Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Can Zhang Date

Genome-wide transcriptome analysis in Fragile X-associated Primary Ovarian Insufficiency (FXPOI) disease

By

Can Zhang

Master of Public Health

Department of Biostatistics and Bioinformatics

_____ [Chair's signature]
Hao Wu, Ph.D.
Committee Chair

_____ [Member's signature]
Tianwei Yu, Ph.D.
Committee Member

Genome-wide transcriptome analysis in Fragile X-associated Primary Ovarian Insufficiency (FXPOI) disease

By

Can Zhang

B.A., Ocean University of China, 2005
Ph.D., University of Florida, 2012

Thesis Committee Chair: Hao Wu, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in department or Biostatistics and Bioinformatics
2019

Abstract

Genome-wide transcriptome analysis in Fragile X-associated Primary Ovarian Insufficiency (FXPOI) disease

By Can Zhang

Fragile X-associated primary ovarian insufficiency (FXPOI) is characterized by reduced function of ovaries that is associated with Fragile X syndrome (FXS). Women with FXPOI often experience premature ovarian failure, infertility, and menopause before age 40, as well as a heightened risk of osteoporosis and cardiovascular disease. Given its impact on infertility and its associated health problems, FXPOI has become an emerging public health topic yet the mechanisms behind FXPOI are largely unknown. In this study, we perform a genome-wide transcriptome analysis in the FXPOI mouse model to detect genes that are dysregulated in FXPOI. Using a differential expression cutoff of FDR<0.05, we find that 195 genes are significantly down-regulated and 80 genes are significantly up-regulated in the FXPOI mouse model. By performing the Gene Ontology analysis, we discover that the down-regulated genes are significantly enriched in steroid hormone regulatory processes, whereas the up-regulated genes are involved in general signaling pathways including stress response, cell communication, etc. We believe this genome-wide study reveals a comprehensive landscape of the genetic architecture of FXPOI, which will provide an excellent opportunity to search for genes involved in the susceptibility to ovarian dysfunction, and improve the chance to develop specific therapeutic targets for FXPOI.

Genome-wide transcriptome analysis in Fragile X-associated Primary Ovarian Insufficiency (FXPOI) disease

By

Can Zhang

B.A., Ocean University of China, 2005
Ph.D., University of Florida, 2012

Thesis Committee Chair: Hao Wu, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Department or Biostatistics and Bioinformatics
2019

Table of Contents

CHAPTER 1
Background and Introduction

Introduction to FXPOI

Fragile X-associated primary ovarian insufficiency (FXPOI) is among a family of disorders caused by a premutation of the fragile X mental retardation 1 gene (*FMR1*). The *FMR1* gene is located on the X chromosome and encodes the FMR1 protein (FMRP) that is essential for normal cognitive development. Normal individuals generally possess 5-54 CGG trinucleotide repeats within the 5' untranslated region (UTR) of the *FMR1* gene. In contrast, individuals carrying premutation alleles have 55-200 CGG repeats, which could lead to the reduced production of FRMP protein (Feng et al., 1995; Sherman, 2002). Twenty percent of women who carry the premutation allele develop hypergonadotropic hypogonadism and cease menstruation prior to the age of 40 (Coffey et al., 2008; Sherman, 2000). Primary ovarian insufficiency (POI) begins with unusually early hormonal changes such as elevated levels of FSH despite normal menstrual cycles. The end stage of POI, or complete cessation of menses before the age of 40, occurs in about 1% of the general population, but up to 20% of *FMR1* pre-mutation carriers, representing a 20-fold increased risk. On average, women who carry the *FMR1* premutation alleles undergo menopause about 5 years earlier than the general population (Sherman, 2000). The effects of the premutation reach beyond reproductive implications to affect all female carriers because in addition to the early loss of fertility, these women face the potential early onset of serious conditions associated with

menopause, including a heightened risk of osteoporosis and cardiovascular disease. However, the molecular mechanism(s) underlying how the *FMR1* premutation alleles disrupt ovarian function and cause the phenotype of POI remain elusive. How the genomic landscape is altered and contributes to FXPOI remain to be determined.

FXPOI significantly impacts public health

Reproductive health is a strong predictor of overall health and wellbeing. One marker of reproductive health is the age at natural menopause. The median age of menopause is ~51±1 years, with 1% of women experiencing menopause prematurely (Palacios et al., 2010). The *FMR1* premutation is an established cause of premature ovarian failure (POF) (Sherman, 2000). POF is defined as 4 months of amenorrhea before age 40 and two follicle stimulating hormone (FSH) levels > 40 MIU/ml. In fragile X research, the term primary ovarian insufficiency (POI) is used to indicate a spectrum of reproductive outcomes that includes not limited to POF, but also occult indicators of the size of the oocyte pool (ovarian age), which may or may not manifest in diminished ovarian function among cycling women.

FXPOI significantly impacts public health. The most immediate and significant consequence of diminished ovarian function is reduced fertility (Allen et al., 2007; Streuli et al., 2009). POF occurs in ~20% of women with the *FMR1* premutation, making the rate of POF in this population ~20 times higher than the general population (De Caro et al., 2008; Sherman, 2000). Taking all women who carry the mutation, on average they go through menopause about five years earlier than those without the mutation (Murray, 2000; Sullivan et al., 2005). Consistently, the frequency of

premutation carriers among women attending reproductive endocrinology clinics for infertility is about 11% of those with familial POF and 3% of those with isolated POF (Sherman et al., 2007). The frequency is highly elevated compared with that in the general population, between 1/150- 1/250 female. Moreover, this makes FXPOI the leading known inherited cause of idiopathic primary ovarian insufficiency.

In addition, the state of early estrogen deficiency observed in FXPOI patients has significant clinical consequences such as an increased risk for low bone density, earlier onset osteoporosis and bone fractures (Gallagher, 2007), impaired endothelial function (Kalantaridou et al., 2004), earlier onset of coronary heart disease (Atsma et al., 2006), and increased cardiovascular mortality and overall mortality (Jacobsen et al., 2003).

Specific Aim and Significance

FXPOI is an understudied manifestation of the *FMR1* premutation. It leads to subfertility and early onset of disorders usually reserved for the aged population. Based on our current knowledge, we know that the three primary risk factors for FXPOI are CGG repeat length, ever smoking, and age at menopause among first-degree relatives (Spath et al., 2011). Beyond these associations, the mechanism behind FXPOI and the modifying factors that influence its onset and severity are unknown. As well as improving the chance to develop specific therapeutic targets for FXPOI, we believe that a comprehensive study of genetic architecture of FXPOI provides an excellent opportunity to define susceptibility genes of ovarian dysfunction, a clinically significant trait leading to subfertility and medical disorders due to early estrogen-deficiency. In this study, we aim to:

1. Characterize the genome-wide transcriptome in FXPOI using the newly characterized mouse model;

2. Perform the Gene Ontology (GO) term enrichment analysis to predict candidate biological processes that modulate the ovarian phenotype.

POI is common and costly. We hope this project can contribute to better understanding of the perturbed biological pathways leading to ovarian dysfunction. Also, this study is timely, as we can garner new knowledge to design a focused search for genes involved in the susceptibility to ovarian dysfunction and infertility. FXPOI has the potential to serve as a model to identify factors that modify, predict, and ameliorate the clinical burden of early diminished ovarian reserve for many women.

Approach

Animal preparation: Both control (wild-type background with no *FMR1* premutation) and FXPOI (disease background with *FMR1* premutation) mice are raised under the same environment and sacrificed at 6 months to dissect ovaries for RNA extraction. For each condition, we collect 3 biological replicates due to resource and budget limitations. We pay attention to represent every experimental condition in each batch to avoid possible confounding factors.

Isolate total RNA from ovary tissues: We will use Trizol to extract RNA from 6-month-old control and FXPOI mouse ovaries respectively. One microgram of total RNA will be used to generated RNA-seq libraries using the Illumina TruSeq RNA Sample Preparation Kit v2 following manufacturer's protocol. Briefly, Poly-A enriched mRNAs from total RNA samples will be reverse transcribed, and the cDNA of each sample will be amplified and

indexed. Qubit Fluorometric Quantitation will be used to determine the library concentration. An Agilent 2100 BioAnalyzer will be used to QC the libraries. 20 pM diluted libraries will be used for sequencing. 50-cycle pair-ended sequencing reactions will be performed using Illumina HiSeq 2000 platform. Image processing and sequence extraction will be conducted following the standard Illumina Pipeline.

Bioinformatic analysis: RNA-seq reads will be aligned to mouse mm9 genome (UCSC genomic browser) using TopHat v2.0.13 (Trapnell et al., 2009) with default parameters, and differential RPKM (Reads Per Kilobase of transcript per Million mapped reads) expression values will be generated using Cuffdiff v2.2.1 (Trapnell et al., 2012). Gene expression will be evaluated by RPKM values between experimental conditions, and genes with a FDR value <0.05 will be considered as significantly differentially expressed.

A basic overview of the main steps in this study is given in Figure 1.

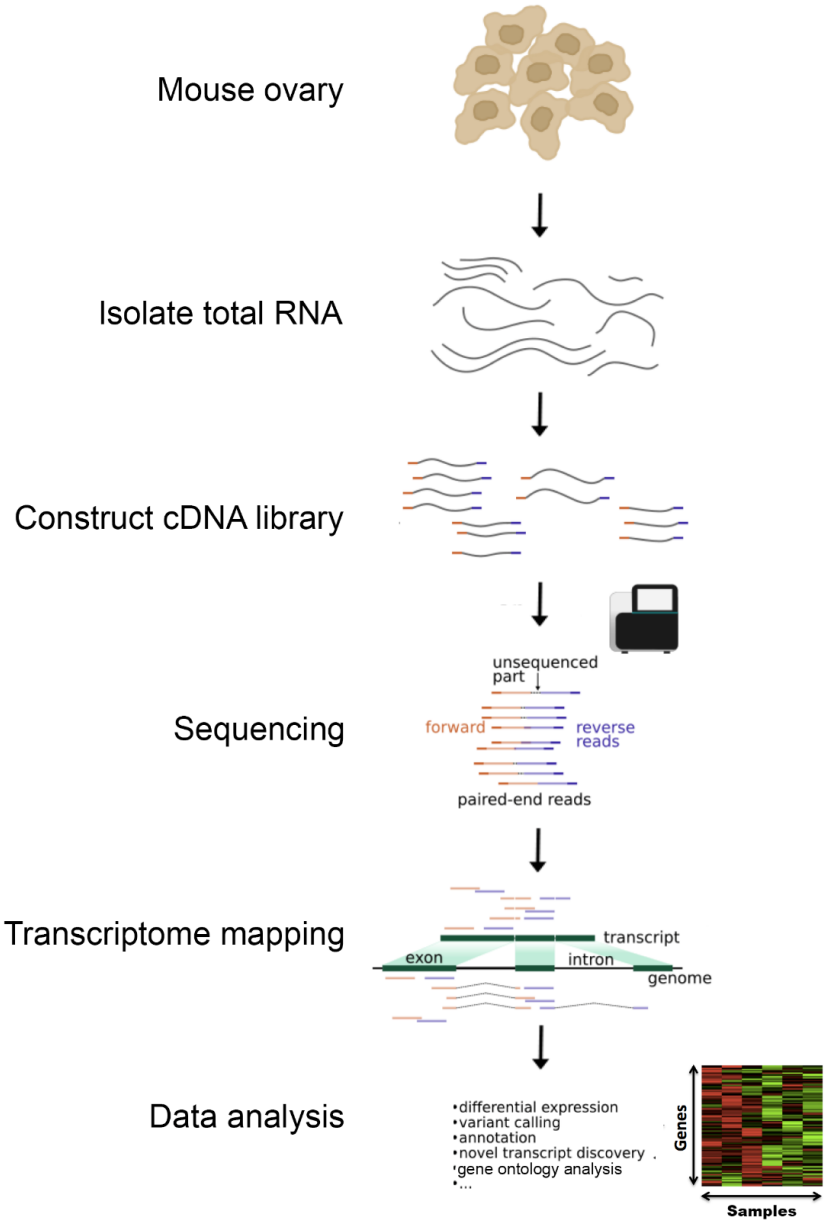

Figure 1. Overview of the main steps in current study.
The total RNA is isolated from mouse ovary. The cDNA library is generated, sequenced and the reads are aligned to mouse mm9 genome. Downstream data analysis is performed to examine differential gene expression between disease and control mice.

CHAPTER 2

Genome-wide transcriptome analysis reveals differential gene expression in FXPOI

RNA-sequencing applications

In order to better understand the molecular mechanism of the phenotypic differences between disease and normal groups, we need to identify differentially expressed genes (DEGs). The idea is to identify a set of genes with altered expressions in disease states, which may contribute to the disease phenotype. The DEGs will provide a basis to further discover the molecular pathways/mechanisms associated with disease pathogenesis, and hopefully shed light on discovering potential therapy target.

RNA-sequencing (RNA-seq) is a technique that utilizes next-generation sequencing platforms to investigate the quantify and sequence of total RNA (transcriptome) in a genome (Metzker, 2010). Comparing to the traditional microarray technology, RNA-seq is more advantageous in many aspects. First, it covers the whole genome instead of known genes, thus provides an unbiased open system to profile transcriptome. Secondly, it measures gene expression at much higher resolution and dynamic range than microarray, yet at a comparable cost (Marioni et al., 2008). RNA-seq has become the most popular option for gene expression study nowadays, and has been widely applied to basic science research, translational clinical research, as well as public health study (Oakeson et al., 2017).

RNA-sequencing differential expression analysis

Although RNA-seq experiment can serve many purposes, one of the most popular practices of RNA-seq is to identify differences in gene expression between two or more groups (for example, diseased group *vs.* normal group). This is also the first aim of the current study, i.e., characterizing the genome-wide transcriptome and find differentially expressed (either up- or down- regulated) genes between FXPOI and control mouse ovaries.

RNA-seq experiments produce enormous volumes of raw sequences containing millions to billions of short (50-150bp) cDNA fragments or "reads". In order to translate the raw data into quantitative measurements of gene expressions, we will need to analyze the raw reads with efficient and statistically principled bioinformatics algorithms. Fortunately, the biostatistics and bioinformatics community has developed many software tools to handle the raw dataset from RNA-seq. After studying published literatures (Trapnell et al., 2013; Trapnell et al., 2012) and trying differential analytic tools, we come up with the following protocol for differential expression analysis (summarized in Figure 2):

First, we will map raw reads for both conditions to the reference genome (mouse mm9 from UCSC genomic browser) using TopHat v2.0.13. In most eukaryotic genome, genes contain both exon (coding-sequence) and intron (noncoding-sequence). Only exons will retain in mRNA during transcription and contribute to protein synthesis. Since the cDNA fragments generated from RNA-seq correspond to mRNA sequence without

intron, for a read spanning an exon boundary, part of the constituent sequence will be separated by tens of thousands of nucleotides in the genome. This raises a potential mapping challenge, since the spanning reads need to be properly aligned in order to accurately count the reading depths. TopHat is a fast read-mapping tool to align RNA-seq reads to genome. It first aligns reads to a reference genome, and then goes through the results to find splicing junctions between exons (Trapnell et al., 2009).

Next, we use Cufflinks to assemble individual transcripts from the reads that have been aligned to the reference genome. The assembly would serve as a uniform basis for further calculation of gene expression levels. The transcript assembly are then sent to Cuffdiff v2.2.1 to calculate the expression level of each gene in both conditions and to test the statistical significance of any change between them. Gene expression change will be evaluated by RPKM (Reads Per Kilobase of transcript per Million mapped reads) values between experimental conditions, and genes with a FDR value <0.05 will be considered as significantly differentially expressed (Trapnell et al., 2012).

Finally, we will use R to generate plots for results display.

This analysis generated mapped reads corresponding to more than 23,000 transcripts (Figure 3). Among these transcripts, using a differential expression cutoff of FDR<0.05, we identified 195 significantly down-regulated genes and 80 significantly up-regulated genes in FXPOI mouse ovaries (Figure 4).

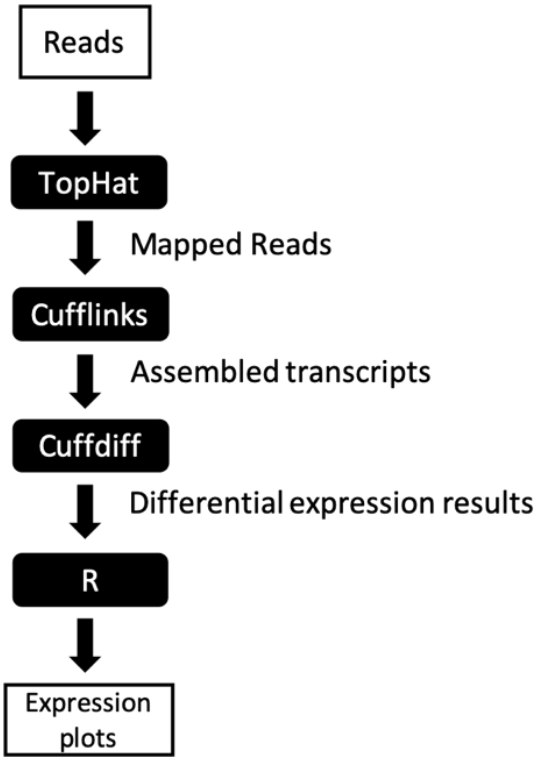

Figure 2. Overview of the RNA-seq differential expression analysis protocol. Raw reads are first mapped to the reference genome using TopHat. The mapped reads are loaded to Cufflinks to produce transcript expressions. The estimated expressions are then analyzed by Cuffdiff to find differentially expressed genes. Finally, R plots help to visualize the data.

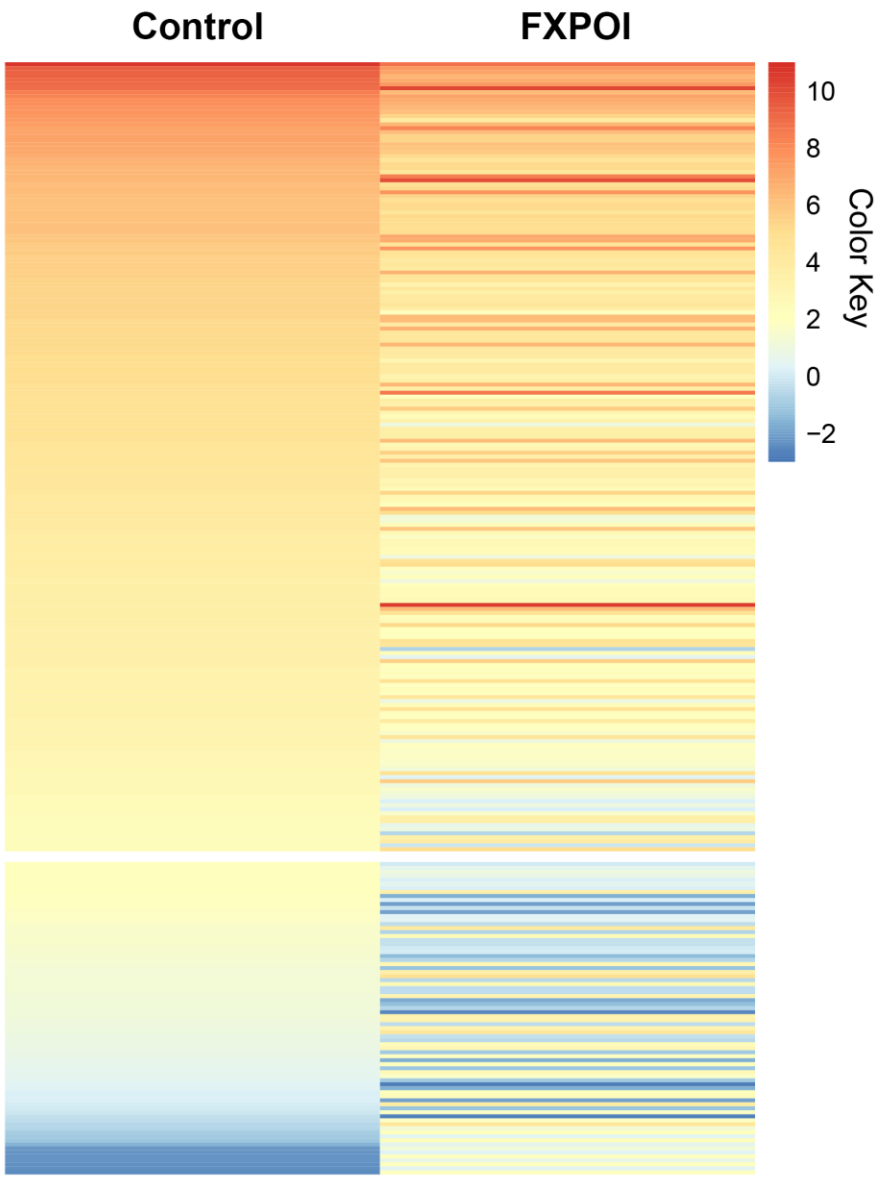

Figure 3. Heat map of the RNA-seq data.
Color-coded heap map illustrating gene expression level for a number of transcripts in control and FXPOI mouse ovaries

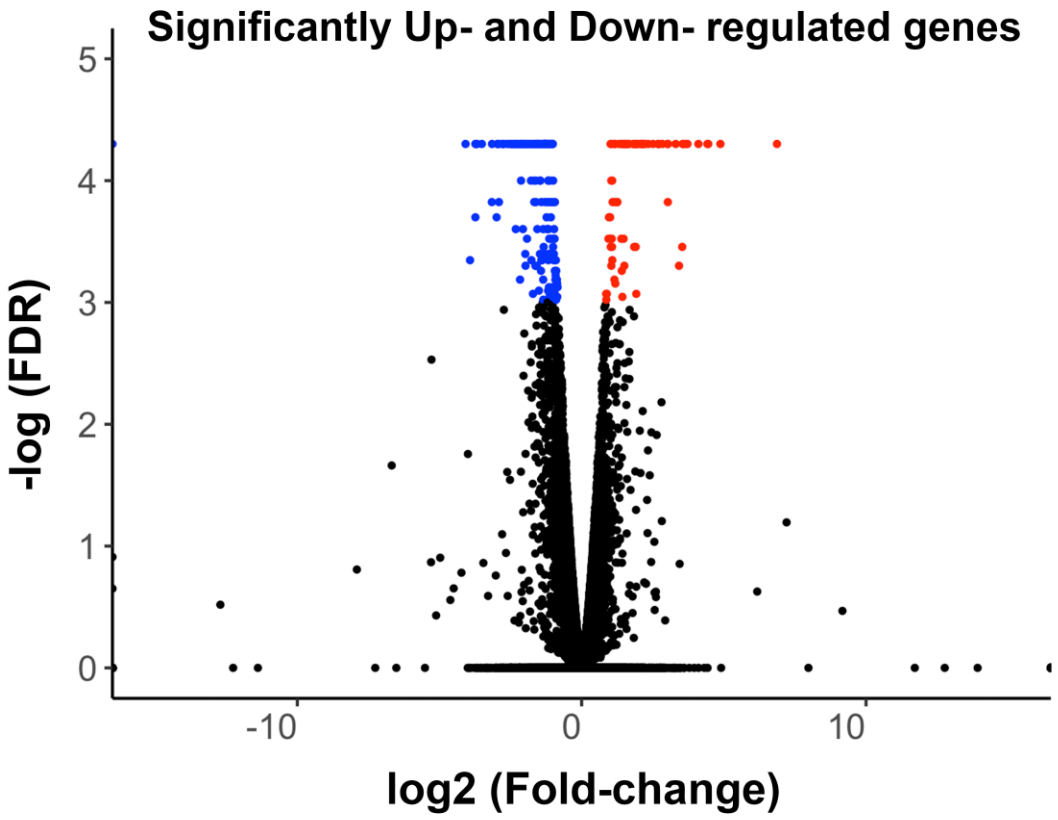

Figure 4. Volcano plot displaying differential gene expression in control and FXPOI mouse ovaries from RNA-seq (n=3).
It plots the negative log of the FDR value on the y-axis and fold change between the two conditions on the x-axis. Up-regulated genes in FXPOI mouse ovaries were highlighted in red, and down-regulated genes in blue.

CHAPTER 3

Gene Ontology analysis identifies dysregulated biological processes in FXPOI

Scientific background of Gene Ontology analysis

The differential gene expression analysis of RNA-seq data returns sets of genes that are either up- or down- regulated. Some genes may work coordinately in the same signaling pathway to conduct certain biological process. Identification of these dysregulated biological processes will help to elucidate the underlying molecular and pathological mechanisms associated with the disease condition. To have an overall understanding of what biological processes are over-represented (or under-represented) in the FXPOI mouse ovary, we performed a functional profile of the gene set through the Gene Ontology (GO) analysis.

The GO analysis provides a system for hierarchically classifying genes by their function, and testing if a GO term is statistically enriched for the given set of genes. Therefore, it is also called term overrepresentation analysis.

We consider the set of up- or down-regulated genes identified in RNA-seq as the input list or test list, and the entire set of genes mapped in RNA-seq as the reference list. Each list is classified into different biological processes (e.g., cell proliferation, stress response, etc.). The statistical test is to answer the following question: for each biological process, are genes in the input list statistically over- or under- represented when compared to the reference list? Fisher's exact test and binomial test are two

common statistical tests used for this purpose (Mi et al., 2013). The false discovery rate (FDR) is calculated and interpreted: a significant FDR value (i.e., FDR<0.05) indicates the biological process in the input list is nonrandom and potentially interesting; whereas a non-significant FDR (i.e., FDR>0.05) suggests the result is random and not worth following. Here we use a FDR value cutoff of 0.05 as a starting point (Mi et al., 2019).

Gene Ontology analysis using the PANTER classification system

There are multiple online tools for the GO analysis. In this work, we used PANTHER Classification System (www.pantherdb.org). PANTHER is a comprehensive system that combines gene classification, function, and statistical tests. We followed the most updated protocol (Mi et al., 2019) and conducted the two analyses using PANTER:

1. Functional classification viewed in gene list
2. Statistical overrepresentation test

GO analysis of the down-regulated genes showed enrichment in several key steroid hormone regulatory terms, including sterol biosynthetic process, sterol metabolic process, and steroid biosynthetic process (Figure 5). In contrast, upregulated genes were enriched in several general biological functions such as stress response, cell proliferation, and signal transduction (Figure 6).

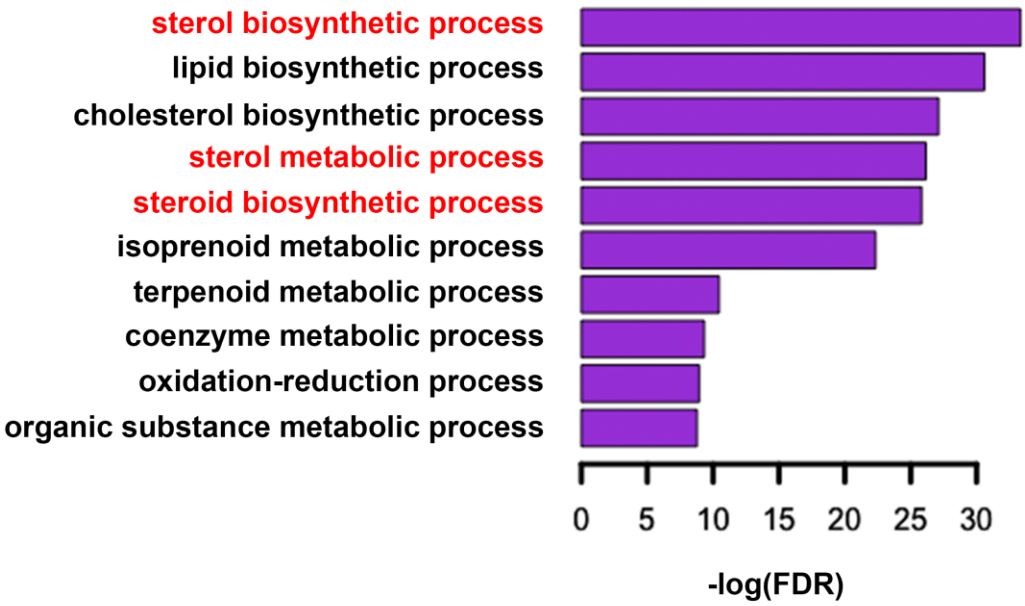

Figure 5. Gene Ontology analysis on a subset of downregulated genes.
FDR<0.05 was applied as the threshold cutoff for significantly overrepresented biological processes and top 10 processes were displayed in the bar graph. Several biological processes involved in hormone signaling were enriched and highlighted in red.

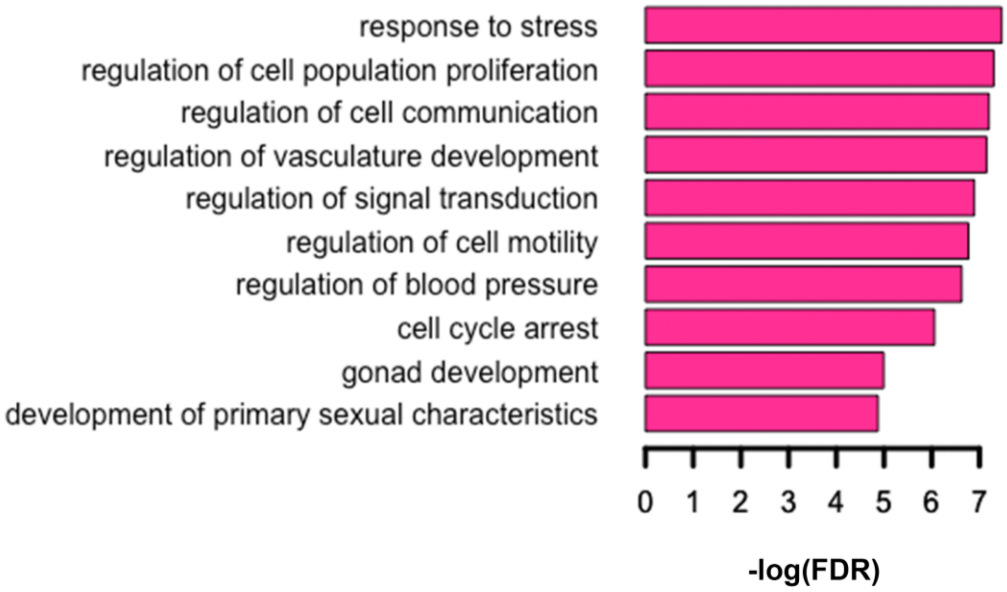

Figure 6. Gene Ontology analysis on a subset of upregulated genes.
FDR<0.05 was applied as the threshold cutoff for significantly overrepresented biological processes and top 10 processes were displayed in the bar graph. Upregulated genes were enriched in general biological processes such as stress response, cell proliferation, and cell cycle arrest.

Chapter 4

Discussion and Perspective

Studies of FXPOI disease in recent years have uncovered more information on its clinical features, diagnostic standards, genetic inheritance, and so on. Yet the underlying molecular and pathologic mechanisms are largely unclear. In this thesis study, we applied a genome-wide transcriptome analysis using the FXPOI mouse model, and discovered the set of genes and biological processes that are dysregulated in FXPOI. Here we show that several steroid hormone biological processes are down-regulated in the FXPOI mouse model. This is in consistent with the previous finding that selective serum hormones, such as FSH, LH, and 17β-estradiol, are alternated in FXPOI patients (Lu et al., 2012), suggesting these biological terms may be involved in the susceptibility to ovarian dysfunction and infertility in FXPOI patients. In fact, hormone (i.e., Estrogen) replacement therapy has been recommended and applied for women with POI by the American Society for Reproductive Medicine and the International Menopause Society.

With the practiced and established bioinformatics pipeline in this thesis work, we would like to apply the whole-genome study on FXPOI patients next. We plan to collaborate with the scientists in the National Fragile X Center an Emory to conduct the following Case-control study: We will recruit 100 women with the FXPOI prior to age 35 and 100 matched control women. We will collect personal background information such as medical history, education levels, and environmental exposures, to characterize potential risk factors. We will then perform the whole genome sequencing to identify

any genetic mutations that may suggest involvement in ovarian dysfunction. In addition, we will also perform whole genome transcriptome analysis to identify altered expression patterns and biological pathways that could contribute to ovarian dysfunction. We hope these studies could provide insight into mechanism of ovarian dysfunction associated with FXPOI.

Reference

Allen, E.G., Sullivan, A.K., Marcus, M., Small, C., Dominguez, C., Epstein, M.P., Charen, K., He, W., Taylor, K.C., and Sherman, S.L. (2007). Examination of reproductive aging milestones among women who carry the FMR1 premutation. Hum Reprod *22*, 2142-2152.

Atsma, F., Bartelink, M.L., Grobbee, D.E., and van der Schouw, Y.T. (2006). Postmenopausal status and early menopause as independent risk factors for cardiovascular disease: a meta-analysis. Menopause *13*, 265-279.

Coffey, S.M., Cook, K., Tartaglia, N., Tassone, F., Nguyen, D.V., Pan, R., Bronsky, H.E., Yuhas, J., Borodyanskaya, M., Grigsby, J.*, et al.* (2008). Expanded clinical phenotype of women with the FMR1 premutation. Am J Med Genet A *146A*, 1009-1016.

De Caro, J.J., Dominguez, C., and Sherman, S.L. (2008). Reproductive health of adolescent girls who carry the FMR1 premutation: expected phenotype based on current knowledge of fragile x-associated primary ovarian insufficiency. Ann N Y Acad Sci *1135*, 99-111.

Feng, Y., Zhang, F., Lokey, L.K., Chastain, J.L., Lakkis, L., Eberhart, D., and Warren, S.T. (1995). Translational suppression by trinucleotide repeat expansion at FMR1. Science *268*, 731-734.

Gallagher, J.C. (2007). Effect of early menopause on bone mineral density and fractures. Menopause *14*, 567-571.

Jacobsen, B.K., Heuch, I., and Kvale, G. (2003). Age at natural menopause and all-cause mortality: A 37-year follow-up of 19,731 Norwegian women. Am J Epidemiol *157*, 923-929.

Kalantaridou, S.N., Naka, K.K., Papanikolaou, E., Kazakos, N., Kravariti, M., Calis, K.A., Paraskevaidis, E.A., Sideris, D.A., Tsatsoulis, A., Chrousos, G.P.*, et al.* (2004). Impaired endothelial function in young women with premature ovarian failure: normalization with hormone therapy. J Clin Endocrinol Metab *89*, 3907-3913.

Lu, C., Lin, L., Tan, H., Wu, H., Sherman, S.L., Gao, F., Jin, P., and Chen, D. (2012). Fragile X premutation RNA is sufficient to cause primary ovarian insufficiency in mice. Hum Mol Genet *21*, 5039-5047.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res *18*, 1509-1517.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat Rev Genet *11*, 31-46.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat Protoc *8*, 1551-1566.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P.D. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc *14*, 703-721.

Murray, A. (2000). Premature ovarian failure and the FMR1 gene. Semin Reprod Med *18*, 59-66.

Oakeson, K.F., Wagner, J.M., Mendenhall, M., Rohrwasser, A., and Atkinson-Dunn, R. (2017). Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerg Infect Dis *23*, 1441-1445.

Palacios, S., Henderson, V.W., Siseles, N., Tan, D., and Villaseca, P. (2010). Age of menopause and impact of climacteric symptoms by geographical region. Climacteric *13*, 419-428.
Sherman, S. (2002). Epidemiology. In Fragile X Syndrome: Diagnosis, Treatment and Research. Baltimore, MD: The Johns Hopkins University Press.
Sherman, S.L. (2000). Premature ovarian failure among fragile X premutation carriers: parent-of-origin effect? Am J Hum Genet *67*, 11-13.
Sherman, S.L., Taylor, K., and Allen, E.G. (2007). FMR1 premutation: a leading cause of inherited ovarian dysfunction. . In: Arrieta I, Penagarikano O, Telez M (eds) Fragile Sites: New Discoveries and Changing Perspectives Nova Science Publishers.
Spath, M.A., Feuth, T.B., Smits, A.P.T., Yntema, H.G., Braat, D.D.M., Thomas, C.M.G., van Kessel, A.G., Sherman, S.L., and Allen, E.G. (2011). Predictors and risk model development for menopausal age in fragile X premutation carriers. Genet Med *13*, 643-650.
Streuli, I., Fraisse, T., Ibecheole, V., Moix, I., Morris, M.A., and de Ziegler, D. (2009). Intermediate and premutation FMR1 alleles in women with occult primary ovarian insufficiency. Fertil Steril *92*, 464-470.
Sullivan, A.K., Marcus, M., Epstein, M.P., Allen, E.G., Anido, A.E., Paquin, J.J., Yadav-Shah, M., and Sherman, S.L. (2005). Association of FMR1 repeat size with ovarian dysfunction. Hum Reprod *20*, 402-412.
Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Appendix

```
setwd("/Users/canzhang/Box Sync/RSPH/Can.Analysis.FXPOI")
library(dplyr)
library(ggplot2)

#Volcano plot with sig diff genes
RNA = read.table("gene_exp.diff", header = T, stringsAsFactors = F)
#RNA = subset(RNA, status == "OK")
#RNA = RNA[RNA$value_1 != 0 & RNA$value_2 !=0,]

#Risk = read.csv("risk.csv",header = T)
#RiskRNA = RNA[RNA$gene_id %in% Risk$Gene,]

tiff('volcano.tiff', units="in", width=5, height=4, res=300, compression = 'lzw')
ggplot() +
 geom_point(data = RNA, aes(RNA$log2.fold_change., -log10(RNA$p_value)), color =
      ifelse(RNA$significant == "yes",
          ifelse(RNA$log2.fold_change. > 0, "red", "blue"), "black"), shape = 16, size = 1, show.legend = F)
+
 #Titles the x and y axes, graph, and legend
 xlab("log2FC") + ylab("-log10(pValue)") + ggtitle("Significantly Up and Down Regulated Genes") +

 #Sets the x and y graph limitations
 xlim(-15,15) + ylim(0,5) +
 theme( plot.title = element_text(size = 12, hjust = 0.5, face = "bold"),
           axis.title = element_text(size = 15, face = "bold"), #Adjusts text properties of the axis titles
           axis.line = element_line(colour = "black"), #Adds axis lines in black color
```

```
        axis.text.x = element_text(angle = 0, hjust = 1, size = 12),
        axis.text.y = element_text(angle = 0, size = 12),
        #axis.ticks = element_blank(), #Removes axis tick marks
        #legend.text = element_text(size = 10), #Adjusts font size of the legend elements
        #legend.title = element_text(size = 12),  #Adjusts font size of the legend title
        #Modifies the position of the legend. c(0,0) represents the bottom left corner of the graph and
         #c(1,1) represents the top right corner of the graph
         #legend.position = c(0.2,0.9),
         panel.background = element_blank(), #Removes gray background
         text = element_text(family = "Arial") #Customizes text font and applies to all text
         #fonts() displays all the available fonts to use
         #Recommended font: Arial
    )
dev.off()

##heatmap with sig diff genes
setwd("/Users/canzhang/Box Sync/RSPH/Can.Analysis.FXPOI")
mat = read.table("sig.diff.gene.txt", header = T, stringsAsFactors = F)
pmat=as.matrix(mat[,2:3])
pmat=pmat[order(-pmat[,1]),]
lpmat=log2(pmat)

library(pheatmap)
library(RColorBrewer)
#customize color
color = colorRampPalette(rev(brewer.pal(n = 7, name ="RdYlBu")))(100)
#adjust color 'Yellow' to center at 0
breaks=c(seq(-3,1.9, length.out = 51)[1:50],2, seq(2.1,11, length.out=51)[2:51])
```

```
pheatmap(lpmat,
        color=color,
        breaks=breaks,
        cluster_row=F,
        cluster_cols=F, # don't cluster columns
        show_rownames=T,
        #treeheight_row=0,
        #treeheight_col=0,
        #cutree_cols=3,
        #cutree_rows=5, # cut hclust result into groups?
        gaps_row=197,
        #gaps_col=c(2,4,6,8), #create white gap
        #annotation_row=anno.rows,
        #annotation_colors=list(
           #GeneClass=c(red='red', blue='blue', green='green')),
           #GeneClass=GeneClass),
        #annotation_legend=T,
        #annotation_names_row=F
)

####bar plot GO-down

expfile="Sig.down.GO.txt"
mat=read.table(expfile,header=T,sep="\t")
lgmat=-log2(mat$FDR)

op <- par(mar = c(5,18,4,2) + 0.1)
barplot(rev(lgmat), col="darkviolet", horiz=TRUE, names.arg=rev(mat$GO), las=1, xlab="-logFDR")
axis(side=1,lwd=3)
```

```
####bar plot GO-up


expfile="Sig.up.GO.txt"
mat=read.table(expfile,header=T,sep="\t")
lgmat=-log2(mat$FDR)


op <- par(mar = c(5,20,4,2) + 0.1)
barplot(rev(lgmat), col="deeppink", horiz=TRUE, names.arg=rev(mat$GO), las=1, xlab="-logFDR")
axis(side=1,lwd=3)
```