ANDROLOGY

ISSN: 2047-2919

ANDROLOGY

ORIGINAL ARTICLE

Correspondence:
Donald F. Conrad, Department of Genetics, Washington University School of Medicine, Campus Box 8232, St. Louis, MO 63110, USA.
E-mail: dconrad@genetics.wustl.edu

Keywords:
fertility genes, machine learning, ovary, systems biology, testis

Received: 4-Jun-2015
Revised: 7-Aug-2015
Accepted: 26-Aug-2015

doi: 10.1111/andr.12109

Improved detection of disease-associated variation by sex-specific characterization and prediction of genes required for fertility

N. R. Y. Ho, N. Huang and D. F. Conrad

Departments of Genetics, and Pathology & Immunology, Washington University School of Medicine, St. Louis, MO, USA

SUMMARY

Despite its great potential, high-throughput functional genomic data are rarely integrated and applied to characterizing the genomic basis of fertility. We obtained and reprocessed over 30 functional genomics datasets from human and mouse germ cells to perform genome-wide prediction of genes underlying various reproductive phenotypes in both species. Genes involved in male fertility are easier to predict than their female analogs. Of the multiple genomic data types examined, protein–protein interactions are by far the most informative for gene prediction, followed by gene expression, and then epigenetic marks. As an application of our predictions, we show that copy number variants (CNVs) disrupting predicted fertility genes are more strongly associated with gonadal dysfunction in male and female case–control cohorts when compared to all gene-disrupting CNVs (OR = 1.64, $p < 1.64 \times 10^{-8}$ vs. OR = 1.25, $p < 4 \times 10^{-6}$). Using gender-specific fertility gene annotations further increased the observed associations (OR = 2.31, $p < 2.2 \times 10^{-16}$). We provide our gene predictions as a resource with this article.

INTRODUCTION

Infertility is a growing problem around the world. In its 2014 infertility white paper, the CDC reported that 12–18% of couples and 9% of men are infertile in the United States (Centers for Disease Control and Prevention, 2014). Infertility is also highly heritable, with heritability estimates ranging from 0.16 to 0.81, with a mean of around 0.3 (Kosova *et al.*, 2010). When looking specifically at male infertility, male relatives of couples treated with intracytoplasmic sperm injection were found to have higher rates of infertility than the general population (Meschede *et al.*, 2000), and up to 10% of cases of azoospermia are clinically attributable to Y-chromosome microdeletions in typical populations of European ancestry (Hotaling & Carrell, 2014; Krausz *et al.*, 2014). Numerous physiological systems are required for the maintenance of human fertility, and genetic studies in mice and humans have played a major role in their dissection. Genes are now known to be involved in the proper formation of male and female gonads and genitalia, neuroendocrine control of gonadal function, paracrine regulation of gamete development, fertilization and implantation (Matzuk & Lamb, 2008). In total, the data suggest there is a significant genetic component to infertility, and, as high-throughput DNA sequencing moves to the clinic, there will be a pressing need to interpret variation of unknown significance in patients with infertility.

There are two classic sources of information on the relative importance of each human gene to normal gonadal function. Human genetic analysis, including genome-wide association studies (GWAS) and pedigree studies, has provided definitive lists of genes involved in human gonadal function, but because of the limited statistical power of these approaches, progress has been slow. To date, only a small handful of loci have been identified as definitively involved in human fertility, and these genes explain only a small proportion of the heritability of fertility (O'Flynn O'Brien *et al.*, 2010; Evian Annual Reproduction (EVAR) Workshop Group 2010, *et al.*, 2011; Kosova *et al.*, 2012; Xu *et al.*, 2013; Zorrilla & Yatsenko, 2013). Second, systematic analysis of mouse mutants is producing a larger list of gonad-essential genes, but this approach also suffers from ascertainment and a relatively slow rate of progress. As of today, a large fraction of the genome remains uncharacterized in human and mouse.

In parallel to this classical genetic work, there has been a recent explosion in the amount of genome-wide genomic data being generated on hundreds of human and mouse cell types, including germ cells, much from the Encyclopedia of DNA

Elements Consortium (ENCODE) (Consortium, 2012). Gene expression, histone modifications and methylation have all been assayed on bulk gonadal tissue, and in some cases, purified germ cells. We hypothesize that while known fertility genes work in a small set of pathways, we can computationally identify genomic features among this high-throughput data that distinguish 'fertility genes', and then apply this knowledge to identify all genes in the genome with similar features.

To accomplish this we apply a technique known as 'supervised learning' (Rogers & Girolami, 2012), where one creates a model for classifying unlabeled objects from studying pre-existing, labeled, training data. In the simplest implementation, the purpose of the classifier is to place unlabeled objects into one of two groups, say 'positive' and 'negative'. This method has had good results for well-studied diseases with a large set of known causative genes in humans (Sun *et al.*, 2011; Singh-Blom *et al.*, 2013) and various tools have been made which use single features to define similarity among genes, ranging from disease ontology terms to protein structure and tissue-specific expression (van Driel *et al.*, 2005; Radivojac *et al.*, 2008; Vanunu *et al.*, 2010; Ballouz *et al.*, 2013).

While the genomic feature information fed to the model greatly influences the results, we had to build our own tool as the existing tools can only use a small subset of the current high-throughput data. We explored the use of a diverse set of high-throughput genomic data types, including protein–protein interaction (PPI) networks, gene co-expression networks, tissue- and cell type-specific expression levels, epigenetic marks and gene conservation. In the process, we obtained over 30 published sequencing-based genomic datasets from human and mouse models and reprocessed them with a uniform pipeline to ensure comparability.

The other key factor that determines accuracy is the size and curation of the training set used to train the model. Larger gene sets that are more specific to a given phenotype improve the performance of supervised learning models. In the case of infertility, there is a relatively small list of genes that have been definitively shown to cause disease in humans. To get around this issue, we used data from mouse knockout lines to augment our training datasets.

We first tested our approach on the mouse genome, classifying genes in general categories like 'reproductive' as well as focusing on specific physiological processes such as 'meiosis arrest' or 'ovulation' for classification. We were able to validate that our classifier performs at a comparable level to other previously published models. In general, there were improvements in classification accuracy for the more specific phenotypes, as quantified by the area under the receiver operator characteristic area under the curve (AUC). We then applied our model to the human genome to classify genes by using the small set of known human fertility genes.

The main product of our work is a list of quantitative predictions about the relevance of each gene in the human and mouse genome to reproductive function. These quantitative summaries can be used as a research resource for hypothesis generation, say in the design of experiments, or the interpretation of human genetic data. We give examples of some uses of our quantitative predictions by showing how they can be used to improve detection and interpretation of pathogenic copy number variants (CNVs) in GWAS of gonadal function.

One other significant result of our work is to provide some insight into the relative importance of the complex and rapidly growing genomic data on reproductive cell types. By evaluating a large amount of genomic data side by side, we were able to make precise statements about the information generically relevant to infertility contained in each of these data types. This is a first attempt at what we believe will be an increasingly visible and routine problem for reproductive biologists: how to computationally integrate human genetic analysis, model organism research and genomic data to precisely predict the reproductive consequences of mutation.

METHODS

Training gene sets

To create the positive and negative trainings sets of mouse genes, we first used the Mouse Phenotypic Alleles database from Jackson Labs Mouse Genome Informatics (MGI) to make a list of unique genes where a knockout mouse had been made and phenotyped for at least one system. From this list, we extracted the genes with an observed reproductive system phenotype (MP:0005389) to make a reproductive positive training set gene list (Table S12). For the negative training set gene list, we took all the other genes from the list that did not have a reproductive system phenotype (Table S10) and further filtered out the genes which caused embryogenesis (MP:0005380) and abnormal survival (MP:0010769) phenotypes, but not the extended life span phenotype (MP:0001661) among the abnormal survival genes (Table S11).

To make the male reproductive training gene list we picked the genes shown to cause abnormal male reproductive morphology (MP:0001145) and physiology (MP:0003698) from the original positive training set gene list (Table S13). Similarly, we used the categories for abnormal female reproductive morphology (MP:0001119) and physiology (MP:0003699) to create our female reproductive training gene list (Table S14). We also evaluated each of 12 subcategories: abnormal female meiosis (MP:0005168), abnormal endometrium morphology (MP:0004896), abnormal spermiogenesis (MP:0001932), azoospermia (MP:0005159), decreased oocyte number (MP:0005431), male meiosis arrest (MP:0008261), oligozoospermia (MP:0002687), teratozoospermia (MP:0005578), abnormal ovulation without superovulation (MP:0001928), abnormal ovulation cycle (MP:0009344), male germ cell apoptosis (MP:0008280) and sperm physiology (MP:0004543), all taken from the same MGI database (Tables S19–S30).

Mouse Genome Informatics's vertebrate homology table was used to translate the negative training sets into human conserved genes. Due to orthology relationships between mouse and human, the sizes of the human training genes set differed slightly from those of mouse, increasing from 3344 genes in mouse to 3406 genes in human.

The human genetic studies' derived positive training gene set was taken using azoospermic/oligospermic gene set (AO) and female infertility gene set (POF) from some review articles (Navarro-Costa *et al.*, 2010; Evian Annual Reproduction (EVAR) Workshop Group 2010, 2011; Hamada *et al.*, 2013; Zorrilla & Yatsenko, 2013). This produced a list of 67 human male fertility genes (Table S16) and 62 human female fertility genes (Table S17). These lists were combined to produce a list of 125

human fertility genes (Table S15). Finally the human male fertility gene list was curated for genes only found in non-obstructive azoospermia to produce a list of 41 genes (Table S18).

Gene similarity features

All genomic feature data were normalized to a mean of 0 and a spread from −1 to 1 for the purposes of being able to compare different features.

Expression properties

Encyclopedia of DNA Elements Consortium (Consortium, 2012; Yue *et al.*, 2014) paired-end RNA-Seq reads were used for differential tissue expression. Mouse liver (ENCSR000AJU and ENCSR216KLZ), heart (ENCBS441FDF and ENCSR000BYQ), testis (ENCSR266ESZ and ENCSR000BYW) and ovary (ENCSR516UNF and ENCSR000BZC) were used. Human liver (ENCSR085HNI and ENCSR000AEU), heart (ENCSR000AHH and ENCSR635GTY), testis (ENCSR693GGB) and ovary (ENCSR046XHI) were used. Mouse spermatogenesis-specific expression was obtained from RNA-Seq paired-end reads of two datasets, Soumillon (GSE43717) (Soumillon *et al.*, 2013) and Hammoud (GSE49624) (Hammoud *et al.*, 2014). Mouse and human oocyte RNA-Seq paired-end reads were taken from Xue (GSE44183) (Xue *et al.*, 2013).

All mouse RNA-Seq fastq files were mapped to the mm9 assembly while human RNA-Seq fastq files were mapped to the hg19 assembly. Alignment was performed using tophat2 (Kim *et al.*, 2013) with the default values and gene expression was summarized using cuffnorm (Trapnell *et al.*, 2013) on the UCSC gene annotations to normalize across the datasets. We used cuffdiff with the default options to determine which genes are differentially expressed.

Histone modification properties

H3K27ac, H3K27me3, H3K4me1 and H3K4me3 histone modification marks for mouse spermatogenesis cell-specific stages were taken from Hammoud (GSE49624). ENCODE was the source for the same histone modification marks for mouse testis (ENCSR000CCU, ENCSR000CGB, ENCSR000CCV and ENCSR000CCW).

We aligned the CHIP-Seq read to the mm9 assembly using Novocraft's novoalign tool with its default options (http://www.novocraft.com). Following that, we used seqMINER (Ye *et al.*, 2011) to map the reads + −5 kb around the transcription start site (TSS). We created several statistical summaries of the distribution of marks around the TSS including mean, standard deviation, kurtosis and skew.

Network properties

Protein–protein interactions were collected from HPRD (Prasad *et al.*, 2009), Reactome (Joshi-Tope *et al.*, 2005) and STRING (Franceschini *et al.*, 2013) and integrated into a single PPI network by mapping interacting entities to HGNC symbols. Measures of network centrality (degree and betweenness) and modularity (cluster coefficient) were calculated using MCL (Enright *et al.*, 2002). Sum of weight of edges were calculated as a measure of proximity to a group of 'seed' genes as described previously (Huang *et al.*, 2010). Seed gene sets that we used to calculate scores included cancer, early development, haploinsufficiency and known reproductive genes that we supplied to the model.

Gene properties

The dN/dS, GERP scores, number of domains, number of exons and length of domains for each gene were downloaded from EnsEMBL version 74.

Linear discriminant analysis model

For each genomic feature, a given gene will have a score normalized to between −1 and 1. For this score, one can calculate the likelihood that a given gene belongs in either the positive training set genes or the negative training set genes based on how similar it is to each group. In order to combine the information from multiple features together we use linear discriminant analysis (LDA). The LDA approach assigns weights to each given feature such that the likelihood variance within each group is minimized and the variance between the positive and negative groups is maximized. Using the results from the LDA we then calculate the χ score for all genes which is a projection of the multidimensional data onto a one-dimensional continuum. We can then pick a threshold χ score to divide the positive and negative groups, thus classifying the genes as either reproductively important (positive) or not (negative).

To do 10-fold cross validation, we first split the positive and negative training sets into 10 random subsets, then training the model using nine of those subsets, leaving one subset for testing. We then plot the false-positive rate (FPR) using the remaining negative subset and the false-negative rate using the remaining positive subset at the various likelihood cutoffs for each possible subset. The receiver operating characteristic (ROC) curve is generated by plotting the average FPR against the false-negative rate of the 10 models.

Human infertility gene deletion analysis

We obtained existing CNV calls from men assayed in our previous study of spermatogenic impairment (Lopes *et al.*, 2013). Using our published, validated CNV calling pipelines, we generated new CNV calls from two female cohorts with extensive reproductive health history, GARNET and SHARE, both of which are components of the Women's Health Initiative (WHI), using data obtained, with permission, from the Database of Genotypes and Phenotypes (dbGAP, Bethesda, MD: National Center for Biotechnology Information, National Library of Medicine). We used the extensive health history data available on each WHI subject to construct a diagnosis that we believe approximates the clinical definition of primary ovarian insufficiency, resulting in a case–control classification for all WHI individuals.

We performed a series of case–control association tests, testing for association between CNV carrier status and disease status. Patients were defined as 'positive' for CNV carrier status if they carry at least one CNV that results in meeting one of the following criteria, depending on the analysis: gene disrupting, fertility gene disrupting or sex-specific fertility gene disrupting, where disrupting means that the CNV is deleted, not duplicated. A patient was otherwise classified as 'negative' for CNV carrier status. We then built a 2×2 contingency table for the case–control status and ran a Fisher's exact test.

We then picked the sex-specific fertility gene-disrupting CNVs that were found only in the cases and not the controls and

extracted the candidate genes from them for further analysis. The three genes that occurred more than once were discussed in the article while the one-off genes were listed in a separate file (Table S31).

RESULTS

We chose to use Linear Discriminant Analysis (LDA) as the modeling framework for our supervised learning classifier. LDA has worked well in the past to generically predict human haploinsufficient genes (Huang *et al.*, 2010). While LDA uses a linear combination of genomic features to make its predictions, the features with the largest separation between training groups are also the ones weighted most when classifying the test set. This provides us with the advantage of being able to consider many different data types, picking only the most informative genomic features (larger difference = more informative). In this study, we considered numerous genomic features such as stage-specific and tissue differential RNA expression data, locus conservation between species and PPI, but only picked the best (3–4) to actually make any given model prediction.

There are two inputs for an LDA model. Aside from the genomic features that the LDA model will use to calculate variance, it also needs examples of the 'positive' and 'negative' genes. Because the ideal, large and well-curated fertility training set is unavailable for humans, we performed our investigations with various gene sets (Fig. 1). This resulted in predictions of sets of genes involved in different reproductive processes, ranging from a category as broad as 'fertility' to something as narrow as 'abnormal ovulation without superovulation'. We have picked the models with the best results to present here, but full results from all the models we tested are presented in the supplement.

A popular measurement of the accuracy of a classification model is the area under the receiver operator curve (AUC), where a larger AUC means the model has a better trade-off between accuracy and specificity. The maximum AUC for a two-category model is 1 (perfect prediction) and the minimum is 0.5 (random guessing). For each set of predictions, we used 10-fold cross validation to test the precision and sensitivity of the LDA models. This method essentially leaves out 10% of the training set for testing and repeats it 10 times, leaving out different genes each time, in order to determine how much error there is in the precision and sensitivity measurements.

MGI genes on mouse genome model

We first constructed models for three broad categories of genes: fertility, male fertility and female fertility, using mouse genetic and genomic data. The AUC for the gender-aspecific model was 0.711, for the male-specific model was 0.741 and for the female-specific model was 0.738, for the 15 212 genes tested in the mouse genome (Fig. 2).

In principle, genes involved in a narrow biological process should be more tightly co-regulated and co-evolving than a set of genes involved in diverse processes; thus, we reasoned that genes involved in narrowly defined processes should be easier to model and predict. On the basis of the phenotype observed in knockout mice, we picked 12 of them that had at least 50 different genes implicated. We then characterized these models on the mouse genome (see Methods section).

The subcategory models all had better ROC AUC than the more general infertility models, but their precision–recall AUCs were not as good (Fig. 3). Among these models we ended up characterizing the results of two, male meiosis arrest and abnormal female meiosis, because they were the only ones that performed better than the more general fertility models in both metrics.

Among all the genomic features for the mouse models that we tested (Figs. S1–S3, S9–S20), we found that PPI distance from genes in the positive set and gonad RNA expression were the most important ones. Some histone modification marks (H3K27ac) also proved to be helpful in building the male infertility prediction models (Fig. 4).

Figure 1 Overview of the study. We set out to assess the utility of functional genomic data for predicting the identity of genes relevant to mammalian fertility using a machine learning approach. We obtained and reprocessed over 30 high-throughput functional genomic datasets from mouse and human models, and used these to annotate all genes in the mouse and human genomes respectively. Using extensive phenotyping data from Jackson Labs Mouse Genome Informatics, we generated a negative training set of genes which were highly unlikely to be related to mammalian fertility. We then created multiple positive gene training sets of genes known to disrupt mammalian fertility, identified in either mouse or human models. We combined each positive gene training set with the negative gene training set and used these to create phenotype-specific gene classifiers using linear discriminant analysis. The accuracy of each classifier was then evaluated using standard statistical approaches.

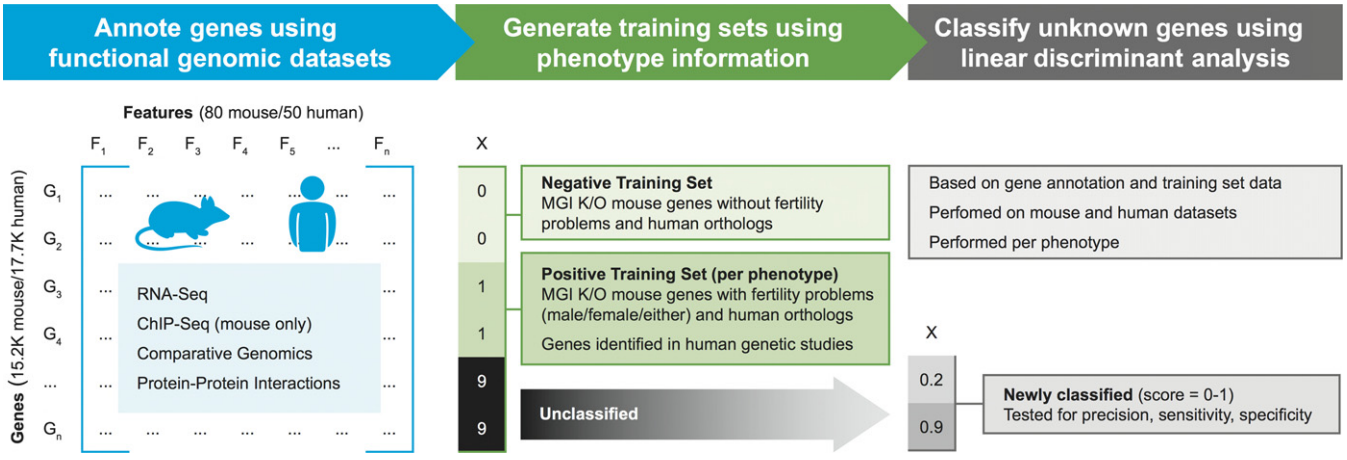


Figure 2 Model performance benchmarks for fertility gene sets. Each figure shows the receiver operator characteristic curves corresponding to classifiers based on functional genomic data derived from mouse (left) or human (right) genomes. The negative training set for each classifier is always the same set of null genes.

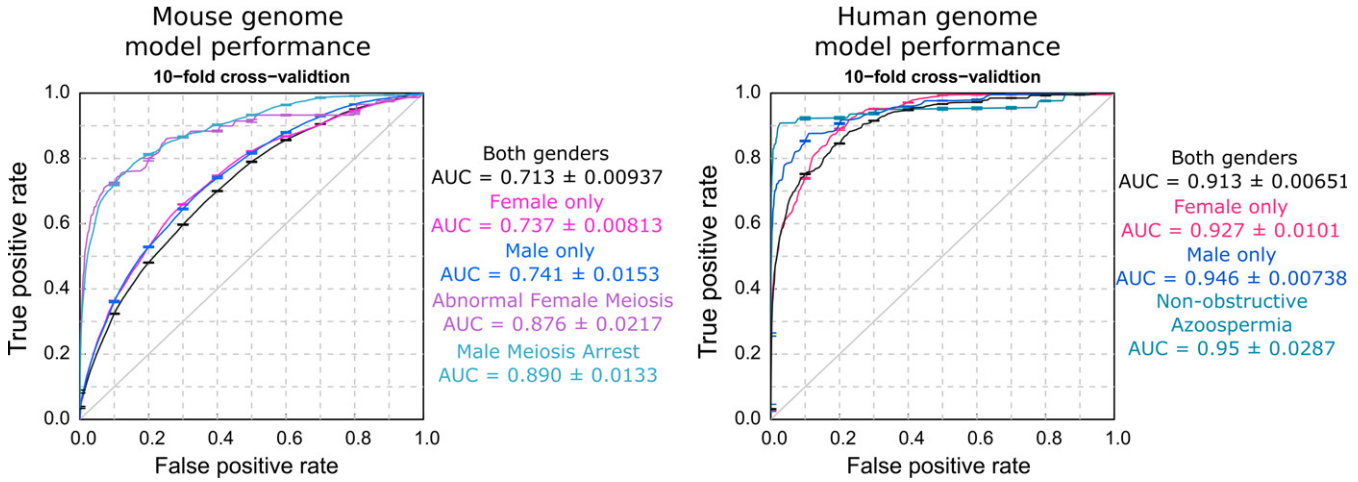
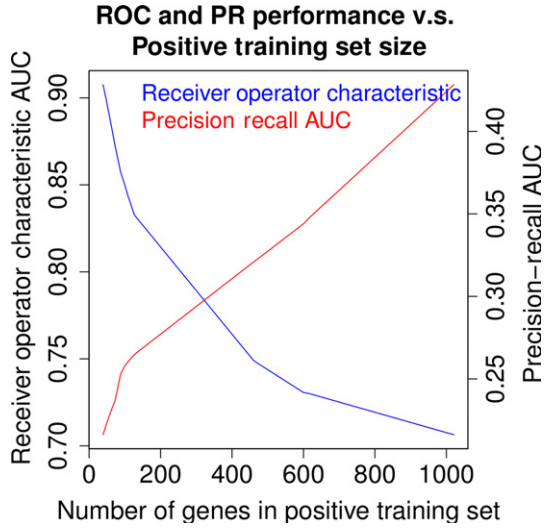


Figure 3 Model performance vs. training set size. For each classification model, we plotted the precision–recall AUC (AUPRC) vs. receiver operator characteristic AUC (AUROC) as two measures of model performance. We fit trend lines to each set of points using LOESS regression, with AUPRC in red and AUROC in blue. In general, AUROC decreases with increasing positive training set size, while AUPRC increases.



Human genetic studies' genes on human genome model

Our approach was working for the large and small mouse-derived training sets, so we reasoned that it may work well with the smaller set of fertility genes implicated in humans. We combined male and female infertility genes that work in various pathways from review articles (Navarro-Costa *et al.*, 2010; Evian Annual Reproduction (EVAR) Workshop Group 2010 *et al.*, 2011; Hamada *et al.*, 2013; Zorrilla & Yatsenko, 2013) to generate four positive training sets (see Methods section). Because it was difficult to find a list of genes proven not to cause fertility problems in humans, we translated the MGI null gene set into conserved human genes and used this as the negative training set.

We got better performance for these models compared to the mouse models, with AUCs of 0.913 for the general fertility model, 0.946 for the male-specific model, and 0.927 for the female-specific model for the 17 758 genes tested in the human genome. The non-obstructive azoospermia model also had a good AUC of 0.95 (Fig. 2).

There were fewer human genomic features available compared to mouse (Figs S4–S7), and among the ones we tested PPI with genes in the positive set was the most important. Other useful features were gene conservation and gonad RNA expression, but to a much smaller extent (Fig. 4).

Model predictions

In order to produce a list of candidate genes likely to modulate fertility, we used a cutoff for the χ score produced by the LDA model, where any gene with a χ score greater than or equal to the cutoff is considered a candidate for being an important gene for reproduction. The cutoff for each model was chosen so that the resulting candidate gene predictions would have at most a 5% FPR (Table 1). This created shortlists of candidate infertility genes numbering between 400 and 900 genes for the mouse genome and between 590 and 1030 genes for the human genome (depending on the phenotype). We provide, as supplemental information, results of our gene classification framework for nine phenotypes (Tables S1–S9). These tables can be used with different cutoffs to produce more stringent/lenient predictions depending on their purpose.

Using model predictions to improve identification of human fertility-associated CNVs

A primary challenge in GWAS is the identification of true disease-associated variation among the background of millions of unassociated variants within a given set of individuals. One strategy for improving detection power is to test only those variants that have strong a priori evidence for contributing to the disease process, such as variants near genes expressed in the tissue(s) of interest. We sought to evaluate how our fertility gene predictions could be used to improve analysis of case–control data from cohort studies of gonadal function.

First, we took data generated from several cohorts of male and female gonadal dysfunction, as well as matched controls,

Figure 4 Selected feature importance to each model. These show the relative importance of the genomic features used to construct the nine models presented toward the model predictions. We show a subset of all the features that we tested presenting only the ones actually used in the model(s). All the other features are presented in the supplement.

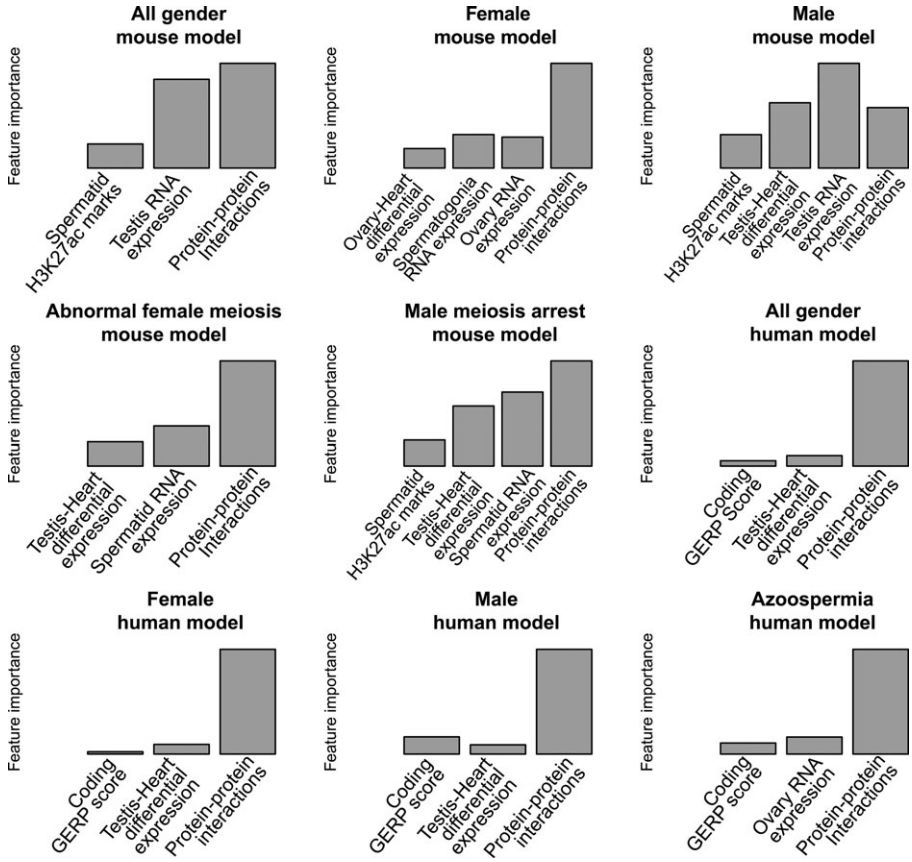


Table 1 Summary of prediction results for nine reproductive gene classifiers

LDA model	Predictive score cutoff	False-positive rate	Sensitivity	Precision	Positive training set size	Negative training set size	No. candidate genes
MGI reproductive set on mouse genome	0.4906	0.05	0.211	0.557	999	3344	558
MGI male reproductive set on mouse genome	0.3585	0.05	0.228	0.449	600		737
MGI female reproductive set on mouse genome	0.4492	0.05	0.247	0.402	458		402
MGI male meiosis arrest set on mouse genome	0.0225	0.05	0.609	0.2	69		867
MGI abnormal female meiosis set on mouse genome	0.00197	0.05	0.692	0.1385	39		876
Human fertility genes on human genome	0.1055	0.05	0.536	0.282	125	3406	638
Human male fertility genes on human genome	0.0002303	0.05	0.612	0.193	67		592
Human female fertility genes on human genome	0.0355	0.05	0.516	0.158	62		696
Human non-obstructive azoospermia genes on human genome	4.737×10^{-32}	0.05	0.659	0.136	41		1030

The table shows the benchmarks using different cutoffs for the chi-squared score produced by the functional gene prediction models. The predictive score cutoff for each model was chosen to result in a 5% false-positive rate. We tested 15 212 genes in the mouse genome and 17 758 genes in the human genome.

previously used for GWAS. We found that an infertile patient had an odds ratio of 1.25 of having a deletion spanning any gene exon compared to a control individual. When we used our all-gender model gene predictions to consider only deletions spanning at least one exon of a candidate gene, the same patients had a slightly higher odds ratio of 1.48 than the controls. Finally, when looking at the male human model predicted infertility genes for the male cases and the female human model

Table 2 Tests of association between gene-disrupting CNVs and infertility

All-gender model scores			Fisher's exact test	
	Positive	Negative	*p*-value	1.64×10^{-8}
Case	313	1558	Odds ratio	1.475309
Control	1792	13 160	95% Confidence interval	1.289827–1.683816
Gender-specific model scores			**Fisher's exact test**	
	Positive	Negative	*p*-value	2.2×10^{-16}
Case	642	1229	Odds ratio	2.307403
Control	2760	12 192	95% Confidence interval	2.075971–2.563118
Deletion CNVs spanning genes			**Fisher's exact test**	
	Positive	Negative	*p*-value	3.663×10^{-6}
Case	953	1005	Odds ratio	1.250934
Control	6447	8505	95% Confidence interval	1.136988–1.376255

Positive/negative status of case–control individuals was determined by taking the highest scoring gene in any CNV in the patient and judging based on the cutoffs determined in Table 1. Positive/negative status for the deletion CNVs spanning genes table was determined by whether the patient had any loss of CNVs that span exons.

Table 3 Candidate genes identified multiple times in gene-disrupting CNVs for infertility cohorts

Gene	Recurrence	Cohort	All-gender chi-squared score	Female chi-squared score	Male chi-squared score
CDY1	2	Azoo	0.999999	5.03E-03	1
CDY1B		Azoo	0.999998	1.36E-04	1
DAZ1		Azoo	1	5.75E-03	1
DAZ2		Azoo	1	3.47E-02	1
DAZ3		Azoo	0.125516	6.74E-07	0.99982
DDX3Y	5	Azoo	0.999991	1.59E-04	1
USP9Y	6	Azoo	1	8.81E-04	1
DMRT1	4	Azoo	1	0.999999	1
PSG5	2	Azoo azoo	0.640456	0.877	0.000839
PRL	2	POI	1	1	1

Known infertility genes have their names bolded.

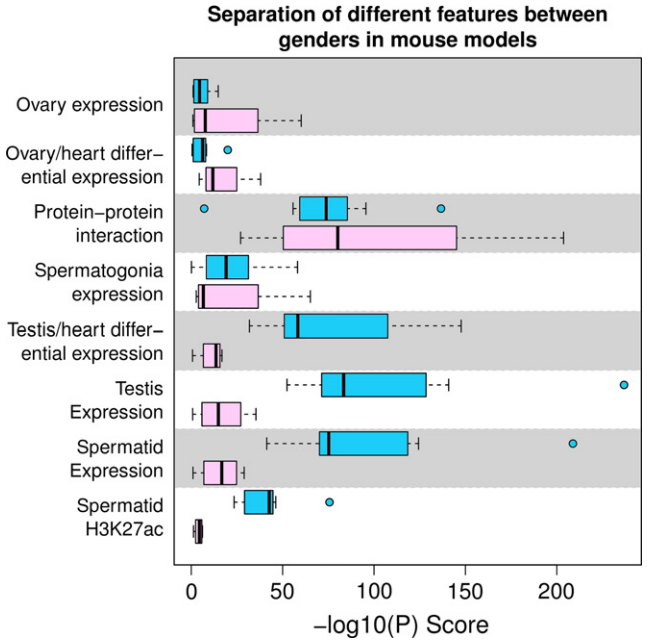
predicted infertility genes for the female cases, we found that cases had a much higher odds ratio of 2.31 of having a deletion in one of the predicted infertility genes than controls (Table 2).

We also looked for genes that were deleted multiple times in either male or female cases alone, but not controls (Table 3). This produced a list of 14 patients in azoospermia and two patients in primary ovarian insufficiency. Eight of our cases of azoospermia had deletions covering known male fertility genes (*CDY1*, *CDY1B*, *DAZ1*, *DAZ2*, *DAZ3*, *DDX3Y* and *USP9Y*), while four of them had deletions covering *DMRT1* (two of them also covered *FOXD4*). The last two patients had deletions covering *PSG5*. For our female cases with primary ovarian insufficiency, we found two patients with deletions covering *PRL*.

DISCUSSION

The premise of this study was that high-throughput functional genomic data from mouse and human germ cells and tissues could be used to identify novel infertility genes. Ideally this would work by finding other genes that are regulated similarly or interact with known infertility genes, thus likely to work in the same molecular pathways. Because pathways are often the basis of genotype–

Figure 5 Functional annotations contain both general and sex-specific information. Many of the functional annotations used to produce our classifiers are obtained from sex-specific germ cells. For the eight top most informative features in our study, we show the relative importance of each feature to the performance of seven male (blue) and five female (pink) classifiers, summarized as a box-and-whiskers plot of the $-\log_{10}(P)$ scores for each feature. A higher $-\log_{10}(P)$ score indicates that the feature better differentiates between the positive and negative training set genes. While features derived from male gonads were typically more sex-biased in their predictive power than features derived from female gonads, it is interesting to note that spermatogonial expression levels appeared to be equally useful for predicting genes involved in both male and female reproductive traits.



phenotype mapping, we expect that disrupting the same pathway in different ways can produce correlated disease phenotypes.

Are the genomic features that were ultimately most informative for our gene classifiers consistent with this hypothesis? It would appear so, given that the PPI distance to reproductive genes was consistently the most significantly separated genomic data features between the positive and negative gene sets. Aside from PPI distance, six of the eight most useful genomic features that we observed were based on germ cell or gonad gene expression levels, and only one was based on an epigenetic mark (Fig. 5). Genomic features derived from male tissues tended to be more informative across multiple models than genomic features derived from female tissues. This could reflect the fact that more high-quality functional genomic data are available on specific developmental subpopulations in male gametogenesis compared to female gametogenesis. This is largely due to technical limitations in isolating and generating data from scarce cellular populations, and we expect that richer female functional genomic datasets will emerge with time and innovation. Intriguingly, some genomic features derived from male gonads were also informative for predicting female infertility genes across a broad range of phenotypes, especially male germ cell and gonad expression levels. We interpret this as underscoring a common set of pathways that are involved in gametogenesis for males and females (probably beyond obvious shared processes such as meiosis).

These results suggest that obtaining high-resolution RNA expression of bulk gonadal tissue and purified germ cell

populations will be the best way to improve fertility gene predictions in both humans and mice. Furthermore, it appeared that RNA expression results that came from a pool of cells were more reliable than the single-cell experiments, leading us to conclude that for single-cell sequencing results to be useful, many replicate experiments from one individual may be needed to get an accurate idea of the average cell expression.

We evaluated our ability to predict genes involved in 15 mouse reproductive phenotypes and 4 human reproductive phenotypes. Each predictive model produced an area under the ROC curve in the range of 0.7–0.9 and area under the precision–recall curve of 0.2–0.4, numbers competitive with many other predictive models that have been reported for disease gene classification (van Driel *et al.*, 2005; Radivojac *et al.*, 2008; Huang *et al.*, 2010; Sun *et al.*, 2011). Interestingly, the gender-specific models slightly outperformed the all-gender model, confirming that while there is a shared molecular basis for infertility in both genders (e.g. defects in meiosis), there are also unique pathways that contribute to fertility in each gender.

While we used a standard cutoff of 5% FPR for all models to identify candidate infertility genes, the two measures of model performance we used were the sensitivity and precision. A high sensitivity means that the model was able to identify most of the known fertility genes among its identified candidate genes, providing confidence that the model is reasonably comprehensive. A high precision means that there are more known fertility genes than non-fertility genes among all the predicted candidate genes, indicating that any given candidate gene is less likely to be a false positive.

Sensitivity was negatively correlated with the size of the positive gene training set (Fig. 3). This could be due, in part, to the method we use to obtain the smaller positive gene training sets, picking genes involved in a certain phenotype and thus similar pathways, making other genes in the small set of pathways easier to identify. However, since the negative gene training set is much larger, it results in the unfortunate side effect of lower precision with shrinking positive training gene set sizes. The false negatives can be attributed to genes that affect fertility by external mechanisms (e.g. insulin reduces fertility by causing diabetes) or genes that have few other genes annotated in their pathways. The false positives are most likely caused by noisy data such as spurious in vitro PPI with little biological function in vivo.

Even with the trade-off between precision and accuracy, we found that using the predicted infertility genes helped improve the odds ratio of cases vs. controls in the human infertility studies (Table 2). This increase in the odds ratios show that our predictions are enriched for true infertility genes relative to a random selection of genes and that the gender-specific model predictions provide the best enrichment.

To highlight how our gene predictions may be used to arbitrate rare CNV association results, we chose the infertility genes that were deleted multiple times in the infertile patients across the case–control studies (Table 3). One such candidate is *DMRT1*, which was deleted in four different patients and is known to affect post-natal testis differentiation in mice (Raymond *et al.*, 2000). We have previously shown that deletions of *DMRT1* are reproducibly associated with spermatogenic impairment in humans (Lopes *et al.*, 2013). We also found the prolactin gene (*PRL*) deleted in two primary ovarian insufficiency patients. Female knockout mice lacking *PRL* are also infertile (Bole-Feysot *et al.*, 1998), and hypoprolactinemia is associated with ovarian dysfunction in humans (Kauppila *et al.*, 1988). Finally, the last candidate gene that we highlight is pregnancy-specific glycoprotein 5 (*PSG5*), which was deleted in two different azoospermic men. This gene is not very well studied, and its homolog in mice (*PSG22*) does not have a knockout line. *PSG5* is expressed at high levels in the testis and is closely related to *PSG1*, which is highly expressed in the placenta. The PSG gene family is prone to recurrent CNV, and based on the low male model chi-squared score (0.0008) we suggest that these two case-specific deletions do not influence male fertility. In addition to these recurrent events, we also provide a list of case-specific singleton gene deletions highlighted by this approach (Table S31).

This example shows application of our human gene predictions, providing a basis for prioritizing large candidate gene lists produced in GWAS for further experiments. In our own data, we have seen how these predictions can be used to hone in on one specific gene when investigating many genes deleted by a large CNVs (Fig. S22). Given that our available case–control studies' genetic data were low resolution, this limited our analysis to large deletions. With the commonplace use of exome sequencing for studies like this in the future it is likely that more of our predicted genes will be implicated.

Functional genomic analysis of mammalian germ cells has historically been limited by the difficulty in working with mammalian gonadal tissue. These tissues are complex cellular mixtures, and large-scale isolation of specific cell types has been a rate-limiting step, especially from ovaries. A primary barrier to follow-up of our large prediction sets is to apply a complementary high-throughput experimental system to test these predictions. To this end, we have been developing a method to perform multiplex shRNA screening directly in mammalian testis, and are in the process of using this to test over a hundred of our top candidates reported here. Our hope is that by tying together high-dimensional computational analysis of mammalian germ cells with novel high-throughput genomic assays in these same cells, we can help usher in a new era of functional genomics for mammalian reproductive biology.

ETHICS STATEMENT

Access and analysis of human genetic data performed in this study were reviewed and approved by the Washington University in St. Louis IRB, under protocols #201109261 and #201107177.

ACKNOWLEDGEMENTS

The authors thank Katinka Vigh-Conrad for assistance with figure preparation and Brad Cairns and Susan Hammoud for sharing their datasets prior to publication. This research was supported by HD078641 from the Eunice Kennedy Shriver National Institute of Child Health and Development and the National Science Scholarship (PhD) training grant from the Agency for Science Technology and Research (A*STAR) of Singapore. This study makes use of data from the Women's Health Initiative, obtained from the Database of Genotypes and Phenotypes (accession numbers phs000200.v9.p3.c1 and phs000200.v9.p3.c2). These are available from http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap.

AUTHOR CONTRIBUTIONS

NRYH, NH and DC designed the study, analyzed data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing financial interests

REFERENCES

Ballouz S, Liu JY, George RA, Bains N, Liu A, Oti M, Gaeta B, Fatkin D & Wouters MA (2013) Gentrepid V2.0: a web server for candidate disease gene prediction. *BMC Bioinformatics* 14, 249.

Bole-Feysot C, Goffin V, Edery M, Binart N & Kelly PA. (1998) Prolactin (PRL) and its receptor: actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocr Rev* 19, 225–268.

Centers for Disease Control and Prevention. (2014) *National Public Health Action Plan for the Detection, Prevention, and Management of Infertility*. Centers for Disease Control and Prevention, Atlanta, Georgia; June 2014.

Consortium E. P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG & Vriend G (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 33, W758–W761.

Enright AJ, Van Dongen S & Ouzounis CA. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575–1584.

Evian Annual Reproduction (EVAR) Workshop Group 2010, Fauser BC, Diedrich K, Bouchard P, Dominguez F, Matzuk M, Franks S, Hamamah S, Simón C, Devroey P, Ezcurra D & Howles CM (2011) Contemporary genetic technologies and female reproduction. *Hum Reprod Update* 17, 829–847.

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C & Jensen LJ (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41, D808–D815.

Hamada AJ, Esteves SC & Agarwal A. (2013) A comprehensive review of genetics and genetic testing in azoospermia. *Clinics (Sao Paulo)* 68 (Suppl. 1), 39–60.

Hammoud SS, Low DH, Yi C, Carrell DT, Guccione E & Cairns BR (2014) Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell* 15, 239–253.

Hotaling J & Carrell DT. (2014) Clinical genetic testing for male factor infertility: current applications and future directions. *Andrology* 2, 339–350.

Huang N, Lee I, Marcotte EM & Hurles ME. (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 6, e1001154.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E & Stein L (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33, D428–D432.

Kauppila A, Martikainen H, Puistola U, Reinila M & Ronnberg L. (1988) Hypoprolactinemia and ovarian function. *Fertil Steril* 49, 437–441.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R & Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14, R36.

Kosova G, Abney M & Ober C. (2010) Colloquium papers: Heritability of reproductive fitness traits in a human population. *Proc Natl Acad Sci USA* 107(Suppl. 1), 1772–1778.

Kosova G, Scott NM, Niederberger C, Prins GS & Ober C. (2012) Genome-wide association study identifies candidate genes for male fertility traits in humans. *Am J Hum Genet* 90, 950–961.

Krausz C, Hoefsloot L, Simoni M, Tuttelmann F, European Academy of Andrology & European Molecular Genetics Quality Network (2014) EAA/EMQN best practice guidelines for molecular diagnosis of Y-chromosomal microdeletions: state-of-the-art 2013. *Andrology* 2, 5–19.

Lopes AM, Aston KI, Thompson E, Carvalho F, Goncalves J, Huang N, Matthiesen R, Noordam MJ, Quintela I, Ramu A, Seabra C, Wilfert AB, Dai J, Downie JM, Fernandes S, Guo X, Sha J, Amorim A, Barros A, Carracedo A, Hu Z, Hurles ME, Moskovtsev S, Ober C, Paduch DA, Schiffman JD, Schlegel PN, Sousa M, Carrell MT & Conrad DF (2013) Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genet* 9, e1003349.

Matzuk MM & Lamb DJ. (2008) The biology of infertility: research advances and clinical challenges. *Nat Med* 14, 1197–1213.

Meschede D, Lemcke B, Behre HM, De Geyter C, Nieschlag E & Horst J (2000) Clustering of male infertility in the families of couples treated with intracytoplasmic sperm injection. *Hum Reprod* 15, 1604–1608.

Navarro-Costa P, Plancha CE & Goncalves J. (2010) Genetic dissection of the AZF regions of the human Y chromosome: thriller or filler for male (in)fertility? *J Biomed Biotechnol* 2010, 936569.

O'Flynn O'Brien KL, Varghese AC & Agarwal A (2010) The genetic causes of male factor infertility: a review. *Fertil Steril* 93, 1–12.

Prasad TS, Kandasamy K & Pandey A. (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol* 577, 67–79.

Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM & Mooney SD (2008) An integrated approach to inferring gene-disease associations in humans. *Proteins* 72, 1030–1037.

Raymond CS, Murphy MW, O'Sullivan MG, Bardwell VJ & Zarkower D. (2000) Dmrt1, a gene related to worm and fly sexual regulators, is required for mammalian testis differentiation. *Genes Dev* 14, 2587–2595.

Rogers S & Girolami M. (2011) *A First Course in Machine Learning*. Chapman and Hall/CRC, Boca Raton.

Singh-Blom UM, Natarajan N, Tewari A, Woods JO, Dhillon IS & Marcotte EM (2013) Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE* 8, e58977.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, Mikkelsen TS & Kaessmann H (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* 3, 2179–2190.

Sun PG, Gao L & Han S. (2011) Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci* 7, 61–73.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL & Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31, 46–53.

Vanunu O, Magger O, Ruppin E, Shlomi T & Sharan R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 6, e1000641.

Xu M, Qin Y, Qu J, Lu C, Wang Y, Wu W, Song L, Wang S, Chen F, Shen H, Sha J, Hu Z, Xia Y & Wang X (2013) Evaluation of five candidate genes from GWAS for association with oligozoospermia in a Han Chinese population. *PLoS ONE* 8, e80374.

Xue Z, Huang K, Cai C, Cai L, Jiang CY, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu JY, Horvath S & Fan G (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593–597.

Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I & Tora L (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* 39, e35.

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, Sandstrom R, Ma Z, Davis C, Pope BD, Shen Y, Pervouchine DD, Djebali S, Thurman RE, Kaul R, Rynes E, Kirilusha A, Marinov GK, Williams BA, Trout D,

Amrhein H, Fisher-Aylor K, Antoshechkin I, DeSalvo G, See LH, Fastuca M, Drenkow J, Zaleski C, Dobin A, Prieto P, Lagarde J, Bussotti G, Tanzer A, Denas O, Li K, Bender MA, Zhang M, Byron R, Groudine MT, McCleary D, Pham L, Ye Z, Kuan S, Edsall L, Wu YC, Rasmussen MD, Bansal MS, Kellis M, Keller CA, Morrissey CS, Mishra T, Jain D, Dogan N, Harris RS, Cayting P, Kawli T, Boyle AP, Euskirchen G, Kundaje A, Lin S, Lin Y, Jansen C, Malladi VS, Cline MS, Erickson DT, Kirkup VM, Learned K, Sloan CA, Rosenbloom KR, Lacerda de Sousa B, Beal K, Pignatelli M, Flicek P, Lian J, Kahveci T, Lee D, Kent WJ, Ramalho Santos M, Herrero J, Notredame C, Johnson A, Vong S, Lee K, Bates D, Neri F, Diegel M, Canfield T, Sabo PJ, Wilken MS, Reh TA, Giste E, Shafer A, Kutyavin T, Haugen E, Dunn D, Reynolds AP, Neph S, Humbert R, Hansen RS, De Bruijn M, Selleri L & Rudensky A, Josefowicz S, Samstein R, Eichler EE, Orkin SH, Levasseur D, Papayannopoulou T, Chang KH, Skoultchi A, Gosh S, Disteche C, Treuting P, Wang Y, Weiss MJ, Blobel GA, Cao X, Zhong S, Wang T, Good PJ, Lowdon RF, Adams LB, Zhou XQ, Pazin MJ, Feingold EA, Wold B, Taylor J, Mortazavi A, Weissman SM, Stamatoyannopoulos JA, Snyder MP, Guigo R, Gingeras TR, Gilbert DM, Hardison RC, Beer MA, Ren B & Mouse ENCODE Consortium (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355–364.

Zorrilla M & Yatsenko AN (2013) The genetics of infertility: current status of the field. *Curr Genet Med Rep* 1, 247–260.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1. Data distribution of all features for MGI reproductive training set in the mouse genome.
Figure S2. Data distribution of all features for MGI male reproductive training set in the mouse genome.
Figure S3. Data distribution of all features for MGI female reproductive training set in the mouse genome.
Figure S4. Data distribution of all features for reproductive training set in the human genome.
Figure S5. Data distribution of all features for male reproductive training set in the human genome.
Figure S6. Data distribution of all features for female reproductive training set in the human genome.
Figure S7. Data distribution of all features for non-obstructive azoospermia training set in the human genome.
Figure S8. Model performance benchmarks without PPI.
Figure S9. Data distribution of all features for MGI abnormal female meiosis reproductive training set in the mouse genome.
Figure S10. Data distribution of all features for MGI abnormal endometrium morphology training set in the mouse genome.
Figure S11. Data distribution of all features for MGI decreased oocyte number reproductive training set in the mouse genome.
Figure S12. Data distribution of all features for MGI abnormal ovulation cycle reproductive training set in the mouse genome.
Figure S13. Data distribution of all features for MGI abnormal ovulation reproductive training set in the mouse genome.
Figure S14. Data distribution of all features for MGI male meiosis arrest reproductive training set in the mouse genome.
Figure S15. Data distribution of all features for MGI azoospermia reproductive training set in the mouse genome.
Figure S16. Data distribution of all features for MGI oligozoospermia reproductive training set in the mouse genome.
Figure S17. Data distribution of all features for MGI male germ cell apoptosis reproductive training set in the mouse genome.
Figure S18. Data distribution of all features for MGI abnormal spermiogenesis reproductive training set in the mouse genome.
Figure S19. Data distribution of all features for MGI sperm physiology reproductive training set in the mouse genome.
Figure S20. Data distribution of all features for MGI teratozoospermia reproductive training set in the mouse genome.
Figure S21. Model performance benchmarks for mouse infertility subset phenotypes.
Figure S22. Examples of finding candidate infertility genes in patient CNVs.
Table S1. MGI reproductive mouse gene scores.
Table S2. MGI male reproductive mouse gene scores.
Table S3. MGI female reproductive mouse gene scores.
Table S4. Human fertility gene scores.
Table S5. Human male fertility gene scores.
Table S6. Human female fertility gene scores.
Table S7. Human non-obstructive azoospermia gene scores.
Table S8. MGI male meiosis arrest mouse gene scores.
Table S9. MGI abnormal female meiosis mouse gene scores.
Table S10. MGI negative genes.
Table S11. MGI survival genes.
Table S12. MGI reproductive genes.
Table S13. MGI male reproductive genes.
Table S14. MGI female reproductive genes.
Table S15. Human infertility genes.
Table S16. Human male infertility genes.
Table S17. Human female infertility genes.
Table S18. Human non-obstructive azoospermia infertility genes.
Table S19. MGI abnormal endometrium morphology genes.
Table S20. MGI abnormal female meiosis genes.
Table S21. MGI abnormal ovulation cycle genes.
Table S22. MGI abnormal ovulation genes.
Table S23. MGI abnormal spermiogenesis genes.
Table S24. MGI azoospermia genes.
Table S25. MGI deceased oocyte number genes.
Table S26. MGI male germ cell apoptosis genes.
Table S27. MGI male meiosis arrest genes.
Table S28. MGI oligozoospermia genes.
Table S29. MGI sperm physiology genes.
Table S30. MGI teratozoospermia genes.
Table S31. Case–control study predicted infertility genes.