Getting to the source:
Measuring the primary response to acute transcription factor perturbation

Thomas Gabriel William Scott
Lusby, Maryland

A.B. Molecular Biology, Princeton University, 2015

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
December 2023

Abstract

Transcription factors (TFs) define and drive cellular identity and state. The coordinate activities of TFs dynamically regulate gene expression physiologically during development and in response to environmental cues. However, transcription is dysregulated in many disease states, including malignancies. Thus it is critical to study the mechanisms by which TFs regulate transcription of their target genes. This dissertation sets out to do so for the TF TRPS1 in a cell line model of luminal breast cancer. Chapter 1 serves as an introduction to breast cancer, the key TFs TRPS1 and estrogen receptor alpha (ER), and the main experimental approach, targeted protein degradation (TPD) to study TF function. Chapter 2 details the methods we use to analyze the data from nascent transcriptional profiling experiments. These methods are most specific to the precision nuclear run-on assay we use, but there are many parallels with the other genome-wide sequencing assays used in the rest of this dissertation. In Chapter 3, we use TPD to demonstrate that acute TRPS1 depletion redistributes ER binding genome-wide, both activating and repressing transcription of genes related to cancer cell fitness. In Chapter 4, we follow up on an initial observation made in the cell lines we generated, that acute TRPS1 depletion in three independent clones activates cholesterol biosynthesis gene transcription. In Chapter 5, I list my contributions to two other published works. Finally, Chapter 6 serves as a discussion of the conclusions drawn from this dissertation and the future directions of this work.

Acknowledgements

To Dr. Michael Guertin — thank you for guiding me through my graduate career. You have convinced me that sequence-specific transcription factors are the most important macromolecules in the cell. You have taught me how to thoughtfully analyze my own data and how to critically interpret published data. You allowed me the freedom to spread my wings and fly off in oftentimes unfruitful directions. You told me early on that we learn more from our failures than our successes, and I can tell you I have learned *a lot* over the past five years. Through it all, you never appeared to lose faith in my ability to succeed, and I drew confidence from our weekly meetings.

To the members of the Guertin Lab — thank you for the training, support, and feedback on my work. In particular, thank you to Dr. Kizhakke Mattada Sathyan for your technical wizardry and for teaching me how to genetically engineer cell lines and the ins and outs of PRO-seq. Thank you also to Drs. Jacob Wolpe and Arun Dutta for your critical feedback during lab meetings, as well as for being there to talk about the experience of graduate school.

To Dr. Daniel Gioeli — thank you for stepping up to serve as my co-mentor these past few years. Our meetings critically developed my abilities to question my assumptions and to articulate my hypotheses clearly. You pushed me to think about the broader biological context and impact to cancer patients beyond the confines of the nucleus. You forced me to provide more rationale for an experiment than simply, "I want to light this thing on fire just to see what happens."

To the members of the Gioeli Lab — thank you for making me feel at home in my new lab environment. Thank you for engaging with my data in lab meetings. Special thank you to Devin Roller for helping me with literally anything I asked about.

To Drs. David Auble, Kevin Janes, and Todd Stukenberg — thank you for serving on my dissertation committee. Your thoughtful critiques helped guide my project from amorphous to concrete.

To the Department of Biochemistry and Molecular Genetics — thank you for providing a strong training environment. The questions I received after presenting my own work in the research club

sparked new insights and directions for my work. Thank you to Dr. Anja Belinsky for your focus on community-building and graduate student morale. Enormous thank you to the administrative staff, especially Bill Garmer, Helen Norfleet-Shiflett, Nancy Rush, and Carolyn Smith, for all you work behind the scenes to make my life easier.

To the Cell and Molecular Biology training grant — thank you for exposing me to a diversity of perspectives from departments around grounds. Thank you to Drs. Todd Stukenberg and Bryce Paschal, as well as Bill Garmer, for providing me with food for thought and actual food.

To the Medical Scientist Training Program — thank you for your support and training throughout this long path towards becoming a physician-scientist. Thank you to Dr. Dean Kedes for admitting me into this program and giving me a chance to pursue this career. Thank you to Dori Williams, Ashley Woodard, and Kate Creveling for ensuring this operation runs smoothly.

To the National Institutes of Health and the United States taxpayers — thank you for the financial support of all the work in this dissertation.

To my former scientific mentors — thank you for believing in me and for advancing my technical skills and scientific thinking. Thank you specifically to Drs. Jose-Miguel Yamal, Donald Rao, Carlos Brody, and James Bradner. I would not have been able to pursue this career without the valuable experience gained in your laboratories.

To my friends — thank you for your support and sometimes necessary distraction from work. I especially thank my MSTP cohort for your *esprit de corps*.

To my family — thank you for the love and support over the years. Thank you to my parents for encouraging me to follow my dreams and for only seeming to care that I am happy in my career path. Thank you to my grandparents for many summers of fun. Thank you to my aunts and uncles for giving me books as presents, even when I thought I was too cool to read for pleasure.

To my partner Brielle — thank you for your constant love and encouragement. I know I can always talk to you when I have had a difficult day, and you make me more optimistic about tomorrow. Thank you for planning our hiking adventures and for buying me more running shorts.

Contents

Abstract i

Acknowledgements ii

Contents v

List of Figures xiii

List of Abbreviations xv

1 Introduction 1

1.1 Breast cancer . . . 1

1.1.1 Epidemiology . . . 1

1.1.2 Genetics . . . 1

1.1.3 Tumor classification . . . 2

1.1.4 ER in breast cancer . . . 3

1.1.5 TRPS1 in breast cancer . . . 4

1.2 Targeted protein degradation to study transcription . . . 5

1.2.1 Rationale and alternative strategies . . . 5

Genetic perturbations . . . 5

Chemical perturbations . . . 6

1.2.2 Chemical genetic systems . . . 7

AID . . . 7

dTAG . . . 7

Other . . . 8

1.2.3 Examples from the literature . . . 8

BET family proteins . . . 8

Chromatin architectural proteins 9
Sequence-specific TFs 11
1.2.4 Tagging considerations and strategies 12
Exogenous or endogenous expression 12
Copy number 13
Cas9-expression 13
Amino- or carboxy-terminal tagging 13
Repair templates 14
Clone isolation 14
Screening clones 15
1.2.5 Best practices 15
Compare basal expression 15
Avoid the "Hook effect" 15
Measure nascent RNA 16
1.2.6 Conclusion 16

2 Processing and evaluating the quality of genome-wide nascent transcription profiling libraries 17

2.1 Preface 17
2.2 Author contributions 17
2.3 Abstract 18
2.4 Introduction 18
2.5 Software and Hardware Requirements 19
2.5.1 Dependencies, Software, and Scripts 19
2.5.2 Hardware 21
2.6 Genome and Annotation Downloads and Processing 22
2.6.1 Reference Genomes 22
2.6.2 Reference Gene Annotation 22

2.7 Processing PRO-seq Data . . . 26
2.7.1 Initialize Variables . . . 26
2.7.2 Preprocessing . . . 27
2.7.3 Processing Reads . . . 27
2.7.4 RNA Degradation Ratio Score . . . 30
2.7.5 Processing for Alignment . . . 30
2.7.6 Remove Reads Aligning to rDNA . . . 31
2.7.7 Genome Alignment . . . 32
2.7.8 rDNA Alignment Rate . . . 32
2.7.9 Mappability rate . . . 33
2.7.10 Complexity and Theoretical Read Depth . . . 33
2.7.11 Run-on Efficiency . . . 36
2.7.12 Estimate Nascent RNA Purity with Exon/Intron Density Ratio . . . 37
2.7.13 Remove intermediate files and zip raw sequencing files . . . 38
2.7.14 Pipeline Automation . . . 40
2.7.15 Plot all QC metrics . . . 46
2.8 Differential Expression with DESeq2 . . . 46
2.9 ER antagonists affect the same genes as ER agonists . . . 49
2.10 Conclusions . . . 51
2.11 Methods . . . 51
2.11.1 Cell culture . . . 51
2.11.2 Cell treatments . . . 51
2.11.3 Cell permeabilization for PRO-seq . . . 52
2.11.4 PRO-seq library preparation . . . 52
2.12 Data Access . . . 53

3 TRPS1 modulates chromatin accessibility to regulate estrogen receptor alpha (ER) binding and ER target gene expression in luminal breast cancer cells 56

3.1 Preface . . . 56
3.2 Author contributions . . . 56
3.3 Abstract . . . 57
3.4 Author Summary . . . 58
3.5 Introduction . . . 58
3.6 Results . . . 60
3.6.1 *TRPS1* is associated with breast cancer incidence and promotes breast cancer cell number accumulation . . . 60
3.6.2 Endogenously degron-tagged TRPS1 is rapidly degraded in T47D cells . 62
3.6.3 TRPS1 directly represses regulatory element activity . . . 63
3.6.4 TRPS1 directly represses transcription of target genes . . . 70
3.6.5 TRPS1 redistributes ER binding to modulate ER target gene transcription 73
3.6.6 TRPS1 activity is associated with breast cancer patient outcomes . . . 79
3.7 Discussion . . . 81
3.8 Methods . . . 87
3.8.1 GWAS and DepMap data visualization . . . 87
3.8.2 Cell culture . . . 88
3.8.3 Plasmid generation for gene editing . . . 88
3.8.4 dTAG-TRPS1 clone generation . . . 89
3.8.5 Western blotting . . . 89
3.8.6 ChIP-seq library preparation . . . 90
3.8.7 ChIP-seq analysis . . . 91
3.8.8 ATAC-seq library preparation . . . 92
3.8.9 ATAC-seq analysis . . . 92
3.8.10 PRO-seq library preparation . . . 93
3.8.11 PRO-seq analysis . . . 94
3.8.12 Genome browser visualization . . . 94
3.8.13 Cell number enumeration . . . 95

3.8.14 TRPS1 activity score and patient outcome stratification 95
3.9 Data Access 95
3.10 Competing Interest Statement 95
3.11 Acknowledgements 96

4 TRPS1 represses transcription of cholesterol biosynthesis genes and is associated with blood cholesterol traits 97

4.1 Preface 97
4.2 Author contributions 97
4.3 Abstract 98
4.4 Introduction 99
4.5 Results 100
4.5.1 TRPS1 represses cholesterol biosynthesis gene transcription 100
4.5.2 *TRPS1* is associated with blood cholesterol traits 100
4.5.3 Lipidomics normalization method affects results 103
4.5.4 TRPS1 depletion does not consistently affect mRNA expression of cholesterol biosynthesis genes 104
4.5.5 TRPS1 depletion does not consistently affect lipid droplet abundance . 104
4.5.6 TRPS1 depletion does not consistently affect cholesterol abundance . . 104
4.5.7 Cholesterol biosynthesis gene transcription is only transiently activated . 107
4.5.8 TRPS1 depletion does not significantly affect cholesterol abundance at earlier time points 108
4.6 Discussion 109
4.7 Methods 112
4.7.1 Cell culture 112
4.7.2 Cell treatments for PRO-seq 112
4.7.3 Cell permeabilization for PRO-seq 113
4.7.4 PRO-seq library preparation 113

4.7.5 PRO-seq analysis . . . 114
4.7.6 GWAS and eQTL data visualization and analysis . . . 114
4.7.7 Lipidomics . . . 114
4.7.8 RT-qPCR . . . 116
4.7.9 Lipid droplet staining . . . 117
4.7.10 Cholesterol assay . . . 117
4.8 Data Access . . . 118
4.9 Competing Interest Statement . . . 118
4.10 Acknowledgements . . . 118

5 Contributions to other projects **119**

5.1 ARF-AID: a rapidly inducible protein degradation system that preserves basal endogenous protein levels . . . 119
5.1.1 Abstract . . . 119
5.1.2 Contribution . . . 119
5.2 The androgen receptor does not directly regulate the transcription of DNA damage response genes . . . 120
5.2.1 Abstract . . . 120
5.2.2 Contribution . . . 120

6 Discussion **121**

6.1 Concordance between initial and steady-state ER activity . . . 121
6.1.1 Conclusions . . . 121
6.1.2 Future directions . . . 123
Which TFs mediate the modulation of ER target gene expression? . . . 123
Does TEAD1 activate ER target genes in the absence of estrogen? . . . 125
Does over-activation of ER target genes upon TRPS1 depletion contribute to the cell number defect? . . . 126
Do RUNX family members augment activation of ER target genes? . . . 128

Do KLF family members repress ER target genes in the absence of hormone? 129
Does ER consistently regulate *ESR1* expression across luminal breast cancer cell lines? . . . 129
6.2 Concordance between chromatin accessibility, ER binding intensity, and ER target gene expression . . . 130
6.2.1 Conclusions . . . 130
6.2.2 Future directions . . . 131
Is decreased chromatin accessibility necessary for the TRPS1-dependent decrease in ER binding intensity? . . . 131
Are the changes in ER target gene expression due to modulation of ER activity? . . . 131
6.3 TF redistribution as a general model . . . 132
6.3.1 Conclusions . . . 132
6.3.2 Future directions . . . 133
Does liganded ER protein abundance influence a redistribution model? . 133
Can we uncouple estrogen-induced transcriptional activation from repression? 134
6.4 Discordance between TRPS1 activity score and *TRPS1* expression . . . 134
6.4.1 Conclusions . . . 134
6.4.2 Future directions . . . 135
Are other *TRPS1* splice isoforms expressed in breast cancer cells? . . . 135
Is TRPS1 post-translationally regulated? . . . 136
6.5 Functional follow-up of GWAS hits . . . 137
6.5.1 Conclusions . . . 137
6.5.2 Future directions . . . 138
Is there an association between the breast cancer associated SNPs and *TRPS1* expression in breast tumor tissue? . . . 138
6.6 Discordance between nascent transcription and downstream assays . . . 138
6.6.1 Conclusions . . . 138

6.6.2 Future directions 139
Can we endogenously tag all *TRPS1* alleles in a cell line with high *TRPS1* expression? 139
Does TRPS1 regulate blood cholesterol traits *in vivo*? 140
6.7 Data Access 140

References **141**

List of Figures

2.1 Library insert size is a measure of RNA degradation 31
2.2 Library complexity captures information about PCR over-amplification and read depth requirements for a sample 34
2.3 Pause index is a measure of nuclear run-on efficiency 38
2.4 Exon density to intron density ratio is a measure of nascent RNA purity 39
2.5 A summary plot illustrates all quality control metrics and their respective recommended thresholds 47
2.6 Differential expression analysis quantifies transcriptomic changes upon treating T47D cells with estrogen for an hour 48
2.7 Principle component analysis of nascent transcription upon acute ER agonism or antagonism 54
2.8 ER antagonists affect the same genes as ER agonists 55

3.1 *TRPS1* is associated with breast cancer incidence and promotes breast cancer cell fitness 61
3.2 Endogenously degron-tagged TRPS1 is rapidly degraded in T47D cells 63
3.3 Fraction of reads in peaks (FRiP) for ChIP-seq libraries 64
3.4 Strand cross-correlation (CC) plots for ChIP-seq libraries 65
3.5 Fragment size distribution plots for ATAC-seq libraries 66
3.6 Plots of signal enrichment around TSS's for ATAC-seq libraries 67
3.7 TRPS1 directly represses regulatory element activity 68
3.8 Quality control metrics for PRO-seq libraries 71
3.9 Bidirectional transcription at TRPS1 peaks increases upon TRPS1 depletion . . 72
3.10 Change in chromatin accessibility at ATAC-seq peaks without bidirectional transcription 73

3.11 TRPS1 directly represses transcription of target genes 74
3.12 Acute estrogen treatment identifies direct ER target genes in T47D cells 75
3.13 TRPS1 redistributes ER binding to modulate ER target gene transcription . . . 77
3.14 Bidirectional transcription at TRPS1 peaks increases upon TRPS1 depletion . . 78
3.15 TRPS1 activity is associated with breast cancer patient outcomes 80
3.16 Bidirectional transcription at TRPS1 peaks increases upon TRPS1 depletion . . 81
3.17 TRPS1 activity is associated with breast cancer patient outcomes specifically for ER-positive and Luminal A tumors 82
3.18 Genes differentially expressed after 24 hours of TRPS1 depletion are associated with breast cancer patient outcomes 83

4.1 TRPS1 represses cholesterol biosynthesis gene transcription 100
4.2 *TRPS1* is associated with blood cholesterol traits 102
4.3 Lipidomics normalization method affects results 103
4.4 TRPS1 depletion does not consistently affect mRNA expression of cholesterol biosynthesis genes 105
4.5 TRPS1 depletion does not consistently affect lipid droplet abundance 106
4.6 TRPS1 depletion does not consistently affect cholesterol abundance 107
4.7 Cholesterol biosynthesis gene transcription is only transiently activated 108
4.8 TRPS1 depletion does not significantly affect cholesterol abundance at earlier time points 109
4.9 Cholesterol assay controls do not demonstrate the expected effects 110

6.1 ER target genes follow one of three expression patterns across media conditions and ER activity modulation 123
6.2 TEAD and RUNX family TFs change in expression between media conditions . 124
6.3 YAP1 binding intensity is increased genome-wide upon acute TRPS1 depletion 127
6.4 Estrogen-activated genes are closer to ER binding sites 133

List of Abbreviations

AID	Auxin-inducble degron
ATAC-seq	Assay for transposase-accessible chromatin with sequencing
BET	Bromodomain and extra-terminal domain
Cas9	CRISPR-associated protein 9
CBFB	Core-binding factor subunit beta
CC	Strand cross-correlation
ChIP-seq	Chromatin immunoprecipitation with sequencing
cDNA	Complementary DNA
CoREST	Corepressor for RE1 silencing transcription factor
CRBN	Cereblon
CRISPR	Clustered regularly interspaced palindromic repeats
CtBP	c-terminal binding protein
CTCF	CCCTC-binding factor
CVD	Cardiovascular disease
dES	Differential enrichment score
dTAG	dTAG-13 and dTAGV -1 at 50nM each
ENSG	Ensembl gene ID
ENST	Ensembl transcript ID
eQTL	Expression quantitative trait locus
ER	Estrogen receptor α
Estrogen	17-β-estradiol
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
FKBP12	12-kiloDalton FK506-binding protein
FRiP	Fraction of reads in peaks
GR	Glucocorticoid receptor
GRO-seq	Global run-on and sequencing
GSEA	Gene set enrichment analysis
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
HDL	High-density lipoprotein
HER2	Human epidermal growth factor receptor 2
HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase
IHC	Immunohistochemistry
ISTD	Internal standard normalization
kb	Kilobase
KLF	Krüppel-like factor
LDL	Low-density lipoprotein
LEXY	Light-inducible nuclear export system
mRNA	Messenger RNA

MST1/2	Mammalian sterile 20-like kinases 1 and 2
NuRD	Nucleosome remodeling and deacetylase
ORA	Over-representation analysis
PCR	Polymerase chain reaction
PE1	Paired end 1
PE2	Paired end 2
Pol II	RNA polymerase II
PR	Progesterone receptor
PRO-seq	Precision run-on sequencing
PCSK9	Proprotein convertase subtilisin/kexin type 9
PQN	Probabilistic quotient normalization
RE	Regulatory element
RT-qPCR	Reverse transcription quantitative PCR
RNAi	RNA interference
RNA-seq	mRNA sequencing
RUNX	Runt-related
SRC-1	Steroid receptor coactivator-1
SMASh	Small molecule–assisted shutoff
SNP	Single nucleotide polymorphism
SP	SP1-like
TAD	Topologically associated domain
TAZ	Transcriptional Activator with PDZ binding domain
TEAD	TEA/ATTS domain
TF	Transcription factor
TNBC	Triple-negative breast cancer
TPD	Targeted protein degradation
VHL	von Hippel-Lindau
TSS	Transcription start site
TT-seq	Transient transcriptome sequencing
UMI	Unique molecular identifier
YAP1	Yes-associated protein
YY1	Yin Yang 1

Chapter 1

Introduction

1.1 Breast cancer

1.1.1 Epidemiology

Breast cancer is one of the most common forms of cancer in the United States, with over 300,000 cases diagnosed and over 40,000 associated deaths each year. This represents a lifetime risk of about 1 in 8 for a diagnosis of breast cancer and about 1 in 40 for death due to breast cancer for women in the United States [1]. Globally, breast cancer is the most common form of cancer among women, with over 2.2 million new cases and over 600,000 deaths each year [2]. Though this disease can affect individuals of any sex and gender, the vast majority of breast cancer patients are female. Breast cancer incidence increases with age, with over 80% of invasive breast cancers in the United States diagnosed in patients over the age of 50 [1]. There are racial disparities in breast cancer outcomes in the United States — White, non-Hispanic individuals have the highest incidence, but Black individuals have the highest mortality rates [1].

1.1.2 Genetics

A small proportion of breast cancers can be attributed to a heritable mutation in a single gene [3]. Germline mutations in *BRCA1* or *BRCA2* decrease the homology-directed repair of DNA damage and lead to increased incidence of breast, ovarian, and fallopian tube cancers [4–6]. Rarer mutations increase the risk of breast cancer as a part of a broader syndrome, including *PTEN* mutation in Cowden Syndrome, *TP53* mutation in Li-Fraumeni Syndrome, *CDH1* mutation in Hereditary Diffuse Gastric Cancer, and *STK11* mutation in Peutz-Jeghers Syndrome [3, 7–10].

However, most breast cancers are sporadic in nature. There are few recurrent somatic mutations in breast cancers, with the most common being *TP53*, *PIK3CA*, and *GATA3* [11]. *TP53*, the most frequently mutated gene in cancer, encodes the tumor suppressor P53 that

promotes DNA repair or apoptosis in response to DNA damage [12]. *PIK3CA* encodes PI3K, an intracellular kinase that canonically responds to growth factors to activate the AKT-mTOR pathway and promote cell growth and proliferation [13]. *GATA3* encodes for a transcription factor (TF) in the same family as TRPS1, the star of this dissertation.

1.1.3 Tumor classification

Far from a monolithic disease, breast cancer has been divided into distinct molecular subtypes based on immunohistochemistry (IHC) or transcriptional patterns, and these inform prognosis and treatment [14, 15].

IHC was the original method to classify breast tumors and is still used today [16]. Generally, tumor sections are stained to assess for the expression of estrogen receptor alpha (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Tumors are classified as ER-positive and/or PR-positive if over 1% of the cells express the corresponding receptor [16]. Over 70% of breast cancer cases are ER-positive [17]. For patients with these tumors, hormonal therapies inhibiting ER activity such as the selective ER modulator tamoxifen are first-line therapeutic options [18].

Similarly, tumors are classified as HER2-positive if over 10% of the cells exhibit circumferential membrane staining [19]. As *ERBB2*, the gene encoding HER2, is often genetically amplified in HER2-positive tumors, HER2-positivity can alternatively be determined via *in situ* hybridization if there are more than six signals per cell [19]. For patients with these tumors, HER2-targeting therapies such as the monoclonal antibody trastuzumab improve outcomes [20].

Tumors lacking signal for any of the above three receptors are termed triple-negative breast cancer (TNBC) and generally carry the worst prognosis among breast cancers [21]. Additional IHC markers, such as cytokeratin 5/6 and epidermal growth factor receptor have been used to further stratify these TNBC tumors [22].

A newer method to classify breast tumors is through gene expression patterns. Multiple gene sets and platforms exist for profiling transcript abundance exist, including PAM50, OncotypeDx, and MammaPrint [23–26]. Based on the PAM50 classification, the five predominant intrinsic subtypes of breast tumors are Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like, sometimes referred to as Claudin-low [23, 27]. The Luminal A subtype generally corresponds to IHC status as ER-positive, PR-positive, and low proliferation index as measured by Ki67 staining and generally carries the best prognosis among breast cancers [28].

1.1.4 ER in breast cancer

The work presented in this dissertation will focus on ER-positive breast cancer using a breast cancer cell line representing the Luminal A molecular subtype of breast cancer.

High lifetime exposure to endogenous estrogen is a strong risk factor for breast cancer incidence [29]. ER is transcription factor (TF) in the nuclear hormone receptor family that is activated by estrogens such as 17-beta-estradiol, a steroid hormone produced in the ovaries and peripheral adipose tissue. In cells that express ER, estrogen binds and activates ER. Upon activation, ER homodimerizes and binds to its cognate DNA motif on DNA and recruits coactivators like steroid receptor coactivator-1 (SRC-1), CBP, and P300 [30–33]. DNA-bound ER increases transcription of genes related to cell growth and proliferation [34–37]. Estrogen signaling through ER is a major driver of cell growth both physiologically in the breast and pathophysiologically in breast cancer.

Accordingly, most medical therapies for ER-positive breast cancer are targeted to decrease the activity of ER. This can be accomplished with endocrine therapy that decreases endogenous estrogen production, inhibits coactivator interaction with ER, or degrades ER [38–41]. While these treatments are initially effective, unfortunately 30% of tumors relapse and become resistant to endocrine therapy [42]. Around 90% of relapsed tumors maintain ER-positivity, and many patients respond to second line therapies targeting ER, indicating that ER activity remains an

important driver of tumor growth [43–46]. Thus, new treatments targeting proteins that act downstream of ER or that regulate ER signaling are needed.

1.1.5 TRPS1 in breast cancer

TRPS1 was first described as the gene mutated in cases of tricho-rhino-phalangeal syndrome, an autosomal dominant disorder characterized by developmental abnormalities of the hair, nose, and fingers [47]. *TRPS1* is crucial for the proper development of several tissues, including hair, bone, and kidney [48, 49]. As with many developmentally important genes co-opted during the process of cancer initiation and progression, *TRPS1* is commonly over-expressed or amplified in breast tumors, both relative to normal tissue and relative to other tumor types [50–54].

TRPS1 over-expression has been shown to increase the colony formation ability of a non-transformed mammary epithelial cell line [55]. In addition, in multiple xenograft mouse models, TRPS1 loss inhibits tumor progression [56–58]. On the other hand, TRPS1 appears to act as a tumor suppressor in genetically engineered mouse models. *TRPS1* was a hit in a transposon mutagenesis screen in a *PTEN*-null mouse model of TNBC and in another transposon mutagenesis screen in a *CDH1*-null mouse model of invasive lobular carcinoma [53, 59].

Like GATA3 referenced above, TRPS1 is in the GATA-family of TFs, which share a conserved GATA-like zinc finger that recognizes (A/T)GATA(A/G) motifs on DNA [60]. Unlike GATA3 and the rest of the members of this family, TRPS1 also has two carboxy-terminal Ikaros-like zinc fingers [61]. These zinc fingers recruit corepressor complexes, including the Nucleosome Remodeling and Deacetylase (NuRD) and corepressor for RE1 silencing transcription factor (CoREST) complexes, to directly repress transcription of TRPS1 target genes [53, 57, 58, 61–63].

TRPS1 knockdown has been shown to influence the activity of other TFs, as well as other cellular processes. TRPS1 was a hit in an unbiased screen to identify repressors of YAP1 activity in a luminal breast cancer cell line, and TRPS1 knockdown led to a genome-wide activation of YAP target genes [57]. Paradoxically, TRPS1 knockdown has been reported to lead to both a

genome-wide repression of ER target genes as well as a genome-wide increase in ER binding [53]. In addition, TRPS1 knockdown in a TNBC cell line led to failures of mitotic progression and chromosome segregation [56]. Consistent with this effect, TRPS1 knockout mouse embryo chondrocytes displayed hyperacetylated histones during mitosis and chromosome segregation defects [64]. Several studies have noted an increase in epithelial to mesenchymal transition and genome instability upon TRPS1 knockdown [65–68].

In sum, extended TRPS1 depletion appears to have pleiotropic effects on many aspects of breast cancer cell and tumor biology. We set out to study the most upstream effects of TRPS1 perturbation, the primary and direct effects of TRPS1 on transcription. To do so, we needed a strategy to rapidly perturb TRPS1 activity. We used targeted protein degradation (TPD), for reasons we describe in the next section.

1.2 Targeted protein degradation to study transcription

The study of transcription has been aided in recent years by the ability to rapidly perturb TFs and other chromatin associated factors using TPD strategies. In this section, we will provide an introduction to this topic, in the hope that the reader will consider adopting these strategies to study their favorite TFs.

1.2.1 Rationale and alternative strategies

To study the function of a gene, various methods have been developed to activate or inactivate its function in the cell. Ideally, a perturbation is rapid, specific, and broadly applicable to any target gene. Traditionally, genetic and chemical approaches have been used, which each have strengths and limitations.

Genetic perturbations

Classically, genetic knockouts and RNA interference (RNAi)-based approaches target the DNA or RNA encoding a gene of interest for knockout or knockdown, respectively. For example, clustered

regularly interspaced palindromic repeats (CRISPR)-based methods use a short guide RNA to target a nuclease like CRISPR associated protein 9 (Cas9) to the 5′-end of the DNA encoding a gene. This creates a double-stranded break, which is repaired in an error-prone fashion, leading to frameshift mutations and effective gene knockout [69–71]. Small interfering RNA or short hairpin RNA can be transfected or transduced into cells to target RNA encoding a gene. This leads to degradation of the RNA via the RNA-induced silencing complex and knockdown of gene expression [72–76]. By using nucleic acids to bind specifically to the DNA or RNA encoding a gene, genetic perturbations can be both specific for the target gene and generalizable to most any gene.

However, these methods tend to take days to affect gene expression, which can lead to undesireable effects. TFs generally regulate transcription of many primary target genes, some of which regulate transcription of secondary effect genes. If a genetic perturbation takes several days to take effect, the resulting transcriptional changes from baseline will include not only the primary effects, but secondary effects and beyond. Furthermore, many lineage specific TFs are essential for cell viability and proliferation [77]. Measuring transcription after extended TF knockdown or knockout will represent a new cellular state, with many genes changing beyond the direct TF target genes.

Chemical perturbations

An orthogonal method to perturb TF activity is through the use of small molecule chemical compounds. A few TFs, such as the estrogen receptor (ER), have natural ligands [78–80]. Transcription-associated kinases, like CDK9, are targetable with enzymatic inhibitors [81, 82]. Transcriptional cofactors with hydrophobic pockets that can be bound with chemical probes, including BRD4, can be inhibited or targeted for proteasomal degradation through the use of TPD [83–85]. This strategy forms the basis for the techniques that are the main focus of this overview. TPD uses heterobifunctional small molecules, with one moiety targeting a protein of interest and another targeting an E3 ubiquitin ligase complex, connected by a flexible linker [86].

Upon addition of the compound, the target protein is brought into physical proximity with the E3 ubiquitin ligase complex and is polyubiquitinated and shuttled to the proteasome for degradation. These chemical perturbations are rapid, limited only by the diffusion of the small molecules into the cell and to their targets [86]. However, the number of TFs with known natural or synthetic ligands is limited, leaving many endogenous TFs untargetable via this approach.

1.2.2 Chemical genetic systems

Thankfully, the field of chemical genetics provides tools to genetically modify genes of interest to render them susceptible to chemical perturbation [87]. By combining the specificity and generalizability of genetic manipulation with the temporal acuity of chemical perturbations, chemical genetics can provide an ideal avenue to study TFs. Many systems have been described, with the most widely used being the auxin-inducble degron (AID) and the degradation tag (dTAG) systems [88, 89].

AID

Auxin is a plant hormone that targets proteins for inducible degradation [90, 91]. Two components of the AID system has been coopted from plants — The AID tag, the IAA17 gene from plants, is fused to a target gene, and the E3 ubiquitin ligase TIR1 is heterologously expressed [88, 92, 93]. Upon the treatment with auxin, the fusion protein is brought into proximity with a TIR1-containing E3 ubiquitin ligase complex. Two limitations of this method are the requirement for an additional genetic manipulation to express TIR1 and a basal level of chronic knockdown in the absence of auxin [94]. The latter limitation has been addressed by several iterations of the AID system including mini-AID and ARF-AID [94–101].

dTAG

The dTAG system uses a similar strategy [89]. A mutant of the prolyl isomerase 12-kiloDalton FK506-binding protein (FKBP12) is fused to a target protein. This mutant allows for a bump-and-hole strategy such that a ligand for wildtype FKBP12 is modified with a sterically bulky

moiety, and the mutant FKBP12 contains a pocket for this moiety [102]. In this way, the modified ligand binds specifically to the fusion mutant FKBP12 protein and not the endogenous wildtype FKBP12. Several compounds can be used to induce proximity between the fusion protein and an endogenous E3 ubiquitin ligase. dTAG-13 contains a derivative of thalidomide that binds to cereblon (CRBN), and dTAG^{V}-1 uses a ligand for von Hippel-Lindau (VHL) [89, 103–105].

Other

A previous iteration of FKBP12-based tagging used a cryptic degron that can be exposed upon the addition of an FKBP12 ligand [106]. Another thalidomide-induced neosubstrate of CRBN, Sal-like protein 4, can be fused to a target protein and degraded upon the addition of 5-hydroxythalidomide [107]. The small molecule–assisted shutoff (SMASh) tag excises itself at baseline but can be retained to degrade the tagged protein upon the addition of HCV protease inhibitors [108]. The generic HaloTag can be functionalized for multiple purposes, such as imaging and TPD [109–113]. Finally, as an example of a non-TPD form of chemical genetics, light-inducible nuclear export system (LEXY)-tagged proteins can be rapidly exported from the nucleus under optogenetic control [114].

1.2.3 Examples from the literature

Here we present examples of how the use of TPD has extended our understanding of the mechanisms of transcriptional regulation by specific sets of chromatin-associated factors.

BET family proteins

The bromodomain and extra-terminal domain (BET) family of proteins that recognize acetylated histones are popular targets in the field of chemical biology, beginning with the chemical probe JQ1 [83]. Many E3 ligases have been coopted for TPD, and BET proteins are often the first targets for proof of principle reporting [84, 115, 116]. With the plethora of data generated from these tool compounds, we have been able to gain new insights into BET protein function in transcriptional regulation. The distribution of BET protein binding across the genome is

asymmetric, with a disproportionate amount of protein occupying a small fraction of the binding sites, often referred to as "super enhancers" [117]. Treatment with BET bromodomain inhibitors like JQ1 predominantly affects transcription of genes controlled by these super enhancers [83]. In contrast, BET protein degradation with the potent degrader dBET6 led to a profound loss of transcriptional elongation genome-wide [85]. The P-TEFb complex, consisting of the cyclin-dependent kinase CDK9 and the cyclin CCNT1, regulates the release of RNA polymerase II (Pol II) into productive elongation [118]. The prototypical BET protein, BRD4, physically interacts with P-TEFb and had previously been thought to contribute to P-TEFb recruitment to DNA [119–121]. However, genome-wide depletion of BET proteins had no effect on CDK9 occupancy at promoters or enhancers or nuclear levels of CCNT1, suggesting that BET proteins regulate P-TEFb activity via a mechanism distinct from recruitment to DNA [85].

While JQ1 inhibits and dBET6 degrades the entire family of BET proteins, individual members can be specifically targeted using inducible degron tags to differentiate their functions. In doing so, we have learned that the genome-wide effects on elongation are specific to BRD4 depletion [122]. Furthermore, the decrease in Pol II occupancy at enhancer regions observed with dBET6 treatment was recapitulated specifically by BRD2 depletion and not by depletion of BRD3 or BRD4 [122]. Finally, while BET bromodomain inhibitors specifically target the ability of BET proteins to recognize acetylated lysines, TPD acutely depletes the entire protein and allows for the assessment of non-bromodomain-dependent functions of the BET family. Indeed, BRD4 depletion and complementation with deletion mutants revealed that BRD4 is able to stimulate transcriptional elongation via its carboxy-terminal region that interacts with P-TEFb without the requirement for its bromodomains [123].

Chromatin architectural proteins

The eukaryotic genome is highly organized in three-dimensional space. The first use of the genome-wide chromosome conformation capture technology Hi-C revealed an example of this organization at the broadest scale, with two main compartments, active and inactive, segregating

in their contact frequencies [124]. On a megabase scale, the genome is divided into topologically associated domains (TADs), alternating regions of the chromosome with contact frequencies much higher within the region than with neighboring regions [125]. The TF CCCTC-binding factor (CTCF) often demarcates the boundaries of these TADs. The motor protein cohesin is also found at these boundaries and has been shown to extrude DNA until it reaches an impasse like a stably-bound CTCF protein [126]. At the kilobase scale, there can be points of contact between two distal regions of chromatin, such as an enhancer and a promoter, that are thought to represent looping [127]. As with TADs, these loops are often bordered by CTCF and cohesin binding sites.

It is difficult to test whether CTCF is causally related to this genome organization, as its knockout in mice is embryonically lethal, and its expression is essential for cell proliferation [128, 129]. However, the rapid depletion of CTCF with the AID system in multiple cell lines demonstrated that CTCF is required for looping and the maintenance of TADs but dispensable for the segregation of the active and inactive compartments of the genome and even transcription of most genes [130, 131]. Furthermore, as with BRD4 above, complementation experiments performed by inducing exogenous expression of wildtype or mutant CTCF while depleting endogenous CTCF allowed for the characterization of mutations in zinc fingers outside of the core DNA binding zinc finger array [132]. Of technical significance, a detailed study of acute CTCF degradation using both sequencing and imaging readouts showed that the kinetics of CTCF depletion differ across the genome, reinforcing the importance of choosing a reasonable time point after inducing degradation [133].

As with CTCF, acute depletion of cohesin reduces looping genome-wide, a result that is rapidly reversible upon re-expression of cohesin, but has only a minor effect on transcription [134]. Together, these studies and others of the chromatin architectural proteins CTCF and cohesin call into question the functional significance of TADS and loops for transcription of most genes. However, one report using acute depletion of the cohesin release factor WAPL produced a model in which cohesin release tends to occur distally to lineage-specific genes and that this

cohesin turnover maintains the correct cohesin dynamics required for the promoter-enhancer contacts to regulate transcription of these lineage-specific genes [135]. Furthermore, depletion of individual subunits of the multiprotein cohesin complex have distinct effects, suggesting cohesin may perform distinct roles across its binding sites [136].

Yin Yang 1 (YY1) is a TF similar to CTCF in that it homodimerizes and promotes enhancer-promoter looping [137, 138]. Named in part for its activating and repressive effects on transcription after chronic knockdown, YY1 was recently perturbed more acutely using TPD [139]. In this study, YY1 depletion greatly reduced enhancer-promoter looping. In addition, and in contrast to CTCF and cohesin depletion, there were large changes to nascent transcription, with both activation and repression of primary response genes. The repression is more easily explained by the loss of activating promoter-enhancer contacts. The authors do not perform much follow-up analysis on the activated genes but attribute this activation to the profound changes in genome architecture inappropriately positioning of transcriptional regulators.

Sequence-specific TFs

The proto-oncogene *MYC* is over-expressed in the majority of human cancers [140]. When bound to canonical E-boxes with its binding partner MAX, MYC regulates many genes that drive cell growth and proliferation. MYC target genes have been queried using multiple methods, including inducible over-expression coupled with nascent RNA profiling, genome-wide binding assays, and expression correlation [141]. These target genes vary across cell types, but a common core set of genes are involved in nucleolar function and ribosomal biogenesis [141]. There are conflicting reports on whether MYC acts as an activator at specific genes, an amplifier of all genes, or a direct repressor at some genes [142–146]. The acute depletion of an AID-tagged MYC within 30 minutes, coupled with nascent RNA sequencing, allowed for the identification of primary MYC-responsive genes [147]. 98% of these genes were repressed, indicating that the direct effect of MYC regulation is transcriptional activation of a fraction of the expressed genes in a cell [147].

OCT4 is one of the four "Yamanaka" factors, the expression of which is sufficient to

reprogram differentiated fibroblasts into induced pluripotent stem cells [148]. However, the genes OCT4 regulates in pluripotent stem cells are difficult to identify, as the half-life of its protein and mRNA are much too long for traditional knockdown methods to isolate the primary effects of depletion [149]. A recent study compared extended knockdown to rapid depletion with TPD and found that only the latter was able to identify that the primary effect of OCT4 on transcription is the activation of pluripotency factors and that the delayed activation of trophoblast-associated genes is a secondary effect of OCT4 depletion [149].

A key takeaway from these and other studies is that these TFs directly activate transcription of their target genes. The growing list of TFs that can be acutely perturbed provides evidence that TFs do not activate some direct targets and repress others. Another theme is that, in contrast to extended knockdown, acute depletion of most sequence-specific TFs affects transcription of a limited number of primary response genes [150–152].

1.2.4 Tagging considerations and strategies

There are several choices to make when generating a strategy for tagging a TF for TPD. Here we provide an overview of some of the key choices and our recommendations for how to approach them.

Exogenous or endogenous expression

The fusion protein can be expressed exogenously, for example by lentiviral transduction before or after knockout of the endogenous gene [153]. However, this involves the cells to go through a period of over- or under-expression of the TF. In addition, the exogenously expressed gene is constitutively expressed and not under the transcriptional regulation of its endogenous locus. These differences can lead to chronic changes in the abundance, localization, or interactions of the protein independent from the acute perturbation we plan to induce [154]. Thus, if possible, we recommend tagging at the endogenous locus. We use CRISPR-Cas9 to induce a double-stranded break at a targeted DNA sequence and provide a repair template for the insertion of the tag [94].

Copy number

When endogenously tagging your TF of interest, it is easier to use a cell line without DNA copy number gain of the gene. Based on our experience, we would not recommend using a cancer cell line with a gene copy number of four, as is the case for *TRPS1* in many luminal breast cancer cell lines. In addition to the efficiency decreasing exponentially with additional alleles, an effect that can be mitigated by scaling up the initial transfection, there is also the possibility of a threshold of DNA breaks within each cell being surpassed, leading to cell cycle arrest or apoptosis [155]. When choosing a cell line, the Cancer Cell Line Encyclopedia hosted on the Cancer Dependency Map project website lists absolute copy number for each gene [77, 156]. However, many cancer cell lines display chromosomal instability, so the risk remains that the cells used in a tagging endeavor may harbor additional copies at the outset or acquire them during clonal isolation. Much of the choice of cell type is driven by the expression of the lineage-specific TF under study. When studying general transcriptional regulatory mechanisms, though, one useful cell line is the chronic myelogenous leukemia cell line HAP1 or relatives thereof, which are haploid for most genes in the genome [157–161].

Cas9-expression

Using a cell line that constitutively or inducibly expresses Cas9 can be helpful because it reduces the necessary genetic material to be transfected. If no such cell line exists but you will be doing substantial genome editing, it may save time to first generate such a clone [162]. The drawbacks are that this step takes time and forces the cells through an additional bottleneck that may skew the downstream results.

Amino- or carboxy-terminal tagging

Though the described inducible degron tags are small, tagging proteins can have unpredictable consequences on their function [163]. If feasible, we would recommend targeting both termini in parallel. If there is a critical protein domain near one terminus, targeting the other terminus may

be more likely to succeed, though there is a flexible linker that can potentially separate the tag and avoid catastrophic interference. Of note, one welcome outcome of targeting the 5′-end of the gene when endogenously tagging is the potential for a knockout of any untagged alleles.

Repair templates

One form of repair template is an additional plasmid with long homology arms [119]. This strategy is helpful for large insertions. We have used this strategy to insert TIR1 into a safe harbor locus [94]. A drawback of this approach is the amount of genetic material that needs to be transfected when using an additional plasmid.

We have had success with a polymerase chain reaction (PCR) product with 50 base pair homology arms [94]. This can be generated from a generic plasmid template with the tag and selection marker. The specificity comes from the PCR primers used. Importantly, these should be ordered with phosphorothioate modifications at the 5′-ends to improve resistance to cellullar exonucleases [164, 165].

A third option is the CRIS-PITCh system, which uses microhomology regions (5-25 base pairs) flanking the insert [89, 166]. The insert is transfected as a part of a plasmid and is excised as linear DNA via Cas9 once in the cell. A new plasmid needs to be generated for each repair template, in contrast to the PCR product method above. However, once generated the plasmid is easier to amplify and does not require PCR and gel purification.

Clone isolation

When isolating single-cell-derived clones, three common strategies are manual colony picking after cell divisions have occurred, limiting dilution of cells shortly after transfection, and fluorescence-activated cell sorting (FACS) [167–169]. Colony picking is more labor intensive, but limiting dilution uses more plates, as there are many empty wells. FACS uses specialized equipment and exposes cells to potential contamination but rapidly isolates single cells into individual wells. Performed shortly after transfection, the GFP from the Cas9-expressing plasmid can be used to

enrich for transfected cells. Performed at a later time point, FACS can also be used to isolate clones expressing fluorescent markers indicating genomic integration.

Screening clones

When screening clones, integration can be assayed via PCR or Western blot. PCR is faster and reveals whether DNA has been inserted, but a Western blot measures the functional outcome at the protein expression level. We recommend screening by Western blot and performing follow-up analysis of the DNA to determine the sequence of each allele.

1.2.5 Best practices

Here we present practical advice on how best to use cell lines with TFs tagged for TPD.

Compare basal expression

As mentioned above, the ideal perturbation of a TF only occurs upon acute induction of degradation. Before the experiment, the tagged cells ideally should have expressed the TF at endogenous levels throughout the tagging process. When presenting acute TF depletion in a Western blot, it is informative for the reader to compare the basal expression to that of parental cells. Depletion from 10% of parental expression to 0% is different than depletion from 100% to 0%, though both would look similar on a Western blot without the reference point of the parental cells. As referenced above, several iterations of the AID system have been developed to reduce basal degradation. A beautiful example of a Western blot establishing the expression of several different tagged proteins across multiple clones can be found in Figure 1B of [122].

Avoid the "Hook effect"

Heterobifunctional molecules like dTAG-13 are subject to the "Hook effect", in which the dose response curve is not monotonic [84, 170, 171]. At high concentrations of compound, both the E3 ubiquitin ligase and the tagged target TF become saturated with low ternary complex formation and decreased degradation rate. As such, it is helpful to initially test a range of doses

at an extended time point, such as 48 hours, to choose a concentration in the center of the range of maximal effect. After this dose has been chosen, a time course experiment can be performed to determine how rapidly the TF can be depleted. Of note, "molecular glue" molecules like auxin cannot bind independently to each of their two targets and so are not subject to the Hook effect [172]. However, it may be wise to use the minimal dose necessary to achieve maximal degradation, as auxin can activate the aryl hydrocarbon receptor as an off-target effect [152]. Intriguingly, a recent report using artificial topological nanostructures allows for multivalent interactions between the proteins of interest and E3 ubiquitin ligases to counteract this "Hook effect" [173].

Measure nascent RNA

Acute TF depletion is best coupled with a rapidly-responsive readout of transcriptional activity. Changes in messenger RNA (mRNA) abundance lag behind changes in transcriptional rate, as they also depend on RNA turnover rates. Waiting until a time point late enough to detect significant changes in mRNA abundance by mRNA sequencing (RNA-seq) limits the benefit of rapid degradation in isolating the primary effects of TF depletion. Instead, nascent RNA sequencing methods, such as precision run-on sequencing (PRO-seq) or transient transcriptome sequencing (TT-seq) [174, 175], detect rapid changes in transcription rate.

1.2.6 Conclusion

We are in an exciting time for the study of TFs. Nuclear hormone receptors are well-studied via rapid activation, and transcription-associated kinases can be acutely inhibited with small molecules. More recently, the mechanisms of general transcription factors and architectural proteins have been elucidated via TPD. However, many hundreds of lineage-specific TFs remain to be studied! We hope that the reader comes away from this section ready to start tagging their favorite TFs.

Chapter 2

Processing and evaluating the quality of genome-wide nascent transcription profiling libraries

2.1 Preface

This chapter is adapted from a manuscript under review at *Methods in Molecular Biology*.

Thomas G. Scott, André L. Martins, Michael J. Guertin

2.2 Author contributions

All authors contributed to the conceptualization of the project and to the methodology. TGS performed the experiments and analyzed the data. TGS and MJG wrote the original draft of the manuscript. All authors reviewed and edited the manuscript.

Processing and evaluating the quality of genome-wide nascent transcription profiling libraries

Thomas G. Scott[a], André L. Martins[b], Michael J. Guertin[b,c]

[a]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America

[b]Center for Cell Analysis and Modeling, University of Connecticut, Farmington, Connecticut, United States of America

[c]Department of Genetics and Genome Sciences, University of Connecticut, Farmington, Connecticut, United States of America

2.3 Abstract

Precision run-on assays (PRO-seq) quantify nascent RNA at single nucleotide resolution with strand specificity. Here we deconstruct a recently published genomic nascent RNA processing pipeline (PEPPRO) into its components and link the analyses to the underlying molecular biology. PRO-seq experiments are evolving and variations can be found throughout the literature. The analyses are presented as individual code chunks with comprehensive details so that users can modify the framework to accommodate different protocols. We present the framework to quantify the following quality control metrics: library complexity, nascent RNA purity, nuclear run-on efficiency, alignment rate, sequencing depth, and RNA degradation.

2.4 Introduction

Genomic nascent RNA profiling assays, such as precision genomic run-on (PRO-seq) [176], quantify the precise position and direction of transcriptionally engaged RNA polymerases. Quantifying nascent RNA complements conventional RNA-seq by directly measuring active transcription. Steady-state RNA levels are influenced by RNA stability, so we can leverage the discordance

between RNA-seq and PRO-seq expression to estimate genome-wide RNA half-lives [177]. Bidirectional transcripts are a hallmark of enhancers and promoters. We can detect these short and unstable transcripts with PRO-seq to directly infer regulatory element activity as an orthogonal approach to chromatin accessibility assays [178, 179]. Similarly to regulatory elements, gene isoforms can vary between cell types and conditions. We can use RNA-seq to define splice variants and PRO-seq to identify differing primary transcript boundaries [180] and transcription start sites [181]. Additionally, PRO-seq sensitively detects immediate changes in transcription without the need for mature RNAs to accumulate or degrade. Lastly, nascent RNA profiling determines RNA polymerase density within all genomic features, such as promoter-proximal regions, gene bodies, and enhancers [152, 182, 183]. Changes in RNA polymerase distribution within these regions can inform on how various treatments and stimuli regulate steps in the transcription cycle [37, 152, 184]. Here, we describe quality control metrics that are used to determine if PRO-seq libraries are worth proceeding with these or other downstream analyses. New genomic nascent RNA-seq methodologies [147, 175, 179, 185, 186] necessitate flexible analysis workflows and standardized quality control metrics [187]. We present the workflow as deconstructed code that can be adapted to fit a diversity of protocols and experimental details.

2.5 Software and Hardware Requirements

Many processes, downloads, and software installations are reused throughout the analyses. Users should periodically check for updated annotations and new software releases.

2.5.1 Dependencies, Software, and Scripts

We present specialized software and scripts herein, but much of the workflow depends upon more general software. These general bioinformatic software tools are well-maintained and documented, so we provide short descriptions and the links below.

- `bedtools`: a comprehensive suite of tools that efficiently perform a wide range of operations on genomic intervals. [188]

- `bowtie2`: aligns sequencing reads to reference sequences. [189]
- `cutadapt`: removes a defined sequence, such as adapter sequence, from sequencing reads. [190]
- `fastq_pair`: outputs only sequencing reads that have a matched paired end read. [191]
- `FLASH`: merges paired end reads by detecting overlapping sequence. [192]
- `fqdedup`: removes duplicated sequences from FASTQ files. [193]
- `samtools`: a suite of tools for parsing and interfacing with high throughput sequencing data files. [194]
- `seqOutBias`: software that parses files and outputs desired formats with the option to correct enzymatic sequence biases. [195]
- `seqtk`: a multifunctional toolkit for processing sequence files, including trimming a defined number of bases from the ends of reads and reverse complementing sequencing reads. [196]
- `sratoolkit`: a suite of tools that interface with data deposited into the Sequence Read Archive.
- `wget`: retrieves files from a wide range of internet protocols.
- **R** packages:
  - `lattice`: graphics plotting package. [197]
  - `DESeq2`: statistical package for quantifying differences in counts-based genomics data. [198]

In addition, we developed the following software and R scripts to facilitate data analysis and graphical output. Below, we use `wget` to retrieve the software and scripts. The command `chmod +x` changes the permissions of the files to executable.

```
github=https://raw.githubusercontent.com/guertinlab
```

```
wget ${github}/fqComplexity/main/fqComplexity
wget ${github}/fqComplexity/main/complexity_pro.R
wget ${github}/Nascent_RNA_Methods/main/insert_size.R
wget ${github}/Nascent_RNA_Methods/main/pause_index.R
wget ${github}/Nascent_RNA_Methods/main/exon_intron_ratio.R
wget ${github}/Nascent_RNA_Methods/main/plot_all_metrics.R
wget ${github}/Nascent_RNA_Methods/main/differential_expression.R

wget ${github}/Nascent_RNA_Methods/main/PRO_normalization
wget ${github}/Nascent_RNA_Methods/main/normalization_factor.R
wget ${github}/Nascent_RNA_Methods/main/normalize_bedGraph.py

chmod +x insert_size.R
chmod +x fqComplexity
chmod +x complexity_pro.R
chmod +x pause_index.R
chmod +x exon_intron_ratio.R
chmod +x plot_all_metrics.R
chmod +x differential_expression.R

chmod +x normalize_bedGraph.py
chmod +x normalization_factor.R
chmod +x PRO_normalization
```

Next, move the software dependencies and R scripts to a directory within the `$PATH` variable.

2.5.2 Hardware

This workflow requires a single-core computer, 8GB of RAM, and 200GB hard drive space. However, more RAM and multiple cores will greatly reduce compute time.

2.6 Genome and Annotation Downloads and Processing

2.6.1 Reference Genomes

PRO-seq experiments have been performed in a variety of organisms, including yeast [199], Drosophila [176, 183], and humans [200]. Analysis of the data requires alignment to a reference genome annotation. The first step is to use `wget` to retrieve the reference genome. Many websites host the assembly data in FASTA format, such as the human genome build 38 shown below retrieved from the UCSC genome browser server [201]. The `gunzip` command unzips the reference genome file, and `bowtie2-build` indexes the file to allow for efficient alignment. The code also retrieves, unzips, and builds the human rDNA reference genome [202] so that we can calculate rDNA alignment rates as a metric for nascent RNA purity.

```
wget https://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.fa.gz
gunzip hg38.fa.gz
bowtie2-build hg38.fa hg38

wget https://github.com/databio/ref_decoy/raw/master/human_rDNA.fa.gz
gunzip human_rDNA.fa.gz
bowtie2-build human_rDNA.fa human_rDNA

# Compute mappability for the given read length and the k-mer that corresponds to
# each possible read alignment position.
# This is the most time-consuming step of the seqOutBias command but can be
# completed once before processing the sequencing data
seqOutBias seqtable hg38.fa --read-size=62
```

2.6.2 Reference Gene Annotation

The quality control metrics outlined herein require the counting of sequence reads that align to three genomic features: exons, introns, and promoter-proximal pause regions. Gene annotations are available from many sources, and we outline retrieval and parsing of GTF files from Ensembl

[203]. The Ensembl website (http://www.ensembl.org/index.html) contains the information for the latest release, which at the time of writing this manuscript is release 104 for hg38. After retrieving and unzipping the file, we parse out all exon 1 annotations. These coordinates include all annotated transcription start sites that correspond to different gene isoforms. Ensembl chromosome numbers do not include the preceding "chr", so the first `sed` command appends "chr" to the chromosome name for downstream compatibility with the reference genome chromosome names. The output is then piped to `awk`, which prints the following fields: chromosome coordinates in columns 1-3, Ensembl transcript ID (ENST), gene name, and strand. Subsequent `sed` commands drop the semicolon and quote characters from the gene and Ensembl IDs while editing the mitochondrial chromosome to match the reference genome, "chrM" replacing "chrMT." Finally, we sort the exon output by the first column (chromosome), then the second column (starting position), in ascending order. The gene annotations are processed similarly, except the Ensembl gene ID (ENSG) replaces ENST in column 4, and we sort the file by the fifth column (gene name).

```
release=104
file=Homo_sapiens.GRCh38.${release}.chr.gtf.gz

wget http://ftp.ensembl.org/pub/release-${release}/gtf/homo_sapiens/${file}
gunzip $file

# extract all exon 1 annotations
grep 'exon_number "1"' Homo_sapiens.GRCh38.${release}.chr.gtf | \
    sed 's/^/chr/' | \
    awk '{OFS="\t";} {print $1,$4,$5,$14,$20,$7}' | \
    sed 's/";//g' | \
    sed 's/"//g' | sed 's/chrMT/chrM/g' | \
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.tss.bed

# extract all exons
```

```
grep 'exon_number' Homo_sapiens.GRCh38.${release}.chr.gtf | \
    sed 's/^/chr/' | \
    awk '{OFS="\t";} {print $1,$4,$5,$14,$20,$7}' | \
    sed 's/";//g' | \
    sed 's/"//g' | sed 's/chrMT/chrM/g' | \
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.all.exons.bed

# extract all complete gene annotations, sorted for use with join
awk '$3 == "gene"' Homo_sapiens.GRCh38.${release}.chr.gtf | \
    sed 's/^/chr/' | \
    awk '{OFS="\t";} {print $1,$4,$5,$10,$14,$7}' | \
    sed 's/";//g' | \
    sed 's/"//g' | sed 's/chrMT/chrM/g' | \
    sort -k5,5 > Homo_sapiens.GRCh38.${release}.bed

# extract all complete gene annotations, sorted for use with bedtools map
awk '$3 == "gene"' Homo_sapiens.GRCh38.${release}.chr.gtf | \
    sed 's/^/chr/' | \
    awk '{OFS="\t";} {print $1,$4,$5,$10,$14,$7}' | \
    sed 's/";//g' | \
    sed 's/"//g' | sed 's/chrMT/chrM/g' | \
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}_sorted.bed
```

The following operations output: 1) a set of exons that excludes all instances of first exons, 2) all potential pause regions for each gene, and 3) all introns. There are many exon 1 gene annotations depending on gene isoforms, and the upstream most annotated TSS is not necessarily the most prominently transcribed isoform. We define the pause window for a gene as position 20 - 120 downstream of the most prominent TSS. The most prominent TSS is determined by calculating the density in this 20 - 120 window for all annotated TSSs for each gene and choosing the TSS upstream of the most RNA-polymerase-dense region for each gene.

In order to define these windows, we use `mergeBed` to collapse all overlapping exon

intervals and then `subtractBed` to exclude all first exon coordinates from the merged exon file. Since `mergeBed` drops the gene name, we use `intersectBed` to reassign gene names to all remaining exons. The `awk` command defines the 100 base pause region window downstream of all transcription start sites based on the gene strand. Lastly, we subtract the exons from the full gene coordinates to produce the intron annotations.

```
# merge exon intervals that overlap each other
mergeBed -s -c 6 -o distinct -i Homo_sapiens.GRCh38.${release}.all.exons.bed | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$2,$4}' |
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.all.exons.merged.bed

# remove all first exons
# (so pause region is excluded from exon / intron density ratio)
subtractBed -s -a Homo_sapiens.GRCh38.${release}.all.exons.merged.bed \
    -b Homo_sapiens.GRCh38.${release}.tss.bed | \
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.no.first.exons.bed

# extract gene names of exons
intersectBed -s -wb -a Homo_sapiens.GRCh38.${release}.no.first.exons.bed \
    -b Homo_sapiens.GRCh38.${release}.bed | \
    awk '{OFS="\t";} {print $1,$2,$3,$11,$4,$4}' | \
    sort -k1,1 -k2,2n >  Homo_sapiens.GRCh38.${release}.no.first.exons.named.bed

# extract the pause region from the first exons,
# position 20 - 120 downstream of the TSS
awk  '{OFS="\t";} $6 == "+" {print $1,$2+20,$2 + 120,$4,$5,$6} \
    $6 == "-" {print $1,$3 - 120,$3 - 20,$4,$5,$6}' \
    Homo_sapiens.GRCh38.${release}.tss.bed  | \
    sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.pause.bed

# define and name all introns
subtractBed -s -a Homo_sapiens.GRCh38.${release}.bed \
```

```
-b Homo_sapiens.GRCh38.${release}.all.exons.merged.bed | \
sort -k1,1 -k2,2n > Homo_sapiens.GRCh38.${release}.introns.bed
```

2.7 Processing PRO-seq Data

2.7.1 Initialize Variables

In order to automate the processing and naming of output files, we conform to a strict naming convention for the FASTQ files: cellType_conditions_replicate_pairedend.fastq.gz. For example, a gzipped paired end 1 (PE1) file from the first replicate of treating T47D cells with DMSO would be: `T47D_DMSO_rep1_PE1.fastq.gz`.

We first initialize six variables:

- `$directory`: location of the sequencing files
- `$filename`: name of the gzipped paired end 1 FASTQ file.
- `$annotation_prefix`: Ensembl gene annotation GTF prefix; this is the user-defined prefix from above.
- `$UMI_length`: length of the UMI on the 5´ end of the paired end 1 read.
- `$read_size`: read length minus UMI length.
- `$cores`: number of cores for parallel processing.
- `$genome`: absolute or relative path to the genome FASTA file.
- `$genome_index`: the basename (including the path) of the genome index files from `bowtie2-build`.
- `$prealign_rdna_index`: the basename (including the path) of the prealign rDNA index files from `bowtie2-build`.
- `$tallymer` and `$table`: outputs of the `seqOutBias` command above for a given genome and read length.

```
directory=/Users/genomicslab/sequencing_run1
filename=T47D_DMSO_rep1_PE1.fastq
annotation_prefix=Homo_sapiens.GRCh38.104
UMI_length=8
read_size=62
cores=6
genome=hg38.fa
genome_index=hg38
prealign_rdna_index=human_rDNA
tallymer=hg38.tal_${read_size}.gtTxt.gz
table=hg38_${read_size}.4.2.2.tbl
```

2.7.2 Preprocessing

Make a working directory, download the files from GEO accession GSE184378, and save the basename as a variable.

```
mkdir -p $directory
cd $directory
fasterq-dump SRR15944159
mv SRR15944159_1.fastq T47D_DMSO_rep1_PE1.fastq
mv SRR15944159_2.fastq T47D_DMSO_rep1_PE2.fastq
filename=T47D_DMSO_rep1_PE1.fastq
name=$(echo $filename | awk -F"_PE1.fastq" '{print $1}')
echo $name
```

2.7.3 Processing Reads

Here we describe processing and analysis of paired end PRO-seq libraries with unique molecular identifiers ligated to the 3′ end of the nascent RNA. The user may need to modify or omit specific steps in order to accommodate their library preparation protocol.

The first processing step is to remove adapter sequences and simultaneously discard reads that have insert sizes of one base. If the adapters ligate directly to one another, the UMI is treated as an insert to `cutadapt` and the effective insert length is equal to the UMI length. The option `-m $((UMI_length+2))` provides a one base buffer and discards reads with a length of less than the UMI length + 2.

The fraction of reads that represent adapter/adapter ligation products is a useful metric to help determine the raw read depth needed to achieve a specified aligned read depth. FASTQ files contain four lines per sequence entry, so we calculate the raw sequencing depth by first using `wc -l` to count the number of lines in the original FASTQ file and using `awk 'print $1/4'` to divide by 4. We perform the same operation on the file containing reads with 0 or 1 base insertions. Finally, we use `$(echo "scale=2 ; $PE1_w_Adapter / $PE1_total" | bc)` to divide the adapter/adapter ligation product value by the total and round to the nearest hundredth.

The proportion of reads that are adapter/adapter ligation products varies widely depending upon whether a size selection was performed in the library preparation. We recently excluded the size selection step from the PRO-seq protocol to reduce bias against small RNA inserts [152, 187]. If no size selection is performed, the adapter/adapter ligation fraction can be quite high resulting in low effective sequencing depth. In a later section we provide a formula for determining the required raw sequencing depth to result in a desired number of concordant aligned reads. We typically recommend further sequencing if all other QC metrics indicate that the data is high quality. However, if more than 80% of the reads are adapter/adapter ligation products, the user should balance the cost of performing another experiment with sequencing uninformative adapter sequences.

The fraction of adapter/adapter reads and this 0.80 threshold is printed to `$name_QC_metrics.txt`. We continue to append all metrics and thresholds to this file and plot the data at the end of the workflow.

```
cutadapt --cores=$cores -m $((UMI_length+2)) -O 1 -a TGGAATTCTCGGGTGCCAAGG \
```

```
    ${name}_PE1.fastq -o ${name}_PE1_noadap.fastq \
    --too-short-output ${name}_PE1_short.fastq > ${name}_PE1_cutadapt.txt
cutadapt --cores=$cores -m $((UMI_length+10)) -O 1 -a GATCGTCGGACTGTAGAACTCTGAAC \
    ${name}_PE2.fastq -o ${name}_PE2_noadap.fastq \
    --too-short-output ${name}_PE2_short.fastq > ${name}_PE2_cutadapt.txt

PE1_total=$(wc -l ${name}_PE1.fastq | awk '{print $1/4}')
PE1_w_Adapter=$(wc -l ${name}_PE1_short.fastq | awk '{print $1/4}')
AAligation=$(echo "scale=2 ; $PE1_w_Adapter / $PE1_total" | bc)

echo -e  "value\texperiment\tthreshold\tmetric" > ${name}_QC_metrics.txt
echo -e "$AAligation\t$name\t0.80\tAdapter/Adapter" >> ${name}_QC_metrics.txt
```

The next step removes reads that are shorter than 10 bases.

```
seqtk seq -L $((UMI_length+10)) ${name}_PE1_noadap.fastq \
    > ${name}_PE1_noadap_trimmed.fastq
```

A proportion of short nascent RNAs from different cells are identical because their 5´ end corresponds to a transcription start site, and their 3´ end is located within a focused promoter-proximal pause region [176]. Therefore, we cannot filter potential PCR duplicates based on whether two independent pairs of reads have identical paired end read alignments. We rely on the presence of the UMI to remove PCR duplicates from the PE1 FASTQ file. We use `fastq_pair` to deduplicate the PE2 read by pairing with the deduplicated PE1 file. In theory, counting the number of reads and providing this number as a table size for `fastq_pair` should make it run more quickly, but we have found, at least recently, that it takes much longer. If this is the case, remove the `-t` option.

```
# remove PCR duplicates
fqdedup -i ${name}_PE1_noadap_trimmed.fastq -o ${name}_PE1_dedup.fastq

# this variable is a near-optimal table size value for fastq_pair:
PE1_noAdapter=$(wc -l ${name}_PE1_dedup.fastq | awk '{print $1/4}')
```

```
# pair FASTQ files
fastq_pair -t $PE1_noAdapter ${name}_PE1_dedup.fastq ${name}_PE2_noadap.fastq
```

2.7.4 RNA Degradation Ratio Score

An abundance of short inserts within a library indicates that RNA degradation occurred. We measure RNA degradation by searching for overlap between paired-end reads with `flash` and plotting the resultant histogram output with `insert_size.R` (Figure 1). RNA starts to protrude from the RNA Polymerase II exit channel at approximately 20 bases in length, so 20 bases of the nascent RNA are protected from degradation during the run-on. Libraries with a substantial amount of degradation after the run-on step are enriched for species in the range 10 - 20. We empirically found that there are fewer reads within the range 10 - 20 than within the range 30 - 40 for high-quality libraries [187]. A degradation ratio of less than 1 indicates a high-quality library. This metric becomes unreliable if the protocol includes size selection to remove adapter/adapter ligation products. Size selection inevitably removes some small RNAs and inflates this ratio.

```
flash -q --compress-prog=gzip --suffix=gz ${name}_PE1_dedup.fastq.paired.fq \
    ${name}_PE2_noadap.fastq.paired.fq -o ${name}
insert_size.R ${name}.hist ${UMI_length}
```

2.7.5 Processing for Alignment

The final processing step reverse complements and removes the UMI from both paired-end reads.

```
seqtk trimfq -b ${UMI_length} ${name}_PE1_dedup.fastq | seqtk seq -r - \
    > ${name}_PE1_processed.fastq
seqtk trimfq -e ${UMI_length} ${name}_PE2_noadap.fastq | seqtk seq -r - \
    > ${name}_PE2_processed.fastq
```

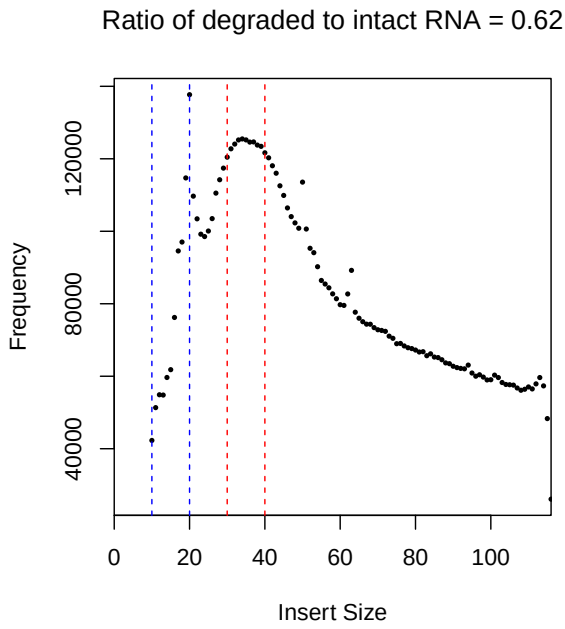

Figure 2.1: Library insert size is a measure of RNA degradation. The plot illustrates the frequency (y-axis) of insert size lengths (x-axis) for the PRO-seq library. The ratio of read counts in the 10 - 20 base range (blue region) to read counts in the 30 - 40 range (red region) is the degradation ratio. High-quality PRO-seq libraries have degradation ratios less than 1.

2.7.6 Remove Reads Aligning to rDNA

While between 70 - 80% of stable RNA is rRNA, generally less than 20% of the nascent RNA arises from rRNA. By first aligning to the rDNA, we can later estimate nascent RNA purity. Any reads that map non-uniquely to both rDNA and non-rDNA regions in the genome result in artifactual spikes at regions in the genome that share homology with the rDNA locus. Before aligning to the genome, we first align reads to rDNA and use `samtools fastq` and the `-f 0x4` flag to specify that only unmapped reads are included in the FASTQ output. We recommend the following site to help understand the meaning of samtools flags: `https://broadinstitute.github.io/picard/explain-flags.html`.

```
bowtie2 -p $((cores-2)) -x $prealign_rdna_index -U ${name}_PE1_processed.fastq \
    2>${name}_bowtie2_rDNA.log | samtools sort -n - | samtools fastq -f 0x4 - \
    > ${name}_PE1.rDNA.fastq
```

```
# This removes PE2-aligned reads with an rDNA-aligned mate
reads=$(wc -l ${name}_PE1.rDNA.fastq | awk '{print $1/4}')
fastq_pair -t $reads ${name}_PE1.rDNA.fastq ${name}_PE2_processed.fastq
```

2.7.7 Genome Alignment

The last processing step for individual libraries is to align to the genome. We invoke the `--rf` flag to account for the fact that we reverse complemented both reads. The `samtools` commands convert the file to a compressed binary BAM format and sort the reads.

```
bowtie2 -p $((cores-2)) --maxins 1000 -x $genome_index --rf \
    -1 ${name}_PE1.rDNA.fastq.paired.fq \
    -2 ${name}_PE2_processed.fastq.paired.fq 2>${name}_bowtie2.log \
    | samtools view -b - | samtools sort - -o ${name}.bam
```

2.7.8 rDNA Alignment Rate

In order to calculate the rDNA alignment rate, we first count the total number of rDNA-aligned reads. Next, we use `samtools view -c -f 0x42` to count the PE1 reads that concordantly align to hg38 and not to rDNA. Lastly, we calculate the fraction of aligned reads that map to the rDNA locus and print it to the QC metrics file.

```
#calculate the total number of rDNA-aligned reads
PE1_prior_rDNA=$(wc -l ${name}_PE1_processed.fastq | awk '{print $1/4}')
PE1_post_rDNA=$(wc -l ${name}_PE1.rDNA.fastq | awk '{print $1/4}')
total_rDNA=$(echo "$(($PE1_prior_rDNA-$PE1_post_rDNA))")

#calculate the total that concordantly align to hg38 and/or rDNA
concordant_pe1=$(samtools view -c -f 0x42 ${name}.bam)
total=$(echo "$(($concordant_pe1+$total_rDNA))")

#rDNA alignment rate
```

```
rDNA_alignment=$(echo "scale=2 ; $total_rDNA / $total" | bc)

echo -e "$rDNA_alignment\t$name\t0.20\trDNA Alignment Rate" \
    >> ${name}_QC_metrics.txt
```

2.7.9 Mappability rate

The majority of reads should map concordantly to the genome. We expect an alignment rate above 80% for high quality libraries. As described above, we use `samtools` and `wc -l` to count concordantly aligned reads in the BAM alignment file and the pre-alignment FASTQ files and then divide these values to calculate the alignment rate. We found that low alignment rates typically arise from either poor quality sequencing or microorganism contamination of reagents/buffers. We recommend using FastQC to determine if the poor alignment is due to a problem with the FASTQ base quality scores [204]. If the user suspects that the poor alignment is due to xenogeneic DNA contamination, we recommend using `BLAST` to query unaligned sequences to genome databases [205]. The user can leverage the sequences to design PCR primers for the contaminating species and test reagents to identify the source.

```
map_pe1=$(samtools view -c -f 0x42 ${name}.bam)
pre_alignment=$(wc -l ${name}_PE1.rDNA.fastq.paired.fq | awk '{print $1/4}')
alignment_rate=$(echo "scale=2 ; $map_pe1 / $pre_alignment" | bc)

echo -e "$alignment_rate\t$name\t0.80\tAlignment Rate" >> ${name}_QC_metrics.txt
```

2.7.10 Complexity and Theoretical Read Depth

The proportion of PCR duplicates in a library affects how many additional raw sequencing reads are required to achieve a target number of concordantly aligned reads. We developed `fqComplexity` to serve two purposes:

1) Calculate the number of reads that are non-PCR duplicates as a metric for complexity. 2) Provide a formula and constants to calculate the theoretical read depth that will result in a

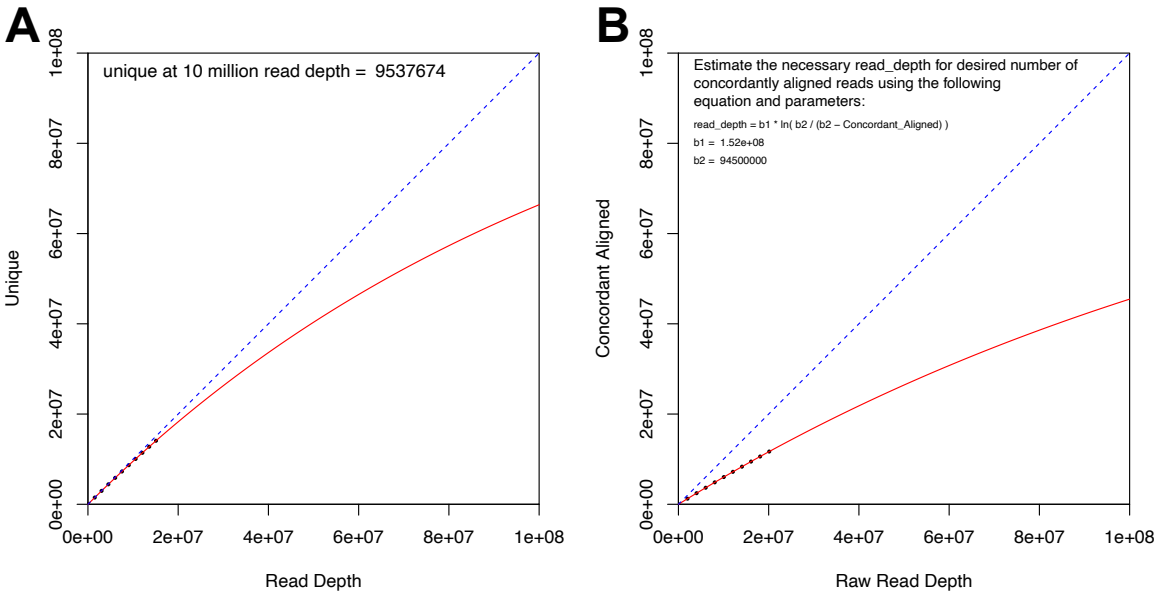

Figure 2.2: Library complexity captures information about PCR over-amplification and read depth requirements for a sample. A) We subsample the pre-processed FASTQ file to the indicated read depths (x-axis) and plot this value against the number of unique subsampled reads (y-axis). The plot includes an asymptotic regression model curve and prints the estimated number of unique reads at a read depth of 10 million. B) We use the fraction of raw PE1 reads that do not contain adapter ligation products or small inserts, the fraction of deduplicated reads that align concordantly to the non-rDNA genome, and the data from panel A to derive the theoretical read depth equation and parameters.

user-defined number of concordant aligned reads.

The proportion of reads that are PCR duplicates is related to read depth. At very low read depth, nearly all reads are unique; at very high read depth, the observed fraction of duplicates approaches the true PCR duplicate rate of the library. We calculate the PCR duplicate rate using the processed FASTQ file without adapter/adapter ligation products or small inserts. The FASTQ file is randomly subsampled to read depths of 10%, 20%, 30%, ..., 100%, and the intermediate files are deduplicated. We print the total and deduplicated counts for each subsample to the `$name_complexity.log` file. The R script fits an asymptotic regression model and plots the model and data (Figure 2.2A). We recommend that at least 75% of reads are unique at a read depth of 10 million.

```
fqComplexity -i ${name}_PE1_noadap_trimmed.fastq
```

Sequencing depth requirements vary depending upon downstream applications and the size/gene density of the genome. We recommend three replicates and over 10 million concordantly aligned reads per replicate for differential expression analysis with human cells. Data-driven approaches to define gene annotations or identify regulatory elements require higher sequencing depth. We need two factors to calculate the raw read depth necessary to achieve a specified target concordantly aligned depth. The first value is the fraction of raw PE1 reads that do not contain adapter/adapter ligation products or small inserts: `$factorX`. The second value is the fraction of deduplicated reads that align concordantly to the non-rDNA genome: `$factorY`. Finally, we run `fqComplexity` and specify the `-x` and `-y` options to fit an asymptotic regression model to the factor-scaled log file. `fqComplexity` searches for and reuses the previous log file to avoid unnecessarily repeating subsampling and deduplication. We use the equation and constants printed in the PDF output (Figure 2.2B) to determine the practicality of increasing depth using the same libraries.

```
PE1_total=$(wc -l ${name}_PE1.fastq | awk '{print $1/4}')
PE1_noadap_trimmed=$(wc -l ${name}_PE1_noadap_trimmed.fastq | awk '{print $1/4}')

factorX=$(echo "scale=2 ; $PE1_noadap_trimmed / $PE1_total" | bc)

echo fraction of reads that are not adapter/adapter ligation products \
    or below 10 base inserts
echo $factorX

# calculate PE1 deduplicated reads
PE1_dedup=$(wc -l ${name}_PE1_dedup.fastq | awk '{print $1/4}')

# divide
factorY=$(echo "scale=2 ; $concordant_pe1 / $PE1_dedup" | bc)

# re-run with factors
```

```
fqComplexity -i ${name}_PE1_noadap_trimmed.fastq -x $factorX -y $factorY
```

2.7.11 Run-on Efficiency

RNA polymerases that are associated with gene bodies efficiently incorporate nucleotides during the run-on reaction under most conditions, but promoter-proximal paused RNA polymerase requires high salt or detergent to run-on efficiently [182, 206]. Therefore, the pause index, or the density of signal in the promoter-proximal pause region divided by density in the gene body, is an indirect measure of run-on efficiency. Since pause windows are user-defined and variable, pause indices can differ substantially based on how they are calculated.

To determine the coverage of PRO-seq signal in genomic intervals, it is convenient to convert the genomic signal to a BED6 file format. Although we are not correcting enzymatic sequence bias in this workflow, we use `seqOutBias` with the `--no-scale` option to convert the BAM file. We include the `--tail-edge` option to realign the end of the read so that the exact position of RNA Polymerase is specified in the BED6 output file. The `--out-split-pairends` option separates all the paired-end reads, and `--stranded` prints strand information in column 6.

```
#convert to bigWig and BED6
seqOutBias scale $table ${name}.bam --no-scale --stranded \
    --bed-stranded-positive --bw=$name.bigWig --bed=$name.bed \
    --out-split-pairends --only-paired --tail-edge \
    --read-size=$read_size --tallymer=$tallymer
#Remove chromosomes not in the gene annotation file and sort for use in mapBed
grep -v "random" ${name}_not_scaled_PE1.bed | grep -v "chrUn" | \
    grep -v "chrEBV" | sort -k1,1 -k2,2n > ${name}_tmp.txt
mv ${name}_tmp.txt ${name}_not_scaled_PE1.bed

#count reads in pause region
mapBed -null "0" -s -a $annotation_prefix.pause.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | sort -k5,5 -k7,7nr | \
```

```
    sort -k5,5 -u > ${name}_pause.bed

#discard anything with chr and strand inconsistencies
join -1 5 -2 5 ${name}_pause.bed $annotation_prefix.bed | \
    awk '{OFS="\t";} $2==$8 && $6==$12 \
    {print $2, $3, $4, $1, $6, $7, $9, $10}' | \
    awk '{OFS="\t";} $5 == "+" {print $1,$2+480,$8,$4,$6,$5} \
    $5 == "-" {print $1,$7,$2 - 380,$4,$6,$5}' | \
    awk  '{OFS="\t";} $3>$2 {print $1,$2,$3,$4,$5,$6}' | \
    sort -k1,1 -k2,2n > ${name}_pause_counts_body_coordinates.bed

#column ten is Pause index
mapBed -null "0" -s -a ${name}_pause_counts_body_coordinates.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$5,$6,$7,$5/100,$7/($3 - $2)}' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$5,$6,$7,$8,$9,$8/$9}' \
    > ${name}_pause_body.bed
```

We use an R script to calculate pause indices and plot the distribution of $\log_{10}$ pause index values as a PDF (Figure 2.3).

```
pause_index.R ${name}_pause_body.bed
```

2.7.12 Estimate Nascent RNA Purity with Exon/Intron Density Ratio

Exon and intron densities within each gene are comparable in nascent RNA-seq data. In contrast, RNA-seq primarily measures mature transcripts, and exon density far exceeds intron density. We can infer mature RNA contamination in PRO-seq libraries if we detect a high exon density to intron density ratio. We exclude contributions from the first exon because pausing occurs in this region and artificially inflates the exon density. The distribution of $\log_{10}$ exon density to intron density ratios is output as a PDF (Figure 2.4). This metric complements rDNA alignment rate to determine nascent RNA purity.

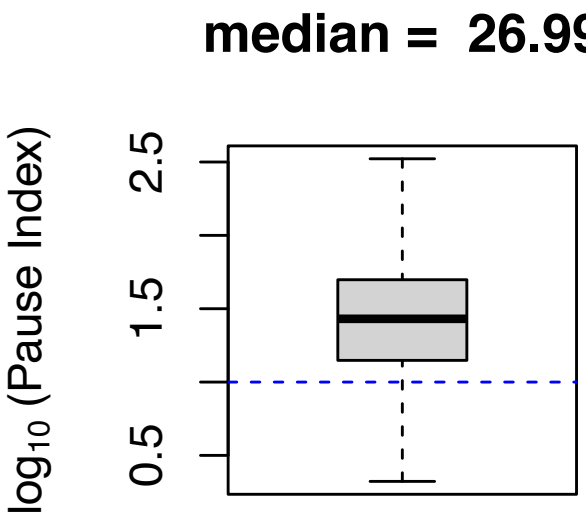

Figure 2.3: Pause index is a measure of nuclear run-on efficiency. The plot illustrates the distribution of $\log_{10}$ pause indices and includes a threshold line at a raw pause index of 10. A median pause index below 10 indicates that the library may be of poor quality.

```
mapBed -null "0" -s -a $annotation_prefix.introns.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$5,$5,$6,$7,($3 - $2)}' \
    > ${name}_intron_counts.bed

mapBed -null "0" -s -a $annotation_prefix.no.first.exons.named.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$4,$6,$7,($3 - $2)}' \
    > ${name}_exon_counts.bed

exon_intron_ratio.R ${name}_exon_counts.bed ${name}_intron_counts.bed
```

2.7.13 Remove intermediate files and zip raw sequencing files

Calculating these quality control metrics necessitates many intermediate files. Many files are unused output from various processing steps or only used briefly. FASTQ files are large and rarely used in downstream analyses, so the following code chunk removes intermediate FASTQ files and compresses the original files.

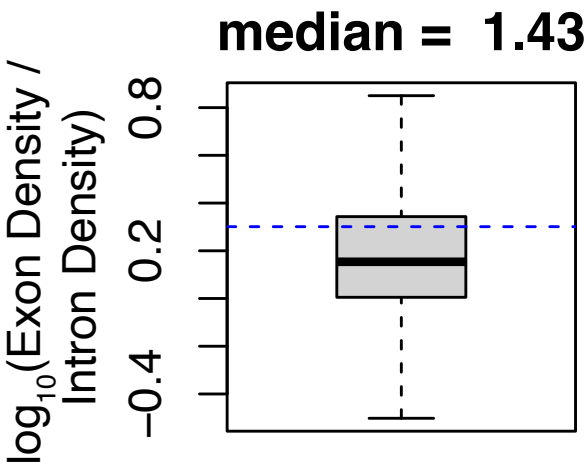

Figure 2.4: Exon density to intron density ratio is a measure of nascent RNA purity. The plot illustrates the distribution of $\log_{10}$ exon density to intron density ratios and includes a threshold line at a raw ratio of 2. A median ratio below 2 indicates mature RNA contamination is low.

```
rm ${name}_PE1_short.fastq
rm ${name}_PE2_short.fastq
rm ${name}_PE1_noadap.fastq
rm ${name}_PE2_noadap.fastq
rm ${name}_PE1_noadap_trimmed.fastq
rm ${name}_PE1_dedup.fastq
rm ${name}_PE1_processed.fastq
rm ${name}_PE2_processed.fastq
rm ${name}_PE1_dedup.fastq.paired.fq
rm ${name}_PE2_noadap.fastq.paired.fq
rm ${name}_PE1_dedup.fastq.single.fq
rm ${name}_PE2_noadap.fastq.single.fq
rm ${name}_PE1.rDNA.fastq.paired.fq
rm ${name}_PE1.rDNA.fastq.single.fq
rm ${name}_PE2_processed.fastq.paired.fq
rm ${name}_PE2_processed.fastq.single.fq
rm ${name}.extendedFrags.fastq.gz
rm ${name}.notCombined_1.fastq.gz
```

```
rm ${name}.notCombined_2.fastq.gz
gzip ${name}_PE1.fastq
gzip ${name}_PE2.fastq
```

2.7.14 Pipeline Automation

We present the deconstructed workflow above because it is helpful to run through the code chunks individually to gain further understanding of each step. A more complete understanding of the processes allows the user to modify steps based on PRO-seq protocol variations. However, automation of routine processing and analysis is more practical once a workflow is established. Below, we provide a shell script loop that will process each set of paired end files in series. This loop can be adapted to perform all processing in parallel using a job scheduler and submission of a batch script for each set of paired end input files.

```
#initialize variables
directory=/Users/genomicslab/sequencing_run1_series
annotation_prefix=Homo_sapiens.GRCh38.104
UMI_length=8
read_size=62
cores=10
genome=hg38.fa
genome_index=hg38
prealign_rdna_index=human_rDNA
tallymer=hg38.tal_${read_size}.gtTxt.gz
table=hg38_${read_size}.4.2.2.tbl

mkdir -p $directory
cd $directory

for i in {59..62}
do
    fasterq-dump SRR159441${i}
```

```
done

mv SRR15944159_1.fastq T47D_Starved_DMSO_rep1_PE1.fastq
mv SRR15944159_2.fastq T47D_Starved_DMSO_rep1_PE2.fastq
mv SRR15944160_1.fastq T47D_Starved_DMSO_rep2_PE1.fastq
mv SRR15944160_2.fastq T47D_Starved_DMSO_rep2_PE2.fastq
mv SRR15944161_1.fastq T47D_Starved_Estrogen_rep1_PE1.fastq
mv SRR15944161_2.fastq T47D_Starved_Estrogen_rep1_PE2.fastq
mv SRR15944162_1.fastq T47D_Starved_Estrogen_rep2_PE1.fastq
mv SRR15944162_2.fastq T47D_Starved_Estrogen_rep2_PE2.fastq

for i in {65..72}
do
    fasterq-dump SRR273143${i}
done

mv SRR27314365_1.fastq T47D_Complete_Tamoxifen_rep2_PE1.fastq
mv SRR27314365_2.fastq T47D_Complete_Tamoxifen_rep2_PE2.fastq
mv SRR27314366_1.fastq T47D_Complete_Tamoxifen_rep1_PE1.fastq
mv SRR27314366_2.fastq T47D_Complete_Tamoxifen_rep1_PE2.fastq
mv SRR27314367_1.fastq T47D_Complete_Raloxifene_rep2_PE1.fastq
mv SRR27314367_2.fastq T47D_Complete_Raloxifene_rep2_PE2.fastq
mv SRR27314368_1.fastq T47D_Complete_Raloxifene_rep1_PE1.fastq
mv SRR27314368_2.fastq T47D_Complete_Raloxifene_rep1_PE2.fastq
mv SRR27314369_1.fastq T47D_Complete_Fulvestrant_rep2_PE1.fastq
mv SRR27314369_2.fastq T47D_Complete_Fulvestrant_rep2_PE2.fastq
mv SRR27314370_1.fastq T47D_Complete_Fulvestrant_rep1_PE1.fastq
mv SRR27314370_2.fastq T47D_Complete_Fulvestrant_rep1_PE2.fastq
mv SRR27314371_1.fastq T47D_Complete_DMSO_rep2_PE1.fastq
mv SRR27314371_2.fastq T47D_Complete_DMSO_rep2_PE2.fastq
mv SRR27314372_1.fastq T47D_Complete_DMSO_rep1_PE1.fastq
```

```
mv SRR27314372_2.fastq T47D_Complete_DMSO_rep1_PE2.fastq


for filename in *PE1.fastq
do
    name=$(echo $filename | awk -F"_PE1.fastq" '{print $1}')
    echo $name
    echo 'removing dual adapter ligations and calculating'
    echo 'the fraction of adapter/adapters in' $name
    cutadapt --cores=$cores -m $((UMI_length+2)) -O 1 -a TGGAATTCTCGGGTGCCAAGG \
        ${name}_PE1.fastq -o ${name}_PE1_noadap.fastq --too-short-output \
        ${name}_PE1_short.fastq > ${name}_PE1_cutadapt.txt
    cutadapt --cores=$cores -m $((UMI_length+10)) -O 1 \
        -a GATCGTCGGACTGTAGAACTCTGAAC ${name}_PE2.fastq \
        -o ${name}_PE2_noadap.fastq --too-short-output \
        ${name}_PE2_short.fastq > ${name}_PE2_cutadapt.txt
    PE1_total=$(wc -l ${name}_PE1.fastq | awk '{print $1/4}')
    PE1_w_Adapter=$(wc -l ${name}_PE1_short.fastq | awk '{print $1/4}')
    AAligation=$(echo "scale=2 ; $PE1_w_Adapter / $PE1_total" | bc)
    echo -e  "value\texperiment\tthreshold\tmetric" > ${name}_QC_metrics.txt
    echo -e "$AAligation\t$name\t0.80\tAdapter/Adapter" >> ${name}_QC_metrics.txt
    echo 'removing short RNA insertions in' $name
    seqtk seq -L $((UMI_length+10)) ${name}_PE1_noadap.fastq \
        > ${name}_PE1_noadap_trimmed.fastq
    echo 'removing PCR duplicates from' $name
    fqdedup -i ${name}_PE1_noadap_trimmed.fastq -o ${name}_PE1_dedup.fastq
    PE1_noAdapter=$(wc -l ${name}_PE1_dedup.fastq | awk '{print $1/4}')
    fastq_pair -t $PE1_noAdapter ${name}_PE1_dedup.fastq ${name}_PE2_noadap.fastq
    echo 'calculating and plotting RNA insert sizes from' $name
    flash -q --compress-prog=gzip --suffix=gz ${name}_PE1_dedup.fastq.paired.fq \
        ${name}_PE2_noadap.fastq.paired.fq -o ${name}
    insert_size.R ${name}.hist ${UMI_length}
```

```
echo 'trimming off the UMI from' $name
seqtk trimfq -b ${UMI_length} ${name}_PE1_dedup.fastq | \
    seqtk seq -r - > ${name}_PE1_processed.fastq
seqtk trimfq -e ${UMI_length} ${name}_PE2_noadap.fastq | \
    seqtk seq -r - > ${name}_PE2_processed.fastq
echo 'aligning' $name 'to rDNA and removing aligned reads'
bowtie2 -p $((cores-2)) -x $prealign_rdna_index \
    -U ${name}_PE1_processed.fastq 2>${name}_bowtie2_rDNA.log | \
    samtools sort -n - | samtools fastq -f 0x4 - \
    > ${name}_PE1.rDNA.fastq
reads=$(wc -l ${name}_PE1.rDNA.fastq | awk '{print $1/4}')
fastq_pair -t $reads ${name}_PE1.rDNA.fastq ${name}_PE2_processed.fastq
echo 'aligning' $name 'to the genome'
bowtie2 -p $((cores-2)) --maxins 1000 -x $genome_index --rf \
    -1 ${name}_PE1.rDNA.fastq.paired.fq \
    -2 ${name}_PE2_processed.fastq.paired.fq 2>${name}_bowtie2.log | \
    samtools view -b - | samtools sort - -o ${name}.bam
PE1_prior_rDNA=$(wc -l ${name}_PE1_processed.fastq | awk '{print $1/4}')
PE1_post_rDNA=$(wc -l ${name}_PE1.rDNA.fastq | awk '{print $1/4}')
total_rDNA=$(echo "$(($PE1_prior_rDNA-$PE1_post_rDNA))")
concordant_pe1=$(samtools view -c -f 0x42 ${name}.bam)
total=$(echo "$(($concordant_pe1+$total_rDNA))")
rDNA_alignment=$(echo "scale=2 ; $total_rDNA / $total" | bc)
echo -e "$rDNA_alignment\t$name\t0.10\trDNA Alignment Rate" \
    >> ${name}_QC_metrics.txt
map_pe1=$(samtools view -c -f 0x42 ${name}.bam)
pre_alignment=$(wc -l ${name}_PE1.rDNA.fastq.paired.fq | \
    awk '{print $1/4}')
alignment_rate=$(echo "scale=2 ; $map_pe1 / $pre_alignment" | bc)
echo -e "$alignment_rate\t$name\t0.80\tAlignment Rate" \
    >> ${name}_QC_metrics.txt
```

```
echo 'plotting and calculating complexity for' $name
fqComplexity -i ${name}_PE1_noadap_trimmed.fastq
echo 'calculating and plotting theoretical sequencing depth'
echo 'to achieve a defined number of concordantly aligned reads for' $name
PE1_total=$(wc -l ${name}_PE1.fastq | awk '{print $1/4}')
PE1_noadap_trimmed=$(wc -l ${name}_PE1_noadap_trimmed.fastq | \
    awk '{print $1/4}')
factorX=$(echo "scale=2 ; $PE1_noadap_trimmed / $PE1_total" | bc)
echo 'fraction of reads that are not adapter/adapter ligation products'
echo 'or below 10 base inserts'
echo $factorX
PE1_dedup=$(wc -l ${name}_PE1_dedup.fastq | awk '{print $1/4}')
factorY=$(echo "scale=2 ; $concordant_pe1 / $PE1_dedup" | bc)
fqComplexity -i ${name}_PE1_noadap_trimmed.fastq -x $factorX -y $factorY
echo 'Separating paired end reads and creating genomic BED and bigWig'
echo 'intensity files for' $name
seqOutBias scale $table ${name}.bam --no-scale --stranded \
    --bed-stranded-positive --bw=$name.bigWig --bed=$name.bed \
    --out-split-pairends --only-paired \
    --tail-edge --read-size=$read_size --tallymer=$tallymer
grep -v "random" ${name}_not_scaled_PE1.bed | grep -v "chrUn" | \
    grep -v "chrEBV" | sort -k1,1 -k2,2n > ${name}_tmp.txt
mv ${name}_tmp.txt ${name}_not_scaled_PE1.bed
mapBed -null "0" -s -a $annotation_prefix.pause.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    sort -k5,5 -k7,7nr | sort -k5,5 -u > ${name}_pause.bed
join -1 5 -2 5 ${name}_pause.bed $annotation_prefix.bed | \
    awk '{OFS="\t";} $2==$8 && $6==$12 \
    {print $2, $3, $4, $1, $6, $7, $9, $10}' | \
    awk '{OFS="\t";} $5 == "+" {print $1,$2+480,$8,$4,$6,$5} \
    $5 == "-" {print $1,$7,$2 - 380,$4,$6,$5}' | \
```

```
    awk  '{OFS="\t";} $3>$2 {print $1,$2,$3,$4,$5,$6}' \
    | sort -k1,1 -k2,2n > ${name}_pause_counts_body_coordinates.bed
mapBed -null "0" -s -a ${name}_pause_counts_body_coordinates.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$5,$6,$7,$5/100,$7/($3 - $2)}' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$5,$6,$7,$8,$9,$8/$9}' \
    > ${name}_pause_body.bed
pause_index.R ${name}_pause_body.bed
echo 'Calculating exon density / intron density'
echo 'as a metric for nascent RNA purity for' $name
mapBed -null "0" -s -a $annotation_prefix.introns.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$5,$5,$6,$7,($3 - $2)}' \
    > ${name}_intron_counts.bed
mapBed -null "0" -s -a $annotation_prefix.no.first.exons.named.bed \
    -b ${name}_not_scaled_PE1.bed | awk '$7>0' | \
    awk '{OFS="\t";} {print $1,$2,$3,$4,$4,$6,$7,($3 - $2)}' \
    > ${name}_exon_counts.bed
exon_intron_ratio.R ${name}_exon_counts.bed ${name}_intron_counts.bed
rm ${name}_PE1_short.fastq
rm ${name}_PE2_short.fastq
rm ${name}_PE1_noadap.fastq
rm ${name}_PE2_noadap.fastq
rm ${name}_PE1_noadap_trimmed.fastq
rm ${name}_PE1_dedup.fastq
rm ${name}_PE1_processed.fastq
rm ${name}_PE2_processed.fastq
rm ${name}_PE1_dedup.fastq.paired.fq
rm ${name}_PE2_noadap.fastq.paired.fq
rm ${name}_PE1_dedup.fastq.single.fq
rm ${name}_PE2_noadap.fastq.single.fq
```

```
    rm ${name}_PE1.rDNA.fastq.paired.fq
    rm ${name}_PE1.rDNA.fastq.single.fq
    rm ${name}_PE2_processed.fastq.paired.fq
    rm ${name}_PE2_processed.fastq.single.fq
    rm ${name}.extendedFrags.fastq.gz
    rm ${name}.notCombined_1.fastq.gz
    rm ${name}.notCombined_2.fastq.gz
done
```

2.7.15 Plot all QC metrics

Individual plots for each quality control metric provide valuable information about the data, but each plot can be summarized as a single informative value. We empirically determined thresholds for each value that constitute acceptable libraries. These thresholds are not absolute and should only be used as guidelines. Below, we concatenate all the summarized metrics for the experiments and plot the results (Figure 2.5) and thresholds. The user can quickly glance at the plot to determine whether the quality control values fall within the acceptable range, which is shaded light green. If values are within the dark pink region, then we recommend looking back at the more detailed quality control plots to diagnose possible issues with the libraries. The user can change the term "Estrogen_treatment_PRO" to a description of their own experiment to name the output file.

```
cat *_QC_metrics.txt | awk '!x[$0]++' > project_QC_metrics.txt


plot_all_metrics.R project_QC_metrics.txt Estrogen_treatment_PRO
```

2.8 Differential Expression with DESeq2

Differential expression analysis is a common first step after routine RNA-seq and PRO-seq data processing. Below we present the `bedtools` command to count reads within gene annotations, and we provide an `R` script for differential expression analysis with `DESeq2`. The script also plots

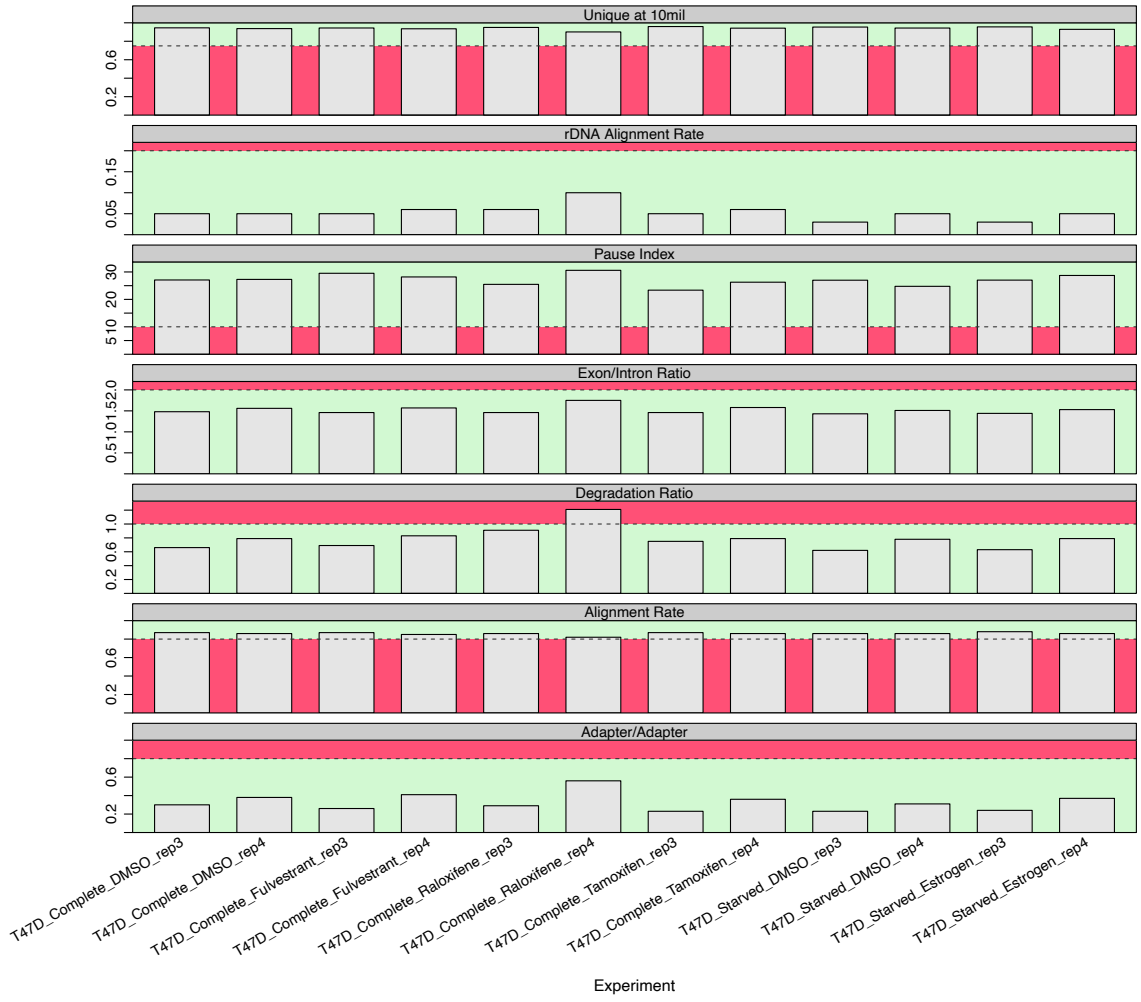

Figure 2.5: A summary plot illustrates all quality control metrics and their respective recommended thresholds. If all quality control values fall within the shaded light green range, then the libraries are likely of high quality. If values are within the dark pink region, then we recommend looking back at the more detailed quality control plots in Figures 2.1-2.4 to diagnose possible issues with the libraries.

the fold change between conditions and mean expression level for each gene. For simplicity, we use the most upstream transcription start site and most downstream transcription termination site for annotations, but there are more accurate methods to define primary transcripts [180, 181]. The `R` script requires three ordered arguments:

1) A file with the signal counts for each gene in every even row. 2) The prefix for the baseline experimental condition for which to compare (often termed "untreated"). 3) Prefix name for the output PDF plot (Figure 2.6).

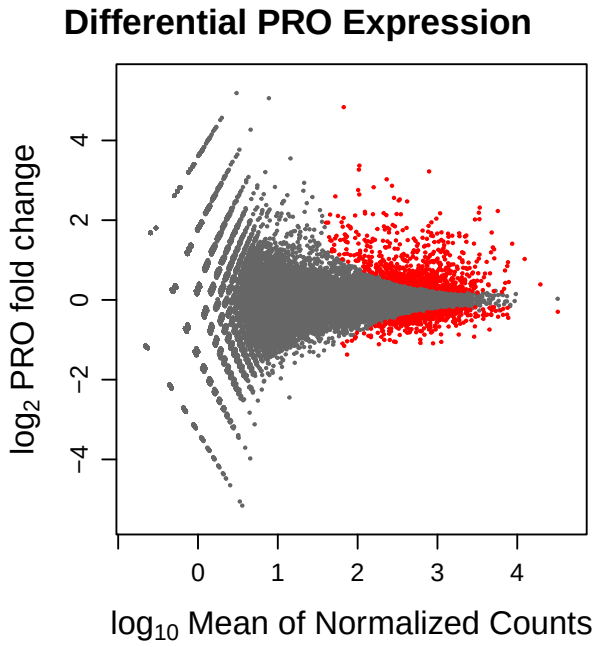

Figure 2.6: Differential expression analysis quantifies transcriptomic changes upon treating T47D cells with estrogen for an hour. Genes in red are classified as activated and repressed based on a false discovery rate of 0.05.

```
for filename in *_not_scaled_PE1.bed
do
    name=$(echo $filename | awk -F"_not_scaled_PE1.bed" '{print $1}')
    echo -e  "\t${name}" > ${name}_gene_counts.txt
     mapBed -null "0" -s -a ${annotation_prefix}_sorted.bed -b $filename | \
        awk '{OFS="\t";} {print $4,$7}' >> ${name}_gene_counts.txt
```

```
done
paste -d'\t' *_gene_counts.txt > Estrogen_treatment_PRO_gene_counts.txt


differential_expression.R \
    Estrogen_treatment_PRO_gene_counts.txt T47D_Starved_DMSO Estrogen_treatment
```

2.9 ER antagonists affect the same genes as ER agonists

The data used in the above analysis were generated from the luminal breast cancer cell line T47D. Cells were either grown in hormone-starved medium and acutely treated with the estrogen receptor (ER) agonist 17-beta-estradiol (estrogen) or grown in complete medium and acutely treated with the ER antagonists fulvestrant, raloxifene, or tamoxifen. The purpose of the experiment was to differentiate acute ER agonism from acute ER antagonism.

As ER is a defining feature of ER-positive breast cancer cells and a necessary TF for these cells' growth and proliferation, there have been multiple studies to identify the downstream effectors of ER signaling, or ER target genes [207]. One feature of ER that makes this endeavor easier is the ability to rapidly induce ER from an inactive state. After hormone starving cells by growing them in charcoal-stripped media for several days, ER activity can be rapidly induced via the addition of estrogen to the media. After this step, changes in transcription were originally measured using tiled microarrays or RNA sequencing [207]. These measurements of RNA levels required time points of many hours to days after the initial perturbation, in order for activated transcripts to accumulate over baseline and repressed transcripts to be degraded.

However, changes in transcription at these time points reflect not only the primary effects of the perturbation but also secondary effects and beyond. For example, if ER activates transcription of another TF, then, once its gene product is transcribed and translated, this TF will regulate additional genes, some of which are ER-independent. A study using cycloheximide to inhibit translation and mitigate these secondary effects estimated that less than 30% of estrogen-responsive genes are direct ER targets [208]. Additionally, various studies have identified

widely divergent sets of ER target genes, with estimates of the number of estrogen-responsive genes ranging from 100-1500 [209, 210]. To avoid these complications, a key step forward was the use of global run-on and sequencing (GRO-seq) to identify ER target genes [37]. GRO-seq directly measures the location of transcriptionally engaged RNA polymerases to determine the effects of a perturbation on nascent transcription, thus allowing for measurements at time points on the order of 30 minutes [211].

Still, all of these studies have measured ER activity after rapid induction from an inactive state. This hormone-starved context is different from the physiological context in which ER normally functions in patient tumors, with fluctuating but ever present levels of estrogen and other hormones. We hypothesized that ER regulates distinct sets of genes initially versus at steady state. This is the case for another rapidly inducible TF, HSF1, the binding sites for which differ between oncogenically transformed cells with constitutive HSF1 activation and the non-transformed progenitor cells after HSF1 activation with heat shock [212]. In the case of ER, differences could be caused by differential expression of chromatin associated proteins in hormone starved versus complete media.

Thus we performed the above experiment to compare the estrogen-activated genes to the antagonist-repressed genes. When comparing the antagonists, we found that the effect sizes were largest for fulvestrant and smallest for tamoxifen. Furthermore, the directions were in the opposite direction and of smaller magnitude than the effects of estrogen. These high level patterns can be observed in the principle component analysis plot (Figure 2.7).

To our surprise, we found no genes that were significantly repressed by the ER antagonists that were not also significantly activated by estrogen. We did, however, find that not all estrogen-activated genes were significantly antagonist-repressed. We wanted to determine whether this was a qualitative difference, in which ER antagonists only repress a subset of estrogen-activated genes, or a quantitative difference, in which ER antagonists repress all estrogen-activated genes with a smaller effect size than the estrogen effect. To do so, we plotted the distribution of effect sizes in the fulvestrant treatment for all estrogen-activated genes (Figure 2.8). We found

that almost all estrogen-activated genes demonstrated a decrease in nascent transcription upon fulvestrant treatment, even though not all these genes were significantly fulvestrant-repressed. These data do not support our initial hypothesis that acute ER antagonism from a high activity state would differ from acute ER agonism from a low activity state.

2.10 Conclusions

We provide standardized metrics and detailed plots that indicate whether libraries are of sufficiently high quality to warrant downstream analysis. The presented analyses provide information about RNA degradation, nascent RNA purity, alignment rate, library complexity, and nuclear run-on efficiency. We deconstruct each analysis and explain the biological rationale of each metric. All code and scripts are presented so that researchers can use this framework to develop their own workflows and pipelines, or as Captain Barbossa succinctly stated: "The code is more of what you'd call *guidelines* than actual rules."

2.11 Methods

2.11.1 Cell culture

T47D cells (RRID:CVCL_0553) (ATCC) were cultured in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (Gemini).

2.11.2 Cell treatments

4 replicates were performed from cells treated and collected at different times in the same day. For the hormone starved samples, $2*10^6$ cells per sample were plated in 10cm dishes 4 days before harvest. The following day, the medium was replaced with phenol red free RPMI 1640 medium (Gibco) supplemented with 10% charcoal-stripped/dextran-treated fetal bovine serum (Cytiva). For the hormone replete samples, $3*10^6$ cells per sample were plated in 10cm dishes 1 day before harvest. On the day of the harvest, the hormone starved cells were treated with 0.1%

DMSO or 1nM estrogen (Sigma) in DMSO for 1 hour. The hormone replete cells were treated with 0.1% DMSO or 100nM fulvestrant (Sigma), raloxifene hydrochloride (Sigma), or tamoxifen (Sigma) in DMSO for 1 hour.

2.11.3 Cell permeabilization for PRO-seq

Cell permeabilization was performed as previously described [213], with modifications. At the time of harvest, cells were scraped in 10mL ice cold PBS and washed in 5mL buffer W (10mM Tris-HCl pH 7.5, 10mM KCl, 150mM sucrose, 5mM MgCltextsubscript2, 0.5mM CaCltextsubscript2, 0.5mM DTT, 0.004U/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)). Cells were permeabilized by incubating with buffer P (10 mM Tris-HCl pH 7.5, KCl 10 mM, 250 mM sucrose , 5 mM MgCltextsubscript2, 1 mM EGTA, 0.05% Tween-20, 0.1% NP40, 0.5 mM DTT, 0.004 units/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)) for 3 minutes on ice. Cells were washed with 10 mL buffer W before being transferred into 1.5mL tubes using wide bore pipette tips. Finally, cells were resuspended in 50µL buffer F (50mM Tris-HCl pH 8, 5mM MgCltextsubscript2, 0.1mM EDTA, 50% Glycerol, 0.5 mM DTT). Cells were snap frozen in liquid nitrogen and stored at -80°C. During these steps, the first two replicates of each condition were lost due to aggressive aspiration instead of careful decanting.

2.11.4 PRO-seq library preparation

PRO-seq libraries were prepared as previously described [152], with modifications. RNA extraction after the run-on reaction was performed with 500µL Trizol LS (Thermo Fisher) followed by 130µL chloroform (Sigma). The equivalent of 1µL of 50µM for each adapter was used. A random eight base unique molecular identifier (UMI) was included at the 5′ end of the adapter ligated to the 3′ end of the nascent RNA. For the reverse transcription reaction, RP1 was used at 100µM and dNTP mix was used at 10mM each. Libraries were amplified by PCR for a total of 10 cycles in 50µL reactions with Phusion polymerase (New England Biolabs).

No PAGE purification was performed to ensure that our libraries were not biased against short nascent RNA insertions.

2.12 Data Access

Raw sequencing files and processed *bigWig* files are available from GEO accession record GSE184378.

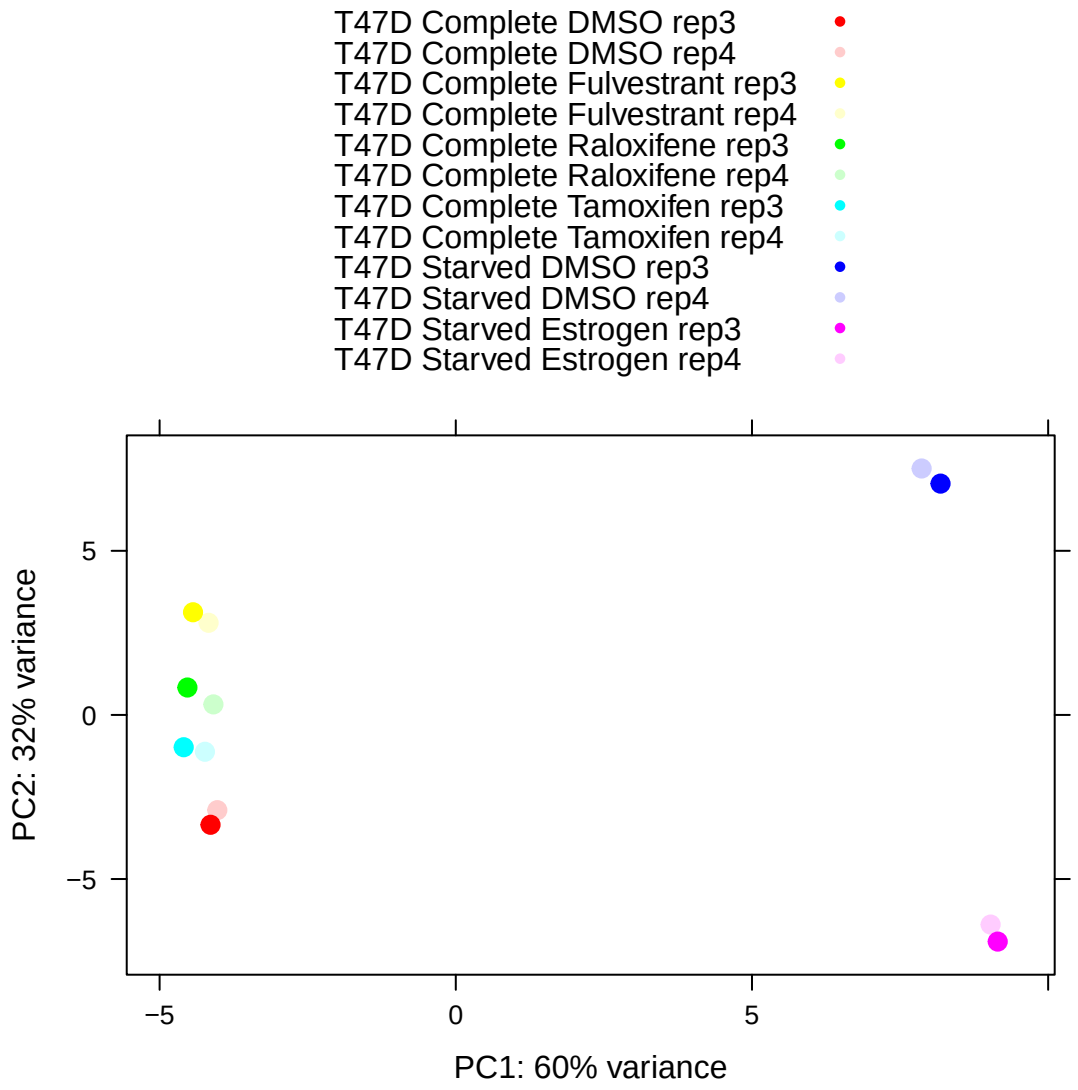

Figure 2.7: Principle component analysis of nascent transcription upon acute ER agonism or antagonism.

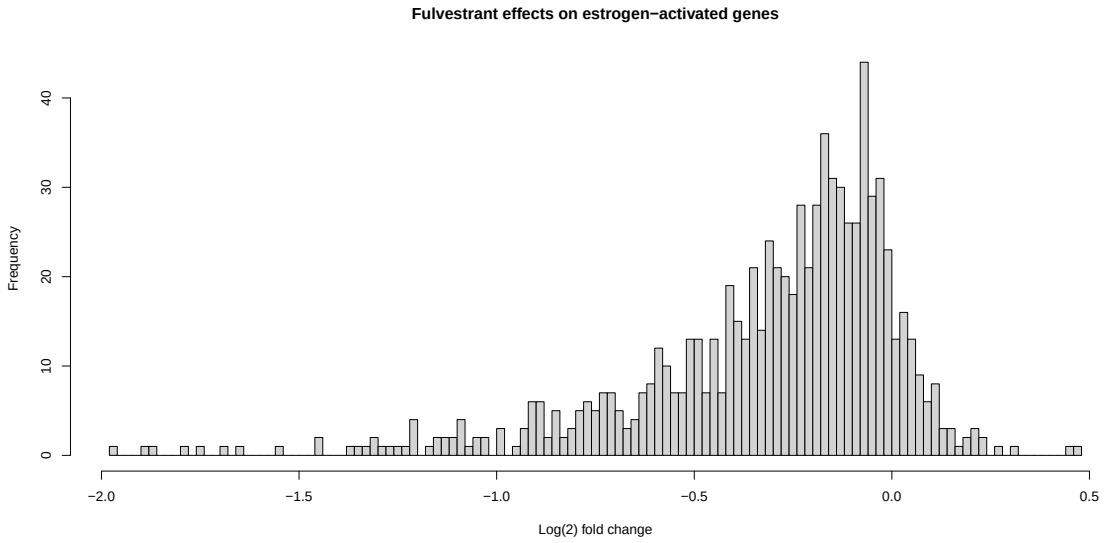

Figure 2.8: ER antagonists affect the same genes as ER agonists. Histogram of fold changes upon one hour of fulvestrant treatment in T47D cells grown in complete media. Genes used in this plot are significantly activated genes upon one hour of estrogen treatment in T47D cells grown in hormone-starved media.

Chapter 3

TRPS1 modulates chromatin accessibility to regulate estrogen receptor alpha (ER) binding and ER target gene expression in luminal breast cancer cells

3.1 Preface

This chapter is adapted from a manuscript under review at *PLOS Genetics* after revision.

Thomas G. Scott, Kizhakke Mattada Sathyan, Daniel Gioeli, Michael J. Guertin

3.2 Author contributions

All authors contributed to the conceptualization of the project. SKM contributed to the methodology. TGS performed the experiments, analyzed the data, and wrote the original draft of the manuscript. All authors reviewed and edited the manuscript.

TRPS1 modulates chromatin accessibility to regulate estrogen receptor alpha (ER) binding and ER target gene expression in luminal breast cancer cells

Thomas G. Scott[a], Kizhakke Mattada Sathyan[b,c], Daniel Gioeli[d,e], Michael J. Guertin[b,c]

[a]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America

[b]Center for Cell Analysis and Modeling, University of Connecticut, Farmington, Connecticut, United States of America

[c]Department of Genetics and Genome Sciences, University of Connecticut, Farmington, Connecticut, United States of America

[d]Department of Microbiology, Immunology, and Cancer, University of Virginia, Charlottesville, Virginia, United States of America

[e]Cancer Center Member, University of Virginia, Charlottesville, Virginia, United States of America

3.3 Abstract

Common genetic variants in the repressive GATA-family transcription factor (TF) *TRPS1* locus are associated with breast cancer risk, and luminal breast cancer cell lines are particularly sensitive to *TRPS1* knockout. We introduced an inducible degron tag into the native *TRPS1* locus within a luminal breast cancer cell line to identify the direct targets of TRPS1 and determine how TRPS1 mechanistically regulates gene expression. We acutely deplete over 80 percent of TRPS1 from chromatin within 30 minutes of inducing degradation. We find that TRPS1 regulates transcription of hundreds of genes, including those related to estrogen signaling. TRPS1 directly regulates chromatin structure, which causes estrogen receptor alpha (ER) to redistribute in the genome. ER redistribution leads to both repression and activation of dozens of ER target genes. Downstream from these primary effects, TRPS1 depletion represses cell cycle-related gene sets and reduces cell doubling rate. Finally, we show that high TRPS1 activity, calculated using a gene expression signature defined by primary TRPS1-regulated genes, is associated with worse

breast cancer patient prognosis. Taken together, these data suggest a model in which TRPS1 modulates the genomic distribution of ER, both activating and repressing transcription of genes related to cancer cell fitness.

3.4 Author Summary

Breast cancer is the most common cancer among women. The majority of cases are luminal, which tend to be estrogen receptor alpha (ER)-positive. ER is well-studied among transcription factors (TFs) because it is ligand-activated. This allows for the rapid induction of ER activity and the identification of primary estrogen-responsive genes. Most TFs have not been so extensively characterized, because their activity is not so rapidly perturbable. TRPS1 is an atypical GATA family TF that is associated with corepressor complexes and transcriptional repression. Here, we use an inducible degron tag system to rapidly deplete endogenously-tagged TRPS1 in luminal breast cancer cells within 30 minutes. We find that TRPS1 directly decreases local chromatin accessibility. This decreases ER binding intensity at TRPS1-proximal ER binding sites. As an indirect effect, ER binding intensity distal to TRPS1 increases in intensity. These effects on ER binding are associated with changes in ER target gene transcription, repressing or activating these genes in concordance with the effect on ER binding intensity. Our work is consistent with a model in which TFs either exclusively activate or exclusively repress transcription of their direct target genes, with indirect primary response genes changing due to the redistribution of limiting activating TFs or coactivators.

3.5 Introduction

Breast cancer is the most frequently diagnosed cancer in women, with an estimated lifetime risk of about 1 in 8 for women in the United States [1]. Far from a monolithic disease, breast tumors can be classified into subtypes based on gene expression, histology, and immunohistochemistry [14, 15]. The most common subtype is luminal breast cancer, which is typically estrogen receptor alpha (ER)-positive [41].

High lifetime exposure to endogenous estrogen is a strong risk factor for breast cancer incidence [29]. Estrogen is a potent hormone that binds to ER, a ligand-activated transcription factor (TF), which then homodimerizes, binds to reverse palindromic pairs of AGGTCA motifs on DNA, and recruits cofactors to activate hundreds of genes that promote cell growth and proliferation [36, 37]. In additional to surgery, radiation, and traditional chemotherapy, endocrine therapies that inhibit endogenous estrogen production or binding to ER provide a significant survival benefit to luminal breast cancer patients [39–41]. However, patients with advanced disease frequently develop resistance to these therapies, though many endocrine therapy-resistant luminal tumors still remain dependent on ER activity [46]. Thus, there is a need to identify additional factors that regulate ER activity or genomic binding and contribute to breast tumor progression.

TRPS1 is a member of the GATA-family of TFs that bind to (A/T)GATA(A/G) motifs on DNA [60]. In contrast to the other six members of the GATA family that activate transcription, TRPS1 directly represses transcription of target genes via its unique IKZF1-like zinc fingers [61]. TRPS1 has been shown to interact with multiple corepressors and lysine deacetylases, including members of the NuRD and coREST complexes, to regulate transcription [53, 57, 58, 62, 63].

TRPS1 was first described as the gene mutated in cases of tricho-rhino-phalangeal syndrome, an autosomal dominant disorder characterized by developmental abnormalities of the hair, nose, and fingers [47]. *TRPS1* is crucial for the proper development of several tissues, including hair, bone, and kidney [48, 49]. As with many developmentally important genes co-opted during the process of cancer initiation and progression, *TRPS1* is commonly over-expressed in breast tumors, both relative to normal tissue and relative to other tumor types [52, 54].

The transcriptional program that TRPS1 regulates in breast cancer is not fully understood. Knockdown of *TRPS1* in various breast cancer cell lines has been shown to increase markers of epithelial to mesenchymal transition and genome instability [65–68]. Additionally, TRPS1 binding sites on chromatin overlap with those of YAP1 and ER, though this is coupled with a genome-wide activation of YAP1 target genes but repression of ER target genes [53, 57].

A key feature of these previous studies is the use of extended knockdown kinetics with traditional RNA interference methods. Days after knockdown, the resultant effects represent not only the primary TRPS1-responsive genes but also secondary and compensatory effects. As such, we do not know which genes TRPS1 directly regulates and whether these genes are important for breast cancer cell growth and proliferation.

In this study, we set out to directly assay the primary effects of TRPS1 on chromatin accessibility, ER binding, and transcription in luminal breast cancer cells. To do so, we acutely depleted TRPS1 protein levels using an inducible degron tag inserted into the endogenous *TRPS1* locus. By performing sensitive genome-wide assays minutes to hours after TRPS1 depletion, we demonstrated that TRPS1 changes chromatin structure, which allows ER to redistribute in the genome. Along with this redistribution, we propose that TRPS1 both directly represses and indirectly activates dozens of ER target genes at baseline. Furthermore, we defined a signature of primary TRPS1-regulated genes that predicts breast cancer patient prognosis.

3.6 Results

3.6.1 *TRPS1* is associated with breast cancer incidence and promotes breast cancer cell number accumulation

A recent genome-wide association study (GWAS) identified 32 novel single nucleotide polymorphisms (SNPs) associated with breast cancer susceptibility [214]. When we queried the NHGRI EBI GWAS Catalog to find published associations with genetic variants within the *TRPS1* genomic locus, we found one of the lead SNPs from this study [215]. Furthermore, in the authors' subtype-specific analysis of the results, the association with this variant was strongest for luminal breast cancers relative to other subtypes [214].

We used LocusZoom to plot the data within this locus (Figure 3.1A). A plot of these data indicates that two sets of SNPs have low linkage disequilibrium with one another, indicating that they are inherited independently and each confer risk. One set of variants is within an intronic

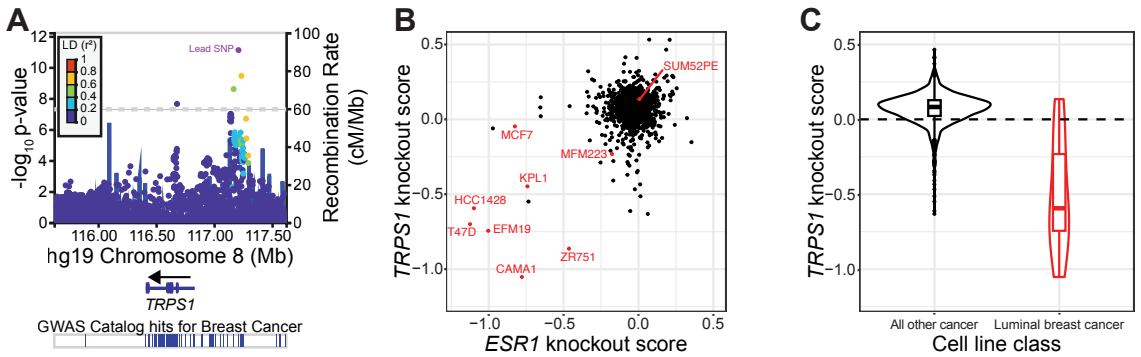

Figure 3.1: *TRPS1* is associated with breast cancer incidence and promotes breast cancer cell fitness. A) LocusZoom plot of the *TRPS1* genomic locus depicting the location and significance of SNPs associated with breast cancer susceptibility. *TRPS1* is the closest gene to two sets of genetic variants in low linkage disequilibrium with one another. Data from [214], generated with summary statistics downloaded from the NHGRI-EBI GWAS Catalog, using LocusZoom [215, 216]. B) Scatter plot of *TRPS1* and ESR1 knockout scores for each gene tested. Scores are normalized such that knockout of a gene with a score of 0 has no effect on cell number, and knockout of a gene with a score of -1 has an effect equal to that of knocking out one of a set of universally essential genes. Luminal breast cancer cell lines are colored in red. The data are from the Cancer Dependency Map project [77]. C) Violin and box and whisker plots of *TRPS1* knockout scores from (B) for luminal breast cancer cell lines versus all other cancer cell lines. Wilcoxon rank sum test p-value of $3.2*10^{-5}$.

region of the *TRPS1* gene, and one is about 400 kilobases upstream from the transcription start site (TSS) of the *TRPS1* gene. There are often many genes in close proximity to the lead SNP in a GWAS, making it difficult to predict which gene mediates the effect on the associated phenotype. However, in this case the nearest gene is almost a megabase upstream from *TRPS1*, suggesting that *TRPS1* contributes to the breast cancer susceptibility associated with one or both of these sets of genetic variants.

Based on this result, we hypothesized that perturbation of *TRPS1* in luminal breast cancer cell lines would affect cell fitness. Using data from the Cancer Dependency Map project, we found that sensitivity to *TRPS1* knockout correlated with sensitivity to knockout of *ESR1*, the gene encoding ER (Figure 3.1B) [77]. Furthermore, while *TRPS1* knockout led to an increase in cell number for most cancer cell lines, luminal breast cancer cell lines were significantly enriched for *TRPS1* dependency (Figure 3.1C).

Taken together, these data indicate that *TRPS1* influences breast cancer incidence and is required for maximal breast cancer cell fitness. Next we sought to determine how TRPS1 regulates its target genes to mediate these breast cancer patient and cellular outcomes.

3.6.2 Endogenously degron-tagged TRPS1 is rapidly degraded in T47D cells

To rapidly deplete TRPS1 and isolate primary TRPS1-regulated genes, we employed the dTAG system for targeted protein degradation [89, 103]. We inserted an inducible degron tag into the endogenous *TRPS1* locus in the luminal breast cancer cell line T47D. We generated three independent clones that express the tagged TRPS1 protein that can be degraded by the addition of the small molecules dTAG-13 and dTAG^{V}-1 at 50nM each (dTAG) (Figure 3.2A). Of note, we depleted around 50% of TRPS1 from whole cell lysates in 10 minutes of treatment with dTAG, as determined by quantitative western blot, with less than 10% detected as soon as 20 minutes and as late as 48 hours after treatment (Figure 3.2B).

To ensure this treatment depleted TRPS1 from chromatin, we performed chromatin immunoprecipitation with sequencing (ChIP-seq) using an anti-HA antibody to recognize the

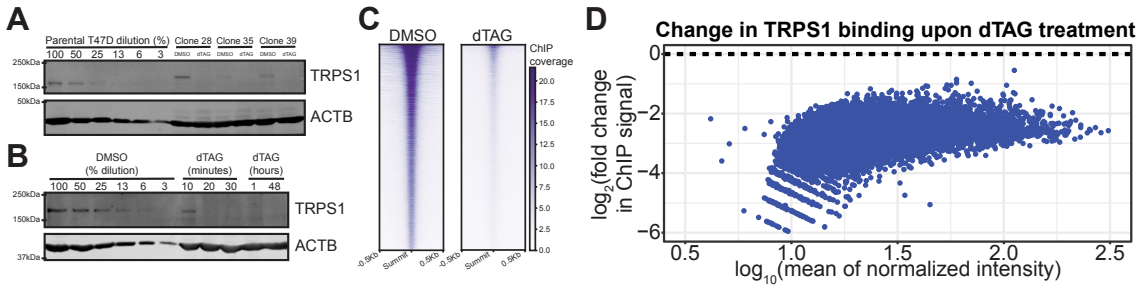

Figure 3.2: Endogenously degron-tagged TRPS1 is rapidly degraded in T47D cells. A) Quantitative Western blot with a serial dilution of the parental T47D cells followed by three independent dTAG-TRPS1 clones treated with DMSO or dTAG-13 and dTAG^V-1 at 50nM each (dTAG) for 2 hours. Membranes were probed with anti-TRPS1 or anti-ACTB antibodies. B) Quantitative Western blot with a serial dilution of dTAG-TRPS1 Clone 28 followed by a time course of treatment with dTAG. Membranes were probed as in (A). C) Heatmap of TRPS1 ChIP-seq peaks, in rows ranked by intensity, in cells treated with DMSO or dTAG for 30 minutes. D) MA plot of TRPS1 ChIP-seq peaks, with fold change values representing binding intensity in the dTAG condition relative to the DMSO condition. All points are colored blue to indicate they are significantly decreased at an FDR of 0.1.

2xHA tag within the degron tag. Our ChIP-seq libraries were of high quality, as measured using the quality control metrics of fraction of reads in peaks (Figure 3.3) and the peak calling-independent strand cross-correlation (Figure 3.4), generated using the ChIPQC R package [217]. We called peaks using MACS2 [218], using all DMSO samples together and all dTAG samples as the control. We observed a genome-wide decrease in TRPS1 binding intensity, with over 80 percent of TRPS1 depleted from chromatin after 30 minutes of dTAG treatment (Figure 3.2C,D).

3.6.3 TRPS1 directly represses regulatory element activity

With this system in hand, we set out to test the effects of TRPS1 depletion on regulatory element (RE) (i.e., enhancer or promoter) activity. To capture the dynamics of chromatin accessibility after TRPS1 depletion, we conducted a time course analysis using the Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq). The time points included dTAG treatments at 30 minutes, 1 hour, 2 hours, 4 hours, and 24 hours, while a DMSO treatment served as the vehicle control, assigned as the zero minute time point. Our ATAC-seq libraries

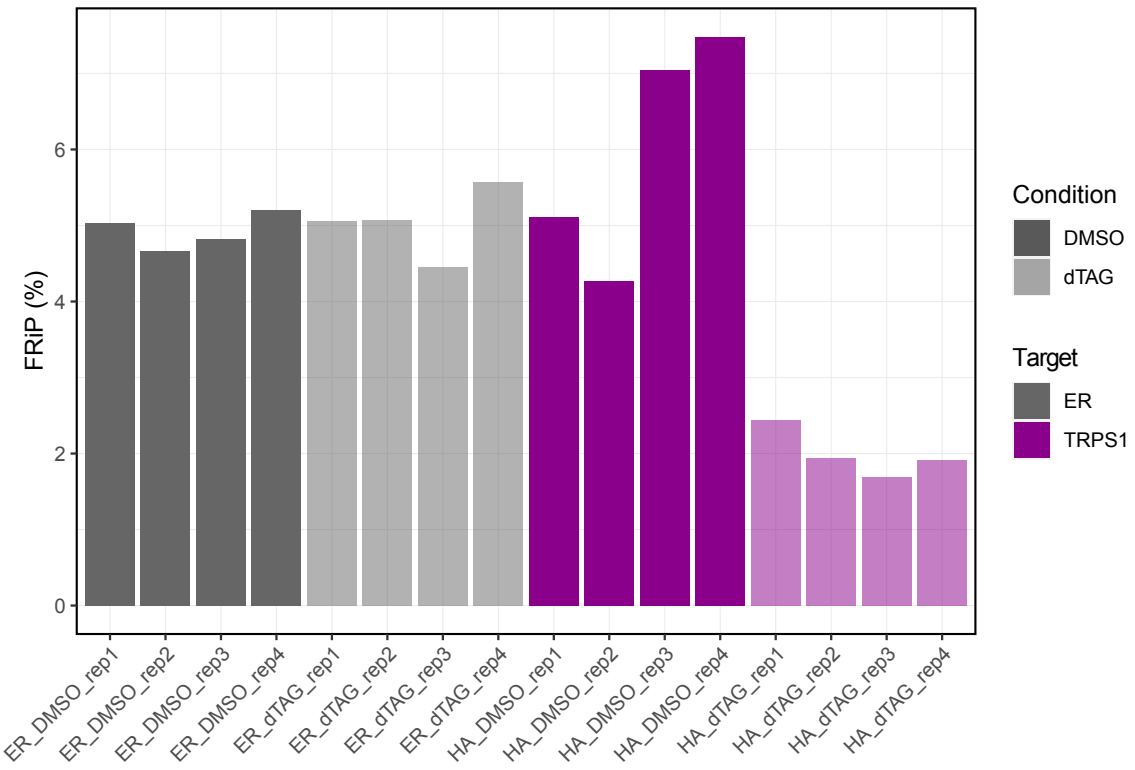

Figure 3.3: Fraction of reads in peaks (FRiP) for ChIP-seq libraries. FRiP scores for each library, calculated using the ChIPQC R package [217].

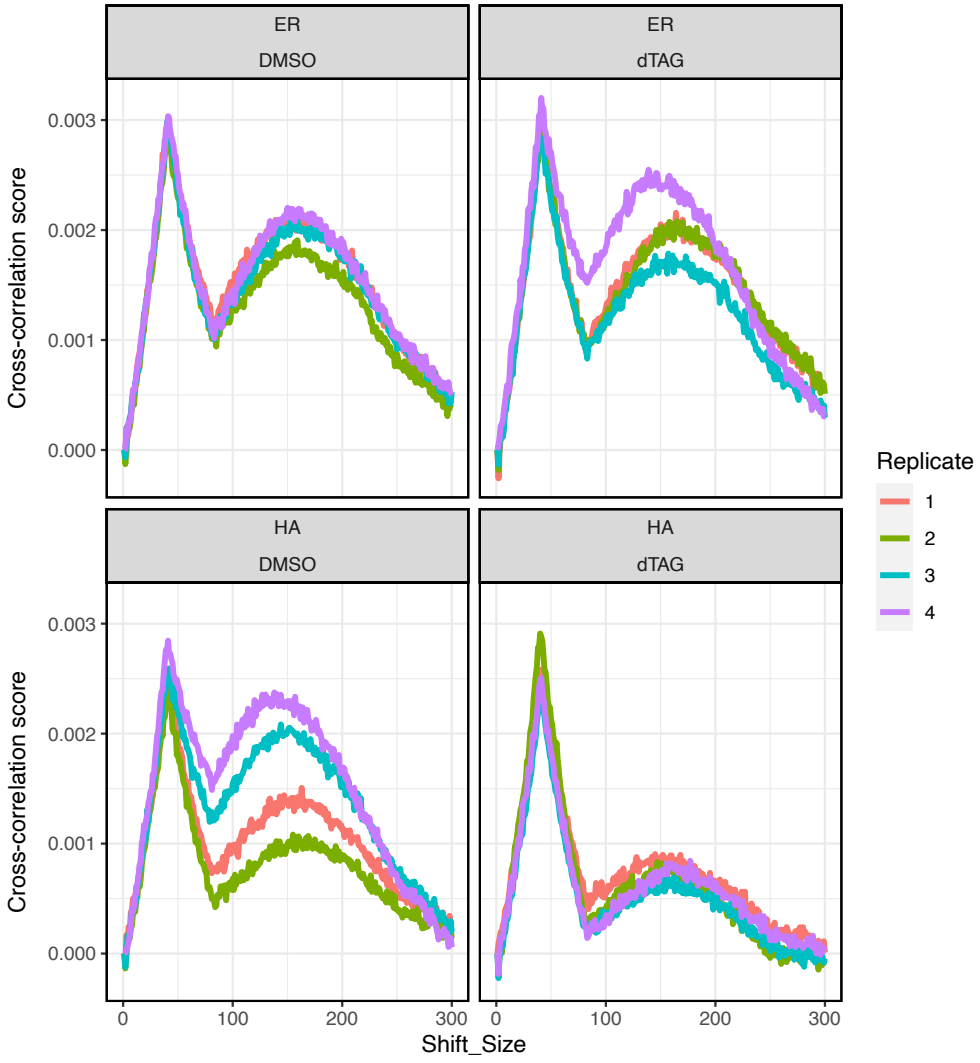

Figure 3.4: Strand cross-correlation (CC) plots for ChIP-seq libraries. CC values for each library, calculated using the ChIPQC R package [217].

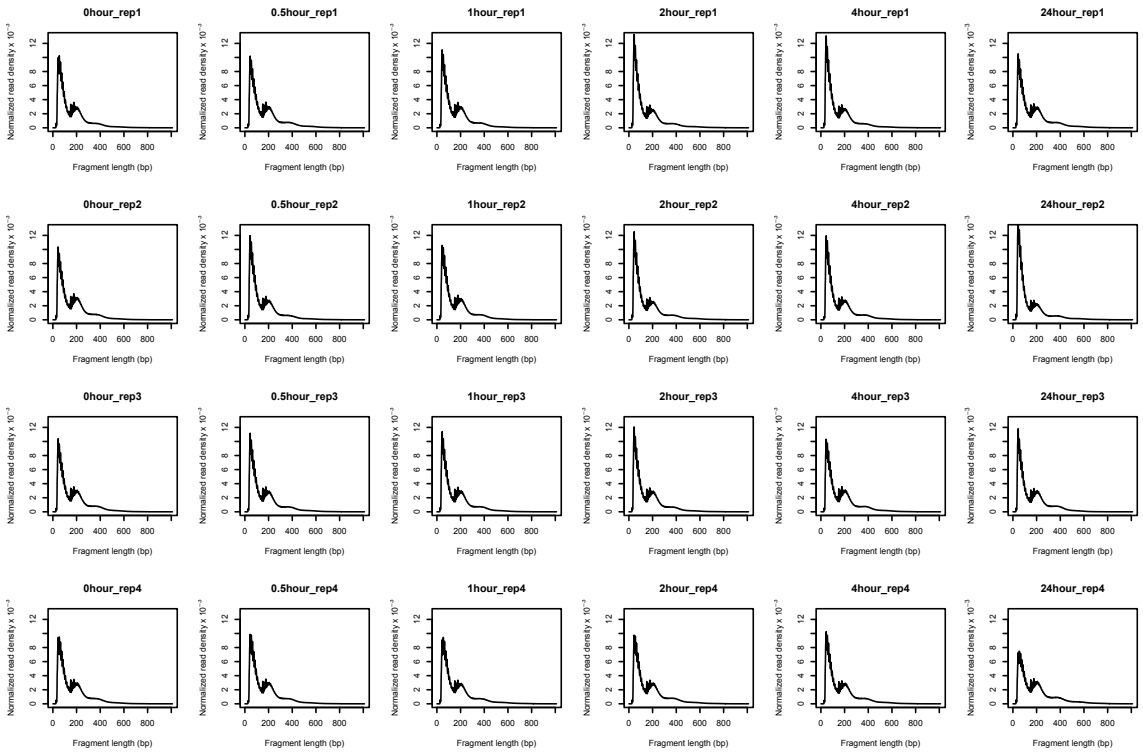

Figure 3.5: Fragment size distribution plots for ATAC-seq libraries. A plot for each library was generated using the ATACseqQC R package [219].

were of high quality, as determined by plotting the distribution of fragment sizes (Figure 3.5) and the enrichment of signal around TSS's (Figure 3.6), generated using the ATACseqQC R package [219]. We generated a consensus peak set using MACS2 [218] for all samples together. At the earliest time point after degradation, our best estimate of the primary effects of TRPS1 depletion, we identify 472 peaks that increased in intensity and 36 that decreased in intensity, at a false discovery rate (FDR) of 0.1. (Figure 3.7A). We hypothesized that the increased peaks result from loss of a direct TRPS1 reduction of chromatin accessibility, with the decreased peaks an indirect effect of the redistribution of limiting cofactors.

To test this hypothesis, we performed *de novo* motif identification in the increased and decreased peaks. We identified a GATA motif in the increased peaks but not the decreased peaks (Figure 3.7B). To explicitly calculate the motif prevalence in each class of peaks, we found individual motif occurrences genome-wide and intersected the peaks with these motif instances.

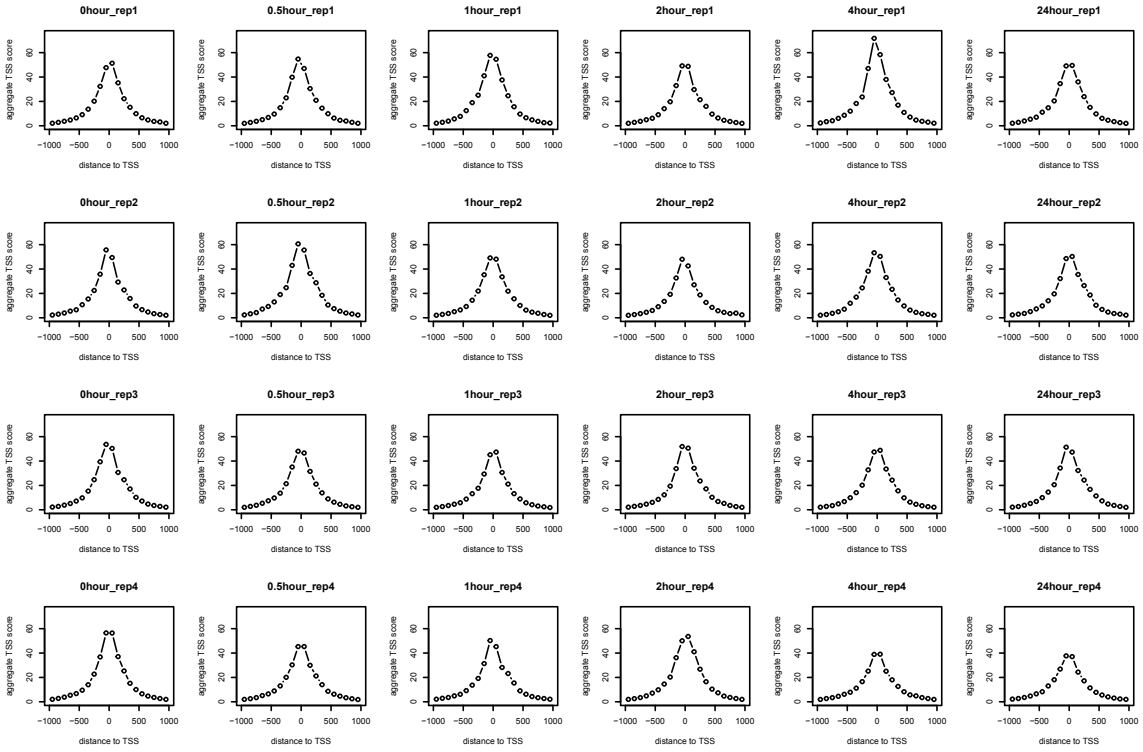

Figure 3.6: Plots of signal enrichment around TSS's for ATAC-seq libraries. A plot for each library was generated using the ATACseqQC R package [219].

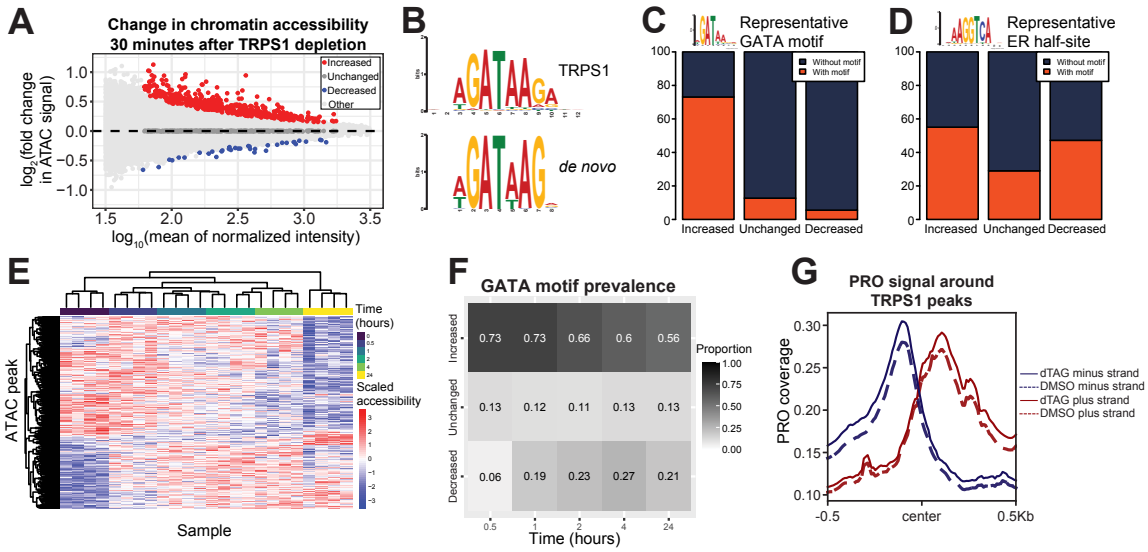

Figure 3.7: TRPS1 directly represses regulatory element activity. A) MA plot of ATAC-seq peaks, with fold change values representing accessibility in the 30 minute dTAG-13 and dTAG^{V}-1 at 50nM each (dTAG) treatment condition relative to the DMSO condition. B) *De novo* motif identified in increased peaks from (A) (below), matched to the TRPS1 motif (above). C) Bar charts of prevalence of a representative GATA motif in increased, unchanged, and decreased peaks from (A). Chi-square test p-value $< 2.2*10^{-16}$. D) Bar charts of prevalence of a representative ER half-site in increased, unchanged, and decreased peaks from (A). Chi-square test p-value $4.5*10^{-15}$. E) Heat map with hierarchical clustering of chromatin accessibility in ATAC-seq peaks that are significantly changed over the time course at an FDR of 0.1. F) Heat map of prevalence of the representative GATA motif in increased, unchanged, and decreased peaks, as in (C), for each time point relative to the DMSO condition. G) Density plot for composite PRO-seq signal across all TRPS1 ChIP-seq peaks, separated by strand and treatment condition.

We found a significant enrichment of a representative GATA motif in the increased peaks over the unchanged and decreased peaks (Figure 3.7C).

To identify additional TFs that might be contributing to changes in chromatin accessibility, we analyzed motif enrichment in both increased peaks relative to unchanged peaks as well as decreased peaks relative to unchanged peaks. To our surprise, we identified nuclear receptor motifs in both classes of peaks. To follow up this analysis, we again calculated motif prevalence in each class of peaks for a representative ER half-site. For this motif, we found a significant enrichment in both increased and decreased peaks relative to unchanged peaks (Figure 3.7D).

We predicted that the enrichment of the GATA motif within increased peaks would wane over time. To test this prediction, we turned to our time course data. We identified ATAC-seq peaks that were significantly changed over the time course at an FDR of 0.1, using a likelihood ratio test within DESeq2 [198] to identify peaks for which including a variable for the time point increased the predictive power of the model over one without this information. We performed hierarchical clustering of these peaks and found that the replicates for each time point clustered together and that the majority of dynamic peaks changed gradually over the time course, with additional clusters displaying different kinetics (Figure 3.7E). We called increased, unchanged, and decreased peaks as in Figure 3.7A for each time point relative to the control condition. We then calculated GATA motif prevalence in each set of peaks as in Figure 3.7C. As we expected, the GATA motif prevalence in the increased peaks decreased over time, but the majority of increased peaks at 24 hours still contained a GATA motif (Figure 3.7F). This is consistent with the primary effects of TRPS1 depletion driving a large proportion of the changes in chromatin accessibility even as late as 24 hours after TRPS1 depletion.

As an orthogonal readout of RE activity, we measured bidirectional transcription around TRPS1 binding sites using precision run-on with sequencing (PRO-seq). Our libraries were of high quality using several quality control metrics (Figure 3.8) [220]. Using a window centered on each summit of TRPS1 ChIP-seq intensity, we observed an increase in bidirectional transcription 30 minutes after TRPS1 depletion (Figure 3.7G).

Since composite profiles have limitations, we chose to look more closely at each TRPS1 peak and determine if increasing bidirectional transcription is a reproducible trend. We counted PRO reads in the window around the TRPS1 peak summits shown in Figure 3.7G and used DESeq2 to identify differentially transcribed regions. We normalized our counts based on the size factors generated from the DESeq2 analysis of the reads in genes from Figure 3.11. As read counts in small putative regulatory elements tend to be fewer than in large genes, it is not surprising that only 9 of these regions have individually statistically significantly higher bidirectional transcription and none with significantly lower transcription, at a false discovery rate of 0.1 (Figure 3.9). However, we did find a TRPS1 cistrome-wide increase in PRO signal, with over 62% of these regions increasing in bidirectional transcription upon TRPS1 depletion. We performed ANOVA on a linear model predicting the logarithm of the normalized PRO reads for each region based on the DMSO or 30 minute dTAG treatment condition across the four replicates, and the p-value for the F-test was $< 2.2*10^{-16}$.

The biological interpretation is consistent with our other results that indicate a role of TRPS1 in chromatin compaction. We would not necessarily expect an increase in bidirectional transcription at each TRPS1 peak upon TRPS1 degradation, but the increase in chromatin accessibility could facilitate downstream factors that promote transcription initiation. As chromatin accessibility and bidirectional transcription can affect one another, we next isolated ATAC-seq peaks without bidirectional transcription, as identified using dREG [178]. As we saw in Figure 3.7A, chromatin accessibility predominantly increased upon TRPS1 depletion (Figure 3.10). Along with our accessibility data and motif analysis, these data indicate that TRPS1 directly represses RE activity primarily via its effect on chromatin accessibility.

3.6.4 TRPS1 directly represses transcription of target genes

Downstream from the changes in RE activity, we measured changes in nascent transcription within genes with PRO-seq over the same time course of TRPS1 depletion as in the ATAC-seq experiment. As with our ATAC-seq time course analysis, we identified genes that were significantly

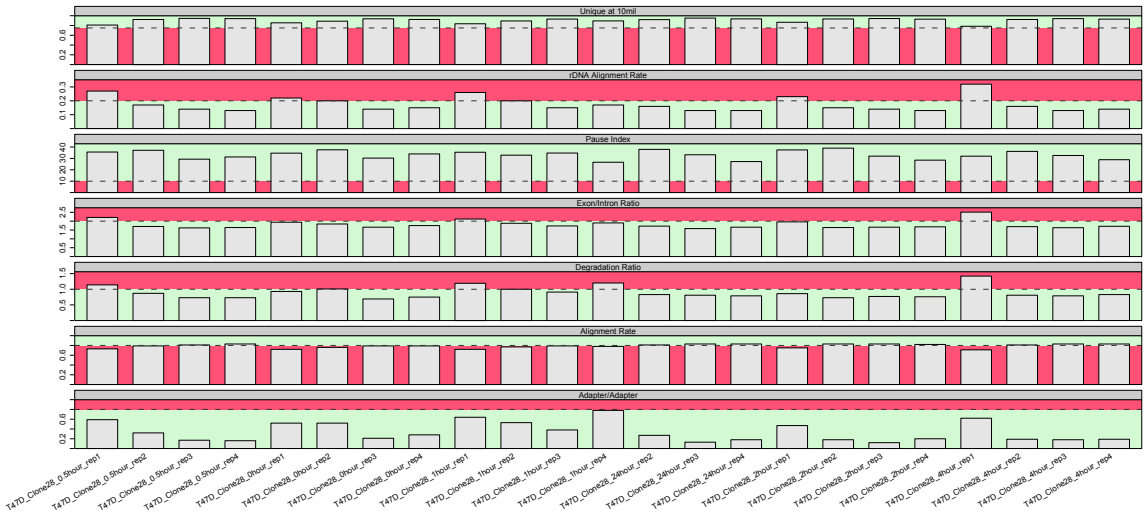

Figure 3.8: Quality control metrics for PRO-seq libraries. Quality control metrics are defined as in [220]. Each metric is a row, and each sample is a column. The green region for each metric is the goal for a high quality library.

changed over the time course at an FDR of 0.1, using a likelihood ratio test within DESeq2 [198] to identify genes for which including a variable for the time point increased the predictive power of the model over one without this information. We identified 1,425 dynamic genes over the time course and performed hierarchical clustering to classify the genes based on their expression kinetics (Figure 3.11A,B). Over-representation analysis (ORA) of the activated genes identified several enriched Hallmark gene sets [221], most prominently cholesterol homeostasis genes (Figure 3.11C). ORA on the repressed genes revealed that the two estrogen response gene sets were the most significantly enriched of the Hallmark gene sets.

We hypothesized that TRPS1 regulates these gene sets by distinct mechanisms. Specifically, we predicted that TRPS1 directly represses the dTAG-activated genes by repressing the activity of REs proximal to these genes. To test this prediction, we measured the distance from the TSS of each gene to the nearest TRPS1 ChIP-seq peak overlapping an increased ATAC-seq peak. By constructing a cumulative distribution function, we found that the activated genes are significantly closer to these activated REs (Figure 3.11E).

In contrast to the genes activated by TRPS1 depletion, we predicted that the effects

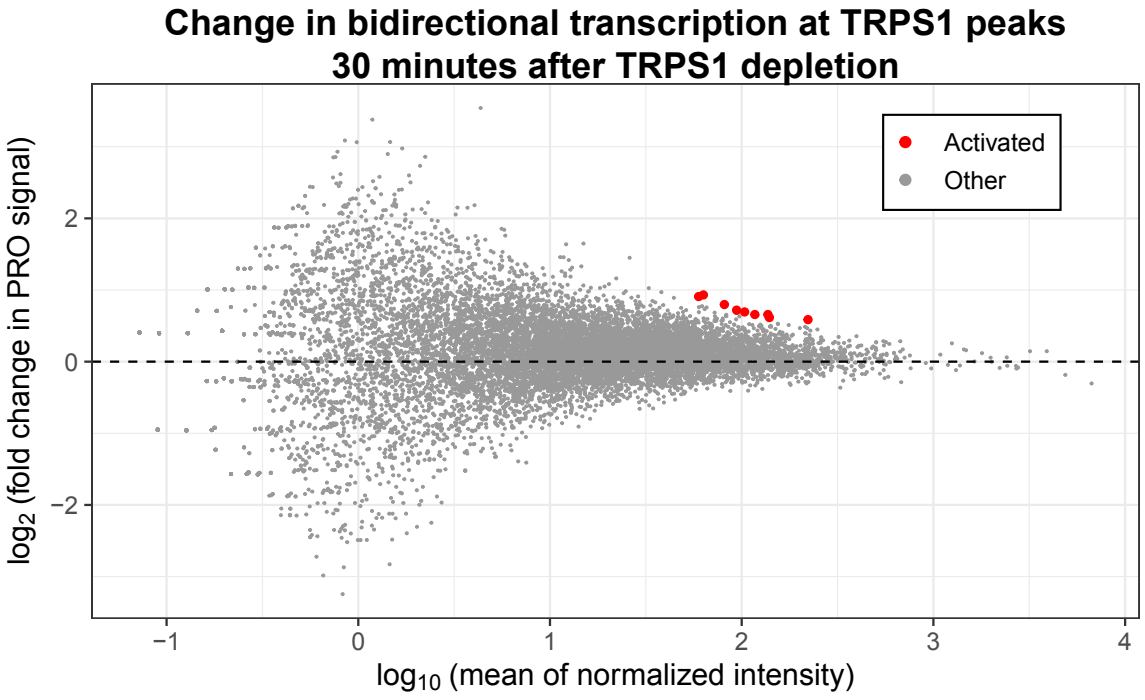

Figure 3.9: Bidirectional transcription at TRPS1 peaks increases upon TRPS1 depletion. MA plot of TRPS1 ChIP-seq peaks from Figure 3.7G, with fold change values representing bidirectional transcription in the 30 minute dTAG-13 and dTAGV -1 at 50nM each (dTAG) treatment condition relative to the DMSO condition. Testing for a TRPS1 cistrome-wide increase in bidirectional transcription, the ANOVA F-test p-value was $< 2.2*10^{-16}$.

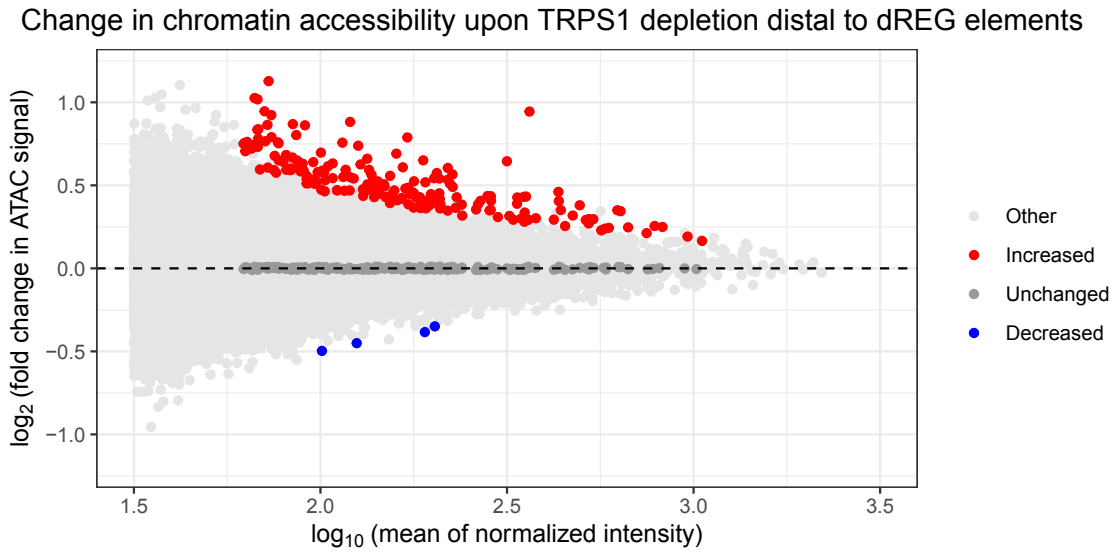

Figure 3.10: Change in chromatin accessibility at ATAC-seq peaks without bidirectional transcription. MA plot of ATAC-seq peaks, with fold change values representing accessibility in the 30 minute dTAG-13 and dTAGV -1 at 50nM each (dTAG) treatment condition relative to the DMSO condition, as in Figure 3.7.

on the genes repressed by TRPS1 depletion are indirect and distal to TRPS1 binding. To test whether TRPS1 directly activates a subset of REs to activate transcription of proximal genes, we measured the distance from the TSS of each gene to the nearest TRPS1 ChIP-seq peak overlapping a decreased ATAC-seq peak. There are few examples of this class of RE, so the distances were much farther, and there was no significant enrichment of repressed genes proximal to these peaks (Figure 3.11F). These data suggest that, while TRPS1 positively and negatively regulates hundreds of primary response genes, TRPS1 only represses transcription of its direct target genes.

3.6.5 TRPS1 redistributes ER binding to modulate ER target gene transcription

Based on the correlation between cancer cell line sensitivity to *TRPS1* knockout and *ESR1* knockout (Figure 3.1B), the enrichment of an ER binding motif in both increased and decreased ATAC-seq peaks (Figure 3.7D), and the over-representation of estrogen response gene sets in the genes repressed by TRPS1 depletion (Figure 3.11D), we focused on ER target genes to explore a

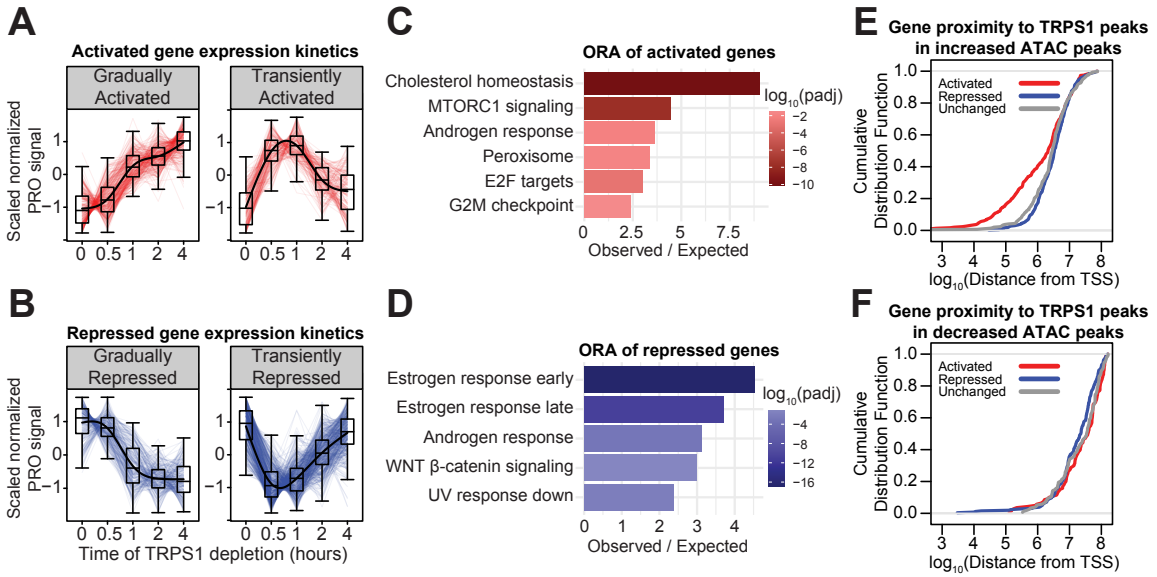

Figure 3.11: TRPS1 directly represses transcription of target genes. A) Kinetic traces of the two major clusters of activated genes over the time course. B) Kinetic traces of the two major clusters of repressed genes over the time course. C) Over-representation analysis of the activated genes from (A). D) Over-representation analysis of the repressed genes from (B). E) Cumulative distribution function plot of the distance from the TSS of each gene to the nearest TRPS1 ChIP-seq peak overlapping an increased ATAC-seq peak, by gene class. Kolmogorov–Smirnov test between activated and unchanged genes: p-value = 0.011. F) Cumulative distribution function plot of the distance from the TSS of each gene to the nearest TRPS1 ChIP-seq peak overlapping an decreased ATAC-seq peak, by gene class. KS test between repressed and unchanged genes: p-value $>$ 0.1.

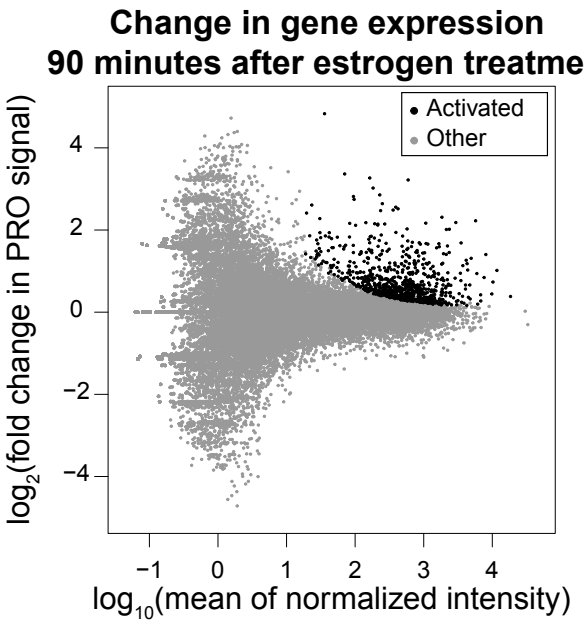

Figure 3.12: Acute estrogen treatment identifies direct ER target genes in T47D cells. MA plot of PRO signal, with fold change values representing transcription in the 90 minute estrogen treatment condition relative to the DMSO condition. Each point represents a gene, and black points represent the estrogen-activated genes that we use in Figure 3.13.

possible mechanism by which TRPS1 indirectly activates transcription. We first defined direct ER target genes using our previously-generated PRO-seq data from parental T47D cells that were hormone starved for three days and then acutely stimulated with estrogen or a DMSO vehicle control for 90 minutes (Figure 3.12) [220]. We exclusively focused on estrogen-activated genes because ER directly activates these genes [37, 222, 223]. 65 ER target genes were activated, and 58 were repressed by acute TRPS1 depletion (Figure 3.13A). To test the robustness of this change in ER target gene transcription, we additionally performed PRO-seq in each of the three independent TRPS1-dTAG clones generated from the parental T47D cells. Indeed, the genes we identified as TRPS1-regulated ER target genes in the one clone used for the time course experiment tend to be regulated in the same direction upon TRPS1 degradation across these three clones, suggesting that this effect is robust across the cell lines in which we can acutely deplete TRPS1 (Figure 3.14).

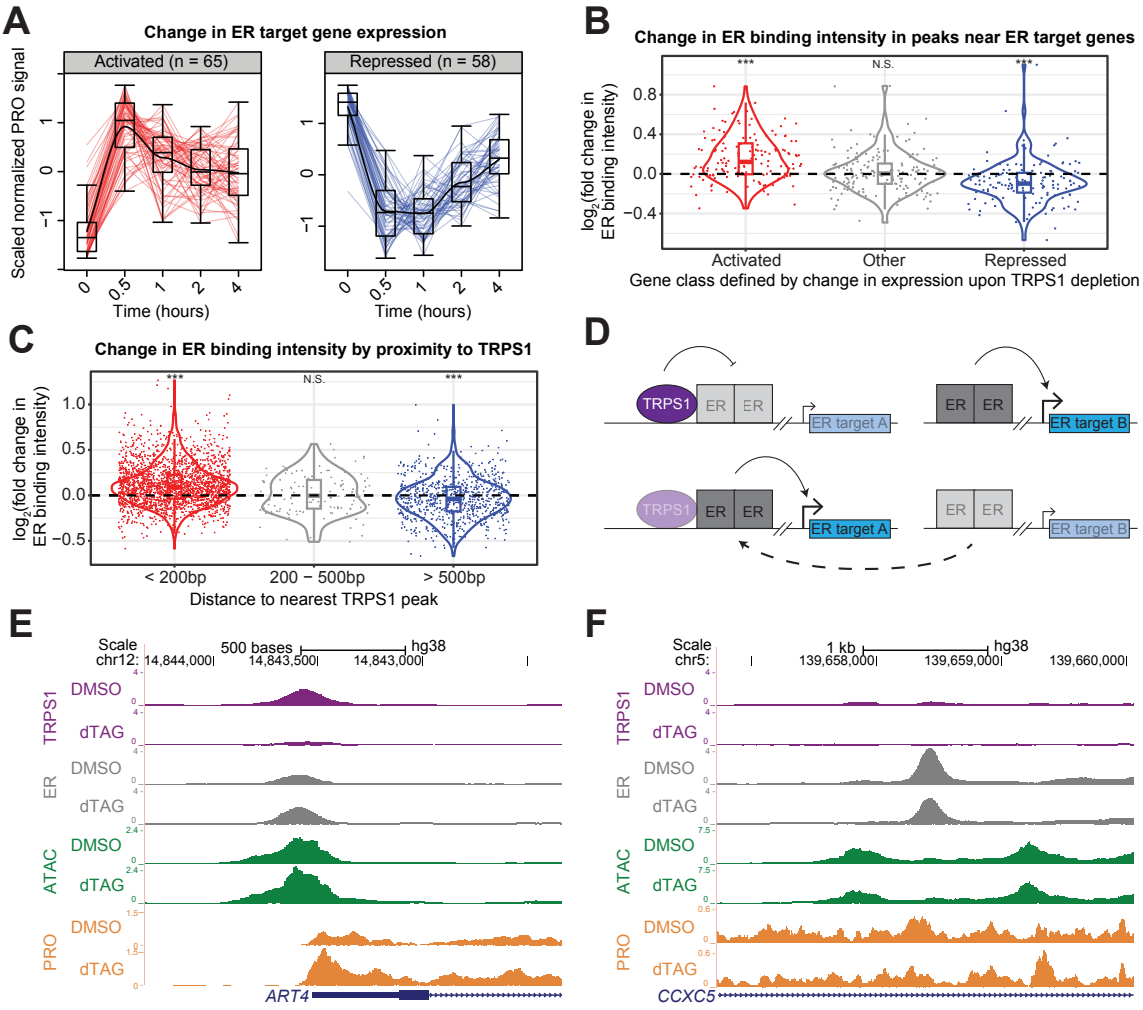
A
Change in ER target gene expression
Activated (n = 65)
Repressed (n = 58)
Scaled normalized PRO signal
Time (hours)
B
Change in ER binding intensity in peaks near ER target genes
log2(fold change in ER binding intensity)
Activated
Other
Repressed
Gene class defined by change in expression upon TRPS1 depletion
C
Change in ER binding intensity by proximity to TRPS1
log2(fold change in ER binding intensity)
< 200bp
200 − 500bp
> 500bp
Distance to nearest TRPS1 peak
D
TRPS1
ER
ER target A
ER target B
E
Scale
chr12:
500 bases
hg38
TRPS1
ER
ATAC
PRO
DMSO
dTAG
ART4
F
Scale
chr5:
1 kb
hg38
CCXC5

Figure 3.13: TRPS1 redistributes ER binding to modulate ER target gene transcription. A) Kinetic traces of activated and repressed ER target genes over the time course. B) Violin and box and whisker plots for ER binding intensity fold change upon TRPS1 depletion at ER ChIP-seq peaks within 100kb of the TSS of each gene, grouped by gene class defined by change in expression upon TRPS1 depletion. (In (B) and (C), *** represents a significant one-sample t-test p-value $< 10^{-3}$. N.S. represents a non-significant p-value > 0.1.) C) Violin and box and whisker plots for ER binding intensity fold change upon TRPS1 depletion at ER ChIP-seq peaks, grouped by summit-to-summit distance to the nearest TRPS1 ChIP-seq peak. D) Model of TRPS1-mediated ER redistribution and modulation of ER target gene transcription. Transparent boxes indicate reduced binding intensity or attenuated transcription. Above, at baseline, TRPS1 directly decreases ER binding intensity proximal to TRPS1, attenuating ER activation of proximal ER target genes. Distal to TRPS1, ER binding intensity is not directly affected by TRPS1, and ER fully activates proximal ER target genes. Below, after TRPS1 depletion, ER binding proximal to TRPS1 increases in intensity, augmenting ER activation of proximal ER target genes. Distal to TRPS1, ER binding intensity is indirectly decreased, as limiting ER molecules are redistributed to TRPS1-proximal regulatory elements, attenuating ER activation of proximal ER target genes. E) ChIP, ATAC, and PRO density around an example increased ER binding site near an activated ER target gene. At this TRPS1-proximal ER binding site, upon dTAG treatment, TRPS1 binding intensity decreases, ER binding intensity increases, chromatin accessibility increases, and gene expression increases. F) ChIP, ATAC, and PRO density around an example decreased ER binding site near an repressed ER target gene. At this TRPS1-distal ER binding site, upon dTAG treatment, ER binding intensity decreases, chromatin accessibility decreases, and gene expression decreases. In (F) and (G), dTAG refers to dTAG-13 and dTAG^{V}-1 at 50nM each.

We hypothesized that these changes in ER target gene transcription are mediated by changes in the genomic distribution of ER binding. We performed ER ChIP-seq to test the prediction that ER binding intensity proximal to dynamic ER target genes would change in concordance with the change in gene transcription. As before, our ChIP-seq libraries were of high quality (Figures 3.3,3.4. We called peaks using MACS2 [218], using all ER samples together and all IgG samples as the control. Consistent with our hypothesis, we found that ER ChIP-seq peaks within a 100 kilobase (kb) window around the TSS of activated genes tended to increase in intensity, and ER ChIP-seq peaks within a 100kb window around the TSS of repressed genes tended to decrease in intensity (Figure 3.13B).

We further hypothesized that only the increased ER binding sites represent a direct effect of TRPS1 activity. Consistent with this hypothesis, we found that ER binding proximal to TRPS1 tends to increase in intensity, and ER binding distal to TRPS1 tends to decrease in intensity (Figure 3.13C). Together, these data suggest a model in which TRPS1 depletion

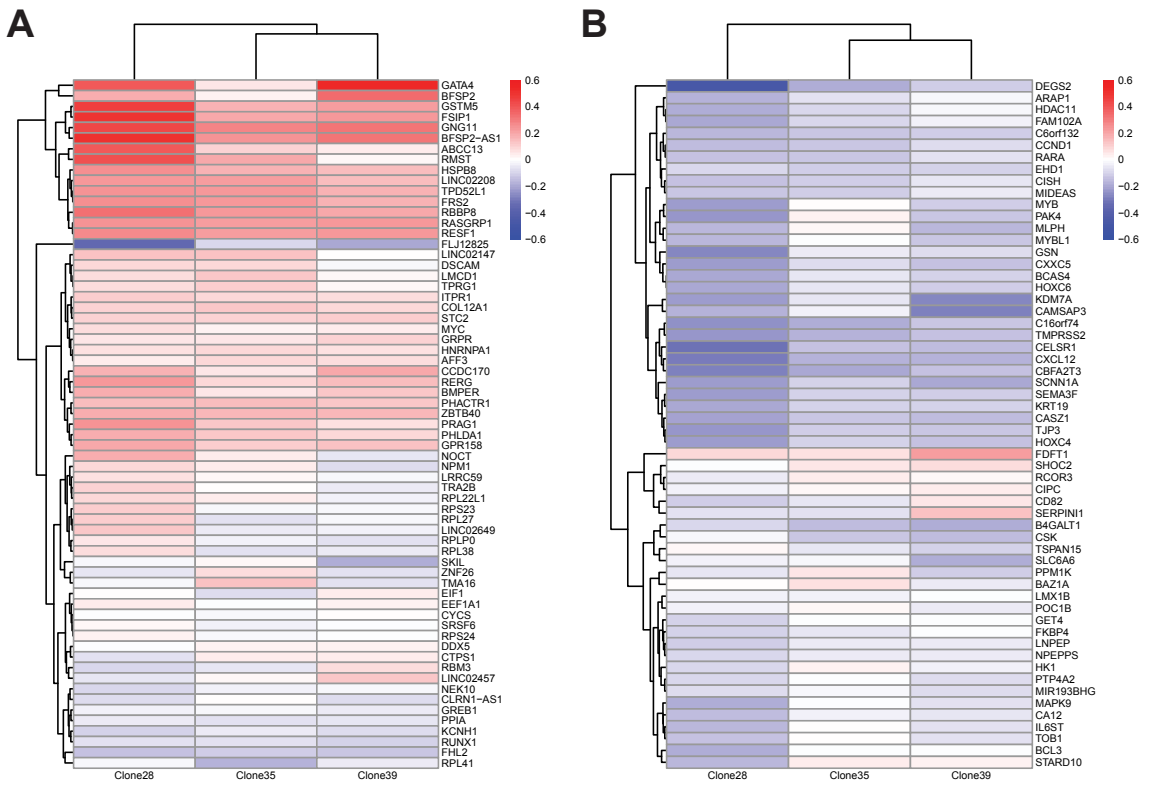

Figure 3.14: The effect of TRPS1 depletion on ER target gene transcription is consistent across three independent clones. Fold changes in normalized PRO signal across three independent clones for ER target genes that are (A) activated or (B) repressed upon TRPS1 depletion, as defined in Figure 3.13A.

redistributes ER binding from TRPS1-distal sites to TRPS1-proximal sites and modulates ER target gene transcription proximal to the dynamic ER binding sites (Figure 3.13D). To illustrate this phenomenon, we provide an example of a TRPS1-proximal, increased ER peak near an activated ER target gene in (Figure 3.13E), and a TRPS1-distal, decreased ER peak near a repressed ER target gene in (Figure 3.13F).

3.6.6 TRPS1 activity is associated with breast cancer patient outcomes

We next sought to connect the primary TRPS1-responsive genes with downstream cellular and patient-related outcomes. We defined a new steady state of transcription with the 24 hour time point after TRPS1 depletion. We ranked genes based on their shrunken fold change in PRO signal (Figure 3.15A). Using this ranking, we performed gene set enrichment analysis with the Hallmark gene sets and found multiple cell-cycle-related gene sets to be negatively enriched, including E2F Targets (Figure 3.15B) [224, 225]. Consistent with this, we observed a significant decrease in cell number doubling rate of T47D dTAG-TRPS1 cells upon TRPS1 depletion (Figure 3.15C). Importantly, the isogenic parental T47D cells do not display a cell number defect with dTAG treatment, so we attribute this effect to TRPS1 depletion and not a non-specific effect of the compounds (Bidirectional transcription at TRPS1 peaks increases upon TRPS1 depletion).

Finally, we calculated a TRPS1 activity score by adapting methods developed by [228, 229]. We used our PRO-seq data to determine a primary TRPS1 regulon based on the differentially expressed genes 30 minutes after TRPS1 depletion. We classified breast cancer patients from the METABRIC cohort as having high TRPS1 activity if both a) TRPS1-repressed genes are negatively enriched and b) TRPS1-activated genes are positively enriched, relative to all other patients in the cohort (example patient in (Figure 3.15D)) [226, 227]. Similarly, we classified patients as having low TRPS1 activity if both a) TRPS1-repressed genes are positively enriched and b) TRPS1-activated genes are negatively enriched. We classified the remaining patients as having intermediate TRPS1 activity. We ranked patients based on their TRPS1 activity and found no association with other clinical covariates (Figure 3.15E). When we stratified patients

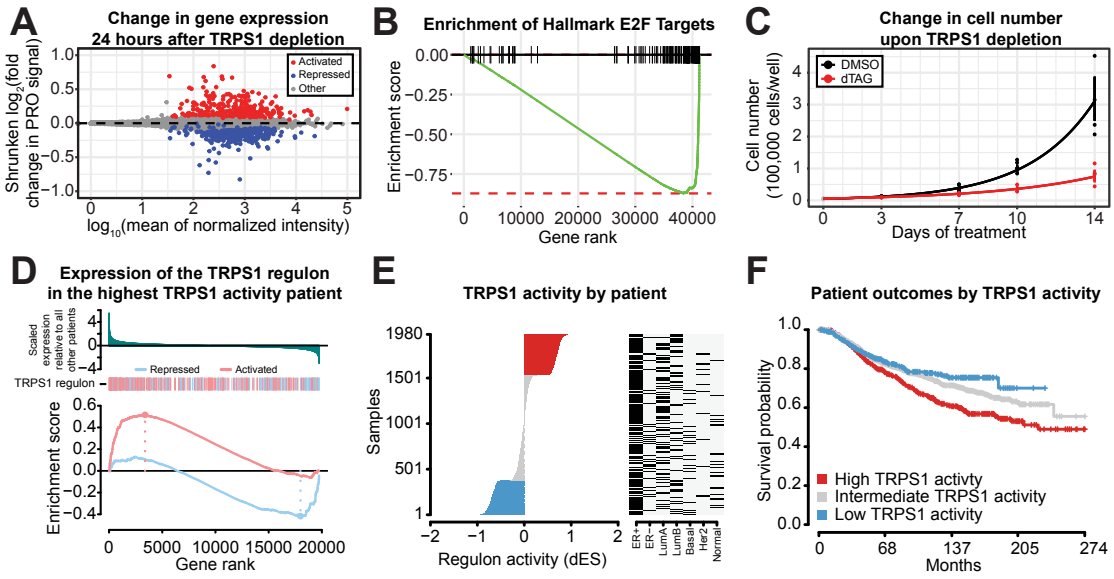

Figure 3.15: TRPS1 activity is associated with breast cancer patient outcomes. A) MA plot of PRO signal in each gene, with shrunken log fold change values representing transcription in the 30 minute dTAG-13 and dTAG^{V}-1 at 50nM each (dTAG) treatment condition relative to the DMSO condition. B) Mountain plot of the Hallmark E2F Targets gene set, using genes ranked by shrunken fold change from (A). A negative enrichment score indicates an enrichment of the gene set among repressed genes. Adjusted p-value $2.5*10^{-19}$. C) Cell number over time of dTAG-TRPS1 cells treated with dTAG or DMSO. Analysis of variance for the coefficient corresponding to the difference in doubling rates between the conditions in a linear model of the logarithm of cell number versus time: p-value $1.1*10^{-5}$. D) Differential enrichment score (dES) calculation for an example patient with the highest TRPS1 activity. Above, genes ranked by scaled expression in this patient relative to all other patients in the METABRIC cohort [226, 227]. Below, gene set enrichment analysis of TRPS1-repressed and TRPS1-activated genes defined by response after 30 minutes of TRPS1 depletion. E) Patients from the METABRIC cohort, ranked by dES as calculated in (D), with classifications of the tumors on the right. F) Kaplan-Meier curves for patients in the METABRIC cohort, stratified by TRPS1 activity as in (E). Logrank p-value $4.99*10^{-4}$.

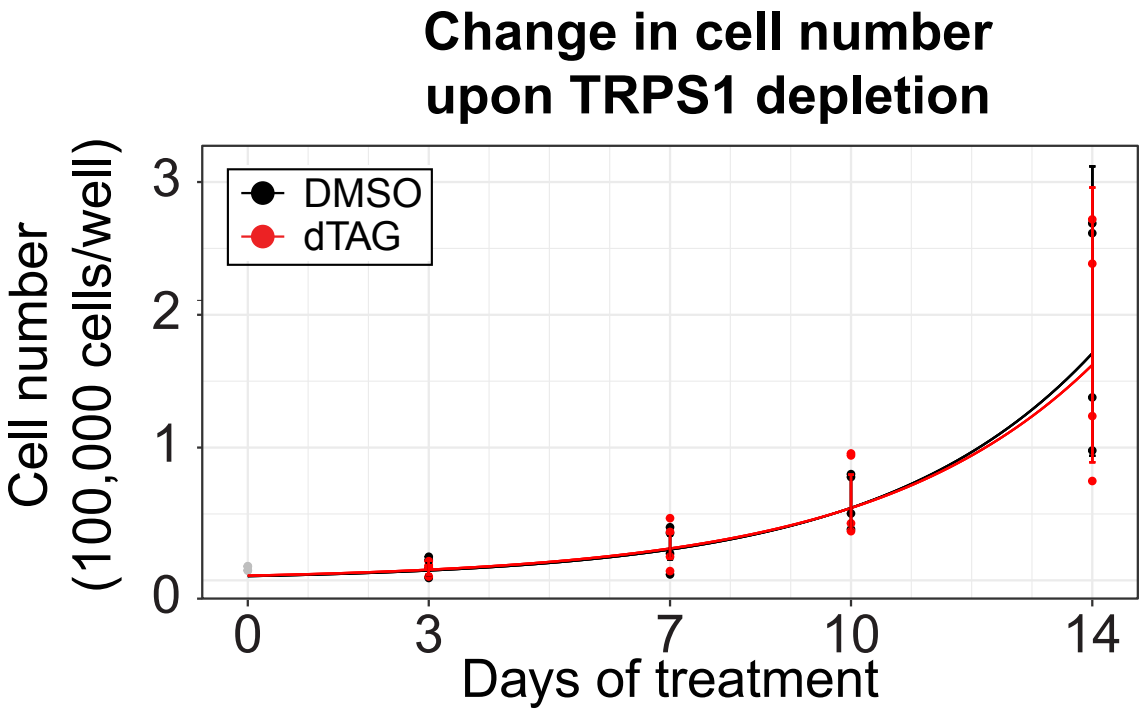

Figure 3.16: T47D cells do not display a cell number defect with dTAG treatment. Cell number over time of parental T47D cells treated with dTAG or DMSO, as in Figure 3.15C.

by TRPS1 activity, we found high TRPS1 activity to be significantly associated with shorter survival time (Logrank p-value $4.99*10^{-4}$) (Figure 3.15F). When we first separated tumors by ER-positivity or by intrinsic subtype, we found that this association was specific for ER-positive tumors and Luminal A tumors (Figure 3.17). We also performed this analysis using genes differentially expressed after 24 hours of TRPS1 depletion (Figure 3.18). However, by this time cell cycle genes dominate, and we speculate that the association with breast cancer patient survival is due to these genes [230]. We believe using the primary response genes offers a unique measure of TRPS1 activity not achievable using previous surrogates.

3.7 Discussion

In this study, we used rapidly inducible targeted protein degradation to systematically determine the primary effects of acute TRPS1 depletion on chromatin accessibility, ER binding, and nascent

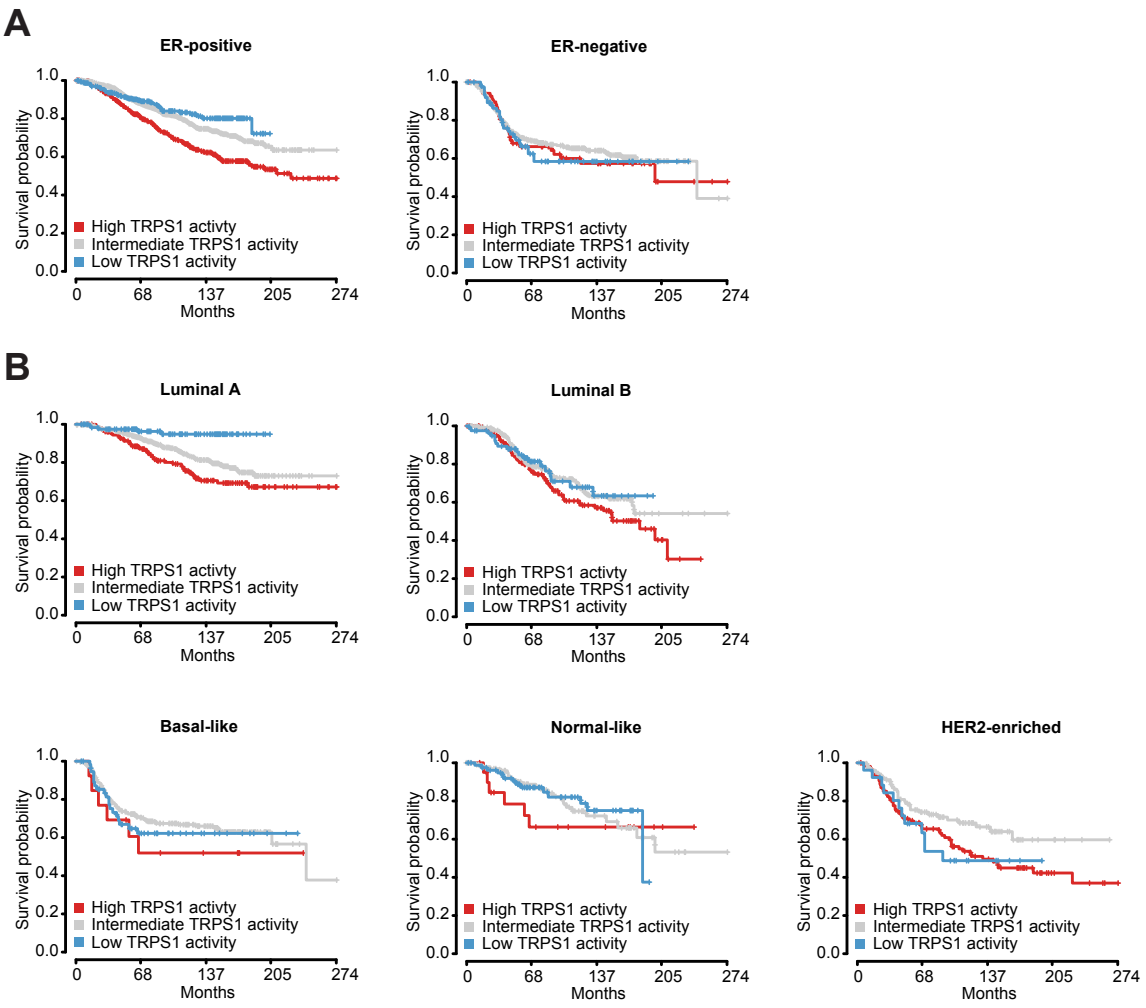

Figure 3.17: TRPS1 activity is associated with breast cancer patient outcomes specifically for ER-positive and Luminal A tumors. A) Kaplan-Meier curves for patients in the METABRIC cohort, stratified by TRPS1 activity as in Figure 3.15F, separated by ER-posivity. Logrank p-value $3.83*10^{-6}$ for ER-positive tumors and not significant for ER-negative tumors. B) Kaplan-Meier curves for patients in the METABRIC cohort, stratified by TRPS1 activity as in Figure 3.15F, separated by intrinsic subtype. Logrank p-value $4.23*10^{-4}$ for Luminal A tumors and not significant for the other subtypes.

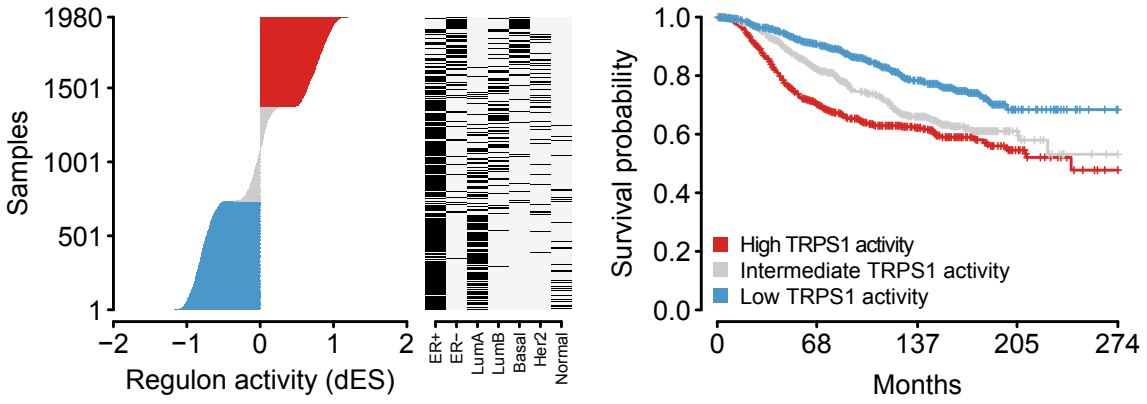

Figure 3.18: Genes differentially expressed after 24 hours of TRPS1 depletion are associated with breast cancer patient outcomes. Kaplan-Meier curves for patients in the METABRIC cohort, stratified by TRPS1 activity as in Figure 3.15F, but using genes differentially expressed after 24 hours of TRPS1 depletion. Logrank p-value $2.09*10^{-13}$.

transcription in a luminal breast cancer cell line. We focused on TRPS1 based on two orthogonal, genome-wide, unbiased assays that implicated *TRPS1* in the processes of breast tumor incidence and breast cancer cell number accumulation.

First, we used the summary statistics from a recent GWAS to plot two sets of common genetic variants in the *TRPS1* locus associated with breast cancer incidence [214]. These genetic variants were independently identified as significantly associated with breast cancer incidence in a previous GWAS [231], but Zhang *et al.* determined that the association was strongest among luminal breast tumors. Second, we analyzed data from the Cancer Dependency Map project and found that sensitivity to *TRPS1* knockout was correlated with sensitivity to *ESR1* knockout and significantly enriched among luminal breast cancer cell lines [77]. Both of these unbiased screens indicate that *TRPS1* contributes to luminal breast cancer cell fitness and led us to the hypothesis that TRPS1 influences ER activity or genomic binding.

As TFs regulate the transcription of many other chromatin-associated factors that themselves regulate RE activity, TF binding, and transcription, we sought to isolate the primary effects of TRPS1 depletion. To do so, we used the dTAG inducible degron tag system to acutely deplete endogenous TRPS1 protein abundance within minutes of induction [89]. This is in

contrast to traditional RNA interference or gene knockout methods, which can take days to deplete the target of interest.

Minutes to hours after TRPS1 depletion, we performed several sensitive, genome-wide assays. ATAC-seq and ChIP-seq can be performed at any time point after a perturbation, as they measure chromatin accessibility and chromatin-associated factor binding, which can change with rapid kinetics. In contrast, changes in messenger RNA abundance accumulate more slowly, with kinetics that depend not only on the rate of nascent transcription but also on the ratio of abundance to synthesis and degradation rates. In contrast, nascent transcriptional profiling measures the immediate change in RNA synthesis rates after a perturbation. Here we use PRO-seq coupled with acute TRPS1 depletion to identify primary TRPS1-responsive genes.

With our cell lines and assays in hand, we first measured changes in chromatin accessibility upon TRPS1 depletion. Consistent with previous studies linking TRPS1 to corepressor complexes, we found that the predominant effect of TRPS1 depletion is an increase in chromatin accessibility and bidirectional transcription at REs [53, 57, 58, 62, 63]. We observed decreasing enrichment of the GATA motif prevalence in increased peaks over time, indicating that our shortest time point was the most specific for isolating the primary effects of TRPS1 depletion. Intriguingly, we identified a significant enrichment of ER half-site motifs in increased as well as decreased ATAC peaks, suggesting that ER binding intensity was changing in a site-specific manner.

We next measured changes in nascent transcription minutes to hours after TRPS1 depletion and clustered the gene responses. Activated genes were enriched for cholesterol homeostasis genes. Of note, several recent GWAS have identified SNPs in the *TRPS1* locus associated with blood cholesterol levels [232–234]. However, as of yet, no mechanistic follow-up studies into the role of TRPS1 in cholesterol biology have been performed. After TRPS1 depletion, repressed genes were enriched for estrogen response gene sets. Consistent with our previous data, activated genes were closer to increased TRPS1-bound ATAC peaks, suggesting TRPS1 directly represses these target genes at steady state. On the other hand, repressed genes were not closer to decreased TRPS1-bound ATAC peaks, suggesting a distinct mechanism of

transcriptional regulation of this gene class.

We hypothesized that at steady state TRPS1 directly represses and indirectly activates its primary response genes. Using ER target genes and ER genomic binding as a case study, we found evidence supporting a model of acute ER redistribution. ER binding sites proximal to TRPS1 tended to increase in intensity upon TRPS1 knockdown, with distal ER binding sites tending to decrease in intensity. Furthermore, genes activated upon TRPS1 depletion were surrounded by ER binding sites that increased in intensity, and repressed genes were near decreased ER binding sites. Taken together, we propose a model in which TRPS1 directly decreases chromatin accessibility at steady state. Upon acute TRPS1 depletion, TRPS1-proximal REs increase in accessibility, an effect which we propose allows ER to redistribute from TRPS1-distal REs. In this proposed model, subsets of ER target genes are activated or repressed by TRPS1 depletion via distinct mechanisms.

First described in the 1980s, the concept of coactivator "squelching" has been debated as a mechanism of indirect activity distal from a TF's genomic binding sites [235, 236]. Squelching has been proposed as a mechanism by which nuclear receptors like ER acutely repress transcription of a subset of primary response genes by competing for limiting coactivators [223, 237–239]. Here we propose not the redistribution of coactivators by an activating TF, but a redistribution of activating TFs themselves via a rapid increase in local chromatin accessibility after the acute depletion of a repressive TF.

Our findings of both increased and decreased ER genomic binding and target gene transcription are distinct from previous studies of the effects of TRPS1 on TF binding and activity. Elster *et al.* used an unbiased screen to identify TRPS1 as a repressor of Yes-associated protein (YAP1) activity in another luminal breast cancer cell line, MCF7 [57]. After TRPS1 knockdown, the authors observed a genome-wide activation of YAP1 target genes. We did not find a YAP1 gene signature among dynamic genes in our PRO-seq data, though we did observe an enrichment of TEA/ATTS domain (TEAD) motifs in increased ATAC peaks, suggesting differences between the cells used in each study and perhaps their baseline YAP1-TEAD activity.

While we would predict that our acute redistribution model is generalizable to other TFs and sets of TRPS1-regulated genes beyond ER and its target genes, it remains possible that TRPS1 modulates the genomic binding intensity of additional TFs in a unidirectional manner via a distinct mechanism.

Serandour *et al.* knocked down *TRPS1* in MCF7 cells and reported both a genome-wide repression of ER target genes as well as a genome-wide increase in ER binding [53]. We did not perform ChIP-seq at a comparable time point to their days-long knockdown, so we cannot directly compare our ER binding data. Our latest PRO-seq time point was 24 hours after TRPS1 depletion, at which time we do not observe a genome-wide repression of ER target genes. This could once again be attributable to a difference in cell lines. However, we would also speculate that the unidirectional and nonconcordant changes in ER binding and target gene expression at later time points could be due to non-primary effects of extended *TRPS1* knockdown.

Finally, we used PRO-seq data from both late and early time points to identify genes that represent cells at a new steady state after TRPS1 depletion, as well as primary TRPS1-responsive genes. After 24 hours of TRPS1 depletion, repressed genes were enriched for cell cycle related genes, consistent with a decrease in cell number doubling rate. Unique to this study, we used primary TRPS1-responsive genes to define a TRPS1 activity score, adapting a method based on predicted TF target genes [228, 229]. Using this method, we were able to stratify breast cancer patients into groups with differing survival probabilities.

Using TRPS1 activity score to classify patients may provide additional insight into the transcriptional program within a patient's tumor that might not be immediately apparent based on previous surrogates for TRPS1 activity. For example, *TRPS1* is frequently amplified in breast tumors, and this amplification is associated with worse prognosis [50, 53]. However, *TRPS1* is often co-amplified along with the rest of the chromosomal segment 8q23–q24, where the proto-oncogene *MYC* resides, making it difficult to discern whether *TRPS1* amplification is a driver of breast cancer progression [240].

In contrast, higher *TRPS1* expression has been associated with better breast cancer

patient outcomes, though its expression is highly correlated with ER and GATA3, both favorable prognostic indicators [52, 241]. As relative TF expression and activity across patients are not always identical, our data uses primary TRPS1-responsive genes as a measure of TRPS1 activity. Our TRPS1 activity score is not correlated with ER-positivity and effectively stratifies patients. Though our patient outcome analysis, as with all similar analyses, describes an association and does not necessarily imply a causative relationship, the direction is consistent with the effect on cell number observed in this study as well as the Cancer Dependency Map, suggesting that TRPS1 drives breast cancer cell number accumulation.

Altogether, we provide a systematic study of the primary effects of rapid TRPS1 depletion in luminal breast cancer cells. We propose a model in which TRPS1 depletion leads to decondensation of local chromatin structure, allowing for the acute redistribution of ER, both activating and repressing subsets of ER target genes. This TRPS1-regulated transcription appears to be relevant for cancer cell fitness, as TRPS1 depletion decreases cell number doubling rate, and high TRPS1 activity is associated with worse breast cancer patient outcomes. These methods of inducible targeted protein degradation coupled with genomic chromatin assays and nascent RNA transcriptional profiling should in principle be applicable to the study of any TF, allowing us to better understand the mechanisms behind the phenotypes associated with additional GWAS hits.

3.8 Methods

3.8.1 GWAS and DepMap data visualization

Summary statistics from [214] were downloaded from the NHGRI-EBI GWAS Catalog [215]. SNPs in the *TRPS1* locus were plotted using LocusZoom [216]. Knockout scores and luminal breast cancer identifiers were downloaded from the Cancer Dependency Map project [77] and plotted using the statistical programming language R [242].

3.8.2 Cell culture

T47D cells (RRID:CVCL_0553) (ATCC) were cultured in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (Gemini) and 10µg/ml insulin from bovine pancreas (Sigma, made as a 1000x solution in 1% aqueous glacial acetic acid).

3.8.3 Plasmid generation for gene editing

DNA for transfection was prepared as previously described [94, 152]. A CRISPR sgRNA (TTATCTTTGCAGATATGGTC) targeting the 5′ end of the *TRPS1* coding sequences was designed using Benchling. The sgRNA was cloned into hSpCas9 plasmid PX458 (Addgene #48138) as previously described [243], using the following primers:

5′-CACCGTTATCTTTGCAGATATGGTC-3′

and

5′-AAACGACCATATCTGCAAAGATAAC-3′ .

A plasmid harboring a synthetic HygR-P2A-2xHA-FKBP_F36V insert was generated with Cold Fusion (System Biosciences), starting with the HygR-P2A-AID cassete in pMGS58 (Addgene #135311) [152] and the Puro-P2A-2xHA-FKBP_F36V casette in (Addgene #91793) [89]. The linear donor was generated by PCR using primers (IDT) that contain 50-nucleotide homology tails and gel-purified. The primers contained 5′ phosphorothioate modifications to increase PCR product stability in the cell [244]. The primers used for making PCR donor fragments were:

5′-G*T*AACTTTCAGATAACACTGTATCTGCCTTTTCCCTTTATCTTTGCAGATATGAAAA
AGCCTGAACTCACCG-3′

and

5′-T*T*CACTTGCAACGTTTCTCAGAGGGGGGTTCTTTTTCCGGACACCTGAACCTGAAC
CTCCAGATCCACCAGATCTTTCCAGTTTTAGAAGCTCCACATCG-3′

with asterisks representing the phosphorothioate modifications.

3.8.4 dTAG-TRPS1 clone generation

Clones were generated as previously described [94, 152], with modifications. An initial round of cloning was performed using puromycin selection, but upon genomic DNA sequencing this clone did not appear to have a dTAG insertion event within *TRPS1*. Nevertheless, this clone was used for a second round of cloning using hygromycin selection. $3*10^6$ cells were plated in 10cm plates. The next day, cells were cotransfected with 15µg of CRISPR/Cas9-sgRNA plasmid and 1.85µg of linear donor PCR product using Lipofectamine 3000 (Thermo Fisher Scientific) in Optimem (Gibco). One day after transfection, the media was replaced. Starting four days after transfection, cells were selected for two weeks with 200µg/mL of Hygromycin B (Invitrogen) with 20% conditioned media, replaced twice per week. Colonies were then grown in 20% conditioned media, replaced twice per week, until they were large enough to be picked and passaged to a 24-well plate. Clones were expanded and frozen at 8 passages after transfection. Integration was tested with Western blotting, PCR, and Sanger sequencing. In each of the three clones, two to three of the four genomic copies of *TRPS1* are knocked out, and only tagged TRPS1 protein is expressed. Details about each of the determined alleles are available at `https://guertinlab.github.io/TRPS1_ER_analysis/Vignette.html#allele-sequencing`.

3.8.5 Western blotting

$8*10^5$ cells per sample were plated in each well of a 6-well plate. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO at various time points and collected simultaneously. At the time of harvest, cells were scraped and lysed in RIPA buffer (1% Nonidet P-40, 1% sodium deoxycholate, 0.1% sodium dodecyl sulfate, 2mM EDTA, 150mM NaCl, 10mM sodium phosphate, 50mM NaF, 50mM Tris pH 7.5), with 100µM benzamidine, 5µg/mL aprotinin, 5µg/mL leupeptin, 1µg/mL pepstatin, 1mM phenylmethylsulfonyl fluoride, and 2mM sodium orthovanadate added fresh. Lysates were sonicated in a Biorupter UCD-200 (Diagenode) on high for 30 seconds on and 30 seconds off for 5 cycles, and clarified by centrifugation at 14,000rpm for 15 min in 4°C. Protein concentration was measured by BCA assay and diluted

to the same concentration. 10× Laemmli buffer was added to a final concentration of 1x, and 2-mercaptoethanol was added to a final concentration of 1%. Samples were boiled at 95°C for 10 minutes, and 30µg of each was loaded into a 10% polyacrylamide gel. Samples were separated by gel electrophoresis and transferred to nitrocellulose membranes. Membranes were incubated in blocking buffer (3% bovine serum albumin, 1X Tris buffered saline) for 1 hour at room temperature with rocking. Primary antibodies (anti-TRPS1, Cell Signaling #17936S, and anti-ACTB, Cell Signaling #3700S) were diluted 1:1,000 in primary buffer (3% bovine serum albumin, 0.1% sodium azide, 0.1% Tween-20, 1X Tris buffered saline) at 4°C with rocking overnight. Fluorescent secondary antibodies were diluted 1:10,000 in secondary buffer (5% bovine serum albumin, 0.1% sodium azide, 0.1% Tween-20, 1X Tris buffered saline) and incubated for 1 hour at room temperature with rocking, and fluorescence was measured (Odyssey, Licor).

3.8.6 ChIP-seq library preparation

$2.4*10^7$ cells per sample were plated across 3 15cm dishes 2 days before harvest. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO for 30 minutes and collected simultaneously. At the time of harvest, cells were fixed with 1% formaldehyde (Sigma) for 10 minutes at 37°C and quenched with 125mM Glycine (Fisher) for 10 minutes at 37°C. Plates were moved to ice, and cells were washed and scraped into ice cold PBS containing Complete EDTA-free Protease Inhibitor Cocktail (Roche). Cells were pelleted in aliquots of $3.6*10^7$ cells, snap frozen in liquid nitrogen, and stored at -80°C. Pellets were thawed, and cells were lysed in 1mL Cell Lysis Buffer (85mM KCl, 0.5%NP40, 5mM PIPES pH 8.0), with protease inhibitor cocktail added fresh, for 10 minutes with rotation at 4°C. Nuclei were pelleted at 3300g at 4°C for 5 minutes and resuspended in 500µL ChIP lysis buffer (0.5% SDS, 10mM EDTA, 50mM Tris-HCl pH 8.1), with protease inhibitor cocktail added fresh, for 10 minutes with rotation at 4°C. Lysates were moved to 15ml polystyrene conical tubes (Falcon) and sonicated in a Biorupter UCD-200 (Diagenode) on high for 30 seconds on and 30 seconds off for 4 sets of 5 cycles. Before each set, ice in the water bath was replaced, and samples were gently vortexed to mix. Sonicated lysates were then move to 1.5ml tubes and clarified by centrifugation at 14,000rpm for 15 min in

4°C. 500µL of the supernatant was diluted into 6.5mL Dilution Buffer (0.01% SDS, 1.1% Triton X-100, 1.2mM EDTA, 167mM NaCl, 16.6mM Tris-HCl pH 8.0), with protease inhibitor cocktail added fresh ($1*10^6$ cells in 200µL). 1ml ($5*10^6$ cells) was aliquoted into each of 3 tubes with antibody (1.25 µg anti-HA, Cell Signaling #3724S, 2.5µg anti-ER, Millipore #06-935, or 2.5µg IgG control, Cell Signaling #2729S), and incubated with end-over-end rotation at 4°C overnight.

50µL Protein A/G Magnetic Beads (Pierce) per sample were washed with bead washing buffer (PBS with 0.1% BSA and 2mM EDTA) and then incubated with samples for 2 hours with rotation at 4°C. The samples were washed once each with low salt immune complex buffer (0.1% SDS, 1% Triton x-100, 2mM EDTA, 150mM NaCl, 20mM Tris HCl pH 8.0), high salt immune complex buffer (0.1% SDS, 1% Triton x-100, 2mM EDTA, 500mM NaCl, 20mM Tris Hcl pH8.0), LiCl immune complex buffer (0.25M LiCl, 1% NP-40, 1% deoxycholate, 1mM EDTA, 10mM Tris-HCl pH8.0), and 1xTE (10mM Tris-HCl, 1mM EDTA pH8.0). Immune complexes were eluted in elution solution, (1% SDS, 0.1M sodium bicarbonate) in a thermomixer for 30 min at 65°C at 1,200rpm. Crosslinks were reversed and proteins were digested with the addition of 200mM NaCl and 2ul Proteinase K in a thermocycler at 65°C for 16 hours. DNA was purified with a Qiaquick PCR cleanup (Qiagen), and libraries were prepared with a NEBNext Ultra II Library Prep Kit (New England Biolabs).

3.8.7 ChIP-seq analysis

Adapters were removed using `cutadapt` [190]. Reads were aligned to the *hg38* genome assembly with `bowtie2` [189]. Duplicate reads were removed, and the remaining reads were sorted into *BAM* files and converted to *bed* format for counting with `samtools` [194]. Reads were also converted to *bigWig* format with `deeptools` [245]. Peaks were called with `MACS2` [218]. Reads were counted in peaks using `bedtools`, and differentially bound peaks were identified with `DESeq2` [188, 198]. Heatmaps were generated with `deeptools`. Peak proximity to and overlap with other features were calculated with `bedtools`.

3.8.8 ATAC-seq library preparation

ATAC-seq libraries were prepared as previously described [246], with modifications. 4 replicates were performed from cells treated and collected at different times in the same day. $4*10^5$ cells per sample were plated in each well of a 6-well plate 2 days before harvest. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO at various time points and collected simultaneously. At the time of harvest, cells were moved to ice and scraped in 1mL ice cold PBS, and 100µL ($\sim$5 x 10^4 cells) were transferred to 1.5 mL tubes. Cells were centrifuged at 500 x *g* for 5 minutes at 4°C, and the pellets were resuspended in 50 µL cold lysis buffer (10mM Tris-HCl, 10mM NaCl, 3mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin, adjusted to pH 7.4) and incubated on ice for 3 minutes. Samples were washed with 1 mL cold wash buffer (10mM Tris-HCl, 10mM NaCl, 3mM $MgCl_2$, 0.1% Tween-20). Cells were centrifuged at 500 x *g* for 10 minutes at 4°C, and pellets were resuspended in the transposition reaction mix (25 µL 2X TD buffer (Illumina), 2.5 µL TDE1 Tn5 transposase (Illumina), 16.5 µL PBS, 0.5 µL 1% Digitonin, 0.5 µL 10% Tween-20, 5 µL nuclease-free water) and incubated in a thermomixer at 37°C and 100rpm for 30 minutes. DNA was extracted with the DNA Clean and Concentrator-5 Kit (Zymo Research). Sequencing adapters were attached to the transposed DNA fragments using NEBNext Ultra II Q5 PCR mix (New England Biolabs), and libraries were amplified with 8 cycles of PCR. PEG-mediated size fractionation [247] was performed on the libraries by mixing SPRIselect beads (Beckman) with each sample at a 0.5:1 ratio, then placing the reaction vessels on a magnetic stand. The right side selected sample was transferred to a new reaction vessel, and more beads were added for a final ratio of 1.8:1. The final size-selected sample was eluted into nuclease-free water. This size selection protocol was repeated to further remove large fragments.

3.8.9 ATAC-seq analysis

Adapters were removed using `cutadapt` [190]. Reads aligning to the mitochondrial genome with `bowtie2` [189] were removed. The remaining reads were aligned to the *hg38* genome assembly with `bowtie2`. Duplicate reads were removed, and the remaining reads were sorted into *BAM*

files with `samtools` [194]. Reads were converted to *bed* format with `seqOutBias` and *bigWig* format with `deeptools` [195, 245]. Accessibility peaks were called with `MACS2` [218]. Reads were counted in peaks using `bedtools`, and differentially accessible peaks were identified with `DESeq2` [188, 198]. *de novo* motif identification was performed on dynamic peaks with `MEME`, and `TOMTOM` was used to match motifs to the HOMER, Jaspar, and Uniprobe TF binding motif databases [248–251]. AME was used to identify motifs enriched in increased or decreased peaks relative to unchanged peaks [252]. `FIMO` and `bedtools` were used to assess motif enrichment around peak summits [253]. Dynamic peaks were clustered into response groups using `DEGreport` [254].

3.8.10 PRO-seq library preparation

Cell permeabilization was performed as previously described [213], with modifications. 4 replicates were performed from cells treated and collected at different times in the same day. For the time course experiment, $8*10^6$ dTAG-TRPS1 Clone 28 cells per sample were plated in 15cm dishes 2 days before harvest. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO at various time points and collected simultaneously. For the three clone experiment, $4*10^6$ cells per sample were plated in 10cm dishes 1 day before harvest. Cells were treated with DMSO or 100nM dTAG-13 in DMSO for 90 minutes and collected simultaneously.

At the time of harvest, cells were scraped in 10mL ice cold PBS and washed in 5mL buffer W (10mM Tris-HCl pH 7.5, 10mM KCl, 150mM sucrose, 5mM $MgCl_2$, 0.5mM $CaCl_2$, 0.5mM DTT, 0.004U/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)). Cells were permeabilized by incubating with buffer P (10 mM Tris-HCl pH 7.5, KCl 10 mM, 250 mM sucrose , 5 mM $MgCl_2$, 1 mM EGTA, 0.05% Tween-20, 0.1% NP40, 0.5 mM DTT, 0.004 units/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)) for 3 minutes on ice. Cells were washed with 10 mL buffer W before being transferred into 1.5mL tubes using wide bore pipette tips. Finally, cells were resuspended in 50µL buffer F (50mM Tris-HCl pH 8, 5mM $MgCl_2$, 0.1mM EDTA, 50% Glycerol, 0.5 mM DTT). Cells were snap frozen in liquid nitrogen and stored at -80°C.

PRO-seq libraries were prepared as previously described [255], with modifications. RNA extraction after the run-on reaction was performed with 500µL Trizol LS (Thermo Fisher) followed by 130µL chloroform (Sigma). The equivalent of 1µL of 50µM for each adapter was used. A random eight base unique molecular identifier (UMI) was included at the 5′ end of the adapter ligated to the 3′ end of the nascent RNA. 37°C incubations were performed with rotation with 1.5mL tubes placed in 50mL conical tubes in a hybridization oven. For the reverse transcription reaction, RP1 was used at 100µM and dNTP mix was used at 10mM each. Libraries were amplified by PCR for a total of 8 cycles in 100µL reactions with Phusion polymerase (New England Biolabs). No PAGE purification was performed to ensure that our libraries were not biased against short nascent RNA insertions.

3.8.11 PRO-seq analysis

Adapters were removed using `cutadapt` [190]. Libraries were deduplicated using `fqdedup` and the 3′ UMIs [193]. UMIs were removed, and reads were reverse complemented with the `seqtk`. Reads aligning to the rDNA genome with `bowtie2` [189] were removed. The remaining reads were aligned, sorted, and convert to *bed* and *bigWig* files with `bowtie2`, `samtools`, `seqOutBias`, and `deeptools`, respectively [194, 195, 245]. Composite profiles around TRPS1 peaks were generated with `deeptools`. Reads were counted in genes using `bedtools`, and differentially expressed genes were identified with `DESeq2` [188, 198]. Dynamic genes were clustered into response groups using `DEGreport` [254]. Over-representation analysis was performed with `enrichr` [256], and gene set enrichment analysis was performed with `fgsea` [257], both using the Hallmark gene sets [221].

3.8.12 Genome browser visualization

Genome browser [258] images were taken from the following session: https://genome.ucsc.edu/s/tgscott/dTAG_TRPS1_ChIP_PRO_ATAC.

3.8.13 Cell number enumeration

$1.25*10^4$ cells per sample were plated in a 24-well plate. The next day (day 0), cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO. Media was replaced, maintaining the treatment condition, every 2 days. Cells were enumerated using a hemocytometer. 2 technical replicates were used on day 0, 2 for each treatment on day 3, and 3 for each treatment on days 7, 10, 14. The technical replicates were merged, and the experiment was performed in 4 biological replicates from different cell passages. The data were imported into R [242] for visualization and statistical analysis. A linear model was fit for the log-transformed cell number and the time. A second linear model was fit that included an interaction term between the time and the treatment condition, representing the effect of treatment on the doubling rate. Analysis of variance was performed on the two models to test for the significance of the interaction term.

3.8.14 TRPS1 activity score and patient outcome stratification

Primary TRPS1-regulated genes were defined based on the 30 minute time point using `DESeq2` [198]. This TRPS1 regulon was then used in `RTN` [228, 229] to define a TRPS1 activity score for each patient within the METABRIC cohort [226, 227].

3.9 Data Access

All analysis details and code are available at https://guertinlab.github.io/TRPS1_ER_analysis/Vignette.html. Raw sequencing files and processed counts and *bigWig* files are available from GEO SuperSeries accession record GSE236176, with SubSeries accession records GSE236175 (ATAC-seq), GSE236174 (ChIP-seq), and GSE236172 (time course PRO-seq), and GSE251772 (three clone PRO-seq).

3.10 Competing Interest Statement

The authors declare no competing interests.

3.11 Acknowledgements

This work was funded by R35-GM128635 to MJG. 5T32GM007267-39 and 5T32GM008136-35 supported TGS. We thank Arun Dutta, Jacob Wolpe, Devin Roller, and Adam Spencer for critical feedback.

Chapter 4

TRPS1 represses transcription of cholesterol biosynthesis genes and is associated with blood cholesterol traits

4.1 Preface

This chapter is unpublished.

Thomas G. Scott, Daniel Gioeli, Michael J. Guertin

4.2 Author contributions

All authors contributed to the conceptualization of the project. TGS performed the experiments, analyzed the data, and wrote the original draft of the manuscript. All authors reviewed and edited the manuscript.

TRPS1 represses transcription of cholesterol biosynthesis genes and is associated with blood cholesterol traits

Thomas G. Scott[a], Daniel Gioeli[b,c], Michael J. Guertin[d,e]

[a]Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia, United States of America

[b]Department of Microbiology, Immunology, and Cancer, University of Virginia, Charlottesville, Virginia, United States of America

[c]Cancer Center Member, University of Virginia, Charlottesville, Virginia, United States of America

[d]Center for Cell Analysis and Modeling, University of Connecticut, Farmington, Connecticut, United States of America

[e]Department of Genetics and Genome Sciences, University of Connecticut, Farmington, Connecticut, United States of America

4.3 Abstract

Cardiovascular disease (CVD) is the leading cause of death worldwide. Inhibition of cholesterol biosynthesis reduces morbidity and mortality due to CVD. Multiple genome-wide association studies (GWAS) have identified associations between common genetic variants and blood cholesterol traits in the human population. We still lack a mechanistic understanding of how most GWAS hits influence their associated traits. TRPS1 is an atypical GATA family transcription factor associated with corepressor complexes and transcriptional repression of its target genes. We previously generated clones in the luminal breast cancer cell line T47D in which can rapidly degrade TRPS1 endogenously fused with an inducible degron tag. Following up on an observation from one of the clones, here we show that acute TRPS1 depletion in three independent clones activates cholesterol biosynthesis gene transcription. We queried GWAS data to demonstrate that common genetic variants in the *TRPS1* locus are associated with blood cholesterol traits. Unexpectedly, we do not observe a change in steady state cholesterol biosynthesis gene mRNA, lipid droplet, or cholesterol abundance after 48 hours of TRPS1 depletion.

4.4 Introduction

Cardiovascular disease (CVD) affects over 6% of the global population and is the leading cause of death worldwide [259]. Blood cholesterol traits, including low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol levels have long been known to be risk factors for CVD [260, 261]. Accordingly, cholesterol-lowering medications such as statins decrease morbidity and mortality due to CVD [262]. The mechanism of action of statins is the inhibition of 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR), the rate-limiting enzyme in cholesterol biosynthesis [263]. Additional therapeutic targets, such as proprotein convertase subtilisin/kexin type 9 (PCSK9) have been identified by studying rare familial cases of hypercholesterolemia [264, 265].

Multiple genome-wide association studies (GWAS) have identified associations between hundreds of clusters of common genetic variants and blood cholesterol traits in the human population [266–269]. These GWAS hits can be used to generate polygenic risk scores to predict cholesterol traits without additional study [270]. However, additional experiments are needed to improve our understanding of the underlying biology regulating cholesterol homeostasis in human physiology.

TRPS1 is a member of the GATA family of transcription factors [60]. Unique among this family, TRPS1 represses transcription of its target genes via its carboxy-terminal IKZF1-like zinc fingers [61]. TRPS1 interacts with corepressors, including members of the NuRD and coREST complexes, to regulate transcription of its target genes [53, 57, 58, 62, 63].

We previously generated three independent clones from the luminal breast cancer cell line T47D in which we endogenously tagged TRPS1 with the dTAG inducible degron tag (Chapter 3). In these cells, we can rapidly degrade TRPS1 by the addition of the small molecule dTAG-13 and dTAG^{V}-1 [89]. When we depleted TRPS1, we noticed that cholesterol biosynthesis genes were over-represented among the activated genes. Here we follow up this observation with additional experiments to test whether TRPS1 regulates additional measures of cholesterol biosynthesis.

4.5 Results

4.5.1 TRPS1 represses cholesterol biosynthesis gene transcription

We first set out to validate our previous observations from one clone in all three clones. We performed precision run-on with sequencing (PRO-seq) to measure nascent transcription after 90 minutes of TRPS1 depletion in four replicates of each of the three clones [174]. We used DESeq2 to identify differentially expressed genes [198]. 224 genes were significantly activated, and 116 genes were repressed, at a false discovery rate (FDR) of 0.1 (Figure 4.1A). We performed over-representation analysis on the activated genes using the Reactome database of gene sets [256, 271, 272] (Figure 4.1B). The most significantly over-represented gene set was Cholesterol Biosynthesis. We specifically queried the significantly activated genes within this gene set and found seven to be activated in each of the three clones upon TRPS1 depletion (Figure 4.1C).

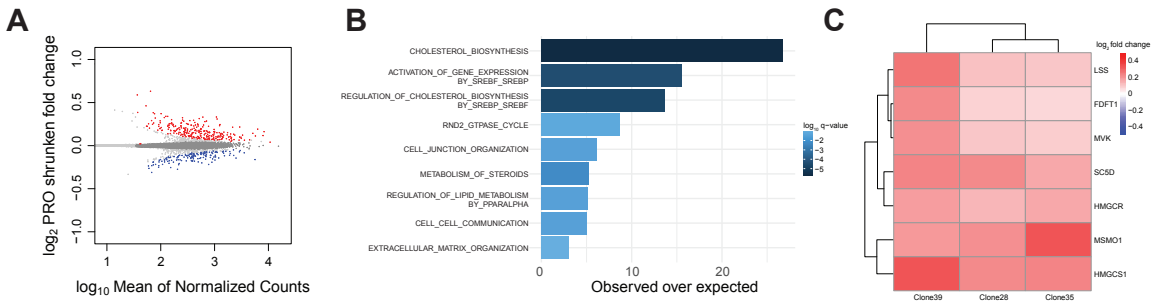

Figure 4.1: TRPS1 represses cholesterol biosynthesis gene transcription. A) MA plot of PRO-seq signal in genes, with shrunken fold change values representing transcription in the 100nM dTAG-13 (TRPS1-depleted) condition relative to the DMSO condition. Red points represent significantly activated genes, and blue points represent significantly repressed genes at an FDR of 0.1. B) Over-representation analysis of the activated genes from (A), using the Reactome database of gene sets. C) Heatmap of PRO signal in the seven significantly activated genes in the Cholesterol Biosynthesis gene set, with fold change values representing transcription in the 100nM dTAG-13 condition relative to the DMSO condition.

4.5.2 *TRPS1* is associated with blood cholesterol traits

To assess the relevance of *TRPS1* to cholesterol regulation in humans, we searched for publicly available GWAS data as orthogonal evidence of the importance of TRPS1 to cholesterol

homeostasis in humans. When we queried the NHGRI EBI GWAS Catalog to find published associations with genetic variants within the *TRPS1* genomic locus, we found several significant single nucleotide polymorphisms (SNPs) associated with blood cholesterol traits [215]. To further investigate these SNPs, we downloaded summary statistics from the UK Biobank [266].

We used LocusZoom to plot the data within this locus (Figure 4.2A) [216]. A plot of these data indicates that SNPs within the *TRPS1* gene that are associated with HDL and LDL cholesterol levels in the blood (Figure 4.2A,B). Furthermore, another cluster of SNPs upstream from *TRPS1* are associated with cholesteryl esters to total lipids ratio in small HDL (Figure 4.2C). We used data from the Genotype-Tissue Expression (GTEx) project, which identified expression quantitative trait loci (eQTL) within this region, and analyzed it with eQTpLot [273–275]. We found that the SNPs that were the most significant in the GWAS were the most significant in the eQTL study. The directions of effect were incongruous, meaning SNPs associated with higher TRPS1 expression were associated with lower values of the cholesterol trait (Figure 4.2C,D,E).

To further test the hypothesis that the same putative causual SNP associated with TRPS1 expression is also associated with the GWAS cholesterol trait, we performed Bayesian colocalization analysis. This method produces posterior probabilities for five hypotheses. H_0: No association with either trait. H_1: Association with trait 1 but not with trait 2. H_2: Association with trait 2 but not with trait 1. H_3: Association with trait 1 and trait 2, with two independent SNPs. H_4: Association with trait 1 and trait 2, with one shared SNP [276]. Starting with standard priors, including P_{12}, the probability that one SNP is associated with both traits, as 10^{-6}, we found the posterior probability for H_4 to be 0.9.

Finally, we searched COLOCdb, a publicly available resource of colocalization analyses with GWAS and QTL studies [277]. We found that SNPs associated with HDL cholesterol colocalized with SNPs associated with methylation levels, chromatin accessibility levels, and histone acetylation levels within the *TRPS1* locus, each with posterior probabilities for $H_4 > 0.85$. Collectively, these analyses suggest that TRPS1 influences cholesterol biology in the human population.

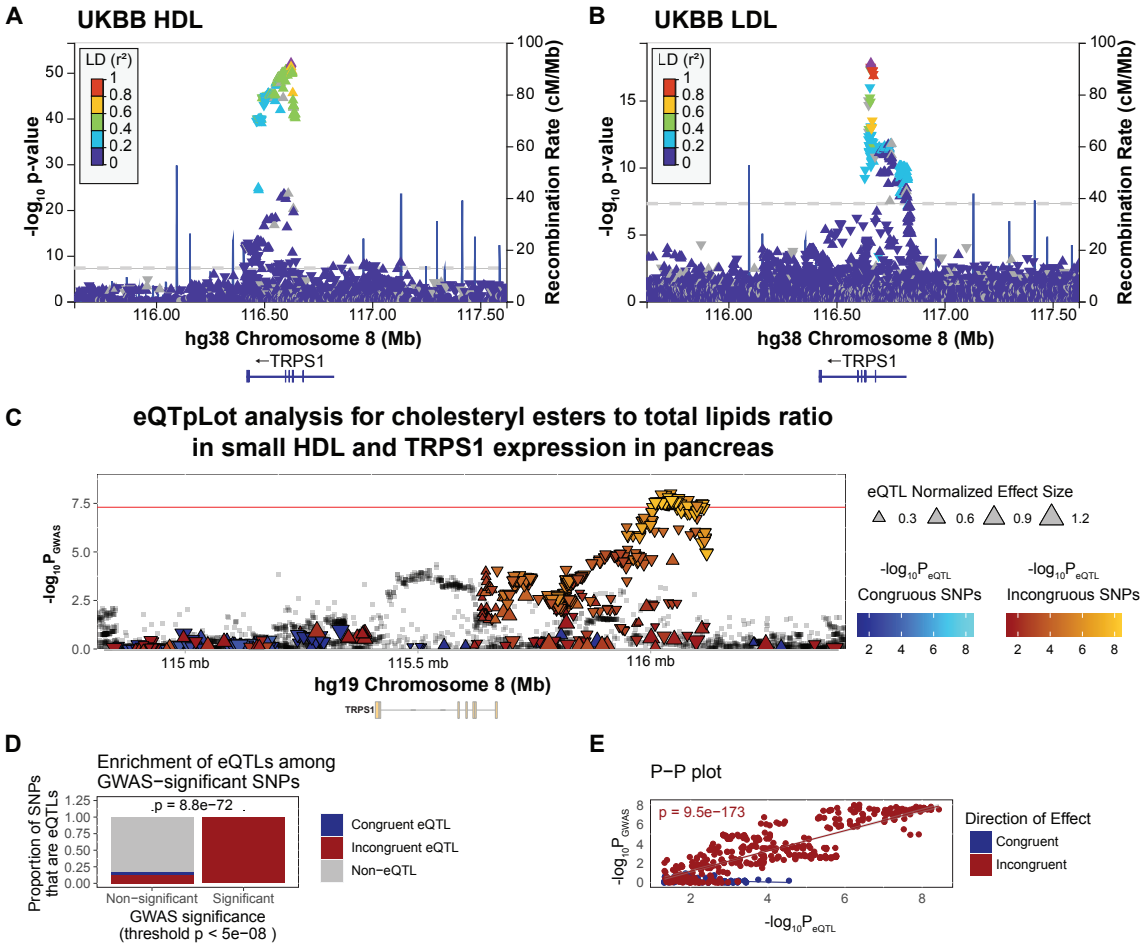

Figure 4.2: *TRPS1* is associated with blood cholesterol traits. A) LocusZoom plot of the *TRPS1* genomic locus depicting the location and significance of SNPs associated with HDL cholesterol levels. B) LocusZoom plot of the *TRPS1* genomic locus depicting the location and significance of SNPs associated with LDL cholesterol levels. C) eQTpLot of the *TRPS1* genomic locus depicting the location and significance of SNPs associated with cholesterol esters to total lipids ratio in small HDL. Triangle size and color represents the normalized effect size and significance of association with TRPS1 expression in pancreas tissue. Congruous refers to the effect size in GWAS and eQTL being in the same direction, and incongruous being in the opposite direction. D) Barchart of the proportions of SNPs that are significant in eQTL, separated by significance in GWAS. E) Scatter plot of p-values for GWAS and p-values for eQTL. Data generated with summary statistics from [266], using LocusZoom and eQTpLot [216, 275].

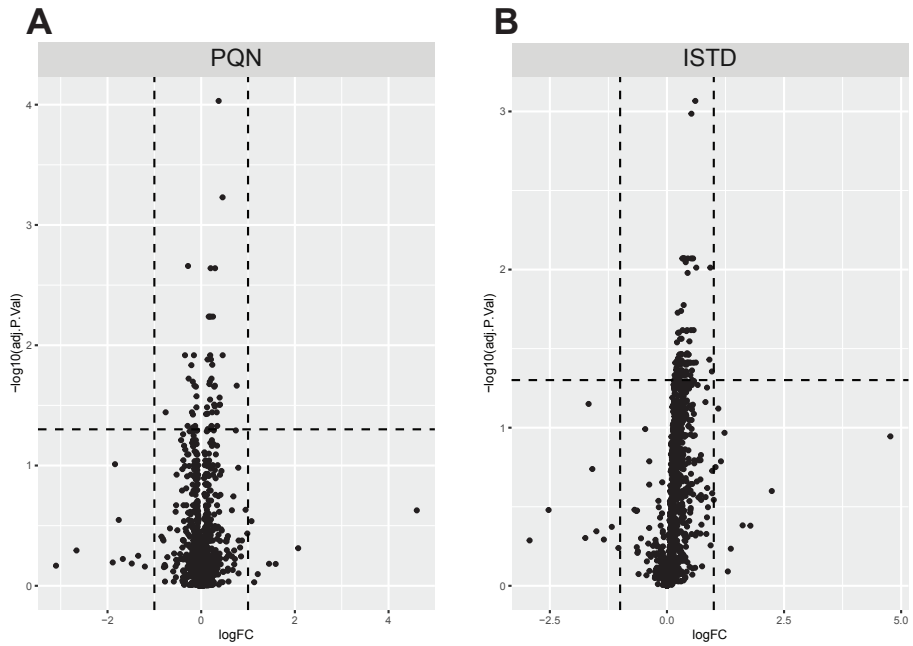

Figure 4.3: Lipidomics normalization method affects results. Volcano plot of lipid species abundance in dTAG-TRPS1 Clone 28 cells after treatment with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for 24 hours, normalized using A) probabilistic quotient normalization (PQN) or B) internal standard normalization (ISTD).

4.5.3 Lipidomics normalization method affects results

Based on these observations, we hypothesized that the abundance of cholesterol and other lipid species in our dTAG-TRPS1 Clone 28 cells would increase upon TRPS1 depletion. To test this hypothesis, we performed an unbiased lipidomics experiment in which cells were treated with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for 24 hours. We provided snap-frozen cell pellets to the Biomolecular Analysis Facility, and they performed ultra-performance liquid chromatography (UPLC) and mass spectrometry (MS) on the extracted lipids from these cell pellets. When these data were normalized using probabilistic quotient normalization (PQN), which assumes there is no lipidome-wide change in lipid species abundance, it appeared that very few lipid species were significantly changing in abundance (Figure 4.3A). However, when we used the spiked-in heavy internal standards, there appeared to be a lipidome-wide increase in lipid species abundance (Figure 4.3B).

4.5.4 TRPS1 depletion does not consistently affect mRNA expression of cholesterol biosynthesis genes

If TRPS1-mediated repression of cholesterol biosynthesis gene transcription decreases cholesterol biosynthesis rate, we would hypothesize that 48 hours of TRPS1 depletion would increase steady state mRNA abundance for these cholesterol biosynthesis genes. To test this hypothesis, we performed reverse transcription quantitative polymerase chain reaction (RT-qPCR) for each of the above genes, as well as *TRPS1* as a positive control and *GAPDH* as a loading control, in each of the three dTAG-TRPS1 clones (Figure 4.4). We grew cells in medium depleted of lipids to increase the endogenous rate of cholesterol biosynthesis. *TRPS1* mRNA expression increased upon TRPS1 depletion, consistent with the PRO-seq data and representing a negative feedback loop in which TRPS1 represses the transcription of its own gene. In contrast with this positive control, mRNA abundance for each of the tested cholesterol biosynthesis genes were not significantly changed upon TRPS1 depletion. These data suggest that the initial increase in nascent transcription of these cholesterol biosynthesis genes does not produce a sustained increase in steady state mRNA abundance of these genes.

4.5.5 TRPS1 depletion does not consistently affect lipid droplet abundance

We hypothesized that an increase in cholesterol biosynthesis would lead to an increase in lipid droplet abundance after 48 hours of TRPS1 depletion. We stained cells with BODIPY to label neutral lipids and DAPI to label nuclei and performed fluorescent microscopy. We used Fiji to quantify lipid droplet area and enumerate nuclei [278]. Across four biological replicates, using the parental T47D cell line as a negative control, TRPS1 depletion did not consistently affect lipid droplet abundance in the dTAG-TRPS1 clone 28 (Figure 4.5).

4.5.6 TRPS1 depletion does not consistently affect cholesterol abundance

We next sought to more directly assay whether TRPS1 depletion affects the abundance of cholesterol in these cells. To do so, we used a luminescent readout in a 96-well plate-based

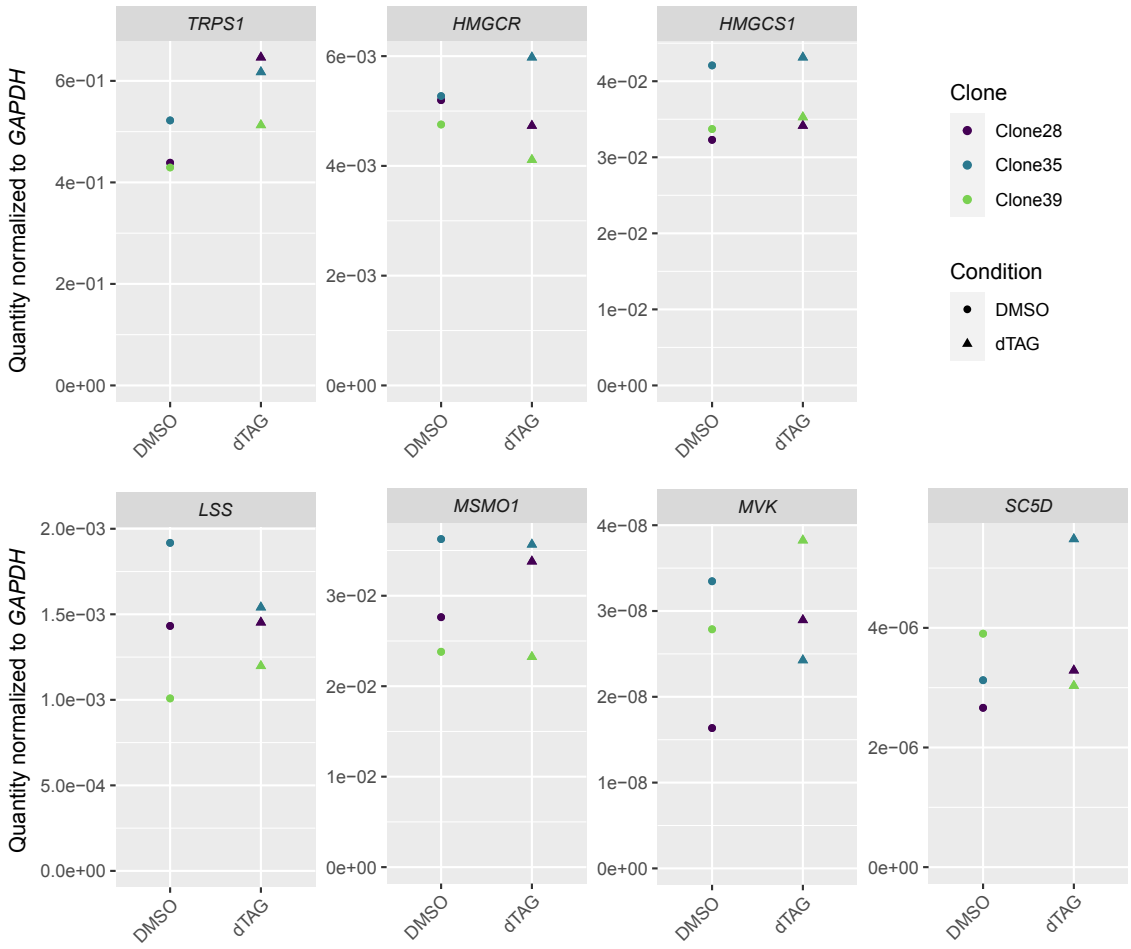

Figure 4.4: TRPS1 depletion does not consistently affect mRNA expression of cholesterol biosynthesis genes. RT-qPCR of cholesterol biosynthesis genes, normalized to *GAPDH* expression. *TRPS1* mRNA is used as a positive control for TRPS1 protein depletion. Cells were treated with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for 48 hours.

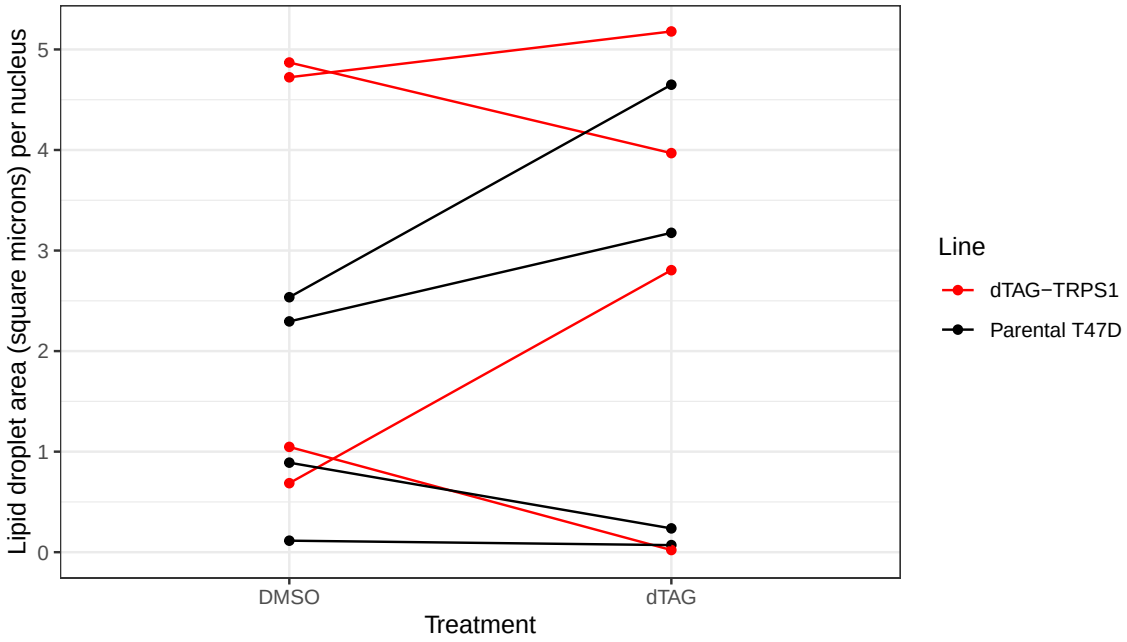

Figure 4.5: TRPS1 depletion does not consistently affect lipid droplet abundance. Quantification of BODIPY staining of the dTAG-TRPS1 clone 28 or the parental T47D cells after treatment with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for 48 hours.

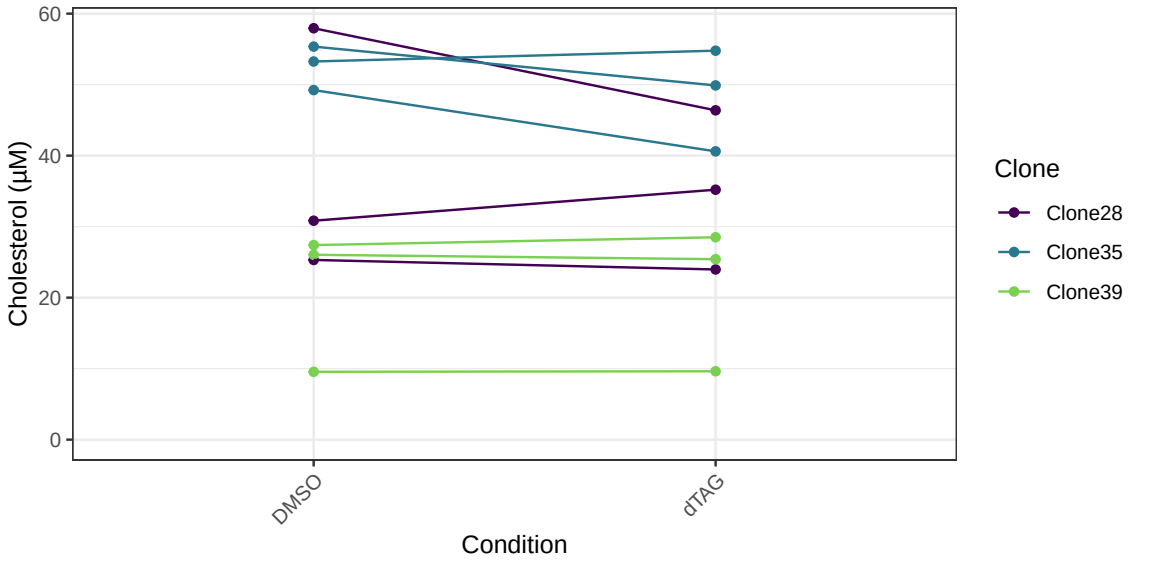

Figure 4.6: TRPS1 depletion does not consistently affect cholesterol abundance. Quantification of cholesterol abundance in each of the three dTAG-TRPS1 clones after treatment with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for 48 hours.

format. We grew cells in medium depleted of lipids to increase the endogenous rate of cholesterol biosynthesis. Upon 48 hours of TRPS1 depletion, we found no consistent change in cholesterol abundance across the three dTAG-TRPS1 clones.

4.5.7 Cholesterol biosynthesis gene transcription is only transiently activated

Based on these negative results, we analyzed the kinetic PRO-seq data we generated after the initial 90 minute TRPS1 depletion experiment from Figure 4.1. We identified genes that were significantly changed over the time course at an FDR of 0.1, using a likelihood ratio test within DESeq2 to identify genes for which including a variable for the time point increased the predictive power of the model over one without this information. Focusing on the Reactome Cholesterol Biosynthesis gene set, we plotted the kinetics of normalized PRO signal for the 16 differentially expressed genes. We found that transcription of each gene was maximal at 2-4 hours after TRPS1 depletion and returned to baseline at our latest time point of 24 hours (Figure 4.7).

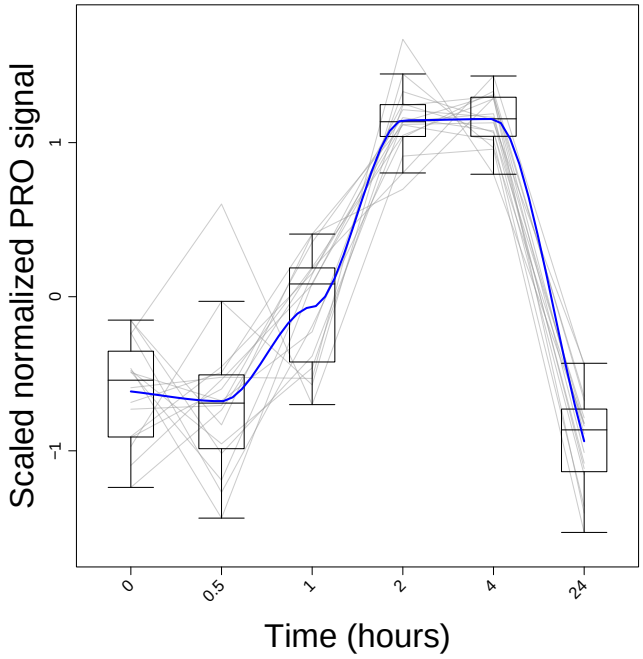

Figure 4.7: Cholesterol biosynthesis gene transcription is only transiently activated. Kinetic traces of dynamic cholesterol biosynthesis gene transcription.

4.5.8 TRPS1 depletion does not significantly affect cholesterol abundance at earlier time points

Based on this finding, we chose to measure cholesterol abundance at earlier time points after TRPS1 depletion. We performed the same cholesterol assay after growing cells in medium depleted of lipids for 48 hours and treating with 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) for various amounts of time. We still found no significant change in cholesterol abundance at any time point, though the measured concentrations were increased after 8 hours of TRPS1 depletion for each of the three clones.

Finally, we turned to control treatments to assess the dynamic range of the assay under our culture conditions. Atorvastatin is a clinically used cholesterol-lowering medication that inhibits HMGCR, the rate-limiting enzyme in the cholesterol biosynthesis pathway, and the enzymatic product of HMGCR is mevalonate [279, 280]. Thus we would expect atorvastatin treatment to decrease cholesterol abundance relative to its DMSO vehicle control and mevalonate

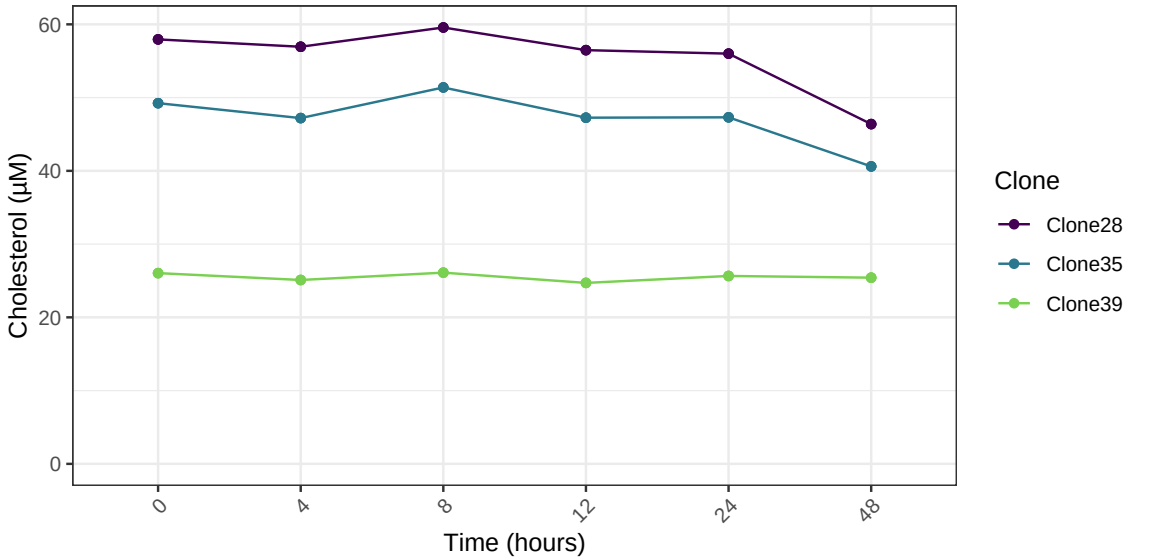

Figure 4.8: TRPS1 depletion does not significantly affect cholesterol abundance at earlier time points. Quantification of cholesterol abundance in each of the three dTAG-TRPS1 clones after treatment with DMSO or 50nM dTAG-13 and 50nM dTAG^{V}-1 in DMSO (dTAG) over a time course of 48 hours.

treatment to increase cholesterol abundance relative to its ethanol vehicle control. However, we only observed a 30% decrease in cholesterol abundance upon atorvastatin treatment and an unexpected decrease in cholesterol abundance upon mevalonate treatment (Figure 4.9). Therefore we do not have high confidence that this assay under our culture conditions is equipped to detect a small increase in cholesterol abundance.

4.6 Discussion

Here we use an inducible degron tag system to measure the early nascent transcriptional changes upon acute depletion of the TF TRPS1 in three independent clones of the luminal breast cancer cell line T47D. We present data demonstrating that 90 minutes of TRPS1 depletion significantly activates transcription of seven cholesterol biosynthesis genes, a proportion of this gene set that is significantly over-represented among all activated genes. In a follow-up kinetic PRO-seq experiment, we found that 16 cholesterol biosynthesis genes were dynamic over the time course,

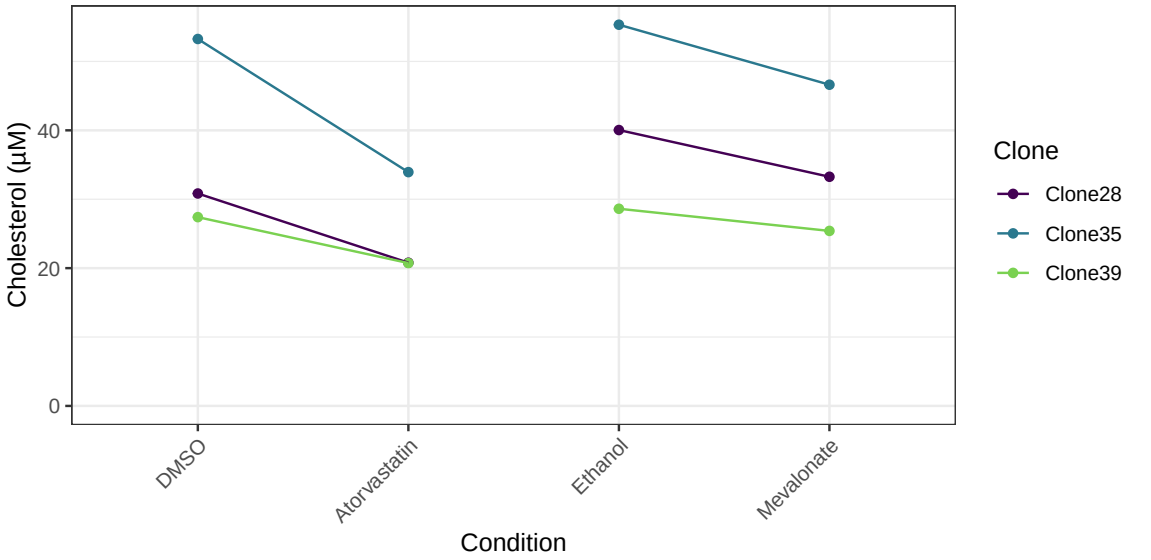

Figure 4.9: Cholesterol assay controls do not demonstrate the expected effects. Quantification of cholesterol abundance in each of the three dTAG-TRPS1 clones after treatment with DMSO, atorvastatin, ethanol, or mevalonate for 48 hours.

with expression changes that tended not to increase at the earliest time point of 30 minutes, to peak at 2-4 hours, and to return to baseline by the latest time point of 24 hours.

We also present an analysis of publicly available GWAS and eQTL data to suggest that TRPS1 regulates blood cholesterol levels in the human population. First, we found that the levels of two common forms of cholesterol, HDL and LDL, are associated with common genetic variants within the *TRPS1* transcriptional unit. Second, we found that the SNPs associated with a more specific cholesterol trait, the ratio of cholesteryl esters to total lipids in small HDL, colocalize with SNPs associated with TRPS1 expression in an eQTL study. These data support our hypothesis that one or more of these common genetic variants regulates these cholesterol traits and that this effect is mediated through an effect on TRPS1 expression.

We do not find evidence of eQTL colocalization with the more proximal SNPs associated with HDL and LDL cholesterol. We speculate that could be attributed to a dearth of significant eQTLs for TRPS1 in the GTEx data we analyzed. The tissue with the most significant eQTLs was pancreas. It is possible that the causal SNP for the HDL or LDL traits affects binding of a

TF that is less expressed in the pancreas tissue studied than in the relevant cell type where the SNP mediates its effect. Increasing samples size of eQTL datasets will eventually give us more power to detect significant associations.

We next performed several follow-up experiments to test our hypothesis that TRPS1 regulates cholesterol biosynthesis in our breast cancer cell lines. We first measured mRNA levels of the genes with activated nascent transcription by RT-qPCR. After 48 hours of TRPS1 depletion, we did not detect a consistent trend in steady state mRNA expression of these genes. We next used fluorescent microscopy to measure lipid droplets, as these can be a depot for neutral lipids like esterified cholesterol. We did not observe a consistent difference in lipid droplet abundance 48 hours after TRPS1 depletion. Finally, we measured cholesterol abundance with a fluorescent, plate-based assay. Once again, we did not find a significant change in cholesterol abundance.

There are several possibilities that could explain these negative results. Of course, steady state cholesterol biosynthesis mRNA levels and cholesterol abundance in these cells might not change in response to TRPS1 depletion. This could be due to feedback mechanisms that strictly maintain cholesterol homeostasis. In fact, in our kinetic PRO-seq experiment, we do see a return to baseline of nascent transcription of these cholesterol biosynthesis genes, which occurs at some point between 4 and 24 hours after TRPS1 depletion. We do not lack confidence in the robustness of the nascent transcriptional data, as we observed the over-representation of cholesterol biosynthesis genes in two separate experiments and three independent clones.

However, even if steady state cholesterol biosynthesis mRNA levels and cholesterol abundance do change, we might not be equipped to detect the small effect sizes. PRO-seq is a sensitive assay, and the fold changes we observed were quite small. Similarly, GWAS data are generated from many thousands of participants to provide statistical power to detect small effect sizes. In our hands, RT-qPCR and lipid droplet staining may be too noisy to detect a weak signal. In our microscopy assay, we found significant field-to-field variability in lipid droplet abundance. If we were to perform this experiment again, capturing many more fields per sample may be helpful.

Finally, the cholesterol assay did not appear so noisy, but the controls indicated a limited dynamic range. Even in lipid-depleted medium, which should increase endogenous cholesterol biosynthesis, treating with atorvastatin, which should significantly inhibit the rate-limiting enzyme in the pathway, only reduced cholesterol abundance by 30% after 48 hours. Furthermore, mevalonate, the product of this enzymatic reaction and a precursor to cholesterol, did not increase cholesterol abundance in our assay. For these reasons, this assay should be further optimized to increase the dynamic range in our cell culture conditions.

In sum, we provide evidence that TRPS1 represses nascent transcription of many cholesterol biosynthesis enzymes and that the *TRPS1* genomic locus is associated with several cholesterol traits in the human population. We would predict that intermediate measures of the cholesterol biosynthesis gene transcripts, proteins, or enzymatic products would be increased upon TRPS1 depletion, though we do not observe this result in the three assays we performed.

4.7 Methods

4.7.1 Cell culture

T47D cells (RRID:CVCL_0553) (ATCC) were cultured in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (Gemini) and 10µg/ml insulin from bovine pancreas (Sigma, made as a 1000x solution in 1% aqueous glacial acetic acid). dTAG-TRPS1 clones were previously generated in (Chapter 3).

4.7.2 Cell treatments for PRO-seq

4 replicates were performed from cells treated and collected at different times in the same day. $4*10^6$ cells per sample were plated in 10cm dishes overnight. For each replicate, cells were treated with DMSO or 100nM dTAG-13 in DMSO for 90 minutes and collected simultaneously.

4.7.3 Cell permeabilization for PRO-seq

Cell permeabilization was performed as previously described [213], with modifications. At the time of harvest, cells were scraped in 10mL ice cold PBS and washed in 5mL buffer W (10mM Tris-HCl pH 7.5, 10mM KCl, 150mM sucrose, 5mM $MgCl_2$, 0.5mM $CaCl_2$, 0.5mM DTT, 0.004U/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)). Cells were permeabilized by incubating with buffer P (10 mM Tris-HCl pH 7.5, KCl 10 mM, 250 mM sucrose , 5 mM $MgCl_2$, 1 mM EGTA, 0.05% Tween-20, 0.1% NP40, 0.5 mM DTT, 0.004 units/mL SUPERaseIN RNase inhibitor (Invitrogen), Complete protease inhibitors (Roche)) for 3 minutes on ice. Cells were washed with 10 mL buffer W before being transferred into 1.5mL tubes using wide bore pipette tips. Finally, cells were resuspended in 50µL buffer F (50mM Tris-HCl pH 8, 5mM $MgCl_2$, 0.1mM EDTA, 50% Glycerol, 0.5 mM DTT). Cells were snap frozen in liquid nitrogen and stored at -80°C.

4.7.4 PRO-seq library preparation

PRO-seq libraries were prepared as previously described [255], with modifications. RNA extraction after the run-on reaction was performed with 500µL Trizol LS (Thermo Fisher) followed by 130µL chloroform (Sigma). The equivalent of 1µL of 50µM for each adapter was used. A random eight base unique molecular identifier (UMI) was included at the 5′-end of the adapter ligated to the 3′-end of the nascent RNA. 37°C incubations were performed with rotation with 1.5mL tubes placed in 50mL conical tubes in a hybridization oven. For the reverse transcription reaction, RP1 was used at 100µM and dNTP mix was used at 10mM each. Libraries were amplified by PCR for a total of 8 cycles in 100µL reactions with Phusion polymerase (New England Biolabs). No PAGE purification was performed to ensure that our libraries were not biased against short nascent RNA insertions.

4.7.5 PRO-seq analysis

Adapters were removed using `cutadapt` [190]. Libraries were deduplicated using `fqdedup` and the 3′ UMIs [193]. UMIs were removed, and reads were reverse complemented with the `seqtk`. Reads aligning to the rDNA genome with `bowtie2` [189] were removed. The remaining reads were aligned, sorted, and convert to *bed* and *bigWig* files with `bowtie2`, `samtools`, `seqOutBias`, and `deeptools`, respectively [194, 195, 245]. Reads were counted in genes using `bedtools`, and differentially expressed genes were identified with `DESeq2` [188, 198]. Over-representation analysis was performed with `clusterProfiler`, using the Reactome gene sets [256, 271, 272].

4.7.6 GWAS and eQTL data visualization and analysis

Summary statistics were downloaded from [266]. SNPs in the *TRPS1* locus were plotted using LocusZoom [216]. eQTL data were downloaded from [273]. Colocalization visualization was performed with [275], and Bayesian analysis was performed with [276].

4.7.7 Lipidomics

5 replicates were performed from cells treated and collected at different times in the same day. 8*10^6 cells per sample were plated in 15cm dishes overnight. For each replicate, cells were treated with DMSO or 50nM dTAG-13 and 50nM dTAG^V-1 in DMSO for 24 hours and collected simultaneously. For each replicate, the plates were moved onto ice, the media was aspirated, and cells were washed with 10ml PBS, scraped, and moved to a 50ml conical tube. The cells were centrifuged at 500g for 5 minutes, resuspended in 1ml PBS, and moved to 1.5ml Protein LoBind Tube (Eppendorf). The cells were centrifuged at 2500g for 1 minute, and the pellets were snap froze in liquid nitrogen, and stored at -80°C.

The following steps were performed by staff at the Biomolecular Analysis Facility, who provided this protocol:

To each tube, 750µL of -20°C cold chloroform:methanol (2:1) mixture was added and vortexed. Cells were broken in a bead beater with steel balls for 3 minutes at an intensity of 5.

Tubes were shaken vigorously for 30 minutes at 4°C in a temperature controlled thermal shaker. Further 400µL of water was added, shaken vigorously, and the top aqueous methanolic phase was recovered as metabolite mixture and transferred in Eppendorf tubes. The lower phase was saved for lipid extraction.

To each tube, 500µL of chloroform phase, 500µL of cold Chloroform:methanol (2:1) mixture was added, vortexed and shaken vigorously for 30 minutes at 4°C in temperature controlled thermal shaker. Further 200µL of water was added, shaken vigorously and lower organic phase was recovered as lipid mixture. A second extraction was performed by adding 500µL of chloroform and 200µL water. The lower organic phase was recovered as lipid extract, stored in glass bottles at -80°C.

10µL of Avanti Splash Lipidomix was added to each sample as internal standard, and then samples were dried under gentle stream of N_2 using a Recti-Vap Evaporator (Thermo Fisher Scientific) at 40°C. The dried lipid extract was reconstituted in 110µL of methanol:isoproponal (1:1). Approx. 100µL was recovered. Samples were transferred to borosilicate glass inserts kept inside a screw-capped glass autosampler vials (Agilent).

MS data was acquired on Thermo Orbitrap IDX MS connected to Vanquish UPLC system. Lipid extract was separated using Ascentis Express C18 (Sigma-Aldrich®, 2.1 x 100 mm, 2.7 µm) operated at 55°C and a flow rate of 260 µL/min Mobile phase A was 60:40 acetonitrile/water and mobile phase B was 90:10 isopropyl alcohol/acetonitrile; both A and B contained 10 mM ammonium formate and 0.1% formic acid.

Mass scan range: 250-2000 at a resolution of 120,000 with a scan range of 1.5 sec. Data dependent MS2 scans were obtained with an Orbitrap resolution of 15,000 and stepped collision HCD energy of 25,30,35 was used. AcquireX workflow was employed for lipid characterization by additional targeted product ion (m/z 184.0733) or neutral loss (fatty acid + NH4) fragmentation CID MS2 and MS3 experiment were performed to provide characterization of PC and TG lipids.

Data analysis was performed on MS-DIAL v4.8. Search type was set to Product search

(for identification of spectra obtained in MS2 measurement). Precursor tolerance was set at 5.0 ppm and product tolerance was set at 5 ppm with a relative intensity cutoff of 1% was used to remove unwanted noise. MS1 tolerance was set to 0.01 Da and MS2 set to 0.025 Da. For peak picking mass slice width was set 0.1. For peak alignment maximum retention time tolerance was set at 0.2 min, MS1 tolerance set to 0.015. Peaks were identified by searching the MS2 spectra in MS-DIAL LipidBlast database (http://prime.psc.riken.jp/compms/msdial/download/lipidblast/LipidMsmsBinaryDB-VS68-FiehnO.lbm2) using a mass tolerance of 0.01 Da for MS1 and 0.05 Da for MSMS with identification score cutoff of 60%. After the search and alignment, peak data was exported out as .txt file. Normalization was performed, and volcano plots were generated using the lipidr package in R [242, 281].

4.7.8 RT-qPCR

$4*10^5$ cells per well for each of the three clones were plated in 6-well plates overnight. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO for 48 hours. RNA was isolated with TRIzol (Thermo Fisher Scientific). RNA concentrations were determined using a NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Fisher Scientific). cDNA was synthesized using Sensifast cDNA synthesis kit (Bioline). qPCR was performed using iTaq Universal SYBR Green Supermix (Biorad). Primer sequences:
GAPDH-F: ACAGTTGCCATGTAGACCCC; *GAPDH*-R: TGGTTGAGCACAGGGTACTT;
TRPS1-F: TCCCTGTTACGGAGGCGTAG; *TRPS1*-R: CGCGTTGCATACATATCCGC;
HMGCR-F: CCGCGACTGCGTTAACTGG; *HMGCR*-R: ACAGAATCCTTGGATCCTCCAGA;
HMGCS1-F: TTGTGCCCGAAGGAGGAAAC; *HMGCS1*-R: GCATGGTGAAAGAGCTGTGTG;
LSS-F: GCGTTATTTGCAGAGTGCCC; *LSS*-R: CCCCAGCAATGTTTTCCTGC;
MSMO1-F: AGTTCATCATGAGTTTCAGGCTCC; *MSMO1*-R: ATGGTCACCCATGCCCAAAG;
MVK-F: CTCTGGGTTGTGGGAGTTGG; *MVK*-R: TACAGCCAGTGCTACCTTGC;
SC5D-F: CTCGCAGCACGGCTTTTCTC; *SC5D*-R: GATCCATCACTTAGCCCCTGC.
The relative standard curve method was used to determine transcriptional fold changes [282, 283]. RNA starting quantities were determined using a standard curve and normalized to the

reference gene *GAPDH*.

4.7.9 Lipid droplet staining

4 replicates were performed from cells of different passages on different days. $4*10^5$ cells per well for each of the parental T47D cells and the dTAG-TRPS1 Clone 28 cells were plated on uncoated cover slips in 6-well plates overnight. Cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO for 48 hours. The wells were washed twice with PBS, fixed with a 4% formaldehyde (Sigma) solution in PBS for 10 minutes, washed twice with PBS, stained with a 1:1000 solution of (1mM BODIPY dye (Invitrogen) in DMSO) in PBS for 30 minutes protected from light, washed twice with PBS, and mounted onto slides with DAPI-containing mounting medium (Vector Labs). The samples were imaged using a confocal fluorescent microscope (Zeiss). Images were analyzed with Fiji [278]. Briefly, images were thresholded and converted to binary. A watershed was applied to separate particles, and particles were analyzed for number, average size, and total area. Total lipid droplet area divided by nuclei number were plotted in R [242].

4.7.10 Cholesterol assay

Replicates were performed from cells of different passages on different days. $4*10^5$ cells per well for each of the dTAG-TRPS1 clones were plated in 6-well plates overnight. For Figure 4.6, cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO for 48 hours. For Figure 4.9, one replicate from Figure 4.6 also had control wells with 10μM atorvastatin (Sigma), ethanol, or 100μM mevalonolactone (Sigma). For Figure 4.8, cells were treated with DMSO or 50nM dTAG-13 and 50nM $dTAG^V$-1 in DMSO at the indicated time points and collected simultaneously. Cholesterol was measured using the Cholesterol-Glo assay (Promega). Cells were washed twice with PBS, lysed for 30 minutes at 37°C, and incubated with detection reagent for 1 hour protected from light. Fluorescence was measured using a plate reader (Biotek). Fluorescence was converted to cholesterol concentration using a standard curve of cholesterol in the same plate and plotted in R [242].

4.8 Data Access

All analysis details and code are available at https://tgscott400.github.io/TRPS1_cholesterol_analysis/Vignette.html. Raw PRO-seq reads and processed counts and *bigWig* files are available from GEO accession record GSE251772.

4.9 Competing Interest Statement

The authors declare no competing interests.

4.10 Acknowledgements

This work was funded by R35-GM128635 to MJG. 5T32GM007267-39 and 5T32GM008136-35 supported TGS. We thank the Biomolecular Analysis Facility for performing the UPLC-MS/MS. We thank Devin Roller for critical feedback.

Chapter 5

Contributions to other projects

5.1 ARF-AID: a rapidly inducible protein degradation system that preserves basal endogenous protein levels

Kizhakke Mattada Sathyan, Thomas G. Scott, Michael J. Guertin.

Current Protocols in Molecular Biology. 2020.

5.1.1 Abstract

Inducible degron systems are widely used to specifically and rapidly deplete proteins of interest in cell lines and organisms. An advantage of inducible degradation is that the biological system under study remains intact and functional until perturbation, a feature that necessitates that the endogenous levels of the protein are maintained. However, endogenous tagging of genes with auxin-inducible degrons (AID) can result in chronic, auxin-independent proteasome-mediated degradation. The ARF-AID (auxin-response factor–auxin-inducible degron) system is a re-engineered auxin-inducible protein degradation system. The additional expression of the ARF-PB1 domain prevents chronic, auxin-independent degradation of AID-tagged proteins while preserving rapid auxin-induced degradation of tagged proteins. Here, we describe the protocol for engineering human cell lines to implement the ARF-AID system for specific and inducible protein degradation. These methods are adaptable and can be extended from cell lines to organisms.

5.1.2 Contribution

I wrote portions of the original draft of the manuscript on CRISPR guide RNA design, genomic DNA primer design, and homology-directed repair template design and edited the manuscript.

5.2 The androgen receptor does not directly regulate the transcription of DNA damage response genes

Sylwia Hasterok*, Thomas G. Scott*, Devin G. Roller, Adam Spencer, Arun B. Dutta, Kizhakke M. Sathyan, Daniel E. Frigo, Michael J. Guertin, Daniel Gioeli

*S. Hasterok and T.G. Scott contributed equally to this article.

Molecular Cancer Research. 2023.

5.2.1 Abstract

The clinical success of combined androgen deprivation therapy (ADT) and radiotherapy (RT) in prostate cancer created interest in understanding the mechanistic links between androgen receptor (AR) signaling and the DNA damage response (DDR). Convergent data have led to a model where AR both regulates, and is regulated by, the DDR. Integral to this model is that the AR regulates the transcription of DDR genes both at a steady state and in response to ionizing radiation (IR). In this study, we sought to determine which immediate transcriptional changes are induced by IR in an AR-dependent manner. Using PRO-seq to quantify changes in nascent RNA transcription in response to IR, the AR antagonist enzalutamide, or the combination of the two, we find that enzalutamide treatment significantly decreased expression of canonical AR target genes but had no effect on DDR gene sets in prostate cancer cells. Surprisingly, we also found that the AR is not a primary regulator of DDR genes either in response to IR or at a steady state in asynchronously growing prostate cancer cells.

5.2.2 Contribution

I analyzed the PRO-seq data, generated figure panels, and reviewed and edited the manuscript.

Chapter 6

Discussion

6.1 Concordance between initial and steady-state ER activity

6.1.1 Conclusions

In Chapter 2, contrary to our hypothesis, we found that acute antagonism of ER from an active state recapitulated the effect that acute agonism of the estrogen receptor (ER) from an inactive state had on nascent transcription, just in the opposite direction and with a smaller magnitude. We had hypothesized that, though there may be some overlap in the gene sets, ER would regulate distinct sets of genes when initially activated than when at a steady state of activation. We predicted that the differential expression of transcription factors (TFs) and cofactors as a part of the primary response to estrogen would feed back to modulate ER binding and ability to recruit coactivators to specific loci. We based this hypothesis in part on the results presented in Figure 3B in [37].

Indeed, much of the nascent transcriptional response to estrogen is transient. However, our results suggest that the delayed repression of the estrogen-activated genes that return to baseline expression over time is not primarily due to a reduction in ER activation of specific genes but rather due to a separate repression that overpowers the ER activation. There does appear to be a general reduction in ER activation of its target genes, as the effect sizes for ER antagonism are smaller than those for ER agonism. This aspect of our results is expected, as estrogen treatment downregulates *ESR1* mRNA expression [284–286].

In addition to this negative feedback, we also predicted that ER would gain the ability to directly activate secondary effect genes that it was not able to activate under the hormone-starved condition. However, this does not appear to be true, suggesting that ER binding and direct activation of target genes is unaffected by the steady-state condition in which it is acting, at least for the two we tested, complete medium and hormone-starved medium. This robustness is

remarkable in the face of many differentially expressed TFs and cofactors and a difference in cell state from non-proliferating to proliferating. This is consistent with the kinetics of ER binding over the first 90 minutes of estrogen treatment, as determined using a quantitative measure with high temporal resolution, which showed relatively constant binding intensity genome-wide [287]. We would predict that this trend would continue at longer time points, with a genome-wide decrease in binding over time as ER expression is downregulated, but without a redistribution of ER binding.

To illustrate the dynamics of estrogen-activated genes, we performed hierarchical clustering as in Figure 3.13A and found that most ER target genes conform to one of three major patterns (Figure 6.1). Over 80% of the estrogen-activated genes are more lowly expressed in complete medium than upon acute ER activation in hormone-starved medium. This is consistent with the previously reported transient activation of most ER target genes [37]. Within this class of genes, Group 5 in our data are more severely repressed to around the baseline levels observed in the hormone-starved condition, and Group 1 are more moderately repressed to intermediate levels. The other class of genes are actually further activated in complete media, though this cannot be attributable to further ER activation of these genes, as ER antagonist treatment does not repress their expression close to the baseline levels observed in the hormone-starved condition.

Based on our above interpretation of the data, we hypothesize that one or more TFs change in activity as a result of a primary transcriptional effect of estrogen treatment and that this change in activity represses the transiently activated ER target genes and further activates the other ER target genes. Furthermore, we would predict that this modulation must be via distinct regulatory elements (REs) from the ER binding sites that mediate the initial estrogen activation. In this way, these genes are still ER target genes that are repressed by ER antagonist treatment.

Figure 6.1: ER target genes follow one of three expression patterns across media conditions and ER activity modulation. Kinetic traces of estrogen-activated genes across the media and drug treatment conditions.

6.1.2 Future directions

Which TFs mediate the modulation of ER target gene expression?

To identify these differentially active TFs, we performed assay for transposase-accessible chromatin with sequencing (ATAC-seq) under the same conditions in which we performed precision run-on sequencing (PRO-seq). We performed an initial analysis of these data before switching focus to our newly-generated dTAG-TRPS1 clones and formed hypotheses that could be further explored. Specifically, we identified differentially accessible ATAC-seq peaks between the complete medium and hormone-starved medium conditions and performed *de novo* motif identification within these peaks. Within the peaks with decreased accessibility in complete media, we identified the TEAD family motif. Within the peaks with increased accessibility in complete media, we identified the ER motif, the SP/KLF family motif, the forkhead-box family motif, and the RUNX family motif.

Among these motifs, the ER motif serves as a positive control, as ER is active in complete medium and inactive in hormone-starved media. Forkhead-box motifs are often found near ER binding sites, and the forkhead-box family member FOXA1 increases chromatin accessibility

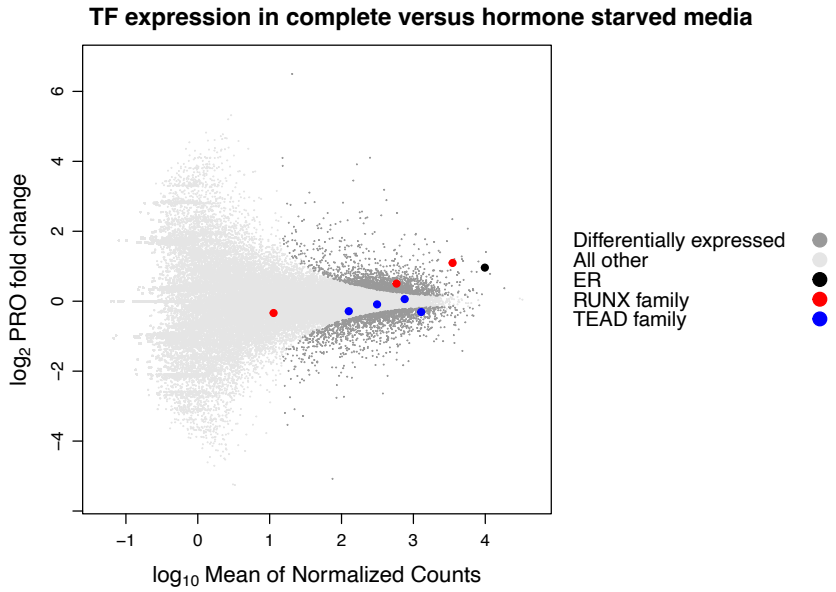

Figure 6.2: TEAD and RUNX family TFs change in expression between media conditions. MA plot of genes, with fold change values representing nascent transcription in the complete medium condition relative to the hormone-starved medium condition.

and ER binding intensity, at least locally [288–290]. However, this motif does not significantly differ in abundance between increased and decreased peaks in our ATAC-seq data, nor is *FOXA1* differentially expressed between complete and hormone-starved media. When performing this analysis, we included unchanged peaks as well, but these are not matched to the dynamic peaks based on accessibility and are in general less accessible regions. In future, we would use the R package MatchIt to pair a control unchanged peak with each activated peak [291]. In this way we can more carefully identify potential examples of TF redistribution, as we did with ER in Chapter 3.

In contrast, the TEAD family motif and the RUNX family motif are specifically enriched in decreased and increased ATAC-seq peaks, respectively. Furthermore, the expression of individual family members change in the expected directions (Figure 6.2). Specifically, *TEAD1* is significantly repressed, and *RUNX1* is most significantly activated, with *RUNX2* less so.

Does TEAD1 activate ER target genes in the absence of estrogen?

The TEA/ATTS domain (TEAD) family of TFs, TEAD1-4, are most well known as the DNA-binding factors of the Hippo signaling pathway which recruit the coactivators Yes-associated protein (YAP1) and Transcriptional Activator with PDZ binding domain (TAZ) to activate their target genes, though more recent work has revealed Hippo pathway-independent regulation of TEAD activity [292, 293]. Based on our data, we hypothesize that TEAD1 positively regulates transcription of a portion of the transiently-activated ER target genes in the hormone-starved condition. In complete medium, *TEAD1* is repressed, leading to an attenuation in expression of these genes, dampening the ER-mediated activation.

This hypothesis generates several predictions that could be tested in future experiments. First, we would confirm that TEAD1 protein abundance is decreased upon hormone starvation. If so, we would determine whether TEAD1 activates a portion of ER target genes, in particular those that are transiently activated by estrogen. As covered in Chapter 1, this could be done with genetic, chemical, or chemical genetic perturbation. As the first two are available and more easily applied to multiple cell lines, we would start with those. RNA interference (RNAi) could be used to knock down *TEAD1* expression in hormone-starved medium, and ER target gene expression could be measured with messenger RNA (mRNA) sequencing (RNA-seq). A repression of ER target gene expression would indicate that TEAD1 positively regulates these genes, but the delayed time point would not discriminate between primary and secondary effects. To address this point, a more acute perturbation could be achieved with chemical inhibitors of TEAD/YAP1 interaction, those these inhibitors would miss any YAP1-independent effects of TEAD1 [294–297].

If our hypothesis is correct, then it would be interesting to study the therapeutic implication of TEAD1 activity modulation in luminal breast cancer. From data in the Cancer Dependency Map project, TEAD1 knockout does not tend to affect luminal breast cancer cell number [77]. However, these experiments were done in complete medium, in which we predict TEAD1 expression is diminished. In a predicted higher TEAD1 activity setting such

as hormone-starved medium or ER antagonist treatment, we would hypothesize that TEAD1 inhibition, predicted to further repress ER target genes, would further decrease proliferation.

Does over-activation of ER target genes upon TRPS1 depletion contribute to the cell number defect?

An alternative, not mutually exclusive, hypothesis would be that the transient nature of the estrogen activation of this set of genes benefits cell fitness and that the activation of TEAD1 in complete medium would lead to over-activation of these ER target genes and a decrease in proliferation. Intriguingly, in Chapter 3, we found an enrichment of TEAD motifs in ATAC-seq peaks that increased in accessibility upon TRPS1 depletion. We hypothesize that TRPS1 depletion increases TEAD1 binding to DNA, compensating for the decrease in *TEAD1* expression, and contributing to the activation of a subset of ER target genes.

Consistent with this hypothesis, our definition of ER target genes based on the PRO-seq data generated in Chapter 2, of which approximately equal numbers were activated or repressed upon TRPS1 depletion, differed from the Hallmark estrogen response gene sets, which were specifically enriched among the genes that were repressed upon TRPS1 depletion. We would predict that the ER target genes activated by TRPS1 depletion would be enriched for the genes that are only transiently activated by estrogen over those that are sustained. In future, we would analyze the data from Chapter 2 as well as the time course data from [37] to test this prediction.

Unfortunately, we were not able to find an antibody that could immunoprecipitate TEAD1 in our hands. We also were unable to generate many reads of high quality chromatin immunoprecipitation with sequencing (ChIP-seq) data for YAP1 to fully address this question. However, our preliminary data suggest that there is a genome-wide increase in YAP1 binding intensity upon TRPS1 depletion (Figure 6.3). These data were normalized to read depth, but additional read depth and the use of the quantitative parallel factor ChIP-seq may be helpful to more convincingly demonstrate a genome-wide increase [298]. If this hypothesis is correct, we would predict that *TEAD1* knockdown or inhibition would rescue the decrease in cell number

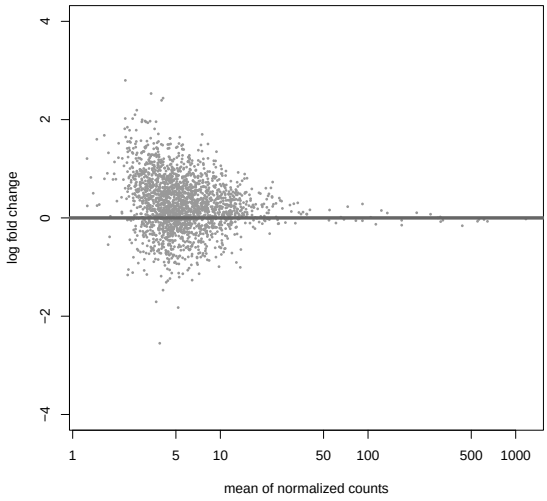

Figure 6.3: YAP1 binding intensity is increased genome-wide upon acute TRPS1 depletion. MA plot of YAP1 binding intensity, with fold change values representing read depth-normalized ChIP-seq reads upon TRPS1 depletion relative to the control condition.

doubling rate upon TRPS1 depletion. Furthermore, if YAP1 activity is not maximal upon TRPS1 depletion, further augmentation of YAP1 activity via inhibition of its upstream inhibitory kinases, mammalian sterile 20-like kinases 1 and 2 (MST1/2), should further decrease cell number doubling rate [299].

Two recent publications connect YAP1 and TEAD with ER and TRPS1. The first identified TRPS1 as a repressor of YAP1-activated transcription in an unbiased screen in the luminal breast cancer cell line MCF-7 [57]. The authors do not test whether TRPS1 depletion increases YAP1 chromatin binding but do find that it increases expression of YAP1-activated genes. Though we do not identify a YAP1 signature among the genes activated upon TRPS1 depletion in our own data, these published results are consistent with our hypothesis that TRPS1 decreases TEAD1 binding intensity genome-wide.

The second publication found that knockdown of a different TEAD family member, *TEAD4*, in hormone-depleted medium, reduced ER occupancy on chromatin and ER target gene expression upon acute estrogen treatment in the same MCF-7 cell line [101]. The authors do

not test whether *TEAD4* knockdown represses these genes in the absence of estrogen or further reduces cell proliferation. Separately, the authors found that YAP1 activation in complete medium reduces cell number doubling rates of both MCF-7 and T47D cells. This is consistent with our hypothesis that increased TEAD1 activity in complete medium over-activates ER target genes that are normally only transiently activated in response to estrogen and that this over-activation reduces cell fitness. However, YAP1 activation also repressed ER protein expression, which is sufficient to reduce cell fitness without our additional mechanism. YAP1-mediated *ESR1* transcriptional repression has been previously reported as a secondary effect of the YAP1-activated corepressor VGLL3 [300, 301]. It would be interesting to test whether increased TEAD1 binding intensity at specific sites, not including REs controlling *ESR1*, induced upon TRPS1 depletion also contributes to the effect on cell number.

Do RUNX family members augment activation of ER target genes?

In our analysis of the ATAC-seq data generated in complete and hormone-starved coniditions, the RUNX family motif was specifically enriched in peaks that increased in intensity in complete medium, and RUNX1 was the most significantly activated RUNX family member in complete medium. The runt-related (RUNX) family of TFs, RUNX1-3, heterodimerize with core-binding factor subunit beta (CBFB) to activate transcription of target genes [302]. They are most well-characterized as tumor suppressors in the hematopoetic niche, but RUNX1 and CBFB are also recurrently mutated in luminal breast cancer [11]. We hypothesize that RUNX1 further activates a subset of ER target genes as a secondary effect of estrogen. As with TEAD1 above, we can test this with genetic and chemical perturbations, specifically with *RUNX1* RNAi-mediated knockdown or with compounds that reduce RUNX binding to DNA [303–305].

We might also predict that RUNX inhibition would decrease cell fitness by repressing a subset of ER target genes that are normally continuously expressed downstream from ER activation. However, we do not observe an effect of knockout of any individual RUNX family member or even of CBFB in the Cancer Dependency Map project data [77]. *RUNX1* knockout

in the mouse mammary epithelium leads to a decrease in ER-positive mature luminal cells, but *RUNX1* knockdown increases proliferation of the luminal breast cancer cell line MCF-7, as well as three-dimensional culture of the basal-like mammary epithelial cell line MCF10A [306–308].

Do KLF family members repress ER target genes in the absence of hormone?

The final motif enriched in ATAC-seq peaks that increased in intensity in complete medium was the SP/KLF family motif. Sp1-like (SP) proteins and Krüppel-like factors (KLFs) are a large and diverse set of TFs that bind to similar DNA sequences but have varied activator or repressor domains [309]. Of these family members, KLF3, KLF7, KLF8, and KLF12 are significantly repressed in complete medium. Of these four, KLF3, KLF8, and KLF12 recruit corepressors in the c-terminal binding protein (CtBP) family, and KLF7 has a putative acidic activator domain that has been inactivated over the course of evolution [310–312]. Thus we hypothesize that these TFs repress ER target genes in hormone-starved medium and that this repression is alleviated as a secondary effect of estrogen treatment via the transcriptional repression of these KLF genes. This hypothesis is more difficult to test than those for the TEAD and RUNX family TFs for two reasons. Four TFs in the family are transcriptionally affected, making a genetic perturbation more challenging, and we are unaware of any chemical activators of KLF proteins.

Does ER consistently regulate *ESR1* expression across luminal breast cancer cell lines?

The last observation we will mention from this set of experiments is the transcriptional activation of the *ESR1* gene upon acute estrogen treatment and in complete medium, compared with the hormone-starved condition. This is contrary to our expectation from MCF-7 cells that *ESR1* mRNA and ER protein abundance are decreased upon estrogen treatment [284–286]. In future we would test the effect of estrogen treatment on ER protein expression across a panel of luminal breast cancer cell lines. If the effect is not consistent, this could have therapeutic implications. For example, in patients with luminal breast cancer treated with hormonal therapy to inhibit ER activity, ER protein downregulation or upregulation could lead to alternative mechanisms

of therapy resistance. Specifically, ER downregulation could facilitate the transition to an ER-negative, ER-independent, and estrogen-independent state, and upregulation could allow for the accumulation of mutations in growth factors or *ESR1* itself, rendering ER estrogen-independent [44, 313, 314].

6.2 Concordance between chromatin accessibility, ER binding intensity, and ER target gene expression

6.2.1 Conclusions

In Chapter 3, we claim that TRPS1 modulates chromatin accessibility to regulate ER binding and ER target gene expression. This is a simple explanation of our data and the one we are inclined to believe. Local chromatin accessibility, in addition to the presence and strength of the ER motif in the DNA, is known to influence ER binding, but ER binding itself also increases local chromatin accessibility [315, 316]. We have not ruled out the possibility that TRPS1 primarily decreases ER binding intensity, with changes in chromatin accessibility at these sites merely a consequence of the loss of ER-recruited coactivators. This possibility could in theory be due to direct competition for DNA binding, though TRPS1 and ER recognize distinct motifs, and we did not observe a conserved close spacing between TRPS1 and ER ChIP-seq peak summits or a longer hybrid motif. Alternatively, the corepressors TRPS1 recruits to chromatin may modify non-histone substrates like ER itself, acetylation of which has been shown to increase the DNA binding of ER [317]. In addition, we have correlated much of the change in ER target gene expression to the change in local ER binding intensity. However, we have not ruled out the possibility that TRPS1 regulates these genes independently from its effects on ER binding. To further test the model we put forth, we propose the following experiments.

6.2.2 Future directions

Is decreased chromatin accessibility necessary for the TRPS1-dependent decrease in ER binding intensity?

It is difficult to experimentally disentangle changes in chromatin accessibility from changes in ER binding intensity. It would require creating a separation-of-function mutant that decreased ER binding intensity via recruitment of certain corepressors yet failed to decrease chromatin accessibility via recruitment of different corepressors. However, the resolution of the protein-interacting interfaces on TRPS1 is quite low, so we would propose generating a panel of small deletions in the carboxy-terminal region of TRPS1. These mutants could be used in co-immunoprecipitation experiments for various corepressive complexes. If slightly different residues are responsible for recruiting different complexes, then we could test whether different corepressors are responsible for the effects of TRPS1 on chromatin accessibility and ER binding intensity.

Are the changes in ER target gene expression due to modulation of ER activity?

To more directly address this question, we can perturb ER function to measure ER activation of these TRPS1-dependent ER target genes. As in Chapter 2, we can either acutely antagonize ER in complete medium or acutely stimulate ER in hormone-starved medium. Complete medium would better match the context of our previous observations, as the rest of the experiments in Chapter 3 were done in this condition. The potential drawback based on our results in Chapter 2 would be a smaller effect size than we might see in the hormone-starved condition. In either case, we can perform PRO-seq in four conditions — with and without TRPS1 activity, and with and without ER activity. Then we can focus on the TRPS1-dependent ER target genes and determine whether the effect size of ER activity on each gene is dependent on TRPS1 activity. Specifically, we predict that the TRPS1-repressed ER target genes will be less responsive to ER activity modulation upon TRPS1 depletion, and that the TRPS1-activated ER target genes will be more responsive to ER activity modulation upon TRPS1 depletion.

6.3 TF redistribution as a general model

6.3.1 Conclusions

In the model of coactivator "squelching", a TF that directly activates target genes can indirectly repress distal genes by recruiting coactivators that are limiting in the cell and competing with other activating TFs that are less able to directly activate their target genes [223, 235–239, 318]. Our ER redistribution model is conceptually "squelching" of a TF, as opposed to coactivators. One unknown factor is whether the protein expression and nuclear distribution of ER influences the propensity for an increase in ER binding intensity at many loci, for example at TRPS1-proximal sites, to decrease ER binding intensity elsewhere, for example at TRPS1-distal sites.

In an alternative model of ER-mediated repression, this activating TF can also directly repress its target genes, recruiting corepressors instead of coactivators at a subset of ER target genes [285, 319–321]. The context specificity, determining which cofactors are recruited by the same TF at different loci, required for such a model has generally not been elucidated. These studies of direct ER-mediated repression have been done at individual loci, and it is difficult to prove or disprove that a specific ER binding site proximal to an estrogen-repressed gene causes direct ER-mediated repression of that gene. Our strategy is to turn to genome-wide assays to identify general mechanisms of transcriptional regulation. Of course there is the possibility that there are exceptions at individual genes. In our own data, we find that estrogen-activated genes tend to be closer to ER binding sites but that estrogen-repressed genes are as a class no closer to ER binding sites than are estrogen-unresponsive genes, matched for expression levels (Figure 6.4). A simple explanation is that ER directly activates its target genes proximal to ER binding sites and does not directly regulate the estrogen-repressed genes. This is similar to our analysis of TRPS1 proximity to TRPS1-dependent genes (Figure 3.11E,F).

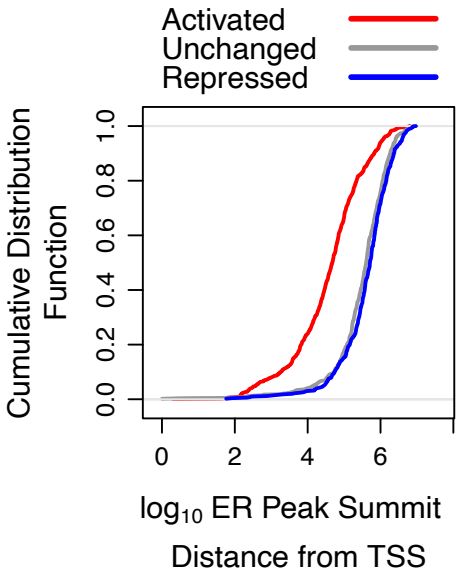

Figure 6.4: Estrogen-activated genes are closer to ER binding sites. Cumulative distribution function plot of proximity to ER ChIP-seq peak summits for genes grouped into classes based on their response to estrogen.

6.3.2 Future directions

Does liganded ER protein abundance influence a redistribution model?

In our TRPS1-mediated ER redistribution model, liganded ER protein is limiting and not acutely replaced on chromatin. We might be able to test this assumption by titrating estrogen in hormone-depleted medium to acutely stimulate an increasing fraction of ER molecules. In contrast to other nuclear receptors, ER is generally already localized to the nucleus in the absence of ligand, and it is difficult to determine based purely on fractionation studies what proportion of ER molecules are bound to DNA [286]. However, a ChIP assay may provide a measure of ER binding enrichment upon stimulation with varying concentrations of estrogen. With a set of concentrations that elicit a range of ER binding intensities, we can acutely deplete TRPS1. We would expect TRPS1-proximal ER binding intensity to increase as before, and the TRPS1-distal ER binding intensity to decrease most drastically in low-estrogen conditions. With increasing liganded ER protein abundance, we would expect the potential for a pool of liganded ER protein not bound to DNA that could fuel a genome-wide increase in ER binding. However, this may

not be the case if most additional liganded ER protein binds to DNA, perhaps at lower affinity ER motifs.

Can we uncouple estrogen-induced transcriptional activation from repression?

To further test a model in which acute ER stimulation leads to the immediate repression of primary response genes via a squelching mechanism, we could change the location of ER binding genome-wide. Previous work has determined specific residues within the "P-box" of the ER DNA binding domain determine the specificity of ER binding to ER motifs as opposed to glucocorticoid receptor (GR) motifs [322, 323]. We have generated HA-tagged expression constructs for wildtype *ESR1* as well as a mutant *ESR1*. We can transfect each construct into cells lacking ER protein expression, for example HEK293T cells, though we would prefer to have physiologically-relevant expression levels of the exogenous ER. Alternatively, we could edit the endogenous *ESR1* locus in a cell line that expresses ER but does not require it for proliferation, such as the ovarian carcinoma cell line SKOV3 [324]. Once we can express both wildtype and mutant ER, we can starve the cells of hormone and acutely treat with estrogen. We would predict that the wildtype ER protein would bind to ER motifs and that the mutant ER protein would bind elsewhere in the genome, to GR motifs. Importantly, the squelching model would predict that the repressed genes would largely overlap between the two ER protein conditions, as the location of ER binding will not affect the repression due to a redistribution of limiting coactivators.

6.4 Discordance between TRPS1 activity score and *TRPS1* expression

6.4.1 Conclusions

In Chapter 3, we propose the use of a score as a readout of TRPS1 activity in breast tumor transcriptomic data and demonstrate that higher TRPS1 activity is associated with worse patient outcome, specifically for patients with tumors of the Luminal A subtype. This score is based

on the primary response genes that are activated or repressed immediately upon acute TRPS1 depletion. At first glance, a simpler measurement of TPRS1 activity might be *TRPS1* expression. However, these measures are complicated by two additional observations. First, *TRPS1* is often co-amplified with the proto-oncogene *MYC*, amplification of which is independently associated with worse outcomes for patients [240]. Second, *TRPS1* expression is also correlated with *ESR1* and *GATA3* expression, which are themselves associated with better outcomes for patients [52, 241]. We would predict that our TRPS1 activity score would correlate with *TRPS1* expression, and yet these two measures of TRPS1 activity are associated with outcomes in the opposite direction. Based on this, we predict that TRPS1 activity is regulated by additional factors beyond simply *TRPS1* expression.

6.4.2 Future directions

Are other *TRPS1* splice isoforms expressed in breast cancer cells?

When the *TRPS1* gene was first cloned, two splice isoforms were noted in the RT-PCR products, one which could in theory produce a protein with 13 additional amino acids in frame just upstream from the start codon [47]. The authors noted that the translation initiation signal differed significantly from the consensus Kozak sequence and predicted that this protein isoform would be less likely to be expressed. Intruiguingly, the authors also found two predominant isoforms via Northern blot in several human fetal tissues, including brain, lung, and kidney. The shorter of the two corresponds to the size of the product of an internal polyadenylation and cleavage site that produced the complementary DNA (cDNA) for an evolutionarily conserved expressed sequence tag clone in their database search. This shorter transcript would be predicted to form a truncated TRPS1 protein lacking its nuclear localization sequences, GATA-like DNA-binding zinc finger, and carboxy-terminal Ikaros-like corepressor-recruiting zinc fingers. However, there would be no stop codon, so it is unclear if this would lead to expression of a truncated protein. If this shorter isoform were expressed, we would predict that it would have no activity in transcriptional regulation via the mechanisms we have previously discussed. In this way, *TRPS1*

expression level would differ from our TRPS1 activity score.

We do not see evidence of this protein isoform in our immunoblots. If it were, using an antibody against the HA tag or against TRPS1, we would expect to see a lower molecular weight band that is depleted by dTAG treatment. However, it remains possible that there are circumstances under which this isoform is expressed in tumors. To address this possibility, we could re-analyze the publicly available RNA-seq data from breast cancer patient tumors and isolate reads from the 3′-end of the full-length transcript. If expression of this part of the mRNA correlated with our TRPS1 activity score, this would suggest that a portion of *TRPS1* expression does not produce a protein product with activity in the way we have described. This would then generate two additional questions — does the truncated protein have an independent function, and what regulates this alternative polyadenylation and cleavage? The first question could be addressed by exogenous expression of the truncated cDNA, potentially coupled with rapid depletion. As this protein product may not directly regulate transcription, we would at this point be searching for a phenotype, for example an effect on cell proliferation.

Is TRPS1 post-translationally regulated?

Our approach to the second question would also address the broader question of whether TRPS1 is post-translationally regulated, such as through post-translational modifications or subcellular localization. Our first choice for prioritizing potential protein interaction partners would be to analyze published data from a recent proximity ligation experiment [57]. Beyond the corepressor complexes, we would look for peptides corresponding to kinases or phosphatases as a starting point. Indeed, in the curated phosphoproteomic database PhosphoSitePlus, there are many reported peptides corresponding to phosphorylated residues on TRPS1 from publicly available data [325–328]. If we were interested in a set of post-translational modifications, we could express mutants of TRPS1 that mimic or prevent phosphorylation. With these mutants, we would assess sub-cellular localization as well as transcriptional activity. In this way, we could determine how post-translational regulation of TRPS1 could uncouple *TRPS1* expression and

TRPS1 activity in transcriptional regulation of the genes we identified as TRPS1-dependent.

6.5 Functional follow-up of GWAS hits

6.5.1 Conclusions

Genome-wide association studies (GWAS), which use powerful methods to identify genetic associations in an unbiased manner. However, only a small fraction of GWAS hits have currently been studied functionally. Recent GWAS have identified common genetic variants in the *TRPS1* locus that are associated with breast cancer incidence and blood cholesterol traits [214, 266]. In this dissertation, we perform experimental follow-up to better understand the mechanisms by which TRPS1 contributes to these phenotypes. We used the GWAS hits as a starting point, assuming that these single nucleotide polymorphisms (SNPs) causally affect the phenotypes and that this effect is mediated via a *cis*-regulatory effect on *TRPS1* expression. For the blood cholesterol traits, we were able to support this assumption with expression quantitative trait loci (eQTL) data. In pancreas tissue, the SNPs associated with lower levels of a specific blood cholesterol trait are associated with higher *TRPS1* expression. This result not only supports our hypothesis that the SNPs regulate blood cholesterol levels via an effect on *TRPS1* expression but also is consistent with the direction of effect we see in our PRO-seq data, in which TRPS1 depletion increases cholesterol biosynthesis gene transcription.

However, for the trait of breast cancer incidence, we were not able to find evidence of colocalization between the SNPs associated with *TRPS1* expression and those associated with the phenotype. This does not rule out the possibility that these SNPs affect breast cancer incidence via changes in *TRPS1* expression. When a SNP in a non-coding region of the genome causally regulates expression of a gene in *cis* via the binding in *trans* of a TF, that TF must be active in the tissue under study for an eQTL to be identified. While breast is a tissue type represented in the Genotype-Tissue Expression (GTEx) project data we used for eQTL analysis, it is possible that normal breast tissue is not the context in which the relevant SNP acts and that only later in tumorigenesis is the relevant TF active.

6.5.2 Future directions

Is there an association between the breast cancer associated SNPs and *TRPS1* expression in breast tumor tissue?

To address this hypothesis, we can analyze the publicly available data from the same METABRIC cohort that we used for our survival analysis or in TCGA breast cancer data [11, 226, 227]. Both of these datasets contain transcriptional data as well as genetic variant calling. An eQTL analysis has been previously performed on these datasets [329]. There are many challenges to using tumor data, including complex genetic changes and heterogeneity in tumor purity. However, we can use the summary statistics from this study to test for colocalization between the SNPs associated with breast cancer incidence and those associated with *TRPS1* expression in breast cancer tumor data. Furthermore, we can isolate tumors based on their intrinsic subtypes to test whether there is evidence for colocalization specifically within Luminal A tumors.

6.6 Discordance between nascent transcription and downstream assays

6.6.1 Conclusions

In Chapter 4, we performed several downstream assays to follow up on our initial observation that nascent transcription of cholesterol biosynthesis genes was activated upon TRPS1 depletion. This is a result that was consistent across three independent clones, as well as for the one clone used for multiple time points in Chapter 3. Along with the GWAS results associating *TRPS1* with blood cholesterol traits, we have evidence to support the hypothesis that TRPS1 regulates cholesterol biosynthesis. However, our effect sizes are small for all our genome-wide sequencing-based data. Lipid droplet staining, RT-qPCR, and the cholesterol assays were quite variable in our hands. Amid this noise, we do not have much power to detect small signals. If cholesterol abundance is changing, it would help to generate a large effect size. Unfortunately, in our three clones in which we tagged TRPS1 with dTAG, two to three of the four *TRPS1*

alleles are knockouts based on cloning and sequencing PCR products of genomic DNA. We have not definitively identified all the alleles yet, but at best we have tagged one to two of the four *TRPS1* alleles in each clone. TRPS1 protein expression of each clone is lower than that of the parental T47D cells.

6.6.2 Future directions

Can we endogenously tag all *TRPS1* alleles in a cell line with high *TRPS1* expression?

To increase the effect size of TRPS1 depletion, it would help to have TRPS1 expression be higher at baseline. While it would be possible to simply exogenously over-express TRPS1, we would be concerned about the physiologic relevance of excessive TRPS1 protein abundance that might lead to differential protein-protein interactions and sub-cellular or even sub-nuclear and genomic localization. Our ideal clone would have endogenous *TRPS1* tagged at all the alleles. Most breast cancer cell lines have more than two copies of *TRPS1*, based on data from the Cancer Dependency Map project [77]. Of these, MCF-7 and CAMA-1 cells have three copies and would be our next best choice. In addition, we would experiment with fluorescent selection for genetic insertion and use three different markers, green, blue, and red fluorescent proteins instead of antibiotic resistance markers [330]. In this way, we can use flow cytometry to isolate cells that have three insertion events for follow-up tagging confirmation. We have previously attempted to transfect multiple repair templates with different antibiotic resistance markers but were unable to isolate any clones that survived even double selection. One other technical note is our use of an ultraviolet lamp when isolating repair template PCR product from agarose gel. In previous successful rounds of tagging, we had used a blue light lamp that allows for visualization of SYBR Safe-stained gels. However, it is possible that excessive ultraviolet exposure damaged our repair templates to a degree incompatible with use in homology-directed repair. For this reason and for technical ease of plasmid amplification and purification over PCR product gel isolation, we would use the CRIS-PITCh system discussed in Chapter 1 in our next round of tagging [166]. With these changes, it may be possible to endogenously tag all *TRPS1* alleles and increase the

resultant effect sizes of TRPS1 depletion.

Does TRPS1 regulate blood cholesterol traits *in vivo*?

TRPS1 is expressed during development in multiple organ systems, and mice lacking the GATA-like DNA-binding zinc finger of *TRPS1* die shortly after birth [48, 61, 331]. However, an inducible knockout mouse has never been generated for *TRPS1*. This would involve generating a *TRPS1* allele flanked by loxP sites and crossing these mice with those with whole body expression of a tamoxifen-inducible Cre recombinase [332]. After generating homozygous animals, we can allow for proper *TRPS1* expression during development and then induce a knockout in adulthood and measure blood cholesterol levels at later time points. Importantly, generating and phenotyping this mouse would provide insight into the suitability of TRPS1 as a therapeutic target in breast cancer.

6.7 Data Access

All analysis details and code are available at https://tgscott400.github.io/ER_antagonist_analysis/Vignette.html. Raw sequencing files and processed counts and *bigWig* files are available from GEO accession records GSE251785 (PRO-seq), GSE251793 (ATAC-seq), and GSE236174 (ChIP-seq).

References

1. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA Cancer J. Clin.* **73,** 17–48 (Jan. 2023).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71,** 209–249 (May 2021).
3. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: a topic in evolution. *Ann. Oncol.* **26,** 1291–1299 (July 2015).
4. Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250,** 1684–1689 (Dec. 1990).
5. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265,** 2088–2090 (Sept. 1994).
6. Venkitaraman, A. R. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. *Science* **343,** 1470–1475 (Mar. 2014).
7. FitzGerald, M. G. *et al.* Germline mutations in PTEN are an infrequent cause of genetic predisposition to breast cancer. *Oncogene* **17,** 727–731 (Aug. 1998).
8. Garber, J. E. *et al.* Follow-up study of twenty-four families with Li-Fraumeni syndrome. *Cancer Res.* **51,** 6094–6097 (Nov. 1991).
9. Pharoah, P. D., Guilford, P., Caldas, C. & International Gastric Cancer Linkage Consortium. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* **121,** 1348–1353 (Dec. 2001).
10. Boardman, L. A. *et al.* Increased risk for cancer in patients with the Peutz-Jeghers syndrome. *Ann. Intern. Med.* **128,** 896–899 (June 1998).
11. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (Oct. 2012).
12. Kastenhuber, E. R. & Lowe, S. W. Putting p53 in Context. *Cell* **170,** 1062–1078 (Sept. 2017).
13. Cantley, L. C. The phosphoinositide 3-kinase pathway. *Science* **296,** 1655–1657 (May 2002).
14. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
15. Cheang, M. C. U. *et al.* Defining breast cancer intrinsic subtypes by quantitative receptor expression. *Oncologist* **20,** 474–482 (May 2015).
16. Hammond, M. E. H. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer (unabridged version). *Arch. Pathol. Lab. Med.* **134,** e48–72 (July 2010).
17. Howlader, N. *et al.* US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J. Natl. Cancer Inst.* **106** (Apr. 2014).
18. Eggemann, H., Altmann, U., Costa, S.-D. & Ignatov, A. Survival benefit of tamoxifen and aromatase inhibitor in male and female breast cancer. *J. Cancer Res. Clin. Oncol.* **144,** 337–341 (Feb. 2018).

19. Wolff, A. C. *et al.* Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* **31,** 3997–4013 (Nov. 2013).
20. Gianni, L. *et al.* Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. *Lancet Oncol.* **13,** 25–32 (Jan. 2012).
21. Aysola, K. *et al.* Triple Negative Breast Cancer - An Overview. *Hereditary Genet* **2013** (2013).
22. Cheang, M. C. U. *et al.* Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. *Clin. Cancer Res.* **14,** 1368–1376 (Mar. 2008).
23. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–1167 (Mar. 2009).
24. Kensler, K. H. *et al.* PAM50 Molecular Intrinsic Subtypes in the Nurses' Health Study Cohorts. *Cancer Epidemiol. Biomarkers Prev.* **28,** 798–806 (Apr. 2019).
25. Van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (Jan. 2002).
26. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351,** 2817–2826 (Dec. 2004).
27. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* **12,** R68 (Sept. 2010).
28. Goldhirsch, A. *et al.* Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. *Ann. Oncol.* **24,** 2206–2223 (Sept. 2013).
29. Henderson, B. E., Ross, R. K., Pike, M. C. & Casagrande, J. T. Endogenous hormones as a major factor in human cancer. *Cancer Res.* **42,** 3232–3239 (Aug. 1982).
30. McInerney, E. M., Tsai, M. J., O'Malley, B. W. & Katzenellenbogen, B. S. Analysis of estrogen receptor transcriptional enhancement by a nuclear hormone receptor coactivator. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 10069–10073 (Sept. 1996).
31. Kamei, Y. *et al.* A CBP integrator complex mediates transcriptional activation and AP-1 inhibition by nuclear receptors. *Cell* **85,** 403–414 (May 1996).
32. Chakravarti, D. *et al.* Role of CBP/P300 in nuclear receptor signalling. en. *Nature* **383,** 99–103 (Sept. 1996).
33. McKenna, N. J. & O'Malley, B. W. Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell* **108,** 465–474 (Feb. 2002).
34. Deroo, B. J. & Korach, K. S. Estrogen receptors and human disease. *J. Clin. Invest.* **116,** 561–570 (Mar. 2006).
35. Hall, J. M., Couse, J. F. & Korach, K. S. The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J. Biol. Chem.* **276,** 36869–36872 (Oct. 2001).
36. Kumar, V. & Chambon, P. The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell* **55,** 145–156 (Oct. 1988).
37. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145,** 622–634 (2011).

38. Jameera Begam, A., Jubie, S. & Nanjan, M. J. Estrogen receptor agonists/antagonists in breast cancer therapy: A critical review. *Bioorg. Chem.* **71,** 257–274 (Apr. 2017).
39. Francis, P. A. *et al.* Tailoring Adjuvant Endocrine Therapy for Premenopausal Breast Cancer. *N. Engl. J. Med.* **379,** 122–137 (July 2018).
40. Gnant, M. *et al.* Zoledronic acid combined with adjuvant endocrine therapy of tamoxifen versus anastrozol plus ovarian function suppression in premenopausal early breast cancer: final analysis of the Austrian Breast and Colorectal Cancer Study Group Trial 12. *Ann. Oncol.* **26,** 313–320 (Feb. 2015).
41. Harbeck, N. *et al.* Breast cancer (Primer). en. *Nature Reviews: Disease Primers; London* **5,** s41572–019 (2019).
42. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365,** 1687–1717 (2005).
43. Clarke, R., Tyson, J. J. & Dixon, J. M. Endocrine resistance in breast cancer–An overview and update. *Mol. Cell. Endocrinol.* **418 Pt 3,** 220–234 (Dec. 2015).
44. Ellis, M. J. *et al.* Outcome prediction for estrogen receptor-positive breast cancer based on postneoadjuvant endocrine therapy tumor characteristics. *J. Natl. Cancer Inst.* **100,** 1380–1388 (Oct. 2008).
45. Carlson, R. W. & Henderson, I. C. Sequential hormonal therapy for metastatic breast cancer after adjuvant tamoxifen or anastrozole. *Breast Cancer Res. Treat.* **80 Suppl 1,** S19–26, discussion S27–8 (2003).
46. Dodwell, D., Wardley, A. & Johnston, S. Postmenopausal advanced breast cancer: options for therapy after tamoxifen and aromatase inhibitors. en. *Breast* **15,** 584–594 (Oct. 2006).
47. Momeni, P. *et al.* Mutations in a new gene, encoding a zinc-finger protein, cause tricho-rhino-phalangeal syndrome type I. *Nat. Genet.* **24,** 71–74 (Jan. 2000).
48. Malik, T. H., Von Stechow, D., Bronson, R. T. & Shivdasani, R. A. Deletion of the GATA domain of TRPS1 causes an absence of facial hair and provides new insights into the bone disorder in inherited tricho-rhino-phalangeal syndromes. *Mol. Cell. Biol.* **22,** 8592–8600 (Dec. 2002).
49. Gai, Z. *et al.* Trps1 functions downstream of Bmp7 in kidney development. *J. Am. Soc. Nephrol.* **20,** 2403–2411 (Nov. 2009).
50. Radvanyi, L. *et al.* The gene associated with trichorhinophalangeal syndrome in humans is overexpressed in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 11005–11010 (Aug. 2005).
51. Wu, L. *et al.* A central role for TRPS1 in the control of cell cycle and cancer development. *Oncotarget* **5,** 7677–7690 (Sept. 2014).
52. Lin, H.-Y., Zeng, D., Liang, Y.-K., Wei, X.-L. & Chen, C.-F. GATA3 and TRPS1 are distinct biomarkers and prognostic factors in breast cancer: database mining for GATA family members in malignancies. *Oncotarget* **8,** 34750–34761 (May 2017).
53. Serandour, A. A., Mohammed, H., Miremadi, A., Mulder, K. W. & Carroll, J. S. TRPS1 regulates oestrogen receptor binding and histone acetylation at enhancers. *Oncogene* **37,** 5281–5291 (Sept. 2018).
54. Ai, D. *et al.* TRPS1: a highly sensitive and specific marker for breast carcinoma, especially for triple-negative breast cancer. *Mod. Pathol.* **34,** 710–719 (Apr. 2021).

55. Sanchez-Garcia, F. *et al.* Integration of genomic data enables selective discovery of breast cancer drivers. *Cell* **159,** 1461–1475 (Dec. 2014).
56. Witwicki, R. M. *et al.* TRPS1 Is a Lineage-Specific Transcriptional Dependency in Breast Cancer. *Cell Rep.* **25,** 1255–1267.e5 (Oct. 2018).
57. Elster, D. *et al.* TRPS1 shapes YAP/TEAD-dependent transcription in breast cancer cells. *Nat. Commun.* **9,** 3115 (Aug. 2018).
58. Wang, Y. *et al.* Tricho-rhino-phalangeal syndrome 1 protein functions as a scaffold required for ubiquitin-specific protease 4-directed histone deacetylase 2 de-ubiquitination and tumor growth. *Breast Cancer Res.* **20,** 83 (Aug. 2018).
59. Kas, S. M. *et al.* Insertional mutagenesis identifies drivers of a novel oncogenic pathway in invasive lobular breast carcinoma. *Nat. Genet.* **49,** 1219–1230 (Aug. 2017).
60. Ko, L. J. & Engel, J. D. DNA-binding specificities of the GATA transcription factor family. *Mol. Cell. Biol.* **13,** 4011–4022 (July 1993).
61. Malik, T. H. *et al.* Transcriptional repression and developmental functions of the atypical vertebrate GATA protein TRPS1. *EMBO J.* **20,** 1715–1725 (Apr. 2001).
62. Wang, Y. *et al.* Atypical GATA transcription factor TRPS1 represses gene expression by recruiting CHD4/NuRD(MTA2) and suppresses cell migration and invasion by repressing TP63 expression. *Oncogenesis* **7,** 96 (Dec. 2018).
63. Cornelissen, L. M. *et al.* TRPS1 acts as a context-dependent regulator of mammary epithelial cell growth/differentiation and breast cancer development. *Genes Dev.* **34,** 179–193 (Feb. 2020).
64. Wuelling, M. *et al.* The multi zinc-finger protein Trps1 acts as a regulator of histone deacetylation during mitosis. *Cell Cycle* **12,** 2219–2232 (July 2013).
65. Huang, J.-Z. *et al.* Down-regulation of TRPS1 stimulates epithelial-mesenchymal transition and metastasis through repression of FOXA1. *J. Pathol.* **239,** 186–196 (June 2016).
66. Hu, J. *et al.* TRPS1 Suppresses Breast Cancer Epithelial-mesenchymal Transition Program as a Negative Regulator of SUZ12. *Transl. Oncol.* **11,** 416–425 (Apr. 2018).
67. Stinson, S. *et al.* TRPS1 targeting by miR-221/222 promotes the epithelial to mesenchymal transition in breast cancer. *Sci. Signal.* **4,** ra41 (June 2011).
68. Yang, J. *et al.* TRPS1 drives heterochromatic origin refiring and cancer genome evolution. *Cell Rep.* **34,** 108814 (Mar. 2021).
69. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (Feb. 2013).
70. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (Feb. 2013).
71. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (Aug. 2012).
72. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411,** 494–498 (May 2001).
73. Caplen, N. J., Parrish, S., Imani, F., Fire, A. & Morgan, R. A. Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 9742–9747 (Aug. 2001).
74. Sui, G. *et al.* A DNA vector-based RNAi technology to suppress gene expression in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 5515–5520 (Apr. 2002).

75. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296,** 550–553 (Apr. 2002).
76. Rivas, F. V. *et al.* Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat. Struct. Mol. Biol.* **12,** 340–349 (Apr. 2005).
77. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49,** 1779–1784 (Dec. 2017).
78. Fuentes, N. & Silveyra, P. Estrogen receptor signaling mechanisms. *Adv. Protein Chem. Struct. Biol.* **116,** 135–170 (Feb. 2019).
79. Jensen, E. V. *et al.* A two-step mechanism for the interaction of estradiol with rat uterus. *Proc. Natl. Acad. Sci. U. S. A.* **59,** 632–638 (Feb. 1968).
80. Jensen, E. V. *et al.* Estrogen-binding substances of target tissues. *Science* **158,** 529–530 (Oct. 1967).
81. Clopper, K. C. & Taatjes, D. J. Chemical inhibitors of transcription-associated kinases. *Curr. Opin. Chem. Biol.* **70,** 102186 (Oct. 2022).
82. Olson, C. M. *et al.* Pharmacological perturbation of CDK9 using selective CDK9 inhibition or degradation. *Nat. Chem. Biol.* **14,** 163–170 (Feb. 2018).
83. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nature* **468,** 1067–1073 (Dec. 2010).
84. Winter, G. E. *et al.* Phthalimide conjugation as a strategy for in vivo target protein degradation. *Science* **348,** 1376–1381 (June 2015).
85. Winter, G. E. *et al.* BET Bromodomain Proteins Function as Master Transcription Elongation Factors Independent of CDK9 Recruitment. *Mol. Cell* **67,** 5–18.e19 (July 2017).
86. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: the past is prologue. *Nat. Rev. Drug Discov.* **21,** 181–200 (Mar. 2022).
87. Schreiber, S. L. Chemical genetics resulting from a passion for synthetic organic chemistry. *Bioorg. Med. Chem.* **6,** 1127–1152 (Aug. 1998).
88. Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T. & Kanemaki, M. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat. Methods* **6,** 917–922 (Dec. 2009).
89. Nabet, B. *et al.* The dTAG system for immediate and target-specific protein degradation. *Nat. Chem. Biol.* **14,** 431–441 (2018).
90. Teale, W. D., Paponov, I. A. & Palme, K. Auxin in action: signalling, transport and the control of plant growth and development. *Nat. Rev. Mol. Cell Biol.* **7,** 847–859 (Nov. 2006).
91. Yu, Z., Zhang, F., Friml, J. & Ding, Z. Auxin signaling: Research advances over the past 30 years. *J. Integr. Plant Biol.* **64,** 371–392 (Feb. 2022).
92. Ouellet, F., Overvoorde, P. J. & Theologis, A. IAA17/AXR3: biochemical insight into an auxin mutant phenotype. *Plant Cell* **13,** 829–841 (Apr. 2001).
93. Ruegger, M. *et al.* The TIR1 protein of Arabidopsis functions in auxin response and is related to human SKP2 and yeast grr1p. *Genes Dev.* **12,** 198–207 (Jan. 1998).

94. Sathyan, K. M., Scott, T. G. & Guertin, M. J. ARF-AID: A Rapidly Inducible Protein Degradation System That Preserves Basal Endogenous Protein Levels. *Curr. Protoc. Mol. Biol.* **132,** e124 (Sept. 2020).
95. Kubota, T., Nishimura, K., Kanemaki, M. T. & Donaldson, A. D. The Elg1 replication factor C-like complex functions in PCNA unloading during DNA replication. *Mol. Cell* **50,** 273–280 (Apr. 2013).
96. Nishimura, K. & Kanemaki, M. T. Rapid depletion of budding yeast proteins via the fusion of an auxin-inducible degron (AID). en. *Curr. Protoc. Cell Biol.* **64,** 20.9.1–16 (Sept. 2014).
97. Li, S., Prasanna, X., Salo, V. T., Vattulainen, I. & Ikonen, E. An efficient auxin-inducible degron system with low basal degradation in human cells. *Nat. Methods* **16,** 866–869 (Sept. 2019).
98. Yesbolatova, A., Natsume, T., Hayashi, K.-I. & Kanemaki, M. T. Generation of conditional auxin-inducible degron (AID) cells and tight control of degron-fused proteins using the degradation inhibitor auxinole. *Methods* **164-165,** 73–80 (July 2019).
99. Yesbolatova, A. *et al.* The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. *Nat. Commun.* **11,** 5701 (Nov. 2020).
100. Tanaka, S. Construction of Tight Conditional Mutants Using the Improved Auxin-Inducible Degron (iAID) Method in the Budding Yeast Saccharomyces cerevisiae. *Methods Mol. Biol.* **2196,** 15–26 (2021).
101. Li, J. *et al.* A One-step strategy to target essential factors with auxin-inducible degron system in mouse embryonic stem cells. *Front Cell Dev Biol* **10,** 964119 (Aug. 2022).
102. Clackson, T. *et al.* Redesigning an FKBP-ligand interface to generate chemical dimerizers with novel specificity. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 10437–10442 (Sept. 1998).
103. Nabet, B. *et al.* Rapid and direct control of target protein levels with VHL-recruiting dTAG molecules. *Nat. Commun.* **11,** 4687 (Sept. 2020).
104. Ito, T. *et al.* Identification of a primary target of thalidomide teratogenicity. *Science* **327,** 1345–1350 (Mar. 2010).
105. Buckley, D. L. *et al.* Small-molecule inhibitors of the interaction between the E3 ligase VHL and HIF1α. *Angew. Chem. Int. Ed Engl.* **51,** 11463–11467 (Nov. 2012).
106. Bonger, K. M., Chen, L.-C., Liu, C. W. & Wandless, T. J. Small-molecule displacement of a cryptic degron causes conditional protein degradation. *Nat. Chem. Biol.* **7,** 531–537 (July 2011).
107. Yamanaka, S. *et al.* An IMiD-induced SALL4 degron system for selective degradation of target proteins. *Commun Biol* **3,** 515 (Sept. 2020).
108. Chung, H. K. *et al.* Tunable and reversible drug control of protein production via a self-excising degron. *Nat. Chem. Biol.* **11,** 713–720 (Sept. 2015).
109. England, C. G., Luo, H. & Cai, W. HaloTag technology: a versatile platform for biomedical applications. *Bioconjug. Chem.* **26,** 975–986 (June 2015).
110. Los, G. V. *et al.* HaloTag: a novel protein labeling technology for cell imaging and protein analysis. *ACS Chem. Biol.* **3,** 373–382 (June 2008).
111. Cook, A., Walterspiel, F. & Deo, C. HaloTag-Based Reporters for Fluorescence Imaging and Biosensing. *Chembiochem* **24,** e202300022 (June 2023).

112. Neklesa, T. K. *et al.* Small-molecule hydrophobic tagging-induced degradation of HaloTag fusion proteins. *Nat. Chem. Biol.* **7,** 538–543 (July 2011).
113. Buckley, D. L. *et al.* HaloPROTACS: Use of Small Molecule PROTACs to Induce Degradation of HaloTag Fusion Proteins. en. *ACS Chem. Biol.* **10,** 1831–1837 (Aug. 2015).
114. Niopek, D., Wehler, P., Roensch, J., Eils, R. & Di Ventura, B. Optogenetic control of nuclear protein export. *Nat. Commun.* **7,** 10624 (Feb. 2016).
115. Zengerle, M., Chan, K.-H. & Ciulli, A. Selective Small Molecule Induced Degradation of the BET Bromodomain Protein BRD4. *ACS Chem. Biol.* **10,** 1770–1777 (Aug. 2015).
116. Li, L. *et al.* In vivo target protein degradation induced by PROTACs based on E3 ligase DCAF15. *Signal Transduct Target Ther* **5,** 129 (July 2020).
117. Chapuy, B. *et al.* Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. *Cancer Cell* **24,** 777–790 (Dec. 2013).
118. Peterlin, B. M. & Price, D. H. Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell* **23,** 297–305 (Aug. 2006).
119. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154,** 1370–1379 (Sept. 2013).
120. Patel, M. C. *et al.* BRD4 coordinates recruitment of pause release factor P-TEFb and the pausing complex NELF/DSIF to regulate transcription elongation of interferon-stimulated genes. *Mol. Cell. Biol.* **33,** 2497–2507 (June 2013).
121. Sakurai, N. *et al.* BRD4 regulates adiponectin gene induction by recruiting the P-TEFb complex to the transcribed region of the gene. *Sci. Rep.* **7,** 11962 (Sept. 2017).
122. Zheng, B. *et al.* Acute perturbation strategies in interrogating RNA polymerase II elongation factor function in gene expression. *Genes Dev.* **35,** 273–285 (Feb. 2021).
123. Zheng, B. *et al.* Distinct layers of BRD4-PTEFb reveal bromodomain-independent function in transcriptional regulation. *Mol. Cell* (July 2023).
124. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (Oct. 2009).
125. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (Apr. 2012).
126. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366,** 1338–1345 (Dec. 2019).
127. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (Dec. 2014).
128. Moore, J. M. *et al.* Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. *PLoS One* **7,** e34915 (Apr. 2012).
129. González-Buendía, E., Pérez-Molina, R., Ayala-Ortega, E., Guerrero, G. & Recillas-Targa, F. in *Cancer Cell Signaling: Methods and Protocols* (ed Robles-Flores, M.) 53–69 (Springer New York, New York, NY, 2014).
130. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169,** 930–944.e22 (May 2017).
131. Hyle, J. *et al.* Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res.* **47,** 6699–6713 (July 2019).
132. Hyle, J. *et al.* Auxin-inducible degron 2 system deciphers functions of CTCF domains in transcriptional regulation. *Genome Biol.* **24,** 14 (Jan. 2023).

133. Luan, J. *et al.* Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep.* **34,** 108783 (Feb. 2021).
134. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171,** 305–320.e24 (Oct. 2017).
135. Liu, N. Q. *et al.* WAPL maintains a cohesin loading cycle to preserve cell-type-specific distal gene regulation. *Nat. Genet.* **53,** 100–109 (Jan. 2021).
136. Casa, V. *et al.* Redundant and specific roles of cohesin STAG subunits in chromatin looping and transcriptional control. *Genome Res.* **30,** 515–527 (Apr. 2020).
137. Shi, Y., Seto, E., Chang, L. S. & Shenk, T. Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. *Cell* **67,** 377–388 (Oct. 1991).
138. Li, M. *et al.* Human cytomegalovirus IE2 drives transcription initiation from a select subset of late infection viral promoters by host RNA polymerase II. *PLoS Pathog.* **16,** e1008402 (Apr. 2020).
139. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171,** 1573–1588.e28 (Dec. 2017).
140. Dang, C. V. MYC on the path to cancer. *Cell* **149,** 22–35 (Mar. 2012).
141. Ji, H. *et al.* Cell-type independent MYC target genes reveal a primordial signature involved in biomass accumulation. *PLoS One* **6,** e26057 (Oct. 2011).
142. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151,** 56–67 (Sept. 2012).
143. Nie, Z. *et al.* c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151,** 68–79 (Sept. 2012).
144. Sabò, A. *et al.* Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. *Nature* **511,** 488–492 (July 2014).
145. Walz, S. *et al.* Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature* **511,** 483–487 (July 2014).
146. Lorenzin, F. *et al.* Different promoter affinities account for specificity in MYC-dependent gene regulation. *Elife* **5** (July 2016).
147. Muhar, M. *et al.* SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science* **360,** 800–805 (2018).
148. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (Aug. 2006).
149. Bates, L. E., Alves, M. R. P. & Silva, J. C. R. Auxin-degron system identifies immediate mechanisms of OCT4. *Stem Cell Reports* **16,** 1818–1831 (July 2021).
150. Sheppard, H. E. *et al.* Targeted brachyury degradation disrupts a highly specific autoregulatory program controlling chordoma cell identity. *Cell Rep Med* **2,** 100188 (Jan. 2021).
151. Stengel, K. R., Ellis, J. D., Spielman, C. L., Bomber, M. L. & Hiebert, S. W. Definition of a small core transcriptional circuit regulated by AML1-ETO. *Mol. Cell* **81,** 530–545.e5 (Feb. 2021).
152. Sathyan, K. M. *et al.* An improved auxin-inducible degron system preserves native protein levels and enables rapid and specific protein depletion. *Genes & development* **33,** 1441–1455 (2019).

153. Erb, M. A. *et al.* Transcription control by the ENL YEATS domain in acute leukaemia. *Nature* **543,** 270–274 (Mar. 2017).
154. Moore, I. & Murphy, A. Validating the location of fluorescent protein fusions in the endomembrane system. *Plant Cell* **21,** 1632–1636 (June 2009).
155. Van den Berg, J. *et al.* A limited number of double-strand DNA breaks is sufficient to delay cell cycle progression. *Nucleic Acids Res.* **46,** 10132–10144 (Nov. 2018).
156. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528,** 84–87 (Dec. 2015).
157. Essletzbichler, P. *et al.* Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. *Genome Res.* **24,** 2059–2065 (Dec. 2014).
158. Carette, J. E. *et al.* Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. *Nat. Biotechnol.* **29,** 542–546 (May 2011).
159. Kotecki, M., Reddy, P. S. & Cochran, B. H. Isolation and characterization of a near-haploid human cell line. *Exp. Cell Res.* **252,** 273–280 (Nov. 1999).
160. Li, Y. & Shuai, L. A versatile genetic tool: haploid cells. *Stem Cell Res. Ther.* **8,** 197 (Sept. 2017).
161. Llargués-Sistac, G., Bonjoch, L. & Castellvi-Bel, S. HAP1, a new revolutionary cell model for gene editing using CRISPR-Cas9. *Front Cell Dev Biol* **11,** 1111488 (Mar. 2023).
162. Liao, J. Q., Zhou, G. & Zhou, Y. in *Induced Pluripotent Stem (iPS) Cells: Methods and Protocols* (eds Nagy, A. & Turksen, K.) 575–588 (Springer US, New York, NY, 2022).
163. Bräuer, M., Zich, M. T., Önder, K. & Müller, N. The influence of commonly used tags on structural propensities and internal dynamics of peptides. *Monatshefte für Chemie - Chemical Monthly* **150,** 913–925 (May 2019).
164. Nikiforov, T. T., Rendle, R. B., Kotewicz, M. L. & Rogers, Y. H. The use of phosphorothioate primers and exonuclease hydrolysis for the preparation of single-stranded PCR products and their detection by solid-phase hybridization. *PCR Methods Appl.* **3,** 285–291 (Apr. 1994).
165. Gilar, M., Belenky, A., Budman, Y., Smisek, D. L. & Cohen, A. S. Study of phosphorothioate modified oligonucleotide resistance to 3 -exonuclease using capillary electrophoresis. *J. Chromatogr. B Biomed. Sci. Appl.* **714,** 13–20 (Aug. 1998).
166. Nakade, S. *et al.* Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nat. Commun.* **5,** 5560 (Nov. 2014).
167. Singh, A. M. An Efficient Protocol for Single-Cell Cloning Human Pluripotent Stem Cells. *Front Cell Dev Biol* **7,** 11 (Jan. 2019).
168. Han, Z., Madhavan, B. K., Kaymak, S., Nawroth, P. & Kumar, V. A Fast and Reliable Method to Generate Pure, Single Cell-derived Clones of Mammalian Cells. *Bio Protoc* **12** (Aug. 2022).
169. Gallagher, C. & Kelly, P. S. Selection of High-Producing Clones Using FACS for CHO Cell Line Development. *Methods Mol. Biol.* **1603,** 143–152 (2017).
170. Douglass Jr, E. F., Miller, C. J., Sparer, G., Shapiro, H. & Spiegel, D. A. A comprehensive mathematical model for three-body binding equilibria. *J. Am. Chem. Soc.* **135,** 6092–6099 (Apr. 2013).

171. Rodbard, D., Feldman, Y., Jaffe, M. L. & Miles, L. E. Kinetics of two-site immunoradiometric ('sandwich') assays-II. Studies on the nature of the 'high-dose hook effect'. *Immunochemistry* **15,** 77–82 (Feb. 1978).
172. Tan, X. *et al.* Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature* **446,** 640–645 (Apr. 2007).
173. Zhang, N.-Y. *et al.* Nano proteolysis targeting chimeras (PROTACs) with anti-hook effect for tumor therapy. en. *Angew. Chem. Weinheim Bergstr. Ger.* **135** (Sept. 2023).
174. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 (Feb. 2013).
175. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352,** 1225–1228 (2016).
176. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 (2013).
177. Blumberg, A. *et al.* Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC biology* **19,** 1–17 (2021).
178. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Identification of regulatory elements from nascent transcription using dREG. *Genome research* **29,** 293–303 (2019).
179. Scruggs, B. S. *et al.* Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Molecular cell* **58,** 1101–1112 (2015).
180. Anderson, W. D., Duarte, F. M., Civelek, M. & Guertin, M. J. Defining data-driven primary transcript annotations with primaryTranscriptAnnotation in R. *Bioinformatics* **36,** 2926–2928 (2020).
181. Zhao, Y. *et al.* Deconvolution of Expression for Nascent RNA Sequencing Data (DENR) Highlights Pre-RNA Isoform Diversity in Human Cells. *bioRxiv* (2021).
182. Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell reports* **2,** 1025–1035 (2012).
183. Duarte, F. M. *et al.* Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. *Genes & development* **30,** 1731–1746 (2016).
184. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Molecular cell* **50,** 212–222 (2013).
185. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).
186. Mayer, A. *et al.* Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161,** 541–554 (2015).
187. Smith, J. P., Dutta, A. B., Sathyan, K. M., Guertin, M. J. & Sheffield, N. C. PEPPRO: quality control and processing of nascent RNA profiling data. *Genome Biology* **22,** 1–17 (2021).
188. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
189. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357–359 (2012).
190. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17,** 10–12 (2011).

191. Edwards, J. A. & Edwards, R. A. Fastq-pair: efficient synchronization of paired-end fastq files. *bioRxiv,* 552885 (2019).
192. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
193. Martins, A. L. & Guertin, M. J. *fqdedup: Remove PCR duplicates from FASTQ files* 2018.
194. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
195. Martins, A. L., Walavalkar, N. M., Anderson, W. D., Zang, C. & Guertin, M. J. Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic acids research* **46,** e9–e9 (2018).
196. Li, H. *et al. Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences* 2013.
197. Sarkar, D. *Lattice: multivariate data visualization with R* (Springer Science & Business Media, 2008).
198. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 1–21 (2014).
199. Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome research* **26,** 799–811 (2016).
200. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* **46,** 1311–1320 (2014).
201. Karolchik, D. *et al.* The UCSC genome browser database. *Nucleic acids research* **31,** 51–54 (2003).
202. Stolarczyk, M., Reuter, V. P., Smith, J. P., Magee, N. E. & Sheffield, N. C. Refgenie: a reference genome resource manager. *GigaScience* **9,** giz149 (2020).
203. Yates, A. D. *et al.* Ensembl 2020. *Nucleic acids research* **48,** D682–D688 (2020).
204. Andrews, S. *et al. FastQC: a quality control tool for high throughput sequence data. 2010* 2017.
205. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25,** 3389–3402 (1997).
206. Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54,** 795–804 (1988).
207. Tavera-Mendoza, L. E., Mader, S. & White, J. H. Genome-wide approaches for identification of nuclear receptor target genes. *Nucl. Recept. Signal.* **4,** e018 (July 2006).
208. Lin, C.-Y. *et al.* Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biol.* **5,** R66 (Aug. 2004).
209. Cheung, E. & Kraus, W. L. Genomic analyses of hormone signaling and gene regulation. *Annu. Rev. Physiol.* **72,** 191–218 (2010).
210. Ioannidis, J. P. A. *et al.* Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41,** 149–155 (Feb. 2009).
211. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (Dec. 2008).

212. Mendillo, M. L. *et al.* HSF1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. *Cell* **150,** 549–562 (2012).
213. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nature protocols* **11,** 1455–1476 (2016).
214. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. en. *Nat. Genet.* **52,** 572–581 (June 2020).
215. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51,** D977–D985 (Jan. 2023).
216. Boughton, A. P. *et al.* LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics* **37,** 3017–3018 (Sept. 2021).
217. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* **5,** 75 (Apr. 2014).
218. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9,** 1–9 (2008).
219. Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19,** 169 (Mar. 2018).
220. Scott, T. G., Martins, A. L. & Guertin, M. J. Processing and evaluating the quality of genome-wide nascent transcription profiling libraries. *bioRxiv,* 2022–12 (2022).
221. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1,** 417–425 (Dec. 2015).
222. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122,** 33–43 (2005).
223. Guertin, MJ, Zhang, X, Coonrod, SA & Hager, GL. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol. Endocrinol.* **28,** 1522–1533 (2014).
224. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (Oct. 2005).
225. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273 (July 2003).
226. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (Apr. 2012).
227. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7,** 11479 (May 2016).
228. Castro, M. A. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48,** 12–21 (Jan. 2016).
229. Fletcher, M. N. C. *et al.* Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.* **4,** 2464 (2013).
230. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* **7,** e1002240 (Oct. 2011).
231. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94 (Nov. 2017).

232. Ripatti, P. *et al.* Polygenic Hyperlipidemias and Coronary Artery Disease Risk. *Circ Genom Precis Med* **13,** e002725 (Apr. 2020).
233. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med.* **17,** e1003062 (Mar. 2020).
234. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53,** 1415–1424 (Oct. 2021).
235. Gill, G. & Ptashne, M. Negative effect of the transcriptional activator GAL4. *Nature* **334,** 721–724 (Aug. 1988).
236. Tasset, D., Tora, L., Fromental, C., Scheer, E. & Chambon, P. Distinct classes of transcriptional activating domains function by different mechanisms. *Cell* **62,** 1177–1187 (Sept. 1990).
237. Meyer, M. E. *et al.* Steroid hormone receptors compete for factors that mediate their enhancer function. *Cell* **57,** 433–442 (May 1989).
238. Bocquel, M. T., Kumar, V., Stricker, C., Chambon, P. & Gronemeyer, H. The contribution of the N- and C-terminal regions of steroid receptors to activation of transcription is both receptor and cell-specific. *Nucleic Acids Res.* **17,** 2581–2595 (Apr. 1989).
239. Schmidt, SF, Larsen, BD, Loft, A & Mandrup, S. Cofactor squelching: Artifact or fact? *Bioessays* **38,** 618–626 (2016).
240. Savinainen, K. J. *et al.* Expression and copy number analysis of TRPS1, EIF3S3 and MYC genes in breast and prostate cancer. *Br. J. Cancer* **90,** 1041–1046 (Mar. 2004).
241. Chen, J. Q. *et al.* Quantitative immunohistochemical analysis and prognostic significance of TRPS-1, a new GATA transcription factor family member, in breast cancer. *Horm. Cancer* **1,** 21–33 (Feb. 2010).
242. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). https://www.R-project.org/.
243. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (Nov. 2013).
244. Zheng, Q. *et al.* Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. *Biotechniques* **57,** 115–124 (Sept. 2014).
245. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42,** W187–91 (July 2014).
246. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17,** 1518–1552 (June 2022).
247. Lis, J. T. in *Methods in enzymology* 347–353 (Elsevier, 1980).
248. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic acids research* **43,** W39–W49 (2015).
249. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38,** 576–589 (2010).
250. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research* **46,** D260–D266 (2018).

251. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic acids research* **43,** D117–D122 (2015).
252. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11,** 165 (Apr. 2010).
253. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (Apr. 2011).
254. Pantano, L. DEGreport: Report of DEG analysis. *New Jersey, NJ: R package version* **1** (2019).
255. Judd, J. *et al.* A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). *bioRxiv,* 2020.05.18.102277 (May 2020).
256. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (Apr. 2013).
257. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv,* 060012 (Feb. 2021).
258. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12,** 996–1006 (2002).
259. Roth, G. A. *et al.* Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70,** 1–25 (July 2017).
260. Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N. & Stokes 3rd, J. Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study. *Ann. Intern. Med.* **55,** 33–50 (July 1961).
261. Wilson, P. W., Abbott, R. D. & Castelli, W. P. High density lipoprotein cholesterol and mortality. The Framingham Heart Study. *Arteriosclerosis* **8,** 737–741 (1988).
262. US Preventive Services Task Force *et al.* Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: US Preventive Services Task Force Recommendation Statement. *JAMA* **328,** 746–753 (Aug. 2022).
263. Endo, A., Kuroda, M. & Tanzawa, K. Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase by ML-236A and ML-236B fungal metabolites, having hypocholesterolemic activity. en. *FEBS Lett.* **72,** 323–326 (Dec. 1976).
264. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat. Genet.* **34,** 154–156 (June 2003).
265. Jaworski, K., Jankowski, P. & Kosior, D. A. PCSK9 inhibitors - from discovery of a single mutation to a groundbreaking therapy of lipid disorders in one decade. *Arch. Med. Sci.* **13,** 914–929 (June 2017).
266. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12,** e1001779 (Mar. 2015).
267. Selvaraj, M. S. *et al.* Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat. Commun.* **13,** 5995 (Oct. 2022).
268. Ma, L. *et al.* Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. *BMC Med. Genet.* **11,** 55 (Apr. 2010).
269. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600,** 675–679 (Dec. 2021).

270. Wu, H. *et al.* Polygenic Risk Score for Low-Density Lipoprotein Cholesterol Is Associated With Risk of Ischemic Heart Disease and Enriches for Individuals With Familial Hypercholesterolemia. *Circ Genom Precis Med* **14,** e003106 (Feb. 2021).
271. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50,** D687–D692 (Jan. 2022).
272. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16,** 284–287 (May 2012).
273. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (June 2013).
274. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20120362 (May 2013).
275. Drivas, T. G., Lucas, A. & Ritchie, M. D. eQTpLot: a user-friendly R package for the visualization of colocalization between eQTL and GWAS signals. *BioData Min.* **14,** 32 (July 2021).
276. Giambartolomei, C. & et. al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics.* https://doi.org/10.1371/journal.pgen.1004383 (2014).
277. Pan, S. *et al.* COLOCdb: a comprehensive resource for multi-model colocalization of complex traits. *Nucleic Acids Res.*
278. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9,** 676–682 (June 2012).
279. Grundy, S. M. *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139,** e1082–e1143 (June 2019).
280. Miziorko, H. M. Enzymes of the mevalonate pathway of isoprenoid biosynthesis. *Arch. Biochem. Biophys.* **505,** 131–143 (Jan. 2011).
281. Mohamed, A., Molendijk, J. & Hill, M. M. lipidr: A Software Tool for Data Mining and Analysis of Lipidomics Datasets. *J. Proteome Res.* **19,** 2890–2897 (July 2020).
282. Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* **25,** 169–193 (Oct. 2000).
283. Wong, M. L. & Medrano, J. F. Real-time PCR for mRNA quantitation. *Biotechniques* **39,** 75–85 (July 2005).
284. Borrás, M. *et al.* Estradiol-induced down-regulation of estrogen receptor. Effect of various modulators of protein synthesis and expression. *J. Steroid Biochem. Mol. Biol.* **48,** 325–336 (Mar. 1994).
285. Ellison-Zelski, S. J., Solodin, N. M. & Alarid, E. T. Repression of ESR1 through actions of estrogen receptor alpha and Sin3A at the proximal promoter. *Mol. Cell. Biol.* **29,** 4949–4958 (Sept. 2009).
286. Kocanova, S., Mazaheri, M., Caze-Subra, S. & Bystricky, K. Ligands specify estrogen receptor alpha nuclear localization and degradation. *BMC Cell Biol.* **11,** 98 (Dec. 2010).
287. Holding, A. N., Cullen, A. E. & Markowetz, F. Genome-wide Estrogen Receptor-α activation is sustained, not cyclical. *Elife* **7** (Nov. 2018).

288. Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J.* **17,** 244–254 (Jan. 1998).
289. Cirillo, L. A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9,** 279–289 (Feb. 2002).
290. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet.* **43,** 27–33 (2011).
291. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* **42,** 1–28 (2011).
292. Meng, Z., Moroishi, T. & Guan, K.-L. Mechanisms of Hippo pathway regulation. *Genes Dev.* **30,** 1–17 (Jan. 2016).
293. Lin, K. C., Park, H. W. & Guan, K.-L. Regulation of the Hippo Pathway Transcription Factor TEAD. *Trends Biochem. Sci.* **42,** 862–872 (Nov. 2017).
294. Pobbati, A. V. *et al.* Targeting the Central Pocket in Human Transcription Factor TEAD as a Potential Cancer Therapeutic Strategy. *Structure* **23,** 2076–2086 (Nov. 2015).
295. Kaneda, A. *et al.* The novel potent TEAD inhibitor, K-975, inhibits YAP1/TAZ-TEAD protein-protein interactions and exerts an anti-tumor effect on malignant pleural mesothelioma. *Am. J. Cancer Res.* **10,** 4399–4415 (Dec. 2020).
296. Sun, Y. *et al.* Pharmacological blockade of TEAD-YAP reveals its therapeutic limitation in cancer cells. *Nat. Commun.* **13,** 6744 (Nov. 2022).
297. Hagenbeek, T. J. *et al.* An allosteric pan-TEAD inhibitor blocks oncogenic YAP/TAZ signaling and overcomes KRAS G12C inhibitor resistance. *Nat Cancer* **4,** 812–828 (June 2023).
298. Guertin, M. J., Cullen, A. E., Markowetz, F. & Holding, A. N. Parallel factor ChIP provides essential internal control for quantitative differential ChIP-seq. *Nucleic Acids Res.* **46** (Apr. 2018).
299. Fan, F. *et al.* Pharmacological targeting of kinases MST1 and MST2 augments tissue repair and regeneration. *Sci. Transl. Med.* **8,** 352ra108 (Aug. 2016).
300. Ma, S. *et al.* Hippo signalling maintains ER expression and ER+ breast cancer growth. *Nature* **591,** E1–E10 (Mar. 2021).
301. Ma, S. *et al.* Transcriptional repression of estrogen receptor alpha by YAP reveals the Hippo pathway as therapeutic target for ER+ breast cancer. *Nat. Commun.* **13,** 1061 (Feb. 2022).
302. De Bruijn, M. & Dzierzak, E. Runx transcription factors in the development and function of the definitive hematopoietic system. *Blood* **129,** 2061–2069 (Apr. 2017).
303. Illendula, A. *et al.* Small Molecule Inhibitor of CBFβ-RUNX Binding for RUNX Transcription Factor Driven Cancers. *EBioMedicine* **8,** 117–131 (June 2016).
304. Kim, M. S., Gernapudi, R., Choi, E. Y., Lapidus, R. G. & Passaniti, A. Characterization of CADD522, a small molecule that inhibits RUNX2-DNA binding and exhibits antitumor activity. *Oncotarget* **8,** 70916–70940 (Sept. 2017).
305. Maeda, R. *et al.* Molecular Characteristics of DNA-Alkylating PI Polyamides Targeting RUNX Transcription Factors. *J. Am. Chem. Soc.* **141,** 4257–4263 (Mar. 2019).

306. Van Bragt, M. P. A., Hu, X., Xie, Y. & Li, Z. RUNX1, a transcription factor mutated in breast cancer, controls the fate of ER-positive mammary luminal cells. *Elife* **3,** e03881 (Nov. 2014).
307. Chimge, N.-O. *et al.* RUNX1 prevents oestrogen-mediated AXIN1 suppression and β-catenin activation in ER-positive breast cancer. *Nat. Commun.* **7,** 10751 (Feb. 2016).
308. Wang, L., Brugge, J. S. & Janes, K. A. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. *Proc. Natl. Acad. Sci. U. S. A.* **108,** E803–12 (Oct. 2011).
309. Kaczynski, J., Cook, T. & Urrutia, R. Sp1- and Krüppel-like transcription factors. *Genome Biol.* **4,** 206 (Feb. 2003).
310. Turner, J. & Crossley, M. Cloning and characterization of mCtBP2, a co-repressor that associates with basic Krüppel-like factor and other mammalian transcriptional regulators. *EMBO J.* **17,** 5129–5140 (Sept. 1998).
311. Van Vliet, J., Turner, J. & Crossley, M. Human Krüppel-like factor 8: a CACCC-box binding protein that associates with CtBP and represses transcription. *Nucleic Acids Res.* **28,** 1955–1962 (May 2000).
312. Piskacek, M., Havelka, M., Jendruchova, K., Knight, A. & Keegan, L. P. The evolution of the 9aaTAD domain in Sp2 proteins: inactivation with valines and intron reservoirs. *Cell. Mol. Life Sci.* **77,** 1793–1810 (May 2020).
313. López-Knowles, E. *et al.* PI3K pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality. *Int. J. Cancer* **126,** 1121–1131 (Mar. 2010).
314. Jeselsohn, R. *et al.* Emergence of constitutively active estrogen receptor-α mutations in pretreated advanced estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **20,** 1757–1767 (2014).
315. Gertz, J. *et al.* Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* **52,** 25–36 (Oct. 2013).
316. Chen, D. *et al.* Nonlinear relationship between chromatin accessibility and estradiol-regulated gene expression. *Oncogene* (Jan. 2021).
317. Kim, M. Y., Woo, E. M., Chong, Y. T. E., Homenko, D. R. & Kraus, W. L. Acetylation of estrogen receptor alpha by p300 at lysines 266 and 268 enhances the deoxyribonucleic acid binding and transactivation activities of the receptor. *Mol. Endocrinol.* **20,** 1479–1493 (July 2006).
318. Gerber, A. N., Newton, R. & Sasse, S. K. Repression of transcription by the glucocorticoid receptor: A parsimonious model for the genomics era. *J. Biol. Chem.,* 100687 (Apr. 2021).
319. Kelley, K. M. M., Rowan, B. G. & Ratnam, M. Modulation of the folate receptor alpha gene by the estrogen receptor: mechanism and implications in tumor targeting. *Cancer Res.* **63,** 2820–2828 (June 2003).
320. Oesterreich, S. *et al.* Estrogen-mediated down-regulation of E-cadherin in breast cancer cells. *Cancer Res.* **63,** 5203–5208 (Sept. 2003).
321. Rajendran, R. R. *et al.* Regulation of nuclear receptor transcriptional activity by a novel DEAD box RNA helicase (DP97). *J. Biol. Chem.* **278,** 4628–4638 (Feb. 2003).

322. Mader, S., Kumar, V., de Verneuil, H. & Chambon, P. Three amino acids of the oestrogen receptor are essential to its ability to distinguish an oestrogen from a glucocorticoid-responsive element. *Nature* **338,** 271–274 (Mar. 1989).
323. Zilliacus, J., Wright, A. P., Carlstedt-Duke, J. & Gustafsson, J. A. Structural determinants of DNA-binding specificity by steroid receptors. *Mol. Endocrinol.* **9,** 389–400 (Apr. 1995).
324. Hua, W., Christianson, T., Rougeot, C., Rochefort, H. & Clinton, G. M. SKOV3 ovarian carcinoma cells have functional estrogen receptor but are growth-resistant to estrogen and antiestrogens. *J. Steroid Biochem. Mol. Biol.* **55,** 279–289 (Dec. 1995).
325. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512–20 (Jan. 2015).
326. Kettenbach, A. N. *et al.* Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Sci. Signal.* **4,** rs5 (June 2011).
327. Mertins, P. *et al.* Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics* **13,** 1690–1704 (July 2014).
328. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534,** 55–62 (June 2016).
329. Geeleher, P. *et al.* Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biol.* **19,** 130 (Sept. 2018).
330. Finley, K. R., Davidson, A. E. & Ekker, S. C. Three-color imaging using fluorescent proteins in living zebrafish embryos. *Biotechniques* **31,** 66–70, 72 (July 2001).
331. Suemoto, H. *et al.* Trps1 regulates proliferation and apoptosis of chondrocytes through Stat3 signaling. *Dev. Biol.* **312,** 572–581 (Dec. 2007).
332. Hayashi, S. & McMahon, A. P. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: a tool for temporally regulated gene activation/inactivation in the mouse. *Dev. Biol.* **244,** 305–318 (Apr. 2002).