

Dissecting the Functional Effects of Non-coding Gene Regulatory Elements

By

Laavanya Sankaranarayanan

University Program in Genetics and Genomics  
Duke University

Defense Date: April 1, 2024

Approved:

Timothy Reddy, Supervisor

Gregory Crawford, Co-Chair

Michael Hauser, Co-Chair

Aravind Asokan

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in  
the University Program of Genetics and Genomics in The Graduate School of  
Duke University  
2024

ABSTRACT

Dissecting the Functional Effects of Non-coding Gene Regulatory Elements

By

Laavanya Sankaranarayanan

University Program in Genetics and Genomics  
Duke University

Defense Date: April 1, 2024

Approved:

Timothy Reddy, Supervisor

Gregory Crawford, Co-Chair

Michael Hauser, Co-Chair

Aravind Asokan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of  
Philosophy in the University Program of Genetics and Genomics in The Graduate School of  
Duke University  
2024

Copyright by  
Laavanya Sankaranarayanan  
2024

## **Abstract**

One of the most beautiful and challenging aspects in biology is deciphering the complexity of the genome, and how it functions or dysfunctions. It is this intricate complexity that is dependent on developmental stages, time of the day, and tissue types that allows for the proper development of an organism comprising of different tissue types with different functions. Amongst the many complexities, I focused on the gene-regulatory functions of the non-coding genome and its relation to diseases including disease risk, severity, and progression. Over the last few decades, there has been an increase in the research of genetic causes underlying several complex, common multifactorial diseases including metabolic and cardiovascular diseases. While these studies have identified genetic risk loci, they have not directly identified the genetic mechanisms behind what causes those diseases. Identifying genetic mechanisms for complex traits has been challenging because most of the variants are located outside of protein-coding regions, and determining the effects of such non-coding variants remains difficult. Previous studies that explore the genetic mechanism of complex diseases have underscored the need to develop new methods to study non-coding regions and to systematically identify the effects of non-coding variants towards a disease.

In this dissertation, I evaluate the hypothesis that non-coding regulatory elements can contribute to disease-relevant traits by altering gene expression levels. I will specifically focus on a common complex disease, polycystic ovary syndrome (PCOS) which is the most prevalent endocrine disorder among menstruating people. Family- and twin-studies have demonstrated a genetic basis to PCOS. Previous studies have identified non-coding genetic variation associated with PCOS risk across populations with different ancestries. However, the functional follow up of these risk

loci has been limited. Therefore, there is a gap in addressing the functional effects of genetic variants and regulatory elements impacting PCOS phenotypes. We identified gene regulatory mechanisms that help explain genetic association with PCOS in several loci using high throughput reporter assays, CRISPR-based epigenome editing, and genetic association analysis. To develop approaches to study regulatory elements, I implemented reporter assays at three different scales to create a framework for the regulatory elements across PCOS risk loci. I also implemented experimental approaches that measured changes in gene expression at single cell levels to identify target genes of regulatory elements identified by the reporter assays. Specifically, we identified regulatory elements across PCOS genetic risk loci in cell models of steroidogenesis, H295R cells and COV434 cells. We then identified regulatory elements that controlled the expression of the gene DENND1A, which altered the levels of testosterone produced by the cell models upon perturbation. Lastly, we quantified the regulatory effects of allele-specific genetic variants from a population of PCOS cases and controls. Taken together, we have identified regulatory elements that could contribute to PCOS pathogenesis. More broadly, my results demonstrate the strengths of combining experimental and statistical approaches to identify molecular mechanisms of genetic risk loci contributing to disease pathogenesis.

## **Dedication**

This body of work is dedicated to:

Chidambaram FGF retrogene Chonksta - Chidi, bestboi, my A dog, my heart dog.

And to:

All those who were told 'no you cannot, and you shouldn't':

You can and you will.

# Contents

Abstract.....	iv
List of Tables .....	xi
List of Figures.....	xi
Acknowledgements.....	xiii
1.Introduction.....	1
1.1 Gene Regulation in mammals .....	1
1.1.1 Overview of the structure of the eukaryotic genome.....	1
1.1.2 Regulation of transcription in the eukaryotic genome.....	2
1.1.3 Regulatory elements in disease development .....	7
1.2 Challenges in functional analysis of disease-causing regulatory elements .....	9
1.2.1 Genome-wide association studies have identified genomic loci associated with diseases .....	9
1.2.2 Challenges in translating results from genome-wide association studies to genetic mechanisms .....	10
1.2.3 Identifying the mechanistic role of GWAS signals .....	12
1.3 From genetic association to biological insights.....	13
1.3.1 Reporter assays measure regulatory activity of a candidate DNA fragment.....	14
1.3.2 CRISPR-Cas9 genome and epigenome editing .....	17
1.4 Polycystic ovary syndrome is a complex genetic disease .....	19
1.4.1 Classification of a reproductive disease based on ovarian morphology .....	20
1.4.2 Public Health Consequences of PCOS .....	20
1.4.3 Challenges in studying PCOS.....	21
1.4.4 Steroids and nuclear receptors in PCOS .....	22

1.5 The genetics basis of PCOS .....	24
1.5.1 PCOS Genome Wide Associate Studies.....	24
1.6 Functional Follow Up of PCOS GWAS have identified several likely candidate genes. ....	27
2. Identifying gene regulatory mechanisms in PCOS GWAS loci .....	31
2.1 Introduction .....	31
2.2 Results.....	33
2.2.1 Measuring the regulatory activity of PCOS-associated regulatory elements .....	33
2.2.2 Regulatory element activity in PCOS GWAS regions corresponds to regions of chromatin accessibility .....	37
2.2.3 PCOS-associated genetic variants fine-mapped to within regulatory elements .....	41
2.2.4 Active STARR-seq regions have increased conservation score .....	44
2.2.5 Activation of PCOS-associated regulatory elements increased DENND1A expression	45
2.2.6 Activation of PCOS-associated regulatory elements increased testosterone production in steroidogenic adrenal cells.....	49
2.3 Discussion .....	50
2.4 Methods.....	53
2.4.1 Data Availability: .....	53
2.4.2 Acknowledgements: .....	53
2.4.3 STARR Seq Assay Library Construction:.....	53
2.4.4 Cell Culture protocol .....	56
2.4.5 PCOS GWAS STARR-seq reporter library construction .....	57
2.4.6 PCOS case-control variant association testing within candidate regulatory regions.....	60
2.4.7 Colocalization testing .....	60
2.4.8 ATAC-Seq.....	61
2.4.9 CRISPR-dCas9 epigenome editing.....	62
2.4.10 RNA isolation and qRT-PCR to measure gene expression levels .....	65

2.4.11 ELISA for measuring testosterone production .....	66
3. Regulatory variants in the DENND1A locus .....	67
3.1 Introduction .....	67
3.2 Results .....	68
3.2.1 DENND1A is implicated in PCOS pathogenesis .....	68
3.2.2 Capture-STARR-seq of the DENND1A locus .....	69
3.2.3 Allele-specific regulatory activity across the DENND1A locus .....	72
3.2.4 Whole genome STARR-seq using donor genomes .....	76
3.3 Discussion .....	83
3.4 Methods .....	84
3.4.1 Cell Culture .....	84
3.4.2 DENND1A enrichment .....	85
3.4.3 Capture STARR-seq .....	85
3.4.4 Capture STARR-seq analysis .....	87
3.4.5 whole genome STARR-seq assay .....	89
4. Developing a probe-based method to quantify gene expression in intact single cells .....	92
4.1 Introduction .....	92
4.2 Results .....	94
4.2.1 HCR Flow-FISH optimization .....	94
4.2.2 Inducible change in gene expression measured using an optimized HCR Flow-FISH protocol for cell lines .....	96
4.2.3 HCR Flow-FISH allows multiplexed readout of target genes .....	98
4.2.4 HCR Flow-FISH captures altered gene expression levels in perturbations relevant to glucocorticoid responses .....	102
4.2.5 HCR Flow-FISH captures altered gene expression levels in perturbations of genomic loci associated with PCOS .....	104

4.3 Discussion .....	109
4.3.1 Other uses of Flow-FISH in genomics research .....	110
4.3.2 Advantages and considerations for implementing HCR Flow-FISH .....	110
4.4 Methods .....	112
4.4.1 Cell Culture .....	112
4.4.2 RPS knockdown in H295R cells .....	113
4.4.3 RNA extraction and qRT-PCR .....	114
4.4.4 H295R-dCas9-p300 & gRNA lentivirus .....	114
4.4.5 Generating stable H295R-dCas9-P300 cell line .....	115
4.4.6 Transduction of gRNA into dCas9-P300 expressing cell lines: .....	115
4.4.7 CRISPR screen DENND1A gRNA pool library design .....	116
4.4.8 CRSIPR screen plasmid pool sequencing .....	117
4.4.9 HCR Flow-FISH .....	118
5. Summary and future directions .....	122
5.1 Summary .....	122
5.2 Future Directions .....	125
5.2.1 Regulatory mechanisms contributing to PCOS .....	125
5.2.2 Mechanisms of regulatory elements in causing human diseases .....	133
6. Conclusion .....	136
Appendix A : Gene regulatory mechanisms in PCOS GWAS loci .....	138
Appendix B : Regulatory variants in DENND1A locus .....	145
Appendix C : Probe-based methods to quantify gene expression .....	152
References .....	154

## List of Tables

Table 1: List of PCOS GWAS loci selected for STARR-seq experiments.....	34
Table 2: Top variants associated with PCOS within STARR-seq regulatory elements. OR = odds ratio, MAF = minor allele frequency. ....	42
Table 3: Colocalization of PCOS-associated variants with eQTL data from GTEx.....	43
Table 4: Candidate allele-specific regulatory variants identified. ....	75
Table 5: HCR Flow-FISH identified changes in gene expression due to Dex treatment in A549 cells.....	102
Table 6: PER1 signal measured by HCR Flow-FISH in A549 cells with enhancer deletion .....	103
Table 7: RPS26 signal measured by HCR Flow-FISH in H295R cells with siRNA perturbation .....	106
Table 8: DENND1A signal measured by HCR Flow-FISH in H295R cells with 2 different gRNA pools targeting regulatory elements in DENND1A locus perturbation .....	109

## List of Figures

Figure 1: Principle of STARR-seq reporter assay. This image of the construct design is adapted from Stark et al., 2015.....	16
Figure 2: Manhattan plot depicting several of the loci associated with PCOS, with the nearest gene indicated. Image adapted from Day et al., 2018.....	25
Figure 3: Graphical representation of the biological relevance of genes near PCOS risk loci. Adapted from Dapas et al., 2022 .....	26
Figure 4 Measuring the regulatory activity in PCOS GWAS loci.....	35
Figure 5: Regulatory elements correspond to regions of chromatin accessibility .....	39
Figure 6: Prioritizing PCOS-associated variants within functional regulatory elements.....	44
Figure 7: Perturbation of regulatory elements in DENND1A impacts testosterone levels.....	48
Figure 8: Principle of probe-based capture of genomic regions .....	71
Figure 9: Fine-mapping variants identified four regulatory variants that are also eQTLs for DENND1A.....	73
Figure 10: Library complexity to estimate whole genome STARR-seq optimization.....	77
Figure 11: Complexity and length of whole genome STARR-seq libraries. ....	78
Figure 12: pseudo log <sub>2</sub> (Fold Change) of regulatory elements called by MACS.....	79
Figure 13: Whole genome STARR-seq regulatory elements correspond to accessible chromatin regions.....	80
Figure 14: Principle of hybridization chain reaction (HCR). Image adapted from Molecular Instruments.....	95
Figure 15: Comparison of intensity of IL11 by HCR Flow-FISH. ....	99
Figure 16: Optimizing HCR Flow-FISH in A549 cells. ....	101
Figure 17: HCR Flow-FISH measures gene expression changes due to genomic editing in A549 cells.....	103
Figure 18: RPS26 signal is decreased by siRNA knockdown measured by HCR Flow-FISH....	106
Figure 19: DENND1A signal is increased by dCas9-p300 perturbation in H295R cells .....	108

## Acknowledgements

Like many dissertations before mine, and likely many after, this is a collection of new thoughts, ideas and questions asked that are somewhat answered. In taking on this monumental task of a doctoral dissertation and presenting it, I would be remiss to not acknowledge the support of people around me.

Firstly, thank you to my mentor, Dr. Timothy Reddy. He has been incredibly supportive of this project and it is with his guidance that has trained me to be the scientist I am today. To my committee, Dr. Aravind Asokan, Dr. Greg Crawford, and Dr. Mike Hauser, thank you for challenging me, for your feedback, and guidance and importantly for supporting me across the challenges of getting the projects moving and for supporting me in my pursuits of gaining teaching experience. Thank you to Dr. Andrea Dunaif and Kelly Brewer at Mt. Sinai for their support. Thank you to Dr. Aleks Babic and Dr. Christine Stracey at Guilford College for the conversations over STEM education, and pedagogy.

To my scientific growth, I owe many thanks to my lab mates. I am grateful to Graham Johnson for mentoring me in scientific methods, and for being the person I would reach out to for questions and ideas. I am also grateful to Alex Barrera, who has taught me how to use bioinformatics tools, and learn and think about computational methods. Sarah, thank you for being a wonderful friend and lab mate to discuss everything from evolution and STEM education to troubleshooting and goss. Thank you to other lab members, Tony D'Ippolito, Young-Sook Kim, Kari Strouse, Apoorva Iyengar, Schuyler Melore, Shauna Morrow, Kuei-Yueh Ko and Revathy Venukuttan for making my lab experience more enjoyable, and scientifically more enriching. And lastly, thank you Tim, for creating a space where I was able to learn and grow in so many ways, for bringing in Goodboi Huck and for also loaning Chidi's first kennel.

Thank you to my Durham family, as we supported each other through everything thrown at us. Ankita, - thank you for all the time we spent together in our home. Iman, Erika, Crystal, Rylee, and Isabel, all of whom are amazing scientists and kindest friends. Maria, Sarah, Kayla, Tiina, Jon, Judy are incredible neighbours who opened their doors, and hearts, and we survived the toughest days of the height of COVID-19 pandemic together. Lauren, Katie and Katelyn, thank you your friendship, and music love. To all the people I have met through agility, thank you for challenging me to learn new things and channel my competitive spirit. Finally, my profound gratitude to my family. Sheetal, Khushi, Kalpana and Meena, you guys always kept me going with our conversations. Raahul was always the first to say something dumb to me, or chuckle at my coding flails, but also always there to help me. To those who have finished this particular journey before – Doctors: Malavika, Nikhila, Akila, Rohini - thank you for encouraging me in the last phase having been through the same. Aditi and Chinx, thank you for sustaining me with enthusiasm for so many different things. My deepest love and gratitude to Frank Willig for being the best source of support and comfort during all the ups and downs of the PhD journey. Your intelligence and curiosity inspired me to learn more, and thank you for enriching my life. And for reminding me about the wonders of macro-biology when molecular biology sometimes just doesn't work. And for always making me laugh asking "what's a guh-was?"

And finally, thank you to Chidi. If you know me, you would not be shocked to this paragraph in here. Chidi, you bring with you an infectious joy about life. I am learning from your confidence in tackling things that are so much bigger than you. I am always impressed by how new things come to you easily and you don't shy away from trying it out. Thank you for listening to lab meetings and for spending hours with me in lab so you can run around outside CIEMAS. Thank you for bringing me joy through your friends, Pacho, Juno, Hudson, Penny, Parker and

Teddy. Thank you to the best PhD buddy. Words are not sufficient to describe how instrumental you are, but I know you can feel my feels.

Lastly, as I've read through some other dissertations, I find myself reading their dedications and acknowledgments. These are the sections that beckon the reader to think about the other scientist as a whole person, a friend and someone who has been on a similar journey and all of the people who helped them get to the finish. This journey is no small feat. If you are reading this section after I have graduated, I hope you take a moment to think about the village supporting you in your doctoral journey.

Thank you.

# **1.Introduction**

In this dissertation, I have focused on expanding our understanding of the role of non-coding gene regulatory mechanisms in contributing to human diseases by developing methods to investigate regulatory elements, regulatory variants, and their effects on regulating gene expression. As a case study, I have implemented experimental and statistical approaches to study the non-coding regulatory elements that are associated with polycystic ovary syndrome (PCOS), a common, complex genetic disease. In this chapter, I will provide an overview of the different processes involved in gene regulation in eukaryotes, and how the non-coding genome affecting gene expression can contribute to human diseases. I then focus on how genetic variation impacts gene regulation and describe the current efforts to functionally identify the mechanisms of non-coding genetic variants in causing a disease or a trait. I also describe some of the challenges and recent advances in the field that have made investigating the functional effects of non-coding genetic variants easier. I will then give an overview of PCOS as a genetic disease and summarize the current studies that have evaluated the non-coding regulatory contributions to PCOS pathogenesis.

## ***1.1 Gene Regulation in mammals***

### **1.1.1 Overview of the structure of the eukaryotic genome**

In eukaryotes, the DNA of an organism is contained inside the nucleus of every cell. In general, there exists a flow of information contained with nucleotide and protein sequences. That is, genetic information encoded in the genome can be transcribed into an mRNA template which can then be translated into proteins as the functional units in the eukaryotes (G. M. Cooper,

2000). Genomic DNA consists of protein-coding sequences called genes, however much of the eukaryotic DNA is comprised of non-coding sequences. These sequences can be interspersed between the genes or also found within the genes as intronic regions that split a protein-coding gene sequence into segments called exons(G. M. Cooper, 2000). In addition, a substantial portion of the eukaryotic genome is comprised of non-coding repetitive sequences of several types scattered throughout the genome. Some of these DNA repeats have evidence of function, while most are inert (Shapiro & Sternberg, 2005). The genomic DNA is linear and compacted by the interactions between the DNA itself and several proteins called histones. This compacted complex is the chromatin where the DNA is wound around the histone proteins (G. M. Cooper, 2000). The basic structural unit of chromatin, the nucleosome, is described as beads on a string referring to the physical appearance of linear DNA compacted around the histones (Kornberg, 1974). Based on the compaction chromatin is categorized as heterochromatin and euchromatin. Heterochromatin is highly organized, dense, and not accessible to proteins like transcription factors (TFs) or RNA polymerase while euchromatin is less compact, and allows for increased transcription factor-binding and increased gene expression (Huisinga et al., 2006). This control of compaction is achieved by the location of the histones and modifications of the histone proteins itself. Post-translational modifications of histones can include acetylation, deacetylation, methylation etc., and can functionally influence gene expression (Bannister & Kouzarides, 2011). Chromatin structure is therefore linked to the control of gene expression in eukaryotes.

### **1.1.2 Regulation of transcription in the eukaryotic genome**

Genes are typically transcribed by RNA polymerase II in eukaryotes. Most genes have promoter sequences at the start of gene transcription start site and act as the binding site for transcription machinery to assembly and then begin the initiation of transcription (Smale &

Kadonaga, 2003). The genes usually have two core promoter elements, the TATA box and the initiator sequence, that act as specific binding sites for general transcription factors (G. M. Cooper, 2000). Early studies sequencing the human genome identified that only about 1-2 % of the complete human genome was considered protein-coding (Hatje et al., 2019). Since then, several studies and consortia have led efforts to characterize the non-coding genome which refocused the idea that these non-coding regions can have functional effects, primarily through gene regulation (Abascal et al., 2020; E. P. Consortium, 2004; Dunham et al., 2012; Kundaje et al., 2015). These studies revealed a complex genomic landscape where the interaction between DNA and proteins led to ranges of interactions across different distance ranges between the linear genomic loci. Importantly, these efforts successfully mapped the state of the epigenome across several cell and tissue types for 3D-interaction maps, histone modifications, transcription factor binding sites and other chromatin features. The datasets from these studies are publicly available, thereby creating a resource for identifying gene regulation and building a comprehensive understanding of different cell types. These studies and on-going projects have yielded an insight into the heterogeneity of genomic states that account for development and stimuli responses. However, there is still a vast amount of the non-coding genome that is less well understood, particularly in respect to human disease and development.

Many genes are controlled by genomic sequences that are at a distance from the transcription start site on a linear scale, sometimes > 10 kb. The genomic sequences that function as regulatory elements are physically separated from the transcription start site of a gene. Such regulatory sequences function as enhancers or repressors, influencing the gene expression of the target gene(s) of that regulatory sequence. Enhancers are *cis*-acting regulatory elements that increase the levels of gene expression, while repressors decrease or inhibit transcription. A single

gene is often regulated by multiple enhancers, and multiple genes can be regulated by a single enhancer (Karnuta & Scacheri, 2018). Enhancers function by the coordinated binding of transcription factors to open chromatin DNA regions comprised of short 20-30 bp sequences called motifs (Maurya, 2021). Enhancers can be highly cell-type specific, and the combination of the enhancer states and transcription factor binding is important for cell identity, lineage determination and differentiation and development (Maurya, 2021). Enhancers can recruit chromatin remodeling complexes that modify the histones, thereby, altering accessibility of the promoter. Repressors can inhibit the transcriptional process through different approaches including by impeding the binding of RNA polymerase to the DNA and by inhibiting the transcription initiation process (Rojo, 2001). Insulators are elements which block the enhancer-promoter interaction by recruiting chromatin structural modifiers and are important to separate the activity of genes with different regulatory requirements (Phillips & Corces, 2009).

While the classification of regulatory elements into enhancers or repressors describes the function of that regulatory element on the transcription of the target gene, the specific function is dependent on several factors and is a source of the complexity of gene regulation. Furthermore, a transcription factor can act as both transcription activators and repressors based on the context. The specific function of a transcription factor in any context is influenced by environmental stimuli, presence of co-activators or co-repressors, developmental stages, cell signaling and self-regulation (Weidemüller et al., 2021). Understanding the full range of transcription factor activity across regulatory elements is one key part of understanding cell signaling and gene regulation.

Regulatory elements for a target gene can be separated from the transcription start site by several kilobases. Despite the distance on a linear scale, one question in the field was how these

regulatory elements have an effect on transcription. These regulatory elements function by permitting the binding of transcription factors that can directly or indirectly have physical interactions with RNA polymerase II or other transcription factors, which is possible through DNA folding or looping (G. M. Cooper, 2000). The DNA looping process brings in close proximity the promoter region with several other regulatory elements, and the proximity is measured in the 3-dimensional space rather than the linear genomic scale. These long-range chromosomal interactions are mediated by proteins such as mediator and cohesin complexes (Kagey et al., 2010). Enhancers can also strengthen association with a gene by promoting interaction between promoter-proximal-transcription factors and chromatin remodeler proteins (Dean, 2011). Measurements of the interaction frequency between a candidate enhancer and the promoter of the target gene can yield insight into enhancer activity and dynamics (Hsiung et al., 2016). Recent studies have converged on a model for chromatin looping called the loop extrusion model (Kaur et al., 2023; Sanborn et al., 2015). Looping of the genomic DNA is achieved by extruding DNA through the ring-like structure of cohesin, and the extrusion continues till it reaches boundary regions marked by the binding of CTCF proteins. Studies of the 3D architecture of the genome identified regions of increased contact within each other called topological associating domains, or TADs (Lieberman-Aiden et al., 2009).

The binding of proteins such as transcription factors to DNA is permissible when the chromatin DNA is not tightly compacted. That open chromatin state allows protein-DNA interaction to occur. Therefore, one indirect way of identifying regions that might have regulatory function due to transcription factor binding is by identifying accessible chromatin regions. Studies of the non-coding genome revealed that regions of the genome that are not bound by proteins or histones are susceptible to DNA digestions, termed accessible or open chromatin. An

enzyme, DNase I, was used in high-throughput genome-wide studies to identify genome regions susceptible to cleavage known as DNase I hypersensitive sites and are regions enriched for transcription factor binding (Song & Crawford, 2010; Thurman et al., 2012). Recent advances in mapping chromatin accessibility have employed several other enzymes to improve the precision of data. One such advancement is the use of ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) which involves the use of hyperactive Tn5 transposase to identify accessible chromatin regions (Buenrostro et al., 2015). Optimization of this technology allowed for accessible chromatin measurements across a cell population of even < 100,000 and single-cell analyses have led to the development of several accessible chromatin maps including single cell maps across different cell and tissue types in different species (Corces et al., 2017; C. Liu et al., 2019; K. Zhang et al., 2021).

Another core component of gene regulation is in modification of the histone proteins that compact chromatin (Lorch et al., 1987). The chromatin is compacted by a set of 8 histone proteins. Each histone octamer is comprised of two sets of H2A, H2B, H3, and H4 proteins that are modified at the N-terminal by several different covalent modifications (Durrin et al., 1991) like acetylation, methylation, etc. Enhancers can exist in several states, which are typically marked by specific histone modifications such as H3K4me1 and H3K27ac for active enhancers and H3K4me1 and H3K27me3 for poised enhancers (Karnuta & Scacheri, 2018). Promoters of transcriptionally active genes are associated with trimethylation of histone-H3 lysine-4 (H3K4me3), while actively transcribing genes usually have higher levels of H3K36me3 and H3K79me3 in the gene body (Dixon et al., 2012; Gates et al., 2017).

### 1.1.3 Regulatory elements in disease development

Disruptions in the sequential flow of information from DNA to proteins can lead to several pathological indications. Focusing on the genome, genetic variation in protein-coding-genes has been shown to cause several Mendelian diseases (Roth & Marson, 2021). Given the essential role of transcriptional regulation in human health and development, misregulation could lead to disease. Gene regulation by regulatory elements is done through precise spatiotemporal control during stages of development, lineage specification, and in response to environmental stimuli (D'Ippolito et al., 2018; Maurya, 2021; Reddy et al., 2009). Gene regulation can happen at largely different genomic scales ranging from chromosomal conformation to chromatin accessibility and single nucleotides.

Transcriptional misregulation has been associated with a diverse set of diseases, including cancer and developmental syndromes (Darnell, 2002; Engelkamp & Heyningen, 1996; López-Bigas et al., 2006; Scherzer et al., 2008). For example, the androgen receptor (AR) gene is amplified in a population of prostate cancer cases (Merson et al., 2014) and increased levels of SNCA gene can contribute to dominant Parkinson's disease (Soldner et al., 2016). Mutations in the regulatory elements can range from chromosomal aberrations on the large scale right down to single nucleotide genetic variants at the single base level and are a major cause of human diseases. Structural variations like chromosomal translocation and inversions of large genomic segments have been identified in cases of cancer and metabolic disorders (D.-H. Lee et al., 2018; Nambiar et al., 2008; Nomura et al., 2018; Puig et al., 2015). Misregulation in the steps involved in chromatin looping can disrupt the organization of the genome called topological associated domains (TADs) leading to diseases collectively termed tadopathies (Matharu & Ahituv, 2015). One study reported a 660 kb deletion in a patient with autosomal dominant adult-onset

leukodystrophy (ADLD). This deletion caused the loss of a TAD, and aberrant enhancer adoption caused an enhancer to regulate a gene, LMNB1, which otherwise would not interact with that enhancer leading to ADLD (Giorgio et al., 2015). In some cases of a neurodevelopmental disorder, Cornelia de Lange Syndrome, there are mutations in the genes of the cohesin complex involved in chromatin loop formation. That change in cohesin leads to an altered chromatin-interaction map across several cell types (Garcia et al., 2021).

Regulatory variation can affect human health by altering any of the steps along the gene expression cascade from DNA to protein. It has been reported that genetic variants within transcription factor binding sites can affect the binding of transcription factors and co-activators or co-repressors based, contextually (Behera et al., 2018). Most causal variants are likely to influence traits by altering gene expression (Kioussis et al., 1983; Lettice et al., 2002; 2008; Rasmussen et al., 2015; Junwei Shi et al., 2013; Smemo et al., 2014). For example, genetic variants identified in an enhancer of IRF6 disrupts transcription factor binding and are associated with cleft lip (Rahimov et al., 2008). Another study identified obesity-associated variants to be regulating IRX3 as an example of long-range regulatory events leading to disease phenotypes (Smemo et al., 2014). For complex, non-monogenic diseases, several studies have associated genomic loci with predisposition to that disease. These genomic loci contain multiple regulatory elements, and these elements can influence several genes (Albert & Kruglyak, 2015; French & Edwards, 2020; Knight, 2005). Complex disease traits are, therefore, likely influenced by the contribution of dysregulation across several regulatory elements with small effects on the phenotype. One challenge of studying complex diseases is that their genetic architecture follows a polygenic rather than a Mendelian model (Visscher & Goddard, 2019), and therefore several genetic loci would have been assayed for their contribution to the disease.

## ***1.2 Challenges in functional analysis of disease-causing regulatory elements***

### **1.2.1 Genome-wide association studies have identified genomic loci associated with diseases**

Common non-communicable diseases are a concern for people and animals. The physical presentations of complex diseases are influenced by the interaction between genetic predisposition and environmental and lifestyle factors (Smith et al., 2005). Genome-wide association studies (GWAS) are designed to map the polygenic nature of common complex diseases by identifying genetic variants that are significantly enriched in a population of individuals with the disease than in a healthy population (Burton et al., 2007). Specifically, GWAS involves correlating allele-frequencies of genetic variant markers across the whole genome with a given trait within a population. These analyses yield risk haplotypes which likely contain some genetic variants that contribute to the disease phenotype. The results from thousands of GWAS for different human traits have identified millions of variants associated with traits, such that for a given trait, there are several genetic variants associated with it (Sollis et al., 2022). One recent model put forward is the omnigenic model which posits that all the genes expressed in a cell contribute to a given trait and that information flows from all the genetic variants to affecting nearby genes in turn affecting other genes (Boyle et al., 2017). The omnigenic model also describes a network where gene contributions towards complex diseases can be classified as core or peripheral genes. Core genes are those for which the gene product has a direct effect on the phenotype while peripheral genes outnumber core genes and contribute to a phenotype indirectly. Therefore, for a complex disease according to the omnigenic model, the phenotypic traits are a result of the expression of both core and peripheral genes. Based on the

omnigenic model, GWAS could be powered to identify new genetic loci that mark core or peripheral genes that contribute to the disease phenotypes with better genotypic approaches.

A database that compiles published genetic variant-to-trait associations called the GWAS Catalog was developed to aid researchers in accessing results across thousands of published GWAS results (Sollis et al., 2022). Currently, the GWAS Catalog, a compilation of completed genome-wide association studies, lists 6,499 studies and 539,949 significant associations as of July 29, 2023 release. The large number of GWAS studies have been facilitated by three key advancements in the genomics field: decreased cost of genotyping-arrays, increase in the number of variants able to be tested including better imputation statistics and reference panels including multi-ancestral genetic data, and increased efforts in statistical analyses to include traits, clustering to understand the heterogeneity of a trait, genome-wide gene-environment interactions (Loos, 2020). Despite these advancements with GWAS, the clinical insights into the disease pathology derived from the GWAS results have been limited.

### **1.2.2 Challenges in translating results from genome-wide association studies to genetic mechanisms**

A key motivation in studying the genetic mechanisms of disease pathogenesis is elucidating biological pathways for therapeutic interventions. That is, identifying the molecular pathways that drive disease phenotypes will enable us to identify potential new targets for drug development research to increase treatment options for that disease. Some complex genetic diseases of unknown etiology are managed clinically using drugs that aim for symptom management and do not cure the disease. Taken together, elucidating the underlying molecular pathway is a key step in developing targeted therapeutics as well as increasing our understanding

of disease diagnosis and causes. While genome-wide association studies have identified several risk loci, there remains a challenge in identifying the mechanism by which genetic variants contribute to the overall phenotype. GWAS results are statistical associations between genetic variants or single nucleotide polymorphisms (SNPs) and the risk of a disease. There are three main challenges that make it difficult to gain insight from GWAS associations to biological mechanisms contributing to a trait or disease phenotype. The first challenge is that most of the disease-associated SNPs are in the non-coding genome (Tam et al., 2019). This means that a large portion of the disease-causing variants are presumed to affect the regulation of the genes (French & Edwards, 2020). Second, genetic variants that are physically nearby on the genome are often inherited together due to co-segregation during meiotic recombination, known as linkage disequilibrium (LD) (Slatkin, 2008). This means that there are multiple variants present in a genomic locus due to this correlation and therefore, the causal genetic variant or variants can be hard to distinguish. Third, the effect of identified variants has not been able to fully account for the narrow sense of genetic heritability estimated for complex traits. That is, the phenotypic variance explained by the identified genetic variants doesn't fully capture all the phenotypic variance that is present, therefore there is still a question of what all the genetic contributions to that trait are. This problem, called the missing heritability problem has been a focus of several recent research studies to understand the molecular insights of complex diseases (Manolio et al., 2009). Explanations for the source of missing heritability likely include a combination of several factors including limited sample size and population (McCarthy et al., 2008), the lack of statistical ability to identify rare variants (Zuk et al., 2014) and gene-environmental interactions (T. Huang & Hu, 2015; Ni et al., 2019).

The identification of causal variants within a disease-associated genomic locus requires follow-up studies of the implicated genomic region, an approach known as fine mapping (Schaid et al., 2018). There can be several complementary approaches, experimental and computational, to be needed to determine which of the linked variants are functional (Edwards et al., 2013). Given that variants in high LD with the lead GWAS variant tend to be enriched in genomic regions with evidence of functionality, it is thought that the causal variants act to influence disease phenotype by altering gene expression patterns (Maurano et al., 2012). However, these genomic variants are in genomic loci that often have multiple genes which makes it challenging to identify which genes are affected in which cell type(s) in the case of a particular disease. Furthermore, the lack of understanding of causal cell types makes validation of GWAS variants difficult.

### **1.2.3 Identifying the mechanistic role of GWAS signals**

With increasing study populations and decreasing costs of sequencing, the number of associations from GWAS has been growing at a rapid pace (Gurdasani et al., 2019) and has necessitated follow-up studies on understanding the biological insight of those associations. The results from these GWAS studies led to the next obvious question in this field - how can we investigate the disease-associated loci to identify genetic variants that can impact cellular phenotypes leading to mechanistic insights into disease pathogenesis (Lichou & Trynka, 2020). A common hypothesis is that non-coding genetic associations increase disease risk via impacts on gene regulatory element activity and downstream gene expression (Gallagher & Chen-Plotkin, 2018). There have been several approaches to build upon the GWAS results in different traits or associations. These studies have used different approaches including statistical (Giambartolomei et al., 2014; Hormozdiari et al., 2014, 2016; Ma et al., 2015; Maurano et al., 2012; Nicolae et al.,

2010; Trynka et al., 2013) and experimental approaches to fine map GWAS signals, and in some cases identify functional variants (Kircher et al., 2019; Kneppers et al., 2022; S. Liu et al., 2017; Myint et al., 2020; Ouwerkerk et al., 2020; P. Zhang et al., 2018). For example, one study identified a single SNP that regulates SORT1 in a liver-specific manner to alter plasma lipid levels, in a GWAS risk locus for low-density lipoprotein cholesterol and myocardial infarction (MI) in humans (Musunuru et al., 2010). Another study focused on maternal hyperglycemia identified variants spanning multiple enhancers to have a coordinated effect on HKDC1 expression (Guo et al., 2015). Notably, detailed cellular or molecular studies are often needed to connect the identified gene regulatory impacts to a disease-relevant phenotype (Musunuru et al., 2010; Ouwerkerk et al., 2020).

### ***1.3 From genetic association to biological insights***

There are thousands of genetic variants reported to be associated with different traits or diseases and the pace of discovering new associations has been assisted by better genotyping technology and decreasing cost (Sollis et al., 2022). However, the pace and scale of the ability to identify genetic mechanisms that contribute to the disease or trait behind the association has been slower in part due to the challenge that necessitates developing and scaling up methods to address that question. Two emerging methods have been substantially advancing the ability to test for functional relevance in disease-associated risk loci. One approach is using high-throughput reporter assays that estimate the ability of genomic regions to act as regulatory elements. A second complementary approach is using CRISPR-Cas9 to make genetic or epigenetic perturbations in cells that recapitulate the effects of non-coding genetic variants.

### **1.3.1 Reporter assays measure regulatory activity of a candidate DNA fragment**

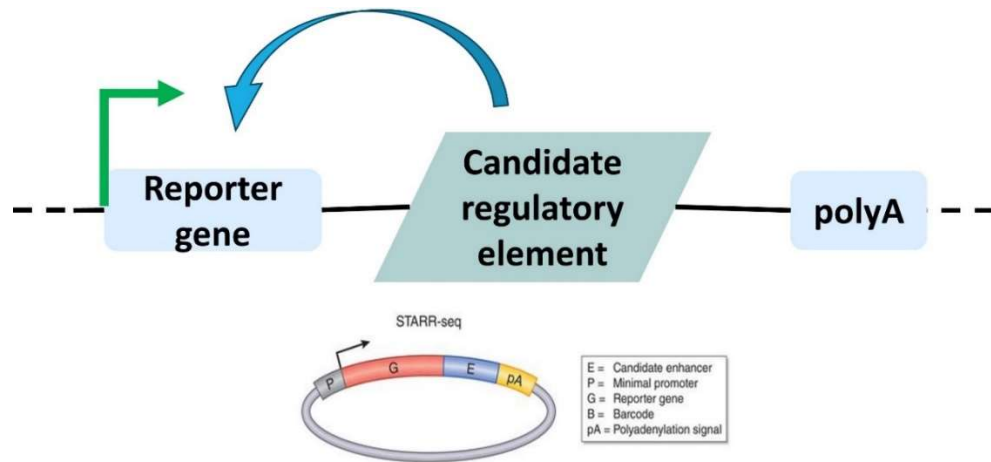
Reporter assays are experimental methods that are typically used to measure the regulatory ability of a given DNA sequence by regulating the expression of a reporter gene. The luciferase assay is a type of reporter assay where the reporter gene is involved in a biochemical reaction to produce bioluminescence (Nair & Baier, 2018). The luciferase assay was used to identify the causal variant from a GWAS association of the 1p13 genomic locus and low-density lipoprotein cholesterol (LDL-C) in humans (Musunuru et al., 2010). This study identifies haplotype-specific differences in transcriptional activity which was caused by a single SNP - rs12740374, and subsequently identified its target gene as SORT1 which has liver-specific expression. Another study focused on two cardiac traits, and combined results from luciferase reporter assays as well as 4-C chromatin conformation assay to identify variants that had allele-specific reporter activity to identify altered enhancer function of the SNPs - rs1743292/rs1772203 (X. Wang et al., 2016). Technological advances in this field have led to several approaches to multiplex and scale the reporter assays from testing one candidate element at a time.

Massively parallel reporter assays (MPRAs) allow for the testing of thousands of genomic regions to identify functional regulatory elements. The principle of this technology was first designed for promoter assays (Patwardhan et al., 2009) where the authors studied the effect of all possible point mutations of three promoter regions on gene expression levels. MPRAs were then used to test for enhancer function. For example, one study analyzed the effect of all possible mutations or deletions of two inducible enhancers on transcriptional activity in the HEK293T cell line (Melnikov et al., 2012). One study identified several variants within candidate cis-regulatory

elements to have an effect on enhancer activity, which are linked to known disease-causing loci (Tewhey et al., 2016). Another study identified 12 regulatory variants from a population-based study cohort from the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study (Vockley et al., 2015). Several other studies have used variants of MPRA to study GWAS loci (Castaldi et al., 2018; Mulvey & Dougherty, 2021; Ouwerkerk et al., 2020; Tewhey et al., 2016). These results underline the importance of tissue-specificity of reporter activity estimation and the cellular effects of changes in target gene expression

One of the MPRA technologies developed is STARR-seq (Muerdter et al., 2015), which stands for self-transcribing active regulatory region sequencing. STARR-seq allows for testing for regulatory activity of any genomic element inserted into the STARR-seq vector using high-throughput sequencing methods. Briefly, the library of regulatory elements to be tested is cloned into the STARR-seq vector which consists of a reporter gene and minimal promoter. Millions of fragments can be cloned into the vector yielding a library of pool of plasmids that contain the STARR-seq vector and the millions of fragments to be tested for reporter activity. The fragment to be tested is inserted into the 3'-UTR of the reporter gene which is followed by the poly-A (polyadenylation) site (Figure 1). This allows for the DNA to be transcribed as part of the mRNA processing when inserted into a cell. Therefore, reporter activity can be measured by comparing the quantities of mRNA transcripts that contain that particular fragment to the quantities of DNA containing that fragment obtained from the library of STARR-seq plasmid. That is, regions that have more RNA transcripts compared to DNA have greater candidate enhancer activity. STARR-seq has been used to generate quantitative enhancer activity across three *Drosophila melanogaster* cell types of developmentally different origins (Arnold et al., 2013). The human genome is 20x larger than the fly genome, and therefore adaptation of STARR-seq has required

both experimental and statistical modifications. One study used libraries of reduced complexity to test selected enhancer candidates (Vanhille et al., 2015). Another study scaled up STARR-seq to test the human genome for glucocorticoid (GC) stimulated changes in regulatory activity and demonstrated the use of STARR-seq data alongside orthogonal genome-wide functional genomics assays to identify putative GC-responsive regulatory elements (Johnson et al., 2018).



**Figure 1: Principle of STARR-seq reporter assay. This image of the construct design is adapted from Stark et al., 2015**

Reporter assays have allowed for testing thousands of candidate regulatory elements in multiple cell lines fairly easily. However, there are some limitations as well. First, most of these assays are episomal in nature where a plasmid is delivered to the cell and doesn't integrate into the host genome. This is problematic since the fragments to be tested do not have chromatin marks or are influenced by the binding of other factors, which is the native context of the genome. Furthermore, the effects of the enhancer–promoter distance and DNA looping are not factors of regulatory activity in MPRA. The fragments tested for regulatory activity in MPRAs have a limited size range either by limitations of plasmid size, or by limitations in fragmenting, synthesizing, or probe-capturing those fragments (200-500 bp) (Inoue & Ahituv, 2015). These limitations are problematic, since regulatory elements could be longer or have combinatorial

activity and thus are missed during testing important sequences. STARR-seq and MPRA results can include a level of false-positive and false-negatives which can vary because of technical problems in experimental approaches, sequencing, and statistical limitations (Gallagher & Chen-Plotkin, 2018). In addition to the experimental challenges, new technologies like STARR-seq have led to the development of appropriate analytical methods for data generated by these experiments. As STARR-seq experiments have progressively employed more complex libraries, one issue is that the coverage on the readout from these experiments is non-uniform, and confounded by biases such as GC content and shearing. Lee et al. (2020) and Kim et al. (2021) have published two different methods that allow for modeling biases in STARR-seq experiments to improve the detection of regulatory elements (Kim et al., 2021; D. Lee et al., 2020).

### **1.3.2 CRISPR-Cas9 genome and epigenome editing**

Genome- or epigenome-editing approaches have been used to dissect GWAS loci to better understand disease etiology. While statistical fine-mapping methods seek to prioritize causal variants, genomic annotation, and high throughput reporter assays can yield information on the prioritization of causal functional elements. For coding variants, the target gene can be inferred directly from its genomic location. However, for non-coding variants, the challenge in identifying the target gene for cis-regulatory elements (CRE) is confounded by the fact that these CREs can affect the transcription of genes over several kb and vary over developmental periods (Rao et al., 2021). Therefore, to investigate the functions of the prioritized CREs and/or causal variants, genome- or epigenome-editing allows us to determine the effects of perturbing those regions in the native genomic context. One of the systems used for perturbation is the CRISPR/Cas system, comprising CRISPR repeat-spacer arrays and Cas proteins, which was identified as part of an adaptive immune system in bacteria (Mojica et al., 2009). Most Cas

proteins are nucleases that cleave the targeted nucleic acid sequence and can be used to knock out a gene. This CRISPR/Cas complex protein also involves a guide RNA (gRNA) which acts as a targeting guide to bind to a specific location on the genome. Mutations in the Cas proteins can produce nuclease-deactivated Cas (dCas). The deactivated Cas protein retains its ability to bind to a target-genomic DNA but cannot cleave the double-stranded DNA. This property of dCas9 was taken advantage of to tether epigenetic effectors to the dCas9 (Gilbert et al., 2014; Konermann et al., 2015; Nakamura et al., 2021) . Tethering gene-regulatory proteins to dCas proteins has expanded the genome editing toolkit to include tools for transcriptional downregulation and upregulation (Fulco et al., 2016; Gilbert et al., 2014; Konermann et al., 2015; Liao et al., 2017; Thakore et al., 2015).

The CRISPR/Cas systems have been used to dissect GWAS loci, to test for target genes or pathways affected in complex diseases. For example, one study employed a genome-editing CRISPR screen and identified that depletion of HRI increases fetal hemoglobin levels, which is a therapeutic potential for hemoglobinopathies (Grevet et al., 2018). One study used CRISPR to disrupt likely variants associated with renal cancer susceptibility and identified gene regulatory regions that decreased the expression of oncogenic genes (Grampp et al., 2016). Another study used the epigenome-editing method by fusing a KRAB domain to dCas9 and identified that LINC00339 is the target gene responsible for the risk locus for osteoporosis (X.-F. Chen et al., 2018). Recent studies have employed high-throughput perturbation or epigenome editing to dissect GWAS loci and successfully identified target genes that contribute towards a disease (Morris et al., 2023; Schnitzler et al., 2024; Wünnemann et al., 2023).

The targeted editing of the epigenome using CRISPR-based approaches has several advantages. By perturbing a locus endogenously, we are testing for the target genes of regulatory elements within the native genomic context of that cell type, thereby capturing the cell-type specificities of gene regulation. Second, because the targeting of the Cas9 protein depends on a gRNA which can be synthesized individually, pools of gRNA can be used to efficiently target a single region or several regions at once. However, there are some disadvantages to this method. First, the range of the effects of epigenome editing is influenced by cell-type contexts (Gilbert et al., 2014). Another confounder is that it has been shown there is a strong correlation between the CRISPR activity scores to the distance from the transcription start site of the target gene (Gilbert et al., 2014). Despite these limitations, there have been successful examples of combining CRISPR-based methods with other functional experimental methods to dissect GWAS loci (Y. A. Cooper et al., 2022; Townsley et al., 2020).

#### ***1.4 Polycystic ovary syndrome is a complex genetic disease***

Specific case studies are instrumental for developing approaches that identify the mechanistic causes underlying genetic associations (French & Edwards, 2020; Gaulton et al., 2023; Heshmatzad et al., 2023; Khurana et al., 2016). One class of understudied complex genetic diseases involves reproductive health (Fabbri-Scallet et al., 2023; Marla et al., 2023; Owen et al., 2023). Polycystic ovary syndrome (PCOS) is a common complex disease that is characterized by reproductive traits including infertility, anovulation, and metabolic traits including obesity and type II diabetes risk and endometrial cancer risk.

### **1.4.1 Classification of a reproductive disease based on ovarian morphology**

The first study to describe the characteristics of polycystic ovary syndrome (PCOS) by Stein and Leventhal in 1953 identified clinical characteristics of PCOS in seven women to include amenorrhea, hirsutism, and polycystic appearance to their ovaries. They also reported that clinical features of menstrual irregularity and infertility could be improved by removing parts of the enlarged ovaries.

It is now understood that PCOS is an endocrine disorder and is a leading cause of infertility. Recent advances in medicine and endocrinology have led to several different criteria for diagnosis of PCOS, which includes quantitative measurements of hormones (Chang & Dunaif, 2021). The estimated incidence ranging from 5% to 15% varies based on the study population and diagnostic criteria (Dapas & Dunaif, 2022). PCOS involves dysregulation of the hypothalamus-pituitary-ovary axis leading to altered hormone levels (Lentscher & Decherney, 2021). In PCOS patients, altered hormone levels lead to infertility via impaired oocyte development and anovulation. Patients with PCOS also experience irregular menstrual cycles, acne, hirsutism, insulin resistance and obesity. There are several different criteria to diagnose PCOS (Chang & Dunaif, 2021). Common symptoms across those criteria include hyperandrogenism and ovulatory dysfunction.

### **1.4.2 Public Health Consequences of PCOS**

The CDC estimates that there are ~5 million people in the US who have been diagnosed with PCOS. Previous studies estimated that estimated healthcare costs for initial evaluation and treatment and PCOS is about 3 billion USD (Ricardo Azziz et al., 2005). However, because of the chronic nature of the disease, a recent report suggested that the healthcare burden for PCOS

should include pregnancy-related and long-term complication-related costs. This addition is reported to increase the average annual PCOS-related costs to be closer to 8 billion USD (Riesterberg et al., 2021). However, even with the high global prevalence of PCOS, it is still a syndrome that is poorly understood in terms of etiology, phenotypic presentation, diagnosis, and estimated prevalence (Ding et al., 2016). Understanding the molecular pathways impacted by PCOS will not only help in understanding the causes of the disease but also aid in improving diagnostic criteria and yield insight into potential avenues of drug development for treating PCOS.

### **1.4.3 Challenges in studying PCOS**

PCOS is a common complex condition in women associated with psychological, reproductive, and metabolic features. Currently, PCOS has no well-established hypothesis for disease pathophysiology. Given the differences in the symptoms and biochemical markers amongst the PCOS subphenotypes, PCOS can be thought to be a multifaceted disease with genetic and environmental factors affecting the disease. Therefore, a major challenge in the diagnosis of PCOS is that there are several different criteria for diagnosis, and there is yet to be a consensus formed amongst endocrinologists on a common definition (Chang & Dunaif, 2021). Thus, given the different criteria, there are several published PCOS patient-cohort studies that have used different diagnostic criteria in recruiting the patient population. Recently, there are studies that are aiming to better define biologically meaningful and discrete subphenotypes of PCOS based on biomarker profiles, responses to treatment, and/or genetic architecture. A recently published example of subphenotyping was performed in people with European ancestry and replicated with cohorts from Greek and Korean ancestry into reproductive, metabolic and indeterminate types (Dapas et al., 2020), based on both biochemical measurements and genetic

analysis. This study highlights the need for better classification of PCOS in the future. Additionally, there are several associated comorbidities for people with PCOS. This includes obesity, insulin resistance, increased cancer risk, non-alcoholic fatty acid liver disease, and cardiovascular diseases (Boudreaux et al., 2006; Diamanti-Kandarakis & Dunaif, 2012; Meyer et al., 2005; Vassilatou, 2014). Given the heterogeneity in phenotypic presentation and diagnosis, there are reports of decreased psychological well-being among PCOS patients (Jedel et al., 2010; Veltman-Verhulst et al., 2012). Another major challenge in treating PCOS is that there is no cure, partly as a result of the heterogeneity in phenotypes, diagnosis, and lack of information in disease etiology. Typically, PCOS cases are treated by symptom management. For example, metformin is a drug typically used for type-2 diabetes as an insulin-sensitizing agent and used to treat metabolic symptoms of PCOS. Clomiphene citrate is a selective estrogen receptor modulator that is used to induce ovulation in people with PCOS while oral contraceptive pills that contain estrogen and progestin are prescribed for regulating menstrual cycles and decreasing endometrial cancer risk (Hayek et al., 2016; Legro et al., 2013). Therefore, there is a large gap in our knowledge in understanding PCOS, and points towards several avenues of future research towards the aim of better diagnosis and treatment for PCOS patients.

#### **1.4.4 Steroids and nuclear receptors in PCOS**

One of the major phenotypes in PCOS is altered testosterone levels. Testosterone is a small molecule sex steroid that is necessary for sex differentiation, development, and fertility in mammals (Nassar & Leslie, 2023). While most testosterone is bound to proteins like sex-hormone-binding-globulin and albumin in the plasma, free testosterone in the blood exerts its effects on various tissues by binding to its cognate nuclear receptor to regulate gene expression.

Nuclear receptors are proteins that function by transducing signals of their cognate ligands when bound to the receptor. Small molecules like steroids and hormones function as ligands that can bind to nuclear receptors, causing the translocation of nuclear receptors into the nucleus and influencing gene expression by up- or down-regulating transcription levels. Type I subfamily 3 of nuclear receptors Type I nuclear receptors include members of subfamily 3, such as the androgen receptor (AR), estrogen receptors, glucocorticoid receptor (GR), and progesterone receptor. Genomic binding studies for these receptors in a cancer model demonstrated a very similar chromatin-binding landscape (Severson et al., 2018) and were shown to bind qualitatively to similar DNA sequences. Despite this similarity, AR and GR have differential effects on the transcriptional regulation of genes (Kulik et al., 2021; Rundlett & Miesfeld, 1995). AR mediates genomic responses through the binding of androgens such as testosterone and GR through the binding of glucocorticoids (GCs) like cortisol. In the ovary, both theca and granulosa cells are important regulators of oocyte development and express AR. Theca cells have receptors that bind to luteinizing hormone (LH) and produce testosterone. The ovarian granulosa cells have receptors for follicle stimulating hormone (FSH) and convert the testosterone produced by theca cells into estrogens. Structural variants in AR have been seen in PCOS cases but not controls (F. Wang et al., 2015), and expression of AR is altered in a sub-group of PCOS cases (Gao et al., 2020). Studies have also shown that AR expression in granulosa cells is crucial for ovarian development (Sen & Hammes, 2010). Recent studies have demonstrated a functional crosstalk between androgens and GR signaling, and further reported GR-based effects on hyperglycemia in mouse models of elevated androgen exposure (S. Li et al., 2023; Spaanderman et al., 2019). Taken together, these studies highlight the importance in understanding receptor function in the context of PCOS, elevated androgen and gene expression. Furthermore, the ability

of using exogenously added ligands that activate AR or GR makes it a useful system for studying the receptor-mediated gene expression responses.

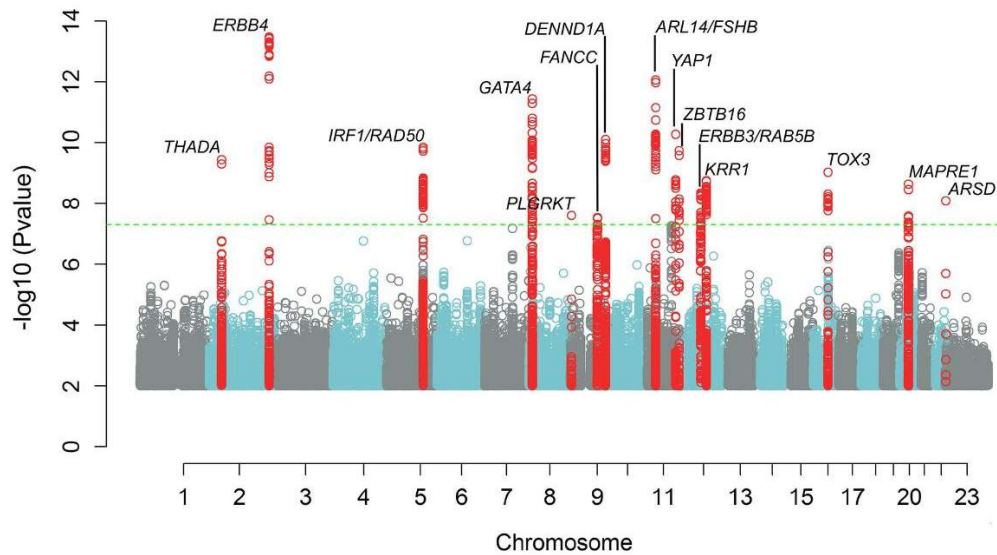
### ***1.5 The genetic basis of PCOS***

The first insight into the genetic nature of PCOS was described based on a familial clustering of PCOS phenotypes (Cooper et al., 1968). While familial clustering supports the role of genetic factors in the development of PCOS, the heterogeneity of PCOS phenotypes present in different PCOS patients even within the same family underscores the importance of the environmental contribution. PCOS is a common and complex trait, with twin studies estimating 70% of the phenotypic variation is explained by genetic variations (Vink et al., 2006). More recent studies have shown similar familial clustering in larger cohorts and support the genetic basis of PCOS (Franks et al., 2008; Govind et al., 1999; Legro et al., 1998; Lunde et al., 1989). Some studies have shown that the quantitative hormone measurements relevant to PCOS phenotypes are also highly heritable with observed familial clustering (Coviello et al., 2011; Franks et al., 2008; J. A. Harris et al., 1998; Ruth et al., 2020; Yildiz et al., 2006).

#### **1.5.1 PCOS Genome Wide Associate Studies**

The estimated heritability and genetic basis for PCOS and its quantifiable phenotypic characteristics provided a rationale for studying the role of genetic variation in contributing to PCOS development. Historically, linkage analysis is one approach to identify genetic markers of known chromosomal location that are co-inherited with the trait of interest to potentially locate disease-causing genes (Bush & Haines, 2010). However, for complex traits like PCOS, linkage testing is not as well suited because of polygenic effects, genetic heterogeneity and smaller effect

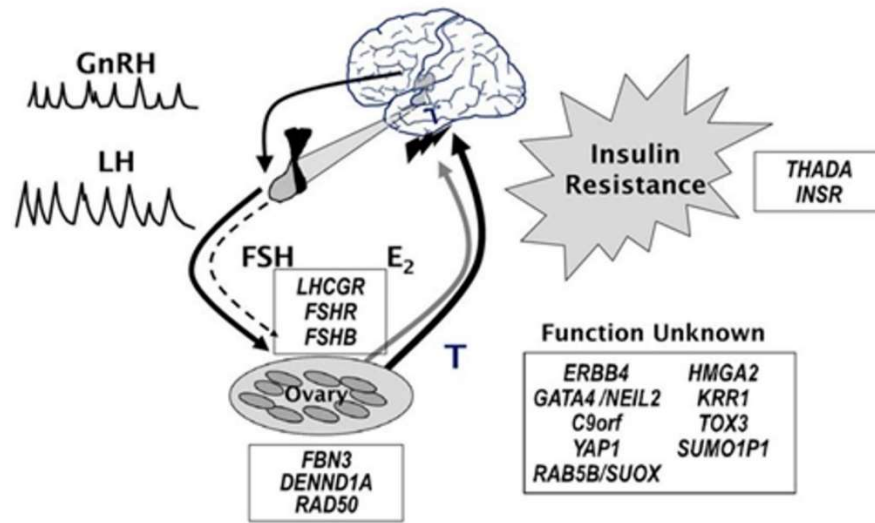
sizes of genetic variants (Bush & Haines, 2010). Genetic association testing has been performed in the cases of complex diseases like PCOS using case-control populations. The extent to which alleles at different loci are correlated in a population is referred to as linkage disequilibrium (LD). Genetic association testing is based upon the hypothesis that likely causal variants are present in the population that might tag through LD with another common variant as a genetic locus associated with the disease. With a large enough sample size, association tests can detect a number of genetic associations with modest effect sizes. That is, the null hypothesis is that in these population-based genetic association tests, variant allele frequencies are not different between case and control populations.



**Figure 2: Manhattan plot depicting several of the loci associated with PCOS, with the nearest gene indicated. Image adapted from Day et al., 2018**

The advent of human genome sequencing and subsequently genotyping array technology allowed the genetic association testing to be scaled up (I. H. G. S. Consortium et al., 2001). Association testing on a genome-wide scale, termed Genome Wide Association Studies (GWAS) has identified thousands of statistically significant associations between genetic variants and several different diseases (Welter et al., 2014). There are 30 genomic loci in the EMBL/EBI

GWAS Catalog that are associated with increased PCOS risk across nine multi-ethnic genome-wide association studies (Figure 2, GWAS catalog accessed 20 Oct 2023). The associated regions encompass genes involved in neuroendocrine, reproductive, and metabolic pathways (Figure 3).



**Figure 3: Graphical representation of the biological relevance of genes near PCOS risk loci. Adapted from Dapas et al., 2022**

Across four large multi-ethnic GWAS there are over 16 genomic loci that have been associated with increased PCOS risk (Z.-J. Chen et al., 2011; F. R. Day et al., 2015; Hayes et al., 2015; Y. Shi et al., 2012). About 12 of these loci have been replicated in more than one GWAS across different ancestries. The associated regions encompass genes that could contribute to PCOS via neuroendocrine, reproductive, and metabolic impacts (Figure 3) (Dapas & Dunaif, 2022). However, since trait-associated SNPs in GWAS are rarely causal variants themselves, but rather are linked to causal variants because of LD, GWAS results have yet to identify causal SNPs and genes towards the development of PCOS (Hirschhorn & Daly, 2005). Furthermore, the global prevalence of PCOS and the identification of the same genetic risk-loci through GWAS in

multiple ancestries point to the idea that PCOS related traits might be selected for from an early evolutionary timeline for human migration (Ricardo Azziz et al., 2011).

### ***1.6 Functional Follow Up of PCOS GWAS have identified several likely candidate genes.***

There is an emerging understanding within the complex trait biology of the interplay between common variants, trait polygenicity and rare variants (Karlsson et al., 2022; Maher, 2008; Manolio et al., 2009). Although rare variants may contribute less to the phenotypic variance of complex traits, disease-associated rare variants can point towards genes implicated in that disease (Momozawa & Mizukami, 2021). In the case of PCOS, there is evidence of both common and rare variants contributing to PCOS pathogenesis (Dapas et al., 2019, 2020). The emerging classification of PCOS into subtypes (Dapas et al., 2019) is an important step in understanding PCOS pathogenesis and the idea that the different traits of PCOS might be affected by different genetic variants, which in turn affect gene regulation in different biological contexts. Furthermore, in relating the PCOS sub-phenotyping genetic results with the omnigenic model of complex diseases, DENND1A can be thought of as a core PCOS gene. Thus, understanding the genetic causes of a complex disease like PCOS involves identifying the gene regulatory mechanisms in several pathways of the neuroendocrine and reproductive systems. In addition, due to cell type and developmental specificities, regulatory regions are harder to identify (Gte. Consortium, 2020; Kim-Hellmuth et al., 2020). To follow up on identifying potential causal variants and causal genes towards PCOS pathogenesis, there have been several different approaches to study it including experimental and statistical methods.

One of the significant loci identified is on chromosome 9 near the DENND1A gene. DENND1A encodes a protein (connecdenn 1) associated with clathrin-coated pits where cell-surface receptors reside. The DENND1A protein is located in the cytoplasm and in the nuclei of theca cells (McAllister et al., 2014). Research has shown that a splice variant of DENND1A is more expressed in PCOS theca cells compared to normal theca cells (Tee et al., 2016). This group also showed that exogenous overexpression of DENND1A led to increased androgen biosynthesis in human cell line models and mice (McAllister et al., 2014; Teves et al., 2020). Mice models with knock out of DENND1A were embryonic lethal (around E 14.5) and also affected development of primordial germ cells in mice (Jingjing Shi et al., 2019). Lastly, a family-based meta-analysis demonstrated a significant association of non-coding rare variants present within DENND1A locus with PCOS patients (Dapas et al., 2019). The function of DENND1A is still yet to be determined in the cellular contexts. ZNF217, a transcription factor, has been shown to be expressed in granulosa cells (Zhai et al., 2020). Recent studies have linked ZNF217 with DENND1A and showed decreased expression of ZNF217 in PCOS theca cells allow for the upregulation of DENND1A (Waterbury et al., 2022). Taking all these results together points to the hypothesis that the DENND1A locus is likely involved in PCOS pathogenesis.

There have been other studies that have tested some of the other PCOS GWAS loci and their relevance to PCOS pathogenesis. For example, FSHB encodes the beta subunit of the follicle-stimulating hormone which is involved in regulating the menstrual cycle and oocyte development. A study functionally dissected the lead GWAS SNP near the FSHB gene in mouse pituitary cells, and directly observed an allele-specific regulatory activity for rs11031006 and affected the transcription levels of FSHB (Bohaczuk et al., 2020). This locus also harbors another gene, ARL14EP. A transcriptome-wide association study identified that the increased expression

of the gene ARL14EP was significantly associated with PCOS risk (Lyle et al., 2023). Related to the hormone signaling pathway of FSHB, another PCOS GWAS risk locus includes the genomic regions surrounding the genes for LHCGR. A case study showed that inactivating mutations of a gene in a PCOS GWAS risk locus, LHCGR, led to oligomenorrhea and infertility (Toledo et al., 1996). Additionally, gene expression studies from adipose tissues identified that LHCGR was overexpressed in PCOS patient tissues compared to controls (Jones et al., 2015). The same paper identified a nominal association of RAB5B and IKZF4 with PCOS, as well as identified differentially methylated sites in this locus. Furthermore, the RAB5B locus harbors the gene RPS26. Loss of RPS26 in mice oocytes led to arrested follicle development and oocyte maturation. A whole exome sequencing study of theca cells from PCOS patients identified several variants associating PCOS with RAB5B, SUOX and ERBB3 in this locus (R. A. Harris et al., 2023). A study by Kulkarni et al., 2019, showed that DENND1A and RAB5B are present in the nucleus of PCOS theca cells, and suggest evidence of colocalization. Taken together, these different studies demonstrate that several other loci are likely contributing to PCOS pathogenesis with as of yet undiscovered genetic mechanisms,

In addition to these experimental methods, there have also been several statistics-based analyses of causal genes and variants for PCOS. A colocalization analysis identified strong evidence between rs227184 and expression levels of SUOX, ERBB3, RPS26 and reported that RPS26 expression is likely influenced by several variants in that locus (Censin et al., 2021). Another study identified RPS26 and NEIL2 as candidate causal genes for PCOS using a mendelian randomization approach using tissue expression data (Sun et al., 2022). However, systematic testing of these loci for the effects on PCOS pathogenesis has not been performed yet, and reflects the challenges in testing causality in the context of heterogeneous nature of PCOS.

The results from the studies highlighted above indicate that there is a genetic basis for PCOS, and that there are several genes likely involved in PCOS. However, one of the questions that remains is how PCOS-associated variants might be contributing to the disease risk through which pathways. In this dissertation, I have explored different approaches to identify and quantify the effects of regulatory elements and allele-specific regulatory variants in contributing to a phenotype. In chapter 2, I will describe our attempt in identifying regulatory elements that alter a key PCOS phenotype of testosterone production. In chapter 3, I will discuss methods to identify genetic variants that have allele-specific regulatory activity, and focus on one locus that is relevant to PCOS. Finally, in chapter 4, I will detail a new method to determine changes in gene expression levels that is agnostic to the type of genomic perturbation introduced as another approach to dissect non-coding regulatory elements.

## **2. Identifying gene regulatory mechanisms in PCOS GWAS loci<sup>1</sup>**

### ***2.1 Introduction***

Polycystic Ovary Syndrome (PCOS) is one of the most common disorders affecting people who menstruate with prevalence rates of 5% to 15%(Dapas & Dunaif, 2022), depending on the diagnostic criteria applied(Chang & Dunaif, 2021). It is the leading cause of anovulatory infertility. PCOS is commonly associated with insulin resistance and obesity, disorders that confer increased risk for type 2 diabetes as well as for other serious cardiometabolic morbidities across the lifespan(Diamanti-Kandarakis & Dunaif, 2012; Rubin et al., 2017). However, the cause(s) of PCOS remains unknown and the disorder is relatively understudied compared to other common medical conditions affecting women(Brakta et al., 2017)

Genetic factors are a major contributor to PCOS. Twin studies estimate that the narrow-sense heritability of PCOS is ~79%(Vink et al., 2006). There are 30 genomic loci that are associated with increased PCOS risk(Z.-J. Chen et al., 2011; Dapas et al., 2020; F. Day et al., 2018; F. R. Day et al., 2015; Hayes et al., 2015; H. Lee et al., 2015; Y. Shi et al., 2012; Sollis et al., 2022; Tyrmi et al., 2021; Yanfei Zhang et al., 2020) (GWAS catalog accessed 20 Oct 2023). The associated regions encompass genes involved in neuroendocrine, reproductive, and metabolic pathways. The functional consequences of noncoding genetic variants associated with complex traits such as PCOS have been exceptionally difficult to elucidate(Visscher et al., 2017; Vockley et al., 2017). One challenge of fine mapping GWAS signals is the difficulty in identifying causal genetic variant(s) from other genetic variants in regions of strong linkage disequilibrium (LD). In general, the lead GWAS SNPs are not the causal variants but are tagging regions of the genome

---

<sup>1</sup> This chapter has been adapted from a primary-author manuscript that is in review (Sankaranarayanan et al., manuscript under review)

containing non-coding pathogenic variants(Uffelmann et al., 2021; Visscher et al., 2017) that contribute to common disease risk by altering regulatory element activity and downstream gene expression(Pai et al., 2015; Roussos et al., 2014). Nevertheless, GWAS have provided considerable insight into PCOS causal pathways. *DENNDIA* was first identified as a PCOS candidate gene in GWAS(Dapas & Dunaif, 2022). *DENNDIA* was subsequently shown to be an important regulator of theca cell androgen biosynthesis where ectopic overexpression led to increased androgen production(Dapas et al., 2019; McAllister et al., 2014; Teves et al., 2020). Collectively, rare variants in *DENNDIA* were associated with PCOS quantitative traits in 50% of affected families(Dapas et al., 2019). Taken together with previous studies indicating that elevated testosterone levels were a consistent endophenotype in sisters of women with PCOS(Legro et al., 1998), these genetic analyses implicate *DENNDIA* as a core gene(Boyle et al., 2017) in PCOS pathogenesis. However, a mechanistic link between the noncoding genome, altered *DENNDIA* expression, and testosterone production has yet to be demonstrated.

The goal of this study is to systematically evaluate the effects of non-coding genomic regions associated with PCOS risk on gene regulatory element activity. To do so, we measured regulatory element activity across PCOS-associated genomic loci and identified genetic variants that alter that activity(Arnold et al., 2013; Johnson et al., 2018; Kheradpour et al., 2013; Muerdter et al., 2017; Tewhey et al., 2016). To measure regulatory activity, we used high-throughput reporter assays because they can quantify the regulatory activity of millions of genomic fragments at once. That scale enables systematic studies of the effects of non-coding variants across megabases of the genome and in many different cell types(Long et al., 2022; Muerdter et al., 2015; Shen et al., 2016; Tewhey et al., 2016; Vockley et al., 2015). To prioritize variants, we used a combination of allele-specific reporter assays, and targeted genetic association within the identified regulatory elements.

As proof of concept that the identified regulatory mechanisms contribute to PCOS, we perturbed PCOS-associated regulatory elements near *DENNDIA* using CRISPR-based epigenome editing (Canver et al., 2017; Klann et al., 2017; Thakore et al., 2016). We found that epigenetic activation of those regulatory elements in an androgen-producing adrenocortical cell model caused both increased *DENNDIA* expression and increased testosterone production. Together, these findings suggest a novel endogenous gene-regulatory mechanism contributing to PCOS; and demonstrate an approach for identifying additional molecular mechanisms of PCOS.

## **2.2 Results**

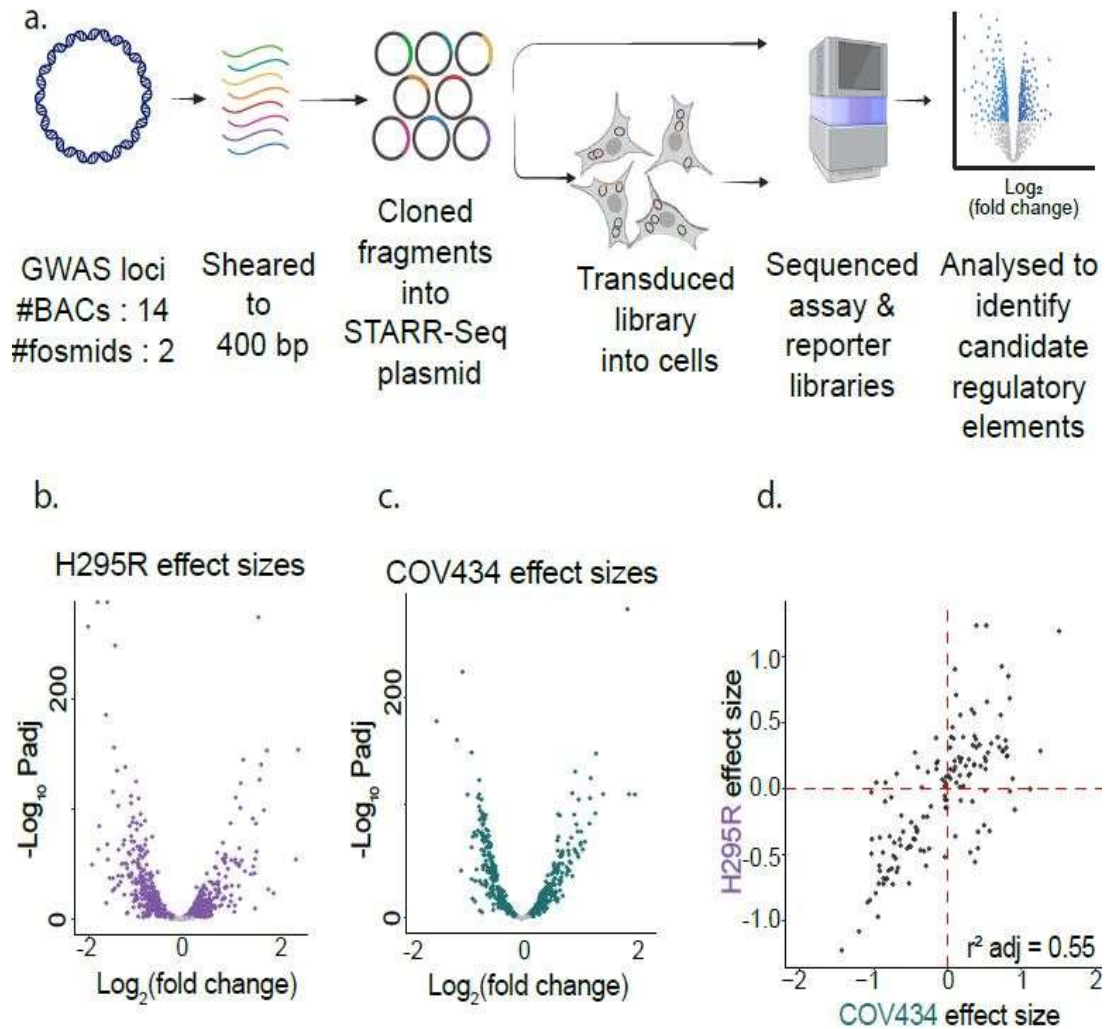
### **2.2.1 Measuring the regulatory activity of PCOS-associated regulatory elements**

To identify gene regulatory elements in which genetic variation can contribute to PCOS risk, we analyzed 14 genetic associations identified in cohorts of European and Han Chinese ancestry at the time of this study (Z.-J. Chen et al., 2011; F. Day et al., 2018; F. R. Day et al., 2015; Hayes et al., 2015; Y. Shi et al., 2012) (Table 1). Those 14 associations include several genes involved in hormone synthesis via the hypothalamic-pituitary-ovarian axis including *FSHR*, *FSHB*, *LHCGR* and *DENNDIA*. We focused on two human cell models: a testosterone-producing adrenal cell line, H295R; and an estradiol-producing ovarian granulosa cell line, COV434 (Lin et al., 2020; McAllister et al., 2019; Yu et al., 2019).

**Table 1: List of PCOS GWAS loci selected for STARR-seq experiments.**

<b>Genes in the locus</b>	<b>Genomic location</b>
<i>LHCGR/GTF2A1L</i>	2p16.3
<i>FSHR</i>	2p16.3
<i>THADA</i>	2p21
<i>DENND1A</i>	9q33.3
<i>RAB5B/SUOX</i>	12q13.2
<i>YAP1</i>	11q22.1
<i>GATA4/NEIL2</i>	8p23.1
<i>AOPEP</i>	9q22.32
<i>KCNA4/FSHB/ARL14EP</i>	11p14.1
<i>RAD50/IRF1</i>	5q31
<i>KRR1</i>	12q21
<i>HMGA2</i>	12q14.3
<i>TOX3</i>	16q12.1
<i>SUMO1P1</i>	20q13.2

To measure regulatory activity in these two cell lines, we used a high-throughput reporter assay known as STARR-seq (Arnold et al., 2013; Muerdter et al., 2017) (Figure 4a). STARR-seq can assay millions of DNA fragments for regulatory activity. STARR-seq assays work through two key libraries – the creation of an input library termed ‘assay library’ and a library of the regulatory effect readout termed ‘reporter library’ in this study. Briefly, the assay library consists of plasmid reporter assays containing diverse DNA fragments of interest. When transfected into cells, the DNA fragments regulate their own transcription into mRNA molecules. Thus, by sequencing the reporter library of the resulting mRNA fragments, one can estimate the regulatory activity of each DNA fragment in the assay library.



**Figure 4 Measuring the regulatory activity in PCOS GWAS loci.**

- a. Overview of targeted STARR-seq method:** We selected bacterial artificial chromosomes (BACs) or fosmids spanning 14 PCOS GWAS loci and sheared them to ~400bp. The sheared fragments were inserted into the digested STARR-seq backbone (Addgene#99296). The resulting plasmid library was sequenced to form the control assay library. For measuring regulatory activity, the plasmid pool was transfected into the respective cell lines (2  $\mu\text{g}$  plasmid pool / 1 million cells). Six hours post transfection, RNA was isolated from the cells, and the STARR-seq transcripts were enriched and sequenced as the output reporter library. Candidate regulatory elements were called using CRADLE(Kim et al., 2021) and effect sizes estimated with DESeq2(Love et al., 2014).
- b. STARR-seq effect size for H295R cells:** The effect size is estimated as pseudo  $\log_2$  (fold change) using DESeq2 on CRADLE-corrected STARR-seq peak calls.
- c. STARR-seq effect size for COV434 cells:** The effect size is estimated as pseudo  $\log_2$  (fold change) using DESeq2 on CRADLE-corrected STARR-seq peak calls.
- d. Comparing effect sizes of shared regulatory elements:** About 145 of the regulatory elements were shared between the two cell lines. The adjusted correlation coefficient is 0.55.

We constructed a STARR-seq assay library that spans 14 PCOS GWAS loci and encompasses 2.9 Mb of the human genome. The assay library includes 179 open chromatin regions identified in H295R and COV434 (Appendix A, Figure S1). The median fragment length in the assay library was 320 bp, and the 260 bp in the reporter library (Appendix A, Figure S2). Assay library covers the target region at a median of >300x (Appendix A, Figure S3) and replicates are highly correlated with Pearson correlation coefficient (Pearson' r) > 0.95 (Appendix A, Figure S4).

We called 956 regulatory elements in the 14 PCOS GWAS loci across the two cell models at a false discovery rate (FDR)  $\leq 0.5\%$  (Kim et al., 2021). Between replicates in the same cell model, the estimated regulatory element activity was highly correlated ( $0.84 \leq r \leq 0.90$ , Appendix A, Figure S5). Much of the observed variation in effect sizes can be attributed to differences between assay and reporter libraries, and differences between cell lines. The strong correlation suggests that the targeted STARR-seq approach robustly estimates regulatory activity for the cell types within the PCOS GWAS loci.

We identified 464 and 585 regulatory elements in COV434 and H295R cells, respectively. In both cell models, about half of the identified regulatory elements had enhancer activity, and half had repressor activity (Love et al., 2014) (Figure 4b and 4c). There were 93 regulatory elements identified in both cell lines. The regulatory activity of those commonly identified elements was highly concordant. The effect sizes in shared regulatory elements were substantially correlated (Pearson's  $r = 0.81$ ,  $p < 2 \times 10^{-16}$ ), and the direction of effects was the same for 85% of shared elements (Figure 4d). The concordance in the direction of effect increased to 93% when we required the regulatory element calls to overlap in the genome by at least 50%

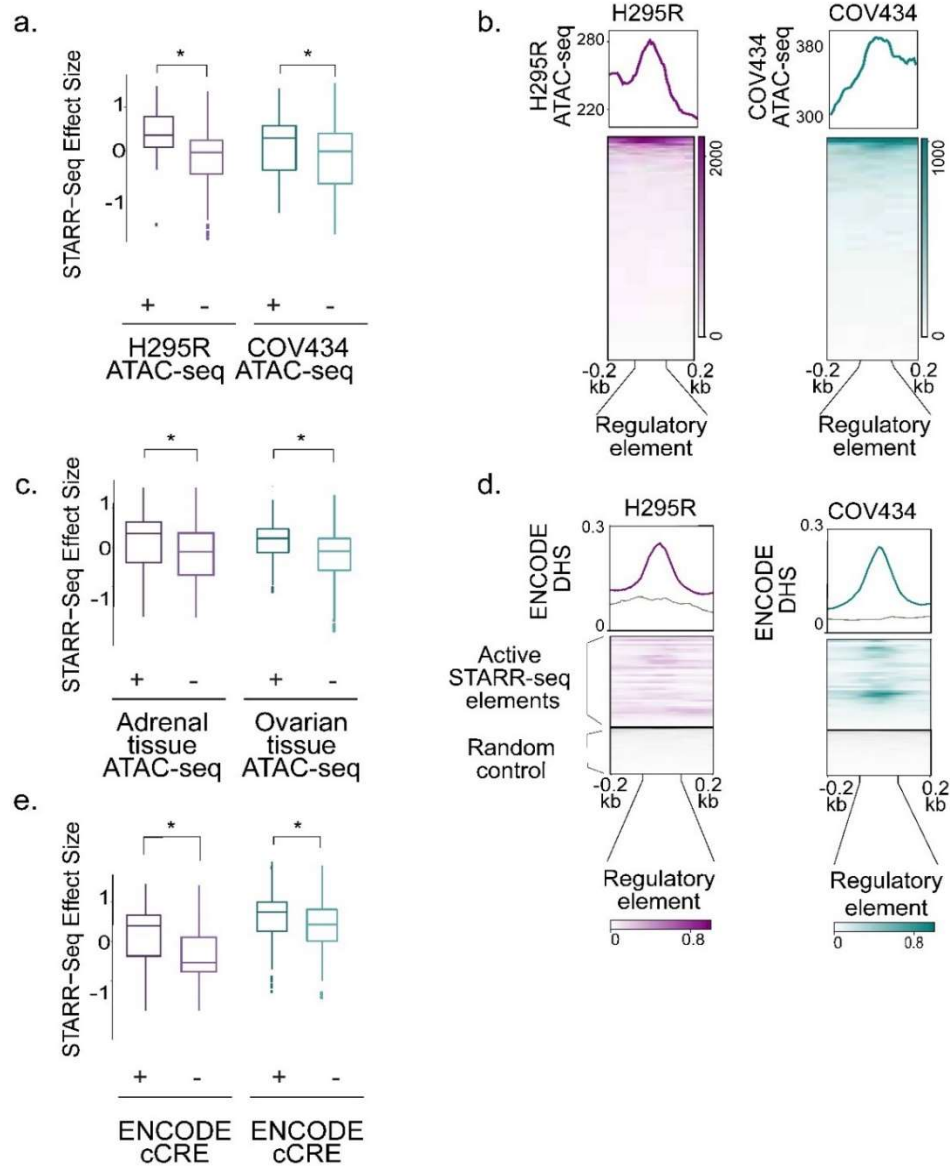
(Pearson's  $r = 0.85$ ,  $p < 2 \times 10^{-16}$ ). To our knowledge, this data set is the largest reporter-assay screen for enhancers in adrenal and ovarian cell models.

### **2.2.2 Regulatory element activity in PCOS GWAS regions corresponds to regions of chromatin accessibility**

Enrichment of genetic associations in tissue specific sites of increased chromatin accessibility can predict causal tissues of disease (Currin et al., 2021; Tehranchi et al., 2019). Of the PCOS associated SNPs in the GWAS catalog across the 14 loci we tested, five of those variants overlapped DNaseI hypersensitive sites (DHS) from the ENCODE consortium (Dunham et al., 2012). To increase confidence that the regulatory elements identified by STARR-seq are active in H295R and COV434 cells, we evaluated whether STARR-seq regulatory elements correspond to chromatin accessibility in the same cell lines. We identified ~73,000 and ~66,000 open chromatin sites in H295R and COV434, respectively, using ATAC-seq. Between 40 and 50% of the open chromatin sites identified in each cell line overlapped sites in the other (Appendix A, Figure S6, S7). Those results reveal a substantial number of chromatin accessible sites shared between cell lines.

There were 116 chromatin accessible sites within the 14 genomic regions we assayed with STARR-seq across each COV434 and H295R cell lines. Of those, 39 (34%) and 37 (32%) had regulatory activity in H295R and COV434 cells, respectively, according to STARR-seq assays. For H295R cells, the overlap between chromatin accessibility and STARR-seq activity is ~4-fold more than what would be expected if STARR-seq sites were randomly distributed across the genomic regions. For COV434, the overlap was ~6-fold more than expected by random (Fisher's exact test,  $p < 2 \times 10^{-4}$  for each). There was also significantly greater regulatory activity

in open chromatin regions in the same cell type or tissue than in regions with less chromatin accessibility (Figure 5a) (Mann-Whitney U,  $p < 10^{-10}$  for H295R,  $p < 0.01$  for COV434). Conversely, there was more chromatin accessibility in regions where we identified regulatory element activity (Figure 5b).



**Figure 5: Regulatory elements correspond to regions of chromatin accessibility**  
**a,c,e.** Candidate regulatory elements in H295R cells and COV434 cells with increasing activity correspond to regions with increased evidence of functionality. STARR-seq regulatory activity is measured across overlap with the respective cell line ATAC-seq (a), GTEx primary tissue ATAC-seq (c) and ENCODE candidate cis-regulatory elements (e).  
**b.** Aggregate profile plots of chromatin accessibility based on ATAC-seq on the respective cell lines centred on the candidate regulatory elements (with increasing and decreasing effect sizes) across 400 bp windows for both cell lines (H295R in purple, COV434 in teal).  
**d.** Aggregate profile plots of chromatin accessibility based on ENCODE DNaseI Hypersensitive sites (DHS) centred on the active candidate regulatory elements across 400 bp windows for both cell lines (H295R in purple, COV434 in teal). Control regions (grey) are randomly generated genomic regions that are chromosome-, length- and GC-matched to the STARR-seq elements.

We also investigated similarities and differences in regulatory activity between H295R and COV434 cells. There were 69 genomic regions that had significant regulatory activity and significant chromatin accessibility in either cell model. Of those, seven had regulatory activity in both cell models. The small overlap was due to differences in statistical power. Specifically, regulatory activity was similar across both cell types ( $\rho = 0.65$ , Appendix A, Figure S8). There was also no with no strong evidence of elements with opposing regulatory activity between cell types. Taken together, the high concordance of regulatory effect size across STARR-seq in H295R and COV434 suggests that regulatory activity is largely similar between the two steroidogenic cell lines.

To relate cell line observations to the corresponding primary tissues, we evaluated if STARR-seq regulatory activity is enriched in chromatin accessible sites in adrenal and ovarian tissues (Dunham et al., 2012; Luo et al., 2019). Approximately 18% of the identified H295R regulatory elements overlap with open chromatin from primary adrenal tissue, and 24% of the identified COV434 regulatory elements overlap with open chromatin from primary ovarian tissue. The overlap is a 2.8 and 3.1-fold enrichment in H295R and COV434, respectively, over what would be expected if regulatory elements were randomly distributed across the assayed regions (Fisher's exact test,  $p$ -value  $< 10^{-7}$  for both). As with our observations in H295R and COV434 cells, regulatory activity is greater in regions of accessible chromatin in primary tissue versus those without accessible chromatin (Figure 5c, Mann-Whitney U,  $p < 10^{-9}$ ). That result indicates that regulatory activity measurements in H295R and COV434 cells correspond to activity in primary adrenal and ovarian cells, respectively.

Regulatory activity in H295R and COV434 cells also corresponds to chromatin accessibility in other tissues. About 50% of the regulatory elements we identified via STARR-seq (n = 296 for H295R, n = 304 for COV434) overlap chromatin accessible sites identified in diverse tissues as part of the ENCODE project (Dunham et al., 2012; Luo et al., 2019). That overlap is 1.7- and 2.7-fold enriched over what would be expected if regulatory activity was randomly distributed across the assayed regions in H295R and COV434, respectively (Appendix A, Figure S10, Mann-Whitney U, corrected p-value <  $10^{-4}$ ). ENCODE DNase hypersensitive sites also had increased activity in STARR-seq regulatory elements (Fisher's exact test p <  $10^{-12}$ , Figure 5d, Appendix A, Figure S9). We observed similar results when focusing on enhancer-like regions defined across diverse cells and tissues by the ENCODE project (Abascal et al., 2020). Specifically, ~30% of the regulatory elements we identified overlap proximal or distal enhancers defined by ENCODE (n = 158 for H295R; n = 207 for COV434); and quantitative estimates of regulatory activity was greater in regions identified as enhancer-like sequences (Figure 5e).

### **2.2.3 PCOS-associated genetic variants fine-mapped to within regulatory elements**

To discover genetic variants that may alter regulatory activity and gene expression, we completed genetic association analyses focused on the regulatory elements we identified (Figure 6a). To identify additional risk variants within these functional regulatory elements, we first tested for genetic associations between common (minor allele frequency [MAF] >1%) single nucleotide polymorphisms (SNPs) and PCOS disease within the regulatory elements we identified. Across a cohort of 983 PCOS cases and 2951 controls (Hayes et al., 2015), we tested 759 SNPs in H295R cells, and 486 in COV434 cells. We found 19 variants that were significantly associated with PCOS at an adjusted p-value (Bonferroni correction) threshold of  $1.15 \times 10^{-4}$  –

2.55 x 10<sup>-4</sup> (Table 2). Of the associated variants, four were in the follicle stimulating hormone subunit beta (*FSHB*) locus, six were in the neighboring *ARL14EP-DR* locus and two were in the *GATA4/NEIL2* locus (Figure 6b and 6c, Table 2). There were four previously identified PCOS-associated risk variants in the regulatory elements we assayed (rs6022786, rs2268361, rs11225154, rs10835638)(Dapas et al., 2020; F. R. Day et al., 2015; Y. Shi et al., 2012) . Of those, only rs6022786 was tested in this analysis, and there was not a significant association with PCOS in our cohort.

**Table 2: Top variants associated with PCOS within STARR-seq regulatory elements. OR = odds ratio, MAF = minor allele frequency.**

CHR	BP	OR	P	MAF	rsID
11	30205987	1.452	3.08E-06	0.4044	rs568394
11	30195456	1.419	1.16E-05	0.4005	rs574096
11	30317914	1.544	1.49E-05	0.1676	rs7929660
11	30323044	1.544	1.49E-05	0.1676	rs4071558
11	30323178	1.544	1.49E-05	0.1676	rs4071559
11	30350068	1.485	1.05E-04	0.1605	rs12363432
11	30353061	1.465	1.14E-04	0.1754	rs7949790
11	30213950	1.452	2.58E-06	0.4189	rs586326
11	30295275	1.545	1.43E-05	0.1675	rs12278989
11	30347873	1.376	5.37E-05	0.4513	rs6484481
11	30359529	1.472	8.99E-05	0.1766	rs10835661
8	11766380	1.408	9.20E-05	0.2597	-
11	30233136	1.359	9.77E-05	0.4534	rs594982
11	30227279	1.358	1.00E-04	0.4517	rs1782509
11	30227537	1.357	1.04E-04	0.4532	rs1716024
11	30237442	1.357	1.07E-04	0.4527	rs615577
11	30199192	0.7338	1.15E-04	0.3929	-
8	11766907	1.405	1.18E-04	0.2518	rs34928882
11	30341554	1.488	9.67E-05	0.1601	rs7926666

To relate those variants to effects on gene expression, we next tested for colocalization(Giambartolomei et al., 2014) between PCOS-associated genetic variation in

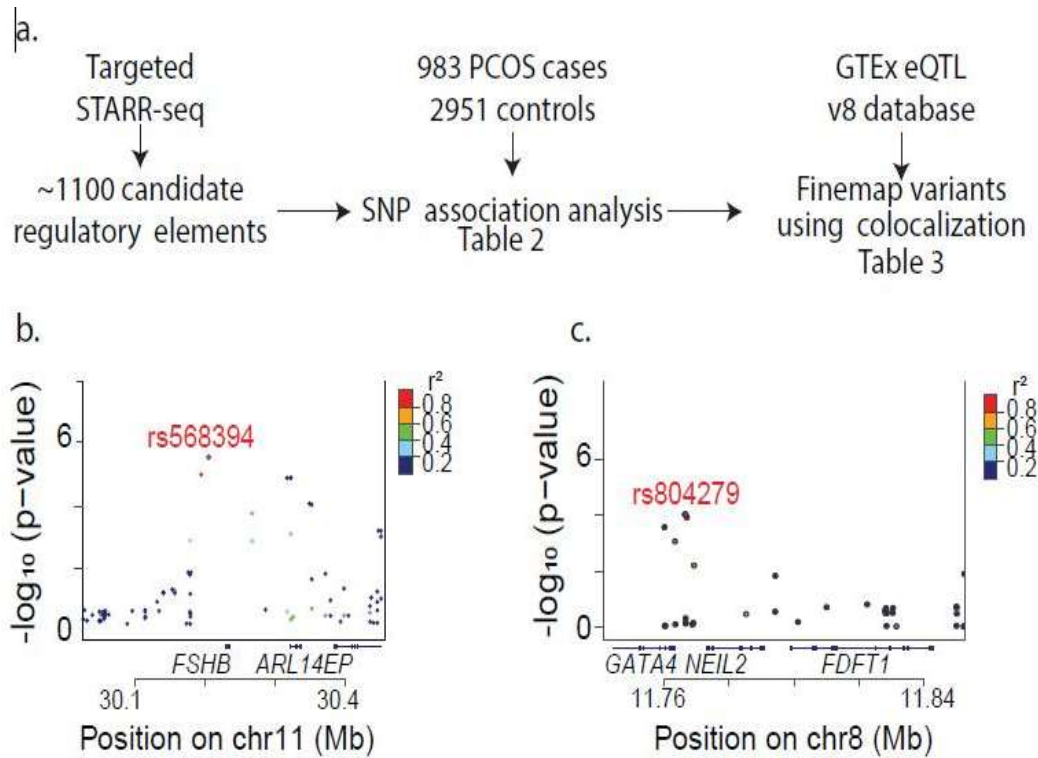
regulatory elements and expression quantitative trait loci (eQTLs) from GTEx(Gte. Consortium, 2020). Specifically, we used significant single tissue-eQTL association for this analysis. We identified seven variants in seven different loci where PCOS association and gene expression association colocalize with a posterior probability > 0.6 (Table 3). We further refined that analysis to only include adrenal and ovarian eQTLs. Focusing on those tissues has the advantage of being relevant to specific PCOS mechanisms but will also have reduced statistical power because there are fewer samples. Via that tissue-specific analysis, we identified four of the same variants. Two of those variants – rs804271 in the *GATA4/NEIL2* locus, and rs11349741 in the *FSHB/ARL14EP* locus – were also significant in the genetic association focused on regulatory elements described above.

**Table 3: Colocalization of PCOS-associated variants with eQTL data from GTEx. We identified 7 variants with high probability of likely causal of both PCOS-risk-association and altered gene expression from eQTL data. Colocalization was done using coloc (R package)**

Variant position	PP.AllTissue	Gene(s) associated from eQTLs
chr8:11769705	1	<i>GATA4</i>
chr12:55999553	1	<i>RPS26/RAB5B/SUOX</i>
chr20:53949370	0.86	<i>BCAS1</i>
chr9:123707451	0.84	<i>DENND1A</i>
chr2:48730442	0.76	<i>LHCGR</i>
chr11:30341402	0.71	<i>ARL14EP/FSHB</i>
chr9:94907391	0.69	<i>C9orf3/AOPEP</i>

There are reasonable biological mechanisms by which the expression of genes in each locus may impact PCOS. *GATA4* codes for a transcriptional factor that is involved in embryogenesis and functional studies showed that deletion of *GATA4* resulted in decreased fertility and granulosa and theca cell proliferation (Efimenko et al., 2013). Meanwhile, *FSHB* is a key peptide hormone for follicular development and thus a strong candidate gene for PCOS.

Together, these analyses further fine map genetic variants that alter regulatory activity and the expression of several genes in PCOS.



**Figure 6: Prioritizing PCOS-associated variants within functional regulatory elements**  
**a.** Association analysis to identify PCOS-associated variants in regulatory elements. We use the candidate regulatory elements from STARR-seq experiments to define the genomic regions of interest. We then performed an association analysis to identify variants associated with PCOS using a cohort of 983 PCOS cases and 2951 controls (results in Table 2). We then colocalized the association analysis results with GTEx eQTL SNPs to identify SNPs and genes as those likely involved in PCOS pathogenesis (results in Table 3).  
**b,c.** Locuszoom plots for PCOS-associated variants in candidate regulatory elements in FSHB/ARL14EP locus (b) and GATA4/NEIL2 locus (c).

#### 2.2.4 Active STARR-seq regions have increased conservation score

Evolutionary conservation is another indicator of biological function that is complementary to chromatin accessibility and STARR-seq. We also anticipate that genes affecting fertility will have strong evolutionary consequences. Previous studies have also reported

that conservation of regulatory elements corresponds to a greater functional role in the organism (Berthelot et al., 2018). Therefore, we investigated patterns of conservation across the regulatory elements we identified. We compared conservation scores of regulatory elements that we identified by STARR-seq across 20 vertebrate species (Siepel et al., 2005). The STARR-seq regulatory elements with enhancer activity had increased conservation score when compared to GC- and length-matched regions on the same chromosome (Appendix A, Figure S9b, Mann-Whitney U,  $p < 0.001$ ). We also observed that the accessible chromatin region identified by ATAC-Seq within COV434 and H295R cells have higher conservation scores when compared to similarly matched genomic regions from the same chromosome. Those results further corroborate the functional importance of the regulatory elements we identified.

### **2.2.5 Activation of PCOS-associated regulatory elements increased DENND1A expression**

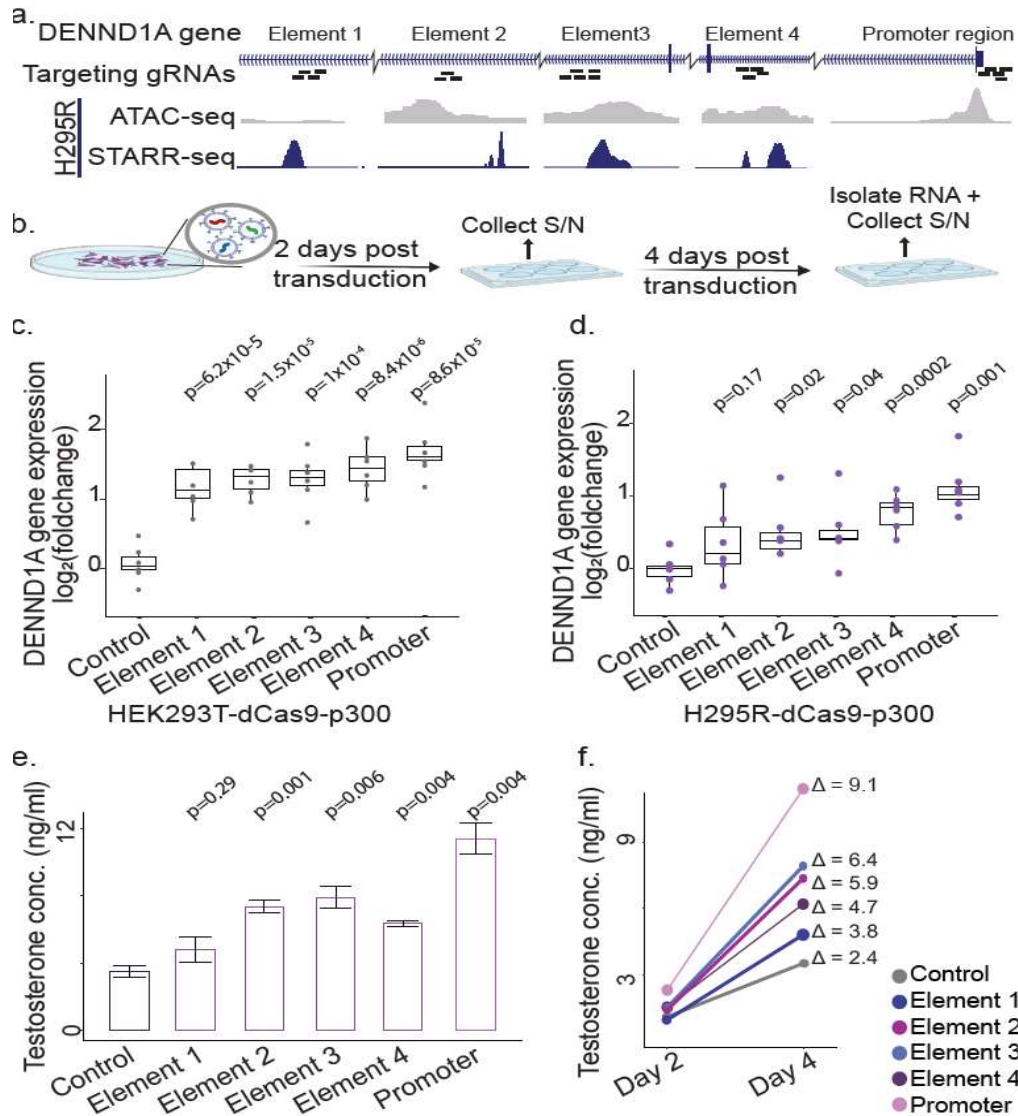
Estimating the effect of regulatory elements on altering gene expression can provide an insight into the underlying mechanisms that contribute to the development of PCOS. While reporter assays like STARR-seq can functionally test for allele-specific regulatory activity, the approach does not identify the target genes of those regulatory elements because the plasmids are not integrated in the genome. One approach to identify target genes of candidate regulatory elements is by epigenomic perturbation of that element. Specifically, a fusion of catalytically inactive Cas9 (dCas9) and histone acetyltransferase domain of P300 is targeted to candidate regulatory elements to measure the effects on the expression of nearby genes (Long et al., 2022). Several studies have demonstrated that dCas9-P300 can act over tens of kilobases, thus allowing the identification of distal gene regulatory elements (L.-F. Chen et al., 2019; Hilton et al., 2015; Klann et al., 2017).

To identify target genes of PCOS-associated gene regulatory regions, we created dCas9-P300-expressing H295R cells. We targeted dCas9-P300 to four candidate regulatory elements within the introns of the *DENNDIA* gene and to the *DENNDIA* promoter labeled “element 1-4” (Figures 7a, S21). We prioritized those four regulatory elements based on their proximity to *DENNDIA* gene, the strength of STARR-seq signal, chromatin accessibility, and the ability to design guide RNAs (gRNAs) targeting dCas9-P300 to the element (Figure 7a).

We designed 5-7 guide RNAs (gRNAs) for each of the four regulatory elements and promoter region. As a negative control, we also designed five guide RNAs to not target any location in the human genome. We made lentiviral pools for each of the four targeted regions and for the negative controls. We then transduced each lentiviral pool into two cell lines, H295R and HEK293T, that were modified to express dCas9-P300. *DENNDIA* is already expressed in both cell lines (average TPM: 20.4 for HEK23T and 15.4 for H295R) (Malm et al., 2020; Scholl et al., 2018), indicating that the gene is not in heterochromatin and thus may be targeted by dCas9-P300 effectively. Finally, we measured the effects on *DENNDIA* expression via qPCR, and levels of testosterone at two time points (Figure 7b). In HEK293T-dCas9-P300 cells (Klann et al., 2017), targeting dCas9-P300 to the *DENNDIA* promoter increased *DENNDIA* expression by 3.2-fold. Targeting dCas9-P300 to the intronic regulatory elements increased *DENNDIA* expression between 2.1-fold and 2.6-fold. The increase in *DENNDIA* expression was statistically significant compared to the effect of the non-targeting gRNAs for the promoter and all four of the regulatory elements after Bonferroni correction for multiple hypothesis testing (Figure 7c,  $\alpha < 0.05$ , t-test).

In H295R-dCas9-P300 cells, we observed a similar trend of increased *DENNDIA* gene expression for on-target CRISPR perturbation compared to the effect of the non-targeting

gRNAs. We observed a 2.2-fold increase in *DENNDIA* expression when targeting dCas9-P300 to the *DENNDIA* promoter. Targeting dCas9-P300 to the intronic regulatory elements increased *DENNDIA* expression between 1.2-fold and 1.6-fold. The increase in *DENNDIA* expression was statistically significant compared to the effect of the non-targeting gRNAs for the promoter and “element 3” of the regulatory elements after Bonferroni correction (Figure 7d,  $\alpha < 0.05$ , t-test).



**Figure 7: Perturbation of regulatory elements in DENND1A impacts testosterone levels**  
**a.** We targeted four candidate regulatory elements (Elements 1-4, also Figure S17) and the promoter regions of DENND1A. The STARR-seq activity (purple) and chromatin accessibility (grey) for these regions are shown. **b.** H295R cells that stably express dCas9-p300 were transduced with lentiviral pools of guide RNAs for each regulatory element. The cell media supernatant was collected 2- and 4-days post transduction for measuring testosterone concentration produced by the cells. RNA was harvested from the cells 4-days post transduction to measure gene expression. **c, d.** Log-fold change of DENND1A expression (GAPDH as control) for HEK293T (**c.**) and H295R (**d.**) cells stably expressing dCas9-p300. A set of 5 non-targeting control guides were designed to not target any part of the human genome as control cell population. **e.** Testosterone concentration (ng/ml) measured in the cell media 4 days post transduction in H295R-dCas9-p300 cells targeted with the specific guide RNA. **f.** Change in testosterone concentration (ng/ml) between day 2 and day 4 post transduction for each H295R-dCas9-p300 cells. The change in testosterone concentration between the two measurements is represented as  $\Delta$  (in ng/ml).

To test for off-target effects for genes in the *DENNDIA* locus, we also measured gene expression changes for *DENNDIA* flanking genes LHX2, CRB2, and STRBP. In both cell lines, we found CRB2 was not expressed and that expression of LHX2 and STRBP was not affected (Appendix A, Figure S10 and S11). We, therefore, inferred that the effects we observed were specific to *DENNDIA*, and that genetic variation in the region likely contributes to PCOS via effects on *DENNDIA* expression.

### **2.2.6 Activation of PCOS-associated regulatory elements increased testosterone production in steroidogenic adrenal cells**

Changes in gene expression levels may alter physiologically relevant phenotypes that can contribute towards disease pathogenesis (Musunuru et al., 2010). To test if endogenous overexpression of *DENNDIA* can alter testosterone production in H295R cells, we overexpressed *DENNDIA* by targeting dCas9-P300 to the *DENNDIA* promoter or distal regulatory elements. We then measured the concentration of testosterone in the cell culture media four days later. Increasing *DENNDIA* expression via activating the promoter caused a 3.2-fold increase in testosterone concentration, while activating three of the four distal regulatory elements individually increased testosterone concentration by between 1.7-fold and 2.2-fold (Figure 7e). The increases in testosterone concentration were statistically significant ( $\alpha < 0.05$ , t-test). As a complementary analysis, we measured the rate of increase in testosterone concentration over the four days post transduction. Overall, the rate of change of testosterone concentration mirrored the levels measured after four days. Specifically, cells with increased *DENNDIA* expression had substantially increased testosterone production between 2- and 4-days post-transduction compared to control-treated samples (Figure 7f). Those results indicate that altered expression of endogenous *DENNDIA* is sufficient to increase androgen biosynthesis in H295R cells.

## 2.3 Discussion

One of the central challenges of complex trait genetics is identifying the causal variants within GWAS susceptibility loci and determining their functional consequences. Here, we have fine mapped PCOS genetic associations to specific gene regulatory elements using a combination of high-throughput reporter assays and genetic analyses. Specifically, we have mapped candidate regulatory elements by testing for the regulatory activity of millions of DNA fragments across 14 PCOS GWAS loci comprising of about 3 Mb of the human genome. We further demonstrated a scalable approach to fine map genetic variants within candidate regulatory elements. We identified novel PCOS-associated genetic variants by performing genetic association tests across genomic regions that we identified as candidate regulatory elements. Together, we demonstrated a generalizable strategy for identifying genetic variants within experimentally identified functional regulatory elements to fine map genetic association loci for complex genetic diseases. As proof-of-concept of the strengths of this approach, we focused on *DENND1A*, a PCOS GWAS candidate gene reported to regulate androgen biosynthesis (McAllister et al., 2014). We showed that manipulating the epigenome of *DENND1A*-proximal regulatory elements caused increased *DENND1A* expression and, subsequently, increased androgen in human adrenal cells. These results extend previous studies identifying a role for *DENND1A* in testosterone production in theca cells, while also demonstrating specific gene regulatory elements wherein genetic variation can alter *DENND1A* expression. Our results demonstrate the advantage of combining high-throughput reporter assays, fine mapped genetic analyses, and targeted epigenome editing to discover novel gene regulatory mechanisms contributing to common human diseases.

The experimental approaches we used have several advantages and limitations. The targeted STARR-seq approach allows us to scale a high throughput reporter assay towards

multiple experimentally harder-to-manipulate cell lines that might not easily facilitate a whole-genome STARR-seq experiment. This approach also allowed us to test for regulatory activity outside context of genetic linkage(Vockley et al., 2015). Furthermore, the ability to capture natural genetic variation present in a pool of genomes allowed us to test for allele-specific regulatory activity across one locus in depth. It is understood that weak effects of non-coding variants contribute to a phenotype through coordinated regulation across several regulatory elements(Corradin et al., 2014). Thus, this approach allows us to identify regulatory elements identified that can to an organismal phenotype through gene expression patterns. A limitation in interpreting STARR-seq results is that the DNA fragments are not assayed in their chromatin context, and thus the effects on target genes in the genome remains unknown. For example, while the two cell lines used in this study have different steroidogenic properties, the concordance observed in the regulatory activity between the two cell lines reflects the fact that regulatory activity measured by STARR-seq is across a model of open chromatin. Though complementary studies of chromatin accessibility, eQTL associations, and evolutionary constraint support results from STARR-seq, definitive measurements require perturbation of the element in its genomic context. As proof of concept overcoming that limitation in this study, we showed that epigenetic activation of the elements in the genome caused increased DENND1A expression and subsequent testosterone production.

Identifying the underlying mechanisms is exceptionally valuable for realizing the potential of GWAS to benefit human health. There are several successful examples testing the effect of regulatory elements and non-coding variants contributing to disease phenotypes. For example, one study identified a SNP that regulates *SORT1* in a liver-specific manner in a GWAS risk locus for low-density lipoprotein cholesterol and myocardial infarction (MI)(Musunuru et al., 2010). Another study focused on maternal hyperglycemia identified variants spanning multiple

enhancers to have a coordinated effect on *HKDCI* expression (Guo et al., 2015). Other studies focused on post-GWAS functional analyses have used different methods including statistical (Giambartolomei et al., 2014; Hormozdiari et al., 2014, 2016; Ma et al., 2015; Maurano et al., 2012; Nicolae et al., 2010; Trynka et al., 2013; Wallace, 2021) and experimental (Kircher et al., 2019; Kneppers et al., 2022; S. Liu et al., 2017; Myint et al., 2020; Ouwerkerk et al., 2020; P. Zhang et al., 2018) approaches to fine map GWAS signals and identify functional variants. Notably, detailed cellular or molecular studies are often needed to connect the identified gene regulatory impacts to a disease relevant phenotype (Musunuru et al., 2010; Ouwerkerk et al., 2020). A challenge that may be difficult to make routine is the molecular follow up on putative causal genes, which is dependent on cell type, function of the genes and assays to measure the function of the gene with respect to the disease phenotype. PCOS, however, is particularly amenable to experimental perturbation since hormone responses are easy to model in cell systems and offer a potential for testing one of the main clinical phenotypes of PCOS. Our results extend the knowledge of non-coding genetic mechanisms of PCOS pathogenesis. Previous experimental studies characterized a highly conserved enhancer regulating *FSHB* expression in mouse pituitary cells (Bohaczuk et al., 2020, 2021); and non-coding variants intronic to *AMHR2*, a receptor for anti-Müllerian hormone (Gorsic et al., 2019). Previous statistical approaches have also nominated common and rare genetic variants altering the expression of *DENND1A* (Dapas et al., 2019), *FSHB*, *ZFP36L2*, *ERBB3*, *RPS26*, *RAD50* (Censin et al., 2021) as potentially contributing to PCOS. Here, we add both a specific gene regulatory mechanism controlling *DENND1A* expression to that body of knowledge, while also demonstrating a general strategy for identifying analogous mechanisms for other PCOS genes.

The candidate regulatory elements that we identified in this study can serve as a framework to identify functional non-coding regions that might contribute to PCOS risk by

harbouring causal variants. The results from our study adds to growing empirical evidence of regulatory regions contributing to complex diseases (Abell et al., 2022; Brandt et al., 2018; Kneppers et al., 2022; Soldner et al., 2016). We expect that future evaluation of the regulatory elements from this study will provide new insights into one of the mechanisms leading to PCOS phenotypes. Broadly, our results demonstrate a scalable approach to study disease-associated regulatory regions towards revealing mechanisms for PCOS.

## ***2.4 Methods***

### **2.4.1 Data Availability:**

All open chromatin regions and STARR-seq results for the studied regions are available via the a UCSC Genome Browser track

(<https://genome.ucsc.edu/s/laavatar/2024%2DPCOS%2DSS%2Dpaper> ).

### **2.4.2 Acknowledgements:**

The Genotype-Tissue Expression (GTEx) Project was supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript was obtained from the GTEx Portal on 7 June 2020.

### **2.4.3 STARR Seq Assay Library Construction:**

#### **2.4.3.1 Selection of GWAS regions for targeted STARR-seq assays**

To select PCOS-associated genomic regions for STARR-seq assays, we identified all genetic variants in linkage disequilibrium (LD,  $r^2 > 0.8$ ) with the 16 genetic variants that were most strongly associated with PCOS or its clinical phenotypes (Z.-J. Chen et al., 2011; F. R. Day

et al., 2015; Hayes et al., 2015; Y. Shi et al., 2012). We then selected bacterial artificial chromosomes (BACs) and fosmids that encompassed all the identified genetic variants. We obtained a total of 18 BACs and 2 fosmids. All BACs and fosmid clones were sourced from BACPAC Genomics, Inc and the source of these clones is Children's Hospital & Research Center at Oakland (CHRCO). The list of BACs and fosmids is detailed in Appendix A, Table S1.

All BACs and fosmids were obtained as clones in *E. coli*. We propagated each bacterial clone in selective conditions. We isolated the BAC DNA using NucleoBond Xtra BAC (Machery-Nagel); and we isolated fosmid DNA using FosmidMAX (Lucigen), following manufacturer's protocols. To validate that the BACs and fosmids were intact and covered the target region, we created Illumina high-throughput sequencing libraries from the isolated DNA using NEBNext Ultra II FS DNA Library Prep. We barcoded the sequencing library for each BAC or fosmid independently, and pooled the resulting libraries for sequencing. We sequenced the pooled libraries on an Illumina MiSeq instrument, and aligned to the human genome. For two of the 16 target regions, the BACs either recombined or the sequencing reads from the BAC aligned to a different genomic region suggesting contamination with another BAC. We removed those two regions from subsequent analysis. The BACs and fosmids for the remaining 14 target regions span ~3 Mb of the human genome.

#### **2.4.3.2 STARR-seq reporter plasmid construction**

To create STARR-seq assay libraries from the BACs and fosmids, we cloned sheared DNA from each BAC into the STARR-seq plasmid. We sheared each BAC or fosmid to ~400 bp DNA fragments using a Covaris S220 sonication instrument. We then ligated custom universal adapters to the resulting DNA fragments using the NEBNext DNA Library Prep protocol (#E6040L) (Appendix A, Table S3). We then amplified the adapted DNA fragments and added

sequences for Gibson assembly into the STARR-seq plasmid using PCR. For the PCR, we used KAPA HiFi HotStart kit (Roche) and the primers TS2SS-F and TS2SS-R (Appendix A, Table S3). The PCR cycling conditions were: 98 °C for 30 s, followed by 10 cycles of 98 °C for 15 s, 64 °C for 30 s, 72 °C for 30 s, with a final extension at 72 °C for 5 min.

We then cloned the fragment libraries into the STARR-seq ORI vector (Addgene#99296). To do so, we first linearized the plasmid using AgeI and SalI (NEB R3552L and NEB R3138L). We then ran the digested plasmid on a 1% agarose gel, confirmed that the linear plasmid was the expected ~3600 bp size, and isolated the linearized plasmid using either the QIAquick Gel Extraction Kit (#28704) or GeneJET Gel Extraction Kit (#K0691). We then cloned the adapted and amplified DNA fragments from the BACs and fosmids into the linearised STARR-seq ORI vector using the NEBuilder HiFi DNA Assembly (#E2621) kit. We ethanol precipitated the products. To do so, we added 0.1X volume 3 M NaOAc and 2.5X volume cold 100% ethanol and stored the mixture at -20 °C overnight. We then pelleted the DNA via centrifugation at 16,000 RCF for 30 min at 4 °C. We washed the pellets with 5 ml cold 70% ethanol, and resuspended them in water. To amplify the resulting plasmid libraries, we electroporated into E. cloni 10G SUPREME Electrocompetent Cells following manufacturer protocol for optimal settings in 1.0 mm cuvette (10 µF, 600 Ohms, 1800 Volts). We grew the plasmids in individual 1 L volumes of LB with carbenicillin for antibiotic selection at 37 °C overnight. We isolated the resulting PCOS GWAS STARR-seq assay plasmids using NucleoBond PC 10000 EF (Machery-Nagel).

To make the final PCOS GWAS STARR-seq assay library, we pooled the individual BAC and fosmid STARR-seq plasmids in equimolar concentration. We validated the size of the plasmid library using the Agilent TapeStation, and quantified the resulting pool using Qubit (Invitrogen).

### **2.4.3.3 PCOS GWAS STARR-seq assay library sequencing**

To estimate the abundance of reads mapping to the regions selected in the assay library, we used Illumina high-throughput sequencing NextSeq 2000 with 50bp paired end sequencing protocol. To prepare the sequencing libraries, we first amplified the STARR-seq regions from a 20 ng pooled plasmid library using KAPA HiFi HotStart kit (Roche). The PCR cycling conditions were: 98 °C for 30 s, followed by 15 cycles of 98 °C for 15 s, 64 °C for 20 s, 72 °C for 30 s, with a final extension at 72 °C for 5 min using 208-F Index7 primers (Appendix A, Table S3). To isolate the final library, we used Axygen Spri Beads (AxyPrep™ Mag PCR Clean-Up Kit) beads at appropriate concentrations based on the manufacturer's manual for an insert size of 400 bp.

### **2.4.4 Cell Culture protocol**

We obtained NCI-H295R cells from ATCC. The cells were cultured in DMEM/F-12 medium (Gibco #21041025) supplemented with 2.5% Nu-Serum (Corning #355100) and 1% ITS+Premix (Corning #354352) and grown as a monolayer at 37 °C, 5%CO<sub>2</sub>. We validated testosterone produced by the cells stimulated with 10 μM forskolin using ELISA following manufacturer's protocol (Cayman Chemicals #582701).

We obtained COV434 cells from ECACC (Sigma–Aldrich #07071909). The cells were cultured in DMEM (Gibco #11965092) supplemented with 2 mM Glutamine and 10% Foetal Bovine Serum (FBS) and grown as a monolayer at 37 °C, 5%CO<sub>2</sub>. We validated estradiol produced by these cells treated with 100 ng/mL follicle stimulating hormone (FSH) and 2.9 μg/mL androstenedione (A4) using ELISA (Cayman Chemicals #501890). All experiments were performed between passages 5 and 12.

#### **2.4.4.1 Nucleofection Optimization**

To transiently introduce the PCOS GWAS STARR-seq assay library into the cell lines, we used electroporation via the Lonza 4D-Nucleofector System. To optimize the electroporation settings for H295R and COV434, we used the Cell Line Optimization 4D-Nucleofector™ X Kit (Lonza #V4XC-9064) following manufacturer's protocol. Based on this optimization, we chose SF-CM-138 for COV434 cells and SF-DN-100 for the H295R cells, with 2 µg of plasmid to every 1 million cells transfected.

#### **2.4.4.2 Transfection of cells**

To test the regulatory potential of PCOS GWAS targeted regions, we first transfected the PCOS GWAS STARR-seq library into both H295R and COV434 cells, and isolated and sequenced the resulting RNA. We isolated the RNA from the cells 6 hours post transfection. We transfected the PCOS GWAS STARR-seq plasmid library into H295R and COV434 based on the nucleofection optimization settings we described using SF Cell Line 4D-Nucleofector® LV Kit L (Lonza #V4LC-2002) following manufacturer's protocol. All the experiments were performed in triplicate for each cell line. For each replicate for each of the cell lines, we used 50 million cells transfected with 100 µg of the PCOS GWAS STARR-seq plasmid library.

#### **2.4.5 PCOS GWAS STARR-seq reporter library construction**

To isolate the PCOS GWAS reporter RNA, we first isolated total RNA followed by enriching for cDNA produced from the PCOS GWAS STARR-seq library plasmid pool.

Six hours post transfection, we rinsed the cells with PBS and dissociated the cells using Trypsin-EDTA 0.25% (Life Technologies). We lysed the cell pellets using RLT buffer (Qiagen) with 2-mercaptoethanol (Sigma). We passed the lysates through a 18-gauge needle ten times and stored at  $-80^{\circ}\text{C}$  before RNA extraction.

#### **2.4.5.1 RNA extraction**

We isolated total RNA using the Qiagen RNeasy Midi kit including the on-column DNaseI digestion step. We treated the isolated total RNA with 1  $\mu\text{L}$  RNase Block (Agilent). We then isolated poly-A RNA using Dynabead Oligo-dT25 beads (Life Technologies) according to the manufacturer's recommended protocol. We treated the poly-A RNA with DNase (TURBO DNase, Invitrogen) and 1  $\mu\text{L}$  RNase Block at  $37^{\circ}\text{C}$  for 30 min before halting the reaction with the DNase inactivation reagent. We then synthesized PCOS GWAS reporter cDNA by reverse transcription using Superscript III (800 U, Life Technologies) following manufacturer's protocol and a STARR-seq specific primer (SSRT-UMI, Appendix A, Table S3).

#### **2.4.5.2 PCOS GWAS STARR-seq reporter construction**

Following synthesis, we treated the cDNA with RNaseA (Sigma) at  $37^{\circ}\text{C}$  for 1 hour. We purified the PCOS GWAS reporter cDNA with SPRI beads (1.5X) and amplified using index-PCR primer and indexed PostSS-Index-5 primers (Appendix A, Table S3) to allow barcoding for sample multiplexing under the following conditions:  $98^{\circ}\text{C}$  for 30 s, followed by 10-12 cycles of  $98^{\circ}\text{C}$  for 10 s,  $64^{\circ}\text{C}$  for 30 s,  $72^{\circ}\text{C}$  for 30 s, with a final extension at  $72^{\circ}\text{C}$  for 5 min. We split each sample into 7 individual PCR amplification reactions in this step. We determined the total number of cycles for amplification using a small portion of that sample in a qPCR protocol and estimating cycle number using  $1/4^{\text{th}}$  the maximum plateau observed in the qPCR. We cleaned the

amplified PCR products using SPRI beads (1.0X) and then validated the length distribution of the PCOS GWAS reporter library on Agilent tape station.

#### **2.4.5.3 PCOS GWAS STARR-seq reporter library sequencing**

Final PCOS GWAS reporter libraries from each replicate experiment were pooled at equimolar 2nM concentrations. We sequenced the PCOS GWAS reporter libraries on Illumina NextSeq 2000 using 50bp PE sequencing.

#### **2.4.5.4 Alignments and STARR-seq analysis**

To estimate regulatory activity in the targeted PCOS GWAS regions, we used the abundance of the fragments expressed as RNA in the reporter library relative to their abundance in the assay library. We first aligned the PCOS GWAS assay library and the PCOS GWAS reporter library individually to the human genome (hg38) using bowtie2. We filtered reads with a quality score of  $Q \geq 30$ , and outside the centromeres and blacklisted regions. These reads were used for the downstream analysis. We used picardtools(Institute, 2019) to mark and call duplicates. RPKM normalized STARR-seq read density was computed at single base pair resolution using deepTools(Ramírez et al., 2016) utility bamCoverage. We used CRADLE(Kim et al., 2021) package to correct biases and call peaks with the following options. We then estimated differential STARR-seq activity across the regions as fold change using DESeq(Love et al., 2014) . We compared PCOS STARR-seq results from both COV434 and H295R cell lines to ATAC-Seq datasets generated for these cell lines. We also compared the peaks to the regulatory regions across ENCODE (V4) for both cell lines and primary tissues.

#### **2.4.6 PCOS case-control variant association testing within candidate regulatory regions**

To identify any association between genetic variants within functional STARR-seq regulatory elements and PCOS, we performed an association test. PCOS association was tested within candidate regulatory regions using 983 PCOS cases and 2951 controls that were genotyped using the Illumina OmniExpress (HumanOmniExpress-12v1\_C) array(Hayes et al., 2015). Genotype imputation was performed using minimac4(Fuchsberger et al., 2015) on the Michigan Imputation Server(Das et al., 2016) for phasing via Eagle(Loh et al., 2016) using the TOPMED freeze 8 reference panel(Das et al., 2018; Taliun et al., 2021). Variants were filtered to remove any SNPs with imputation quality ( $R^2$ ) less than 0.8 and restricted to STARR-seq regions of regulatory activity. Single variant associations were carried out using PLINK(Purcell et al., 2007) on common variants (minor allele frequency [MAF]>1%) using logistic regression with PCOS as the outcome variable and age, BMI, and five principal components (PCs)(Price et al., 2006) as covariates. To control for false-positive discoveries, results were adjusted for Bonferroni correction thresholds.

#### **2.4.7 Colocalization testing**

To test for association between two datasets to identify likely causal SNP between two traits, we used a bayesian colocalization method(Giambartolomei et al., 2014). For the PCOS-associated variants, we used the list of variants and its associated statistics from the above result for all variants with  $P < 0.3$  (Table 3). We used the standard options for the colocalization testing. For the eQTL dataset, we used publicly available expression quantitative association data from the GTEx consortium GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2, accessed on June 7, 2020). The GTEx dataset contains cis-eQTL data from ~900 American donors of mostly

European Ancestry (~85%) across 49 tissues and of varying ages. We applied coloc, a Bayesian test for colocalization to identify the probability of a shared causal signal between the PCOS-regulatory element-associated variant and eQTL variants. We used the coloc.abf() function in the coloc R package with the default assignment of prior probabilities for a SNP being associated with each trait from the Coloc package. All analyses with a colocalization posterior probability (PP.4) > 0.3 using eQTL data from all tissues, adrenal tissue and ovarian tissue were reported in Table 4.

#### **2.4.8 ATAC-Seq**

To identify the accessible chromatin within the H295R and COV434 cells, we performed ATAC-Seq(Corces et al., 2017) in duplicate as described below.

We harvested 50,000 viable cells for each replicate. COV434 cells were additionally incubated in TURBO DNase (Invitrogen, #AM2238) for 1 hour at 37 °C. We then incubated the cells with 50 µL cold ATAC-RSB with 0.1% NP40, 0.1% Tween20, 0.01% Digitonin and incubated on ice for 3 minutes. We washed the cells with 1ml cold ATAC RSB with 0.1% Tween 20 and pelleted. We resuspended the cell pellets in the transposition mixture comprising of 25 µL TD buffer, 2.5 µL transposase, 16.5 µL PCS, 0.5 µL 1% digitonin, 0.5 µL 10% Tween 20 and 5 µL H2O and incubated in a thermomixer at 37 °C for 30 minutes. We cleaned up the DNA using MinElute Reaction Cleanup Kit (Qiagen, #28204). We amplified the resulting DNA using an ATAC-Universal primer and an ATAC-barcode primer (Appendix A, Table S3) and cleaned it using SPRI beads. We sequenced the ATAC seq libraries on Illumina NextSeq 550, 50 bp PE sequencing at the Duke Genomics core.

### **2.4.8.1 ATAC-seq preprocessing and alignment.**

ATAC-seq libraries for H295R and COV434 cell lines were individually aligned to the human genome (hg38). Each cell line had 2 biological replicates, and >40 million reads were generated per sample. Sequencing data quality was assessed with FastQC, and adapters were trimmed with Trimmomatic. Trimmed reads were aligned to the GRCh38 genome using Bowtie(Langmead & Salzberg, 2012) reporting only alignments having no more than two mismatches, discarding multi-mapping reads(-v 2 --best --strata -m 1). Reads mapping to the ENCODE hg38 blacklisted regions (<https://www.encodeproject.org/files/ENCFF356LFX>; manually curated regions with anomalous signal across multiple genomic assays and cell types) were removed using bedtools2 intersect(Quinlan & Hall, 2010) (v2.25.0). Properly paired reads were then filtered to exclude presumed PCR duplicates using Picard MarkDuplicates (v1.130; <http://broadinstitute.github.io/picard/>). Reads were then used to generate reads per million (RPM) counts of bigWig files for visualization using deeptools bamCoverage(Ramírez et al., 2014) (v3.0.1). Peaks were called using MACS2 with an FDR cutoff 0.1. We used the ENCODE ATAC-Seq standards for analysing the dataset we generated. We generated Transcription Start Site enrichment values using GRCh38 Refseq TSS annotation and used the cutoff of >7 for high quality data (Figure S8).

## **2.4.9 CRISPR-dCas9 epigenome editing**

### **2.4.9.1 GuideRNA (gRNA) design and gRNA plasmid synthesis**

Four candidate regulatory elements were identified from the targeted STARR-seq results with coordinates listed in Table 9. The regions were selected based on STARR-seq effect, chromatin accessibility and ability to design guides considering genomic sequence and PAM restrictions.

To design the guide oligos, we used Guidescan2(Perez et al., 2017), with “specificity” filter > 0.2. We had a total of 21 gRNAs, across four candidate regulatory elements and *DENND1A* promoter region (Appendix A, Table S2), with each regulatory element comprising of 5-7 guides targeting that element. For the negative control, we designed a set of five guides that did not have any targets in the human genome. Each gRNA oligo was synthesized as individual oligos that were then processed as described below to make pooled gRNA plasmids.

To make the gRNA plasmids, we followed the outline of the CROP-Seq protocol(Datlinger et al., 2017). First, we prepared the gRNA plasmid backbone by digesting CROPseq-Guide-Puro plasmid from Addgene (#86708) using BsmBI. We ran the digested product on 1% agarose gel, and we purified the 8.3 kb fragment using GeneJET Gel Extraction Kit (#K0691). To prepare the gRNA oligos for insertion into the plasmid, for each gRNA oligo synthesized, we first converted it to a double stranded oligo using Primers ssds-F and ssds-R (Appendix A, Table S3). We then cloned each double-stranded gRNA oligo into the digested CROPseq-Guide-Puro vector using NEBuilder HiFi DNA Assembly (#E2621) kit. The plasmid products were purified with QIAquick PCR Purification Kit (Qiagen #28104).

To make the pooled plasmids, we pooled (equimolar) each plasmid product for each regulatory element, or promoter region, or negative control. To amplify the plasmid pools, we electroporated each pool into Lucigen Endura Cells (Lucigen #60242-2) following manufacturer protocol for optimal settings in 1.0 mm cuvette (25  $\mu$ F, 200  $\Omega$ , 1.5 kV). We grew the plasmids in individual 25 mL volumes of LB with carbenicillin for antibiotic selection at 37 °C overnight and isolated the gRNA plasmid pools using Qiagen Midi Prep (Qiagen #12143) following manufacturer’s protocols. Each purified plasmid pool was then used to prepare lentiviral particles.

### 2.4.9.2 Lentivirus production

To test the target gene of the identified STARR-seq regulatory elements, we used CRISPRa to perturb the selected candidate regulatory elements. First, we designed a stable cell line expressing a Cas protein. To do so, we used a catalytically inactive Cas9 (dCas9) fused with the P300 domain of histone acetyltransferase (dCas9-P300). This dCas9-p300 can act as a transcriptional activator when combined with targeting guide RNA (Klann et al., 2017).

To make stable dCas9-P300 cell lines, we generated lentivirus expressing dCas9-p300. Briefly, we combined the following plasmids: dCas9-p300 (Addgene #83889), psMD2.G (Addgene #12259) and psPAX2 (Addgene #12260) with Lipofectamine 3000 (Invitrogen #L3000001) and lipofected into HEK293T cells (ATCC #CRL-3216™) according to the manufacturer's protocol. After 14 to 20 hours, transfection media was exchanged with fresh media. We then harvested viral supernatant at 24 and 48 hours post lipofection. We concentrated the viral supernatant at 1/100x using LentiX Concentrator (Clontech #631232) following the manufacturer's protocols.

To produce lentivirus for individual gRNAs, we transfected HEK293T cells with an equimolar pool of gRNA plasmids for each regulatory element, psPAX2, and psMD2.G using Lipofectamine 3000 following the manufacturer's instructions. We harvested media containing the produced lentivirus at 24 and 48 hours later and concentrated the viral supernatant at 1/100x using LentiX Concentrator (Clontech #631232) following the manufacturer's protocols.

HEK293T cell line with stable dCas9-P300 expression: We received HEK293T-dCas9-P300 cell line (Klann et al., 2017) from Dr. Charles Gersbach. We followed the published culture and growth conditions for 293T cells.

#### **2.4.9.3 Generating stable H295R-dCas9-P300 cell line**

To make stable H295R cells expressing dCas9-P300, we transduced the concentrated lentiviral particles containing dCas9-p300 into H295R cells with a multiplicity of infection of 5.0 using 6 µg/ml of polybrene (EMD Millipore Corporation #TR-1003-G). Additionally, we selected for the transduced cells using 0.5 µg/mL of puromycin (Gibco #A1113803). We confirmed the expression of dCas9-p300 in H295R cells using qRT-PCR.

#### **2.4.9.4 Transduction of gRNA into dCas9-P300 expressing cell lines:**

To test the effect of P300 on the targeted regulatory elements, we transduced each lentiviral pool for the regulatory elements, *DENND1A* promoter region and negative control in two cell lines (HEK293T and H295R) with stable dCas9-P300 expression. We transduced the cells during seeding in a 12-well or 6-well plate supplemented with 6 µg/ml of polybrene for H295R cells and 4 µg/mL of polybrene for HEK293T cells across 6 replicates for each pool (EMD Millipore Corporation #TR-1003-G). We changed the media on the cells 24 hours after transduction.

#### **2.4.10 RNA isolation and qRT-PCR to measure gene expression levels**

To measure any changes in gene expression levels due to the CRISPRa perturbation, we used qRT-PCR. First, we harvested RNA from each replicate 4 days post transduction with the gRNA lentivirus pool using RNeasy Mini Kit (Qiagen #4004) following manufacturer's protocol including the DNase treatment. We quantified the RNA using Qubit (Invitrogen) and used 500 ng of RNA for each sample for subsequent cDNA synthesis. For the cDNA synthesis, we used Superscript III (800 U, Life Technologies) with Oligo dT primers following manufacturer's

protocol (Thermo Fisher #18418012). Following cDNA synthesis, we performed qRT-PCR using that cDNA, TaqMan™ Fast Advanced Master Mix for qPCR (Thermo Fisher #4444556), and TaqMan™ Gene Expression Assays (DENND1A, CRB2, LHX2 and STRBP and GAPDH). The qPCR analysis was performed using the  $2^{-\Delta\Delta CT}$  method in R, using GAPDH as the internal control. All the fold change is reported as  $\log(2^{-\Delta\Delta CT})$  compared to the negative (non-targeting gRNA) control. Each sample was measured in triplicate for the qRT-PCR.

#### **2.4.11 ELISA for measuring testosterone production**

To measure changes in testosterone production, we collected the supernatant from the gRNA pool transduced H295R cells two- and four- days post transduction. We measured the amount of testosterone produced using ELISA (Cayman Chemicals #582701) according to the manufacturer's protocols. All samples were measured in triplicate. The absorbance of the compound was measured at 405-420 nm using the GloMax Discover System (Promega). Fold-change reported is based on the negative (non-targeting gRNA) control.

### **3. Regulatory variants in the DENND1A locus**

#### ***3.1 Introduction***

As described in chapter 1, one of the central challenges of complex trait genetics is identifying the causal variants within GWAS susceptibility loci and determining their functional consequences. The identification of eQTLs, expression quantitative trait loci revealed the pervasive nature of genetic variant effects on gene regulation (Stranger & Raj, 2013), and that genetic variants within accessible chromatin regions explain a significant proportion of the eQTL signals (Degner et al., 2012). More recently, several studies have shown that an enrichment of genetic variants in disease-associated loci overlap gene regulatory elements from complementary datasets (U. Consortium et al., 2019; Schmidt et al., 2015; Trynka et al., 2013). Following the evidence supporting the contributions of non-coding genetic variants towards phenotypes, there have been several advancements in strategies for identifying disease-associated non-coding genetic variants and yielded successful examples to link variants to disease (Guo et al., 2015; Maurano et al., 2012). However, there remain several challenges towards making those approaches systematic, thus enabling routine identification of specific causal variants, genes, and tissues.

Recent advances in modifying high-throughput reporter assays like STARR-seq have allowed for the direct identification of allele-specific regulatory variants (Vockley et al., 2015). To study the role of regulatory elements in diseases, several high throughput reporter assays have been performed that have used some fraction of genomic DNA from a disease cohort. While synthesizing oligonucleotides for STARR-seq is one approach, DNA synthesis is limited in the length and number of oligonucleotides that can be generated (Mangan et al., 2022). Previous studies have assayed putative regulatory elements directly by amplifying the desired genomic

regions from donor genomes to identify effects of regulatory variants specific to that study population (Vockley et al., 2015; P. Zhang et al., 2018). One advantage of this approach is that haplotypes are maintained when the target regions are amplified by custom amplicon sequencing, thereby capturing the effects of common and rare variants in each genome. Using amplicon sequencing allowed for an effective approach to identify causal variants since the candidate regulatory elements are assayed by STARR-seq independently of each other. A main disadvantage of this approach is that while amplicon generation is largely successful for smaller test regions, it is technically not practical for scaling up to larger regions like DENND1A, which is a large gene spanning ~550 kb on chromosome 9. An alternative to amplicon-based approaches is to directly capture specific regions of the genome using oligonucleotide probes (P. B. Chen et al., 2022; C.-C. F. Huang et al., 2021). Here, we used that capture-based STARR-seq approach to selectively enrich and measure regulatory activity across the DENND1A locus from 5 individual genomes. We then implemented a whole genome STARR-seq approach and identified over 140,000 regulatory elements used in an adrenal cell model specifically using donor genomic DNA from a PCOS case-control cohort. The whole genome STARR-seq approach allows for testing for regulatory variants across the whole genome and as a case study, we identified X regulatory variants in the DENND1A locus.

## ***3.2 Results***

### **3.2.1 DENND1A is implicated in PCOS pathogenesis**

The DENND1A locus harbours reproducible genetic associations with PCOS. DENND1A is a guanine nucleotide exchange factor involved in clathrin-mediated endocytosis. The expression of DENND1A has been implicated in androgen biosynthesis, including in H295R cell model (Teves et al., 2020). Furthermore, DENND1A knockout in mice is embryonic lethal,

and affects follicle development in the ovary (Jingjing Shi et al., 2019). Furthermore, a recent study identified several non-coding rare variants in the *DENND1A* locus that are associated to PCOS, and increased ratios of luteinizing and follicle stimulating hormone levels in those PCOS cases (Dapas et al., 2019). Therefore, to test the role of genetic variants in contributing to PCOS, we focused on the *DENND1A* locus in H295R cells.

### **3.2.1.1 Overview of the *DENND1A* genomic locus**

We successfully identified hundreds of regulatory elements in two cells lines as described in chapter 2. However, there are currently no reports of experimentally measuring the effects of genetic variants on *DENND1A* locus. From our association and colocalization analysis in chapter 2, one of the top hits implicated *DENND1A* as a likely causal gene for PCOS. As proof of concept that the regulatory elements we identify are relevant to PCOS, we focused on mechanisms contributing to altered expression of genes in the *DENND1A* locus. We identified 38 candidate regulatory elements between the second and sixth introns of *DENND1A* spanning ~180 kb of the genome (Figure 9a). Several of the identified regulatory elements overlap regions called as candidate cis regulatory elements (cCRE) through ENCODE or in regions with increased chromatin accessibility in H295R and COV434. Indeed, for most of these candidate regulatory elements, there are common variants in linkage disequilibrium with the lead GWAS SNPs.

### **3.2.2 Capture-STARR-seq of the *DENND1A* locus<sup>2</sup>**

In chapter 2, I described our results in identifying regulatory elements using a combination of bacterial artificial chromosomes and fosmids and scaled to assay cross 14 PCOS

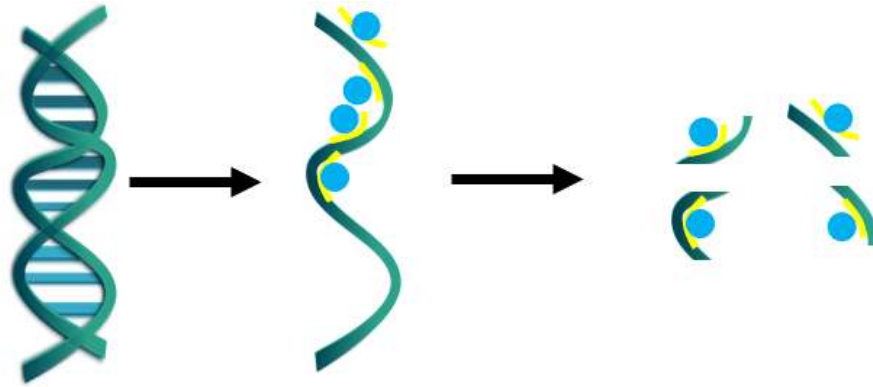
---

<sup>2</sup> This subsection has been adapted from a primary-author manuscript that is in review (Sankaranarayanan et al., manuscript under review)

GWAS risk loci. However, this approach doesn't capture genetic variants present in the human population to then identify what are the genetic variants that display allele-specific regulatory activity. Therefore, to be able to assay donor genotypes using STARR-seq, we first enriched for the DENND1A locus from donor genomes using oligonucleotide probes. Briefly, the targeted regions are captured using RNA probes that are biotinylated which allows for using streptavidin bead-based separation. We enriched for the DENND1A locus from 3 µg of five donor genomes, and used the enriched genomic DNA as the input material for the capture STARR-seq experiments described.

To capture the genetic variation, present in the DENND1A locus, we enriched for the genomic regions spanning the DENND1A locus from three individuals of European ancestry and two of Han Chinese ancestry from the 1000 genomes project (Appendix B, Table S1). We selected those samples as healthy donor genomes, based on the availability of DNA, sex of the donor, and included the two ancestries to reflect the ancestry of the GWA studies that identified variants in the DENND1A locus to be associated with PCOS.

Specifically, to enrich for the DENND1A locus, we used a probe-based capture method (Agilent Sure Select, Figure 8, Figure 9b), and inserted those fragments into the STARR-seq plasmid. We estimated 75% of the captured fragments were indeed captured from the DENND1A locus. There have been several studies that used targeted enrichment of genomic DNA for the subsequent assay of choice (Peterson et al., 2019; Zeineldin et al., 2023)



**Figure 8: Principle of probe-based capture of genomic regions**

To measure the effects of genetic variation across *DENND1A* on gene expression, we used the *DENND1A*-enriched assay library in H295R cells using STARR-seq. In total, we assayed ~700,000 unique DNA fragments (Appendix B, Figure S2). The average length of the targeted fragments is 200 bp (Appendix B, Figure S3). The assay library covered the *DENND1A* gene locus at a median coverage of 140x and the libraries were sequenced to a depth of ~350 million reads using 150 bp PE sequencing. The assay libraries were highly concordant, while the measure of log fold change between the reporter and assay libraries was moderately concordant among replicates (Pearson's  $r > 0.72$ , Appendix B, Figure S4)

### **3.2.2.1 Probe based enrichment of *DENND1A* locus captured ~600 heterozygous variants**

In the locus that we assayed across the five genomes, there were a total of 1005 SNPs that were present as heterozygous SNPs in the pool. We filtered the list of variants to only include those SNPs that did not have indels and any 'unknowns' in the file, the final list of SNPs was 952. Of those 952 SNPs, 630 SNPs had read count  $> 5$  in the sequencing readout of the assay and reporter libraries after using WASP for allele-aware alignment (Geijn et al., 2015). We then tested these 630 SNPs for regulatory effects. The capture STARR-seq libraries covered  $> 100X$  of

the target genomic region making it suitable to identify allele-specific regulatory variants using a bayesian statistical approach

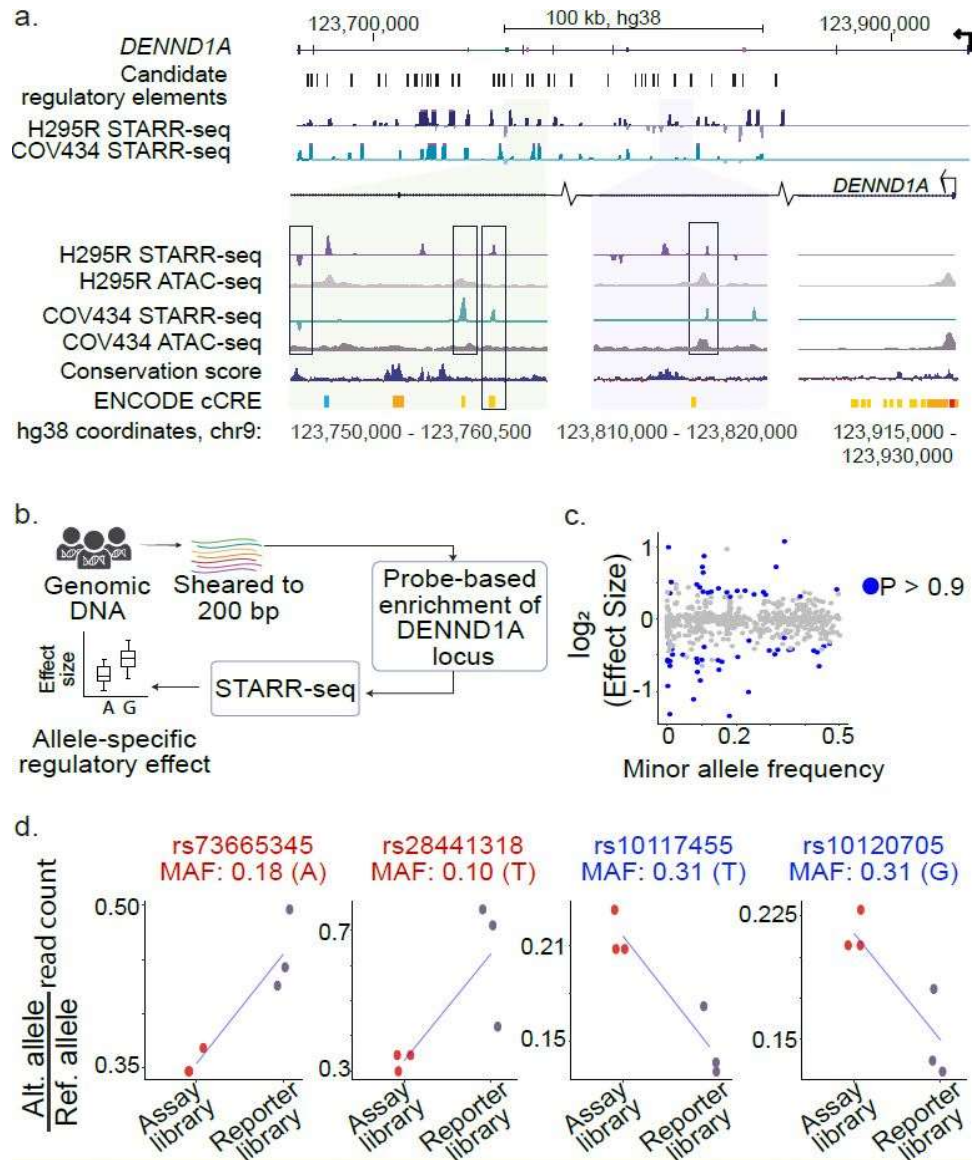
### 3.2.3 Allele-specific regulatory activity across the *DENND1A* locus <sup>3</sup>

One of the major goals of functionally evaluating regulatory variants is to determine genetic effects on regulatory element activity that may explain genetic associations with phenotypes. To estimate the allele specific regulatory effects, we used a Bayesian approach, BIRD, that identifies differences in the relative abundance of alleles in the assay library and in the expressed reporter library. Of the 623 variants we assayed in the targeted locus, 62 had allele specific regulatory activity with a posterior probability,  $P_{\text{reg}} > 0.90$ . On average, the identified variants altered regulatory activity by 40% (Appendix B, Figure S5); and minor alleles overall had less regulatory activity (chi-squared = 6.9, p-value = 0.009). We observed a modest correlation between the absolute effect size and the minor allele-frequency of the selected variants as determined by the 1000 Genomes project ( $\rho = -0.36$ ,  $p = 0.005$ , Fig 9c).

Of the 62 identified regulatory variants we identified, 24 were eQTLs for *DENND1A* ( $n = 11$ ) or flanking genes *CRB2*, *RABGAP1* or *STRBP* ( $n = 14$ ). Of those variants, 12 also overlapped open chromatin sites or candidate cis-regulatory elements identified by ENCODE (Table 4).

---

<sup>3</sup> This subsection has been adapted from a primary-author manuscript that is in review (Sankaranarayanan et al., manuscript under review)



**Figure 9: Fine-mapping variants identified four regulatory variants that DENND1A eQTLs.**  
**a.** Candidate regulatory elements in *DENND1A* locus identified in H295R (purple track) and COV434 (teal track) cell lines. Each STARR-seq track is reported as assay (input) subtracted reporter (output) libraries. **b.** Overview of enriched *DENND1A*-STARR-seq method. Genomes from five individuals from the 1000 Genomes Project were sheared to 200 bp. We enriched the target *DENND1A* locus using RNA-probes. The custom probes were designed to span the *DENND1A* locus at 2x tiling density. These enriched fragments were then subject to the STARR-seq protocol. Allele-specific regulatory effect was estimated using the BIRD model. **c.** Distribution of allele-specific effect sizes as estimated by BIRD against minor allele frequencies. The estimated significant SNPs with allele-specific regulatory activity ( $P > 0.9$ ) are in blue. **d.** SNPs in red - alternate allele has increased regulatory activity while SNPs in blue - reference allele has increased regulatory activity.

Furthermore, the lead variant from colocalization analyses, rs10117940 (Table 3) was also identified in allele-specific analysis with an effect size of 1.299 ( $p = 0.731$ ). The variant rs10117940 is in LD with two STARR-seq regulatory variants (rs28441318 and rs73665345) and a PCOS-associated rare variant (rs78012023) ( $0.32 < r^2 < 0.65$ ;  $0.5 D' > 0.9$ ). Taken together, these results indicate several loci within the *DENND1A* gene may contribute to PCOS phenotypes by altering *DENND1A* gene expression.

**Table 4: Candidate allele-specific regulatory variants identified.**  
**The effect size is measured using the BIRD model. Overlap with genomic data indicates whether that variant overlaps open chromatin as measured by ATAC-seq or DNase-seq, in proximal enhancer like elements (pELS) or distal enhancer like elements (dELS) from ENCODE. The eQTL data was obtained from GTEx data.**

<b>rsID</b>	<b>Effect size</b>	<b>Minor Allele Frequency</b>	<b>Posterior probability</b>	<b>Overlap with orthogonal genomic data</b>	<b>eQTL gene</b>
rs4466467	0.56	0.096	0.97	DNase-seq	<i>CRB2</i>
rs10985986	1.32	0.226	0.96	pELS	<i>CRB2</i>
rs10117455	0.71	0.312	0.99	ATAC-seq	<i>DENND1A</i>
rs10120705	0.73	0.311	0.98	ATAC-seq	<i>DENND1A</i>
rs28441318	1.66	0.102	0.96	ATAC-seq	<i>DENND1A</i>
rs73665345	1.24	0.183	0.95	dELS	<i>DENND1A</i>
rs10760295	0.77	0.454	0.91	dELS	<i>DENND1A</i>
rs10760271	0.81	0.237	0.91	dELS	<i>RABGAP1</i>
rs10114139	0.73	0.435	0.91	dELS	<i>STRBP</i>
rs62579936	1.30	0.104	0.91	dELS	<i>STRBP, CRB2</i>
rs10156609	1.57	0.104	0.99	pELS	<i>STRBP, CRB2</i>
rs2491351	0.79	0.471	0.97	dELS	<i>STRBP, CRB2, DENND1A</i>

### **3.2.4 Whole genome STARR-seq using donor genomes<sup>4</sup>**

Using the whole genome for STARR-seq assays has two main advantages over the capture STARR-seq assays described in the previous section. First, we observed a large variation in the coverage across the DENND1A locus, reflecting the biases of the probe-based capture method. Current methods to analyse STARR-seq data like CRADLE are currently underpowered to account for significant biases arising due to the probe-capture method. Second, using the whole genome method will allow us to test for regulatory effects of variants across the whole genome, not just the captured locus.

#### **3.2.4.1 Optimizing whole genome STARR-seq for low quantities of donor genomic DNA**

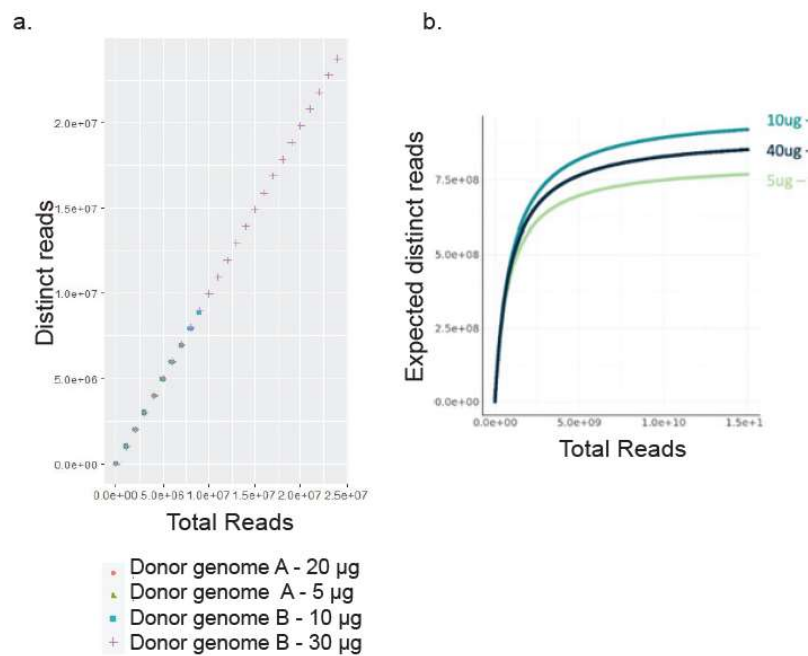
Previous human whole genome STARR-seq studies have used a minimum of 40 µg of genomic DNA to build the assay library (Johnson et al., 2018). However, one of the challenges in being able to use donor genomic DNA from a population cohort is the quantity. There is usually a limited amount of DNA extracted from donor blood sample, and is not a renewable resource. In the case of the donor genomes from the PCOS case control cohort, we identified that most of the total genomic DNA content was between 2 -5 µg. Of the 17 samples that we had initially received, 4 samples were removed because the DNA was degraded, or the amount of DNA was too low (< 0.5 µg). Therefore, we first optimized the STARR-seq protocol in three key steps to maximize the donor genome coverage in the assay and reporter libraries. First, for each genomic DNA, for cloning the fragmented DNA into the STARR-seq vector, we split each sample into 10 parallel reactions. Second, we cloned each donor genome separately into the vector, and amplified each STARR-seq plasmid from a donor genome separately. Third, for each PCR and cDNA

---

<sup>4</sup> Shauna Morrow assisted with optimization and library preparation of the low input whole genome STARR-seq library preparation.

conversion reaction, we split one reaction into 8 reactions, to maximize capturing the complexity of the fragments. For each library, we measured the complexity using pre-seq estimates to quantify the number of unique fragments in the library as a metric for the optimization approach.

By optimizing the whole genome STARR-seq approach, we built STARR-seq assay and reporter libraries from thirteen donor genomes. Here, we have demonstrated the first whole genome STARR-seq assay using a population based cohort of PCOS cases and controls.

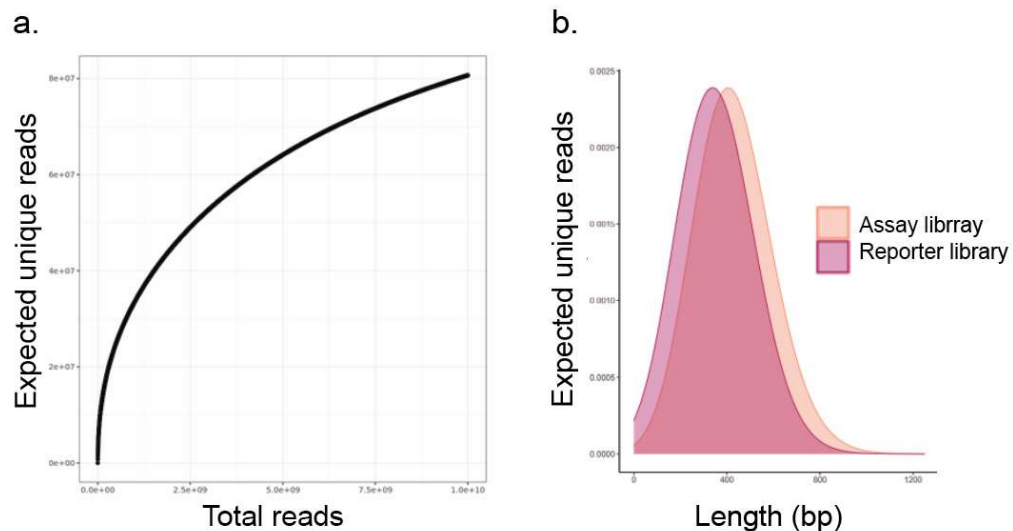


**Figure 10: Library complexity to estimate whole genome STARR-seq optimization. a) Pre-seq complexity estimates for different input samples. b) Extrapolated pre-seq complexity estimates for different input quantities for whole genome STARR-seq**

### 3.2.4.2 Genome-wide measurement of regulatory elements across a population of thirteen individuals from a PCOS case-control cohort

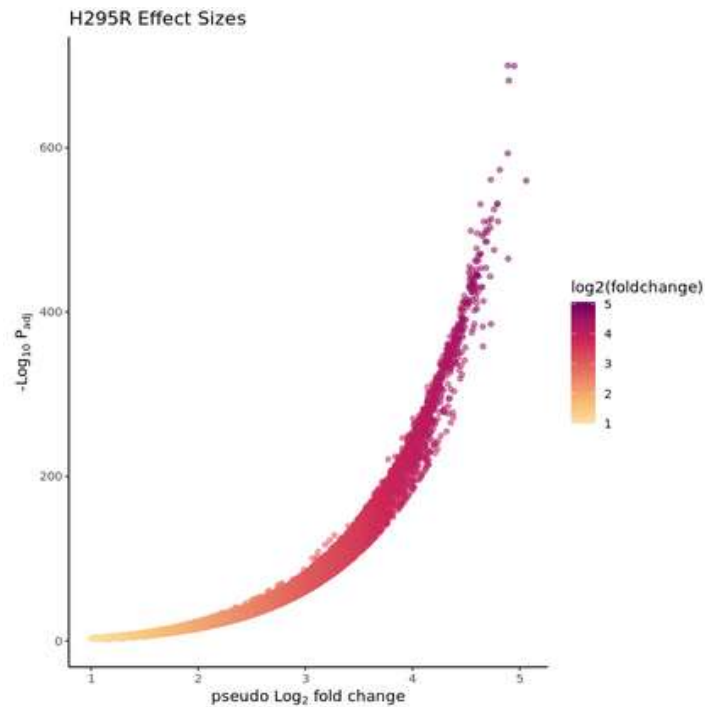
To quantify regulatory activity of regulatory elements in the adrenal cell model across the whole genome, we used STARR-seq assay using whole donor genomes for building the assay

library. We constructed a STARR-seq assay library from 13 individual donor genomes, including 6 PCOS cases and 7 controls. The donor genomes were obtained from a population who had their whole genomes sequenced. Within this population, several rare noncoding variants have been identified, including several in the DENND1A locus that are associated with PCOS (Dapas et al., 2019). Each sample was sequenced on Illumina NovaSeq with S4 across all lanes yielding ~1.5-2 billion reads per sample and the assays libraries across the samples are all highly correlated (Pearson's  $r > 0.8$  Appendix B, S6). To generate the reporter library, we pooled the genomes of four or five individuals and sequenced the resulting pool to have ~80 million unique fragments and  $>10X$  coverage across the whole genome per each sample (Figure 11a). Across the assay and reporter libraries, the median fragment size was 417 bp and 367 bp respectively (Figure 11b), which is comparable to the average size of open chromatin sites identified with ATAC-seq or DNase-seq. The reporter libraries correlated with Pearson's  $r > 0.6$  (Appendix B, S7).



**Figure 11: Complexity and length of whole genome STARR-seq libraries.**  
a) Expected unique reads per library estimates  $> 10X$  coverage across the genome. b) average length of fragments in the assay and reporter libraries.

To measure the effect of regulatory elements across the whole genome, we compared the sequencing reads in the reporter libraries compared to the assay libraries. Specifically, we identified ~146,000 regions with increased regulatory activity using MACS2 peak caller (FDR < 0.001, Figure 12). MACS2 is widely used for identifying ChIP-seq peaks and is a widely used peak calling algorithm (Yong Zhang et al., 2008).



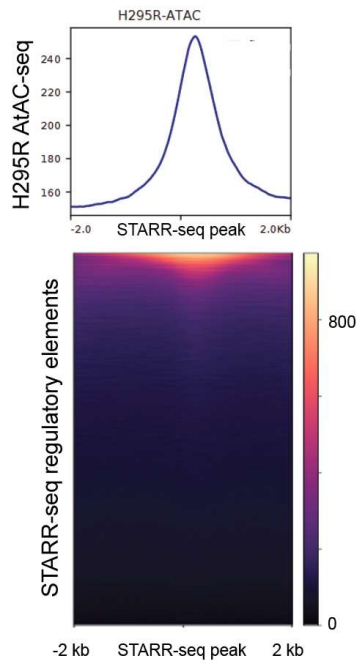
**Figure 12: pseudo log<sub>2</sub>(Fold Change) of regulatory elements called by MACS**

### 3.2.4.3 Regulatory element activity is enriched across regions of chromatin accessibility

STARR-seq being an episomal assay identifies candidate regulatory elements outside of the native genomic context. Furthermore, while MACS2 is useful for analysis of whole genome STARR-seq, one main limitation is that MACS2 does not consider replicate variance, leading to false positives. Therefore, to increase confidence in the regulatory elements called as having a functional effect in the cells, we evaluated whether STARR-seq regulatory elements

corresponded to chromatin accessibility in the same cell line as well as other cell lines from ENCODE.

About 9500 (13%) of the ~73,000 accessible chromatin sites in H295R were found to have regulatory activity measured by STARR-seq. We observed an enrichment of STARR-seq measured regulatory activity across accessible chromatin sites, and the overlap between chromatin accessibility and STARR-seq activity is ~2-fold more than what would be expected if STARR-seq identified regions were randomly distributed across the genomic regions (Chi-square test,  $p < 0.001$ , Figure 13). Regulatory activity in H295R also corresponds to chromatin accessibility in other tissues. About 64% of the regulatory elements we identified via STARR-seq overlap accessible chromatin sites identified in diverse tissues as part of the ENCODE project with an 1.7-fold overlap enrichment over what would be expected if regulatory activity was randomly distributed across the assayed regions in H295R (Chi-square test,  $p < 0.001$ ).



**Figure 13: Whole genome STARR-seq regulatory elements correspond to accessible chromatin regions**

We also observed a distribution of the regulatory elements called identified and ENCODE candidate cis-regulatory elements (cCRE) classification (Abascal et al., 2020). Specifically, ~28% of the regulatory elements we identified overlap proximal or distal enhancers defined by ENCODE (n = 41513); and ~1400 regions overlap promoter like sequences and 0.9% overlap CTCF binding sites (n = 1460) (Appendix B, Figure S8). Taken together, these results demonstrate that the candidate regulatory elements identified by whole genome STARR-seq and MAC2 peak calls are enriched for regions with evidence of functionality from orthogonal genomic datasets.

#### **3.2.4.4 Regulatory variants identified in the DENND1A locus**

One key advantage of using donor genomes in the STARR-seq assay is the ability to use the natural genetic variation present in those genomes to estimate the effect sizes of allele-specific regulatory activity of those variants. Notably from the population cohort in the PCOS study ~50% of the families with PCOS cases sequences had rare variants in DENND1A collectively (Dapas et al., 2019). However, it was determined that the instance of each individual variant was typically identified in only one – two families (Dapas et al., 2019). Therefore, the pooling strategy we employed would ensure that even if only one copy of the variant was present in a pool of genomes, that variant would still be at a frequency of 1:8 or 1:10 in the reporter libraries. There are 2209 heterogenetic variants in our cohort and were largely observed at their expected frequencies in the STARR-seq assay library, indicating little bias or bottlenecking during library production (Appendix B, Figure S9). We measured allele-specific regulatory activity for ~2200 variants across the STARR-seq reporter library.

To identify variants with significant allele specific activity, we used a bayesian estimation method, BIRD (Majoros et al., 2019). We identified 453 variants (20%) for which the posterior probability of there being a difference between the allele ratio between the reporter and assay libraries was greater than 0.95 and 785 variants with a posterior probability of 0.9 (Appendix B, Figure S10). We then examined whether these identified regulatory variants are involved within regulatory regions. Specifically, 6 overlapped within STARR-seq identified regulatory elements (Appendix B, Figure S11).

#### **3.2.4.5 Comparing the regulatory effects determined by capture STARR-seq and whole genome STARR-seq**

We identified 445 regulatory elements that overlapped between the targeted STARR-seq described in chapter 2 and the whole genome STARR-seq in this chapter. The overlap is enriched by 20X when considering overlap between the BAC STARR-seq and random control for the whole genome STARR-seq (Chi-square test,  $p < 0.001$ ).

In performing whole genome STARR-seq, one of the main advantages is in reducing bias that comes from targeting specific loci. For example, with the targeted BAC STARR-seq, one of the challenges in analysis is due to the nature of BAC fragmentation, recombination and uneven coverage of the BAC regions. This necessitates several normalization processes during the analysis steps and because of the nature of selecting BACS, it would require extensive custom normalization that is catered towards each experimental setup. In performing the capture STARR-seq experiments, while we were able to identify candidate regulatory variants within a locus, the biases of the capture process inherently affected the ability to call peaks to identify regulatory elements, as well as a large variation in the coverage across the genetic variants.

### ***3.3 Discussion***

In this work, we optimized a high-throughput empirical approach to measure the regulatory effects of genetic variants directly from the genomic DNA of individuals from a population-based study cohort. We map candidate regulatory elements in an adrenal cell model, expanding the work from chapter 2, which targeted only specific disease associated loci. We estimated about 140,000 regulatory elements across the adrenal cell line. We further identified ~400 genetic variants within functional regulatory elements that acted as allele-specific regulatory variants.

These results add to previous research on the contributions of non-coding genetic variants to human diseases. This approach has several advantages. The main advantage in using the whole genome based approach from a population cohort, is that in addition to offsetting the biases introduced by targeted methods, we can investigate variants and haplotypes that are present specifically across the population cohort that may not be present in existing SNP databases. Furthermore, using this approach, we are able to test for regulatory effects of rare variants present in the study population. Rare variants although form one answer to the missing heritability model, burden testing for rare variants may or may not have supported statistical association with the disease due to the inherent low frequency of observing rare variants. In addition to empirically measuring the regulatory effect of these rare variants, this method allows us to maintain the genetic linkage across each fragment that is being tested. In maintaining the linkage, one advantage is that we are measuring regulatory activity that is influenced by the native haplotype present within those fragments from donor genomes. Furthermore, while an early limitation was the quantities of donor genomes, once the STARR-seq plasmid library has been made, it can be

further amplified in the future to maintain a renewable stock of the plasmid library. This would then be useful for testing for regulatory activity and regulatory variant effects in several other cell lines that are implicated in the disease, for example - in metabolic tissues such as adipocytes or liver cells. This approach currently has two key disadvantages. First, the ability to call variants using BIRD needs to be tested on identified rare variants, and whether the assumptions of the model still hold for several rare variants identified in a case control cohort. Second, there is limitation in performing such high-throughput reporter assays on a large scale is a factor of the computational time and scalability for large datasets, particularly when running processes that are memory intensive such as de-duplication. If the data are not deduplicated using the information from the fragments and UMIs, it is likely that the analysis of reads mapping to variants might be biased.

The candidate regulatory variants and regulatory elements that we identified in this study will serve as a framework to further investigate the role of regulatory variants in causing PCOS phenotypes. Taken together, we have demonstrated two different approaches to quantify the regulatory effect of genetic variants in the DENND1A locus that is implicated in PCOS pathogenesis.

### ***3.4 Methods***

#### **3.4.1 Cell Culture**

We obtained NCI-H295R cells from ATCC. The cells were cultured in DMEM/F-12 medium (Gibco #21041025) supplemented with 2.5% Nu-Serum (Corning #355100) and 1% ITS+Premix (Corning #354352) and grown as a monolayer at 37 °C, 5% CO<sub>2</sub>. The H295R-dCas9

cells were grown under the above described conditions. All experiments were performed between passages 5 and 15.

### **3.4.2 DENND1A enrichment**

We focused on variants present in the *DENND1A* locus, a region that spans the entire *DENND1A* gene and 100 kb upstream and downstream of the gene. For target enrichment of the *DENND1A* locus, we used targeting oligonucleotide probes. We first sheared each genome separately to ~200 bp using Covaris (S220). We then used Agilent SureSelect Custom DNA Target Enrichment Probes to enrich the region around *DENND1A* (hg 38: chr9:123279654-124030107). To design the custom probes, we optimized for tiling the target region at 2x the density using probes using the Balanced parameter in the Sure Design. We then reran the probe design on the regions missed by Sure Design in the first round to yield a final set of probes tiles across the target region. We followed the Agilent SelectXT2 custom (Cat# 5190-4846) protocol to enrich the target regions in each genome, however we modified the protocol at the adaptor ligation steps. We used a custom adaptor (SS\_Adaptor) and amplified the resulting oligo fragments using TS2SS-F and TS2SS-R primers (Appendix A, Table S3).

### **3.4.3 Capture STARR-seq**

To create the *DENND1A* STARR-seq assay libraries from the five genomes, we cloned the sheared and *DENND1A*-locus enriched DNA fragments. We cloned the amplified and enriched fragments into the linearised STARR-seq vector using NEBuilder HiFi DNA Assembly (#E2621). We ethanol precipitated the products. To do so, we added 0.1X volume 3 M NaOAc and 2.5X volume cold 100% ethanol, and stored the mixture at -20 °C overnight. We then

pelleted the DNA via centrifugation at 16,000 RCF for 30 min at 4 °C. We washed the pellets with 5 ml cold 70% ethanol, and resuspended in water.

#### **3.4.3.1 DENND1A locus STARR-seq reporter plasmid construction**

To create the DENND1A STARR-seq assay libraries from the five genomes, we cloned the sheared and DENND1A-locus enriched DNA fragments. We cloned the amplified and enriched fragments into the linearised STARR-seq vector using NEBuilder HiFi DNA Assembly (#E2621). We ethanol precipitated the products. To do so, we added 0.1X volume 3 M NaOAc and 2.5X volume cold 100% ethanol, and stored the mixture at –20 °C overnight. We then pelleted the DNA via centrifugation at 16,000 RCF for 30 min at 4 °C. We washed the pellets with 5 mL cold 70% ethanol, and resuspended in water.

We then pooled the plasmids from each genome in equimolar concentrations. We amplified the pooled plasmids by transfecting the plasmids into E. coli 10G Electrocompetent Cells following manufacturer protocol for optimal settings (1.0 mm cuvette, 10 µF, 600 Ohms, 1800 Volts). We subsequently isolated the plasmids using Qiagen Plasmid Kit, GigaPrep (Qiagen #12191), and quantified it using Qubit and validated the length of the pooled library on a 1% agarose gel. This purified, pooled plasmid is our DENND1A-locus STARR-seq library that was used for DENND1A-locus STARR-seq experiments.

#### **3.4.3.2 DENND1A locus STARR-seq assay library sequencing**

To estimate the abundance of reads mapping to the variant loci selected in each assay library, we used Illumina high-throughput sequencing NextSeq 2000 with 50 bp paired end sequencing protocol. We sequenced 3 replicates of the DENND1A-locus STARR-seq assay

library using the amplified the STARR-seq assay fragments from the pooled library using 208-F Index7 primers (Appendix A, Table S3).

#### **3.4.3.3 DENND1A locus enriched STARR-seq assay**

To test for effects of variants in the targeted DENND1A locus, we transfected the DENND1A locus STARR-seq assay library into H295R cells, and isolated and sequenced the resulting RNA similar to the methods described previously. All experiments were performed in triplicate for each cell line. For each replicate for each of the cell lines, we used 70 million H295R cells transfected with 140  $\mu$ g DENND1A locus STARR-seq assay library using the Lonza Nucleofector (setting SF-DN-100). We isolated the RNA from the cells 6 hours post transfection.

#### **3.4.3.4 DENND1A locus STARR-seq reporter library construction**

To isolate the DENND1A locus reporter RNA, we first isolated total RNA followed by enriching for cDNA produced from the DENND1A-locus STARR-seq plasmid library. We used the same protocol as described for the PCOS GWAS reporter library construction. We pooled the DENND1A locus reporter libraries from each replicate at equimolar 2nM concentrations. We then sequenced the DENND1A locus reporter libraries on Illumina NextSeq 2000 using 75bp PE sequencing.

#### **3.4.4 Capture STARR-seq analysis**

To identify variants that have allele specific regulatory activity, we compared the ratio of reads mapping to the alternate allele versus reference allele in each assay library and reporter library. If the ratio of reads mapping to alternate allele versus reference allele was higher in the

reporter library compared to the assay library, that variant was called as having increased regulatory activity of the alternate allele.

To do so, we first obtained a list of the variants present in the *DENNDIA* locus in the pool. We obtained the VCF for these samples from the 1000 Genomes Project (PMID). We filtered the variants in the targeted *DENNDIA* locus, to only include those SNPs present as heterogeneous within the pool of five genomes we used. The final list of ~600 variants was then used for the regulatory variant analysis.

To compare reads mapping to each allele in both the reporter and assay libraries, we first aligned *DENNDIA* locus enriched STARR-seq libraries (assay library and reporter library) individually aligned to the human genome (hg38) using WASP (Geijn et al., 2015) and bowtie2 (Langmead & Salzberg, 2012). Reads with a quality score of  $Q \geq 30$ , and outside the centromeres and blacklisted regions were used for downstream analysis. We used picardtools to mark and call duplicates (Institute, 2019). RPKM normalized STARR-seq read density was computed at single base pair resolution using deepTools utility bamCoverage (Ramírez et al., 2016). We then assigned reads mapping to each variant for each sequenced sample using samtools mpileup (H. Li, 2011).

To estimate the regulatory effect of variants, we used BIRD (Majoros et al., 2019). BIRD is a bayesian statistical framework for analysis of regulatory variants and uses bayesian priors to identify allele-specific regulatory effects, and identifies variants that have a high probability of being a regulatory variant with an effect size, theta. We used the standard options for BIRD and set the regulatory effect threshold as 1.2.

### **3.4.5 whole genome STARR-seq assay**

#### **3.4.5.1 STARR-seq reporter plasmid construction and assay library sequencing**

To create the whole genome STARR-seq assay libraries from the thirteen donor genomes, we followed similar methods as described in chapter 1. The donor genomic DNA was first quantified using Agilent TapeStation. We sheared the donor genomes using CovarisS220 with the setting for 450-500 bp at Duke Microbiome Core Facility. We used between 2-5 $\mu$ g of the sheared genomic DNA for the cloning and STARR-seq plasmid assembly. We cloned the sheared DNA fragments into the linearised STARR-seq vector using NEBuilder HiFi DNA Assembly (#E2621). The Gibson Assembly reactions for each sample were split as 4 pmol of the insert in 10 reactions. We ethanol precipitated the products and to do so, we added 0.1X volume 3 M NaOAc and 2.5X volume cold 100% ethanol, and stored the mixture at  $-20^{\circ}\text{C}$  overnight. We then pelleted the DNA via centrifugation at 16,000 RCF for 30 min at  $4^{\circ}\text{C}$ . We washed the pellets with 5 ml cold 70% ethanol, and resuspended in water. Each plasmid library from one donor genome was amplified by transfecting the plasmids into E. coli 10G Electrocompetent Cells following manufacturer protocol for optimal settings (1.0 mm cuvette, 10  $\mu$ F, 600 Ohms, 1800 Volts). We subsequently isolated the plasmids using Qiagen Plasmid Kit, GigaPrep (Qiagen #12191). All PCR amplification steps were split into 4 wells per sample. To estimate the reads across each genome the resulting plasmids from the thirteen donor-genomes were pooled in equimolar ratio, amplified using the 208-F and index primers, and we used Illumina NovaSeq (P4, all lanes) with 150bp paired end sequencing. The sequencing was performed at Duke Sequencing and Genomic Technologies core.

### 3.4.5.2 STARR-seq reporter assay library preparation

To build the whole genome STARR-seq reporter library, we transfected the pooled plasmids into H295R cells. Specifically, we made 3 equimolar pools comprising of 4, 4 and 5 donor genome plasmid libraries for the transfection. Each library was transfected into a total of ~300 million cells, using the Lonza Nucleofector setting as described earlier. We isolated the RNA from the cells 6 hours post transfection, first by pelleting the cells dissociated with Trypsin-EDTA 0.25% (Life Technologies). We lysed the cell pellets using RLT buffer (Qiagen) with 2-mercaptoethanol (Sigma). We passed the lysates through a 18-gauge needle ten times and stored at  $-80^{\circ}\text{C}$  before RNA extraction. We isolated total RNA using the Qiagen RNeasy Midi kit as previously described in chapter 2 and quantified and estimated the quality via RIN values (RIN > 8.5 was used) using Agilent Tape Station. We used ~120  $\mu\text{g}$  of total RNA for the cDNA synthesis and STARR-seq reporter cDNA preparation. We synthesized reporter cDNA by reverse transcription using Superscript III (800 U, Life Technologies) following manufacturer's protocol and a STARR-seq specific primer (SSRT-UMI, Appendix A, Table S3) and each sample was split into 4 reactions. Following cDNA synthesis, we prepared the reporter library as described in chapter 2, such that each replicate was amplified by splitting into 16 wells, for the last step of amplifying using index-PCR primer and indexed PostSS-Index-5 primers. For all sequencing libraries, we determined the total number of cycles for amplification using a small portion of that sample in a qPCR protocol and estimating cycle number using 1/4th the maximum plateau observed in the qPCR. Final whole genome STARR-seq reporter libraries were sequenced using Illumina NovaSeq (P4, all lanes) with 150bp paired end sequencing. The sequencing was performed at Duke Sequencing and Genomic Technologies core.

### 3.4.5.3 STARR-seq reporter assay analysis and regulatory variant estimation

STARR-seq input library and output libraries were individually aligned to the human genome assembly hg38 with Bowtie2, using the following parameters: bowtie2 -X 2000 sensitive. Only properly paired alignments with a MAPQ (mapping quality) score  $\geq 30$  outside hg38 centromeres, gap, and blacklist regions were retained in downstream analyses. Regions were called individually for each sample using merged STARR-seq input alignments as controls with the MACS2 package using the following parameters: -f BAMPE -g hs -ratio -q 0.10. We tested for differential STARR-seq activity across the union region set by fitting negative binomial models using DESeq 2 (Anders & Huber, 2010). We obtained a list of the variants present in the DENND1A locus across the 13 donor genomes. We filtered the variants in the targeted DENND1A locus, to only include those SNPs present as heterogeneous within the pool of thirteen genomes. The final list of ~2000 variants was then used for the regulatory variant analysis. We then assigned reads mapping to each variant from each sequenced sample using samtools mpileup (Danecek et al., 2021). To estimate the regulatory effect of variants, we used BIRD (Majoros et al., 2019). BIRD is a bayesian statistical framework for analysis of regulatory variants and uses bayesian priors to identify allele-specific regulatory effects, and identifies variants that have a high probability of being a regulatory variant with an effect size, theta. We used the standard options for BIRD and set the regulatory effect threshold as 1.25.

## **4. Developing a probe-based method to quantify gene expression in intact single cells**

### ***4.1 Introduction***

Gene regulation plays a crucial role in maintaining various biological functions and response to different stimuli. Though there have been several large efforts in steps to better understand the genomics of gene regulation, the dynamics of regulation is less well understood (Consortium, 2012, Roadmap Epigenomics et al., 2015). A deeper understanding of these processes will provide fundamental insight into how mechanisms of gene regulation could contribute towards disease (Degner et al., 2012, Trynka et al., 2013). While genetic fine-mapping and usage of reporter assays can help identify regulatory regions, methods to quantify the function of regulatory elements on target genes is essential (Ray et al., 2020; Sanjana et al., 2016; Tewhey et al., 2016; Vockley et al., 2017)

One promising approach to functionally characterizing an endogenous regulatory element endogenous is through epigenome-editing. The advent of CRISPR tools (Clustered Regularly Interspaced Short Palindromic Repeats) has since evolved its use in not only gene editing but also gene regulation. CRISPR activation (CRISPRa) and CRISPR interference (CRISPRi) are two such innovations that have opened new avenues for exploring gene function and understanding disease mechanisms. Specifically, perturbation of their predicted activity using CRISPR–dCas9 (dCas9) tethered to a transcription activator or repressor. Typically, the readout is measured in terms of an easily observable cellular phenotype like growth (Fulco et al., 2016; Sanjana et al., 2016), but this leads to limitations based on what cell types and phenotypes can be measured and if gene expression is measured by flow cytometry, it requires characterized antibodies for

successful detection (Klann et al., 2017). Simultaneously, the development of highly multiplexed imaging techniques has propelled our capacity to visualize and quantify gene expression within intact cells and tissues. A more direct and generalizable approach recently published called HCR Flow-FISH (Reilly et al., 2021) combines the use of RNA-probes that can bind to a gene transcript and is tethered to a fluorescent dye that can be used in flow cytometry. Hybridization Chain Reaction Flow with Fluorescence In Situ Hybridization (HCR Flow-FISH) combines the signal amplification capability of HCR with the spatial resolution of FISH. In HCR Flow-FISH a set of fluorescently labeled DNA probes, designed to hybridize to target RNA sequences, undergoes chain reaction amplification through alternating hybridization events. This results in the formation of long fluorescently labeled polymers, which can be quantified using flow cytometry or imaged with fluorescence microscopy. This chain reaction allows for the amplification of the signal or fluorescence and therefore allows the detection of genes expressed at low levels.

By combining the ability to manipulate gene expression with the capacity to measure RNA transcripts at single-cell resolution using flow-cytometry, we can investigate the functional consequences of gene regulation within complex biological systems and disease states. In this study, I describe establishing and scaling up HCR in two cell lines as a potential method for further functional analysis in GWAS loci. As proof of concept, we performed the first set of studies to measuring the gene expression response to glucocorticoids in a cell model. We then tested HCR Flow-FISH using H295R cells to build a cell model system for future perturbation studies across PCOS GWAS loci to identify target genes of PCOS-associated regulatory elements. We were able to use HCR Flow-FISH to measure changes in gene expression as a response to several different kinds of perturbations. These results highlight the scalable and

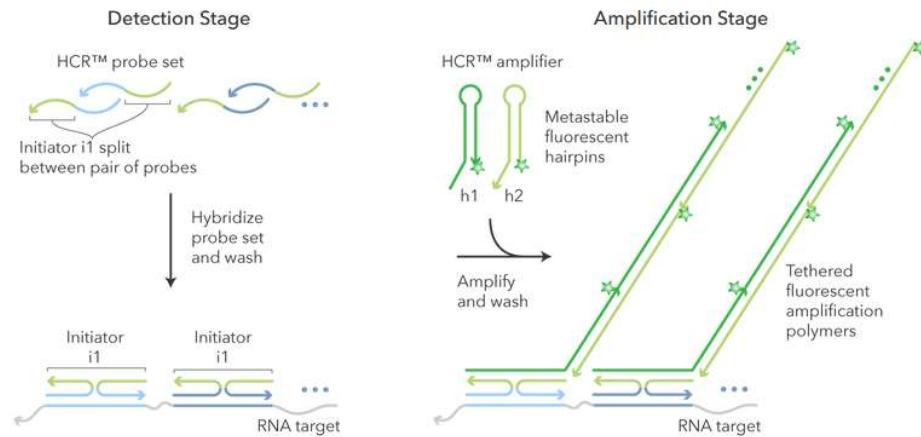
generalizable potential of HCR across different types of experiments. In this chapter, I describe several iterations to build a scalable approach to measure gene expression levels using HCR Flow-FISH.

## ***4.2 Results***

### **4.2.1 HCR Flow-FISH optimization**

Protocols for implementing HCR were first developed for imaging RNA through fluorescence in situ hybridization on fixed and mounted tissues and organisms (H. M. T. Choi et al., 2016) that were quantified using microscopy methods (H. M. Choi et al., 2014; H. M. T. Choi et al., 2018). HCR works based on the principle of complementary binding to mRNA. Probes are designed to bind the mRNA transcript of a gene. Each probe is designed to contain initiator sequences that are necessary for the amplification process. Once the probe is bound to the target mRNA, in the amplification step, a set of hairpin probes tagged with a fluorophore are added to the sample. The hairpin probes bind to the specific initiator sequences on the probes (Figure 14). The design of the hairpins is such that upon binding to the initiator sequence, the hairpins can bind to each other sequentially, thereby amplifying the fluorescence signal from the fluorophore. The intensity of the fluorescence signal is directly correlated to the amount of mRNA present that is bound by the target probes, thereby providing a fluorescence based approach to quantify the abundance of mRNA in a given sample.

### HCR™ RNA-FISH: How It Works



**Figure 14: Principle of hybridization chain reaction (HCR). Image adapted from Molecular Instruments**

At the time of optimization, HCR had not been used to measure gene expression in samples using flow cytometry. Therefore, to quantify gene expression levels in cell lines, I optimized the HCR protocol in three key steps : i) fine-tuning centrifugation speeds and durations for cell handling, ii) enhancing probe binding kinetics by adjusting of incubation times, and iii) adapting the protocol to incorporate control sample collection at a critical juncture.

First, since we were starting with cell lines and not fixed tissues or mounted organisms, we modified the fixing and permeabilizing step. We used 4% formaldehyde with incubation and rotations to fix the cells for 1 hour similar to ChIP protocols (Baranello et al., 2016). One key factor was altering the centrifugation steps since the cells were to be suspended in a solution for the final read out on the flow cytometer. To optimize the centrifugation, on average the speeds had to be increased to account for cell-swelling due to addition of formamide in the HCR probe and hairpin buffers. Second, we tested the incubation time for probe binding to be 4 hours vs 16 hours or overnight, and observed that with 4 hour binding, we did not see any change in fluorescence intensity. Therefore, for all the experiments described below, we used ~16 hours

incubation time for the probe hybridization. We also modified all incubation steps to include use of a rotator to account for the fact that the cells tended to settle at the bottom of the tube. And lastly, to allow for the judicious use of the cells in an experiment and to increase the number of cells used for the actual fluorescence experiment, we subsampled a portion of the cell population prior to the amplification step to use a negative control. We tested subsampling at 50% and 10% of the initial cell population, and identified that subsampling at 10% of starting with 5 million cells still allowed for robust gating strategy for flow cytometry measurements. Therefore, we reserved 10% of the cell population before the amplification phase, subjecting them to incubation in 1X PBS until the final assay step for the experiments described below. Furthermore, using the subsampled cells as a negative control allowed us to control for changes in cell sizes as a factor of the HCR protocol.

The approach of using flow cytometry and HCR to measure changes in gene expression was recently published and termed HCR Flow-FISH (Reilly et al., 2021).

#### **4.2.2 Inducible change in gene expression measured using an optimized HCR Flow-FISH protocol for cell lines**

The inducible activity of glucocorticoid receptor (GR) upon binding with glucocorticoids (GCs) to cause gene-regulatory responses makes it a tractable model to study gene regulation at putative regulatory elements (Reddy et al., 2009). Importantly, GR activation by GCs, including dexamethasone (Dex) can cause both activating and repressing transcriptional responses (Reddy et al., 2009). For this study, we selected 3 genes, PER1, IL11 and SF3B1. One of the most well characterized and reproducible GR responsive genes is PER1 (Reddy et al., 2012), shown to be induced in A549 cells when treated with Dex, reaching maximum expression level at 4 hours post Dex-treatment. In contrast, unstimulated A549 epithelial-like cells produced modest amounts of

IL-11. However, when A549 cells upon Dex treatment, demonstrate a decrease in IL-11 transcripts, as an example of GC-induced repressive signaling (J. Wang et al., 1999). SF3B1 was chosen as a reference gene that is stable even with Dex stimulation (Carmona et al., 2017; Reddy et al., 2009). Through this experiment, we aim to gain insights into the dynamic interplay between GR activation and gene regulation.

The three genes we selected can be classified as low, middle or high expressing under the Dex-induced or uninduced conditions (low; TPM  $\leq$  10; middle TPM  $\leq$  100; high TPM  $\leq$  250). For each of the three target genes, we were observed robust signal using HCR. PER1 expression increases by  $\sim$ 8 fold after 4 hours of Dex treatment. We observed  $\sim$  1.7 fold change as measured using median fluorescence intensity through HCR. There is very little PER1 expression prior to Dex treatment, which is in line with previous observations of the dynamics of PER1 expression (McDowell et al., 2018). Furthermore, we show that Dex treatment doesn't affect the signal in unamplified control cells. The fold change in IL-11 as measured by RNAseq is  $\sim$  6-fold. We measured  $\sim$ 1.4 fold change decrease in IL-11 measured by HCR with Dex treatment (Table 5).

Additionally, our analysis revealed distinct flow cytometry distribution profiles for the upregulated gene, PER1, and the downregulated gene, IL11. Notably, we employed identical fluorescent dyes and hairpins for both genes to minimize potential confounding variables. We also observed a difference in the flow-cytometry distribution profiles for an enhanced gene, PER1 and for a repressed gene IL-11. Upon induction by GR, PER1 exhibited a discernible rightward shift in distribution, with the proportion of PER1-positive cells increasing from 5% to 25% (Table 5). Conversely, in the absence of induction, the elevated expression levels of IL-11 appeared to be

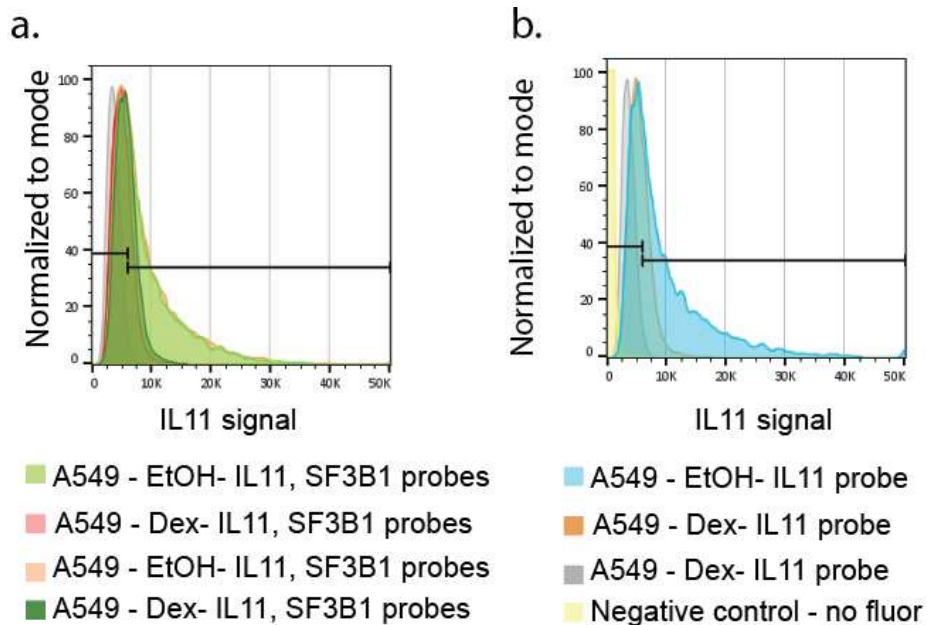
confined to a subset of cells, evident from the right-sided tail of the distribution. One explanation for the skew could be the combinatorial aspect of other gene regulatory factors in the control of IL-11 in the uninduced state. Such combinatorial regulation underscores the complexity in controlling gene expression dynamics, and further investigation into the underlying molecular mechanisms of IL-11 regulation under basal conditions can help resolve the signal.

### **4.2.3 HCR Flow-FISH allows multiplexed readout of target genes**

One of the key advantages of HCR is in the ability to have a multiplexed readout, limited by the fluorescence probes available and filters on a flow-cytometer. At the time of these studies, there were six different amplification hairpins available through Molecular Instruments, with the option of different fluorescent tags bound to this. Consequently, in our experimental setup, the compatibility of each target gene probe with specific hairpins ensured that amplification occurred exclusively with the corresponding hairpin conjugated to a particular fluorescent dye. This strategic design allowed for the simultaneous quantification of multiple target genes, optimizing experimental efficiency and resource utilization. In doing so, we were able to simultaneously measure the levels of PER1 and SF3B1 and IL-11 and SF3B together, respectively.

Specifically, we designed the probes for IL11 and PER1 to allow for binding to the amplification hairpin B2, and the probes for SF3B1 allowed for binding of the amplification hairpin B3. We designed the B2 hairpins to include fluorophore AF-488 with excitation wavelength 490 nm and emission wavelength 525 nm. We used AF-647 fluorophore for the B3 hairpins which had excitation wavelength 650 nm and emission wavelength 671 nm. The fluorophores were chosen to avoid spectral overlap between the emission and excitation wavelength range to minimize fluorescence spillover. The hairpins were chosen such that we

could combine both B2 and B3 hairpins that would bind with the respective sequences in the targeting probes, and the amplification of one hairpin does not interfere with the amplification of the other hairpin. We observed that the fluorescence signal distribution of IL11 is not affected by the addition of the probes for SF3B1, and was not impacted by exogenous treatment with Dex (Figure 15).



**Figure 15: Comparison of intensity of IL11 by HCR Flow-FISH.**  
**a) IL11 signal across samples multiplexed with both IL11 and SF3B1 probes. b) IL11 signal across samples with only IL11 probe**

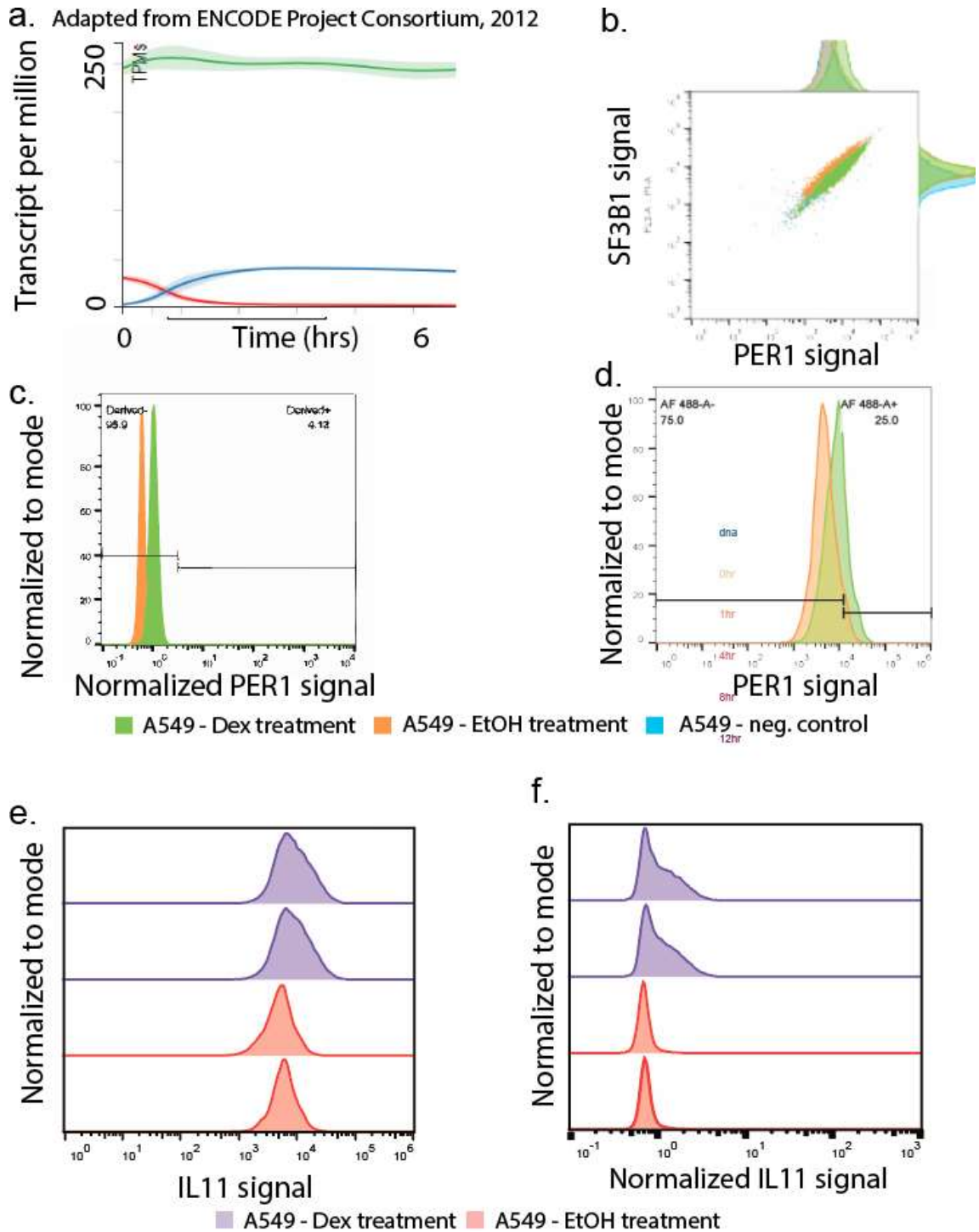
#### 4.2.3.1 Normalization approach using a housekeeping gene

An essential step that allows direct comparisons of gene expression levels across experiments is normalization of the signal to correct for various sources of variability in measurement of the signal. For example, when measuring gene expression using RT-qPCR, normalization is done by using the bulk signal for an internal reference gene or housekeeping

gene. For measuring gene expression levels using RNA-seq, normalization methods include using library size to account for differences in sequencing depth and by using the distribution of read counts for non-differentially expressed genes or exogenously added controls (Anders & Huber, 2010; Trapnell et al., 2010; Vandesompele et al., 2002).

Therefore, to normalize the signal intensity readout in HCR Flow-FISH, we first identified the distributions of multiple gene expression signals and with measurements of cell size measured as side scatter – area (SSC-A) using flow cytometer data. We observed a robust correlation between cell size and the SF3B1 signal, as well as a similar correlation between PER1 signal and cell size. We also observed a correlation between the SF3B1 signal intensity and PER1 signal intensity (Figure 16a). Therefore, leveraging this correlation, we used the SF3B1 signal as an internal control to normalize the signals of PER1 and IL11.

To normalize the signal for the target genes, for each cell we divided the signal from the target gene to the signal from the housekeeping gene on a log scale. This normalization strategy significantly improved signal resolution, allowing for a more accurate determination of fold changes in gene expression attributed to Dex treatment. Furthermore, by normalizing against SF3B1, we can increase the interpretability of gene expression analyses, and compare across experiments.



**Figure 16: Optimizing HCR Flow-FISH in A549 cells.**  
 a) Average RNA-seq TPM for IL11 (red), SF3B1 (green) and PER1 (blue). b) Scatterplot comparing the intensity of PER1 signal to SF3B1 signal. c, d) PER1 signal in A549 cells with

**Dex or EtOH treatment and with (c) or without normalization (d). e,f) IL11 signal in A549 cells with Dex or EtOH treatment and with (f) or without normalization (e).**

**Table 5: HCR Flow-FISH identified changes in gene expression due to Dex treatment in A549 cells**

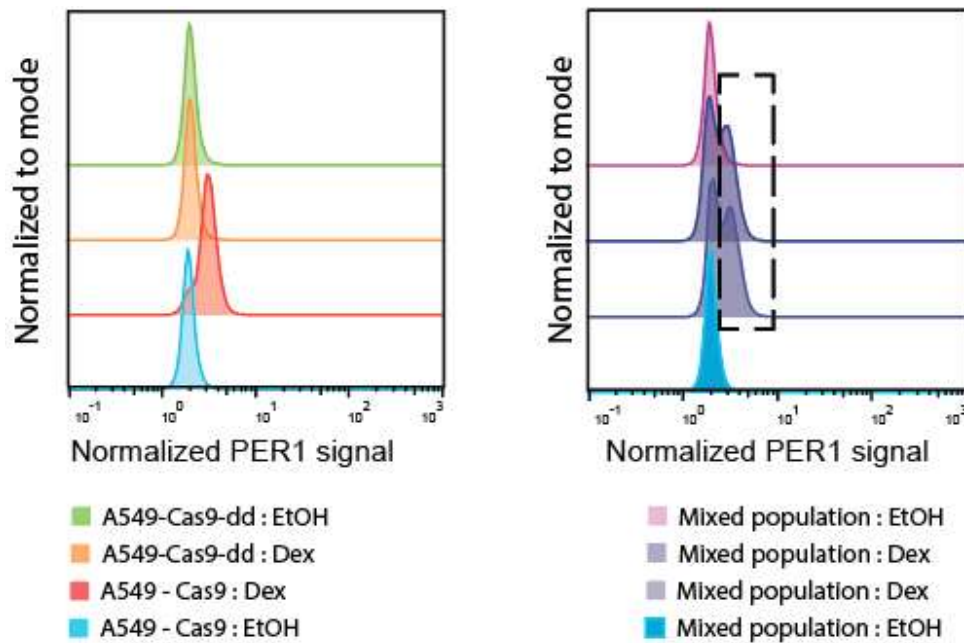
<b>HCR probe</b>	<b>Measurement</b>	<b>% Positive in EtOH treated</b>	<b>% positive in Dex treated</b>
<b>IL11</b>	<i>Unnormalized signal</i>	15.90%	1.31%
	<i>Normalized signal</i>	29.50%	1.43
<b>PER1</b>	<i>Unnormalized signal</i>	6.60%	25%
	<i>Normalized signal</i>	4%	94.60%

#### **4.2.4 HCR Flow-FISH captures altered gene expression levels in perturbations relevant to glucocorticoid responses**

##### **4.2.4.1 Genome editing of A549 to create Dex-unresponsive PER1 expression**

We next tested the ability of using HCR as a tool to measure perturbations of regulatory elements. We first examined the effects of deleting a putative GR binding site, identified as an enhancer of PER1 gene in A549 cells in collaboration with Dewran Kocak . Briefly, gRNAs targeting a putative GR binding site, that also showed evidence of regulatory activity through STARR-seq (Johnson et al., 2018) was transduced into A549-Cas9 cells which are A549 cells stably expressing Cas9 protein. The ~400 bp deletion was confirmed through PCR. This deletion effectively removed the putative GR binding site, thereby likely disrupting the regulatory activity associated with the enhancer element.

To test the effect of the deletion of the putative enhancer on PER1 expression, we measured the gene expression level of PER1 using HCR. We successfully measured that this deletion ablated GC-induced PER1 expression in the A549-Ca9-dd cells compared to the control which was A549-Cas9 cells (Figure 17, Table 6). This observation underscores the role of the deleted regulatory element in mediating GC-induced PER1 expression. These experiments highlight the ability of using HCR to measure gene expression responses to regulatory element perturbation.



**Figure 17: HCR Flow-FISH measures gene expression changes due to genomic editing in A549 cells.**

**Table 6: PER1 signal measured by HCR Flow-FISH in A549 cells with enhancer deletion**

HCR probe	Normalized signal measured in	% positive in EtOH treated	% positive in Dex treated
PER1	<i>A549-Cas9</i>	6.51%	79.8%
	<i>A549-Cas9-dd</i>	8.73%	8.53%

We then explored the feasibility of sorting cell populations based on gene expression levels using HCR Flow-FISH and subsequently genotyping them. To do so, we generated a mixed population comprising equal proportions of A549-Cas9 and A549-Cas9-dd cells (Appendix C, Figure S2), representing different genotypes with distinct PER1 expression profiles. We then subjected this mixed population to HCR Flow-FISH to quantify and sort based on PER1 expression levels. We demonstrated that HCR is able to successfully capture a bimodal distribution of PER1 expression upon Dex treatment. On average, we estimated ~48.1% cells as PER1 positive in the Dex condition compared to 7% in the EtOH condition. We then validated our approach for sorting cell populations based on PER1 expression for genotyping purposes. We isolated up to 20,000 cells from PER1-positive and PER1-negative bins and extracted genomic DNA. We were able to recover genomic DNA, and a PCR of the PER1 enhancer region verified the successful deletion of the enhancer in the subset of the PER1-negative cells.

Overall, we demonstrate a scalable approach to measure changes in gene expression levels that can be combined with CRISPR-based perturbations to test the role of regulatory elements.

#### **4.2.5 HCR Flow-FISH captures altered gene expression levels in perturbations of genomic loci associated with PCOS**

The ability to measure gene expression levels in single cell using flow cytometry is a major advantage of using HCR Flow-FISH to identify target genes of regulatory elements. To establish HCR Flow-FISH as an approach to investigate PCOS GWAS risk loci, we first tested the ability to measure changes in gene expression H295R cells using the PER1 locus that we established in A549 cells previously. Studies have reported that H295R cells are also responsive to glucocorticoid stimulation and express PER1 gene upon stimulation with Dex. We observed a

1.3X fold change in PER1 expression as measured by HCR Flow-FISH in H295R cells upon Dex treatment (Appendix C, S3).

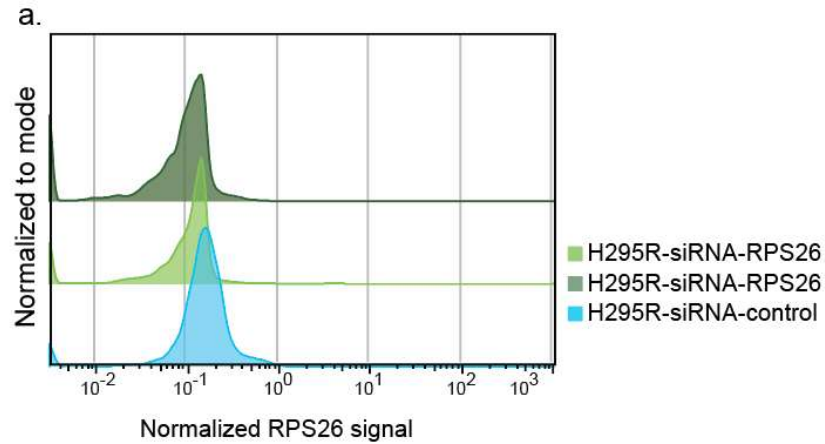
To then test the method in PCOS GWAS risk loci, we tested the methods in two loci that are associated with PCOS risk from our studies in chapter 2 and 3. First, we focused on the RPS26 locus which has been associated with PCOS and essential for oocyte development (Censin et al., 2021; X.-M. Liu et al., 2018; Sun et al., 2022). This locus on chromosome 12 also harbours one of the strongest colocalization signals for rs1081975 which is intronic to SUOX gene is RPS26 from our results in chapter 2. We then expanded the method to test for measuring changes in DENND1A expression due to the perturbation of regulatory elements identified in the DENND1A locus described in chapter 2.

#### **4.2.5.1 Transient knockdown of RPS26 using siRNA**

To establish HCR Flow-FISH as a model to investigate PCOS GWAS risk loci, we first tested the ability to measure changes in gene expression H295R cells due to siRNA mediated transcription knock down. For this study, we selected RPS26 as the target gene and GAPDH as the housekeeping gene, both known for their high expression levels - TPM for RPS26 is ~500 and TPM for GAPDH is ~3800 (Scholl et al., 2018). siRNA knockdown is a widely employed strategy for transiently suppressing gene expression by harnessing the RNA interference (RNAi) pathway. By introducing synthetic siRNA molecules complementary to target mRNA sequences, specific genes can be selectively silenced.

We observed a modest decrease in RPS26 expression in these cells when treated with siRNA targeting RPS26 compared to a siRNA negative control. When measured using RT-qPCR

on these cells, targeting RPS26 with an siRNA reduced the gene expression by 0.015x (Figure 18, Table 7). The fold change measured using geometric means was 0.55x using the flow cytometric geometric means. This observed difference underscores the utility of HCR as a complementary approach for gene expression analysis, while also highlighting the differences in the sensitivity and dynamics of these two techniques.



**Figure 18: RPS26 signal is decreased by siRNA knockdown measured by HCR Flow-FISH**

**Table 7: RPS26 signal measured by HCR Flow-FISH in H295R cells with siRNA perturbation**

HCR probe	Measurement	siRNA Neg control	siRNA RPS26
RPS26	<i>Unnormalized signal</i>	21.50%	19.90%
	<i>Normalized signal</i>	15.30%	5.7%

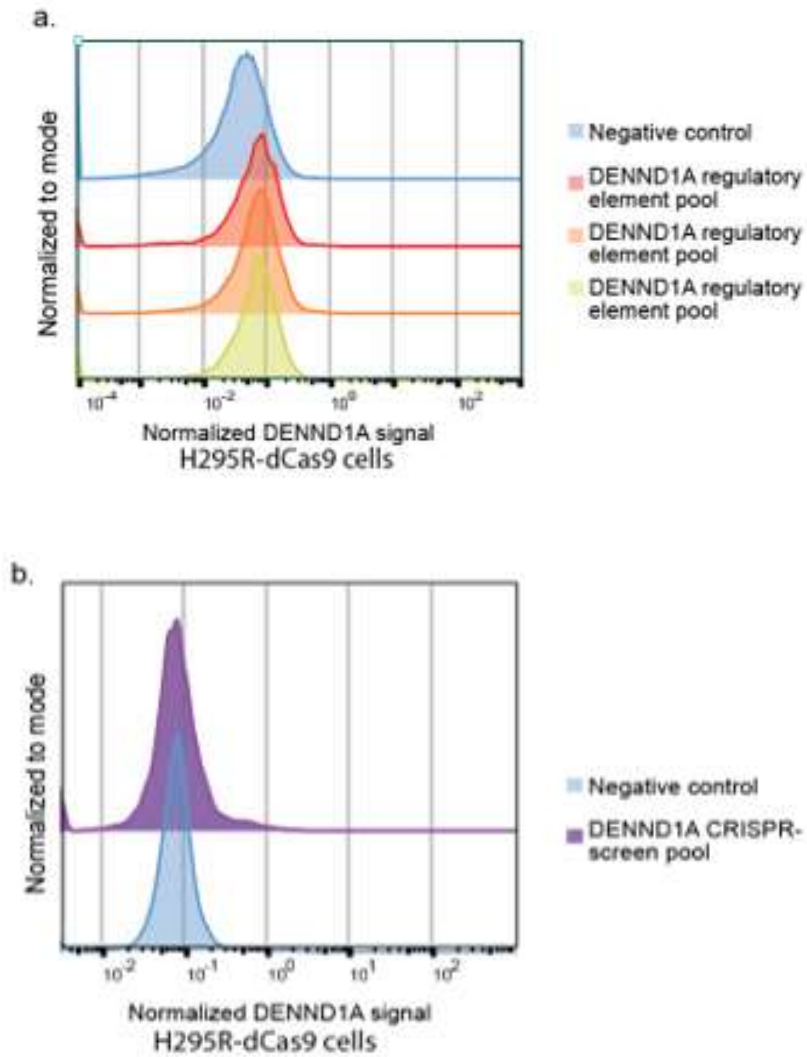
#### 4.2.5.2 Epigenomic perturbation in H295R effected DENND1A levels measured by qPCR

The CRISPR-Cas9 system has been extensively utilized for targeted gene or targeted enhancer knockouts, while the CRISPR-dCas9 system fused to transcriptional activators recruit endogenous transcription machinery to specific genomic loci. We tested the potential of HCR Flow-FISH to measure epigenomic perturbations, such as targeting regulatory elements with

dCas9-P300 to increase target gene expression levels. To do so, we used the H295R-dCas9-P300 cell lines that were created as described in chapter 2 of this dissertation. To measure changes in DENND1A expression using HCR Flow-FISH, we pooled the three cell lines each transduced with the lentivirus for regulatory elements 2, 3, and 4 described in chapter 4. The regulatory elements were identified to impact DENND1A gene expression described in chapter 2. We observed an increase in DENND1A-positive signal, increasing from 5.28% to 14.5%, with a fold change of 2.2X when measured using geometric means, while the average fold change measured by qPCR was 1.2X across the three regulatory elements (Figure 19a, Table 8). Overall, we were able to identify increased DENND1A expression level through targeted epigenomic perturbations.

One advantage in using a flow cytometry based readout is that the signal is measured at the single cell level, and therefore allows for sorting cell populations based on gating on the fluorescence signal as a proxy for sorting cells with increased expression of DENND1A. This ability could then be optimized to identify functional regulatory elements from a pool of candidate regulatory elements. To test the potential of using HCR Flow-FISH as a readout for CRISPRa-screen, we designed a guide library targeting candidate regulatory elements in DENND1A locus. We designed and tested a pool of gRNAs targeting candidate regulatory elements measured in the DENND1A locus. The total number of targeting gRNAs was 1200 and the total number of non-targeting gRNAs was 500. We observed a modest increase in DENND1A-positive signal from 6% to 11.8%, with a fold change of 1.02x when measured using geometric means (Figure 19b, Table 8). Ongoing work in optimizing HCR Flow-FISH demonstrates the potential in using HCR to identify functional regulatory elements and target

genes. Overall, our results highlight the versatility of HCR Flow-FISH in testing the functional consequences of regulatory element perturbations.



**Figure 19: DENND1A signal is increased by dCas9-p300 perturbation in H295R cells**

**Table 8: DENND1A signal measured by HCR Flow-FISH in H295R cells with 2 different gRNA pools targeting regulatory elements in DENND1A locus perturbation**

HCR probe	Normalized signal measured in	With gRNAs	Negative control
DENND1A	<i>H295R-dCas9-p300 - 3 gRNA pools</i>	5.28%	14.50%
	<i>H295R-dCas9-p300 - 1500 gRNAs library</i>	6.00%	11.80%

### 4.3 Discussion

In this chapter, I have reported on the optimization of HCR Flow-FISH protocol as a tool for measuring the target genes for regulatory elements. Broadly, one of the challenging aspects in deciphering regulatory elements has been identifying the target gene(s) of those regulatory elements. Some of the challenges include the tissue-type and developmental stage-specificity of regulatory elements. I have optimized HCR Flow-FISH to measure changes in target gene expression levels as a result of different types of perturbation methods. Specifically, we tested the ability of HCR Flow-FISH in detecting gene expression changes due to environmental/drug perturbation, siRNA perturbation and both CRISPR-based genomic and epigenomic perturbations. Taking the results together, we show that this method is agnostic to the type of perturbation applied to the cell system. Overall, we have demonstrated a generalizable, and scalable method to measure changes in gene expression levels

### **4.3.1 Other uses of Flow-FISH in genomics research**

HCR Flow-FISH has a major advantage in the ability to measure readout from a multiplexed high-throughput perturbation approach. Previously, several studies have used flow cytometry to measure changes in gene expression levels by using phenotypic measurements, an antibody against the target gene product or creating a protein-tagged target gene (like FLAG tag) to use an antibody specific to that protein tag (Black et al., 2020; Kulsuptrakul et al., 2021; Sanchez et al., 2021; Xiang et al., 2019). HCR Flow-FISH, however, is based on designing probes that bind to the RNA transcript and uses hairpin amplification to generate the fluorescent signal, thereby avoiding the biases of antibody availability for the target gene. Additionally, HCR Flow-FISH also offsets the limitations of using cell death or other growth phenotypes as measurements for screens.

There is potential for future HCR-FlowFISH experiments to yield high-quality results using a wider range of potential gRNAs to build more comprehensive screens and improve interpretability. Furthermore, the ability to multiplex target genes will allow for the simultaneous measurement of other genes in the locus to test for the effects of regulatory elements on multiple genes. Additionally, there is potential in using HCR Flow-FISH to identify splice-variant specific transcripts with judicious probe design restrictions.

### **4.3.2 Advantages and considerations for implementing HCR Flow-FISH**

We successfully assessed HCR Flow-FISH on two different adherent cell lines to measure changes in gene expression levels. There are, however, some limitations to this method. While HCR Flow-FISH allows for the identification of target genes of several regulatory

elements, it currently lacks the resolution for identifying the functional effects of single nucleotide genetic variants. Another challenge in the widespread use of epigenome editing with HCR is in the delivery challenges dictated by the experimental cell or tissue type. Some of the challenges specific to the cell type nature can be overcome with more thorough optimization for that specific cell model. Over the course of developing and optimizing the protocol, there have been other studies and changes across the HCR research community that need to be taken into account for designing HCR studies.

First, at the time of these experiments, the number of probes per target transcript was limited to 10, set by the probe design company, Molecular Instruments. In fact, the number of probes we obtained for the RPS26 transcript totaled to 6. Recently, Molecular Instruments has expanded the number of probes per target gene, scaling up to 40. Furthermore, a recent study detailing optimization efforts for HCR identified that higher probe number and probe concentration allowed for better signal-to-noise ratio for transcripts that typically have less 100 TPM in RNA-seq data (Reilly et al., 2021; Schwarzkopf et al., 2020).

Second, for the H295R cell studies, we used GAPDH as a control. However, GAPDH is ubiquitously expressed at very high levels (TPM > 3000), and our data show that the probe concentration used in the protocol was not sufficient to tag all the transcripts, thereby acting as a limiting reagent. To offset this limitation, in our analyses, we subset to only the GAPDH-positive cells to measure gene expression of other target genes. A solution to this would be either testing newer probe design approaches and/or identifying a different housekeeping gene.

Third, one of the technical challenges in our experiments was in the ability to sort cells using flow cytometry methods after the HCR protocol. The technical challenges included clumping of cell prior to the sort which can clog the fluid lines in the sorter. Even with filtering the cells, we observed clogs. We also observed a significant drop in the total number of cells after the gating strategy (> 50%) to sort as target gene-positive or target gene-negative. This affected our ability to successfully isolate genomic DNA and identify gRNAs that were enriched in the high versus low bins of the cytometer. Furthermore, the cell number limitations as a result of the gating strategy required starting the experiment with a much higher population of cells (> 10M).

Lastly, one of the main challenges in these experiments was the generation of a clonal population of H295R cells with the dCas9-p300 protein and the gRNA libraries. H295R cells were easily stressed by antibiotic selection and single-cell isolation was ineffective. One strategy around this is to redesign the plasmids to include a fluorescent tag as an approach to sort out cells that only have both the Cas proteins and the gRNAs.

## ***4.4 Methods***

### **4.4.1 Cell Culture**

A549 cells were obtained from ATCC. A single seed stock of A549s was first plated into a 15 cm dish and grown under standard culture conditions using Ham's F-12K (Kaighn's) Medium, 10% FBS, 1% penicillin-streptomycin. Once confluence was reached (approximately 25 M cells), 500  $\mu$ M Dex in 100% EtOH was added to a final concentration of 100 nM for 4 hours. All treatments were performed in parallel with an equal volume (0.02% [vol/vol]) of ethanol to control for solvent effects.

A549-Cas9 and A549-Cas9-dd cells were a gift from Dewran Kocak and the Gersbach lab at Duke University. A549-Cas9 cells were produced by the stable integration of Cas9 protein from a lentivirus containing the Sp-Cas9 plasmid (#39312). To generate the A549-Cas9-dd cells, gRNA were designed targeting to GR binding regions in the PER1 locus. These regions were identified as enhancers for PER1 expression induced by GC stimulation (22801371, 30575722). A549-Cas9 expressing cells were transduced with gRNA oligos. The targeting gRNAs were transfected into 15 cm<sup>2</sup> plates of A549 cells at ~75% confluence using Lipofectamine 3000 (Thermo Fisher Scientific) scaling the manufacturer's recommended protocol. Genomic DNA was harvested from the cells and cell lines containing a knock out of the two GR regions and was confirmed using qPCR. A549-Cas9 and A549-Cas9-dd cells were then grown and maintained under the above described conditions for the rest of the experiments

We obtained NCI-H295R cells from ATCC. The cells were cultured in DMEM/F-12 medium (Gibco #21041025) supplemented with 2.5% Nu-Serum (Corning #355100) and 1% ITS+Premix (Corning #354352) and grown as a monolayer at 37 °C, 5%CO<sub>2</sub>. The H295R-dCas9 cells were grown under the above described conditions. All experiments were performed between passages 5 and 20.

#### **4.4.2 RPS knockdown in H295R cells**

To generate the knockdown H295R cells, we used siRNAs. SiRNA targeting RPS26 (Thermo Fisher #142221) and siRNA Negative control (Thermo Fisher #4390843) were ordered. To insert the siRNA into H295R cells, we transiently transfected the siRNA to the cells via the Lonza 4D-Nucleofector System. To 82 µl of the SF solution, we added 18 µl of the supplement and siRNA solution to a final concentration of 75 mM. 2 M H295R cells were transfected with

this 100  $\mu$ l solution using the SF-DN100 setting. The cells were incubated post transfection for 10 minutes, and then gently aspirated and plated into 6 well plates. Media was changed on the cells 18 hours post transfection, and the cells were harvested for HCR and RNA isolation 48 hours post transfection.

#### **4.4.3 RNA extraction and qRT-PCR**

We isolated total RNA from 0.6 M cells using the Qiagen RNeasy Midi kit including the on-column DNaseI digestion step. We treated the isolated total RNA with 1  $\mu$ L RNase Block (Agilent). We then synthesized cDNA by reverse transcription using Superscript III (800 U, Life Technologies) following manufacturer's protocol and an oligo-dT primer (Thermo Fisher #18418012). The RNA quality was assessed using RIN values  $> 8$  using an Agilent Tape Station. Each cDNA sample was subject to qRT-PCR with an AppliedBiosystems StepOnePlus system using TaqMan Assay probes (RPS26 Hs00762561\_s1 and GAPDH Hs02786624\_g1 ) and TaqMan Master Mix (Applied Biosystems #4444556). GAPDH was chosen as the internal reference, and the relative fold expression of RPS26 was calculated using the comparative critical threshold (Ct) method with normalization to the GAPDH gene based on the  $2^{-\Delta\Delta Ct}$  calculation. The qPCR analysis was performed in R. Each sample was measured in triplicate for the qRT-PCR.

#### **4.4.4 H295R-dCas9-p300 & gRNA lentivirus**

To make stable dCas9-P300 cell lines, we generated lentivirus expressing dCas9-p300. Briefly, we combined the following plasmids: dCas9-p300 (Addgene #83889), psMD2.G (Addgene #12259) and psPAX2 (Addgene #12260) with Lipofectamine 3000 (Invitrogen

#L3000001) and lipofected into HEK293T cells (ATCC #CRL-3216™) according to the manufacturer's protocol. After 14 to 20 hours, transfection media was exchanged with fresh media. We then harvested viral supernatant at 24 and 48 hours post lipofection. We concentrated the viral supernatant at 1/100x using LentiX Concentrator (Clontech #631232) following the manufacturer's protocols.

To produce lentivirus for individual gRNAs, we transfected HEK293T cells with an equimolar pool of gRNA plasmids for each regulatory element, psPAX2, and pMD2.G using Lipofectamine 3000 following the manufacturer's instructions. We harvested media containing the produced lentivirus at 24 and 48 hours later and concentrated the viral supernatant at 1/100x using LentiX Concentrator (Clontech #631232) following the manufacturer's protocols.

#### **4.4.5 Generating stable H295R-dCas9-P300 cell line**

To make stable H295R cells expressing dCas9-P300, we transduced the concentrated lentiviral particles containing dCas9-p300 into H295R cells with a multiplicity of infection of 5.0 using 6 µg/ml of polybrene (EMD Millipore Corporation #TR-1003-G). Additionally, we selected for the transduced cells using 0.5 µg/mL of puromycin (Gibco #A1113803). We confirmed the expression of dCas9-p300 in H295R cells using qRT-PCR.

#### **4.4.6 Transduction of gRNA into dCas9-P300 expressing cell lines:**

To test the effect of P300 on the targeted regulatory elements, we transduced each lentiviral pool for the regulatory elements, *DENND1A* promoter region and negative control in two cell lines (HEK293T and H295R) with stable dCas9-P300 expression. We transduced the cells with the lentiviral pools targeting regulatory elements 2,3 and 4 as described in chapter 2.

We transduced the cells during seeding in a 6-well plate supplemented with 6 µg/ml of polybrene for H295R cells across three replicates (EMD Millipore Corporation #TR-1003-G). We changed the media on the cells 24 hours after transduction and harvested the cells for HCR protocol 4 days post transduction. For transducing the larger gRNA library, we used a similar approach, however, we started the experiments with transducing ~10M cells.

#### **4.4.7 CRISPR screen DENND1A gRNA pool library design**

We used the ~38 candidate regulatory elements identified using STARR-seq in chapter 2 as genomic coordinates for designing gRNA. We also designed a set of guides that did not have any targets in the human genome, and included gRNAs in the DENND1A locus that were within DHS

The regions gRNA library included 528 scrambled guides that do not target any part of the human genome, 20 guides targeting the DENND1A promoter region, 1113 gRNAs targeting closed chromatin regions in H295R in the DENND1A locus, and 822 gRNAs targeting candidate regulatory elements identified in H295R cells. The regions were selected based on the ability to design guides considering genomic sequence and PAM restrictions.

To design the guide oligos, we used Guidescan2 (28263296) with specificity filter > 0.2. The synthetic gRNA library was ordered as a pool of all the gRNA from Twist Biosciences. To make the gRNA plasmids, we followed the outline of the CROP-Seq protocol (28099430). First, we prepared the gRNA plasmid backbone by digesting CROPseq-Guide-Puro plasmid from Addgene (#86708) using BsmBI. We ran the digested product on 1% agarose gel, and we purified the 8.3 kb fragment using GeneJET Gel Extraction Kit (#K0691). The gRNA pool was diluted in 11 µL water. To prepare the gRNA oligos for insertion into the plasmid, for each gRNA oligo

synthesized, we first converted it to a double stranded oligo using Primers ssds-F and ssds-R (Appendix A, Table S3). We amplified the guides using PCR with cycling conditions: 98 °C for 3 mins, followed by 10 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 10 s, with a final extension at 72 °C for 2 min. The PCR products were cleaned using Axygen Beads at 1.3x volumes. We then cloned gRNA oligo pool into the digested CROPseq-Guide-Puro vector using NEBuilder HiFi DNA Assembly (#E2621) kit, and split the reaction into 8 individual assemblies. The plasmid pool was purified using ethanol precipitation. We added 0.1X volume 3 M NaOAc and 2.5X volume cold 100% ethanol, and stored the mixture at -20 °C overnight. We then pelleted the DNA via centrifugation at 16,000 RCF for 30 min at 4 °C. We washed the pellets with 5 ml cold 70% ethanol, and resuspended in water to make the final gRNA plasmid library pool. This gRNA plasmid library was used to design lentiviral particles for transduction into H295R-dCas9-p300 cells.

#### **4.4.8 CRSIPR screen plasmid pool sequencing**

To verify that there are no biases in the gRNA fragments inserted into the CROP-seq backbone, we measured the sequencing reads mapping to each guide. Briefly, the gRNA inserts from the CROP -seq backbone was amplified. To verify the distribution of gRNA oligos in this library, 20 µL of 20 ng/µL dilution of this pool was amplified for sequencing on the Illumina MiSeq, single end sequencing of 26 base pairs, using custom primers. To prepare the sequencing libraries, we amplified the library using KAPA HiFi HotStart kit (Roche). The PCR cycling conditions were: 98 °C for 30 s, followed by 15 cycles of 98 °C for 15 s, 64 °C for 20 s, 72 °C for 30 s, with a final extension at 72 °C for 5 min. To isolate the final library, we used Axygen Spri Beads (AxyPrep™ Mag PCR Clean-Up Kit) beads at appropriate concentrations based on the manufacturer's manual.

The resulting sequences were aligned to a custom reference genome that we generated using the sequences of all the gRNAs ordered. The reads were aligned using bowtie2 (with the following options: `--norc \ -p 8 \ --no-unal \ --end-to-end \ --trim3 1 \ -D 16 -R 3 -N 1 -L 20 \`). We used samtools to create counts of reads mapping to each guide. We confirmed an even distribution of the gRNAs across the four categories (DENND1A- targeting candidate regulatory regions, NegCtrl - targeting closed chromatin regions in H295R, Promoter - targeting DENND1A promoter region, scrm - guides that do not target any part of the human genome). Results are displayed in appendix 4.

#### **4.4.9 HCR Flow-FISH**

Gene specific HCR probes and fluorescently labeled hairpins were purchased from Molecular Instruments using the sequences and targeted against the messenger RNA transcript most abundant in A549 cells. Buffers were optimized for HCR–FlowFISH and prepared in the lab using RNase-free practices. The procedure below has worked for starting cell quantities of 2M or 5M cells. All samples were prepared using low-binding plasticware (Eppendorf) where possible. All values below are listed per 5 M of cell aliquot. All centrifugations were performed at 500G for 5 min unless otherwise noted.

##### **4.4.9.1 Cell harvesting**

A549 or H295R cells grown to ~90-100% confluency were first washed with 1X PBS. We then dissociated the cells using Trypsin-EDTA 0.25% (Life Technologies). The cell pellets were harvested by centrifugation at 300 rpm for 3 minutes.

#### **4.4.9.2 Cell fixing and permeabilization**

The cell pellets harvested were resuspended in 1 ml of 4% formaldehyde in PBS with Tween 20 (PBST) (1× PBS, 0.1% Tween 20) and incubated at room temperature for 1 h with rotation. After centrifugation, we washed the cell pellet with 0.5 ml of PBST and repeated for a total of 4x PBST washes. For the last wash, the cells were suspended in PBST, and incubated at room temperature with rotation for 5 minutes.

#### **4.4.9.3 Probe hybridization**

We pelleted the cells and resuspended the cell pellets in 400 µl of pre-warmed probe hybridization buffer (30% formamide, 5× sodium chloride sodium citrate (SSC), 9 mM of citric acid, pH 6.0, 0.1% Tween 20, 50 µg ml<sup>-1</sup> heparin, 1× Denhardt's solution and 10% low molecular weight dextran sulfate). The cells were incubated in this buffer at 37 °C, in a shaker incubator for 30 minutes. During the incubation, we prepared a 1X probe solution by combining 2µl of 1uM probe stock with 100 µl of the probe hybridization buffer. We then added the probe solution to the cell sample and incubated overnight (15–20 h) at 37 °C with rotation.

For multiplexed experiments, each probe was diluted in the probe hybridization buffer separately at volumes that would total to 100 µl. For example, for multiplexing probes for PER1 and SF3B1, 2µl of each probe was mixed with 50 µl of the probe hybridization buffer. We added 50 µl of each probe solution to the cells in the probe-hybridization buffer totaling to 100 µl.

#### **4.4.9.4 Excess probe washing**

After incubating with the probes overnight, we added 1000 µl probe wash buffer (30% formamide, 5× SSC, 9 mM of citric acid, pH 6.0, 0.1% Tween 20 and 50 µg ml<sup>-1</sup> heparin) pre-warmed to 37 °C to each sample, followed by incubation at room temperature for 10 minutes with

rotation. We then centrifuged the samples at 800-900 G for 5 mins in a swing-bucket centrifuge, with the longer spins necessary for the effect of cell-swelling caused due to the formamide the probe-hybridization buffer. We then resuspended the pellets in 500  $\mu$ l probe wash buffer and repeated the wash 4x times. After the final wash, we resuspended the cells in 500  $\mu$ l of SSC with Tween 20 (SSCT) wash buffer (5 $\times$  SSC, 0.1% Tween), and incubated at room temperature for 5 mins with rotation. We then took out ~10% of the cells from 2 different samples as the unamplified control cells. The control cells were resuspended in 1X PBS and subject to the same temperature and rotations condition as the other cells.

#### **4.4.9.5 Signal amplification**

We pelleted the cells in SSCT and resuspended them in 150  $\mu$ l amplification buffer (5 $\times$  SSC 0.1% Tween, 10% low molecular weight dextran sulfate) pre-warmed to 37  $^{\circ}$ C. The cells were incubated at room temperature for 30mins with rotations. During this incubation, we prepared the amplification hairpin solution. We heated 5 $\mu$ l of each hairpin stock per sample (H1 & H2) separately to 95  $^{\circ}$ C for 90s, and then cooled down at room temperature in the dark for 30 mins. 100  $\mu$ l of the pre-warmed amplification buffer was added to the now mixed (5  $\mu$ l of the corresponding H1 & H2 are combined) hairpin solution. We then added 100  $\mu$ l of the amplification buffer containing the probe hairpins to the incubating cell solutions. The cells with the hairpins were then incubated at 30  $^{\circ}$ C in the dark with rotation for 2-4 hours.

#### **4.4.9.6 Excess hairpin washing**

After incubation with the hairpins, we added 1000  $\mu$ l SSCT to the cells, and then centrifuged. We then resuspended the cells in 500  $\mu$ l SSCT and centrifuged. This SSCT wash was repeated 5x times. After the last wash, the cells were suspended in 1X PBS, and filtered through a

<40  $\mu\text{m}$  nylon mesh filter. The cells suspended in PBS were stored at 4 °C before using the flow-cytometer.

#### **4.4.9.7 Flow cytometric analysis**

Cells were sorted with a Sony Sorter SH8000, using a 130- $\mu\text{M}$  chip (Sony), using the 405 nm, 488 nm and 638 nm lasers. Cells were first gated for live, single cells and the fluorescence from the hairpin was read out using appropriate channels and compensation filters. We used a minimum of 2 replicates for a total of 10000 cells across all events. The data was analyzed using FlowJo v.10.10 (FlowJo LLC).

Normalization: For subsequent HCR experiments where the target gene measurement was multiplexed with an HCR target for a housekeeping gene, the signal intensity was measured as  $(\text{Intensity for target gene})/(\text{Intensity for housekeeping gene})$  and the histograms for this normalized signal was plotted in log scale. For the experiments where GAPDH is used a control, to account for a limiting HCR probe to gene transcript level in the cells, normalization was done on GAPDH positive cells (after gating for live cells and singlets) and all the results shown are for this subpopulation.

## 5. Summary and future directions

### 5.1 Summary

In this work, we followed up on previously reported genetic associations to polycystic ovary syndrome and identified potential gene regulatory mechanisms contributing to PCOS phenotypes. Among the challenges in has been the need for improved technological and statistical methods that can yield insight into how genetic associations can give information on mechanistic models of diseases or traits. With sequencing as the readout for the STARR-seq assays, there is a tradeoff between the sequencing depth and coverage and the complexity of the regions being sequenced. That is a smaller library can be sequenced deeper to yield higher confidence in the readout for the same amount sequencing reads than a larger library.

To implement the STARR-seq assay to identify regulatory elements in PCOS-associated loci, we first devised an approach using bacterial artificial chromosomes and fosmids to capture the potential regions of interest and assay them for regulatory activity. Using this approach, we identified ~1000 regulatory elements across PCOS associated loci as described in chapter 2 while sequencing to a depth of ~350 million reads using 50 bp PE sequencing. However, while this method had the resolution to be able to identify regulatory elements, using BACs and fosmids restricted the ability to identify allele-specific regulatory variants. To be able to use STARR-seq to quantify regulatory effects of genetic variants, we first designed the capture strategy to enrich for a targeted region using limited started material of genomic DNA. The capture method allowed us to be able to identify variants in a pool of five genomes while sequencing to a depth of ~300 million reads using 150 bp PE sequencing for coverage across the locus at > 100X. However, the statistical methods we used in peak calling for STARR-seq currently do not account for biases that were introduced due to the capture process in generating the assay library. To offset that bias

when using donor genomes, we further optimized the STARR-seq protocol to be able to use donor genomic DNA at low quantities (2-5  $\mu\text{g}$ ) as the input for the assay library as described in chapter 3. However, to increase the coverage across the whole genome, the libraries were sequenced to a depth of  $\sim 10$  billion reads with 150 bp PE sequencing, with  $> 20\text{X}$  coverage across the genome for each sample. We identified several candidate regulatory elements in an adrenal and ovarian granulosa cell model and this method can serve as a framework to identify regulatory elements in PCOS risk loci across several cell types.

To identify genetic variants that could be contributing to PCOS pathogenesis, we used two different methods. First, in chapter 2, we described a statistical variant association approach that identified novel PCOS-associated variants that were present in candidate regulatory elements identified. We further fine-mapped likely causal variants using a colocalization method. However, this method doesn't consider the variants present within the experimental assay or rare variants identified in previous PCOS studies. Additionally, the association analysis method currently is limited to common variants that were genotyped in the GWAS cohort. To measure regulatory effects of genetic variants experimentally, we used methods described in chapter 3, to be able to use STARR-seq to identify regulatory variants using the BIRD statistical method. Using donor genomic DNA allows us to identify regulatory variants from a naturally occurring population cohort including PCOS cases and controls. Furthermore, we can assay the regulatory effects of rare variants in the population. Expanding on this analysis will enable us to test the hypothesis of rare variants contributing to complex common genetic phenotypes.

A key advantage in using STARR-seq is the parallelization it offers in identifying regulatory elements and regulatory variants that could be adapted and tested across different cell

lines. Identifying the target genes of regulatory elements is crucial as it yields insight into the mechanistic understanding of the trait or disease. To identify target genes of regulatory elements we used two approaches, a statistical approach, and an experimental approach. First, we used a bayesian approach, colocalization, to prioritize candidate causal genes based on eQTL data. One challenge is that there is a discrepancy in the number of different tissue samples in the GTEx database which affects the statistical power to identify likely causal genes using this method. However, our colocalization results from chapter 2 further added evidence to the DENND1A locus contributing to PCOS pathogenesis. To identify target genes experimentally, we focused on the regulatory elements identified in the DENND1A locus. In this study, we developed a cell based assay that would allow for a systematic testing of candidate regulatory elements and their effects on testosterone production as a hallmark PCOS phenotype.

The CRISPR-perturbation results in chapter 2 demonstrate a scalable method in using this cell model to dissect the regulatory elements that contribute to PCOS phenotypes via altered gene expression levels. To facilitate scaling up of the epigenome-editing experiments, I also developed and optimized a flow cytometry based approach, HCR Flow-FISH, that can capture changes in gene expression at the single-cell. HCR Flow-FISH has advantages by using probes binding to the mRNA that accounts for biases introduced by endogenous fluor-tagging of genes or fluorescent antibodies that can bind to the target genes. Furthermore, we demonstrated the ability of HCR to be generalizable across two different cell models and types of genomic perturbations. We expect future optimization efforts to incorporate better probe design and offsetting technical challenges in cell model manipulation to scale up HCR Flow-FISH.

Overall, in this dissertation I have implemented several approaches to characterize the non-coding regulatory elements as one major step in understanding how gene regulation contributes to PCOS pathogenesis. More broadly, these methods can serve as an approach to investigate regulatory elements and their target genes in different contexts.

## ***5.2 Future Directions***

The work in this dissertation highlights the importance of studying the role of non-coding regulatory mechanisms in contributing to disease. With the pace of identifying new disease-associated genetic variants, the functional follow up of those variants becomes necessary in addition to developing technologies better suited to study those regulatory variants at scale. The results from our studies pose important questions as avenues for future research in both understanding general mechanisms of regulatory elements in human diseases, as well as following up on PCOS-specifically.

### **5.2.1 Regulatory mechanisms contributing to PCOS**

#### **5.2.1.1 How does DENND1A expression affect testosterone production levels?**

In our results, we identified putative regulatory elements within the 9q33.3 locus that regulate the expression of DENND1A. Results from our studies and previous studies demonstrated that DENND1A expression levels affect testosterone produced in ovarian theca cells and adrenal cell model (McAllister et al., 2014). In addition to elevated ovarian testosterone production, there is evidence of increased adrenal androgen in people with PCOS (R Azziz et al., 1998). Recently, studies showed that forced overexpression of DENND1A in a cell model increased the CYP17A1 mRNA levels (McAllister et al., 2014; Teves et al., 2020). Notably,

CYP17A1 codes for a protein that is involved in the biochemical conversion of cholesterol to testosterone. This insight in increased levels of CYP17A1 might be one possible model of increased testosterone levels. Based on previous studies, DENND1A levels impacts the levels of CYP11A1 and CYP17A1 genes which are involved in the steroidogenic pathway. Such inferential questions could be answered using epigenomic modifiers like dCas9-p300 and dCas9-KRAB in conjunction to test the role of both silencing and activating this gene.

There is yet to be studies that detail how increased DENND1A levels might lead to increased CYP17A1 levels. DENND1A is not a transcription factor, therefore it doesn't directly bind to DNA to regulate gene expression levels. One possible mechanism is based on the role of DENND1A in clathrin mediated endocytosis (CME). CME is a process by which cells control the specific uptake of certain substances like hormones or metabolites mediated by the cell-surface receptors that have specific cognate binding towards specific ligands. This mechanism is implicated in transducing signals from the cell periphery to the nucleus (Pyrzynska et al., 2009). One possible mechanism connecting DENND1A to upregulation of CYP17A1 could be involved in receptor-activation and recycling of the luteinizing hormone chorionic gonadotropin receptor (LHCGR) which is an upstream component in ovarian steroidogenesis (Richards, 1994), which is also a PCOS-associated risk locus. There is one study that identified protein-colocalization between DENND1A, LHCGR and RAB5B in theca cells (Kulkarni et al., 2019). RAB5B is one the genes on a reproducible PCOS GWAS risk locus on chromosome 12, which also includes the genes SUOX, ERBB3 and RPS26. It is known that RAB5B is involved in the early endosome process while RAB35 is involved in recycling endosome processes (J. Zhang et al., 2022). Previous studies have shown that the addition of molecules that stimulate cyclic-AMP (cAMP) leads to increased testosterone production in H295R cells (Haggard et al., 2017; Maglich et al.,

2014). Therefore, one hypothesis is that DENND1A-mediated endocytosis affects the cAMP signaling pathway which in turn, affects the expression of the genes involved in steroidogenesis (Kallen et al., 1998; Tremblay et al., 2002). However, there haven't been definitive conclusions on the cargo for DENND1A mediated endocytosis. It is also unclear how DENND1A-mediated endocytosis affects the dynamics of receptor signaling. To address these questions, one option would be to combine CRISPR-based DENND1A perturbation and identify all gene expression changes within the steroidogenic pathway downstream of DENND1A perturbation using RNA-seq. One additional possibility is identifying transcription factors whose expression levels might be altered by altering DENND1A expression levels, indicating targets for identifying transcription factors that promote steroidogenesis in the cells. This avenue of thought could lead to transcription factor binding experiments and chromatin interaction maps to identify transcription factor-promoter interactions and dynamics that might lead to another insight into how DENND1A functions. To study the endocytosis process itself, methods such as targeted live cell imaging of DENND1A, RAB35 and RAB5b, and hypothesized receptors involved in this process is one way to identify the interaction between DENND1A and receptor mediated endocytosis (Ecker et al., 2021; Krishnan et al., 2022). Furthermore, to test if increased testosterone levels due to increased DENND1A expression occurs via the cAMP signaling pathway, one way to test that is to include conditions that perturb the cAMP pathway. In contrast to the cAMP-dependent pathway, there have been other reports of cAMP-independent regulators of steroidogenesis observed in different cell types (Stocco et al., 2005). While we have identified regulatory elements that alter DENND1A expression levels, it is possible that these regulatory elements have different effects or even no effects in other cell types. Therefore, one key consideration is the ability to identify whether DENND1A expression is altered in other cell types that likely contribute to PCOS including ovarian theca and granulosa cells.

Much of the previous research implicating DENND1A expression on testosterone production has focused on a DENND1A variant, DENND1Av.2 which was first identified in theca cells of PCOS cases (Tee et al., 2016). Notably, the canonical variant and DENND1A v.2 are largely the same till ~exon 19. Therefore, if both isoforms are expressed, and share the same promoter region, the promoter perturbations that we performed would likely increase transcript levels of both isoforms. By implementing this method, we can then ask questions that connect the genetic variants that we identified from the whole genome POPSTARR method and regulatory elements to altered DENND1A expression levels. Specifically, is the V2 isoform expression in H295R cells and, do the genetic variants identified within regulatory elements within the locus affect the expression of one isoform over the other? One caveat to this approach is that if the V2 isoform is not produced by H295R cells, we would need to identify other candidate cell models. RNA-seq from H295R cells currently do not distinguish any isoforms of DENND1A. HCR Flow-FISH works by designing probes that binding specifically to the target mRNA. A modification to the probe design might allow for the simultaneous measurement of multiple splice transcripts of a given gene by using the multiplexing capability of HCR Flow-FISH. Consideration for probe design (and subsequent analysis) should include the length of transcript, the specific portions of the mRNA that differ between the isoforms, and possible off-target binding of the HCR probes. By combining epigenome editing and the results from our STARR-seq studies, we can use HCR Flow-FISH to quantify levels of DENND1A isoform and its role in steroidogenesis.

Taken together, there is still a gap in understanding how DENND1A might affect testosterone levels, and highlight the need for detailed experimental methods to dissect the mechanism.

### 5.2.1.2 Opportunities to test for mechanisms in other loci contributing to PCOS

In this dissertation, while we focused on DENND1A as a proof-of-concept to test for regulatory elements that could drive PCOS phenotypes, we have also mapped several hundred candidate regulatory elements across PCOS GWAS loci. In this section, I will highlight two other loci that could likely contribute to PCOS pathogenesis.

Notably, from our studies, as well as previous studies (Censin et al., 2021; Sun et al., 2022), RPS26 locus was identified to contain candidate causal genes for PCOS pathophysiology. This locus includes the genes : RPS26, SUOX, RAB5B and ERBB3. Gonadotropins have been shown to upregulate ERBB3 gene expression levels in ovarian granulosa cells. ERBB3 is involved in granulosa cell development and survival (Chowdhury et al., 2017; Mukherjee & Roy, 2013). One study that demonstrated the role of RPS26 in oocyte growth and follicle development, specifically reporting the premature ovarian failure like response to RPS26 knockout in mice (X.-M. Liu et al., 2018). Notably, from this study, the loss of RPS26 identified several genes that were up- or down-regulated, with the caveat that there was global decrease in mRNA expression levels. Furthermore, in this locus, RAB5B is involved in early endosome processes (J. Zhang et al., 2022) which I hypothesize might play a role in receptor recycling in the previous section with DENND1A. Additionally, Hi-C data from ENCODE from ovarian tissue demonstrated that this locus has increased chromatin interactions, while a recent study found evidence of an association between androgen, DHEA production by theca cells, and ten SNPs in a haplotype on chromosome 12 (R. A. Harris et al., 2023). Taken together, these findings along with our results support a conclusion that locus 12q13.2 contains important genetic determinants of PCOS. However, there has been no conclusive evidence on the target genes(s) of the regulatory elements in this locus. Similar to our experiments on DENND1A described in chapter 2, I expect

epigenome-editing coupled with measuring gene expression levels would aid in characterizing the regulatory environment in this locus. Furthermore, with several candidate genes in this locus, using several different cell types to identify target gene(s) would aid in characterizing any cell-type or tissue-specific effects of those regulatory elements.

Another locus that is of importance is the GATA4/NEIL2 locus where we identified several regulatory elements as well as new SNPs associated with PCOS. While the role of NEIL2 in DNA base excision repair and how it relates to cancer have been studied, including functional variants, the role of NEIL2 as related to PCOS is as yet under-explored (Hua & Sweasy, 2023). The group of GATA factors are evolutionarily conserved and share a highly conserved zinc finger DNA-binding domain. Specifically, GATA4, is a transcription factor that has been implicated in ovarian development and steroidogenesis (Anttonen et al., 2003; George et al., 2015; Tremblay & Viger, 1999, 2001). Additionally, the loss of GATA4 in mice has been shown to impair ovarian follicle development, impaired ovarian response to gonadotropins and granulosa cell proliferation (Efimenko et al., 2013; Kyrönlähti et al., 2011). To dissect the role of this locus in PCOS pathogenesis, in addition to similar epigenome-editing measurements, we could also use methods to quantify transcription factor binding sites of GATA4 in different tissues, and in response to environmental stimuli such as excess androgen to identify the downstream effects of GATA4.

With these two loci as examples, I have highlighted the need for the targeted follow up of regulatory elements to determine the mechanism behind PCOS pathogenesis. Between the results described in this dissertations, there are several other loci with regulatory elements that require a mechanistic follow up for their role in contributing to PCOS phenotypes. Understanding disease

etiology would also assist in better understanding the molecular mechanisms behind the sub-phenotyping of PCOS cases.

### **5.2.1.3 Role of PCOS associated genetic variants in contributing to developmental phenotypes**

In our whole genome STARR-seq approach, we were able to successfully identify thousands of regions with regulatory activity, we then focused on one specific locus to estimate the effects of allele-specific regulatory activity. Given the importance of DENND1A in androgenesis, one avenue of study can be the effects of the variants in both development and testosterone production. To study the role of DENND1A variants in cell development, one approach could be engineering the specific regulatory variants via CRISPR-based genomic editing into biologically more relevant tissues. While theca cells isolated from donors have been used in previous studies of DENND1A and PCOS, the cells are not easy to harvest and propagate in culture. Therefore, one option is to use induced pluripotent stem cells (iPSCs). There are some recent advances in in-vitro differentiation of iPSCs into cells from the female reproductive tract (Smela et al., 2023; Wu et al., 2023). This approach of combining differentiation of iPSCs along with genome editing of DENND1A variants will allow us to test both the development of theca cells as well as the ability of theca cells to produce testosterone. Importantly, this approach will yield a significant advance in the biological context of gene regulation in theca cells which is directly relevant to PCOS.

### **5.2.1.4 Identifying PCOS-associated rare variants from whole genome STARR-seq**

With the advent of GWAS and genetic variant-association testing, several studies have identified that the common SNPs associated with the trait do not fully explain the heritability

estimates of that trait (Manolio et al., 2009). One reason for missing heritability could be due to undiscovered variants including rare variants that are found at low allele frequencies in the population (Zuk et al., 2012). A recent study identified rare variants that contributed significantly towards the phenotypic trait using a population cohort of over 20,000 people (Wainschtein et al., 2022). Rare variants have been identified to be associated with PCOS implicated genes such as DENND1A, AMHR, BMP6 and AOPEP (Dapas et al., 2020).

Therefore, one major avenue of future study is to estimate the effect sizes of the rare variants that are associated with PCOS-cases when compared to controls. One approach to do this using bayesian approach, BIRD described in chapter 3, would be to model the regulatory effects in cases and controls separately and then use an appropriate statistical measure to identify which of those variant effects are more likely due to the PCOS status. This will test not just for functional rare variants, but also identify the variants that are PCOS-associated. In order to test for the functional effects of rare variants, one approach could be in association testing to identify whether a group of identified regulatory rare variants are associated with different PCOS traits. For example, a comparative statistical analysis can be designed to identify whether regulatory variants are clustered based on PCOS subtypes. However, it is possible that for the subtyping analysis, the sample size that we have in this study is limited, but it highlights the case for expanding the STARR-seq method across more samples. More broadly, there is evidence from our data and previous studies that suggests that estimating the functional mechanistic role of the rare variants associated with PCOS represents a next step in dissecting the genetic architecture of PCOS.

## 5.2.2 Mechanisms of regulatory elements in causing human diseases

### 5.2.1.2 How do regulatory elements work together?

Different cells display differential gene expression through a complex gene regulatory network comprising of cis-regulatory elements, trans-acting regulators and the target genes. There is evidence that multiple enhancers interact with promoters to regulate transcription of a gene (T. F. Consortium et al., 2014; Fulco et al., 2016; G. Li et al., 2012), however, it is unclear what the quantitative effects of multiple enhancers on target gene expression are. In one study using mouse embryos, it was observed that deletion of some enhancers displayed redundant activity, which may be spatio-temporally delineated across different cell types (Osterwalder et al., 2018). Another study demonstrated that an additive model of controlling gene expression in adult mice (Lam et al., 2015). However, there are still unanswered questions in both global- and locus-specific questions of how interactions between enhancers, silencers and promoters affect transcription.

One question is do multiple enhancers interact co-operatively or does each regulatory element function individually to contribute to the overall phenotype? We can then ask the question as to whether the regulatory elements follow an additive model, synergistic model or perhaps a redundancy role. The DENND1A locus can be used to determine the combinatorial role of enhancers within this region in determining expression of DENND1A and the nearby genes. One advantage of using this locus in the adrenal cell is the effect of DENND1A on testosterone production, which is an easily measurable phenotype. By combining the data from the locus in our whole genome STARR-seq and accessible chromatin regions, we can determine regions for targeted perturbation and systematically dissect how different regulatory elements contribute to

the overall phenotype. An example of a model to test for interaction could be a generalized linear model where the coefficient describes the action of each regulatory element, and for the combinations of regulatory elements perturbed. By focusing on a singular locus, the main advantage is that this will allow for testing the effects on a deeper scale than if we were to do a single-cell RNA seq across all the perturbations. Other approaches can combine methods to determine higher-order chromatin interactions with epigenomic editing and single cell read outs to elucidate the coordination among regulatory elements in eliciting a gene expression response.

#### **5.2.2.1 How do nuclear receptors play a role in human diseases?**

Nuclear receptors are a class of transcription factors that regulate several biological processes like metabolism, reproduction, and inflammation responses by modulating gene expression levels which is initiated through the process of binding with their cognate ligands. Aberrant nuclear receptor signaling is known to contribute to several diseases including several types of cancer, diabetes and adrenal hypoplasia congenita (Tenbaum & Baniahmad, 1997; Weikum et al., 2018).

Both glucocorticoid receptor (GR) and androgen receptor (AR) have been extensively studied for their roles in transcriptional regulation, particularly as activating transcription factors, increasing target gene expression levels. While there are known targets of GR- or AR-mediated gene repressions, the mechanisms of these methods are not as well understood (D'Ippolito et al., 2018; Grosse et al., 2012). One of the models for nuclear receptor mediated repression is one in which GR or AR binds to DNA-bound transcription factors and recruit co-repressors for the transcriptional repression activity (Avenant et al., 2010; Reddy et al., 2009). To dissect the transcriptional responses of GR and AR, one approach is to identify the other transcription factors

that confer specific transcriptional responses of specific genes upon GR or AR activation. Another possible avenue of study is to dissect the crosstalk between AR and GR signaling pathways. Cross talk between GR and AR has also been reported in adipocytes and pancreatic  $\beta$  cells (Harada et al., 2015; Hartig et al., 2012). While not much is known about GR-AR crosstalk, we can use the H295R cell model to explore whether activation of GR or AR exogenously affects the expression of the other nuclear receptor respectively. One main advantage of better understanding receptor action is in being able to identify specific pathways that are targetable for drug development.

## 6. Conclusion

One of the key challenges in investigating the genetic basis of complex genetic diseases has been identifying causal variants and causal genes of a trait or disease. In this dissertation, I have advanced studies of the role of non-coding genetic variation in contributing to human diseases. I have developed approaches that allow for identifying regulatory elements and measuring the regulatory effects of genetic variants on gene expression. Specifically, we used the methods to test for the role of regulatory elements in contributing to polycystic ovary syndrome and androgen synthesis. I have described several complementary approaches that have increased our understanding of regulatory element function across PCOS genetic risk loci using cell models relevant to PCOS pathogenesis.

In this dissertation, I have demonstrated scaling up of experiments to work with cell models relevant to PCOS which necessitates several optimization efforts. This method also underscores the need for complementary research in developing better cell models for understanding human diseases. This work also demonstrates the challenges in following up on genetic associations. First, there remains a large task in identifying the molecular function of functional genetic variants which is complicated by availability of tissue types and experimental techniques. Second, for complex diseases like PCOS, there likely interactions between several pathways that are disrupted by genetic variants, and there remains a lot of work to be done to set up appropriate methods to understand how variants act collectively at the level of the organism. Third, the ability to discover genetic-variant-to-trait associations have been increasing with increased access to technology, decreasing cost of genotyping and better genotyping arrays that include more variants; while the approaches to functionally identify the genetic mechanisms underlying the associations is increasing at a slower pace. The ability to implement such high

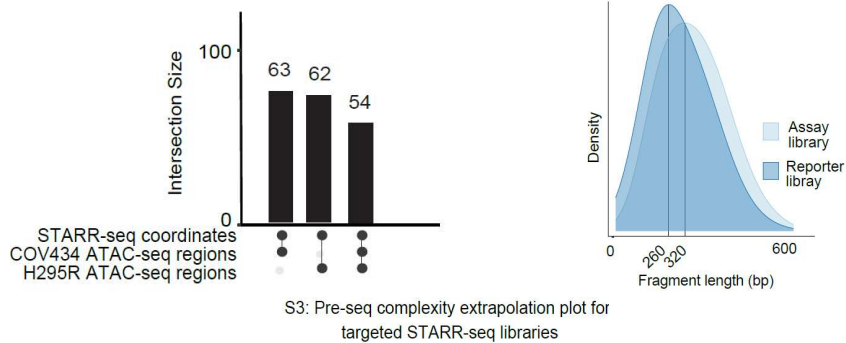
throughput assays as described in this dissertation across multiple labs will greatly aid in dissecting the non-coding regulatory variants. By combining advancements in technology with organismal- and cellular-biology expertise of different research groups to implement the methods in diverse tissue types, we can advance our understanding of diseases.

Collectively, our findings highlight the importance of GWAS follow-up using a combination of experimental and statistical approaches to dissect the molecular mechanisms that can contribute to disease risk. More broadly, continued advancement of studies such as those described in this dissertation will bring us closer to making these types of experiments routine. This advancement will facilitate more mechanistic discoveries related to diseases, and will place a premium on domain-specific knowledge about gene function and organismal biology.

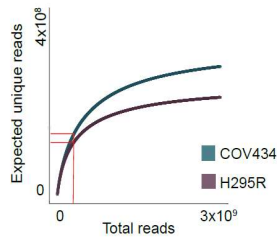
# Appendix A : Gene regulatory mechanisms in PCOS GWAS loci

S1: Selected STARR-seq regions that have an accessible chromatin regions (ATAC-seq)

S2: Distribution of lengths of fragments in assay and reporter libraries in targeted STARR-seq experiments



S3: Pre-seq complexity extrapolation plot for targeted STARR-seq libraries

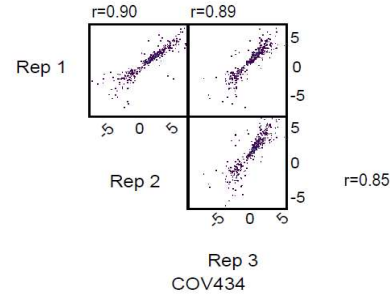
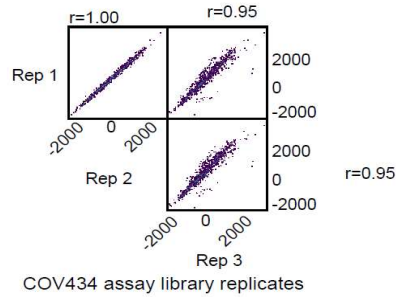
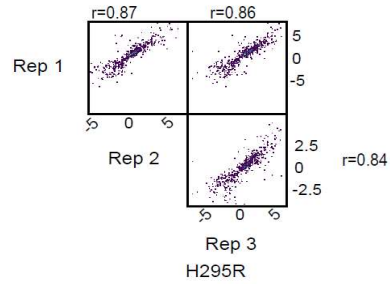
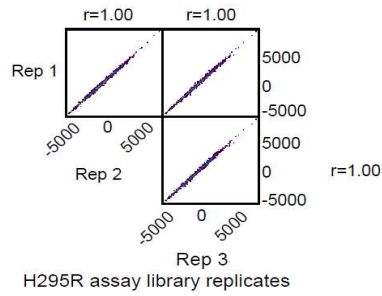


**S1:** Number of chromatin accessible regions within the selected coordinates for targeted STARR-seq.

**S2:** of the fragments (bp) present in the assay (light blue) and reporter (dark blue) libraries.

**S3:** The number of expected unique fragments of the reporter libraries for H295R (purple) and COV434 cells (teal) estimated using preseq. At  $3 \times 10^9$  total reads, expected unique reads would be  $2.8 \times 10^8$  for H295R and  $3.2 \times 10^8$  for COV434 respectively.

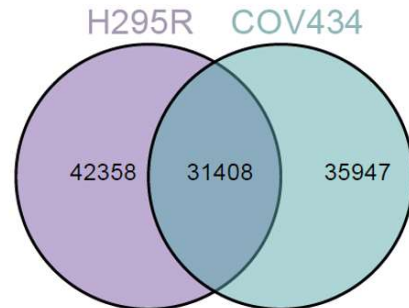
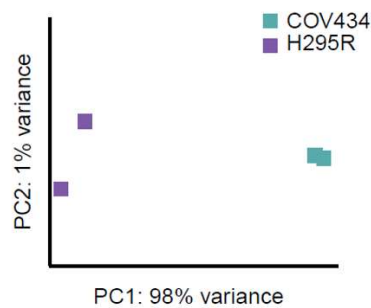
S4: Pearson correlation for STARR-seq assay libraries S5: Pearson correlation of log2 (fold change) of reporter libraries to assay libraries



**S4:** Correlation across assay libraries for STARR-seq: Scatterplots comparing log10-transformed CRADLE corrected read counts across chromatin accessible regions across three replicates of STARR-seq assay libraries.

**S5:** Correlation across assay and reporter libraries for STARR-seq: Scatterplots comparing log2-transformed reporter libraries over assay libraries as log2-fold changes for H295R cells (top) and COV434 cells (bottom) after CRADLE correction.

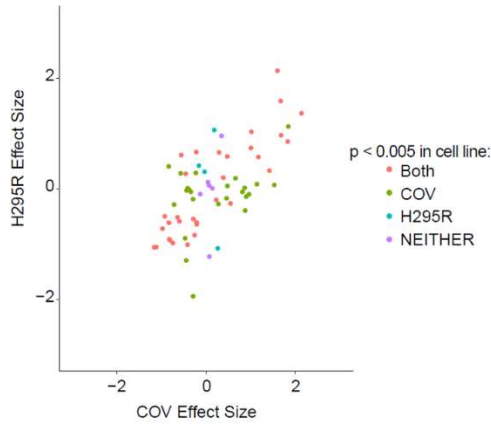
S6: PCA of ATAC-seq libraries for COV434 and H295R S7: Accessible chromatin regions in H295R and COV434 using ATAC-seq



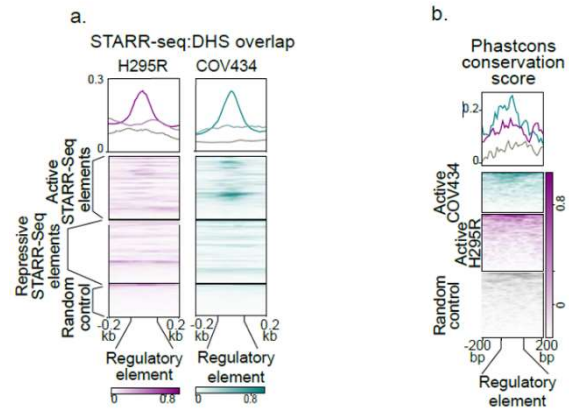
**S6:** Principal component analysis for ATAC-seq libraries of COV434 and H295R cells.

**S7:** Venn diagram of the overlap between Tn5 accessible chromatin regions in COV434 (teal) and H295R cells (purple).

S8: Effect sizes across open chromatin regions



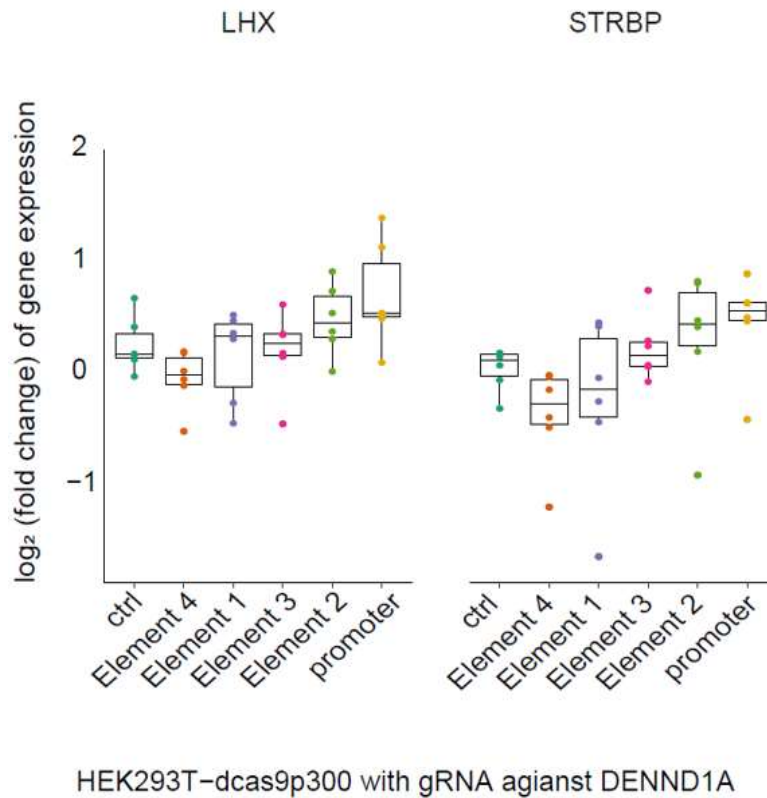
S9: Active STARR-seq regulatory elements correspond to regulatory activity in multiple cell types and have increased conservation score



**S8:** Scatter plot of effect sizes measured using DESeq for both COV434 and H295R cells. The regions were selected as chromatin accessible regions that also were called in CRADLE as a peak and had regulatory activity measured in at least one cell line.

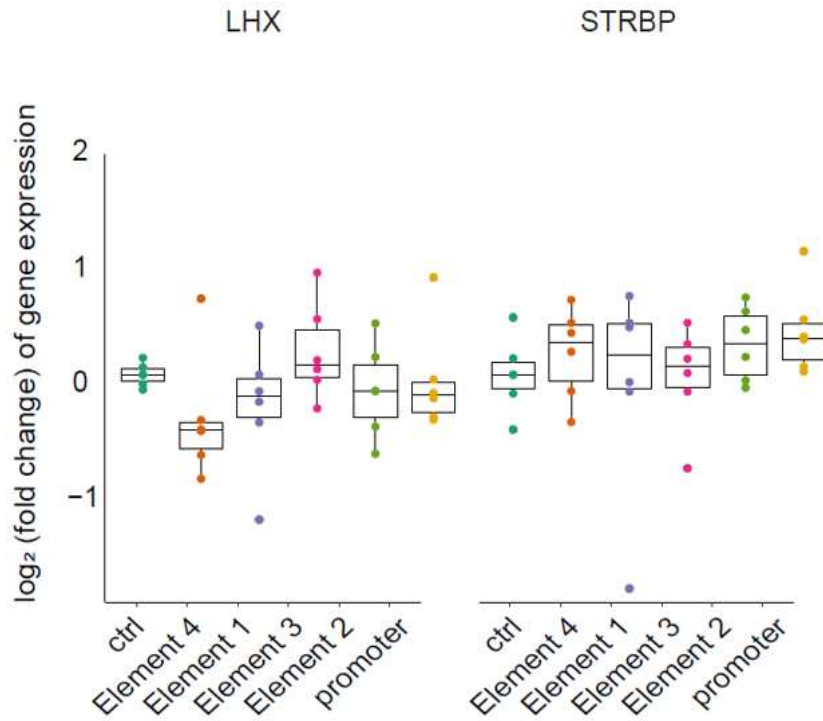
**S9:** (a). Aggregate profile plots of chromatin accessibility based on ENCODE DNaseI Hypersensitive sites (DHS) centred on the active candidate regulatory elements across 400 bp windows for both cell lines (H295R in purple, COV434 in teal). Control regions (grey) are randomly generated genomic regions that are chromosome-, length- and GC-matched to the STARR-seq elements. (b). Aggregate profile plots of conservation score (phastcons) across active STARR-seq regulatory elements in H295R (purple) and COV434 (teal). The control regions are in grey and are random GC- and length-matched to the STARR-seq regions.

S10: Gene expression for STRBP and LHX (GAPDH control)  
via RT-qPCR in HEK293T-dCas9-p300



**S10:** Log-fold change of LHX expression (left) and STRBP expression (right) with GAPDH as control for HEK293T cells stably expressing dCas9-p300. A set of 3 non-targeting control guides were designed to not target any part of the human genome as control cell population.

S11: Gene expression for STRBP and LHX (GAPDH control)  
via RT-qPCR in H295R-dCas9-p300



H295R-dcas9p300 with gRNA against DENND1A

**S11:** Log-fold change of LHX expression (left) and STRBP expression (right) with GAPDH as control for H295R cells stably expressing dCas9-p300. A set of 3 non-targeting control guides were designed to not target any part of the human genome as control cell population.

**Table S1:** Genomic coordinates for BAC STARR-seq (hg38).

<b>BAC/fosmid ID</b>	<b>PCOS GWAS locus</b>	<b>chr</b>	<b>Start</b>	<b>End</b>
RP11-352H17	LHCGR	chr2	48567960	48773736
RP11-825O6	FSHR	chr2	48947410	49120420
RP11-17L5	THADA	chr2	43201351	43356486
RP11-831F19	THADA	chr2	43348866	43512262
RP11-677O13	RAD50/IRF1	chr5	132440666	132624332
RP11-118D21	GATA4/NEIL2	chr8	11702874	11884058
RP11-959L16	DENND1A	chr9	123670418	123860422
RP11-885N4	AOPEP	chr9	94874999	95023646
RP11 640G3	YAP1	chr11	102141887	102311386
RP11-1023M3	RAB5B/SUOX	chr12	55873705	56034531
RP11-1030M2	KRR1	chr12	75372807	75564639
RP11-462A13	HMGA2	chr12	65796499	65964907
RP11-261N13	TOX3	chr16	52296365	52442860
RP11-159F20	SUMO1P1	chr20	53800706	53974652
CH17-438B10	FSHB	chr11	30024172	30225507
CH17-267K24	FSHB	chr11	30272821	30482841
WI2-1214M9	FSHB	chr11	30239629	30280497
WI2-2586K13	FSHB	chr11	30211547	30250398

**Table S2:** Regions targeted for epigenomic perturbation using CRISPR-dCas9-p300.

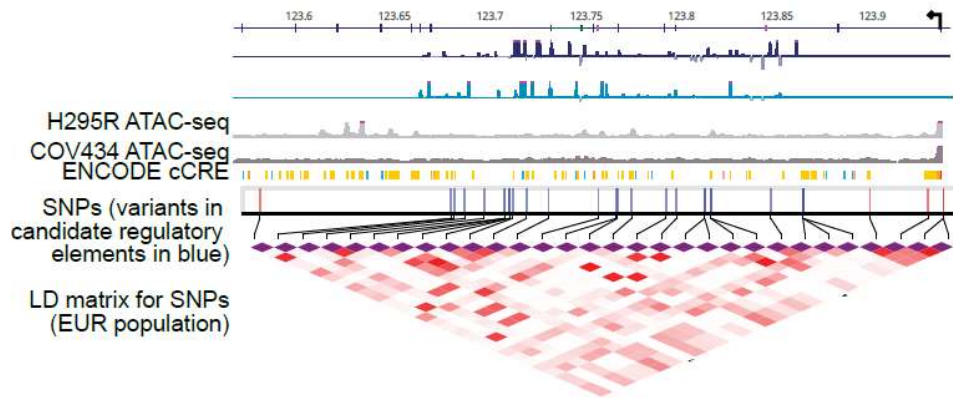
	<b>Genomic coordinates (hg38)</b>		
<b>gRNA pool name</b>	<b>chromosome</b>	<b>start</b>	<b>end</b>
Element 1	chr9	123704717	123705190
Element 2	chr9	123724455	123724767
Element 3	chr9	123752897	123753207
Element 4	chr9	123770099	123770523
Promoter	chr9	123929973	123930568

**Table S3:** Oligonucleotides and primers used.

Primer name	Sequence (5'->3')
SS_Adaptor_1	ACACTCTTCCCTACACGACGCTCTCCGATCT
SS_Adaptor_2	[Phos]GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
TS2SS-F	TAGAGCATGCACCGGACACACTCTTCCCTACACGACGCTCTCCGATCT
TS2SS-R	GGCCGAATTCGTCGAGTGACTGGAGTTCAGACGTGTGCTCTCCGATCT
208-F	AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTCCGATCT
Index7	CAAGCAGAAGACGGCATAACGAGATNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTCCGATC
SSRT-UMI	CAAGCAGAAGACGGCATAACGAGATNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTCCG*A*T*C
Index-PCR	CAAGCAGAAGACGGCATAACGA*G*A*T
PostSS-Index-5	AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNACACTCTTCCCTACAC*G*A*C
ssds-F	TAACTTGAAAGTATTTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG
ssds-R	GTTGATAACGGACTAGCCTTATTTAAACTTGCTATGCTGTTTCCAGCATAGCTCTTAAAC
ATAC-universal	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
ATAC-barcode	CAAGCAGAAGACGGCATAACGAGATNNNNNNNNNGTCTCGTGGGCTCGGAGATGT

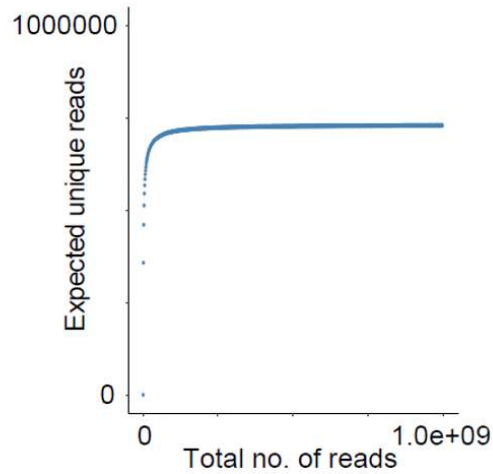
## Appendix B : Regulatory variants in DENND1A locus

S1 : Genome browser map of regulatory elements identified in DENND1A locus including LD map



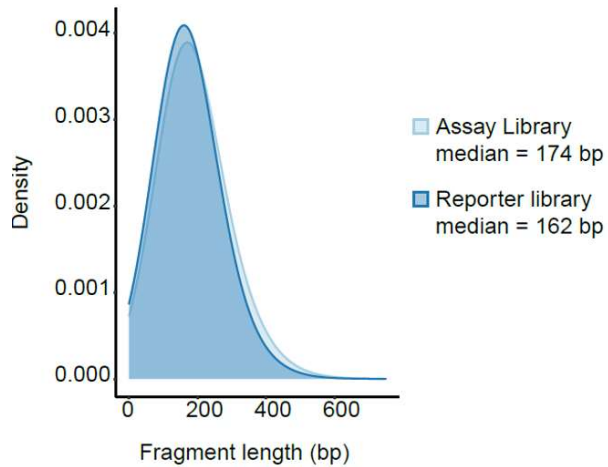
**S1** : Genome browser overview of the regulatory elements identified in DENND1A locus. Tracks shown : H295R STARR-seq (purple), COV434 STARR-seq (teal), H295R and COV434 ATAC-seq (grey), ENCODE cCRE : yellow and orange : distal and prximal Enhancer like elements, ref : promoter like elements. SNPs identified within the candidate regulatory elements are shown in blue vertical bars. The linkage disequilibrium heatmap was generated using EUR data from the 1000 Genomes Project

S2: Pre-seq complexity extrapolation plot for DENND1A-enriched STARR-seq



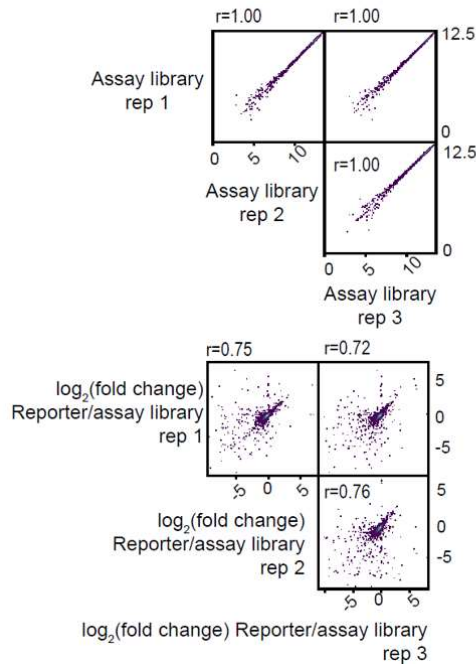
S2: DENND1A enriched STARR-seq library complexity: The number of expected unique fragments of the reporter libraries in H295R cells estimated using preseq. At  $1 \times 10^8$  total reads, expected unique reads would be  $\sim 0.7 \times 10^6$  unique reads averaging at 140x coverage per base pair.

S3: Distribution of lengths of fragments in assay and reporter libraries in DENND1A-enriched STARR-seq experiments



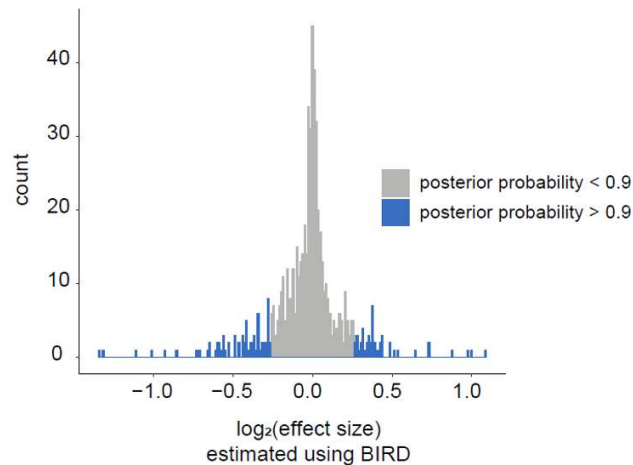
S3: Fragment length distribution in DENND1A-enriched STARR-seq: Length of the fragments (bp) present in the assay (light blue) and reporter (dark blue) libraries.

**S4:** Assay and log(fold change) of reporter library correlation plot for DENND1A-enriched STARR-seq



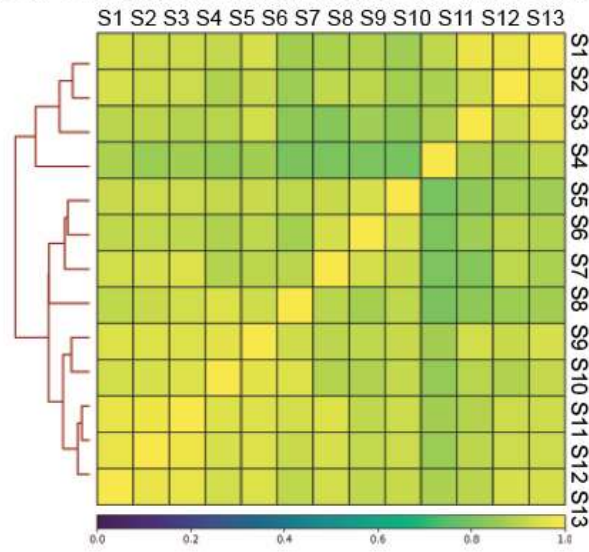
**S4:** Pearson correlation for the assay (input) libraries in H295R cells (top). Scatterplots comparing log<sub>10</sub>-transformed read counts across chromatin accessible regions across three replicates of assay libraries. Scatterplots comparing log<sub>2</sub>-transformed reporter libraries over assay libraries in H295R cells (bottom).

**S5:** Distribution of effect sizes calculated for variants in DENND1A locus



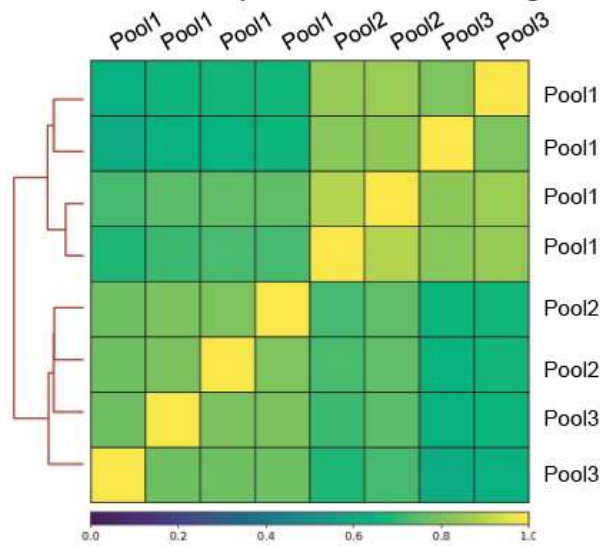
**S5:** Distribution of log<sub>2</sub>-transformed effect sizes for variants in DENND1A locus estimated using BIRD model: The variants in blue have a posterior probability of being a regulatory variant > 0.9.

**S6: Pearson correlation of input libraries for whole genome STARR-seq**



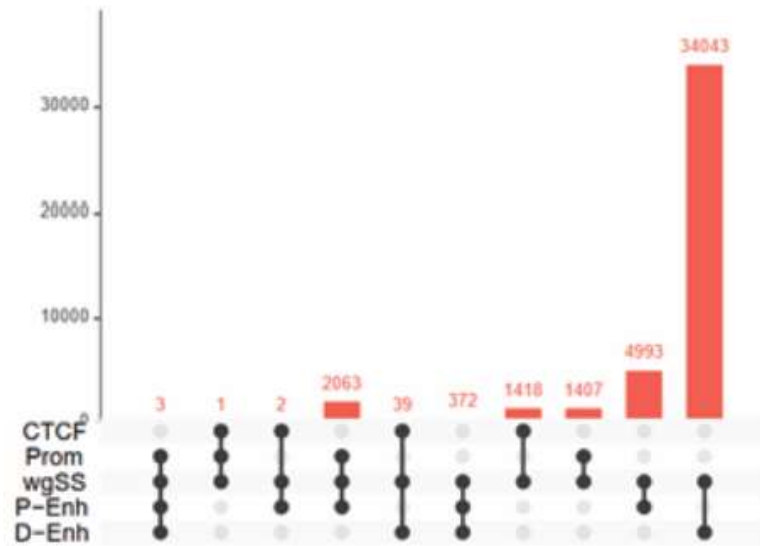
**S6:** Correlation coefficient across 13 whole genome STARR-seq assay libraries

**S7: Pearson correlation of output libraries for whole genome STARR-seq**



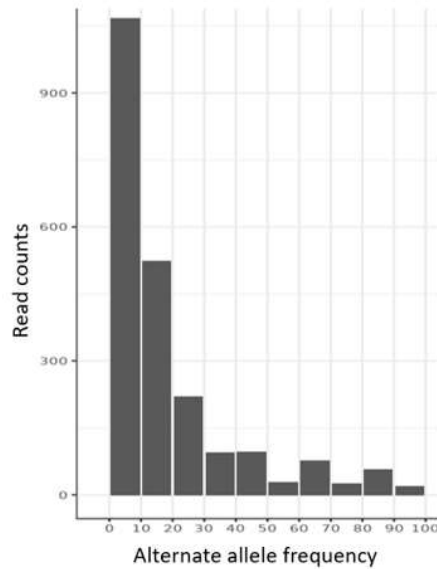
**S7:** Correlation coefficient across pooled whole genome STARR-seq reporter libraries

**S8: Upset plot of the overlap between whole genome candidate regulatory elements and ENCODE cCRE**

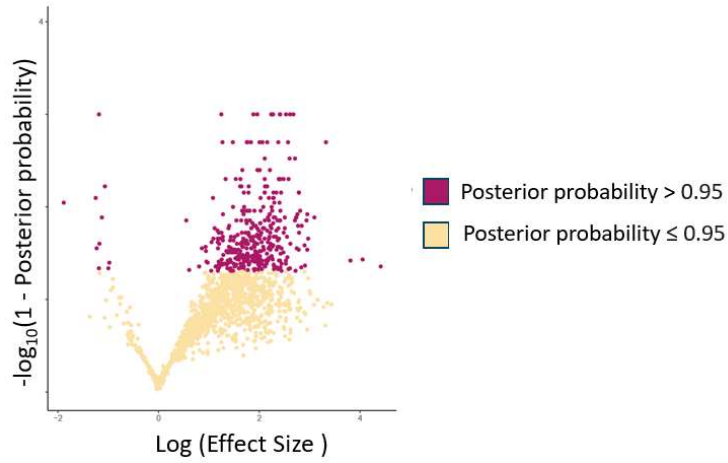


**S8: Overlap between candidate regulatory elements identified by STARR-seq and ENCODE cCRE**

**S9: Distribution of allele frequencies in DENND1A locus**

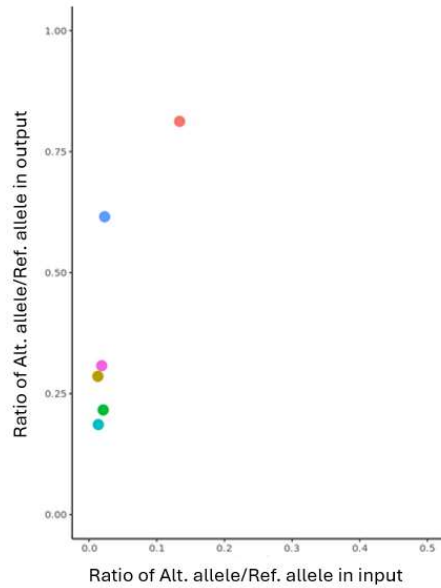


**S9: Distribution of allele-frequencies of variants in thirteen PCOS case control cohorts assay libraries for whole genome STARR-seq**



**S10:** Scatterplot of log(effect sizes) of the variants in DENND1A locus measured using BIRD and whole genome STARR-seq

**S11 :** Regulatory variants that overlap regulatory elements in DENND1A



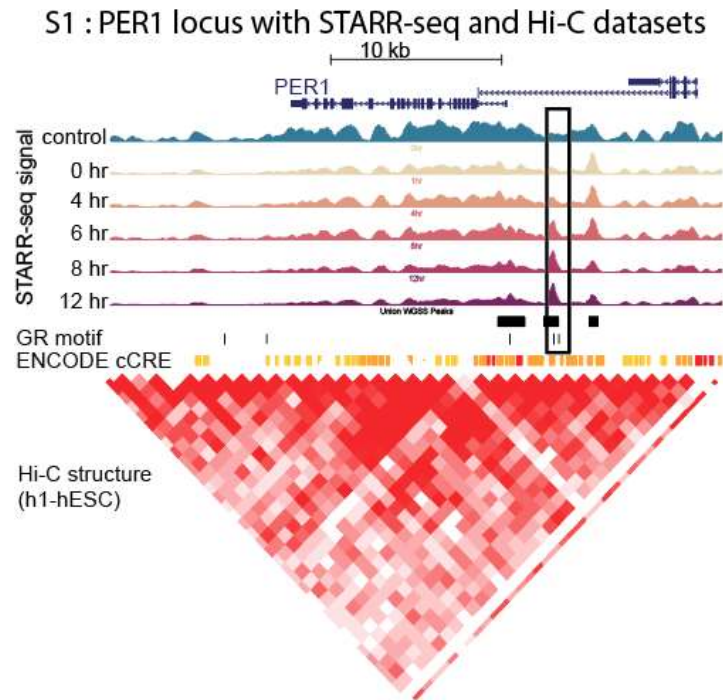
**S11:** Scatterplot ratio of readcounts mapping to alternate allele compared to reference allele for the assay library (x axis) and reporter library (y axis). Each point is a regulatory variant

overlapping regulatory element measured by STARR-seq. The SNPs are genotyped from patient samples, and not all of them have a common SNP rsID.

**Table S1:** Samples for capture STARR-seq obtained from the 1000 Genomes Project

<b>1KGP names</b>	<b>Ancestry</b>
NA12616	CEPH
NA06989	CEPH
NA12832	CEPH
HG00410	CHS
HG00452	CHS

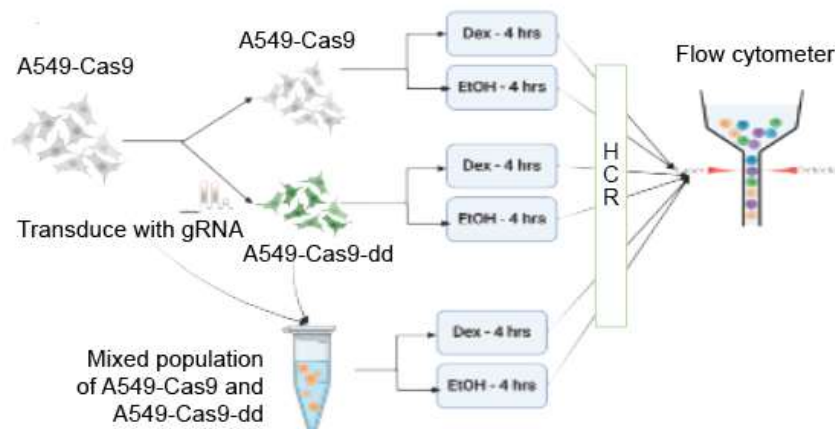
## Appendix C : Probe-based methods to quantify gene expression



Adapted from Johnson et al., 2018

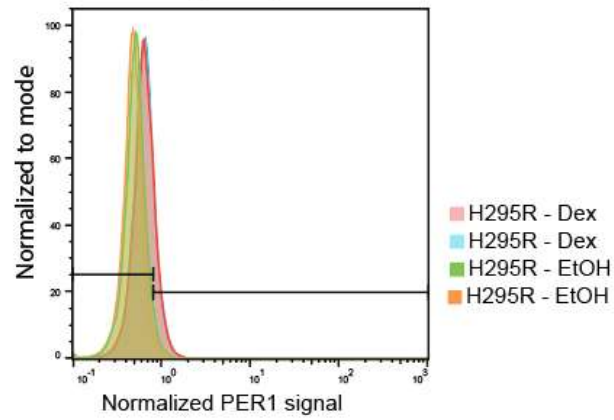
**S1** : Genome browser screenshot of the PER1 locus in A549 cells. Adapted from Johnson et al., 2018. The region highlighted in black indicates the GR binding site edited for PER1 ablation experiments.

### S2 : Generation of mixed A549-cas9 for HCR Flow-FISH



**S2**: Overview of the method to test HCR Flow-FISH in a mixed population of A549-dCas9 cells with or without the putative PER1 binding site edited out.

### S3: PER1 signal upon Dex stimulation in H295R



**S3:** PER1 signal in H295R measured by STARR-seq. Samples were treated with Dex or ethanol (EtOH) for four hours for stimulation response measurement.

## References

- Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Ai, R., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., ... Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818), 699–710. <https://doi.org/10.1038/s41586-020-2493-4>
- Abell, N. S., DeGorter, M. K., Gloude-mans, M. J., Greenwald, E., Smith, K. S., He, Z. & Montgomery, S. B. (2022). Multiple causal variants underlie genetic associations in humans. *Science*, 375(6586), 1247–1254. <https://doi.org/10.1126/science.abj5117>
- Albert, F. W. & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4), 197–212. <https://doi.org/10.1038/nrg3891>
- Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106. <https://doi.org/10.1186/gb-2010-11-10-r106>
- Anttonen, M., Ketola, I., Parviainen, H., Pusa, A.-K. & Heikinheimo, M. (2003). FOG-2 and GATA-4 Are Coexpressed in the Mouse Ovary and Can Modulate Mullerian-Inhibiting Substance Expression. *Biology of Reproduction*, 68(4), 1333–1340. <https://doi.org/10.1095/biolreprod.102.008599>
- Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M. & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, 339(6123), 1074–1077. <https://doi.org/10.1126/science.1232542>
- Avenant, C., Ronacher, K., Stubrud, E., Louw, A. & Hapgood, J. P. (2010). Role of ligand-dependent GR phosphorylation and half-life in determination of ligand-specific transcriptional activity. *Molecular and Cellular Endocrinology*, 327(1–2), 72–88. <https://doi.org/10.1016/j.mce.2010.06.007>
- Azziz, R., Black, V., Hines, G. A., Fox, L. M. & Boots, L. R. (1998). Adrenal Androgen Excess in the Polycystic Ovary Syndrome: Sensitivity and Responsivity of the Hypothalamic-Pituitary-Adrenal Axis. *The Journal of Clinical Endocrinology & Metabolism*, 83(7), 2317–2323. <https://doi.org/10.1210/jcem.83.7.4948>
- Azziz, Ricardo, Dumesic, D. A. & Goodarzi, M. O. (2011). Polycystic ovary syndrome: an ancient disorder? *Fertility and Sterility*, 95(5), 1544–1548. <https://doi.org/10.1016/j.fertnstert.2010.09.032>

- Azziz, Ricardo, Marin, C., Hoq, L., Badamgarav, E. & Song, P. (2005). Health Care-Related Economic Burden of the Polycystic Ovary Syndrome during the Reproductive Life Span. *The Journal of Clinical Endocrinology & Metabolism*, 90(8), 4650–4658. <https://doi.org/10.1210/jc.2005-0628>
- Bannister, A. J. & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3), 381–395. <https://doi.org/10.1038/cr.2011.22>
- Baranello, L., Kouzine, F., Sanford, S. & Levens, D. (2016). ChIP bias as a function of cross-linking time. *Chromosome Research*, 24(2), 175–181. <https://doi.org/10.1007/s10577-015-9509-1>
- Behera, V., Evans, P., Face, C. J., Hamagami, N., Sankaranarayanan, L., Keller, C. A., Giardine, B., Tan, K., Hardison, R. C., Shi, J. & Blobel, G. A. (2018). Exploiting genetic variation to uncover rules of transcription factor binding and chromatin accessibility. *Nature Communications*, 9(1), 782. <https://doi.org/10.1038/s41467-018-03082-6>
- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. (2018). Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nature Ecology & Evolution*, 2(1), 152–163. <https://doi.org/10.1038/s41559-017-0377-2>
- Black, J. B., McCutcheon, S. R., Dube, S., Barrera, A., Klann, T. S., Rice, G. A., Adkar, S. S., Soderling, S. H., Reddy, T. E. & Gersbach, C. A. (2020). Master Regulators and Cofactors of Human Neuronal Cell Fate Specification Identified by CRISPR Gene Activation Screens. *Cell Reports*, 33(9), 108460. <https://doi.org/10.1016/j.celrep.2020.108460>
- Bohaczuk, S. C., Cassin, J., Slaiwa, T. I., Thackray, V. G. & Mellon, P. L. (2021). Distal Enhancer Potentiates Activin- and GnRH-Induced Transcription of FSHB. *Endocrinology*, 162(7). <https://doi.org/10.1210/endocr/bqab069>
- Bohaczuk, S. C., Thackray, V. G., Shen, J., Skowronska-Krawczyk, D. & Mellon, P. L. (2020). FSHB Transcription is Regulated by a Novel 5' Distal Enhancer with a Fertility-Associated Single Nucleotide Polymorphism. *Endocrinology*. <https://doi.org/10.1210/endocr/bqaa181>
- Boudreaux, M. Y., Talbott, E. O., Kip, K. E., Brooks, M. M. & Witchel, S. F. (2006). Risk of T2DM and impaired fasting glucose among pcos subjects: Results of an 8-year follow-up. *Current Diabetes Reports*, 6(1), 77–83. <https://doi.org/10.1007/s11892-006-0056-1>
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7), 1177–1186. <https://doi.org/10.1016/j.cell.2017.05.038>
- Brakta, S., Lizneva, D., Mykhalchenko, K., Imam, A., Walker, W., Diamond, M. P. & Azziz, R. (2017). Perspectives on Polycystic Ovary Syndrome: Is Polycystic Ovary Syndrome Research Underfunded? *The Journal of Clinical Endocrinology & Metabolism*, 102(12), 4421–4427. <https://doi.org/10.1210/jc.2017-01415>

- Brandt, M. M., Meddens, C. A., Louzao-Martinez, L., Dungen, N. A. M. van den, Lansu, N. R., Nieuwenhuis, E. E. S., Duncker, D. J., Verhaar, M. C., Joles, J. A., Mokry, M. & Cheng, C. (2018). Chromatin Conformation Links Distal Target Genes to CKD Loci. *Journal of the American Society of Nephrology*, 29(2), 462–476. <https://doi.org/10.1681/asn.2016080875>
- Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1), 21.29.1-21.29.9. <https://doi.org/10.1002/0471142727.mb2129s109>
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Burton, P. R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., ... Worthington, J. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. <https://doi.org/10.1038/nature05911>
- Bush, W. S. & Haines, J. (2010). Overview of Linkage Analysis in Complex Traits. *Current Protocols in Human Genetics*, 64(1), 1.9.1-1.9.18. <https://doi.org/10.1002/0471142905.hg0109s64>
- Canver, M. C., Bauer, D. E. & Orkin, S. H. (2017). Functional interrogation of non-coding DNA through CRISPR genome editing. *Methods*, 121, 118–129. <https://doi.org/10.1016/j.ymeth.2017.03.008>
- Carmona, R., Arroyo, M., Jiménez-Quesada, M. J., Seoane, P., Zafra, A., Larrosa, R., Alché, J. de D. & Claros, M. G. (2017). Automated identification of reference genes based on RNA-seq data. *BioMedical Engineering OnLine*, 16(Suppl 1), 65. <https://doi.org/10.1186/s12938-017-0356-5>
- Castaldi, P. J., Guo, F., Qiao, D., Du, F., Naing, Z. Z. C., Li, Y., Pham, B., Mikkelsen, T. S., Cho, M. H., Silverman, E. K. & Zhou, X. (2018). Identification of Functional Variants in the FAM13A Chronic Obstructive Pulmonary Disease Genome-Wide Association Study Locus by Massively Parallel Reporter Assays. *American Journal of Respiratory and Critical Care Medicine*, 199(1), 52–61. <https://doi.org/10.1164/rccm.201802-0337oc>
- Censin, J. C., Bovijn, J., Holmes, M. V. & Lindgren, C. M. (2021). Colocalization analysis of polycystic ovary syndrome to identify potential disease-mediating genes and proteins. *European Journal of Human Genetics*, 1–9. <https://doi.org/10.1038/s41431-021-00835-8>
- Chang, S. & Dunaif, A. (2021). Diagnosis of Polycystic Ovary Syndrome Which Criteria to Use and When? *Endocrinology and Metabolism Clinics of North America*, 50(1), 11–23. <https://doi.org/10.1016/j.ecl.2020.10.002>
- Chen, L.-F., Lin, Y. T., Gallegos, D. A., Hazlett, M. F., Gómez-Schiavon, M., Yang, M. G., Kalmeta, B., Zhou, A. S., Holtzman, L., Gersbach, C. A., Grandl, J., Buchler, N. E. & West, A. E. (2019). Enhancer Histone Acetylation Modulates Transcriptional Bursting Dynamics of

Neuronal Activity-Inducible Genes. *Cell Reports*, 26(5), 1174-1188.e5.  
<https://doi.org/10.1016/j.celrep.2019.01.032>

- Chen, P. B., Fiaux, P. C., Zhang, K., Li, B., Kubo, N., Jiang, S., Hu, R., Roohofada, E., Wu, S., Wang, M., Wang, W., McVicker, G., Mischel, P. S. & Ren, B. (2022). Systematic discovery and functional dissection of enhancers needed for cancer cell fitness and proliferation. *Cell Reports*, 41(6), 111630. <https://doi.org/10.1016/j.celrep.2022.111630>
- Chen, X.-F., Zhu, D.-L., Yang, M., Hu, W.-X., Duan, Y.-Y., Lu, B.-J., Rong, Y., Dong, S.-S., Hao, R.-H., Chen, J.-B., Chen, Y.-X., Yao, S., Thynn, H. N., Guo, Y. & Yang, T.-L. (2018). An Osteoporosis Risk SNP at 1p36.12 Acts as an Allele-Specific Enhancer to Modulate LINC00339 Expression via Long-Range Loop Formation. *The American Journal of Human Genetics*, 102(5), 776–793. <https://doi.org/10.1016/j.ajhg.2018.03.001>
- Chen, Z.-J., Zhao, H., He, L., Shi, Y., Qin, Y., Shi, Y., Li, Z., You, L., Zhao, J., Liu, J., Liang, X., Zhao, X., Zhao, J., Sun, Y., Zhang, B., Jiang, H., Zhao, D., Bian, Y., Gao, X., ... Zhao, Y. (2011). Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nature Genetics*, 43(1), 55. <https://doi.org/10.1038/ng.732>
- Choi, H. M., Beck, V. A. & Pierce, N. A. (2014). Next-Generation in Situ Hybridization Chain Reaction: Higher Gain, Lower Cost, Greater Durability. *ACS Nano*, 8(5), 4284–4294. <https://doi.org/10.1021/nn405717p>
- Choi, H. M. T., Calvert, C. R., Husain, N., Huss, D., Barsi, J. C., Deverman, B. E., Hunter, R. C., Kato, M., Lee, S. M., Abelin, A. C. T., Rosenthal, A. Z., Akbari, O. S., Li, Y., Hay, B. A., Sternberg, P. W., Patterson, P. H., Davidson, E. H., Mazmanian, S. K., Prober, D. A., ... Pierce, N. A. (2016). Mapping a multiplexed zoo of mRNA expression. *Development*, 143(19), 3632–3637. <https://doi.org/10.1242/dev.140137>
- Choi, H. M. T., Schwarzkopf, M., Fornace, M. E., Acharya, A., Artavanis, G., Stegmaier, J., Cunha, A. & Pierce, N. A. (2018). Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development*, 145(12), dev165753. <https://doi.org/10.1242/dev.165753>
- Chowdhury, I., Branch, A., Mehrabi, S., Ford, B. D. & Thompson, W. E. (2017). Gonadotropin-Dependent Neuregulin-1 Signaling Regulates Female Rat Ovarian Granulosa Cell Survival. *Endocrinology*, 158(10), 3647–3660. <https://doi.org/10.1210/en.2017-00065>
- Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306(5696), 636–640. <https://doi.org/10.1126/science.1105136>
- Consortium, Gte. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509), 1318–1330. <https://doi.org/10.1126/science.aaz1776>

- Consortium, I. H. G. S., Research:, W. I. for B. R., Center for Genome, Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., ... Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. <https://doi.org/10.1038/35057062>
- Consortium, T. F., Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., ... Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. <https://doi.org/10.1038/nature12787>
- Consortium, U., Iotchkova, V., Ritchie, G. R., Geihs, M., Morganella, S., Min, J. L., Walter, K., Timpson, N., Dunham, I., Birney, E. & Soranzo, N. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature Genetics*, *51*(2), 343–353. <https://doi.org/10.1038/s41588-018-0322-6>
- Cooper, G. M. (2000). *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates. <https://www.ncbi.nlm.nih.gov/books/NBK9839/>
- Cooper, Y. A., Teyssier, N., Dräger, N. M., Guo, Q., Davis, J. E., Sattler, S. M., Yang, Z., Patel, A., Wu, S., Kosuri, S., Coppola, G., Kampmann, M. & Geschwind, D. H. (2022). Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science*, *377*(6608), eabi8654. <https://doi.org/10.1126/science.abi8654>
- Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., Kathiria, A., Cho, S. W., Mumbach, M. R., Carter, A. C., Kasowski, M., Orloff, L. A., Risca, V. I., Kundaje, A., Khavari, P. A., ... Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, *14*(10), 959–962. <https://doi.org/10.1038/nmeth.4396>
- Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal·lari, R., Lupien, M., Markowitz, S. & Scacheri, P. C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research*, *24*(1), 1–13. <https://doi.org/10.1101/gr.164079.113>
- Coviello, A. D., Zhuang, W. V., Lunetta, K. L., Bhasin, S., Ulloor, J., Zhang, A., Karasik, D., Kiel, D. P., Vasan, R. S. & Murabito, J. M. (2011). Circulating Testosterone and SHBG Concentrations Are Heritable in Women: The Framingham Heart Study. *The Journal of Clinical Endocrinology & Metabolism*, *96*(9), E1491–E1495. <https://doi.org/10.1210/jc.2011-0050>
- Currin, K. W., Erdos, M. R., Narisu, N., Rai, V., Vadlamudi, S., Perrin, H. J., Idol, J. R., Yan, T., Albanus, R. D., Broadaway, K. A., Etheridge, A. S., Bonnycastle, L. L., Orchard, P., Didion, J. P., Chaudhry, A. S., Program, N. C. S., Barnabas, B. B., Black, S., Bouffard, G. G., ...

- Mohlke, K. L. (2021). Genetic effects on liver chromatin accessibility identify disease regulatory variants. *The American Journal of Human Genetics*, *108*(7), 1169–1189. <https://doi.org/10.1016/j.ajhg.2021.05.001>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. <https://doi.org/10.1093/gigascience/giab008>
- Dapas, M. & Dunaif, A. (2022). Deconstructing a Syndrome: Genomic Insights Into PCOS Causal Mechanisms and Classification. *Endocrine Reviews*, *43*(6), 927–965. <https://doi.org/10.1210/endrev/bnac001>
- Dapas, M., Lin, F. T. J., Nadkarni, G. N., Sisk, R., Legro, R. S., Urbanek, M., Hayes, M. G. & Dunaif, A. (2020). Distinct subtypes of polycystic ovary syndrome with novel genetic associations: An unsupervised, phenotypic clustering analysis. *PLoS Medicine*, *17*(6), e1003132. <https://doi.org/10.1371/journal.pmed.1003132>
- Dapas, M., Sisk, R., Legro, R. S., Urbanek, M., Dunaif, A. & Hayes, M. (2019). Family-based quantitative trait meta-analysis implicates rare noncoding variants in DENND1A in polycystic ovary syndrome. *The Journal of Clinical Endocrinology and Metabolism*. <https://doi.org/10.1210/jc.2018-02496>
- Darnell, J. E. (2002). Transcription factors as targets for cancer therapy. *Nature Reviews Cancer*, *2*(10), 740–749. <https://doi.org/10.1038/nrc906>
- Das, S., Abecasis, G. R. & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*(1), 1–24. <https://doi.org/10.1146/annurev-genom-083117-021602>
- Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. <https://doi.org/10.1038/ng.3656>
- Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D. & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, *14*(3), 297–301. <https://doi.org/10.1038/nmeth.4177>
- Day, F., Karaderi, T., Jones, M. R., Meun, C., He, C., Drong, A., Kraft, P., Lin, N., Huang, H., Broer, L., Magi, R., Saxena, R., Laisk, T., Urbanek, M., Hayes, G. M., Thorleifsson, G., Fernandez-Tajés, J., Mahajan, A., Mullin, B. H., ... Welt, C. K. (2018). Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLOS Genetics*, *14*(12), e1007813. <https://doi.org/10.1371/journal.pgen.1007813>

- Day, F. R., Hinds, D. A., Tung, J. Y., Stolk, L., Styrkarsdottir, U., Saxena, R., Bjornes, A., Broer, L., Dunger, D. B., Halldorsson, B. V., Lawlor, D. A., Laval, G., Mathieson, I., McCardle, W. L., Louwers, Y., Meun, C., Ring, S., Scott, R. A., Sulem, P., ... Perry, J. R. (2015). Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nature Communications*, 6(1), 8464. <https://doi.org/10.1038/ncomms9464>
- Dean, A. (2011). In the loop: long range chromatin interactions and gene regulation. *Briefings in Functional Genomics*, 10(1), 3–10. <https://doi.org/10.1093/bfgp/elq033>
- Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., Leon, S. D., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y. & Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385), 390–394. <https://doi.org/10.1038/nature10808>
- Diamanti-Kandarakis, E. & Dunaif, A. (2012). Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. *Endocrine Reviews*, 33(6), 981–1030. <https://doi.org/10.1210/er.2011-1034>
- Ding, T., Baio, G., Hardiman, P. J., Petersen, I. & Sammon, C. (2016). Diagnosis and management of polycystic ovary syndrome in the UK (2004–2014): a retrospective cohort study. *BMJ Open*, 6(7), e012461. <https://doi.org/10.1136/bmjopen-2016-012461>
- D’Ippolito, A. M., McDowell, I. C., Barrera, A., Hong, L. K., Leichter, S. M., Bartelt, L. C., Vockley, C. M., Majoros, W. H., Safi, A., Song, L., Gersbach, C. A., Crawford, G. E. & Reddy, T. E. (2018). Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids. *Cell Systems*, 7(2), 146-160.e7. <https://doi.org/10.1016/j.cels.2018.06.007>
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376–380. <https://doi.org/10.1038/nature11082>
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., ... Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. <https://doi.org/10.1038/nature11247>
- Durrin, L. K., Mann, R. K., Kayne, P. S. & Grunstein, M. (1991). Yeast histone H4 N-terminal sequence is required for promoter activation in vivo. *Cell*, 65(6), 1023–1031. [https://doi.org/10.1016/0092-8674\(91\)90554-c](https://doi.org/10.1016/0092-8674(91)90554-c)
- Ecker, M., Redpath, G. M. I., Nicovich, P. R. & Rossy, J. (2021). Quantitative visualization of endocytic trafficking through photoactivation of fluorescent proteins. *Molecular Biology of the Cell*, 32(9), 892–902. <https://doi.org/10.1091/mbc.e20-10-0669>

- Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. (2013). Beyond GWASs: Illuminating the Dark Road from Association to Function. *The American Journal of Human Genetics*, 93(5), 779–797. <https://doi.org/10.1016/j.ajhg.2013.10.012>
- Efimenko, E., Padua, M. B., Manuylov, N. L., Fox, S. C., Morse, D. A. & Tevosian, S. G. (2013). The transcription factor GATA4 is required for follicular development and normal ovarian function. *Developmental Biology*, 381(1), 144–158. <https://doi.org/10.1016/j.ydbio.2013.06.004>
- Engelkamp, D. & Heyningen, V. van. (1996). Transcription factors in disease. *Current Opinion in Genetics & Development*, 6(3), 334–342. [https://doi.org/10.1016/s0959-437x\(96\)80011-6](https://doi.org/10.1016/s0959-437x(96)80011-6)
- Fabbri-Scaliet, H., Werner, R., Guaragna, M. S., Andrade, J. G. R. de, Maciel-Guerra, A. T., Hornig, N. C., Hiort, O., Guerra-Júnior, G. & Mello, M. P. de. (2023). Can Non-Coding NR5A1 Gene Variants Explain Phenotypes of Disorders of Sex Development? *Sexual Development*, 16(4), 252–260. <https://doi.org/10.1159/000524956>
- Franks, S., Webber, L. J., Goh, M., Valentine, A., White, D. M., Conway, G. S., Wiltshire, S. & McCarthy, M. I. (2008). Ovarian Morphology Is a Marker of Heritable Biochemical Traits in Sisters with Polycystic Ovaries. *The Journal of Clinical Endocrinology & Metabolism*, 93(9), 3396–3402. <https://doi.org/10.1210/jc.2008-0369>
- French, J. D. & Edwards, S. L. (2020). The Role of Noncoding Variants in Heritable Disease. *Trends in Genetics*, 36(11), 880–891. <https://doi.org/10.1016/j.tig.2020.07.004>
- Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, 31(5), 782–784. <https://doi.org/10.1093/bioinformatics/btu704>
- Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S. & Engreitz, J. M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, 354(6313), 769–773. <https://doi.org/10.1126/science.aag2445>
- Gallagher, M. D. & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*, 102(5), 717–730. <https://doi.org/10.1016/j.ajhg.2018.04.002>
- Gao, X., Liu, Y., Lv, Y., Huang, T., Lu, G., Liu, H. & Zhao, S. (2020). Role of Androgen Receptor for Reconsidering the “True” Polycystic Ovarian Morphology in PCOS. *Scientific Reports*, 10(1), 8993. <https://doi.org/10.1038/s41598-020-65890-5>
- García, P., Fernández-Hernández, R., Cuadrado, A., Coca, I., Gómez, A., Maqueda, M., Latorre-Pellicer, A., Puisac, B., Ramos, F. J., Sandoval, J., Esteller, M., Mosquera, J. L., Rodríguez, J., Pié, J., Losada, A. & Queralt, E. (2021). Disruption of NIPBL/Scp2 in Cornelia de Lange Syndrome provokes cohesin genome-wide redistribution with an impact in the transcriptome. *Nature Communications*, 12(1), 4551. <https://doi.org/10.1038/s41467-021-24808-z>

- Gates, L. A., Foulds, C. E. & O'Malley, B. W. (2017). Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends in Biochemical Sciences*, 42(12), 977–989. <https://doi.org/10.1016/j.tibs.2017.10.004>
- Gaulton, K. J., Preissl, S. & Ren, B. (2023). Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nature Reviews Genetics*, 24(8), 516–534. <https://doi.org/10.1038/s41576-023-00598-6>
- Geijn, B. van de, McVicker, G., Gilad, Y. & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11), 1061–1063. <https://doi.org/10.1038/nmeth.3582>
- George, R. M., Hahn, K. L., Rawls, A., Viger, R. S. & Wilson-Rawls, J. (2015). Notch signaling represses GATA4-induced expression of genes involved in steroid biosynthesis. *Reproduction*, 150(4), 383–394. <https://doi.org/10.1530/rep-15-0226>
- Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10(5), e1004383. <https://doi.org/10.1371/journal.pgen.1004383>
- Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M. & Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159(3), 647–661. <https://doi.org/10.1016/j.cell.2014.09.029>
- Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Gregorio, E. D., Imperiale, D., Vaala, G., Stamoulis, G., Santoni, F., Atzori, C., Gasparini, L., Ferrera, D., Canale, C., Guipponi, M., Pennacchio, L. A., Antonarakis, S. E., Brussino, A. & Brusco, A. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Human Molecular Genetics*, 24(11), 3143–3154. <https://doi.org/10.1093/hmg/ddv065>
- Gorsic, L. K., Dapas, M., Legro, R. S., Hayes, M. & Urbanek, M. (2019). Functional Genetic Variation in the Anti-Müllerian Hormone Pathway in Women With Polycystic Ovary Syndrome. *The Journal of Clinical Endocrinology and Metabolism*, 104(7), 2855–2874. <https://doi.org/10.1210/jc.2018-02178>
- Govind, A., Obhrai, M. S. & Clayton, R. N. (1999). Polycystic Ovaries Are Inherited as an Autosomal Dominant Trait: Analysis of 29 Polycystic Ovary Syndrome and 10 Control Families. *The Journal of Clinical Endocrinology & Metabolism*, 84(1), 38–43. <https://doi.org/10.1210/jcem.84.1.5382>
- Grampp, S., Platt, J. L., Lauer, V., Salama, R., Kranz, F., Neumann, V. K., Wach, S., Stöhr, C., Hartmann, A., Eckardt, K.-U., Ratcliffe, P. J., Mole, D. R. & Schödel, J. (2016). Genetic

variation at the 8q24.21 renal cancer susceptibility locus affects HIF binding to a MYC enhancer. *Nature Communications*, 7(1), 13183. <https://doi.org/10.1038/ncomms13183>

- Grevet, J. D., Lan, X., Hamagami, N., Edwards, C. R., Sankaranarayanan, L., Ji, X., Bhardwaj, S. K., Face, C. J., Posocco, D. F., Abdulmalik, O., Keller, C. A., Giardine, B., Sidoli, S., Garcia, B. A., Chou, S. T., Liebhaber, S. A., Hardison, R. C., Shi, J. & Blobel, G. A. (2018). Domain-focused CRISPR screen identifies HRI as a fetal hemoglobin regulator in human erythroid cells. *Science*, 361(6399), 285–290. <https://doi.org/10.1126/science.aaa0932>
- Grosse, A., Bartsch, S. & Baniahmad, A. (2012). Androgen receptor-mediated gene repression. *Molecular and Cellular Endocrinology*, 352(1–2), 46–56. <https://doi.org/10.1016/j.mce.2011.06.032>
- Guo, C., Ludvik, A. E., Arlotto, M. E., Hayes, M. G., Armstrong, L. L., Scholtens, D. M., Brown, C. D., Newgard, C. B., Becker, T. C., Layden, B. T., Lowe, W. L. & Reddy, T. E. (2015). Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. *Nature Communications*, 6(1), 6069. <https://doi.org/10.1038/ncomms7069>
- Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nature Reviews Genetics*, 20(9), 520–535. <https://doi.org/10.1038/s41576-019-0144-0>
- Haggard, D. E., Karmaus, A. L., Martin, M. T., Judson, R. S., Setzer, R. W. & Friedman, K. P. (2017). High-Throughput H295R Steroidogenesis Assay: Utility as an Alternative and a Statistical Approach to Characterize Effects on Steroidogenesis. *Toxicological Sciences*, 162(2), 509–534. <https://doi.org/10.1093/toxsci/kfx274>
- Harada, N., Katsuki, T., Takahashi, Y., Masuda, T., Yoshinaga, M., Adachi, T., Izawa, T., Kuwamura, M., Nakano, Y., Yamaji, R. & Inui, H. (2015). Androgen Receptor Silences Thioredoxin-interacting Protein and Competitively Inhibits Glucocorticoid Receptor-Mediated Apoptosis in Pancreatic  $\beta$ -Cells. *Journal of Cellular Biochemistry*, 116(6), 998–1006. <https://doi.org/10.1002/jcb.25054>
- Harris, J. A., Vernon, P. A. & Boomsma, D. I. (1998). The Heritability of Testosterone: A Study of Dutch Adolescent Twins and Their Parents. *Behavior Genetics*, 28(3), 165–171. <https://doi.org/10.1023/a:1021466929053>
- Harris, R. A., Archer, K. J., Goodarzi, M. O., York, T. P., Rogers, J., Dunaif, A., McAllister, J. M. & Strauss, J. F. (2023). Loci on chromosome 12q13.2 encompassing ERBB3, PA2G4 and RAB5B are associated with polycystic ovary syndrome. *Gene*, 852, 147062. <https://doi.org/10.1016/j.gene.2022.147062>
- Hartig, S. M., He, B., Newberg, J. Y., Ochsner, S. A., Loose, D. S., Lanz, R. B., McKenna, N. J., Buehrer, B. M., McGuire, S. E., Marcelli, M. & Mancini, M. A. (2012). Feed-Forward

Inhibition of Androgen Receptor Activity by Glucocorticoid Action in Human Adipocytes. *Chemistry & Biology*, 19(9), 1126–1141. <https://doi.org/10.1016/j.chembiol.2012.07.020>

Hatje, K., Mühlhausen, S., Simm, D. & Kollmar, M. (2019). The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *BioEssays*, 41(11), e1900066. <https://doi.org/10.1002/bies.201900066>

Hayek, S. E., Bitar, L., Hamdar, L. H., Mirza, F. G. & Daoud, G. (2016). Poly Cystic Ovarian Syndrome: An Updated Overview. *Frontiers in Physiology*, 7, 124. <https://doi.org/10.3389/fphys.2016.00124>

Hayes, G. M., Urbanek, M., Ehrmann, D. A., Armstrong, L. L., Lee, J., Sisk, R., Karaderi, T., Barber, T. M., McCarthy, M. I., Franks, S., Lindgren, C. M., Welt, C. K., Diamanti-Kandarakis, E., Panidis, D., Goodarzi, M. O., Azziz, R., Zhang, Y., James, R. G., Olivier, M., ... Dunaif, A. (2015). Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nature Communications*, 6(1), 7502. <https://doi.org/10.1038/ncomms8502>

Heshmatzad, K., Naderi, N., Maleki, M., Abbasi, S., Ghasemi, S., Ashrafi, N., Fazelifar, A. F., Mahdavi, M. & Kalayinia, S. (2023). Role of non-coding variants in cardiovascular disease. *Journal of Cellular and Molecular Medicine*, 27(12), 1621–1636. <https://doi.org/10.1111/jcmm.17762>

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E. & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5), 510–517. <https://doi.org/10.1038/nbt.3199>

Hirschhorn, J. N. & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2), 95–108. <https://doi.org/10.1038/nrg1521>

Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2), 497–508. <https://doi.org/10.1534/genetics.114.167908>

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankaranarayanan, S., Pasaniuc, B. & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, 99(6), 1245–1260. <https://doi.org/10.1016/j.ajhg.2016.10.003>

Hsiung, C. C.-S., Bartman, C. R., Huang, P., Ginart, P., Stonestrom, A. J., Keller, C. A., Face, C., Jahn, K. S., Evans, P., Sankaranarayanan, L., Giardine, B., Hardison, R. C., Raj, A. & Blobel, G. A. (2016). A hyperactive transcriptional state marks genome reactivation at the mitosis–G1 transition. *Genes & Development*, 30(12), 1423–1439. <https://doi.org/10.1101/gad.280859.116>

- Hua, A. B. & Sweasy, J. B. (2023). Functional roles and cancer variants of the bifunctional glycosylase NEIL2. *Environmental and Molecular Mutagenesis*.  
<https://doi.org/10.1002/em.22555>
- Huang, C.-C. F., Lingadahalli, S., Morova, T., Ozturan, D., Hu, E., Yu, I. P. L., Linder, S., Hoogstraat, M., Stelloo, S., Sar, F., Poel, H. van der, Altintas, U. B., Saffarzadeh, M., Bihan, S. L., McConeghy, B., Gokbayrak, B., Feng, F. Y., Gleave, M. E., Bergman, A. M., ... Lack, N. A. (2021). Functional mapping of androgen receptor enhancer activity. *Genome Biology*, 22(1), 149. <https://doi.org/10.1186/s13059-021-02339-6>
- Huang, T. & Hu, F. B. (2015). Gene-environment interactions and obesity: recent developments and future directions. *BMC Medical Genomics*, 8(Suppl 1), S2. <https://doi.org/10.1186/1755-8794-8-s1-s2>
- Huisinga, K. L., Brower-Toland, B. & Elgin, S. C. R. (2006). The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, 115(2), 110–122.  
<https://doi.org/10.1007/s00412-006-0052-x>
- Inoue, F. & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3), 159–164. <https://doi.org/10.1016/j.ygeno.2015.06.005>
- Institute, B. (2019). *Picard Toolkit*. <https://broadinstitute.github.io/picard/>;
- Jedel, E., Waern, M., Gustafson, D., Landén, M., Eriksson, E., Holm, G., Nilsson, L., Lind, A.-K., Janson, P. O. & Stener-Victorin, E. (2010). Anxiety and depression symptoms in women with polycystic ovary syndrome compared with controls matched for body mass index. *Human Reproduction*, 25(2), 450–456. <https://doi.org/10.1093/humrep/dep384>
- Johnson, G. D., Barrera, A., McDowell, I. C., D’Ippolito, A. M., Majoros, W. H., Vockley, C. M., Wang, X., Allen, A. S. & Reddy, T. E. (2018). Human genome-wide measurement of drug-responsive regulatory activity. *Nature Communications*, 9(1), 5317.  
<https://doi.org/10.1038/s41467-018-07607-x>
- Jones, M. R., Brower, M. A., Xu, N., Cui, J., Mengesha, E., Chen, Y.-D. I., Taylor, K. D., Azziz, R. & Goodarzi, M. O. (2015). Systems Genetics Reveals the Functional Context of PCOS Loci and Identifies Genetic and Molecular Mechanisms of Disease Heterogeneity. *PLOS Genetics*, 11(8), e1005455. <https://doi.org/10.1371/journal.pgen.1005455>
- Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., Berkum, N. L. van, Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J. & Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314), 430–435. <https://doi.org/10.1038/nature09380>
- Kallen, C. B., Arakane, F., Christenson, L. K., Watari, H., Devoto, L. & Strauss, J. F. (1998). Unveiling the mechanism of action and regulation of the steroidogenic acute regulatory

protein. *Molecular and Cellular Endocrinology*, 145(1–2), 39–45.  
[https://doi.org/10.1016/s0303-7207\(98\)00167-1](https://doi.org/10.1016/s0303-7207(98)00167-1)

- Karlsson, O., Domingue, B. W., Kim, R. & Subramanian, S. V. (2022). Estimating heritability of height without zygoty information for twins under five years in low- and middle-income countries: An application of normal finite mixture distribution models. *SSM - Population Health*, 17, 101043. <https://doi.org/10.1016/j.ssmph.2022.101043>
- Karnuta, J. M. & Scacheri, P. C. (2018). Enhancers: bridging the gap between gene control and human disease. *Human Molecular Genetics*, 27(R2), R219–R227.  
<https://doi.org/10.1093/hmg/ddy167>
- Kaur, P., Lu, X., Xu, Q., Irvin, E. M., Pappas, C., Zhang, H., Finkelstein, I. J., Shi, Z., Tao, Y. J., Yu, H. & Wang, H. (2023). High-speed AFM imaging reveals DNA capture and loop extrusion dynamics by cohesin-NIPBL. *Journal of Biological Chemistry*, 299(11), 105296.  
<https://doi.org/10.1016/j.jbc.2023.105296>
- Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S. & Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5), 800–811.  
<https://doi.org/10.1101/gr.144899.112>
- Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A. & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2), 93–108.  
<https://doi.org/10.1038/nrg.2015.17>
- Kim, Y.-S., Johnson, G. D., Seo, J., Barrera, A., Majoros, W. H., Ochoa, A., Cowart, T. N., Allen, A. S. & Reddy, T. E. (2021). Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Research*, 31(5), gr.269209.120.  
<https://doi.org/10.1101/gr.269209.120>
- Kim-Hellmuth, S., Aguet, F., Oliva, M., Muñoz-Aguirre, M., Kasela, S., Wucher, V., Castel, S. E., Hamel, A. R., Viñuela, A., Roberts, A. L., Mangul, S., Wen, X., Wang, G., Barbeira, A. N., Garrido-Martín, D., Nadel, B. B., Zou, Y., Bonazzola, R., Quan, J., ... Volpi, S. (2020). Cell type-specific genetic regulation of gene expression across human tissues. *Science*, 369(6509), eaaz8528. <https://doi.org/10.1126/science.aaz8528>
- Kioussis, D., Vanin, E., deLange, T., Flavell, R. A. & Grosveld, F. G. (1983).  $\beta$ -Globin gene inactivation by DNA translocation in  $\gamma\beta$ -thalassaemi. *Nature*, 306(5944), 662–666.  
<https://doi.org/10.1038/306662a0>
- Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., Costello, J. F., Shendure, J. & Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications*, 10(1), 3583.  
<https://doi.org/10.1038/s41467-019-11526-w>

- Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., Crawford, G. E., Reddy, T. E. & Gersbach, C. A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology*, *35*(6), 561–568. <https://doi.org/10.1038/nbt.3853>
- Kneppers, J., Severson, T. M., Siefert, J. C., Schol, P., Joosten, S. E. P., Yu, I. P. L., Huang, C.-C. F., Morova, T., Altıntaş, U. B., Giambartolomei, C., Seo, J.-H., Baca, S. C., Carneiro, I., Emberly, E., Pasaniuc, B., Jerónimo, C., Henrique, R., Freedman, M. L., Wessels, L. F. A., ... Zwart, W. (2022). Extensive androgen receptor enhancer heterogeneity in primary prostate cancers underlies transcriptional diversity and metastatic potential. *Nature Communications*, *13*(1), 7367. <https://doi.org/10.1038/s41467-022-35135-2>
- Knight, J. C. (2005). Regulatory polymorphisms underlying complex disease traits. *Journal of Molecular Medicine*, *83*(2), 97–109. <https://doi.org/10.1007/s00109-004-0603-7>
- Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O. & Zhang, F. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588. <https://doi.org/10.1038/nature14136>
- Kornberg, R. D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA. *Science*, *184*(4139), 868–871. <https://doi.org/10.1126/science.184.4139.868>
- Krishnan, N., Swoger, M., Rathbun, L. I., Fioramonti, P. J., Freshour, J., Bates, M., Patteson, A. E. & Hehnly, H. (2022). Rab11 endosomes and Pericentrin coordinate centrosome movement during pre-abscission in vivo. *Life Science Alliance*, *5*(7), e202201362. <https://doi.org/10.26508/lsa.202201362>
- Kulik, M., Bothe, M., Kibar, G., Fuchs, A., Schöne, S., Prekovic, S., Mayayo-Peralta, I., Chung, H.-R., Zwart, W., Helsen, C., Claessens, F. & Meijnsing, S. H. (2021). Androgen and glucocorticoid receptor direct distinct transcriptional programs by receptor-specific and shared DNA binding sites. *Nucleic Acids Research*, *49*(7), gkab185-. <https://doi.org/10.1093/nar/gkab185>
- Kulkarni, R., Teves, M. E., Han, A. X., McAllister, J. M. & Strauss, J. F. (2019). Co-Localization of Polycystic Ovary Syndrome Candidate Gene Products in Theca Cells Suggests Novel Signaling Pathways. *Journal of the Endocrine Society*, *3*(12), 2204–2223. <https://doi.org/10.1210/js.2019-00169>
- Kulsuptrakul, J., Wang, R., Meyers, N. L., Ott, M. & Puschnik, A. S. (2021). A genome-wide CRISPR screen identifies UFMylation and TRAMP-like complexes as host factors required for hepatitis A virus infection. *Cell Reports*, *34*(11), 108859. <https://doi.org/10.1016/j.celrep.2021.108859>
- Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D.,

- Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., ... Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. <https://doi.org/10.1038/nature14248>
- Kyrönlahti, A., Vetter, M., Euler, R., Bielinska, M., Jay, P. Y., Anttonen, M., Heikinheimo, M. & Wilson, D. B. (2011). GATA4 Deficiency Impairs Ovarian Function in Adult Mice. *Biology of Reproduction*, *84*(5), 1033–1044. <https://doi.org/10.1095/biolreprod.110.086850>
- Lam, D. D., Souza, F. S. J. de, Nasif, S., Yamashita, M., López-Leal, R., Otero-Corchon, V., Meece, K., Sampath, H., Mercer, A. J., Wardlaw, S. L., Rubinstein, M. & Low, M. J. (2015). Partially Redundant Enhancers Cooperatively Maintain Mammalian Pomc Expression Above a Critical Functional Threshold. *PLoS Genetics*, *11*(2), e1004935. <https://doi.org/10.1371/journal.pgen.1004935>
- Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K. P. & Gerstein, M. (2020). STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biology*, *21*(1), 298. <https://doi.org/10.1186/s13059-020-02194-x>
- Lee, D.-H., Park, C.-J., Jang, S., Cho, Y.-U., Seo, J. J., Im, H. J., Koh, K.-N., Cho, K. J., Song, J. S. & Seo, E.-J. (2018). Clinical and Cytogenetic Profiles of Rhabdomyosarcoma with Bone Marrow Involvement in Korean Children: A 15-Year Single-Institution Experience. *Annals of Laboratory Medicine*, *38*(2), 132–138. <https://doi.org/10.3343/alm.2018.38.2.132>
- Lee, H., Oh, J.-Y., Sung, Y.-A., Chung, H., Kim, H.-L., Kim, G. S., Cho, Y. S. & Kim, J. T. (2015). Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome. *Human Reproduction*, *30*(3), 723–731. <https://doi.org/10.1093/humrep/deu352>
- Legro, R. S., Arslanian, S. A., Ehrmann, D. A., Hoeger, K. M., Murad, M., Pasquali, R., Welt, C. K. & Society, E. (2013). Diagnosis and treatment of polycystic ovary syndrome: an Endocrine Society clinical practice guideline. *The Journal of Clinical Endocrinology and Metabolism*, *98*(12), 4565–4592. <https://doi.org/10.1210/jc.2013-2350>
- Legro, R. S., Driscoll, D., Strauss, J. F., Fox, J. & Dunaif, A. (1998). Evidence for a genetic basis for hyperandrogenemia in polycystic ovary syndrome. *Proceedings of the National Academy of Sciences*, *95*(25), 14956–14960. <https://doi.org/10.1073/pnas.95.25.14956>
- Lenthcher, Jessica A. & Decherney, A. H. (2021). Clinical Presentation and Diagnosis of Polycystic Ovarian Syndrome. *Clinical Obstetrics and Gynecology*, *64*(1), 3–11. <https://doi.org/10.1097/grf.0000000000000563>
- Lettice, L. A., Horikoshi, T., Heaney, S. J. H., Baren, M. J. van, Linde, H. C. van der, Breedveld, G. J., Joesse, M., Akarsu, N., Oostra, B. A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E.,

- Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S. W., Heutink, P., ... Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences*, 99(11), 7548–7553. <https://doi.org/10.1073/pnas.112212199>
- Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., ... Ruan, Y. (2012). Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, 148(1–2), 84–98. <https://doi.org/10.1016/j.cell.2011.12.014>
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, S., Ying, Z., Gentenaar, M., Rensen, P. C. N., Kooijman, S., Visser, J. A., Meijer, O. C. & Kroon, J. (2023). Glucocorticoid Receptor Antagonism Improves Glucose Metabolism in a Mouse Model of Polycystic Ovary Syndrome. *Journal of the Endocrine Society*, 8(1), bvad162. <https://doi.org/10.1210/jendso/bvad162>
- Liao, H.-K., Hatanaka, F., Araoka, T., Reddy, P., Wu, M.-Z., Sui, Y., Yamauchi, T., Sakurai, M., O’Keefe, D. D., Núñez-Delicado, E., Guillen, P., Campistol, J. M., Wu, C.-J., Lu, L.-F., Esteban, C. R. & Belmonte, J. C. I. (2017). In Vivo Target Gene Activation via CRISPR/Cas9-Mediated Trans-epigenetic Modulation. *Cell*, 171(7), 1495-1507.e15. <https://doi.org/10.1016/j.cell.2017.10.025>
- Lichou, F. & Trynka, G. (2020). Functional studies of GWAS variants are gaining momentum. *Nature Communications*, 11(1), 6283. <https://doi.org/10.1038/s41467-020-20188-y>
- Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950), 289–293. <https://doi.org/10.1126/science.1181369>
- Lin, J., Huang, H., Lin, L., Li, W. & Huang, J. (2020). MiR-23a induced the activation of CDC42/PAK1 pathway and cell cycle arrest in human cov434 cells by targeting FGD4. *Journal of Ovarian Research*, 13(1), 90. <https://doi.org/10.1186/s13048-020-00686-9>
- Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., Xia, J., Cheng, M., Yuan, Y., Zhang, P., Li, J., Feng, T., Chen, A., Zhang, W., Chen, F., Shang, Z., Zhang, X., Peters, B. A. & Liu, L. (2019). An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific Data*, 6(1), 65. <https://doi.org/10.1038/s41597-019-0071-0>

- Liu, S., Liu, Y., Zhang, Q., Wu, J., Liang, J., Yu, S., Wei, G.-H., White, K. P. & Wang, X. (2017). Systematic identification of regulatory variants associated with cancer risk. *Genome Biology*, 18(1), 194. <https://doi.org/10.1186/s13059-017-1322-z>
- Liu, X.-M., Yan, M.-Q., Ji, S.-Y., Sha, Q.-Q., Huang, T., Zhao, H., Liu, H.-B., Fan, H.-Y. & Chen, Z.-J. (2018). Loss of oocyte Rps26 in mice arrests oocyte growth and causes premature ovarian failure. *Cell Death & Disease*, 9(12), 1144. <https://doi.org/10.1038/s41419-018-1196-3>
- Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R. & Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11), 1443–1448. <https://doi.org/10.1038/ng.3679>
- Long, E., Yin, J., Funderburk, K. M., Xu, M., Feng, J., Kane, A., Zhang, T., Myers, T., Golden, A., Thakur, R., Kong, H., Jessop, L., Kim, E. Y., Jones, K., Chari, R., Machiela, M. J., Yu, K., Consortium, M. M.-A., Iles, M. M., ... Choi, J. (2022). Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. *The American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2022.11.006>
- Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1), 5900. <https://doi.org/10.1038/s41467-020-19653-5>
- López-Bigas, N., Blencowe, B. J. & Ouzounis, C. A. (2006). Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics*, 22(3), 269–277. <https://doi.org/10.1093/bioinformatics/bti781>
- Lorch, Y., LaPointe, J. W. & Kornberg, R. D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49(2), 203–210. [https://doi.org/10.1016/0092-8674\(87\)90561-7](https://doi.org/10.1016/0092-8674(87)90561-7)
- Love, M. I., Huber, W. & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Lunde, O., Magnus, P., Sandvik, L. & Høglø, S. (1989). Familial Clustering in the Polycystic Ovarian Syndrome. *Gynecologic and Obstetric Investigation*, 28(1), 23–30. <https://doi.org/10.1159/000293493>
- Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., ... Cherry, J. M. (2019). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, 48(D1), D882–D889. <https://doi.org/10.1093/nar/gkz1062>

- Lyle, S. M., Ahmed, S., Elliott, J. E., Stener-Victorin, E., Nachtigal, M. W. & Drögemöller, B. I. (2023). Transcriptome-wide association analyses identify an association between ARL14EP and polycystic ovary syndrome. *Journal of Human Genetics*, 1–7. <https://doi.org/10.1038/s10038-023-01120-w>
- Ma, M., Ru, Y., Chuang, L.-S., Hsu, N.-Y., Shi, L.-S., Hakenberg, J., Cheng, W.-Y., Uzilov, A., Ding, W., Glicksberg, B. S. & Chen, R. (2015). Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics*, 16(Suppl 8), S3. <https://doi.org/10.1186/1471-2164-16-s8-s3>
- Maglich, J. M., Kuhn, M., Chapin, R. E. & Pletcher, M. T. (2014). More than just hormones: H295R cells as predictors of reproductive toxicity. *Reproductive Toxicology*, 45, 77–86. <https://doi.org/10.1016/j.reprotox.2013.12.009>
- Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218), 18–21. <https://doi.org/10.1038/456018a>
- Majoros, W. H., Kim, Y.-S., Barrera, A., Li, F., Wang, X., Cunningham, S. J., Johnson, G. D., Guo, C., Lowe, W. L., Scholtens, D. M., Hayes, G. M., Reddy, T. E. & Allen, A. S. (2019). Bayesian Estimation of Genetic Regulatory Effects in High-throughput Reporter Assays. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btz545>
- Malm, M., Saghaleyni, R., Lundqvist, M., Giudici, M., Chotteau, V., Field, R., Varley, P. G., Hatton, D., Grassi, L., Svensson, T., Nielsen, J. & Rockberg, J. (2020). Evolution from adherent to suspension: systems biology of HEK293 cell line development. *Scientific Reports*, 10(1), 18996. <https://doi.org/10.1038/s41598-020-76137-8>
- Mangan, R. J., Alsina, F. C., Mosti, F., Sotelo-Fonseca, J. E., Snellings, D. A., Au, E. H., Carvalho, J., Sathyan, L., Johnson, G. D., Reddy, T. E., Silver, D. L. & Lowe, C. B. (2022). Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell*, 185(24), 4587–4603.e23. <https://doi.org/10.1016/j.cell.2022.10.016>
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753. <https://doi.org/10.1038/nature08494>
- Marla, S., Mortlock, S., Yoon, S., Crawford, J., Andersen, S., Mueller, M. D., McKinnon, B., Nguyen, Q. & Montgomery, G. W. (2023). Global Analysis of Transcription Start Sites and Enhancers in Endometrial Stromal Cells and Differences Associated with Endometriosis. *Cells*, 12(13), 1736. <https://doi.org/10.3390/cells12131736>
- Matharu, N. & Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS Genetics*, 11(12), e1005640. <https://doi.org/10.1371/journal.pgen.1005640>

- Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutuyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., ... Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099), 1190–1195. <https://doi.org/10.1126/science.1222794>
- Maurya, S. S. (2021). Role of Enhancers in Development and Diseases. *Epigenomes*, 5(4), 21. <https://doi.org/10.3390/epigenomes5040021>
- McAllister, J. M., Han, A. X., Modi, B. P., Teves, M. E., Mavodza, G. R., Anderson, Z. L., Shen, T., Christenson, L. K., Archer, K. J. & Strauss, J. F. (2019). MicroRNA Profiling Reveals miRNA-130b-3p Mediates DENND1A Variant 2 Expression and Androgen Biosynthesis. *Endocrinology*, 160(8), 1964–1981. <https://doi.org/10.1210/en.2019-00013>
- McAllister, J. M., Modi, B., Miller, B. A., Biegler, J., Bruggeman, R., Legro, R. S. & Strauss, J. F. (2014). Overexpression of a DENND1A isoform produces a polycystic ovary syndrome theca phenotype. *Proceedings of the National Academy of Sciences*, 111(15), E1519–E1527. <https://doi.org/10.1073/pnas.1400574111>
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A. & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5), 356–369. <https://doi.org/10.1038/nrg2344>
- McDowell, I. C., Barrera, A., D'Ippolito, A. M., Vockley, C. M., Hong, L. K., Leichter, S. M., Bartelt, L. C., Majoros, W. H., Song, L., Safi, A., Koçak, D. D., Gersbach, C. A., Hartemink, A. J., Crawford, G. E., Engelhardt, B. E. & Reddy, T. E. (2018). Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Research*, 28(9), 1272–1284. <https://doi.org/10.1101/gr.233346.117>
- Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., Kellis, M., Lander, E. S. & Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology*, 30(3), 271–277. <https://doi.org/10.1038/nbt.2137>
- Merson, S., Yang, Z. H., Brewer, D., Olmos, D., Eichholz, A., McCarthy, F., Fisher, G., Kovacs, G., Berney, D. M., Foster, C. S., Møller, H., Scardino, P., Cuzick, J., Cooper, C. S., Clark, J. P. & Group, T. P. (2014). Focal amplification of the androgen receptor gene in hormone-naive human prostate cancer. *British Journal of Cancer*, 110(6), 1655–1662. <https://doi.org/10.1038/bjc.2014.13>
- Meyer, C., McGrath, B. P., Cameron, J., Kotsopoulos, D. & Teede, H. J. (2005). Vascular Dysfunction and Metabolic Parameters in Polycystic Ovary Syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 90(8), 4630–4635. <https://doi.org/10.1210/jc.2004-1487>

- Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, *155*(3), 733–740. <https://doi.org/10.1099/mic.0.023960-0>
- Momozawa, Y. & Mizukami, K. (2021). Unique roles of rare variants in the genetics of complex diseases in humans. *Journal of Human Genetics*, *66*(1), 11–23. <https://doi.org/10.1038/s10038-020-00845-2>
- Morris, J. A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D. A., Hao, S., Mimitou, E. P., Smibert, P., Roeder, K., Katsevich, E., Lappalainen, T. & Sanjana, N. E. (2023). Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science*, *380*(6646), eadh7699. <https://doi.org/10.1126/science.adh7699>
- Muerdter, F., Boryń, Ł. M. & Arnold, C. D. (2015). STARR-seq — Principles and applications. *Genomics*, *106*(3), 145–150. <https://doi.org/10.1016/j.ygeno.2015.06.001>
- Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. D. & Stark, A. (2017). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods*, *15*(2), 141. <https://doi.org/10.1038/nmeth.4534>
- Mukherjee, A. & Roy, S. K. (2013). Expression of ErbB3-Binding Protein-1 (EBP1) during Primordial Follicle Formation: Role of Estradiol-17β. *PLoS ONE*, *8*(6), e67068. <https://doi.org/10.1371/journal.pone.0067068>
- Mulvey, B. & Dougherty, J. D. (2021). Transcriptional-regulatory convergence across functional MDD risk variants identified by massively parallel reporter assays. *Translational Psychiatry*, *11*(1), 403. <https://doi.org/10.1038/s41398-021-01493-6>
- Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., Pirruccello, J. P., Muchmore, B., Prokunina-Olsson, L., Hall, J. L., Schadt, E. E., Morales, C. R., Lund-Katz, S., Phillips, M. C., Wong, J., ... Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, *466*(7307), 714–719. <https://doi.org/10.1038/nature09266>
- Myint, L., Wang, R., Boukas, L., Hansen, K. D., Goff, L. A. & Avramopoulos, D. (2020). A screen of 1,049 schizophrenia and 30 Alzheimer’s-associated variants for regulatory potential. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *183*(1), 61–73. <https://doi.org/10.1002/ajmg.b.32761>
- Nair, A. K. & Baier, L. J. (2018). Disease Gene Identification, Methods and Protocols. *Methods in Molecular Biology*, *1706*, 303–319. [https://doi.org/10.1007/978-1-4939-7471-9\\_17](https://doi.org/10.1007/978-1-4939-7471-9_17)

- Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. (2021). CRISPR technologies for precise epigenome editing. *Nature Cell Biology*, 23(1), 11–22. <https://doi.org/10.1038/s41556-020-00620-7>
- Nambiar, M., Kari, V. & Raghavan, S. C. (2008). Chromosomal translocations in cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1786(2), 139–152. <https://doi.org/10.1016/j.bbcan.2008.07.005>
- Nassar, G. N. & Leslie, S. W. (2023). *Physiology, Testosterone*. StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK526128/>
- Ni, G., Werf, J. van der, Zhou, X., Hyppönen, E., Wray, N. R. & Lee, S. H. (2019). Genotype–covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. *Nature Communications*, 10(1), 2239. <https://doi.org/10.1038/s41467-019-10128-w>
- Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. & Cox, N. J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics*, 6(4), e1000888. <https://doi.org/10.1371/journal.pgen.1000888>
- Nomura, T., Suzuki, S., Miyauchi, T., Takeda, M., Shinkuma, S., Fujita, Y., Nishie, W., Akiyama, M. & Shimizu, H. (2018). Chromosomal inversions as a hidden disease-modifying factor for somatic recombination phenotypes. *JCI Insight*, 3(6), e97595. <https://doi.org/10.1172/jci.insight.97595>
- Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A. & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, 554(7691), 239–243. <https://doi.org/10.1038/nature25461>
- Ouwerkerk, A. F. van, Bosada, F. M., Liu, J., Zhang, J., Duijvenboden, K. van, Chaffin, M., Tucker, N. R., Pijnappels, D., Ellinor, P. T., Barnett, P., Vries, A. A. F. de & Christoffels, V. M. (2020). Identification of Functional Variant Enhancers Associated With Atrial Fibrillation. *Circulation Research*, 127(2), 229–243. <https://doi.org/10.1161/circresaha.119.316006>
- Owen, D. M., Kwon, M., Huang, X., Nagari, A., Nandu, T. & Kraus, W. L. (2023). Genome-wide identification of transcriptional enhancers during human placental development and association with function, differentiation, and disease. *Biology of Reproduction*, 109(6), 965–981. <https://doi.org/10.1093/biolre/ioad119>
- Pai, A. A., Pritchard, J. K. & Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics*, 11(1), e1004857. <https://doi.org/10.1371/journal.pgen.1004857>

- Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D. & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27(12), 1173–1175. <https://doi.org/10.1038/nbt.1589>
- Perez, A. R., Pritykin, Y., Vidigal, J. A., Chhangawala, S., Zamparo, L., Leslie, C. S. & Ventura, A. (2017). GuideScan software for improved single and paired CRISPR guide RNA design. *Nature Biotechnology*, 35(4), nbt.3804. <https://doi.org/10.1038/nbt.3804>
- Peterson, E. J., Bailo, R., Rothchild, A. C., Arrieta-Ortiz, M. L., Kaur, A., Pan, M., Mai, D., Abidi, A. A., Cooper, C., Aderem, A., Bhatt, A. & Baliga, N. S. (2019). Path-seq identifies an essential mycolate remodeling program for mycobacterial host adaptation. *Molecular Systems Biology*, 15(3), e8584. <https://doi.org/10.15252/msb.20188584>
- Phillips, J. E. & Corces, V. G. (2009). CTCF: Master Weaver of the Genome. *Cell*, 137(7), 1194–1211. <https://doi.org/10.1016/j.cell.2009.06.001>
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909. <https://doi.org/10.1038/ng1847>
- Puig, M., Casillas, S., Villatoro, S. & Cáceres, M. (2015). Human inversions and their functional consequences. *Briefings in Functional Genomics*, 14(5), 369–379. <https://doi.org/10.1093/bfgp/elv020>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. de, Daly, M. J. & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Pyrzynska, B., Pilecka, I. & Miaczynska, M. (2009). Endocytic proteins in the regulation of nuclear signaling, transcription and tumorigenesis. *Molecular Oncology*, 3(4), 321–338. <https://doi.org/10.1016/j.molonc.2009.06.001>
- Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Rahimov, F., Program, N. C. S., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., Domann, F. E., Govil, M., Christensen, K., Bille, C., Melbye, M., Jugessur, A., Lie, R. T., Wilcox, A. J., Fitzpatrick, D. R., Green, E. D., Mossey, P. A., Little, J., ... Murray, J. C. (2008). Disruption of an AP-2 $\alpha$  binding site in an IRF6 enhancer is associated with cleft lip. *Nature Genetics*, 40(11), 1341–1347. <https://doi.org/10.1038/ng.242>
- Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, 42(W1), W187–W191. <https://doi.org/10.1093/nar/gku365>

- 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. <https://doi.org/10.1093/nar/gkw257>
- Rao, S., Yao, Y. & Bauer, D. E. (2021). Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Medicine*, *13*(1), 41. <https://doi.org/10.1186/s13073-021-00857-3>
- Rasmussen, K. D., Jia, G., Johansen, J. V., Pedersen, M. T., Rapin, N., Bagger, F. O., Porse, B. T., Bernard, O. A., Christensen, J. & Helin, K. (2015). Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. *Genes & Development*, *29*(9), 910–922. <https://doi.org/10.1101/gad.260174.115>
- Ray, J. P., Boer, C. G. de, Fulco, C. P., Lareau, C. A., Kanai, M., Ulirsch, J. C., Tewhey, R., Ludwig, L. S., Reilly, S. K., Bergman, D. T., Engreitz, J. M., Issner, R., Finucane, H. K., Lander, E. S., Regev, A. & Hacohen, N. (2020). Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. *Nature Communications*, *11*(1), 1237. <https://doi.org/10.1038/s41467-020-15022-4>
- Reddy, T. E., Gertz, J., Crawford, G. E., Garabedian, M. J. & Myers, R. M. (2012). The Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes. *Molecular and Cellular Biology*, *32*(18), 3756–3767. <https://doi.org/10.1128/mcb.00062-12>
- Reddy, T. E., Pauli, F., Sprouse, R. O., Neff, N. F., Newberry, K. M., Garabedian, M. J. & Myers, R. M. (2009). Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Research*, *19*(12), 2163–2171. <https://doi.org/10.1101/gr.097022.109>
- Reilly, S. K., Gosai, S. J., Gutierrez, A., Mackay-Smith, A., Ulirsch, J. C., Kanai, M., Mouri, K., Berenzy, D., Kales, S., Butler, G. M., Gladden-Young, A., Bhuiyan, R. M., Stitzel, M. L., Finucane, H. K., Sabeti, P. C. & Tewhey, R. (2021). Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH. *Nature Genetics*, *53*(8), 1166–1176. <https://doi.org/10.1038/s41588-021-00900-4>
- Richards, J. S. (1994). Hormonal Control of Gene Expression in the Ovary. *Endocrine Reviews*, *15*(6), 725–751. <https://doi.org/10.1210/edrv-15-6-725>
- Riesterberg, C., Jagasia, A., Markovic, D., Buyalos, R. P. & Azziz, R. (2021). Health Care-Related Economic Burden of Polycystic Ovary Syndrome in the United States: Pregnancy-Related and Long-Term Health Consequences. *The Journal of Clinical Endocrinology & Metabolism*, *107*(2), dgab613-. <https://doi.org/10.1210/clinem/dgab613>
- Rojo, F. (2001). Mechanisms of transcriptional repression. *Current Opinion in Microbiology*, *4*(2), 145–151. [https://doi.org/10.1016/s1369-5274\(00\)00180-6](https://doi.org/10.1016/s1369-5274(00)00180-6)

- Roth, T. L. & Marson, A. (2021). Genetic Disease and Therapy. *Annual Review of Pathology: Mechanisms of Disease*, 16(1), 145–166. <https://doi.org/10.1146/annurev-pathmechdis-012419-032626>
- Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., Stahl, E. A., Georgakopoulos, A., Ruderfer, D. M., Charney, A., Okada, Y., Siminovitch, K. A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Plenge, R. M., Raychaudhuri, S., Fromer, M., ... Sklar, P. (2014). A Role for Noncoding Variation in Schizophrenia. *Cell Reports*, 9(4), 1417–1429. <https://doi.org/10.1016/j.celrep.2014.10.015>
- Rubin, K. H., Glintborg, D., Nybo, M., Abrahamsen, B. & Andersen, M. (2017). Development and Risk Factors of Type 2 Diabetes in a Nationwide Population of Women With Polycystic Ovary Syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 102(10), 3848–3857. <https://doi.org/10.1210/jc.2017-01354>
- Rundlett, S. E. & Miesfeld, R. L. (1995). Quantitative differences in androgen and glucocorticoid receptor DNA binding properties contribute to receptor-selective transcriptional regulation. *Molecular and Cellular Endocrinology*, 109(1), 1–10. [https://doi.org/10.1016/0303-7207\(95\)03477-o](https://doi.org/10.1016/0303-7207(95)03477-o)
- Ruth, K. S., Day, F. R., Tyrrell, J., Thompson, D. J., Wood, A. R., Mahajan, A., Beaumont, R. N., Wittemans, L., Martin, S., Busch, A. S., Erzurumluoglu, M. A., Hollis, B., O'Mara, T. A., Consortium, E., McCarthy, M. I., Langenberg, C., Easton, D. F., Wareham, N. J., Burgess, S., ... Perry, J. R. (2020). Using human genetics to understand the disease impacts of testosterone in men and women. *Nature Medicine*, 26(2), 252–258. <https://doi.org/10.1038/s41591-020-0751-5>
- Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S. & Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47), E6456–E6465. <https://doi.org/10.1073/pnas.1518552112>
- Sanchez, C. G., Acker, C. M., Gray, A., Varadarajan, M., Song, C., Cochran, N. R., Paula, S., Lindeman, A., An, S., McAllister, G., Alford, J., Reece-Hoyes, J., Russ, C., Craig, L., Capre, K., Doherty, C., Hoffman, G. R., Luchansky, S. J., Polydoro, M., ... Elwood, F. (2021). Genome-wide CRISPR screen identifies protein pathways modulating tau protein levels in neurons. *Communications Biology*, 4(1), 736. <https://doi.org/10.1038/s42003-021-02272-1>
- Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A. & Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*, 353(6307), 1545–1549. <https://doi.org/10.1126/science.aaf7613>
- Schaid, D. J., Chen, W. & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8), 491–504. <https://doi.org/10.1038/s41576-018-0016-z>

- Scherzer, C. R., Grass, J. A., Liao, Z., Pepivani, I., Zheng, B., Eklund, A. C., Ney, P. A., Ng, J., McGoldrick, M., Mollenhauer, B., Bresnick, E. H. & Schlossmacher, M. G. (2008). GATA transcription factors directly regulate the Parkinson's disease-linked gene  $\alpha$ -synuclein. *Proceedings of the National Academy of Sciences*, *105*(31), 10907–10912. <https://doi.org/10.1073/pnas.0802437105>
- Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E. & Willer, C. J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, *31*(16), 2601–2606. <https://doi.org/10.1093/bioinformatics/btv201>
- Schnitzler, G. R., Kang, H., Fang, S., Angom, R. S., Lee-Kim, V. S., Ma, X. R., Zhou, R., Zeng, T., Guo, K., Taylor, M. S., Vellarikkal, S. K., Barry, A. E., Sias-Garcia, O., Bloemendal, A., Munson, G., Guckelberger, P., Nguyen, T. H., Bergman, D. T., Hinshaw, S., ... Engreitz, J. M. (2024). Convergence of coronary artery disease genes onto endothelial cell programs. *Nature*, 1–9. <https://doi.org/10.1038/s41586-024-07022-x>
- Scholl, U. I., Stölting, G., Schewe, J., Thiel, A., Tan, H., Nelson-Williams, C., Vichot, A. A., Jin, S. C., Loring, E., Untiet, V., Yoo, T., Choi, J., Xu, S., Wu, A., Kirchner, M., Mertins, P., Rump, L. C., Onder, A. M., Gamble, C., ... Lifton, R. P. (2018). CLCN2 chloride channel mutations in familial hyperaldosteronism type II. *Nature Genetics*, *50*(3), 349–354. <https://doi.org/10.1038/s41588-018-0048-5>
- Schwarzkopf, M., Choi, H. M. T. & Pierce, N. A. (2020). In Situ Hybridization Protocols. *Methods in Molecular Biology*, *2148*, 127–141. [https://doi.org/10.1007/978-1-0716-0623-0\\_8](https://doi.org/10.1007/978-1-0716-0623-0_8)
- Sen, A. & Hammes, S. R. (2010). Granulosa Cell-Specific Androgen Receptors Are Critical Regulators of Ovarian Development and Function. *Molecular Endocrinology*, *24*(7), 1393–1403. <https://doi.org/10.1210/me.2010-0006>
- Severson, T. M., Kim, Y., Joosten, S. E. P., Schuurman, K., Groep, P. van der, Moelans, C. B., Hoeve, N. D. ter, Manson, Q. F., Martens, J. W., Deurzen, C. H. M. van, Barbe, E., Hedenfalk, I., Bult, P., Smit, V. T. H. B. M., Linn, S. C., Diest, P. J. van, Wessels, L. & Zwart, W. (2018). Characterizing steroid hormone receptor chromatin binding landscapes in male and female breast cancer. *Nature Communications*, *9*(1), 482. <https://doi.org/10.1038/s41467-018-02856-2>
- Shapiro, J. A. & Sternberg, R. (2005). Why repetitive DNA is essential to genome function. *Biological Reviews*, *80*(2), 227–250. <https://doi.org/10.1017/s1464793104006657>
- Shen, S. Q., Myers, C. A., Hughes, A. E., Byrne, L. C., Flannery, J. G. & Corbo, J. C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Research*, *26*(2), 238–255. <https://doi.org/10.1101/gr.193789.115>

- Shi, Jingjing, Gao, Q., Cao, Y. & Fu, J. (2019). *Dennd1a*, a susceptibility gene for polycystic ovary syndrome, is essential for mouse embryogenesis. *Developmental Dynamics*, 248(5), 351–362. <https://doi.org/10.1002/dvdy.28>
- Shi, Junwei, Whyte, W. A., Zepeda-Mendoza, C. J., Milazzo, J. P., Shen, C., Roe, J.-S., Minder, J. L., Mercan, F., Wang, E., Eckersley-Maslin, M. A., Campbell, A. E., Kawaoka, S., Shareef, S., Zhu, Z., Kendall, J., Muhar, M., Haslinger, C., Yu, M., Roeder, R. G., ... Vakoc, C. R. (2013). Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes & Development*, 27(24), 2648–2662. <https://doi.org/10.1101/gad.232710.113>
- Shi, Y., Zhao, H., Shi, Y., Cao, Y., Yang, D., Li, Z., Zhang, B., Liang, X., Li, T., Chen, J., Shen, J., Zhao, J., You, L., Gao, X., Zhu, D., Zhao, X., Yan, Y., Qin, Y., Li, W., ... Chen, Z.-J. (2012). Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nature Genetics*, 44(9), 1020. <https://doi.org/10.1038/ng.2384>
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W. & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8), 1034–1050. <https://doi.org/10.1101/gr.3715005>
- Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6), 477–485. <https://doi.org/10.1038/nrg2361>
- Smale, S. T. & Kadonaga, J. T. (2003). THE RNA POLYMERASE II CORE PROMOTER. *Annual Review of Biochemistry*, 72(1), 449–479. <https://doi.org/10.1146/annurev.biochem.72.121801.161520>
- Smela, M. D. P., Kramme, C. C., Fortuna, P. R., Adams, J. L., Su, R., Dong, E., Kobayashi, M., Brix, G., Kavirayuni, V. S., Tysinger, E., Kohman, R. E., Shioda, T., Chatterjee, P. & Church, G. M. (2023). Directed differentiation of human iPSCs to functional ovarian granulosa-like cells via transcription factor overexpression. *ELife*, 12, e83291. <https://doi.org/10.7554/elife.83291>
- Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puvion, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H.-K., Naranjo, S., Acemel, R. D., ... Nóbrega, M. A. (2014). Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature*, 507(7492), 371–375. <https://doi.org/10.1038/nature13138>
- Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J. & Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *The Lancet*, 366(9495), 1484–1498. [https://doi.org/10.1016/s0140-6736\(05\)67601-5](https://doi.org/10.1016/s0140-6736(05)67601-5)

- Soldner, F., Stelzer, Y., Shivalila, C. S., Abraham, B. J., Latourelle, J. C., Barrasa, M. I., Goldmann, J., Myers, R. H., Young, R. A. & Jaenisch, R. (2016). Parkinson-associated risk variant in distal enhancer of  $\alpha$ -synuclein modulates target gene expression. *Nature*, *533*(7601), 95–99. <https://doi.org/10.1038/nature17939>
- Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., ... Harris, L. W. (2022). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, *51*(D1), D977–D985. <https://doi.org/10.1093/nar/gkac1010>
- Song, L. & Crawford, G. E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384. <https://doi.org/10.1101/pdb.prot5384>
- Spaanderman, D. C. E., Nixon, M., Buurstedde, J. C., Sips, H. H. C. M., Schilperoort, M., Kuipers, E. N., Backer, E. A., Kooijman, S., Rensen, P. C. N., Homer, N. Z. M., Walker, B. R., Meijer, O. C. & Kroon, J. (2019). Androgens modulate glucocorticoid receptor activity in adipose tissue and liver. *Journal of Endocrinology*, *240*(1), 51–63. <https://doi.org/10.1530/joe-18-0503>
- Stocco, D. M., Wang, X., Jo, Y. & Manna, P. R. (2005). Multiple Signaling Pathways Regulating Steroidogenesis and Steroidogenic Acute Regulatory Protein Expression: More Complicated than We Thought. *Molecular Endocrinology*, *19*(11), 2647–2659. <https://doi.org/10.1210/me.2004-0532>
- Stranger, B. E. & Raj, T. (2013). Genetics of human gene expression. *Current Opinion in Genetics & Development*, *23*(6), 627–634. <https://doi.org/10.1016/j.gde.2013.10.004>
- Sun, Q., Gao, Y., Yang, J., Lu, J., Feng, W. & Yang, W. (2022). Mendelian Randomization Analysis Identified Potential Genes Pleiotropically Associated with Polycystic Ovary Syndrome. *Reproductive Sciences*, *29*(3), 1028–1037. <https://doi.org/10.1007/s43032-021-00776-z>
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., ... Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. <https://doi.org/10.1038/s41586-021-03205-y>
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. <https://doi.org/10.1038/s41576-019-0127-1>
- Tee, M. K., Speek, M., Legeza, B., Modi, B., Teves, M. E., McAllister, J. M., Strauss, J. F. & Miller, W. L. (2016). Alternative splicing of DENND1A, a PCOS candidate gene, generates

- variant 2. *Molecular and Cellular Endocrinology*, 434, 25–35.  
<https://doi.org/10.1016/j.mce.2016.06.011>
- Tehranchi, A., Hie, B., Dacre, M., Kaplow, I., Pettie, K., Combs, P. & Fraser, H. B. (2019). Fine-mapping cis-regulatory variants in diverse human populations. *ELife*, 8, e39595.  
<https://doi.org/10.7554/elife.39595>
- Tenbaum, S. & Baniahmad, A. (1997). Nuclear receptors: Structure, function and involvement in disease. *The International Journal of Biochemistry & Cell Biology*, 29(12), 1325–1341.  
[https://doi.org/10.1016/s1357-2725\(97\)00087-3](https://doi.org/10.1016/s1357-2725(97)00087-3)
- Teves, M. E., Modi, B. P., Kulkarni, R., Han, A. X., Marks, J. S., Subler, M. A., Windle, J., Newall, J. M., McAllister, J. M. & Strauss, J. F. (2020). Human DENND1A.V2 Drives Cyp17a1 Expression and Androgen Production in Mouse Ovaries and Adrenals. *International Journal of Molecular Sciences*, 21(7), 2545. <https://doi.org/10.3390/ijms21072545>
- Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F. & Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, 165(6), 1519–1529. <https://doi.org/10.1016/j.cell.2016.04.027>
- Thakore, P. I., Black, J. B., Hilton, I. B. & Gersbach, C. A. (2016). Editing the epigenome: technologies for programmable transcription and epigenetic modulation. *Nature Methods*, 13(2), 127–137. <https://doi.org/10.1038/nmeth.3733>
- Thakore, P. I., D’Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E. & Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods*, 12(12), 1143–1149. <https://doi.org/10.1038/nmeth.3630>
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., ... Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414), 75–82. <https://doi.org/10.1038/nature11232>
- Toledo, S. P., Brunner, H. G., Kraaij, R., Post, M., Dahia, P. L., Hayashida, C. Y. & Themmen, A. P. K. H. (1996). An inactivating mutation of the luteinizing hormone receptor causes amenorrhea in a 46,XX female. *The Journal of Clinical Endocrinology & Metabolism*, 81(11), 3850–3854. <https://doi.org/10.1210/jcem.81.11.8923827>
- Townsend, K. G., Brennand, K. J. & Huckins, L. M. (2020). Massively parallel techniques for cataloguing the regulome of the human brain. *Nature Neuroscience*, 23(12), 1509–1521.  
<https://doi.org/10.1038/s41593-020-00740-1>

- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J. van, Salzberg, S. L., Wold, B. J. & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), 511–515. <https://doi.org/10.1038/nbt.1621>
- Tremblay, J. J., Hamel, F. & Viger, R. S. (2002). Protein Kinase A-Dependent Cooperation between GATA and CCAAT/Enhancer-Binding Protein Transcription Factors Regulates Steroidogenic Acute Regulatory Protein Promoter Activity. *Endocrinology*, 143(10), 3935–3945. <https://doi.org/10.1210/en.2002-220413>
- Tremblay, J. J. & Viger, R. S. (1999). Transcription Factor GATA-4 Enhances Müllerian Inhibiting Substance Gene Transcription through a Direct Interaction with the Nuclear Receptor SF-1. *Molecular Endocrinology*, 13(8), 1388–1401. <https://doi.org/10.1210/mend.13.8.0330>
- Tremblay, J. J. & Viger, R. S. (2001). GATA Factors Differentially Activate Multiple Gonadal Promoters through Conserved GATA Regulatory Elements. *Endocrinology*, 142(3), 977–986. <https://doi.org/10.1210/endo.142.3.7995>
- Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S. & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45(2), 124–130. <https://doi.org/10.1038/ng.2504>
- Tyrmi, J. S., Arffman, R. K., Pujol-Gualdo, N., Kurra, V., Morin-Papunen, L., Sliz, E., Consortium, E. B. R. T. F., Piltonen, T. T., Laisk, T., Kettunen, J. & Laivuori, H. (2021). Leveraging Northern European population history: novel low-frequency variants for polycystic ovary syndrome. *Human Reproduction*, 37(2), 352–365. <https://doi.org/10.1093/humrep/deab250>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., Vries, J. de, Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 59. <https://doi.org/10.1038/s43586-021-00056-9>
- Vandesompele, J., Preter, K. D., Pattyn, F., Poppe, B., Roy, N. V., Paepe, A. D. & Speleman, F. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, 3(7), research0034.1. <https://doi.org/10.1186/gb-2002-3-7-research0034>
- Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., Ballester, B., Andrau, J. C. & Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, 6(1), 6905. <https://doi.org/10.1038/ncomms7905>
- Vassilatou, E. (2014). Nonalcoholic fatty liver disease and polycystic ovary syndrome. *World Journal of Gastroenterology*, 20(26), 8351–8363. <https://doi.org/10.3748/wjg.v20.i26.8351>

- Veltman-Verhulst, S. M., Boivin, J., Eijkemans, M. J. C. & Fauser, B. J. C. M. (2012). Emotional distress is a common risk in women with polycystic ovary syndrome: a systematic review and meta-analysis of 28 studies. *Human Reproduction Update*, *18*(6), 638–651. <https://doi.org/10.1093/humupd/dms029>
- Vink, J., Sadrzadeh, S., Lambalk, C. & Boomsma, D. (2006). Heritability of Polycystic Ovary Syndrome in a Dutch Twin-Family Study. *The Journal of Clinical Endocrinology & Metabolism*, *91*(6), 2100–2104. <https://doi.org/10.1210/jc.2005-1494>
- Visscher, P. M. & Goddard, M. E. (2019). From R.A. Fisher’s 1918 Paper to GWAS a Century Later. *Genetics*, *211*(4), 1125–1130. <https://doi.org/10.1534/genetics.118.301594>
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101*(1), 5–22. <https://doi.org/10.1016/j.ajhg.2017.06.005>
- Vockley, C. M., Barrera, A. & Reddy, T. E. (2017). Decoding the role of regulatory element polymorphisms in complex disease. *Current Opinion in Genetics & Development*, *43*, 38–45. <https://doi.org/10.1016/j.gde.2016.10.007>
- Vockley, C. M., Guo, C., Majoros, W. H., Nodzenski, M., Scholtens, D. M., Hayes, G. M., Lowe, W. L. & Reddy, T. E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Research*, *25*(8), 1206–1214. <https://doi.org/10.1101/gr.190090.115>
- Wainschtein, P., Jain, D., Zheng, Z., Aslibekyan, S., Becker, D., Bi, W., Brody, J., Carlson, J. C., Correa, A., Du, M. M., Fernandez-Rhodes, L., Ferrier, K. R., Graff, M., Guo, X., He, J., Heard-Costa, N. L., Highland, H. M., Hirschhorn, J. N., Howard-Claudio, C. M., ... Visscher, P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, *54*(3), 263–273. <https://doi.org/10.1038/s41588-021-00997-7>
- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genetics*, *17*(9), e1009440. <https://doi.org/10.1371/journal.pgen.1009440>
- Wang, F., Pan, J., Liu, Y., Meng, Q., Lv, P., Qu, F., Ding, G.-L., Klausen, C., Leung, P. C., Chan, H., Yao, W., Zhou, C.-Y., Shi, B., Zhang, J., Sheng, J. & Huang, H. (2015). Alternative splicing of the androgen receptor in polycystic ovary syndrome. *Proceedings of the National Academy of Sciences*, *112*(15), 4743–4748. <https://doi.org/10.1073/pnas.1418216112>
- Wang, J., Zhu, Z., Nolfo, R. & Elias, J. A. (1999). Dexamethasone regulation of lung epithelial cell and fibroblast interleukin-11 production. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, *276*(1), L175–L185. <https://doi.org/10.1152/ajplung.1999.276.1.1175>

- Wang, X., Tucker, N. R., Rizki, G., Mills, R., Krijger, P. H., Wit, E. de, Subramanian, V., Bartell, E., Nguyen, X.-X., Ye, J., Leyton-Mange, J., Dolmatova, E. V., Harst, P. van der, Laat, W. de, Ellinor, P. T., Newton-Cheh, C., Milan, D. J., Kellis, M. & Boyer, L. A. (2016). Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *ELife*, *5*, e10557. <https://doi.org/10.7554/elife.10557>
- Waterbury, J. S., Teves, M. E., Gaynor, A., Han, A. X., Mavodza, G., Newell, J., Strauss, J. F. & McAllister, J. M. (2022). The PCOS GWAS Candidate Gene ZNF217 Influences Theca Cell Expression of DENND1A.V2, CYP17A1, and Androgen Production. *Journal of the Endocrine Society*, *6*(7), bvac078. <https://doi.org/10.1210/jendso/bvac078>
- Weidemüller, P., Kholmatov, M., Petsalaki, E. & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. *PROTEOMICS*, *21*(23–24), e2000034. <https://doi.org/10.1002/pmic.202000034>
- Weikum, E. R., Liu, X. & Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Science*, *27*(11), 1876–1892. <https://doi.org/10.1002/pro.3496>
- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(D1), D1001–D1006. <https://doi.org/10.1093/nar/gkt1229>
- Wu, G. M. J., Chen, A. C. H., Yeung, W. S. B. & Lee, Y. L. (2023). Current progress on in vitro differentiation of ovarian follicles from pluripotent stem cells. *Frontiers in Cell and Developmental Biology*, *11*, 1166351. <https://doi.org/10.3389/fcell.2023.1166351>
- Wünnemann, F., Tadjó, T. F., Beaudoin, M., Lalonde, S., Lo, K. S., Kleinstiver, B. P. & Lettre, G. (2023). Multimodal CRISPR perturbations of GWAS loci associated with coronary artery disease in vascular endothelial cells. *PLOS Genetics*, *19*(3), e1010680. <https://doi.org/10.1371/journal.pgen.1010680>
- Xiang, X., Li, C., Chen, X., Dou, H., Li, Y., Zhang, X. & Luo, Y. (2019). CRISPR Gene Editing, Methods and Protocols. *Methods in Molecular Biology*, *1961*, 255–269. [https://doi.org/10.1007/978-1-4939-9170-9\\_16](https://doi.org/10.1007/978-1-4939-9170-9_16)
- Yildiz, B. O., Goodarzi, M. O., Guo, X., Rotter, J. I. & Azziz, R. (2006). Heritability of dehydroepiandrosterone sulfate in women with polycystic ovary syndrome and their sisters. *Fertility and Sterility*, *86*(6), 1688–1693. <https://doi.org/10.1016/j.fertnstert.2006.05.045>
- Yu, J., Liu, Y., Zhang, D., Zhai, D., Song, L., Cai, Z. & Yu, C. (2019). Baicalin inhibits recruitment of GATA1 to the HSD3B2 promoter and reverses hyperandrogenism of PCOS. *The Journal of Endocrinology*. <https://doi.org/10.1530/joe-18-0678>

- Zeineldin, M., Camp, P., Farrell, D., Lehman, K. & Thacker, T. (2023). Whole genome sequencing of *Mycobacterium bovis* directly from clinical tissue samples without culture. *Frontiers in Microbiology*, *14*, 1141651. <https://doi.org/10.3389/fmicb.2023.1141651>
- Zhai, J., Li, S., Cheng, X., Chen, Z., Li, W. & Du, Y. (2020). A candidate pathogenic gene, zinc finger gene 217 (ZNF217), may contribute to polycystic ovary syndrome through prostaglandin E2. *Acta Obstetricia et Gynecologica Scandinavica*, *99*(1), 119–126. <https://doi.org/10.1111/aogs.13719>
- Zhang, J., Jiang, Z. & Shi, A. (2022). Rab GTPases: The principal players in crafting the regulatory landscape of endosomal trafficking. *Computational and Structural Biotechnology Journal*, *20*, 4464–4472. <https://doi.org/10.1016/j.csbj.2022.08.016>
- Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S. & Ren, B. (2021). A single-cell atlas of chromatin accessibility in the human genome. *Cell*, *184*(24), 5985-6001.e19. <https://doi.org/10.1016/j.cell.2021.10.024>
- Zhang, P., Xia, J.-H., Zhu, J., Gao, P., Tian, Y.-J., Du, M., Guo, Y.-C., Suleman, S., Zhang, Q., Kohli, M., Tillmans, L. S., Thibodeau, S. N., French, A. J., Cerhan, J. R., Wang, L.-D., Wei, G.-H. & Wang, L. (2018). High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nature Communications*, *9*(1), 2022. <https://doi.org/10.1038/s41467-018-04451-x>
- Zhang, Yanfei, Ho, K., Keaton, J. M., Hartzel, D. N., Day, F., Justice, A. E., Josyula, N. S., Pendergrass, S. A., Actkins, K., Davis, L. K., Edwards, D. R. V., Holohan, B., Ramirez, A., Stanaway, I. B., Crosslin, D. R., Jarvik, G. P., Sleiman, P., Hakonarson, H., Williams, M. S. & Lee, M. T. M. (2020). A genome-wide association study of polycystic ovary syndrome identified from electronic health records. *American Journal of Obstetrics and Gynecology*, *223*(4), 559.e1-559.e21. <https://doi.org/10.1016/j.ajog.2020.04.004>
- Zhang, Yong, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. <https://doi.org/10.1186/gb-2008-9-9-r137>
- Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, *109*(4), 1193–1198. <https://doi.org/10.1073/pnas.1119675109>
- Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. & Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, *111*(4), E455–E464. <https://doi.org/10.1073/pnas.1322563111>