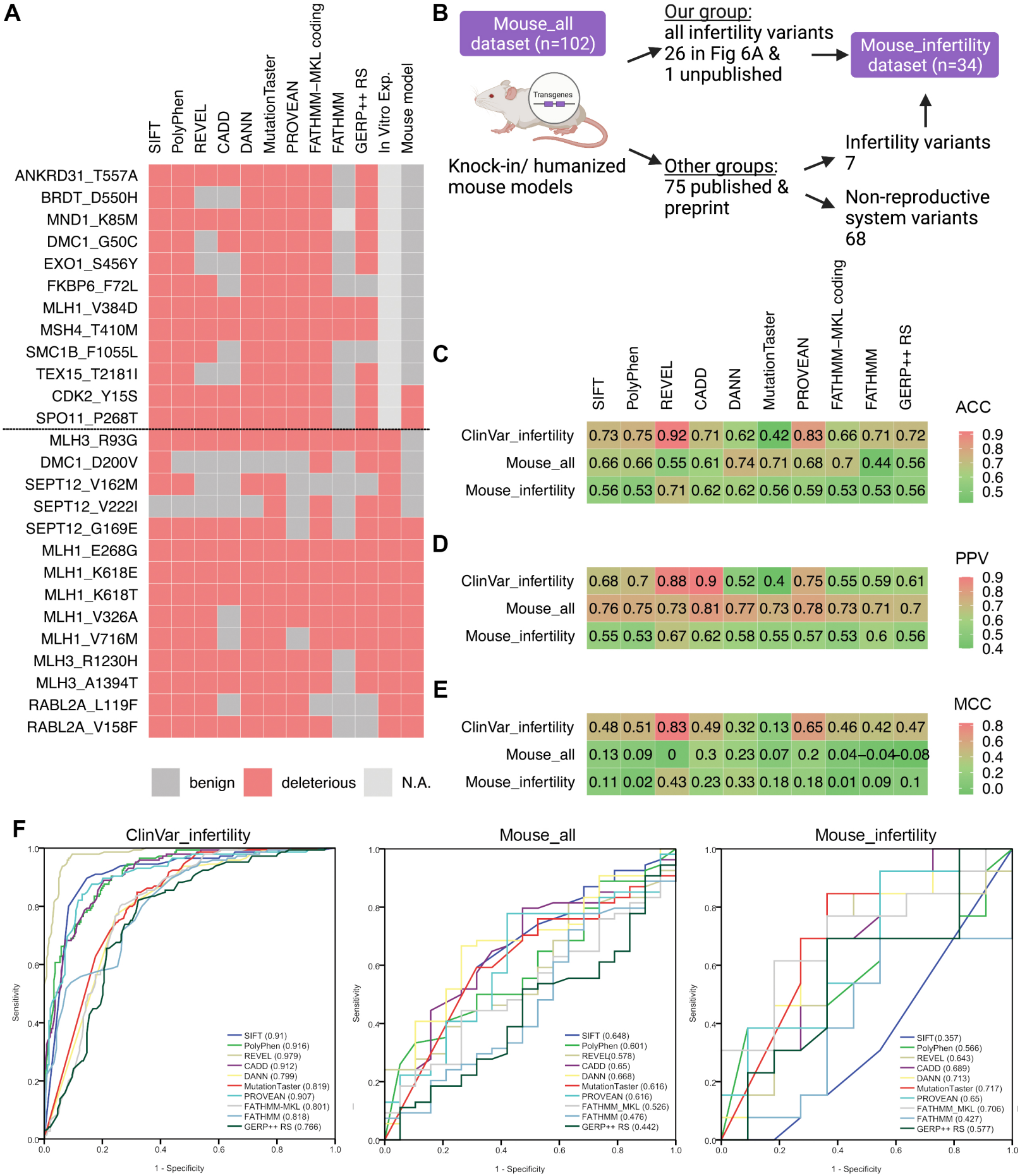

C
SEPT12
T111K
V162M
G169E
C184S
V222I
SEPT12<>SEPT1
SEPT12<>SEPT5
Disruption
No Disruption
N.A.

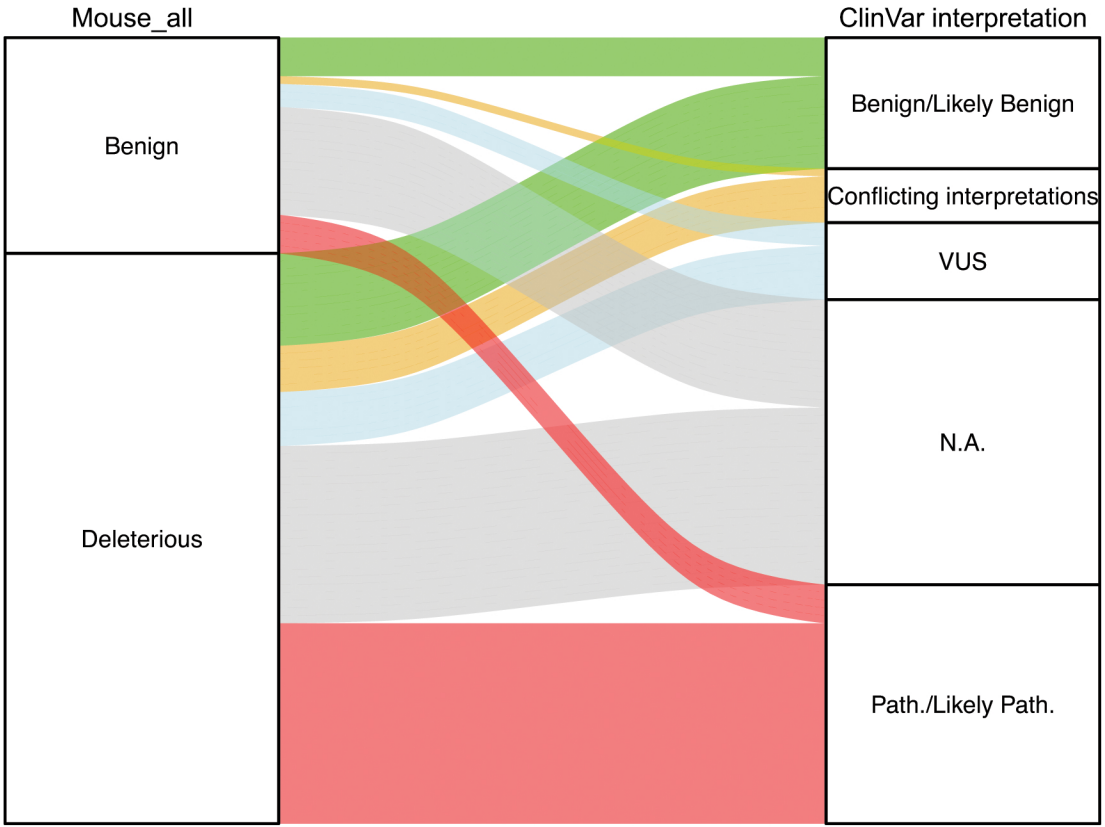

A
Female
Male
Female
Male
Female
Male
Female
Male
Litter size
p=0.7997
p=0.4429
p=0.7203
p=0.5575
p=0.7152
p=0.4897
p=0.5015
p=0.9071
+/+ GC/GC
+/+ GC/GC
+/+ FL/FL
+/+ FL/FL
+/+ SY/SY
+/+ SY/SY
+/+ TM/TM
+/+ TM/TM
Dmc1
Fkbp6
Exo1
Msh4
B
Testis weight (mg)
p=0.6796
p=0.3346
p=0.6337
p=0.4801
+/+GC/GC
+/+ FL/FL
+/+SY/SY
+/+TM/TM
C
Sperm number (10^7)
p=0.682
p=0.0789
p=0.1054
p=0.2652
+/+GC/GC
+/+ FL/FL
+/+ SY/SY
+/+TM/TM
Dmc1
Fkbp6
Exo1
Msh4
D
Dmc1
Fkbp6
Exo1
Msh4
+/+
GC/GC
-/-
+/+
FL/FL
+/+
SY/SY
+/+
TM/TM


A
Female
Male
Litter size
p=0.294
p=0.741
p=0.941
p=0.00086
DH/+
DH/DH
+/-
-/-
B
Testis weight (mg)
p=0.0299
p=0.0002
p<0.0001
DH/-
C
Sperm number (10^7)
p=0.2203
p=0.7612
p<0.0001
D
Testis
Cauda
DH/+
DH/DH
DH/-
+/-
-/-
E
TUNEL+ cells/tubule
p<0.0001
p=0.9612
p<0.0001
p=0.4394
F
TUNEL+ Tubule/section
p<0.0001
p=0.05778
p<0.0001
p=0.67
G
H3K9me3/Sycp3
Ctrl
-/-
DH/DH
DH/-
Late pachynema
Early diplonema
Late diplonema


A
Female
Male
p=0.3944
p=0.138
Litter size
9
6
3
TA/+
TA/TA
TA/+
TA/TA
B
p=0.858
Testis weight (mg)
125
100
75
50
25
0
TA/+
TA/TA
C
p=0.3758
Sperm number (10^7)
2.0
1.5
1.0
0.5
0.0
TA/+
TA/TA
E
TA/+
TA/TA
D
TA/+
TA/TA
Testis
Cauda
F
3wk
Ctrl
TA/TA
G
3wk
12wk
p=0.1699
p=0.2877
Follicle number
2000
1500
1000
500
0
Ctrl
TA/TA
Ctrl
TA/TA
Primordial
Growing
Total


A
Litter size
15
10
5
0
p=0.9587
p=0.2467
+/+
VM/VM
VI/VI
B
Testis weight (mg)
100
50
0
p=0.9732
p=0.4293
+/+
VM/VM
VI/VI
C
Sperm number (10^7)
3
2
1
0
p=0.8782
p=0.8901
+/+
VM/VM
VI/VI
D
+/+
VM/VM
VI/VI


A
SIFT
PolyPhen
REVEL
CADD
DANN
MutationTaster
PROVEAN
FATHMM–MKL coding
FATHMM
GERP++ RS
In Vitro Exp.
Mouse model
ANKRD31_T557A
BRDT_D550H
MND1_K85M
DMC1_G50C
EXO1_S456Y
FKBP6_F72L
MLH1_V384D
MSH4_T410M
SMC1B_F1055L
TEX15_T2181I
CDK2_Y15S
SPO11_P268T
MLH3_R93G
DMC1_D200V
SEPT12_V162M
SEPT12_V222I
SEPT12_G169E
MLH1_E268G
MLH1_K618E
MLH1_K618T
MLH1_V326A
MLH1_V716M
MLH3_R1230H
MLH3_A1394T
RABL2A_L119F
RABL2A_V158F
benign
deleterious
N.A.
B
Mouse_all dataset (n=102)
Transgenes
Knock-in/ humanized mouse models
Our group:
all infertility variants
26 in Fig 6A &
1 unpublished
Mouse_infertility dataset (n=34)
Other groups:
75 published &
preprint
Infertility variants
7
Non-reproductive system variants
68
C
ACC
ClinVar_infertility 0.73 0.75 0.92 0.71 0.62 0.42 0.83 0.66 0.71 0.72
Mouse_all 0.66 0.66 0.55 0.61 0.74 0.71 0.68 0.7 0.44 0.56
Mouse_infertility 0.56 0.53 0.71 0.62 0.62 0.56 0.59 0.53 0.53 0.56
D
PPV
ClinVar_infertility 0.68 0.7 0.88 0.9 0.52 0.4 0.75 0.55 0.59 0.61
Mouse_all 0.76 0.75 0.73 0.81 0.77 0.73 0.78 0.73 0.71 0.7
Mouse_infertility 0.55 0.53 0.67 0.62 0.58 0.55 0.57 0.53 0.6 0.56
E
MCC
ClinVar_infertility 0.48 0.51 0.83 0.49 0.32 0.13 0.65 0.46 0.42 0.47
Mouse_all 0.13 0.09 0 0.3 0.23 0.07 0.2 0.04 –0.04 –0.08
Mouse_infertility 0.11 0.02 0.43 0.23 0.33 0.18 0.18 0.01 0.09 0.1
F
ClinVar_infertility
SIFT (0.91)
PolyPhen (0.916)
REVEL (0.979)
CADD (0.912)
DANN (0.799)
MutationTaster (0.819)
PROVEAN (0.907)
FATHMM-MKL (0.801)
FATHMM (0.818)
GERP++ RS (0.766)
Mouse_all
SIFT (0.648)
PolyPhen (0.601)
REVEL(0.578)
CADD (0.65)
DANN (0.668)
MutationTaster (0.616)
PROVEAN (0.616)
FATHMM_MKL (0.526)
FATHMM (0.476)
GERP++ RS (0.442)
Mouse_infertility
SIFT(0.357)
PolyPhen (0.566)
REVEL (0.643)
CADD (0.689)
DANN (0.713)
MutationTaster (0.717)
PROVEAN (0.65)
FATHMM_MKL (0.706)
FATHMM (0.427)
GERP++ RS (0.577)
Sensitivity
1 - Specificity


Mouse_all
ClinVar interpretation
Benign
Deleterious
Benign/Likely Benign
Conflicting interpretations
VUS
N.A.
Path./Likely Path.