# Characterizing Gene Expression in Human Tissues to Better Understand Sex Differences in Health and Disease

by

Kimberly Carol Olney

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2021 by the
Graduate Supervisory Committee:

Melissa Wilson, Chair
Nicholas Banovich
Kenneth Buetow
Katie Hinde

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

The regulation of gene expression, timing, location, and amount of a given project, ultimately affects the cellular structure and function. More broadly, gene regulation is the basis for cellular differentiation and development. However, gene expression is not uniform among individuals and varies greatly between genetic males and females. Males are hemizygous for the X chromosome, whereas females have two X chromosome copies. Contributing to the sex differences in gene expression between males and females are the sex chromosomes, X and Y. Gene expression differences on the autosomes and the X chromosome between males (46, XY) and females (46, XX) may help inform on the mechanisms of sex differences in human health and disease. For example, XX females are more likely to suffer from autoimmune diseases, and genetic XY males are more likely to develop cancer. Characterizing sex-specific gene expression among human tissues will help inform the molecular mechanisms driving sex differences in human health and disease. This dissertation covers a range of critical aspects in gene expression. In chapter 1, I will introduce a method to align RNA-Seq reads to a sex chromosome complement informed reference genome that considers the X and Y chromosomes' shared evolutionary history. Using this approach, I show that more genes are called as sex differentially expressed in several human adult tissues compared to a default reference alignment. In chapter 2, I characterize gene expression in an early formed tissue, the human placenta. The placenta is the DNA of the developing fetus and is typically XY male or XX female. There are well-documented sex differences in pregnancy complications, yet, surprisingly, there is no observable sex difference in expression of innate immune genes, suggesting expression of these genes is conserved. In

chapter 3, I investigate gene expression in breast cancer cell lines. Cancer arises in part due to the disruption of gene expression. Here I show 19 tumor suppressor genes become upregulated in response to a synthetic protein treatment. In chapter 4, I discuss gene and allele-specific expression in *Nasonia* jewel wasp. Chapter 4 is a replication and extension study and discusses the importance of reproducibility.

# DEDICATION

To my friends and family, who supported me in my pursuits. To my grandmother, who told me to pursue a career that I enjoyed. To my father, Dan Olney, who taught me responsibility and resilience. To my mother, Carol Olney, who taught me kindness and forgiveness.

To my partner, Jason Kennedy, who accompanied me selflessly through this intense journey, every hardship, and every success.

To Miss Claudia for being purfect and always being there when I needed a hug.

To my comrades (aka cohort) for being excellent companions. I look forward to our lifelong friendship.

To my mentors, who showed me the great value of intellectual curiosity.

For my fellow EBoard members, graduate students, undergraduates, and everyone I have interacted with during my studies. Thank you.

ACKNOWLEDGMENTS

My advisor, Dr. Melissa Wilson, deserves special recognition for her role in believing me. Melissa is dedicated to helping others pursue their goals. She is incredibly knowledgeable in many areas of research and life and is always willing to share her knowledge and wisdom; I am a better scientist and person for having been under her tutelage.

I want to acknowledge that my committee members and collaborators shared their expertise and time to enhance my research. I am forever grateful for the grant writing inspiration from Dr. Katie Hinde, the detailed explanations of gene regulatory networks from Dr. Kenneth Buetow, and the thoughtful responses on my differential expression analysis from Dr. Nicholas Banovich.

For every graduate student, post-doc, and undergraduate that took time to proof-read my drafts and listen to presentations, thank you for the support.

I want to thank the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health for funding part of my graduate studies. To my generous donors, Dr. and Mrs. Robert Spetzler, with the ARCS Foundation, thank you. Your kindness propelled my research career forward. I would not be where I am today without your support.

"*It is **not** the most intellectual of **the species that survives**; it is **not the strongest** that **survives**, but **the species that survives** is the one that is able best to adapt and adjust to the changing environment in which it finds itself.*" – 1963 Professor Leon C. Megginson summary of Darwin's Origin of Species.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

# PREFACE

Misregulation of gene expression can profoundly affect the cellular structure and function of a cell and is the basis for many diseases (Lee & Young, 2013). Complicating this is that even among healthy samples, gene expression is not uniform and can vary significantly between genetic males (46, XY) and genetic females (46, XX) (Gershoni & Pietrokovski, 2017; Kassam, Wu, Yang, Visscher, & McRae, 2019; Lopes-Ramos, Chen, et al., 2020). Furthermore, many major diseases show differences in susceptibility and risk of mortality between the sexes that may be driven by differences in gene expression (Khramtsova, Davis, & Stranger, 2019; Lopes-Ramos et al., n.d.; Lopes-Ramos, Quackenbush, & DeMeo, 2020). A major challenge in biology is determining the genetic and molecular mechanisms that underlie phenotypic differences between males and females (Khramtsova et al., 2019).

Genetic males and genetic females have very different body morphology and biochemical processes; yet, males and females share highly similar genomes, only differing on the sex chromosomes X and Y ( 2001). Sexual dimorphisms arise in part due to differences in autosomal (1-22) and X chromosome gene expression between males (46, XY) and females (46, XX) (Isensee & Ruiz Noppinger, 2007; Rinn & Snyder, 2005). Furthermore, sex differences in gene expression may contribute to the observed sex differences in human health and disease (Gershoni & Pietrokovski, 2017). For example, genetic females (46, XX) are more likely to have autoimmune diseases and adverse reactions to vaccines (Angum, Khan, Kaler, Siddiqui, & Hussain, 2020; Klein, Marriott, & Fish, 2015). Genetic males (46, XY) are more likely to develop cancers and suffer

from neurological disorders (Dorak & Karpuzoglu, 2012; May, Adesina, McGillivray, & Rinehart, 2019; Zagni, Simoni, & Colombo, 2016). Additionally, there is evidence that, in adult tissue types, genetic female samples show greater innate and adaptive immune responses than males who are hemizygous for the X chromosome (Jaillon, Berthenet, & Garlanda, 2019; Klein et al., 2015). Jailon et al. 2019 argue that it is largely the sex chromosomes that drive these observed differences in innate immune response between males and females. Klein et al. 2015 suggest that it is a combination of both differences in the expression on the X chromosome and sex hormones. There is evidence, at least in chickens, of sex differences in gene expression before developing gonads (Ayers et al., 2013), suggesting these differences arise via sex-specific and sex-biased gene regulation early in development. This dissertation helps expand our understanding of how differences in gene expression on the autosomes and the X chromosome contribute to sex differences in innate immune expression between genetic males and females. Characterizing sex differences in gene expression among human tissues focusing on the sex chromosomes and the role of immune-related genes will expand our understanding of the molecular mechanisms driving sex differences in human health and disease.

The development of techniques that quantify transcript abundance in a high-throughput way has led to advancements in studying gene expression. Ribonucleic acid sequencing (RNA-Seq) studies, in particular, have produced valuable classifications of differences in transcript levels among populations and individuals. These studies also show that inter-individual differences in gene expression are often highly heritable, making transcriptomic data a vital tool to better understanding gene expression as a measurable phenotype. Gene expression mechanisms are highly complex and tightly

regulated processes by which a gene's information is converted into structures and functions by producing a biologically functional molecule of either protein or RNA (ribonucleic acid). The phenotypes of an organism result from the expression of genes by the synthesis of proteins that control the organism's appearance and function; thus, regulation and control of the timing, location, and amount of gene expression can have profound effects on the functional phenotypes (Jobling, Hurles, & Tyler-Smith, 2013). The fundamental units of gene transcription (including polymerases) and transcriptional regulation (e.g., enhancers, promoters) are conserved across eukaryotes. Still, there are species-specific, tissue-specific, and sex-specific variations. This dissertation will investigate several questions about gene expression focusing on sex differences in gene expression in human tissues.

In the first chapter, I review some of the technical and biological challenges to quantifying gene expression differences on the X and Y chromosomes. In the following chapters, I characterize and investigate questions relating to gene expression. Specifically, the lack of sex differences in immune gene expression among uncomplicated placentas will be described in chapter 2. Chapter 3 will shift gears from sex differences to quantifying gene expression patterns in breast cancer, a sex-biased disease. Finally, chapter 4 discusses the inheritance of gene expression patterns using the *Nasonia* jewel wasp as the model organism. Each chapter has either already been published, is under review, or in preparation for publication. In this introduction, I will provide an overview of each chapter.

Chapter 1 will discuss the technical and biological challenges of studying gene expression on the X and Y chromosomes. The mammalian X and Y chromosomes were

once homologous autosomes that could undergo recombination (Charlesworth, 1991; Lahn & Page, 1999). Due to recombination suppression on the Y chromosome, they no longer recombine along the entire length except for the two pseudoautosomal regions (PARs), PAR1 and PAR2, located on the sex chromosomes, X and Y (Charlesworth, 1991; Lahn & Page, 1999; Ross et al., 2005). During sequencing mapping, the shared sequence homology between the human X and Y chromosome will routinely cause mis-mapping between these two chromosomes, reducing the accuracy of transcription estimates of sex-linked genes. To overcome this, I have helped develop a protocol for aligning XY and XX samples to a sex chromosome complement informed reference. We tested the effects of using reference genomes and reference transcriptomes informed by the sex chromosome complement of the sample's genome on the measurements of RNA-Seq abundance and sex differences in expression using various alignment tools and human tissues. Employing a sex chromosome complement approach, we detect more sex differentially expressed genes that otherwise would have been missed (Olney, Brotman, Andrews, Valverde-Vesling, & Wilson, 2020). Additionally, we found that all scenarios showed higher expression estimates on the X chromosome when sequences were mapped to a sex chromosome complement reference (Olney et al., 2020). This chapter was published in *BMC Biology of Sex Differences* in 2020 and has been cited by six published papers.

In chapter 2, I characterize sex differences in an early formed tissue, the placenta. In humans, sex differences in gene expression occur prior to the development of gonads due to genetic differences between genetics males (46, XY) and genetic females (46, XX) (Mamsen et al., 2017; Rey, Josso, & Racine, 2020). The human placenta, which

shares the genotype of the fetus (Sood, Zehnder, Druzin, & Brown, 2006), forms within the first several days of gastrulation (Turco & Moffett, 2019) and plays a critical role in healthy fetal development by regulating nutrition and protecting the developing fetus from the mother's immune system (PrabhuDas et al., 2015). Poor placentation may lead to severe pregnancy complications, including preterm birth (Liu, Li, & Zhang, 2017; Melamed, Yogev, & Glezerman, 2010), intrauterine growth restriction (Broere-Brown et al., 2020), Preeclampsia (Global Pregnancy Collaboration: et al., 2017), and subchorionic hemorrhage (Liu et al., 2017), which show a sex difference in incidence. Here I have characterized gene expression from the late first trimester and full-term placentas from male and female offspring to comprehensively understand the molecular mechanisms of expression in uncomplicated pregnancies. We show that in uncomplicated placentas, gene expression for innate immune genes is conserved. The placenta is immunologically privileged (Kanellopoulos-Langevin, Caucheteux, Verbeke, & Ojcius, 2003); we, therefore, hypothesize that misregulation of immune genes may lead to pregnancy complications. This chapter is being prepared for submission.

Chapter 3 switches focus from sex differences in gene expression to quantifying gene expression in a sex-biased disease, breast cancer (Greif, Pezzi, Klimberg, Bailey, & Zuraek, 2012; Rubin et al., 2020). Breast cancer arises due to the disruption of gene expression (Chial, 2008; Sørlie et al., 2001). This chapter examines breast cancer gene expression in response to treatment with a synthetic Polycomb-based Transcription Factor (PcTF). We hypothesized that the synthetic protein would up-regulate suppressed genes and improve the expression state of the breast cancer cells to show healthy expression levels as found in non-cancer breast cells. I determined which genes were

differentially expressed (both up-regulated and down-regulated) after the breast cell lines had been treated with PcTF compared to the untreated control cells. I intersected a list of known tumor suppressor genes (TSGene 2.0) (Zhao, Sun, & Zhao, 2013) with the genes identified as being up-regulated post-treatment across all three breast cancer cell lines. My research identified 19 tumor suppressor genes that become up-regulated in response to the synthetic protein (Olney, Nyer, Vargas, Wilson Sayres, & Haynes, 2018). I then computed a co-expression network analysis, which identified 15 transcription factors that are likely regulating the tumor suppressor genes. My research furthers our understanding of how the synthetic protein binds to transcription factors. Our results have implications for breast cancer treatment and have helped build our knowledge of gene regulation mechanisms. Our work led to a patent on the synthetic protein and a publication in the journal *BMC Systems Biology*, published in 2018 and has been cited by six publications.

Chapter 4 focuses on gene expression in Nasonia jewel wasp, a haplodiploid species. Nasonia species have a sex-specific haploid-diploid system where males and females are haploid and diploid, respectively (Beukeboom & van de Zande, 2010). In diploid cells, the paternal and maternal alleles are, on average, equally expressed. There are exceptions from this in which some genes express the maternal or paternal allele copy exclusively (Reik & Walter, 2001). This phenomenon, known as genomic imprinting, is common among eutherian mammals and some plant species (Ishida & Moore, 2013; Lawson, Cheverud, & Wolf, 2013; Moore & Haig, 1991; Reik & Walter, 2001). We processed RNA-Seq transcriptome data from highly inbred jewel wasp species *Nasonia vitripennis* and *Nasonia giraulti* and the reciprocal F1 hybrids. I developed computational scripts to scan the genomes of F1 hybrids to identify allele-specific and biased allele

expression. Our results show that expression is primarily inherited in a species of origin manner in *Nasonia*, which furthers our understanding of the inheritance of allele-specific expression. Furthermore, this is a replication and extension study using previously reported data, replicating the results, and extending these findings using different individuals and sequencing technology. Our results from both datasets demonstrate a species-of-origin effect in Nasonia F1 hybrids. This work has been submitted for publication to *PLOS Biology* and is available as a preprint on *BioRxiv*.

The final 5th chapter summarizes the previous chapters. Specifically, there is a discussion of the scientific impact and the future directions of this research. Overall, this dissertation covers a range of critical aspects in gene expression and highlights the importance of sex differences in gene expression in understanding human health.

CHAPTER 1

# Reference Genome and Transcriptome Informed by the Sex Chromosome Complement of the Sample Increase Ability to Detect Sex Differences in Gene Expression from RNA-Seq Data

(Previously published as Olney, K.C., Brotman, S.M., Andrews, J.P., Valverde-Velsing, V.A., Wilson, M.A)

Biol Sex Differ 11, 42 (2020). https://doi.org/10.1186/s13293-020-00312-9

## ABSTRACT

Human X and Y chromosomes share an evolutionary origin and, as a consequence, sequence similarity. We investigated whether sequence homology between the X and Y chromosomes affects alignment of RNA-Seq reads and estimates of differential expression. We tested the effects of using reference genomes and reference transcriptomes informed by the sex chromosome complement of the sample's genome on measurements of RNA-Seq abundance and sex differences in expression. The default genome includes the entire human reference genome (GRCh38), including the entire sequence of the X and Y chromosomes. We created two sex chromosome complement informed reference genomes. One sex chromosome complement informed reference genome was used for samples that lacked a Y chromosome; for this reference genome version, we hard-masked the entire Y chromosome. For the other sex chromosome complement informed reference genome, to be used for samples with a Y chromosome, we hard-masked only the pseudoautosomal regions of the Y chromosome, because these

regions are duplicated identically in the reference genome on the X chromosome. We analyzed transcript abundance in the whole blood, brain cortex, breast, liver, and thyroid tissues from 20 genetic female (46, XX) and 20 genetic male (46, XY) samples. Each sample was aligned twice; once to the default reference genome and then independently aligned to a reference genome informed by the sex chromosome complement of the sample, repeated using two different read aligners, HISAT and STAR. We then quantified sex differences in gene expression using featureCounts to get the raw count estimates followed by Limma/Voom for normalization and differential expression. We additionally created sex chromosome complement informed transcriptome references for use in pseudo-alignment using Salmon. Transcript abundance was quantified twice for each sample; once to the default target transcripts and then independently to target transcripts informed by the sex chromosome complement of the sample. We show that regardless of the choice of read aligner, using an alignment protocol informed by the sex chromosome complement of the sample results in higher expression estimates on the pseudoautosomal regions of the X chromosome in both genetic male and genetic female samples, as well as an increased number of unique genes being called as differentially expressed between the sexes. We additionally show that using a pseudo-alignment approach informed on the sex chromosome complement of the sample eliminates Y-linked expression in female XX samples.

**Author Summary**

The human X and Y chromosomes share an evolutionary origin and sequence homology, including regions of 100% identity; this sequence homology can result in reads misaligning between the sex chromosomes, X and Y. We hypothesized that misalignment of reads on the sex chromosomes would confound estimates of transcript abundance if the sex chromosome complement of the sample is not accounted for during the alignment step. For example, because of shared sequence similarity, X-linked reads could misalign to the Y chromosome. This is expected to result in reduced expression for regions between X and Y that share high levels of homology. For this reason, we tested the effect of using a default reference genome versus a reference genome informed by the sex chromosome complement of the sample on estimates of transcript abundance in human RNA-Seq samples from whole blood, brain cortex, breast, liver, and thyroid tissues of 20 genetic female (46, XX) and 20 genetic male (46, XY) samples. We found that using a reference genome with the sex chromosome complement of the sample resulted in higher measurements of X-linked gene transcription for both male and female samples and more differentially expressed genes on the X and Y chromosomes. We additionally investigated the use of a sex chromosome complement informed transcriptome reference index for alignment free quantification protocols. We observed no Y-linked expression in female XX samples only when the transcript quantification was performed using a transcriptome reference index informed on the sex chromosome complement of the sample. We recommend that future studies requiring aligning RNA-Seq reads to a reference genome or pseudo-alignment with a transcriptome reference

should consider the sex chromosome complement of their samples prior to running default pipelines.

## Background

Sex differences in aspects of human biology, such as development, physiology, metabolism, and disease susceptibility are partially driven by sex specific gene regulation (Arnold et al., 2012; Khramtsova et al., 2018; Raznahan et al., 2018; Traglia et al., 2017). There are reported sex differences in gene expression across human tissues (Gershoni & Pietrokovski, 2017; Goldstein et al., 2014; Shi et al., 2016) and while some may be attributed to hormones and environment, there are documented genome-wide sex differences in expression based solely on the sex chromosome complement (Arnold & Chen, 2009). However, accounting for the sex chromosome complement of the sample in quantifying gene expression has been limited due to shared sequence homology between the sex chromosomes, X and Y, that can confound gene expression estimates.

The X and Y chromosomes share an evolutionary origin: mammalian X and Y chromosomes originated from a pair of indistinguishable autosomes ~180-210 million years ago that acquired the sex-determining genes (Charlesworth, 1991; Lahn & Page, 1999; Ross et al., 2005). The human X and Y chromosomes formed in two different segments: a) one that is shared across all mammals called the X-conserved region (XCR) and b) the X-added region (XAR) that is shared across all eutherian animals (Ross et al., 2005). The sex chromosomes, X and Y, previously recombined along their entire lengths, but due to recombination suppression from Y chromosome-specific inversions (Lahn & Page, 1999; Pandey et al., 2013), now only recombine at the tips in the pseudoautosomal regions (PAR) PAR1 and PAR2 (Charlesworth, 1991; Lahn & Page, 1999; Ross et al., 2005). PAR1 is ~2.78 million bases (Mb) and PAR2 is ~0.33 Mb; these sequences are 100% identical between X and Y (Aken et al., 2017; Charchar et al., 2003; Ross et al.,

2005) (Figure 1A). The PAR1 is a remnant of the XAR Ross et al. 2005) and shared among eutherians, while the PAR2 is recently added and human-specific (Charchar et al., 2003). Other regions of high sequence similarity between X and Y include the X-transposed-region (XTR) with 98.78% homology (Veerappa et al., 2013) (Figure 1A). The XTR formed from an X chromosome to Y chromosome duplication event following the human-chimpanzee divergence (Ross et al., 2005; Skaletsky et al., 2003). Thus, the evolution of the X and Y chromosomes has resulted in a pair of chromosomes that are diverged, but still share some regions of high sequence similarity.

To infer which genes or transcripts are expressed, RNA-Seq reads can be aligned to a reference genome. The abundance of reads mapped to a transcript is reflective of the amount of expression of that transcript. RNA-Seq methods rely on aligning reads to an available high quality reference genome sequence, but this remains a challenge due to the intrinsic complexity in the transcriptome of regions with a high level of homology (Piskol et al., 2013). By default, the GRCh38 version of the human reference genome includes both the X and Y chromosomes, which is used to align RNA-Seq reads from both male XY and female XX samples. It is known that sequence reads from DNA will misalign along the sex chromosomes affecting downstream analyses (Webster et al., 2019). However, this has not been tested using RNA-Seq data and the effects on differential expression analysis are not known. Considering the increasing number of human RNA-Seq consortium datasets (e.g., the Genotype-Tissue Expression project (GTEx) (GTEx Consortium, 2015), The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013), Geuvadis project (Lappalainen et al., 2013), and Simons Genome Diversity Project (Mallick et al., 2016)), there is an urgent need to understand how

aligning to a default reference genome that includes both X and Y may affect estimates of gene expression on the sex chromosomes (Khramtsova et al., 2018; Tukiainen et al., 2016). We hypothesize that regions of high sequence similarity will result in misaligning of RNA-Seq reads and reduced expression estimates (Figure 1A & B).

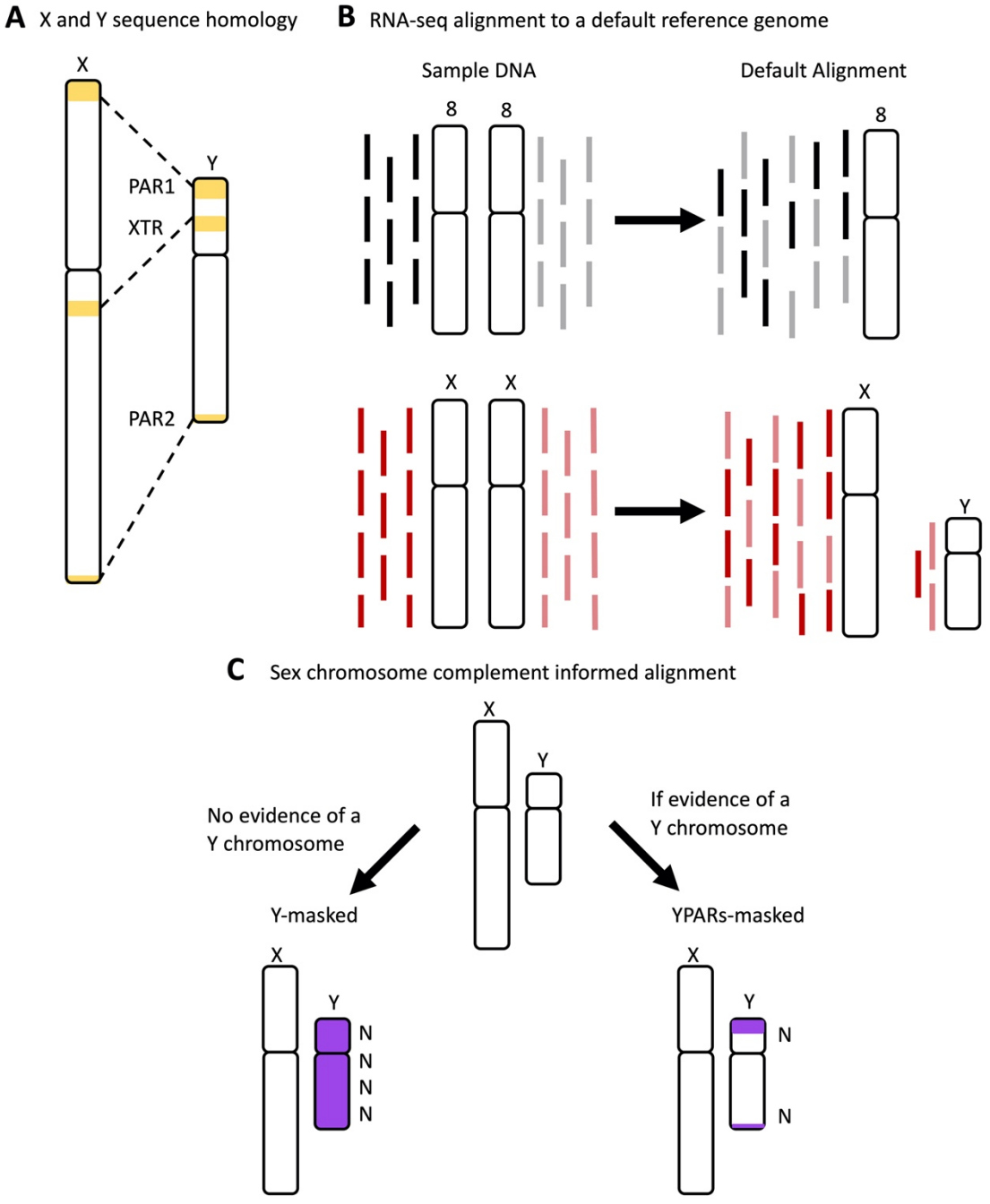
Here, we tested the effect of sex chromosome complement informed read alignment on the quantified levels of gene expression and the ability to detect sex-biased gene expression. We utilized data from the GTEx project, focusing on five tissues, whole blood, brain cortex, breast, liver, and thyroid, which are known to exhibit sex differences in gene expression (Gershoni and Pietrokovski 2017; R. Li and Singh 2014; de Perrot et al. 2000; Melé et al. 2015; Mayne et al. 2016). Many genes have been reported to be differentially expressed between male and female brain samples (Gershoni & Pietrokovski, 2017; Goldstein et al., 2014; Shi et al., 2016) and differential expression in blood samples between males and females has also been documented (Gershoni & Pietrokovski, 2017; Goldstein et al., 2014). An analysis of all GTEx tissue samples reported that breast mammary gland tissues are the most sex differentially expressed tissue (Gershoni & Pietrokovski, 2017). It has also been reported that there are sex disparities in thyroid cancer (Rahbari et al., 2010) and liver cancer (Natri et al., 2019; Naugler et al., 2007) suggesting possible sex differences in gene expression. We used whole blood, brain cortex, breast, liver, and thyroid tissues from 20 genetic male (46, XY) and 20 genetic female (46, XX) individuals for a total of 200 samples evenly distributed among tissues. Male and female samples, for each tissue, were age-matched between the sexes and only included samples of age 55 to 70. We aligned all samples to a default reference genome that includes both the X and Y chromosomes and to a reference

genome that is informed on the sex chromosome complement of the genome: Male XY samples were aligned to a reference genome that includes both the X and Y chromosome, where the Y chromosome PAR1 and PAR2 are hard-masked with Ns (Figure 1C) so that reads will align uniquely to the X PAR sequences. Conversely, female XX samples were aligned to a reference genome where the entirety of the Y chromosome is hard-masked (Figure 1C). We tested two different read aligners, HISAT (Kim et al., 2015) and STAR (Dobin et al., 2013), to account for variation between alignment methods and measured differential expression using Limma/Voom (Law et al., 2014). We found that using a sex chromosome complement informed reference genome for aligning RNA-Seq reads increased expression estimates on the pseudoautosomal regions of the X chromosome in both male XY and female XX samples and uniquely identified differentially expressed genes.

We additionally investigated the effect of transcriptome references on pseudo-alignment methods. We quantified abundance using Salmon (Patro et al., 2017) in male and female brain cortex samples twice, once to a default reference transcriptome index that includes both the X and Y chromosome linked transcripts and to a reference transcriptome index that is informed on the sex chromosome complement of the sample. We found that using a sex chromosome complement informed reference transcriptome index for RNA-Seq pseudo-alignment quantification eliminated Y-linked expression estimates in female XX samples, that were observed in the default approach.

Regardless of alignment or pseudo-alignment approach, we recommended carefully considering the annotations of the sex chromosomes in the references used, as

theses will affect quantifications and differential expression estimates, especially of sex chromosome linked genes.



*Chapter 1. Figure 1. Homology Between the Human X and Y Chromosomes Where Misaligning Could Occur.* (A) High sequence homology exists between the human X and Y chromosomes in three regions: 100% sequence identity for the pseudoautosomal regions (PARs), PAR1, and PAR2, and ~ 99% sequence homology in the X-transposed region (XTR). The X chromosome PAR1 is ~ 2.78 million bases (Mb) extending from X:10,001 to 2,781,479 and the X chromosome PAR2 is ~ 0.33 Mb extending from X:155,701,383 to 156,030,895. The X chromosome PAR1 and PAR2 are identical in sequence to the Y chromosome PAR1 Y:10,001 to 2,781,479 and PAR2 Y:56,887,903 to 57,217,415. (B) Using a standard alignment approach will result in reads misaligning between regions of high sequence homology on the sex chromosomes. (C) Using a reference genome that is informed by the genetic sex of the sample may help to reduce misaligning between the X and Y chromosomes. In humans, samples without evidence of a Y chromosome should be aligned to a Y-masked reference genome, and samples with evidence of a Y should be aligned to a YPAR-masked reference genome.

## Methods

***Building Sex Chromosome Complement Informed Reference Genomes.*** All GRCh38.p10 unmasked genomic DNA sequences, including autosomes 1-22, X, Y, mitochondrial DNA (mtDNA), and contigs were downloaded from ensembl.org release 92 (Aken et al., 2017). The default reference genome here includes all 22 autosomes, mtDNA, the X chromosome, the Y chromosome, and contigs. For the two sex chromosome complement informed reference assemblies, we included all 22 autosomes, mtDNA, and contigs from the default reference and a) one with the Y chromosome either hard-masked for the "Y-masked reference genome" or b) one with the pseudoautosomal regions, PAR1 and PAR2, hard-masked on the Y chromosome for "YPARs-masked reference genome" (Figure 1C). Hard-masking with Ns will force reads to not align to those masked regions in the genome. Masking the entire Y chromosome for the sex chromosome complement informed reference genome, Y-masked, was accomplished by changing all the Y chromosome nucleotides [ATGC] to N using a sed command in linux. YPARs-masked was created by hard-masking the Y PAR1: 6001-2699520 and the Y

PAR2: 154931044-155260560 regions. The GRCh38.p10 Y PAR1 and Y PAR2 chromosome start and end location was defined using Ensembl GRCh38 Y PAR definitions (Aken et al., 2017). After creating the Y chromosome PAR1 and PAR2 masked fasta files, we concatenated all the Y chromosome regions together to create a YPARs-masked reference genome. After creating the GRCh38.p10 default reference genome and the two sex chromosome complement informed reference genomes, we indexed the reference genomes and created a dictionary for each using HISAT version 2.1.0 (Kim et al., 2015) hisat2-build -f option and STAR version 2.5.2 (Dobin et al., 2013), using option --genomeDir and --sjdbGTFfile. Reference genome indexing was followed by picard tools version 1.119 CreateSequenceDictionary (*Broadinstitute/Picard*, 2014/2020), which created a dictionary for each reference genome (Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq).

***Building Sex Chromosome Complement Informed Transcriptome Index.***

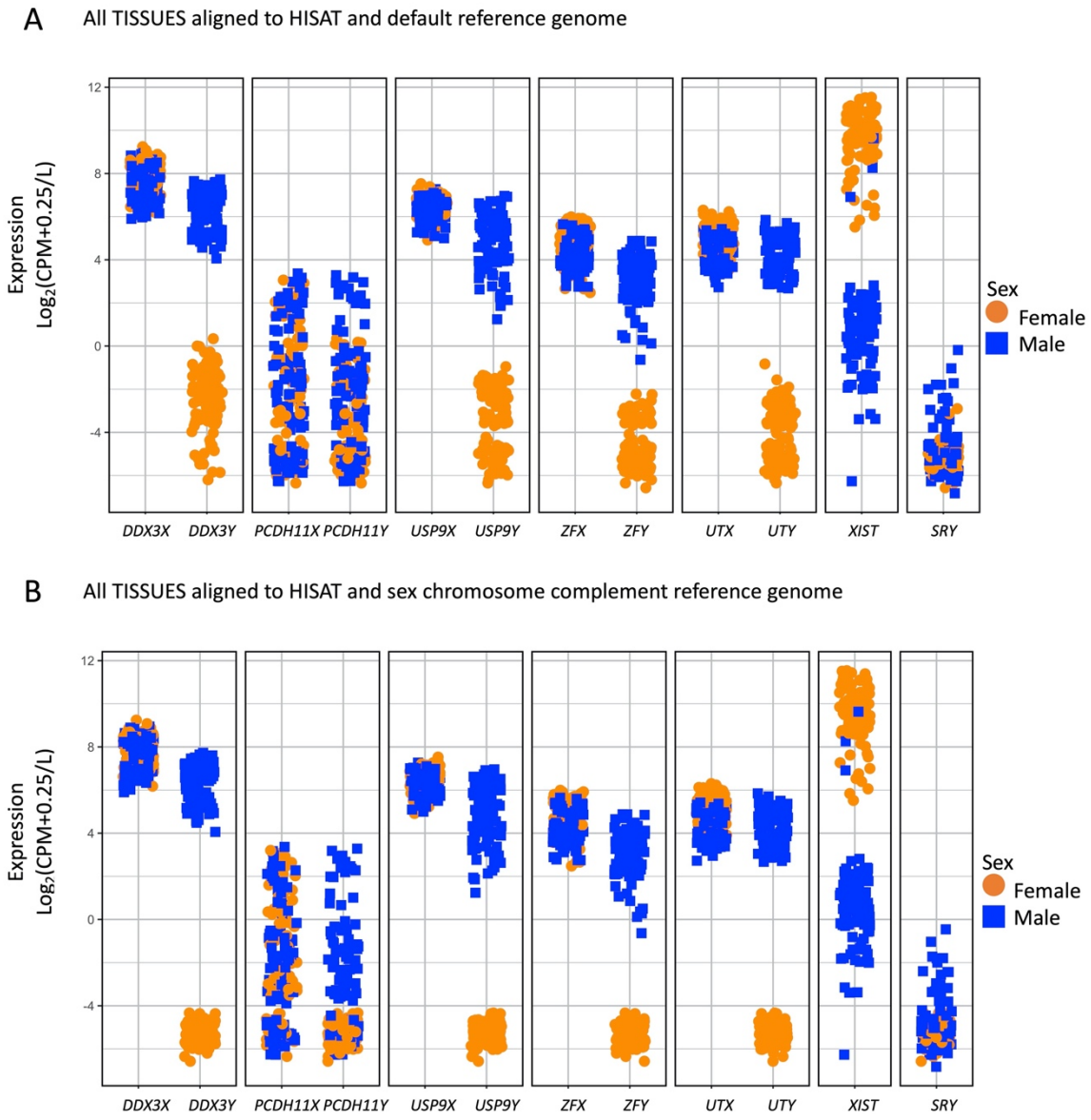
Ensembl's GRCh38.p10 cDNA reference transcriptome fasta consists of transcript sequences resulting from Ensemble gene predictions. Ensembl's cDNA was downloaded from ensembl.org release 92 (Aken et al., 2017). The default transcriptome reference includes 199,234 transcripts which includes autosomal, mtDNA, X chromosome, Y chromosome and contig transcripts. The default Ensembl cDNA does not contain Y chromosome PAR linked transcript sequences, it only contains the X chromosome PAR sequence transcripts. For the sex chromosome complement informed reference transcriptome index, we included all 22 autosomes, mtDNA, X, and contigs from the default cDNA transcriptome and we hard-masked all available Y chromosome linked transcript sequences. Hard-masking the Y chromosome linked transcripts was

accomplished by changing all the Y chromosome nucleotides [ATGC] to N using a sed command in linux. After downloading the GRCh38.p10 default reference transcriptome and creating the Y-masked sex chromosome complement informed reference transcriptome fasta files, we then generated a decoy-aware transcriptome for each transcriptome reference. For generating the default decoy-aware reference transcriptome, we used the default genome as the decoy sequence. This was accomplished by concatenating the default genome fasta to the end of the default transcriptome fasta to populate the decoy file with the chromosome names, as suggested by Salmon (Patro et al., 2017). The default transcriptome fasta and the default decoy file were then used to create the mapping-based index using the Salmon version 1.2.0 index function (Patro et al., 2017). The Y-masked decoy-aware transcriptome fasta was generated by concatenating the Y-masked genome fasta to the end of the Y-masked transcriptome fasta to populate the decoy file with the chromosome names. The Y-masked transcriptome fasta and the decoy file were then used as inputs for generating the Y-masked mapping-based index using the salmon index function. For both the default and the Y-masked mapping-based index, a k-mer of 31 was used as this was suggested to work well for reads of 75bp.

In addition to the Ensembl reference, we investigated the effects of a sex chromosome complement reference transcriptome index using the gencode transcript reference fasta GRCh38.p12 that contains 206,694 transcripts which includes autosomal, mtDNA, X, Y and contigs. The gencode transcriptome reference includes both the X and Y PAR transcripts (J et al., 2012). Following the same parameters for the Ensembl decoy-aware transcriptome, we created two gencode sex chromosome complement decoy-aware

transcriptome references, in addition to a default gencode decoy-aware transcriptome reference. The pipeline is available on GitHub, https://github.com/SexChrLab/XY_RNAseq.

***RNA-Seq Samples.*** From the Genotyping-Tissue Expression (GTEx) Project data, we downloaded SRA files for whole blood, brain cortex, breast, liver, and thyroid tissues from 20 separate genetic female (46, XX) and 20 separate genetic male (46, XY) individuals (Consortium, 2015; GTEx Consortium, 2015) that were age matched between the sexes and ranged from age 55 to 70 (Additional file 1 & 2). Age matching exactly was accomplished using the matchit function in the R package MatchIt (Ho et al. 2011). The GTEx data is described and available through dbGaP under accession phs000424.v6.p1; we received approval to access this data under dbGaP accession #8834. GTEx RNA-Seq samples were sequenced to 76bp reads and the median coverage was ~82 million (M) reads (Consortium, 2015). Although information about the genetic sex of the samples was provided in the GTEx summary downloads, it was additionally investigated by examining the gene expression of select genes that are known to be differentially expressed between the sexes or are known X-Y homologous genes: *DDX3X*, *DDX3Y*, *PCDH11X*, *PCDH11Y*, *USP9X*, *USP9Y*, *ZFX*, *ZFY*, *UTX*, *UTY*, *XIST*, and *SRY* (Figure 2; Additional file 3 & 4).




*Chapter 1. Figure 2. Genetic Sex of RNA-Seq Samples.* We investigated the gene expression, log2(CPM + 0.25/L), of XY homologous genes (DDX3X/Y, PCDH11X/Y, USP9X/Y, ZFX/Y, UTX/Y); XIST; and SRY in all samples from all tissues analyzed here from genetic males (blue squares) and genetic females (orange circles) a) when aligned to a default reference genome and b) when aligned to a sex chromosome complement informed reference genome, using HISAT as the read aligner.

***RNA-Seq Trimming and Quality Filtering.*** RNA-Seq sample data was converted from sequence read archive (sra) format to the paired-end FASTQ format using the SRA

toolkit (Leinonen et al., 2011). Quality of the samples' raw sequencing reads was examined using FastQC (Andrews, n.d.) and MultiQC . Subsequently, adapter sequences were removed using Trimmomatic version 0.36 (Bolger et al., 2014). More specifically, reads were trimmed to remove bases with a quality score less than 10 for the leading strand and less than 25 for the trailing strand, applying a sliding window of 4 with a mean PHRED quality of 30 required in the window and a minimum read length of 40 bases.

***RNA-Seq Read Alignment.*** Following trimming, paired RNA-Seq reads from all samples were aligned to the default reference genome. Unpaired RNA-Seq reads were not used for alignment. Reads from the female (46, XX) samples were aligned to the Y-masked reference genome and reads from male (46, XY) individuals were aligned to the YPARs-masked reference genome. Read alignment was performed using HISAT version 2.1.0 (Kim et al., 2015), keeping all parameters the same, only changing the reference genome used, as described above. Read alignment was additionally performed using STAR version 2.5.2 (Dobin et al., 2013), where all samples were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement, keeping all parameters the same (Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq).

***Processing of RNA-Seq Alignment Files.*** Aligned RNA-Seq samples from HISAT and STAR were output in Sequence Alignment Map (SAM) format and converted to Binary Alignment Map (BAM) format using bamtools version 2.4.0 (Li et al., 2009). Summaries on the BAM files including the number of reads mapped were computed using bamtools version 2.4.0 package (Barnett et al., 2011). RNA-Seq BAM files were indexed, sorted, duplicates were marked, and read groups added using

bamtools, samtools, and Picard (Barnett et al., 2011; *Broadinstitute/Picard*, 2014/2020; Li et al., 2009). All RNA-Seq BAM files were indexed using the default reference genome using Picard ReorderSam (*Broadinstitute/Picard*, 2014/2020), this was done so that all samples would include all chromosomes in the index files. Aligning XX samples to a Y-masked reference genome using HISAT indexes would result in no Y chromosome information in the aligned BAM and BAM index bai files. For downstream analysis, some tools require that all samples have the same chromosomes, which is why we hard-masked rather than removed. Reindexing the BAM files to the default reference genome does not alter the read alignment, and thus does not alter our comparison between default and sex chromosome complement informed alignment.

***Gene Expression Level Quantification.*** Read counts for each gene across all autosomes, sex chromosomes, mtDNA, and contigs were generated using featureCounts version 1.5.2 (Liao et al., 2014) for all aligned and processed RNA-Seq BAM files. Female XX samples when aligned to a sex chromosome complement informed reference genome will show zero counts for Y-linked genes, but will still include those genes in the raw counts file. This is an essential step for downstream differential expression analysis between males and females to keep the total genes the same between the sexes for comparison. Only rows that matched gene feature type in Ensembl Homo_sapiens.GRCh38.89.gtf gene annotation (Aken et al., 2017) were included for read counting. There are 2,283 genes annotated on the X chromosome and a total of 56,571 genes across the entire genome for GRCh38 version of the human reference genome (Aken et al., 2017). Only primary alignments were counted and specified using the --primary option in featureCounts.

***RNA-seq Quantification for Transcriptome Index.*** Transcript quantification for trimmed paired RNA-seq brain cortex samples were estimated twice, once to a default decoy-aware reference transcriptome index and once to a sex chromosome complement informed decoy-aware reference transcriptome index using Salmon with the –validateMappings flag. Salmon's –validateMappings adopts a scheme for finding protentional mapping loci of a read using a chain algorithm introduced in minimap2 (Li, 2018). Transcript quantification for female (46, XX) samples was estimated using a Y-masked reference transcriptome index and male (46, XY) transcript quantification was estimated using a Y PAR masked reference transcriptome index when the Y PAR sequence information was available for the transcriptome build. This was repeated for both the Ensembl and the gencode cDNA transcriptome builds, keeping all parameters the same, only changing the reference transcriptome index used, as described above.

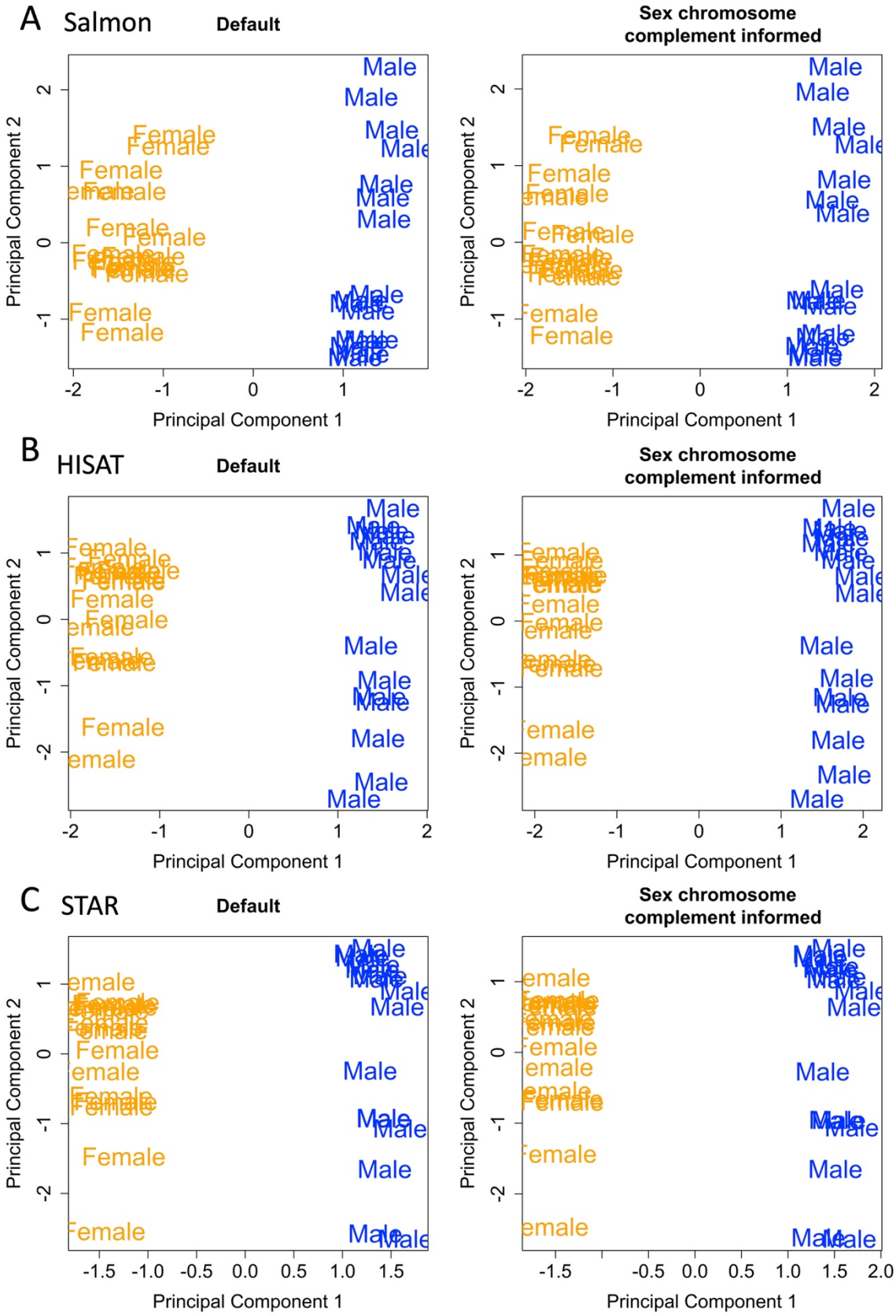
***DGEList Object.*** Differential expression analysis was performed using the limma/voom pipeline (Law et al., 2014) which has been shown to be a robust differential expression software package (Costa-Silva et al., 2017; Seyednasrollah et al., 2015) for both reference-based and pseudo-alignment quantification. Quantified read counts from each sample for the reference-based quantification were generated from featureCounts were combined into a count matrix, each row representing a unique gene ID and each column representing the gene counts for each unique sample. This was repeated for each tissue type and read into R using the DGEList function in the R limma package (Love et al., 2014). A sample-level information file related to the genetic sex of the sample, male or female, and the reference genome used for alignment, default or sex chromosome

complement informed, was created and corresponds to the columns of the count matrix described above.

Pseudo-aligned transcript read counts from each brain cortex sample quantified using Salmon were combined into a count matrix using tximport (Soneson et al., 2015) with each row representing a unique transcript ID and each column representing the transcript counts for each unique sample. To create length scaled transcripts per million (TPM) values to pass into limma, tximport function lengthScaledTPM was employed (Soneson et al., 2015). The reference assembly annotation file was read into R using tximport function makeTxDbFromGFF. Following this, a  key of the transcript ID corresponding to the gene ID was created was created using the keys function (Soneson et al., 2015). Gene level TPM values were then generated using the tx2gene function. This was repeated for the Ensembl and the gencode default and sex chromosome complement informed transcriptome quantification estimates.

***Multidimensional Scaling.*** Multidimensional Scaling (MDS) was performed using the DGEList-object containing gene expression count information for each sample. MDS plots were generated using the plotMDS function in in the R limma package (Law et al., 2014). The distance between each pair of samples is shown as the $\log_2$ fold change between the samples. The analysis was done for each tissue separately using all shared common variable genes for dimensions (dim) 1 & 2 and dim 2 & 3. Samples that did not cluster with reported sex or clustered in unexpected ways in either dim1, 2 or 3 were removed from all downstream analysis (Additional file 5). MDS plots for each tissue containing the samples that were used for quality control are located in Additional file 6. Briefly, one male XY whole blood did not cluster with any of the other samples and was

removed. One female XX breast sample clustered with the opposite sex and was thus removed. In brain cortex, three male XY brain cortex samples didn't cluster neatly with the other male XY samples in dim 1 & 2 were thus removed. Another male brain cortex sample, although clustered with other male samples, had the lowest number of sequencing remaining after trimming for quality, 23.9M, and thus was also removed. To keep the number of samples in each sex roughly equal, four female XX brain cortex samples were randomly selected for removal. For liver and thyroid tissue, no samples appeared to cluster in any unexpected ways and thus no liver or thyroid tissue samples were removed. For all aligners the first component of variation in the MDS plot is explained by the sex of the sample (Figure 3).


A
Salmon
Default
Sex chromosome complement informed
B
HISAT
Default
Sex chromosome complement informed
C
STAR
Default
Sex chromosome complement informed
Principal Component 1
Principal Component 2
Female
Male

*Chapter 1. Figure 3. Multidimensional Scaling for the Top 100 Most Variable Genes*. We investigated multidimensional scaling for the top 100 common variable genes in the brain cortex samples. (A) Salmon pseudo-alignment with Ensembl transcriptome reference, (B) HISAT read aligner, and (C) STAR read aligner when quantifying using both the default and the sex chromosome complement informed references. Most variation in the data is explained by the sex of the sample.

***Differential Expression.*** Using edgeR (Robinson et al., 2010), raw counts were normalized to adjust for compositional differences between the RNA-Seq libraries using the voom normalize quantile function, which normalizes the reads by the method of trimmed mean of values (TMM) (Law et al., 2014). Counts were then transformed to $\log_2$(CPM+0.25/L), where CPM is counts per million, L is library size, and 0.25 is a prior count to avoid taking the log of zero (Robinson et al., 2010). For this dataset, the average library size is about 79.76 million, therefore L is 79.76. Thus, the minimum $\log_2$(CPM+0.25/L) value for each sample, representing zero transcripts, is $\log_2(0+0.25/15) = -8.32$.

A mean minimum of 1 CPM, or the equivalent of 0 in $\log_2$(CPM+2/L), in at least one sex per tissue comparison was required for the gene to be kept for downstream analysis. A CPM value of 1 was used in our analysis to separate expressed genes from unexpressed genes, meaning that in a library size of ~79.76 million reads, there are at least an average of 79 counts in at least one sex. After filtering for a minimum CPM, 53,804 out of the 56,571 quantified genes were retained for the whole blood samples, 53,822 for brain cortex, 54,184 for breast, 53,830 for liver, and 53,848 for thyroid. A linear model was fitted to the DGEList-object, which contains the filtered and normalized gene counts for each sample, using the limma lmfit function which will fit a separate model to the expression values for each gene (Law et al., 2014).

For differential expression analysis a design matrix containing the genetic sex of the sample (male or female) and which reference genome the sample was aligned to (default or sex chromosome complement informed) was created for each tissue type for contrasts of pairwise comparisons between the sexes. Pairwise contrasts were generated using limma makecontrasts function (Law et al., 2014). We identified genes that exhibited significant expression differences defined using an Benjamini-Hochberg adjusted p-value cutoff that is less than 0.01 (1%) to account for multiple testing in pairwise comparisons between conditions using limma/voom decideTests vebayesfit (Law et al., 2014). A conservative adjusted p-value cutoff of less than 0.01 was chosen to be highly confident in the genes that were called as differentially expressed when comparing between reference genomes used for alignment. Pipeline available on GitHub, https://github.com/SexChrLab/XY_RNAseq.

***GO Analysis.*** We examined differences and similarities in gene enrichment terms between the differentially expressed genes obtained from the differential expression analyses of the samples aligned to the default and sex chromosome complement informed reference genomes, to investigate if the biological interpretation would change depending on the reference genome the samples were aligned to. We investigated gene ontology enrichment for lists of genes that were identified as showing overexpression in one sex versus the other sex for whole blood, brain cortex, breast, liver, and thyroid samples (adjusted p-value < 0.01). We used the GOrilla webtool, which utilizes a hypergeometric distribution to identify enriched GO terms (Eden et al., 2009). A modified Fisher exact p-value cutoff < 0.001 was used to select significantly enriched terms (Eden et al., 2009).

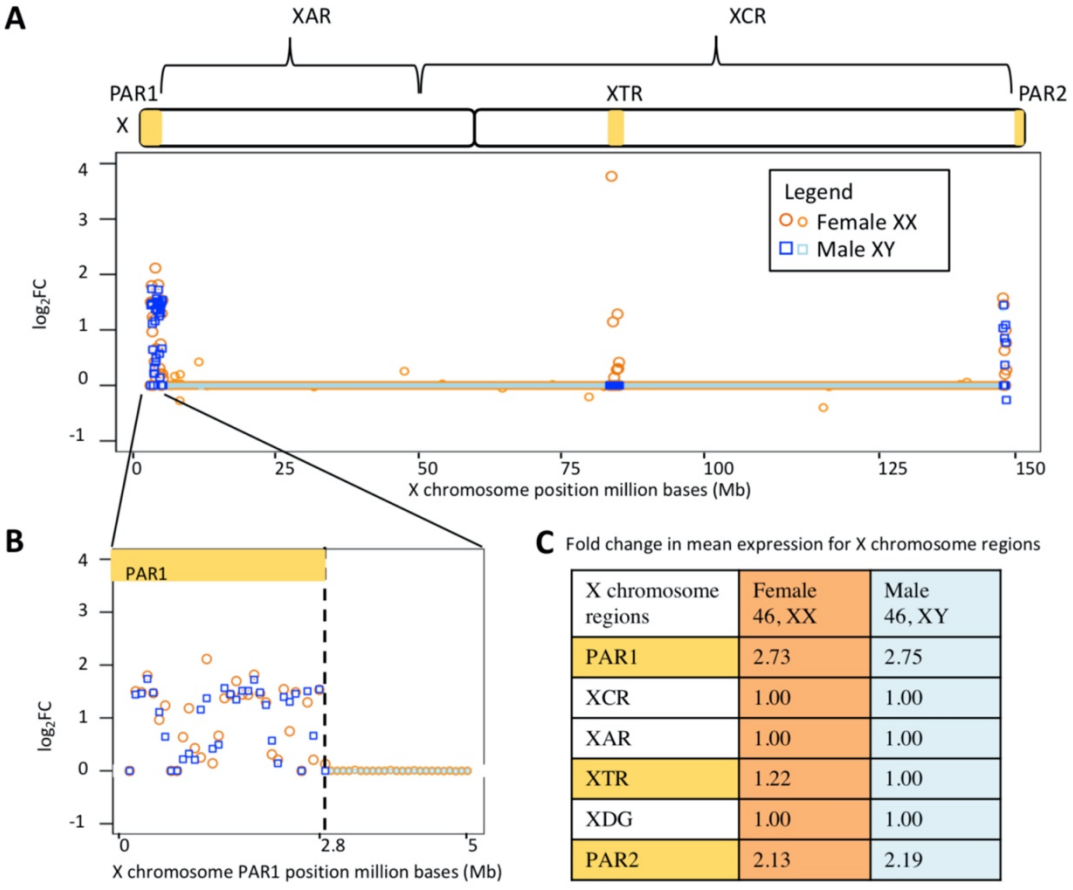
## Results

***RNA-Seq Reads Aligned to Autosomes Do Not Vary Much Between Reference Genomes.*** We compared total mapped reads when reads were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement. Reads mapped across the whole genome, including the sex chromosomes, decreased when samples were aligned to a reference genome informed on the sex chromosome complement, paired t-test p-value < 0.05 (Additional files 7 - 9). This was true regardless of the read aligner used, HISAT or STAR, or of the sex of the sample, XY or XX. To test the effects of realignment on an autosome, we selected chromosome 8, because of its similar size to chromosome X. Overall, there is a slight mean increase in reads mapping to chromosome 8 when samples are aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome (Additional file 9). For female XX samples, the mean increase in reads mapping for chromosome 8 was 42.2 reads for whole blood, 50.25 for brain cortex, 109.9 for breast, 68.5 for liver, and 98.2 for thyroid (Additional file 9), which was significant using a paired t-test, p-value < 0.05 in all tissues (Additional file 9). Male XY samples also showed a mean increase in reads mapping for chromosome 8. The mean increase in reads mapping to chromosome 8 for male whole blood samples was 0.84, 2.38 for brain cortex, 5.58 for breast, 3.2 for liver, and 5 for thyroid (Additional file 9). There was a significant increase, p-value < 0.05 paired t-test, for reads mapping to chromosome 8 for male brain cortex, breast, liver, and thyroid samples. There was no significant increase in reads mapping for male whole blood for chromosome 8 (Additional file 9).

***Reads Aligned to the X Chromosome Increase in Both XX and XY Samples When Using a Sex Chromosome Complement Informed Reference Genome.*** We found that when reads were aligned to a reference genome informed by the sex chromosome complement for both male XY and female XX tissue samples, reads on the X chromosome increased by ~0.12% when aligned using HISAT. For all tissues and both sexes we observe an average increase of 1,991 reads on chromosome X. We observe an increase in reads mapping to the X chromosome for all tissues and for each sex, which was significant using a paired t-test, p-value < 0.05 (Additional file 9). Reads on the Y chromosome decreased 100% (67,033 reads on average) across all female XX samples and by ~57.32% (69,947 reads on average) across all male XY samples when aligned using HISAT (Additional file 7 & 9). Similar increases in X chromosome and decreases in Y chromosome reads when aligned to a sex chromosome complement informed reference were observed when STAR was used as the read aligner for both male XY and female XX samples (Additional file 8 & 9).

***Aligning to a Sex Chromosome Complement Informed Reference Genome Increases the X chromosome PAR1 and PAR2 Expression.*** We next explored the effect of changes in read alignment on gene expression. There was an increase in pseudoautosomal regions, PAR1 and PAR2, expression when reads were aligned to a reference genome informed on the sex chromosome complement for both male XY and female XX samples (Additional file 10 & 11). We found an average of 2.73 $\log_2$ fold increase in expression in PAR1 expression for female XX brain cortex samples and 2.75 $\log_2$ fold increase in expression in PAR1 for male XY brain cortex samples using HISAT (Figure 4). The X-transposed region, XTR, in female XX brain cortex samples showed a

1.22 $log_2$ fold increase in expression and no change in male XY brain cortex samples. PAR2 showed an average of 2.13 $log_2$ fold increase for female XX brain cortex samples and 2.19 $log_2$ fold increase in PAR2 for male XY brain cortex samples using HISAT, with similar results for STAR read aligner (Additional file 10 & 11). Complete lists of the $log_2$(CPM+0.25/L) values for each X chromosomal gene and each gene within the whole genome for male XY and female XX samples are in Additional file 12 available on Dryad for download under https://doi.org/10.5061/dryad.xksn02vbv.




| X chromosome regions | Female 46, XX | Male 46, XY |
|---|---|---|
| PAR1 | 2.73 | 2.75 |
| XCR | 1.00 | 1.00 |
| XAR | 1.00 | 1.00 |
| XTR | 1.22 | 1.00 |
| XDG | 1.00 | 1.00 |
| PAR2 | 2.13 | 2.19 |

*Chapter 1. Figure 4. X Chromosome RNA-Seq Alignment Differences in the Brain Cortex.* We plot log2 fold change (FC) across (A) the entire X chromosome and (B) the first 5 million bases (Mb) and show (C) the average fold change in large genomic regions on the X chromosome between the aligning brain cortex using HISAT to the default

genome and aligning to a sex chromosome complement informed reference genome. For log2 FC, a value less than zero indicates that the gene showed higher expression when aligned to a default reference genome, while values above zero indicate that the gene shows higher expression when aligned to a reference genome informed by the sex chromosome complement of the sample. Samples from genetic females are plotted in orange circles, while samples from males are plotted in blue squares. Darker shades indicate which gene points are in PAR1, XTR, and PAR2 while lighter shades are used for genes outside of those regions.

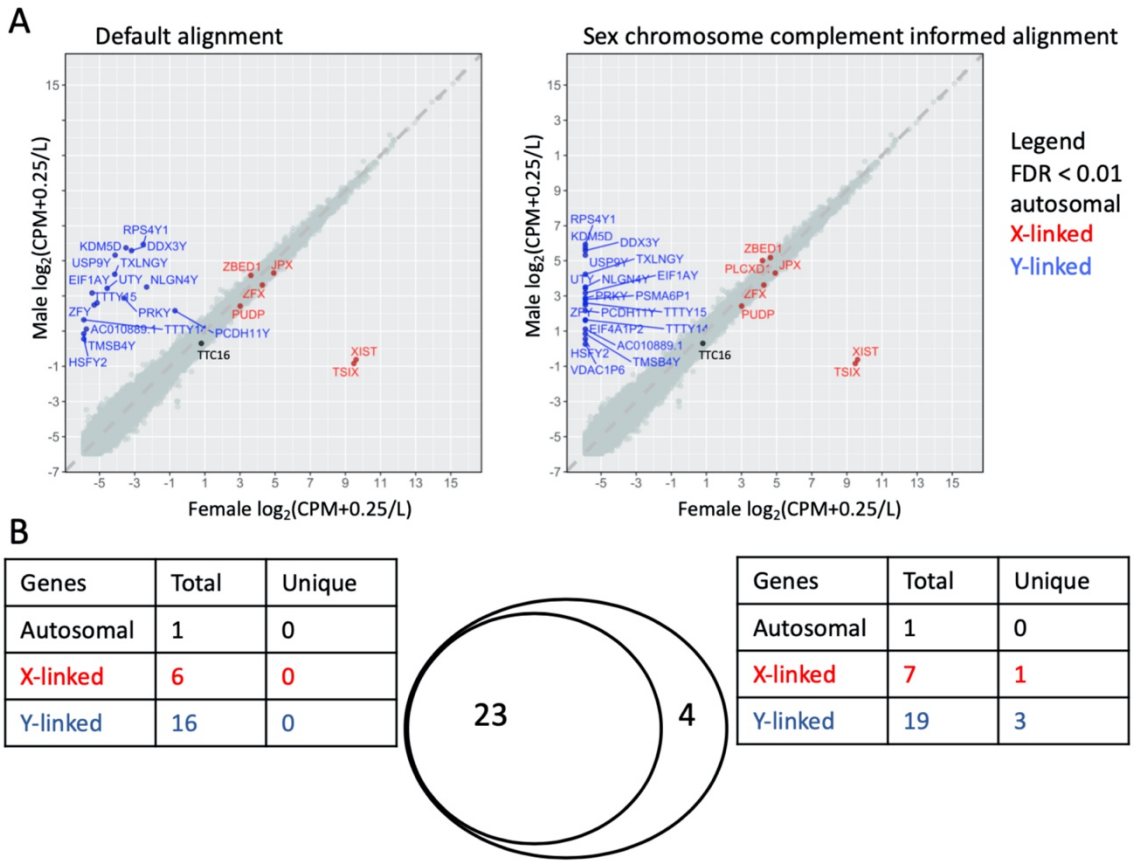
### *Regions Outside the PARs and XTR Show Little Difference in Expression Between Reference Genomes.*

Intriguingly, regions outside the PARs on the X chromosome were less affected by the choice of reference genome. Across the entire X-conserved region, we observed practically no change in estimates of gene expression between the default and sex chromosome complement informed references (e.g., a 0.99 $\log_2$ fold in male thyroid samples, and 1.00 $\log_2$ fold change in female brain cortex samples, essentially showing no difference (Additional file 10 & 11). Additionally, X and Y homologous genes (*AMELX*, *ARSD*, *ARSE*, *ARSF*, *CASK*, *GYG2*, *HSFX1*, *HSFX2*, *NLGN4X*, *OFD1*, *PCDH11X*, *PRKX*, *RBMX*, *RPS4X*, *SOX3*, *STS*, *TBL1X*, *TGIF2LX*, *TMSB4X*, *TSPYL2*, *USP9X*, *VCX*, *VCX2*, *VCX3A*, *VCX3B*, *ZFX*) showed little to no increase in expression when aligned to a sex chromosome complement informed reference genome compared to aligning to a default reference genome (Additional file 13). *PCDH11X* showed the highest increase in expression for all tissues regardless of read aligner. The $\log_2$ fold increase in expression for *PCDH11X* for female samples when aligned using HISAT was 0.4, 0.28, 0.33, 0.16, and 0.16 for whole blood, brain cortex, breast, liver, and thyroid, respectively. Other X and Y homologous genes sometimes increased in expression depending on the tissue and sometimes there was no change in expression (Additional file 13). Next to *PCDH11X*, the most increase in expression in an

X and Y homologous genes was *VCX3B, NLGN4X,* and *VCX3A*. *NLGN4X* in whole blood showed a 0.14 $\log_2$ fold increase when aligned using HISAT. *VCX3B* showed a 0.2 $\log_2$ fold increase in brain, *NLGN4X* showed a 0.04 $\log_2$ fold increase in breast, *VCX3A* showed a 0.07 $\log_2$ fold increase in liver, and *VCX3B* showed a 0.04 $\log_2$ fold increase in thyroid, when aligned using HISAT (Additional file 13).

***A Sex Chromosome Complement Informed Reference Genome Increases the Ability to Detect Sex Differences in Gene Expression.*** We next investigated how this would affect gene differential expression between the sexes. Generally, we find that more genes are differentially expressed on the sex chromosomes between the sexes when the sex chromosome complements are taken into account. The number of differentially expressed genes on the autosomes remained the same or increased. At a conservative Benjamini-Hochberg adjusted p-value of $< 0.01$ and aligning with HISAT, we find 4 new genes (3 Y-linked and 1 X-linked) that are only called as differentially expressed between the sexes in the brain cortex when aligned to reference genomes informed on the sex chromosome complement (Figure 5; Additional file 14). We observed similar trends in changes for differential expression between male XY and female XX for whole blood, breast, liver, and thyroid samples using either HISAT or STAR as the aligner (Additional file 14). For example, in whole blood, 3 additional genes are called as being differentially expressed between the sexes using HISAT, while 1 additional gene is called differentially expressed when aligned using STAR. Additionally, when taking sex chromosome complement into account, the number of genes called as differentially expressed between the sexes for the breast samples increased by 13 genes (8 autosomal, 3 X-linked and 2 Y-linked) using HISAT and by 8 genes using STAR (6 autosomal and 2 X-linked)

(Additional file 14 & 15). For all tissues, no genes were uniquely called as being differentially expressed between the sexes when aligned to a default reference genome compared to a reference genome informed on the sex chromosome complement (Additional file 14 & 15). Rather, only when samples were aligned to a sex chromosome complement did, we observe an increase in the genes called as being differentially expressed (Figure 5; Additional file 14 & 15).




| Genes | Total | Unique |
|---|---|---|
| Autosomal | 1 | 0 |
| X-linked | 6 | 0 |
| Y-linked | 16 | 0 |

| Genes | Total | Unique |
|---|---|---|
| Autosomal | 1 | 0 |
| X-linked | 7 | 1 |
| Y-linked | 19 | 3 |

*Chapter 1. Figure 5. Sex Chromosome Complement Informed Alignment Calls More Sex-linked Genes as Being Differentially Expressed.* (A) Sex differences in the gene expression, log2(CPM + 0.25/L), between the twenty samples from genetic males and females are shown when aligning all samples to the default reference genome (left) and a reference genome informed on the sex chromosome complement (right) for the brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p

value < 0.01, are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. (B) We show the overlap between genes that are called as differentially expressed when all samples are aligned to the default genome, and genes that are called as differentially expressed when aligned to a sex chromosome complement informed genome. When samples were aligned to a reference genome informed on the sex chromosome complement, 27 genes were called as differentially expressed between the sexes, of which 4 were uniquely called in the sex chromosome complement informed alignment. There were no genes that were uniquely called as differentially expressed when aligned to a default reference genome.

***Increase in Gene Enrichment Pathways When Samples are Aligned to a Sex Chromosome Complement Informed Reference Genome.*** A sex chromosome complement informed reference genome increases the ability to detect genes as differentially expressed between the sexes and thus alters gene enrichment results. When the thyroid samples were aligned using a sex chromosome complement informed reference genome using HISAT, genes up-regulated in male XY samples still show enrichment for positive regulation of transcription from RNA polymerase II (found when aligning to a default reference genome), but additionally find postsynaptic membrane assembly, postsynaptic membrane organization, and vocalization behavior (Additional file 16). These additional GO enrichments in the male XY thyroid samples involve *NRXN1* and *NLGN4Y* genes, both of these genes are located on the Y chromosome. GO enrichment analysis of genes that are more highly expressed in female liver compared to male liver samples, when samples were aligned to a default reference genome using HISAT, were genes involved in modification histone lysine demethylation (Additional file 16). However, when these samples were aligned to a sex chromosome complement informed reference genome, genes upregulated in females were enriched for histone lysine demethylation as well as negative regulation of endopeptidase activity, negative

regulation of peptidase activity, cytoplasmic actin-based contraction involved in cell motility (Additional file 16). These additional GO enrichments in the female XX liver samples include the involvement of *KDM6A, DDX3X,* and *VIL1. KDM6A, DDX3X* are X-linked and *VIL1* is on chromosome 2. Whole blood, brain cortex, male liver, and female thyroid samples showed no difference in GO enrichment pathways when using a default reference genome compared to a sex chromosome complement reference genome for alignment when using HISAT with similar results for STAR as the read aligner (Additional file 17). Thus, while there won't always be a difference, aligning to a sex chromosome complement informed reference genome can increase ability to detect enriched pathways.

***Using Sex-linked Genes Alone is Inefficient for Determining the Sex Chromosome Complement of a Sample.*** The sex of each sample used in this analysis was provided in the GTEx manifest. We investigated the expression of genes that could be used to infer the sex of the sample. We studied X and Y homologous genes (*DDX3X/Y*, *PCDH11X/Y*, *USP9X/Y*, *ZFX/Y*, *UTX/Y*), *XIST*, and *SRY* gene expression in male and female whole blood, brain cortex, breast, liver, and thyroid (Figure 2; Additional file 3 & 4). Both males and females are expected to show expression for the X-linked homologs, whereas only XY samples should show expression of the Y-linked homologs. Further, *XIST* expression should only be observed in XX samples and *SRY* should only be expressed in samples with a Y chromosome. Using the default reference genome for aligning samples, we observed a small number of reads aligning to the Y-linked genes in female XX samples, but also observed clustering by sex for *DDX3Y*, *USP9Y*, *ZFY*, and *UTY* gene expression (Figure 2). Male XY samples showed expression

for *DDX3X*, *DDX3Y*, *USP9X*, *ZFX*, and *UTX* (greater than 5 $\log_2$(CPM+2/L). Female XX samples showed expression for *XIST* (greater than 4.0 $\log_2$(CPM+2/L) and male XY samples showed little to no expression for *XIST* (less than 0 $\log_2$(CPM+2/L) with the exception of 2 male whole blood samples and 1 male liver sample, which showed greater than 5 $\log_2$(CPM+2/L) expression). In contrast to the default reference genome, when aligned to a sex chromosome complement informed reference genome, samples cluster more distinctly by sex for *DDX3Y*, *USP9Y*, ZFY, and *UTY*, all showing at least a 4 $\log_2$(CPM+2/L) difference between the sexes (Figure 2; Additional file 3 & 4). *SRY* is predominantly expressed in the testis (Albrecht et al., 2003; Turner et al., 2011) and typically one would expect *SRY* to show male-specific expression. In our set, we did not observe *SRY* expressed in any sample, and so it could not be used to differentiate between XX and XY samples (Figure 2, Additional file 3 & 14). In contrast, the X-linked gene *XIST* was differentially expressed between genetic males and genetic females in both genome alignments (default and sex chromosome complement informed) for the whole blood, brain cortex, breast, liver, and thyroid samples with the exception of 3 male XY samples. *XIST* expression is important in the X chromosome inactivation process (Carrel & Willard, 2005) and serves to distinguish samples with one X chromosome from those with more than one X chromosome (Tukiainen et al., 2016). However, this does not inform about whether the sample has a Y chromosome or not. For X-Y homologous genes, we do not find sex differences in read alignment with either default or sex chromosome complement informed for the X-linked homolog. When aligned to a default reference genome, female XX samples showed some expression for homologous Y-

linked genes, but only presence/absence of Y-linked reads alone is insufficient to determine sex chromosome complement of the sample (Figure 2, Additional file 3).

***No Y-linked Transcript Expression in Female XX Samples When Quantification was Estimated Using a Transcriptome Index Informed on the Sex Chromosome Complement.*** A pseudo-alignment shows similar effects of the reference to that of an alignment approach (Figure 5, Additional files 18 & 19). We observed no Y-linked expression in female XX samples when transcript quantification was estimated using a Y-masked sex chromosome complement reference transcriptome index. This was true for both the Ensembl and gencode pseudo-alignment with a sex chromosome complement reference transcriptome index (Additional files 18 & 19). Interestingly, there was a large difference between the Ensembl and gencode reference files. The transcript IDs in the transcriptome cDNA fasta and the transcript IDs in the annotation file are not one-to-one for the Ensembl assembly (Shanrong Zhao & Zhang, 2015). There are 190,432 transcript sequences in the Ensembl cDNA fasta file but there are 199,234 transcripts in the Ensembl annotation file. Notably, Ensembl's cDNA reference transcriptome fastas does not contain known transcripts such as the XIST transcripts (Eyras et al., 2004). The Ensembl reference transcriptome fasta also does not contain the Y PARs transcript sequences, it only contains the X PAR transcript sequences. In contrast, the gencode cDNA reference transcriptome fasta and annotation file both contain 206,694 sequences, including the Y PARs. Regardless of using an Ensembl or gencode transcriptome, female XX sample show Y-linked expression when using a default refence transcriptome index for pseudo-alignment, however the changes

necessary for making a sex chromosome complement informed reference are different for the two builds.

**Discussion**

For accuracy, the sex chromosome complement of the sample should be taken into account when aligning RNA-Seq reads to reduce misaligning sequences. Neither Ensembl or Gencode human reference genomes are correct for aligning both XX and XY samples. The Ensembl GRCh38 human reference genome includes all 22 autosomes, mtDNA, the X chromosome, the Y chromosome with the Y PARs masked, and contigs (Aken et al., 2017). The Gencode hg19 human reference genome includes everything with no sequences masked (Harrow et al., 2012).

Measurements of X chromosome expression increase for both male XY and female XX whole blood, brain cortex, breast, liver, and thyroid samples when aligned to a sex chromosome complement informed reference genome versus aligning to a default reference genome (Figure 4). While we see increases in measured expression for PAR1 and PAR2 genes in both males and females, we only observe a difference in measured XTR expression in females. This is because while the PARs are 100% identical between the X and Y and so one copy (here we mask the Y-linked copy) should be masked, the XTR is not hard-masked in the YPARs-masked reference genome. The XTR is not identical between the X and Y; it shares 98.78% homology between X and Y but no longer recombines between X and Y (Veerappa et al., 2013) (Figure 1A) and because of this divergence, is therefore not hard-masked when aligning male XY samples. Tukiainen et al., (2016) and others have shown that PAR1 genes have a male bias in expression

(Tukiainen et al., 2016). Our findings here support this regardless if the samples were aligned to a default or a sex chromosome complement reference genome (Additional file 11 & 12). Differential expression results changed when using a sex chromosome complement informed alignment compared to using a default alignment. When aligned to a default reference genome, due to sequence similarity, some reads from female XX samples aligned to the Y chromosome (Figure 2; Figure 5). However, when aligned to a reference genome informed by the sex chromosome complement, female XX samples no longer showed Y-linked gene expression, and more Y-linked genes were called as being differentially expressed between the sexes (Figure 2; Figure 5; Additional file 12 & 15). This suggests that if using a default reference genome for aligning RNA-Seq reads, one would miss some Y-linked genes as differentially expressed between the sexes (Figure 5). Furthermore, these Y-linked genes serve in various important biological processes, thus altering the functional interpretation of the sex differences (Additional file 16 & 17). Only when samples were aligned to a sex chromosome complement reference genome did we observe more genes called as differentially expressed between the sexes (Additional file 14). An increase in genes called differentially expressed additionally alters the GO analysis results (Additional file 16 & 17). When samples were aligned to a default reference genome we sometimes missed GO pathways or misinterpreted which were the top pathways.

The choice of read aligner has long been known to give slightly differing results of differential expression due to the differences in the alignment algorithms (Conesa et al., 2016; Costa-Silva et al., 2017). Differences between HISAT and STAR could be contributed to differences in default parameters for handling multi-aligning reads (Kim et

al., 2015). We show that regardless of choice of read aligner, HISAT or STAR, we observe similar results. Sample size has also long been known to alter differential expression analysis (Ching et al., 2014; Lamarre et al., 2018; Shilin Zhao et al., 2018). We therefore additionally replicated our findings in a smaller sample size of 3 male XY compared to 3 female XX samples for whole blood and brain cortex tissue and where the samples were randomly selected and confirmed the results from the larger sample size (Additional file 20).

In addition to reference-based quantification, we tested whether quantifying sex-linked reads with a pseudo-aligner would be affected by using a sex chromosome complement reference. Previous studies have shown that reference-based alignment is not necessary for high-quality estimation of transcript levels (Zielezinski et al., 2017). However, we observed expression estimates for Y-linked transcripts in female XX samples when using a default reference transcriptome index for pseudo-alignment quantification estimates. In contrast, when a sex chromosome complement informed reference transcriptome index was used, we observed no Y-linked expression in female XX samples. Salmon, and other alignment-free tools such as Kallisto (Bray et al., 2015) and Sailfish (R et al., 2014), build an index of k-mers from a reference transcriptome. The k-mer transcriptome index is used to group pseudoalignments belonging to the same set of transcripts to directly estimate the expression of each transcript. A k-mer alignment free approach is faster and less demanding than alignment protocols (Zielezinski et al., 2017); however, a sex chromosome complement informed transcriptome index should be carefully considered because even a k-mer approach is not sensitive to regions that are 100% identical in sequence. Additionally, alignment-free methods are not as robust in

quantifying expression estimates for small RNAs and lowly-expressed genes (Wu et al., 2018).

The choice of reference transcriptome or reference genome can also give slightly differing results of differential expression due to the difference in which transcripts are included in the transcriptome (Shanrong Zhao & Zhang, 2015). The Ensembl cDNA does not include the Y PAR linked transcripts whereas the gencode transcriptome fasta includes both the X and Y PARs. The Ensembl transcriptome does not include non-coding RNAs, such as *XIST* transcripts. The *XIST* gene is called as being up-regulated in the female XX samples for all tissues and all comparisons except for when transcript expression was estimated using the Ensembl reference transcriptome (Additional file 15, 18, & 19). Given the current builds, for RNA-seq projects interested in sex chromosome linked transcript expression, we suggest that researchers use a gencode sex chromosome complement informed reference transcriptome index.

Ideally, one would use DNA to confirm presence or absence of the Y chromosome, but if DNA sequence was not generated, one would need to confirm the genetic sex of the sample by assessing expression estimates for X-linked and Y-linked genes. To more carefully investigate the ability to use gene expression to infer sex chromosome complement of the sample, we examined the gene expression for a select set of X-Y homologous genes, as well as *XIST* and *SRY* that are known to be differentially expressed between the sexes (Figure 2, Additional file 13). The samples broadly segregated by sex for Y-linked gene expression using default alignment. However, the pattern was messy for each individual Y-linked gene. Thus, if inferring sex from RNA-Seq data, we recommend using the estimated expression of multiple X-Y homologous

genes and *XIST* to infer the genetic sex of the sample. Samples should be aligned to a default reference genome first to look at the expression for several Y-specific genes to determine if the sample is XY or XX. Then samples should be realigned to the appropriate sex chromosome complement informed reference genome. Independently assessing sex chromosome complement of samples becomes increasingly important as karyotypically XY individuals are known to have lost the Y chromosome in particular tissues sampled, as shown in Alzheimer Disease (Dumanski et al., 2016), age-related macular degeneration (Grassmann et al., 2019), and in the blood of aging individuals (Forsberg, 2017), but should not have *XIST* expression. However, *XIST* may not be a sufficient marker alone to infer sex chromosome complement, especially in cancer in samples from XX individuals, where the inactive X can become reactivated (Chaligné et al., 2015). Self-reported sex may not match the sex chromosome complement of the samples, even in karyotypic individuals.

**Conclusion**

Here we show that aligning RNA-Seq reads to a sex chromosome complement informed reference genome will change the results of the analysis compared to aligning reads to a default reference genome. We previously observed that a sex chromosome complement informed alignment is important for DNA as well (Webster et al., 2019). A sex chromosome complement informed approach is needed for a sensitive and specific analysis of gene expression on the sex chromosomes (Khramtsova et al., 2018). A sex chromosome complement informed reference alignment resulted in increased expression of the PARs of the X chromosome for both male XY and female XX samples. We further

found different genes called as differentially expressed between the sexes and identified sex differences in gene pathways that were missed when samples were aligned to a default reference genome.

## Perspectives and Significance

The accurate alignment and pseudo-alignment of the short RNA-Seq reads to the reference genome or reference transcriptome is essential for drawing reliable conclusions from differential expression data analysis on the sex chromosomes. We strongly urge studies using RNA-Seq to carefully consider the genetic sex of the sample when quantifying reads, and provide a framework for doing so in the future (https://github.com/SexChrLab/XY_RNAseq).

## Supplementary Information

Supplemental tables and figures are located in chapter 1. appendices A.

CHAPTER 2

# Sex Differential Gene Expression in the Late First Trimester and in Term Human Placentas is Replicated in Adult Tissues


## ABSTRACT

Early life exposures during pregnancy may be predictive of lifelong health outcomes. Further, pregnancy complications vary based on the fetus's genetic sex, suggesting sex differences in the placenta function and gene expression. Yet, sex differences in gene expression within the placenta at different time points throughout pregnancy and comparisons to adult tissues remains poorly characterized. Here, we collect and characterize sex differences in gene expression in term placentas (term $\geq$ 36.6 weeks; 23 male XY and 27 female XX). We then compare sex differences in term placentas with previously collected first trimester placenta samples and with 42 non-reproductive adult tissues. We identify 268 and 53 sex differentially expressed genes, adjusted p-value < 0.05, in the uncomplicated late first trimester and term placentas, respectively. Genes more highly expressed in female placentas involve translational initiation, regulation of sister chromatid cohesion, histone lysine demethylation, and RNA binding. Genes more highly expressed in male placentas were identified to be involved in histone lysine demethylation, protein demethylation, and cellular glucuronidation regulation. Next, we found that sex differential gene expression in the term placenta is highly correlated with sex differences in gene expression in 42 non-reproductive adult tissues (*r* ranged from 0.892 to 0.957, p-value < 0.01). Although the

above observation is largely driven by sex-linked genes, we do observe some positivie significant correlations for sex differences in gene expression for autosomal genes between term placentas and adult brain regions. In general, we find that sex differences in expression are conserved in late first trimester and term placentas, as well as in adult non-reproductive tissues.

## Background

Early life exposures during pregnancy may be predictive of lifelong health outcomes (Alur 2019). Additionally, there are sex differences in the incidence of adverse adult health outcomes correlated with sex differences during pregnancy (Alur 2019). For example, maternal obesity is associated with obesity in male offspring but not in females at one year of age (Bridgman et al. 2018). Pregnancy complicated by acute asthma led to intrauterine growth restriction (IUGR) status or preterm delivery when the pregnancy was carrying a male fetus, and if the pregnancy was with a female fetus, there was reduced growth but not to the extent of leading to IUGR (Clifton 2010). Preterm birth is a strong predictor of adverse health outcomes later in life (Farooqi et al. 2006), and pregnancies with a male fetus are more likely to be preterm than pregnancies with a female fetus (McGregor et al. 1992; Ito et al. 2017; Peacock et al. 2012).

Pregnancy complications vary in incidence based on genetic sex of the developing fetus. Some pregnancy complications are more common in male-bearing pregnancies, such as subchorionic hemorrhage (Cuestas, Bas, and Pautasso 2009), delivery by cesarean section (Zeitlin et al. 2002), preterm birth (Zeitlin et al. 2002), and term preeclampsia (Vatten and Skjaerven 2004), while others are more common in female-bearing pregnancies, such as intrauterine growth restriction (IUGR) (Sheiner 2007; Melamed, Yogev, and Glezerman 2010), and preterm preeclampsia (Global Pregnancy Collaboration: et al. 2017; Vatten and Skjaerven 2004). Pregnancy complications often involve improper placenta function, which may be driven by changes in gene expression (Kaartokallio et al. 2015; Oros et al. 2017; Lekva et al. 2016; Sheikh, Satoskar, and Bhartiya 2001). The placenta shares the genotype of the developing fetus, which is

typically XY male or XX female. Sex differences in placenta gene regulation may drive the observed sex differences in pregnancy complications (Gonzalez et al. 2018). For example, several studies have shown that placenta gene expression differs between healthy term placentas from those characterized by preterm birth (Kaartokallio et al. 2015) and from placentas from offspring with IUGR (Sheikh, Satoskar, and Bhartiya 2001).

Male XY and female XX fetuses respond differently to the same intrauterine environment, regulated in part by placental gene expression (Gonzalez et al. 2018). Gonzalez et al. 2018 found 58 sex differentially expressed genes in the late first trimester (11.5 - 13.5 weeks) placentas, many of which are located on the sex chromosomes, X and Y. In adult tissues, Lopes-Ramos et al. 2020 showed that most autosomal genes that are sex differentially expressed are tissue-dependent, and sex differentially expressed genes common among many tissues were enriched for sex chromosome genes (Camila M. Lopes-Ramos et al. 2020). Here we examine sex differences in gene expression across the life span.

We generate RNA and DNA from 30 male and 30 female term, ≥ 36.6 weeks, placentas from uncomplicated births. We compare sex differences in term uncomplicated placentas with late first trimester placenta samples (Gonzalez et al. 2018), and adult tissues, to better understand the development of sex differential expression across the life span. We find that sex differences in gene expression in term placentas are correlated with sex differences in the late first trimester placentas. Additionally, sex differences in gene expression on the sex chromosomes in the placenta show similar sex differences in expression in adult tissues.

## Methods

***Samples.*** We collected 60 term, ≥ 36.6 weeks, placentas from uncomplicated pregnancies; 30 from assigned female at birth and 30 from assigned male at birth. The placenta samples here were carefully selected to represent the fetal component of the placenta. Three samples were obtained from each placenta, one for whole exome sequencing, and two tissue samples from opposing quadrants for RNA sequencing (RNAseq) for a total of 120 RNAseq placenta samples. The placentas were collected and sequenced at two different times, with 12 male and 12 female placentas in the first batch and 18 male and 18 female placentas in the second batch. All placenta samples were collected immediately following live birth via cesarean section (CS) except for one male placenta, which was collected following spontaneous vaginal delivery, (SVD), sample ID YPOPS0007M. However, the spontaneous vaginal delivery sample was removed from the study due to failed GC content (Additional Table 1).

***RNAseq Data Processing.*** RNAseq libraries were constructed using Illumina TruSeq reverse forward stranded RiboZero library prep to deplete cytoplasmic polyadenylated tails. Samples were sequenced to 50 million (M) 2 x 100 bp paired-end reads. Samples were checked for quality using FastQC version 0.11.8 (Andrews 2010) and aggregated using MutliQC version 0.9 (Ewels et al. 2016a). RNAseq data were trimmed to remove Illumina universal sequence adapters and to only include paired reads with a PHRED score of ≥ 30, minimum base-pair length of 75, and average read quality of 20 using bbduk as part of bbmap version 38.22 (Bushnell 2014). Post trimming quality was again checked using FastQC version 0.11.8 (Andrews 2010) and MutliQC version

0.9 (Ewels et al. 2016b). Post trimming samples had an average of 35.18M and median of 35M reads (Additional Table 2).

All RNAseq samples were aligned to Gencode GRCh38.p12 human reference genome informed on the sex chromosome complement of the sample (Olney et al. 2020a; Webster et al. 2019a) using HISAT2 version 2.1.0 for alignment (Kim, Langmead, and Salzberg 2015a) and SubRead FeatureCounts version 1.5.2 for quantification (Liao, Smyth, and Shi 2014a). Briefly, the sex chromosome complement of the sample was first checked by investigating the expression of five Y-linked and one X-linked genes *EIF1AY, KDM5D, UTY, DDX3Y, RPS4Y1,* and *XIST.* A sample with presence of a Y chromosome will show expression for *EIF1AY, KDM5D, UTY, DDX3Y,* and *RPS4Y1* Y-linked genes. Samples with at least two X chromosomes will show expression for *XIST* (Additional Figure 1). Samples with at least two X chromosomes will show expression for *XIST.* Samples with the presence of the Y chromosome will show expression for Y-linked genes. Samples with no evidence of a Y chromosome were aligned to a reference genome with the entire Y chromosome masked with Ns to avoid mis-mapping of homologous X-Y sequence reads (Olney et al. 2020b). Samples with evidence of a Y chromosome were aligned to a reference genome with the Y chromosome pseudoautosomal regions (PARs) masked out as those regions are replicated 100% on the X chromosome PARs. We followed the XY_RNAseq readme (Olney et al. 2020) to utilize a Ymasked, and YPARs masked reference genome and HISAT -x index function to create two sex chromosome complement reference indexes used for alignment. HISAT2 alignment was performed with the following parameters, --dta for downstream transcriptome assembly, --rna-strandness RF to indicate the sequences are reverse

forward, --phred 33 encoding, and pair-end alignment. RNAseq alignment files were then sorted, read groups were added, duplicates were marked, and files were indexed using bamtools 2.5.1 (Barnett et al. 2011) and Picard 2.9.2 ("Picard Tools - By Broad Institute" n.d.). FeatureCounts was employed using --primary to only use primary alignments and -p 2 to specify the minimum number of consensus reads from the same pair, suggested for paired-end read data (Liao, Smyth, and Shi 2014b). FeatureCounts uses the gene annotation file to infer exon-exon junctions from connecting each pair of neighboring exons from the same gene (Liao, Smyth, and Shi 2014). FeatureCounts was run for the gene level using -g gene_name. There are 57,133 genes in the Gencode GRCh38.p12 human reference genome used in this analysis.

***Exome Data Processing.*** We used FastQC version 0.11.8 (Andrews 2010) and MutliQC version 0.9 (Ewels et al. 2016) for visualizing quality for whole-exome data. We trimmed adapters using bbduk as part of bbmap version 38.22 (Bushnell 2014) with the following parameters: qtrim=rl trimq=30 minlen=75 maq=20. We used bwa-mem version 0.7.17 (Li 2013) to align the whole exome samples. Samples were aligned to a sex chromosome complement reference; see RNAseq data processing for more details (Olney et al. 2020d; Webster et al. 2019b). Post alignment, PCR duplicates were marked using Picard version 2.18.27 ("Picard Tools - By Broad Institute" n.d.). To genotype variants, we used GATK version 4.1.0.0 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). We first used GATK's HaplotypeCaller to generate GVCF files. Second, we combined GVCF from 60 individuals, 30 male XY, and 30 female XX, using GATK's CombineGVCFs.

***Late First Trimester Placentas.*** Late first trimester, 10.5 - 13.5 weeks, human placenta RNAseq samples from Gonzalez et al. 2018 (Gonzalez et al. 2018) were downloaded from NCBI GEO Accession GSE109082 using fastq-dump -I --split-files (Leinonen et al. 2011). There are 17 female XX and 22 male XY placenta samples in this data set, all of which self-reported as white (Gonzalez et al. 2018). GSE109082 were processed similarly as the full-term uncomplicated placentas with one exception, trimming for quality. GSE109082 transcriptome samples before trimming had an average of 22.53M 2 x75 bp paired-end reads. GSE109082 paired-end reads were checked for quality using FastQC version 0.11.8 (Andrews 2010) and MutliQC version 0.9 (Ewels et al. 2016d). GSE109082 samples were then trimmed to remove Illumina sequence adapters and only included paired reads with a PHRED score of ≥ 25, a minimum base-pair length of 40, and average read quality of 10 using bbduk 38.22 (Bushnell 2014). Post-trimming quality was checked using Fastqc and MutliQC. Post trimming samples had an average of 20.6M and a median of 18.8M reads (Additional Table 2). Reads were then aligned to a sex chromosome complement Gencode GRCh38.p12 reference genome using HISAT2 (Kim, Langmead, and Salzberg 2015b). Gene level counts were obtained using SubRead FeatureCounts (Liao, Smyth, and Shi 2014d). The first trimester placentas were reverse forward sequencing, the same as the term placentas presented here. The HISAT2 and FeatureCounts parameters were the same for both the late first trimester (Gonzalez et al. 2018) and the full-term placentas.

***Multidimensional Scaling.*** Multidimensional scaling (MDS) was employed on the expression data to determine expression similarity among samples. MDS of the expression counts following subRead FeatureCounts was generated using plotMDS of the

limma package (Law et al. 2014). PlotMDS is a slightly modified MDS that plots the transcript expression profiles on a two-dimensional scatterplot so that distances on the plot approximate the typical $\log_2$ fold changes between the samples. MDS plots were generated using the gene.selection parameter and selecting "common" for all shared genes. This was repeated for the top 100 genes that show the most extensive standard deviations between samples (Additional Figure 2). The full-term placentas were sequenced at two different time points in batch 1 and batch 2. Before clustering with MDS, we accounted for batch effects using the removeBatchEffect part of the limma package (Law et al. 2014b). This was for visualization purposes only. Batch was included as a covariant in the linear model downstream; see Differential expression.

***Excluding RNAseq Samples.*** Samples that failed quality control (QC) were removed from downstream analysis. Samples were removed that had less than 12.5M or higher than 90M sequences remaining after trimming. If more than 30% of the reads deviate from the sum of the deviations from the normal distribution of the per-sequence GC content as defined by the FASTQC report, then the sample was removed (Additional Table 3). Samples were also excluded that clustered with the opposite sex of the reported sex assigned at birth (Additional Figure 2 & Table 3). There are 23 male XY and 27 female XX full-term placentas that passed QC and were included in the downstream analyses. All 17 female XX and 22 male XY first trimester placentas from Gonzalez et al. 2018 passed QC and were kept for downstream analysis.

***Subject Demographic Analysis.*** All term, $\geq$ 36.6 weeks, samples in our data set have a self-reported race of either Asian, Black, White, or Unknown (Additional Table 1). Additionally, we inferred population ancestry from the variants obtained from the

whole-exome data using Peddy (Pedersen and Quinlan 2017) (Additional Figure 3 & Additional Table 1). The resulting outputs of the PCA analysis in Peddy yielded principal components (PC) that were used to assign predicted ancestry. PC1 and PC2 were used later downstream as covariates in the linear model for the differential expression analysis. We did not infer population ancestry from the first trimester GSE109082 placentas as this data set only includes RNAseq data, and not DNA.

***Quantify Technical and Biological Variation in RNAseq Expression Data***. Utilizing variancePartition (Hoffman and Schadt 2016), a linear mixed model was employed to quantify variation in each expression trait attribute. Variation within gestational age (GA), sequencing lane, sex, reported race, and birth weight was examined. Variation in placenta expression for maternal clinical data, including parity, gravidity, pre-pregnancy body mass index (BMI), and maternal age, were also examined (Additional Figure 4). We did not run variancePartition for the first trimester placentas as we lack clinical data for this sample set. We additionally examined sex differences for clinical information for full-term placentas for maternal age at delivery, pre-pregnancy BMI, gravidity and parity, gestational age, method of conception, self-reported race, and birth weight. Sex differences for continuous variables were tested using a t-test, p-value < 0.05 (Additional Table 4 & Additional Figure 5).

***X and Y Gametolog Gene Expression.*** A list of X and Y gametology genes were curated from a combination of Skaletsy et al. 2003 and Godfrey et al. 2020 (Godfrey et al. 2020a; Skaletsky et al. 2003) (Additional Table 5). In samples determined to have a Y chromosome, the CPM value of the X-linked gametolog and the Y-linked gametolog were summed and included in a single value under the X-linked gametolog label. Then

we compared expression between XX female X-linked gametology gene expression to XY male X-linked plus Y-linked gametology gene expression using a Wilcox rank-sum, p-value < 0.05 (Additional Table 5).

***Differential Expression.*** Sex differential expression analysis between male XY and female XX placentas was performed using the limma/voom pipeline (Law et al. 2014). Quantified read counts from each sample generated from the SubRead featureCounts were combined into a count matrix, with each row representing a unique gene id, and each column representing the gene counts for each unique sample. Using the DGEList function in the limma package the counts matrix and a tab-delimited file containing sample ID, sex, race, batch, lane, GA, parity, maternal age, gravidity, pre-pregnancy BMI, birth weight, PC1 and PC2 from the whole exome data were read into R (Additional Table 1). Technical replicates from within a placenta were summed together using sumTechReps function in version 3.14.0 (Robinson, McCarthy, and Smyth 2010). Normalization factors were then calculated using the calcNormFactors function in EdgeR (Robinson, McCarthy, and Smyth 2010). After normalization for library and effective gene length, we filter out lowly expressed genes. A minimum of 1 Fragments Per Kilobase Million (FPKM) in at least one group being compared, was required for the gene to be kept for downstream analysis. Then we run Trimmed Means Method (TMM) to account for library size variation between samples (Robinson and Oshlack 2010). Counts were then transformed to $\log_2(CPM+0.25/L)$, where CPM is counts per million, L is library size, and 0.25 is a prior count to avoid taking the log of zero (Law et al. 2014d). For each comparison of interest, a model was created to compare between the groups where each coefficient corresponds to a group mean. The model matrix with batch and

lane was generated before using voom, because voom uses variances of the model residuals (Law et al. 2014e). The model matrix for the term placentas included batch, birth weight, lane, PC1 and PC2. There were no covariances to add to the model matrix for the first trimester placentas. For each differential expression analysis, a linear model was fitted to the DGEList-object, which contained the normalization factors for each gene count for each sample, using the limma lmfit function which will fit a separate model to the expression values for each gene (Law et al. 2014f). Comparisons between groups were then obtained as contrasts of the fitted linear model. An empirical Bayes approach was applied to smooth the standard errors. Genes are defined as being differentially expressed between groups when the adjusted p-value is $\leq 0.05$ using a Benjamini-Hochberg false discovery rate (Law et al. 2014g) (Figure 1).

***Quantifying Sex Differences for Innate Immune Gene Expression.*** Differential expression between male XY and female XX placentas for 979 innate immune genes, as defined by InnateDB (Breuer et al. 2012) (Additional Table 6), was performed using the limma/voom pipeline (Law et al. 2014h). We repeated this analysis for both the term, $\geq$ 36.6 weeks, and late first trimester, 11.5 - 13.5 weeks, placentas. The model matrix for term placentas was the same when looking at the differential expression for the whole transcriptome. Only genes determined to be expressed in at least one sex were included in the analysis (see Methods). In addition to differential expression, we generated an MDS plot on the expression data for only the innate immune genes to determine expression similarity among samples (Figure 2).

***Gene Function and Enrichment Network Analysis.*** We examined differences and similarities in gene enrichment terms between the sex differentially expressed genes

obtained from the differential expression analyses of the samples from the term uncomplicated pregnancies and the samples from the late first trimester (Gonzalez et al. 2018) (Additional Table 7). We used the GOrilla webtool, which utilizes a hypergeometric distribution to identify enriched GO terms (Eden et al. 2009, 2007), with an adjusted Fisher exact p-value cutoff < 0.05 to select significantly enriched terms. Additionally, we looked at each sex differentially expressed gene from the first trimester and term placenta comparisons using genecards.org and a literature review of genome-wide association studies and expression quantitative trait loci (eQTL) to investigate if that gene is involved in known diseases or disorders, particularly with known pregnancy complications (Additional Table 7).

***Sex Differences in Adult GTEx Tissues.*** To determine if sex differences in gene expression within the placenta are correlated with sex differences in adult tissues, we computed the coefficient of correlation, *r*, of the $\log_2$ female-to-male expression ratios for sex differentially expressed genes found in the placenta to 42 non-reproductive adult Genotype-Tissue Expression (GTEx) tissues (Carithers et al. 2015) (Figure 3). For each of the 42 non-reproductive adult GTEx tissues, we computed the $\log_2$ female-to-male expression ratios from the reported Transcripts Per Kilobase Million (TPM) counts version 2017-06-06_v8(Carithers et al. 2015) (Additional Table 8). For each sex differentially expressed gene and all genes expressed in the placenta, we computed the correlation between the placenta $\log_2$ female-to-male expression ratios to each GTEx tissue (Figure 3).

Data processing pipeline available on GitHub, https://github.com/SexChrLab/Placenta_Sex_Diff.

## Results

***Multidimensional Scaling Reveals Outlier Samples.*** We identified outlier samples to remove from downstream analyzes using a modified Multidimensional Scaling plot for RNA and Principal component analysis for DNA. Multidimensional Scaling (MDS) analysis was accomplished for the term placentas and first trimester placentas, to determine if samples cluster by genetic sex. MDS of the term placentas show that the first dimension (dim) is explained by genetic sex. One female XX placenta clustered with the male XY placentas and was removed from the downstream analysis (Additional Figure 2 & Additional Table 3).

***Population Ancestry Inferred From Whole Exome Data.*** Principal component analysis of the term placenta whole exome data shows the samples separated by reported inferred ancestry, in many cases by not all (Additional Table 1). The self-reported race and ethnicity for the placenta samples included 14 Asian, 20 Black, 4 Hispanic, 20 White, and 2 unknown (Additional Table 1). The ancestry prediction estimates that the population ancestry of the samples is: 14 Asain (5 South Asain, 3 East Asain, 1 European, and 2 unknown), 20 Black (17 African, 3 unknown), 4 Hispanic (1 European, 3 unknown), 20 white (15 European, 1 American, 1 South Asain, 1 African, 2 unknown) and 2 unknown (1 American, 1 South Asain). To account for population ancestry differences among the samples, PC1, and PC2 from the whole exome data was included as covariances in the model for sex differential expression analysis (see Methods).

***Clinical Data Shows Little Difference Between the Sexes.*** There is no observable difference between the sexes for clinical data. Birth weight showed some differences between the sexes, with a female mean of 3318 grams and male mean of 3593 grams (t-

test p-value =0.056; Additional Figure 5 & Additional Table 4). There was no significant difference in maternal age or pre-pregnancy body mass index (BMI) for women who carried a male XY versus women who carried a female XX (t-test p-values = 0.39, and 0.73, respectively; Additional Figure 5 & Additional Table 4). Maternal age at delivery ranged from age 22 to 45 years old, and pre-pregnancy BMI ranged from 19.40 to 66.30 (Additional Table 1). Gravidity, the number of pregnancies, and parity, the number of pregnancies reaching higher than 20 weeks did not show a significant difference in women that carried a male XY or female XX pregnancy (t-test p-value 0.43 and 0.61, respectively; Additional Figure 5 & Additional Table 4). Gravidity ranged from 1 to 9, and parity ranged from 0 to 4; the current pregnancy at the moment the data was collected was not included in the parity counts. Gestational age ranged from 36.6 to 41.1 weeks, with no difference between women who carried a male XY versus a female XX pregnancy (t-test p-value = 0.90). Nearly all of the term placentas collected for this study were spontaneous methods of conception, except one male placenta collected from an in vitro fertilization pregnancy and one female placenta collected from intrauterine insemination (Additional Table 1). Only birth weight between the male XX and female XX showed a slight difference (t-test p-value =0.056; Additional Figure 5 & Additional Table 4). Overall, we observe little sex difference in clinical characteristics among the samples collected for this study.

***Variation in the Data and Biological Characteristics Identified.*** Multiple sources of biological and technical variation in the term placenta transcriptome expression were examined. Variance Partitioner was performed for the term placenta RNAseq samples (see methods; Additional Figure 4). We include batch and lane as

covariance in the model for the placenta differential expression analysis. We do not observe other clinical characteristics driving variance between samples with the exception of birth weight which shows some difference between the sexes (Additional Figure 5 & Additional Table 4). Thus, we report the results from the model that includes batch, lane, PC1, PC2, and birth weight.

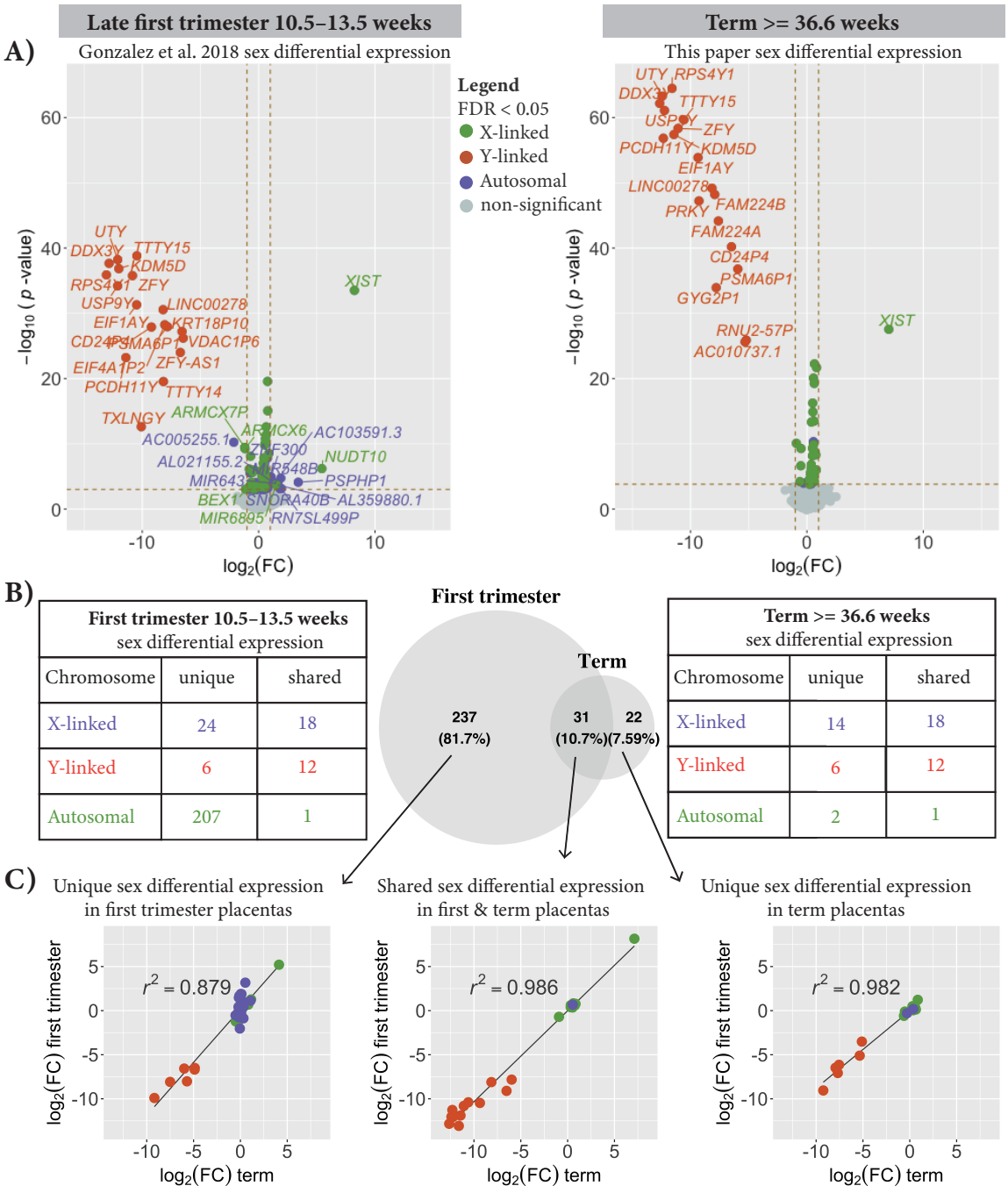
***Sex Differential Expression From Male XY and Female XX Term Uncomplicated Human Placentas.*** We observe 14,441 genes expressed with an FPKM > 1 in at least all the male XY or all female XX placenta samples (Additional Table 9). 53 genes are sex differentially expressed, with an adjusted p-value < 0.05 (Figure 1 & Additional Table 7). Thirty genes showed higher expression in the female XX placentas, and 23 genes showed higher expression in the male XY placentas. Of the 30 genes that are more highly expressed in the female XX placentas than the male placentas, these genes included 28 X-linked and two autosomal genes (*EIF2S3B*, and *EIF1AXP1*). Eighteen of the genes that are more highly expressed in the male XY placentas are Y-linked, four are X-linked (*CD99, RPS6KA6, VDAC1P1, VAMP7*), and one is autosomal (*PRKCE*) (Figure 1 & Additional Table 7).

***Sex Differential Expression Within First Trimester Placentas.*** There are 13,502 genes expressed with an FPKM > 1 in at least the male XY or female XX late first trimester placenta Gonzalez et al. 2018 samples that were reprocessed using new tools (see Methods) (Gonzalez et al. 2018). We identified 268 genes with a significant differential expression between male and female placentas with an adjusted p-value < 0.05 (Figure 1 & Additional Table 7). One hundred eighty genes showed higher expression in the female XX placentas, and 88 genes showed higher expression in the

male XY placentas. Of the 268 sex differentially expressed genes observed in late first trimester placentas, 208 are located on the autosomes (1-22), 42 are X-linked and 18 are Y-linked. Gonzalez et al. 2018 reported 58 genes to be sex differentially expressed in the late first trimester placentas (Gonzalez et al. 2018). Of those 58 genes, 45 are also called as sex differentially expressed in the re-processing of the data using new tools (Additional Table 10).

***Sex Differential Expression Shared Between First Trimester and Term Placentas.*** We find more sex differences in gene expression in late first trimester placentas compared to term placentas. The first trimester (Gonzalez et al. 2018) and term placenta RNAseq samples were processed using the same tools and only differ in the min length for trimming and the covariant added to the linear model for computing sex differential expression (see Methods). Of the 268 genes that are sex differentially expressed in the first trimester placentas, 31 or 10.7% are shared with the genes identified sex differentially expressed in the term placentas (Figure 1). Although there are more genes called as sex differentially expressed in the late first trimester placentas given our adjusted p-value threshold < 0.05, the $\log_2$ female-to-male expression ratio for these genes is highly correlated between first trimester and term placentas (Figure 1C). For the 237 genes that were uniquely called sex differentially expressed in the first trimester placentas, the correlation of coefficients, *r*, for the $\log_2$ female-to-male expression ratio between first trimester and term placentas is 0.879. There are 31 genes that are called sex differentially expressed in both the first trimester and the term placentas. The *r* for the $\log_2$ female-to-male expression ratio between the first trimester and term placentas for the 31 genes shared between placenta datasets is $r = 0.986$. Twenty-two genes are uniquely

called sex differentially expressed in the term placentas compared to the first trimester placentas; again, the $r$ for the $\log_2$ female-to-male expression ratio between the first trimester and term placentas for these genes is positive ($r = 0.982$). Although we observe more sex differential expression in late first trimester placentas, we also observe a high correlation in the $\log_2$ female-to-male expression ratio between first trimester (Gonzalez et al. 2018) and term placentas (Figure 1C).




*Chapter 2. Figure 1. Sex Differential Gene Expression in the Late First Trimester and Term Placentas*. Sex differences in gene expression, $\log_2$(CPM + 0.25/L), between 17 female and 22 male late first trimester (10.5 - 13.5 weeks) placentas on the left and 27 female and 23 male term (>= 36.6 weeks) placentas shown on the right (A). Each point represents a gene. Genes that are sex differentially expressed, adjusted p-value < 0.05, are indicated in purple for autosomal, orange for Y-linked, and green for X-linked. The number of unique sex differentially expressed genes and shared between the late first

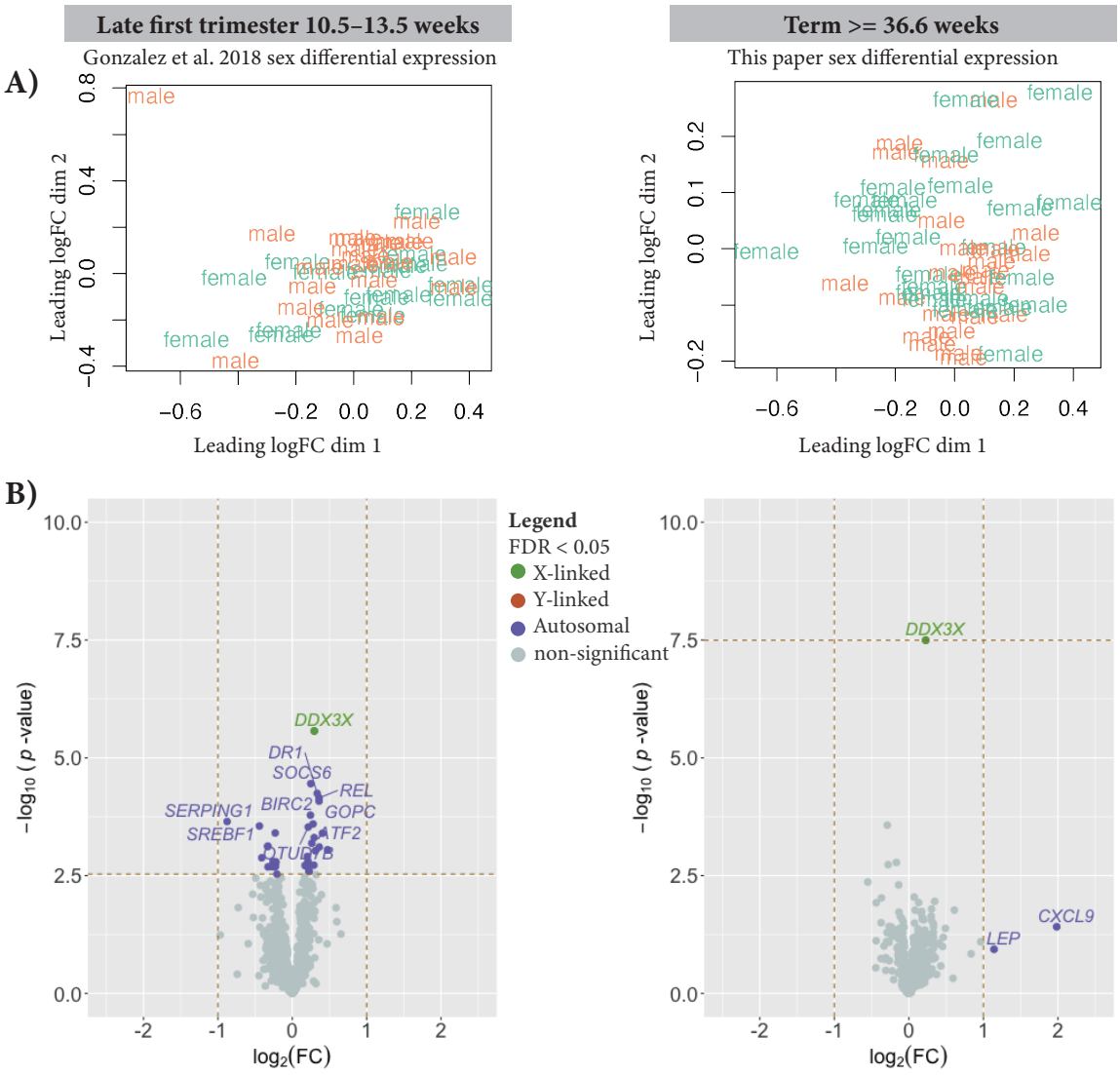
trimester and the term placentas is shown in (B). More genes are sex differentially expressed in the later first trimester (10.5 - 13.5 weeks) than in the term (>= 36.6 weeks) placentas. There are 237 genes that are uniquely called as sex differentially expressed in the late first trimester placentas that are not called as sex differentially expressed in the term placentas; however, the $\log_2$ female-to-male expression ratio for those genes are highly correlated between the later first trimester (Y-axis) and the term (X-axis) placentas $r^2 = 0.879$ (C). There is also a high correlation for the 31 sex differentially expressed genes that are called in both the late first trimester placentas and the term placentas, $r^2 = 0.986$, and for the 22 genes uniquely called in the term placentas $r^2 = 0.982$.

***Gene Enrichment of Sex Differentially Expressed Genes in the Human Placenta are Driven by Sex-linked Genes.*** We investigated gene ontology enrichment for genes that were identified as showing overexpression in one sex versus the other sex. We did this for placentas collected from term uncomplicated pregnancies and from the late first trimester (Gonzalez et al. 2018) (adjusted p-value < 0.05) (Additional Table 7). Genes upregulated in male XY term placentas are involved in histone lysine and protein demethylation processes and histone demethylase activity, driven by Y-linked genes, including *UTY* and *KDM5D*. Genes upregulated in female XX term placentas are involved in translational initiation and regulation of sister chromatid cohesion, driven mainly by X-linked genes including *RPS4X, DDX3X, NAA10* and *HDAC8*. In the late first trimester placentas, genes up-regulated in male placentas are involved in positive regulation of anoikis. Genes up-regulated in the female late first trimester placentas are involved in organonitrogen compound catabolic, branched-chain amino acid catabolic, and positive regulation of protein K63-linked ubiquitination processes. Additionally, the gene function for each sex differentially expressed gene was looked up using genecards.org and a literature review to investigate if that gene is involved in known pregnancy complications (Additional Table 7).

***Lack of Sex Differences in Expression of Immune and Immune Modulator Genes.*** Expression of innate immune genes shows little to no difference in expression between the sexes in placentas from term uncomplicated pregnancies (Figure 2). To examine sex differences in immune expression among placentas, we obtained a list of 979 innate immune genes from InnateDB, a publicly available database of the genes, proteins, experimentally-verified interactions, and signaling pathways involved in the innate immune response of humans (Breuer et al. 2012) (Additional Table 6). Of the 979 innate immune genes reported from InnateDB, 628 are expressed in the term placentas. Unlike the MDS of all genes that show a clear cluster by sex (Additional Figure 2), the MDS of only the innate immune genes shows no distinguishable pattern (Figure 2A). Furthermore, sex differential expression of the 628 innate immune genes in term placentas, only *DDX3X* shows a difference in expression (adjusted p-value < 0.05) (Figure 2). However, *DDX3X* is a gametologous gene with a Y-linked copy, *DDX3Y*. To further investigate this, for samples determined to have a Y chromosome, the count value of the X-linked gametolog and the Y-linked gametolog were summed, then we re-ran the differential expression analysis. After summing expression of the gametologs in males, *DDX3X* still shows a difference in expression between male XY and female XX placentas, but the fold change decreases from a $\log_2$ female-to-male expression ratio of 0.55 and to 0.23 (Additional Table 5 & Additional Figure 6). Although not significantly different in expression (adjusted p-value > 0.05), *CXCL9* and *LEP* show a 3.99 and 2.27, respectively, fold change higher expression in female XX compared to males XY term uncomplicated placentas. With the exception of *DDX3X*, no innate immune genes show a

difference in expression between the sexes for term uncomplicated placentas (Figure 2 & Additional Table 6).

Unlike the term placentas that only show *DDX3X* as sex differentially expressed for innate immune genes, the late first trimester placentas show 37 innate immune genes as differentially expressed between the sexes (adjusted p-value < 0.05) (Additional Table 6). Of the 979 innate immune genes reported from InnateDB, 626 are expressed in the late first trimester placentas (see Methods). There is no clear clustering of the samples for an MDS plot of the first trimester placentas samples when only examining innate immune genes (Figure 2A). Like the term placentas, *DDX3X* also shows a difference in expression between female and male late first trimester placentas (adjusted p-value < 0.05) and shows fold change of 1.23, even after summing the X and Y-linked expression for male samples (Figure 2B & Additional Table 5 & 6). Additionally, *SERPING1* also shows a fold change in expression greater than one, $\log_2$FC 0.88 or 1.84 fold change, showing greater expression male XY than in female XX placentas (Figure 2B). Overall, we observe a lack of sex differences in expression for innate immune genes among term uncomplicated placentas but observe 37 innate immune genes to be sex differentially expressed in the late first trimester placentas (Figure 2 & Additional Table 6).




*Chapter 2. Figure 2. Sex Differences in Gene Expression for Innate Immune Genes.* Of the 979 innate immune genes from InnateDB, 625 genes are expressed in the late first trimester placentas, and 628 are expressed in the term placentas. (A) MDS plot shows no clustering by genetic sex in either the late first trimester (left) or term (right) placentas. (B) Volcano plot of the sex differential expression for late first trimester placentas (left) and term placentas (right). Each point represents a gene. Genes that are sex differentially expressed, adjusted p-value < 0.05, are indicated in purple for autosomal, orange for Y-linked, and green for X-linked.

### *Female-to-Male Gene Expression Ratios in the Placenta are Correlated with Adult Tissues.*

Sex differences in gene expression in the human placenta are correlated

with adult tissues. The correlation of the $\log_2$ female-to-male expression ratios for 243 sex differentially expressed genes found in the placenta (late first trimester or term) adjusted p-value < 0.05 (Figure 1 & Additional Table 7) independently to 42 non-reproductive adult GTEx tissues (Carithers et al. 2015) ranged from an *r* of 0.892 to 0.982 (Figure 3). There are 290 genes sex differentially expressed in the placenta (late first trimester or term) adjusted p-value < 0.05 (Figure 1 & Additional Table 7), but we only have gene TPM count data from GTEx tissues for 243 of the 290 genes (Additional Table 8). We include sex differentially expressed found in either late first trimester or term placentas as the female-to-male expression ratio for these genes were already identified to be highly correlated among placentas (Figure 1C). For the 243 sex differentially expressed genes found in the placenta (late first trimester or term) adjusted p-value < 0.05 and has count data from GTEx, the tissue with the lowest correlation between term placentas and adult tissues is the minor salivary gland with an $r = 0.892$ and p-value < 0.01 (Figure 3). The adult tissue with the highest correlation to that of term placentas in the $\log_2$ female-to-male expression is the frontal brain cortex with an *r* of 0.957 and p-value < 0.01 (Figure 3). The $\log_2$ female-to-male expression ratio correlation for placenta sex differentially expressed genes between term placentas and adult tissue brain regions ranged from *r* of 0.923 to 0.957, p-value < 0.01 (Figure 3). We observe a positive correlation in the $\log_2$ female-to-male expression between term placentas and independently to 42 non-reproductive adult GTEx tissues for genes identified to be sex differentially expressed in the placenta (Figure 1).

The high *r* of the $\log_2$ female-to-male expression between term placenta to adult tissues for genes found to be sex differentially expressed in the placenta is largely driven

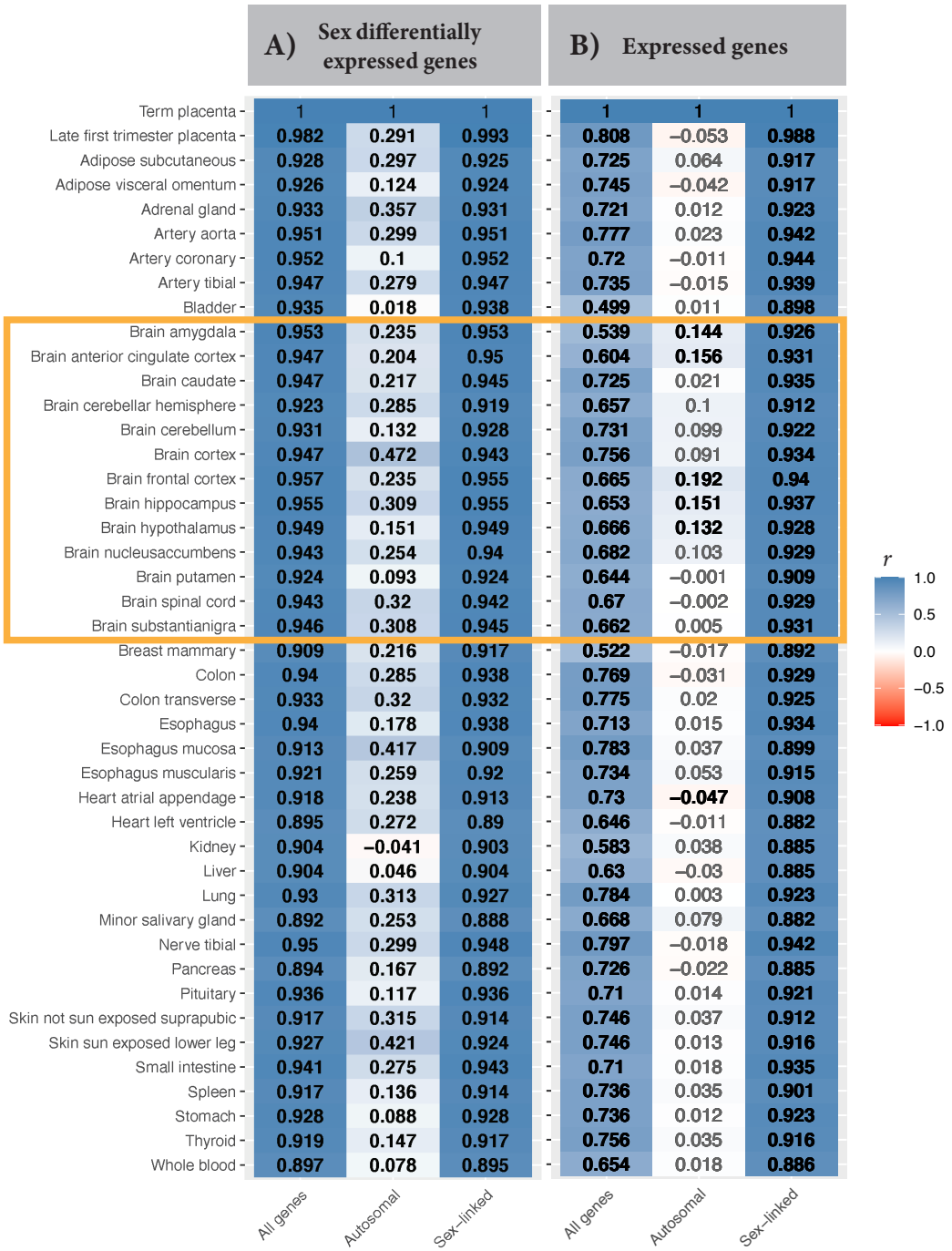
by sex-linked genes (Figure 3). We separated the 243 sex differentially expressed genes found in the placenta (late first trimester or term) adjusted p-value < 0.05 (Figure 1 & Additional Table 7) and included count information for GTEx tissues into autosomal (1-22 & MT) and sex-linked (X or Y-linked) genes. Of the 243 sex differentially expressed genes in the placenta, 182 are autosomal or MT, and the remaining 61 are on the sex chromosomes, X or Y (Additional Table 7). When only looking at the 182 autosomal sex differentially expressed genes, the correlation of the $\log_2$ female-to-male expression ratios decreases between term and late first trimester placentas, and between term placentas independently to 42 non-reproductive adult GTEx tissues (Figure 3). The tissue with the lowest correlation for when only looking at the 182 autosomal sex differentially expressed genes found in the placenta between term placenta and adult tissues is the kidney with an *r* of -.041 p-value < 0.01 (Figure 3). The adult tissue with this highest correlation for looking at the 182 autosomal sex differentially expressed genes is the brain cortex, with an *r* of 0.472 with a p-value < 0.01 (Figure 3). When we repeat this for the 61 sex-linked sex differentially expressed genes, we observe an increase in the correlations in the $\log_2$ female-to-male expression between term placentas and independently to 42 non-reproductive adult GTEx tissues (Figure 3). The tissue with the lowest correlation is the minor salivary gland, *r* of 0.888, p-value < 0.01. The adult tissue with the highest correlation with term placenta is tied between the brain hippocampus and frontal cortex, each with an *r* of 0.955 & p-value < 0.01(Figure 3). In summary, all of the correlations between the term placenta and independently to 42 non-reproductive adult GTEx tissues are all positively correlated when only looking at the 61 sex differentially expressed genes that are on the sex chromosomes, X or Y, compared to looking at the

182 autosomal sex differentially expressed genes that range from negative to positive correlations (Figure 3). The high *r* of the $\log_2$ female-to-male expression between the placenta to adult tissues for all sex differentially expressed genes, autosomal and sex-linked, 243 genes is driven by sex-linked genes (Figure 3). The above results are only for genes that are sex differentially expressed in the placenta (Additional Table 7); next, we investigated the overall $\log_2$ female-to-male expression ratio correlation between tissues for all expressed genes.

The correlation of the $\log_2$ female-to-male expression ratio in term placentas for all genes expressed in both the late first trimester and the term placentas (11,179 genes) independently to 42 adult GTEx tissues shows some correlation but not as strong of a correlation when only looking at sex differentially expressed genes as described above (Figure 3 & Additional Table 8 & 9). The correlation for all expressed genes (11,179 genes) shows a range of *r* of 0.499 to 0.797 between term placentas and independently to 42 adult GTEx tissues (Figure 3). The tissue with the lowest correlation between term placentas and adult GTEx tissues is the bladder with an *r* of 0.499, p-value < 0.01 (Figure 3). The highest correlation between term placenta and adult tissue for looking at all expressed genes is the nerve tibial with an *r* of 0.797, p-value < 0.01 (Figure 3). The correlation of the $\log_2$ female-to-male expression ratio for all placenta expressed genes is highest between term placentas and late first trimester placentas with an *r* of 0.808, p-value < 0.01 (Figure 3). The correlation of the $\log_2$ female-to-male expression ratio for all genes expressed in both the late first trimester and the term placentas (11,179 genes) independently compared to 42 adult GTEx tissues are all positive and are all significant,

p-value < 0.01 (Figure 3). Next, we repeated this analysis by separating the 11,179 genes expressed in the placenta into autosomal and sex-linked.

Sex-linked genes primarily drive the high positive correlations in the $\log_2$ female-to-male expression ratio for all genes expressed in the placenta between term placentas and independently to 42 non-reproductive adult tissues (Figure 3). When we separate the 11,179 expressed genes found in the placenta and have count information from GTEx, into autosomal (1-22 & MT) and sex-linked (X and Y), there are 10,762 autosomal genes and 417 sex-linked genes (Additional Table 8 & 9). When only looking at the 10,762 autosomal genes, the correlation in the $\log_2$ female-to-male expression ratio between term placentas and independently to 42 adult GTEx tissues ranges from an *r* of -0.042 to 0.192 (Figure 3). Even when comparing term placentas to late first trimester placentas, the *r* is only -0.053 for the placenta's autosomal genes (10,762 genes). Interestingly, the $\log_2$ female-to-male expression ratio correlation between term placentas to adult brain regions for only autosomal genes shows some significant positive correlations (Figure 3). The brain amygdala, anterior cingulate cortex, frontal cortex, hippocampus, and hypothalamus all show positive significant, p-value < 0.01, correlations in $\log_2$ female-to-male expression independently to term placentas (Figure 3). When we look at only sex-linked genes (417 genes), the correlations in the $\log_2$ female-to-male expression ratio between term placentas and independently to 42 non-reproductive adult tissues are all positive and significant, ranging from *r* of 0.882 to 0.942, p-value < 0.01 (Figure 3). Thus, sex differences in gene expression for sex-linked genes is correlated between term placenta and adult tissues (Figure 3).




*Chapter 2. Figure 3. Coefficient Correlation, r, in the Log$_2$ Female-to-male Expression Ratios Between Term Placenta, Late First Trimester Placentas, and 42 Non-reproductive Adult GTEx Tissues.* (A) 243 sex differentially expressed genes in the placenta (late first trimester or term) adjusted p-value < 0.05 and contains count information for GTEx

tissues. Of the 243 sex differentially expressed genes, 182 are autosomal and 61 are sex-linked. (B) 11,179 expressed genes in the placenta and contains count information in GTEx includes 10,762 autosomal genes, and 417 that are sex-linked. Black and bold indicates a significant correlation, p-value < 0.01.

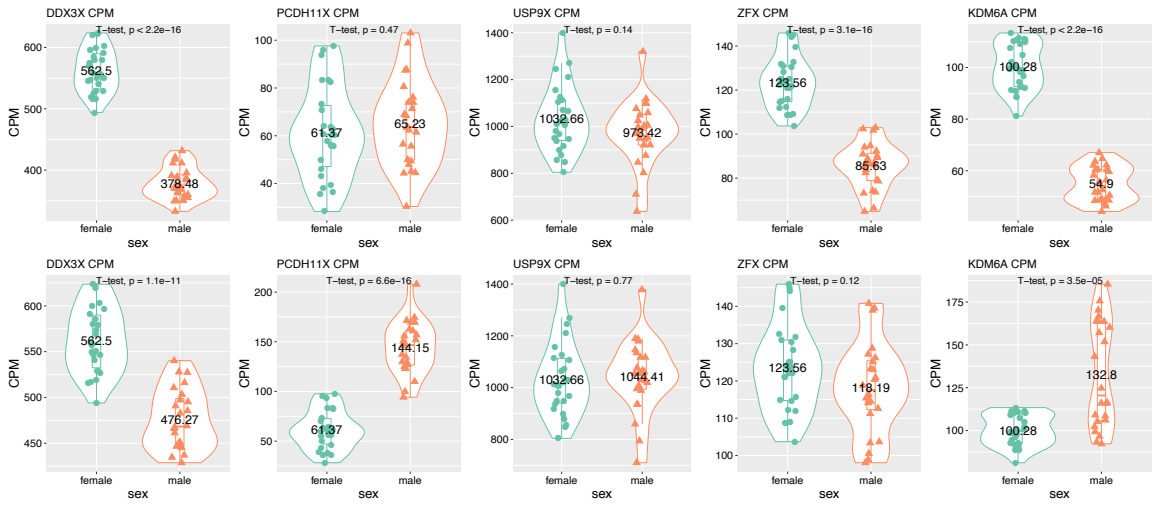
***Sex Differences in Expression for X-linked Gametolog Genes.*** X-linked gametolog genes show a difference in expression between female and male placentas (Figure 4). However, when we consider the evolutionary history of sex-linked genes, the direction of bias either stays the same, is no longer sex differentially expressed, or the direction of bias flips (Figure 4 & Additional Table 5). A list of X and Y gametology genes was curated from Skaletsy et al. 2003 and Godfrey et al. 2020 for a total of 23 gametolog genes (Godfrey et al. 2020b; Skaletsky et al. 2003) (Additional Table 5). Of the 23 gametolog genes, 14 are expressed in the late first trimester and term placenta samples (FPKM > 1 in at least all the male or all the female samples; see Methods) (Additional Table 9).

In the late first trimester placentas, 7 out of the 14 X-linked gametolog genes show higher expression in females compared to males (*DDX3X, ZFX, KDM6A, PRKX, PRS4K, EIF1AX*) (Additional Table 5). When we take the sum expression of the X and Y-linked copy for male samples, we no longer see a sex difference in expression for *ZFX, PRKX,* and *EIF1AX,* Wilcoxon test p-value > 0.05 (Additional Table 5). Three sex differentially expressed genes*, DDX3X, KDM5C, and RPS4X,* continue to show higher expression in female compared to male late first trimester placentas. On the other hand, *KDM6A* flips direction and shows higher expression male compared to female late first-trimester placentas when we take the sum X + Y-linked expression in males, Wilcoxon test p-value < 0.01 (Additional Table 5). Additionally, *PCDH11X* changed from showing

no sex difference in expression to showing higher male expression compared to female late first trimester placentas, Wilcoxon test p-value < 0.01 (Figure 4 & Additional Table 5). Of the 14 X-linked gametolog genes that have expression in uncomplicated late first trimester placentas, 3 continue to show higher expression in female placentas, 3 no longer show a sex difference in expression, 1 gene changes from showing no sex difference to showing higher expression in males, and 1 gene flips direction from showing higher expression in females to now showing higher expression in males (Additional Table 5). In term placentas, we observe a similar pattern when we sum the X and Y-linked expression for males.

The same 7 X-linked gametolog genes that show higher expression in the female late first-trimester placentas also show higher expression in the female term placentas compared to male term placentas (Additional Table 5). When we take the sum expression of the X and Y-linked copy for male term samples, we no longer see a sex difference in expression for *ZFX, PRKX,* and *RPS4X,* Wilcoxon test p-value > 0.05 (Additional Table 5). This differs from the late first trimester placentas that show higher expression in female placentas for *RPS4X* regardless if the analysis is compared to male X-linked expression or male X + Y-linked expression (Additional Table 5). In the term placentas, *RPS4X* is sex differentially expressed, showing higher expression in females, but compared to males X + Y-linked expression; we longer observe a sex difference in expression for this gene, Wilcoxon test p-value > 0.267 (Additional Table 5). Three sex differentially expressed genes*, DDX3X, KDM5C,* and *EIF1AX,* continue to show higher expression in female compared to male term placentas, Wilcoxon test p-value < 0.01 (Additional Table 5). Like in the late first trimester placentas, *KDM6A* flips direction and

shows higher expression male compared to female term placentas when we take the sum X + Y-linked expression in males, Wilcoxon test p-value < 0.01 (Additional Table 5). Again, just like in the late first trimester placentas, *PCDH11X* changed from showing no sex difference in expression to showing higher expression in male compared to female term placentas when we sum the X and Y-linked expression for males, Wilcoxon test p-value < 0.01 (Figure 4 & Additional Table 5). Of the 14 X-linked gametolog genes that have expression in uncomplicated term placentas, 3 continue to show higher expression in female placentas, 3 no longer show a sex difference in expression, 1 changes from showing no sex difference to showing higher expression in males, and 1 gene flips direction from showing higher expression in females to now showing higher expression in males (Additional Table 5). In summary, the directional bias of X-linked gametolog genes may change if the X and Y-linked expression values are summed for male XY samples (Figure 4 & Additional Table 5).




*Chapter 2. Figure 4. Sex Differences in Expression for X-linked Gametolog Genes*. Top row is female X-linked expression compared to male X-linked expression. The bottom row is female X-linked expression compared to male X + Y-linked expression. There is a

significant difference in male XY to female XX expression for *ZFX* and *KDM6A* (*UTX*) when only looking at the X chromosome CPM expression value. When we add the Y chromosome-linked CPM expression count for these genes for male samples, there is no longer a difference in expression between males XY and females XX for *ZFX*. *KDM6A*, on the other hand, flips the bias; it now shows males as having significantly higher expression than females. *PCDH11X*, when adding Y-linked CPM expression, shows a significantly higher expression than females. T-test to see if there is a difference between the female CPM and the male CPM for each gene, p-value < 0.05.

## Discussion

***Sex Differences in Gene Expression in Term Placentas are Replicated Among Tissues.*** We observe a positive correlation in the $\log_2$ female-to-male expression ratio for sex differentially expressed genes between the late first trimester (Gonzalez et al. 2018) and term placentas, and between term placentas and 42 non-reproductive adult GTEx tissues (Figure 3). Previous work has compared first trimester and term placentas and found thousands of genes to be differentially expressed (Sitras et al. 2012). However, the study appeared to not separate the genes based on chromosomal location (Sitras et al. 2012). When we look at all genes located on the sex chromosomes, X & Y, the $\log_2$ female-to-male expression ratio is positively correlated between late first trimester and term placentas, as well as adult tissues (Figure 3). When this is repeated for only autosomal genes, 1-22 and MT, we do not observe the same positive correlation between late first trimester, term placentas, and adult tissues (Figure 3). These findings suggest that sex differences in gene expression for sex-linked genes develop early in embryonic tissue and are replicated in adult tissues. Sex differential expression for autosomal genes may be more tissue-dependent, as previously suggested by Lopes-Ramos et al. 2020. Lopes-Ramos et al. 2020 found that sex differentially expressed genes common among adult tissues were enriched for sex chromosome genes, and sex differences for autosomal

genes were tissue-specific (C. M. Lopes-Ramos et al. 2020). In summary, we observe a positive and significant correlation in the $\log_2$ female-to-male expression ratio for sex-linked genes between term placentas and adult tissues (Figure 3).

***Gene Enrichment of Sexually Dimorphic Genes Reveals Genes that may be Involved in Pregnancy Complications.*** Sex differentially expressed genes may be involved in biological pathways related to pregnancy complications. In term placentas, genes upregulated in XY males are involved in histone lysine and protein demethylation processes and histone demethylase activity, driven by Y-linked genes, including *UTY* and *KDM5D* (Additional Table 7). A review of ruminant placenta gene targeting found histone lysine demethylase 1A and androgen signaling to be involved in gene networks for cell proliferation and angiogenesis (Hord et al. 2020). The authors also note previous studies that have examined exposure to testosterone during pregnancy leading to ovarian dysfunction and low-birth-weight for female offspring, suggesting that increased androgen signaling dysregulates fetal development, at least for female offspring (Hord et al. 2020). Further studies are needed to better understand the role of histone demethylase activity and androgen signaling in sex differences in human pregnancy health and complications. Genes upregulated in XX female term placentas are involved in translational initiation and regulation of sister chromatid cohesion, driven mainly by X-linked genes including: *NAA10*. *NAA10* is involved in the process of post-translational protein modifications and mutations in *NAA10* are known to cause Ogden syndrome which may lead to growth failure (Lee et al. 2017) (Additional Table 7). In a *Naa10* mouse knockout study, the authors report placental insufficiency that contributed to embryonic and neonatal lethality (Lee et al. 2017). Additionally, *Naa10* mouse knockouts

showed low birth weight and postnatal growth failure compared to control mice (Lee et al. 2017). Loss of *NAA10* plays a role in developmental of cardiovascular and growth defects in humans and mice (Lee et al. 2017; Wu and Lyon 2018). In our study, *NAA10* is upregulated in female compared to male term uncomplicated placentas (Additional Table 7). More research is needed to understand if sex differences in expression for *NAA10* are involved in sex differences in development.

Using new tools, we replicated the Gonzalez et al. 2018 late first trimester placenta gene enrichment analysis (Gonzalez et al. 2018). We found genes upregulated in male XY compared to female XX late first trimester placentas are involved in positive regulation of anoikis. Genes upregulated in the female late first trimester placentas are involved in organonitrogen compound catabolic, branched-chain amino acid catabolic, positive regulation of protein K63-linked ubiquitination processes. *NUDT10* was shown to be enriched in these biological processes and is upregulated in female XX late first trimester placentas and was previously reported in the Gonzalez et al. 2018 study as well (Gonzalez et al. 2018). The role of *NUD10* in placenta function remains to be further explored. Additionally, of the 58 genes previously identified as sex differentially expressed in the late first trimester placentas from Gonzalez et al. 2018, we found 45 of those genes to also be sex differentially expressed in the samples using different tools to process the data. Overall, we replicate the findings from Gonzalez et al. 2018 and we identified 210 additional genes to be sex differentially expressed among the late first trimester placentas and we annotate if those genes have been reported in pregnancy complications (Additional Table 7 & 10).

***Lack of Sex Differences in Immune Gene Expression***. The placenta is an immune modulator in the uterine environment interacting with the maternal decidua cells to promote an immunosuppressive environment for maintaining fetal tolerance (PrabhuDas et al. 2015; Xin et al. 2014). The placenta promotes inflammation response with up-regulation of pro-inflammatory cytokines during early implantation (PrabhuDas et al. 2015). For example, placentas from preeclampsia pregnancies have been reported to show lower expression of immune protein *CD74* and enrichment for IL-1-signaling pathway compared to uncomplicated placentas (Przybyl et al. 2016). Maymon et al. 2018 showed that patients with preterm labor showed higher concentrations of the immune protein *CXCR3* and its ligands *CXCL9* and *CXCL10* in amniotic fluid compared to term in labor and term not in labor (Maymon et al. 2018). Karjalainen et al. 2015 similarly found higher expression for *CXCR3* in the preterm cord blood samples compared to term cord blood (Karjalainen et al. 2015). Immune gene expression within the placenta plays a role in maintaining pregnancy to term (PrabhuDas et al. 2015; Xin et al. 2014); we therefore sought to characterize sex differences in immune expression among late first trimester (Gonzalez et al. 2018) and term uncomplicated placentas to expand on previously reported sex differences among uncomplicated placentas such as those reported in Gonzalez et al. 2018 and Sood et al. 2006.

We observe little sex differences for innate immune genes among uncomplicated term placentas. In the term placentas, when looking at sex differences for only innate immune genes, only *DDX3X* showed a difference in expression between the sexes with higher expression in females compared to males, adjusted p-value $< 0.05$ (Figure 2). *DDX3X* is essential in cell cycle control, and loss of *Ddx3x* in male mice resulted in early

post-implantation lethality (Chen et al. 2016). In female mice, inactivation of a paternal *Ddx3x* copy resulted in placental abnormalities and embryonic lethality (Chen et al. 2016). Together with the findings reported here, suggest that expression of *DDX3X* in the placenta may be critical for proper placental development. *DDX3X* may show higher expression in female compared to male placentas because *DDX3X* escapes X chromosome inactivation in female uncomplicated placentas; showing expression for both the maternal and paternal gene copy (Phung et al., n.d.). Although not differentially expressed between the sexes, adjusted p-value > 0.05, *CXCL9* and *LEP* show a fold change higher expression in females compared to male term placentas (Figure 2). In the term placentas, the female mean CPM is 10.38 and the male mean CPM is 3.86 for the immune protein *CXCL9* (Additional Table 6). The observed fold change difference between females and males for *CXCL9* among the uncomplicated term placentas appears to be largely driven by one female sample, OBG0178, with a CPM expression of 160 for *CXCL9* (Additional Figure 7 & Table 6). *CXCL9* is thought to be involved in T cell trafficking (Tokunaga et al. 2018; Ochiai et al. 2015) and the promotion of inflammation within the mater-fetal interface (Nancy and Erlebacher 2014). Maymon et al. 2018 showed that patients with preterm labor showed higher concentrations of *CXCR3* and its ligands *CXCL9* and *CXCL10* in amniotic fluid compared to term (Maymon et al. 2018). Higher expression of *CXCL9* in female compared to male placentas in the samples analyzed here may be reflect sex differences in expression for inflammation response, though further investigation is needed. *LEP* additionally shows a fold change higher expression in female compared to male term uncomplicated placentas (Additional Table 6). *LEP* plays a major role in regulating energy homeostasis; additionally,

hypomethylation of *LEP* in the placenta has been observed in early onset preeclampsia compared to controls (Hogg et al. 2013). There are known sex differences in the incidence of preeclampsia. Term preeclampsia is more common in male-bearing pregnancies compared to female-bearing pregnancies (Vatten and Skjaerven 2004), while preterm preeclampsia is more common in female compared to male-bearing pregnancies (Global Pregnancy Collaboration: et al. 2017; Vatten and Skjaerven 2004). Sex differences in expression for *LEP* may help to explain sex differences in incidence of preeclampsia, but we did not test this and further investigation is needed. Overall, with the exception of *DDX3X,* we observe a lack of sex differences in expression for innate immune genes among uncomplicated term placentas (Figure 2 & Additional Table 6) suggesting expression of these genes may be important for maintaining pregnancy to term.

In the late first trimester placentas (Gonzalez et al. 2018), 37 innate immune genes are differentially expressed between the sexes, adjusted p-value < 0.05, including *DDX3X* (Additional Table 6). Of the 37 innate immune genes differentially expressed between the sexes, adjusted p-value < 0.05, one gene also showed a fold change difference in expression between the sexes, *SERPING1*. *SERPING1* showed 1.84 fold change in higher expression in male XY compared to female XX placentas. *SERPING1* encodes for a highly glycosylated protein and is involved in inhibiting C1r and C1a of the complement component. Epigenetic alterations of genes in the SERPIN superfamily have been described in preeclampsia (Chelbi et al. 2007; Blanch et al. 2003). It has also been suggested that *SERPING1* may be involved in the placental circulatory function, and misregulation of *SERPING1* could lead to placental diseases (Vaiman et al. 2005).

Overall, we observe sex differences in expression for innate immune genes in the late first trimester placentas but not in the term placentas (Figure 2).

***Limitations of the Study.*** All term, late first trimester placentas (Gonzalez et al. 2018), and adult tissues (Carithers et al. 2015) used in this study are from different research groups using different sequencing tools and approaches. Here we study sex differences in term placentas and compare with sex differences in late first trimester placentas (Gonzalez et al. 2018) and adult tissues (Carithers et al. 2015). The late first trimester (Gonzalez et al. 2018) and term placentas were collected at different times using different sequencing approaches; we, therefore, focus on sex differences in each data set separately to study replication of sex differences at different time points.

## Perspectives and Significance

In summary, we find sex differences in gene expression in early developed placenta tissue. However, there is a lack of sex differences in gene expression for innate immune genes among uncomplicated term placentas, suggesting expression of these genes may be involved in sustaining a pregnancy to term. The expression ratio between females and males for sex differentially expressed in term uncomplicated placentas are replicated in adult tissues. Sex differences in gene expression develop early and are observed in adult tissues.

## Supplementary Information

Supplemental tables and figures are located in chapter 2. appendices B.

CHAPTER 3

# The Synthetic Histone-Binding Regulator Protein PcTF Activates Interferon Genes in Breast Cancer Cells

(Previously published Olney, K.C., Nyer, D.B., Vargas, D.A., Wilson, M.A., Haynes, K.A., The synthetic histone-binding regulator protein PcTF activates interferon genes in breast cancer cells. BMC Syst Biol 12, 83 (2018). https://doi.org/10.1186/s12918-018-0608-4)

## ABSTRCT

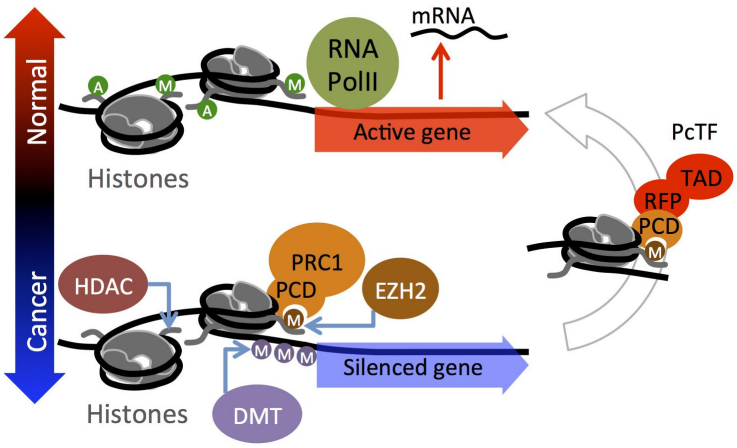
Mounting evidence from genome-wide studies of cancer show that chromatin-mediated epigenetic silencing at large cohorts of genes is strongly linked to a poor prognosis. This mechanism is thought to prevent cell differentiation and enable evasion of the immune system. Drugging the cancer epigenome with small molecule inhibitors to release silenced genes from the repressed state has emerged as a powerful approach for cancer research and drug development. Targets of these inhibitors include chromatin-modifying enzymes that can acquire drug-resistant mutations. In order to directly target a generally conserved feature, elevated trimethyl-lysine 27 on histone H3 (H3K27me3), we developed the Polycomb-based Transcription Factor (PcTF), a fusion activator that targets methyl-histone marks via its N-terminal H3K27me3-binding motif, and co-regulates sets of silenced genes. Here, we report transcriptome profiling analyses of PcTF-treated breast cancer model cell lines. We identified a set of 19 PcTF-upregulated genes, or PUGs, that were consistent across three distinct breast cancer cell lines. These genes associated with the interferon response pathway. Our results demonstrate for the

first time a chromatin-mediated interferon-related transcriptional response driven by an engineered fusion protein that physically links repressive histone marks with active transcription.

## Background

In addition to DNA lesions, disruption of chromatin at non-mutated genes can support the progression of cancer. Chromatin is a dynamic network of interacting proteins, DNA, and RNA that organizes chromosomes within cell nuclei. These interactions regulate gene transcription and coordinate distinct, genome-wide expression profiles in different cell types. Chromatin mediates epigenetic inheritance (Margueron & Reinberg, 2010; Richards & Elgin, 2002) by regulating expression states that persist through cellular mitosis and across generations of sexually reproducing organisms (Cavalli & Paro, 1998; Roemer et al., 1997). Posttranslational modifications (PTMs) of histones within nucleosomes, the fundamental subunits of chromatin, play a central role in the epigenetic regulation of genes that control cell differentiation (Kim & Orkin, 2011; Sparmann & van Lohuizen, 2006). Several landmark studies have revealed that hyperactivity of the histone-methyltransferase enhancer of zeste 1 and 2 (EZH1, EZH2), which generates the histone PTM H3K27me3, is a feature shared by many types of cancer (recently reviewed in (Wang et al., 2015)). In breast cancer, elevated EZH2 has been linked to cell proliferation and metastasis (Alford et al., 2012; Chang et al., 2011) and a poor prognosis for breast cancer patients (Collett et al., 2006; Kleer et al., 2003; Niida et al., 2009; Peña-Llopis et al., 2016). In stem cells and cancer cells, EZH2 generates H3K27me3 mark at nucleosomes (Fig. 1) near the promoters of developmental genes, represses transcription, and thus prevents differentiation to support the proliferative state in stem cells or neoplasia in cancer (reviewed in (Kim & Orkin, 2011)). Polycomb Repressive Complex 1 (PRC1, also known as PRC1.2 or PRC1.4(Gao et al., 2012)) binds to the H3K27me3 mark through the polycomb chromodomain (PCD) motif

of the CBX protein to stabilize the repressed state. Silencing is reinforced by other chromatin regulators including histone deacetylase (HDAC) and DNA methyltransferase (DMT) (Easwaran et al., 2012) (Fig. 1).




*Chapter 3. Figure 1. Reversal of a Cancer-associated Epigenetic State Via the PcTF Fusion Protein.* The lower half of the cartoon depicts the accumulation of repressive chromatin at a developmental gene. EZH2 generates H3K27me3, which is recognized by the PCD fold in the CBX protein of Polycomb Repressive Complex 1 (PRC1). Silencing is re-enforced by histone deacetylase (HDAC), and DNA methyltransferase (DMT) activity. The fusion protein PcTF contains an N-terminal PCD fold (cloned from CBX8) that binds H3K27me3 and stimulates transcription via its C-terminal activator domain to restore the active state (right side of the cartoon). A, acetylation; M, methylation; green circle, activation-associated PTM; orange or purple circle, repression-associated PTM; RFP, red fluorescent protein tag; TAD, transcriptional activation domain VP64.

The PRC module is a group of genes that is regulated by H3K27me3 and Polycomb transcriptional regulators (Bracken & Helin, 2009; Jene-Sanz et al., 2013). Relatively high expression or upregulation of PRC module genes is associated with a non-proliferative state, cell adhesion, organ development, and normal anatomical structure morphogenesis (Jene-Sanz et al., 2013). Knockdown (depletion) of chromatin proteins (reviewed in (Bracken & Helin, 2009; Dawson & Kouzarides, 2012)) and

inhibition of Polycomb proteins with low molecular weight compounds (Simhadri et al., 2014a; Stuckey et al., 2016a; Tabet et al., 2013a) and peptides (Simhadri et al., 2014b; Stuckey et al., 2016b; Tabet et al., 2013) stimulates expression of developmental genes and perturbs cancer-associated cell behavior. The interferon (IFN) pathway is often highly represented among silenced genes in cancer. IFN gene activity has been linked to apoptosis (Bouker et al., 2005; J. Lee et al., 2006) and triggers the body's immune system to attack cancer cells (Dunn & Rao, 2017; Ikeda et al., 2002). Decreased expression and increased levels of repressive epigenetic marks (e.g., DNA methylation) have been detected at IFN genes in Li–Fraumeni fibroblasts (39 of 85 silenced genes) (Kulaeva et al., 2003), colon carcinomas (McGough et al., 2008), and triple negative breast cancers (Teschendorff et al., 2007a; H. Xu et al., 2014a). Transgenic overexpression of *IFN1* in MCF7 breast cancer xenografts perturbs tumor growth in nude mice (Bouker et al., 2005). Treatment of cancerous cells with broad-acting epigenetic inhibitors of DNA methyltransferase (DNMTi) and histone deacetylase (HDACi) leads to activation of IFN genes which induces an arrest of cancer cell proliferation or sensitize cancer cells to immunotherapy (Dunn & Rao, 2017; Li et al., 2014; Stone et al., 2017).

The use of the FDA-approved DNA methyltransferase inhibitors (e.g., 5-azacytidine) to treat cancer, as well as the success of other epigenetic interventions in clinical trials (Biancotto et al., 2010; Mani & Herceg, 2010) demonstrates that chromatin is a druggable target in cancer. Certain limitations of epigenetic inhibitor compounds could encumber complete efficacy of epigenetic therapy. Inhibitors do not interact directly with modified histones, indirectly activate silenced genes by blocking repressors, generate incomplete conversion of silenced chromatin into active chromatin (McGarvey

et al., 2006, 2007), interact with off-target proteins outside of the nucleus (Su et al., 2005), and do not affect resistant Polycomb protein mutants (Fujiwara et al., 2014; Ueda et al., 2014; B. Xu et al., 2015). These limitations could be addressed by technologies that directly target H3K27me3 within the chromatin fiber. H3K27me3 is a highly conserved feature in cancers (Wang et al., 2015). Even in cases where H3K27 becomes mutated to methionine in one allele (Schwartzentruber et al., 2012; Wu et al., 2012), methylation of the wild-type copy of H3K27 is still present at repressed loci in cancer cells(K.-M. Chan et al., 2013; K. M. Chan et al., 2013).

Our group developed a fusion protein called Polycomb-based Transcription Factor (PcTF), which specifically binds H3K27me3 (Tekel et al., 2017) and recruits endogenous transcription factors to PRC-silenced genes (Fig. 1). In bone, brain, and blood-cancer derived cell lines, PcTF expression stimulates transcriptional activation of several anti-oncogenesis genes (Nyer et al., 2017). PcTF-mediated activation leads to the eventual loss of the silencing mark H3K27me3 and elevation of the active mark H3K4me3 at the tumor suppressor locus *CASZ1*.

To explore the therapeutic potential of fusion protein-mediated epigenetic interventions, we sought to investigate the behavior of PcTF in breast cancer cells lines that have been established as models for tumorigenesis (Goodspeed et al., 2016; Lacroix & Leclercq, 2004; Neve et al., 2006). Here, we extend our investigation of PcTF activity to three breast cancer-relevant cell lines. First, we investigated the transcription profiles of predicted PRC module genes in drug-responsive (MCF-7, BT-474) and unresponsive triple negative (BT-549) breast cancer cell lines. Receptor-negative BT-549 cells have a transcription profile and histology similar to aggressive tumor cells from patient samples

(Lehmann et al., 2011; Tseng et al., 2017). Overexpression of PcTF in transfected breast cancer cells led to the upregulation of dozens of genes, including a common set of 19 genes in the interferon response pathway, as early as 24 hours after transfection. The transcriptome of BT-549 (triple-negative) showed the highest degree of PcTF-sensitivity. We observed that PcTF-sensitive genes are associated with a bivalent chromatin environment and moderate levels of basal transcription. Interestingly, these PcTF-sensitive genes do not overlap with very strongly repressed, PRC-enriched loci. This discovery provides new mechanistic insights into the state of genes that are poised for transcriptional activation via PcTF.

## Results

***Differential Regulation of Genes in Breast Cancer Cell Lines.*** To determine expression levels of predicted PRC module genes, we profiled the transcriptomes of three breast cancer cell lines and the non-invasive, basal B cell line MCF10A (Kenny et al., 2007; Nagaraja et al., 2006) using next-generation deep sequencing of total RNA (RNA-seq). MCF7, BT-474, and BT-549 represent luminal A, luminal B, and basal B subtypes of breast cancer, respectively (Table 1) (Neve et al., 2006). Previous studies have shown that gene expression profiles distinguish two major categories of cancer cell lines, luminal and basal, in patient-derived samples (T. Sorlie et al., 2001; Therese Sorlie et al., 2003). The basal class exhibits a stem-cell like expression profile (Ben-Porath et al., 2008), which is consistent with high levels of Polycomb-mediated repression at genes involved in development and differentiation (Boyer et al., 2006; T. I. Lee et al., 2006). Levels of the repressor protein EZH2 and the histone modification that it generates

(H3K27me3) are elevated in MCF7, BT-474, and BT-549 compared to non-metastatic cells such as MCF10A (Table 1). A mechanistic link between Polycomb-mediated repression and tumor aggressiveness has been supported by a study where stimulation of the phosphoinositide 3-kinase (PI3K) signaling pathway, which induces a metastatic phenotype in MCF10A, is accompanied by increased H3K27me3 at several target genes (Lin et al., 2008; Zuo et al., 2011). We hypothesized that known Polycomb-repressed genes (the PRC module) would be down-regulated in the cancerous cell lines compared to MCF10A.

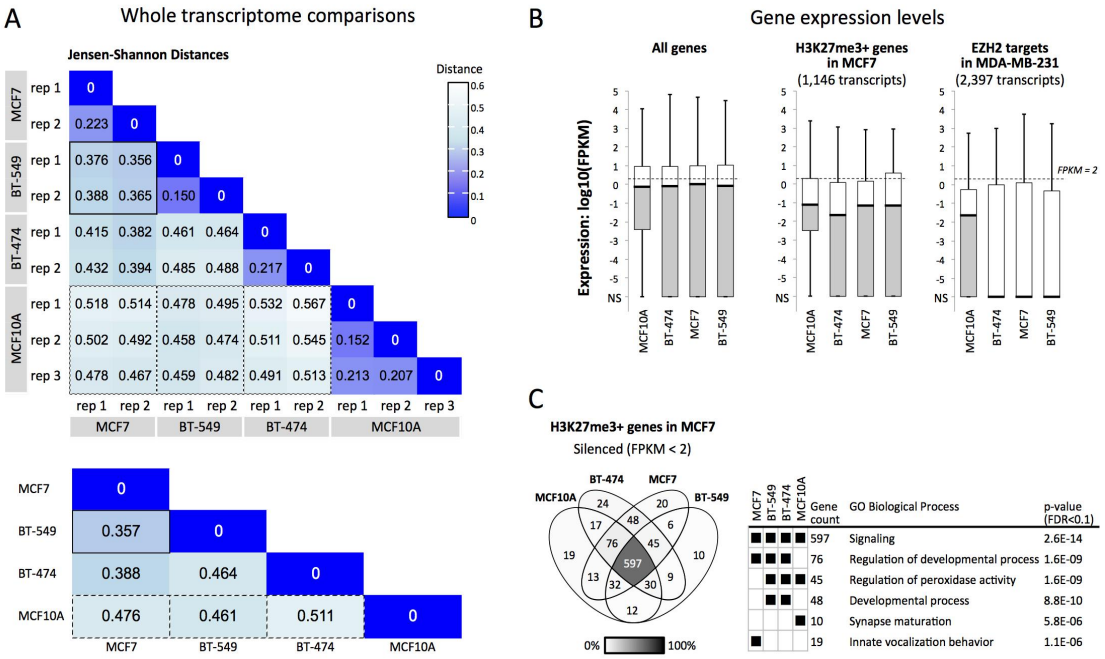
*Chapter 3. Table 1. Descriptions of the Breast Tissue-derived Cell Lines Used in this Study.* ATCC = American Tissue Culture Center ID. Molecular subtype and marker expression status are from Neve et. al 2006 (Neve et al., 2006): Estrogen receptor presence or absence (ER+/-), Progesterone receptor presence or absence (PR+/-), HER2 overexpression (HER2+), and TP53 mutation ($TP53^{M}$). EZH2 and H3K27me3 were shown to be elevated compared to non-metastatic fibroblasts (a) (Leroy et al., 2013), LNCaP (b) (Ren et al., 2012), MCF10A (c) (Chang et al., 2011; Derfoul et al., 2011; Dong et al., 2014), and HMEC (d).

| Cell line | ATCC | Sub-type | Markers (Neve et al., 2006) | EZH2 | H3K27me3 |
|---|---|---|---|---|---|
| MCF7 | HTB-22 | Luminal A | ER+, PR+ | Elevated[a,b,c] (Derfoul et al., 2011; Leroy et al., 2013; Ren et al., 2012) | Elevated[a] (Leroy et al., 2013; Zuo et al., 2011) |
| BT-474 | HTB-20 | Luminal B | ER+, PR+, HER2+ | Elevated[c] (Dong et al., 2014) | Elevated[d] (Zuo et al., 2011) |

| | | | | | |
|---|---|---|---|---|---|
| BT-549 | HTB-122 | Basal B, claudin-low | ER-, PR-, *TP53*$^{M}$ | Elevated[c] (Chang et al., 2011) | Elevated[d] (Zuo et al., 2011) |
| MCF10A | CRL-10317 | Non-invasive/ Basal B | ER-, PR- | n/a | n/a |

Comparison of the expression profiles in untreated cells showed that the three-breast cancer model cell lines were transcriptionally dissimilar to the control cell line MCF10A and that BT-549 and MCF7 were more similar to each other than either were to BT-474. Expression levels (FPKM values) across 63,286 gene protein coding transcripts (GRCh38 reference genome) were used to calculate Jensen-Shannon Divergence (JSD) (Methods and Fig. 2A). JSD values correspond to the similarity of the probability distributions of transcript levels for two RNA-seq experiments. Expression values for biological replicates showed the highest similarities (smallest distances) within cell types (Fig. 2A, upper grid). The largest distances were observed between MCF10A and the three cancer cell types: 0.461 for BT-549, 0.476 for MCF7, and 0.511 for BT-474 (Fig. 2A, lower grid). A similarly high JS distance was observed for BT-549 versus BT-474 (JSD = 0.464), suggesting that these cancer cell lines are transcriptionally distinct. BT-549 and MCF7 showed the highest similarity, with a cumulative JSD of 0.357. This observation contrasts with other reports where BT-549 and MCF7 are described as transcriptionally and phenotypically different (Kenny et al., 2007; Seals et al., 2005).

Differences in transcription profiling methods, RNA-seq used here and the DNA oligomer microarray chip used by others, may underlie the different outcomes.




*Chapter 3. Figure 2. Comparisons of Transcription Profiles of Three Model Breast Cancer Lines (MCF7, BT-549, BT-474) and a Control Non-cancer Line (MCF10A).* (A) Jensen-Shannon Divergence (JSD) values were calculated as the similarity of the probability distributions of expression levels (FPKM values) for 63,286 total transcripts, which include 22,267 protein-coding transcripts. In the lower grid, cummeRbund (Trapnell et al., 2012) was used to consolidate replicates and to calculate overall JSD between cell types. Solid border, BT-549 vs. MCF7, smallest JSD; dashed border, JSD's for MCF10A vs. cancer cell lines. (B) The boxplots show gene expression values (center line, median; lower and upper boxes, 25th and 75th percentiles; lower and upper whiskers, minimum and maximum) for all protein-coding transcripts (22,267), H3K27me3-positive (1,146) or EZH2-positive (2,397) protein-coding loci. NS, no signal. (C) The Venn diagram includes HGNC symbols of genes that are H3K27me3-positive (middle box plot, panel B) and are silenced (FPKM < 2) in at least one cell type. GO term enrichment p-values are shown only for subsets where FDR < 0.1.

Differential expression between cell lines for individual genes (Fig. S1) followed similar trends as those observed for the global JSD analysis. We used an expression

comparison algorithm (Cuffdiff (Trapnell et al., 2013)) to identify genes that were differentially expressed (2-fold or greater difference in expression, $q$ value ≤ 0.05) or similarly expressed (less than 2-fold difference, $q$ value ≤ 0.05) between cell types. Comparisons that included MCF10A showed the highest numbers of differentially expressed genes, as well as the lowest numbers of similarly expressed genes. This result further supports transcriptional differences between the cancerous cell lines and MCF10A (Fig. S1).

Next, we determined expression levels within groups of predicted PRC-regulated genes and observed that expression within these subsets is lower in the three cancer cell types than in MCF10A. We used data from other breast cancer cell line studies of MCF7 and MDA-MB-231 to classify a subset of PRC target genes based on H3K27me3 enrichment or binding of EZH2, an enzyme that generates the H3K27me3 mark (see Methods). Only 245 gene IDs were shared between the H3K27me3 and EZH2 subsets. Although these two groups are mostly distinct, both showed low median expression values (FPKM < 2), which suggests epigenetic repression (Fig. 2B). Median expression levels of predicted PRC module genes were reduced in the cancer cell lines compared to the non-cancer cell line. The H3K27me3-marked subset showed median log10(FPKM) values for BT-474 (-1.66), MCF7 (-1.16), and BT-549 (-1.15) that were slightly lower than MCF10A (-1.10) (Fig. 2B, middle plot). The median FPKM values for EZH2 targets were dramatically lower (zero signal) in the cancer cell lines, while the median value was higher (-1.65) for MCF10A (Fig. 2B, right). Overall, H3K27me3 and EZH2 enrichments from two breast cancer cell lines (MCF7 and MDA-MB-231) correspond to relatively

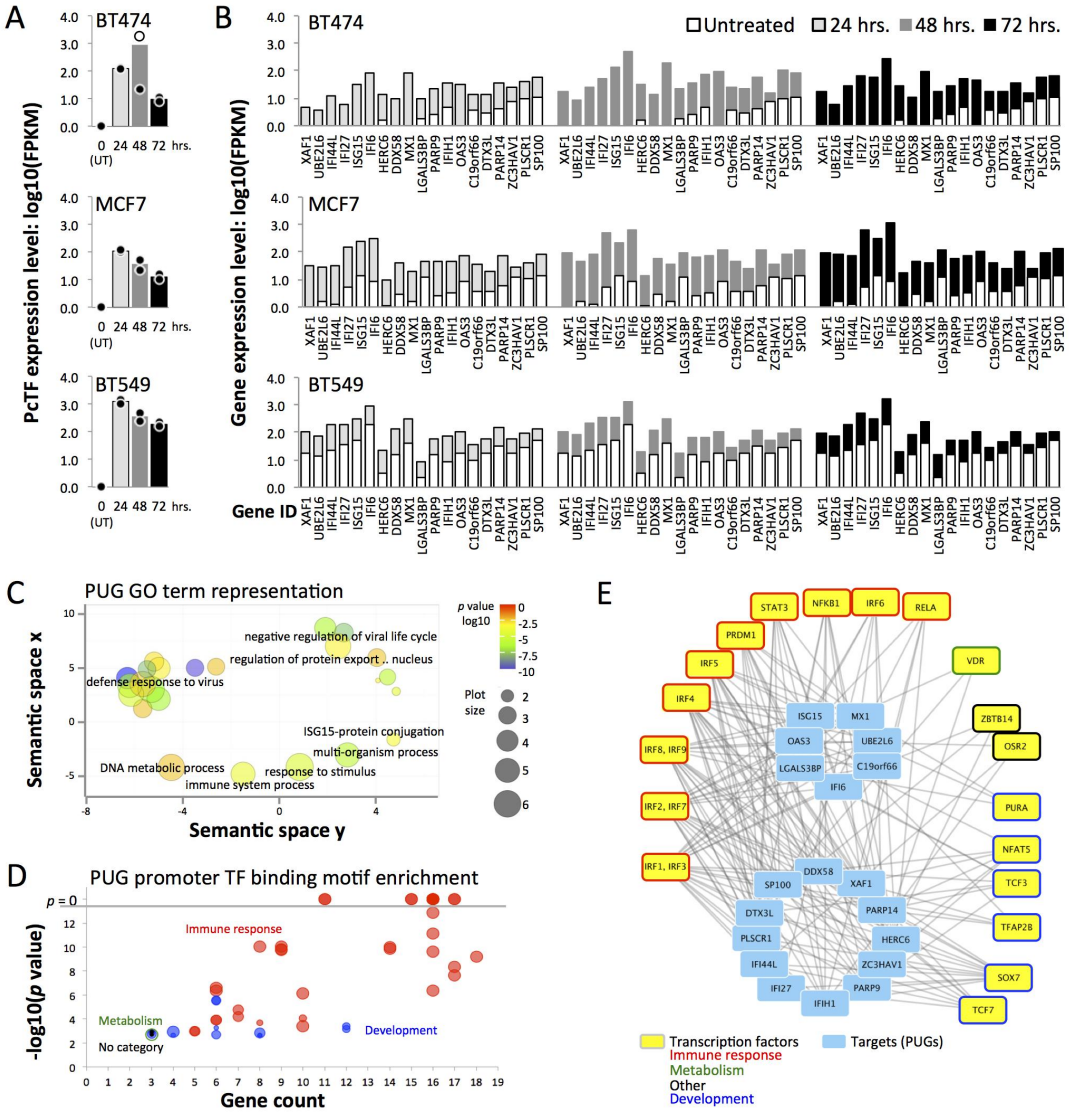
low expression in all three breast cancer cell lines studied here. This result is consistent with the roles of H3K27me3 and EZH2 in cancer-associated gene silencing.

To determine whether individual predicted PRC target genes were similarly regulated across cell lines, we compared two groups of genes that were categorized by expression level: silenced (FPKM < 2) (Gonzàlez-Porta et al., 2013; Rupp et al., 2017) or expressed (FPKM ≥ 2) (Fig. S2). In each cell type, genes with silenced expression levels included 70.2% - 79.3% of the H3K27me3-marked loci (Fig. S2) and 78.4% - 82.2% of the EZH2-enriched loci. About one quarter of the genes (17.8% - 29.8%) showed some expression (FPKM ≥ 2) and only 16.7% - 8.2% were expressed at FPKM ≥ 10. The set of 45 H3K27me3-enriched repressed genes shared by the three cancer cell lines BT-474, BT-549, and MCF7 (Table S1) shows strong representation of the gene ontology processes "regulation of peroxidase activity" (GOrilla(Eden et al., 2009), $p$ = 5.84E-6, FDR = 8.85E-2; Fig. 2C) and "ectoderm development" (Panther(Mi et al., 2017), $p$ = 1.07E-4, FDR = 2.61E-2). The silencing of lipoxygenase (*ALOXE3*) and inhibitor of peroxidase (*LRRK2*) may contribute to elevated pro-cancer COX-mediated peroxidase activity (Fürstenberger et al., 2006; Jardim et al., 2013). Low levels of *ALOXE3*, *ADRB2*, *BNC1*, *BTC*, *CCNO*, *ETV4, MCIDAS*, *PID1*, *SPRR2D*, and *ZBTB16* are consistent with the epigenetic repression of pro-differentiation pathways in cancer cells. We hypothesized that these PRC-module genes would become activated in the presence of the synthetic regulator PcTF, which interacts with the repressive H3K27me3 mark.

***PcTF-sensitive Interferon Response Genes are Shared Across Three Cancer Cell Types.*** We investigated changes in the transcriptomes of PcTF-expressing breast cancer cells over time. We transfected cells with PcTF-encoding plasmid DNA

(previously described (Nyer et al., 2017)) and allowed them to grow for 24, 48, and 72 hours before extracting total RNA for sequencing. RNA-seq reads were aligned to a human reference genome GRCh38 that included the coding region for PcTF (see Methods). No reads aligned to the PcTF coding sequence in control, untransfected cells. In the transfected cells, PcTF expression levels were highest at 24 hours and decreased 1.6 to 5.5-fold every 24 hours (Fig. 3A). We observed a similar trend with other cancer cell lines in a previous study (Nyer et al., 2017). One outlier sample, a replicate for BT-474 cells expressing PcTF for 48 hours, had a markedly different PcTF expression level (Fig. 3A) and genome-wide transcription profile (Fig. S3) and was therefore omitted from further analyses.

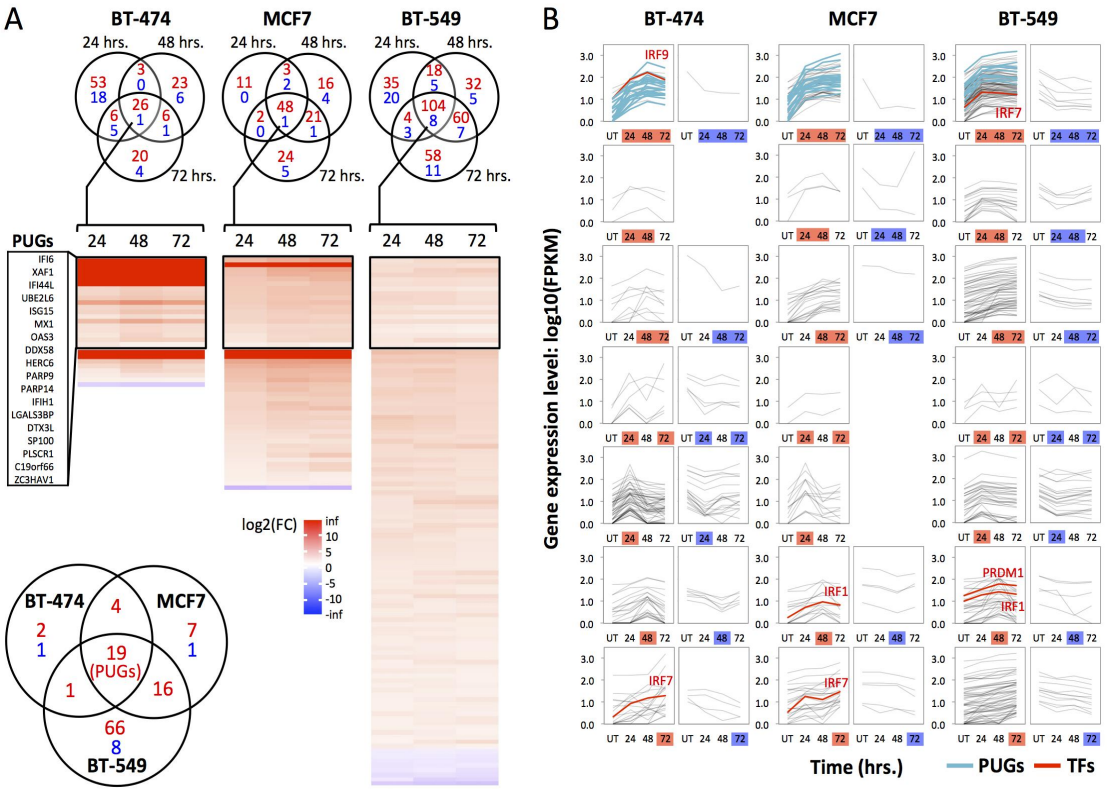
Nineteen genes were upregulated at least 2-fold ($q$ value $\leq 0.05$) at all time points in all three cell lines (Fig. 3B): *C19orf66*, *DDX58, DTX3L, HERC6, IFI27, IFI44L, IFI6, IFIH1, ISG15, LGALS3BP, MX1, OAS1, OAS3, PARP9, PARP14, PLSCR1, SP100, UBE2L6*, and *XAF1*. Here, we refer to this subset PcTF-upregulated genes, or PUGs. The most significantly enriched GO terms for this set include "defense response to virus" and "negative regulation of viral life cycle" (Fig. 3C). An investigation of regulator motif enrichment at the promoters of PUGs revealed significant overrepresentation of transcription factors involved in immune response and tissue development processes (Fig. 3D). Fifteen of the 22 transcription factors showed detectable levels of expression in all three cell lines (Fig. S4). *IRF1*, *IRF7*, *IRF9*, and *PRDM1* showed significant upregulation (FC $\geq 2$, q $\leq 0.05$) in PcTF-expressing cells. Promoter motifs for *IRF1* and *IRF3* were present at all 19 PUGs (Fig. 3E). Therefore, regulation of PUGs may be driven by PcTF-mediated activation of IRF1.




*Chapter 3. Figure 3. PcTF-expressing Breast Tissue-derived Cell Lines Show Upregulation of Interferon (IFN) Pathway Genes.* (A) Charts show log10(FPKM) of PcTF for untransfected cells (UT) and at 24, 48, and 72 hours following transfection of each cell line. The outlier for BT-474 (48 hrs, replicate 1) was omitted from subsequent analyses. Dots, each replicate library; bars, mean of values from the two replicates. (B) Mean log10(FPKM) values are shown for 19 Polycomb-upregulated genes genes (PUGs; FC $\geq$ 2, q $\leq$ 0.05 at all time points in all three cell lines), sorted from lowest to highest average expression level in untreated cells. (C) Gene ontology (GO) Biological Process term enrichment for the 19 PUGs is represented the bubble chart. GO clusters and representative terms (black labels) are plotted based on semantic similarities in the underlying GOA database. (D) Overrepresentation of transcription factor (TF) binding motifs (Plaisier et al., 2016) at the promoters of PUGs (p-value < 0.05/19.0, Bonferroni

correction). (E) Transcription factors (outermost boxes) associated with promoter motifs from panel D are shown in the network graph.

Different subsets of genes were up- or down-regulated at least two-fold ($q$ value ≤ 0.05) early, late, or across all time points during PcTF expression (Fig. 4). Of the genes that showed at least a two-fold change in either direction, the vast majority were up-regulated (Fig. 4A). We also observed that depending on the cell line, two or three predicted regulators of PUGs, including *IRF1*, *IRF7*, *IRF9*, and *PRDM1*, became significantly upregulated (Fig. 4B). This result suggests that the IFN response might be mediated through upregulation of master regulators. Thus, PcTF may target silenced chromatin at *IRF1*, *IRF7*, *IRF9*, and *PRDM1* and not necessarily at PUGs.

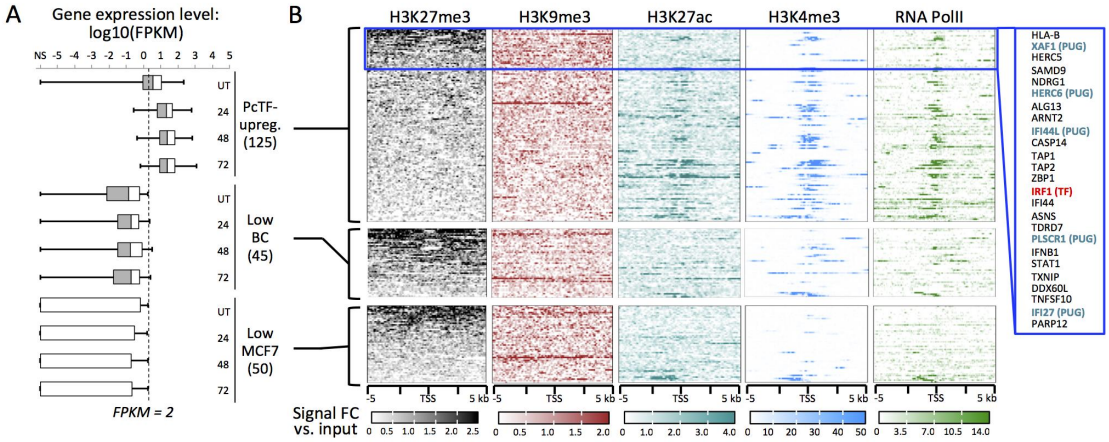


*Chapter 3. Figure 4. PcTF-sensitive Genes Include Cell-type Specific Groups in Addition to PUGs.* (A) The Venn diagrams show genes with expression levels that changed at least 2-fold in either direction (q value ≤ 0.05) at one or more time points in PcTF-expressing cells versus untransfected cells. Red, up-regulated; blue, down-regulated. The heat maps show fold-change (log2(FC)) values for genes that significantly changed (q ≤ 0.05) at all three time points (center regions of the Venn diagrams). The lower left Venn diagram compares these genes between cell types. (B) Expression profiles (log10(FPKM)) of cells before (UT, untransfected), and 24, 48, or 72 hours after PcTF transfection for all genes with expression levels that changed at least 2-fold in either direction (see Venn diagrams in panel A).

Our results also show that the PcTF-activated genes had virtually no overlap with the 45 H3K27me3-enriched, silenced genes (FPKM < 2) shared by the three cancer cell lines (Fig. 2C, Table S1). Only one of these 45 genes, *PID1*, became upregulated in any cell line (BT-549 at 48 and 72 hours). In this study we observed that the genes that were up-regulated came from the pool of low- to moderate-expressing genes. So far, our results suggest that PcTF-mediated activation requires a moderate level of basal expression at the target gene. This idea may be counterintuitive since H3K27me3 mark, the target of PcTF (Tekel et al., 2017), is essential for transcriptional repression according to the model for Polycomb-mediated regulation, which is supported by a wealth of data (Simon & Kingston, 2009). However, a recent study using genome-wide ChIP-seq and transcription profiles in murine cells showed that H3K27me3 was enriched at genes with low levels of expression and depleted at completely silenced genes, and highly expressed genes (Berrozpe et al., 2017). We were prompted to investigate whether the chromatin features at PcTF-activated genes might reflect a low to moderate expression state.

***PcTF-sensitive Loci Bear Repression- and Activation-associated Chromatin Features.*** To investigate the contribution of local chromatin states to PcTF-mediated gene regulation, we analyzed histone modifications and RNA polymerase II enrichment

at PcTF-upregulated genes in MCF7. Here, we utilized the extensive public ChIP-seq data that is available for the MCF7 cell line to investigate chromatin features. The 125 genes that were significantly upregulated (FC ≥ 2, $q \leq 0.05$) at one, two, or all time points in MCF7 (see Fig. 4B) showed a range of H3K27me3 mean enrichment values across 10 kb centered around each transcriptional start site (Fig. 5A). Consistent with PUGs, the 106 additional upregulated genes showed significant overrepresentation of interferon response-related processes (GO biological process "type I interferon signaling pathway," $p$ = 4.08E-28, FDR = 6.21E-24).




*Chapter 3. Figure 5. Comparison of Chromatin Features at PcTF-activated and Non-activated Genes in MCF7.* (A) Box plots show expression levels (center line, median; left and right boxes, 25th and 75th percentiles; left and right whiskers, minimum and maximum) in untreated and PcTF-treated cells (24, 48, and 72 hrs) for each of the following gene subsets: PcTF-upreg., 125 genes that are upregulated (FC ≥ 2, p ≤ 0.05) in MCF7-expressing cells at one or more time points; Low BC, 45 H3K27me3-enriched genes that are repressed (FPKM < 2) in all three cancer cell lines (see Fig. 2C); Low MCF7, 50 genes that are repressed (FPKM < 2) in MCF7. TSS plots show ChIP signals of silencing-associated (H3K27me3, H3K9me3) and activation-associated (H3K27ac, H3K4me3) histone modifications, as well as RNA Polymerase II. Genes within the top 20% of mean values for H3K27me3-enrichment (within 10 kb) are highlighted (blue box).

Genes within the highest 20% of mean values for H3K27me3 included the predicted regulator *IRF1* (Fig. 3D, E) and 5 of the 19 PUGs. Other PcTF-responsive genes that lack the H3K27 methylation mark might represent downstream targets of the products expressed from targets of PcTF. Mean enrichments of H3K9me3 (Fig. 5A), a modification that is frequently found at constitutive pericentric heterochromatin and non-coding DNA (K. A. Haynes et al., 2004; Lachner et al., 2004; Nishibuchi & Déjardin, 2017), showed no pattern that resembled H3K27me3. PcTF-responsive genes tended to be distributed along chromosome arms rather than concentrated near centromeres (Fig. S4). This suggests that PcTF target sites coincide more closely with the distribution of facultative chromatin and epigenetically-regulated cell development genes (Boyer et al., 2006; Wiles & Selker, 2017).

Enrichments for the features associated with active expression, H3K27ac, H3K4me3, and RNA Pol II were stronger at PcTF-responsive genes than at PcTF non-responsive genes (Fig. 5B). Regions containing PcTF-activated genes include interspersed peaks of H3K27me3 and H3K4me3 (Fig. S5), which is characteristic of bivalent domains that are poised for activation (Easwaran et al., 2012; Zaidi et al., 2017). We conclude that under the conditions tested here, strongly repressed genes are resistant to PcTF-mediated activation while an intermediate regulatory state, where silent and active marks are present, supports PcTF activity.

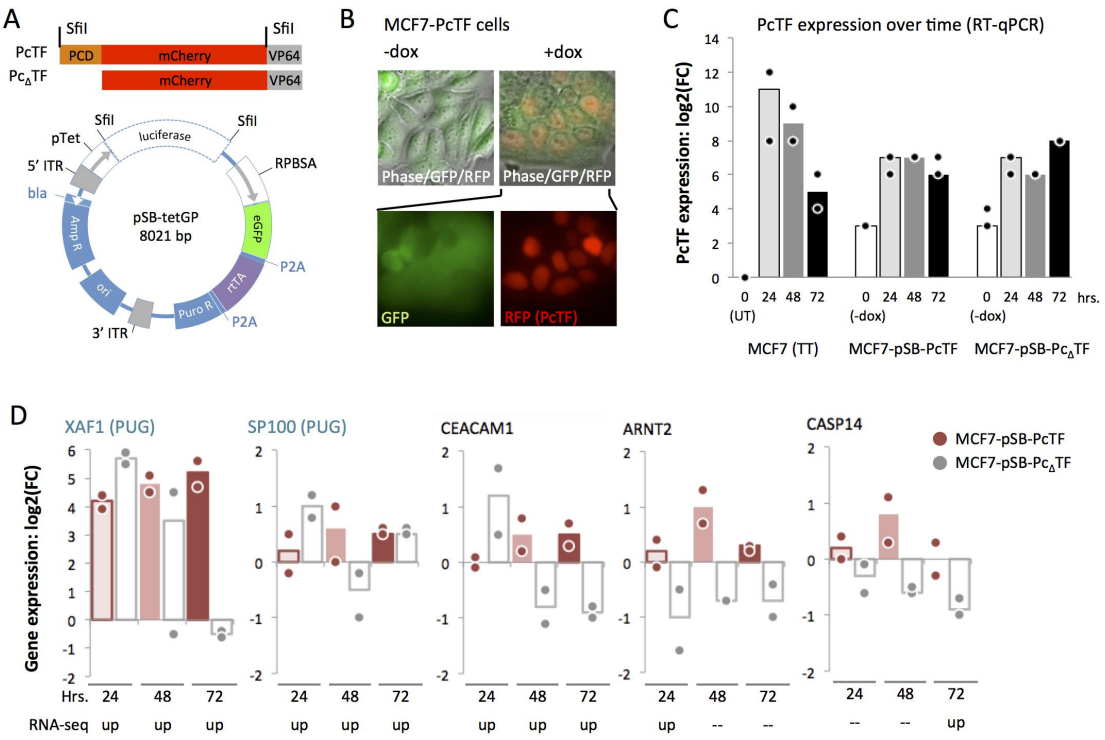
Two substantially different mechanisms might account for the results observed so far. First, target gene activation may depend upon PcTF's interaction with and disruption of silenced chromatin. In previous work, we established that PcTF activity requires the histone-binding PCD domain (Karmella A. Haynes & Silver, 2011; Nyer et al., 2017) and

the presence of H3K27me3 near the target gene (Karmella A. Haynes & Silver, 2011) to disrupt epigenetic silencing. Work reported by others demonstrated activation of interferon networks through the disruption of chromatin-mediated repression with small molecule inhibitors. Treatment of breast cancer cell lines (including BT-474 and MCF7) with DNA methyltransferase (5-azacitidine) led to activation of *DDX58*, *IFI27*, *IFI6*, *IFIH1*, *ISG15*, *MX1*, *OAS3*, *UBE2L6*, *XAF1* (9 of the 19 PUGs), and other genes (Li et al., 2014). Furthermore, inhibitors of histone deacetylase, a class of enzymes that support repressed chromatin, stimulate rapid activation of interferon (IFN) genes in human and mouse cells (Leonova et al., 2018).

Second, introduction of foreign nucleic acids into the cells could have indirectly stimulated the interferon response via sequence non-specific effects (Fiszer-Kierzkowska et al., 2011; Huerfano et al., 2013; Jacobsen et al., 2009; Olejniczak et al., 2010; Sledz et al., 2003) without interaction of PcTF with chromatin. Microarray-based transcriptome profiling of MCF7 cells transfected with Lipofectamine-pM1-MT vector complexes showed upregulation of *HERC6*, *IFIH1*, *ISG15*, *LGALS3BP, MX1*, *OAS3, PLSCR1,* and *UBE2L6* (Jacobsen et al., 2009), which represent 8 of the 19 PUGs. Small RNA-induced knockdown of GAPDH in renal carcinoma cells was accompanied by increased expression of *IFI6*, *OAS3*, and *UBE2L6* (Sledz et al., 2003). *MX1, IRF1 and IRF7* became activated following electroporation (nucleofection) of NIH3T3 cells with control empty plasmids pcDNA3.1 (the origin of the plasmids used in our study), phGF, and pEGFP-N1 (Huerfano et al., 2013). To investigate nonspecific effects from foreign nucleic acids, we used reverse transcription followed by quantitative PCR to measure

expression levels of PcTF-responsive genes in cells that expressed a truncated version of PcTF as a control, as described in the following section.

***Foreign RNA from a PcTF-deletion Mutant is Insufficient for Sustained Expression of XAF1 in MCF7.*** We asked whether the presence of the PcTF transgene and its transcribed RNA were responsible for the consistent interferon response in breast cancer cells. Using transient transfections, we had established that PcTF-mediated activation of genes could be detected over background at multiple time points. However, in this experiment PcTF levels decreased over time (Fig. 3A), which prevents us from distinguishing time- versus dose-dependent effects on gene regulation. Therefore, we constructed stable transgenic cell lines to enable constant expression of the fusion protein over time. We were able to generate viable, transgenic lines from MCF7 cells. Expression of PcTF or a control fusion protein that lacks the histone-binding domain ($Pc_{\Delta}TF$) was placed under the control of the rtTA activator, which binds to the *pTet* promoter in the presence of doxycycline (dox) (Fig. 6A). Expression of rtTA was indicated by constitutive GFP expression, and inducible nuclear localization sequence-tagged PcTF was detected as an RFP signal after treatment with doxycycline (Fig. 6B). We used reverse transcription followed by quantitative PCR (RT-qPCR) to measure the expression levels of PcTF and a subset of PcTF-sensitive genes that were identified in the RNA-seq experiment.




*Chapter 3. Figure 6. RT-qPCR Analysis of Gene Expression in Stable, Transgenic PcTF-Expressing Cells.* (A) SfiI-flanked PcTF or Pc$_\Delta$TF constructs (top) were cloned into the pSBtet-GP expression vector (bottom), resulting in the replacement of the luciferase reporter with fusion protein ORFs. (B) Fluorescence microscopy of the MCF7-PcTF transgenic cell line. (C) Time course RT-qPCR for PcTF. (D) Time course RT-qPCR for select genes. For all RT-qPCR experiments n = two cDNA libraries from independent transfections or dox treatments. FC, fold change relative to "no dox" controls, calculated as double delta $C_p$ (see Methods).

RT-qPCR using a universal mCherry-specific primer set confirmed that PcTF expression levels decreased over time in transiently transfected cells (Fig. 6C) as observed for FPKM values from the RNA-seq experiment (Fig. 3A). The stable transgenic cells showed low levels of fusion protein mRNA in the initial uninduced (-dox) state compared to untransfected MCF7 cells. Exposure to 1 µg/mL dox increased PcTF and Pc$_\Delta$TF levels by an order of magnitude. These levels were slightly higher than the PcTF expression levels observed in transiently transfected cells at the 72-hour time

point, and remained relatively constant over time. Fold-change (compared to untransfected cells) remained within values of 67 - 192 at 24, 48, and 72 hours.

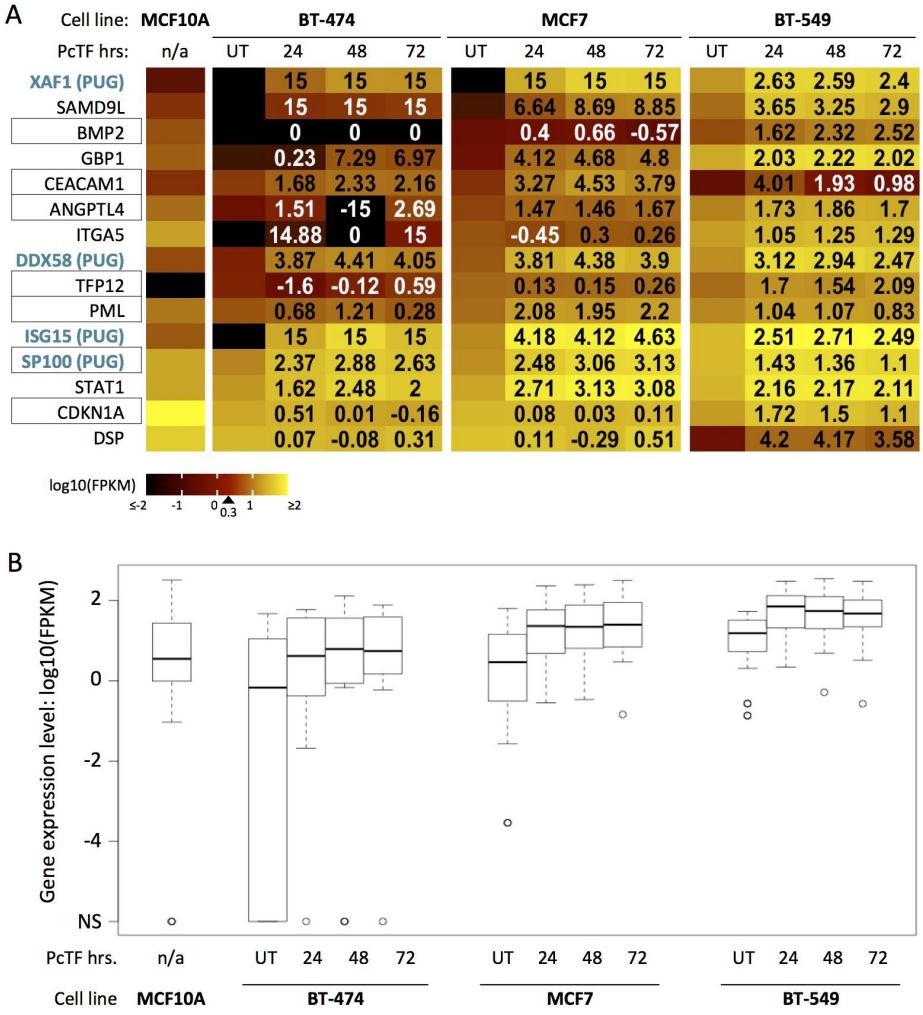
For RT-qPCR analysis of PcTF-sensitive targets, we were able to design and validate specific assays for a subset of genes that were significantly upregulated at one or more time points in MCF7, including two PUGs (*XAF1*, *SP100*) and others. *XAF1* was the most strongly upregulated across all three time points (18 to 36-fold) (Fig. 6D). The other five genes showed slight upregulation in response to dox-induced PcTF expression. The weaker response of these genes compared to *XAF1* could be explained by a smaller dynamic range, where there is little difference between the basal versus activated expression level. Furthermore, these genes may have been slightly upregulated prior to dox treatment since PcTF was detected at low levels before induction (Fig. 6C).
At the 24 hour time point, *XAF1*, *SP100*, and *CEACAM1* became up-regulated in truncation-expressing cells, suggesting an initial nonspecific response to transgenic $Pc_{\Delta}TF$ RNA. At 48 and 72 hours, gene expression decreased in the presence of $Pc_{\Delta}TF$. Over time, expression remained upregulated in the presence of PcTF compared to $Pc_{\Delta}TF$ at *XAF1*, *CEACAM1*, and *ARNT2*. Overall, these results suggest that for certain genes (*XAF1*, *CEACAM1*, and *ARNT2*), maintenance of the PcTF-induced activated state requires interaction with chromatin through the H3K27me3-binding PCD motif.

***Tumor Suppressor and BRCA Pathway Genes Become Upregulated in PcTF-Expressing Cells.*** To explore the clinical implications of PcTF-mediated transcriptional regulation, we determined the representation of known tumor suppressor genes amongst PcTF-responsive loci. For this analysis we used a tumor suppressor gene set that includes 983 candidate anti-cancer targets that are down-regulated in tumor samples (Methods).

Of these, 589 include BRCA human tumor suppressor genes (TSGs) that are repressed in invasive carcinoma samples compared to normal tissue samples (Min Zhao Jingchun Sun, 2013; Zhao et al., 2015). The genes were classified as tumor suppressors based on text-mining of cancer research literature, and manual assessment of relevant cancer types and molecular pathways (TSGene 2.0) (Min Zhao Jingchun Sun, 2013; Zhao et al., 2015). To identify TSGs that are upregulated in response to PcTF, we compared the upregulated subset (FC $\geq 2$, $q \leq 0.05$) to the 983 candidate anti-cancer genes identified by TSGene 2.0. Fifteen of the 983 TSGs were upregulated across all three time points in at least one of the cell lines (Fig. 7A). Information from genecards.org (Rebhan et al., 1997) further validated the association of these 15 genes with tumor suppressor activity. Of the fifteen upregulated TSGs, seven belong to the breast cancer susceptibility (BRCA) pathway: *CDKN1A*, *PML*, *ANGPTL4*, *CEACAM1*, *BMP2*, *SP100*, *TFPI2*.

Cell line comparisons of RNA-seq FPKM values for the fifteen tumor suppressor genes showed that median expression was lower in untreated BT-474 and MCF7 than in the non-cancerous MCF10A cell line (Fig. 7B). This result is consistent with the idea that epigenetic repression of TSGs supports a cancerous cell phenotype. In PcTF-expressing cells, the median expression of the fifteen tumor suppressor genes was increased at all time points compared to the untreated samples for each cancer cell line (Fig. 7B). Interestingly, the median FPKM value for the 15 TSGs was higher in BT-549 than in MCF10A. Closer examination of the individual genes revealed that expression levels for *BMP2*, *CEACAM1*, *CDKN1A*, *DSP* are lower in BT-549 than in MCF10A (Fig. 7A). These genes become upregulated in PcTF-expressing cells. These results demonstrate

that PcTF stimulates conversion of the expression state of several tumor suppressor genes from silenced to active.




*Chapter 3. Figure 7. Tumor Suppressor Genes Show Increased Expression in PcTF-Expressing Cancer Cell Lines.* (A) Individual log10(FPKM) (color scale) for each of the tumor suppressor genes in A. Black boxes highlight BRCA pathway genes. Genes are sorted from lowest to highest expression in untreated MCF7 cells. Numbers in the PcTF-treatment columns show log2 fold change values compared to UT. 15, infinite positive fold change where no expression was detected in untreated cells; -15, infinite negative fold-change where no expression was detected in treated cells. (B) Box plots show expression values (center line, median; lower and upper boxes, 25th and 75th percentiles; lower and upper whiskers, minimum and maximum) across three time points (24, 48, and

72 hours) for fifteen tumor suppressor genes where upregulation was at last two-fold ($q \leq 0.05$) relative to the untreated control (UT) in at least one of the cell lines.

**Discussion**

As the importance of global chromatin-mediated dysregulation in oncogenesis is coming to light, scientists are becoming more interested in using inhibitors to block master regulators of repressive chromatin (i.e., HDACs, DNMTs, HMTs(Dawson & Kouzarides, 2012; Dunn & Rao, 2017; Li et al., 2014; Mani & Herceg, 2010; Stone et al., 2017)) to investigate and treat cancer. This approach has been recently described as "macrogenomic engineering"(Almassalha et al., 2017). A key advantage of broad epigenetic manipulation is that it is DNA sequence-agnostic; the therapeutic effect potentially does not require *a priori* knowledge of patient-specific sequence variations at a candidate target gene or genes. Cancer tissues often accumulate extensive DNA lesions, from small insertions and deletions to large chromosome rearrangements. Therefore, editing or activating single targets may not be effective in some cells. In this report we present a synthetic approach to macrogenomic engineering, a fusion protein that physically bridges a chromatin feature at silenced genes (H3K27me3) with proteins that drive gene activation. Our previous studies have established that PcTF specifically interacts with H3K27me3 *in vitro* (Tekel et al., 2017), and drives the activation of hundreds of repressed loci including master regulators and tumor suppressors in bone, blood, and brain cancer derived model cell lines (Nyer et al., 2017). In our current report, we discovered a core set of interferon-pathway-related genes that responded to PcTF in three distinct breast cancer cell lines.

Several factors can contribute to transcriptomic variations in breast cancer subtypes, such as differences in the abundance of wild type or mutated transcription factors, mutations that impact the stability and turnover of RNA transcripts, and dysregulation of histone-modifying enzymes (Peña-Llopis et al., 2016). It is important to determine the relationship between phenotypic subclasses and transcription profiles (Guo et al., 2017; Jene-Sanz et al., 2013; Kenny et al., 2007; Seals et al., 2005) to elucidate cancer mechanisms and drug targets for more effective treatments. Establishing a link between transcriptomes and phenotypes may require further research. We observed that the transcription profile of BT-549 (invasive basal B) is more similar to MCF7 (luminal) than either were to BT-474 (luminal). In contrast, other reports have shown clear distinctions between the transcription profiles and phenotypes of BT-549 and MCF7 (Kenny et al., 2007; Seals et al., 2005). Differences in transcript profiling methods, our RNA-seq and JSD analysis versus the DNA oligomer arrays used by others, may account for this conflicting result. Further, we acknowledge that the JSD may be driven by a few genes with high expression and high variance, which could account for some of the patterns.

Diversity of breast cancer cell transcriptomes poses a formidable challenge for the development of drugs that target specific proteins, genes, and pathways. Our results demonstrate that activation of a common set of genes can be achieved by direct targeting of H3K27me3 with a fusion activator (PcTF) in three distinct model breast cancer cell lines that show distinct basal gene-expression levels. The 19 common PcTF-upregulated genes (PUGs) show significant overrepresentation of the GO biological processes “defense response to virus” and “negative regulation of viral life cycle.” A larger set of

125 genes that are upregulated at any time point in MCF7 (Fig. 4, 5) are associated with "type I interferon signaling pathway". Enrichments of H3K27me3 signals near the promoters of five PUGs (*XAF1*, *HERC6*, *IFI44L*, *PLSCR1*, *IFI27*) and a predicted regulator of all 19 PUGs (*IRF1*), suggest that PcTF accumulates near these promoters and recruits transcriptional activation machinery as demonstrated for *CASZ1* in a previous study(Nyer et al., 2017). Another potential mechanism for stimulation of the IFN pathway is epigenetic de-repression of endogenous retroviral dsRNA production, as observed during treatments with inhibitors against DNA methyltransferases histone deacetylases (Brocks et al., 2017; Chiappinelli et al., 2015; Roulois et al., 2015). It has been proposed that this process mimics a viral infection that makes the cancer cell a target for destruction by the immune system or immunotherapies (Classon et al., 2017). While many H3K27me3-enriched genes were upregulated in MCF7, many were non-responsive under the conditions tested here (up to 72 hours of PcTF expression). At PcTF-responsive genes, levels of H3K4me3 and H3K27ac were higher than at silenced non-responsive genes. Therefore, the chromatin at PcTF-responsive genes may support a low or intermediate expression state. Berrozpe et al. recently reported that Polycomb complexes preferentially accumulate at weakly expressed genes rather than strongly silenced or strongly expressed genes (Berrozpe et al., 2017). In our experiments, specific PRC-regulated genes may have been expressed at low to intermediate levels and then further upregulated upon exposure to PcTF. Our analysis of PcTF-regulated genes and chromatin states paves the way for future studies to further resolve chromatin features that distinguish regulatable PRC-repressed genes in cancer cells.

So far, low molecular weight compounds are the predominant method for epigenetic research and interventions. Their ease of delivery, orally or intravenously, make these compounds a very attractive approach for *in vivo* studies and cancer treatment. However, small compounds have a very limited range of biological activity, *e.g.* as ligands for specific proteins, compared to macromolecules. Transgenic and synthetic transcription factors expand the repertoire of epigenetic drug activity by allowing selective control of therapeutic genes in cancer cells (Beltran et al., 2007; Falke et al., 2003; Kwilas et al., 2015; Lara et al., 2012). Protein expression often relies on inefficient and possibly mutagenic nucleic acid delivery, which poses a significant barrier for many potential synthetic biologics. Recent advances in large molecule carriers such as cell penetrating peptides (Akishiba et al., 2017; Essafi et al., 2011; Staahl et al., 2017) provide a positive outlook for cellular delivery of purified proteins.

**Conclusions**

In conclusion, we have demonstrated that PcTF stimulates broad changes in expression, reminiscent of the effects observed for small-molecule epigenetic drugs, that could disrupt the immune evasion phenotype of cancer. Activation of IFN pathways genes has important implications for cancer research and therapy. Other studies have linked high levels of expression from interferon pathway genes with a non-cancerous phenotype. In breast cancer, expression of an immune response gene subgroup, which includes *ISG15*, *MX1*, and other interferon genes, has been associated with improved prognosis in triple negative breast cancers (Teschendorff et al., 2007b; H. Xu et al., 2014b). It will be eventually important to determine if PcTF proteins meet or exceed the

efficacy of low molecular weight epigenetic drugs in tumor and patient-derived models. At present, PcTF and its variants (Tekel et al., 2017) represent a new exploration space for rationally designed epigenetic interventions.

## Materials and Methods

***DNA Constructs.*** Plasmids were constructed to express fusion proteins either constitutively or in the presence of doxycycline. The plasmid for constitutive expression of PcTF, hPCD-TF_MV2 (KAH126), was constructed as previously described (Karmella A. Haynes & Silver, 2011). The doxycycline-inducible transgene PcTF_pSBtet-GP was constructed by ligating 50 ng of PCR amplified, SfiI-digested PcTF fragment with a SfiI-linearized pSBtet-GP vector (Kowarz et al., 2015) (Addgene #60495) at a ratio of 5 insert to 1 vector in a 10 uL reaction (1 uL 10x buffer, 1 uL T4 ligase). The same procedure was used to build constructs for dox-inducible Pc$_{\Delta}$TF expression. Primers used for the PCR amplification step are as follows: Forward 5'-tgaaGGCCTCTGAGGCCaattcgcggccgcatctaga, Reverse 5'-gcttGGCCTGACAGGCtgcagcggccgctactagt. Template-binding sequences are underscored. Adjacent nucleotides were designed to add *SfiI* restriction sites (uppercase) to each end. The full annotated sequences of all plasmids reported here are available online at Benchling - Hayneslab: Synthetic Chromatin Actuators (https://benchling.com/hayneslab/f/S0I0WLoRFK-synthetic-chromatin-actuators/).

***Cell Culture and Transfection.*** MCF7 (ATCC HTB-22) cells were cultured in Eagle's Minimal Essential Medium supplemented with 0.01 mg/mL human recombinant insulin, 10% fetal bovine serum, and 1% penicillin and streptomyicn. BT-474 cells

(ATCC HTB-20) were cultured in ATCC Hybri-Care Medium supplemented with 1.5 g/L sodium bicarbonate, 10% fetal bovine serum, and 1% penicillin and streptomycin. BT-549 cells (ATCC HTB-122) were cultured in RPMI-1640 Medium supplemented with 0.0008 mg/mL human recombinant insulin, 10% fetal bovine serum, and 1% penicillin and streptomycin. MCF-10A cells (ATCC CRL-10317) were cultured in Mammary Epithelial Cell Growth Medium (Mammary Epithelial Cell Basal Medium and BulletKit supplements, except gentamycin-amphotericin B mix), supplemented with 100 ng/mL cholera toxin. Cells were grown at 37 °C in a humidified $CO_2$ incubator. PcTF-expressing MCF7, BT-474, and BT-549 cells were generated by transfecting $5x10^5$ cells in 6-well plates with DNA/Lipofectamine complexes: 2 µg of hPCD-TF_MV2 plasmid DNA, 7.5 µl of Lipofectamine LTX (Invitrogen), 2.5 PLUS reagent, 570 µl OptiMEM. Control cells were mock-transfected with DNA-free water. Transfected cells were grown in pen/strep-free growth medium for 18 hrs. The transfection medium was replaced with fresh, pen/strep-supplemented medium and cells were grown for up to 72 hrs.

***Generation of Stable Cell Lines.*** To generate doxycycline-inducible cell lines, MCF7 cells were transfected with the transposase-expressing plasmid SB100X and either hPCD-TF_pSBtet-GP or TF_pSBtet-GP (19:1 molar ratio of pSB to SB100X), under the same conditions as described above. After 24 hrs, the transfection medium was replaced with fresh, puromycin-supplemented medium (0.5 µg/mL). Cells were then grown until cell cultures were >90% GFP-positive as measured by flow cytometry. Total culture time was 2-3 weeks per cell line.

***Preparation of Total mRNA.*** Total messenger RNA was extracted from ~90% confluent cells (~$1-2x10^6$). Adherent cells were lysed directly in culture plates with 500

µl TRIzol. TRIzol cell lysates were extracted with 100 µl chloroform and centrifuged at 12,000 xg for 15 min. at 4°C. RNA was column-purified from the aqueous phase (Qiagen RNeasy Mini kit 74104).

***Reverse Transcription PCR Followed by Quantitative PCR (RT-qPCR).*** SuperScript III (Invitrogen) was used to generate cDNA from 2.0 µg of RNA. Real-time quantitative PCR reactions (15 µl each) contained 1x LightCycler 480 Probes Master Mix (Roche), 2.25 pmol of primers (see Supplemental Table 1 for sequences), and 2 µl of a 1:10 cDNA dilution (1:1000 dilution for GAPDH and mCh). The real time PCR program was run as follows: Pre-incubation, ramp at 4.4°C*sec$^{-1}$ to 95°C, hold 10 min.; Amplification, 45 cycles (ramp at 4.4°C*sec$^{-1}$ to 95°C, hold 10 sec., ramp at 2.2°C*sec$^{-1}$ to 60°C, hold 30 sec., single acquisition); Cooling, ramp at 2.2°C*sec$^{-1}$ to 40°C, hold 30 sec. Crossing point ($C_p$) values, the first peak of the second derivative of fluorescence over cycle number, were calculated by the Roche LightCycler 480 software. Expression level was calculated as *delta* $C_p$ = 2^[$C_p$ *GAPDH* - $C_p$ *experimental gene*]. Fold change was determined as *double delta* $C_p$ = *delta* $C_p$ *treated cells* / *delta* $C_p$ *mock* for PcTF expression levels (Fig. 3C), or as *double delta* $C_p$ = $C_p$ *dox treated cells* / *delta* $C_p$ *no dox* for gene expression levels in the stable cell lines (Fig. 3D).

***Transcriptome Profiling with RNA-seq.*** RNA-seq was performed using two biological replicates per cell type, treatment, and time point for transiently transfected cells and three replicates for untransfected MCF10A. Total RNA was prepared as described for RT-qPCR. 50 ng of total RNA was used to prepare cDNA via single primer isothermal amplification using the Ovation RNA-Seq System (Nugen 7102-A01) and automated on the Apollo 324 liquid handler (Wafergen). cDNA was sheared to

approximately 300 bp fragments using the Covaris M220 ultrasonicator. Libraries were generated using Kapa Biosystem's library preparation kit (KK8201). In separate reactions, fragments from each replicate sample were end-repaired, A-tailed, and ligated to index and adapter fragments (Bioo, 520999). The adapter-ligated molecules were cleaned using AMPure beads (Agencourt Bioscience/Beckman Coulter, A63883), and amplified with Kapa's HIFI enzyme. The library was analyzed on an Agilent Bioanalyzer, and quantified by qPCR (KAPA Library Quantification Kit, KK4835) before multiplex pooling and sequencing on a Hiseq 2000 platform (Illumina) at the ASU CLAS Genomics Core facility. Samples were sequenced at 8 per lane to generate an average of 2.5E+07 reads per sample. Read values ranged from 5.7E+06 (minimum) to 1.11E+08 (maximum) per sample.

***Transcriptome Analysis.*** RNA-seq reads were quality-checked before and after trimming and filtering using FastQC (Andrews, 2010). TrimmomaticSE was used to clip bases that were below the PHRED-scaled threshold quality of 10 at the 5' end and 25 at the trailing 3' end of each read for all samples (Bolger et al., 2014). A sliding window of 4 bases was used to clip reads when the average quality per base dropped below 30. Reads of less than 50 bp were removed. A combined reference genome index and dictionary for GRCH38.p7 (1-22, X, MT, and non-chromosomal sequences) (Harrow et al., 2012) that included the full coding region of the synthetic PcTF protein were created using Spliced Transcripts Alignment to Reference (STARv2.5.2b) (Dobin et al., 2013) and the picard tools (version 1.1.19) (*Picard Tools*, 2003). Trimmed RNA-seq reads were mapped, and splice junctions extracted, using STARv2.5.2b read aligner (Dobin et al., 2013). Bamtools2.4.0 (Barnett et al., 2011) was used to check alignment quality using the

'stats' command. Mapped reads in BAM format were sorted, duplicates were marked, read groups were added, and the files were indexed using the Bamtools 2.4.0 package. CuffDiff, a program in the Cufflinks package (Trapnell et al., 2012), was used to identify genes and transcripts that expressed significant changes in pairwise comparisons between conditions. Fastq and differential expression analysis files are available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database (Accession GSE103520, release date September 8, 2017). CummeRbund (Trapnell et al., 2012) was used to calculate distances between features and to generate graphs and charts (JSD plots). R ggplot2 (Harrow et al., 2012; Warnes et al., 2016) and VennDiagrams (Chen & Boutros, 2011) were used to generate heat maps and Venn diagrams respectively. The entire workflow is provided as a readme file at: https://github.com/WilsonSayresLab/PcTF_differential_expression

***Bioinformatics Analyses and Sources of Publicly Shared Data.*** Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) data: For the results shown in Figure 1B, H3K27me3 data for MCF7 cells was downloaded from the ENCODE project (accession UCSC-ENCODE-hg19:wgEncodeEH002922) (ENCODE Project Consortium, 2012). We classified genes with a ChIP-seq peak within 5000 bp up or downstream of the transcription start site as H3K27me3-positive (1,146 protein-coding transcripts). EZH2-enriched genes (2,397 protein-coding transcripts) for MDA-MB-231 (Jene-Sanz et al., 2013) were provided as a list from E. Benevolenskaya (unpublished). For the results shown in Figure 5 and S6, MCF7 ChIP-seq data (from the P. Farnham, J. Stamatoyannopoulos, and V. Iyer labs) was downloaded from the ENCODE project (ENCODE Project Consortium, 2012): H3K27me3 (ENCFF081UQC.bigWig),

H3K9me3 (ENCFF754TEC.bigWig), H3K27ac (ENCFF986ZEW.bigWig), H3K4me3 (ENCFF530LJW.bigWig), and RNA PolII (ENCFF690CUE.bam) and used to generate plots using DeepTools (Ramírez et al., 2016) (computeMatrix, plotProfile, plotHeatmap) in the Galaxy online platform at usegalaxy.org (Afgan et al., 2016). Prior to plotting, the RNA PolII data was converted to bigWig format using bamCoverage. Gene ontology term enrichment: GOrilla analysis used the following parameters: organism, Homo sapiens; mode, target and background ranked list of genes; ontology, process; p-value threshold = 10.0E-3) (Eden et al., 2009). The background ranked list is available at https://github.com/WilsonSayresLab/PcTF_differential_expression. Panther analysis used the following parameters: analysis type, PANTHER Overrepresentation Test (Released 20171205); annotation version, PANTHER version 13.1 Released 2018-02-03; reference List, Homo sapiens (all genes in database); annotation data set, PANTHER GO-Slim biological process. Figure 3C was generated using REViGO (Supek et al., 2011) and GOrilla. Unique differentially expressed genes were analyzed using GeneCards (Rebhan et al., 1997). Promoter motif analysis: The script TF_targets was downloaded from https://github.com/cplaisier/TF_targets and used to find enriched transcription factor target sites that were determined by empirical evidence from chromatin studies across 68 cell lines(Plaisier et al., 2016). Tumor suppressor genes: The results in Figure 7 are based on human tumor suppressor genes (983 total) that are reported to show lower expressed in cancer samples of the Cancer Genome Atlas (TCGA) compared to the TCGA normal tissue samples was downloaded from https://bioinfo.uth.edu/TSGene/download.cgi. Of these 983 genes, 589 are breast cancer specific (Min Zhao Jingchun Sun, 2013; Zhao et al., 2015).

## Supplementary Information

Supplemental tables and figures are located in chapter 3. appendices C.

CHAPTER 4

# Lack of Parent-of-Origin Effects in *Nasonia* Jewel Wasp: A Replication and Extension Study

(Preprint as Olney, K.C., Gibson, J.D., Natri, H.M., Underwood, A., Gadau, J., Wilson, M.A)

BioRxiv (2021). https://doi.org/10.1101/2021.02.11.430138

## ABSTRACT

In diploid cells, the paternal and maternal alleles are, on average, equally expressed. There are exceptions from this: a small number of genes express the maternal or paternal allele copy exclusively. This phenomenon, known as genomic imprinting, is common among eutherian mammals and some plant species; however, genomic imprinting in species with haplodiploid sex determination is not well characterized. Previous work reported no parent-of-origin effects in the hybrids of closely related haplodiploid *Nasonia vitripennis* and *Nasonia giraulti* jewel wasps, suggesting a lack of epigenetic reprogramming during embryogenesis in these species. Here, we replicate the gene expression dataset and observations using different individuals and sequencing technology, as well as reproduce these findings using the previously published RNA sequence data following our data analysis strategy. The major difference from the previous dataset is that they used an introgression strain as one of the parents and we found several loci that resisted introgression in that strain. Our results from both datasets demonstrate a species-of-origin effect, rather than a parent-of-origin effect. We present a reproducible workflow that others may use for replicating the results. Overall, we reproduced the original report of no parent-

of-origin effects in the haplodiploid *Nasonia* using the original data with our new processing and analysis pipeline and replicated these results with our newly generated data.

## Introduction

Parent-of-origin effects occur when there is a biased expression (or completely monoallelic expression) of alleles inherited from the two parents (Ishida & Moore, 2013; Reik & Walter, 2001). Monoallelic gene expression in the offspring is hypothesized to be primarily the result of genetic conflict between parents over resource allocation in the offspring (Isles, Davies, & Wilkinson, 2006; Moore & Haig, 1991). In mammals, the mechanism of these parent-of-origin effects occurs via inherited methylation of one allele (Lawson et al., 2013; Reik & Walter, 2001). In insects, the relationship between methylation of genomic DNA and the expression of the gene that it encodes is not as well characterized but studies of social insects showed that there is a positive correlation of DNA methylation of gene bodies and gene expression (Yan et al., 2015).

Honey bees have been a focal group for investigation of parent-of-origin effects in insects due to differences in the kinship between queens, males, and workers (Haig, 1992; Queller, 2003). Multiple mating by queens results in low paternal relatedness between workers and should lead to intragenomic conflict over worker reproduction (laying unfertilized eggs to produce males), and ultimately should favor the biased expression of paternal alleles that promote worker reproduction (Galbraith et al., 2016). Utilizing a cross between European (*Apis mellifera ligustica*) and Africanized honey bees, Galbraith et al. 2016 identified genes exhibiting a pattern of biased paternal allele overexpression in worker reproductive tissue from colonies that were queenless and broodless, a colony condition that promotes worker reproduction (Galbraith et al., 2016). Smith et al. 2020 found a similar pattern of paternal allele overexpression in diploid (worker-destined) eggs in a cross between two African subspecies, *A.m. scutellate* and *A.m. capensis* (Smith et

al., 2020). In reciprocal crosses of European (*A.m. ligustica* and *A.m. carnica*) and Africanized honey bees reared in colonies containing both brood and a queen, Kocher et al. 2015 instead found parent-of-origin effects in gene expression that were largely overexpressing the maternal allele in both directions of the cross (Kocher et al., 2015). These studies provide evidence for parent-of-origin effects in the honey bee, a eusocial Hymenoptera. The Kocher et al. 2016 dataset also exhibited asymmetric maternal allelic bias in which the paternal allele was silenced, but only in hybrids with Africanized fathers (Joshua D. Gibson, Arechavaleta-Velasco, Tsuruda, & Hunt, 2015). This set of biased genes was enriched for mitochondrial-localizing proteins and is overrepresented in loci associated with aggressive behavior in previous studies (Hunt, 2007; Hunt, Guzmán-Novoa, Fondrk, & Page, 1998). Interestingly, these same crosses exhibit high aggression in the direction of the cross with the Africanized father but not in the reciprocal cross (Shorter, Arechavaleta-Velasco, Robles-Rios, & Hunt, 2012), and aggression and brain oxidative metabolic rate appears to be linked in honey bees (Alaux et al., 2009). This study points toward a potential role of allelic bias and nuclear-mitochondrial genetic interactions in wide crosses of honey bees.

The parasitoid wasp genus *Nasonia* has emerged as an excellent model for studying genomic imprinting in Hymenoptera. Like honey bees and all Hymenoptera, *Nasonia* has a haplodiploid sex-determination system in which females are diploid, developing from fertilized eggs, and males are haploid, developing from unfertilized eggs. However, it serves as a strong contrast to studying parent-of-origin effects in the eusocial Hymenoptera as *Nasonia* is solitary and singly-mated, which should result in less genomic conflict and therefore less selective pressure for genomic imprinting based

on kinship. By studying allelic expression biases in this system, we can better assess genomic imprinting in the absence of kin selection and the potential contribution of nuclear-mitochondrial interactions to biased allelic expression. *Nasonia* is well-suited for these kinds of studies as two closely related species of *Nasonia* - *N. vitripennis* and *N. giraulti* - that diverged ~1 million years ago (Mya) and show a synonymous coding divergence of ~3% (Werren et al., 2010), can still produce viable and fertile offspring (Breeuwer & Werren, 1995). Highly inbred laboratory populations of *N. vitripennis* and *N. giraulti* with reduced polymorphism provide an ideal system for identifying parent-of-origin effects in hybrid offspring (Wang, Werren, & Clark, 2016). However, the species do show genetic variation and incompatibilities, such that recombinant F2 males (from unfertilized eggs of F1 hybrid females) suffer asymmetric hybrid breakdown in which 50% to 80% of the offspring die during development (Breeuwer & Werren, 1995). The mortality is dependent on the direction of the cross and those with *N. giraulti* maternity (cytoplasm) have the highest level of mortality. Nuclear-mitochondrial incompatibilities have been implicated in this and candidate loci have been identified (Gadau, Page, & Werren, 1999; J. D. Gibson, Niehuis, Peirson, Cash, & Gadau, 2013; Niehuis, Judson, & Gadau, 2008). Despite this high level of mortality in F2 males, there is no obvious difference in mortality of the F1 mothers of these males and non-hybrid females, further highlighting this as an excellent system in which to test the potential role of allelic expression bias in mitigating hybrid dysfunction.

Wang et al. 2016 used genome-wide DNA methylation and transcriptome-wide gene expression data from 11 individuals to test whether differences in DNA methylation drive the differences in gene expression between *N. vitripennis* and *N. giraulti,* and

whether there are any parent-of-origin effects (parental imprinting and allele-specific expression) (Wang et al., 2016). They used reciprocal crosses of these two species and found no parent-of-origin effects, suggesting a lack of genomic imprinting. Unlike the work in honey bees, however, there have not been multiple independent investigations of evidence for parent-of-origin effects in *Nasonia*.

Reproducibility is a major concern in science, particularly for the biological and medical sciences (Baker, 2016; Casadevall & Fang, 2010). To replicate is to make an exact copy. To reproduce is to make something similar to something else. Reports have shown that significant factors contributing to irreproducible research include selective reporting, unavailable code and methods, low statistical power, poor experimental design, and raw data not available from the original lab (Baker, 2015, 2016; Freedman & Inglese, 2014). In RNAseq experiments, raw counts are transformed into gene or isoform counts, which requires an *in silico* bioinformatics pipeline (Simoneau, Dumontier, Gosselin, & Scott, 2019). These pipelines are modular and parameterized according to the experimental setup (Simoneau et al., 2019). The choice of software, parameters used, and biological references can alter the results. In RNAseq, filters can also improve the robustness of differential expression calls and consistency across sites and platforms (Su et al., 2014). There is no, and there may never be, a defined optimal RNAseq processing pipeline from raw sequencing files to meaningful gene or isoform counts. Thus, the same data can be processed in a multitude of ways by the choice of software, parameters, and references used (Simoneau et al., 2019). Given the exact same inputs, software, and parameters, one can reproduce the analysis if the authors provide this documentation and make explicit the information related to the data transformation used to the RNAseq data

(Simoneau et al., 2019). In the case of the Wang et al. 2016, the methods and experimental design were exceptionally well documented, and the authors made available their raw data (Wang et al., 2016).

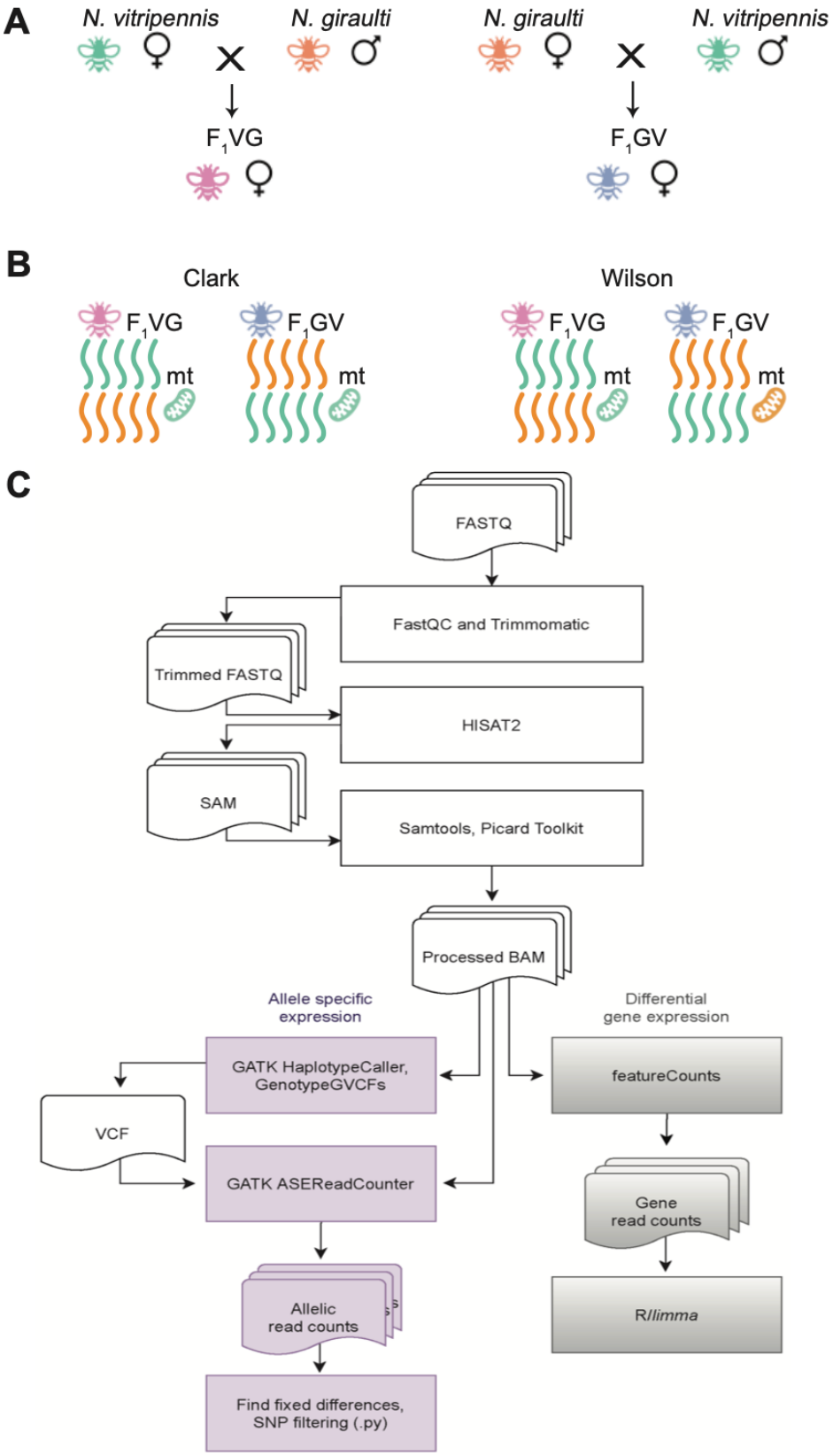
To address whether the Wang et al. 2016 findings of lack of parent-of-origin effects in *Nasonia* may be replicated and reproduced, we conducted two sets of analyses. We first downloaded the raw data from 11 individuals (Wang et al., 2016) and replicated differential expression (DE) and allele-specific expression (ASE) analyses. This allowed us to characterize species differences in gene expression, hybrid effects relative to each maternal and paternal line, and possible parent-of-origin effects using new alignment methods and software. Second, we reproduced the experimental setup with new individuals, generated transcriptome-wide expression levels of 12 *Nasonia* individuals (parental strains and reciprocal hybrids), named here as the Wilson data using similar, but not identical strains as the Wang et al. 2016 samples, which we named as the R16A Clark data. The Wilson data, reported here, used the standard *N. giraulti* strain (RV2Xu). The R16A Clark *N. giraulti* differs from the RV2Xu strain in that it has a nuclear *N. giraulti* genome introgressed into a *N. vitripennis* cytoplasm which harbour *N. vitripennis* mitochondria. Both studies used the same highly inbred standard *N. vitripennis* strain, ASymCx. We completed the above analyses to test for robust reproducibility in biased allele and parent-of-origin effects in *Nasonia*. In this analysis, we processed both the R16A Clark and Wilson data using the same software and thresholds, starting with the raw FASTQ files. While we detect some differences in the specific differentially expressed genes between the two datasets, our study reproduces and confirms the main conclusions of the Wang et al. 2016 study: we observe similar trends in the DE and ASE

genes, and we detect no parent-of-origin effects in *Nasonia* hybrids, indicating a validation of the lack of epigenetic reprogramming during embryogenesis in this taxa (Wang et al., 2016). We make available the bioinformatics processing and analysis pipeline used for both the R16A Clark and Wilson datasets for easily replicating the results reported here: https://github.com/SexChrLab/Nasonia. Finally, during the process of reproducing these results, we extend them to show potential interactions between the mtDNA and autosomal genome that were not apparent in the original study.

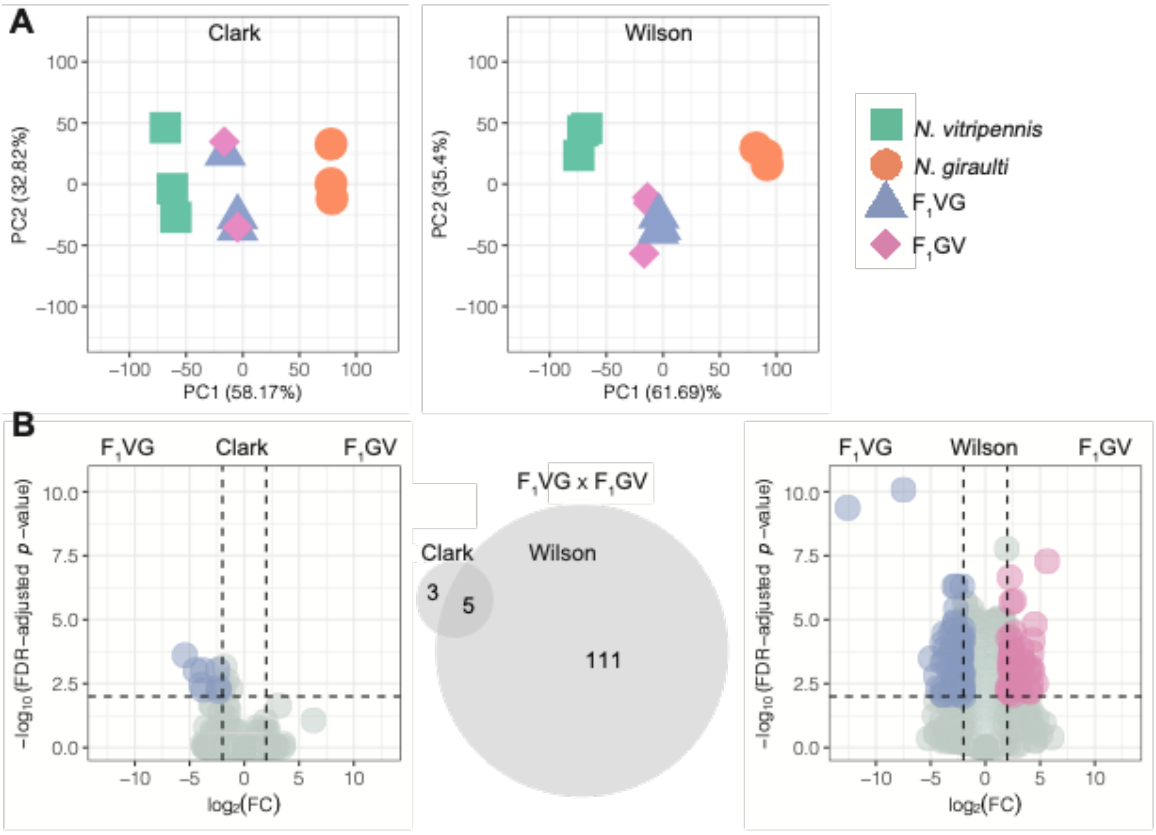
## Results

***Samples Cluster by Species and Hybrid in R16A Clark and Wilson Datasets.*** We used Principal Component Analysis (PCA) of gene expression data to explore the overall structure of the two datasets, R16A Clark and Wilson. Although the reciprocal hybrids from the two datasets are slightly different Figure 1B, in both sets, samples from the two species (strains) form separate clusters, with the clustering of the hybrid samples between them Figure 2A. The first PC explains most of the gene expression variation in both datasets, with proportions of variance explained 58.17% in R16A Clark and 61.69% in the Wilson data. Further, despite differences in experimental protocols, the transcriptome-wide gene expression measurements across the different crosses and species are highly correlated between the R16A Clark and Wilson dataset, Figure 3. There is a difference in the mean RNAseq library size between the two datasets. The mean RNAseq library size for the R16A Clark samples is 48,893,872 base pairs (bp) (SD=11,603,536) and the Wilson samples is 16,518,955 bp (SD=3,205,303),

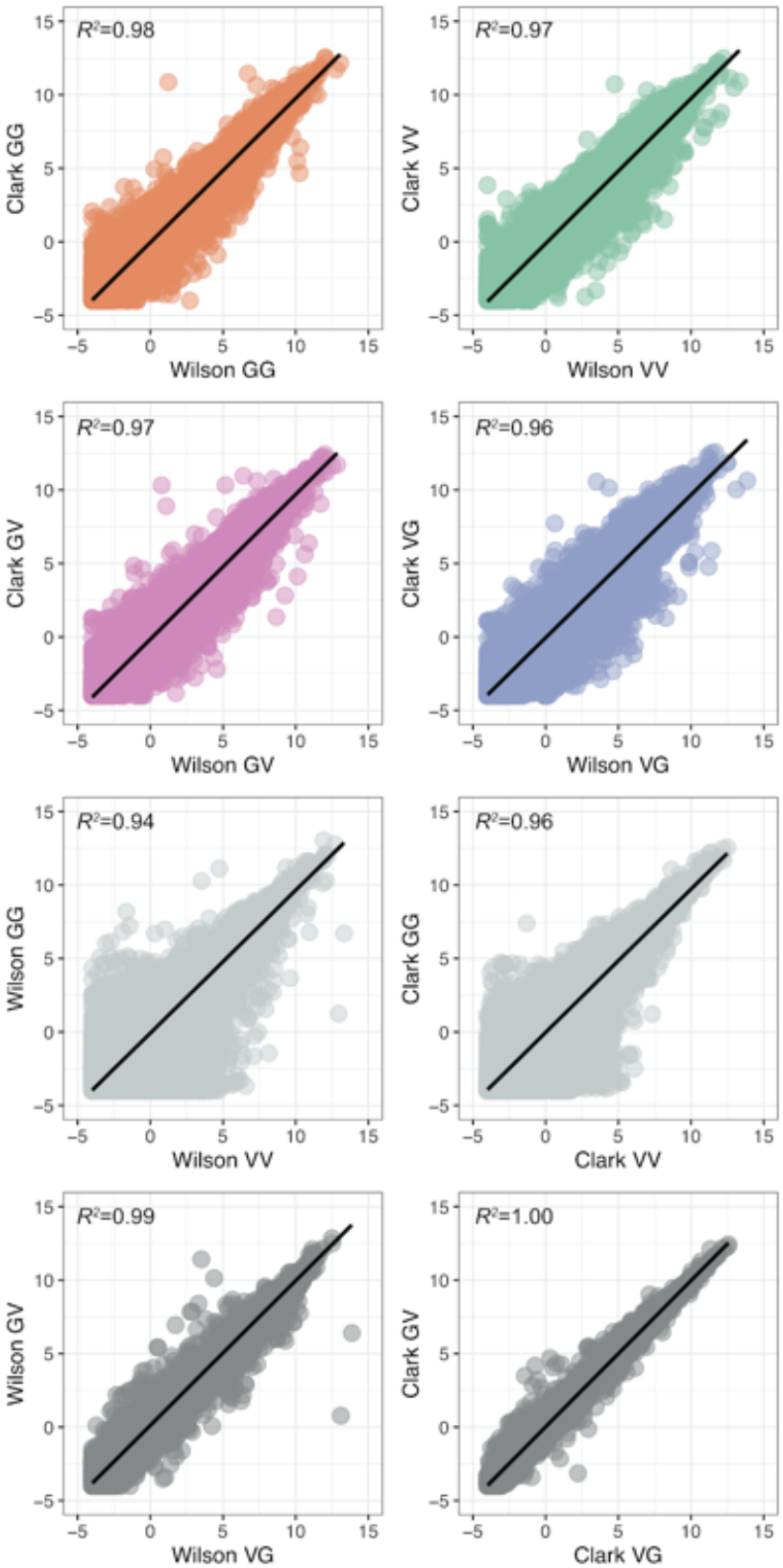
Supplemental 1 Table. Overall, we observe that most of the variation in the data is explained by species and hybrids.


A
N. vitripennis
N. girault
F1VG
N. girault
N. vitripennis
F1GV
B
Clark
F1VG
mt
F1GV
mt
Wilson
F1VG
mt
F1GV
mt
C
FASTQ
FastQC and Trimmomatic
Trimmed FASTQ
HISAT2
SAM
Samtools, Picard Toolkit
Processed BAM
Allele specific expression
Differential gene expression
GATK HaplotypeCaller, GenotypeGVCFs
featureCounts
VCF
GATK ASEReadCounter
Gene read counts
Allelic read counts
R/limma
Find fixed differences, SNP filtering (.py)

*Chapter 4. Figure 1. Experimental Design.* A: A schematic illustration of the reciprocal $F_1$ crosses. B: Schematic illustration of the hybrids nuclear and mitochondrial genomic make up. All hybrids are heterozygous at every nuclear locus for their two parent's alleles. The R16A Clark hybrids have *N. vitripennis* mitochondria, regardless of maternal species. The Wilson hybrids have their maternal species mitochondria. C: Overview of the data processing and analysis workflow.




*Chapter 4. Figure 2. Multidimensional Scaling and Differential Expression.* A: Gene expression PCA based on all expressed genes (mean FPKM $\geq 0.5$ across three biological replicates in at least one sample group) in the R16A Clark and Wilson datasets when taking the average between the *N. vitripennis* and pseudo *N. giraulti* reference genomes. B: Volcano plots of differentially expressed genes between the two reciprocal hybrids in the R16A Clark and Wilson datasets. Significance thresholds of an FDR-adjusted p-value $\leq 0.01$ and an absolute $\log_2FC \geq 2$ are indicated. A Venn diagram shows the overlap of the significant DEGs.


R²=0.98
Clark GG
Wilson GG
R²=0.97
Clark VV
Wilson VV
R²=0.97
Clark GV
Wilson GV
R²=0.96
Clark VG
Wilson VG
R²=0.94
Wilson GG
Wilson VV
R²=0.96
Clark GG
Clark VV
R²=0.99
Wilson GV
Wilson VG
R²=1.00
Clark GV
Clark VG

*Chapter 4. Figure 3.* Gene Expression Correlation. Gene expression correlation between the Wilson and R16A Clark datasets, as well as between species and between reciprocal hybrids within each dataset. Mean logCPM expression of each quantified gene in each cross and dataset is shown. Pearson's correlation $R^2$ is indicated.

***Species and Hybrid Differences in Gene Expression Between Closely Related N. vitripennis and N. giraulti.*** We detect more differentially expressed genes (DEGs) in the Wilson dataset despite the smaller library sizes, particularly in the comparison involving the hybrid samples (Figure 2B). We called DEGs, FDR ≤ 0.01, and absolute $\log_2$ fold change ≥ 2, between the different species and crosses within both datasets (Figure 2B and Supplemental 1 Figure). In the *N. vitripennis* (VV) x *N. giraulti* (GG) comparison, we identify 799 and 1,001 DEGs in the R16A Clark and Wilson datasets, respectively. We observe a 45.5% overlap of these DEGs between the datasets (Supplemental 1 Figure). As expected, we detect fewer DEGs in the comparisons involving the hybrids (Figure 1B). We detect only small differences in the numbers of DEGs called in the R16A Clark and Wilson datasets when examining hybrid effects relative to each maternal line (Supplemental 1 Figure). However, these DEGs show little overlap between the datasets, with the proportions of overlapping DEGs in VVxVG, VVxGV, GGxVG, and GGxGV, comparisons being 24.1%, 16.2%, 39%, and 31.6%, respectively.

There is a notable difference in the number of DEGs called between VG and GV hybrids between the R16A Clark and Wilson datasets. The R16A Clark data used an introgression strain of *N. giraulti*, R16A, that has a nuclear genome derived from *N. giraulti* but maintains *N. vitripennis* mitochondria, therefore the R16A Clark hybrids all have the same genetic makeup whereas the Wilson reciprocal hybrids have the same nuclear genome but different cytoplasms; yet, we do see eight genes called as

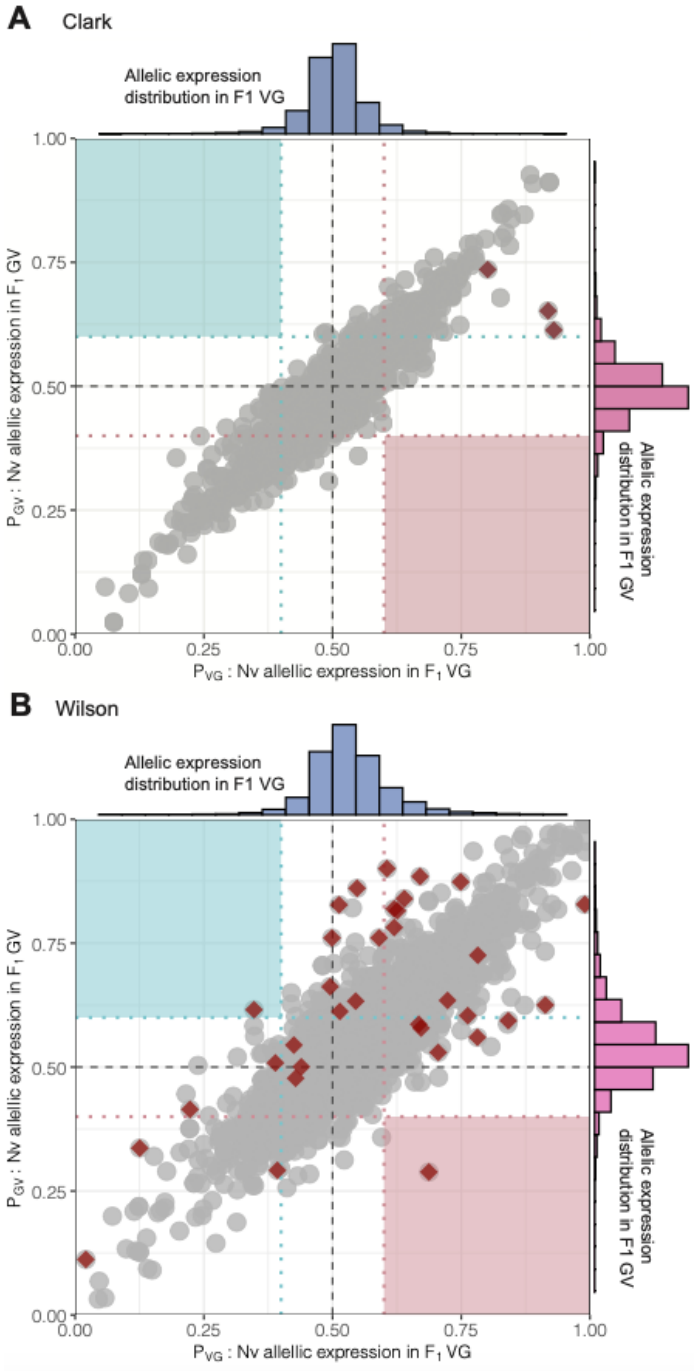
differentially expressed between the VG and GV hybrids in the R16A Clark data. Three of the eight genes in the R16A Clark data (LOC116416025, LOC116416106, LOC116417553) were only called as differentially expressed between the VG and GV hybrids in the R16A Clark dataset and weren't called as differentially expressed in the Wilson dataset. The other five genes (LOC107981401, LOC100114950, LOC116415892, LOC103317241, LOC107981942) were called as differentially expressed between the VG and GV in both datasets. In the Wilson data, we called 116 DEGs, 111 of which are uniquely to the Wilson data set. The original Wang et al. 2016 publication did not investigate differential expression between the hybrids (Wang et al., 2016). Here we report a new way of looking at the data, and despite the same genetic makeup between the hybrids in the R16A Clark data, we do observe differential expression between the hybrids, and five of those eight genes are also called as differentially expressed in the Wilson data.

Four (LOC107981401, LOC100114950, LOC116415892, and LOC103317241) out of the five DEGs shared between the data sets are uncharacterized proteins located on Chr 1, Chr 2, and Chr 4. To gain insight into the possible functions of these genes, we used NCBIs BLASTp excluding *Nasonia* (Johnson et al., 2008; NCBI Resource Coordinators & NCBI Resource Coordinators, 2017) to find regions of similarity between these sequences and characterized sequences. We observe several significant hits to different insects including *Drosophila* suggesting that these proteins have at least some conservation in insects over > 300 million years. The fifth shared DEG, LOC107981942, located on chromosome 1, is annotated as a zinc finger BED domain-containing protein 1. An NCBI Conserved Domain Search

(https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) using these protein sequences uncovered no significant hits with LOC100114950, LOC116415892, and LOC103317241. However, LOC107981401 and LOC107981942 show significant hits for transposase domain superfamilies cl24015 and cl04853, respectively. The role of these proteins in *Nasonia* remains unclear.

***Lack of Parent-of-Origin Effects in Nasonia Hybrids.*** We used allele-specific expression (ASE) analyses to detect parent-of-origin effects — indicated by allelic bias — in *Nasonia* hybrids. The inference of genomic imprinting for each dataset was limited to those sites that meet our filtering criteria (see Methods). We find 107,206 and 115,490 sites to be fixed and different between VV and GG samples, in the R16A Clark and Wilson datasets, respectively. Limiting the analysis to only fixed and different sites, there are 6,377 and 7,164 genes with at least 2 informative SNPs in the reciprocal hybrids in the R16A Clark data set and Wilson datasets, respectively. Using this approach, we find no evidence of genomic imprinting in whole adult female samples of *Nasonia* in the R16A Clark data (Figure 4A). But for the Wilson data we found two genes that show a pattern of expression consistent with genomic imprinting: CPR35 and LOC103315494. In the VG hybrid, CPR35 shows a bias towards the paternally inherited *N. giraulti* allele at an allele ratio of 65.3% and in the GV hybrid towards the paternally inherited *N. vitripennis* allele, with an allele ratio of 62% (Supplemental 2 Table). CPR35 is a cuticular protein in the RR family member 35. Similarly, LOC103315494 shows bias towards the paternally inherited allele with allele ratios of 65.26% and 61.58% in VG and GV, respectively (Supplemental 2 Table). Although both imprinted genes, *CPR35* and

*LOC103315494,* fall below the mean depth and average number of SNPs per gene, both genes are above the thresholds applied here (Supplemental 3 Table).



*Chapter 4. Figure 4. Lack of Parent-of-Origin Expression in $F_1$ Hybrids.* Scatterplots of the expression of the *N. vitripennis* alleles in the two reciprocal hybrids, VG (x-axis) and GV (y-axis), in the R16A Clark (A) and Wilson (B) datasets. Genes with at least two informative SNPs with a minimum depth of 30 were used (R16A Clark = 6,377, Wilson = 7,164). Genes exhibiting a significant difference in allelic bias between the hybrids (Fisher's exact test, FDR-adj. $p<0.05$) are highlighted in red. Paternally imprinted genes are expected to appear in the upper left corner (light blue box), and maternally imprinted genes in the lower right corner (light pink box). Histograms of the *N. vitripennis* allele expression are shown for VG (blue) and GV (pink).

***Allele-specific Expression Differences in Nasonia Hybrids.*** We find three genes with higher expression of the *N. vitripennis* allele in both hybrids, in both datasets, indicative of *cis*-regulatory effects. The genes LOC100123729, LOC100123734, and LOC100113683 show consistent differences in allelic expression between VG and GV hybrids (FDR-$p \leq 0.05$) in both datasets, but the ratio of the *N. vitripennis* allele differs between the hybrids (Supplemental 2 Table). In the R16A Clark dataset: LOC100123729 in the VG hybrids the *N. vitripennis* allele accounts for 93% of the reads, whereas in the GV hybrids this ratio is 61%. In the Wilson dataset, both hybrids showed higher expression of the *N. vitripennis* allele. In the Wilson data, the *N. vitripennis* allele ratio was 61% in VG and 90% in GV. LOC100123729 is located on chromosome 2 and encodes the protein Nasonin-3, which plays a role in inhibiting host insect melanization (Tian, Wang, Ye, & Zhu, 2010). Also on chromosome 2 is LOC100123734, annotated as cadherin-23, which is involved in cell attachment by interacting with other proteins in the cell membrane. Both hybrids in both datasets show a higher expression for the *N. vitripennis* allele for LOC100123734. In the R16A Clark data, the ratio of the *N. vitripennis* allele in VG was 92% and in GV 65%. In the Wilson data, the VG hybrids showed less expression for the *N. vitripennis* allele than the GV hybrids, at a ratio of 64% and 84% of the reads, respectively. Finally, LOC100113683, which is located on

chromosome 4, and is annotated as a general odorant-binding protein 56d also shows more expression for the *N. vitripennis* allele in both datasets and both hybrids (80.13% and 73.54% for VG and GV in R16A Clark, 78.22% and 72.57% in Wilson). Odorant binding proteins are thought to be involved in the stimulation of the odorant receptors by binding and transporting odorants which activate the olfactory signal transduction pathway (He et al., 2020).

***R16A Strain Retains N. vitripennis Alleles.*** R16A is a strain produced by backcrossing an *N. vitripennis* female to an *N. giraulti* male and repeating that for 16 generations (Breeuwer & Werren, 1995). This should give a complete *N. giraulti* nuclear genome with *N. vitripennis* mitochondria. However, we identified two regions in the R16A strain that still show *N. vitripennis* alleles and named them R16A non-introgressed locus 1 and R16A non-introgressed locus 2 (Supplemental 4 Table). Each region is identified by a single marker that retains the *N. vitripennis* allele. Locus 1 contains 44 genes and Locus 2 contains 14 genes. Both of these regions are found on Chromosome 1, and Locus 2 lies within the confidence intervals of the mortality locus for *N. vitripennis* maternity hybrids identified by Niehuis et al. 2008 (Niehuis et al., 2008) (i.e., F2 recombinant hybrids with a *N. vitripennis* cytoplasm showed a significant transmission ratio distortion at this region favoring the *N. vitripennis* allele). R16 A non-introgressed locus 1 harbors a mitochondrial ribosomal gene (39 S ribosomal protein 38) which is a good candidate gene for causing its retention in R16A despite intensive introgression. It would also explain the observed nuc-cytoplasmic effect in F2 recombinant males in a vitripennis cytoplasm, despite the fact that R16A was used as a giraulti parental line in Gadau et al. (1999) (Gadau et al., 1999). Gadau et al. interestingly also mapped one of

the nuc-cytoplasmic incompatibility loci to chromosome 1 (called LG1 in the manuscript) (Gadau et al., 1999). Mutations in mitochondrial ribosomal proteins in humans have severe effects (Sylvester, Fischel-Ghodsian, Mougey, & O'Brien, 2004).

***Expression of Genes in Regions Associated with Hybrid Mortality or Nuclear-mitochondrial Incompatibility.*** We compared the location of genes with either significant differential gene expression or significant differences in allele-specific expression between VG and GV hybrids to the location of previously identified mortality-associated loci. Three of the five genes that were called as differentially expressed between VG and GV hybrids in both the R16A Clark and Wilson data sets (Supplemental 5 Table) are located within mortality associated loci. LOC103317241 is located within a locus on Chr 2 that is associated with mortality in VG hybrids, and LOC107981401 and LOC100114950 are within a locus on Chr 4 that is associated with mortality in GV hybrids. Moreover, two of the three genes showing consistent allele-specific expression in the two data sets are located near one another in the mortality-associated locus on Chr. 2 (LOC100123729 and LOC100123734). None of the genes that are differentially expressed or that exhibit allele-specific expression are located within the 2 loci that retain the *N. vitripennis* genotype in the R16A Clark strain, nor did we find any overlap of these gene sets with either the oxidative phosphorylation or the mitochondrial ribosomal proteins.

## Discussion

We successfully replicate the findings from Wang et al. 2016, showing a lack of parent-of-origin effects in *Nasonia* transcriptomes (Wang et al., 2016). This replication

occurs independently in a different laboratory, with different *Nasonia* individuals derived from a slightly different cross, different bioinformatic pipelines, and sequencing technology. Our results from both the reanalyzed R16A Clark and Wilson datasets could only demonstrate a species-of-origin effect but no parent-of-origin effect within *Nasonia* F1 female hybrids, which may have explained the lack of mortality in the F1 females relative to the F2 recombinant hybrid males. The larger number of differentially expressed genes between the two parental species in our study relative to the Wang et al (2016) (Wang et al., 2016) (1001 vs 799) is most likely the result of using a standard *N. giraulti* strain (RV2Xu) rather than an introgression strain (R16A) where the nuclear genome of *N. giraulti* was introgressed into a *N. vitripennis* cytoplasm. Additionally, we found genomic regions that resisted introgression in the R16A *Nasonia* strains utilized by Wang et al. 2016 (Wang et al., 2016). Furthermore, we present a reproducible workflow for processing raw RNA sequence samples to call differential expression and allele-specific expression openly available on the GitHub page: https://github.com/SexChrLab/Nasonia.

***Differences Between the R16A Clark and Wilson Datasets.*** The primary difference between the R16A Clark cross and the Wilson cross is the *N. giraulti* strain choice Figure 1B. The new crosses presented here used the strain Rv2X(u), which is a pure *N. giraulti* strain that was used for sequencing the genome (Werren et al., 2010). Wang et al. 2016 used an introgression strain, R16A, which has a largely *N. giraulti* nuclear genome with an *N. vitripennis* cytoplasm (Wang et al., 2016). This strain was produced by mating an *N. vitripennis* female with an *N. giraulti* male, and then repeatedly backcrossing the strain to *N. giraulti* males for a further 15 generations

(Breeuwer & Werren, 1995). Hence, both sets of hybrids should be heterozygous at every nuclear locus for species specific markers (though see above for two non-introgressed regions); however, both reciprocal R16A Clark hybrids have *N. vitripennis* mitochondria while the new hybrids have their maternal species' mitochondria. This means that in addition to looking at parent-of-origin effects, our new crosses are uniquely suited to investigate allelic expression biases in the context of nuclear-mitochondrial incompatibility and hybrid dysfunction.

***Observed Differences in Hybrids Between Data Sets.*** We observe substantially more DEGs between the hybrids, VG and GV, in the Wilson data set compared to the R16A Clark data set. The smaller number of DEGs detected in the R16A Clark data in this particular comparison is likely partially due to the one excluded $F_1$GV sample (see Materials and methods). Another likely contributing factor is the differences in one parental strain between the Wilson and R16A Clark data sets. The Wilson data presented here consist of inbred parental *N. vitripennis* (strain AsymCX) VV and *N. giraulti* (strain RV2Xu) GG lines, and reciprocal F1 crosses. This cross differs from the R16A Clark data, which used the same *N. vitripennis* strain but rather than a normal *N. giraulti* strain they used the introgression strain, R16A, that has a nuclear genome derived from *N. giraulti* and a cytoplasm/mitochondria derived from *N. vitripennis* (see R16A section). Despite these differences, of the eight genes that are differentially expressed between the VG and GV hybrids. five are shared between both data sets. Although we were not specifically looking for this, we found that three of the five genes showing differential expression in both data sets as well as two of the three genes showing allele (species)-specific expression in both data sets are located in previously identified loci that are

associated with the observed F2 recombinant male hybrid breakdown from the same crosses (Gadau et al., 1999; Niehuis et al., 2008). These findings point towards an involvement of cis regulatory elements in the genetic architecture of the F2-hybrid male breakdown in *Nasonia*. The finding that, despite using different strains of wasps, we are still able to identify genes associated with these hybrid defects, which bolsters our confidence in further pursuing these genes in our investigation of the genetic architecture of hybrid barriers in *Nasonia*.

***The Choice of Reference and Tools Does Not Alter Main Findings.*** The authors of the Wang et al. 2016 paper used different computational tools for trimming and alignment than the current study (Wang et al., 2016). Additionally, in Wang et al. 2016, the RNAseq reads were aligned to both an *N. vitripennis* and *N. giraulti* reference genome (Wang et al., 2016); whereas here, we created a pseudo *N. giraulti* reference genome from the fixed and differentiated sites between the inbred *N. vitripennis* and *N. giraulti* parental lines. Often, different tools and statistical approaches result in different findings (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013; Schaarschmidt, Fischer, Zuther, & Hincha, 2020); however, despite different approaches, we observe the same pattern as what was originally reported in Wang et al. 2016 (Wang et al., 2016), a lack of parent-of-origin expression in *Nasonia.*

***A Reproducible Workflow for Investigating Genome Imprinting.*** Significant factors contributing to irreproducible research include selective reporting, unavailable code and methods, low statistical power, poor experimental design, and raw data not available from the original lab (Baker, 2016). We replicate a robust experimental design (current study) initially presented in the Wang et al. (2016) (Wang et al., 2016) and

present a new workflow for calling DE and ASE in those two independent but analog *Nasonia* datasets. Both datasets are publicly available for download on the short read archive (SRA) PRJNA260391 and PRJNA613065, respectively. In our analyses of the Wilson data and reanalysis of the R16A data, we corroborated the original findings from Wang et al. 2016 (Wang et al., 2016). There are no parent-of-origin effects in *Nasonia*. All dependencies for data processing are provided as a Conda environment, allowing for seamless replication. All code is openly available on GitHub https://github.com/SexChrLab/Nasonia.

## Materials and methods

***Nasonia vitripennis and Nasonia giraulti Inbred and Reciprocal F1 Hybrid Datasets.*** RNA sequence (RNAseq) samples for 4 female samples each from parental species, *N. vitripennis* (VV) and *N. giraulti* (GG), and from each reciprocal F1 cross ($F_1$VG, female hybrids with *N. vitripennis* mothers, and $F_1$GV, female hybrids with *N. giraulti* mothers), as shown in Figure 1A, were obtained from a 2016 publication (Wang et al., 2016) from SRA PRJNA299670. We refer to the data from (Wang et al., 2016) as R16A Clark. One $F_1$GV RNAseq sample from the R16A Clark dataset (SRR2773798) was excluded due to low quality, as in the original publication (Wang et al., 2016).

The newly generated crosses consisted of 12 RNAseq samples of inbred isofemale lines of parental *N. vitripennis* (strain AsymCX) VV and *N. giraulti* (strain RV2Xu) GG lines, and reciprocal F1 crosses $F_1$VG, and $F_1$GV. (Figure 1A). Whole transcriptome for these samples is available on SRA PRJNA613065. This cross differs from the R16A Clark data, which used the same *N. vitripennis* strain but rather than a

standard *N. giraulti* strain used an introgression strain, R16A, that has a nuclear genome derived from *N. giraulti* and a cytoplasm/mitochondria derived from *N. vitripennis* (see R16A section below) Figure 1B. Total RNA was extracted from a pool of four 48 hour post-eclosion adult females using a Qiagen RNeasy Plus Mini kit (Qiagen, CA). RNA-seq libraries were prepared with 2µg of total RNA using the Illumina Stranded mRNA library prep kit and were sequenced on a HiSeq2500 instrument following standard Illumina protocols. Three biological replicates were generated for each parent and hybrid, with 100-bp paired-end reads per replicate. Sample IDs, parent cross information, and SRA bioproject accession numbers for R16A Clark and Wilson datasets are listed in Supplemental 1 Table.

***Quality Control.*** Raw sequence data from both datasets were processed and analyzed according to the workflow presented in Figure 1C. The quality of the FASTQ files was assessed before and after trimming using FastQC v0.11 (Andrews, 2010) and MultiQC v1.0 (Ewels, Magnusson, Lundin, & Käller, 2016). Reads were trimmed to remove bases with a quality score less than 10 for the leading and trailing stand, applying a sliding window of 4 with a minimum mean PHRED quality of 15 in the window and a minimum read length of 80 bases, and adapters were removed using Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014). Pre- and post-trimming multiQC reports for the R16A Clark and Wilson datasets are available on the GitHub page: https://github.com/SexChrLab/Nasonia.

***Variant Calling.*** For variant calling, BAM files were preprocessed by adding read groups with Picard's AddOrReplaceReadGroups and by marking duplicates with Picard's MarkDuplicates (https://github.com/broadinstitute/picard). Variants were called using

GATK (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013) and the scatter-gather approach: Sample genotype likelihoods were called with HaplotypeCaller minimum base quality of 2. The resulting gVCFs were merged with CombineGVCFs, and joint genotyping across all samples was carried out with GenotypeGVCFs with a minimum confidence threshold of 10.

***Pseudo N. giraulti Reference Genome Assembly.*** To create a pseudo *N. giraulti* reference genome, fixed differences in the homozygous *N. giraulti* and *N. vitripennis* variant call file (VCF) files were identified using a custom Python script, available on the GitHub page: https://github.com/SexChrLab/Nasonia. Briefly, a site was considered to be fixed and different if it was homozygous for the *N. vitripennis* reference allele among all three of the biological VV samples and homozygous alternate among all three of the biological GG samples. Only homozygous sites were included, as the *N. giraulti* and *N. vitripennis* lines are highly inbred. The filtered sites were then used to create a pseudo *N. giraulti* reference sequence with the FastaAlternateReferenceMaker function in GATK version 3.8 (available at: http://www.broadinstitute.org/gatk/). Reference bases in the *N. vitripennis* genome were replaced with the alternate SNP base at variant positions. Following a similar protocol for comparison, we now aligned reads in each sample to the pseudo *N. giraulti* genome reference with HISAT2 version 2.1.0, and performed identical preprocessing steps prior to variant calling with GATK version 3.8 HaplotypeCaller.

***RNAseq Alignment and Gene Expression Level Quantification.*** Trimmed sequence reads were mapped to the NCBI *N. vitripennis* reference genome (assembly accession GCF_009193385.2), as well as the pseudo *N. giraulti* reference using HISAT2 (Kim, Langmead, & Salzberg, 2015). The resulting SAM sequence alignment files were

converted to BAM, and coordinates were sorted and indexed with samtools 1.8 (Li et al., 2009). RNAseq read counts were quantified from the *N. vitripennis* as well as the custom *N. giraulti* alignments using Subread featureCounts (Liao, Smyth, & Shi, 2014) with the *N. vitripennis* gene annotation.

***Inference of Differential Gene Expression.*** Differential expression (DE) analyses were carried out by linear modeling as implemented in the R package *limma* (Ritchie et al., 2015). An average of the reads mapped to each gene in the *N. vitripennis* and the pseudo *N. giraulti* genome references were used in the DE analyses. Counts were filtered to remove lowly expressed genes by retaining genes with a mean FPKM $\leq 0.5$ in at least one sample group (VV, GG, VG, or GV). Normalization of expression estimates was accomplished by calculating the trimmed mean of M-values (TMM) with edgeR (Robinson, McCarthy, & Smyth, 2010). The voom method (Law, Chen, Shi, & Smyth, 2014) was then employed to normalize expression intensities by generating a weight for each observation. Gene expression is then reported as log counts per million (logCPM). Gene expression correlation between datasets and between species within each dataset was assessed using Pearson's correlation of mean logCPM values of each gene. Dimensionality reduction of the filtered and normalized gene expression data was carried out using scaled and centered PCA with the *prcomp*() function of base R. Differential expression analysis with voom was carried out for each pairwise comparison between strains (VV, GG, VG, and GV) for each data set. We identified genes that exhibited significant expression differences with an adjusted *p*-value of $\leq 0.01$ and an absolute $\log_2$ fold-change ($\log_2$FC) $\leq 2$.

***Analysis of Allele-specific Expression in Reciprocal F1 Hybrids.*** Allele-specific expression (ASE) levels were obtained using GATK ASEReadCounter (McKenna et al., 2010) with a minimum mapping quality of 10, minimum base quality of 2, and a minimum depth of 30. Only sites with a fixed difference between inbred VV and GG for both R16A Clark and Wilson datasets were used for downstream analysis of allele-specific expression. Allele counts obtained from GATK ASEReadCounter were intersected with the *N. vitripennis* gene annotation file using bedtools version 2.24.0 (Quinlan & Hall, 2010); the resulting output contained allele counts for each SNP and corresponding gene information. The F1 hybrids' allele counts with gene information was read into R and then filtered to only include genes with at least two SNPs with minimum depth of 30. We counted the number of allele-counts for the reference allele (*N. vitripennis*) and alternative (*N. giraulti*) allele at polymorphic SNP positions. We quantified the number of SNPs in each hybrid replicate that 1) showed a bias towards the allele that came from the *N. vitripennis* parent, 2) showed a bias towards the allele that came from the *N. giraulti* parent, and 3) showed no difference (ND) in an expression of its parental alleles. The significance of allelic bias was determined using Fisher's exact test. Significant genes were selected using a Benjamini-Hochberg false discovery rate FDR-adjusted *p*-value threshold of 0.05. As *Nasonia* are haplodiploid, all ASE analyses were carried out on the diploid female hybrids.

***Identifying Loci Associated with Hybrid Mortality.*** *Nasonia* recombinant F2 hybrid males (haploid sons of F1 female hybrids) suffer mortality during development that differs between VG and GV hybrids (Breeuwer & Werren, 1995). Niehuis et al. 2008 identified four genomic regions associated with this mortality (i.e., regions in which one

parent species' alleles are underrepresented due to mortality during development); three are associated with mortality in hybrids with *N. vitripennis* maternity and one is associated with hybrids with *N. giraulti* maternity (Niehuis et al., 2008). Gibson et al. 2013 later identified a second locus related to mortality in the hybrids with *N. giraulti* maternity (J. D. Gibson et al., 2013). Given that the F2 hybrid females analyzed here experience far less mortality than their haploid male offspring, we hypothesized that these diploid females may use biased allelic expression to rescue themselves from the mortality. To compare our results with these previous studies, we had to map the previous loci to the latest *Nasonia* assembly (PSR1.1, (Benetta et al., n.d.)). Niehuis et al. 2008 defined their candidate loci based on the genetic distance along the chromosome (centimorgans) (Niehuis et al., 2008). The physical locations of the markers along the chromosomes were later identified by Niehius et al. 2010 (Niehuis et al., 2010). Using the genetic distances between these markers in both the 2008 and 2010 Niehuis *et al*. studies (Niehuis et al., 2010, 2008), we calculated the conversion ratio between the genetic distances in these two studies (Supplemental 6 Table). We then converted those 2008 genetic distances that correspond to the 95% Confidence Intervals for these loci to the genetic distances reported by Niehuis et al. 2010 (Niehuis et al., 2010), which used an Illumina Goldengate Genotyping Array (Illumina Inc., San Diego, USA) to produce a more complete and much higher resolution genetic map of *Nasonia*. This array uses Single Nucleotide Polymorphisms (SNPs) to genotype samples at ~1500 loci, which allowed us to identify SNP markers that closely bound the mortality loci from the 2008 study. Gibson et al. 2013 used the same genotyping array, so this conversion was unnecessary for converting the second mortality locus in *N. giraulti* maternity hybrids (J.

D. Gibson et al., 2013). We used the 100bp of sequence flanking each SNP marker to perform a BLAST search of the PSR1.1 assembly and to identify their positions. We then used all of the PSR1.1 annotated genes within these loci to look for enrichment of genes showing biased expression. Mortality loci and genomic location are reported in Supplemental 4 Table.

***Additional Gene Categories of Interest.*** Previous work has identified potential classes of genes that may be involved in nuclear-mitochondrial incompatibilities in *Nasonia,* the oxidative phosphorylation genes (Joshua D. Gibson, Niehuis, Verrelli, & Gadau, 2010) and the mitochondrial ribosomal proteins (Burton & Barreto, 2012). We used the annotated gene sets from these studies to test for enrichment of genes with biased allelic expression. Lists of the genes of interest and their genomic location is reported in Supplemental 4 Table.

***Analysis of R16A Strain.*** In order to assess whether the introgression of the *N. giraulti* nuclear genome into the R16A Clark strain is complete, we analyzed two samples of the R16A strain using the Illumina Goldengate Genotyping Array used in Niehuis *et al*. 2010 (Niehuis et al., 2010). We searched for SNP markers that retained the *N. vitripennis* allele and only considered markers that consistently identified the proper allele in both parent species controls and that were consistent across both R16A samples, leaving 1378 markers. We defined a locus as all of the sequences between the two markers that flank a marker showing the *N. vitripennis* allele (Supplemental 2 Table). As above, we performed a BLAST search of the PSR1.1 assembly to identify the positions of these markers. We identified all genes from the PSR1.1 assembly that lie between the flanking markers and further analyzed their expression patterns.

Scripts and gene lists used to analyze these data are publicly available on GitHub, https://github.com/SexChrLab/Nasonia.

## Supplementary Information

Supplemental tables and figures are located in chapter 4. appendices D.

CHAPTER 5

# Conclusions

## Major Contributions of Dissertation

***Chapter 1. A Sex Chromosome Complement Alignment Approach.*** We inferred if an RNAseq sample has Y chromosome expression by aligning the reads to a reference genome that includes both the X and Y chromosome. If a sample was determined to not have a Y chromosome, we aligned that sample to a reference with the Y chromosome hard masked with Ns. Sample's determined to have a Y chromosome were aligned to a reference with the Y PARs hard masked. Using the sex chromosome complement approach compared to a default alignment that includes both the X and Y chromosomes, we observe an increase in X chromosome expression estimates in both female XX and male XY samples. We urge studies using RNA-Seq to carefully consider the genetic sex of the sample when quantifying reads and we provide a framework for doing so in the future (https://github.com/SexChrLab/XY_RNAseq).

***Chapter 2. Characterization of Sex Differences in Gene Expression in Human Placentas.*** We find that there are sex differences in gene expression in uncomplicated term placentas. However, we observe that gene expression for innate immune genes is not significantly different between the sexes in term (greater than or equal to 36.6 weeks), uncomplicated placentas. We show that most sex differentially expressed genes in term placentas are located on the sex chromosomes, X and Y. Further, we find that the female-to-male expression ratio for sex differentially expressed genes, autosomal and sex-linked genes, is correlated between term placentas, late first trimester placentas, and adult

tissues. Finally, we show that the correlation of female-to-male expression ratio between term, late first trimester placentas, and adult tissues is strongest for sex-linked genes. This suggests that sex differences in gene expression on the sex chromosomes, X and Y, are replicated across the life span. Code available at: https://github.com/SexChrLab/Placenta_Sex_Diff.

***Chapter 3. Breast Cancer in Response to Synthetic Histone-binding Regulator Protein.*** We observe 19 tumor suppressor genes become up-regulated in response to the treatment across three distinct breast cancer cell lines. We demonstrate a chromatin-mediated transcriptional response driven by an engineered fusion protein that physically links repressive histone marks with active transcription. Our results have implications for breast cancer treatment by up-regulating tumor suppressor genes. Code available at: https://github.com/WilsonSayresLab/PcTF_differential_expression.

***Chapter 4 Lack of Parent-Of-Origin Expression in Nasonia Jewel Wasp: a Replication and Extension Study.*** We show a lack of parent-of-origin expression within Nasonia hybrids. Nasonia hybrids do not show expression bias for the maternal or paternal derived allele, instead we observed species-of-origin expression. We replicate the results initially presented in Wang et al. 2016 that show a lack of parent-of-origin expression within Nasonia hybrids and we extend the findings using similar but different Nasonia hybrids to look at gene expression differences between hybrids. We offer a reproducible workflow for investigating genomic imprinting (https://github.com/SexChrLab/Nasonia).

# REFERENCES


Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., & Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research, 44(W1), W3–W10.

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., … Flicek, P. (2017). Ensembl 2017. Nucleic Acids Research, 45(D1), D635–D642. https://doi.org/10.1093/nar/gkw1104

Akishiba, M., Takeuchi, T., Kawaguchi, Y., Sakamoto, K., Yu, H.-H., Nakase, I., Takatani-Nakase, T., Madani, F., Gräslund, A., & Futaki, S. (2017). Cytosolic antibody delivery by lipid-sensitive endosomolytic peptide. Nature Chemistry, 9(8), 751–761.

Alaux, C., Sinha, S., Hasadsri, L., Hunt, G. J., Guzmán-Novoa, E., DeGrandi-Hoffman, G., … Robinson, G. E. (2009). Honey bee aggression supports a link between gene regulation and behavioral evolution. Proceedings of the National Academy of Sciences of the United States of America, 106(36), 15400–15405.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). An Overview of Gene Control. Garland Science.

Albrecht, K. H., Young, M., Washburn, L. L., & Eicher, E. M. (2003). Sry expression level and protein isoform differences play a role in abnormal testis development in C57BL/6J mice carrying certain Sry alleles. Genetics, 164(1), 277–288.

Alford, S. H., Toy, K., Merajver, S. D., & Kleer, C. G. (2012). Increased risk for distant metastasis in patients with familial early-stage breast cancer and high EZH2 expression. Breast Cancer Research and Treatment, 132(2), 429–437.

Almassalha, L. M., Bauer, G. M., Wu, W., Cherkezyan, L., Zhang, D., Kendra, A., Gladstein, S., Chandler, J. E., VanDerway, D., Seagle, B.-L. L., Ugolkov, A., Billadeau, D. D., O'Halloran, T. V., Mazar, A. P., Roy, H. K., Szleifer, I., Shahabi, S., & Backman, V. (2017). Macrogenomic engineering via modulation of the scaling of chromatin packing density. Nature Biomedical Engineering, 1(11), 902–913.

Alur, Pradeep. 2019. "Sex Differences in Nutrition, Growth, and Metabolism in Preterm Infants." Frontiers in Pediatrics 7 (February): 22.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Andrews, S. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].” Online. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Angum, F., Khan, T., Kaler, J., Siddiqui, L., & Hussain, A. (2020). The Prevalence of Autoimmune Disorders in Women: A Narrative Review. Cureus, 12(5), e8094.

Arnold, A. P., & Chen, X. (2009). What does the “four core genotypes” mouse model tell us about sex differences in the brain and other tissues? Frontiers in Neuroendocrinology, 30(1), 1–9. https://doi.org/10.1016/j.yfrne.2008.11.001

Arnold, A. P., Chen, X., & Itoh, Y. (2012). What a difference an X or Y makes: Sex chromosomes, gene dose, and epigenetics in sexual differentiation. Handbook of Experimental Pharmacology, 214, 67–88. https://doi.org/10.1007/978-3-642-30726-3_4

Ayers, K. L., Davidson, N. M., Demiyah, D., Roeszler, K. N., Grützner, F., Sinclair, A. H., … Smith, C. A. (2013). RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. Genome Biology, Vol. 14, p. R26. doi:10.1186/gb-2013-14-3-r26

Ayers, K. L., Davidson, N. M., Demiyah, D., Roeszler, K. N., Grützner, F., Sinclair, A. H., Oshlack, A., & Smith, C. A. (2013). RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. In Genome Biology (Vol. 14, Issue 3, p. R26). https://doi.org/10.1186/gb-2013-14-3-r26

Baker, M. (2015). Irreproducible biology research costs put at $28 billion per year. Nature, 533. Retrieved from http://www.target-biomed.de/resources/Irreproducible-biology-research.pdf

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, Vol. 533, pp. 452–454. doi:10.1038/533452a

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics , 27(12), 1691–1692.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., & Marth, G. T. (2011). BamTools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics, 27(12), 1691–1692. https://doi.org/10.1093/bioinformatics/btr174

Barnett, Derek W., Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth. 2011. “BamTools: A C++ API and Toolkit for Analyzing and Managing BAM Files.” Bioinformatics 27 (12): 1691–92.

Beltran, A., Parikh, S., Liu, Y., Cuevas, B. D., Johnson, G. L., Futscher, B. W., & Blancafort, P. (2007). Re-activation of a dormant tumor suppressor gene maspin by designed transcription factors. Oncogene, 26(19), 2791–2798.

Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., & Weinberg, R. A. (2008). An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nature Genetics, 40(5), 499–507.

Benetta, E. D., Antoshechkin, I., Yang, T., Nguyen, H. Q. M., Ferree, P. M., & Akbari, O. S. (n.d.). Genome Elimination Mediated by Gene Expression from a Selfish Chromosome. doi:10.1101/793273

Berrozpe, G., Bryant, G. O., Warpinski, K., Spagna, D., Narayan, S., Shah, S., & Ptashne, M. (2017). Polycomb Responds to Low Levels of Transcription. Cell Reports, 20(4), 785–793.

Beukeboom, L. W., & van de Zande, L. (2010). Genetics of sex determination in the haplodiploid wasp Nasonia vitripennis (Hymenoptera: Chalcidoidea). Journal of Genetics, 89(3), 333–339.

Biancotto, C., Frigè, G., & Minucci, S. (2010). Histone Modification Therapy of Cancer. In Advances in Genetics (pp. 341–386).

Blanch, Alvaro, Olga Roche, Eduardo López-Granados, Gumersindo Fontán, and Margarita López-Trascasa. 2003. “Erratum: Detection of C1 Inhibitor (SERPING1/C1NH) Mutations in Exon 8 in Patients with Hereditary Angioedema: Evidence for 10 Novel Mutations.” Human Mutation. https://doi.org/10.1002/humu.9105.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics , 30(15), 2114–2120.

Bouker, K. B., Skaar, T. C., Riggins, R. B., Harburger, D. S., Fernandez, D. R., Zwart, A., Wang, A., & Clarke, R. (2005). Interferon regulatory factor-1 (IRF-1) exhibits tumor suppressor activities in breast cancer associated with caspase activation and induction of apoptosis. Carcinogenesis, 26(9), 1527–1535.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Vidal, M., Gifford, D. K., Young, R. A., & Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature, 441(7091), 349–353.

Bracken, A. P., & Helin, K. (2009). Polycomb group proteins: navigators of lineage pathways led astray in cancer. Nature Reviews. Cancer, 9(11), 773–784.

Bray, N., Pimentel, H., Melsted, P., & Pachter, L. (2015). Near-optimal RNA-Seq quantification. ArXiv:1505.02710 [Cs, q-Bio]. http://arxiv.org/abs/1505.02710

Breeuwer, J. A. J., & Werren, J. H. (1995). HYBRID BREAKDOWN BETWEEN TWO HAPLODIPLOID SPECIES: THE ROLE OF NUCLEAR AND CYTOPLASMIC GENES. Evolution; International Journal of Organic Evolution, 49(4), 705–717.

Breuer, Karin, Amir K. Foroushani, Matthew R. Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L. Winsor, Robert E. W. Hancock, Fiona S. L. Brinkman, and David J. Lynn. 2012. "InnateDB: Systems Biology of Innate Immunity and beyond—Recent Updates and Continuing Curation." Nucleic Acids Research 41 (D1): D1228–33.

Bridgman, S. L., M. B. Azad, R. R. Persaud, R. S. Chari, A. B. Becker, M. R. Sears, P. J. Mandhane, et al. 2018. "Impact of Maternal Pre-Pregnancy Overweight on Infant Overweight at 1 Year of Age: Associations and Sex-Specific Differences." Pediatric Obesity. https://doi.org/10.1111/ijpo.12291.

Broadinstitute/picard. (2020). [Java]. Broad Institute. https://github.com/broadinstitute/picard (Original work published 2014)

Brocks, D., Schmidt, C. R., Daskalakis, M., Jang, H. S., Shah, N. M., Li, D., Li, J., Zhang, B., Hou, Y., Laudato, S., Lipka, D. B., Schott, J., Bierhoff, H., Assenov, Y., Helf, M., Ressnerova, A., Islam, M. S., Lindroth, A. M., Haas, S., … Plass, C. (2017). DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nature Genetics, 49(7), 1052–1060.

Broere-Brown, Z. A., Adank, M. C., Benschop, L., Tielemans, M., Muka, T., Gonçalves, R., … Schalekamp-Timmermans, S. (2020). Fetal sex and maternal pregnancy outcomes: a systematic review and meta-analysis. Biology of Sex Differences, 11(1), 26.

Burton, R. S., & Barreto, F. S. (2012). A disproportionate role for mtDNA in Dobzhansky-Muller incompatibilities? Molecular Ecology, Vol. 21, pp. 4942–4957. doi:10.1111/mec.12006

Bushnell, Brian. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner." LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166-bbmap-fast-accurate-splice-aware-aligner.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

Carithers, Latarsha J., Kristin Ardlie, Mary Barcus, Philip A. Branton, Angela Britton, Stephen A. Buia, Carolyn C. Compton, et al. 2015. “A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project.” Biopreservation and Biobanking 13 (5): 311–19.

Carrel, L., & Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature, 434(7031), 400–404. https://doi.org/10.1038/nature03479

Casadevall, A., & Fang, F. C. (2010). Reproducible science. Infection and Immunity, 78(12), 4972–4975.

Cavalli, G., & Paro, R. (1998). The Drosophila Fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. Cell, 93(4), 505–518.

Chaligné, R., Popova, T., Mendoza-Parra, M.-A., Saleem, M.-A. M., Gentien, D., Ban, K., Piolot, T., Leroy, O., Mariani, O., Gronemeyer, H., Vincent-Salomon, A., Stern, M.-H., & Heard, E. (2015). The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. Genome Research, 25(4), 488–503. https://doi.org/10.1101/gr.185926.114

Chan, K. M., Han, J., Fang, D., Gan, H., & Zhang, Z. (2013). A lesson learned from the H3.3K27M mutation found in pediatric glioma: a new approach to the study of the function of histone modifications in vivo? Cell Cycle , 12(16), 2546–2552.

Chan, K.-M., Fang, D., Gan, H., Hashizume, R., Yu, C., Schroeder, M., Gupta, N., Mueller, S., James, C. D., Jenkins, R., Sarkaria, J., & Zhang, Z. (2013). The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. Genes & Development, 27(9), 985–990.

Chang, C.-J., Yang, J.-Y., Xia, W., Chen, C.-T., Xie, X., Chao, C.-H., Woodward, W. A., Hsu, J.-M., Hortobagyi, G. N., & Hung, M.-C. (2011). EZH2 promotes expansion of breast tumor initiating cells through activation of RAF1-β-catenin signaling. Cancer Cell, 19(1), 86–100.

Charchar, F. J., Svartman, M., El-Mogharbel, N., Ventura, M., Kirby, P., Matarazzo, M. R., Ciccodicola, A., Rocchi, M., D’Esposito, M., & Graves, J. A. M. (2003). Complex Events in the Evolution of the Human Pseudoautosomal Region 2 (PAR2). Genome Research, 13(2), 281–286. https://doi.org/10.1101/gr.390503

Charlesworth, B. (1991). The evolution of sex chromosomes. Science (New York, N.Y.), 251(4997), 1030–1033. https://doi.org/10.1126/science.1998119

Chelbi, Sonia T., Françoise Mondon, Hélène Jammes, Christophe Buffat, Thérèse-Marie Mignot, Jorg Tost, Florence Busato, et al. 2007. “Expressional and Epigenetic Alterations of Placental Serine Protease Inhibitors.” Hypertension. https://doi.org/10.1161/01.hyp.0000250831.52876.cb.

Chen, Chia-Yu, Chieh-Hsiang Chan, Chun-Ming Chen, Yin-Shuan Tsai, Tsung-Yuan Tsai, Yan-Hwa Wu Lee, and Li-Ru You. 2016. “Targeted Inactivation of Murine Ddx3x: Essential Roles of Ddx3x in Placentation and Embryogenesis.” Human Molecular Genetics 25 (14): 2905–22.

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics, 12, 35.

Chial, H. (2008). Genetic regulation of cancer. Nature Education, 1(1), 67.

Chiappinelli, K. B., Strissel, P. L., Desrichard, A., Li, H., Henke, C., Akman, B., Hein, A., Rote, N. S., Cope, L. M., Snyder, A., Makarov, V., Budhu, S., Buhu, S., Slamon, D. J., Wolchok, J. D., Pardoll, D. M., Beckmann, M. W., Zahnow, C. A., Merghoub, T., … Strick, R. (2015). Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. Cell, 162(5), 974–986.

Ching, T., Huang, S., & Garmire, L. X. (2014). Power analysis and sample size estimation for RNA-Seq differential expression. RNA (New York, N.Y.), 20(11), 1684–1696. https://doi.org/10.1261/rna.046011.114

Classon, M., LaMarco, K., & De Carvalho, D. D. (2017). Drug-induced activation of “junk” DNA - A path to combat cancer therapy resistance? Oncoscience, 4(9-10), 115–116.

Clifton, V. L. 2010. “Review: Sex and the Human Placenta: Mediating Differential Strategies of Fetal Growth and Survival.” Placenta. https://doi.org/10.1016/j.placenta.2009.11.010.

Collett, K., Eide, G. E., Arnes, J., Stefansson, I. M., Eide, J., Braaten, A., Aas, T., Otte, A. P., & Akslen, L. A. (2006). Expression of enhancer of zeste homologue 2 is significantly associated with increased tumor cell proliferation and is a marker of aggressive breast cancer. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 12(4), 1168–1174.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Consortium, T. Gte. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, 348(6235), 648–660. https://doi.org/10.1126/science.1262110

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. PloS One, 12(12), e0190152. https://doi.org/10.1371/journal.pone.0190152

Cuestas, Eduardo, Jose Bas, and Josefina Pautasso. 2009. “Sex Differences in Intraventricular Hemorrhage Rates among Very Low Birth Weight Newborns.” Gender Medicine 6 (2): 376–82.

Dawson, M. A., & Kouzarides, T. (2012). Cancer epigenetics: from mechanism to therapy. Cell, 150(1), 12–27.

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. PloS One, 8(12), e85024.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data.” Nature Genetics 43 (5): 491–98.

Derfoul, A., Juan, A. H., Difilippantonio, M. J., Palanisamy, N., Ried, T., & Sartorelli, V. (2011). Decreased microRNA-214 levels in breast cancer cells coincides with increased cell proliferation, invasion and accumulation of the Polycomb Ezh2 methyltransferase. Carcinogenesis, 32(11), 1607–1614.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England), 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dong, M., Fan, X.-J., Chen, Z.-H., Wang, T.-T., Li, X., Chen, J., Lin, Q., Wen, J.-Y., Ma, X.-K., Wei, L., Ruan, D.-Y., Lin, Z.-X., Liu, Q., Wu, X.-Y., & Wan, X.-B. (2014). Aberrant expression of enhancer of zeste homologue 2, correlated with HIF-1α, refines relapse risk and predicts poor outcome for breast cancer. Oncology Reports, 32(3), 1101–1107.

Dorak, M. T., & Karpuzoglu, E. (2012). Gender differences in cancer susceptibility: an inadequately addressed issue. Frontiers in Genetics, 3, 268.

Du, Ming-Lun Kang, Guo-Liang Xu, et al. 2017. “The Role of N-α-Acetyltransferase 10 Protein in DNA Methylation and Genomic Imprinting.” Molecular Cell 68 (1): 89-103.e7.

Dumanski, J. P., Lambert, J.-C., Rasi, C., Giedraitis, V., Davies, H., Grenier-Boley, B., Lindgren, C. M., Campion, D., Dufouil, C., European Alzheimer’s Disease Initiative Investigators, Pasquier, F., Amouyel, P., Lannfelt, L., Ingelsson, M., Kilander, L., Lind, L., & Forsberg, L. A. (2016). Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. American Journal of Human Genetics, 98(6), 1208–1219. https://doi.org/10.1016/j.ajhg.2016.05.014

Dunn, J., & Rao, S. (2017). Epigenetics and immunotherapy: The current state of play. Molecular Immunology, 87, 227–239.

Easwaran, H., Johnstone, S. E., Van Neste, L., Ohm, J., Mosbruger, T., Wang, Q., Aryee, M. J., Joyce, P., Ahuja, N., Weisenberger, D., Collisson, E., Zhu, J., Yegnasubramanian, S., Matsui, W., & Baylin, S. B. (2012). A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Research, 22(5), 837–849.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48. https://doi.org/10.1186/1471-2105-10-48

Eden, Eran, Doron Lipson, Sivan Yogev, and Zohar Yakhini. 2007. “Discovering Motifs in Ranked Lists of DNA Sequences.” PLoS Computational Biology 3 (3): e39.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. 2009. “GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists.” BMC Bioinformatics 10 (February): 48.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57–74.

Essafi, M., Baudot, A. D., Mouska, X., Cassuto, J.-P., Ticchioni, M., & Deckert, M. (2011). Cell-penetrating TAT-FOXO3 fusion proteins induce apoptotic cell death in leukemic cells. Molecular Cancer Therapeutics, 10(1), 37–46.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics , 32(19), 3047–3048.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016a. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." Bioinformatics 32 (19): 3047–48.

Exploring the biological contributions to human health: does sex matter? (2001). Journal of Women's Health & Gender-Based Medicine, 10(5), 433–439.

Eyras, E., Caccamo, M., Curwen, V., & Clamp, M. (2004). ESTGenes: Alternative Splicing From ESTs in Ensembl. Genome Research, 14(5), 976–987. https://doi.org/10.1101/gr.1862204

Falke, D., Fisher, M., Ye, D., & Juliano, R. L. (2003). Design of artificial transcription factors to selectively regulate the pro-apoptotic bax gene. Nucleic Acids Research, 31(3), e10.

Farooqi, Aijaz, Bruno Hägglöf, Gunnar Sedin, Leif Gothefors, and Fredrik Serenius. 2006. "Chronic Conditions, Functional Limitations, and Special Health Care Needs in 10- to 12-Year-Old Children Born at 23 to 25 Weeks' Gestation in the 1990s: A Swedish National Prospective Follow-up Study." Pediatrics 118 (5): e1466-77.

Fiszer-Kierzkowska, A., Vydra, N., Wysocka-Wycisk, A., Kronekova, Z., Jarząb, M., Lisowska, K. M., & Krawczyk, Z. (2011). Liposome-based DNA carriers may induce cellular stress response and change gene expression pattern in transfected cells. BMC Molecular Biology, 12, 27.

Forsberg, L. A. (2017). Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men. Human Genetics, 136(5), 657–663. https://doi.org/10.1007/s00439-017-1799-2

Freedman, L. P., & Inglese, J. (2014). The Increasing Urgency for Standards in Basic Biologic Research. Cancer Research, Vol. 74, pp. 4024–4029. doi:10.1158/0008-5472.can-14-0925

Fujiwara, T., Saitoh, H., Inoue, A., Kobayashi, M., Okitsu, Y., Katsuoka, Y., Fukuhara, N., Onishi, Y., Ishizawa, K., Ichinohasama, R., & Harigae, H. (2014). 3-Deazaneplanocin A (DZNep), an inhibitor of S-adenosylmethionine-dependent methyltransferase, promotes erythroid differentiation. The Journal of Biological Chemistry, 289(12), 8121–8134.

Fürstenberger, G., Krieg, P., Müller-Decker, K., & Habenicht, A. J. R. (2006). What are cyclooxygenases and lipoxygenases doing in the driver's seat of carcinogenesis? International Journal of Cancer, 119(10), 2247–2254.

Gadau, J., Page, R. E., Jr, & Werren, J. H. (1999). Mapping of hybrid incompatibility loci in Nasonia. Genetics, 153(4), 1731–1741.

Galbraith, D. A., Kocher, S. D., Glenn, T., Albert, I., Hunt, G. J., Strassmann, J. E., … Grozinger, C. M. (2016). Testing the kinship theory of intragenomic conflict in honey bees (Apis mellifera). Proceedings of the National Academy of Sciences of the United States of America, 113(4), 1020–1025.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., & Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Molecular Cell, 45(3), 344–356.

Gershoni, M., & Pietrokovski, S. (2017). The landscape of sex-differential transcriptome and its consequent selection in human adults. BMC Biology, 15(1), 7. https://doi.org/10.1186/s12915-017-0352-z

Gibson, J. D., Niehuis, O., Peirson, B. R. E., Cash, E. I., & Gadau, J. (2013). Genetic and developmental basis of F2 hybrid breakdown in Nasonia parasitoid wasps. Evolution; International Journal of Organic Evolution, 67(7), 2124–2132.

Gibson, Joshua D., Arechavaleta-Velasco, M. E., Tsuruda, J. M., & Hunt, G. J. (2015). Biased Allele Expression and Aggression in Hybrid Honeybees may be Influenced by Inappropriate Nuclear-Cytoplasmic Signaling. Frontiers in Genetics, Vol. 6. doi:10.3389/fgene.2015.00343

Gibson, Joshua D., Niehuis, O., Verrelli, B. C., & Gadau, J. (2010). Contrasting patterns of selective constraints in nuclear-encoded genes of the oxidative phosphorylation pathway in holometabolous insects and their possible role in hybrid breakdown in Nasonia. Heredity, 104(3), 310–317.

Global Pregnancy Collaboration:, Sarah Schalekamp-Timmermans, Lidia R. Arends, Elin Alsaker, Lucy Chappell, Stefan Hansson, Nina K. Harsem, et al. 2017. "Fetal Sex-Specific Differences in Gestational Age at Delivery in Pre-Eclampsia: A Meta-Analysis." International Journal of Epidemiology 46 (2): 632–42.

Global Pregnancy Collaboration:, Schalekamp-Timmermans, S., Arends, L. R., Alsaker, E., Chappell, L., Hansson, S., … Steegers, E. A. (2017). Fetal sex-specific differences in gestational age at delivery in pre-eclampsia: a meta-analysis. International Journal of Epidemiology, 46(2), 632–642.

Godfrey, Alexander K., Sahin Naqvi, Lukáš Chmátal, Joel M. Chick, Richard N. Mitchell, Steven P. Gygi, Helen Skaletsky, and David C. Page. 2020a. “Quantitative Analysis of Y-Chromosome Gene Expression

Goldstein, J. M., Holsen, L., Handa, R., & Tobet, S. (2014). Fetal hormonal programming of sex differences in depression: Linking women’s mental health with sex differences in the brain across the lifespan. Frontiers in Neuroscience, 8. https://doi.org/10.3389/fnins.2014.00247

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., & Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biology, 14(7), R70.

Gonzalez, Tania L., Tianyanxin Sun, Alexander F. Koeppel, Bora Lee, Erica T. Wang, Charles R. Farber, Stephen S. Rich, et al. 2018a. “Sex Differences in the Late First Trimester Human Placenta Transcriptome.” Biology of Sex Differences 9 (1): 4.

Goodspeed, A., Heiser, L. M., Gray, J. W., & Costello, J. C. (2016). Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. Molecular Cancer Research: MCR, 14(1), 3–13.

Grassmann, F., Kiel, C., den Hollander, A. I., Weeks, D. E., Lotery, A., Cipriani, V., Weber, B. H. F., & International Age-related Macular Degeneration Genomics Consortium (IAMDGC). (2019). Y chromosome mosaicism is associated with age-related macular degeneration. European Journal of Human Genetics: EJHG, 27(1), 36–41. https://doi.org/10.1038/s41431-018-0238-8

Greif, J. M., Pezzi, C. M., Klimberg, V. S., Bailey, L., & Zuraek, M. (2012). Gender differences in breast cancer: analysis of 13,000 breast cancers in men from the National Cancer Data Base. Annals of Surgical Oncology, 19(10), 3199–3204.

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (New York, N.Y.), 348(6235), 648–660. https://doi.org/10.1126/science.1262110

Guo, X., Xiao, H., Guo, S., Dong, L., & Chen, J. (2017). Identification of breast cancer mechanism based on weighted gene coexpression network analysis. Cancer Gene Therapy. https://doi.org/10.1038/cgt.2017.23

Haig, D. (1992). Intragenomic conflict and the evolution of eusociality. Journal of Theoretical Biology, 156(3), 401–403.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., …

Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research, 22(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Haynes, K. A., & Silver, P. A. (2011). Synthetic reversal of epigenetic silencing. The Journal of Biological Chemistry, 286(31), 27176–27182.

Haynes, K. A., Leibovitch, B. A., Rangwala, S. H., Craig, C., & Elgin, S. C. R. (2004). Analyzing heterochromatin formation using chromosome 4 of Drosophila melanogaster. Cold Spring Harbor Symposia on Quantitative Biology, 69, 267–272.

He, Y., Wang, K., Zeng, Y., Guo, Z., Zhang, Y., Wu, Q., & Wang, S. (2020). Analysis of the antennal transcriptome and odorant-binding protein expression profiles of the parasitoid wasp Encarsia formosa. Genomics, Vol. 112, pp. 2291–2301. doi:10.1016/j.ygeno.2019.12.025

Hoffman, Gabriel E., and Eric E. Schadt. 2016. "VariancePartition: Interpreting Drivers of Variation in Complex Gene Expression Studies." BMC Bioinformatics 17 (1): 483.

Hogg, Kirsten, John D. Blair, Peter von Dadelszen, and Wendy P. Robinson. 2013. "Hypomethylation of the LEP Gene in Placenta and Elevated Maternal Leptin Concentration in Early Onset Pre-Eclampsia." Molecular and Cellular Endocrinology 367 (1–2): 64–73.

Hord, Taylor Kimberly, Agata Maria Parsons Aubone, Asghar Ali, Hayley Nicole Templeton, River Evans, Jason Edward Bruemmer, Quinton Alexander Winger, and Gerrit Jerry Bouma. 2020. "Placenta Specific Gene Targeting to Study Histone Lysine Demethylase and Androgen Signaling in Ruminant Placenta." Animal Reproduction / Colegio Brasileiro de Reproducao Animal 17 (3): e20200069.

Huerfano, S., Ryabchenko, B., & Forstová, J. (2013). Nucleofection of expression vectors induces a robust interferon response and inhibition of cell proliferation. DNA and Cell Biology, 32(8), 467–479.

Hunt, G. J. (2007). Flight and fight: A comparative view of the neurophysiology and genetics of honey bee defensive behavior. Journal of Insect Physiology, 53(5), 399–410.

Hunt, G. J., Guzmán-Novoa, E., Fondrk, M. K., & Page, R. E., Jr. (1998). Quantitative trait loci for honey bee stinging behavior and body size. Genetics, 148(3), 1203–1213.

Ikeda, H., Old, L. J., & Schreiber, R. D. (2002). The roles of IFN gamma in protection against tumor development and cancer immunoediting. Cytokine & Growth Factor Reviews, 13(2), 95–109.

Isensee, J., & Ruiz Noppinger, P. (2007). Sexually dimorphic gene expression in mammalian somatic tissue. Gender Medicine, 4 Suppl B, S75-95.

Ishida, M., & Moore, G. E. (2013). The role of imprinted genes in humans. Molecular Aspects of Medicine, 34(4), 826–840.

Isles, A. R., Davies, W., & Wilkinson, L. S. (2006). Genomic imprinting and the social brain. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 361(1476), 2229–2237.

Ito, Masato, Masanori Tamura, Fumihiko Namba, and Neonatal Research Network of Japan. 2017. "Role of Sex in Morbidity and Mortality of Very Premature Neonates." Pediatrics International: Official Journal of the Japan Pediatric Society 59 (8): 898–905.

J, H., A, F., Jm, G., E, T., M, D., F, K., Bl, A., D, B., A, Z., S, S., I, B., A, B., V, B., T, H., M, K., G, M., J, R., G, D.-R., G, S., … Tj, H. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research, 22(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Jacobsen, L., Calvin, S., & Lobenhofer, E. (2009). Transcriptional effects of transfection: the potential for misinterpretation of gene expression data generated from transiently transfected cells. BioTechniques, 47(1), 617–624.

Jaillon, S., Berthenet, K., & Garlanda, C. (2019). Sexual Dimorphism in Innate Immunity. Clinical Reviews in Allergy & Immunology, 56(3), 308–321.

Jardim, B. V., Moschetta, M. G., Leonel, C., Gelaleti, G. B., Regiani, V. R., Ferreira, L. C., Lopes, J. R., & Zuccari, D. A. P. de C. (2013). Glutathione and glutathione peroxidase expression in breast cancer: an immunohistochemical and molecular study. Oncology Reports, 30(3), 1119–1128.

Jene-Sanz, A., Varaljai, R., Vilkova, A. V., Khramtsova, G. F., Khramtsov, A. I., Olopade, O. I., Lopez-Bigas, N., & Benevolenskaya, E. V. (2013). Expression of Polycomb Targets Predicts Breast Cancer Prognosis. Molecular and Cellular Biology, 33(19), 3951–3961.

Jobling, M., Hurles, M., & Tyler-Smith, C. (2013). Human Evolutionary Genetics: Origins, Peoples & Disease. Garland Science.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic Acids Research, Vol. 36, pp. W5–W9. doi:10.1093/nar/gkn201

Kanellopoulos-Langevin, C., Caucheteux, S. M., Verbeke, P., & Ojcius, D. M. (2003). Tolerance of the fetus by the maternal immune system: role of inflammatory mediators at the feto-maternal interface. Reproductive Biology and Endocrinology: RB&E, 1, 121.

Karjalainen, Minna K., Marja Ojaniemi, Antti M. Haapalainen, Mari Mahlman, Annamari Salminen, Johanna M. Huusko, Tomi A. Määttä, et al. 2015. "CXCR3 Polymorphism and Expression Associate with Spontaneous Preterm Birth." Journal of Immunology 195 (5): 2187–98.

Kassam, I., Wu, Y., Yang, J., Visscher, P. M., & McRae, A. F. (2019). Tissue-specific sex differences in human gene expression. Human Molecular Genetics, 28(17), 2976–2986.

Kenny, P. A., Lee, G. Y., Myers, C. A., Neve, R. M., Semeiks, J. R., Spellman, P. T., Lorenz, K., Lee, E. H., Barcellos-Hoff, M. H., Petersen, O. W., Gray, J. W., & Bissell, M. J. (2007). The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. Molecular Oncology, 1(1), 84–96.

Khramtsova, E. A., Davis, L. K., & Stranger, B. E. (2019). The role of sex in the genomics of human complex traits. Nature Reviews. Genetics, 20(3), 173–190.

Khramtsova, E., Davis, L., & Stranger, B. (2018). The role of sex in the genomics of human complex traits. Nature Reviews Genetics, 20. https://doi.org/10.1038/s41576-018-0083-1

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nature Methods, 12(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015a. "HISAT: A Fast Spliced Aligner with Low Memory Requirements." Nature Methods 12 (4): 357–60.

Kim, J., & Orkin, S. H. (2011). Embryonic stem cell-specific signatures in cancer: insights into genomic regulatory networks and implications for medicine. Genome Medicine, 3(11), 75.

Kleer, C. G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S. A., Ghosh, D., Sewalt, R. G. A. B., Otte, A. P., Hayes, D. F., Sabel, M. S., Livant, D., Weiss, S. J., Rubin, M. A., & Chinnaiyan, A. M. (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proceedings of the National Academy of Sciences of the United States of America, 100(20), 11606–11611.

Klein, S. L., Marriott, I., & Fish, E. N. (2015). Sex-based differences in immune function and responses to vaccination. Transactions of the Royal Society of Tropical Medicine and Hygiene, 109(1), 9–15.

Kocher, S. D., Tsuruda, J. M., Gibson, J. D., Emore, C. M., Arechavaleta-Velasco, M. E., Queller, D. C., … Hunt, G. J. (2015). A Search for Parent-of-Origin Effects on Honey Bee Gene Expression. G3 , 5(8), 1657–1662.

Kowarz, E., Löscher, D., & Marschalek, R. (2015). Optimized Sleeping Beauty transposons rapidly generate stable transgenic cell lines. Biotechnology Journal, 10(4), 647–653.

Kulaeva, O. I., Draghici, S., Tang, L., Kraniak, J. M., Land, S. J., & Tainsky, M. A. (2003). Epigenetic silencing of multiple interferon pathway genes after cellular immortalization. Oncogene, 22(26), 4118–4127.

Kwilas, A. R., Ardiani, A., Dirmeier, U., Wottawah, C., Schlom, J., & Hodge, J. W. (2015). A poxviral-based cancer vaccine the transcription factor twist inhibits primary tumor growth and metastases in a model of metastatic breast cancer and improves survival in a spontaneous prostate cancer model. Oncotarget, 6(29), 28194–28210.

Lachner, M., Sengupta, R., Schotta, G., & Jenuwein, T. (2004). Trilogies of histone lysine methylation as epigenetic landmarks of the eukaryotic genome. Cold Spring Harbor Symposia on Quantitative Biology, 69, 209–218.

Lacroix, M., & Leclercq, G. (2004). Relevance of breast cancer cell lines as models for breast tumours: an update. Breast Cancer Research and Treatment, 83(3), 249–289.

Lahn, B. T., & Page, D. C. (1999). Four evolutionary strata on the human X chromosome. Science (New York, N.Y.), 286(5441), 964–967. https://doi.org/10.1126/science.286.5441.964

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., Le Berre-Anton, V., Bouzayen, M., & Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. Frontiers in Plant Science, 9, 108. https://doi.org/10.3389/fpls.2018.00108

Lappalainen, T., Sammeth, M., Friedländer, M. R., ’t Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlöf, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506–511. https://doi.org/10.1038/nature12531

Lara, H., Wang, Y., Beltran, A. S., Juárez-Moreno, K., Yuan, X., Kato, S., Leisewitz, A. V., Cuello Fredes, M., Licea, A. F., Connolly, D. C., Huang, L., & Blancafort, P. (2012). Targeting serous epithelial ovarian cancer with designer zinc finger transcription factors. The Journal of Biological Chemistry, 287(35), 29873–29886.

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 15(2), R29.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014a. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." Genome Biology. https://doi.org/10.1186/gb-2014-15-2-r29.

Lawson, H. A., Cheverud, J. M., & Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. Nature Reviews. Genetics, 14(9), 609–617.

Lee, J., Wang, A., Hu, Q., Lu, S., & Dong, Z. (2006). Adenovirus-mediated interferon-β gene transfer inhibits angiogenesis in and progression of orthotopic tumors of human prostate cancer cells in nude mice. International Journal of Oncology. https://doi.org/10.3892/ijo.29.6.1405

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. Cell, 152(6), 1237–1251.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K.-I., Koseki, H., Fuchikami, T., Abe, K., Murray, H. L., Zucker, J. P., Yuan, B., Bell, G. W., Herbolsheimer, E., Hannett, N. M., … Young, R. A. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell, 125(2), 301–313.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of Clinical Investigation, 121(7), 2750–2767.

Leinonen, R., H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration. 2011. "The Sequence Read Archive." Nucleic Acids Research. https://doi.org/10.1093/nar/gkq1019.

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Research, 39(Database issue), D19–D21. https://doi.org/10.1093/nar/gkq1019

Leonova, K., Safina, A., Nesher, E., Sandlesh, P., Pratt, R., Burkhart, C., Lipchick, B., Gitlin, I., Frangou, C., Koman, I., Wang, J., Kirsanov, K., Yakubovskaya, M. G., Gudkov, A. V., & Gurova, K. (2018). TRAIN (Transcription of Repeats Activates INterferon) in response to chromatin destabilization induced by small molecules in mammalian cells. eLife, 7. https://doi.org/10.7554/eLife.30842

Leroy, G., Dimaggio, P. A., Chan, E. Y., Zee, B. M., Blanco, M. A., Bryant, B., Flaniken, I. Z., Liu, S., Kang, Y., Trojer, P., & Garcia, B. A. (2013). A quantitative atlas of histone modification signatures from human cancer cells. Epigenetics & Chromatin, 6(1), 20.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Chiappinelli, K. B., Guzzetta, A. A., Easwaran, H., Yen, R.-W. C., Vatapalli, R., Topper, M. J., Luo, J., Connolly, R. M., Azad, N. S., Stearns, V., Pardoll, D. M., Davidson, N., Jones, P. A., Slamon, D. J., Baylin, S. B., Zahnow, C. A., & Ahuja, N. (2014). Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. Oncotarget, 5(3), 587–598.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics , 25(16), 2078–2079. Retrieved from PMC. (PMC2723002)

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” ArXiv [q-Bio.GN]. arXiv. http://arxiv.org/abs/1303.3997.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England), 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lin, H.-J. L., Zuo, T., Lin, C.-H., Kuo, C. T., Liyanarachchi, S., Sun, S., Shen, R., Deatherage, D. E., Potter, D., Asamoto, L., Lin, S., Yan, P. S., Cheng, A.-L., Ostrowski, M. C., & Huang, T. H.-M. (2008). Breast cancer-associated fibroblasts confer AKT1-mediated epigenetic silencing of Cystatin M in epithelial cells. Cancer Research, 68(24), 10257–10266.

Liu, Y., Li, G., & Zhang, W. (2017). Effect of fetal gender on pregnancy outcomes in Northern China. The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, 30(7), 858–863.

Lopes-Ramos, C. M., C. Y. Chen, M. L. Kuijjer, and J. N. Paulson. 2020. “Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues.” Cell Reports. https://www.sciencedirect.com/science/article/pii/S2211124720307762.

Lopes-Ramos, C. M., Kuijjer, M. L., Ogino, S., Fuchs, C., DeMeo, D. L., Glass, K., & Quackenbush, J. (n.d.). Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism processes. https://doi.org/10.1101/277186

Lopes-Ramos, C. M., Quackenbush, J., & DeMeo, D. L. (2020). Genome-Wide Sex and Gender Differences in Cancer. Frontiers in Oncology, 10, 597788.

Lopes-Ramos, Camila M., Cho-Yi Chen, Marieke L. Kuijjer, Joseph N. Paulson, Abhijeet R. Sonawane, Maud Fagny, John Platig, Kimberly Glass, John Quackenbush, and Dawn L. DeMeo. 2020. "Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues." Cell Reports 31 (12): 107795.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., … Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 538(7624), 201–206. https://doi.org/10.1038/nature18964

Mamsen, L. S., Ernst, E. H., Borup, R., Larsen, A., Olesen, R. H., Ernst, E., … Andersen, C. Y. (2017). Temporal expression pattern of genes during the period of sex differentiation in human embryonic gonads. Scientific Reports, 7(1), 15961.

Mani, S., & Herceg, Z. (2010). DNA Demethylating Agents and Epigenetic Therapy of Cancer. In Advances in Genetics (pp. 327–340).

Margueron, R., & Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. Nature Reviews. Genetics, 11(4), 285–296.

May, T., Adesina, I., McGillivray, J., & Rinehart, N. J. (2019). Sex differences in neurodevelopmental disorders. Current Opinion in Neurology, 32(4), 622–626.

Maymon, Eli, Roberto Romero, Gaurav Bhatti, Piya Chaemsaithong, Nardhy Gomez-Lopez, Bogdan Panaitescu, Noppadol Chaiyasit, et al. 2018. "Chronic Inflammatory Lesions of the Placenta Are Associated with an Up-Regulation of Amniotic Fluid CXCR3: A Marker of Allograft Rejection." Journal of Perinatal Medicine 46 (2): 123–37.

McGarvey, K. M., Fahrner, J. A., Greene, E., Martens, J., Jenuwein, T., & Baylin, S. B. (2006). Silenced Tumor Suppressor Genes Reactivated by DNA Demethylation Do Not Return to a Fully Euchromatic Chromatin State. Cancer Research, 66(7), 3541–3549.

McGarvey, K. M., Greene, E., Fahrner, J. A., Jenuwein, T., & Baylin, S. B. (2007). DNA methylation and complete transcriptional silencing of cancer genes persist after depletion of EZH2. Cancer Research, 67(11), 5097–5102.

McGough, J. M., Yang, D., Huang, S., Georgi, D., Hewitt, S. M., Röcken, C., Tänzer, M., Ebert, M. P. A., & Liu, K. (2008). DNA methylation represses IFN-gamma-induced and signal transducer and activator of transcription 1-mediated IFN regulatory factor 8 activation in colon carcinoma cells. Molecular Cancer Research: MCR, 6(12), 1841–1851.

McGregor, James, Marilyn Leff, Miriam Orleans, and Anna Baron. 1992. "Fetal Gender Differences in Preterm Birth: Findings in a North American Cohort." American Journal of Perinatology. https://doi.org/10.1055/s-2007-994668.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." Genome Research 20 (9): 1297–1303.

Melamed, N., Yogev, Y., & Glezerman, M. (2010). Fetal gender and pregnancy outcome. The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, 23(4), 338–344.

Melamed, Nir, Yariv Yogev, and Marek Glezerman. 2010. "Fetal Gender and Pregnancy Outcome." The Journal of Maternal-Fetal & Neonatal Medicine: The Official Journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians 23 (4): 338–44.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Research, 45(D1), D183–D189.

Min Zhao Jingchun Sun. (2013). TSGene: a web resource for tumor suppressor genes. Nucleic Acids Research, Database issue, D970–D976.

Moore, T., & Haig, D. (1991). Genomic imprinting in mammalian development: a parental tug-of-war. Trends in Genetics: TIG, 7(2), 45–49.

Nagaraja, G. M., Othman, M., Fox, B. P., Alsaber, R., Pellegrino, C. M., Zeng, Y., Khanna, R., Tamburini, P., Swaroop, A., & Kandpal, R. P. (2006). Gene expression signatures and biomarkers of noninvasive and invasive breast cancer cells: comprehensive profiles by representational difference analysis, microarrays and proteomics. Oncogene, 25(16), 2328–2338.

Nancy, Patrice, and Adrian Erlebacher. 2014. “T Cell Behavior at the Maternal-Fetal Interface.” The International Journal of Developmental Biology 58 (2–4): 189–98.

Natri, H. M., Wilson, M. A., & Buetow, K. H. (2019). Distinct molecular etiologies of male and female hepatocellular carcinoma. BMC Cancer, 19(1), 951. https://doi.org/10.1186/s12885-019-6167-2

Naugler, W. E., Sakurai, T., Kim, S., Maeda, S., Kim, K., Elsharkawy, A. M., & Karin, M. (2007). Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. Science (New York, N.Y.), 317(5834), 121–124. https://doi.org/10.1126/science.1140485

NCBI Resource Coordinators, & NCBI Resource Coordinators. (2017). Database Resources of the National Center for Biotechnology Information. Nucleic Acids Research, Vol. 45, pp. D12–D17. doi:10.1093/nar/gkw1071

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L., Stilwell, J. L., Pinkel, D., Albertson, D. G., … Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell, 10(6), 515–527.

Niehuis, O., Gibson, J. D., Rosenberg, M. S., Pannebakker, B. A., Koevoets, T., Judson, A. K., … Gadau, J. (2010). Recombination and its impact on the genome of the haplodiploid parasitoid wasp Nasonia. PloS One, 5(1), e8597.

Niehuis, O., Judson, A. K., & Gadau, J. (2008). Cytonuclear genic incompatibilities cause increased mortality in male F2 hybrids of Nasonia giraulti and N. vitripennis. Genetics, 178(1), 413–426.

Niida, A., Smith, A. D., Imoto, S., Aburatani, H., Zhang, M. Q., & Akiyama, T. (2009). Gene set-based module discovery in the breast cancer transcriptome. BMC Bioinformatics, 10, 71.

Nishibuchi, G., & Déjardin, J. (2017). The molecular basis of the organization of repetitive DNA-containing constitutive heterochromatin in mammals. Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology, 25(1), 77–87.

Nyer, D. B., Daer, R. M., Vargas, D., Hom, C., & Haynes, K. A. (2017). Regulation of cancer epigenomes with a histone-binding synthetic transcription factor. Npj Genomic Medicine, 2(1). https://doi.org/10.1038/s41525-016-0002-3

Ochiai, Eri, Qila Sa, Morgan Brogli, Tomoya Kudo, Xisheng Wang, Jitender P. Dubey, and Yasuhiro Suzuki. 2015. “CXCL9 Is Important for Recruiting Immune T Cells into the Brain and Inducing an Accumulation of the T Cells to the Areas of Tachyzoite Proliferation to Prevent Reactivation of Chronic Cerebral Infection with Toxoplasma Gondii.” The American Journal of Pathology. https://doi.org/10.1016/j.ajpath.2014.10.003.

Olejniczak, M., Galka, P., & Krzyzosiak, W. J. (2010). Sequence-non-specific effects of RNA interference triggers and microRNA regulators. Nucleic Acids Research, 38(1), 1–16.

Olney, K. C., Brotman, S. M., Andrews, J. P., Valverde-Vesling, V. A., & Wilson, M. A. (2020). Reference genome and transcriptome informed by the sex chromosome complement of the sample increase ability to detect sex differences in gene expression from RNA-Seq data. Biology of Sex Differences, 11(1), 1–18.

Olney, K. C., Nyer, D. B., Vargas, D. A., Wilson Sayres, M. A., & Haynes, K. A. (2018). The synthetic histone-binding regulator protein PcTF activates interferon genes in breast cancer cells. BMC Systems Biology, 12(1), 83.

Olney, Kimberly C., Sarah M. Brotman, Jocelyn P. Andrews, Valeria A. Valverde-Vesling, and Melissa A. Wilson. 2020a. “Reference Genome and Transcriptome Informed by the Sex Chromosome Complement of the Sample Increase Ability to Detect Sex Differences in Gene Expression from RNA-Seq Data.” Biology of Sex Differences 11 (1): 42.

Pandey, R. S., Wilson Sayres, M. A., & Azad, R. K. (2013). Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. Genome Biology and Evolution, 5(10), 1863–1871. https://doi.org/10.1093/gbe/evt139

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417–419. https://doi.org/10.1038/nmeth.4197

Peacock, Janet L., Louise Marston, Neil Marlow, Sandra A. Calvert, and Anne Greenough. 2012. “Neonatal and Infant Outcome in Boys and Girls Born Very Prematurely.” Pediatric Research 71 (3): 305–10.

Pedersen, Brent S., and Aaron R. Quinlan. 2017. “Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy.” The American Journal of Human Genetics. https://doi.org/10.1016/j.ajhg.2017.01.017.

Peña-Llopis, S., Wan, Y., & Martinez, E. D. (2016). Unique epigenetic gene profiles define human breast cancers with poor prognosis. Oncotarget, 7(52), 85819–85831.

Phung, Tanya N., Kimberly C. Olney, Michelle Silasi, Lauren Perley, Jane O'Bryan, Harvey J. Kliman, and Melissa A. Wilson. n.d. "X Chromosome Inactivation in the Human Placenta Is Patchy and Distinct from Adult Tissues." https://doi.org/10.1101/785105.

Picard Tools. (2003). Broad Institute. http://broadinstitute.github.io/picard/

Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. American Journal of Human Genetics, 93(4), 641–651. https://doi.org/10.1016/j.ajhg.2013.08.008

Plaisier, C. L., O'Brien, S., Bernard, B., Reynolds, S., Simon, Z., Toledo, C. M., Ding, Y., Reiss, D. J., Paddison, P. J., & Baliga, N. S. (2016). Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. Cell Systems, 3(2), 172–186.

PrabhuDas, M., Bonney, E., Caron, K., Dey, S., Erlebacher, A., Fazleabas, A., … Yoshinaga, K. (2015). Immune mechanisms at the maternal-fetal interface: perspectives and challenges. Nature Immunology, 16(4), 328–334.

Przybyl, Lukasz, Nadine Haase, Michaela Golic, Julianna Rugor, Maria Emilia Solano, Petra Clara Arck, Martin Gauster, et al. 2016. "CD74-Downregulation of Placental Macrophage-Trophoblastic Interactions in Preeclampsia." Circulation Research 119 (1): 55–68.

Queller, D. C. (2003). Theory of genomic imprinting conflict in social insects. BMC Evolutionary Biology, 3, 15.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, Vol. 26, pp. 841–842. doi:10.1093/bioinformatics/btq033

R, P., Sm, M., & C, K. (2014, May). Sailfish Enables Alignment-Free Isoform Quantification From RNA-seq Reads Using Lightweight Algorithms. Nature Biotechnology; Nat Biotechnol. https://doi.org/10.1038/nbt.2862

Rahbari, R., Zhang, L., & Kebebew, E. (2010). Thyroid cancer gender disparity. Future Oncology (London, England), 6(11), 1771–1779. https://doi.org/10.2217/fon.10.127

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research, 44(W1), W160–W165.

Raznahan, A., Parikshak, N. N., Chandran, V., Blumenthal, J. D., Clasen, L. S., Alexander-Bloch, A. F., Zinn, A. R., Wangsa, D., Wise, J., Murphy, D. G. M., Bolton, P.

F., Ried, T., Ross, J., Giedd, J. N., & Geschwind, D. H. (2018). Sex-chromosome dosage effects on gene expression in humans. Proceedings of the National Academy of Sciences of the United States of America, 115(28), 7398–7403. https://doi.org/10.1073/pnas.1802889115

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., & Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends in Genetics: TIG, 13(4), 163.

Reik, W., & Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nature Reviews. Genetics, 2(1), 21–32.

Ren, G., Baritaki, S., Marathe, H., Feng, J., Park, S., Beach, S., Bazeley, P. S., Beshir, A. B., Fenteany, G., Mehra, R., Daignault, S., Al-Mulla, F., Keller, E., Bonavida, B., de la Serna, I., & Yeung, K. C. (2012). Polycomb protein EZH2 regulates tumor invasion via the transcriptional repression of the metastasis suppressor RKIP in breast and prostate cancer. Cancer Research, 72(12), 3091–3104.

Rey, R., Josso, N., & Racine, C. (2020). Sexual Differentiation. In K. R. Feingold, B. Anawalt, A. Boyce, G. Chrousos, W. W. de Herder, K. Dungan, … D. P. Wilson (Eds.), Endotext. South Dartmouth (MA): MDText.com, Inc.

Richards, E. J., & Elgin, S. C. R. (2002). Epigenetic codes for heterochromatin formation and silencing: rounding up the usual suspects. Cell, 108(4), 489–500.

Rinn, J. L., & Snyder, M. (2005). Sexual dimorphism in mammalian gene expression. Trends in Genetics: TIG, 21(5), 298–305.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” Genome Biology 11 (3): R25.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” Bioinformatics 26 (1): 139–40.

Rockowitz, S., Liu, S., Liu, Y., Li, F., Vedadi, M., Frye, S. V., Garcia, B. A., Zheng, D., Jin, J., &

Roemer, I., Reik, W., Dean, W., & Klose, J. (1997). Epigenetic inheritance in the mouse. Current Biology: CB, 7(4), 277–280.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G. R., Burrows, C., Bird, C. P., Frankish, A., Lovell, F. L., Howe, K. L., Ashurst, J. L., Fulton, R. S., Sudbrak, R., Wen, G., Jones, M. C., Hurles, M. E., … Bentley, D. R. (2005). The DNA sequence of the human X chromosome. Nature, 434(7031), 325–337. https://doi.org/10.1038/nature03440

Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S. Y., Han, H., Liang, G., Jones, P. A., Pugh, T. J., O'Brien, C., & De Carvalho, D. D. (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell, 162(5), 961–973.

Rubin, J. B., Lagas, J. S., Broestl, L., Sponagel, J., Rockwell, N., Rhee, G., … Luo, J. (2020). Sex differences in cancer mechanisms. Biology of Sex Differences, 11(1), 17.

Rupp, S. M., Webster, T. H., Olney, K. C., Hutchins, E. D., Kusumi, K., & Wilson Sayres, M. A. (2017). Evolution of Dosage Compensation in Anolis carolinensis, a Reptile with XX/XY Chromosomal Sex Determination. Genome Biology and Evolution, 9(1), 231–240.

Schaarschmidt, S., Fischer, A., Zuther, E., & Hincha, D. K. (2020). Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana. International Journal of Molecular Sciences, 21(5). doi:10.3390/ijms21051720

Schwartzentruber, J., Korshunov, A., Liu, X.-Y., Jones, D. T. W., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A. M., Quang, D.-A. K., Tönjes, M., Hovestadt, V., Albrecht, S., Kool, M., Nantel, A., Konermann, C., Lindroth, A., Jäger, N., Rausch, T., Ryzhova, M., … Jabado, N. (2012). Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature, 482(7384), 226–231.

Seals, D. F., Azucena, E. F., Jr, Pass, I., Tesfay, L., Gordon, R., Woodrow, M., Resau, J. H., & Courtneidge, S. A. (2005). The adaptor protein Tks5/Fish is required for podosome formation and function, and for the protease-driven invasion of cancer cells. Cancer Cell, 7(2), 155–165.

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics, 16(1), 59–70. https://doi.org/10.1093/bib/bbt086

Sheiner, Eyal. 2007. "The Relationship between Fetal Gender and Pregnancy Outcome." Archives of Gynecology and Obstetrics 275 (5): 317–19.

Shi, L., Zhang, Z., & Su, B. (2016). Sex Biased Gene Expression Profiling of Human Brains at Major Developmental Stages. Scientific Reports, 6, 21181. https://doi.org/10.1038/srep21181

Shorter, J. R., Arechavaleta-Velasco, M., Robles-Rios, C., & Hunt, G. J. (2012). A Genetic Analysis of the Stinging and Guarding Behaviors of the Honey Bee. Behavior Genetics, Vol. 42, pp. 663–674. doi:10.1007/s10519-012-9530-5

Simhadri, C., Daze, K. D., Douglas, S. F., Quon, T. T. H., Dev, A., Gignac, M. C., Peng, F., Heller, M., Boulanger, M. J., Wulff, J. E., & Hof, F. (2014a). Chromodomain antagonists that target the polycomb-group methyllysine reader protein chromobox homolog 7 (CBX7). Journal of Medicinal Chemistry, 57(7), 2874–2883.

Simon, J. A., & Kingston, R. E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nature Reviews. Molecular Cell Biology, 10(10), 697–708.

Simoneau, J., Dumontier, S., Gosselin, R., & Scott, M. S. (2019). Current RNA-seq methodology reporting limits reproducibility. Briefings in Bioinformatics. doi:10.1093/bib/bbz124

Sitras, Vasilis, Christopher Fenton, Ruth Paulssen, Åse Vårtun, and G. Acharya. 2012. "Differences in Gene Expression between First and Third Trimester Human Placenta: A Microarray Study." PloS One 7 (3): e33294.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T., Chinwalla, A., Delehaunty, A., Delehaunty, K., Du, H., Fewell, G., Fulton, L., Fulton, R., Graves, T., Hou, S.-F., … Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature, 423(6942), 825–837. https://doi.org/10.1038/nature01722

Skaletsky, Helen, Tomoko Kuroda-Kawaguchi, Patrick J. Minx, Holland S. Cordum, Ladeana Hillier, Laura G. Brown, Sjoerd Repping, et al. 2003a. "The Male-Specific Region of the Human Y Chromosome Is a Mosaic of Discrete Sequence Classes." Nature 423 (6942): 825–37.

Sledz, C. A., Holko, M., de Veer, M. J., Silverman, R. H., & Williams, B. R. G. (2003). Activation of the interferon system by short-interfering RNAs. Nature Cell Biology, 5(9), 834–839.

Smith, N. M. A., Yagound, B., Remnant, E. J., Foster, C. S. P., Buchmann, G., Allsopp, M. H., … Oldroyd, B. P. (2020). Paternally-biased gene expression follows kin-selected predictions in female honey bee embryos. Molecular Ecology, 29(8), 1523–1533.

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Research, 4, 1521. https://doi.org/10.12688/f1000research.7563.2

Sood, R., Zehnder, J. L., Druzin, M. L., & Brown, P. O. (2006). Gene expression patterns in human placenta. Proceedings of the National Academy of Sciences of the United States of America, 103(14), 5478–5483.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., … Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America, 98(19), 10869–10874.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E., & Borresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences, 98(19), 10869–10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lønning, P. E., Brown, P. O., Børresen-Dale, A.-L., & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proceedings of the National Academy of Sciences of the United States of America, 100(14), 8418–8423.

Sparmann, A., & van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. Nature Reviews. Cancer, 6(11), 846–856.

Staahl, B. T., Benekareddy, M., Coulon-Bainier, C., Banfal, A. A., Floor, S. N., Sabo, J. K., Urnes, C., Munares, G. A., Ghosh, A., & Doudna, J. A. (2017). Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. Nature Biotechnology, 35(5), 431–434.

Stone, M. L., Chiappinelli, K. B., Li, H., Murphy, L. M., Travers, M. E., Topper, M. J., Mathios, D., Lim, M., Shih, I.-M., Wang, T.-L., Hung, C.-F., Bhargava, V., Wiehagen, K. R., Cowley, G. S., Bachman, K. E., Strick, R., Strissel, P. L., Baylin, S. B., & Zahnow, C. A. (2017). Epigenetic therapy activates type I interferon signaling in murine ovarian cancer to reduce immunosuppression and tumor burden. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.1712514114

Stuckey, J. I., Dickson, B. M., Cheng, N., Liu, Y., Norris, J. L., Cholensky, S. H., Tempel, W., Qin, S., Huber, K. G., Sagum, C., Black, K., Li, F., Huang, X.-P., Roth, B. L., Baughman, B. M., Senisterra, G., Pattenden, S. G., Vedadi, M., Brown, P. J., … Frye,

S. V. (2016a). A cellular chemical probe targeting the chromodomains of Polycomb repressive complex 1. Nature Chemical Biology, 12(3), 180–187.

Su, I.-H., Dobenecker, M.-W., Dickinson, E., Oser, M., Basavaraj, A., Marqueron, R., Viale, A., Reinberg, D., Wülfing, C., & Tarakhovsky, A. (2005). Polycomb group protein ezh2 controls actin polymerization and cell signaling. Cell, 121(3), 425–436.

Su, Z., Łabaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., … Others. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology, 32(9), 903–914.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS One, 6(7), e21800.

Sylvester, J. E., Fischel-Ghodsian, N., Mougey, E. B., & O'Brien, T. W. (2004). Mitochondrial ribosomal proteins: candidate genes for mitochondrial disease. Genetics in Medicine: Official Journal of the American College of Medical Genetics, 6(2), 73–80.

Tabet, S., Douglas, S. F., Daze, K. D., Garnett, G. A. E., Allen, K. J. H., Abrioux, E. M. M., Quon, T. T. H., Wulff, J. E., & Hof, F. (2013a). Synthetic trimethyllysine receptors that bind histone 3, trimethyllysine 27 (H3K27me3) and disrupt its interaction with the epigenetic reader protein CBX7. Bioorganic & Medicinal Chemistry, 21(22), 7004–7010.

Tekel, S. J., Vargas, D. A., Song, L., LaBaer, J., & Haynes, K. A. (2017). Tandem histone-binding domains enhance the activity of a synthetic chromatin effector. https://doi.org/10.1101/145730

Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007a). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biology, 8(8), R157.

Tian, C., Wang, L., Ye, G., & Zhu, S. (2010). Inhibition of melanization by a Nasonia defensin-like peptide: implications for host immune suppression. Journal of Insect Physiology, 56(12), 1857–1862.

Tokunaga, Ryuma, Wu Zhang, Madiha Naseem, Alberto Puccini, Martin D. Berger, Shivani Soni, Michelle McSkane, Hideo Baba, and Heinz-Josef Lenz. 2018. "CXCL9, CXCL10, CXCL11/CXCR3 Axis for Immune Activation – A Target for Novel Cancer Therapy." Cancer Treatment Reviews. https://doi.org/10.1016/j.ctrv.2017.11.007.

Traglia, M., Bseiso, D., Gusev, A., Adviento, B., Park, D. S., Mefford, J. A., Zaitlen, N., & Weiss, L. A. (2017). Genetic Mechanisms Leading to Sex Differences Across Common Diseases and Anthropometric Traits. Genetics, 205(2), 979–992. https://doi.org/10.1534/genetics.116.193623

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology, 31(1), 46–53.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols, 7(3), 562–578.

Tseng, L.-M., Chiu, J.-H., Liu, C.-Y., Tsai, Y.-F., Wang, Y.-L., Yang, C.-W., & Shyr, Y.-M. (2017). A comparison of the molecular subtypes of triple-negative breast cancer among non-Asian and Taiwanese women. Breast Cancer Research and Treatment, 163(2), 241–254.

Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., Consortium, Gte., Lappalainen, T., Regev, A., Ardlie, K. G., … MacArthur, D. G. (2016). Landscape of X chromosome inactivation across human tissues. BioRxiv, 073957. https://doi.org/10.1101/073957

Turco, M. Y., & Moffett, A. (2019). Development of the human placenta. Development , 146(22). doi:10.1242/dev.163428

Turner, M. E., Ely, D., Prokop, J., & Milsted, A. (2011). Sry, more than testis determination? American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 301(3), R561–R571. https://doi.org/10.1152/ajpregu.00645.2010

Ueda, K., Yoshimi, A., Kagoya, Y., Nishikawa, S., Marquez, V. E., Nakagawa, M., & Kurokawa, M. (2014). Inhibition of histone methyltransferase EZH2 depletes leukemia stem cell of mixed lineage leukemia fusion leukemia through upregulation of p16. Cancer Science, 105(5), 512–519.

Vaiman, Daniel, Françoise Mondon, Alexandra Garcès-Duran, Thérèse-Marie Mignot, Brigitte Robert, Régis Rebourcet, Hélène Jammes, et al. 2005. “Hypoxia-Activated Genes from Early Placenta Are Elevated in Preeclampsia, but Not in Intra-Uterine Growth Retardation.” BMC Genomics 6 (August): 111.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.], 43, 11.10.1-33.

Van der Auwera, Geraldine A., Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ

Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.] 43: 11.10.1-33.

Vatten, Lars J., and Rolv Skjaerven. 2004. "Offspring Sex and Pregnancy Outcome by Length of Gestation." Early Human Development 76 (1): 47–54.

Veerappa, A. M., Padakannaya, P., & Ramachandra, N. B. (2013). Copy number variation-based polymorphism in a new pseudoautosomal region 3 (PAR3) of a human X-chromosome-transposed region (XTR) in the Y chromosome. Functional & Integrative Genomics, 13(3), 285–293. https://doi.org/10.1007/s10142-013-0323-6

Wang, G. G. (2015). Selective inhibition of EZH2 and EZH1 enzymatic activity by a small molecule suppresses MLL-rearranged leukemia. Blood, 125(2), 346–357.

Wang, W., Qin, J.-J., Voruganti, S., Nag, S., Zhou, J., & Zhang, R. (2015). Polycomb Group (PcG) Proteins and Human Cancers: Multifaceted Functions and Therapeutic Implications. Medicinal Research Reviews, 35(6), 1220–1267.

Wang, X., Werren, J. H., & Clark, A. G. (2016). Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. PLOS Biology, Vol. 14, p. e1002500. doi:10.1371/journal.pbio.1002500

Warnes, M. G. R., Bolker, B., Bonebakker, L., & Gentleman, R. (2016). Package "gplots". Various R Programming Tools for Plotting Data.

Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., Whitford, W., & Wilson, M. A. (2019). Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. GigaScience, 8(7). https://doi.org/10.1093/gigascience/giz074

Webster, Timothy H., Madeline Couse, Bruno M. Grande, Eric Karlins, Tanya N. Phung, Phillip A. Richmond, Whitney Whitford, and Melissa A. Wilson. 2019a. "Identifying, Understanding, and Correcting Technical Artifacts on the Sex Chromosomes in next-Generation Sequencing Data." GigaScience. https://doi.org/10.1093/gigascience/giz074.

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., … Gibbs, R. A. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science, 327(5963), 343–348.

Wiles, E. T., & Selker, E. U. (2017). H3K27 methylation: a promiscuous repressive chromatin mark. Current Opinion in Genetics & Development, 43, 31–37.

Wu, D. C., Yao, J., Ho, K. S., Lambowitz, A. M., & Wilke, C. O. (2018). Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics, 19(1), 510. https://doi.org/10.1186/s12864-018-4869-5

Wu, G., Broniscer, A., McEachron, T. A., Lu, C., Paugh, B. S., Becksfort, J., Qu, C., Ding, L., Huether, R., Parker, M., Zhang, J., Gajjar, A., Dyer, M. A., Mullighan, C. G., Gilbertson, R. J., Mardis, E. R., Wilson, R. K., Downing, J. R., Ellison, D. W., … St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project. (2012). Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. Nature Genetics, 44(3), 251–253.

Wu, Yiyang, and Gholson J. Lyon. 2018. "NAA10-Related Syndrome." Experimental & Molecular Medicine 50 (7): 1–10.

Xin, Lijun, James M. Ertelt, Jared H. Rowe, Tony T. Jiang, Jeremy M. Kinder, Vandana Chaturvedi, Shokrollah Elahi, and Sing Sing Way. 2014. "Cutting Edge: Committed Th1 CD4+ T Cell Differentiation Blocks Pregnancy-Induced Foxp3 Expression with Antigen-Specific Fetal Loss." Journal of Immunology 192 (7): 2970–74.

Xu, B., On, D. M., Ma, A., Parton, T., Konze, K. D., Pattenden, S. G., Allison, D. F., Cai, L.,

Xu, H., Xian, J., Vire, E., McKinney, S., Wei, V., Wong, J., Tong, R., Kouzarides, T., Caldas, C., & Aparicio, S. (2014a). Up-regulation of the interferon-related genes in BRCA2 knockout epithelial cells. The Journal of Pathology, 234(3), 386–397.

Yan, H., Bonasio, R., Simola, D. F., Liebig, J., Berger, S. L., & Reinberg, D. (2015). DNA methylation in social insects: how epigenetics can control behavior and longevity. Annual Review of Entomology, 60, 435–452.

Zagni, E., Simoni, L., & Colombo, D. (2016). Sex and Gender Differences in Central Nervous System-Related Disorders. Neuroscience Journal, 2016, 2827090.

Zaidi, S. K., Frietze, S. E., Gordon, J. A., Heath, J. L., Messier, T., Hong, D., Boyd, J. R., Kang, M., Imbalzano, A. N., Lian, J. B., Stein, J. L., & Stein, G. S. (2017). Bivalent Epigenetic Control of Oncofetal Gene Expression in Cancer. Molecular and Cellular Biology, 37(23). https://doi.org/10.1128/MCB.00352-17

Zeitlin, Jennifer, Marie-Josèphe Saurel-Cubizolles, Jaques De Mouzon, Lucile Rivera, Pierre-Yves Ancel, Béatrice Blondel, and Monique Kaminski. 2002. "Fetal Sex and Preterm Birth: Are Males at Greater Risk?" Human Reproduction 17 (10): 2762–68.

Zhao, M., Kim, P., Mitra, R., Zhao, J., & Zhao, Z. (2015). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Research, 44(D1), D1023–D1031.

Zhao, M., Sun, J., & Zhao, Z. (2013). TSGene: a web resource for tumor suppressor genes. Nucleic Acids Research, 41(Database issue), D970-6.

Zhao, Shanrong, & Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics, 16(1), 97. https://doi.org/10.1186/s12864-015-1308-8

Zhao, Shilin, Li, C.-I., Guo, Y., Sheng, Q., & Shyr, Y. (2018). RnaSeqSampleSize: Real data based sample size estimation for RNA sequencing. BMC Bioinformatics, 19(1), 191. https://doi.org/10.1186/s12859-018-2191-5

Zielezinski, A., Vinga, S., Almeida, J., & Karlowski, W. M. (2017). Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biology, 18(1), 186. https://doi.org/10.1186/s13059-017-1319-7

Zuo, T., Liu, T.-M., Lan, X., Weng, Y.-I., Shen, R., Gu, F., Huang, Y.-W., Liyanarachchi, S., Deatherage, D. E., Hsu, P.-Y., Taslim, C., Ramaswamy, B., Shapiro, C. L., Lin, H.-J. L., Cheng, A. S. L., Jin, V. X., & Huang, T. H.-M. (2011). Epigenetic silencing mediated through activated PI3K/AKT signaling in breast cancer. Cancer Research, 71(5), 1752–1762.

# APPENDIX A

## CHAPTER 1. SUPPLEMENTAL TABLES AND FIGURES.

Any operating systems (e.g., MAC or Windows) should be sufficient to view the materials in this appendix.

Chapter 1 supplemental tables and figures are in the zipped folder “APPENDIX A.”

Additional file 1: Sample IDs. RNA-Seq whole blood, brain cortex, breast, liver, and thyroid tissue samples from 20 genetic female (46, XX) and 20 genetic male (46, XY) individuals were downloaded from the Genotype-Tissue Expression (GTEx) project [19] for a total of 200 RNA-Seq tissue samples.(41K, txt)

Additional file 2: Histogram of sample reported age. For each tissue, whole blood, brain cortex, breast, liver, and thyroid, male XY and female XX samples were age matched perfectly between age 55 to 70. Females are shown in blue and males are shown in lime green. Since the samples were aged perfectly the histogram bars show only the overlap of female and male samples is a mix color of the blue and lime green.(22K, pdf)

Additional file 3: Genetic sex of RNA-Seq samples when aligned using STAR. Gene expression log2(CPM + 0.25/L) for select XY homologous genes (DDX3X/Y, PCDH11X/Y, USP9X/Y, ZFX/Y, UTX/Y) and XIST and SRY when reads were aligned to a default reference genome A), and for B) when reads were aligned to a sex chromosome complement informed reference using STAR. Male XY whole blood, brain cortex, breast, liver, and thyroid samples are shown in blue squares and female XX in orange circles.(5.2M, pdf)

Additional file 4: Genetic sex of RNA-Seq samples per tissue. Gene expression log2(CPM + 0.25/L) for select XY homologous genes (DDX3X/Y, PCDH11X/Y, USP9X/Y, ZFX/Y, UTX/Y) and XIST and SRY when reads were aligned to a default reference genome A), and for B) when reads were aligned to a sex chromosome complement informed reference using HISAT and C) and D), for when the reads were

aligned using STAR. Male XY whole blood, brain cortex, breast, liver, and thyroid samples are shown in blue squares and female XX in orange circles.(47M, pdf)

Additional file 5: List of samples that were removed from downstream analysis. Samples that did not cluster with the reported sex or clustered in unexpected ways were removed from the differential expression analysis. One male XY whole blood, 4 female XX and 4 male XY brain cortex, and one female XX breast sample were removed.(9.0K, xlsx)

Additional file 6: Multidimensional Scaling plots. We investigated multidimensional scaling for all shared common variable genes for dimensions 1 and 2, and for dimensions 2 and 3 in each tissue. The most variation in each tissue is explained by the aligner C.aligner. The second most variation in each tissue is explained by the sex of the sample A.sex.(1.2M, pdf)

Additional file 7: HISAT mapped reads bar plot. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for significant difference in total reads mapped for the whole transcriptome, chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately. Red * indicate a significant, p-value <0.05, difference in average mapped reads, NS is no significant differences.(4.2M, pdf)

Additional file 8: STAR mapped reads bar plot. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for

significant difference in total reads mapped for the whole transcriptome, chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately. Red * indicate a significant, p-value <0.05, difference in average mapped reads, NS is no significant differences.(3.7M, pdf)

Additional file 9: Paired t-test for mapped reads in default compared to sex chromosome complement reference genome. Mean difference in expression for average total reads mapped for each tissue and each sex when aligned to a sex chromosome informed versus a default reference genome. Paired t-test to test for significant difference in total reads mapped for the whole transcriptome (WT), chromosome 8, and chromosome X. Nonparametric Wilcox single rank sum test was used to test for significant difference in total reads mapped on the Y chromosome for male samples in each tissue separately.(13K, xlsx)

Additional file 10: X chromosome expression differences between default and sex chromosome complement informed alignment. X chromosome gene expression differences between default and sex chromosome complement informed alignment. Increase in expression when aligned to a sex chromosome complement informed reference genome is a log2 fold change (FC) > 0. A decrease in expression when aligned to a sex chromosome complement informed reference genome is log2 FC < 0. Female XX samples are indicated by red and pink circles for PAR1, XTR, PAR2 genes, and for all other X chromosome genes respectively. Blue and light blue squares represent male XY samples. Blue squares indicate which gene points are in PAR1, XTR, and PAR2, and light blue squares are for genes outside of those regions. Differences in X chromosome

expression between reference genomes default and sex chromosome complement for male XY and female XX samples aligned using HISAT for the whole X chromosome and the first 5 Mb are shown for the whole blood (A and B, respectively), brain cortex (E and F, respectively), breast (I and J, respectively), liver (M and N, respectively), and thyroid (Q and R, respectively). Differences in X chromosome expression between reference genomes for male XY and female XX samples aligned using STAR for the whole X chromosome and the first 5 Mb are shown for the whole blood (C and D, respectively), brain cortex (G and H, respectively), breast (K and L, respectively), liver (O and P, respectively), and thyroid (S and T, respectively).(1.3M, pdf)

Additional file 11: X chromosome regions mean and median expression values. X chromosome regions PAR1, PAR2, XTR, XDG, XAR, XCR mean and median CPM expression for male XY and female XX samples for each tissue separately when aligned to a default or sex chromosome complement informed reference genome using either HISAT and STAR. Paired t-test was used to test for significant differences in expression. XTR and XAR show a significant increase, p-value $<0.05$, in female expression for each tissue type. XTR and XAR additionally show a significant increase, p-value $<0.05$, in male expression for liver and thyroid. PAR2 shows a significant increase, p-value $<0.05$, in female liver expression. Additionally reported fold change in mean expression when using a sex chromosome complement informed compared to a default reference genome. The mean fold change in expression either increased or stayed the same ranging from 2.8 to 0.999 fold increase in expression. Finally, mean male over mean female expression was reported for each X chromosome region for each tissue. Mean male over mean

female expression decreases for XTR when using a sex chromosome complement reference genome for each tissue.(44K, xlsx)

Additional file 12: Whole genome gene expression values per sample, aligner and reference genome used for alignment. CPM values for male XY and female XX whole blood, brain cortex, breast, liver and thyroid samples when aligned to a default and sex chromosome complement informed reference genome for the whole genome (1-22, mtDNA, X, Y and non-chromosomal).(16K, docx)

Additional file 13: Gene expression for XY homologous genes. X chromosome expression for 26 X and Y homologous genes (AMELX, ARSD, ARSE, ARSF, CASK, GYG2, HSFX1, HSFX2, NLGN4X, OFD1, PCDH11X, PRKX, RBMX, RPS4X, SOX3, STS, TBL1X, TGIF2LX, TMSB4X, TSPYL2, USP9X, VCX, VCX2, VCX3A, VCX3B, ZFX). Difference in gene expression for when male XY and female XX samples were aligned to a default and sex chromosome complement informed reference genome for each tissue. Little to no difference in gene expression between default and sex chromosome complement informed reference genome alignment was observed for 25 of the 26 X and Y homologous genes for both male XY and female XX samples using either HISAT or STAR. The log2 fold increase in expression for PCDH11X when aligned using HISAT was 0.4, 0.28, 0.33, 0.16, and 0.16 for whole blood, brain cortex, breast, liver, and thyroid, respectively. The greatest increase in expression was observed for PCDH11X in female whole blood at a log2 fold increase of 0.4.(86K, xlsx)

Additional file 14: Differentially expressed genes between the sexes that were uniquely and jointly called between reference genomes. Genes that are differentially expressed between the sexes, male XY and female XX, for whole blood, brain cortex,

breast, liver, and thyroid samples. Differentially expressed genes that are uniquely called when using either the default or sex chromosome complement informed reference genome and differentially expressed genes that were jointly called between the reference genomes.(28K, xlsx)

Additional file 15: Gene expression differences between male XY and female XX samples. Sex differences in gene expression for whole blood, brain cortex, breast, liver, and thyroid samples for when samples were aligned to a default reference genome and to a reference genome informed on the sex chromosome complement. Showing sex differences in gene expression between reference genomes used for alignment and for when samples were aligned using HISAT and STAR.(70M, pdf)

Additional file 16: GO analysis of differentially expressed genes in female and male samples with HISAT aligner. Gene enrichment analysis of genes that are more highly expressed in one sex versus the other sex for each tissue, whole blood, brain cortex, breast, liver and thyroid, when samples were aligned to a default or sex chromosome complement informed reference genome using HISAT.(661K, txt)

Additional file 17: GO analysis of differentially expressed genes in female and male samples with STAR aligner. Gene enrichment analysis of genes that are more highly expressed in one sex versus the other sex for each tissue, whole blood, brain cortex, breast, liver and thyroid, when samples were aligned to a default or sex chromosome complement informed reference genome using STAR.(708K, txt)

Additional file 18: Sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A) Sex differences in gene expression, log2(CPM + 0.25/L), between the sixteen samples from genetic males

and females are shown when aligning all samples to the default Ensembl reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. B) We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex chromosome complement, 14 genes were called as differentially expressed between the sexes. PLCXD1 was uniquely called as differentially expressed when aligned to a default reference genome. Ensembl sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A) Sex differences in gene expression, log2(CPM + 0.25/L), between the sixteen samples from genetic males and females are shown when aligning all samples to the default Ensembl reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. B) We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex

chromosome complement, 14 genes were called as differentially expressed between the sexes. PLCXD1 was uniquely called as differentially expressed when aligned to a default reference genome.(6.2M, pdf)

Additional file 19: Gencode sex chromosome complement informed transcriptome reference eliminates Y-linked expression in female XX samples. A) Sex differences in gene expression, log2(CPM + 0.25/L), between the sixteen samples from genetic males and females are shown when aligning all samples to the default gencode reference transcriptome (left) and a reference transcriptome informed on the sex chromosome complement (right) for brain cortex. Each point represents a gene. Genes that are differentially expressed, adjusted p-value <0.01 are indicated in black for autosomal genes, blue for Y-linked genes, and red for X-linked genes. B) We show overlap between genes that are called as differentially expressed when all samples are pseudo-aligned to the default transcriptome, and genes that are called as differentially expressed when pseudo-aligned to a sex chromosome complement informed transcriptome reference. When samples were aligned to a reference transcriptome informed on the sex chromosome complement, 17 genes were called as differentially expressed between the sexes. ZBED1 was uniquely called as differentially expressed when aligned to a default reference genome.(6.0M, pdf)

Additional file 20: 3 male XY and 3 female XX brain cortex and whole blood differential expression analysis. Replicated analysis in a smaller sample size of 3 male XY compared to 3 female XX samples for whole blood and brain cortex tissue. Samples were randomly selected, and confirm the results from the larger sample size.(5.2M, docx)

# APPENDIX B

## CHAPTER 2. SUPPLEMENTAL TABLES AND FIGURES.

Any operating systems (e.g., MAC or Windows) should be sufficient to view the materials in this appendix.

Chapter 2 supplemental tables and figures are in the zipped folder "APPENDIX B."

Additional Figure 1. Sample sex check. Violin jitter plot of counts per million (CPM) expression for each placenta sample for *EIF1AY, KDM5D, UTY, DDX3Y,* and *RPS4Y1* Y-linked genes, and *XIST,* X-linked gene. Samples with at least two X chromosomes will show expression for *XIST.* Samples with the presence of the Y chromosome will show expression for Y-linked genes.

Additional Figure 2. Multidimensional scaling plots reveal outlier samples. Multidimensional scaling (MDS) for all genes (left) and top too genes on the (right) for (A) late first trimester placentas (Gonzalez et al. 2018), (B) term placentas, (C) term placentas excluding failed samples.

Additional Figure 3. Population ancestry inference. Population ancestry was inferred from whole-exome sequencing for each term placenta.

Additional Figure 4. Variation in expression trait attributes. Variation within gestational age (GA), sequencing lane, sex, reported race, and birth weight was examined. Variation in placenta expression for maternal clinical data, including parity, gravidity, pre-pregnancy body mass index (BMI), and maternal age, were also examined.

Additional Figure 5. Sex differences for clinical attributes. Sex differences for clinical information for term placentas for maternal age at delivery, pre-pregnancy BMI, gravidity and parity, gestational age, and birth weight. Sex differences for continuous variables were tested using a t-test, p-value $< 0.05$.

Additional Figure 6. Sex differences in expression for gametolog genes. There is a significant difference in male XY to female XX expression for *ZFX* and *KDM6A* (*UTX*) when only looking at the X chromosome CPM expression value. When we add the Y

chromosome-linked CPM expression count for these genes for male samples, there is no longer a difference in expression between males XY and females XX for *ZFX*. *KDM6A*, on the other hand, flips the bias; it now shows males as having significantly higher expression than females. *PCDH11X*, when adding Y-linked CPM expression, shows a significantly higher expression than females. T-test to see if there is a difference between the female CPM and the male CPM for each gene, p-value $< 0.05$.

Additional Figure 7. Sex differences in expression for innate immune genes.

Additional Table 1. Sample clinical information. Clinical and sequence information for each full-term placenta sample.

Additional Table 2. Post-trimming sample sequence information. Million sequences, percent of duplicate sequences, and percent QC content remaining after quality trimming.

Additional Table 3. Samples removed from downstream analysis. Samples were removed that had less than 12.5M or higher than 90M sequences remaining after trimming. If more than 30% of the reads deviate from the sum of the deviations from the normal distribution of the per-sequence GC content as defined by the FASTQC report, then the sample was removed. If a sample clustered with opposite sex from the reported sex for that sample, than that sample was removed.

Additional Table 4. Sex differences for clinical attributes. Sex differences for clinical information for term placentas for maternal age at delivery, pre-pregnancy BMI, gravidity and parity, gestational age, and birth weight. Ratio of variances in the female and male samples is reported. Female mean and male mean for each clinical attribute is

additionally reported. Sex differences for continuous variables were tested using a t-test, p-value < 0.05.

Additional Table 5. X and Y gametology gene list. A list of X and Y gametology genes were curated from a combination of Skaletsy et al. 2003 and Godfrey et al. 2020 (Godfrey et al. 2020; Skaletsky et al. 2003). FPKM expression for X-linked copy and Y-linked copy for all samples. In samples determined to have a Y chromosome, the FPKM value of the X-linked gametology and the Y-linked gametology were summed. expression between XX female X-linked gametology gene expression to XY male X-linked plus Y-linked gametology gene expression using a Wilcox rank-sum, p-value ≤ 0.05.

Additional Table 6. Sex differences in innate immune genes. 979 innate immune games from InnateDB. Placenta CPM expression values for expressed innate immune genes in the late first trimester and term placentas. Sex differences in late first trimester and term placentas, adjusted p-value ≤ 0.05.

Additional Table 7. Sex differentially expressed genes. Sex differentially expressed genes in the late first trimester and term placentas, adjusted p-value < 0.05.

Additional Table 8. GTEx female and male mean TPM expression values. Female and male mean TPM expression for 42 non-reproductive adult GTEx tissues. TPM expression for each gene obtained from counts version 2017-06-06_v8 (Carithers et al. 2015).

Additional Table 9. Gene FPKM and CPM values for late first trimester and term placentas.

Additional Table 10. Overlap in sex differentially expressed genes in Gonzalez vs reprocessing.

# APPENDIX C

## CHAPTER 3. SUPPLEMENTAL TABLES AND FIGURES.

Any operating systems (e.g., MAC or Windows) should be sufficient to view the materials in this appendix.

Chapter 3 supplemental tables and figures are in the zipped folder "APPENDIX C."

Figure S1. Comparisons of gene sets that are differentially or similarly expressed in BT-474, MCF7, BT-549, and MCF10A.

Figure S2. Comparisons, by cell line, of expressed and silenced genes within PRC-modules.

Figure S3. Jensen Shannon divergence analyses of transcription profiling data (RNA-seq) for all PcTF-treated and untreated cell samples.

Figure S4. Detailed view of the transcription factor (TF) binding motif overrepresentation plot from Figure 3D.

Figure S5. Expression levels of putative regulators of PUGs.

Figure S6. Chromosome plot of PcTF-responsive genes that were identified in the RNA-seq experiment.

Figure S7. Detailed view of MCF7 ChIP-seq signals.

Table S1. The set of 45 H3K27me3-enriched, repressed (FPKM < 2) genes shared by the three cancer cell lines.

Table S2. TF motif enrichment analysis results for the data shown in Figure 3D.

Table S2. Primers used to generate the RT-qPCR results shown in Figure 6.

# APPENDIX D

## CHAPTER 4. SUPPLEMENTAL TABLES AND FIGURES.

Any operating systems (e.g., MAC or Windows) should be sufficient to view the materials in this appendix.

Chapter 4 supplemental tables and figures are in the zipped folder "APPENDIX D."

Supplemental 1 Figure. Volcano plots for differential expression and venn diagram of DEGs between the datasets when taking the average of the counts when aligned to *N. vitripennis* and to pseudo *N. giraulti* reference genome. Volcano plots of DEGs detected between the different comparisons involving *N. vitripennis*, *N. giraulti*, and the two reciprocal $F_1$ hybrids in the R16A Clark (left side) and Wilson (right side) datasets. Venn diagrams of the overlap of significant DEGs in each comparison is shown.

Supplemental 1 Table. Sample identifiers. The samples for each dataset used in the project are provided here. Samples from this study are uploaded at https://www.ncbi.nlm.nih.gov/sra/PRJNA613065.

Supplemental 2 Table. Allele-specific expression differences between hybrids. The number of allele-counts for the reference allele (*N. vitripennis*) and alternative (*N. giraulti*) allele at polymorphic SNPs within a gene. Minimum of two SNPs for a gene to be included. The significance of allelic bias was determined using Fisher's exact test. Significant genes were selected using a Benjamini-Hochberg false discovery rate FDR-adjusted *p*-value threshold of 0.05.

Supplemental 3 Table. Mean and median allele and gene depth for Wilson dataset. Mean and median allele and gene depth for each GV and VG sample in the Wilson data set. Number of SNPs for all genes, *CPR35,* and *LOC103315494.*

Supplemental 4 Table. Genomic location of mortality loci and gene sets of interest. Previously reported loci associated with mortality in *Nasonia* hybrids. 95% Confidence Intervals of loci identified in Niehuis et al. 2008 were converted to genetic distances along the chromosomes and the closest SNP markers from Niehius et al. 2010 were identified

(Niehuis et al., 2010e, 2008). SNP markers for the locus identified in Gibson et al. 2013 were used directly (J. D. Gibson et al., 2013). The SNP marker locations in the PSR1.1 assembly were found via BLAST and all genes within the bounds of these markers are included. The two non-introgressed regions from the R16A strain are included as well as genes from two mitochondria-associated pathways, the oxidative phosphorylation pathway (Joshua D. Gibson et al., 2010) and the mitochondrial ribosomal proteins (Burton & Barreto, 2012).

Supplemental 5 Table. Directional bias of differentially expressed genes between VG and GV in Clark and Wilson datasets. Five genes that were called as differentially expressed between VG and GV hybrids in both the Clark and Wilson data sets.

Supplemental 6 Table. Locus conversion calculations. Calculations for converting the genetic map positions (centimorgan, cM) of mortality loci identified by Niehuis et al. 2008 to the physical chromosomal positions of the latest genome assembly (PSR1.1) (Niehuis et al., 2008).

# APPENDIX E

PERMISSION FROM CO-AUTHORS.

The chapter titled “Reference Genome and Transcriptome Informed by the Sex Chromosome Complement of the Sample Increase Ability to Detect Sex Differences in Gene Expression from RNA-Seq Data” was published earlier – 2020 - in the journal BMC Biology of Sex Differences. The paper had five contributing authors. C. Kimberly Olney was the first author. The original publication can be found at: https://doi.org/10.1186/s13293-020-00312-9. Brotman, S.M., Andrews, J.P., Valverde-Velsing, V.A., Wilson, M.A have all consented for the publication to be included in this dissertation by C. Kimberly Olney.

The chapter titled “The Synthetic Histone-Binding Regulator Protein PcTF Activates Interferon Genes in Breast Cancer Cells” was published earlier – 2018 - in the journal BMC Systems Biology. The paper had five contributing authors. C. Kimberly Olney was the first author. The original publication can be found at: https://doi.org/10.1186/s12918-018-0608-4. The corresponding author, Haynes, K.A. has consented for the publication to be included in this dissertation by C. Kimberly Olney.

The chapter titled “Lack of Parent-of-Origin Effects in *Nasonia* Jewel Wasp: a Replication and Extension Study” was published earlier – 2021 on BioRxiv. The paper had six contributing authors. C. Kimberly Olney was the first author. The original publication can be found at: https://doi.org/10.1101/2021.02.11.430138. All co-authors, Gibson, J.D., Natri, H.M., Underwood, A., Gadau, J., Wilson, M.A, have consented for the publication to be included in this dissertation by C. Kimberly Olney.