Spring 5-15-2016

Application of Genomic Technologies to Study Infertility

Nicholas Rui Yuan Ho
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Donald Conrad, Chair
Barak Cohen
Joseph Dougherty
John Edwards
Liang Ma

Application of Genomic Technologies to Study Infertility

by
Nicholas Rui Yuan Ho

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2016
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

This research was supported by the National Science Scholarship (PhD) training grant from the Agency for Science Technology and Research (A*STAR) of Singapore. We thank the Genome Technology Access Center in the Department of Genetics at Washington University School of Medicine for help with Illumina sequencing. The Center is partially supported by NCI Cancer Center Support Grant #P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant# UL1 TR000448 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

Nicholas Rui Yuan Ho

Washington University in St. Louis

May 2016

ABSTRACT OF THE DISSERTATION

Application of Genomic Technologies to Study Infertility

by

Nicholas Rui Yuan Ho

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2016

Donald Conrad, Chair

An estimated one in eight couples in the United States are diagnosed with infertility. There is a significant genetic contribution to infertility, with estimates of heritability ranging from 0.2 to 0.5. We know surprisingly little about the genetic causes, with only slightly more than a hundred genes known to cause human infertility. I have been translating recent advances in genomics to study infertility in a more efficient manner, in order to improve our knowledge of the genetic causes. By using high throughput genomics and proteomics datasets from other groups, I was able to feed that into a machine learning algorithm to predict novel fertility function genes. While not perfect, this computational model performs comparably to other publish prediction models. In order to test the top predicted fertility genes I also developed an experimental technique to simultaneously screen up to hundreds of genes for spermatogenesis function *in vivo* in mice. This method is based off of RNAi, and I was able to benchmark its performance to demonstrate that it performed comparably to other benchmarked RNAi screens in flies. I then used this method to test the top 26 predicted spermatogenesis genes and showed that most of them (24/26) have an important role in spermatogenesis. Using this technique, other groups can

screen genes for spermatogenesis function in a fraction of the time and cost compared to the traditional approach of generating knockout mouse lines. Finally, I describe the progress I have made in using genetic engineering to rescue spermatogenesis in mice. By analyzing the missteps I have made in delivering constitutively expressed transgenes and CRISPR genes into mouse testes, I describe the probably reasons for my failure and how to implement future experiments to get more success.

Introduction

Application of genomic technologies to study infertility

Nicholas Rui Yuan Ho

Infertility is defined as the inability of a couple to achieve pregnancy after a year or more of regular unprotected sexual intercourse. This reproductive disease presents its phenotype as a couple, but the cause may be found in either partner. Approximately half of the cases of infertility are attributed to the male and half to the female partner in the couple.

In its 2014 infertility white paper, the CDC reported that 12-18% of couples and 9% of men are infertile in the United States[1]. Infertility is also highly heritable, with heritability estimates ranging from 0.16 to 0.81, with a mean of around 0.3[2]. 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[3], 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 information of male and female gonads and genitalia, neuroendocrine control of gonadal function, paracrine regulation of gamete development, fertilization and implantation[6]. In total, the data suggests there is a significant genetic component to infertility.

Given its high prevalence, surprisingly little is known about the genetic causes of this disease in humans. 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]. This is due to the traditional method of infertility gene ascertainment, phenotyping knockout mice, which is time-consuming, low-throughput, and expensive. Ongoing generation and analysis of mouse mutants from places like Knockout Mouse Project and various other investigators have slowly produced a larger list of gonad-essential genes, but this is still far from

comprehensive. As high-throughput DNA sequencing moves to the clinic, there will be a deluge of data generated about human infertility. However interpreting variations across the whole genome of infertile patients is a difficult problem at best and it will be almost impossible to verify all candidate infertility genes via traditional methods.

My research has been focused on increasing the efficiency of discovering candidate genes and verifying their function *in-vivo*. To help identify infertility genes in patients, I will show that I can use available high throughput data about human genes that have been generated by other groups to come up with a set of infertility candidate genes based on co-regulation, expression and protein-protein interactions. I will also demonstrate a new experimental method which can be used to screen a panel of genes *in-vivo* in mouse testis. This method can be used as a primary screen for male infertility genes that are identified in humans, reducing the cost and speeding up the verification process. Finally I discuss the work I have performed using genetic engineering to attempt to rescue spermatogenesis in mice.

References

1. National Public Health Action Plan for the Detection, Prevention and Management of Infertility.
2. Kosova, G., Abney, M. & Ober, C. Heritability of reproductive fitness traits in a human population. Proc. Natl. Acad. Sci. 107, 1772–1778 (2010).
3. Meschede, D. et al. Clustering of male infertility in the families of couples treated with intracytoplasmic sperm injection. Hum. Reprod. Oxf. Engl. 15, 1604–1608 (2000).
4. Krausz, C., Hoefsloot, L., Simoni, M. & Tüttelmann, F. EAA/EMQN best practice guidelines for molecular diagnosis of Y-chromosomal microdeletions: state-of-the-art 2013. Andrology 2, 5–19 (2014).
5. Hotaling, J. & Carrell, D. T. Clinical genetic testing for male factor infertility: current applications and future directions. Andrology 2, 339–350 (2014).
6. Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. Nat. Med. 14, 1197–1213 (2008).
7. Zorrilla, M. & Yatsenko, A. N. The Genetics of Infertility: Current Status of the Field. Curr. Genet. Med. Rep. 1, 247–260 (2013).
8. Fauser, B. C. J. M. et al. Contemporary genetic technologies and female reproduction. Hum. Reprod. Update 17, 829–847 (2011).
9. Xu, M. et al. Evaluation of Five Candidate Genes from GWAS for Association with Oligozoospermia in a Han Chinese Population. PLoS ONE 8, e80374 (2013).
10. Kosova, G., Scott, N. M., Niederberger, C., Prins, G. S. & Ober, C. Genome-wide association study identifies candidate genes for male fertility traits in humans. Am. J. Hum. Genet. 90, 950–961 (2012).
11. O'Flynn O'Brien, K. L., Varghese, A. C. & Agarwal, A. The genetic causes of male factor infertility: A review. Fertil. Steril. 93, 1–12 (2010).

Chapter 1

Computational Fertility Gene Candidate Identification

Nicholas Rui Yuan Ho, Ni Huang, Donald F Conrad

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

*Andrology **3,** 1140–1149 (2015).*

Chapter 1 - Introduction

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)[1,2]. Gene expression, histone modifications, and methylation have all been assayed on bulk gonadal tissue, and in some cases, purified germ cells. I hypothesized that since known fertility genes work in a small set of pathways, I can computationally identify genomic features among genomic high throughput data that distinguish "fertility genes". All genes in the genome with similar features are likely to be similarly regulated and are probably working in the same pathways. These genes are likely to also cause fertility problems when mutated.

To accomplish this I apply a technique known as "supervised learning", creating 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[3,4] 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[5–8].

Since the genomic feature information fed to the model greatly influences the results, I had to build my own tool since the existing tools can only use a small subset of the current high throughput data. I 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 I 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 which work in various pathways. To get around this issue, I used data from mouse knockout lines to augment my training datasets.

I first tested my approach on the better annotated 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. I was able to validate that my 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). I then applied my model to the human genome to classify genes by using the small set of known human fertility genes.

The main product of this 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. I show some uses of the 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 my 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, I was 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 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.

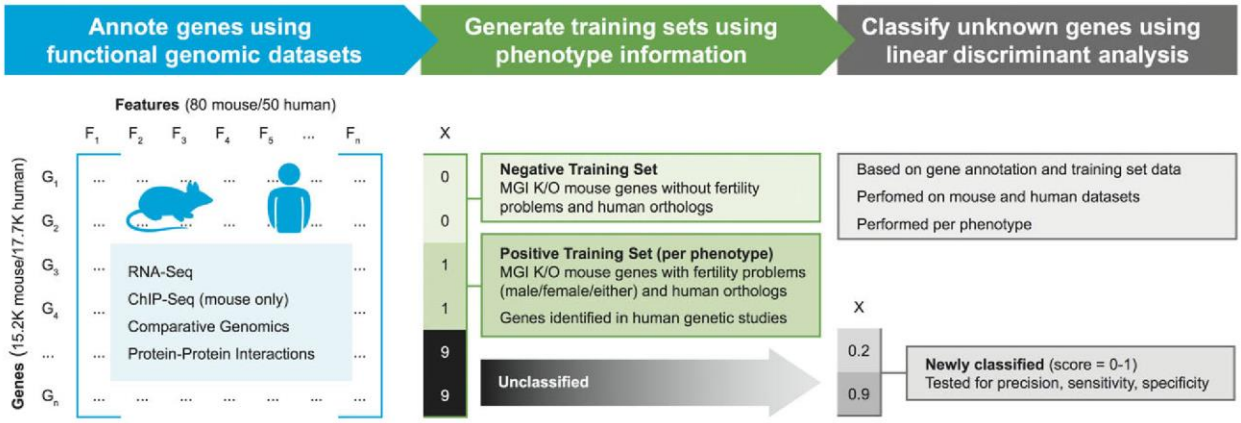


Figure 1.1 Overview of the study.
I set out to assess the utility of functional genomic data for predicting the identity of genes relevant to mammalian fertility using a machine learning approach. I obtained and reprocessed over 30 high-throughput functional genomic datasets from mouse and human, and used these to annotate all genes in the mouse and human genomes, respectively. Using extensive phenotyping data from Jackson Labs Mouse Genome Informatics (MGI), I generated a negative training set of genes which were highly unlikely to be related to mammalian fertility. I then created multiple positive gene training sets of genes known to disrupt mammalian fertility, identified in either mouse or human. I 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.

Chapter 1 - Results

I tried various modeling frameworks before settling on using Linear Discriminant Analysis (LDA) for my supervised learning classifier. LDA has worked well in the past to generically predict human haploinsufficient genes[9]. 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 me 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 I 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.

To ensure that the data generated by different experiments were comparable, I downloaded the raw sequencing reads for the ChIP-seq and RNA-Seq experiments and remapped and quantified them using the same pipeline. PPI scores were generated by determining proximity to different gene sets such as reproductive genes and cancer genes **(Methods).**

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, I performed my investigations with various gene sets (**Figure 1.1**). 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". I have

picked the models with the best results to present here, but I discuss all the models I tested 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 I 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

I 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 1.2**).

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 I reasoned that genes involved in narrowly defined processes should be easier to model and predict. Based on the phenotype observed in knockout mice, I picked 12 of them that had at least 50 different genes implicated. I then characterized these models on the mouse genome. **(Methods)**

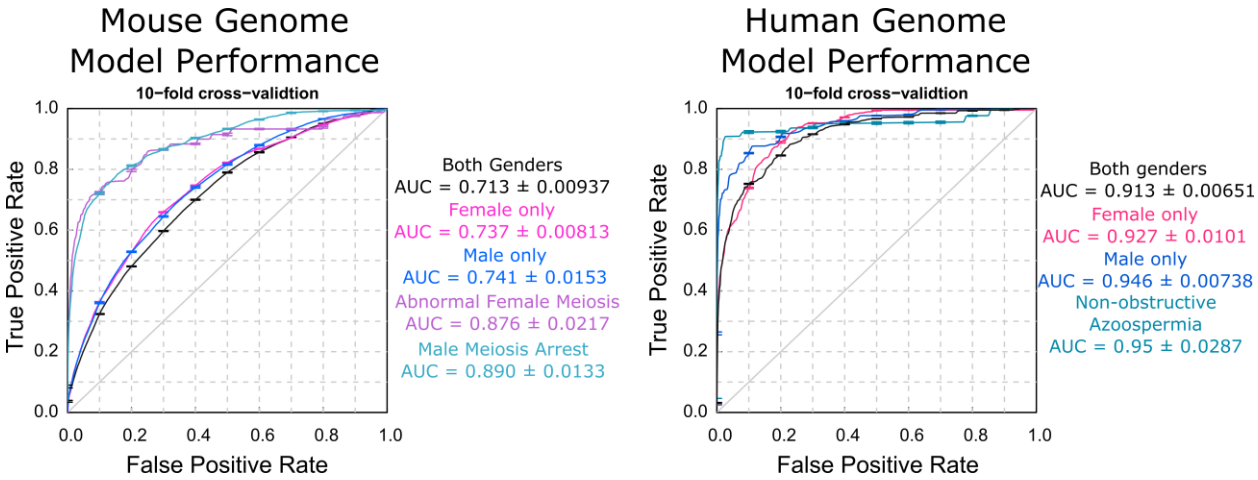


Figure 1.2: Model performance benchmarks for fertility gene sets

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.

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 1.3**). Among these models I 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 I tested **(Figures 1.S1-1.S3, 1.S9-1.S20)**, I found that PPI with 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 1.4)**.

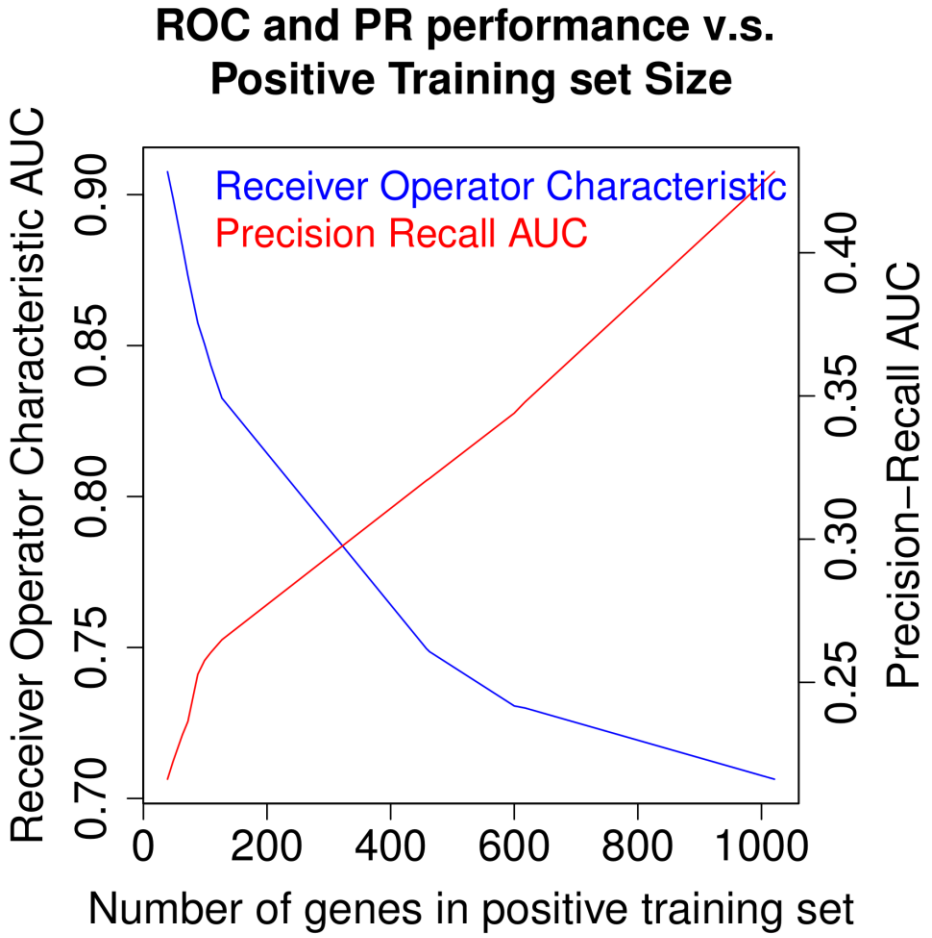


Figure 1.3: Model performance vs training set size
For each classification model, I plotted the Precision-recall AUC (AUPRC) versus Receiver Operator Characteristic AUC (AUROC) as two measures of model performance. I fit trend lines to each set of points using loess regression, with AUPRC in red and AUROC in blue. In general, AUROC decreases with increasing positive training set size, while AUPRC increases.

Human genetic studies' genes on human genome model

Because my approach was working for the large and small mouse derived training sets I reasoned that it may also work just as well with the smaller set of fertility genes implicated in humans. I combined male and female infertility genes that work in various pathways from

review articles 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, I translated the MGI null gene set into conserved human genes and used this as the negative training set.

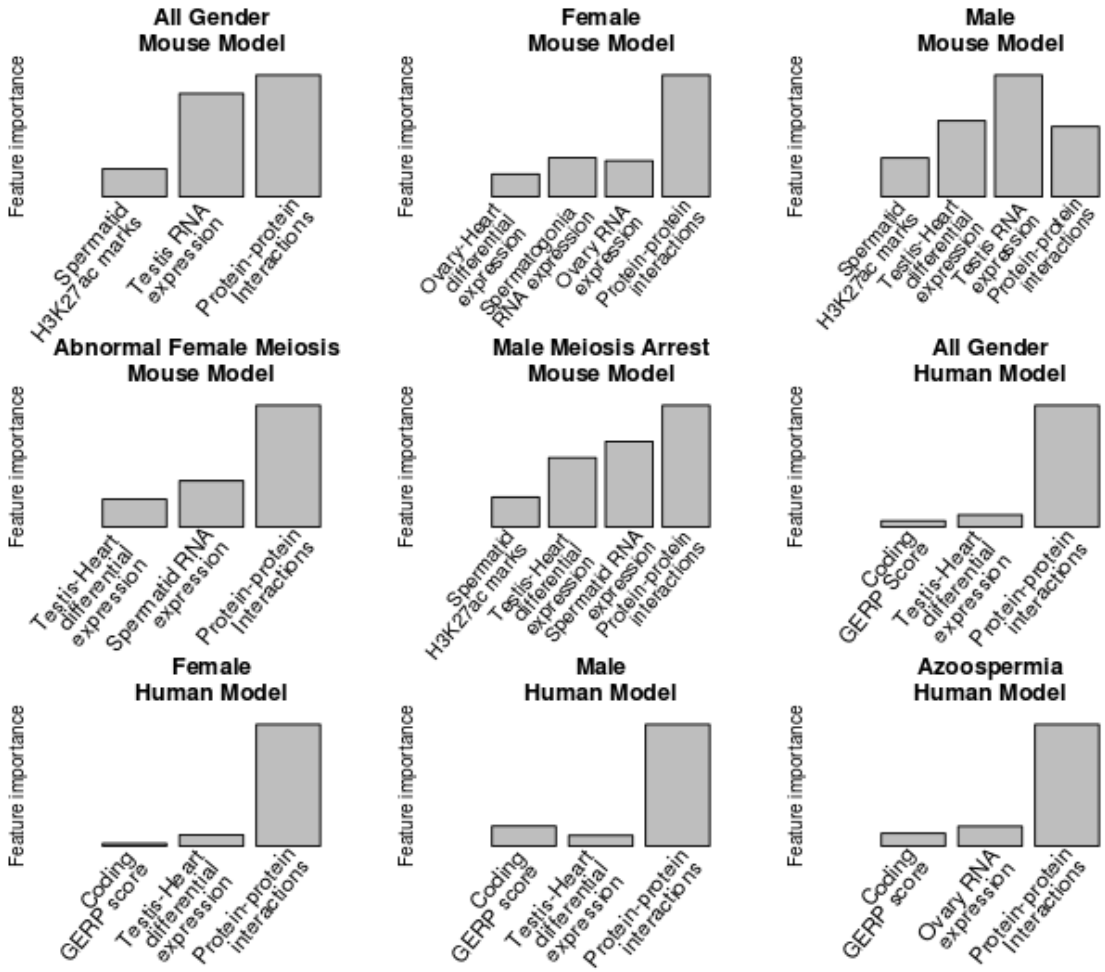


Figure 1.4: Selected feature Importance to each model
These show the relative importance of the genomic features used to construct the nine models presented towards the model predictions. I show a subset of all the features that I tested, presenting only the ones actually used in the model(s). All the other features are presented in the supplement.

I 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 1.2)**.

There were fewer human genomic features available compared to mouse **(Figures 1.S4 – 1.S7)**, and among the ones I 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 1.4)**.

LDA Model	*Predictive Score cutoff*	*False Positive Rate*	*Sensitivity*	*Precision*	*Positive Training Set Size*	*Negative Training Set Size*	*# Candidate Genes*
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

Table 1.1: Summary of prediction results for 9 reproductive gene classifiers
This shows the benchmarks using different cutoffs for the Chi score produced by the functional gene prediction models. The predictive score cutoff tries to get close to a 5% false positive rate. We tested 15,212 genes in the mouse genome and 17,758 genes in the human genome.

Model predictions

In order to produce a list of candidate genes likely to modulate fertility, I 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 1.1**) 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).

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. I sought to evaluate how my fertility gene predictions could be used to improve analysis of case-control data from cohort studies of gonadal function.

First I took data generated from several cohorts of male and female gonadal dysfunction, as well as matched controls, previously used for genomewide association studies (GWAS). I 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 I used my 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 predicted infertility genes for the female cases, I 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 1.2)**

I also looked at candidate genes that were deleted multiple times in either male or female cases alone, but not controls **(Table 1.3)**. This produced a list of 14 patients in azoospermia and 2 patients in primary ovarian insufficiency. Eight of the 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 the female cases with primary ovarian insufficiency, I found 2 patients with deletions covering PRL.

Chapter 1 - 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, I expect that disrupting the same pathway in different ways can produce correlated disease phenotypes.

Are the genomic features that were ultimately most informative for my 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 I observed were based on germ cell or gonad gene expression levels, and only one was based on an epigenetic mark (**Figure 1.5**). 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 I 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. I 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 getting high resolution RNA expression of various germline cell types and gonads will be the best way to improve fertility gene predictions in both humans and mice. Furthermore, it looks like the 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 it need to be repeated many times to get an accurate idea of the average cell expression.

I evaluated my 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[3,6,7,9]. 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.

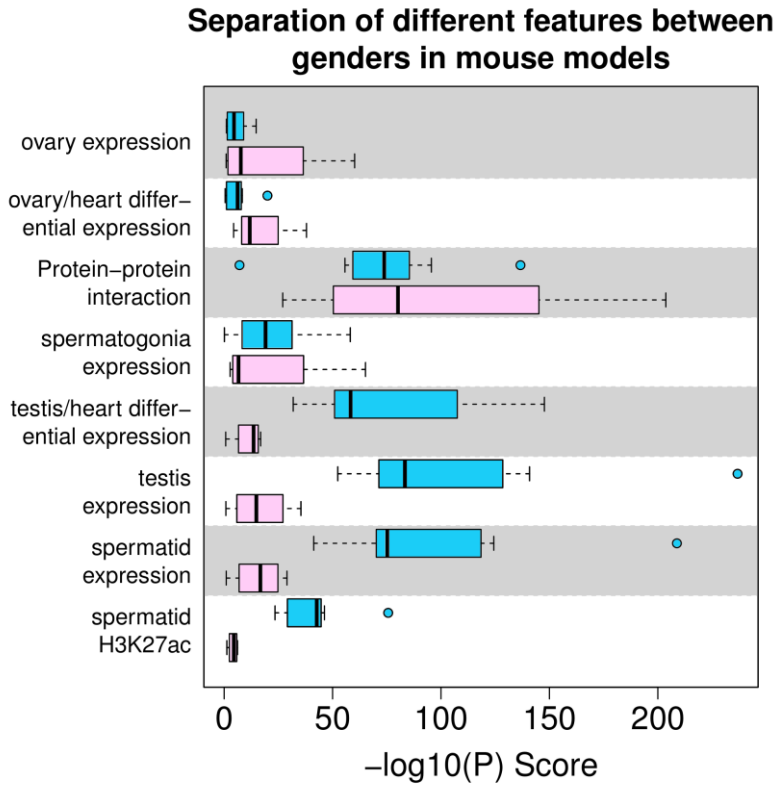


Figure 1.5: Functional annotations contain both general and sex-specific information.
Many of the functional annotations used to produce my classifiers are obtained from sex-specific germ cells. For the top 8 most informative features in my study, I show the relative importance of each feature to the 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.

Since I used a standard cutoff of 5% false positive rate for all models to identify candidate infertility genes, the two measures of model performance I 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, giving you 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, letting you trust 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 5)**. This could be due, in part, to the method I use to get 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, I found that using the predicted infertility genes helped improve the odds ratio of cases versus controls in the human infertility studies **(Table 1.2)**. This increase in the odds ratios show that my 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.

All-gender model scores			Fisher's exact test	
	Positive	Negative	*p*-value	1.64×10^{-8}
Case	313	1,558	Odds ratio	1.475309
Control	1,792	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	1,229	Odds ratio	2.307403
Control	2,760	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	1,005	Odds ratio	1.250934
Control	6,447	8,505	95% Confidence interval	1.136988 – 1.376255

Table 1.2: Tests of association between gene-disrupting CNVs and infertility.
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. The all gender model score uses the MGI reproductive set on human genome model score of the genes for both male and female cohorts. The gender specific model score uses the MGI male reproductive set on human genome model scores for the male cohort patients and the MGI female reproductive set on human genome model scores for the female cohort patients. 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

To highlight the best predictions, I chose the infertility genes that were deleted multiple times in the infertile patients across the case-control studies **(Table 1.3)**. One such candidate is DMRT1, which was deleted in 4 different patients and is known to affect post-natal testis differentiation in mice[10] and is associated with male infertility[11]. I also found PRL deleted in 2 primary ovarian insufficiency patients. Female knockout mice lacking PRL are also infertile (Males are fine)[12], and overexpression of this gene cause many detrimental effects in humans including female infertility[13]. Finally, the last candidate gene that I highlight is PSG5, which was found to be

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 created yet. PSG5 is expressed at high levels in the testis and is closely related to PSG1 which is highly expressed in the placenta. Further studies on this gene may show an important role in male fertility.

Gene	Recurrence	Cohort	All-gender chi-squared score	Female chi-squared score	Male chi-squared score
CDY1	2	Azoo	0.999999	5.03×10^{-3}	1
CDY1B		Azoo	0.999998	1.36×10^{-4}	1
DAZ1		Azoo	1	5.75×10^{-3}	1
DAZ2		Azoo	1	3.47×10^{-2}	1
DAZ3		Azoo	0.125516	6.74×10^{-7}	0.99982
DDX3Y	5	Azoo	0.999991	1.59×10^{-4}	1
USP9Y	6	Azoo	1	8.81×10^{-4}	1
DMRT1	4	Azoo	1	0.999999	1
PSG5	2	Azoo	0.640456	0.877	8.39×10^{-4}
PRL	2	POI	1	1	1

Table 1.3: Candidate genes identified multiple times in gene-disrupting CNVs for infertility cohorts.
Known infertility genes have their names bolded.

This example shows application of my human gene predictions, providing a basis for prioritizing large candidate gene lists produced in GWAS for further experiments. In my own data, I 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 1.S22)**. Given that my available case-control studies' genetic data was low resolution, this limited my analysis to large deletions. With the increasing commonplace use of exome sequencing for studies like this in the future it is likely that more of my predicted genes will be implicated. Furthermore, such data can be used to refine my

predictions by noting which of my predicted infertility genes have recurrent mutations more frequently in the cases than the controls.

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 my large prediction sets is to apply a complementary high-throughput experimental system to test these predictions. To this end, I 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 my top candidates reported here. My hope is that by tying together high-dimensional computational analysis of mammalian germ cells with novel high-throughput genomic assays in these same cells, I can help usher in a new era of functional genomics for mammalian reproductive biology.

Chapter 1 - Methods

Training gene sets

To create the positive and negative trainings sets of mouse genes, I 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, I extracted the genes with an observed reproductive system phenotype (MP:0005389) to make a reproductive positive training set gene list. For the negative training set gene list, I took all the other genes from the list that did not have a reproductive system phenotype 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.

To make the male reproductive training gene list I picked the genes shown to cause abnormal male reproductive morphology (MP:0001145) and physiology (MP:0003698) from the original positive training set gene list. Similarly I used the categories for abnormal female reproductive morphology (MP:0001119) and physiology (MP:0003699) to create my female reproductive training gene list. I 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.

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[14–17] . This produced a list of 67 human male fertility genes and 62 human female fertility genes.

These lists were combined to produce a list of 125 human fertility genes. Finally the human male fertility gene list was curated for genes only found in non-obstructive azoospermia to produce a list of 41 genes.

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[1,2] 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)[18] and Hammoud (GSE49624)[19]. Mouse and human oocyte RNA-Seq paired end reads were taken from Xue (GSE44183)[20].

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[21] with the default values and gene expression was summarized using cuffnorm[22] on the UCSC gene annotations to normalize across the datasets. I 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)

I aligned the CHIP-Seq read to the mm9 assembly using Novocraft's novoalign tool with its default options (http://www.novocraft.com). Following that, I used seqminer[23] to map the reads + -5kb around the transcription start site (TSS). I 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 HPRD[24], Reactome[25], and STRING[26] 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[27]. Sum of weight of edges were calculated as a measure of proximity to a group of 'seed' genes as described previously[9]. Seed gene sets that I used to calculate scores included cancer, early development, haploinsufficiency and known reproductive genes that I 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, I 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 I used 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 I then calculated the χ score for all genes which is a projection of the multidimensional data onto a one dimensional continuum. I 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, I 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. I 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

I obtained existing Copy Number Variant (CNV) calls from men assayed in my previous study of spermatogenic impairment[28]. Using published, validated CNV calling pipelines, I 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). I used the extensive health history data available on each WHI subject to construct a diagnosis that I

believe approximates the clinical definition of primary ovarian insufficiency, resulting in a case/control classification for all WHI individuals.

I 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 the carry at least one CNV that results in meeting one 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. I then built a 2X2 contingency table for the case-control status and ran a Fisher's exact test.

I 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 separately**.**

Note: All supplemental tables (Table S1-S31) can be found in the supplement of the paper:

Ho, N. R. Y., Huang, N. & Conrad, D. F. Improved detection of disease-associated variation by sex-specific characterization and prediction of genes required for fertility. *Andrology* **3,** 1140–1149 (2015).

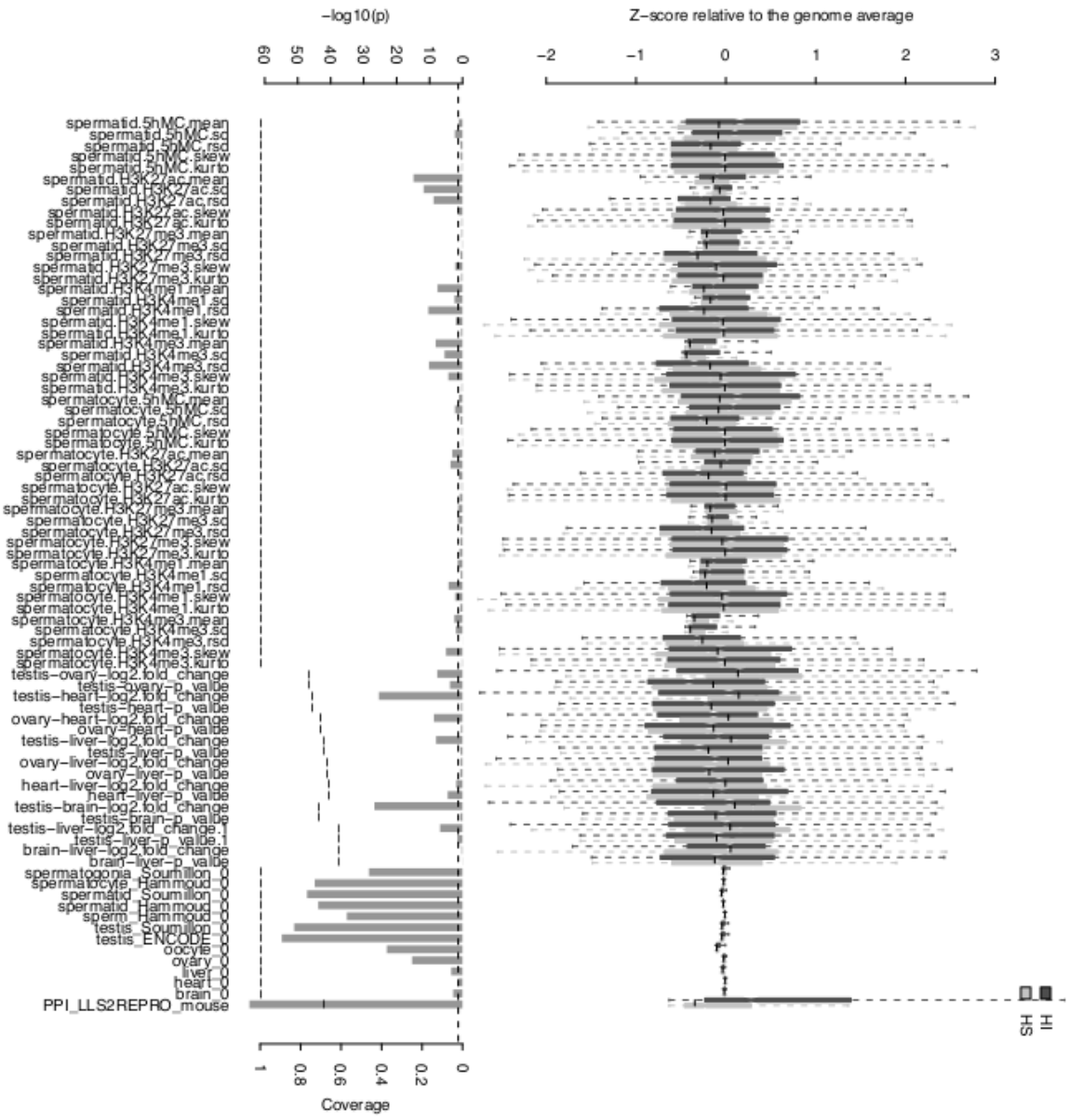


Figure 1.S1: Data distribution of all features for MGI reproductive training set in the mouse genome

The right plot shows the spread of the renormalized data when using positive and negative training sets. HI indicates positive training set genes and HS shows the distribution for negative control genes. The left bar plot shows how likely it is that the data came from the same distribution, where a higher –log10P signifies that it is more likely that the 2 features have different data distributions in the positive and negative training sets. The line in the left plot shows the percentage of the genome that the feature covers. The positive training set is the MGI reproductive gene set and the negative training set is the MGI null gene set.

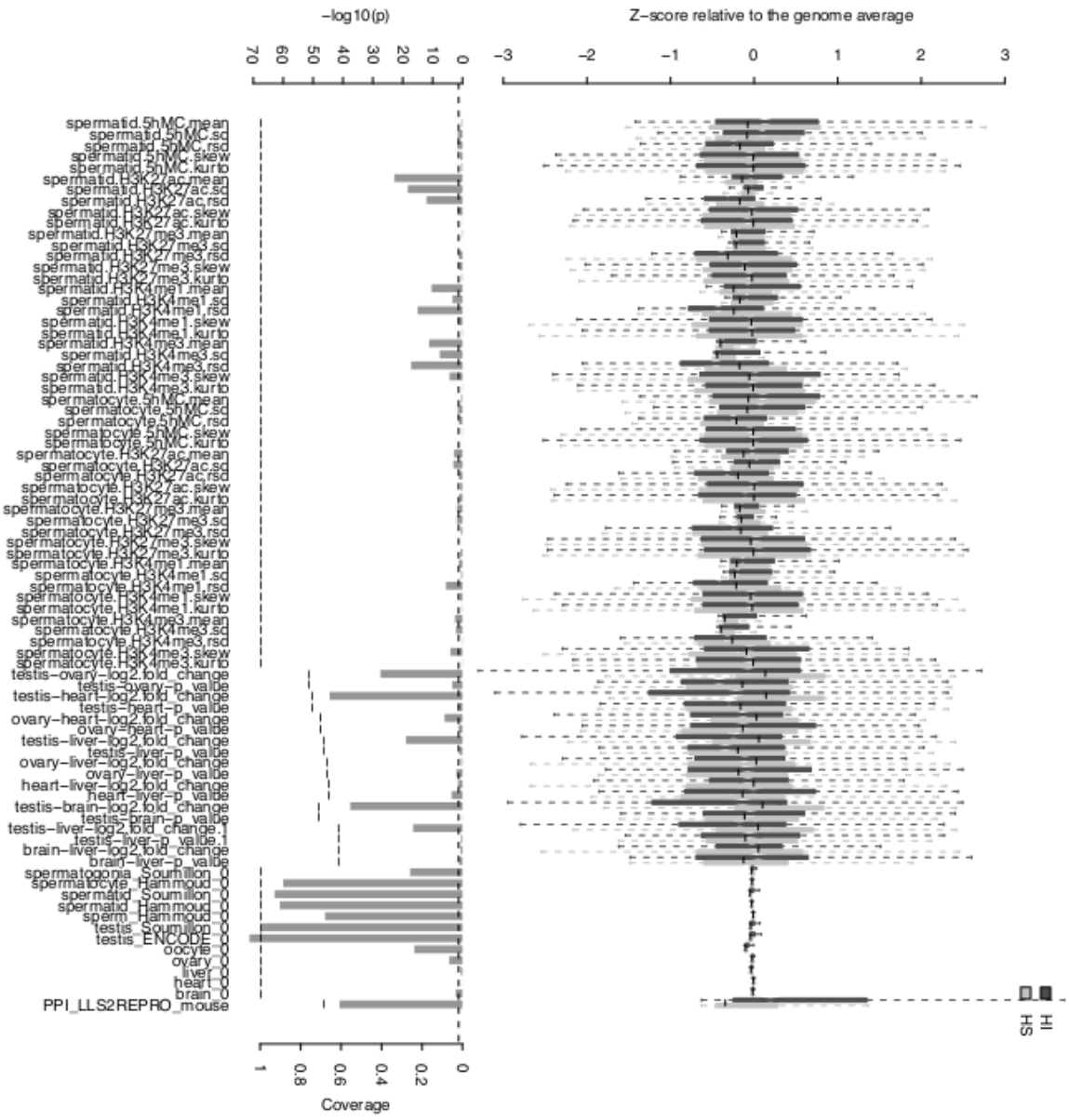


Figure 1.S2: Data distribution of all features for MGI male reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI male specific reproductive gene set and the negative training set is the MGI null gene set.

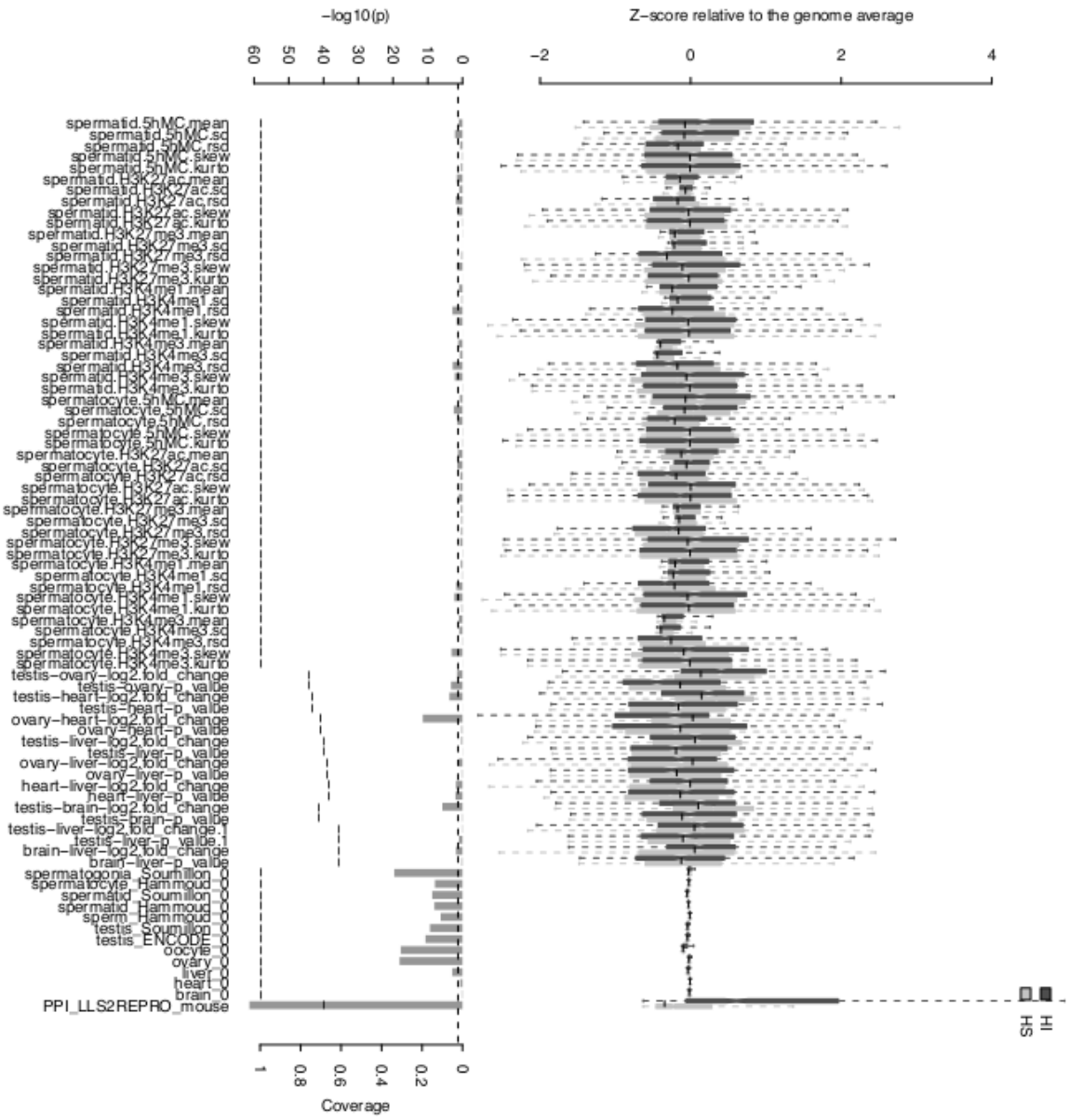


Figure 1.S3: Data distribution of all features for MGI female reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI female specific reproductive gene set and the negative training set is the MGI null gene set.

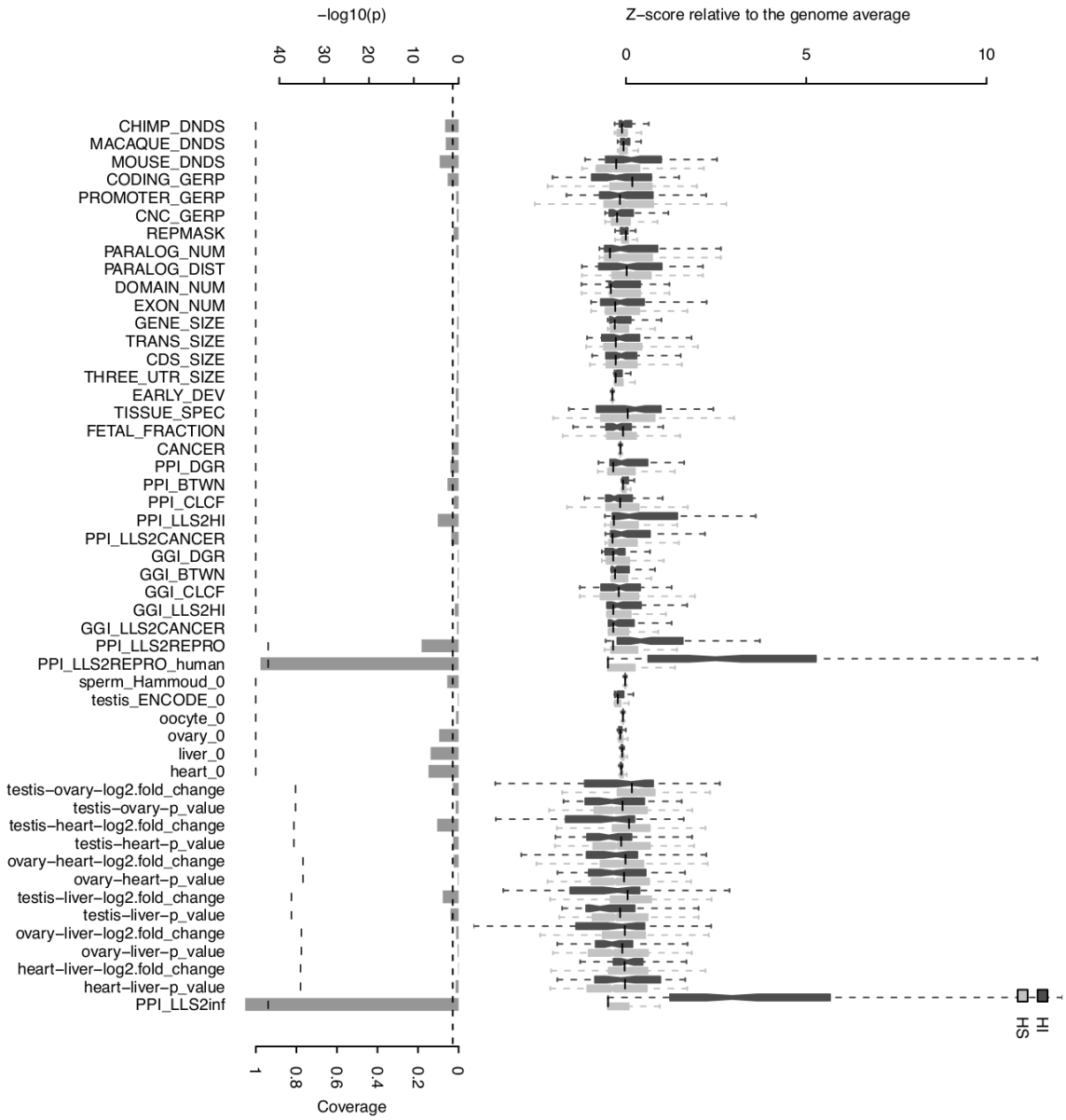


Figure 1.S4: Data distribution of all features for reproductive training set in the human genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the human fertility gene set and the negative training set is the MGI null gene set for genes that are conserved between mice and humans.

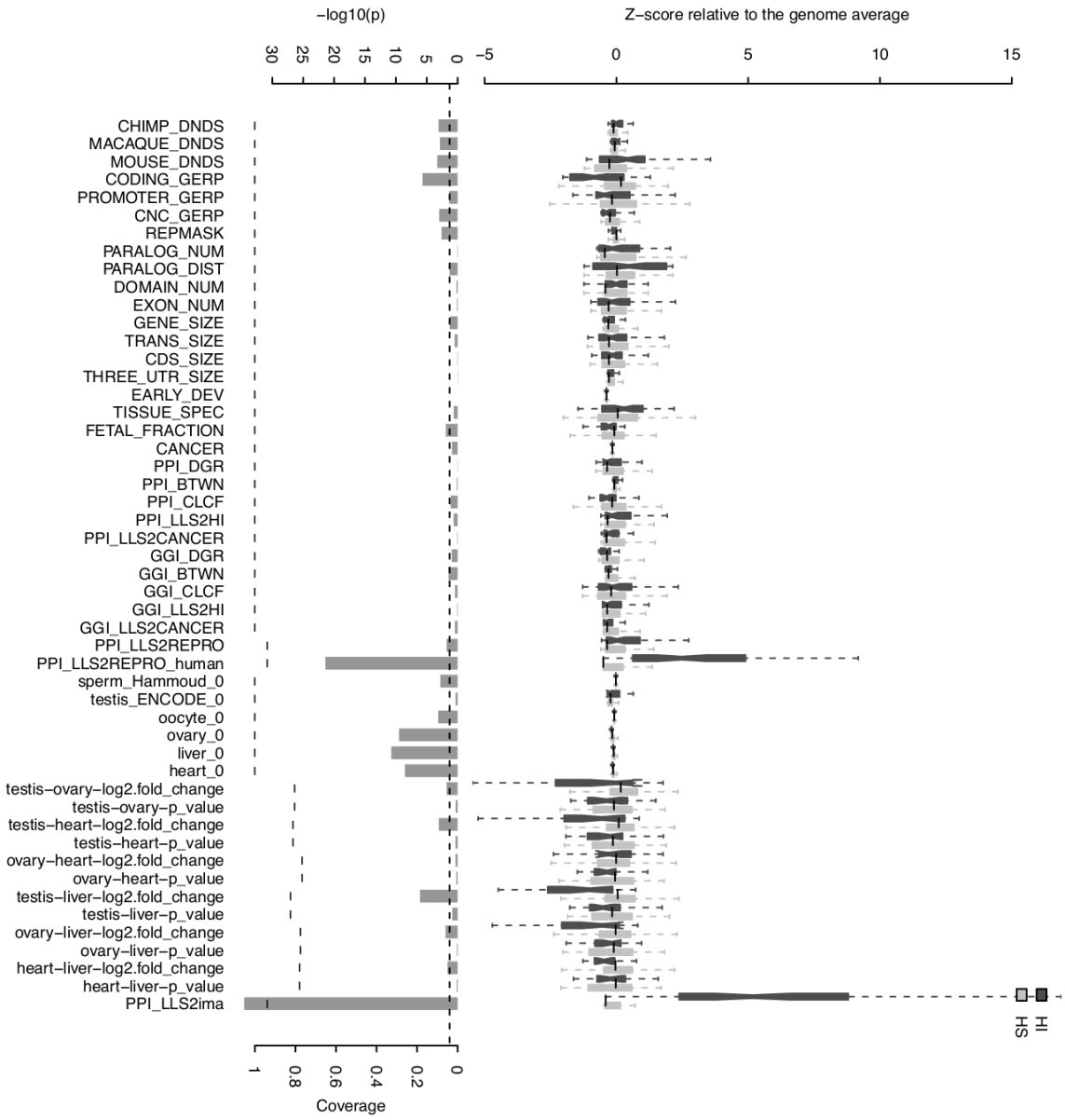


Figure 1.S5: Data distribution of all features for male reproductive training set in the human genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the male specific reproductive gene set and the negative training set is the MGI null gene set for genes that are conserved between mice and humans.

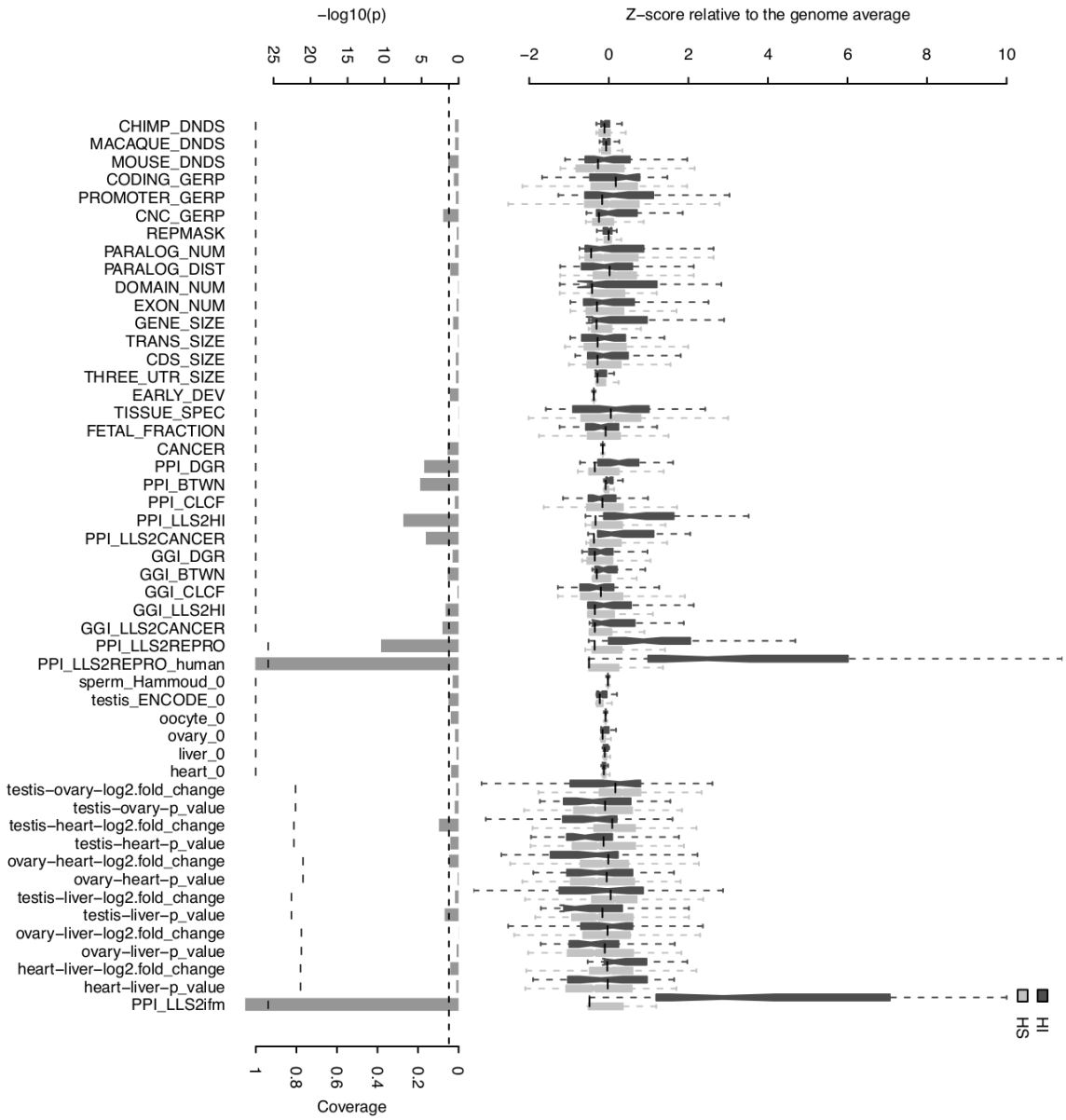


Figure 1.S6: Data distribution of all features for female reproductive training set in the human genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the female specific reproductive gene set and the negative training set is the MGI null gene set for genes that are conserved between mice and humans.

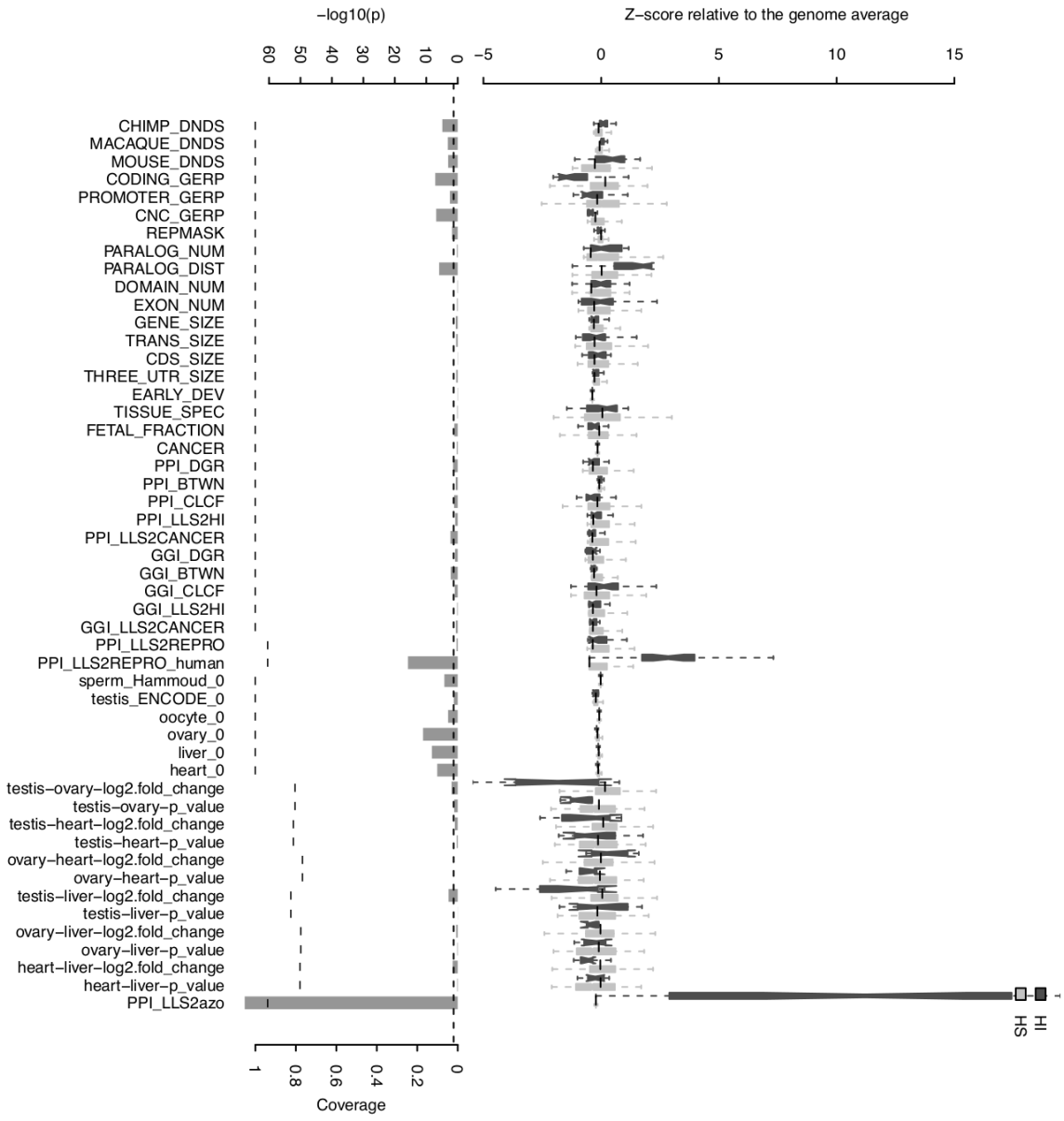


Figure 1.S7: Data distribution of all features for non-obstructive azoospermia training set in the human genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the non-obstructive azoospermia gene set and the negative training set is the MGI null gene set for genes that are conserved between mice and humans.

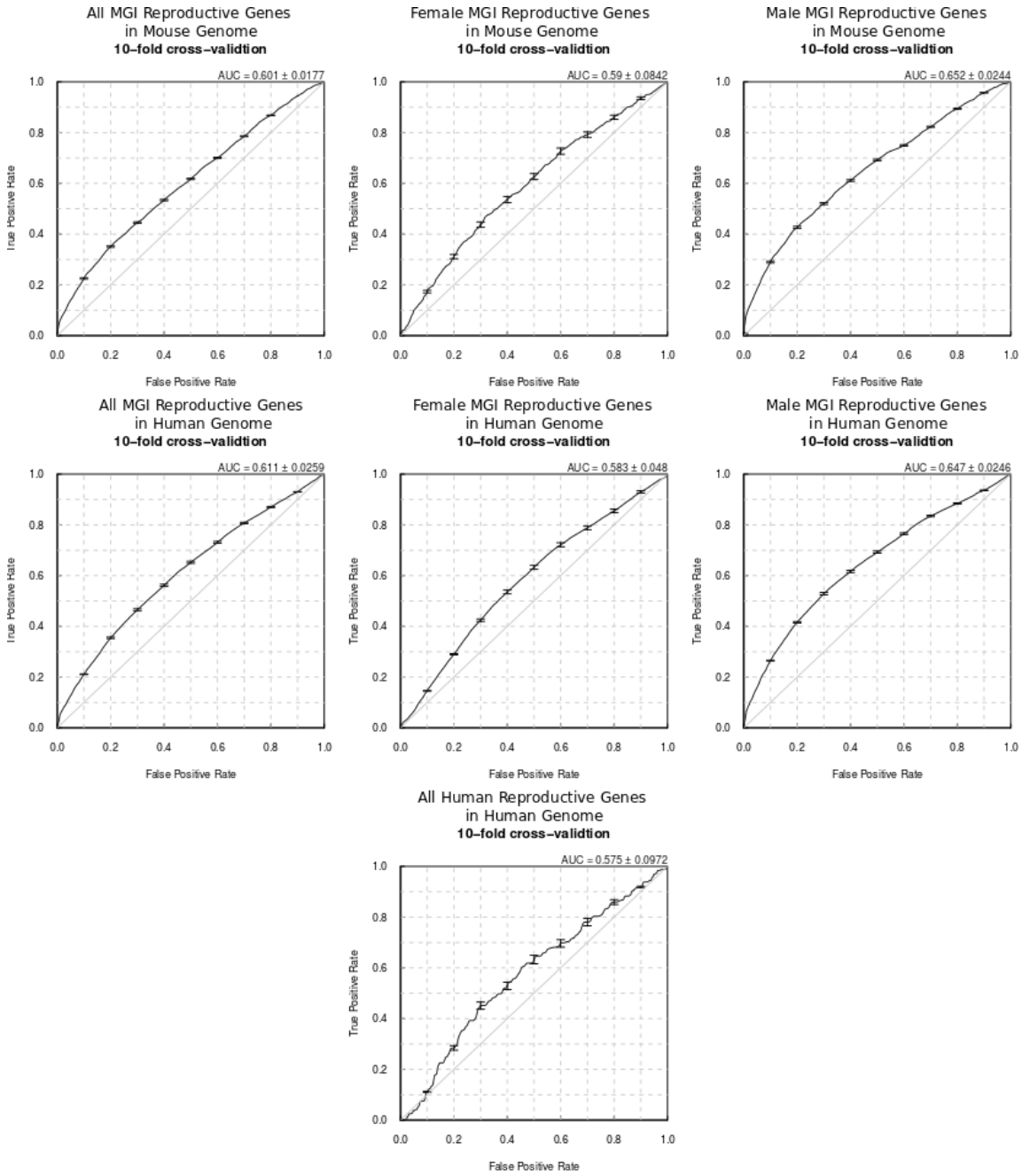


Figure 1.S8: Model performance benchmarks without PPI

Each figure shows the receiver operator curve for the LDA model classifying the test set genes correctly based on χ score cutoffs. Each LDA model has the PPI feature removed from the list of features used to generate the χ scores. The negative training set used is always MGI null genes.
All MGI genes: Positive set is MGI male reproductive gene training set
Female MGI genes: Positive set is MGI female reproductive gene training set
Male MGI genes: Positive set is MGI reproductive gene training set
Human Reproductive genes: Positive set is reproductive genes implicated by human GWAS studies

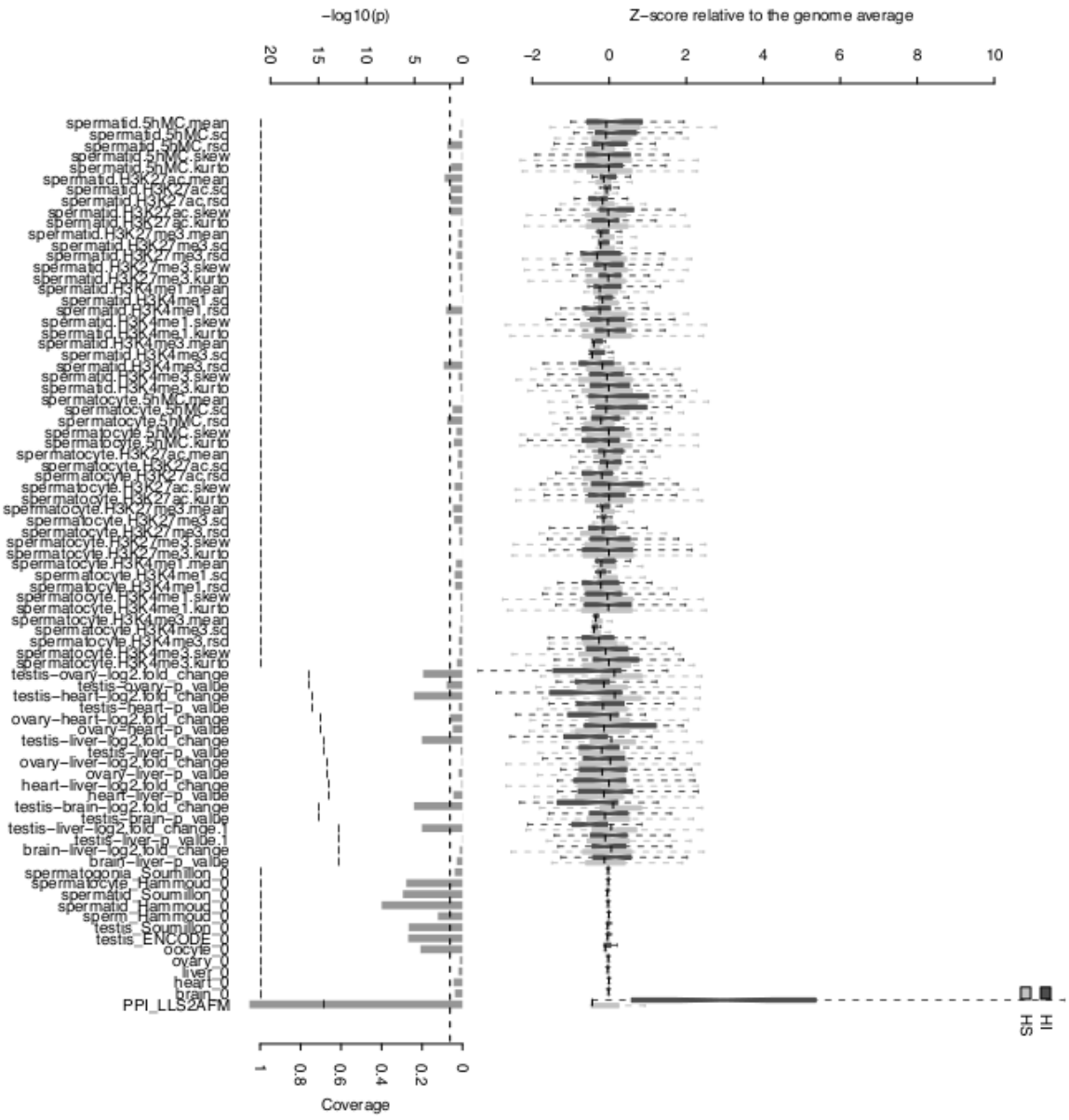


Figure 1.S9: Data distribution of all features for MGI abnormal female meiosis reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI abnormal female meiosis specific reproductive gene set and the negative training set is the MGI null gene set.

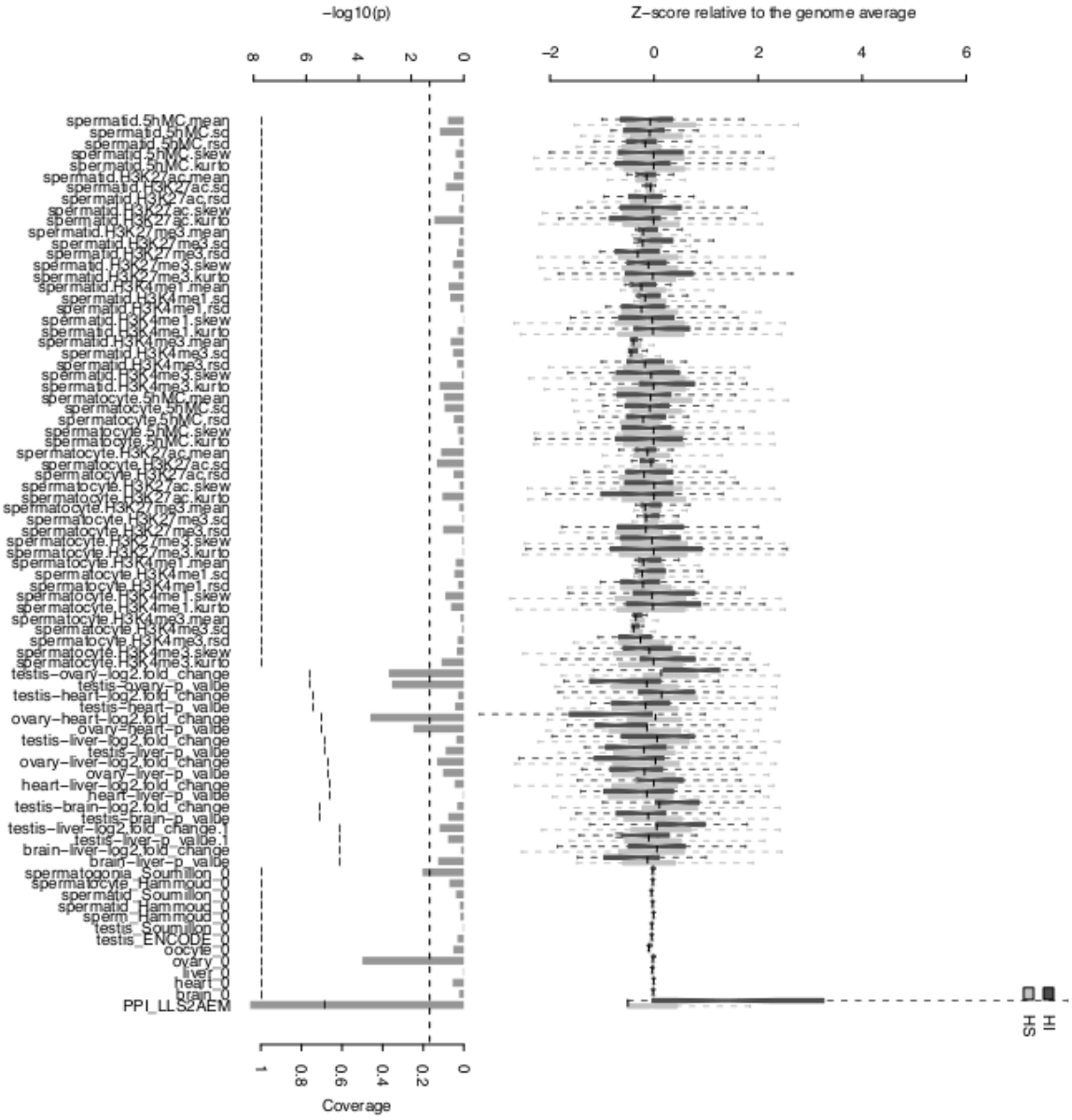


Figure 1.S10: Data distribution of all features for MGI abnormal endometrium morphology training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI abnormal endometrium morphology specific reproductive gene set and the negative training set is the MGI null gene set.

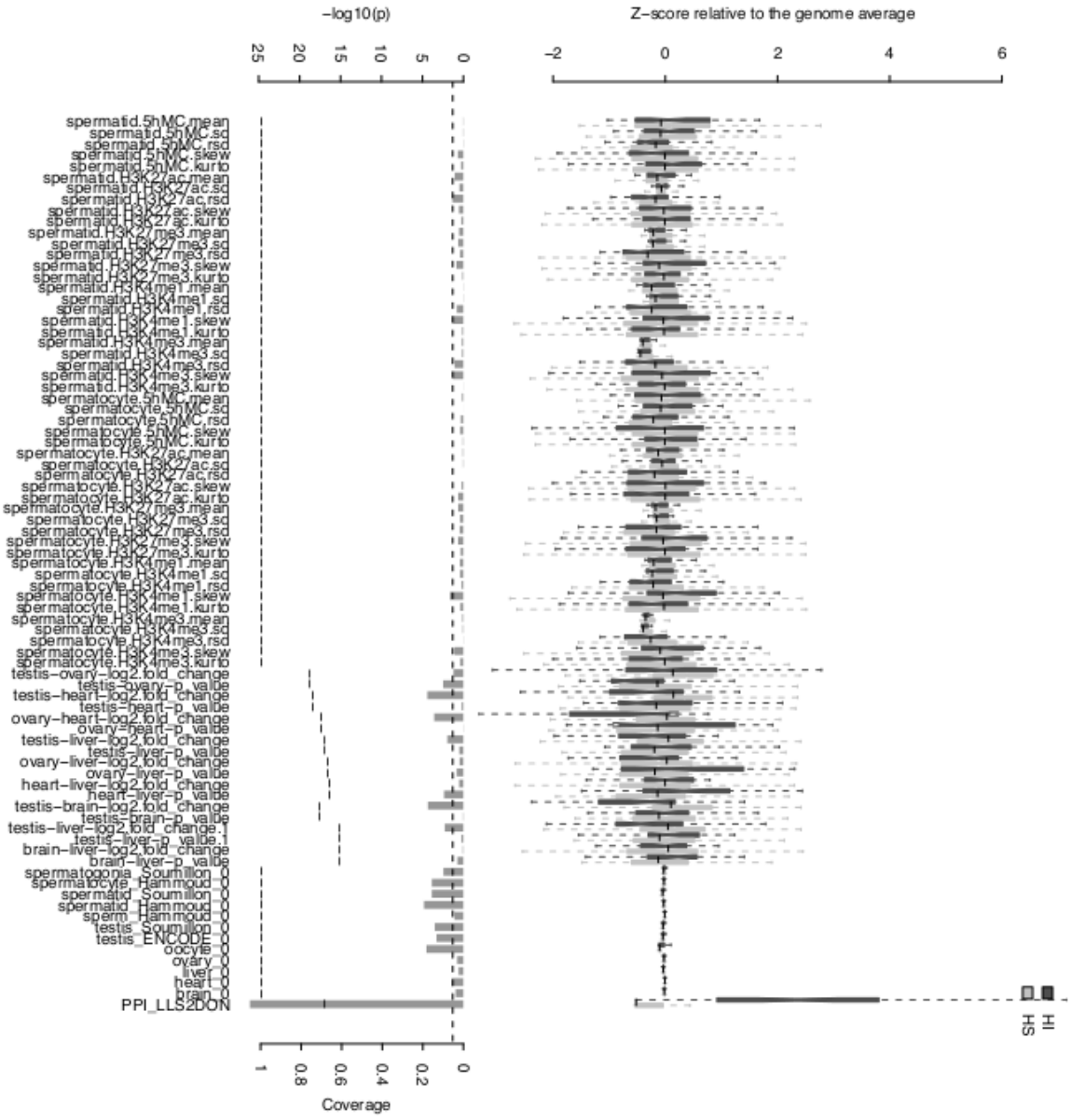


Figure 1.S11: Data distribution of all features for MGI decreased oocyte number reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI decreased oocyte number specific reproductive gene set and the negative training set is the MGI null gene set.

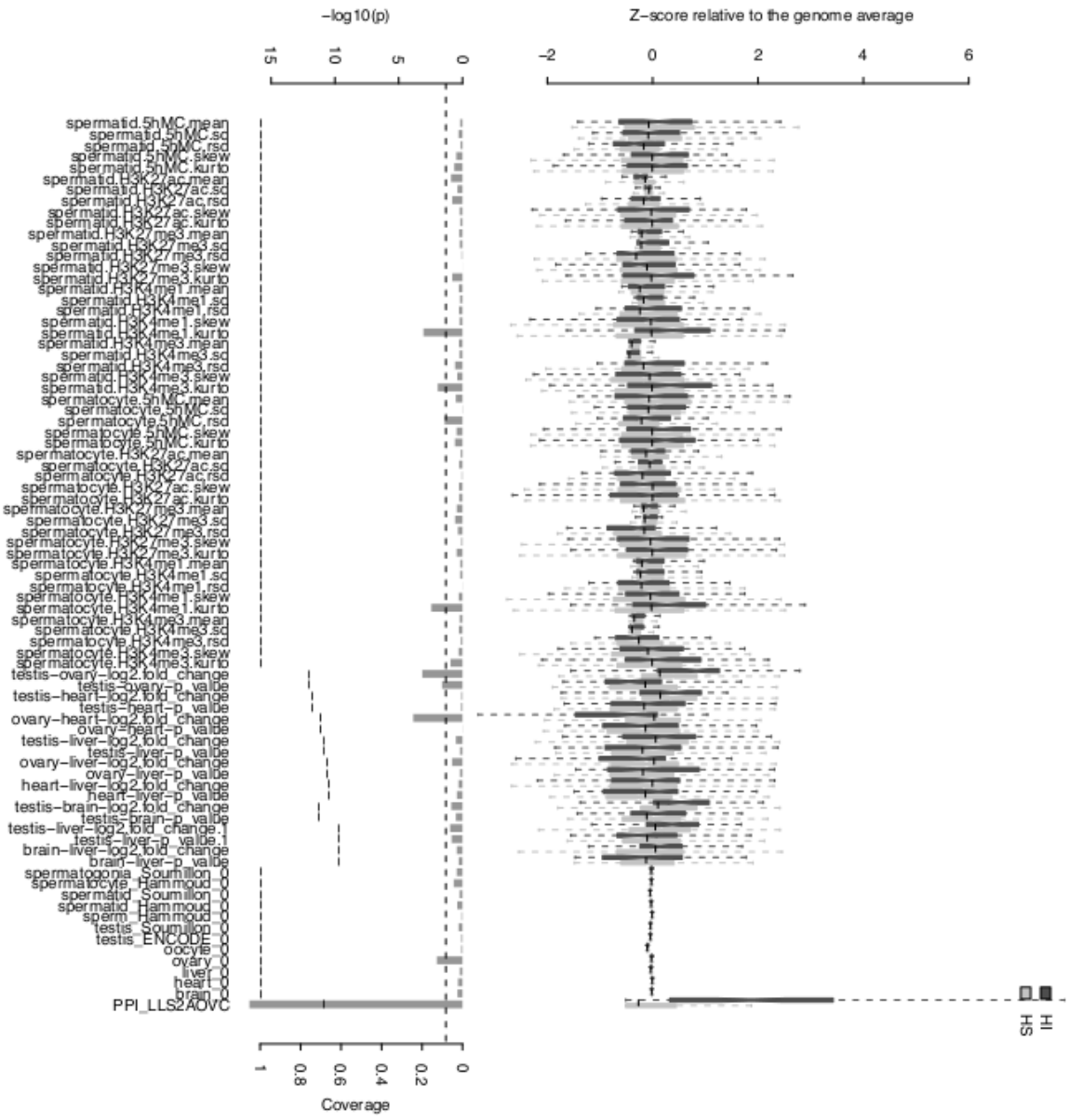


Figure 1.S12: Data distribution of all features for MGI abnormal ovulation cycle reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI abnormal ovulation cycle specific reproductive gene set and the negative training set is the MGI null gene set.

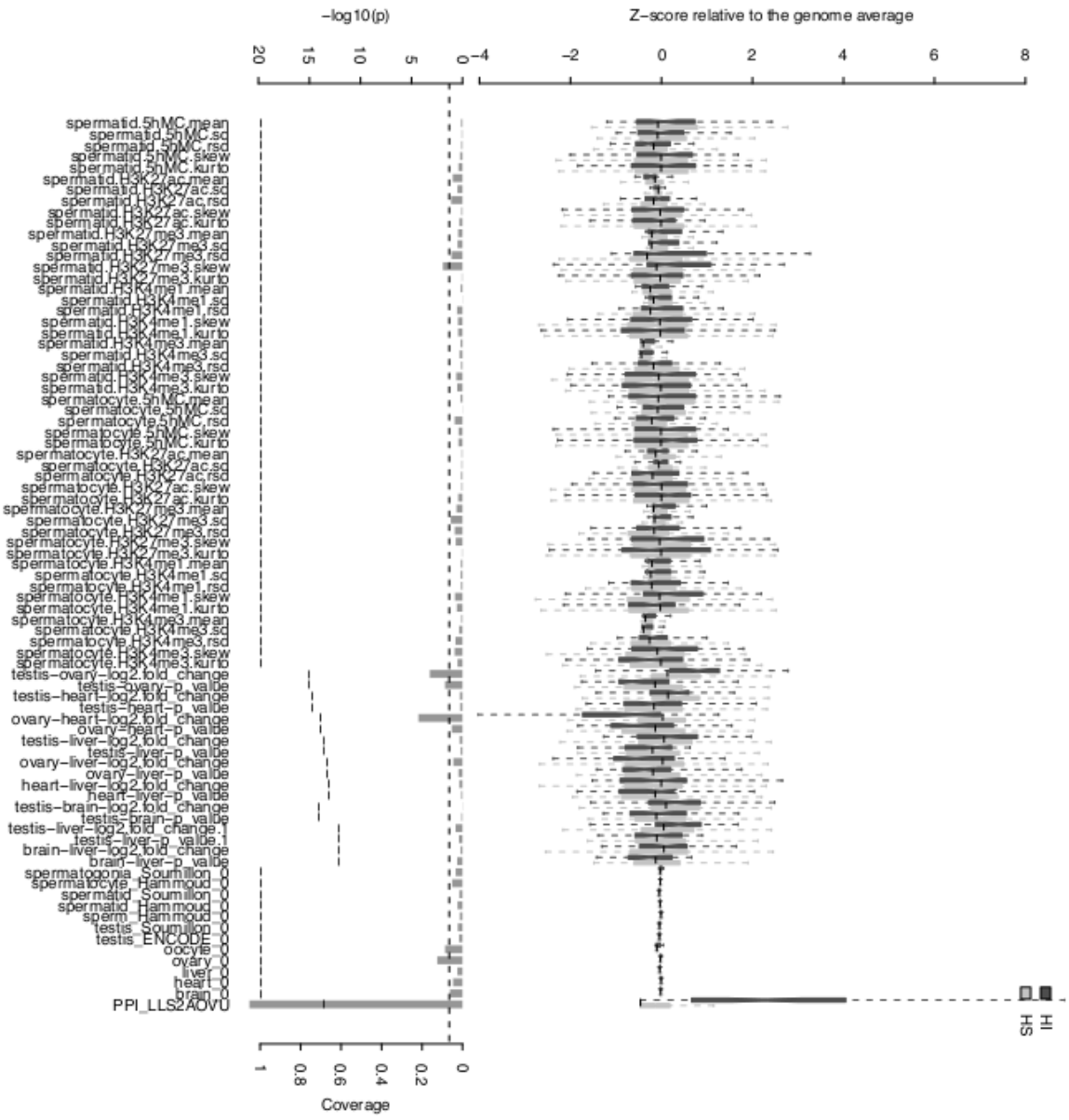


Figure 1.S13: Data distribution of all features for MGI abnormal ovulation reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI abnormal ovulation specific reproductive gene set and the negative training set is the MGI null gene set.

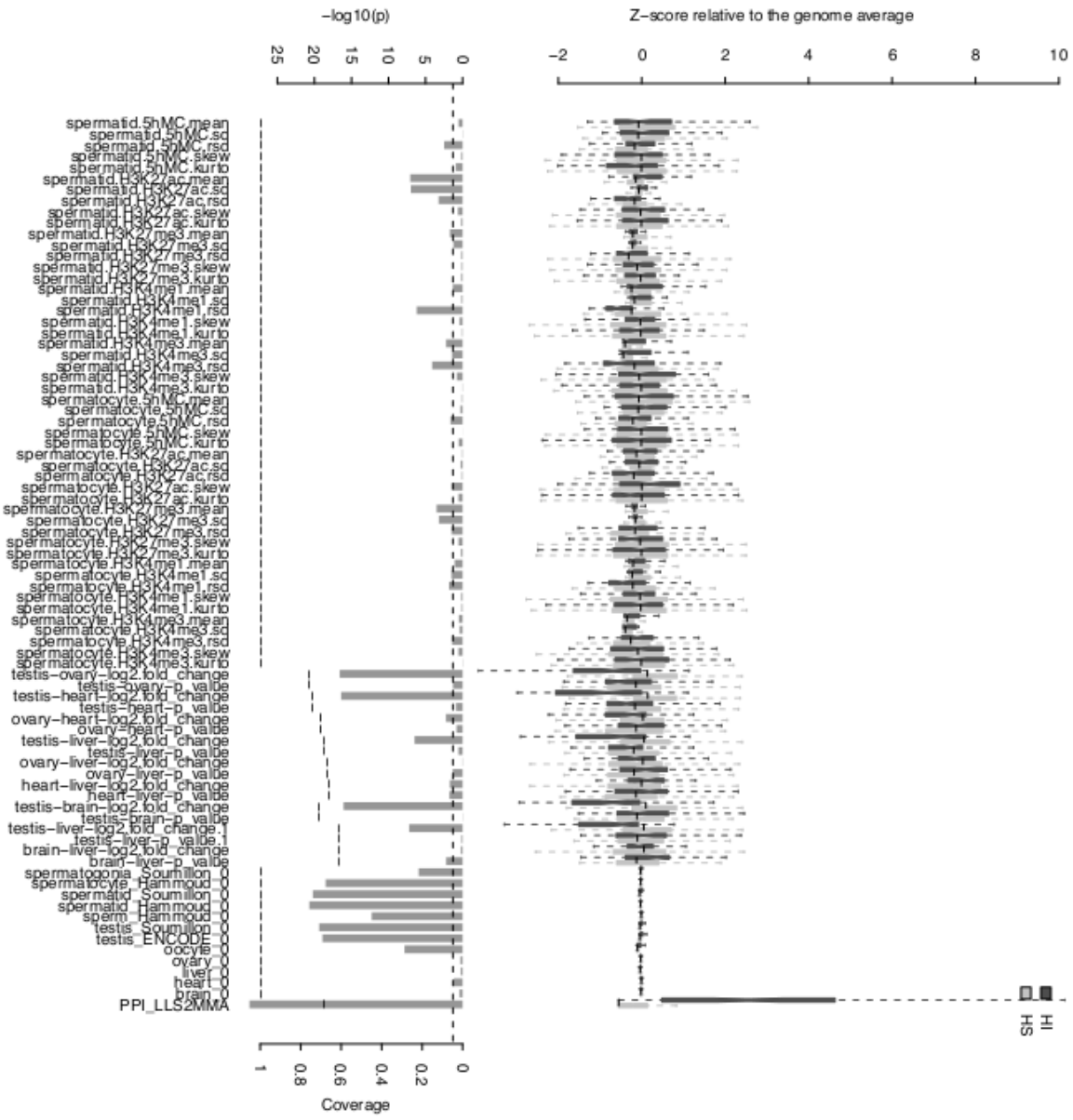


Figure 1.S14: Data distribution of all features for MGI male meiosis arrest reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI male meiosis arrest specific reproductive gene set and the negative training set is the MGI null gene set.

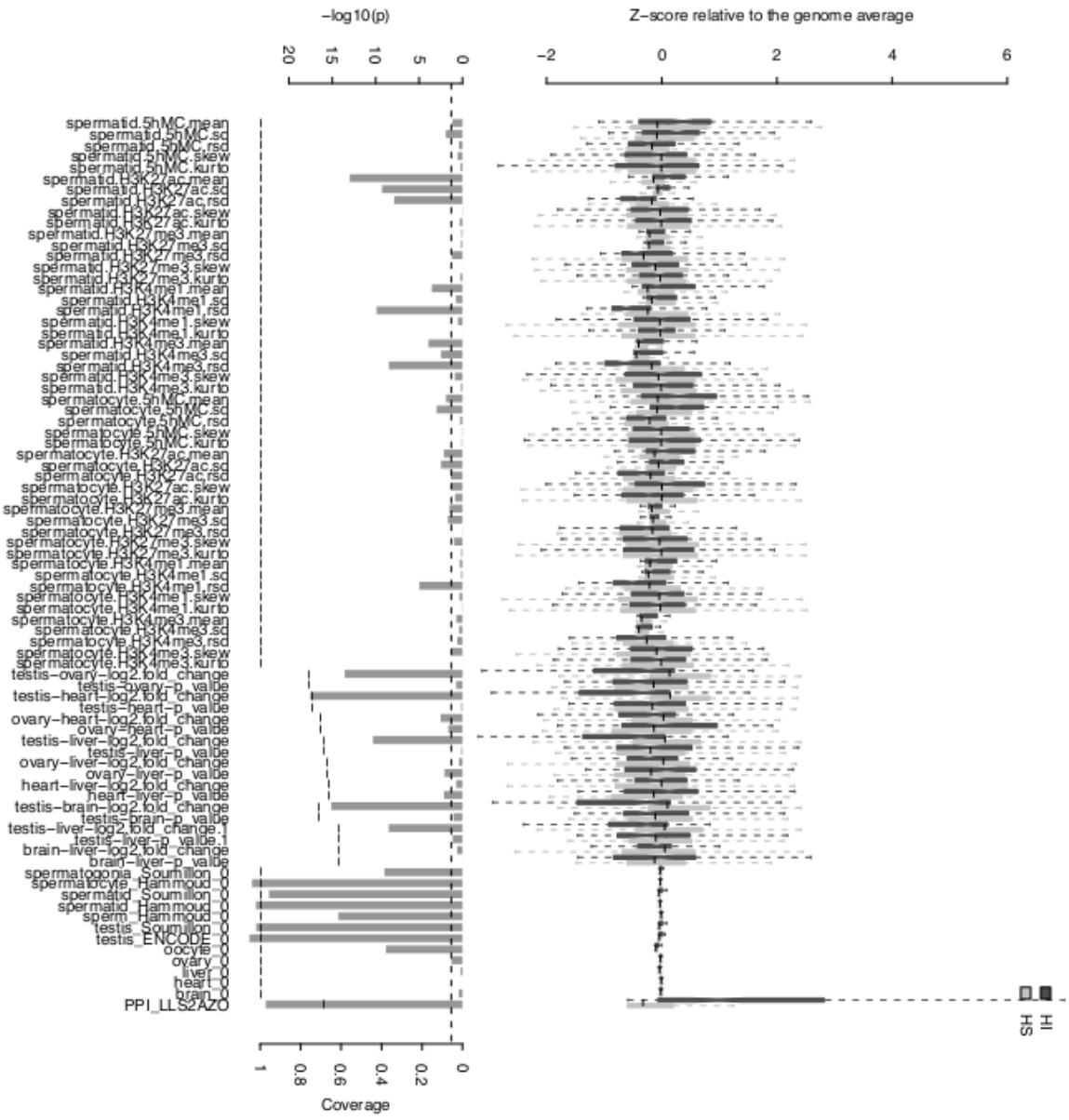


Figure 1.S15: Data distribution of all features for MGI azoospermia reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI azoospermia specific reproductive gene set and the negative training set is the MGI null gene set.

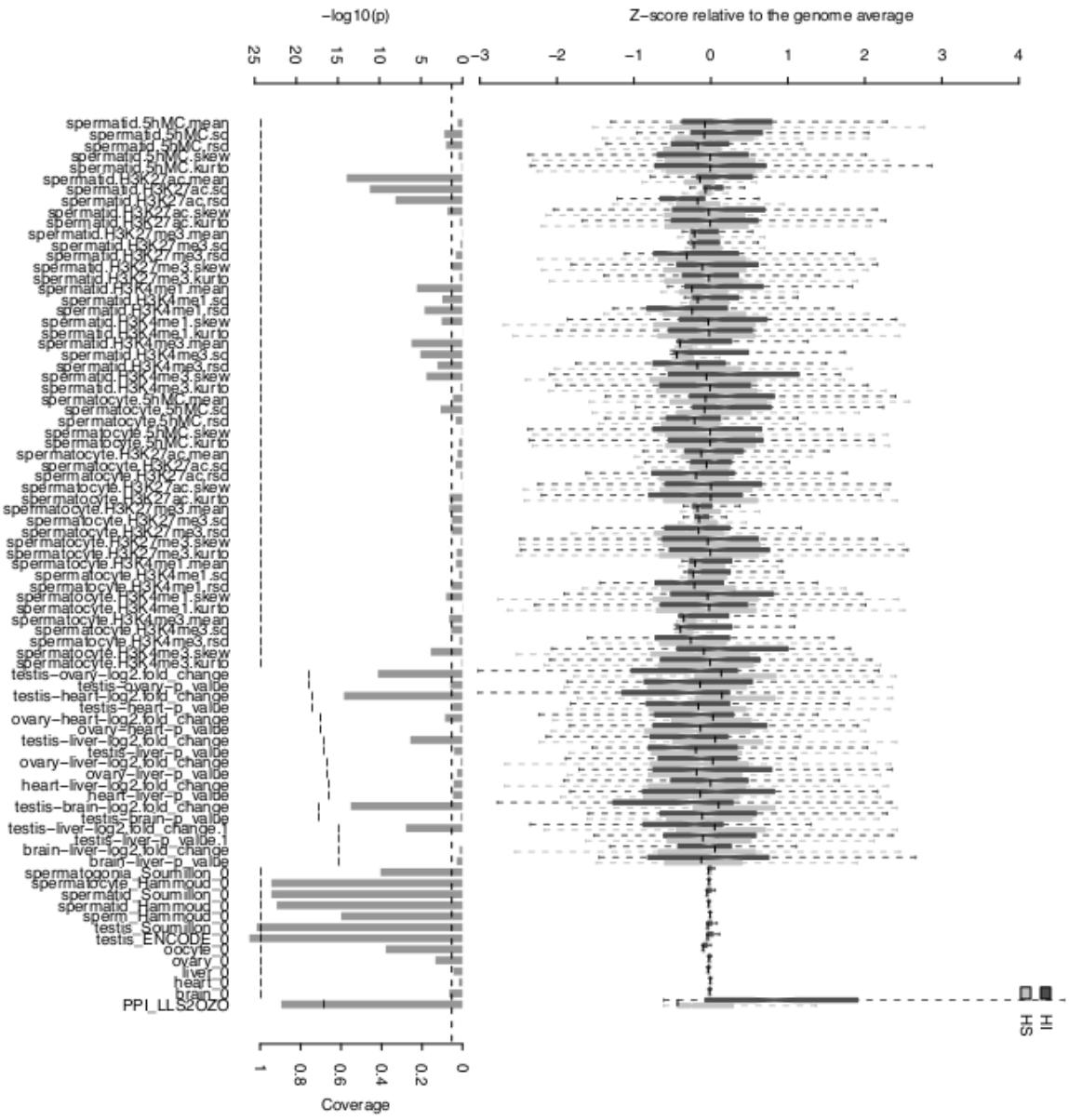


Figure 1.S16: Data distribution of all features for MGI oligozoospermia reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI oligozoospermia specific reproductive gene set and the negative training set is the MGI null gene set.

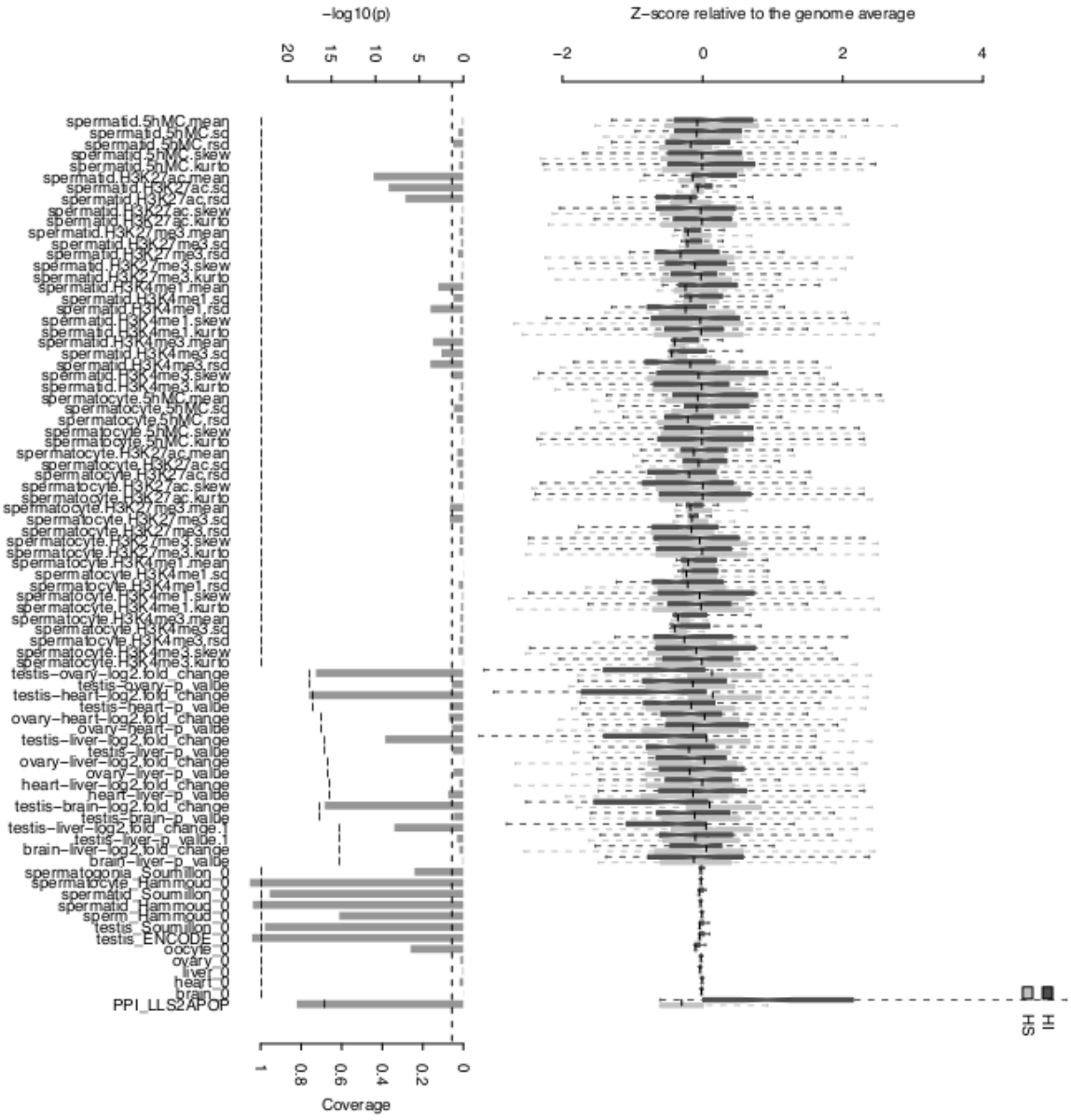


Figure 1.S17: Data distribution of all features for MGI male germ cell apoptosis reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI male germ cell apoptosis specific reproductive gene set and the negative training set is the MGI null gene set.

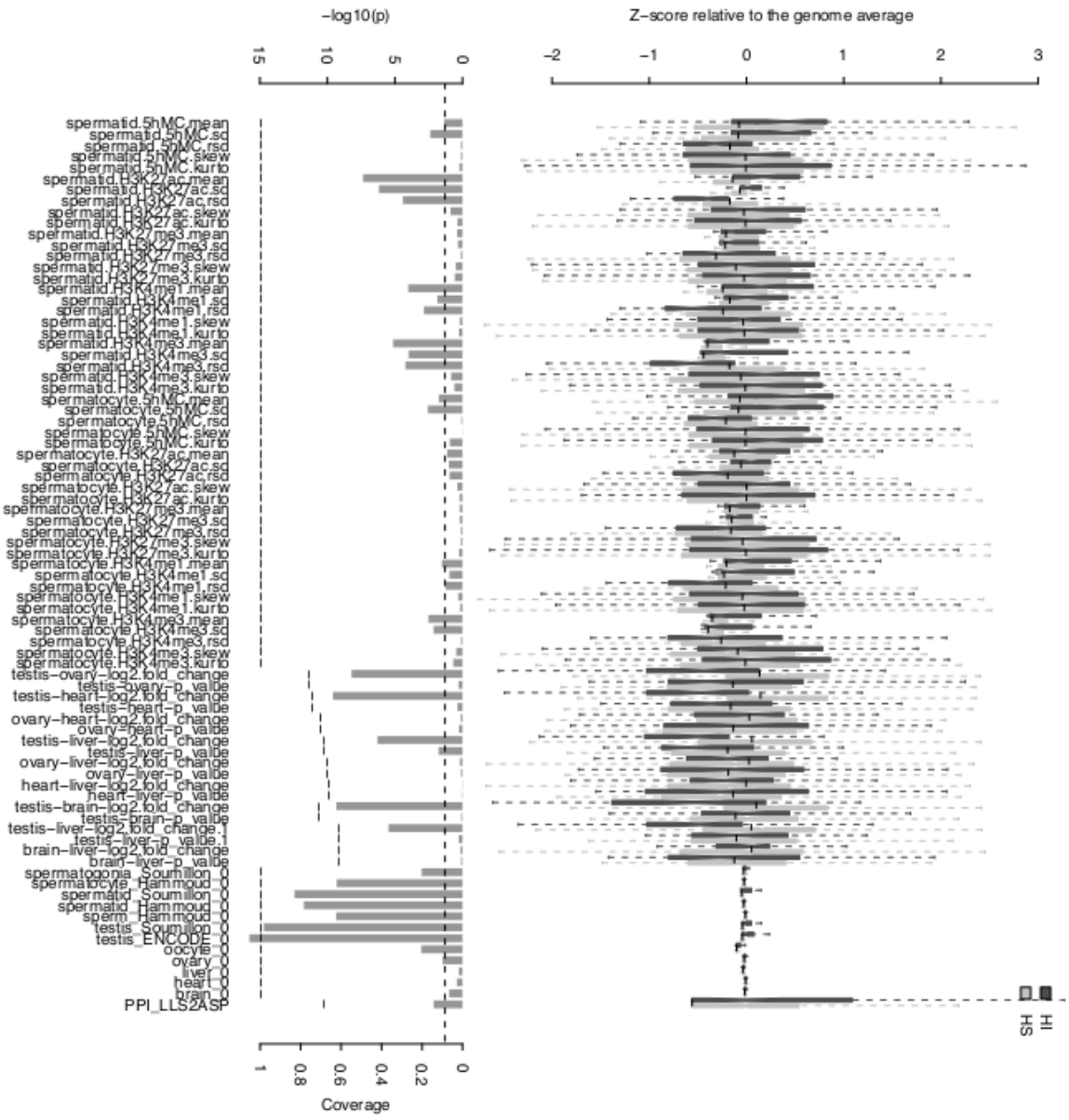


Figure 1.S18: Data distribution of all features for MGI abnormal spermiogenesis reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI abnormal spermiogenesis specific reproductive gene set and the negative training set is the MGI null gene set.

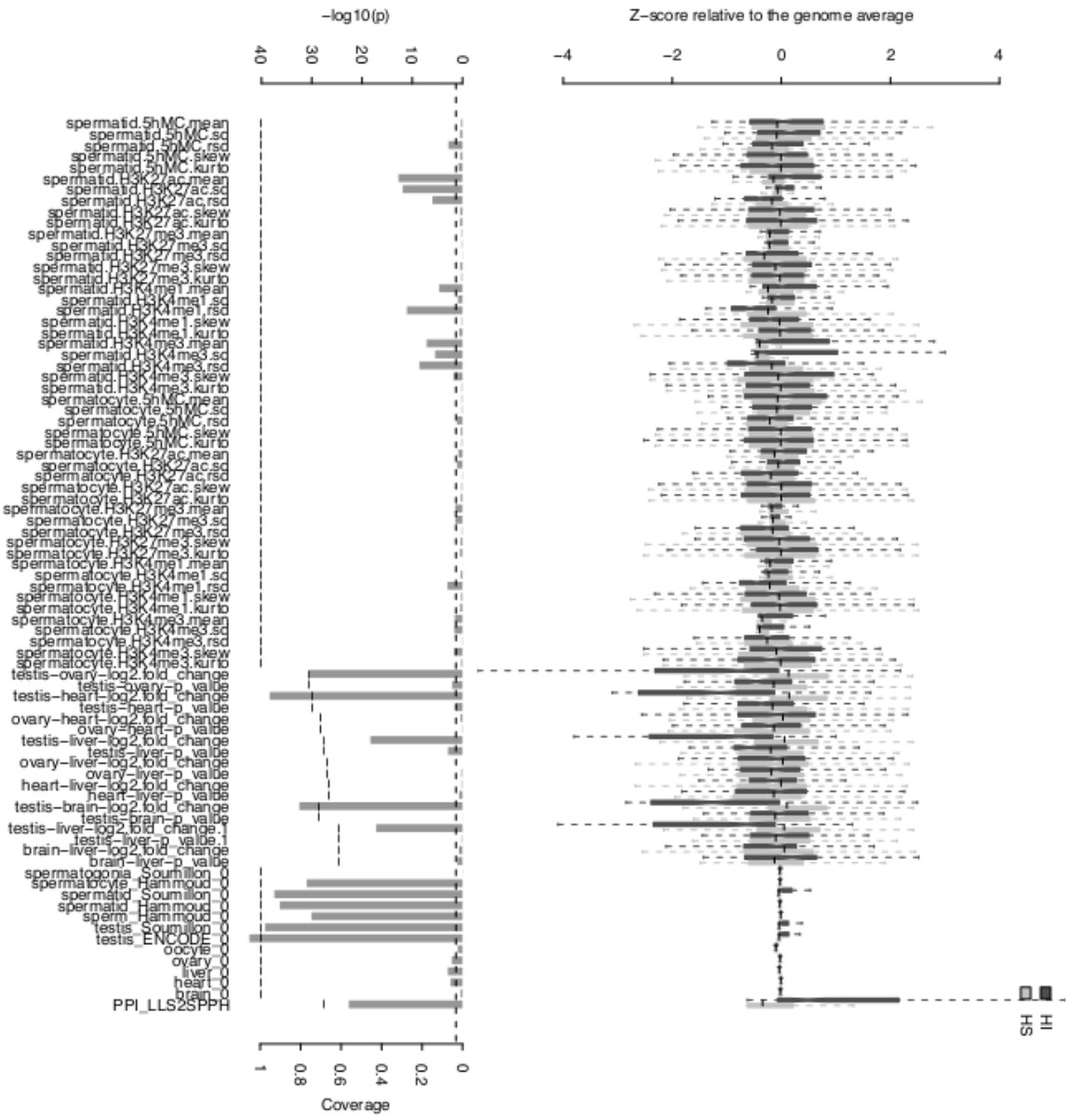


Figure 1.S19: Data distribution of all features for MGI sperm physiology reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI sperm physiology specific reproductive gene set and the negative training set is the MGI null gene set.

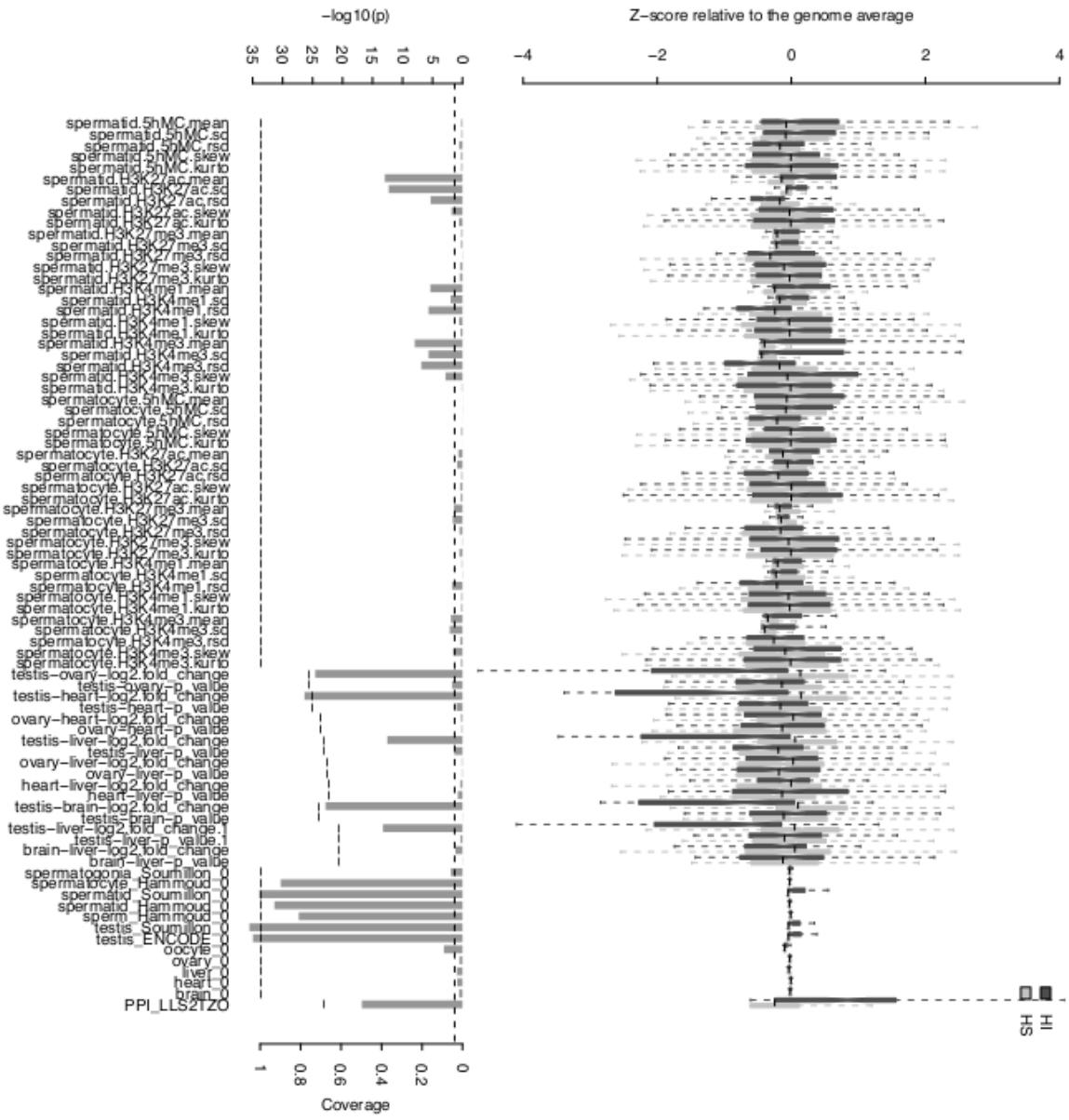


Figure 1.S20: Data distribution of all features for MGI teratozoospermia reproductive training set in the mouse genome

The figure is laid out the same way as Figure S1. Here, the positive training set used is the MGI teratozoospermia specific reproductive gene set and the negative training set is the MGI null gene set.

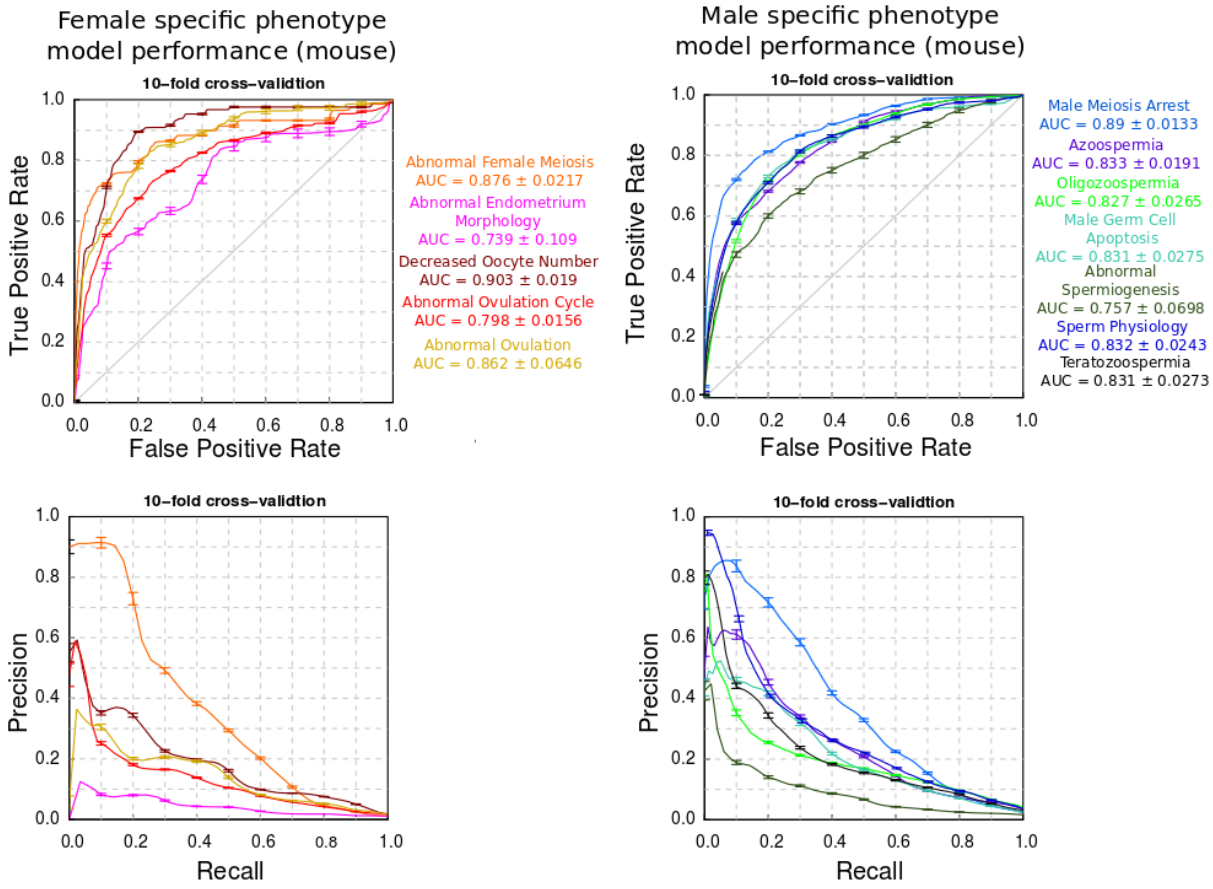


Figure 1.S21: Model performance benchmarks for mouse infertility subset phenotypes
On the top are the figures for the receiver operator curve for the LDA model classifying the test set genes correctly based on χ score cutoffs. The negative training set used is always MGI null genes. On the bottom are the precision recall curves for the same LDA models. On the left are the specific infertility phenotypes that affect females while the male specific infertility phenotypes are on the right.

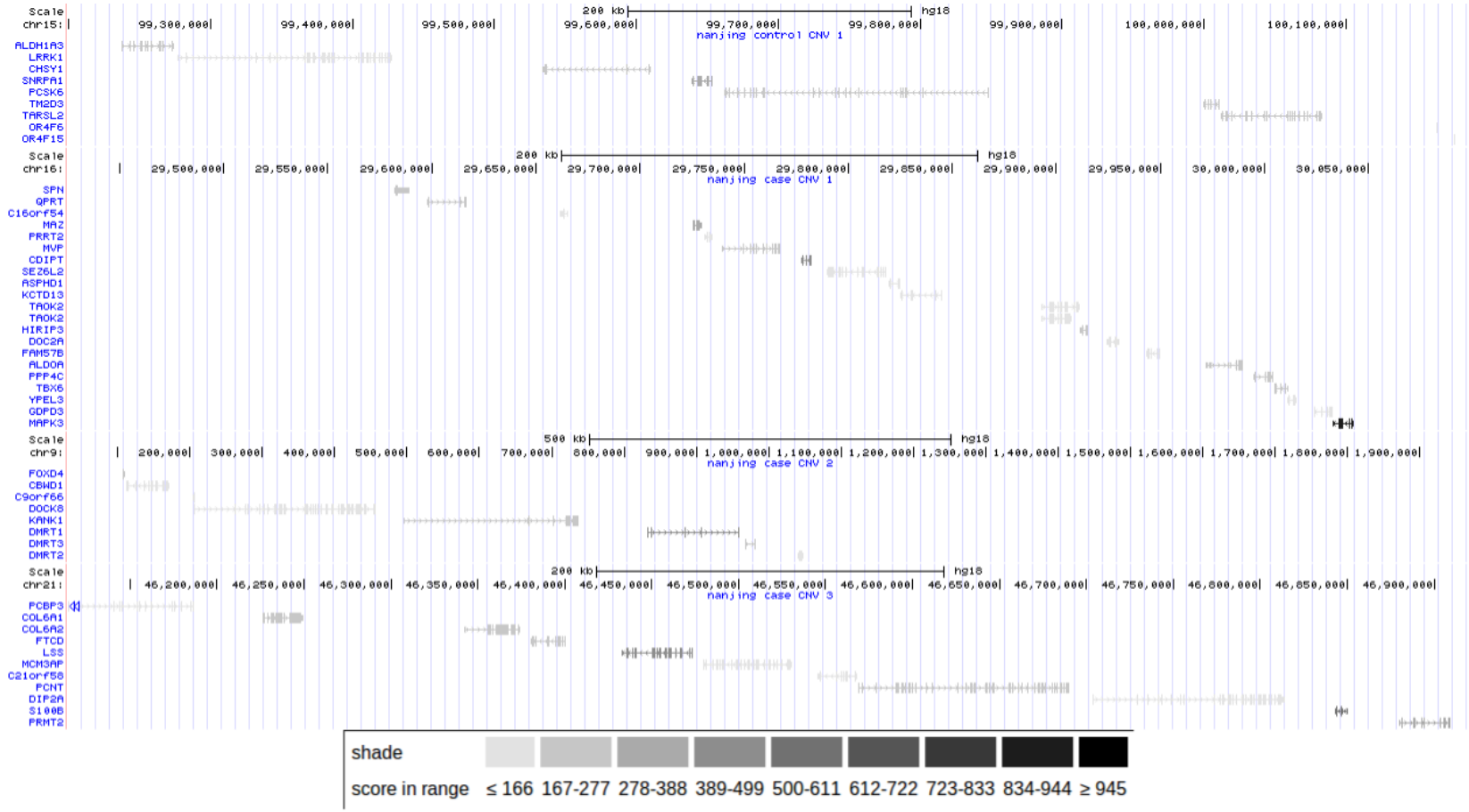


Figure 1.S22: Examples of finding candidate infertility genes in patient CNVs

Each figure shows the scores for each gene in large CNVs found in a Nanjing azoospermia GWAS study. The top figure is a large deletion found in a control patient while the bottom three figures are for large deletions found in different case patients. This shows that while not every large CNV will have at least one potential candidate, in some CNVs there are candidate genes which are predicted to have a high likelihood of being causative for infertility. For example, MAPK3 in the second figure and LSS and S100B in the bottom figure.

References

1. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).
2. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489, 57–74 (2012).
3. Sun, P. G. Prediction of Human Disease-Related Gene Clusters by Clustering Analysis. Int. J. Biol. Sci. 61–73 (2011). doi:10.7150/ijbs.7.61
4. Singh-Blom, U. M. et al. Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses. PLoS ONE 8, e58977 (2013).
5. Ballouz, S. et al. Gentrepid V2.0: a web server for candidate disease gene prediction. BMC Bioinformatics 14, 249 (2013).
6. Radivojac, P. et al. An integrated approach to inferring gene-disease associations in humans. Proteins 72, 1030–1037 (2008).
7. van Driel, M. A. et al. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res. 33, W758–761 (2005).
8. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. PLoS Comput Biol 6, e1000641 (2010).
9. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and Predicting Haploinsufficiency in the Human Genome. PLoS Genet 6, e1001154 (2010).
10. Raymond, C. S., Murphy, M. W., O'Sullivan, M. G., Bardwell, V. J. & Zarkower, D. Dmrt1, a gene related to worm and fly sexual regulators, is required for mammalian testis differentiation. Genes Dev. 14, 2587–2595 (2000).
11. Tewes, A.-C., Ledig, S., Tüttelmann, F., Kliesch, S. & Wieacker, P. DMRT1 mutations are rarely associated with male infertility. Fertil. Steril. 102, 816–820.e3 (2014).
12. Bole-Feysot, C., Goffin, V., Edery, M., Binart, N. & Kelly, P. A. Prolactin (PRL) and Its Receptor: Actions, Signal Transduction Pathways and Phenotypes Observed in PRL Receptor Knockout Mice. Endocr. Rev. 19, 225–268 (1998).
13. McNeilly, A. S., Glasier, A., Jonassen, J. & Howie, P. W. Evidence for direct inhibition of ovarian function by prolactin. J. Reprod. Fertil. 65, 559–569 (1982).
14. Zorrilla, M. & Yatsenko, A. N. The Genetics of Infertility: Current Status of the Field. Curr. Genet. Med. Rep. 1, 247–260 (2013).
15. Hamada, A. J., Esteves, S. C. & Agarwal, A. A comprehensive review of genetics and genetic testing in azoospermia. Clinics 68, 39–60 (2013).
16. Navarro-Costa, P., Plancha, C. E. & Gonçalves, J. Genetic Dissection of the AZF Regions of the Human Y Chromosome: Thriller or Filler for Male (In)fertility? J. Biomed. Biotechnol. 2010, (2010).
17. Evian Annual Reproduction (EVAR) Workshop Group 2010 et al. Contemporary genetic technologies and female reproduction. Hum. Reprod. Update 17, 829–847 (2011).

18. Soumillon, M. et al. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Rep. 3, 2179–2190 (2013).
19. Hammoud, S. S. et al. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell 15, 239–253 (2014).
20. Xue, Z. et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597 (2013).
21. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013).
22. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 31, 46–53 (2013).
23. Ye, T. et al. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res. 39, e35 (2011).
24. Prasad, T. S. K. et al. Human Protein Reference Database—2009 update. Nucleic Acids Res. 37, D767–D772 (2009).
25. Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33, D428–D432 (2005).
26. Franceschini, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41, D808–D815 (2013).
27. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J. Matrix Anal. Appl. 30, 121–141 (2008).
28. Lopes, A. M. et al. Human Spermatogenic Failure Purges Deleterious Mutation Load from the Autosomes and Both Sex Chromosomes, including the Gene DMRT1. PLoS Genet 9, e1003349 (2013).

Chapter 2

Experimental *in vivo* genetics screen for spermatogenesis function

Nicholas Rui Yuan Ho, Abul Usmani, Yan Yin, Liang Ma, Donald F Conrad

To be submitted

Chapter 2 - Introduction

Spermatogenesis requires the activation of many different pathways at various time points, varying from general processes like metabolism, cell cycle, meiosis, and transcription, to very specific functions like membrane capacitation and cell-cell recognition proteins[1,2]. Many of the pathways are even distinct from somatic cells despite having the same function[3]. Due to the complexity of this process, I expect that there should be numerous genetic defects that cause infertility. Pathogenic mutations that impair fertility are unlikely to be inherited (and thus recurrent), making it difficult to identify fertility genes via recurrent mutations in families. However as next generation sequencing becomes more affordable, Genome Wide Association Studies (GWAS) are helping to identify genes which are associated with infertility by finding recurrent mutations in unrelated individuals[4,5].

Unfortunately it is currently difficult to verify candidate fertility gene functions because the existing *in vitro* model systems for complete spermatogenesis are technically challenging to perform[6–8]. Generation of knockout mouse models has thus been the most popular tool to verify candidate pathogenic genes. This has been translated to humans to advance our knowledge and develop treatments for human infertility. However, costs a few thousand dollars and between months to years to characterize a single gene by generating a knockout mouse line. As large scale genomic studies become more commonplace, the gap between implicated and verified fertility genes will only get larger if this remains as the verification method of choice.

To address this problem, I have developed a quick, simple, and inexpensive method to screen numerous genes simultaneously *in* vivo for spermatogenesis function. The basis for the method

lays in RNA interference (RNAi) screens, which are commonly used to efficiently elucidate gene function. This approach has been used *in vitro*[9,10] in cell lines or *in vivo*[11–16] in other tissues in mice to discover important genes for various different biological processes.

The process of *in vitro* RNAi screens is relatively simple; cultured cells are transfected with the miRNA or RNAi expression construct using a highly effective transfection reagent (e.g. chemical, viral) and treated cells are then selected for a trait of interest (e.g. survival, response to stimulus). Following that, the cells are harvested and the RNAi in the cells with and without the selection pressure are quantified to determine the differences between them. To produce reproducible results it is important to have a sufficient percentage of cells infected. *In vivo* transfection rates tend to be orders of magnitude lower than *in vitro* systems for the same transfection reagent[17]. This leads to a trade-off between the number of biological replicates and the cost of the transfection reagent. To avoid the issue of low transfection rates, some groups have transfected certain cells *in vitro* and then transplanted them into recipient mice and performed selections in the xenografted models[13–16]. Spermatogonial stem cells have been transplanted into sterile donor testes to restore fertility in various species[18–20]. However this stem cell transplantation is a difficult technique to perform and there is a low number of unique stem cells that actually successfully transplant per mouse. This creates a bottleneck that makes cell transplantation inappropriate for a screening study, since screening relies on having a large number of independent transfection events in order to produce statistically significant results. Recently, a small number of studies have shown that one can use viruses to transfect mouse tissues *in vivo* with a sufficiently high transfection rate for multiplex selection *in vivo*[11,12].

Intrigued by this concept, I adapted a recently developed approach to transfect the germ cells in the testis[21] to render it compatible with linear DNA libraries expressing small hairpin RNA (shRNA) **(Figures 2.1a, 2.1b)**. The technique uses a buffered salt solution to generate an osmotic gradient which drives water and the dissolved DNA into the germ cells of the testis at a reasonably high rate. A similar approach using electroporation has been shown to work for a single shRNA in spermatogenesis[22].

I demonstrate the feasibility of using this low cost transfection method in mouse testes to screen multiple genes simultaneously for functional importance in spermatogenesis. By carefully designing the pilot study, I was also able to benchmark this system to prove the importance of large numbers of biological replicates and quantify the limits of this system. I also applied this method to establish the functional importance of twenty six uncharacterized genes that I previously predicted to be important for infertility via machine learning[23].

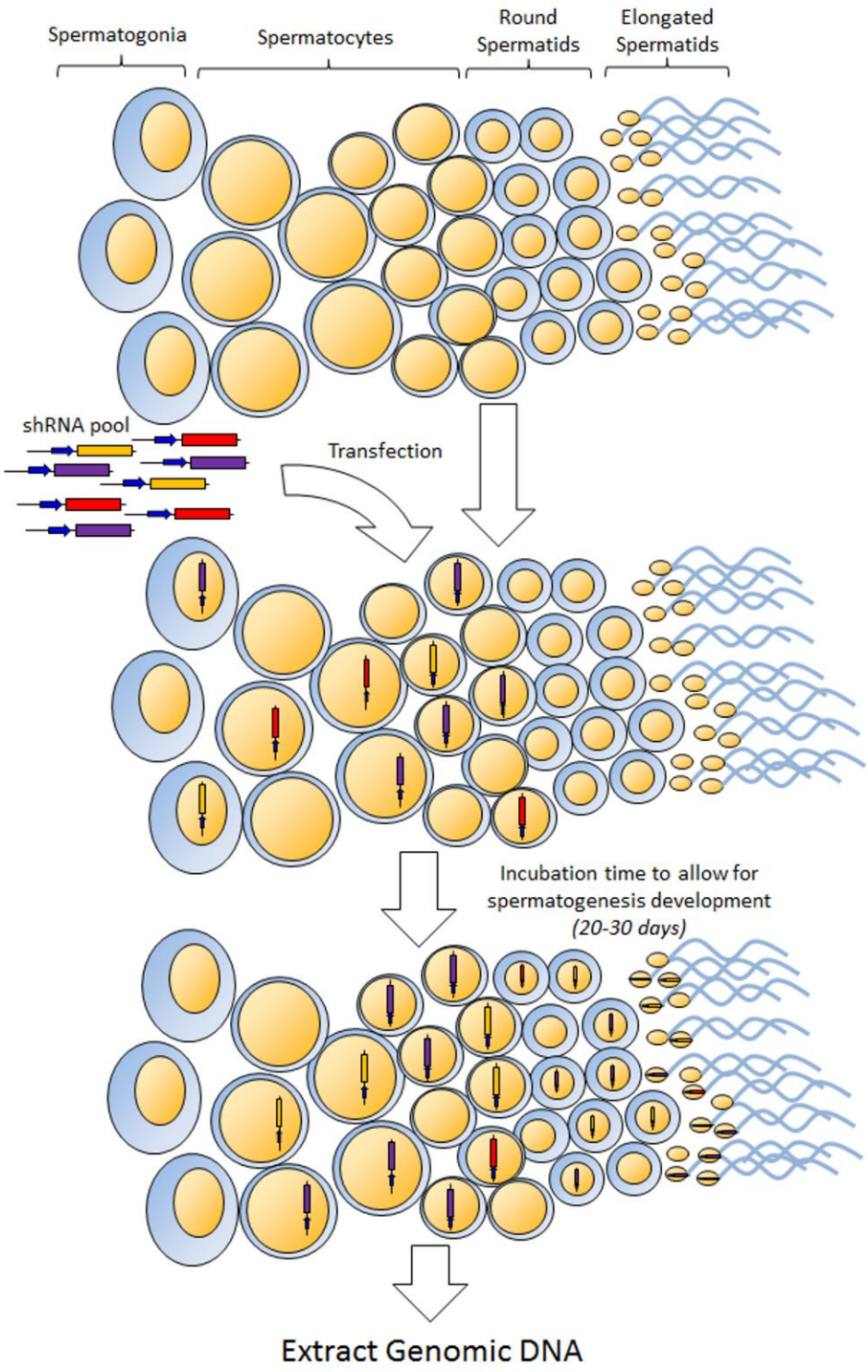


Figure 2.1a: Overview of Experimental approach (Transfection and Selection)

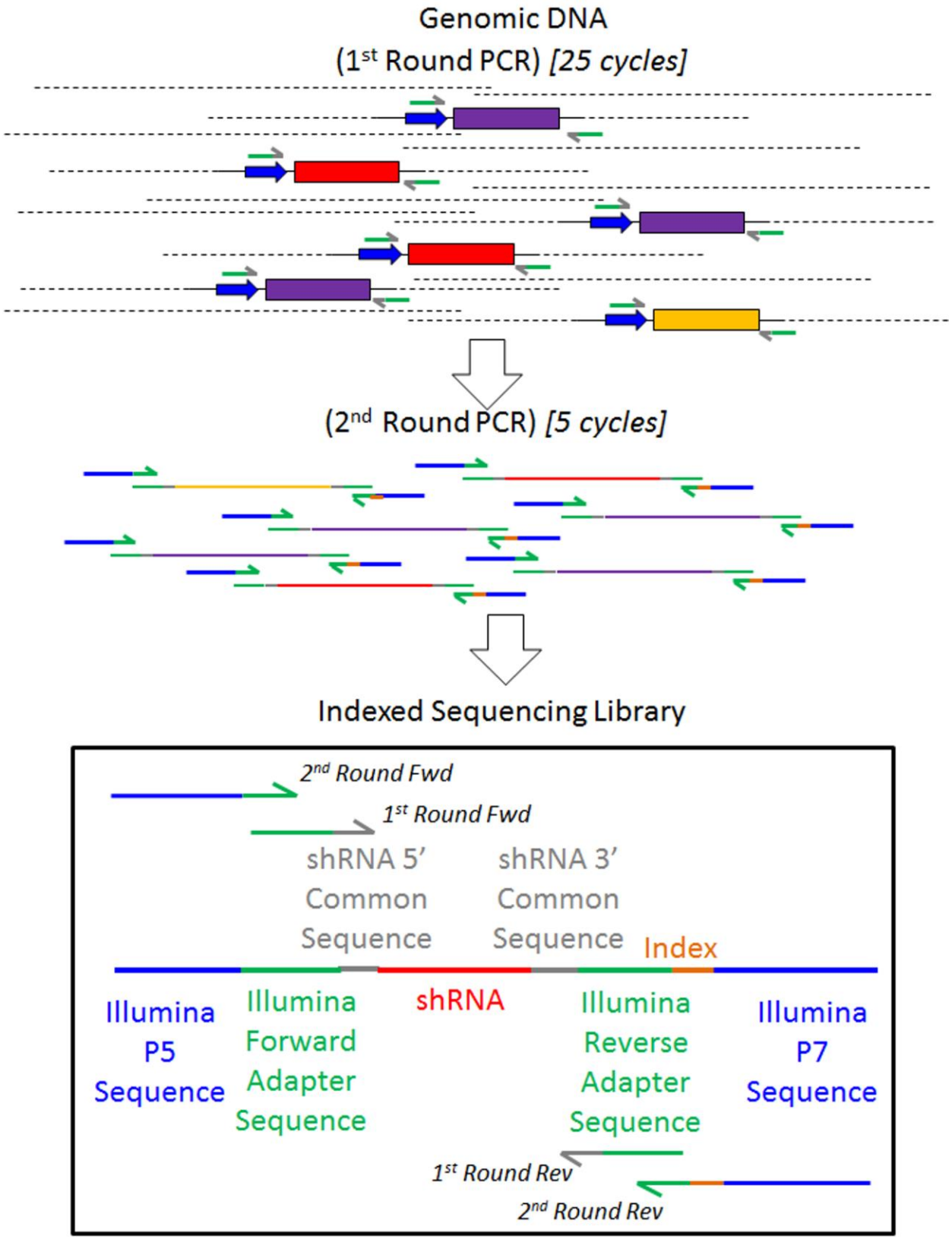


Figure 2.1b: Overview of Experimental approach (Sequencing)

Chapter 2 - Results

Pilot shRNA screen

As a proof-of-concept for this technique, we designed a pilot shRNA pool targeting 4 classes of genes: i) sixteen genes that have been shown to cause assorted sperm development problems when knocked out (BAX, CSF, KIT, PIN1, CPEB1, GNPAT, MLH3, SPO11, CIB1, MAP7, PYGO2, TBPL1, SH2B1, TSN, SIRT1 & VDAC3); ii) five genes that have been characterized in knockout mice but have not been linked to spermatogenesis defects (MMP3, SYT4, TFF3, TNFSF4, TYRP1); iii) three genes that have no knockout mice made and are not expressed in mouse testis (APOC4, LCE1I, SCRG1); and iv) one gene that causes spermatogenesis failure when overexpressed but not reported to affect spermatogenesis when knocked out (VAMP7) **(Table 2.1a)**. Our pool contained one hundred and nineteen unique RNAi with a mode of five RNAi per targeted gene.

Instead of miRNA, I used small hairpin RNA (shRNA) expressing DNA sequences to induce knockdown of genes. By integrating the expression cassette into the genome, I got stable expression of the RNAi construct in transfected cells. I experimented with two different transfection methods, Tris-HCl with naked DNA and lentiviral infection. A single injection of a high titer (10^9 Tu/ml) lentivirus produced a testis with a higher infection rate than an injection of Tris-HCl with 15µg of DNA. However, the single injection of lentivirus produced an infection rate was lower than five injections of the Tris-HCl DNA **(Table 2.S1)**. Due to the lower infection rate, I observed low reproducibility and more dropout of shRNA samples in the single DNA and viral injection testes samples compared to the five DNA injection testes **(Data not shown)**. Since the cost of performing five DNA injections is significantly lower than even one high titer

lentivirus injection (about $50 versus $250), I decided to proceed with that method rather than attempt multiple lentiviral injections.

shRNA	effect	p.value
BAX.1	1.491617	0.000233
BAX.2	0.644068	0.013938
BAX.3	0.685645	0.168694
BAX.4	-0.31448	8.51E-06
BAX.5	0.505972	0.093362
CSF1.1	0.407182	0.72313
CSF1.2	-0.07141	0.001372
CSF1.3	0.89958	0.001235
CSF1.4	-0.34712	2.03E-05
CSF1.5	-0.14532	0.000956
KIT.1	-0.79134	8.43E-07
KIT.2	0.528862	0.163166
KIT.3	0.020387	0.00143
KIT.4	0.602483	0.124486
PIN1.1	0.953665	0.005774
PIN1.2	0.018609	0.000796
PIN1.3	0.998499	0.000163
PIN1.4	0.477457	0.838727
PIN1.5	-0.28984	1.27E-05
CPEB1.1	0.704452	0.014293
CPEB1.2	0.129264	0.005881
CPEB1.3	-0.12752	0.000212
CPEB1.5	-0.38412	6.24E-06
GNPAT.1	-0.28982	1.97E-05
GNPAT.2	-0.12214	6.43E-05
GNPAT.3	-0.20153	0.0002
GNPAT.5	0.050146	0.001491
MLH3.1	0.910026	0.001184
MLH3.2	-0.27059	0.000129
MLH3.3	0.008727	0.003498
MLH3.4	0.614311	0.058491
MLH3.5	-0.49274	6.06E-06
SPO11.1	0.857416	0.030513
SPO11.2	0.47177	0.734457
SPO11.3	-0.16073	5.17E-05
SPO11.4	-0.38517	1.57E-05
SPO11.5	-0.29551	2.83E-05
CIB1.1	-0.17637	3.64E-05
CIB1.2	0.80193	0.005022
CIB1.3	0.010345	0.001795
CIB1.4	0.319618	0.912368
CIB1.5	0.750991	0.002993
MTAP7.1	0.663119	0.016325
MTAP7.2	1.050326	0.000131
MTAP7.3	0.542848	0.477663
MTAP7.4	0.670612	0.352283
MTAP7.5	-1.24653	5.12E-07
PYGO2.1	0.524609	0.208095
PYGO2.2	0.051773	0.004198
PYGO2.3	-0.31263	2.29E-05
PYGO2.4	0.111512	0.024783
PYGO2.5	0.512162	0.151658
TBPL1.1	0.777065	0.014903
TBPL1.2	-0.51976	2.61E-06
TBPL1.3	0.195934	0.04384
TBPL1.4	0.231028	0.060116
TBPL1.5	0.518287	0.446464
SH2B1.1	0.765325	0.002878
SH2B1.2	0.933765	0.000451
SH2B1.3	0.769163	0.002822
SH2B1.4	1.154908	0.000446
SH2B1.5	-0.17093	6.43E-05
TSN.1	-0.25881	9.15E-05
TSN.2	0.496608	0.052717
TSN.3	0.021181	0.001147
TSN.4	0.411955	0.834017
TSN.5	-0.07945	0.000632
SIRT1.1	0.25307	0.098167
SIRT1.2	-0.44504	7.50E-06
SIRT1.3	-0.30681	9.52E-06
SIRT1.4	-0.07802	0.000331
SIRT1.5	-0.06165	0.000355
VDAC3.1	0.307419	0.182138
VDAC3.2	0.005379	0.001009
VDAC3.3	-0.73389	2.43E-06
VDAC3.4	0.821601	0.005827
VDAC3.5	0.430334	0.977149
APOC4.1	0.519536	0.597742
APOC4.3	0.447127	0.213038
APOC4.4	0.831132	0.134781
APOC4.5	0.330404	0.861175
APOC4.6	0.650361	0.283096
LCE1I.1	0.476054	0.162711
LCE1I.2	0.651527	0.040335
LCE1I.4	0.845046	0.005963
LCE1I.5	0.761141	0.019769
MMP3.1	-0.59078	2.08E-06
MMP3.2	0.180992	0.194799
MMP3.3	0.686643	0.013764
MMP3.4	0.451533	0.583173
SCRG1.1	0.544988	0.091899
SCRG1.2	1.06915	0.002892
SCRG1.3	0.419343	0.995189
SCRG1.4	0.553699	0.106069
SYT4.1	0.45195	0.276389
SYT4.2	0.531955	0.04626
SYT4.3	0.273697	0.726521
SYT4.4	0.05074	0.00406
SYT4.5	0.683667	0.032538
TFF3.1	0.82922	0.162711
TFF3.2	0.343532	0.844622
TFF3.3	0.829876	0.012541
TFF3.4	0.116672	0.017221
TFF3.5	0.176176	0.093067
TNFSF4.1	-0.53715	1.87E-06
TNFSF4.2	0.580482	0.026477
TNFSF4.3	0.450666	0.679552
TNFSF4.4	-0.04875	0.00037
TNFSF4.5	0.353803	0.692855
TYRP1.1	0.061815	0.043682
TYRP1.2	0.145296	0.009266
TYRP1.3	0.275121	0.923136
VAMP7.1	-0.36238	9.46E-06
VAMP7.2	0.270951	0.254401
VAMP7.3	0.820457	0.001899
VAMP7.4	0.228109	0.090742
VAMP7.5	-0.52979	5.93E-06
VAMP7.6	0.299056	0.657602
VAMP7.7	0.407084	0.566728

Table 2.1a: Pilot shRNA pool screen effect sizes and p values (testis)

A combined survival/differentiation selection pressure was applied on the germ cells in the testis by waiting for a sufficient length of time (20 days or slightly longer than half of a murine spermatogenesis cycle) after transfection before quantification. I assumed that any DNA injected into a mouse testis would not transfect the other testis because it would first have to pass through a large part of the rest of the body via the circulatory system. Indeed, the Pearson correlation of the shRNA pools between different testes in the same mouse was not significantly different than between testes from different mice. This allowed us to use the two testes in a mouse as different biological replicates or even to test different shRNA pools, halving the mouse requirements for our experiments.

To analyze the data, I used a non-parametric Wilcoxon Rank Sum Test. I expected that many shRNAs would be non-functional, but some of the shRNAs against important spermatogenesis genes would be effective and knockdown expression, causing the cell to either arrest developmentally or undergo apoptosis. Those shRNAs would be depleted in the testis relative to non-functional shRNAs. Naively, you would compare the fold change of the shRNA across biological replicates and compare it to the overall population fold changes. However, that would rely on the ratio of effective shRNAs to the non-functional shRNAs to remain low. I instead spiked in a small number of negative control shRNAs into the pool to act as our miner's canary; I compared the fold change of any shRNA against only the negative control shRNA fold changes. Due to the low transfection efficiency (1-3%) **(Table S1)**, multiplicity of infection, a common issue in RNAi screens, was not a worry in the analysis.

Of the sixteen genes that have been reported to cause sperm development problems when knocked out in mice, twelve of them had at least two shRNAs be significantly depleted ($P \leq 0.01$) in the testis compared to the injected pool. This produces a 25% false negative rate (4/16). One of the eight negative control genes [groups (ii) and (iii)], TNFSF4 (moderately expressed in the testis), also passed our screen, producing a false positive rate of 12.5% (1/8). **(Figure 2.2)**

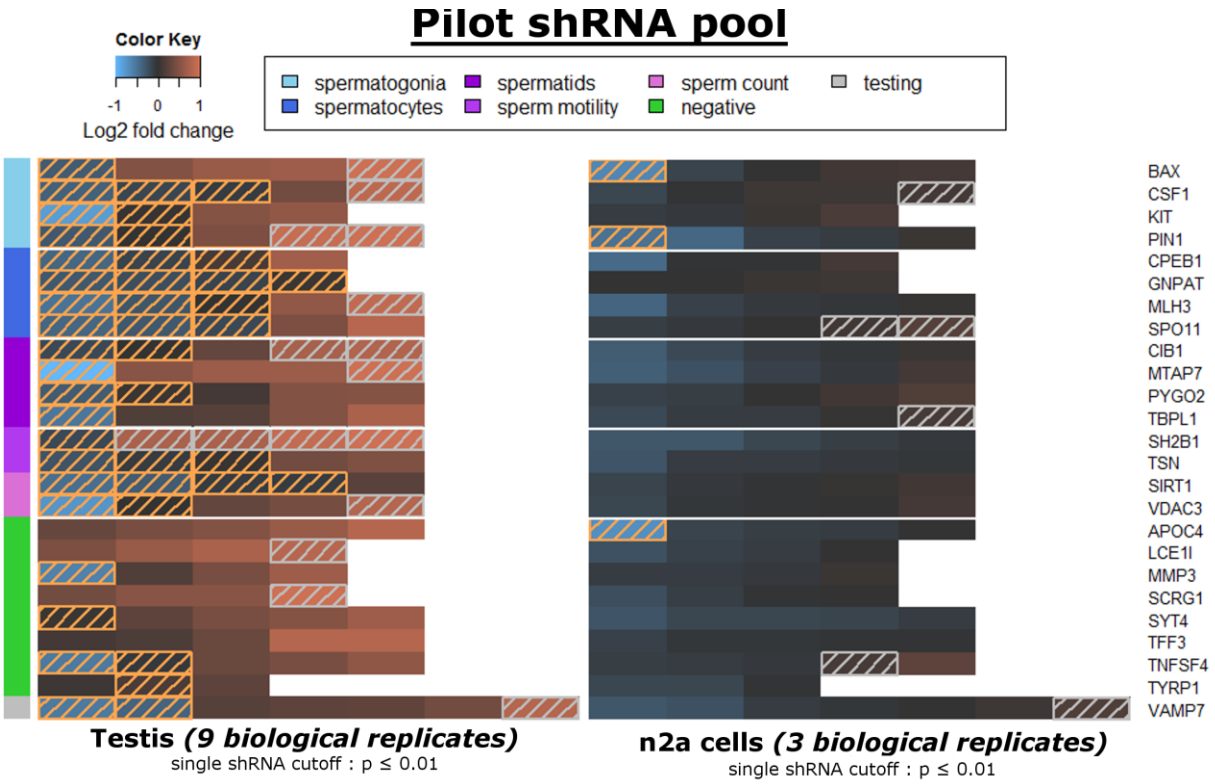


Figure 2.2: Pilot shRNA pool screens results
Each cell shows the log_2 fold change of a shRNA against the target gene on the right from what was injected until after incubation in the mouse testis. Each row is arranged in ascending order based on the log_2 fold change. If a cell is shaded, it means that it passes the significance threshold, which is annotated below the respective figure. Cells are shaded orange for significant depletion and white for significant enrichment. Predicted gene function on different stages of spermatogenesis is annotated to the left using a color bar.

VAMP7, has been shown to cause spermatogenic failure when overexpressed in mice[24], and the two studies examining the knockout mouse focused more on behavior and neurons than the reproductive system[25,26]. Because of these lines of evidence, I thought it was possible that

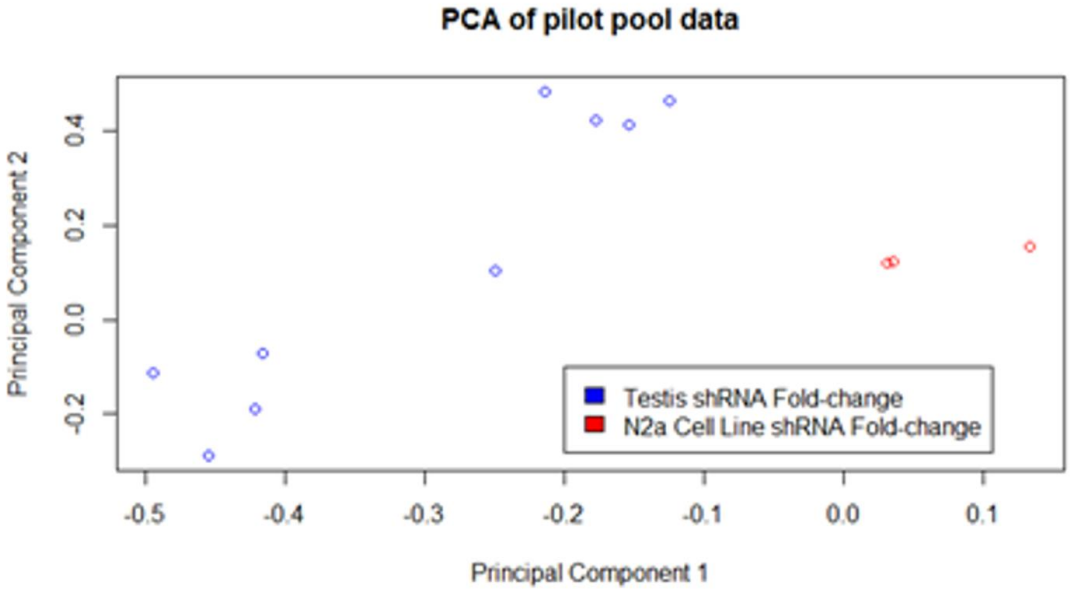
VAMP7 might also cause subfertility when knocked out in mice. VAMP7 did pass our knockdown screen, providing strong evidence for that hypothesis.

Our criterion of requiring any given gene to have two different shRNAs to be significantly depleted might limit how low the false negative rate can go, since if a given gene has only one effective shRNA in the pool it can never pass. Indeed if we look at the false negatives in our pool, all of them (BAX, MTAP7, and SH2B1) simply had only one shRNA that was depleted (albeit very strongly). However, if we were to remove this requirement, our false positive rate would drastically increase to 50% (4/8). Since this is a method for large scale screening, missing some true positives is preferable to implicating excess false positives. Another way of resolving this issue may be to increase the average number of shRNAs per gene (5) that is used in the pool.

One drawback of this approach is that both genes that are important for spermatogenesis and genes that are required for cell survival will be highlighted. By transfecting the same shRNA pool into an unrelated cell line and performing a survival screen on the cells, one should be able to highlight only genes needed for survival. A simple elimination filter on the original highlighted group of genes will then produce a list of genes that only affect spermatogenesis.

I transfected three separate wells of Neuro-2a cells (N2a), prepared sequencing libraries, and analyzed the data through the same pipeline as the testis samples **(Methods)**. The normalized read counts of the shRNA pool before and after the transfection and incubation was not well correlated with the testis read counts **(Figure 2.3)**. Furthermore none of the shRNAs that were significantly depleted in the testes samples were also significantly depleted in the cell line

(Figure 2.2). I thus concluded that the genes with multiple shRNAs depleted in testes affect spermatogenesis and not survival.



Pilot shRNA pool (Spearman Correlations)

	Testis	N2a cells
Testis	0.6639 – 0.9628	-0.1940 – -0.8340
N2a cells	-0.1940 – -0.8340	0.7393 – 0.9238

Figure 2.3: Clustering and correlations fold changes of the shRNAs across testes and cell lines

Benchmarking the performance

The first question with RNAi screening experimental designs is how many biological replicates to use. Increasing the number of biological replicates enables the discovery of genes with subtler effects, but there must be a point of diminishing returns for any given effect size, where more biological replicates do not significantly increase discovery power. I found that with more stringent significance thresholds and less biological replicates, the minimum detectable effect size decreased. For a lenient ($P \leq 0.1$), standard ($P \leq 0.05$), and stringent($P \leq 0.01$) p value

thresholds there was little improvement in detectable effect size beyond 3-4, 4-5, and 6-7 biological replicates respectively **(Figure 2.4)**. In order to reduce the chance of off-target effects, I required any given gene to have two different significantly depleted shRNAs before the gene was determined to have a spermatogenesis function. Thus, the P value for the lenient cutoff was really 0.01 (0.1^2) and the stringent cutoff was 0.0001 (0.01^2) (each shRNA works independently of each other in different cells, so the likelihood of two of them being depleted is the square of the cutoff).

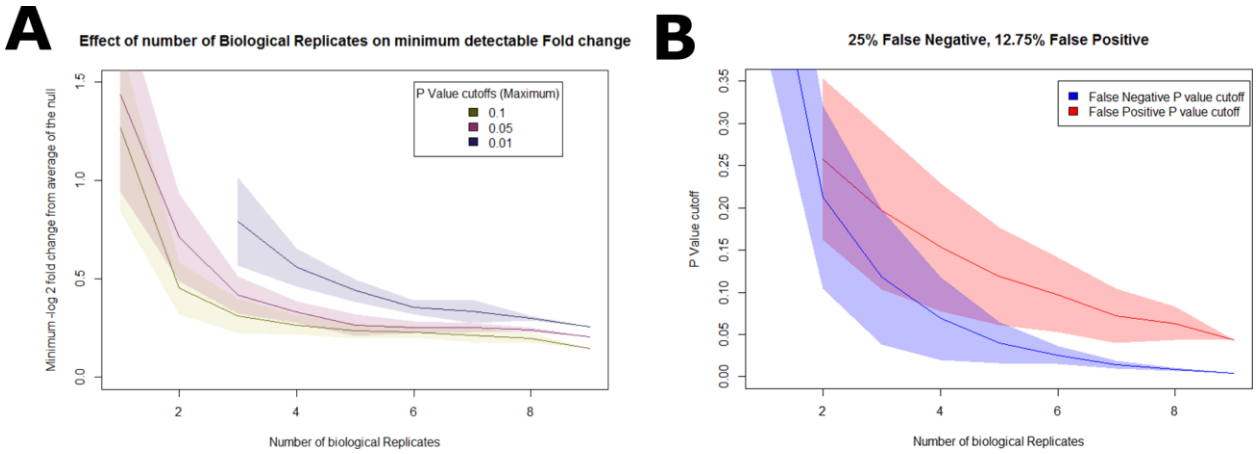


Figure 2.4: Screening method performance benchmarks
Figure A shows the effect of biological replicates on minimum detectable shRNA fold change. The yellow line plots the median minimum log_2 fold change that was observed by any shRNA that passed the p value threshold of $p \leq 0.1$ in the pilot pool for varying amounts of biological replicates. In the lighter shaded area it shows 1 standard deviation from that observed minimum log_2 fold change. In purple shows a similar plot for $p \leq 0.05$ and in blue shows the plot for $p \leq 0.01$. Note that the blue starts at 3 biological replicates because there were no observations that passed that threshold before then. B shows the p value cutoffs for varying number of biological replicatesIn red there is the highest p-value (non-inclusive) that can be used to maintain the false positive rate, in blue is the lowest p-value (inclusive) that can be used to maintain the false negative rate. The shaded area behind each line represents the standard deviation for the cutoff values.

Another reason to use more biological replicates is to minimize the false positives and false negatives. Among the significantly depleted shRNAs, I determined the largest p value of a called

gene and the lowest p value from an uncalled gene with varying numbers of biological replicates **(Figure 2.4)**. The gap between the two values is an indicator of the likelihood of obtaining more false positives and negatives; a larger gap indicating a lower likelihood. Six replicates was the minimum number to avoid any overlap between the two p-value cutoffs within 1 standard deviation.

Based on these analyses, I determined that 6-7 biological replicates were the optimal amount. It was possible to use fewer biological replicates, but less stringent criteria needed to be applied. Importantly, the fold change of each shRNA remained consistent across the biological replicates, but increasing the biological replicates allowed us to call smaller fold enrichment scores as significant. Using fewer biological replicates also sometimes slightly increased the number of false positives and/or negatives (depending on the combination of replicates used).

Screening for uncharacterized predicted fertility genes

Next I implemented this screening technique for the top candidate spermatogenesis genes that I had previously identified[23] **(Methods)**. In this new screen I used a pool comprising of one hundred and thirty shRNAs against twenty-six candidate genes and fifteen of the previously validated negative control shRNAs. Of the twenty-six candidate genes, there were four genes which are expressed early in spermatogenesis (ALPI, POLA1, RFC1, RRM1), fifteen genes which start expression in the middle of spermatogenesis (CRISP2, GSTM5, HRASLS5, KLHDC3, LDHAL6B, PGAM2, PHF7, PHKG2, RFC2, SFI1, SPATA4, TAF9, TCP1, TCP11, ZMYND10), and seven genes which are only expressed late in spermatogenesis (4933411K16Rik, ACTL7B, GSG1, MEA1, SPA17, SPZ1, UGT1A1) **(Table 2.1b)**. I visualized

the functional relationships of these genes with other known spermatogenesis genes **(Figure 2.S2)**.

shRNA	effect	p.value
Actl7b.1	0.015889	0.000258
Actl7b.2	-0.45539	1.09E-05
Actl7b.3	-0.35963	1.15E-05
Actl7b.4	-0.1669	1.68E-05
Actl7b.5	-0.31619	1.15E-05
Alpi.1	1.099163	0.057542
Alpi.2	-0.46416	1.09E-05
Alpi.3	0.197698	0.002453
Alpi.4	0.193738	1.42E-02
Alpi.5	-0.4806	1.09E-05
Crisp2.1	-0.16214	3.94E-05
Crisp2.2	-0.82978	1.03E-05
Crisp2.3	0.342999	0.118144
Gsg1.1	-0.14535	5.10E-05
Gsg1.2	-0.1723	6.59E-05
Gsg1.3	0.569979	0.463421
Gsg1.4	0.419867	0.112592
Gsg1.5	0.272286	0.039837
Gstm5.1	0.172666	3.23E-03
Gstm5.2	0.007751	0.000214
Gstm5.3	-0.10902	5.95E-05
Gstm5.4	0.035054	0.000224
Gstm5.5	0.157049	2.01E-03
Hrasls5.1	0.047816	4.48E-04
Hrasls5.2	-0.07538	6.59E-05
Hrasls5.3	0.642644	0.932945
Hrasls5.4	0.593571	0.763794
Hrasls5.5	0.110306	0.002265
Mea1.1	0.307722	0.052962
Mea1.2	0.885945	0.290161
Mea1.3	0.061647	0.001927
Mea1.4	-0.62207	1.03E-05
Mea1.5	0.83258	0.295677
Pgam2.1	-0.08581	0.000204
Pgam2.2	0.45828	3.74E-01
Pgam2.3	0.505186	3.61E-01
Pgam2.4	0.465959	4.63E-01
Pgam2.5	-0.80073	1.03E-05
Pgam2.6	-0.05413	0.000109
Pgam2.7	1.217661	0.010103
Pola1.1	-0.09564	4.61E-05
Pola1.2	-0.19154	4.61E-05
Pola1.3	0.333183	0.046006
Pola1.4	0.416431	0.198391
Pola1.5	0.109785	0.001507
Rfc4.1	0.290864	0.012413
Rfc4.2	0.308573	0.033374
Rfc4.3	0.300248	0.038691
Rfc4.4	-0.26783	1.28E-05

shRNA	effect	p.value
Rfc4.5	-0.36645	1.15E-05
Sfi1.1	0.410412	0.152605
Sfi1.2	-0.85623	1.03E-05
Sfi1.3	-0.60622	1.03E-05
Sfi1.4	-0.58489	1.09E-05
Sfi1.5	0.212698	0.051504
Spa17.1	-0.41977	1.78E-05
Spa17.2	0.328957	0.079271
Spa17.3	0.136686	0.009098
Spa17.4	-0.37063	1.21E-05
Spa17.5	-0.6222	1.03E-05
Spata4.1	0.722841	0.441724
Spata4.2	0.022258	0.000297
Spata4.3	-0.39323	1.15E-05
Spata4.4	-0.46324	1.09E-05
Tcp1.1	0.265696	0.016216
Tcp1.2	-0.29274	1.35E-05
Tcp1.3	-0.36721	1.21E-05
Tcp1.4	-0.16012	1.59E-05
Tcp1.5	-0.03193	0.000109
Tcp11.1	-0.23818	4.85E-05
Tcp11.2	0.159007	0.004732
Tcp11.3	-0.81735	1.03E-05
Tcp11.4	0.168265	0.007352
Tcp11.5	0.93258	0.092411
4933411K16Rik.1	-0.03061	0.000177
4933411K16Rik.2	0.081464	0.000833
4933411K16Rik.3	-0.55283	1.09E-05
4933411K16Rik.4	0.565777	0.84749
4933411K16Rik.5	0.567327	0.324306
Klhdc3.1	0.282169	0.023835
Klhdc3.2	0.134523	0.005099
Klhdc3.3	0.1164	0.002453
Klhdc3.4	0.161291	0.004389
Klhdc3.5	0.096172	0.004913
Klhdc3.6	0.817036	0.373742
Ldhal6b.1	-0.37889	1.15E-05
Ldhal6b.2	-0.56104	1.03E-05
Ldhal6b.3	0.368932	0.238809
Ldhal6b.4	0.448284	0.21569
Ldhal6b.5	0.487051	0.301263
Ldhal6b.6	1.212021	0.01518
Phf7.1	-0.0998	2.59E-05
Phf7.2	-0.13254	1.88E-05
Phf7.3	0.026095	0.00067
Phf7.4	-0.06238	9.36E-05
Phf7.5	0.167073	0.001705
Phkg2.1	-0.18144	2.88E-05
Phkg2.2	-0.8624	1.03E-05

shRNA	effect	p.value
Phkg2.3	-0.52569	1.03E-05
Phkg2.4	0.278501	0.019082
Phkg2.5	0.50972	0.772978
Rfc2.1	0.196714	0.014684
Rfc2.2	0.643957	0.952076
Rfc2.3	0.111148	0.001851
Rfc2.4	-0.0854	8.91E-05
Rfc2.5	0.585693	0.980821
Rrm1.1	-0.12047	5.66E-05
Rrm1.2	0.464185	0.206909
Rrm1.3	0.337198	0.064156
Rrm1.4	0.482805	0.301263
Rrm1.5	-0.37315	1.21E-05
Spz1.1	-0.10458	9.84E-05
Spz1.2	0.229612	0.008783
Spz1.3	0.883214	0.243635
Spz1.4	-0.35202	1.09E-05
Spz1.5	0.040797	0.000764
Taf9.1	0.073342	0.000428
Taf9.2	0.925805	0.020348
Taf9.3	0.054157	0.002265
Taf9.4	0.12965	0.107247
Taf9.5	-0.32276	1.35E-05
Ugt1a1.1	-0.29293	1.28E-05
Ugt1a1.2	-0.64813	1.03E-05
Ugt1a1.3	0.959077	0.079271
Ugt1a1.4	0.650001	0.99041
Ugt1a1.5	-0.39812	1.09E-05
Zmynd10.1	0.442424	0.224737
Zmynd10.2	-0.79687	0.000469
Zmynd10.3	-1.03077	6.26E-05
Zmynd10.4	-0.35061	1.21E-05
Apoc4.1	0.784977	0.116731
Apoc4.2	0.509159	0.643522
Apoc4.3	1.430087	0.005195
Apoc4.4	0.677533	0.364126
Apoc4.5	0.177995	0.032876
Syt4.1	0.43407	0.321357
Syt4.2	0.713099	0.266192
Syt4.3	0.106947	0.014929
Tff3.1	1.22015	0.015432
Tff3.2	0.765601	0.678362
Tff3.3	0.729007	0.28201
Tff3.4	0.200792	0.045352
Tff3.5	0.29828	0.213464
SHC202	0.621188	0.805361
SHC216	0.59584	0.438158

Table 2.1b: Predicted genes' shRNA pool screen effect sizes and p values (testis)

Similar to the previous pool, we also transfected this pool into N2a cells to eliminate the possibility that these genes are required for general cell survival. Only two genes (MEA1 and TAF9) had at least two shRNAs that were significantly depleted ($P \leq 0.01$) **(Figure 2.5)**. Overall, the normalized read counts of the shRNA pool in N2a cells were not well correlated with the testis read counts **(Figure 2.6)**. Given this, we concluded that most of the genes with multiple shRNAs depleted in testes (24/26) affect spermatogenesis and not survival.

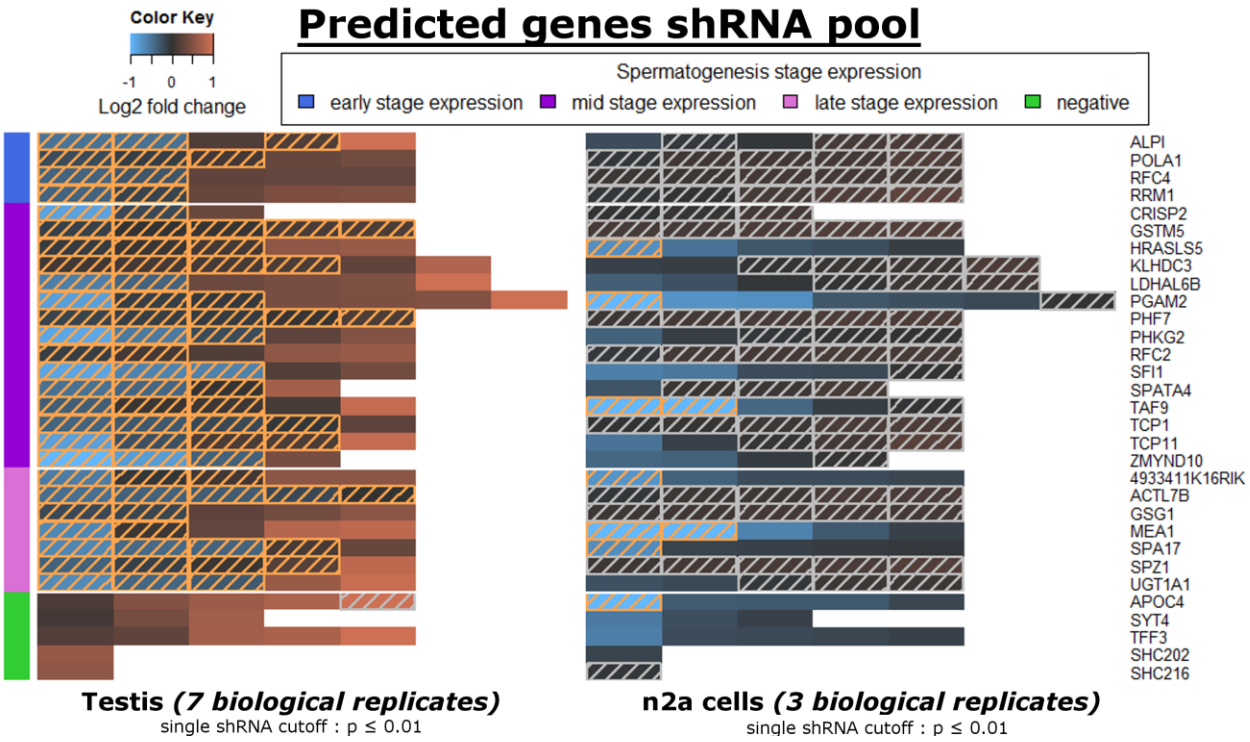
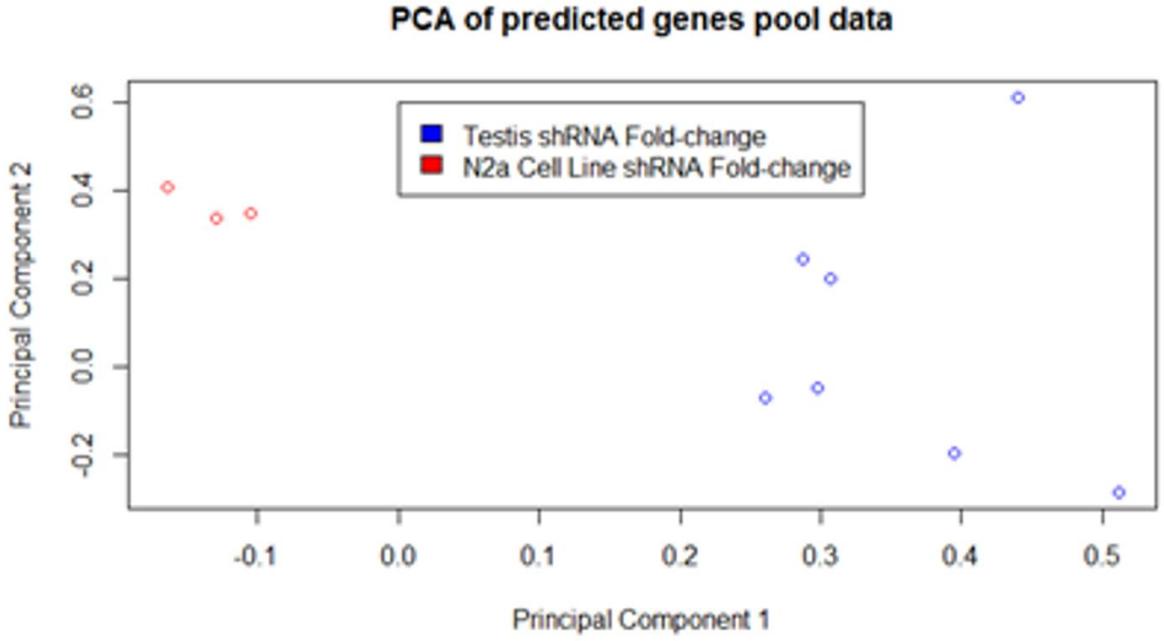


Figure 2.5: Predicted genes' shRNA pool screens results
See Figure 2.3 for plot explanation

Given that I tested the top candidates in a list of close to a thousand, it is not too surprising that most of the candidate genes passed the screen. Furthermore, most of the genes are predicted to have functions that are closely related to meiosis. Seven genes are thought to be involved in cell cycle (ACTL7B, KLHDC3, POLA1, RFC2, RFC4, RRM1, SFI1), seven in metabolism (ALPI, GSTM5, HRASLS5, LDHAL6B, PGAM2, PHKG2, UGT1A1), three transcription factors

(OHF7, SPZ1, ZMYND10), two protein folding (TCP1, TCP11), one cell binding in sperm (SPA17), and one involved in capacitation of the membrane (CRISP2).



Predicted genes shRNA pool (Spearman Correlations)

	Testis	N2a cells
Testis	0.5493 – 0.9555	-0.1915 – -0.6617
N2a cells	-0.1915 – -0.6617	0.9593 – 0.9848

Figure 2.6: Clustering and correlations fold changes of the shRNAs across testes and cell lines

I performed a functional pathway analysis of these genes and found four functional networks. All twenty five tested genes were linked to the twenty associated genes in the co-expressed network, which tests for similar expression levels across different conditions in various published Gene Expression Omnibus (GEO) datasets. In the predicted network, which uses protein interactions and orthologous functional relationships from other organisms I found three isolated networks. The largest network links sixteen of the tested genes with seventeen other genes, fourteen of

which are annotated with GO terms related to sperm function. The next largest network groups six of the genes together. Four of the genes were not functionally associated with any other sperm function gene. The co-localized network identifies genes expressed in the same tissue. This produced a large network linking thirteen of the tested genes with fifteen associated genes and two smaller networks linking three and two tested genes respectively. Lastly I had the shared protein domain network which links genes if their product has a common protein domain. This was the sparsest network with three tested genes linked to each other and two tested genes linked to one spermatogenesis associated gene each. **(Figure S2)**.

Chapter 2 - Discussion

The performance of our direct *in vivo* screen is comparable to other benchmarked RNAi screens performed in other model systems. Our false negative rate of 25% is up to the standards of various studies benchmarking RNAi screens for different pathways in *drosophila melanogaster*[27–30] where it was reported to be between 13% and 50%. Our false positive rate of 12.5% also compares favorably to the same study[30] which found their false positive rates to be between 7% and 18%.

Over half (65/121) the shRNAs in the pilot pool and about a third of the shRNAs (48/145) in the predicted gene pool were reported by Sigma-Aldrich to be validated in various cell lines **(Supplemental Table 2.1 & 2.2)**. There was at least one validated shRNA for 72% (18/25) of the pilot pool genes and 45% (13/29) of the predicted pool genes. While not all the validated

shRNAs were consistently significantly depleted in the two studies, many of them were, giving us confidence that the signal I observed was not caused by off-target effects.

Going forward, I think that it will be important to create a method to verify spermatogenesis genes without having to make knockout mice. I am working on adapting the transfection protocol to accomplish this. I am also interested in looking at the functions of the twenty one genes identified by this screen, especially the three genes of unknown molecular function (4933411K16Rik, GSG1, and SPATA4). These genes could work in novel pathways, providing new insights about spermatogenesis.

I was extremely conservative with the multiplexity of the pool in this study. In order to produce reproducible signals I ensured that the number of cells was orders of magnitude larger than the number of shRNAs. In this manner I could be confident that no shRNA would be underrepresented in the final pool due to transfection efficiency. Even with the relatively low transfection rate, I was able to get consistent signals using pools consisting of up to approximately 150 shRNAs. It required about 2% of the total genomic DNA extracted from a testis to make a library and MiSeq to sequence the libraries. It is conceivable to scale up the number of shRNAs in the pool up to ten times without changing any other parameters other than using ten times more of the genomic DNA to make sequencing libraries and changing to the HiSeq system for sequencing which will produce ten times more sequencing reads in one sequencing lane. Taking advantage of this increase in throughput, one could both increase the number of genes screened and increase the number of shRNAs used per gene to potentially reduce the false negative rate.

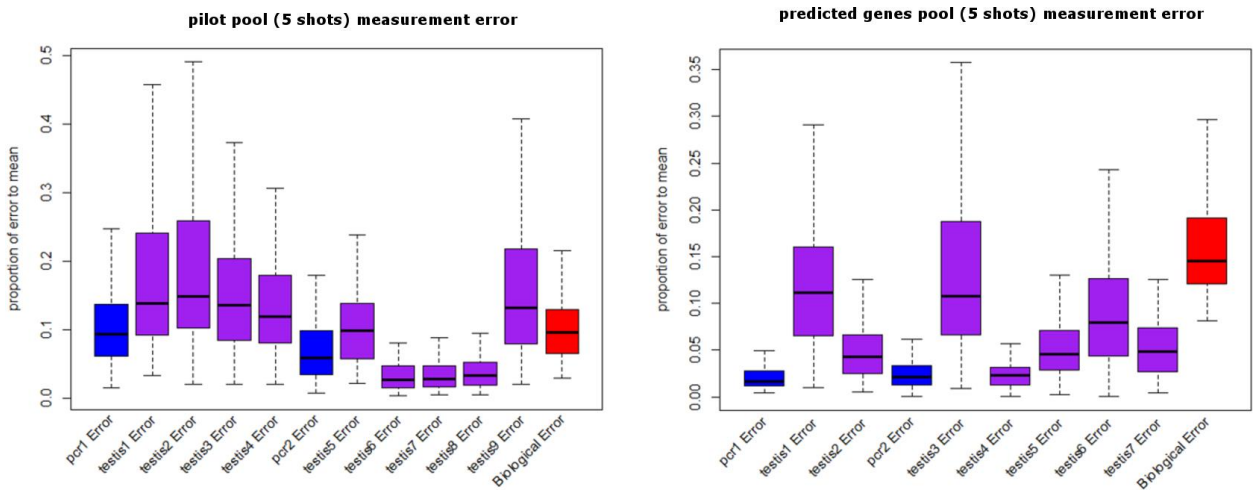


Figure 2.7: Technical and biological noise in the shRNA pools

This visualizes standard error as a proportion of the median normalized count values of each shRNA within each sample in box plot form. In blue are the errors for libraries prepared from the PCR product injected into the testes. Due to the large numbers of shRNA, this is a good quantification of the technical noise. In red are the errors for the median counts across all testes samples, a measure for the biological noise. In purple are the errors within a testis sample, measuring the sum of the biological and technical noise.

To make this protocol accessible I used a total of 31 cycles of PCR to prepare the sequencing libraries. This produced nearly two orders of magnitude more material than was required for sequencing on one Illumina lane, raising concerns about overamplification. However I found that the technical noise inherent in each sample was within the same scale as the noise between biological samples **(Figure 2.7)**, meaning it does not cause too much problem in the downstream analyses. If technical noise is a concern, it is possible to reduce the number of PCR cycles in the first step of library preparation and still have sufficient amounts of DNA for sequencing. However a more sensitive method of DNA quantification such as a Bioanalyzer chip will be required in that case.

I made an attempt to decipher the functional effects of the genes via this same technique by using cell stage and functional separation (FACS and sperm motility assays respectively). Unfortunately because the number of cells I could retrieve in this manner was limited, I was unable to prepare sequencing libraries from the subpopulations. If I could transfect a majority (60-80%) of the cells (perhaps by multiple lentivirus injections), or if it was possible to sort and retrieve large numbers of cells (10s of millions), direct functional assays using an RNAi pool may be possible. More technically challenging protocols such as efferent duct injection or in-vitro organogenesis may be required to enable such future experiments.

For other groups intending to use this technique, I would like to provide some words of caution. Firstly, the fold changes are a relative measure of the strength of the RNAi relative to each other RNAi in the pool. Furthermore, fold change convolutes gene functional effect and RNAi knockdown efficiency. As such, I would advise against using the extent of the fold change of a RNAi to rank gene functional effects. Finally, the analysis I used requires some known negative controls spiked into the pool. I recommend using the ones against the 7 negative control genes that did not produce false positives in the pilot pool. Our experimental protocol (**Supplementary Protocol 2.1, 2.2**) and analysis pipeline are provided to make it easy for any other interested groups to try this technique on shRNA pools of their own design against genes of their interest. Based on our benchmarks, I would recommend that they use at least 6-7 biological replicates for good confidence in their results.

Via the traditional method of making knockout mice to validate gene function, it would have been unreasonably time consuming and expensive to test all twenty-six candidate genes, despite

them being the very top candidates in a list of near a thousand genes. Using our screening technique I was able to quickly produce experimental evidence for their functional effect. I expect this technique to be useful to help narrow down the large lists of genes that will be generated from large scale exome studies of infertility that are currently underway.

Chapter 2 - Methods

Gene Selection

For the initial pool, I used data from the Mouse Genome Database[31] from Jackson Labs (MGI) to create a list of genes that affect the male reproductive system when knocked out. I then used a list of genes that have been knocked out and not reported to cause any male reproductive defects to use as negative controls.

For the predicted spermatogenesis gene pool, picked the top 30 candidates from each of the mouse predicted infertility gene models[23] and filtered it to keep only the genes that were not reported in MGI to have any knockout mice line made. I then used shRNAs against three of the known negative genes from the pilot pools together with two scrambled non-mammalian sequences to use as negative controls.

shRNA Pool Preparation

I used RNAi from the MISSION® TRC-Mm 1.5 and 2.0 (Mouse) obtained from Sigma-Aldrich **(Supplemental Table 2.1 & 2.2)** ordered as shRNA plasmids. The shRNA expression cassette from the plasmids was amplified from the plasmid pool by PCR and purified using AMPure XP beads **(Supplementary Protocol 2.2)**. These purified amplicons were pooled and then used for

injection into mouse testis and transfection into cell lines. An aliquot of this mixture was sequenced on MiSeq to determine initial shRNA pool composition.

Mouse Testis Transfection

I performed the experiments using C57BL/6 mice generated in house between 28-32 days of age. All mice were maintained under pathogen-free conditions and all animal experiments were approved by Washington University's Animal Studies Committee. Each mouse received bilateral intra-testicular DNA injections five times, spaced 3-4 days apart **(Supplementary Protocol 2.1)**. Following the injections, the mice were allowed to recover until 20 days after the third injection, when testes were dissected. Genomic DNA from the whole testis was extracted using Qiagen DNeasy Blood and Tissue kit.

Cell Line Transfection

N2a cells were maintained in Minimum Essential Media (MEM) supplemented with 10% Fetal Bovine Serum (FBS) until they reached 60% confluency in 6 well cell culture plates. Each well of cells was transfected with 2.5µg shRNA pool DNA using Lipofectamine® 3000 (Life Technologies) following manufacturer's instructions. The cells were then incubated at 37°C for 2 days, with daily media replacement, before the genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit.

Illumina Sequencing Library Preparation

I used a custom protocol to amplify the shRNA sequences in the genomic DNA samples **(Supplementary Protocol 2.2)**. This protocol used 2 rounds of PCR amplification to prepare the

sequencing library instead of ligation followed by PCR amplification. I started with 2µg of genomic DNA to survey the genomes of enough cells in order to reduce the likelihood of dropout or PCR jackpotting; common artifacts when testing low numbers of cells.

Each biological sample had between three to five separate sequencing libraries prepared with different indices using different aliquots of genomic DNA to quantify technical noise. Sequencing libraries were then pooled and run in a lane of Illumina MiSeq 2x150bp to obtain an average of at least 3,000 reads per unique shRNA in the library.

Statistical Analysis

Mapping of reads to shRNAs was done by aligning each read to the unique half of the hairpin sequence with no mismatches. A table of read counts for each shRNA was generated to determine significant enrichment/depletion. There was no significant difference in the counts between paired-end reads when they were mapped separately. I minimized technical noise by using the median value of technical replicates as the true count for each shRNA.

To determine significant depletion/enrichment of shRNAs, I used a custom R script. I started by normalizing the shRNA count data to number of reads per shRNA per million reads in the sequencing library. I then calculated the log 2 fold enrichment of each shRNA in the testis relative to the initial DNA pool. The fold changes of different experiments using the same shRNA pool design were always normalized to the sequencing counts of the actual injected material. These fold changes were then merged to produce more biological replicates for a given shRNA pool design. Finally, I performed a Wilcoxon Rank Sum Test for each shRNAs' fold

enrichment across biological replicates against the fold enrichment of shRNAs against genes which are not known to affect spermatogenesis to calculate the likelihood that the shRNA was significantly depleted or enriched compared to the null. Any shRNA that had a p value smaller than the cutoff was determined to be significant.

Network Analysis

I used Cytoscape[32] with the GeneMANIA plugin[33] with the default settings to visualize the functional network of the tested predicted genes in the supplement.

Supplementary Protocol 2.1: Mouse Testis DNA transfection

Materials Needed:

- 1M Tris-Hcl pH7.0
- Pure, DNase and RNase free water
- DNA to be injected
- Anasthesia*
- 29 gauge Insulin Needle (Terumo: SS10M2913)
- 701N Syringe, Cemented Needle, 26s Gauge (Hamilton: 80300)
- 70% Ethanol (denatured is fine)
- 100% pure Ethanol
- Kimwipes

Step 1: Prepare the DNA mixture

If using linear DNA

Dilute the DNA in Tris-Hcl and water to get 15µg of DNA in 20µl of **150mM** Tris-HCL pH 7.0

If using circular DNA

Dilute the DNA in Tris-Hcl and water to get 15µg of DNA in 20µl of **125mM** Tris-HCL pH 7.0

Step 2: Anesthetize the mouse

I used a mixture of *(final concentrations)* Ketamine (10mg/ml), Xylazine (1mg/ml), and Glycopyrrolate (2µg/ml) diluted in sterile PBS. This was injected intraperitoneally using a 29 gauge needle, with 10µl of anesthesia used per 1g of mouse body weight.

Alternative general anesthesia methods such as isoflurane can also be used

Step 3: Injection of material

First, wipe down the inferior torso (where the testis are) with 70% ethanol and a kimwipe. (This lattens the fur and prevents it from interfering with the injection.

Nest, feel for one of the mouse testis and get a good grip on it between your fingers.

Using the 701N syringe, pipette 10µl of the DNA mixture form step 1 and inject it slowly through the skin into the anterior end of the testis. (This should take between 30s to 60s to finish injecting. I find that slower rates of injection lead to better transfection rates) Repeat this with the other 10µl using the same syringe into the posterior end of the testis.

You can repeat the same procedure for the other testis.

Following Injection, tap the testes gently about ten times with your finger to ensure that the DNA mixture is spread throughout the testis.

At the end, clean the syringe by pulling up 100% ethanol through it three times and wiping down the needle using a kimwipe and 100% ethanol.

To prevent contamination, I like to use different 701N syringes for different DNA mixtures. However, the same syringe can be used on multiple different testes if you are using an identical DNA mixture for all of them.

I spaced the injections 3-4 days apart (twice a week) to allow the testis to heal from the prior injection.

Supplementary Protocol 2.2: shRNA pool DNA Preparation (For Injection)

Materials Needed:

- Q5 Hot Start DNA polymerase* [NEB: M0493L]
- 5M Betaine [Sigma-Aldrich: B0300-1VL/ B0300-5VL]
- 10mM dNTPs [Promega: U1511/U1515]
- shRNA pool [Sigma-Aldrich: Mission® shRNA Library]
- PCR Primers: *(5' to 3')*

pLKO U6 shRNA F	GAGGGCCTATTTCCCATGATTCC
pLKO U6 shRNA R	GTGGATGAATACTGCCATTTGTCTC

- Ampure XP beads [Beckman Coulter: A63880/A63881/A63882]
- 70% Ethanol
- Magnetic separation rack for 1.5ml microcentrifuge tubes

**(I have found Q5 to be more sensitive and specific than Taq or Phusion, and this improvement is necessary for amplifying from low amounts of sample.)*

Step 1: Mix shRNA pool

I ordered shRNAs in plasmid form in the 96-well plate format. Taking 5µl of the plasmid each shRNA I wanted in the pool I made a pool of over a hundred shRNAs at a final concentration of around 20ng/µl. This was mixed via a 10 second vortex and spun down briefly to collect all the liquid at the bottom of the tube.

Step 2: PCR Amplification of shRNA pool

Each PCR reaction produces ~ 3-4µg of DNA, meaning around 20 reactions are needed for enough material for 1 testis (15µg/injection X 5 injections = 75µg).

Per reaction, mix:

10µM F and R Primers	0.4µM	2µl
5X Q5 Buffer	1X	10µl
10mM dNTPs	0.2mM	1µl
Q5 Hotstart Polymerase	4 Units	2µl
5M Betaine	1.5M	15µl
DNA sample	~20ng	1µl
Distilled H_2O		19µl
		50µl

PCR Conditions	
98°C for 3 minutes	
98°C for 30 seconds	35 Cycles
60°C for 30 seconds	
72°C for 30 seconds	
72°C for 10 minutes	
4°C hold	

Expected product size is 343bp

Step 3: Clean-up of PCR products

(Adapted from AMpure XP manufacturer's protocol)

1. Pool up to 10 reactions in a 1.5ml microcentrifuge tube.
2. Add 1.8X volume of AMpureXP beads to the mixture (i.e. 900µl beads for a 500µl reaction mix) and incubate for 5 minutes at room temperature
3. Place tubes in the magnetic separation rack for at least 1 minute.
4. Without disturbing the beads on the side of the tube, pipette out the liquid leaving up to 20µl behind.
5. While still on the rack, add 1ml of 70% ethanol to each tube to wash the beads.
6. Incubate for 1 minute at room temperature.
7. Pipette out all the liquid from each tube.
8. Repeat Steps 5 to 7.
9. Air dry the beads for 2 minutes at room temperature
10. Remove tubes from rack and add 200µl of water/elution buffer to the beads, pipetting up and down to ensure all the beads are suspended. Sequentially take the same 200µl bead/water mixture and pipette into the other tubes, until there are 6 tubes worth of beads suspended in the water.
11. Incubate at room temperature for a minimum of 5 minutes.
12. Place tubes in magnetic separation rack and let it sit for 5 minutes.
13. Without disturbing the beads on the side of the tube, pipette out the 200µl of liquid and put into a new tube.
14. Quantify the concentration of DNA using a spectrophotometer.

Supplementary Protocol 2.2: shRNA pool Sequencing Library Preparation

Materials Needed:

- Q5 Hot Start DNA polymerase* [NEB: M0493L]
- 5M Betaine [Sigma-Aldrich: B0300-1VL/ B0300-5VL]
- 10mM dNTPs [Promega: U1511/U1515]
- DNA *(input/genomic)*
- PCR Primers: *(5' to 3')*

  (Step 1 PCR primers)

pLKO F	CTCTTTCCCTACACGACGCTCTTCCGATCT NNNN CTTTATATATCTTGTGGAAAGGACGA
pLKO R	CTGGAGTTCAGACGTGTGCTCTTCCGATCT NNNNNN TGGATGAATACTGCCATTTGTCTC

  Note: NNNN stands for four and NNNNNN stands for six random nucleotides. These are a mixture of 25% of each base and are required to avoid QC errors for Illumina sequencing since the library has a low complexity at the non-shRNA regions.

  (Step 2 PCR primers)

PE PCR F	AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCT
PE PCR R	CAAGCAGAAGACGGCATACGAGAT XXXXXXXX GTGACTGGAGTTCAGACGTGTGCTCTTCCG

  Note: XXXXXXXX stands for the reverse complement of the index sequence used for multiplexing. This can be between 6-8 base pairs in length. Different R primers can be ordered and used with the F primer. If double indexing, add the reverse complement of the second index to the space in the F primer.

- (Optional) Qiagen Minelute Kit [Qiagen: 28004]
- Ampure XP beads [Beckman Coulter: A63880/A63881/A63882]
- 70% Ethanol
- 96-well magnetic separation plate

**(I have found Q5 to be more sensitive and specific than Taq or Phusion, and this improvement is necessary for amplifying from low amounts of sample.)*

Step 1: PCR amplification of shRNA sequences from sample

Per reaction, mix:

10µM pLKO F and R primers	0.4µM	1µl
5X Q5 Buffer	1X	5µl
10mM dNTPs	0.2mM	0.5µl
Q5 Hotstart Polymerase	2 Units	1µl
5M Betaine	1.5M	7.5µl
Genomic DNA sample	~2µg	X µl
Distilled H_2O		10-X µl
		25µl

PCR Conditions

98°C for 3 minutes

98°C for 30 seconds
60°C for 30 seconds
72°C for 30 seconds
} 25 Cycles

72°C for 10 minutes

4°C hold

**Note: If amplifying from input DNA pool, use ~100ng of sample.*
Expected product size is 196bp

(Optional) Step 2: Purification of DNA sample

This step removes excess primers which can inhibit the PCR reaction in step 3
Use the Qiagen Minelute kit with the manufacturer's protocol to purify, eluting the sample in 10µl of EB.
Alternative: Use the AMpure XP beads to purify the sample similar to step 4, scaling the volume of beads added to the 25µl PCR reaction volume (45µl of beads) and eluting in 20µl of water/elution buffer.

Step 3: Illumina Library Preparation using PCR

Per reaction, mix:

10µM PE PCR F and R primers	0.4µM	2µl
5X Q5 Buffer	1X	10µl
10mM dNTPs	0.2mM	1µl
Q5 Hotstart Polymerase	4 Units	2µl
5M Betaine	1.5M	15µl
Step 1 DNA sample	200-500ng	X µl
Distilled H_2O		20-X µl
		50µl

PCR Conditions

98°C for 3 minutes

98°C for 30 seconds
60°C for 30 seconds
72°C for 30 seconds
} 6 Cycles

72°C for 10 minutes

4°C hold

**Note: If planning to run different samples in the same sequencing lane, ensure that the F and R primer combination of indices are different for each sample so that you can demultiplex during analysis.*
If skipping Step 2, X should be half of the volume of the step 1 reaction (12.5µl)
Expected product size is 260bp

Step 4: Illumina Sequencing Library Cleanup

(Adapted from AMpure XP manufacturer's protocol)

1. Add 1.8X volume of AMpureXP beads to each reaction (i.e. 90µl beads for a 50µl reaction) and incubate for 5 minutes at room temperature
2. Place tubes in the magnetic separation rack for at least 1 minute.
3. Without disturbing the beads on the side of the tube, pipette out the liquid leaving up to 20µl behind.
4. While still on the rack, add 200µl of 70% ethanol to each tube to wash the beads.
5. Incubate for 1 minute at room temperature.
6. Pipette out all the liquid from each tube.
7. Repeat Steps 4 to 6.
8. Air dry the beads for 2 minutes at room temperature
9. Remove tubes from rack and add 40µl of water/elution buffer to the beads, pipetting up and down to ensure all the beads are suspended.
10. Incubate at room temperature for a minimum of 5 minutes.
11. Place tubes in magnetic separation rack and let it sit for 5 minutes.
12. Without disturbing the beads on the side of the tube, pipette out the 40µl of liquid and put into a new tube.
13. Quantify the concentration of DNA using a spectrophotometer.

**Note: Different samples can be pooled to produce a pool withan equal amount of DNA per sample and run on one Illumina HiSeq/MiSeq sequencing lane if they are uniquely indexed.*

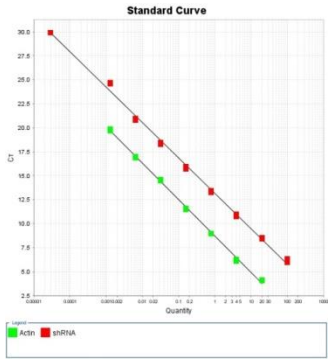


Figure 2.S1: Standard curve for Actin and shRNA qPCR primers
Slope for the actin curve (green) is -3.797 with an R^2 of 0.998, corresponding to 83.374% primer efficiency.
Slope for the shRNA curve (red) is -3.691 with an R^2 of 0.996, corresponding to 86.614% primer efficiency.

Sample	Mean shRNA qPCR cycle	Standard Deviation shRNA qPCR cycle	Mean Actin qPCR cycle	Standard Deviation Actin qPCR cycle	Transfection Rate
lentivirus 1 shot rep1	25.48693275	0.980911136	15.87373199	0.650952438	0.001229082
lentivirus 1 shot rep2	30.67140961	0.239385545	20.69246101	0.140888497	0.000953849
DNA 1 shot rep1	29.62903252	0.755530495	16.19665559	0.961291686	0.0000870752
DNA 1 shot rep2	34.09828949	0.743369937	22.191576	0.202929765	0.000250706
DNA 5 shots rep1	26.41898537	0.645285487	20.50484753	0.522136739	0.015962821
DNA 5 shots rep2	25.82338905	0.411470205	21.09821939	0.416632252	0.036393577
DNA 5 shots rep3	26.44046402	0.415599784	19.91761208	0.464417369	0.010468114
DNA 5 shots rep4	25.54398956	0.581989333	19.63961601	0.557033598	0.016071226

Table 2.S1: Infection Rates for various injection conditions
Each sample was prepared using at least 3 qPCR replicates for each target (Actin, shRNA). The formula to calculate the transfection rate was

$$\frac{1}{2^{shRNA\ cycle\ count - Actin\ cycle\ count}} \times \frac{Actin\ Primer\ efficiency}{shRNA\ Primer\ efficiency}$$

Assuming each cell has 1 copy of actin (haploid germ cell), this should provide the transfection rate of the shRNA in the test.

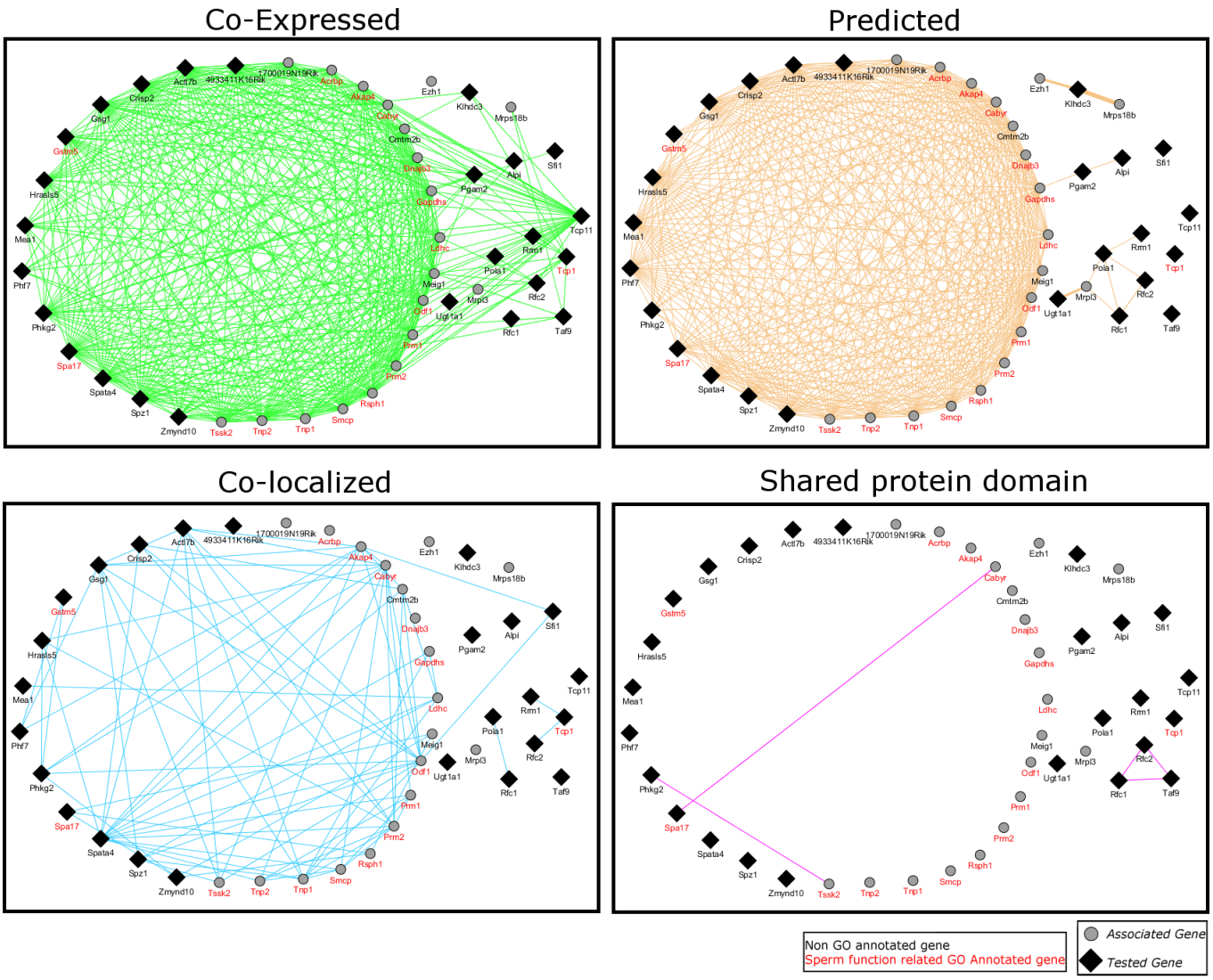


Figure 2.S2: Functional networks for predicted genes

The black diamond nodes are genes that I tested in the predicted shRNA pool while gray circle nodes are genes predicted to be functionally related. Red gene names mean that the shRNA are annotated with GO terms that are related to sperm function, while black gene names have no such annotation. The four figures with identical nodes but different color lines indicate which GeneMANIA mouse network links the genes.

TRC.Id	Clone.Name	Region	Gene	RefSeq.Id	Validated	Cell.line	Validate.Knockdown	Target.Seq	TRC.Version	shRNA ID
TRCN0000272981	NM_007527.3-409s21c1	CDS	BAX	NM_007527	Yes	Hepa 1-6	0.75	GCAGCTGACATGTTTGCTGAT	2	BAX.1
TRCN0000272982	NM_007527.3-323s21c1	CDS	BAX	NM_007527	Yes	Hepa 1-6	0.88	GAGATGAACTGGACAGCAATA	2	BAX.2
TRCN0000273037	NM_007527.3-631s21c1	CDS	BAX	NM_007527	Yes	Hepa 1-6	0.86	TGGCAGACAGTGACCATCTTT	2	BAX.3
TRCN0000273038	NM_007527.3-459s21c1	CDS	BAX	NM_007527	Yes	Hepa 1-6	0.87	CCTCTTCTACTTTGCTAGCAA	2	BAX.4
TRCN0000273039	NM_007527.3-488s21c1	CDS	BAX	NM_007527	Yes	Hepa 1-6	0.9	TCAAGGCCCTGTGCACTAAAG	2	BAX.5
TRCN0000305672	NM_007778.4-972s21c1	CDS	CSF1	NM_007778	Yes	Hepa 1-6	0.84	GCAACTGCCTGTACCCTAAAG	2	CSF1.1
TRCN0000305733	NM_007778.4-634s21c1	CDS	CSF1	NM_007778	Yes	Hepa 1-6	0.89	TGATCCTGTTTGCTACCTAAA	2	CSF1.2
TRCN0000305735	NM_007778.4-686s21c1	CDS	CSF1	NM_007778	Yes	Hepa 1-6	0.92	GATGAGACCATGCGCTTTAAA	2	CSF1.3
TRCN0000324336	NM_007778.4-1939s21c1	CDS	CSF1	NM_007778	Yes	Hepa 1-6	0.79	CCTCCTGTTCTACAAGTGGAA	2	CSF1.4
TRCN0000324337	NM_007778.4-2187s21c1	3UTR	CSF1	NM_007778	Yes	Hepa 1-6	0.88	GCCTACCAAGACTGGATGAAA	2	CSF1.5
TRCN0000361294	NM_021099.3-3417s21c1	3UTR	KIT	NM_021099			NA	CCTTAATGATGGGAGATATAT	2	KIT.1
TRCN0000361295	NM_021099.3-1198s21c1	CDS	KIT	NM_021099			NA	ACTTCGCCTGACCAGATTAAA	2	KIT.2
TRCN0000361296	NM_021099.3-256s21c1	CDS	KIT	NM_021099			NA	ATGGACTTTCAAGACCTATTT	2	KIT.3
TRCN0000368020	NM_021099.3-1318s21c1	CDS	KIT	NM_021099			NA	GACGTACGACAGGCTCATAAA	2	KIT.4
TRCN0000321124	NM_023371.3-136s21c1	CDS	PIN1	NM_023371	Yes	NIH/3T3	0.87	TGTTGGAGGCAGCAGCAAGAA	2	PIN1.1
TRCN0000321125	NM_023371.3-73s21c1	CDS	PIN1	NM_023371			NA	GGTGTACTACTTCAATCACAT	2	PIN1.2
TRCN0000321191	NM_023371.3-396s21c1	CDS	PIN1	NM_023371	Yes	NIH/3T3	0.89	GAGGTCAGATGCAGAAACCAT	2	PIN1.3
TRCN0000321192	NM_023371.3-413s21c1	CDS	PIN1	NM_023371	Yes	NIH/3T3	0.89	CCATTTGAGGATGCGTCGTTT	2	PIN1.4
TRCN0000321193	NM_023371.3-883s21c1	3UTR	PIN1	NM_023371	Yes	NIH/3T3	0.8	CCTACGCACCTTCCATTAAAT	2	PIN1.5
TRCN0000240542	NM_007755.4-972s21c1	CDS	CPEB1	NM_007755			NA	GAGGCGTTCCTTGGGATATTA	2	CPEB1.1
TRCN0000240543	NM_007755.4-1862s21c1	3UTR	CPEB1	NM_007755			NA	GTCTTTGTTTCTGCACTAATT	2	CPEB1.2
TRCN0000240544	NM_007755.4-1368s21c1	CDS	CPEB1	NM_007755	Yes	B16-F0	0.94	CCATCTTGAATGACCTATTTG	2	CPEB1.3
TRCN0000240545	NM_007755.4-627s21c1	CDS	CPEB1	NM_007755			NA	TGATTTCAAGCCTTCGCATTT	2	CPEB1.4
TRCN0000240546	NM_007755.4-172s21c1	CDS	CPEB1	NM_007755	Yes	B16-F0	0.83	AGTCTGTACAACACCTATAAA	2	CPEB1.5
TRCN0000277249	NM_010322.3-1022s21c1	CDS	GNPAT	NM_010322	Yes	Hepa 1-6	0.97	GATACCTACTTTGTCCCAATT	2	GNPAT.1
TRCN0000277250	NM_010322.3-1931s21c1	CDS	GNPAT	NM_010322	Yes	Hepa 1-6	0.89	GTGGAATCATATCAGTTACTT	2	GNPAT.2
TRCN0000277251	NM_010322.3-1226s21c1	CDS	GNPAT	NM_010322	Yes	Hepa 1-6	0.91	CTCAATCGGAACACGTATAAC	2	GNPAT.3
TRCN0000277252	NM_010322.3-437s21c1	CDS	GNPAT	NM_010322			NA	GAAGAGATCAACTATGTCATT	2	GNPAT.4
TRCN0000277292	NM_010322.3-2704s21c1	3UTR	GNPAT	NM_010322	Yes	Hepa 1-6	0.75	AGGACGTTCATGTCTAGATTA	2	GNPAT.5
TRCN0000239426	NM_175337.1-291s21c1	CDS	MLH3	NM_175337			NA	GAAGGTGGGAAACCGGTATTT	2	MLH3.1
TRCN0000239427	NM_175337.1-518s21c1	CDS	MLH3	NM_175337	Yes	Hepa 1-6	0.77	GGACTACAGTAACGGTCTATA	2	MLH3.2
TRCN0000239428	NM_175337.1-988s21c1	CDS	MLH3	NM_175337	Yes	Hepa 1-6	0.63	TCAGAACTCCACGGGATATAT	2	MLH3.3
TRCN0000239429	NM_175337.1-1264s21c1	CDS	MLH3	NM_175337	Yes	Hepa 1-6	0.74	AGTTTCCGGGAAGCGTGTAAT	2	MLH3.4
TRCN0000244279	NM_175337.1-4463s21c1	3UTR	MLH3	NM_175337	Yes	Hepa 1-6	0.66	CTCAAGCCTAAGGGTAGTTTA	2	MLH3.5
TRCN0000271121	NM_012046.2-1224s21c1	CDS	SPO11	NM_012046			NA	GGTTTGGAGGATGGATCTAAA	2	SPO11.1
TRCN0000271122	NM_012046.2-719s21c1	CDS	SPO11	NM_012046	Yes	Hepa 1-6	0.8	GTCGAGAAGGATGCAACATTT	2	SPO11.2
TRCN0000271124	NM_012046.2-443s21c1	CDS	SPO11	NM_012046			NA	GCAACCAAGAGAGACATATAC	2	SPO11.3
TRCN0000271125	NM_012046.2-900s21c1	CDS	SPO11	NM_012046			NA	TCGAGATAATGTGCATCTATA	2	SPO11.4
TRCN0000271173	NM_012046.2-1323s21c1	3UTR	SPO11	NM_012046			NA	TCTTAGGTATGCAATGGTAAA	2	SPO11.5
TRCN0000328391	NM_011870.4-366s21c1	CDS	CIB1	NM_011870			NA	CCAGACATCAAGTCACACTAT	2	CIB1.1
TRCN0000328395	NM_011870.4-503s21c1	CDS	CIB1	NM_011870			NA	GAAGCAGCTGATTGACAATAT	2	CIB1.2
TRCN0000328454	NM_011870.4-602s21c1	CDS	CIB1	NM_011870			NA	CTTTGCCAGCTCCTTTAAGAT	2	CIB1.3
TRCN0000328455	NM_011870.4-642s21c1	3UTR	CIB1	NM_011870			NA	AGTACCAACATCCTGTCCAAG	2	CIB1.4
TRCN0000328456	NM_011870.4-548s21c1	CDS	CIB1	NM_011870			NA	GGATGGGACCATCAATCTTTC	2	CIB1.5
TRCN0000306401	NM_008635.2-959s21c1	CDS	MTAP7	NM_008635	Yes	Hepa 1-6	0.6	CATGGACTAGCGAGCCATAAA	2	MTAP7.1
TRCN0000306402	NM_008635.2-1876s21c1	CDS	MTAP7	NM_008635	Yes	Hepa 1-6	0.73	GAGGACAGAGACCGCTGATAA	2	MTAP7.2
TRCN0000306403	NM_008635.2-2720s21c1	3UTR	MTAP7	NM_008635			NA	AGAGTGAGCGGAAGGTATTTA	2	MTAP7.3
TRCN0000306462	NM_008635.2-2009s21c1	CDS	MTAP7	NM_008635	Yes	Hepa 1-6	0.67	AGCCCACATGGAGTCGCTTTA	2	MTAP7.4
TRCN0000354145	NM_008635.2-696s21c1	CDS	MTAP7	NM_008635	Yes	Hepa 1-6	0.85	CCTCGTCTGCAACTTTGCTAA	2	MTAP7.5

TRCN0000238987	NM_026869.2-287s21c1	CDS	PYGO2	NM_026869	Yes	3T3-L1	0.86	GATCATCTGGTCGCTTCTAAC	2	PYGO2.1
TRCN0000238988	NM_026869.2-225s21c1	CDS	PYGO2	NM_026869	Yes	3T3-L1	0.76	ATACTCAGGGTCCTGCATATT	2	PYGO2.2
TRCN0000238989	NM_026869.2-1654s21c1	3UTR	PYGO2	NM_026869	Yes	3T3-L1	0.8	CCAACACCCGTGCCTACAATA	2	PYGO2.3
TRCN0000238990	NM_026869.2-1070s21c1	CDS	PYGO2	NM_026869	Yes	3T3-L1	0.67	GCCTGCCGTAGTGAGGTAAAT	2	PYGO2.4
TRCN0000257050	NM_026869.2-1287s21c1	CDS	PYGO2	NM_026869	Yes	3T3-L1	0.59	ACGATGGGTGACTCTAGTACC	2	PYGO2.5
TRCN0000329174	NM_011603.5-716s21c1	CDS	TBPL1	NM_011603	Yes	3T3-L1	0.81	AGATCCGTTTGCCAGAATTTA	2	TBPL1.1
TRCN0000329247	NM_011603.5-535s21c1	CDS	TBPL1	NM_011603	Yes	3T3-L1	0.62	CCTAGAATTACAGCTACAATT	2	TBPL1.2
TRCN0000329248	NM_011603.5-791s21c1	CDS	TBPL1	NM_011603	Yes	3T3-L1	0.82	GCTATCGGATAAAGTCTCTAA	2	TBPL1.3
TRCN0000375313	NM_011603.5-498s21c1	CDS	TBPL1	NM_011603	Yes	3T3-L1	0.65	GCGTGATGTTGGGAAAGTATT	2	TBPL1.4
TRCN0000375314	NM_011603.5-1310s21c1	3UTR	TBPL1	NM_011603	Yes	3T3-L1	0.75	GGGCACCAAAGAACCTGTAAA	2	TBPL1.5
TRCN0000247807	NM_011363.2-2184s21c1	CDS	SH2B1	NM_011363	Yes	Hepa 1-6	0.82	GCATGCTCTCTCGACTCAAAG	2	SH2B1.1
TRCN0000247808	NM_011363.2-2322s21c1	CDS	SH2B1	NM_011363	Yes	Hepa 1-6	0.65	ACCTGCGCTTGTCACTAAATG	2	SH2B1.2
TRCN0000247809	NM_011363.2-3164s21c1	3UTR	SH2B1	NM_011363	Yes	Hepa 1-6	0.61	AGGTTCATGAGCCCTGTTAAG	2	SH2B1.3
TRCN0000247810	NM_011363.2-1221s21c1	CDS	SH2B1	NM_011363	Yes	Hepa 1-6	0.59	TTGGTAGGGCATTGGCTAATG	2	SH2B1.4
TRCN0000247811	NM_011363.2-1553s21c1	CDS	SH2B1	NM_011363	Yes	Hepa 1-6	0.81	AGCATTCCCTGCTCTACTATT	2	SH2B1.5
TRCN0000336125	NM_011650.3-252s21c1	CDS	TSN	NM_011650	Yes	NIH/3T3	0.61	AGAACATTTCAGTACAGTAAA	2	TSN.1
TRCN0000336126	NM_011650.3-209s21c1	CDS	TSN	NM_011650	Yes	NIH/3T3	0.73	GTACTGGATTTCAGGACATTC	2	TSN.2
TRCN0000336148	NM_011650.3-478s21c1	CDS	TSN	NM_011650			NA	GAAGATTATCTCTCAGGAGTT	2	TSN.3
TRCN0000336183	NM_011650.3-421s21c1	CDS	TSN	NM_011650	Yes	NIH/3T3	0.7	GTTACAGAGATTCTTGGCATT	2	TSN.4
TRCN0000336185	NM_011650.3-990s21c1	3UTR	TSN	NM_011650	Yes	NIH/3T3	0.74	TGCCGTGTTGTCGTCGTATTA	2	TSN.5
TRCN0000306512	NM_019812.2-1370s21c1	CDS	SIRT1	NM_019812	Yes	Hepa 1-6	0.94	AGTGAGACCAGTAGCACTAAT	2	SIRT1.1
TRCN0000306518	NM_019812.2-2617s21c1	3UTR	SIRT1	NM_019812			NA	CTAGACCAAAGAATGGTATTT	2	SIRT1.2
TRCN0000326966	NM_019812.2-921s21c1	CDS	SIRT1	NM_019812	Yes	Hepa 1-6	0.93	GCCATGTTTGATATTGAGTAT	2	SIRT1.3
TRCN0000327027	NM_019812.2-1944s21c1	CDS	SIRT1	NM_019812	Yes	Hepa 1-6	0.86	GAGGGTAATCAATACCTGTTT	2	SIRT1.4
TRCN0000327028	NM_019812.2-1659s21c1	CDS	SIRT1	NM_019812			NA	CCTGAAAGAACTGTACCACAA	2	SIRT1.5
TRCN0000231561	NM_011696.1-243s21c1	CDS	VDAC3	NM_011696	Yes	Hepa 1-6	0.83	GGCAACCTAGAGACCAAATAT	2	VDAC3.1
TRCN0000231562	NM_011696.1-367s21c1	CDS	VDAC3	NM_011696	Yes	Hepa 1-6	0.67	TGACTCTTGATACCATATTTG	2	VDAC3.2
TRCN0000231563	NM_011696.1-450s21c1	CDS	VDAC3	NM_011696	Yes	Hepa 1-6	0.82	CTCGGCAGTAATGTTGATATA	2	VDAC3.3
TRCN0000231564	NM_011696.1-613s21c1	CDS	VDAC3	NM_011696	Yes	Hepa 1-6	0.85	AGCTGCATACTCACGTGAATG	2	VDAC3.4
TRCN0000231565	NM_011696.1-1149s21c1	3UTR	VDAC3	NM_011696	Yes	Hepa 1-6	0.93	TTGAGTTCTGCAGAGTTAATT	2	VDAC3.5
TRCN0000352166	NM_007385.2-64s21c1	CDS	APOC4	NM_007385			NA	GACCTGCCATCAGTCTCCCTT	2	APOC4.1
TRCN0000352168	NM_007385.2-278s21c1	CDS	APOC4	NM_007385			NA	CCTACTATGAAGATCACCTGA	2	APOC4.2
TRCN0000352241	NM_007385.2-101s21c1	CDS	APOC4	NM_007385			NA	TCAGCTTTGTAGCATCCATGT	2	APOC4.3
TRCN0000363992	NM_007385.2-252s21c1	CDS	APOC4	NM_007385			NA	TGGAGCTGTCCAGGGCTTTAT	2	APOC4.4
TRCN0000376019	NM_007385.2-212s21c1	CDS	APOC4	NM_007385			NA	TGACCAGAACCAGGGACAGAT	2	APOC4.5
TRCN0000376083	NM_007385.2-117s21c1	CDS	APOC4	NM_007385			NA	CATGTCTACAGAAAGCCTGAG	2	APOC4.6
TRCN0000247072	NM_029667.2-530s21c1	3UTR	LCE1I	NM_029667	Yes	Hepa 1-6	0.97	TGAGGAAGACTTCAGACAAAT	2	LCE1I.1
TRCN0000247073	NM_029667.2-191s21c1	CDS	LCE1I	NM_029667			NA	GTGTCTTCCTGCTGTAGCTTG	2	LCE1I.2
TRCN0000247074	NM_029667.2-549s21c1	3UTR	LCE1I	NM_029667			NA	ATAGTGCAGGAGGAGCACATG	2	LCE1I.3
TRCN0000247075	NM_029667.2-561s21c1	3UTR	LCE1I	NM_029667			NA	GAGCACATGCTCAGAAGATTC	2	LCE1I.4
TRCN0000247076	NM_029667.2-359s21c1	CDS	LCE1I	NM_029667			NA	AGCTCTGGATGCTGTAGCAGT	2	LCE1I.5
TRCN0000335079	NM_010809.1-1479s21c1	CDS	MMP3	NM_010809	Yes	3T3-L1	0.95	CCCACATATTGAAGAGCAATA	2	MMP3.1
TRCN0000335080	NM_010809.1-1615s21c1	3UTR	MMP3	NM_010809	Yes	3T3-L1	0.96	GCAGAACCAAACAGGAGCTAT	2	MMP3.2
TRCN0000348408	NM_010809.1-1445s21c1	CDS	MMP3	NM_010809			NA	CAGTTGGAATTTGACCCAAAT	2	MMP3.3
TRCN0000348486	NM_010809.1-1184s21c1	CDS	MMP3	NM_010809	Yes	3T3-L1	0.61	GAGCTAGCAGGTTATCCTAAA	2	MMP3.4
TRCN0000250833	NM_009136.3-459s21c1	CDS	SCRG1	NM_009136			NA	AGATGTCTTCTTTGGACCAAA	2	SCRG1.1
TRCN0000250834	NM_009136.3-423s21c1	CDS	SCRG1	NM_009136	Yes	B16-F0	0.94	CTACTGCAACTTCAGCGAACT	2	SCRG1.2
TRCN0000250835	NM_009136.3-521s21c1	3UTR	SCRG1	NM_009136	Yes	B16-F0	0.91	CCTTGCACTCTGGAGAACATG	2	SCRG1.3
TRCN0000250836	NM_009136.3-308s21c1	CDS	SCRG1	NM_009136	Yes	B16-F0	0.89	AGTTGCTAAAGGATCGCAATT	2	SCRG1.4
TRCN0000379710	NM_009308.3-739s21c1	CDS	SYT4	NM_009308			NA	ACTTCGAGAAGAAAGCATTTG	2	SYT4.1

TRCN0000381498	NM_009308.3-481s21c1	CDS	SYT4	NM_009308			NA	AGAGTGAAGTGAAGGGTAAAG	2	SYT4.2
TRCN0000381737	NM_009308.3-1196s21c1	CDS	SYT4	NM_009308			NA	ATCTGATGTGTCTGGACTTTC	2	SYT4.3
TRCN0000382380	NM_009308.3-388s21c1	CDS	SYT4	NM_009308			NA	AGACTCCTCCATACAAGTTTG	2	SYT4.4
TRCN0000382514	NM_009308.3-1393s21c1	CDS	SYT4	NM_009308			NA	GATCCCGAAATGAGGTGATTG	2	SYT4.5
TRCN0000303144	NM_011575.2-129s21c1	CDS	TFF3	NM_011575			NA	CAATGTATGGTGCCGGCAAAT	2	TFF3.1
TRCN0000303214	NM_011575.2-172s21c1	CDS	TFF3	NM_011575			NA	CCTCTGTCACATCGGAGCAGT	2	TFF3.2
TRCN0000303215	NM_011575.2-236s21c1	CDS	TFF3	NM_011575			NA	GCCCTGGTGCTTCAAACCTCT	2	TFF3.3
TRCN0000303219	NM_011575.2-212s21c1	CDS	TFF3	NM_011575			NA	CTTTGACTCCAGTATCCCAAA	2	TFF3.4
TRCN0000331879	NM_011575.2-89s21c1	CDS	TFF3	NM_011575			NA	CTCTGGGATAGCTGCAGATTA	2	TFF3.5
TRCN0000313482	NM_009452.2-838s21c1	3UTR	TNFSF4	NM_009452			NA	TTCCTCACTCAGGGATATTTA	2	TNFSF4.1
TRCN0000349384	NM_009452.2-411s21c1	CDS	TNFSF4	NM_009452			NA	GCAGAACAATTCGGTTGTCAT	2	TNFSF4.2
TRCN0000349385	NM_009452.2-543s21c1	CDS	TNFSF4	NM_009452			NA	CGATGGTCGAAGGATTGTCTT	2	TNFSF4.3
TRCN0000349440	NM_009452.2-363s21c1	CDS	TNFSF4	NM_009452			NA	GCAACTATTCATCAGCTCATA	2	TNFSF4.4
TRCN0000349887	NM_009452.2-631s21c1	CDS	TNFSF4	NM_009452			NA	TGCGAACACCTCCAGATAAAT	2	TNFSF4.5
TRCN0000419335	NM_031202.2-1939s21c1	3UTR	TYRP1	NM_031202			NA	ACACAGCTGTCAACCGTATTT	2	TYRP1.1
TRCN0000432060	NM_031202.2-809s21c1	CDS	TYRP1	NM_031202			NA	TGAGAACATTTCCGTTTATAA	2	TYRP1.2
TRCN0000440816	NM_031202.2-352s21c1	CDS	TYRP1	NM_031202			NA	CACGAGAGTGTGCCAATATTG	2	TYRP1.3
TRCN0000336014	NM_011515.4-321s21c1	CDS	VAMP7	NM_011515	Yes	Hepa 1-6	0.56	TTTGTATCACTGATGATGATT	2	VAMP7.1
TRCN0000336075	NM_011515.4-524s21c1	CDS	VAMP7	NM_011515	Yes	Hepa 1-6	0.92	GCACAAGTGGATGAACTGAAA	2	VAMP7.2
TRCN0000336077	NM_011515.4-400s21c1	CDS	VAMP7	NM_011515			NA	TTACGGTTCAAGAGCACAAAC	2	VAMP7.3
TRCN0000353291	NM_011515.4-934s21c1	3UTR	VAMP7	NM_011515			NA	CTTTGCCTGTCATATAGTTTG	2	VAMP7.4
TRCN0000353419	NM_011515.4-422s21c1	CDS	VAMP7	NM_011515			NA	GCACTTCCTTATGCTATGAAT	2	VAMP7.5
TRCN0000380436	NM_011515.4-467s21c1	CDS	VAMP7	NM_011515			NA	GCACAACTGAAGCATCACTCT	2	VAMP7.6
TRCN0000380733	NM_011515.4-493s21c1	CDS	VAMP7	NM_011515			NA	TAAGAGCCTAGACAAAGTGAT	2	VAMP7.7

Supplemental Table 2.1: shRNA pool design for pilot pool

TRC Id	Clone Name	Region	Gene	RefSeq Id	Validated	Cell line	Validate Knockdown	Target Seq	TRC Version	shRNA ID
TRCN0000091848	NM_025271.1-439s1c1	CDS	ACTL7B	NM_025271				CCAGAACATCTGGGAGTACAT	1	Actl7b.1
TRCN0000091849	NM_025271.1-270s1c1	CDS	ACTL7B	NM_025271				CCACCTATTTCATCTCCTCTA	1	Actl7b.2
TRCN0000091850	NM_025271.1-269s1c1	CDS	ACTL7B	NM_025271				CCCACCTATTTCATCTCCTCT	1	Actl7b.3
TRCN0000091851	NM_025271.1-607s1c1	CDS	ACTL7B	NM_025271				GCTGCTGTCCATCTACTCATA	1	Actl7b.4
TRCN0000091852	NM_025271.1-608s1c1	CDS	ACTL7B	NM_025271				CTGCTGTCCATCTACTCATAT	1	Actl7b.5
TRCN0000081088	XM_129951.4-1743s1c1	3UTR	ALPI	XM_129951				CCATAGATTTCCTGAGCCCAA	1	Alpi.1
TRCN0000081089	XM_129951.4-967s1c1	CDS	ALPI	XM_129951				CCACAAGGCTTCTACCTCTTT	1	Alpi.2
TRCN0000081090	XM_129951.4-1213s1c1	CDS	ALPI	XM_129951				GACAAATCCTACACCTCCATT	1	Alpi.3
TRCN0000081091	XM_129951.4-861s1c1	CDS	ALPI	XM_129951				CCTCTTTGAGCCAACAGAAAT	1	Alpi.4
TRCN0000081092	XM_129951.4-45s1c1	CDS	ALPI	XM_129951				CCATCTGTCCTTTGGTATCAT	1	Alpi.5
TRCN0000106366	NM_009420.1-713s1c1	CDS	CRISP2	NM_009420				GCCATTACTGTCCTATGGGTA	1	Crisp2.1
TRCN0000106368	NM_009420.1-286s1c1	CDS	CRISP2	NM_009420				CCAGACTTTACTTCTTTGTTA	1	Crisp2.2
TRCN0000106369	NM_009420.1-776s1c1	CDS	CRISP2	NM_009420				CTTGTGCTAGTTGTCCCAATA	1	Crisp2.3
TRCN0000173792	NM_010352.1-1121s1c1	3UTR	GSG1	NM_010352				CGCTCTGTCTCTGAAGCTATT	1	Gsg1.1
TRCN0000175281	NM_010352.1-1160s1c1	3UTR	GSG1	NM_010352				CAGGACAAAGAATTTCAACAA	1	Gsg1.2
TRCN0000175307	NM_010352.1-477s1c1	CDS	GSG1	NM_010352				CGTTTCATTGAACTCACACCA	1	Gsg1.3
TRCN0000175876	NM_010352.1-861s1c1	CDS	GSG1	NM_010352				GAGACCACACTCTTGGAATTA	1	Gsg1.4
TRCN0000193913	NM_010352.1-179s1c1	CDS	GSG1	NM_010352				CTTCATTTCTGCCATCCTCAA	1	Gsg1.5
TRCN0000103220	NM_010360.1-449s1c1	CDS	GSTM5	NM_010360	Yes	Hepa 1-6	0.91	GCTACAATTCTAACCACGAAA	1	Gstm5.1
TRCN0000103221	NM_010360.1-507s1c1	CDS	GSTM5	NM_010360	Yes	Hepa 1-6	0.96	GCTGAAACAATTCTCATTGTT	1	Gstm5.2
TRCN0000103222	NM_010360.1-720s1c1	CDS	GSTM5	NM_010360	Yes	Hepa 1-6	0.93	GATGCCAATCAATAACAAGAT	1	Gstm5.3
TRCN0000103223	NM_010360.1-593s1c1	CDS	GSTM5	NM_010360	Yes	Hepa 1-6	0.95	TCTTGGATCAGAACCGTATAT	1	Gstm5.4
TRCN0000103224	NM_010360.1-592s1c1	CDS	GSTM5	NM_010360	Yes	Hepa 1-6	0.87	GTCTTGGATCAGAACCGTATA	1	Gstm5.5
TRCN0000251025	NM_025731.2-794s21c1	CDS	HRASLS5	NM_025731				TTGTCAATGACCTCAGATATG	2	Hrasls5.1
TRCN0000251026	NM_025731.2-407s21c1	CDS	HRASLS5	NM_025731				AGCTGATTCCAACATCAAATT	2	Hrasls5.2
TRCN0000251027	NM_025731.2-518s21c1	CDS	HRASLS5	NM_025731				ATGAACATTGGGCCATCTATG	2	Hrasls5.3
TRCN0000251028	NM_025731.2-310s21c1	CDS	HRASLS5	NM_025731				CTCAAGCAAGACCGCCGATTA	2	Hrasls5.4
TRCN0000251029	NM_025731.2-755s21c1	CDS	HRASLS5	NM_025731				TAGTTCAGTACAGCCTAATTG	2	Hrasls5.5
TRCN0000317640	NM_010787.1-545s21c1	CDS	MEA1	NM_010787	Yes	NIH/3T3	0.97	GTGGGAAGATGTGGTACAGAA	2	Mea1.1
TRCN0000317719	NM_010787.1-192s21c1	CDS	MEA1	NM_010787	Yes	NIH/3T3	0.93	AGCAGTGAAGAACCCGAGGAA	2	Mea1.2
TRCN0000317720	NM_010787.1-453s21c1	CDS	MEA1	NM_010787	Yes	NIH/3T3	0.81	GAACATGTAGAGCTGGTGAAA	2	Mea1.3
TRCN0000317721	NM_010787.1-658s21c1	3UTR	MEA1	NM_010787	Yes	NIH/3T3	0.86	GACTAACAACTCTGGTCTTAA	2	Mea1.4
TRCN0000319526	NM_010787.1-420s21c1	CDS	MEA1	NM_010787	Yes	NIH/3T3	0.98	TTGAGCAGCCACAGCTCTATC	2	Mea1.5
TRCN0000336483	NM_018870.3-309s21c1	CDS	PGAM2	NM_018870	Yes	NIH/3T3	0.61	TTGGACCATCCTGGATGTTAC	2	Pgam2.1
TRCN0000336553	NM_018870.3-134s21c1	CDS	PGAM2	NM_018870				TTCGCCACGGTGAGAGCTTAT	2	Pgam2.2
TRCN0000336554	NM_018870.3-387s21c1	CDS	PGAM2	NM_018870				TGGCCTCACAGGCCTCAATAA	2	Pgam2.3
TRCN0000336556	NM_018870.3-637s21c1	CDS	PGAM2	NM_018870				GCTGGCCAGAGAGTGCTTATT	2	Pgam2.4
TRCN0000375078	NM_018870.3-476s21c1	CDS	PGAM2	NM_018870				CACCACCCATGGATGAGAAAC	2	Pgam2.5
TRCN0000375079	NM_018870.3-558s21c1	CDS	PGAM2	NM_018870	Yes	NIH/3T3	0.84	GCCTACCTGTGAAAGTCTCAA	2	Pgam2.6
TRCN0000379130	NM_018870.3-446s21c1	CDS	PGAM2	NM_018870				AGATCTGGAGGCGTTCCTTTG	2	Pgam2.7
TRCN0000071228	NM_008892.1-2630s1c1	CDS	POLA1	NM_008892				CCTGGATTTCAACAGTTTATA	1	Pola1.1
TRCN0000071229	NM_008892.1-1583s1c1	CDS	POLA1	NM_008892	Yes	Hepa 1-6	0.96	GCCAATCAGTTGGTGTAAATT	1	Pola1.2
TRCN0000071230	NM_008892.1-3970s1c1	CDS	POLA1	NM_008892	Yes	Hepa 1-6	0.76	CCAGTTTGTATCGTTGCAGTA	1	Pola1.3

TRCN0000071231	NM_008892.1-279s1c1	CDS	POLA1	NM_008892	Yes	Hepa 1-6	0.66	CGTCAGGATGATGACTGGATT	1	Pola1.4
TRCN0000071232	NM_008892.1-2766s1c1	CDS	POLA1	NM_008892	Yes	Hepa 1-6	0.62	CCAAACTTAGAGATGGGCATT	1	Pola1.5
TRCN0000111395	NM_145480.1-928s1c1	CDS	RFC4	NM_145480	Yes	B16-F0	0.75	GCTGTGGTAAAGAACCTCATA	1	Rfc4.1
TRCN0000111396	NM_145480.1-772s1c1	CDS	RFC4	NM_145480	Yes	B16-F0	0.83	CGGAAAGCCATCACATTTCTT	1	Rfc4.2
TRCN0000111397	NM_145480.1-868s1c1	CDS	RFC4	NM_145480	Yes	B16-F0	0.84	GCTGCAACCATTGATGGAATA	1	Rfc4.3
TRCN0000111398	NM_145480.1-658s1c1	CDS	RFC4	NM_145480	Yes	B16-F0	0.93	CCTCTGTCAGATAAGATTCAA	1	Rfc4.4
TRCN0000111399	NM_145480.1-479s1c1	CDS	RFC4	NM_145480	Yes	B16-F0	0.87	GTCCTCCCTTTAAGATTGTAA	1	Rfc4.5
TRCN0000247868	NM_030207.2-636s21c1	CDS	SFI1	NM_030207				ATGAGGAAGAGGTTCCGAATA	2	Sfi1.1
TRCN0000247869	NM_030207.2-1674s21c1	CDS	SFI1	NM_030207				ACATTAGAGAAGCAAGTATTT	2	Sfi1.2
TRCN0000247870	NM_030207.2-2304s21c1	CDS	SFI1	NM_030207				GTTACTTCAGTGCAGATATAT	2	Sfi1.3
TRCN0000247871	NM_030207.2-695s21c1	CDS	SFI1	NM_030207				CTGGAAGTCCTGGTTGATATA	2	Sfi1.4
TRCN0000247872	NM_030207.2-1547s21c1	CDS	SFI1	NM_030207				GGCCAGAGCAGATGGTCATTT	2	Sfi1.5
TRCN0000306509	NM_011449.1-341s21c1	CDS	SPA17	NM_011449	Yes	Hepa 1-6	0.97	TGTGAACAAGAATTAGCTAAG	2	Spa17.1
TRCN0000306510	NM_011449.1-371s21c1	CDS	SPA17	NM_011449	Yes	Hepa 1-6	0.91	AGAGAAGAAACACCAGTCACT	2	Spa17.2
TRCN0000306511	NM_011449.1-394s21c1	CDS	SPA17	NM_011449	Yes	Hepa 1-6	0.88	CTTCGAGGAGTCTACTGAGGA	2	Spa17.3
TRCN0000326953	NM_011449.1-187s21c1	CDS	SPA17	NM_011449	Yes	Hepa 1-6	0.5	ACCGGACAATATACCAGCTTT	2	Spa17.4
TRCN0000327012	NM_011449.1-295s21c1	CDS	SPA17	NM_011449	Yes	Hepa 1-6	0.74	CTTCTATAACAACCACGCATT	2	Spa17.5
TRCN0000196151	NM_133711.2-926s1c1	3UTR	SPATA4	NM_133711				GTGAAGGACATGGAGGAAGTA	1	Spata4.1
TRCN0000215431	NM_133711.3-455s1c1	CDS	SPATA4	NM_133711				GGAGATTTACACTTTACTAAC	1	Spata4.2
TRCN0000215604	NM_133711.3-603s1c1	CDS	SPATA4	NM_133711				GAACTACTAAGCAATCCTAAT	1	Spata4.3
TRCN0000215867	NM_133711.3-477s1c1	CDS	SPATA4	NM_133711				CATCAAGAAATTAGAAGTATC	1	Spata4.4
TRCN0000120437	NM_013686.1-821s1c1	CDS	TCP1	NM_013686				CCTGAGAAATTGGACCAAATT	1	Tcp1.1
TRCN0000120438	NM_013686.1-1220s1c1	CDS	TCP1	NM_013686				CGCTCTTTACATGATGCTCTT	1	Tcp1.2
TRCN0000120439	NM_013686.1-617s1c1	CDS	TCP1	NM_013686	Yes	Hepa 1-6	0.94	CGCTATCCAATCAATTCTGTT	1	Tcp1.3
TRCN0000120440	NM_013686.1-395s1c1	CDS	TCP1	NM_013686				ACATCAGTTATTAGTGGCTAT	1	Tcp1.4
TRCN0000120441	NM_013686.1-1397s1c1	CDS	TCP1	NM_013686				CCTAATACACTGGCAGTGAAT	1	Tcp1.5
TRCN0000106380	NM_013687.2-1698s1c1	CDS	TCP11	NM_013687				CCCGTATTACACCGAGATCTT	1	Tcp11.1
TRCN0000106381	NM_013687.2-1545s1c1	CDS	TCP11	NM_013687				GCGTATACACTTGTTCCTCAA	1	Tcp11.2
TRCN0000106382	NM_013687.2-732s1c1	CDS	TCP11	NM_013687				CTCCTACCTCTCCAAGTATAT	1	Tcp11.3
TRCN0000106383	NM_013687.2-460s1c1	CDS	TCP11	NM_013687				CTTGATTGTCAGTTGGAAGAA	1	Tcp11.4
TRCN0000106384	NM_013687.2-1180s1c1	CDS	TCP11	NM_013687				GAAGAGTTTCCTGAAACCCTA	1	Tcp11.5
TRCN0000346802	NM_025752.2-631s21c1	CDS	4933411K16RIK	NM_025752				TCCACAACTCACCGTCGTAAA	2	4933411K16Rik.1
TRCN0000346863	NM_025752.2-281s21c1	CDS	4933411K16RIK	NM_025752				CAAATGACAGCTGGATCAAAT	2	4933411K16Rik.2
TRCN0000346865	NM_025752.2-840s21c1	CDS	4933411K16RIK	NM_025752				CAACCAGAGTCATCCCGTTTA	2	4933411K16Rik.3
TRCN0000346866	NM_025752.2-1204s21c1	3UTR	4933411K16RIK	NM_025752				TAAGAGACTGCCAACCAAATT	2	4933411K16Rik.4
TRCN0000346867	NM_025752.2-1064s21c1	CDS	4933411K16RIK	NM_025752				CAAGCCACCAGAGCCATAATC	2	4933411K16Rik.5
TRCN0000177244	NM_027910.1-1376s1c1	CDS	KLHDC3	NM_027910	Yes	B16-F0	0.96	GAATTTGACCTCATAGATCAT	1	Klhdc3.1
TRCN0000177891	NM_027910.1-736s1c1	CDS	KLHDC3	NM_027910				CGCCTTTGATGTCAATACTCA	1	Klhdc3.2
TRCN0000178618	NM_027910.1-1142s1c1	CDS	KLHDC3	NM_027910	Yes	B16-F0	0.79	CATTCAGCCTTTGGCTACAAT	1	Klhdc3.3
TRCN0000182094	NM_027910.1-823s1c1	CDS	KLHDC3	NM_027910	Yes	B16-F0	0.79	CCTGGGCAAGATCATGTACAT	1	Klhdc3.4
TRCN0000200040	NM_027910.1-909s1c1	CDS	KLHDC3	NM_027910				CCATGACATGGGCTCTTGTTT	1	Klhdc3.5
TRCN0000217511	NM_027910.2-1023s1c1	CDS	KLHDC3	NM_027910				GGCCATTTCATTCCAACAATG	1	Klhdc3.6
TRCN0000176986	NM_175349.2-405s1c1	CDS	LDHAL6B	NM_175349				GCCCTTGTTGATAATAATGAA	1	Ldhal6b.1
TRCN0000177866	NM_175349.2-1264s1c1	3UTR	LDHAL6B	NM_175349				CCCTTCTAAAGATACCGAAGA	1	Ldhal6b.2
TRCN0000181671	NM_175349.2-150s1c1	CDS	LDHAL6B	NM_175349				GTGAGTACAACCAGGGTAGAT	1	Ldhal6b.3

TRCN0000181693	NM_175349.2-1091s1c1	CDS	LDHAL6B	NM_175349				CCAAGGTCTCTACGGAATCAA	1	Ldhal6b.4
TRCN0000198493	NM_175349.2-266s1c1	CDS	LDHAL6B	NM_175349				GACCGTGAAGGGTGAACTTAT	1	Ldhal6b.5
TRCN0000216958	NM_175349.2-579s1c1	CDS	LDHAL6B	NM_175349				CTGAATTTAGTCCAGCGAAAC	1	Ldhal6b.6
TRCN0000082218	NM_027949.1-1760s1c1	3UTR	PHF7	NM_027949				CCCAGGACAGTGAGATACAAA	1	Phf7.1
TRCN0000082219	NM_027949.1-1142s1c1	CDS	PHF7	NM_027949	Yes	Hepa 1-6	0.87	CGCAAGTGTATCCAGAAATAT	1	Phf7.2
TRCN0000082220	NM_027949.1-1456s1c1	CDS	PHF7	NM_027949				CAGCAAGAAATGGGAATGTAA	1	Phf7.3
TRCN0000082221	NM_027949.1-1087s1c1	CDS	PHF7	NM_027949				CCGAACAAGTGTTGAGAACAT	1	Phf7.4
TRCN0000082222	NM_027949.1-739s1c1	CDS	PHF7	NM_027949				AGACAATCTTTGTGTCCATTA	1	Phf7.5
TRCN0000024369	NM_026888.1-363s1c1	CDS	PHKG2	NM_026888	Yes	3T3-L1	0.82	CGAGTCTTCTAGCTTCATGTT	1	Phkg2.1
TRCN0000024370	NM_026888.1-1060s1c1	CDS	PHKG2	NM_026888	Yes	3T3-L1	0.66	CCACTAACTAAGAATGCACTA	1	Phkg2.2
TRCN0000024371	NM_026888.1-552s1c1	CDS	PHKG2	NM_026888	Yes	3T3-L1	0.59	CCTAGATGACAATATGCAGAT	1	Phkg2.3
TRCN0000024372	NM_026888.1-783s1c1	CDS	PHKG2	NM_026888	Yes	3T3-L1	0.62	CCAAATCCTGATGCTACGCAT	1	Phkg2.4
TRCN0000024373	NM_026888.1-295s1c1	CDS	PHKG2	NM_026888	Yes	3T3-L1	0.84	CGGCGAGAGATGCACATTCTT	1	Phkg2.5
TRCN0000111365	NM_020022.2-1207s1c1	3UTR	RFC2	NM_020022	Yes	Hepa 1-6	0.9	CAGCTAGAACATGCTCACTTT	1	Rfc2.1
TRCN0000111366	NM_020022.2-592s1c1	CDS	RFC2	NM_020022	Yes	Hepa 1-6	0.92	GCGTGTAATGCTTCAGACAAA	1	Rfc2.2
TRCN0000111367	NM_020022.2-1003s1c1	CDS	RFC2	NM_020022	Yes	Hepa 1-6	0.85	CCGATGGCTGAATACTTGAAA	1	Rfc2.3
TRCN0000111368	NM_020022.2-800s1c1	CDS	RFC2	NM_020022	Yes	Hepa 1-6	0.95	CCTTCTCAGGATTTGGCTATA	1	Rfc2.4
TRCN0000111369	NM_020022.2-390s1c1	CDS	RFC2	NM_020022	Yes	Hepa 1-6	0.81	CCTGGAACTCAATGCCTCAAA	1	Rfc2.5
TRCN0000042393	NM_009103.2-2730s1c1	3UTR	RRM1	NM_009103				GCCAGCTTTGATATTAGGAAT	1	Rrm1.1
TRCN0000042394	NM_009103.2-402s1c1	CDS	RRM1	NM_009103				GCCGTCTCTAACTTGCACAAA	1	Rrm1.2
TRCN0000042395	NM_009103.2-1134s1c1	CDS	RRM1	NM_009103				CGCGATCTCTTCTTTGCACTT	1	Rrm1.3
TRCN0000042396	NM_009103.2-2345s1c1	CDS	RRM1	NM_009103	Yes	Hepa 1-6	0.9	GCAGGGTTTAAAGACTGGAAT	1	Rrm1.4
TRCN0000042397	NM_009103.2-559s1c1	CDS	RRM1	NM_009103	Yes	Hepa 1-6	0.92	CCATTATCTATGACCGAGATT	1	Rrm1.5
TRCN0000081553	NM_030237.2-1348s1c1	3UTR	SPZ1	NM_030237				CGCCAGAAGCAGATACAACTT	1	Spz1.1
TRCN0000081554	NM_030237.2-1238s1c1	CDS	SPZ1	NM_030237				CCCGGATGTTTGCTTTACTAA	1	Spz1.2
TRCN0000081555	NM_030237.2-874s1c1	CDS	SPZ1	NM_030237				GCCAGAACGAAACCCAAGAAA	1	Spz1.3
TRCN0000081556	NM_030237.2-356s1c1	CDS	SPZ1	NM_030237				CCCACCACCAAAGAATAGCAT	1	Spz1.4
TRCN0000081557	NM_030237.2-1015s1c1	CDS	SPZ1	NM_030237				CAGCCCTGCTAGAGAATGAAT	1	Spz1.5
TRCN0000244259	NM_027592.2-153s21c1	CDS	TAF9	NM_027592	Yes	Hepa 1-6	0.99	ATCAGGCCTGAAGTACGTTAA	2	Taf9.1
TRCN0000244260	NM_027592.2-560s21c1	CDS	TAF9	NM_027592	Yes	Hepa 1-6	0.98	GGATTGAGCAGTGGGTCAAAG	2	Taf9.2
TRCN0000244261	NM_027592.2-626s21c1	3UTR	TAF9	NM_027592	Yes	Hepa 1-6	0.97	GTCCATCGTCCCAGATCTTAG	2	Taf9.3
TRCN0000244387	NM_027592.2-251s21c1	CDS	TAF9	NM_027592	Yes	Hepa 1-6	0.98	ATAGGGTTGTCGATGAGTTAG	2	Taf9.4
TRCN0000244388	NM_027592.2-328s21c1	CDS	TAF9	NM_027592	Yes	Hepa 1-6	0.98	CCCGAACGCTGGTTTCATATA	2	Taf9.5
TRCN0000093944	NM_201645.1-2005s1c1	3UTR	UGT1A1	NM_201645				CCAGTGTTAGTCATTCTTCAT	1	Ugt1a1.1
TRCN0000093945	NM_201645.1-548s1c1	CDS	UGT1A1	NM_201645				CCCACTGTGTACTTCTTGAAT	1	Ugt1a1.2
TRCN0000093946	NM_201645.1-439s1c1	CDS	UGT1A1	NM_201645				GCTGCACAATGCCGAGTTTAT	1	Ugt1a1.3
TRCN0000093947	NM_201645.1-133s1c1	CDS	UGT1A1	NM_201645				GAGGCTGTTAGTGTTCCCTAT	1	Ugt1a1.4
TRCN0000093948	NM_201645.1-1174s1c1	CDS	UGT1A1	NM_201645				CCGTGGTATTTATGAAGGAAT	1	Ugt1a1.5
TRCN0000362868	NM_053253.3-1209s21c1	CDS	ZMYND10	NM_053253				GCACCAGCTTCAGCACGTATT	2	Zmynd10.1
TRCN0000362869	NM_053253.3-549s21c1	CDS	ZMYND10	NM_053253				CTTGGACCTAGTAGACTATTG	2	Zmynd10.2
TRCN0000362870	NM_053253.3-670s21c1	CDS	ZMYND10	NM_053253				CAGGCGGAGATGATGGAATTT	2	Zmynd10.3
TRCN0000362871	NM_053253.3-283s21c1	CDS	ZMYND10	NM_053253				GCCATCCTTGATGCAACTATC	2	Zmynd10.4
TRCN0000352166	NM_007385.2-64s21c1	CDS	APOC4	NM_007385			NA	GACCTGCCATCAGTCTCCCTT	2	Apoc4.1
TRCN0000352241	NM_007385.2-101s21c1	CDS	APOC4	NM_007385			NA	TCAGCTTTGTAGCATCCATGT	2	Apoc4.2
TRCN0000363992	NM_007385.2-252s21c1	CDS	APOC4	NM_007385			NA	TGGAGCTGTCCAGGGCTTTAT	2	Apoc4.3

TRCN0000376019	NM_007385.2-212s21c1	CDS	APOC4	NM_007385			NA	TGACCAGAACCAGGGACAGAT	2	Apoc4.4
TRCN0000376083	NM_007385.2-117s21c1	CDS	APOC4	NM_007385			NA	CATGTCTACAGAAAGCCTGAG	2	Apoc4.5
TRCN0000379710	NM_009308.3-739s21c1	CDS	SYT4	NM_009308			NA	ACTTCGAGAAGAAAGCATTTG	2	Syt4.1
TRCN0000381498	NM_009308.3-481s21c1	CDS	SYT4	NM_009308			NA	AGAGTGAAGTGAAGGGTAAAG	2	Syt4.2
TRCN0000382380	NM_009308.3-388s21c1	CDS	SYT4	NM_009308			NA	AGACTCCTCCATACAAGTTTG	2	Syt4.3
TRCN0000303144	NM_011575.2-129s21c1	CDS	TFF3	NM_011575			NA	CAATGTATGGTGCCGGCAAAT	2	Tff3.1
TRCN0000303214	NM_011575.2-172s21c1	CDS	TFF3	NM_011575			NA	CCTCTGTCACATCGGAGCAGT	2	Tff3.2
TRCN0000303215	NM_011575.2-236s21c1	CDS	TFF3	NM_011575			NA	GCCCTGGTGCTTCAAACCTCT	2	Tff3.3
TRCN0000303219	NM_011575.2-212s21c1	CDS	TFF3	NM_011575			NA	CTTTGACTCCAGTATCCCAAA	2	Tff3.4
TRCN0000331879	NM_011575.2-89s21c1	CDS	TFF3	NM_011575			NA	CTCTGGGATAGCTGCAGATTA	2	Tff3.5
SHC202								CAACAAGATGAAGAGCACCAA	2	SHC202
SHC216								GCGCGATAGCGCTAATAATTT	2	SHC216

Supplemental Table 2.2: shRNA pool design for predicted genes' pool

References

1. Jan, S. Z. et al. Molecular control of rodent spermatogenesis. Biochim. Biophys. Acta BBA - Mol. Basis Dis. 1822, 1838–1850 (2012).
2. He, Z., Kokkinaki, M. & Dym, M. Signaling molecules and pathways regulating the fate of spermatogonial stem cells. Microsc. Res. Tech. 72, 586–595 (2009).
3. Clemente, E. J., Furlong, R. A., Loveland, K. L. & Affara, N. A. Gene expression study in the juvenile mouse testis: identification of stage-specific molecular pathways during spermatogenesis. Mamm. Genome 17, 956–975 (2006).
4. Hu, Z. et al. A genome-wide association study in Chinese men identifies three risk loci for non-obstructive azoospermia. Nat. Genet. 44, 183–186 (2012).
5. Dohle, G. R. et al. Genetic risk factors in infertile men with severe oligozoospermia and azoospermia. Hum. Reprod. 17, 13–16 (2002).
6. Dores, C. & Dobrinski, I. De novo morphogenesis of testis tissue: an improved bioassay to investigate the role of VEGF-165 during testis formation. Reprod. Camb. Engl. 148, 109–117 (2014).
7. Sato, T. et al. In vitro production of functional sperm in cultured neonatal mouse testes. Nature 471, 504–507 (2011).
8. Stukenborg, J.-B. et al. New horizons for in vitro spermatogenesis? An update on novel three-dimensional culture systems as tools for meiotic and post-meiotic differentiation of testicular germ cells. Mol. Hum. Reprod. 15, 521–529 (2009).
9. Luo, B. et al. Highly parallel identification of essential genes in cancer cells. Proc. Natl. Acad. Sci. 105, 20380–20385 (2008).
10. Zuber, J. et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. Nature 478, 524–528 (2011).
11. Wuestefeld, T. et al. A Direct In Vivo RNAi Screen Identifies MKK4 as a Key Regulator of Liver Regeneration. Cell 153, 389–401 (2013).
12. Beronja, S. et al. RNAi screens in mice identify physiological regulators of oncogenic growth. Nature 501, 185–190 (2013).
13. Zuber, J. et al. An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. Genes Dev. 25, 1628–1640 (2011).
14. Zender, L. et al. An Oncogenomics-Based In Vivo RNAi Screen Identifies Tumor Suppressors in Liver Cancer. Cell 135, 852–864 (2008).
15. Bric, A. et al. Functional Identification of Tumor-Suppressor Genes through an In Vivo RNA Interference Screen in a Mouse Lymphoma Model. Cancer Cell 16, 324–335 (2009).
16. Meacham, C. E., Ho, E. E., Dubrovsky, E., Gertler, F. B. & Hemann, M. T. In vivo RNAi screening identifies regulators of actin dynamics as key determinants of lymphoma progression. Nat. Genet. 41, 1133–1137 (2009).
17. Li, S. & Huang, L. Nonviral gene therapy: promises and challenges. Gene Ther. 7, 31–34 (2000).

18. Ryu, B.-Y. et al. Efficient Generation of Transgenic Rats Through the Male Germline Using Lentiviral Transduction and Transplantation of Spermatogonial Stem Cells. J. Androl. 28, 353–360 (2007).
19. Brinster, R. L. & Zimmermann, J. W. Spermatogenesis following male germ-cell transplantation. Proc. Natl. Acad. Sci. U. S. A. 91, 11298–11302 (1994).
20. Hermann, B. P. et al. Spermatogonial Stem Cell Transplantation into Rhesus Testes Regenerates Spermatogenesis Producing Functional Sperm. Cell Stem Cell 11, 715–726 (2012).
21. Usmani, A. et al. A non-surgical approach for male germ cell mediated gene transmission through transgenesis. Sci. Rep. 3, (2013).
22. Shoji, M., Chuma, S., Yoshida, K., Morita, T. & Nakatsuji, N. RNA interference during spermatogenesis in mice. Dev. Biol. 282, 524–534 (2005).
23. Ho, N. R. Y., Huang, N. & Conrad, D. F. Improved detection of disease-associated variation by sex-specific characterization and prediction of genes required for fertility. Andrology 3, 1140–1149 (2015).
24. Tannour-Louet, M. et al. Increased gene copy number of VAMP7 disrupts human male urogenital development through altered estrogen action. Nat. Med. 20, 715–724 (2014).
25. Sato, M. et al. The Role of VAMP7/TI-VAMP in Cell Polarity and Lysosomal Exocytosis in vivo. Traffic 12, 1383–1393 (2011).
26. Danglot, L. et al. Absence of TI-VAMP/Vamp7 Leads to Increased Anxiety in Mice. J. Neurosci. 32, 1962–1968 (2012).
27. DasGupta, R. et al. A case study of the reproducibility of transcriptional reporter cell-based RNAi screens in Drosophila. Genome Biol. 8, R203 (2007).
28. Booker, M. et al. False negative rates in Drosophila cell-based RNAi screens: a case study. BMC Genomics 12, 50 (2011).
29. Hao, L. et al. Limited Agreement of Independent RNAi Screens for Virus-Required Host Genes Owes More to False-Negative than False-Positive Factors. PLoS Comput Biol 9, e1003235 (2013).
30. Liu, T., Sims, D. & Baum, B. Parallel RNAi screens across different cell lines identify generic and cell type-specific regulators of actin organization and cell morphology. Genome Biol. 10, R26 (2009).
31. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 43, D726–D736 (2015).
32. Shannon, P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498–2504 (2003).
33. Montojo, J. et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics 26, 2927–2928 (2010).

Chapter 3

Genetic engineering to rescue fertility

Nicholas Rui Yuan Ho, Abul Usmani, Michael Rieger, Nathan Kopp, Prabagaran Esaky, Masato Hoshi, Joseph Dougherty, Donald F Conrad

Chapter 3 - Introduction

While identification of fertility genes is important, the eventual goal of reproductive research is to fix the defects and restore fertility in patients. I envision two approaches to accomplish this goal. The first naïve, method is to deliver a plasmid expressing a transgene copy of the defective gene into the germ cells. The alternative approach is to fix the defective copy of the gene in the genome of the germ cells.

Transgenic gene expression works by placing a copy of a gene (without introns) downstream of a ubiquitous promoter (e.g. CMV) and integrating it into the genome of a cell in the hope that this will compensate for the defective endogenous gene. Breeding a transgenic mouse with a knockout mouse has successfully rescued spermatogenesis in their progeny[1–3]. Direct delivery of a transgene expression construct has also successfully improved spermatogenesis in knockout mice[4–6], but there have been many more reports about such methods inhibiting spermatogenesis instead[7–10].

Given the possibility of transgene delivery backfiring, I also explored the possibility of fixing the endogenous "broken" copy of the gene. CRISPR/cas9 is the most promising system at present. It uses a bacterial protein (cas9) to make double stranded DNA breaks based on a 26 base pair sequence targeting RNA. The endogenous mammalian DNA repair mechanisms then randomly repair the break by inserting or deleting random base pairs. If a homologous sequence is present, it will instead repair the broken DNA using the homolog as a template. By directly injecting the cas9 mRNA, targeting RNA, and single stranded homologous DNA into the nucleus of mouse embryos, you can make transgenic mice with relatively high efficiency[11,12].

In this chapter, I will present the work I have done to adapt the direct DNA delivery technique described in the previous chapter to work with these two approaches. To perform transgene delivery, I selected mlh3 as the gene to study because there was an available knockout mouse that had been shown to have completely arrested spermatogenesis. I transfected the knockout and wild-type mouse testes with an expression plasmid that had a cDNA copy of mlh3 cloned into it. This was sufficient to determine the ability of transgene delivery to rescue spermatogenesis or if it had deleterious effects instead. For endogenous gene correction I used the all-in-one plasmid pioneered by Feng Zhang's group[11] to determine if it would be sufficient to cause double stranded breaks in the male germ cells, the first step in correcting defecting genes.

Chapter 3 - Results

Transgene delivery

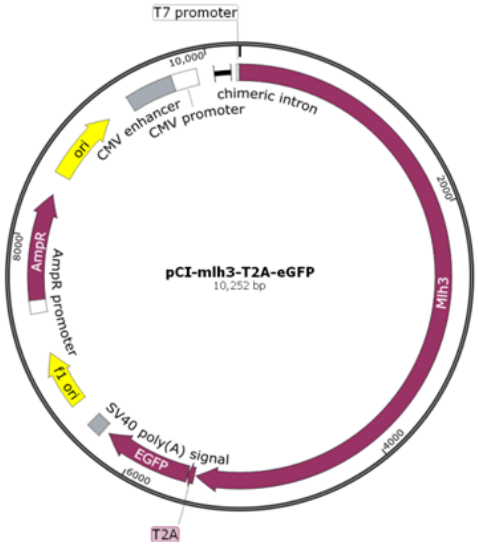


Figure 3.1: Plasmid for mlh3 rescue

Sequence of plasmid is in Data 3.S1.

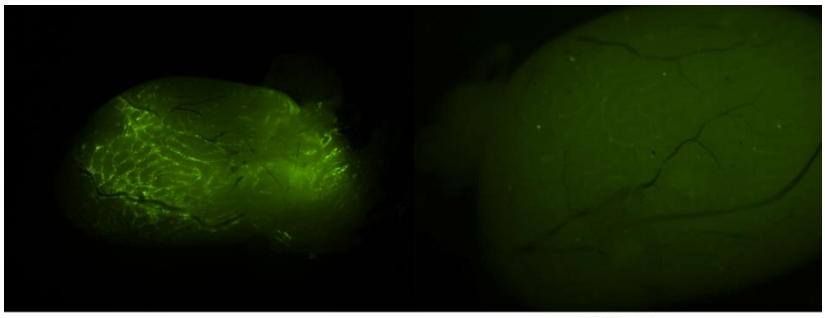


Figure 3.2: Expression of eGFP in testis after injection

I injected a plasmid expressing mlh3 and eGFP into the testis of mlh3 lnockout mice and WT C57Bl6 mice **(Figure 3.1)** (I used GFP as a marker for mlh3 expression). While GFP expression in WT testis peaked 3 days after the injection, the expression rapidly decreased, becoming almost undetectable after 50 days **(Figure 3.2)**. The injected mlh3 knockout mice were bred with WT female mice to see if any offspring could be produced. However, after 6 months of breeding there were no offspring. This data suggests that the plasmid was not effective at rescuing spermatogenesis.

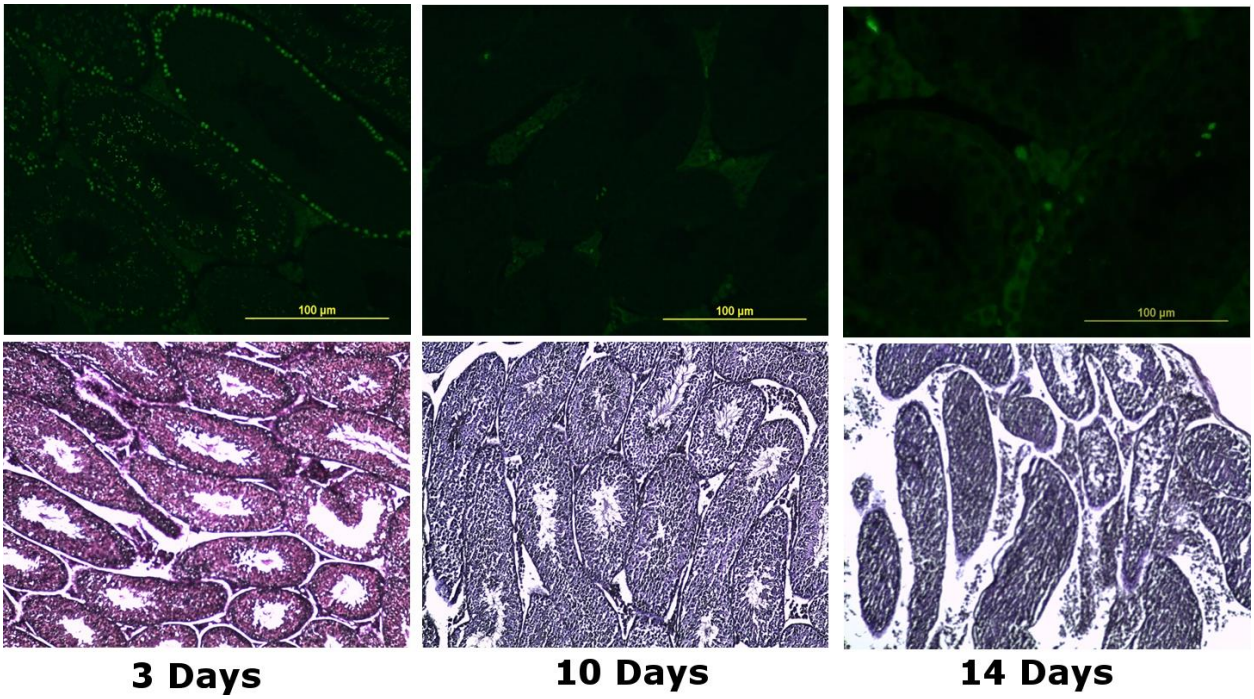


Figure 3.3: Morphology of testis after mlh3 expression plasmid injection
The top row show the eGFP expression for the testis while the bottom row shows the hematoxylin and Eosin staining for the same sample.

To explain why this would be the case I performed histology of the WT testis after plasmid injection at different time points. We found that eGFP expression was almost undetectable 10 days after the injection and the morphology of the testis indicated severe impairment of spermatogenesis **(Figure 3.3)**. This data seems to suggest that overexpression of mlh3 actually has a detrimental effect on spermatogenesis.

Endogenous gene repair

Since my attempts at spermatogenesis rescue by transgene expression had failed, I worked on an alternative approach to rescue spermatogenesis instead. I hoped to accomplish this by 'fixing' the deleterious mutation in the affected gene to the wild-type sequence using the CRISPR/cas9 system developed by Feng Zhang **(Figure 3.4)**.

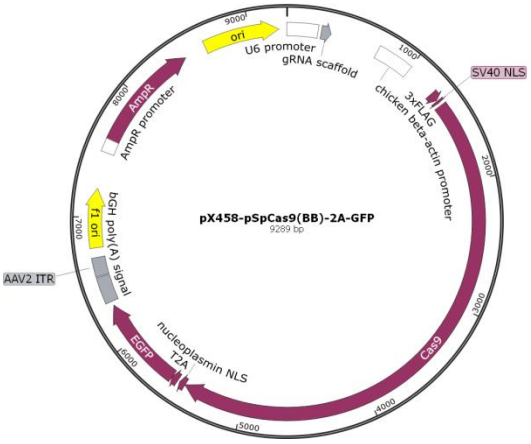


Figure 3.4: CRISPR/cas9 plasmid system

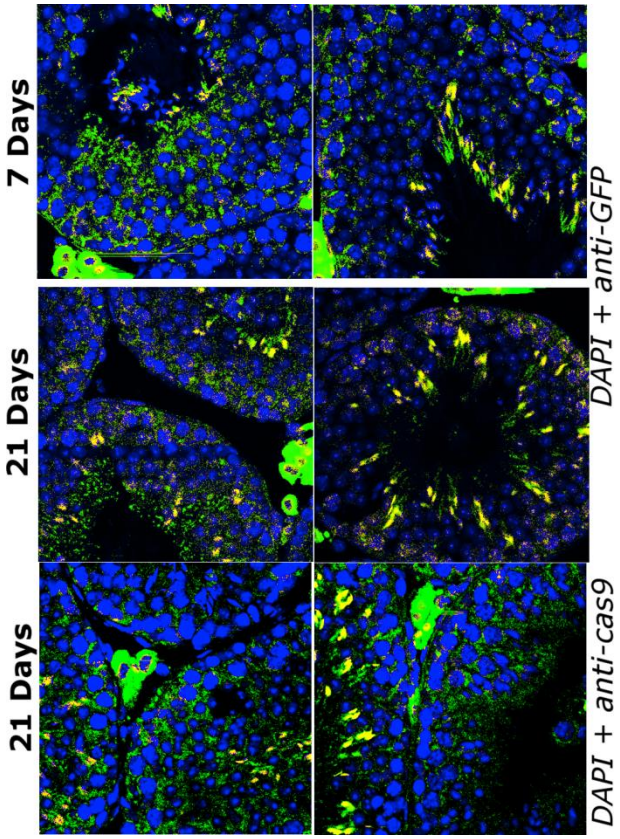


Figure 3.5: Immunofluorescence of cas9 plasmid injected testes

Blue is DAPI stain, Green is eGFP (top two rows) or cas9 (bottom row) antibody

The first benchmark we used was to detect if cas9 could be delivered and consistently expressed over a prolonged period. We were able to detect eGFP and cas9 expression even three weeks

after one injection of the plasmid into wild type mouse testes **(Figure 3.5)**. Between 10%-30% of the tubules showed evidence of cas9/GFP expression.

I then made four small guide RNA constructs against tyr (tyrosinase precursor), a gene when knocked out in mice produces a white coat color phenotype (v.s. normal black color). I called them tyr136, tyr184, tyr211 and tyr237 small guide RNAs after the nucleotide position in the gene where it targets. These guide RNAs were cloned into the cas9 plasmid and the constructs were injected into WT mouse testis which were then bred with female tyr KO mice from Jackson labs.

Since no white offspring were produced **(Table 3.1)**, I considered two possible explanations; either the four small guide RNA constructs are not effective or that it could be creating mutations at such a low rate that the breeding study was insufficient to detect the changes. To resolve this question I transfected the same constructs into N2a cell lines and WT mouse testes and used their genomic DNA to prepare a deep sequencing library across the tyr locus. I grouped substitutions, insertions, and deletions as mutations when calculating the mutation rate.

	Tyr136	Tyr184	Tyr211	Tyr237
Black offspring	36	8	31	27
White offspring	0	0	0	0

Table 3.1: Offspring coat colors from the injected WT X tyr KO cross

Black/White offspring indicates that the directed mutation failed or succeeded respectively.

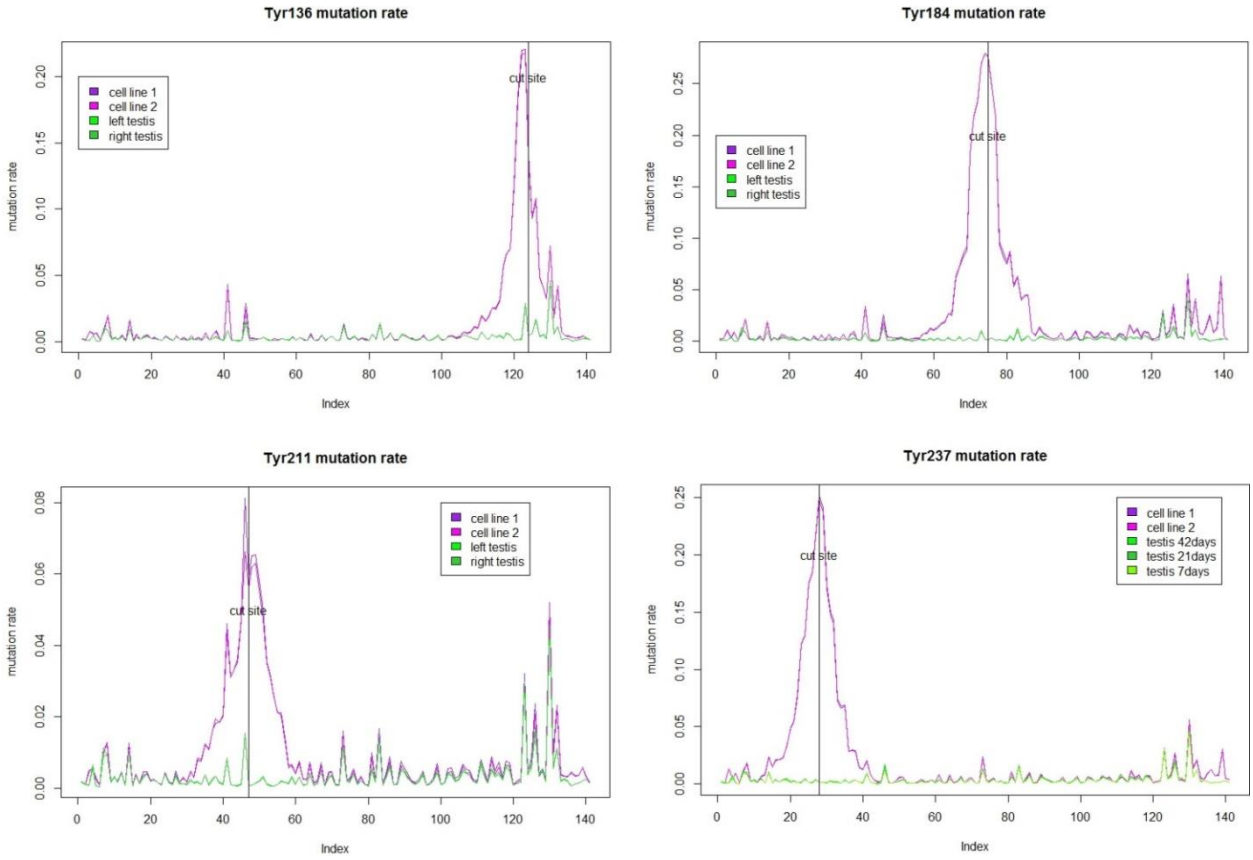


Figure 3.6: Mutation rate of tyr constructs in N2a cell line and mouse testes

The x axis shows the coordinates of section of the tyr gene that was targeted, with a vertical line indicating the targeted cut site. The y axis indicates the mutation rate.

The mutation rate of the constructs in the N2a cell line ranged between 8-28%. However, the exact same constructs had no detectable mutations in the testes **(Figure 3.6)**. Since previous experiments show that cas9 is delivered and expressed in the testes for prolonged periods of time, I eliminated the possibility of it being a plasmid delivery or expression issue.

Because *in vitro* cell line transfection is more efficient than *in vivo* testis transfection I considered that this might be a dosage issue. I took the tyr237 construct and injected it into WT mouse testes up to ten times, spaced three to four days apart, in order to boost the *in vivo* transfection rate. Furthermore, I also FACS sorted one WT testis that had undergone ten

injections to isolate a million eGFP positive and negative cells. This was performed to isolate the cells with cas9 expression to see if it would increase the sensitivity for detecting mutation rates. The samples' genomic DNA was once again used to prepare a deep sequencing library to determine the mutation rate.

I found that the mutation rate at the target site was undetectable under all conditions **(Figure 3.7)**. Although GFP positive cells appear to have an elevated mutation rate downstream of the target site, this is probably due to stochastic selection of cells elevating natural variants by chance rather than evidence of cas9 activity.

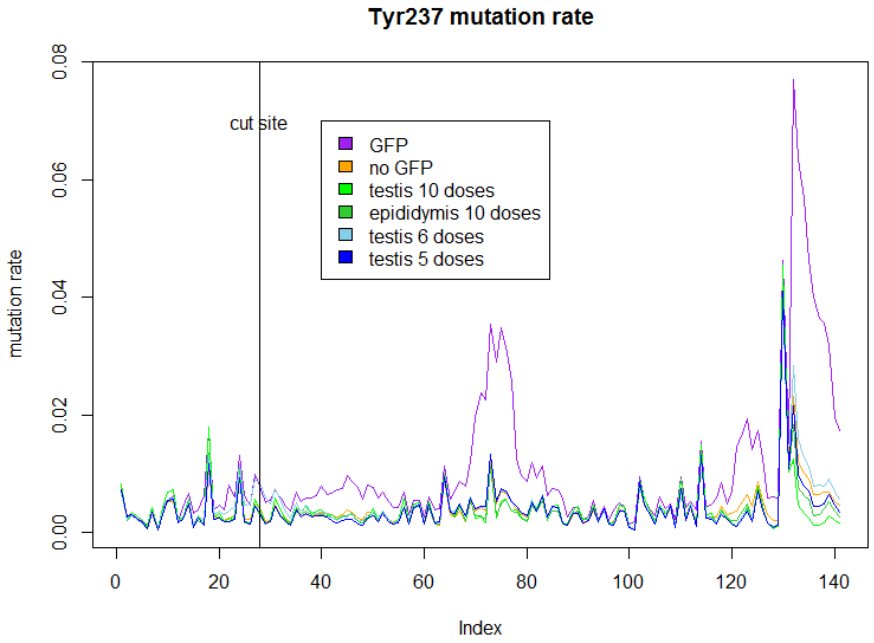
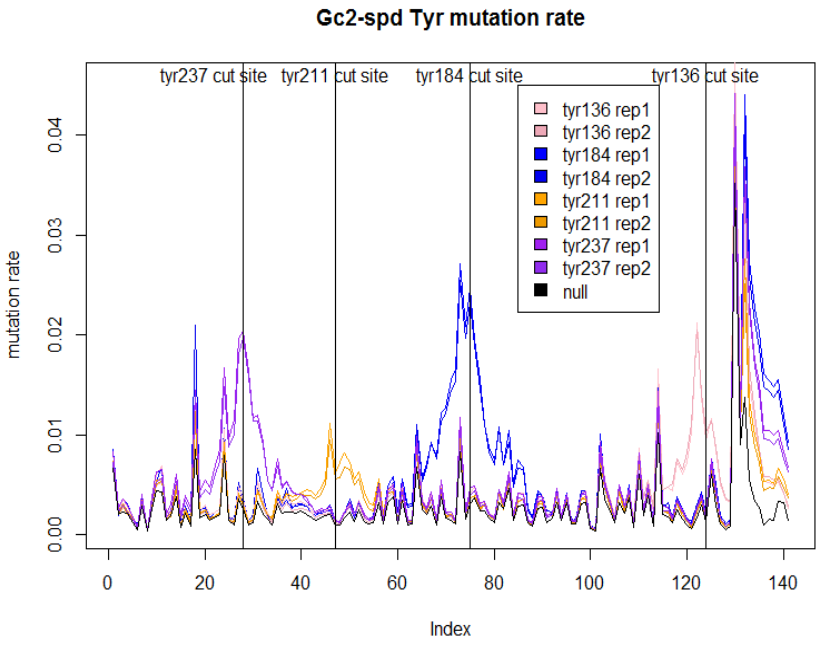


Figure 3.7: Mutation rate of tyr237 cas9 constructs in mouse testes (Higher dosage)
See Figure 3.6 for graph axis explanation

Another possible explanation for why the cas9 constructs were not working is that male germ cells have some mechanism that inhibits cas9 function (Either an active pathway or DNA packing). I decided to use a mouse spermatocyte cell line, Gc2-spd as the model system to

answer this question. If this hypothesis is true, we would expect the mutation rate using the same constructs to be much lower than in the N2a cell line despite the higher transfection rate.

There was a detectable mutation rate in the Gc2-spd cell line for all the constructs, but it was up to ten times lower than in the N2a cell lines, despite increasing the amount of DNA and transfection reagent **(Figure 3.8)**. A previous attempt using identical DNA and transfection reagent amounts in the Gc2-spd cell as the N2a cells did not create any detectable mutation rate **(Data not shown).**



This data suggests that there is some mechanism in the germ cell that inhibits cas9 activity.

Figure 3.8: Mutation rate of tyr cas9 constructs in Gc2-spd cell line

See Figure 3.6 for graph axis explanation

Chapter 3 - Discussion

Overexpressing transgenic copies of genes can have just as much deleterious consequences as knocking them out in the genome[7–10]. Studies that rescue fertility by transgenic expression attempt to ensure that not too many copies of the transgene are present in the genome, presumably to avoid the deleterious effects of massive oversexpression[5,6]. This suggests that a naïve transfection of a pool of transgenic genes is unlikely to boost spermatogenesis and might even cause detrimental effects.

To fix infertility due to spermatogenesis problems, we should be fixing the endogenous 'broken' genes instead. I have shown that the previously presented low cost *in vivo* male germ cell transfection method cannot be used to work with the CRISPR/cas9 system. The guide RNA constructs also showed lower efficiency when transfected into a spermatocyte cell line compared to a neuronal cell line, despite an increase in the transfection reagents.

Why did the tyr constructs work in the Gc2-spd cell line when it did not work in the testes? My hypothesis is that the Gc2-spd cell line contains proportionally more stem cells than the testis tissue does. Other groups have shown that it is possible to manipulate the genomes of spermatogonial stem cells[13]. My hypothesis is that if you can introduce cas9 before the cells develop into germ cells, it will be possible to edit their genomes.

If I could increase the *in vivo* transfection rate it would increase the chances of delivering the CRISPR constructs into primordial germ cells, where they could work. To accomplish this,

future work will have to involve more technically challenging protocols such as efferent duct injection and/or lentivirus infection.

If a working approach could be found that fixes spermatogenesis in mice via genetic engineering, we could also use the same system to make new transgenic animals for other disease models. More experiments will also be needed to ensure the safety and minimize off-target effects, but there is the potential to use the system to repair all sorts of genetic defects, not just the ones that affect fertility.

Chapter 3 - Methods

Plasmid cloning

The mlh3 overexpression plasmid was cloned using the sequence obtained from Mammalian Gene Collection (MGC:100285) with the pCI mammalian expression plasmid as the backbone (Promega: E1731).

CRISPR/cas9 constructs were cloned into the pX458 plasmid backbone (Addgene: 48138) following the previously published protocol[14]. Primers used for the constructs can be found in **Table 3.S1**.

Cell line transfection

N2a cells were maintained in Minimum Essential Media (MEM) supplemented with 10% Fetal Bovine Serum (FBS) until they reached 60% confluency in 6 well cell culture plates. Each well of cells was transfected with 1µg of plasmid DNA using 3.75µl of Lipofectamine® 3000 (Life

Technologies: L3000008). The cells were then incubated at 37°C for 2 days, with daily media replacement, before the genomic DNA was collected.

Gc2-spd cells were treated the same way and the N2a cells, with the only change being each well was transfected with of 2.5µg of plasmid DNA and 7.5µl of Lipofectamine® 3000 instead.

Cell genomic DNA was extracted using DNeasy blood and Tissue Kit (Qiagen: 69504) using the manufacturer's suggested protocol.

DNA Library preparation

Extracted genomic DNA was amplified using a Q5 hot-start high fidelity polymerase (NEB: M0493L) using the manufacturer's protocol and custom primers (Table S2). This PCR product was purified using Minelute columns (Qiagen: 28004). The overhang from the custom primers was used to attach Illumina sequencing adapters and indices via a second round of PCR. This PCR product was purified using Agencourt AmPure XP beads (Beckman-Coulter: A63880) following the manufacturer's protocol. Each library was run on a 3% agarose gel to ensure that the library size was around 300bp. Libraries from multiple samples were pooled and run on a single lane of Illumina MiSeq.

Data Analysis

Paired end reads were mapped to the mouse mm10 genome assembly using STAR[15]. Reads across the tyr locus was verified to make up more than 95% of the total reads in the sample. These reads were condensed using samtools' mpileup command, using the –A option to increase the max depth to 2,000,000[16]. This data was then run through a custom analysis script to calculate and plot the mutation rate at each base pair in the locus. A mutation was classified as

any base at that position which did not match the reference genome, including insertions and deletions.

LOCUS Exported 10252 bp ds-DNA circular SYN 08-JAN-2015
DEFINITION .
ACCESSION .
VERSION .
KEYWORDS Untitled
SOURCE synthetic DNA construct
ORGANISM synthetic DNA construct
REFERENCE 1 (bases 1 to 10252)
AUTHORS .
TITLE Direct Submission
JOURNAL Exported Thursday, Jan 08, 2015 from SnapGene Viewer 2.5.0
http://www.snapgene.com
FEATURES Location/Qualifiers
source 1..10252
/organism="synthetic DNA construct"
/mol_type="other DNA"
CDS 1..5481
/codon_start=1
/gene="Mlh3"
/product="mutL homolog 3 (E coli)"
/note="Mlh3"
/note="color: #993366"; direction: RIGHT
/db_xref="GI:51259774"
/db_xref="GeneID:217716"
/db_xref="MGI:1353455"
/protein_id="AAH79861.1"
/translation="MIRCLSDDVKTKLRSGLAISSLGQCVEELTLNSIDAEATCVAIRV
NMETFQVQVIDNGLGMAGDDVEKVGNRYFTSKCHSVRDLENPAFYGFRGEALASIADMA
GAVEISSKKNTTLKTFVKMFQNGKALATHEADLTRPSVGTTVTVYNLFYQFPVRRKSMD
PRLEFEKVRQRVEALSLMHPSISFSLRNDVSGSMVLQLPKTKDICSRFCQIYGLGKSQK
LREIRFKYKEFEFSGYISSEAHYNKNMQFLFVNRRLVLRTKLHKLIDFLLRKESIICRP
KNGSASRQMNSSPRHRSASELHGIYVINVQCPFCEYDVCIEPAKTLIEFQSWDTVLICI
QEGVKRFLKQEKLFVELSGEDIKEFNEDNGFSLFGTTLQTHVSTHEKCDQSSFREACNK
ILDSYEMFNLQSKAVKRIATLENKTRQNPGDSETIRKKTVGSLYTDASDGPCYSKSVES
VLQDSNNSAYLEPRVSEEEVAKTSHSGENEKWKKSFLENKTSGRIHETSPKMFSSPIQM
HHLLEEREADLEMQTISSTVNVMAANIPQNNDIPSQLEKWKDAPEVGCQPLPFETTLLR
VRGTQRKKERRKKEPSSRGRVNVFSYGQVKLCSTGFITHVVQSEHAKSTETEHSFKNYA
RPGPVSAQETFGKRTHHAIETPDSSDLTSTLSKESSQPPNKRFCRTNTGYGTENKPVAT
DDNLALFQESCKESHTDRLLPDASSFPWCRYVSDGCRKIDKRGSFKQVVRRKLSLRSQV
GSLEKFKRQYGKVSSSLDTEKDNNTEVRTHLDPQNEPDVLLKDKSHLDMSDGCEITTVE
HSETCQPLSPILYPEKILFSKEDRLEQMPHLRESPITLEELSHCNRKADVEKSAASLAS
KLSKLKDSEKEMQTVGMTGHTSELPDSNPSWKDNSQCTRLDLDFCELLKNKLEKIESDM
LPMADSATEDGPINKNSELHPNNTTDDTEKPETPLLFPCNDSKISRDSDVLIRTSEQPT
GNPDSVGKVIMSQVEDGIGSQGGVCPQGDESKARSCSKNEPNAHCMDWQQHFDVTLGRM
VYINRMTGLSTFVAPTDDLHTACTKDLTTVAVDVLLGNDAVDAAAAAVSEPLQSLFSEW
SNPVFARYPEVAVDVSSGQAESLAVKIHNVLYPYRFTKEMIHSVKVLQQVDNKFIACLM
STRMDEDGRTGGNLLVLVDQHAAHERIRLEQLITDSYEKQDPQSAGRKKLLSSTIIPPL
AITVSEEQRRLLRSYHKHLEDLGLELLFPDASDSLILVGKVPLCFVEREASELRRGRST
VTKSIVEELIREQVELLQTTGGIQGTLPLTVQKVLASQACHGAIKFNDRLSLEESCRLI
EALSLSQLPFQCAHGRPSMLPLADLDHLEQEKQVKPNLAKLRKMVRAWHLFGKTEQNLQ
QPIRPCEPP"

CDS 5493..5546
/codon_start=1
/product="2A peptide from Thosea asigna virus capsid protein"
/note="T2A"
/note="Eukaryotic ribosomes fail to insert a peptide bond between the Gly and Pro residues, yielding separate polypeptides."
/translation="EGRGSLLTCGDVEENPGP"
CDS 5547..6260
/codon_start=1
/product="enhanced GFP"
/note="EGFP"
/note="mammalian codon-optimized"
/translation="VSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLK
FICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDG
NYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKV
NFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLE
FVTAAGITLGMDELYK"
polyA_signal 6308..6429
/note="SV40 poly(A) signal"
/note="SV40 polyadenylation signal"
rep_origin 6610..7065
/direction=RIGHT
/note="f1 ori"
/note="f1 bacteriophage origin of replication; arrow indicates direction of (+) strand synthesis"
promoter 7397..7501
/gene="bla"
/note="AmpR promoter"
CDS 7502..8362
/codon_start=1
/gene="bla"
/product="beta-lactamase"
/note="AmpR"
/note="confers resistance to ampicillin, carbenicillin, and related antibiotics"
/translation="MSIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAEDQLGARVGYI
ELDLNSGKILESFRPEERFPMMSTFKVLLCGAVLSRIDAGQEQLGRRIHYSQNDLVEYS
PVTEKHLTDGMTVRELCSAAITMSDNTAANLLLTTIGGPKELTAFLHNMGDHVTRLDRW
EPELNEAIPNDERDTTMPVAMATTLRKLLTGELLTLASRQQLIDWMEADKVAGPLLRSA
LPAGWFIADKSGAGERGSRGIIAALGPDGKPSRIVVIYTTGSQATMDERNRQIAEIGAS
LIKHW"
rep_origin 8533..9121
/direction=RIGHT
/note="ori"
/note="high-copy-number ColE1/pMB1/pBR322/pUC origin of replication"
enhancer 9332..9711
/note="CMV enhancer"
/note="human cytomegalovirus immediate early enhancer"

```
     promoter        9712..9915
                     /note="CMV promoter"
                     /note="human cytomegalovirus (CMV) immediate early
                     promoter"
     intron          10051..10183
                     /note="chimeric intron"
                     /note="chimera between introns from human beta-globin and
                     immunoglobulin heavy chain genes"
     promoter        10228..10246
                     /note="T7 promoter"
                     /note="promoter for bacteriophage T7 RNA polymerase"
ORIGIN
        1 ggcgcgccga attcggcacg agggtgagag ttgatggagg agattcgagg tgattatttt
       61 ccagtcagag aaggaagcca gtgtctgcca ctccctccac atgtgtccct gccatgatca
      121 ggtgtctatc agatgacgta aaaaccaagt tgcgttccgg tttagccata agctccttgg
      181 gccagtgtgt tgaagaactt acccttaaca gtattgatgc tgaagcaaca tgtgtggcca
      241 tcagagtgaa tatggaaacc ttccaagttc aagtgataga caatggactt gggatggcgg
      301 gggacgatgt agagaaggtg ggaaaccggt attttactag taaatgccac tcagtgcggg
      361 acttggagaa cccagcattt tatggcttcc gaggagaggc cttggcaagt atagccgaca
      421 tggctggtgc tgtggagatt tcatccaaga aaaacacaac actgaaaacc tttgtgaaaa
      481 tgtttcagaa tggaaaagcc cttgccaccc atgaggctga tttgaccaga ccaagtgtgg
      541 ggactacagt aacggtctat aacctgtttt accagtttcc tgtgcggagg aaaagcatgg
      601 atcctagact agagtttgag aaagttcggc agagggtaga agctctctca cttatgcacc
      661 cctccatttc tttctctttg aggaacgatg tatctggatc catggttctt cagctcccta
      721 aaaccaaaga catatgctct cgattctgtc aaatttacgg attgggcaag tcccaaaagt
      781 taagagaaat acgttttaaa tacaaggaat ttgagttcag tggctacatc agctctgaag
      841 cacactacaa taagaatatg cagtttttgt ttgtgaacag aagactagtt ttaagaacaa
      901 agttgcataa acttattgac tttttattaa gaaaagaaag cattatatgc aggccaaaga
      961 atggctctgc cagtaggcaa atgaattcaa gtcctcgaca ccgttctgcc tcagaactcc
     1021 acgggatata tgtaatcaat gtgcagtgcc ctttctgtga gtatgatgtc tgcatagagc
     1081 cagccaaaac tctgattgag tttcagagct gggataccgt gttgatttgt attcaggaag
     1141 gagtaaaaag gtttttaaag caagaaaaat tatttgtaga attatcaggt gaagatatta
     1201 aggaatttaa tgaagataat ggttttagtt tgtttggcac gactcttcag acacatgtgt
     1261 ctactcatga gaagtgtgac cagagcagtt tccgggaagc gtgtaataaa attctggatt
     1321 cctatgaaat gtttaatttg cagtcaaaag ctgtgaaaag aatagctact ctagaaaata
     1381 aaaccagaca aaaccctggt gattcagaaa ctatcagaaa aaagacagtg ggctcattgt
     1441 acacagatgc atcggatggc ccgtgctata gtaaatcggt agagtctgtt ttacaggaca
     1501 gcaacaacag tgcttactta gaaccgaggg tgtcagaaga agaggtagcc aaaacatcac
     1561 actccggaga aaatgagaaa tggaaaaaat ctttttgga aaataagact tcaggaagga
     1621 tacatgaaac cagtccaaaa atgttttcaa gccccatcca aatgcatcac ctccttgagg
     1681 agagagaggc agatctggaa atgcagacaa taagtagtac tgttaatgtc atggctgcca
     1741 acattcccca aaataatgac attccgagtc aactggagaa atggaaagat gctcctgaag
     1801 tggggtgcca acctctgcct tttgagacaa ccttattaag ggtacggggt actcagagga
     1861 agaaggaaag aaggaaaaag gagcccagta gtcgtggaag agtaaatgtt tttagttatg
     1921 gacaagttaa attatgctcc actggcttta taactcatgt agtacaaagt gagcacgcta
     1981 aatcaactga aacagaacat tcatttaaaa attatgctcg acctggtcct gtaagtgccc
     2041 aagaaacatt tggaaaaaga acacaccatg caattgagac tccagacagc agtgatttaa
     2101 caagcacttt aagtaaagaa tccagtcaac cgcccaacaa aaggttttgc agaacaaata
     2161 caggttacgg gacagagaac aaacctgtag caacagatga caacttggct ctttttcagg
     2221 aaagctgtaa agaatcacac acagatcgcc ttttgcctga tgcatcctcc ttcccatggt
     2281 gtagatatgt ttccgatggt tgtaggaaaa tagataaaag gggttccttc aaacaggtag
```

2341 tccgtaggaa gctaagcttg cgttcacaag tagggtcttt agagaagttt aagaggcagt
2401 atgggaaggt cagcagttcc ctagatacag aaaaggataa taacactgaa gtcaggactc
2461 atcttgatcc tcaaaatgaa cccgatgttc ttctgaaaga caagagccac ttagatatgt
2521 ctgatggttg tgagatcact actgtggagc acagtgagac ttgtcaacca ttaagtccca
2581 tcctgtaccc agaaaaaatt ttattttcca aagaagatcg cttagaacag atgcctcatt
2641 tgagagaaag tcctataact ctggaagaat tatctcactg taacagaaaa gctgatgttg
2701 agaagtccgc tgcatcactg gcttctaaat tatccaaact aaaggattct gaaaaagaga
2761 tgcaaaccgt ggggatgaca ggtcatacta gtgaacttcc agattcaaat cccagttgga
2821 aagataatag ccagtgcact aggttagact tggatttttg tgaattatta aaaaacaaac
2881 ttgaaaaaat agagagtgat atgcttccaa tggcagattc tgccacagag gatggtccca
2941 tcaataaaaa cagtgaacta catcctaaca atacaacgga tgacacagag aaaccagaaa
3001 ctcctttgct gttcccctgt aatgattcta aaatcagcag agattcagat gttcttatca
3061 gaacttcaga acaacctaca ggaaaccctg actctgtcgg taaagtgata atgagtcagg
3121 tagaggatgg cattggcagc caaggtggag tctgtcccca gggtgatgaa tctaaggcaa
3181 gatcttgttc caaaaatgaa ccaaacgcac actgtatgga ttggcagcag cattttgatg
3241 taaccctggg aagaatggtt tacatcaaca gaatgacagg acttagcaca ttcgttgctc
3301 caactgacga ccttcatact gcttgtacta aagatctgac aactgtggct gtggatgtcc
3361 tgcttgggaa tgatgctgtg gatgctgctg ctgctgctgt cagtgaaccc cttcagtctc
3421 tgttttcaga atggagcaat ccagtgtttg ctcgataccc agaggttgct gttgatgtca
3481 gcagtggcca ggctgagagc ttagccgtta aaattcacaa cgtcctgtat ccctatcgct
3541 tcaccaaaga gatgattcac tcagtgaagg ttctccagca agtggataac aagtttattg
3601 cctgcttaat gagcacgagg atggatgagg atggccgaac aggtggaaac ctgttagtcc
3661 tggtggacca gcatgctgcc catgaacgca ttcgtttgga gcagcttatt actgattcct
3721 atgagaaaca agatccacaa agcgctggcc ggaagaaatt attgtcttcc acaataatcc
3781 ctccactggc aatcaccgtg tcagaggaac aaaggagact cttacggtct taccacaaac
3841 atttagaaga tctggggctt gagttgctct ttccagatgc tagtgattct ctgatcctgg
3901 tgggaaaagt gccgctctgc tttgtagaga gagaagctag tgagcttcga agaggacgct
3961 ctactgtgac taagagtatt gtggaggaat taattcgaga acaagtggag ctgctccaga
4021 ccacaggagg tatccaaggg acactgccac tgactgtcca gaaggtgttg gcctcccagg
4081 cctgccatgg ggctattaag tttaatgatc gtctgagcct agaagagagc tgccgcctca
4141 tcgaagctct gtccttgtcc cagctgccat ttcagtgtgc tcatgggaga ccctcaatgc
4201 tgcccttagc tgacctggac cacttggagc aggaaaaaca ggttaaaccc aaccttgcta
4261 aacttcgcaa aatggttcgt gcctggcatc tctttggaaa aacagagcag aacctgcagc
4321 agcctatacg tccttgtgag cctccatgag gagaggattc tggagtgtaa ggagacaagg
4381 gagtgccgtg catcccgagc aggagcagtg cagctgtggg caggtcggcg gccctgagcg
4441 ggctggcaca tcagtccccc tgagcagatg gagcaggcac gtgcactcaa gcctaagggt
4501 agtttatttc tttgcatcca tgcacacagg agcttgacat ataataccta tcttttgtaa
4561 gttgatttag tgataaaatg taatgatttt gtaattggtg agttggctta tgtttgaggg
4621 gcgcagctat tgtttttagc agttttcccc agcctctcag tttatattac gtgaggatgc
4681 taagccctaa gcgctggtct gctcttctct gagcccctgg ctctgcccct ccccatccat
4741 ttctcttttg catttgtctc cttcacttca tacctctgct tcttcacatt gtgctttaca
4801 gacttacggt gtttctctgc tcattataaa aatatttccc gccaggcagc ggtggctcac
4861 acctttagtc ccagtacttg ggaggcagag gcaggtggat ttctgatttc aaggccagcc
4921 tggtctgcaa agtgagttcc aggacagcca cagggctaca aagagaaacc ctgtctcaaa
4981 aaaaccaaaa caacaacaac aacaacaaaa cttcctgatg tcttccagag agactaaatt
5041 atattaggga ttaaaagtta tttatagctg ggtgtagtga catatacctg taatcctagt
5101 acctgggagg ctgagacaag tccgtgacag ggcaccattt gcctaggctg gatatatagc
5161 aagaccttga ctcaaaataa ataagtaaac tatagacaaa gagagacaca aagacagaat
5221 agagaaagtt tgaaaagaat ttttttaatt ctcttggtaa cctggctgtc ctggaactgg
5281 aactcagaga tccccctttg ctgcctccga gtgctgggat taacggtgcg cgccaccact
5341 gcctggcaag aaaaattaca accctgtccc tggtttttt gatagtctta ctggttttta

5401 aaaccagagg tgaatgtttc tattctgaac taataaaaca ctaaaaaata aaaaaaaaaa
5461 aaaaaaaaaa aggctctctc agcggccgcg gagagggcag aggaagtctg ctaacatgcg
5521 gtgacgtcga ggagaatcct ggcccagtga gcaagggcga ggagctgttc accggggtgg
5581 tgcccatcct ggtcgagctg gacggcgacg taaacggcca caagttcagc gtgtccggcg
5641 agggcgaggg cgatgccacc tacggcaagc tgaccctgaa gttcatctgc accaccggca
5701 agctgcccgt gccctggccc accctcgtga ccaccctgac ctacggcgtg cagtgcttca
5761 gccgctaccc cgaccacatg aagcagcacg acttcttcaa gtccgccatg cccgaaggct
5821 acgtccagga gcgcaccatc ttcttcaagg acgacggcaa ctacaagacc cgcgccgagg
5881 tgaagttcga gggcgacacc ctggtgaacc gcatcgagct gaagggcatc gacttcaagg
5941 aggacggcaa catcctgggg cacaagctgg agtacaacta caacagccac aacgtctata
6001 tcatggccga caagcagaag aacggcatca aggtgaactt caagatccgc cacaacatcg
6061 aggacggcag cgtgcagctc gccgaccact accagcagaa cacccccatc ggcgacggcc
6121 ccgtgctgct gcccgacaac cactacctga gcacccagtc cgccctgagc aaagacccca
6181 acgagaagcg cgatcacatg gtcctgctgg agttcgtgac cgccgccggg atcactctcg
6241 gcatggacga gctgtacaag gaattctaac tagagctcgc tgatcacccg ggttcgagca
6301 gacatgataa gatacattga tgagtttgga caaaccacaa ctagaatgca gtgaaaaaaa
6361 tgctttattt gtgaaatttg tgatgctatt gctttatttg taaccattat aagctgcaat
6421 aaacaagtta acaacaacaa ttgcattcat tttatgtttc aggttcaggg ggagatgtgg
6481 gaggtttttt aaagcaagta aaacctctac aaatgtggta aaatcgataa ggatccgggc
6541 tggcgtaata gcgaagaggc ccgcaccgat cgcccttccc aacagttgcg cagcctgaat
6601 ggcgaatgga cgcgccctgt agcggcgcat taagcgcggc gggtgtggtg gttacgcgca
6661 gcgtgaccgc tacacttgcc agcgccctag cgcccgctcc tttcgctttc ttcccttcct
6721 ttctcgccac gttcgccggc tttccccgtc aagctctaaa tcgggggctc cctttagggt
6781 tccgatttag tgctttacgg cacctcgacc ccaaaaaact tgattagggt gatggttcac
6841 gtagtgggcc atcgccctga tagacggttt ttcgcccttt gacgttggag tccacgttct
6901 ttaatagtgg actcttgttc caaactggaa caacactcaa ccctatctcg gtctattctt
6961 ttgatttata agggattttg ccgatttcgg cctattggtt aaaaaatgag ctgatttaac
7021 aaaaatttaa cgcgaatttt aacaaaatat taacgcttac aatttcctga tgcggtattt
7081 tctccttacg catctgtgcg gtatttcaca ccgcatatgg tgcactctca gtacaatctg
7141 ctctgatgcc gcatagttaa gccagccccg acacccgcca acacccgctg acgcgccctg
7201 acgggcttgt ctgctcccgg catccgctta cagacaagct gtgaccgtct ccgggagctg
7261 catgtgtcag aggttttcac cgtcatcacc gaaacgcgcg agacgaaagg gcctcgtgat
7321 acgcctattt ttataggtta atgtcatgat aataatggtt tcttagacgt caggtggcac
7381 ttttcgggga aatgtgcgcg gaacccctat ttgtttattt ttctaaatac attcaaatat
7441 gtatccgctc atgagacaat aaccctgata aatgcttcaa taatattgaa aaaggaagag
7501 tatgagtatt caacatttcc gtgtcgccct tattcccttt tttgcggcat tttgccttcc
7561 tgtttttgct cacccagaaa cgctggtgaa agtaaaagat gctgaagatc agttgggtgc
7621 acgagtgggt tacatcgaac tggatctcaa cagcggtaag atccttgaga gttttcgccc
7681 cgaagaacgt tttccaatga tgagcacttt taaagttctg ctatgtggcg cggtattatc
7741 ccgtattgac gccgggcaag agcaactcgg tcgccgcata cactattctc agaatgactt
7801 ggttgagtac tcaccagtca cagaaaagca tcttacggat ggcatgacag taagagaatt
7861 atgcagtgct gccataacca tgagtgataa cactgcggcc aacttacttc tgacaacgat
7921 cggaggaccg aaggagctaa ccgctttttt gcacaacatg ggggatcatg taactcgcct
7981 tgatcgttgg gaaccggagc tgaatgaagc cataccaaac gacgagcgtg acaccacgat
8041 gcctgtagca atggcaacaa cgttgcgcaa actattaact ggcgaactac ttactctagc
8101 ttcccggcaa caattaatag actggatgga ggcggataaa gttgcaggac cacttctgcg
8161 ctcggccctt ccggctggct ggtttattgc tgataaatct ggagccggtg agcgtgggtc
8221 tcgcggtatc attgcagcac tggggccaga tggtaagccc tcccgtatcg tagttatcta
8281 cacgacgggg agtcaggcaa ctatggatga acgaaataga cagatcgctg agataggtgc
8341 ctcactgatt aagcattggt aactgtcaga ccaagtttac tcatatatac tttagattga
8401 tttaaaactt cattttaat ttaaaaggat ctaggtgaag atccttttg ataatctcat

8461 gaccaaaatc ccttaacgtg agttttcgtt ccactgagcg tcagaccccg tagaaaagat
8521 caaaggatct tcttgagatc ctttttttct gcgcgtaatc tgctgcttgc aaacaaaaaa
8581 accaccgcta ccagcggtgg tttgtttgcc ggatcaagag ctaccaactc tttttccgaa
8641 ggtaactggc ttcagcagag cgcagatacc aaatactgtt cttctagtgt agccgtagtt
8701 aggccaccac ttcaagaact ctgtagcacc gcctacatac ctcgctctgc taatcctgtt
8761 accagtggct gctgccagtg gcgataagtc gtgtcttacc gggttggact caagacgata
8821 gttaccggat aaggcgcagc ggtcgggctg aacgggggggt tcgtgcacac agcccagctt
8881 ggagcgaacg acctacaccg aactgagata cctacagcgt gagctatgag aaagcgccac
8941 gcttcccgaa gggagaaagg cggacaggta tccggtaagc ggcagggtcg gaacaggaga
9001 gcgcacgagg gagcttccag ggggaaacgc ctggtatctt tatagtcctg tcgggtttcg
9061 ccacctctga cttgagcgtc gatttttgtg atgctcgtca gggggggcgga gcctatggaa
9121 aaacgccagc aacgcggcct ttttacggtt cctggccttt tgctggcctt ttgctcacat
9181 ggctcgacag atcttcaata ttggccatta gccatattat tcattggtta tatagcataa
9241 atcaatattg gctattggcc attgcatacg ttgtatctat atcataatat gtacatttat
9301 attggctcat gtccaatatg accgccatgt tggcattgat tattgactag ttattaatag
9361 taatcaatta cggggtcatt agttcatagc ccatatatgg agttccgcgt tacataactt
9421 acggtaaatg gcccgcctgg ctgaccgccc aacgaccccc gcccattgac gtcaataatg
9481 acgtatgttc ccatagtaac gccaataggg actttccatt gacgtcaatg ggtggagtat
9541 ttacggtaaa ctgcccactt ggcagtacat caagtgtatc atatgccaag tccgcccccc
9601 attgacgtca atgacggtaa atggcccgcc tggcattatg cccagtacat gaccttacgg
9661 gactttccta cttggcagta catctacgta ttagtcatcg ctattaccat ggtgatgcgg
9721 ttttggcagt acaccaatgg gcgtggatag cggtttgact cacggggatt tccaagtctc
9781 caccccattg acgtcaatgg gagtttgttt tggcaccaaa atcaacggga ctttccaaaa
9841 tgtcgtaata accccgcccc gttgacgcaa atgggcggta ggcgtgtacg gtgggaggtc
9901 tatataagca gagctcgttt agtgaaccgt cagatcacta gaagctttat tgcggtagtt
9961 tatcacagtt aaattgctaa cgcagtcagt gcttctgaca caacagtctc gaacttaagc
10021 tgcagaagtt ggtcgtgagg cactgggcag gtaagtatca aggttacaag acaggtttaa
10081 ggagaccaat agaaactggg cttgtcgaga cagagaagac tcttgcgttt ctgataggca
10141 cctattggtc ttactgacat ccactttgcc tttctctcca caggtgtcca ctcccagttc
10201 aattacagct cttaaggcta gagtacttaa tacgactcac tatagggtcg ac
//

Data 3.S1: mlh3 overexpression plasmid seqeuence genBank DNA file

tyr 211 sgRNA F	caccg ACGGTCATCCACCCCTTTGA
tyr 184 sgRNA F	caccg CTGAGGTCCAGATGGTGCAC
tyr 136 sgRNA F	caccg TCTGCCTGAAAGCTGGCCGC
tyr 237 sgRNA F	cacc GGGTGGATGACCGTGAGTCC

tyr 211 sgRNA R	aaac TCAAAGGGGTGGATGACCGT c
tyr 184 sgRNA R	aaac GTGCACCATCTGGACCTCAG c
tyr 136 sgRNA R	aaac GCGGCCAGCTTTCAGGCAGA c
tyr 237 sgRNA R	aaac GGACTCACGGTCATCCACCC

Table 3.S1: Primer sequences used in CRISPR/cas9 construct cloning
These constructs were cloned into the pX458 plasmid created by Feng Zhang for mammalian CRISPR/cas9 engineering.

tyr F Ill Adapter	CTCTTTCCCTACACGACGCTCTTCCGATCT NNNN CACCATGGATGGGTGATGGG
tyr R Ill Adapter	CTGGAGTTCAGACGTGTGCTCTTCCGATCT NNNN TGAGCACTGGCAGGTCCTAT
PE PCR F	AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT
PE PCR R	CAAGCAGAAGACGGCATACGAGAT XXXXXX GTGACTGGAGTTCAGACGTGTGCTCTTCCG

Table 3.S2: Primers used for genomic locus amplification and Illumina sequencing library preparation
The first two primers are the pair used for gDNA amplification of the tyr locus. The last two primers are the pair used for the second round of PCR for Illumina library preparation. XXXXXX in PE PCR R represents any six bases which can be used for sample indexing, which allows for pooling of samples in a single illumine lane.

References

1. Vogel, T., Speed, R. M., Ross, A. & Cooke, H. J. Partial rescue of the Dazl knockout mouse by the human DAZL gene. Mol. Hum. Reprod. 8, 797–804 (2002).
2. Walker, W. H. et al. Restoration of Spermatogenesis and Male Fertility Using an Androgen Receptor Transgene. PLoS ONE 10, e0120783 (2015).
3. Wu, X. et al. Transgene-Mediated Rescue of Spermatogenesis in Cldn11-Null Mice. Biol. Reprod. 86, (2012).
4. Yomogida, K., Yagura, Y. & Nishimune, Y. Electroporated Transgene-Rescued Spermatogenesis in Infertile Mutant Mice with a Sertoli Cell Defect. Biol. Reprod. 67, 712–717 (2002).
5. Higgy, N. A., Zackson, S. L. & van der Hoorn, F. A. Cell interactions in testis development: overexpression of c-mos in spermatocytes leads to increased germ cell proliferation. Dev. Genet. 16, 190–200 (1995).
6. Li, X., Mao, Z., Wu, M. & Xia, J. Rescuing Infertility of Pick1 Knockout Mice by Generating Testis-specific Transgenic Mice via Testicular Infection. Sci. Rep. 3, (2013).
7. Yamamoto, C. M. et al. Impairment of Spermatogenesis in Transgenic Mice With Selective Overexpression of Bcl-2 in the Somatic Cells of the Testis. J. Androl. 22, 981–991 (2001).
8. Huang, Z., Rivas, B. & Agoulnik, A. I. NOTCH1 Gain of Function in Germ Cells Causes Failure of Spermatogenesis in Male Mice. PLoS ONE 8, e71213 (2013).
9. Yan, W. et al. Overexpression of Bcl-w in the Testis Disrupts Spermatogenesis: Revelation of a Role of BCL-W in Male Germ Cell Cycle Control. Mol. Endocrinol. 17, 1868–1879 (2003).
10. Wang, Y.-L. et al. Overexpression of ubiquitin carboxyl-terminal hydrolase L1 arrests spermatogenesis in transgenic mice. Mol. Reprod. Dev. 73, 40–49 (2006).
11. Wang, H. et al. One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. Cell 153, 910–918 (2013).
12. Cong, L. et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. Science 339, 819–823 (2013).
13. Wu, Y. et al. Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. Cell Res. 25, 67–79 (2015).
14. Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8, 2281–2308 (2013).
15. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics bts635 (2012). doi:10.1093/bioinformatics/bts635
16. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma. Oxf. Engl. 27, 2987–2993 (2011).

Summary

Application of genomic technologies to study infertility

Nicholas Rui Yuan Ho

Science is commonly described as a collection of facts. While the scientific body of knowledge is important and should be referenced, the core of the scientific method requires creative application of these facts to test new hypotheses; in the process we discover more about the world and add to the body of knowledge. As Albert Einstein famously said, “Any fool can know. The point is to understand.” In this work, I have applied the knowledge of others in new ways to better understand the processes underlying fertility and more specifically spermatogenesis.

I first took the high throughput data on various germ and somatic cell stages, sequencing and protein interaction, which were generated by other groups. Using machine learning I was able to figure out which ones would be the most informative for differentiating between fertility and non-fertility genes. This was then used to produce lists of a few hundred genes which I predict to affect various fertility functions like spermatogenesis and oogenesis in both mice and humans. With better quality sequencing data for human germ cells and oogenesis stages, I think better predictions could be made, but the existing list should be reasonably accurate.

In order to test some of my predictions I then developed an experimental protocol to screen genes for spermatogenesis function *in vivo*. By transfecting a pool of small hairpin RNA expression cassettes into mouse testicular germ cells, I was able to affordably test up to 29 different genes simultaneously in a month. The performance of this method compares favorably to other benchmarked RNAi screens and with minimal adjustments this protocol could be used to test up to 300 genes simultaneously in one experiment. Twenty one of the top twenty six predicted spermatogenesis genes passed the experimental screen, providing confidence for more costly follow-up studies, especially for three genes with unknown molecular function (4933411K16Rik, GSG1, and SPATA4).

Finally I detail the progress I have made towards using genetic engineering to repair dysfunctional spermatogenesis in mice. From the first flawed attempts to overexpress transgenes in knockout mice to the current experiments using CRISPR to change genotypes in mouse germ cells, I have found various ways to fail in direct *in vivo* genetic engineering. However those failures have provided hints on different approaches to achieve better results which we are currently carrying out.

Most of my projects and research interests lie in taking new technological innovations and using them for novel applications. As a result, much of my work would not have been possible without the efforts of many other people. Obviously the data and experimental systems developed by other groups have been pivotal in inspiring all of the projects that I have described. Less obvious contributions have included the comments and mentorship of numerous professors, post-docs, and fellow graduate students in helping me understand the various technologies and their capabilities and limitations so that I could freely adjust them for my desired application. I hope that this work inspires other research projects which will add to the immense and yet insufficient body of scientific knowledge.

“If I have seen further, it is by standing on the shoulders of giants.” – Isaac Newton