Published in final edited form as:

Andrology. 2015 November ; 3(6): 1140–1149. doi:10.1111/andr.12109.

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

Nicholas Rui Yuan Ho, Ni Huang, and Donald F. Conrad*

Department of Genetics, Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA

Abstract

Despite its great potential, high-throughput functional genomic data is 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 genomewide 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 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, $p < 1.64 \times 10^{-8}$ versus 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 paper.

Keywords

testis; ovary; machine learning; fertility genes; systems biology

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 States¹.

*Corresponding author: Dr. Donald F. Conrad, Department of Genetics, Washington University School of Medicine, Campus Box 8232, St. Louis MO, 63110, USA; Email: dconrad@genetics.wustl.edu

Ethics Statement

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

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). There 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

Infertility is also highly heritable, with heritability estimates ranging from 0.16 to 0.81, with a mean of around 0.3². 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 population³, and up to 10% of cases of azoospermia are clinically attributable to Y chromosome microdeletions in typical populations of European ancestry^{4,5}. 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 implantation⁶. In total, the data suggests 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 genomewide association studies and pedigree studies, has provided definitive lists of genes involved in human gonadal function, but due to 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^{7–11}. 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 genomewide genomic data being generated on hundreds of human and mouse cell types, including germ cells, much from the Encyclopedia of DNA Elements Consortium (ENCODE)¹². Gene expression, histone modifications, and methylation have all been assayed on bulk gonadal tissue, and in some cases, purified germ cells. We hypothesize that since 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”¹³, 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^{14,15} and various tools have been made which use single features to define similarity amongst genes, ranging from disease ontology terms to protein structure and tissue-specific expression^{16–19}.

Since the genomic feature information fed to the model greatly influences the results, we had to build our own tool since 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 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 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 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 genomewide association studies 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 S19–30).

MGI'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 3,344 genes in mouse to 3,406 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^{7,8,20,21}. 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—ENCODE^{12,22} paired-end RNA-Seq reads were used for differential tissue expression. Mouse liver (ENCSR000AJU & ENCSR216KLZ), heart (ENCBS441FDF & ENCSR000BYQ), testis (ENCSR266ESZ & ENCSR000BYW) and ovary (ENCSR516UNF & ENCSR000BZC) were used. Human liver (ENCSR085HNI & ENCSR000AEU), heart (ENCSR000AHH & 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)²³ and Hammoud (GSE49624)²⁴. Mouse and human oocyte RNA-Seq paired end reads were taken from Xue (GSE44183)²⁵.

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 tophat²⁶

with the default values and gene expression was summarized using cuffnorm²⁷ 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 seqminer²⁸ to map the reads + -5kb 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 (PPI) were collected from HPRD²⁹, Reactome³⁰, and STRING³¹ 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 MCL³². Sum of weight of edges were calculated as a measure of proximity to a group of 'seed' genes as described previously³³. 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 9 of those subsets, leaving 1 subset for testing. We then plot the false positive rate 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 false positive rate against the false negative rate of the 10 models.

Human Infertility Gene Deletion Analysis

We obtained existing Copy Number Variant (CNV) calls from men assayed in our previous study of spermatogenic impairment³⁴. 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 2x2 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 3 genes that occurred more than once were discussed in the paper 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 genes²⁸. Since 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, protein-protein interactions (PPI), but only picked the best 3–4 to actually make any given model prediction.

There are two inputs for an LDA model. Apart 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 (Figure 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 ten 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, 0.741 for the male specific model, and 0.738 for the female specific model for the 15,212 genes tested in the mouse genome (Figure 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. Based on 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 (**Methods**).

The subcategory models all had better ROC curve AUC than the more general infertility models, but their precision recall curve AUCs were not as good (Figure 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 (Figures 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 (Figure 5).

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^{7,20,21,35} to generate four positive training sets (**Methods**). 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 (Figure 2).

There were fewer human genomic features available compared to mouse (Figures 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 (Figure 5).

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% false positive rate (FPR). (Table 1) This created shortlists of candidate infertility genes numbering between 400 – 900 genes for the mouse genome and between 590 – 1030 genes for the human genome (depending on the phenotype). We provide, as supplemental information, results of our gene classification framework for 9 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 genomewide association studies is the identification of true disease-associated variation amongst 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, previously used for genomewide association studies (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 males cases and the female human model 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 2 patients in primary ovarian insufficiency. Eight of our cases of azoospermia had deletions covering known male fertility genes (*CDY1*, *CDY1B*, *DAZ1*, *DAZ2*, *DAZ3*, *DDX3Y*,

USP9Y), while 4 of them had deletions covering *DMRT1* (2 of them also covered *FOXD4*). The last two patients had deletions covering *PSG5*. For our female cases with primary ovarian insufficiency, we found 2 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-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, 6 of the 8 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 (Figure 4). 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^{14,17,18,33}. 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.

Since we used a standard cutoff of 5% false positive rate 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 to the size of the positive gene training set (Figure 4). 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* protein-protein interactions 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 versus 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 4 different patients and is known to affect post-natal testis differentiation in mice³⁶. We have previously shown that deletions of *DMRT1* are reproducibly associated with spermatogenic impairment in humans³⁷. We also found the prolactin gene (*PRL*) deleted in two primary ovarian insufficiency patients. Female knockout mice lacking *PRL* are also infertile³⁸, and hypoprolactinemia is associated with ovarian dysfunction in humans³⁹. Finally, the last candidate gene that we highlight is pregnancy-specific glycoprotein 5 (*PSG5*), which was deleted in 2 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 copy number variation, and based on the low male-model chi 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 (Figure S22). Given that our

available case-control studies' genetic data was 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 of 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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.

References

1. Centers for Disease Control and Prevention. National Public Health Action Plan for the Detection, Prevention, and Management of Infertility. Centers for Disease Control and Prevention; 2014.
2. Kosova G, Abney M, Ober C. Colloquium papers: Heritability of reproductive fitness traits in a human population. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107(Suppl 1):1772–1778. DOI: 10.1073/pnas.0906196106 [PubMed: 19822755]
3. Meschede D, et al. Clustering of male infertility in the families of couples treated with intracytoplasmic sperm injection. *Hum Reprod*. 2000; 15:1604–1608. [PubMed: 10875874]
4. Krausz C, et al. EAA/EMQN best practice guidelines for molecular diagnosis of Y-chromosomal microdeletions: state-of-the-art 2013. *Andrology*. 2014; 2:5–19. DOI: 10.1111/j.2047-2927.2013.00173.x [PubMed: 24357628]
5. Hotaling J, Carrell DT. Clinical genetic testing for male factor infertility: current applications and future directions. *Andrology*. 2014; 2:339–350. DOI: 10.1111/j.2047-2927.2014.00200.x [PubMed: 24711280]
6. Matzuk MM, Lamb DJ. The biology of infertility: research advances and clinical challenges. *Nat Med*. 2008; 14:1197–1213. DOI: 10.1038/nm.f.1895 [PubMed: 18989307]
7. Zorrilla M, Yatsenko AN. The Genetics of Infertility: Current Status of the Field. *Curr Genet Med Rep*. 2013:1.
8. Evian Annual Reproduction Workshop G et al. Contemporary genetic technologies and female reproduction. *Hum Reprod Update*. 2011; 17:829–847. DOI: 10.1093/humupd/dmr033 [PubMed: 21896560]
9. Xu M, et al. Evaluation of five candidate genes from GWAS for association with oligozoospermia in a Han Chinese population. *PLoS One*. 2013; 8:e80374. [PubMed: 24303009]

10. Kosova G, Scott NM, Niederberger C, Prins GS, Ober C. Genome-wide association study identifies candidate genes for male fertility traits in humans. *Am J Hum Genet.* 2012; 90:950–961. DOI: 10.1016/j.ajhg.2012.04.016 [PubMed: 22633400]
11. O'Flynn O'Brien KL, Varghese AC, Agarwal A. The genetic causes of male factor infertility: a review. *Fertil Steril.* 2010; 93:1–12. DOI: 10.1016/j.fertnstert.2009.10.045 [PubMed: 20103481]
12. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012; 489:57–74. DOI: 10.1038/nature11247 [PubMed: 22955616]
13. Rogers, S.; Girolami, M. A first course in machine learning. CRC Press; 2012.
14. Sun PG, Gao L, Han S. Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci.* 2011; 7:61–73. [PubMed: 21278917]
15. Singh-Blom UM, et al. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS One.* 2013; 8:e58977. [PubMed: 23650495]
16. Ballouz S, et al. Gentrepid V2.0: a web server for candidate disease gene prediction. *BMC Bioinformatics.* 2013; 14:249. [PubMed: 23947436]
17. Radivojac P, et al. An integrated approach to inferring gene-disease associations in humans. *Proteins.* 2008; 72:1030–1037. DOI: 10.1002/prot.21989 [PubMed: 18300252]
18. van Driel MA, et al. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.* 2005; 33:W758–761. DOI: 10.1093/nar/gki435 [PubMed: 15980578]
19. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol.* 2010; 6:e1000641. [PubMed: 20090828]
20. Hamada AJ, Esteves SC, Agarwal A. A comprehensive review of genetics and genetic testing in azoospermia. *Clinics (Sao Paulo).* 2013; 68(Suppl 1):39–60. [PubMed: 23503954]
21. Navarro-Costa P, Plancha CE, Goncalves J. Genetic dissection of the AZF regions of the human Y chromosome: thriller or filler for male (in)fertility? *J Biomed Biotechnol.* 2010; 2010:936569. [PubMed: 20671934]
22. Yue F, et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature.* 2014; 515:355–364. DOI: 10.1038/nature13992 [PubMed: 25409824]
23. Soumillon M, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* 2013; 3:2179–2190. DOI: 10.1016/j.celrep.2013.05.031 [PubMed: 23791531]
24. Hammoud SS, et al. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell.* 2014; 15:239–253. DOI: 10.1016/j.stem.2014.04.006 [PubMed: 24835570]
25. Xue Z, et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature.* 2013; 500:593–597. DOI: 10.1038/nature12364 [PubMed: 23892778]
26. Kim D, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013; 14:R36. [PubMed: 23618408]
27. Trapnell C, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 2013; 31:46–53. DOI: 10.1038/nbt.2450 [PubMed: 23222703]
28. Ye T, et al. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* 2011; 39:e35. [PubMed: 21177645]
29. Prasad TS, Kandasamy K, Pandey A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol.* 2009; 577:67–79. DOI: 10.1007/978-1-60761-232-2_6 [PubMed: 19718509]
30. Joshi-Tope G, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005; 33:D428–432. DOI: 10.1093/nar/gki072 [PubMed: 15608231]
31. Franceschini A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013; 41:D808–815. DOI: 10.1093/nar/gks1094 [PubMed: 23203871]
32. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002; 30:1575–1584. [PubMed: 11917018]

33. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* 2010; 6:e1001154. [PubMed: 20976243]
34. Lopes AM, et al. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLOS Genetics.* 2013; 9:e1003349. [PubMed: 23555275]
35. Fauser BCJM, et al. Contemporary genetic technologies and female reproduction. *Human Reproduction Update.* 2011; 17:829–847. DOI: 10.1093/humupd/dmr033 [PubMed: 21896560]
36. Raymond CS, Murphy MW, O'Sullivan MG, Bardwell VJ, Zarkower D. Dmrt1, a gene related to worm and fly sexual regulators, is required for mammalian testis differentiation. *Genes Dev.* 2000; 14:2587–2595. [PubMed: 11040213]
37. Lopes AM, et al. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genet.* 2013; 9:e1003349. [PubMed: 23555275]
38. Bole-Feysot C, Goffin V, Edery M, Binart N, Kelly PA. Prolactin (PRL) and its receptor: actions, signal transduction pathways and phenotypes observed in PRL receptor knockout mice. *Endocr Rev.* 1998; 19:225–268. DOI: 10.1210/edrv.19.3.0334 [PubMed: 9626554]
39. Kauppila A, Martikainen H, Puistola U, Reinila M, Ronnberg L. Hypoprolactinemia and ovarian function. *Fertil Steril.* 1988; 49:437–441. [PubMed: 3342895]

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, and used these to annotate all genes in the mouse and human genomes, respectively. Using extensive phenotyping data from Jackson Labs Mouse Genome Informatics (MGI), 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. 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 (ROC) 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 MGI null genes.

Figure 3. Model performance vs training set size

For each classification model, we plotted the Precision-recall AUC (AUPRC) versus 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.

Figure 4. 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 top 8 most informative features in our study, we show the relative importance of each feature to the performance of 7 male (blue) and 5 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.

Figure 5. Selected feature Importance to each model

These show the relative importance of the genomic features used to construct the nine models presented towards 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 9 reproductive gene classifiers

<i>LDA Model</i>	<i>Predictive Score cutoff</i>	<i>False Positive Rate</i>	<i>Sensitivity</i>	<i>Precision</i>	<i>Positive Training Set Size</i>	<i>Negative Training Set Size</i>	<i># Candidate Genes</i>
MGI reproductive set on mouse genome	0.4906	0.05	0.211	0.557	999	3,344	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	3,406	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

This shows the benchmarks using different cutoffs for the Chi 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.

Table 2

Tests of association between gene-disrupting CNVs and infertility.

<i>All Gender model Scores</i>			<u>Fisher's Exact Test</u>	
	positive	negative	p-value	
case	313	1558	odds ratio	1.64×10^{-8}
				1.475309
control	1792	13160	95% confidence interval	$1.289827 - 1.683816$
<i>Gender specific model scores</i>			<u>Fisher's Exact Test</u>	
	positive	negative	p-value	
case	642	1229	odds ratio	2.2×10^{-16}
				2.307403
control	2760	12192	95% confidence interval	$2.075971 - 2.563118$
<i>Deletion CNVs spanning genes</i>			<u>Fisher's Exact Test</u>	
	positive	negative	p-value	
case	953	1005	odds ratio	3.663×10^{-6}
				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 copy number CNVs that span exons

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

Gene	Recurrence	Cohort	All gender chi score	Female chi score	Male chi score
<i>CDY1</i>	2	azo0	0.999999	5.03E-03	1
<i>CDY1B</i>		azo0	0.999998	1.36E-04	1
<i>DAZ1</i>		azo0	1	5.75E-03	1
<i>DAZ2</i>		azo0	1	3.47E-02	1
<i>DAZ3</i>		azo0	0.125516	6.74E-07	0.99982
<i>DDX3Y</i>	5	azo0	0.999991	1.59E-04	1
<i>USP9Y</i>	6	azo0	1	8.81E-04	1
<i>DMRT1</i>	4	azo0	1	0.999999	1
<i>PSG5</i>	2	azo0 azo0	0.640456	0.877	0.000839
<i>PRL</i>	2	POI	1	1	1

Known infertility genes have their names bolded.