Discovery and Investigation of Long Non-Coding RNAs in Cancer

by

Yashar S. Niknafs

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Cellular and Molecular Biology)
in the University of Michigan
2017

Doctoral Committee:

Professor Arul M. Chinnaiyan, Chair
Assistant Professor Stephanie Bielas
Assistant Professor Hui Jiang
Adjunct Assistant Professor Daniel R. Rhodes
Professor Muneesh Tewari
Professor Thomas E. Wilson

© Yashar Niknafs 2017
All Rights Reserved
yniknafs@umich.edu
ORCID iD: orcid.org/0000-0002-6279-1886

ACKNOWLEDGEMENTS

Completing the work detailed within this dissertation has been a truly humbling, exhilarating, and trying experience. I am so grateful to have had the unbelievable opportunities I've had working at the Michigan Center for Translational Pathology. I would like to thank my advisor Arul Chinnaiyan, for pushing me, challenging me, believing in me, encouraging me, and providing a research environment I could only dream of prior to joining his lab.

Matthew Iyer was another mentor to me during my graduate training, and I will be forever indebted to him for his patience and confidence in my drive, determination, and ability. Matthew taught me pretty much everything I know about programming and bioinformatics. I came to the Chinnaiyan lab with zero experience in programming, and he took me under his wing and I am so grateful for that. Becoming a bioinformatician has been an incredible journey, and I would not want it any other way, and without Matthew none of this would have happened. He continues to be a college and friend that I respect deeply, and I look forward to many years of continued work and friendship.

Rohit Malik was another mentor to me throughout my PhD. He is a remarkable person and a remarkable scientist. His ability to collaborate and support others is unmatched, and in him I found a mentor to guide me in planning experiments and visioning for my projects, but also found a friend who made lab fun and tolerable during the trying times. Thank you, Rohit.

I would also like to thank other players along the way who were instrumental in me getting to this point. John Prensner was my mentor during my first rotation in Arul's lab in 2011. He was my first exposure to the lab, and I am grateful to his guidance and mentorship. Felix Feng became somewhat of a second mentor to me. I am happy to call him a friend and colleague, and without him my PhD would have not been anywhere near as successful or enjoyable. I want to thank Christine Betts, Lori Scott, Xuhong Cao, and Karen Giles for making the lab an easy place to do science. Any hurdles have been surmounted and I have always felt supported by you all and am very grateful for that.

Finally, I would like to thank my family and friends. My mother and father have inspired me to be curious my whole life, and have supported me unwaveringly. They make me confident and make me believe I can do anything, and for that I am so grateful. I want to thank my sister, Shadi for always being by my side. My best friend Trevor Yoak for always having my back unconditionally. And other friends who have been invaluable to me along this journey including Ron Siebenaler, Steve Kregel, Jessica Waninger, Udit Singhal, Sethu Pitchaiya, Tanya Ranginwala, Andrew Adan, Steve Sun, John Donkersloot, and Anna Owensby.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii
LIST OF TABLES xi
LIST OF FIGURES xii
LIST OF APPENDICES xv
LIST OF ABBREVIATIONS xvi
ABSTRACT xvii
CHAPTER 1 Introduction 1
1.1 Cancer Genomics 1
1.2 Power of RNA-seq in cancer genomics 2
1.3 History of the discovery of non-coding RNAs 2
1.4 Functional RNAs belie the Central Dogma 4
1.5 Discovery and cataloging of lncRNAs 5
1.6 Transcriptome reconstruction from short-read RNA-seq 6
1.6 Transcriptome meta-assembly 7
1.7 Nomination of discovered lncRNAs for functional investigation 8
1.8 Methods for functional interrogation of lncRNAs 10
CHAPTER 2 The Landscape of Long Noncoding RNAs in the Human Transcriptome 12
2.1 Abstract 12
2.2 Introduction 13
2.3 Results 14

2.3.1 An expanded landscape of human transcription 14
2.3.2 Coding potential assessment of long RNA transcripts 17
2.3.3 Characterization of long RNAs 18
2.3.4 LncRNAs harboring conserved elements 18
2.3.5 LncRNAs overlapping disease-associated SNPs 19
2.3.6 Differential Expression Analysis 20
2.3.7 Characterization of differentially expressed lncRNAs 22
2.4 Discussion 25
2.5 Methods 27
2.5.1 High performance computing 27
2.5.2 RNA-seq data processing 27
2.5.3 Overview of transcriptome reconstruction 27
2.5.4 Filtration of noise contamination 28
2.5.5 Transcriptome meta-assembly 28
2.5.6 Merging of meta-assemblies 29
2.5.7 Comparisons of MiTranscriptome with reference catalogs 29
2.5.8 Prediction of transcripts of unknown coding potential (TUCP) 29
2.5.9 Proteomics analysis 30
2.5.10 Confidence scoring system 31
2.5.11 Validation of lncRNA transcript by qRT-PCR 31
2.5.12 Cell lines and reagents 32
2.5.13 Evidence for active regulation of transcriptional start sites 33
2.5.14 Conservation analysis 33
2.5.15 GWAS analysis 34
2.5.16 Transcript expression estimation 35
2.5.17 Transcript expression enrichment analysis 35
2.5.18 Benchmarking SSEA performance using microarray gene signatures 36
2.5.19 Discovery of lineage-specific and cancer-specific transcripts 37
2.5.20 Guilt-by-association GSEA analysis 38

2.6 Acknowledgements 38
2.7 Author Contributions 39
CHAPTER 3 TACO Produces Robust Multi-Sample Transcriptome Assemblies from RNA-Seq 50
3.1 Abstract 50
3.2 Introduction 50
3.3 Results 52
3.4 Discussion 55
3.5 Methods 56
3.5.1 Genome-guided assembly 56
3.5.2 Meta-assembly 56
3.5.3 Performance assessment 57
3.5.4 Sample size performance assessment 58
3.5.5 Isoform fraction cutoff assessment 58
3.5.6 Comparison to tissue PacBio long-read sequencing 58
3.5.7 Comparison to brain long-read sequencing 59
3.5.8 Performance of highest expressed transcripts 59
3.5.9 Assessment of performance varying change point parameters 60
3.5.10 Visualization of meta-assemblies 60
3.5.11 Overview of TACO meta-assembly approach 60
3.5.12 Details of TACO algorithm 61
3.6 Acknowledgements 71
3.7 Author Contributions 71
CHAPTER 4 The LncRNA Landscape of Breast Cancer Reveals a Role for DSCAM-AS1 in Breast Cancer Progression 76
4.1 Abstract 76
4.2 Introduction 77
4.3 Results 78
4.3.1 Identification of ER- and breast cancer-associated lncRNAs 78
4.3.2 Characterization of *DSCAM-AS1* 80

4.3.3 *DSCAM-AS1* is implicated in cancer aggression 81
4.3.4 Role of hnRNPL in *DSCAM-AS1* mechanism 84
4.3.5 Role of *DSCAM-AS1* in tamoxifen resistance 86
4.4 Discussion 89
4.5 Methods 90
4.5.1 Cell lines and cell culture 90
4.5.2 Cell proliferation assay 90
4.5.3 Cell viability assay 91
4.5.4 Soft agar colony formation assay 91
4.5.5 Quantitative RT-PCR assay 91
4.5.6 Estrogen and tamoxifen treatment 92
4.5.7 Subcellular fractionation 92
4.5.8 Knockdown and overexpression studies 92
4.5.9 *In vitro* RNA binding assay 93
4.5.10 Mass spectrometry 94
4.5.11 Mass spectrometry data analysis 94
4.5.12 RNA immunoprecipitation 95
4.5.13 Invasion assay 96
4.5.14 Antibodies and immunoblot analyses 96
4.5.15 Chromatin immunoprecipitation (ChIP) 96
4.5.16 ChIP-seq library construction and sequencing analysis 97
4.5.17 ChIP-seq peak promoter overlap 97
4.5.18 Coding potential scoring 98
4.5.19 Xenograft analysis 98
4.5.20 RNA-seq data processing 99
4.5.21 RNA-seq transcript expression estimation 99
4.5.22 Breast cancer tissue expression heatmap generation 99
4.5.23 RNA-seq differential expression testing 100
4.5.24 Associations with Oncomine clinical signatures 100

4.5.25 Association of correlation signatures with Oncomine concepts 101

4.5.26 Survival analysis with TCGA breast data 101

4.5.27 Tissue expression level percentile metric 102

4.5.28 RNA-sequencing library preparation 102

4.5.29 Gene Set Enrichment Analysis 102

4.5.30 *In silico* binding prediction 103

4.5.31 Rapid Amplification of cDNA Ends (RACE) 103

4.5.32 Single-molecule fluorescence *in situ* hybridization (smFISH) 104

4.6 Acknowledgements 104

4.7 Author Contributions 105

CHAPTER 5 Oncogenic Role of *THOR*, a Conserved Cancer/Testis Long Noncoding RNA 114

5.1 Abstract 114

5.2 Introduction 114

5.3 Results 116

5.3.1 Discovery of *THOR*, a conserved lncRNA expressed in the testis 116

5.3.2 Transcriptional *THOR* homologues exist in the mouse and zebrafish 117

5.3.3 *THOR* exhibits an evolutionarily conserved expression pattern in normal tissues 118

5.3.4 Expression and functional implication of *THOR* in human cancers 118

5.3.5 Conservation of the *THOR*-IGF2BP1 interaction 120

5.3.6 *THOR* regulates IGF2BP1's target mRNAs levels 122

5.3.7 *THOR* knockout zebrafish exhibit fertilization defects and resistance to melanoma formation 124

5.3.8 Human *THOR* enhances the onset of melanoma in zebrafish 126

5.4 Discussion 127

5.5 Methods 129

5.5.1 RNA-seq data processing 129

5.5.2 Conservation analysis 129

5.5.3 Cell lines 130

5.5.4 Tissue Samples **130**
5.5.5 RNA isolation and cDNA synthesis **131**
5.5.6 Quantitative Real-time PCR **131**
5.5.7 RNA-ligase-mediated Rapid Amplification of cDNA Ends (RACE) **132**
5.5.7 siRNA Knockdown Experiments **132**
5.5.8 Overexpression Studies **132**
5.5.9 Cell Proliferation Assays **133**
5.5.10 Murine Subcutaneous *In Vivo* Models **133**
5.5.11 RNA *In Situ H*ybridization in Testis **133**
5.5.12 RNA *In Situ H*ybridization in Testis **134**
5.5.13 Western Blot Analysis **134**
5.5.14 RNA Immunoprecipitation (RIP) Assay **135**
5.5.15 RNA Pulldown Assay **135**
5.5.16 Mass Spectrometry **135**
5.5.17 MS Data Analysis **136**
5.5.18 RNA-protein Interaction Assay **137**
5.5.19 RNA Endogenous Degradation Assay **137**
5.5.20 Anchorage-independent soft agar colony formation Assay **137**
5.5.21 RNA-seq Data Processing **138**
5.5.22 RNA-Seq Differential Expression Testing **138**
5.5.23 Gene signature GSEA analysis **138**
5.5.24 Zebrafish Maintenance **139**
5.5.25 Cas9 Target Site Design, Vector Construction and *in vitro* RNA Synthesis **139**
5.5.26 Microinjection of Zebrafish Embryos **139**
5.5.27 DNA Isolation and PCR Analysis for Identifying Deletion **139**
5.5.28 Germ cell sorting **140**
5.5.29 Generation of THOR Knockout Zebrafish **141**
5.5.30 Zebrafish Mosaic Melanoma Model **141**

5.6 Acknowledgements 142
5.7 Author Contributions 143
CHAPTER 6 Concluding Remarks and Future Directions 155
6.1 RNA-seq driven transcript discovery 155
6.2 Transcriptome meta-assembly 156
6.3 Functional interrogation of lncRNAs 157
APPENDICES 159
LITERATURE CITED 223

LIST OF TABLES

Table 2.1 Summary of lineage and/or cancer- specific lncRNAs nominated 47

Table A.1 RNA-seq datasets 180

Table A.2 Software packages 181

Table A.3 Confidence tiers 182

Table A.4 Validation information 183

Table A.5 SSEA Statistics 184

Table A.6 MiTranscriptome Overlap With Intergenic Cancer GWAS SNPs 185

Table A.7 Reported Functional lncRNAs 186

Table C.1 RACE-determined isofrom sequences for *DSCAM-AS1* 209

Table C.2 Oligonucleotide Primers 210

Table C.3 sh/siRNA Sequence Used 211

Table C.4 Antibodies Used 211

LIST OF FIGURES

Figure 2.1 ***Ab initio*** **transcriptome assembly reveals an expansive landscape of human transcription 40**

Figure 2.2 **Characterization of the MiTranscriptome assembly..... 41**

Figure 2.3 **Analysis of conservation in lncRNAs..... 42**

Figure 2.4 **Methodology for discovering cancer-associated lncRNAs..... 43**

Figure 2.5 **Discovery of lineage-associated and cancer-associated lncRNAs in the MiTranscriptome compendium..... 45**

Figure 3.1 **Schematic of meta-assembly workflow for TACO..... 72**

Figure 3.2 **Assessment of TACO performance..... 73**

Figure 3.3 **Examples of TACO performance..... 75**

Figure 4.1 **Identification of ER and breast cancer associated lncRNAs 106**

Figure 4.2 **Characterization of *DSCAM-AS1*.....108**

Figure 4.3 ***DSCAM-AS1*** **is implicated in cancer aggression clinically and in cell lines 109**

Figure 4.4 **Physical and functional relationship of *DSCAM-AS1* with hnRNPL ..111**

Figure 4.5 ***DSCAM-AS1*** **is implicated in tamoxifen resistance 113**

Figure 5.1 ***THOR*** **is a conserved testis specific lncRNA.....144**

Figure 5.2 ***THOR*** **is expressed in cancers and potentiates tumorigenesis.....146**

Figure 5.3 **Conserved interaction of *THOR* and IGF2BP1 .....148**

Figure 5.4 **Interrogation of the functional relationship of *THOR* and IGF2BP1...150**

Figure 5.5 **Shared transcriptional regulation by *THOR* and *IGF2BP1* 151**

Figure 5.6 ***THOR*** **regulates melanoma onset in zebrafish.....152**

Figure 5.7 **Summary Schematic 154**

Figure A.1 **Curation and processing of samples in the MiTranscriptome compendium.....159**

Figure A.2 Transfrag filtering 161
Figure A.3 Meta-assembly 163
Figure A.4 Characterization of unannotated transcripts 165
Figure A.5 MiTranscriptome characterization 166
Figure A.6 Validation of lncRNA transcripts 168
Figure A.7 Validation of lncRNA transcripts, continued 170
Figure A.8 Classification of transcripts of unknown coding potential 172
Figure A.9 Enrichment of MiTranscriptome assembly for disease-associated regions 174
Figure A.10 Discovery of lineage-associated and cancer-associated transcripts 176
Figure A.11 Lineage-specific and cancer-specific transcripts 178
Figure A.12 Examples of cancer and/or lineage associated transcripts 179
Figure B.1 Schematic example of TACO splice and path graph generation 187
Figure B.2 Schematic depicting the change point detection method 188
Figure B.3 Assessment of TACO performance at different isoform fraction cutoffs 189
Figure B.4 Assessment of TACO performance using long-read sequencing reference 190
Figure B.5 Assessment of TACO performance for the highest expressed transcripts 191
Figure B.6 Assessment of read-through transcription 192
Figure B.7 Assessment of expression filtering on meta-assembly performance 193
Figure B.8 Change point detection parameters 194
Figure C.1 Discovery of basal specific lncRNAs 195
Figure C.2 Characterization of *DSCAM-AS1* 197
Figure C.3 *DSCAM-AS1* clinical associations 199
Figure C.4 *DSCAM-AS1* knockdown and overexpression mediate cancer phenotype 201
Figure C.5 HnNRPL binds to and phenocopies *DSCAM-AS1* 203
Figure C.6 Binding of hnRNPL localized to region at 3′ end of *DSCAM-AS1* 204

Figure C.7 ***DSCAM-AS1*** **and hnRNPL mediate estrogen independent growth in breast cancer cells 206**

Figure C.8 Representative full Western blot images presented in Figure 4.4c,g ... 208

Figure D.1 Conservation of protein-coding genes and expression of CRNDE in normal tissue 212

Figure D.2 Genomic organization of the long non-coding RNA ***THOR*** **..... 213**

Figure D.3 RACE verified ***THOR*** **sequence 214**

Figure D.4 ***THOR*** **knockdown efficiency and proliferation effects 216**

Figure D.5 ***THOR*** **cellular localization 217**

Figure D.6 Expression of ***IGF2BP1*** **and** ***IGF2BP3*** **in cancer and normal tissue 218**

Figure D.7 ***THOR*** **interaction with** ***IGF2BP*** **proteins 219**

Figure D.8 Comparing GSEA performance for knockdown of ***THOR***, ***IGF2BP1***, **and** ***HUR*** **..... 220**

Figure D.9 ***THOR*** **genetic model in zebrafish contributes to melanoma development 221**

Figure D.10 Interrogation of developmental sperm stages 222

LIST OF APPENDICES

APPENDIX A **Supplemental Data for Chapter 2** **159**
APPENDIX B **Supplemental Data for Chapter 3** **187**
APPENDIX C **Supplemental Data for Chapter 4** **195**
APPENDIX D **Supplemental Data for Chapter 5** **212**

LIST OF ABBREVIATIONS

CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary DNA
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
DNA	Deoxyribonucleic acid
ER	Estrogen Receptor
ENCODE	Encyclopedia of DNA Elements
FPKM	Fragments per Kilobase Transcript per Million Reads
GSEA	Gene Set Enrichment Analysis
hg19	Human Genome revision 19
HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts
ICGC	International Cancer Genome Consortium
IGV	Integrated Genomics Viewer
lncRNA	Long Non-coding RNA
MS	Mass Spectrometry
PCA3	Prostate Cancer Associated 3
PCR	Polymerase Chain Reaction
PSA	Prostate-Specific Antigen
RACE	Rapid Amplification of cDNA Ends
RNA-Seq	RNA Sequencing
SChLAP1	Second Chromosome Locus Associated with Prostate-1
SSEA	Sample Set Enrichment Analysis
STAR	Spliced Transcripts Alignment to a Reference
TACO	Transcriptome Assemblies Combined into One
TCGA	The Cancer Genome Atlas
THOR	Testis-associated Highly-conserved Oncogenic lncRNA
UTR	Untranslated Region

ABSTRACT

Cancer is a leading cause of death worldwide, and its etiology and mechanistic underpinnings have been the target of extensive investigation. As a disease resulting from genetic aberrations, extensive effort has been put forth to understand the genomic landscape of cancer. Recently, large multi-institutional consortia have genomically profiled thousands of tumors from a myriad of different cancer types. Extensive investigation of the cancer transcriptome bears the potential to expand our understanding of the potential players in oncogenesis, and thus the goal of this dissertation was to leverage large-scale RNA-seq datasets to discovery novel transcripts, and to subsequently investigate the function of two lncRNA candidates.

In chapter two, we present the MiTranscriptome, the most comprehensive annotation of the human transcriptome to date. Through a large bioinformatics effort, we curated over 6,500 RNA-seq samples from a myriad of tumor types and tissues, and employed a meta-assembly approach to generate a consensus transcriptome we termed MiTranscriptome. Many of these lncRNAs exhibited cancer and tissue specific expression patterns, and these findings have been made available in an online resource (mitranscriptome.org).

In chapter three, we describe TACO, a novel tool designed to produce highly accurate and robust transcriptome meta-assemblies from large numbers of RNA-seq libraries. This tool dramatically outperforms its competitors in its ability to accurately recapitulate the splicing patterns and transcription start and end sites of meta-

assembled transcripts. TACO enables generation of a robust consensus transcriptome from large quantities of RNA-seq samples, which is critical for the identification of novel transcripts from RNA-seq data.

In chapter four, we present a study interrogating the transcriptional landscape of breast cancer, specifically focusing on discovery of novel transcripts associated with estrogen receptor (ER) driven disease. We identify DSCAM-AS1 as an ER-regulated lncRNA, and investigate its role in cancer progression of ER-positive breast tumors, implicating it with the development of tamoxifen resistance.

Finally, in chapter five, we present a study functionally investigating the lncRNA THOR, which has a cancer/testis expression pattern, and is also highly conserved with analogous isoforms in the mouse and zebrafish. We show that THOR functions through an interaction with the protein IGF2BP1, and generate zebrafish animal models to show that THOR plays a critical role in fertility and oncogenesis.

CHAPTER 1

Introduction

1.1 Cancer Genomics

Cancers are the second leading cause of death in the United States after heart disease, with over 1.2 million new cases and 600,000 cancer-related deaths estimated to occur in 2016 (i.e. greater than one in four deaths in America will be caused by cancers)[1]. Since the late nineteenth and early twentieth century, cancer has been known to be a disease of the genome. Every cancer is known to be the result of an aberration in the DNA sequence of the genomes of cancer cells[2]. Consequently, the focus of cancer research has largely been to identify the genomic abnormalities giving rise to tumors[3]. Many of the landmark discoveries in the field of cancer genomics have utilized recurrent somatic mutations[4] and copy number alterations[5] to distinguish "driver" aberrations that are key players in oncogenesis from "passenger" aberrations. With recent advancements in high-throughput genomics technologies, it is now possible to study DNA and RNA sequence and epigenetic markers (e.g. methylation patterns) on a large scale and at a rapid rate[6]. Multiple large consortia have been collaborating to analyze and make publicly available data for a vast number of tumors of many subtypes, with The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) serving as leaders in this endeavor. Since their first comprehensive cancer genomics study of glioblastoma[7], the TCGA have utilized their data from thousands of patient samples to identify many novel driver genes, pathways, and mutations implicated in many individual cancer types[7–20]. Recently, mutational landscapes and copy number

alteration status at a pan-cancer level have been studied[21–24]. While much focus has been on genomic structural variants, these consortia have generated an abundance of RNA-sequencing (RNA-seq) data that provide an opportunity to look further into transcriptional etiologies of cancer that may be missed in standard genomics analysis workflows.

1.2 Power of RNA-seq in cancer genomics

Studying the transcriptional output of cells provides key insight into underlying genetic changes in cancers and the mechanisms by which these changes contribute to carcinogenesis. Much progress has been made studying somatic mutations[4], copy number aberrations[5,25], and rearrangements/fusions[26] at a DNA level, but studying the complete set of transcripts in a cell (i.e. the transcriptome) provides specific insight into a myriad of functional molecules of the genome that may be playing a role in cancer biology. Profiling transcription in cells at a high-throughput scale has become possible with the advent of RNA-Seq[27], which provides a modality to study all cellular RNA, including unannotated transcripts that may or may not have the capacity to generate protein. Unlike methods confined to monitoring predefined molecular lesions, RNA-seq is uniquely capable of capturing the all the cellular RNA products in an unbiased manner. Thus, RNA-seq is capable of discovery of novel transcriptional elements and it has vastly expanded our understanding of the functional elements on the human genome[28–30].

1.3 History of the discovery of non-coding RNAs

As stated above, RNA-seq has enabled detection of widespread transcription through the human genome that was previously overlooked. Prior to the sequencing of the

human genome, estimates of the number of genes (defined in this context as DNA elements with the potential to generate functional elements in cells[31]) in the human genome were in the range of 100,000 – 300,000[32,33]. At this time, no technology existed to comprehensively profile RNA in cells, and without knowledge of the genome sequence, scientist simply had no vision on the scale of transcription in cells. Upon completion of Human Genome Project, researchers were now equipped with vision of the DNA sequence, which permitted sequence pattern detection of areas likely to encode protein[34,35]. These analyses suggested a much more conservative estimate for the number of genes in the human genome in the range of 20,000-30,000. These estimates, however, were of the number of DNA elements with the capacity of generating protein in cells, and without the capacity to profile RNA comprehensively, researchers had no vision of the potential for transcription occurring in places outside of regions of canonical protein-coding genes.

The advent of microarray technology provided the first glimpse into transcription in human cells in a high-throughput manner. While microarrays were and continue to be leveraged to perform powerful analyses of transcription in cells, they are limited by the fact that they can only assess expression for RNAs that hybridize to the probes on their microarray chips. Pre-defined genes and probe sequences must be known in order for microarrays to assess expression. Nevertheless, even in the early stages of the technology, some microarray platforms included probes mapping to regions where protein coding genes were not expected to exist. Detection of expression in these probes initiated a piqued interest in the possibility of widespread transcription of the human genome in regions that may not encode protein[36]. However, as microarrays are limited by pre-defined probes, detection of transcription in these non-protein probes was only suggestive of non-protein transcription.

Next-generation sequencing opened the door for genomic profiling at a rapid rate and at a large-scale. In a seminal study, ChIP-seq of various histone marks associated with transcription was leveraged to identify a myriad of transcripts that do not encode protein in the mouse[37]. With a growing knowledge that transcription is occurring in the genome in a fashion that is similar to that of protein-coding mRNAs, the prevalence, abundance, and functional potential of these long non-coding RNAs (lncRNAs) became open-ended and important avenues of research. Finally, with the advent of RNA-seq, vision of the entire transcriptome was provided, which led to the discovery of many thousands of long non-coding RNAs in human genome[28,29]. Returning to the definition of a gene as a union of DNA-derived functional elements in cells necessitates further functional investigation of these novel transcripts for understanding of how they are involved in the mechanistic underpinnings of the cell.

1.4 Functional RNAs belie the Central Dogma

LncRNAs have been broadly defined as non-coding RNAs over ~200 nucleotides long. Many of these molecules have been shown to be essentially identical to protein coding mRNAs, exhibiting RNA polymerase II mediated transcription, splicing, 5' capping, and polyadenylation[38]. Following their emergence, much work has been done to show that lncRNAs do in fact function as crucial components of human biology, however, given the youth of the field, their role in cellular and molecular biology and the mechanism through which they act are still poorly understood[38,39]. Many recent discoveries have shown that lncRNAs are implicated in development[40,41], cellular function[42], and various pathological systems including cancer[30,39,43–46].

The lncRNA *Xist*, one of the most widely studied lncRNAs, is critical for X chromosome inactivation in women, and mutations in the its promoter have been shown to be associated with can cause X-linked disease including Duchenne muscular dystrophy[47]. Several human lncRNAs have already been shown to be integral players in cancer progression[5]. For example, *HOTAIR* (HOX antisense intergenic RNA) and *ANRIL* (antisense non-coding RNA in the INK4 locus) promote oncogenesis by coordinating Polycomb Repressive Complex 2 (PRC2) mediated transcriptional repression[7,8]. Additionally, *SCHLAP1* (Second Chromosome Locus Associated with Prostate-1) confers prostate cancer aggressiveness by antagonizing the function of the tumor-suppressive SWI/SNF nucleosome remodeling complex[9,10]. These and other compelling anecdotes may represent a vast and underappreciated body of cancer-associated lncRNAs that have the potential to advance our understanding of cancer initiation and progression.

1.5 Discovery and cataloging of lncRNAs

As stated above, RNA-seq provides a means for systematic discovery of transcribed elements, and has already led to a deeper appreciation of the intricate nature of transcription by facilitating discovery of lncRNAs both located in intergenic 'gene deserts' and overlapping protein-coding loci[4]. The aligned sequence data generated by RNA-Seq experiments can be used to predict full-length transcripts *in silico* by a procedure called *ab initio* transcriptome assembly[11,12]. *Ab initio* assembly provides an unbiased modality for gene discovery. Despite efforts to catalog human lncRNAs in this manner[28], several lines of evidence suggest that our current knowledge of cancer-associated lncRNAs remains inadequate. First, reported discrepancies between independent lncRNA cataloguing efforts suggest that lncRNA annotations are fragmented or incomplete[13]. Second, previous studies largely avoided the annotation of

monoexonic transcripts and intragenic lncRNAs due to the added complexity of transcriptional reconstruction in these regions[14]. Third, the rapid co-evolution of high-throughput sequencing technologies and bioinformatics algorithms now enables more accurate transcript reconstruction compared to previous efforts[11]. Fourth, unbiased cataloguing efforts have thus far been confined to select cell lines, individual cancer types, or relatively small cohorts[9,14]. However, cancers possess highly heterogeneous gene expression patterns and detecting recurrent expression of subtype-specific lncRNAs will likely require analysis of much larger tumor cohorts.

While established efforts to annotate high fidelity gene models by a combination of manual and automated systems rely primarily on cDNA evidence[48–50], several studies using RNA-Seq data have described thousands of novel long RNAs, suggesting that widely used reference gene catalogs are far from complete. This annotation gap has been widened further by the rapid sequencing of thousands of high quality RNA-Seq datasets by consortia such as TCGA, ICGC, ENCODE, the Genotype-Tissue Expression project (GTEx), the Cancer Cell Line Encyclopedia (CCLE). Given the limitations of current reference annotation projects, new emphasis has been placed on developing computational transcriptome assembly methods to utilize these remarkable RNA-Seq datasets for transcript discovery[49].

1.6 Transcriptome reconstruction from short-read RNA-seq

Much of the focus in the field of transcriptome reconstruction has been directed towards the problem of predicting full-length transcripts from the short sequence reads generated by a single RNA-Seq experiment. Approaches to this problem can be classified as genome-guided assembly, in which reads are first aligned to a reference genome before being assembling into transcripts[51,52] or *de novo* assembly, which

reconstructs transcripts without a reference genome[53,54]. Studies assessing the performance of transcript reconstruction methods have shown that while several tools achieve high accuracy at the nucleotide or splice junction level, they do not reliably predict the splicing patterns of full-length transcripts[55]. Barriers to accurate transcriptome reconstruction include sources of biological variability such as intron retentions, alternative polyadenylation, and read-through transcription, as well as noise introduced during the process of library preparation, sequencing, and spliced read alignment.

1.6 Transcriptome meta-assembly

Although transcriptome reconstruction for standalone experiments remains an active area of investigation, it has limited utility for the downstream analyses of transcriptional dynamics across many samples. For this purpose, an essential step is to first derive a consensus transcriptome from multiple input datasets. One approach to generating a multi-sample consensus transcriptome is to concatenate the sequence reads from different RNA-Seq samples together and run an existing genome-guided or *de novo* assembly program on the aggregated reads[53]. This aggregation approach comes at the cost of compounding biological and technical noise and forces the use of a single set of parameters, such as the fragment length distribution, for samples that may have been prepared under varying conditions. Furthermore, assembling an aggregated set of reads incurs significant additional computational and memory overhead, limiting its feasibility for large numbers of samples.

In contrast to aggregation, a scalable approach for deriving a consensus transcriptome from multiple samples is to compute transcriptome assemblies for individual datasets and then merge the assemblies in a hierarchical fashion. We term this merging

procedure “meta-assembly” because in its application it is an “assembly of assemblies”. Two genome-guided meta-assembly methods have been made available as part of RNA-Seq analysis protocols: the Cuffmerge utility within the Cufflinks package[51], and a merge mode within the StringTie software (hereafter denoted ‘StringTie-merge’)[52], however, rigorous performance assessment of these meta-assembly tools has yet to be demonstrated.

In principle, merging RNA-Seq datasets from different cell types would augment sensitivity for detecting known transcripts. However, we demonstrate that in practice the performance of both Cuffmerge and StringTie-merge deteriorates as the number of input datasets increases, suggesting that these meta-assembly methods do not adequately mitigate sources of error in predicted transcripts. As more and more assemblies are merged, the amount of incorrectly predicted transcript structures increases and confounds the meta-assembly process. Incompletely processed RNA and DNA contamination are among the primarily culprits giving rise to spurious transcript prediction. While assembly tools have features in place to overcome these complicating factors, they are far from perfect. Therefore, meta-assembly methodology must be designed to overcome intron retention, aberrantly long 3′ and 5′ ends, and read-through transcripts that concatenate multiple neighboring genes on the same strand. Improvements in this meta-assembly approach will provide robust transcript structure predictions which will facilitate facile downstream analysis and investigation of any novel transcripts.

1.7 Nomination of discovered lncRNAs for functional investigation

Through leveraging the large amount of available RNA-seq data, a comprehensive and extensive human transcriptome can be produced via the meta-assembly process

described above. As we will show in Chapter 2, there are potentially tens-of-thousands of novel transcripts in the human genome, and at this time high-through assessment of their function is not possible. Nevertheless, functional interrogation of these molecules is needed for their assimilation into the accepted body of functional cellular elements. Thus, some prioritization is necessary to identify the most exciting lncRNA candidates for further investigation.

In the context of cancer, we and others have employed a number of prioritization parameters to identify the lncRNA candidates most likely to possess robust cellular function. These measures include cancer specific expression pattern[30] and affiliation with other potential driver-aberrations in cancers. There has been some investigation of lncRNAs that are associated with recurrent copy number alterations in cancer[56]. There have been very few reported studies of recurrently mutated lncRNAs. *MALAT1* is a lncRNA that has been widely studied in the context of cancer, and while it has been reported to harbor a large number of somatic mutations, the functional implications of this are unclear[23]. Another approach has been to pursue a hypothesis-driven, targeted interrogation of lncRNAs associated with some feature of interest. With a prior hypothesis, experiments can be tailored to reveal transcripts associated with various experimental conditions. LncRNAs have been discovered in this way to be associated with various cellular function including genome stability[57] and DNA damage signaling[58]. With a prior interest in hormone-regulated cancers, we similarly set out on a hypothesis-driven approach to lncRNA discovery, interrogating the association of lncRNAs with estrogen status in breast cancers (Chapter 4).

Evolutionary conservation is another metric that can be used to prioritize lncRNAs. While there are differing reports on the degree to which lncRNAs are evolutionarily conserved[28,29,37], there are unequivocally a subpopulation of lncRNAs with evolutionary

conservation that present an interesting class of potentially functional molecules. Multiple studies have interrogated the function of evolutionarily conserved lncRNAs, assessing function in zebrafish and mice[41,59]. The capacity to interrogate the function of these lncRNAs in both human cell line models, and in animal models presents a compelling avenue for functional interrogation. We functionally investigate a highly conserved lncRNA exhibiting a cancer/testis expression pattern in Chapter 5.

1.8 Methods for functional interrogation of lncRNAs

As the lncRNA field is still in its nascent stages, much of our understanding of lncRNA function is anecdotal and the summation of multiple insular studies functionally characterizing a single lncRNA[38]. While it has been difficult to ascertain trends and patterns predictive of lncRNA pattern, there have been identification of recurrent mechanisms through which lncRNAs have been shown to function in cells. LncRNAs have been shown to have a myriad of functions including acting as molecular scaffolds bringing protein complexes together, recruiting various RNAs and/or proteins to genomic regions for epigenetic reprogramming, facilitating chromatin architecture in enhancer regions, and acting as molecular decoys in preventing functional binding interactions between protein, RNA, and DNA[38,39]. A common theme to all of these functional roles is within the binding of lncRNAs with other RNAs, with DNA, and with protein. Thus, a major means through which the function of lncRNAs has been elucidated is in interrogation of lncRNA biding interactions. LncRNA-protein interactions have been interrogated in multiple ways. Addition of *in vitro* biotinylated RNA species followed by pulldown and mass spectrometry has been a method commonly employed to identify the protein binding partners for lncRNAs[60], but this method is limited by the need for exogenous transcription and may not capture the endogenous interactions robustly. Consequently, methods have been developed

wherein nucleic acid probes that have been tagged are designed to tile the RNA of interest are added to cell lysate and subsequently pulled-down for interrogation of the lncRNA binding interactions[61,62]. Employing different crosslinking methods prior to pulldown has been shown to enable profiling of endogenous lncRNA-DNA interactions[61,62], lncRNA-RNA interactions[63], and lncRNA-protein interactions[64,65].

In the context of cancer, assessing the impact of modulating the expression level of lncRNAs on various cancer phenotypes has become a common method to ascertain oncogenic relevance. A myriad of cancer assays can be used as a functional readout including proliferation assays, invasion assays, and soft agar colony formation. Additionally, implantation of cell lines in mouse xenograft models has also been leveraged as a more robust measure of oncogenic function for lncRNAs[66]. All of these experiments, however, are dependent upon modifying the expression level of lncRNAs. While standard RNA interference (RNAi) methods including siRNA and shRNA have commonly employed to assess function of lncRNAs, there is significant concern that off target affects could falsely suggest function in an RNA species that is not otherwise functional[66]. A myriad of methods has been employed to minimize off target affects. Antisense oligonucleotides have become an alternative RNAi approach with fewer and orthogonal off target affects to siRNA and shRNA[67]. Additionally CRISPR based approached have been employed, knocking out various regions of the lncRNA genomic locus.

Ultimately, through rigorous functional assessment of lncRNAs in cell line and animal models, we have begun to expand our understanding of the functional players in cells. Thus, much work will be needed to be done in order to continue to discover and functionally investigate the many transcribed lncRNAs in the human genome.

CHAPTER 2

The Landscape of Long Noncoding RNAs in the Human Transcriptome

Citation: Yashar S. Niknafs, Matthew K. Iyer, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R. Barrette, John R. Prensner, Joseph R. Evans, Shuang Zhao, Anton Poliakov**,** Xuhong Cao, Saravana M. Dhanasekaran, Yi-Mi Wu, Dan R. Robinson, David G. Beer, Felix Y. Feng, Hariharan K. Iyer, Arul M. The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nature Genetics* (2015).[68]

This manuscript has been cited over 350 times as of May, 2017. Comprehensive annotation of long non-coding RNAs in the human transcriptome remains an active area of research, and many researchers have utilizsed the MiTranscriptime for further analyses.

2.1 Abstract

Long non-coding RNAs (lncRNAs) are emerging as important regulators of tissue physiology and disease processes including cancer. In order to delineate genome-wide lncRNA expression, we curated 7,256 RNA-Seq libraries from tumors, normal tissues, and cell lines comprising over 43 terabases of sequence from 25 independent studies. We applied *ab initio* assembly methodology to this dataset, yielding a consensus human transcriptome of 91,013 expressed genes. Over 68% (58,648) of genes were classified as lncRNAs, of which 79% (48,952) were previously unannotated. About 1% (597) of the lncRNAs harbored ultraconserved elements and 7% (3,900) overlapped disease-

associated single nucleotide polymorphisms (SNPs). To prioritize lineage-specific, disease-associated lncRNA expression we employed non-parametric differential expression testing and nominated 7,942 lineage- or cancer-associated lncRNA genes. The lncRNA landscape characterized here may shed light into normal biology and cancer pathogenesis, and be valuable for future biomarker development.

2.2 Introduction

Cancers are a leading cause of morbidity and mortality worldwide, with over 14 million new cases and 8 million deaths in 2012[69]. To improve our understanding of cancer pathogenesis, ongoing large-scale efforts led by The Cancer Genome Atlas (TCGA) consortium are using high-throughput molecular profiling strategies to characterize genetic, epigenetic, and transcriptional changes[23,70]. However, efforts to interpret these data have mainly focused on protein-coding genes, despite definitive evidence that transcription of the non-coding genome produces functional RNAs[71]. In particular, long non-coding RNAs (lncRNAs) have been implicated in biological, developmental, and pathological processes and act through mechanisms such as chromatin reprogramming, *cis*-regulation at enhancers, and post-transcriptional regulation of mRNA processing[38,39].

The emergence of high-throughput RNA sequencing (RNA-Seq) technology provides a revolutionary means for systematic discovery of transcriptional units. Indeed, RNA-Seq has led to a deeper appreciation of the intricate nature of transcription by revealing a milieu of lncRNAs both located in intergenic 'gene deserts' and overlapping protein-coding loci[71]. The aligned sequence data generated by RNA-Seq experiments can be used to predict full-length transcripts *in silico* with *ab initio* transcriptome assembly[55,72]. *Ab initio* assembly provides an unbiased modality for gene discovery, and has been successful in pinpointing novel cancer-associated lncRNAs[30]. Here, we utilized a

compendium of 7,256 RNA-Seq libraries to comprehensively interrogate the human transcriptome, identifying 58,648 lncRNA genes. Moreover, we leveraged our dataset to identify myriad lncRNAs associated with 27 tissue and cancer types. By uncovering this expansive landscape of tissue- and cancer-associated lncRNAs, we provide the scientific community a powerful starting point to begin investigating their biological relevance.

2.3 Results

2.3.1 An expanded landscape of human transcription

We attempted to capture the spectrum of human transcriptional diversity by curating 25 independent datasets totaling 7,256 poly-A+ RNA-Seq libraries, including 5,847 from TCGA, 928 from the Michigan Center for Translational Pathology (MCTP), 67 from the Encyclopedia of DNA Elements (ENCODE), and 414 from other public datasets (**Figure A.1a** and **Table A.1**). We developed an automated transcriptome assembly pipeline and employed it to process the raw sequencing datasets into *ab initio* transcriptome assemblies (**Figure A.1b, Table A.2**). This bioinformatics pipeline utilized approximately 1,870 core-months (average 0.26 core-months per library) on high-performance computing environments.

Collectively the RNA-Seq data constituted 493 billion fragments; individual libraries averaged 67.9M total fragments and 55.5M successful alignments to human chromosomes. On average 86% of aligned bases from individual libraries corresponded to annotated RefSeq exons, while the remaining 14% fell within introns or intergenic space[73]. We applied coarse quality control measures to account for variations in sequencing throughput, run quality, and RNA content by removing 753 libraries with

(1) fewer than 20 million total fragments, (2) fewer than 20 million total aligned reads, (3) read length less than 48bp, or (4) fewer than 50% of aligned bases corresponding to RefSeq genes (**Figures A.1c, d**). After coarse filtration, we obtained approximately 391 billion aligned fragments (43.69 terabases of sequence) to use for subsequent analysis. The set of 6,503 libraries passing quality control filters included 6,280 datasets from human tissues and 223 samples from cell lines. Of the tissue libraries, 5,298 originated from primary tumor specimens, 281 from metastases, and 701 from normal or benign adjacent tissues (**Figure A.1e**). We subsequently refer to this set of samples as the MiTranscriptome compendium.

To permit sensitive detection of lineage-specific transcription we partitioned the libraries into 18 cohorts by organ system (**Figure 2.1a**), performed cohort-wise filtering and meta-assembly, before re-merging the data (**Figure 2.1b**). We developed and employed computational methods to filter library-specific background noise and predict the most likely isoforms from the assemblies of transcript fragments (transfrags) (**Figure 2.1b**). Our filtering approach utilized transcript abundance and recurrence information to differentiate robust transcription from incompletely processed RNA or genomic DNA contamination[4]. This stringent approach eliminated the vast majority (>96%) of unannotated transfrags in the compendium (**Figure A.2a-f**). The remaining transfrags were collapsed into full-length transcript predictions using a greedy dynamic programming algorithm (**Figure A.3a,b**). For example, in the chromosome 12 locus containing *HOTAIR* and *HOXC11*, the algorithm consolidated 7,471 raw transfrags into 17 transcripts, including ones that accurately matched annotated *HOTAIR* and *HOXC11* isoforms (**Figure A.3c**). After merging meta-assemblies from 18 organ system cohorts, we established a consensus set of 384,066 predicted transcripts that we designated as the MiTranscriptome assembly (**Figure 2.1b**).

To characterize the MiTranscriptome we compared it to reference catalogs from RefSeq (Dec, 2013)[73], UCSC (Dec, 2013)[74], GENCODE (Release 19)[48], and intergenic lncRNA predictions from the previous cataloguing study by Cabili *et al.*[28]. We observed increases in exons, splice sites, transcripts, and genes of 29%, 52%, 95%, and 57%, respectively, relative to GENCODE, the largest of the reference catalogs (**Figure 2.1c**). In terms of well-annotated genes, the assembly demonstrated high sensitivity at the nucleotide and splice site level, recovering 94% and 93% of RefSeq nucleotides and splice sites, respectively (**Figure A.4a,b**). However, detection of precise RefSeq splicing patterns, an ongoing challenge for *in silico* transcriptome reconstruction methods[55], was just 31%. Unannotated transcripts were defined as lacking strand-specific nucleotide overlap with reference transcripts (RefSeq, UCSC, and GENCODE). While the fraction of transcripts overlapping annotated genes was high in individual cohorts (range 62-88%, mean 75%), the fraction of annotated genes within the entire MiTranscriptome was just 46%, alluding to the presence of much unannotated transcription unique to specific lineages (**Figure A.4c**).

To assess the robustness of the MiTranscriptome we stratified transcripts into confidence tiers based on annotation status, the presence of annotated splice junctions, and mono- or multi-exonic structure (**Table A.3**). Using the empirical cumulative distribution function derived from annotated transcript expression levels, we assigned confidence scores to unannotated transcripts (**Figure A.5a**). Next, we performed qRT-PCR validations of 100 unannotated transcripts (38 mono-exonic, 62 multi-exonic) with modest expression (i.e. FPKM > 1.0) in at least one of the lung, prostate, or breast cancer cell lines A549, LNCaP, or MCF7, respectively. To assess false positives arising from background levels of genomic DNA control reactions without reverse transcriptase were also included. Of the 100 lncRNAs tested, 95 had significantly higher expression in the appropriate cell line relative to control (Student's t-test, p-value < 0.05, **Figure A.6**),

and showed high correlation between qRT-PCR and RNA-Seq expression profiles (**Figure A.7a**). In addition we also performed independent Sanger sequence verification of 18 amplicons that were highly expressed in the three cell lines (**Table A.4**, **Figure A.7b,c**).

2.3.2 Coding potential assessment of long RNA transcripts

To facilitate further study of the assembly we classified transcripts into one of five categories: (1) Protein-coding, (2) Read-through (implying a transcript overlapped multiple separate annotated genes), (3) Pseudogene, (4) lncRNA, and (5) Transcript of Unknown Coding Potential (TUCP) (**Figure A.8a**). The TUCP classification was originally suggested by Cabili *et al*.[28] and pertains to long RNAs with *in silico* evidence of coding potential. The ability to predict coding potential from sequence features alone has important implications for *ab initio* transcript annotation. Here, we predicted TUCPs by incorporating two methods: (1) predictions from the Coding Potential Assessment Tool (CPAT)[75], which analyzes the sequence features of transcript open reading frames (ORFs), and (2) presence of a known Pfam domain[76] within a transcript ORF (**Figure A.8b-h**).

Remarkably, over sixty percent of MiTranscriptome genes were classified as either lncRNAs or TUCPs (59% lncRNAs, 3.5% TUCPs, **Figure 2.2a**). The majority of lncRNAs and TUCPs were unannotated relative to RefSeq, UCSC, and GENCODE genes (79% and 66%, respectively) and located within intergenic regions (72% and 60%, respectively) (**Figure 2.2b**). Interestingly, 5,248 transcripts overlapping annotated lncRNAs were flagged as TUCPs, suggesting that previous annotation attempts identified ostensibly non-coding fragments of transcripts possessing robust ORFs. For

example, in a chromosome 16 intergenic locus we detected transcripts harboring a 418 amino acid ORF containing 29 exons overlapping three independent genes annotated by GENCODE as lncRNAs (LINC00514, LA16c-380H5.3, LA16c-380H5.4), suggesting that some annotated lncRNAs may in fact be inaccurate partial representations of a larger protein-coding gene (**Figure 2.2c**). To further investigate coding potential we searched a large human proteomics dataset derived from benign tissue samples[77] for peptides uniquely mapping to TUCP ORFs and noted 268 such genes. Given these intriguing results we anticipate that future integration of proteomics data from tumor tissues will strengthen our TUCP predictions.

2.3.3 Characterization of long RNAs

LncRNA and TUCP genes tended to have fewer exons than read-through or protein coding genes, but we nevertheless observed appreciable alternative splicing for all classes of transcripts[28,78] (**Figure A.5b**). Furthermore, we observed that lncRNAs and TUCPs were expressed at lower levels than read-through or protein-coding transcripts, which is consistent with previous studies[28,78–80] (**Figure 2.2d**). To further corroborate active transcription of the lncRNAs and TUCPs, we intersected intervals surrounding the TSSs with ECNODE ChipSeq data for histone 3 lysine 4 trimethylation (H3K4me3) ChIP-Seq, RNA polymerase II (PolII) binding sites, and DNase hypersensitivity data from 13 cell lines[81,82]. Maximal enrichment of these marks at their TSSs, but not in randomly shuffled control regions, suggested that the assembled lncRNA and TUCP transcripts possess actively regulated promoters (**Figure 2.2e-g**).

2.3.4 LncRNAs harboring conserved elements

The evolutionary conservation of lncRNAs has been a topic of ongoing conversation, with several reports suggesting that lncRNAs are modestly conserved[28,78,80,83]. In agreement with previous reports we observed increases in both transcript and promoter conservation levels for lncRNAs and TUCPs relative to random control regions (**Figure A.5c-f**). Shifts in the cumulative distributions of lncRNA and TUCP transcripts were greater for annotated transcripts relative to unannotated transcripts. This difference may reflect discovery bias towards highly conserved genes detectable across multiple model systems. Moreover, the subtle conservation increases we observe for lncRNAs suggest, at least in humans, that lncRNA conservation might be an exceptional phenomenon rather than a general one. Therefore, we specifically delineated 3,309 lncRNAs (5.6% of all lncRNAs) harboring markedly higher base-wise conservation relative to random intergenic regions to enable focused study of these transcripts (**Figure 2.3a** and **Figure A.5e**). In addition, an intriguing aspect of the non-coding genome includes ultraconserved elements (UCE), which are stretches of DNA >200nt with nearly perfect sequence identity across multiple organisms[83,84]. We delineated 597 intergenic lncRNAs (1.2% of all intergenic lncRNAs) harboring UCEs and designated these as Highly Conserved Long Intergenic Non-Coding RNAs (HICLINCs) (**Figure A.5h**). For example, *THCAT126*, a previously unannotated intergenic lncRNA on chromosome 2q24, contains elements in its final exons that are conserved in nearly all vertebrates including zebrafish (**Figure 2.3b**). Moreover, *THCAT126* is expressed widely across many tissue types, including thyroid cancer (**Figure 2.3c**). Highly conserved lncRNAs such as *THCAT126* (and other cancer-associated HICLINCs described below) provide an exciting avenue for *in vivo* study of the role of lncRNAs in development and carcinogenesis.

2.3.5 LncRNAs overlapping disease-associated SNPs

To investigate the relationship of the MiTranscriptome assembly with disease-associated regions of the genome, we assessed overlap of transcripts in the assembly with 11,194 unique disease associated single nucleotide polymorphisms (SNPs) from a catalog of genome-wide association studies (GWAS)[85]. MiTranscriptome exons and transcripts overlapped 2,586 and 9,770 GWAS SNPs compared to just 1,096 and 7,050 SNPs overlapping reference transcripts, respectively (**Figure A.9a,b**). Altogether transcripts in the assembly overlapped 2,881 formerly intergenic SNPs located within 'gene deserts', and only lacked 161 GWAS SNPs overlapping annotated genes. We tested for the possibility that the increased overlap with GWAS SNPs occurred at a rate above chance and observed that both MiTranscriptome transcripts and exons were significantly enriched for GWAS SNPs relative to random SNPs chosen from the same chip platform (paired t-test, p-value, 5.25e-135 and 1.15e-199, respectively, **Figure A.9c**). Moreover, unannotated intergenic lncRNAs and TUCPs were also significantly enriched for disease-associated regions, with exons more highly enriched than full-length transcripts (paired t-test, p-values 9.90e-78 and 5.50e-50, for whole transcript and exon, respectively, **Figure A.9d**). These data argue that a rigorous reevaluation of allele-specific gene expression regulation in regions proximal to GWAS SNPs may yield informative biological associations with the new lncRNAs identified in this study.

2.3.6 Differential Expression Analysis

Our large-scale transcriptome reconstruction process unveiled tremendous transcriptional complexity highlighted by the presence of thousands of uncharacterized lncRNAs and TUCPs. To prioritize disease-associated and lineage-specific transcription, we developed a non-parametric method for differential expression testing called Sample Set Enrichment Analysis (SSEA). SSEA adapts the weighted Kolmorgorov-Smirnoff-like tests used by Gene Set Enrichment Analysis (GSEA)[86] to discover

transcript expression changes between two groups of samples. The non-parametric nature of this method permits sensitive detection of differential expression within heterogeneous sample populations (e.g., tumor sub-types). We performed 50 differential expression analyses including various cancer or normal lineage types (i.e., one cancer or lineage type versus all other MiTranscriptome samples), and cancer versus normal comparisons within a single tissue type (**Figure 2.4a**). Collectively, SSEA detected over two million significant associations (FDR < 1e-3 for cancer versus normal analyses and FDR < 1e-7 for lineage analyses) involving 267,726 MiTranscriptome transcripts (**Table A.5**). To validate the enrichment testing approach we assessed its ability to rediscover known biomarkers up-regulated and down-regulated in prostate and breast cancers. We assessed the concordance between the top 1% positively and negatively enriched genes from each cancer type with cancer gene signatures obtained from the Oncomine database of microarray studies[17,87–92]. A heatmap of the odds ratios of the gene signature associations revealed striking agreement between SSEA and the other studies, with SSEA often demonstrating equal or better concordance to each microarray study than comparison between microarray studies (**Figure 2.4b**). Thus, isoform-level differential expression testing from the MiTranscritpome *ab initio* assembly of RNA-Seq data recapitulated the results from cancer microarray gene expression studies, supporting the SSEA method as a viable tool for detection of differential expression.

To further credential the enrichment testing approach, we assessed the ability to detect positive control lncRNAs and protein-coding genes in breast and prostate cancers. For example, SSEA correctly identified the oncogenic lncRNA *HOTAIR*, estrogen receptor 1 (*ESR1*), and GATA binding protein 3 (*GATA3*) as highly positively enriched in breast cancers, and accurately nominated the tumor suppressor lncRNA *MEG3* and the metastasis suppressor *LIFR*[93] as highly negatively enriched[17,46,94,95] (**Figure 2.4c-e**).

Similarly, in prostate cancers SSEA detected differential expression of lncRNAs and protein-coding genes consistent with the literature (**Figure 2.4f**). Notably, the known prostate cancer lncRNAs Prostate Cancer Antigen-3 (*PCA3*) and *SChLAP1* were strikingly enriched in a cancer-specific and prostate-specific manner relative to all other sample set analyses (**Figure 2.3g,h**)[43,89]. Overall the ability of the enrichment testing approach to rediscover known cancer genes in an unbiased fashion indicates its utility for analysis of the cancer association and lineage specificity within the panorama of uncharacterized transcription unveiled by MiTranscriptome.

2.3.7 Characterization of differentially expressed lncRNAs

To extend our study beyond known cancer genes, we mined the enrichment test results for lineage-specific and cancer-specific transcripts in an unbiased manner. Lineage specificity was assayed using sample sets for each cancer or tissue type compared to all other samples in the MiTranscriptome compendium (**Figure 2.4a,** "Cancer Types/Normal Types"), and SSEA results were utilized to determine the degree of enrichment for each transcript in the various cancer and tissue types. Unsupervised clustering of transcript percentile ranks for the top 1% of transcripts in each lineage demonstrated distinct lineage specific signatures while also suggesting relationships among lineages and between cancer and normal sets from the same lineage (**Figure A.10a**). Examples of closely related lineage clusters include blood cancers (acute myeloid leukemia (AML), chronic myeloid leukemia (CML), and myeloproliferative neoplasia (MPN)), brain cancers (lower grade glioma (LGG) and glioblastome multiforme (GBM)), and muscle tissue (cardiac and skeletal). Additionally, a cluster comprising cervical cancer, head and neck cancer and normal lineages, lung squamous cell cancer, and bladder cancer emerged and suggested that primarily squamous (and transitional) cell carcinomas from distant primary sites share important gene expression

relationships. Intriguingly, unsupervised clustering of only the lncRNAs in the top 1% of the SSEA analysis for lineage association recapitulated all of these relationships, indicating the capacity for lncRNAs to independently identify cancer and normal lineages (**Figure 2.5a**).

Next, we investigated the dimension of cancer-specific transcriptional dynamics in twelve tissues with ample numbers of both cancer and normal samples (**Figure 2.4a,** "Cancer vs. Normal"). Similar to above, unsupervised clustering of the top 1% cancer-associated lncRNAs demonstrated highly specific signatures for each cancer type, with the exception of lung and kidney cancers (**Figure 2.5b** and **Figure A.10b**). Lung squamous cell carcinomas (LUSC) and adenocarcinomas (LUAD) clustered together and shared numerous transcripts with cancer association. Similarly, renal clear cell (KIRC) and papillary cell (KIRP) carcinomas exhibited highly overlapping signatures, while renal chromophobe carcinomas (KICH) remained distinct from KIRC and KIRP. Finally, we intersected results from lineage and cancer analyses. With extensive further evaluation, such transcripts may have translational potential for use in non-invasive clinical tests, particularly for cancers that lack reliable biomarkers. Notable examples included the prostate-specific lncRNAs *PCA3* and *SChLAP1* presented earlier (**Figure 2.4g,h**). A myriad of lncRNAs were detected as being lineage and cancer associated (i.e. in the top 5% of both analyses) for each of the cancer types analyzed (**Figure 2.5c** and **Figure A.11a**). A direct comparison of lncRNAs and protein-coding transcripts revealed that both annotated and unannotated lncRNAs have the potential to perform at a comparable level to protein-coding genes, supporting a role for lncRNAs in augmenting tissue and cancer specificity gene signatures (**Figure 2.5d** and **Figure A.11b,c**).

We applied stringent statistical cutoffs to nominate 7,942 lncRNA or TUCP genes (11,478 transcripts) with as cancer associated, lineage associated, or both. Transcripts

meeting the stringent cutoffs in the cancer versus normal analyses were designated as having "cancer association". Those transcripts meeting stringent cutoffs for lineage specificity in non-cancerous tissue (e.g. heart, skeletal muscle, embryonic stem cells) and in cancers lacking RNA-Seq data for benign tissue were designated as "lineage associated". Moreover, transcripts meeting the cutoffs for both the cancer versus normal and lineage specificity analyses were designated as having "cancer and lineage association" (**Table 2.1).** Transcripts with significant association in just one tissue type were given names according to that tissue type (**Table 2.1**), and transcripts with associations in multiple tissues were named "Cancer Associated Transcripts" (CATs). An additional 545 lncRNA genes (1634 transcripts) that possessed ultraconserved elements but did not meet the stringent lineage and cancer association criteria were designated as HICLINCs (Highly Conserved Long Intergenic Non-Coding RNA). Of these 8,487 lncRNAs, 7,804 did not possess an official gene symbol according to the HUGO Gene Nomenclature Committee[93], and were thus named according to the convention described in **Table 2.1**.

To infer putative roles for cancer or lineage associated lncRNAs in oncogenesis, we curated 2,078 MSigDB gene sets into categories corresponding to biological function (angiogenesis/hypoxia, metastasis, proliferation/cell-cycle, cell adhesion, DNA damage/repair) or signatures from gene expression profiling[86]. We constructed an expression correlation matrix between lncRNAs and protein-coding genes and employed a "guilt by association" analysis whereby the correlation data was processed by GSEA to generate a matrix of the association of each lncRNA with each gene set, capturing over 14,000 transcripts with significant associations[37].
To allow the scientific community to explore our discoveries, we developed an online portal featuring detailed characteristics of the nominated transcripts (mitranscriptome.org), and present several examples of intriguing lncRNAs here. First,

the lncRNA Breast Cancer Associated Transcript-49 (*BRCAT49*) is a breast cancer- and lineage-associated lncRNA (**Figure 2.5d**) located ~45kb downstream of the intergenic breast cancer SNP rs13387042 that has been implicated by multiple GWAS studies (**Figure 2.5e,f**)[96–100]. *BRCAT49* provides a possible target for explaining the breast cancer association of this genomic region, and would be a candidate for intergenic expression quantitative trait loci (eQTL) analysis. Further interrogation of the relationship with GWAS SNPs was also performed, and all transcripts within 50kb of a GWAS SNP implicated in a disease site for which the lncRNAs was identified as having a significant association are reported in **Table A.6.** Second, the lncRNA we termed Melanoma Associated Transcript-6 (*MEAT6*) was found to be in the 99.8th percentile in the melanoma lineage SSEA analysis (**Figure 2.5a**). Genomic investigation delineated *MEAT6* as a partially annotated transcriptional variant of the lncRNA *AK090788* on chromosome 6q26 (**Figure A.12a**). However, *MEAT6* utilizes an alternative start site and upstream exons absent from reference catalogs. Expression of *MEAT6* isoforms using the novel start site were highly specific to the melanoma samples in the MiTranscriptome cohort (**Figure 2.5g**); however, isoforms lacking the *MEAT6* start site had a dramatically different pan-cancer expression profile with almost no expression in melanoma (**Figure A.12b**). Additional examples of expression profiles for cancer- or lineage-specific lncRNAs in other tissue types are displayed in **Figure A.12c,d**. The examples shown here are indeed representative, and we anticipate that an abundance of uncharacterized transcription with biological and translational potential can be leveraged using our discoveries here and our online resource.

2.4 Discussion

Here, we discovered and characterized an expanded landscape of transcription via unbiased transcriptome reconstruction from thousands of tumors, normal tissues, and

cell lines. Our work utilizes several orders-of-magnitude more RNA-seq data (~100 fold) than previous RNA-seq lncRNA discovery efforts and vastly increases the universe of known transcripts in both normal tissues and cancer. The unprecedented breadth (6,503 samples) and depth (>43 Terabases of sequence) of our compendia enabled sensitive detection of robust transcription and specific filtration of background noise. The lncRNAs in our assembly (58,648 genes, often with multiple isoforms) far outnumber entries in current lncRNA databases (<16,000 genes), implying that reference transcript annotations may be fragmented or otherwise incomplete[28,78,101–104]. Moreover, our assembly indicates that the genomic diversity of lncRNAs eclipses coding transcripts (i.e., nearly 60,000 lncRNA genes versus approximately 30,000 protein coding genes), a disparity that may grow as additional diseases and cell types are sequenced and more lncRNAs are discovered.

Multiple lines of *in silico* evidence support the biological and functional relevance of MiTranscriptome transcripts, including robust expression, protein-coding potential (for TUCPs), high conservation, active regulation at promoters, proximity to disease-associated genomic polymorphisms, correlation with protein-coding gene signatures, lineage-specificity, and cancer-specificity. Moreover, many lncRNAs independently identified by this study have been previously validated and mechanistically linked to carcinogenesis (**Table A.7**)[43,46,79,105,106]. Regardless of their functional contributions, uncharacterized MiTranscriptome transcripts could serve as future cancer biomarkers.

Although the central dogma remains a core tenet of cellular and molecular biology, the appreciation of lncRNAs as functional genomic elements that defy the central dogma may be essential for fully understanding biology and disease. Taken together, our results indicate that the vastness and complexity of lncRNA transcription has been grossly underappreciated, and that myriad lncRNAs are associated with carcinogenesis.

We anticipate that the MiTranscriptome assembly and lncRNAs identified by this study, as well as the computational tools developed herein, will provide a foundation for lncRNA genomics, biomarker development, and the delineation of cancer disease mechanisms.

2.5 Methods

2.5.1 High performance computing

Computational analysis was performed using the Flux high-performance computer cluster hosted by the Advanced Research Computing (ARC) at the University of Michigan.

2.5.2 RNA-seq data processing

A comprehensive RNA-Seq analysis pipeline was employed on all samples (**Figure A.1b**). The analysis pipeline provided sequence quality metrics, filtering of contaminant reads, fragment size estimation, strand-specific library type estimation, spliced alignment of reads to the human reference genome (version hg19/GRCh37), alignment performance metrics, generation of visualization tracks for genome browsers, and *ab initio* transcript assembly. The third-party tools used to process RNA-Seq data were selected based on computational performance, ease-of-use, user and community support, and experience (**Table A.2**).

2.5.3 Overview of transcriptome reconstruction

To merge *ab initio* assembled transcript fragments (transfrags) into a consensus transcriptome we developed and utilized a bioinformatics method that (1) classifies and filters sources of background noise in individual libraries and (2) reassembles transfrags weighted by their expression levels from multiple libraries into a consensus transcriptome.

2.5.4 Filtration of noise contamination

We controlled for alignment artifacts and poorly assembled transcripts by clipping very short first or last exons (< 15bp) and excluding short transfrags (≤ 250bp). We removed noise due genomic DNA contamination and incompletely processed RNA using a machine learning method. The method models the empirical distributions of relative transcript abundance and recurrence (number of independent samples in which the transcript was observed). From this model the method determines optimal library-specific thresholds for distinguishing annotated from unannotated transcription as a proxy for signal versus background noise, respectively.

2.5.5 Transcriptome meta-assembly

We created directed acyclic splicing graphs where nodes in the graph reflected contiguous exonic regions and edges corresponded to splicing possibilities (**Figure A.3a**). Nodes in the splicing graph with relatively low abundance were then pruned. We then incorporated partial path information inherent in transfrags spanning multiple exons by building splicing pattern graphs that subsumed the original splice graphs (**Figure A.3b**). The splicing pattern graph is a type of *De Bruijn* graph where each node represents a contiguous path of length *k* through the splice graph, and edges connect paths with *k-1* nodes in common. The algorithm finds and reports a set of highly

abundant transcripts by iteratively traversing the graph using dynamic programming in a greedy fashion.

2.5.6 Merging of meta-assemblies

To merge meta-assemblies from 18 cohorts we used the Cuffmerge tool[51], which produced a final transcriptome GTF file.

2.5.7 Comparisons of MiTranscriptome with reference catalogs

The exons, splice sites, and splicing patterns of all assembled transcripts were compared to RefSeq, UCSC, GENCODE (version 19), and the merged union of all three reference catalogs using custom python scripts. Sensitivity and precision values were computed using the number of shared strand-specific transcribed bases, introns, and splicing patterns. Precision was also computed for the subset of *ab initio* transcripts that overlapped any part of a reference transcript. Transcripts that overlapped a reference transcript on the same strand were designated annotated. When an *ab initio* transcript matched multiple reference transcripts, a best match was chosen using the following criteria: (1) matching splicing pattern, (2) fraction of shared introns, and (3) fraction of shared transcribed bases. The biotype (protein, read-through, pseudogene, or lncRNA) for annotated transcripts was imputed from the best matching reference transcript. Annotated lncRNAs and unannotated transcripts were reclassified as either lncRNAs or TUCPs.

2.5.8 Prediction of transcripts of unknown coding potential (TUCP)

We predicted coding potential by integrating two sources of evidence: (1) predictions from the alignment-free Coding Potential Assessment Tool (CPAT)[75] and (2) searches for Pfam 27.0 matches[76]. CPAT determines the coding probability of transcript sequences using a logistic regression model built from ORF size, Fickett TESTCODE statistic[107], and hexamer usage bias. We chose a CPAT probability cutoff by repeatedly randomly sampling 100,000 each of putative non-coding and protein-coding transcripts and optimizing on the balanced accuracy (average of sensitivity and specificity) metric (**Figure A.8b,c**). The average area-under-the-curve (AUC) across 100 iterations was 0.9310 (minimum 0.9302, maximum 0.9320), and the average optimal probability cutoff was 0.5242 (minimum 0.5090, maximum 0.5482). This cutoff value achieved accurate discrimination of lncRNAs and protein-coding genes (sensitivity: 0.84, specificity: 0.95, FDR: 0.076). Of the putative non-coding transcripts 9,903 (5.3%) exceeded the CPAT cutoff and met the criteria for TUCP. As additional evidence of coding potential we scanned all transcripts for Pfam A or B domains across the three translated reading frames for stranded transcripts and six frames for monoexonic transcripts of unknown strand. We designated putative non-coding transcripts with either a Pfam domain or a positive CPAT prediction as TUCP.

2.5.9 Proteomics analysis

We obtained the following Thermo files (in the RAW format) from a recent study mapping the human proteome[77]: Adult_Kidney_Gel_Elite_55, Adult_Liver_Gel_Elite_56, Adult_Pancreas_Gel_Elite_60, Adult_Rectum_Gel_Elite_63, Adult_Urinarybladder_Gel_Elite_40, Fetal_Brain_Gel_Velos_16, Adult_Lung_Gel_Elite_56, and Adult_Prostate_Gel_Elite_62. The Thermo files were transformed into mzXML using MSConverter[108] and interrogated against human UniProt database V.15.11 using the X!tandem search engine. The database was

concatenated with all possible open reading frames longer than 7 amino acids from the lncRNAs and with reversed sequences for determination of false discovery rate (FDR). The X!Tandem search parameters were: fully tryptic cleavage, parent mass error 5 ppm, fragment mass error 0.5 Da, 2 allowed missed cleavages. Fixed modifications: Cys carbamidomethylation. Variable modifications: Met oxidation. X!Tandem output files were processed by PeptideProphet and ProteinProphet. Data was filtered at peptide probability 0.5 and protein probability 0.9 to ensure protein FDR < 1%.

2.5.10 Confidence scoring system

After assembly of the MiTranscriptome, transcripts were subjected to an additional confidence evaluation. lncRNAs in the MiTranscriptome were categorized into tiers based on their annotation status and the degree of matching of splice junctions to the reference annotation (**Table A.3**). Tier 1 transcripts are all annotated and tier 2 transcripts are unannotated. An empirical cumulative distribution function (eCDF) was developed by profiling the second highest expression value (across all 6,503 samples) for each tier 1 transcript. The second highest value was used to control for outlier expression. The eCDF was using to compute confidence scores for tier 2 transcripts using the same expression summary statistic.

2.5.11 Validation of lncRNA transcript by qRT-PCR

We chose 150 lncRNAs with at least 1 FPKM expression in either A549, LNCaP, or MCF7 cells for biological validation. For each transcript, primer pairs were designed using the Primer-BLAST tool[109]. Primer pairs with the following parameters were selected: (1) amplicon length between 80-140 bp (2) primer GC content between 35-65%, and (3) primer length greater than 20 bp. Primers were blasted against the human

genome to ensure specificity to our target gene, and primers designed against multi-exonic transcripts spanned exon junctions. Regions of any transcript that directly overlapped an exon on the antisense strand were avoided. Primer pairs meeting these criteria could be designed for 100 out of 150 lncRNAs (38 monoexonic and 62 multiexonic). All oligonucleotide primers were obtained from Integrated DNA Technologies (Coralville, IA) and are listed in **Table A.4**.

RNA was isolated from A549, LNCaP and MCF7 cells in Trizol (Invitrogen) using the RNeasy Mini Kit (Qiagen). Equal amount of RNA was converted into cDNA using random primers and the Superscript III reverse transcription system (Invitrogen). Quantitative real-time PCR (qPCR) was performed using Power SYBR Green Mastermix (Applied Biosystems, Foster City, CA) on an Applied Biosystems 7900HT Real-Time PCR System. The housekeeping genes, *CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3* were used as loading controls[55]. Data was normalized first to housekeeping genes and then to the median value of all samples using the delta-delta Ct method and plotted as fold change over median. To ensure the specificity of the primers, 20 amplicons were further analyzed by Sanger sequencing.

2.5.12 Cell lines and reagents

All cell lines were obtained from the American Type Culture Collection (Manassas, VA). Cell lines were maintained using standard conditions. Specifically, A549 were grown in F-12K plus 10% fetal bovine serum (FBS), LNCaP in RMPI1640 (Invitrogen) plus 10% FBS and 1% penicillin-streptomycin, and MCF7 in Eagle's Minimum Essential Media (EMEM) plus 10% FBS. All of the cell lines were grown at 37°C degrees in a 5% CO_2 cell culture incubator. To ensure identity, cell lines were genotyped at the University of Michigan Sequencing Core using Profiler Plus (Applied Biosystems) and

compared with the short tandem repeat (STR) profiles of respective cell lines available in the STR Profile Database (ATCC). All of the cell lines were routinely tested and found to be free of *Mycoplasma* contamination.

2.5.13 Evidence for active regulation of transcriptional start sites

To conduct analysis of TSS intervals ENCODE project datasets were downloaded from the UCSC Genome Browser[74]. For H3K4me3 analysis we used the Encode Project Broad Institute H3K4me3 ChIP-Seq peaks for cell lines GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HSMMtube, HUVEC, K562, NH-A, NHDF-Ad, NHEK, and NHLF[56]. For RNA polymerase II analysis we used POL2RA binding sites from the Encode Project Uniform TFBS master file version 3 for any of the cell lines with H3k4me3 data[81]. Finally, for the DNase hypersensitivity analysis the Encode Project combined UW and Duke DNaseI hypersensitivity regions were downloaded as a master file from EMBL-EBI, and filtered for any of the cell lines with H3k4me3 data. Peak enrichment files (BED format) were aggregated across all cell lines. Intervals of +/- 10 kilobases surrounding unique MiTranscriptome TSSs were generated using BEDTools 'slop' tool[110]. To control for expression, TSSs were filtered for transcripts not expressed in any of the cell lines (FPKM < 0.1). Basewise peak coverage was generated for each TSS interval using the BEDTools 'coverage' function and summarized across subsets of TSSs. Summed per-base coverage histograms were normalized by dividing by the number of expressed TSSs.

2.5.14 Conservation analysis

The evolutionary conservation of transcripts in our assembly was studied using two metrics: (1) the fraction of significantly conserved bases ($p \leq 0.01$, phyloP algorithm),

and (2) the maximally conserved 200nt sliding window (phastCons scores averaged within each window). The former captures independently conserved elements within a transcript regardless of position, and the latter captures contiguous regions of high conservation. The 200nt sliding window size was chosen to aid in discovery of putative ultraconserved elements[84]. As a negative control we measured the conservation of non-transcribed regions using these metrics by randomly sampling contiguous length-matched intervals from intergenic and intronic space. Non-transcribed interval sampling was restricted to regions with valid 46-way conservation data.

The fractional basewise conservation and contiguous window conservation metrics were used to nominate highly conserved and ultraconserved transcripts, respectively. In both cases cutoffs for significant transcripts were determined by controlling the rate of observing elements with similar conservation levels within non-transcribed intergenic space at a level of 0.01. For fractional basewise conservation a score of 0.0947 (9.5% of transcript bases conserved at phyloP p-value < 0.01) corresponded to a false discovery rate < 0.01. At this cutoff the sensitivity for detecting protein-coding transcripts was 0.67. For contiguous sliding window conservation an average PhastCons probability of 0.9986 corresponded to a false discovery rate < 0.01. At this cutoff the sensitivity for detecting true positive ultraconserved non-coding elements downloaded from UCNEbase was 0.69[111]. Applying these criteria to our assembly yielded 6,034 lncRNAs (3.4%) and 541 TUCPs (4.7%) with significant basewise conservation levels. Additionally, 1,686 lncRNAs (0.96%) and 121 TUCPs (0.01%) harbored contiguous ultraconserved regions.

2.5.15 GWAS analysis

A list of GWAS SNPs was obtained from the National Human Genome Research Institute's GWAS catalog (accessed Jan 6, 2014)[85]. SNP haplotypes were excluded from the SNP overlap analysis, and a list of 11,194 unique SNPs was obtained. The merged union of the RefSeq, UCSC, and GENCODE catalogs was used as a reference for comparison with MiTranscriptome.

2.5.16 Transcript expression estimation

Expression levels (FPKM) of the transcripts in the assembly were determined using Cufflinks (version 2.02 and 2.1.1)[112]. Normalized abundance estimates (FPKM) were computed for all MiTranscriptome transcripts, converted into approximate fragment count values, and aggregated into a matrix of expression data (**Figure 2.4a**). Library size factors for expression normalization were computed by applying the geometric normalization method described by Anders and Huber[113].

2.5.17 Transcript expression enrichment analysis

To analyze differential expression of transcripts relative to sample phenotypes we developed a method called Sample Set Enrichment Analysis (SSEA). SSEA performs weighted KS-tests using normalized count data vectors as weights. To convert count values into weights for a single KS-test the following steps are performed: (1) raw count values are normalized by library-specific size factors, (2) normalized count values are "resampled" from a Poisson distribution (lambda equals the observed count value) to mimic the effect of technical replication, and (3) random Poisson noise (by default, lambda equals 1) is added to the normalized, resampled count values to destabilize zero-valued counts and break ties. A power transform (exponential or logarithmic) is then applied to the weights (by default, a log-transformation is applied after

incrementing normalized count values by 1). The choice of power transformation influences the relative importance of precision versus recall during enrichment testing. For example, users aiming to discover genes new in molecular subtypes of a disease would prioritize precision over sensitivity, whereas a user aiming to discover ideal biomarkers may value sensitivity over precision. Following count data normalization and power transformation, SSEA performs the weighted KS-test procedure described in GSEA[86,114]. The resulting enrichment score (ES) statistic describes the strength of association between the weights and the sample set.

To control for random sampling bias in count values (*e.g.* "shot noise") SSEA performs repeated enrichment tests using resampled count values to mimic observations from technical replicates and uses the median enrichment score (by default, 100 tests are performed). The basis for Poisson resampling as a legitimate model for technical replication was established by Marioni *et al*.[114] To test for significance, SSEA performs enrichment tests using randomly shuffled sample labels to derive a set of null enrichment scores with the same sign as the observed score (by default, 1000 null enrichment scores are computed). The nominal p value reported is the relative rank of the observed enrichment score within the null enrichment scores. To control for multiple hypothesis testing, SSEA maintains the null normalized enrichment score (NES) distributions for all transcripts in a sample set, and uses the null NES distribution to compute FDR q values in the same manner as proposed by Subramanian *et al*.[86]

2.5.18 Benchmarking SSEA performance using microarray gene signatures

Gene signatures for the top 1% of overexpressed and underexpressed genes from three prostate cancer[87–89] and three breast cancer[17,90,94] microarray studies were obtained using Oncomine[92]. The top 1% gene signatures as detected by SSEA in the MiTranscriptome

breast and prostate cohorts were determined using prostate cancer versus normal and breast cancer versus normal sample sets (**Figure 2.4a**). Given that the MiTranscriptome was produced from an *ab initio* assembly, transcript identity was assigned to the annotated reference gene with the greatest degree of concordance, where degree of splicing agreement was prioritized over degree of exonic same-stranded overlap. The most-enriched isoform for each gene was used to produce a gene signature.

Degree of overlap for all combinations of the 16 gene sets tested (3 published breast up-regulated sets, 3 published breast down-regulated sets, 3 published prostate up-regulated sets, 3 published prostate down-regulated sets, 1 SSEA-determined prostate up-regulated set, 1 SSEA-determined prostate down-regulated set, 1 SSEA-determined breast up-regulated and 1 SSEA-determined breast down-regulated set) was determined by calculating an odds ratio and performing a Fisher's exact test for each gene set pair. Each comparison was restricted to the set of genes assessed by both profiling platforms. Microarray chip annotation files were downloaded from the Molecular Signatures Database (MSigDB)[115]. The set of all annotated genes (relative to RefSeq, UCSC, and GENCODE) was used as the annotation file for MiTranscriptome. Unsupervised hierarchical clustering of the heatmap data was performed using the 'euclidean' distance measure and the 'complete' agglomeration method.

2.5.19 Discovery of lineage-specific and cancer-specific transcripts

To generate enrichment test data for unsupervised clustering, we ranked transcripts within each SSEA sample set by normalized enrichment score (NES) and assigned fractional ranks (e.g. a fractional rank of 0.95 implies the transcript ranked in the top 5th percentile of all transcripts in the sample set). Only significant results (FDR < 1e-7 for lineage analysis and FDR < 1e-3 for cancer versus normal analysis) were used.

Unsupervised clustering was performed using Pearson correlation of log-transformed fractional ranks as a distance metric and Ward's method. Transcripts that were significantly associated with multiple sample sets were grouped with the most strongly associated sample set. Heatmaps were produced using the 'heatmap.2' function from the 'gplots' package in R.

2.5.20 Guilt-by-association GSEA analysis

For each cancer and/or lineage associated lncRNA, expression levels of the target lncRNA were correlated to the expression of all protein-coding genes across all samples in the associated tissue cohort. For cancer cohorts (e.g. breast, prostate), correlations were performed (Spearman) using only the cancer samples (i.e. normal samples were excluded). The protein-coding genes were then ranked by the Rho value, and used in a weighted, pre-ranked GSEA analysis against a collection of cancer associated gene sets from MSigDB. Significant associations were determined for any gene set having an FWER p-value below 0.001.

2.6 Acknowledgements

We thank Brock Palen and Jeremy Hallum for technical assistance with the high performance computing cluster, Sameek Roychowdhury for reviewing the manuscript, the University of Michigan DNA Sequencing Core for Sanger sequencing, and K. Giles for critically reading manuscript and submission of documents. This work was supported in part by the NIH Prostate Specialized Program of Research Excellence grant P50CA69568, the Early Detection Research Network grant UO1 CA111275, US National Institutes of Health R01CA132874, RO1 CA154365 (D.G.B, A.M.C.), and the

Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by the Prostate Cancer Foundation and the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor and a Taubman Scholar of the University of Michigan. R.M. was supported by a Prostate Cancer Foundation Young Investigator Award and the Department of Defense Post-doctoral Fellowship W81XWH-13-1-0284. Y.S.N. is supported by a University of Michigan Cellular & Molecular Biology NIGMS Training Grant.

2.7 Author Contributions

M.K.I., Y.S.N. and A.M.C. conceived the study and analyses. M.K.I. processed RNA-seq data and performed *ab initio* assembly. M.K.I. and Y.S.N. performed data processing and data analysis with assistance from T.R.B., R.M., A.S., Y.H., J.E., S.Z., J.R.P., and F.Y.F.; R.M., U.S., A.S., and Y.H. performed qPCR validations; M.K.I. and Y.S.N. developed SSEA with help of H.K.I.; D.G.B. contributed primary samples. D.R.R., W.Y. and S.M.D. generated RNA-seq libraries and X.C. performed the sequencing. M.K.I., Y.S.N., and A.S. developed the web resource; T.R.B. provided systems administration, data storage, high-performance computing, and networking support. A.P. performed the proteomics analysis. M.K.I, Y.S.N., and A.M.C. wrote the manuscript. All authors discussed results and commented on the manuscript.

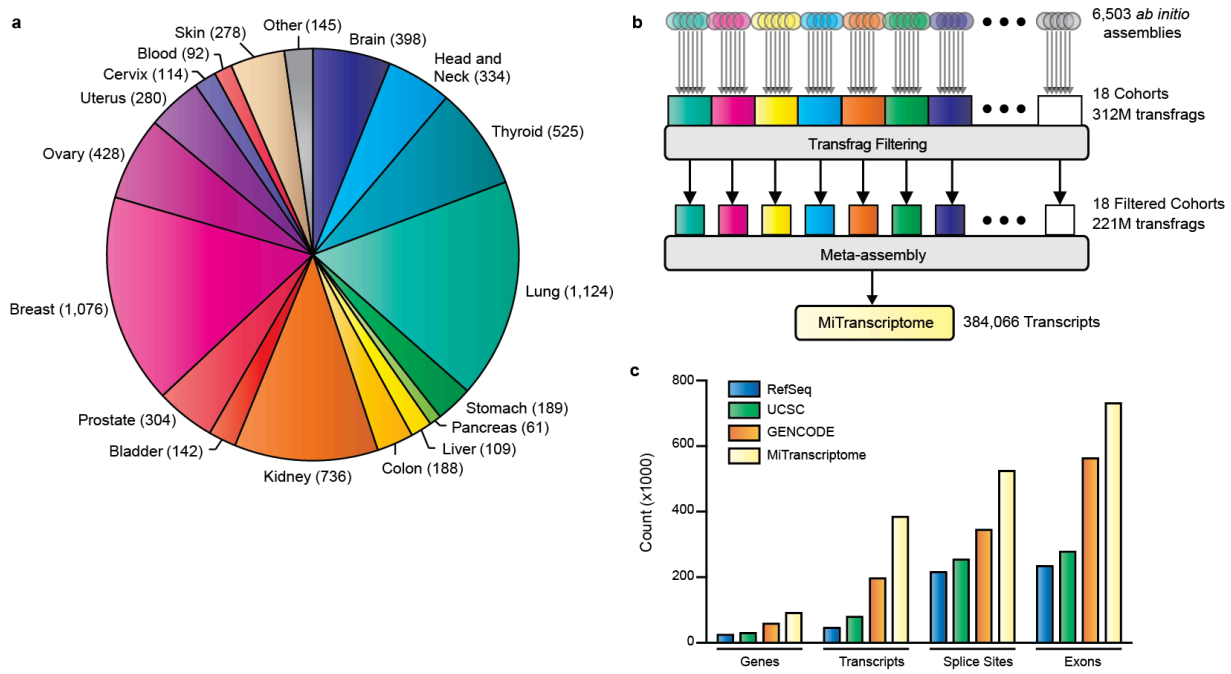


Figure 2.1 ***Ab initio*** **transcriptome assembly reveals an expansive landscape of human transcription.** (**a**) Pie chart showing composition and cohort sizes for transcriptome reconstruction. The 6,503 RNA-Seq libraries were categorized into 18 cohorts by organ system. Organ systems with relatively few libraries were grouped together as 'other'. (**b**) Workflow diagram for transcriptome reconstruction. *Ab initio* assembly was carried out on each RNA-Seq library yielding transcript fragments (transfrags) predictions that may represent full or partial length transcripts. *Ab initio* assemblies were grouped by cohort and filtered to remove unreliable transfrags. Meta-assembly was performed on filtered transfrags for each cohort. Finally, transcripts from individual cohorts were merged to produce a consensus MiTranscriptome assembly. (**c**) Bar chart comparing exons, splice sites, transcripts, and genes in the MiTranscriptome assembly with the RefSeq (Dec, 2013), UCSC (Dec, 2013) and GENCODE (release 19) catalogs.

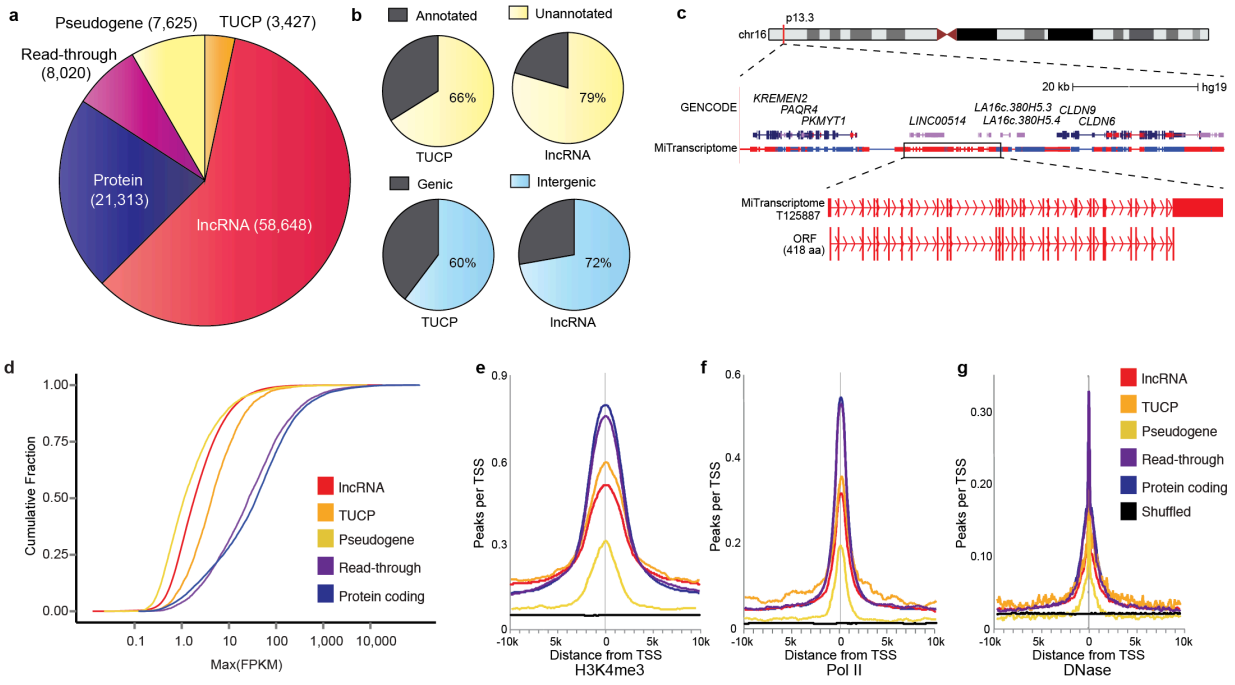


Figure 2.2 Characterization of the MiTranscriptome assembly. (**a**) Pie chart of composition and quantities of lncRNA, transcripts of unknown coding potential (TUCP), expressed pseudogene, read-through, and protein-coding genes in the MiTranscriptome assembly. (**b**) Pie charts of number of lncRNAs and TUCP genes (top) unannotated versus annotated relative to reference catalogs and (bottom) intragenic versus intergenic. (**c**) Genomic view of the chromosome 16p13.3 locus. Protein coding genes (*PKMYT1* to *CLDN9*) border an intergenic region containing GENCODE lncRNA genes *LINC00514* and *LA16c.380H5*. MiTranscriptome transcripts encompassing these genes are shown in a dense view, and (bottom) an individual isoform containing a 29-exon, 418aa ORF is highlighted. This ORF spans multiple GENCODE lncRNAs. (**d**) Empirical cumulative distribution plot comparing the maximum expression (FPKM) of the major isoform of each gene across gene categories. (**e**, **f**, and **g**) Plots of aggregated ENCODE ChIP-Seq data from 13 cell lines at 10kb intervals surrounding expressed transcription start sites (FPKM > 0.1) for (**e**) H3K4me3, (**f**) RNA polymerase II (Pol II), and (**g**) DNase hypersensitivity.

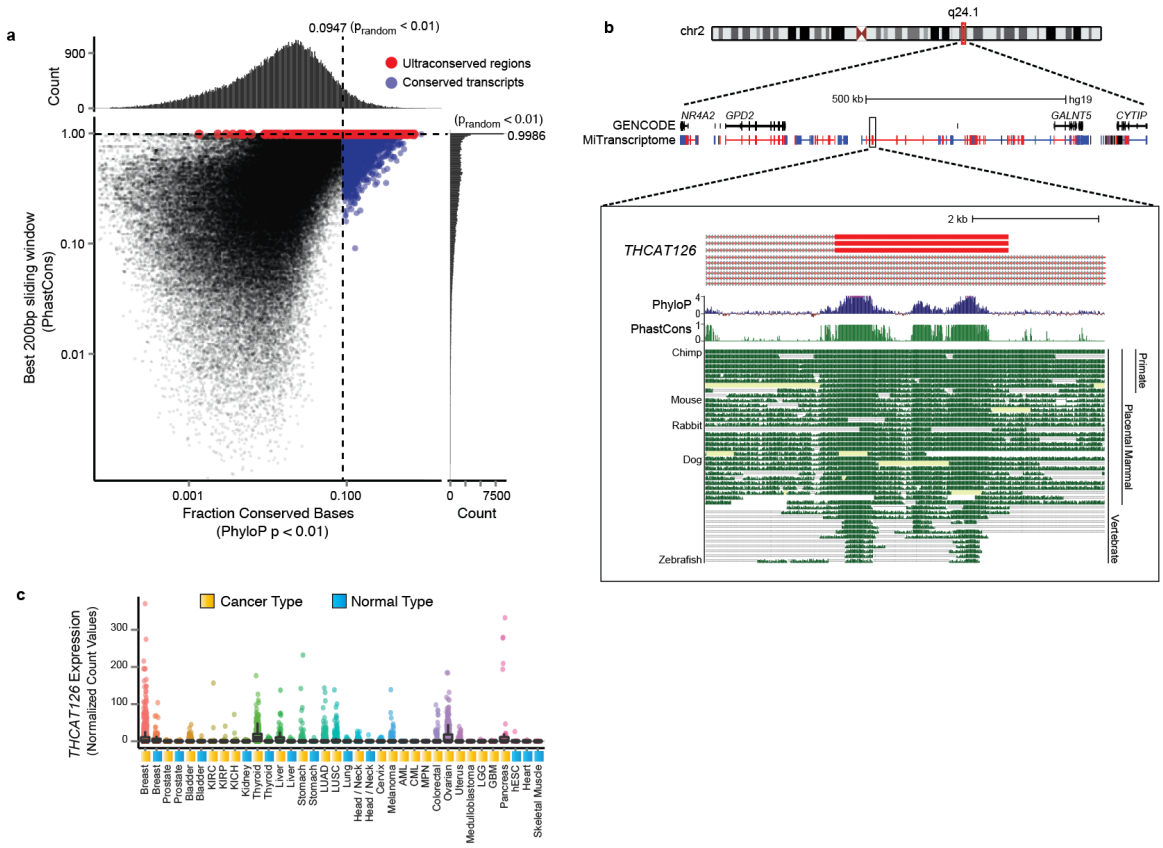


Figure 2.3 Analysis of conservation in lncRNAs. (**a**) Scatter plot with marginal histograms depicting the distribution of full transcript conservation levels (x axis) and maximal 200bp window conservation levels (y axis) for lncRNA and TUCP transcripts. Full transcript conservation levels were measured using the fraction of conserved bases (PhyloP $p < 0.01$). Sliding window conservation levels were measured using the average PhastCons score across 200bp regions along the transcript. Blue points indicate transcripts that were conserved relative to random non-transcribed intergenic control regions (false positive rate < 0.01). Red points indicate transcripts with 200bp windows that meet the criteria for 'ultraconserved' regions. Marginal histograms depict the distribution of scores along both axes. Scores of zero were omitted from the plot. (**b**) Genomic view of chromosome 2q24.1 locus. Protein coding genes *GALNT5* and *GPD2* flank an intergenic region with no annotated transcripts. MiTranscriptome transcripts are shown in a dense view populating this intergenic space. Blue and red color represents positive and negative strand transcripts, respectively (color scheme applies to all subsequent genomic views). Most zoomed view (bottom) depicts a highly conserved exon from the lncRNA *THCAT126*. Multiz alignment of 46 vertebrate species depicted as well as the per base PhyloP and PhastCons conservation score. (**c**) Expression data for *THCAT126* across all MiTranscriptome cancer and normal tissue type cohorts.

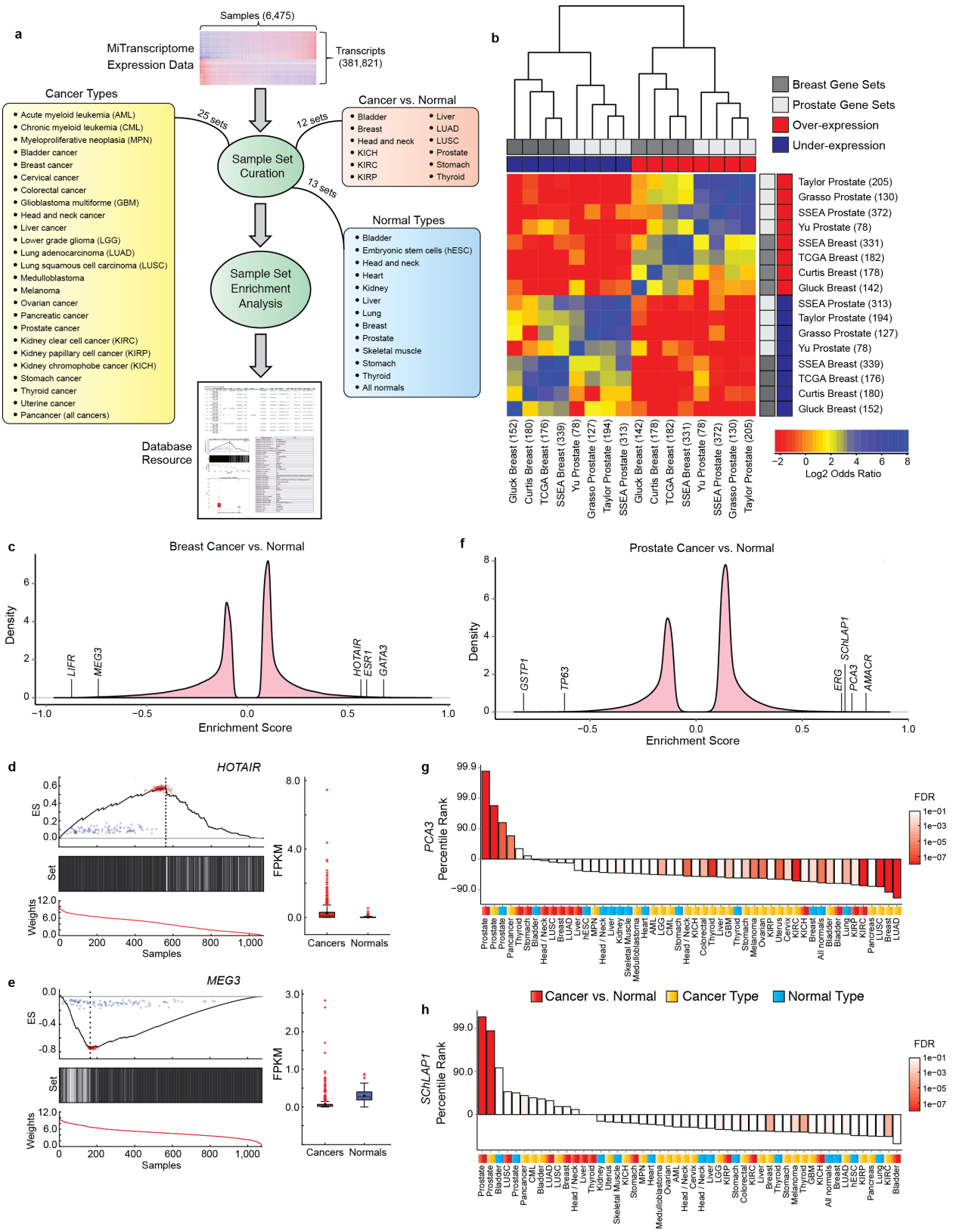


Figure 2.4 Methodology for discovering cancer-associated lncRNAs.

(**a**) Samples were grouped into 50 different sample sets in three categories: (1) cancer type, (2) normal type, and (3) cancer versus normal. Enrichment testing was performed using SSEA, and significant transcripts were imported into an online resource. (**b**) Heatmap showing concordance of SSEA algorithm with prostate and breast cancer gene signatures obtained from the Oncomine database. The top 1% over-expressed and under-expressed genes from each analysis were compared using Fisher's Exact Tests. (**c**) Enrichment score density plots for breast cancers versus normal samples. (**d** and **e**) Enrichment and expression plots for lncRNAs (**d**) *HOTAIR* and (**e**) *MEG3*. Subplots include: (*top*) running ES across all samples (dotted line: max/min ES, red points: Poisson resamplings of fragment counts, blue points: random permutations of the sample labels). (*middle*) Black bars (cancers) or white bars (normals). (*bottom*) Rank-ordered normalized expression values. Adjacent boxplots (interquartile range and median shown by box and whiskers) depict transcript expression (FPKM) in cancers and normals. 967 and 109 patients in the breast cancer and normal groups, respectively. (**f**) Enrichment score density plots for prostate cancers versus normal samples. (**g** and **h**) Bar plots of percentile ranks for prostate cancer-specific lncRNAs (**g**) *PCA3* and (**h**) *SChLAP1* across Cancer vs. Normal (red), Cancer Type (gold) and Normal Type (blue) sample sets. Bar colors depict statistical significance (FDR).

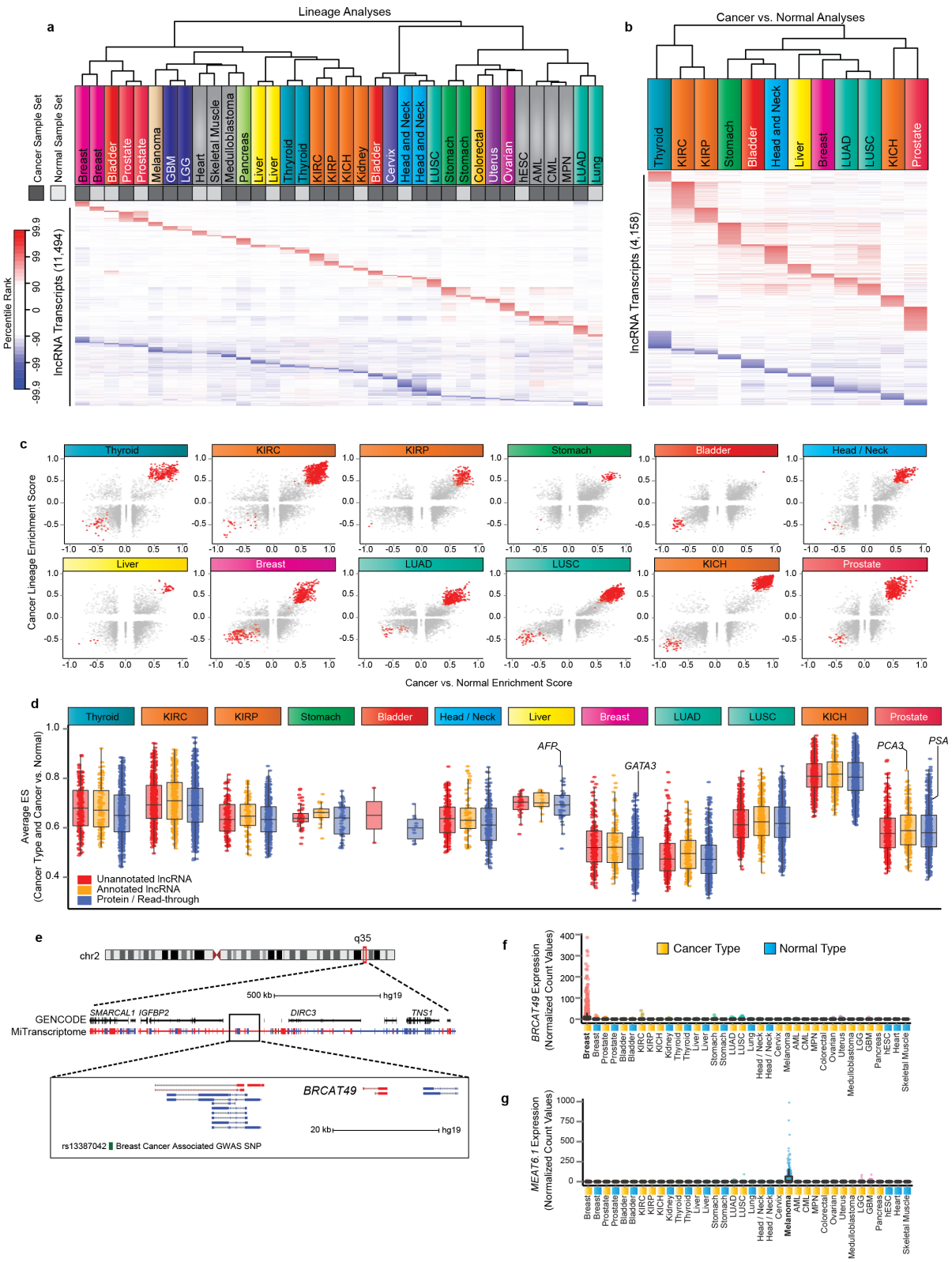


Figure 2.5 Discovery of lineage-associated and cancer-associated lncRNAs in the MiTranscriptome compendium. (**a**) Heatmap of lineage-specific lncRNAs. Each column represents a sample set from one of 25 cancer (dark grey) and normal (light

grey) lineages and each row represents an individual lncRNA transcript. All transcripts were statistically significant (FDR < 1e-7) and ranked in the top 1% most positively or negatively enriched transcripts within at least one sample set. The heatmap color spectrum corresponds to percentile ranks, with under-expressed transcripts (blue) and over-expressed transcripts (red). (**b**) Heatmap of cancer-specific lncRNAs nominated by SSEA Cancer vs. Normal analysis of 12 cancer types (columns). All transcripts were statistically significant (FDR < 1e-3) and ranked in the top 1% most positively or negatively enriched transcripts within at least one sample set. (**c**) Scatter plots showing enrichment score for Cancer vs. Normal (*x* axis) and Cancer Lineage (*y* axis) for all lineage-specific and cancer-associated lncRNA transcripts across 12 cancer types. Red points indicate transcripts meeting the percentile cutoffs for cancer- and lineage-association. (**d**) Boxplot comparing the performance of cancer- and lineage-associated lncRNAs across 12 cancer types. The average of the lineage and cancer versus normal ES is plotted on the *y* axis. (**e**) Genomic view of chromosome 2q35 locus. Most zoomed view (bottom) depicts BRCAT49, a breast lineage and breast cancer specific lncRNA. Breast cancer associated GWAS SNP, rs13387042, is depicted in green. (**f**) Expression data for BRCAT49 across all MiTranscriptome cancer and normal tissue type cohorts. (**g**) Expression data for MEAT6 across all MiTranscriptome cancer and normal tissue type cohorts.

Table 2.1 Summary of lineage and/or cancer- specific lncRNAs nominated.
For each of the 27 tissue types (rows), the table lists numbers of lncRNA genes associated with tissue and/or cancer type, enriched for conserved nucleotides, containing ultraconserved elements, or classified as TUCP. Cancer and tissue specific lncRNAs only delineated for tissue types for which there was a sufficient number of matched normal to perform a cancer versus normal analysis (NA reported otherwise).

Tissue/Cancer Type (Naming Convention)	**Total Associated Non-Coding Transcripts**	**# Cancer- & Tissue-Specific**	**# Conserved**	**# Containing Ultraconserved Element**	**# Classified as TUCP**
Acute Myelogenous Leukemia Associated Transcripts (AMATs)	373	NA	29	13	26
Bladder Cancer Associated Transcripts (BLCATs)	61	0	9	2	5
Breast Cancer Associated Transcripts (BRCATs)	1115	134	82	27	76
Cervical Cancer Associated Transcripts (CVATs)	162	NA	12	2	13
Chronic Myelogenous Leukemia Associated Transcripts (CMATs)	157	NA	16	3	11
Colorectal Cancer Associated Transcripts (CRATs)	163	NA	29	4	17
Glioblastoma Multiforme Associated Transcripts (GBATs)	161	NA	11	2	22
Head and Neck Cancer Associated Transcripts (HNCATs)	766	5	45	15	68

Heart Tissue Associated Transcripts (HRATs)	170	NA	16	1	12
Human Embryonic Stem Cells Associated Transcripts (ESATs)	205	NA	10	0	20
Chromophobe Renal Cell Carcinoma Associated Transcripts (KCHCATs)	1050	52	64	20	92
Renal Clear Cell Carcinoma Associated Transcripts (KCCATs)	1429	215	84	26	123
Renal Papillary Cell Carcinoma Associated Transcripts (KPCATs)	474	0	41	8	38
Low Grade Glioma Associated Transcripts (LGATs)	265	NA	31	10	23
Liver Cancer Associated Transcripts (LVCATs)	250	0	18	1	20
Lung Adenocarcinoma Associated Transcripts (LACATs)	953	19	64	19	61
Lung Squamous Cell Carcinoma Associated Transcripts (LSCATs)	1014	10	70	23	58
Medulloblastoma Associated Transcripts (MBATs)	312	NA	26	3	33
Melanoma Associated Transcripts (MEATs)	339	NA	24	2	34
Myeloproliferative	101	NA	12	1	8

Neoplasia Associated Transcripts (MPATs)					
Ovarian Cancer Associated Transcripts (OVATs)	163	NA	37	12	30
Pancreatic Cancer Associated Transcripts (PNATs)	247	NA	27	4	22
Prostate Cancer Associated Transcripts (PRCATs)	727	38	49	14	62
Skeletal Muscle Tissue Associated Transcripts (SMATs)	123	NA	5	1	11
Stomach Cancer Associated Transcripts (STCATs)	95	0	10	1	10
Thyroid Cancer Associated Transcripts (THCATs)	1289	80	73	21	111
Uterine Endometrial Carcinoma Associated Transcripts (UTATs)	183	NA	31	1	16

CHAPTER 3

TACO Produces Robust Multi-Sample Transcriptome Assemblies from RNA-Seq

Citation: Yashar S. Niknafs, Balaji Pandian, Hariharan K. Iyer, Arul M. Chinnaiyan, Matthew K. Iyer. TACO produces robust multi-sample transcriptome assemblies from RNA-seq. *Nature Methods* (2017)[116].

3.1 Abstract

Accurate transcript structure and abundance inference from RNA-Seq data is foundational for molecular discovery. Here we present TACO, a computational method to reconstruct a consensus transcriptome from multiple RNA-Seq datasets. TACO employs novel change-point detection to demarcate transcript start and end sites, leading to dramatically improved reconstruction accuracy compared to other tools in its class. The tool is available at http://tacorna.github.io and can be readily incorporated into RNA-Seq analysis workflows.

3.2 Introduction

High-throughput RNA sequencing (RNA-Seq) has enabled a deep understanding of the transcriptome[68,71,117]. While efforts to annotate high fidelity gene models by manual and automated systems have relied primarily on low-throughput sequencing methods[48,50,73],

several studies using RNA-Seq have described an expansive transcriptome, suggesting that reference gene catalogs are far from complete[28,68,78]. This annotation gap has been widened further by the rapid sequencing of thousands of high quality RNA-Seq datasets by consortia such as TCGA[118], ICGC[119], and GTEX[120]. Limitations of current reference annotation projects necessitate development of computational transcriptome assembly methods to utilize these large-scale RNA-Seq datasets for transcript discovery[73].

Transcriptome reconstruction for single-sample experiments remains an active area of investigation. Although current reconstruction methods achieve high accuracy at the nucleotide or splice junction level, studies have shown that they do not reliably predict the splicing patterns of full-length transcripts[55]. Moreover, single-sample transcriptome reconstruction has limited utility for the downstream analyses of transcriptional dynamics across many samples. To address this issue, a consensus transcriptome from multiple input datasets must be constructed, which can be done by merging individual transcriptome reconstructions in a hierarchical fashion. We term this procedure "meta-assembly".

Current software for meta-assembly include the Cuffmerge utility within the Cufflinks package[112] and a merge mode within the StringTie software (hereafter denoted 'StringTie-merge')[52]. We demonstrate that the performance of both Cuffmerge and StringTie-merge deteriorates as the number of input datasets increases, yielding myriad incorrectly predicted transcript structures with intron retentions, aberrantly long 3' and 5' ends, and read-through transcripts that concatenate multiple neighboring genes on the same strand. These errors may originate from contamination of the input libraries with incompletely processed RNA or genomic DNA[78,121], or be propagated from

genome-guided assembly. To remain scalable, meta-assembly methodology must be designed to mitigate these sources of error.

Here we present a new meta-assembly method, TACO (**T**ranscriptome **A**ssemblies **C**ombined into **O**ne), as a robust solution for leveraging the vast RNA-Seq data landscape for transcript structure prediction. To prepare data for TACO, sequence reads are aligned to a reference genome by a spliced alignment tool. Then, genome-guided transcript assembly is performed which serve as input to TACO (**Figure 3.1**). As a pure meta-assembler TACO remains agnostic to the alignment and assembly strategy used, and therefore can be readily incorporated into existing RNA-Seq analysis protocols such as the Tuxedo suite[122]. The software, written in Python and C, and its documentation can be obtained at http://tacorna.github.io.

3.3 Results

Building on existing transcriptome assembly methods, TACO approaches meta-assembly by modeling alternatively spliced genes using directed acyclic graphs, or splice graphs, built from the input transcripts (**Figure B.1a** and **Methods**). To harness the splicing pattern information present in the input transcripts, TACO constructs path graphs wherein each node corresponds to a sequence of consecutive splice junctions (**Figure B.1b**). TACO then applies dynamic programming to enumerate the set of isoforms with highest abundance. To mitigate the deleterious effects of incorrectly assembled input transcripts, TACO employs change point detection via binary segmentation to identify points of significant change of expression in the basewise transcript abundance landscape (**Figure B.2**). As evidenced by the performance tests

detailed below, change point detection effectively breaks apart read-through transcripts and accurately delineates transcript start and end sites.

To assess the performance of TACO compared to other tools, we initially computed the precision, recall (i.e. sensitivity), and F-measure (harmonic mean of precision and recall) at the nucleotide, splice junction, and transcript splicing pattern (i.e. isoform) level while varying the number of input transcriptomes. We used RNA-seq data from the CCLE[123] and used GENCODE v24 as a reference standard. TACO outperformed Cuffmerge and StringTie-merge in its ability to predict bases covered, splice junctions, and splicing patterns for runs with more 10 samples (**Figure 3.2a**). The performance of Cuffmerge and StringTie-merge decreased as the number of samples increased, whereas TACO was robust to sample size. The most prominent advantage of TACO was its ability to identify correct splicing patterns, which is the most challenging goal of transcriptome assembly[52,55]. Predicting correct isoforms is essential for downstream analysis, as abundance estimation applied to incorrectly assembled transcripts will likely be inaccurate. TACO achieved splicing pattern precision of approximately 30%, compared to about 5% for each of the other tools (**Figure 3.2a**). Of note, the reported splicing pattern precision of individual sample transcriptome reconstruction tools is around 30%[52,55], suggesting that even when merging 500 samples, TACO displays little loss of accuracy.

All three merging tools offer a key parameter, termed the "isoform fraction cutoff", for filtering out minor isoforms based on their abundance relative to the major isoform. Adjusting this parameter results in a trade-off between precision and recall. Thus, by modulating this parameter we assessed performance the tools' full range of precision and recall (**Figure 3.2b** and **Figure B.3**). TACO outperformed Cuffmerge and StringTie-merge across the spectrum of isoform fraction cutoffs, achieving an average precision (a

statistic that serves as a surrogate "area under the curve" for precision-recall plots) of 0.53, 0.78, and 0.21 for bases covered, splice junctions, and splicing patterns, respectively, compared to 0.33, 0.75, and 0.06 for Cuffmerge and 0.41, 0.75, and 0.09 for StringTie-merge. Most prominent was TACO's ability to reconstruct correct transcript splicing patterns with high precision (>50% for high isoform fraction cutoffs).

It is important to note that the selection of reference standard can heavily impact precision metrics, introducing potential biases depending on the nature of the reference selected. In addition to GENCODE, we also tested performance using long-read RNA-seq data as a reference standard, and the superior performance of TACO was consistent in these additional analyses (**Figure B.4**).

We assessed how well the meta-assembly tools correctly prioritized transcripts that are highly expressed. Meta-assemblies using all three tools were generated, and transcript abundance from these assembled transcripts was quantified. We then measured each tool's precision, recall, and F-measure on subsets of the top *N* most highly abundant transcripts, varying *N* from 500 to 30,000. TACO possessed a striking ability to predict the nucleotides, splice junctions, and splicing patterns of highly expressed transcripts, achieving an average precision of 0.45, 0.59, and 0.17, respectively, compared to 0.35, 0.49, 0.01 for Cuffmerge and 0.41, 0.54, and 0.03 for StringTie. Notably, TACO attained a splicing pattern precision of 76.8% for the top 5,000 most highly abundant transcripts, dramatically better than the 30% and 16.2% precision of StringTie-merge and Cuffmerge, respectively (**Figure 3.2c** and **Figure B.5**).

In order to directly assess the efficacy of change point detection, we quantified the fraction of read-through genes produced by each tool (**Figure B.6**). At low isoform fraction cutoffs, ~10% and ~7.5% of the genes produced by StringTie-merge and

Cuffmerge, respectively, contained multiple unique GENCODE genes (i.e., "read-throughs"), while only ~2.5% of TACO transcripts were read-throughs. This finding supports the effectiveness of change point detection in detecting transcription start and end sites. Reduction of read-through transcription is of potential benefit to downstream analyses, as read counting is particularly susceptible to the deleterious effects of read-through transcription, especially if the underlying genes encompassed by the read-through are highly discrepant in expression value. It is important to note, however, that some of this read-through transcription may be real transcriptional events.

As evidenced by the performance assessment above, TACO retains robust accuracy even when processing a large number of input datasets, producing higher-fidelity transcripts with read-throughs and spurious isoforms compared to Cuffmerge and StringTie-merge. We observed a representative example in the 3p21 genomic locus, which harbors three genes in close proximity on the negative strand (**Figure 3.3**). Although gene-dense regions are particularly challenging for meta-assembly, the change point detection and other algorithmic advances of TACO enable it to cleanly disambiguate all three genes, whereas Cuffmerge and StringTie-merge produce dozens of spurious isoforms, many of which are read-throughs. This problem is present when merging 50 samples and is exacerbated as the number of merged samples rises to 500 (**Figure 3.3**).

3.4 Discussion

While extraordinary efforts of sequencing consortia have made thousands of high quality RNA-Seq datasets from diverse tissue types and disease states available to the scientific community, RNA-Seq continues to be an underutilized source of evidence for

expanding reference transcript catalogs. The lack of a robust computational method to accurately coalesce large numbers of input datasets into a consensus transcriptome has been a major impediment to the use of RNA-Seq for this purpose. Given its accuracy and scalability, TACO has potential to facilitate progress in this area, and we expect that its adoption will help unravel the complexity of the transcriptome.

3.5 Methods

3.5.1 Genome-guided assembly

FASTQ files were obtained from DBGap (accession number phs000178) using the Cancer Genomics Hub (https://cghub.ucsc.edu). Reads were aligned to human genome version GRCh38/hg38 using STAR[124] version 2.4.2 with the following parameters: --alignSJDBoverhangMin 3, --alignIntronMin 20, --alignSJoverhangMin 8, --alignMatesGapMax 1000000, --alignIntronMax 1000000, --scoreGenomicLengthLog2scale 0, --outSAMmode NoQS --outFilterType BySJout. Cufflinks[112] (version 2.2.1) and StringTie[52] (version 1.2.2) were utilized to produce transcriptome assemblies from the BAM alignments generated by STAR as detailed above.

3.5.2 Meta-assembly

Performance of Cuffmerge[112] (version 2.2.1), StringTie-merge (version 1.2.2), and TACO were tested. Cuffmerge was run using default parameters except for the isoform fraction, "--min-isoform-fraction", which was adjusted according to the appropriate analyses. StringTie was run with the analogous parameter ("-f") for isoform fraction

cutoff. For all analyses unless otherwise indicated, StringTie was run with the "-T" and "-F" parameters set to zero and TACO was run with the *--filter-min-expr* set to zero. These parameters filter the input transcripts using a provided expression cutoff. While StringTie-merge and TACO provide an option to filter input assemblies by expression level, Cufflinks does not provide this option, and so all analyses reported above were done with the expression filter option turned off for these tools. Comparing the performance of StringTie-merge and TACO run with a 1 FPKM (the StringTie-merge default) filtration of input transcripts did not reveal substantial differences in the tools' performance. StringTie-merge did display an increase in performance at the base level, but TACO remained a superior tool for all metrics (base, splice junctions, splicing patterns) (**Figure B.7**).

3.5.3 Performance assessment

The performance of all meta-assemblies were assessed using the "gffcompare" utility version 0.9.5 (https://ccb.jhu.edu/software/stringtie/gff.shtml). All meta-assemblies were compared to GENCODE version 24 (level 1 protein-coding genes). Only poly-exonic transcripts were utilized in performance assessment ("-M" and "-N" gffcompare flags), and precision-correction was performed by only utilizing test transcripts overlapping reference ("-Q" flag in gffcompare). The average precision metric was utilized as a surrogate measure of the AUC for the precision / recall curve[125] and was calculated as follows:

$$\sum_{k=1}^{N} P(k)\Delta r(k)$$

where N is the number of points samples, $P(k)$ is the precision at point k, and Δr is the change in recall that occurs between point $k-1$ and k.

3.5.4 Sample size performance assessment

Performance of all three meta-assembly tools was measured at varying numbers of input assemblies. Twenty batches of different sizes ranging from 1 to 500 were selected randomly from the 935 CCLE samples. All three tools were run using their default setting for the isoform fraction cutoff parameter (i.e., 0.05 for TACO, 0.05 for Cuffmerge, 0.01 for StringTie-merge).

3.5.5 Isoform fraction cutoff assessment

Cuffmerge, StringTie, and TACO were tested at varying ranges of isoform fraction cutoffs as described above. Meta-assembly was performed on the 55 breast cancer cell lines in the CCLE. To test the full range of performance for all tools the following isoform fraction cutoffs were used: 0.001, 0.00115139, 0.0013257, 0.0015264, 0.00175748, 0.00202354, 0.00232989, 0.00268261, 0.00308873, 0.00355633, 0.00409472, 0.00471462, 0.00542837, 0.00625017, 0.00719638, 0.00828584, 0.00954024, 0.01098453, 0.01264748, 0.01456218, 0.01676675, 0.01930507, 0.02222767, 0.02559271, 0.02946719, 0.03392823, 0.03906463, 0.04497862, 0.05178793, 0.05962811, 0.06865521, 0.07904892, 0.09101613, 0.10479506, 0.12065998, 0.1389267 , 0.15995881, 0.18417497, 0.21205722, 0.24416056, 0.28112402, 0.32368339, 0.37268581, 0.42910671, 0.49406917, 0.5688663 , 0.65498696, 0.75414543, 0.86831549, and 0.999.

3.5.6 Comparison to tissue PacBio long-read sequencing

Long-read PacBio RNA sequencing data was used as a comprehensive reference standard for meta-assembly performance assessment. Consensus split mapped

molecules (CSMMs) for aligned PacBio reads from the Sharon et al. study[126] were obtained from http://stanford.edu/~htilgner/2013_NBT_paper/pacBio.index.html in GFF format. All GFF attributes arising from the same PacBio read were given a common transcript ID for conversion to GTF format to be utilized in reference statistics. Additonally, the *gffread* utility (part of the Cufflinks suite)[112] was used to collapse overlapping CSMMs. Of the 20 tissue and organ types represented in the Sharon et al. study, 17 were present in the GTEX dataset. Three samples from each tissue type were selected at random, and fastq files for these samples were downloaded from dbGAP (phs000424). RNA-seq data was aligned as described above and Cufflinks was used to obtain transcriptome assemblies to be used as input for meta-assembly. Assessment of performance for the meta-assemblies was performed as described above using the *gffcompare* utility.

3.5.7 Comparison to brain long-read sequencing

CSMMs from long-read SLR-RNA-seq of multiple brain samples was used as an additional comprehensive reference standard. CSMMs from the Tilgner et al. study[127] were obtained from http://stanford.edu/~htilgner/2014_humanMouseBrain_SLR_RNA_Seq/index_SLRseq.html in GFF format and converted to GTF format as described above. Fifty benign brain RNA-seq samples from GTEX were selected at random, and fastq files for these samples were downloaded from dbGAP (phs000424). RNA-seq data was aligned as described above and Cufflinks was used to obtain transcriptome assemblies to be used as input for meta-assembly. Assessment of performance for the meta-assemblies was performed as described above using the *gffcompare* utility.

3.5.8 Performance of highest expressed transcripts

In order to assess the performance of Cuffmerge, StringTie, and TACO, for the highest expressed transcripts, transcript abundance was estimated for the meta-assemblies produced from the merging of the 55 breast cancer CCLE samples using an isoform fraction cutoff of 0.05. The Kallisto RNA-seq isoform abundance estimation tool[128] was used to calculate isoform abundance for all transcripts in the meta-assemblies produced by the three tools. Expression across the 55 samples was calculated as the sum of the TPM values reported by Kallisto for each transcript across all samples. The highest 100, 500, 1000, 1500, 2000, 3000, 5000, 7000, 10000, 12000, 15000, 18000, 20000, 22000, 25000, 27000, and 30000 transcripts were utilized to test performance of the tools.

3.5.9 Assessment of performance varying change point parameters

Performance of TACO was assessed by comparing to the GENCODE protein-coding genes as described above at different values for the change point p-value cutoff and the change point fold-change cutoff. The following p-value cutoffs were tested: 1, 0.5, 0.25, 0.1, 0.05, 0.01, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7. The following fold-change cutoffs used: 1, 0.95, 0.9, 0.85, 0.7, 0.6, 0.5, 0.35, 0.25, 0.125, 0.0625, 0.03125.

3.5.10 Visualization of meta-assemblies

Examples of the meta-assembly performance were visualized using IGV[129]. For the chromosome 3p21 locus the samples from run "0" were utilized for batch sizes of 50, 100, 500. For all other loci, run "0" was used for a batch size of 100.

3.5.11 Overview of TACO meta-assembly approach

We developed TACO as a software package written in Python and C. The software adapts and builds upon methods that were developed previously by our bioinformatics group for the purposes of meta-assembly, and utilized for the MiTranscriptome project[68]. TACO accepts as input a set GTF files containing transcripts assembled from individual libraries. Transcripts can be filtered based on length in base pairs (*--min-transfrag-length*), with the TACO default set to filtering transcripts less than 200bp in length. TACO aggregates and sorts the input GTF files, and then parses the aggregated set of transcripts into independent loci with non-overlapping genomic coordinates. Within each locus, TACO reassigns unstranded mono-exonic transcripts to either the positive or negative strand whenever there is unambiguous supporting evidence of strandedness from the other transcripts in the locus. Transcripts on the positive and negative strand are then processed independently. TACO then clusters sets of overlapping transcripts on the same strand into splice graphs. Nodes in the splice graph comprise contiguous transcribed regions not interrupted by splicing. Once nodes are identified, the summed expression profile across each node is utilized to perform change point detection. For each change point identified, a new edge is created to the source or sink node, representing a potential transcript start or end site. Details of the various components of the TACO algorithm are found below. For further details, the TACO source code can be found at http://tacorna.github.io.

3.5.12 Details of TACO algorithm

Major Steps:

1. Aggregate: Prior to meta-assembly, the input GTF files are merged and transcripts are sorted by chromosome and position. Filters for transcript length (*--filter_min_length*) and expression (*--filter_min_expr*) are applied.

2. Locus Identification: Transcripts in merged GTF file are partitioned into loci, where a locus is defined as a collection of input transcripts with overlapping genomic coordinates. Groups of transcripts from independent loci are then assembled in parallel.
3. Strand Imputation: TACO attempts to impute the strand for any unstranded transcripts included in the input data. See below for further details.
4. Determine Locus Expression: Stranded basewise expression of each locus is the summed expression from the input transcripts for each base determined by the individual transcriptome assembler utilized to generate the input transcriptome. Of note, TACO is compatible with any expression unit (as long as that unit is consistent for all samples used) and the unit of expression can be specified with the *--gtf-expr-attr* flag.
5. Build Splice Graphs: Transcripts within each locus are partitioned into splice graphs. A splice graph is a directed acyclic graph representing the stranded transcribed genomic regions of the input transcripts. See below for further details.
6. Change Point Detection: Change points are identified within each node in the splice graph (see definition below). Binary segmentation is used to recursively identify multiple points of change meeting significance criteria in each node's expression profile. Upon identification of all change points in each node of a splice graph, the splice graph is updated with new connections to the source and/or sink. See below for further details.
7. Build Path Graph: A specialized graph structure called the path graph is derived from the splice graph. The path graph retains structural information from the input transcripts and encapsulates splicing pattern information into the graph. See below for further details.

8. Predict Isoforms: The path graph is iteratively traversed using a dynamic programming approach, yielding the most highly expressed isoforms in the gene. Users may specify how exhaustively the algorithm predicts isoforms using the *--isoform-frac* and/or *--max-isoforms* command line options.
9. Report output

Step Details:

Strand Imputation

Unstranded transcript assemblies may arise from RNA-Seq data if the underlying reads are unstranded and the transcript is monoexonic. TACO attempts to impute the strand of each unstranded transcript using the following steps:

For each unstranded transcript *T*:

- If all stranded transcripts overlapping *T* are on the positive strand, then assign *T* to the positive strand
- If all stranded transcripts overlapping *T* are on the negative strand, then assign *T* to the negative strand
- If there are transcripts overlapping *T* on both the positive and negative strands, then the strand of *T* is deemed ambiguous and is not imputed
- If no stranded transcripts overlap *T*, then the strand of *T* is deemed ambiguous and is not imputed.
- By default, transcripts with ambiguous strand are not assembled unless the *--assemble-unstranded* option is enabled.

Splice Graph

TACO partitions the input transcripts within a locus by strand, and builds splice graphs from sets of overlapping transcripts within the locus. A splice graph is a directed acyclic graph representing the stranded transcribed genomic regions and splice junctions of the input transcripts.

Properties*:*

- Node: a contiguous exonic genomic interval with no internal splice donors or acceptors in the input transcripts
- Edge: connects nodes x and y if an input transcript contains x and y consecutively
- Expression: The expression of each node is determined by summing the expression of input transcripts that contain the node.

Creation*:*

1. Define Nodes:
    - To define nodes, we first iterate through the input transcripts and define the set of node boundaries. Nodes are bounded by splice donor/acceptor sites and transcriptional start/stop sites (e.g. sites where the summed expression of the input transcripts changes from zero to nonzero).
2. Define Edges:
    - After defining the set nodes in the splice graph, we iterate through the input transcripts in a second pass. For each transcript, we map its exons across node boundaries, thus representing the transcript as a sequence of nodes. We then add the edges inferred by this sequence to the set of all edges in the splice graph.

Change Point Detection

Change point detection is performed at the level of each node within a splice graph. Using the basewise expression for each node, recursive binary segmentation[130] of the change point detection algorithm is performed until potential change points fail to meet significance criteria. Once change points are detected for all nodes in a splice graph, the splice graph is updated to account for the new connections to the source and sink.

Definitions*:*

- Expression vector: the input for the change point detection algorithm. Initially, the expression vector is the number of bases of the entire splice graph node, whose values are the basewise expression of that node. In subsequent recursive iterations of the algorithm, the expression vector represents the basewise expression of the node segment being tested, whose boundaries are defined by previously identified change points.
- Mean squared error (MSE):

$$MSE(m) = \sum_{i=1}^{m} (X_i - \overline{X_1})^2 + \sum_{i=m+1}^{l} (X_i - \overline{X_2})^2$$

    - Where X is the expression vector being tested, m is the index of point being tested in the expression vector, X_i is the expression at index i of the expression vector, $\overline{X1}$ is the mean expression for bases 0 to m in the expression vector, l is the length of X, and $\overline{X2}$ is the mean expression for bases $m + 1$ to l
- Significance criteria for potential change points:
    - Mann-Whitney U (MWU) test must meet defined p-value threshold (--*change-point-pvalue*). The MWU compares the expression values only at points of change on either side of a potential change point.

        - *E.g.*, if the expression on either side of the potential change point are X_1 = [15, 15, 15, 15, 15, 14, 14, 14, 13, 13, 12, 12, 10] and X_2 = [10, 10, 10, 9, 9, 8, 8, 8, 8, 7, 7, 4, 4, 2, 2, 2, 3, 2, 2], the MWU will compare only the values where expression values changed X'_1= [15, 14, 13, 12, 10] and X'_2= [10, 9, 8, 7, 4, 2, 3, 2]
    - Fold-change comparing the mean expression on either side of the change point must meet a defined threshold (*--change-point-fold-change*).
    - Default settings for change point p-value and fold-change were determined by testing the performance of TACO at varying p-values and fold-changes (**Figure B.8**).

Algorithm*:*

For a given expression vector *X*:

- If the length of $X < 20$, do not perform change point detection
- Identify the base/index within *X* with the minimum *MSE* as a potential change point, *PC*.
- Test whether *PC* meets significance criteria
- If *PC* meets criteria:
    - Append point of change, *C*, to list of change points for a given node
    - Utilizing the expression on either side of *C* as a new expression vector to be fed back into the change point detection algorithm, recursively test for other change points (a.k.a, binary segmentation)
        - As change points are often gradual and sloping, in order to prevent calling multiple change points along one slope, when a change point is identified, all bases on the same slope as the identified change point are removed from the expression vector used for future binary segmentation iterations

- If *PC* does not meet significance criteria, stop change point detection binary segmentation

Path Graph

After detecting change points in a splice graph, TACO builds a new graph structure called a path graph to encapsulate sequences of consecutive splice junctions. The data structure of the path graph is reminiscent of the De Bruijn graph model used for *de novo* assembly.

Properties*:*

- Node: a sequence of *k* nodes from the splice graph that appears in at least one input transcript. We subsequently refer to nodes in the path graph as "subpaths".
- Edge: connects subpaths *x* and *y* if the last *k-1* nodes in subpath *x* are equal to the first *k-1* nodes in subpath *y*.
- Expression: the expression at each node equals the summed expression of all of the input transcripts that contain the node
- Source: subpaths originating at a transcriptional start site are connected to a common source node
- Sink: subpaths terminating at a transcriptional stop site are connected to a common sink node

Algorithm*:*

Creation of a path graph with subpaths of length *k*:

- For each input transcript *T:*
    - Define the sequence of splice graph nodes represented by *T* as *NT*
    - If the length of *NT* is less than *k*:

        - If the first node of *NT* is a transcriptional start site and the last node of *NT* is a transcriptional stop site:
            - Add *NT* to the path graph as a single node with connections to the source and sink regardless of the length of *NT* or the value of *k*.
        - Else:
            - Append *NT* to a list of short transcripts, to be dealt with later
    - Else:
        - Add nodes NT[0..*k*], NT[1..*k*+1], NT[2..*k*+2], etc. to the path graph, with edges connecting consecutive nodes.
- Determine and remove unreachable nodes from the path graph:
    - If a node in the path graph is unreachable from either the source node or the sink node, remove it from the path graph
- Attempt to rescue short input transcripts
    - For each short input transcript *ST* (length of the sequences of nodes *NT* less than *k*):
        - Determine all the nodes in the path graph that contain *ST*
            - This is done by first creating a suffix array index of all of the nodes in the path graph and then aligning the sequence *ST* to this index. Further details are available in the online source code.
        - Allocate the expression of *T* to all matching nodes proportionally using the relative expression of each matching node.
- Extend transcripts that do not begin at a start site and end at a stop site:
    - If the first node of *NT* is *not* a transcriptional start site:
        - Allocate the expression of *T* to all predecessor nodes of *NT* in a weighted fashion (see below). This is implemented using breadth

first search starting at *NT[0]*, traversing the graph in reverse until all possible paths to the source node are visited.

- If the last node if *NT* is *not* a transcriptional stop site:
  - Allocate the expression of *T* to all successor nodes of *NT* in a weighted fashion (see below). This is implemented using breadth first search starting at *NT*[-1], traversing the graph in the forward direction until all possible paths to the sink node are visited.
- Expression is allocated to neighbor nodes (either predecessor or successor nodes depending on whether *NT* is being extended from its start site or stop site, respectively) in a weighted fashion, by using the fractional expression levels of all neighbor nodes. For example, if NT[0] with an expression level of 12 is being extended upstream to two possible neighbor (e.g. predecessor) nodes *NA* and *NB* with expression levels of 30 and 60, respectively, *NA* and *NB* will be allocated 4 and 8 expression units, respectively.
- This procedure ensures that the summed expression at the source node equals the summed expression at the sink node.

Choice of k for creating the Path Graph

The choice of *k* for path graph construction poses an interesting tradeoff. As *k* decreases, the chances of assembling transcripts that do not represent the input data increases. In this case, even if input transcript predictions are incorrect the algorithm could still assemble correctly by utilizing expression information. By contrast, as *k* increases, the assembled transcripts will be increasingly constrained to precisely match paths in the input transcripts, and less dependent upon expression information in the input data. Optimal performance could theoretically be achieved by selecting a *k* that balances this tradeoff[131]. Intuitively, the number of nodes in a path graph could be used as a

surrogate measure for graph complexity. With more nodes (i.e., more subpaths) comes an increase in the potential diversity within the graph from which transcript isoforms are generated. Therefore, we select a value of *k* by maximizing the number of nodes in the ensuing path graph (efficient selection of the *k* yielding the maximum numbers of nodes is done using a bisection algorithm implementation). In practice, we observed that for most problems, plotting the value of *k* versus the number of nodes in the path graph follows a gaussian distribution, wherein there is a k_{opt} that results in a maximum of potential subpaths to traverse when generating transcript isoforms.

Predict Isoforms

Isoforms are predicted using a dynamic programming algorithm that traverses the nodes of path graph to find the most highly expressed isoform in the following steps:

1. Predict isoform, *I*, and its expression, $EXPR_I$, that represents the most highly expressed path in the current path graph via dynamic programming.

a. The first isoform identified is designated I_{max}, with expression, $EXPR_{max}$. The absolute expression value at which to stop predicting additional isoforms is determined by multiplying $EXPR_{max}$ by the value of the command line parameter *--isoform-frac*.

2. Subtract $EXPR_I$ from all of the nodes contained within *I*
3. Repeat steps 1-2 to predict isoforms *I* and their expression values $EXPR_I$, terminating when $EXPR_I$ is less than $EXPR_{max}$ x *isoform-frac*

The dynamic programming algorithm is implemented as follows:

- For each node *i* in the path graph, initialize the following state variables:
    - Let *min_expr[i]* equal 0.0
    - Let *prev[i]* equal NULL
- For each node *i* in the path graph, ordered by topological sort:
    - For each successor node *j* of *i*:

- If (*prev[j]* is NULL) or *min*(*min_expr[i], expr[j]*) > *min_expr[j]:*
    - Let *min_expr[j]* equal *min*(*min_expr[i], expr[j]*)
    - Let *prev[j]* equal *i*

A traceback loop then reconstructs the highest expression isoform by starting at the *SINK* node and prepending nodes in the *prev* vector until reaching the *SOURCE*.

3.6 Acknowledgements

This work was supported in part by the NIH Prostate Specialized Program of Research Excellence grant P50CA186786 and the Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by the Prostate Cancer Foundation and the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor and a Taubman Scholar of the University of Michigan.

3.7 Author Contributions

M.K.I. designed the core of TACO method with assistance from Y.S.N, B.P., and H.K.I. The change point detection method was developed by Y.S.N. and M.K.I. Code optimization was performed by M.K.I., B.P., and Y.S.N. Performance benchmarking was performed by Y.S.N. Y.S.N., A.M.C., and M.K.I. wrote the manuscript. A.M.C. supervised development of the tool and guided the project to completion. All authors read and approved the final manuscript.

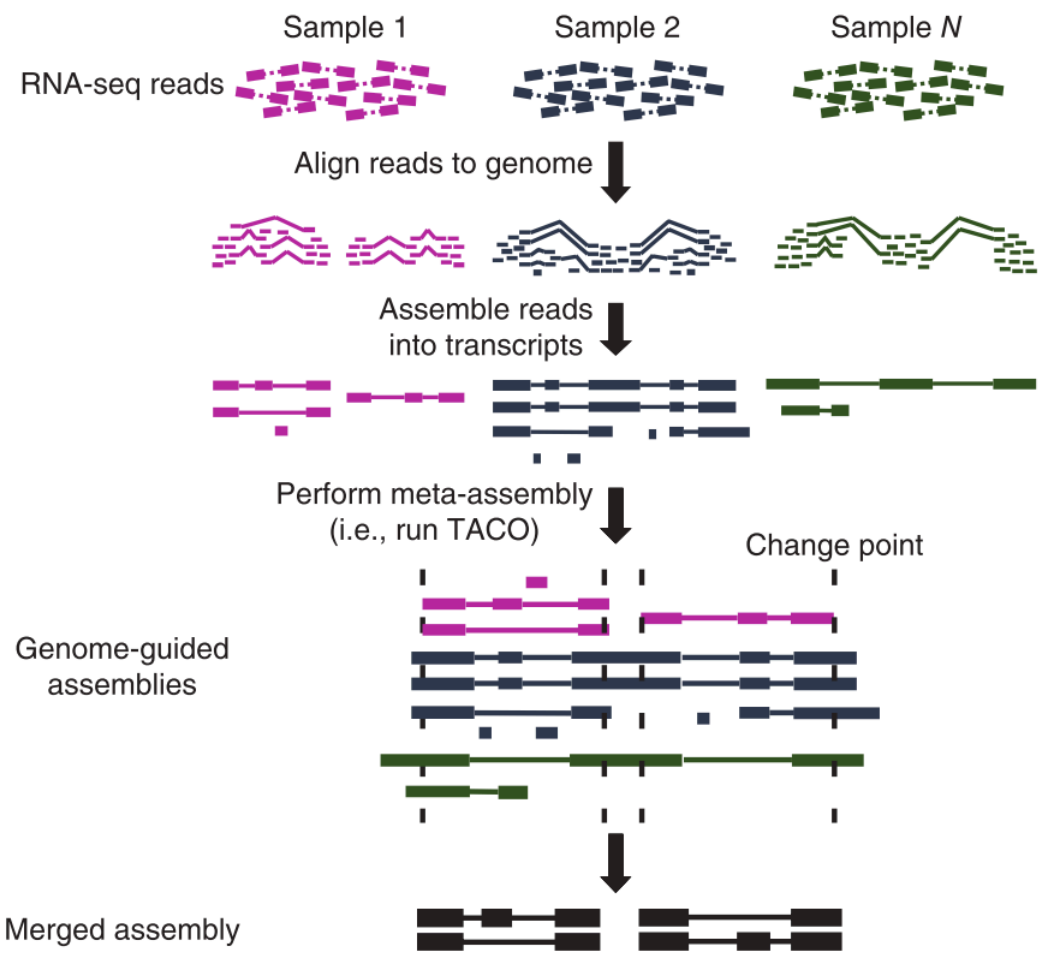


Figure 3.1 Schematic of meta-assembly workflow for TACO. Schematic detailing the transcriptome meta-assembly workflow for TACO. Reads are initially aligned to the genome. *Ab initio* assembly is then performed, generating a transcriptome assembly for each input sample. These transcriptome assemblies are then merged into a meta-assembly using the TACO tool, which leverages change point detection and a dynamic programming algorithm to generate robust transcript isoforms from the underlying network of splicing patterns.

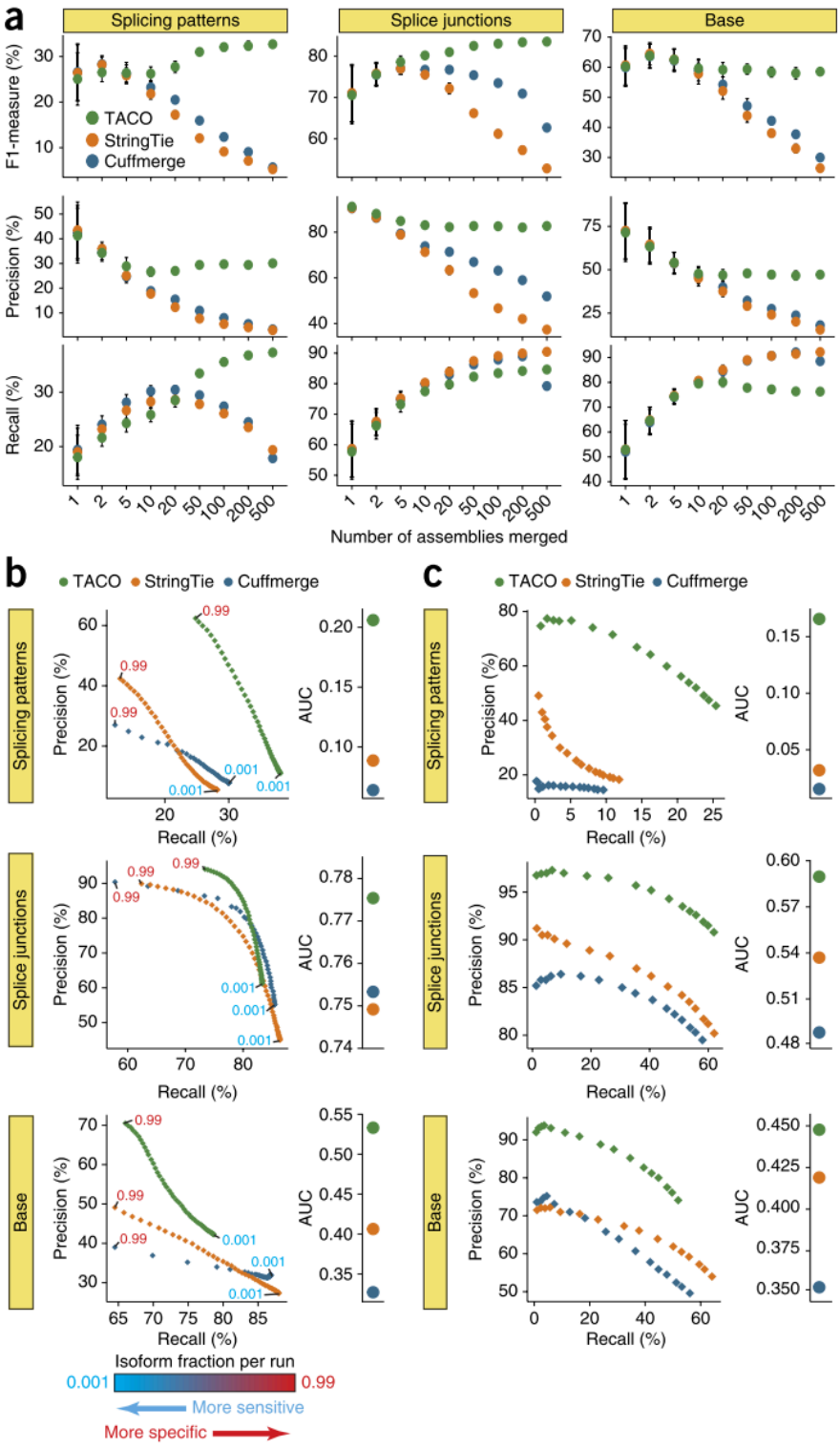


Figure 3.2 Assessment of TACO performance.

a. Performance metrics for TACO, Cuffmerge, and Stringtie when merging different numbers of input assemblies. Recall (i.e., sensitivity), precision and the F-measure for

all three tools were assessed for splicing patterns, splice junctions, and bases. Points represent the mean statistic across the 20 runs, error bars represent the 95% confidence interval. **b,c.** Precision-recall plots (left) and bar plots depicting the average precision (right) depicting performance for the three tools merging 55 CCLE breast cancer cell lines (**a**) at 50 different isoform fraction cutoffs ranging from 0.001-0.999, and (**b**) for the highest expressed transcripts in the meta-assemblies. Points represent statistics for the top *N* transcripts, with *N* ranging from 500-30,000.

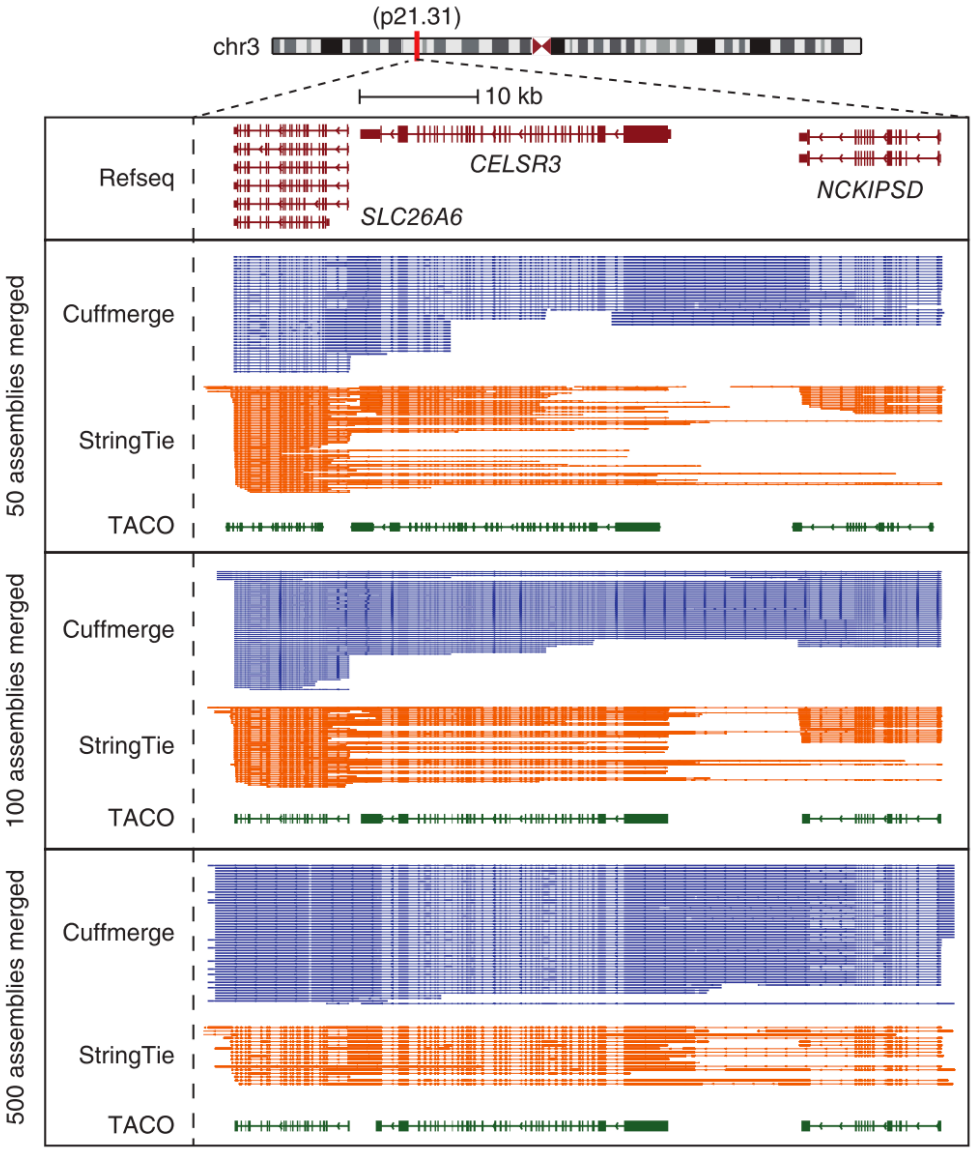


Figure 3.3 Examples of TACO performance.
Examples of TACO performance. The 3p21 genomic locus is depicted. Assembly of the genes SLC26A6, CELSR3, and NCKIPSD are shown for 50, 100, and 500 samples merged. The assembly produced by Cuffmerge is shown in blue, Stringtie in orange, and TACO in green. The Refseq reference annotation is shown above in red.

CHAPTER 4

The LncRNA Landscape of Breast Cancer Reveals a Role for DSCAM-AS1 in Breast Cancer Progression

Citation: Yashar S. Niknafs, Sumin Han, Teng Ma, Corey Speers, Chao Zhang, Kari Wilder-Romans, Matthew K. Iyer, Sethuramasundaram Pitchiaya, Rohit Malik, Yasuyuki Hosono, John R, Prensner, Anton Poliakov, Udit Singhal, Lanbo Xiao, Steven Kregel, Ronald F. Siebenaler, Shuang G. Zhao, Michael Uhl, Alexander Gawronski, Daniel F. Hayes, Lori J. Pierce, Xuhong Cao, Rolf Backofen, Cenk S. Sahinalp, James M. Rae, Arul M. Chinnaiyan, Felix Y. Feng. The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nature Communications* (2016)[60].

4.1 Abstract

Molecular classification of cancers into subtypes has resulted in an advance in our understanding of tumor biology and treatment response across multiple tumor types. However, to date, cancer profiling has largely focused on protein-coding genes, which comprise less than 1% of the genome. Here, we leverage a compendium of 58,648 long noncoding RNAs (lncRNAs) to subtype 947 breast cancer samples. We show that lncRNA-based profiling categorizes breast tumors by their known molecular subtypes in breast cancer. We identify a cohort of breast cancer-associated and estrogen-regulated lncRNAs, and investigate the role of the top prioritized ER-regulated

lncRNA, *DSCAM-AS1*. We demonstrate that *DSCAM-AS1* mediates tumor progression and tamoxifen resistance and identify hnRNPL as an interacting protein involved in the mechanism of *DSCAM-AS1* action. By highlighting the role of *DSCAM-AS1* in breast cancer biology and treatment resistance, this study provides insight into the potential clinical implications of lncRNAs in breast cancer.

4.2 Introduction

Long noncoding RNAs (LncRNAs) have recently been implicated in a variety of biological processes, including carcinogenesis and tumor growth[30,38,39,43,46,132]. Operating through a myriad of mechanisms[38], lncRNAs have challenged the central dogma of molecular biology as prominent functional RNA molecules. In order to investigate the role of lncRNAs in breast cancer, we interrogated the expression of lncRNAs across an RNA-sequencing (RNA-seq) breast tissue cohort comprised of 947 breast samples[68,133]. Previously, in a large-scale *ab initio* meta-assembly study from 6,503 RNA-seq libraries, we discovered approximately 45,000 of unannotated human lncRNAs[68], and this assembly was utilized for the present study. Building upon prior work that has begun to investigate the role of lncRNAs in breast cancer[134], we set out to perform a comprehensive analysis of breast cancer tissue RNA-sequencing data to identify the lncRNAs potentially involved in breast cancer.

Patients with estrogen receptor (ER) positive breast cancer have better prognosis than those with ER-negative disease, based on both a more indolent natural history but perhaps more importantly due to effective anti-estrogen, also designated "endocrine," therapy[135]. Despite the efficacy of endocrine therapy, however, the majority of breast cancer deaths occur in women with ER-positive breast cancers, because the incidence of ER-positive vs.-negative disease is much higher (approximately 80 vs. 20%), and

because a substantial fraction of women either have inherent or acquired endocrine therapy-resistant disease [136].

Taken together, these considerations highlight the pressing need to understand the biology of the ER-driven breast cancers and their mechanism of resistance to endocrine therapy. The mechanism through which ER mediates cancer initiation and progression is an area of intense scientific investigation[137–139]take , that remains incompletely understood. In this regard, while substantial research has been focused on ER abnormalities, such as mutations in the gene encoding for ER (*ESR1*)[139,140] and on the co-existing activation pathways that might mediate resistance, such as HER2[141], only a few cell line-focused studies have performed cursory analysis of ER-regulated non-coding RNAs[142–145]. Therefore, we set out to perform a comprehensive discovery and investigation of those lncRNAs that are driven by estrogen in breast cancers drawing from our large human tissue RNA-seq cohort.

4.3 Results

4.3.1 Identification of ER- and breast cancer-associated lncRNAs

We initially focused on those most differentially expressed in breast cancers in comparison to benign adjacent tissue, utilizing a non-parametric differential expression tool for RNA-seq called Sample Set Enrichment Analysis (SSEA)[68]. After applying an expression filter (at least 1 FPKM expression in the breast samples in the top 5% based on gene expression level), we identified 437 of the most differentially expressed lncRNAs in breast cancer. Interestingly, unsupervised hierarchical clustering of the samples based on expression of these lncRNAs across all breast cancer samples largely separated out the breast cancer samples by PAM50 subtypes[146,147], suggesting that

lncRNAs may be contributing to the distinct biology of these subtypes (**Figure 4.1a**). While lncRNA expression was unable to distinguish between the ER-driven luminal A and luminal B subtypes, the luminal subtypes were well separated from the HER2, basal, and normal subtypes (**Figure 4.1a**). In addition to separating out the clinical subtypes of breast cancer, the lncRNAs themselves separated into three distinct clusters. The first cluster (**Figure 4.1a**, "Luminal") contains lncRNAs overexpressed mostly in luminal A and luminal B samples, with little expression in samples of the other subtypes, and little expression in normal samples. The next cluster contains lncRNAs upregulated across all breast cancer samples (**Figure 4.1a**, "Upregulated"), and this cluster included the known breast cancer lncRNA, *HOTAIR*. The third cluster (**Figure 4.1a**, "Downregulated") contains lncRNAs downregulated in breast cancers. The lncRNAs in the luminal cluster present a particularly intriguing class of potentially estrogen-responsive lncRNAs.

Using the 947 breast tumor RNA-seq samples, we identified lncRNAs differentially expressed in ER-positive versus ER-negative breast tumors (**Figure 4.1b**). As expected, the expression of lncRNAs differentially expressed in ER-positive tumors separated the luminal tumors from the basal and HER2 upon unsupervised hierarchical clustering (**Figure 4.1b**). Quite interestingly, a number of lncRNAs that were downregulated in ER-positive samples exhibited increased expression in the basal samples (**Figure 1b**, "Basal lncRNAs"). While these basal lncRNAs were identified in an ER-positive versus ER-negative cancer analysis, a number of them also exhibit low expression in normal breast tissue (**Figure C.1**). Given that a paucity of known driver genes exist for basal breast cancers and that these tumors are the most clinically aggressive, these basal-specific lncRNAs may represent an exciting future area for basal breast cancer biology.

We set out to investigate potentially oncogenic ER-regulated lncRNAs by intersecting the lncRNAs upregulated in both the cancer versus normal (**Figure 4.1a**) and ER-positive versus ER-negative (**Figure 4.1b**) analyses. Sixty-three lncRNAs were upregulated in both the cancer versus normal analysis and the ER-positive versus ER-negative analysis (**Figure 4.1c**). To prioritize the most biologically and clinically relevant lncRNAs, we focused on lncRNAs most highly expressed in breast cancer tissues, and those most directly regulated by ER, based on ER binding to the targets' promoter as well as the degree of induction of expression following estrogen stimulation in breast cancer cells (**Figure 4.1c** and **Figure C.2a**). This approach nominated *DSCAM-AS1* as a lncRNA expressed at a very high level in breast cancer tissues, containing ER promoter binding, and exhibiting the strongest estrogen induction in MCF7 and T47D cells by both RNA-seq and qPCR validation (**Figure 4.1c** and **Figure C.2a**). We thus selected *DSCAM-AS1* for further investigation.

4.3.2 Characterization of *DSCAM-AS1*

DSCAM-AS1 has been previously reported to be involved in the proliferation of a luminal breast cancer cell line[145]. It exhibits a highly cancer-specific expression pattern, mostly in breast cancer and lung adenocarcinoma, in transcriptome sequencing data from a cohort of 6,503 cancer and normal tissues and cell lines from the TCGA and the Michigan Center for Translation Pathology (MCTP)[68] (**Figure 4.2a**). Supporting its association with ER biology, *DSCAM-AS1* expression is highly enriched (Student's t-test, p-value < 10e-5) in ER-positive tumors among the breast cancer samples in this RNA-seq cohort with ER-status determined by IHC (**Figure 4.2b**). Additionally, analysis of RNA-seq performed on 50 breast cancer cell lines[148] revealed that expression of *DSCAM-AS1* is highly specific to ER-positive cell lines (**Figure 4.2c** and **Figure C.2b**). Further supporting the association of ER with *DSCAM-AS1*, ESR1 ChIP-seq in both

MCF7 and T47D identified ER binding to the *DSCAM-AS1* promoter following estrogen stimulation (**Figure 4.2d**), and this finding was confirmed by ChIP-qPCR of the *DSCAM-AS1* promoter (**Figure C.2c**). The isoforms of *DSCAM-AS1* in MCF7 cells were identified using 3′ and 5′ RACE (**Figure 4.2d** and **Table C.1**). *DSCAM-AS1* expression is induced in both MCF7 and T47D cells after estrogen stimulation, and this induction is reversed with addition of tamoxifen, corroborating that ER is in fact regulating the expression of this lncRNA (**Figure 4.2e**). Expression of known ER-regulated protein-coding genes *GREB1* and *PGR* follow the same pattern of response to estrogen, while the lncRNA *MALAT1*, serving as a negative control, is not induced by estrogen (**Figure 4.2e**). In addition to being estrogen-responsive, *DSCAM-AS1* expression is present in both the cytoplasm and nucleus at nearly identical fractions in both MCF7 and T47D cells (**Figure C.2d**), and the identity of *DSCAM-AS1* as a non-coding gene was confirmed using the CPAT tool[75] (**Figure C.2e**). We used smFISH to further dissect the subcellular localization and gene expression levels of DSCSM-AS1 in breast cancer cells. To this end, we designed probes that targeted all potential isoforms of the transcript, the sequences predicted by RACE. Upon staining, we found that each MCF7 cell expressed ~800 copies of the DSCAM-AS1 transcript, almost half as much as the expression level of GAPDH (**Figure C.2f,g**). Moreover the lncRNA was almost equally distributed between the nucleus and cytosol (~42% nuclear, **Figure C.2h**), whereas GAPDH was predominantly cytoplasmic (~7% nuclear, **Figure C.2h**). While the abundance of DSCAM-AS1 was lower in T47D cells (~260 molecules / cell, **Figure C.2i,j**), the relative expression level (compared to GAPDH) and the subcellular localization pattern were very similar to those observed in MCF7 cells (**Figure C.2k**).

4.3.3 *DSCAM-AS1* is implicated in cancer aggression

We next investigated the clinical relevance of *DSCAM-AS1*. Given that *DSCAM-AS1* is a lncRNA, its expression is not measured by most traditionally used microarrays, which are the primary high-throughput platforms annotated with reliable clinical outcomes in breast cancer[94]. As a surrogate, we employed a guilt-by-association analysis to interrogate the clinical relevance of those genes most correlated to *DSCAM-AS1*. Given that *DSCAM-AS1* is an ER-regulated lncRNA, correlation was performed using only ER-positive breast cancers, to ascertain clinical relevance in the breast cancer samples in which *DSCAM-AS1* would be enriched and most relevant. We obtained a number of breast cancer clinical datasets from Oncomine[149] containing gene expression sets associated with the presence of cancer (versus normal tissue), high clinical grade, recurrence, survival, and metastasis[94,146,147,149–162]. We assessed for the overlap between these gene sets with the genes most positively or negatively correlated to *DSCAM-AS1*. *DSCAM-AS1* positively correlated genes were significantly associated with clinical signatures associated with increased cancer aggression, tamoxifen resistance, higher grade, stage, and metastasis (**Figure 4.3a,b**). Similarly, the *DSCAM-AS1* negatively correlated genes associated with clinical signatures that portended a more favorable clinical outcome (**Figure C.3a,b**). For many of the clinical concepts, *DSCAM-AS1*-positively correlated genes displayed a clinical association comparable to those genes most correlated to *EZH2*, a gene known to be a marker of clinical aggressiveness in breast cancer[163], while genes correlated to other lncRNAs expressed in breast tissue, such as *HOTAIR, MALAT1, and NEAT1*, showed modest-to-no association (**Figure 4.3b** and **Figure C.3b**). Additionally, performing a Gene Set Enrichment Analysis (GSEA)[164] on all genes correlated to *DSCAM-AS1* yielded significant association with a myriad of breast cancer, cancer aggressiveness, and ER- and tamoxifen-associated gene signatures (**Figure C.3c**). While ER-positive breast cancers typical portend better clinical outcomes[147], among the luminal breast cancers, *DSCAM-AS1* is expressed significantly higher in luminal B, a clinical subtype containing most of the clinically aggressive ER-

positive breast cancers[146,147] (**Figure 4.3c**). Despite these associations of clinical aggression with DSCAM-AS1, in a survival analysis of the ER-positive TCGA breast samples, expression of DSCAM-AS1 was not significantly associated with clinical outcome (**Figure C.3d**). Definitive assessment of survival in this cohort, however, will likely require more robust and longer-term clinical curation of the TCGA breast samples.

We then studied the role of *DSCAM-AS1* on oncogenic phenotypes in ER-positive breast cancer cell lines. In MCF7 and T47D cells, stable knockdown of *DSCAM-AS1* was achieved using shRNA approaches. *DSCAM-AS1* knockdown reduced the proliferative ability of both cell lines (**Figure 4.3d**), diminished the ability of these cells to invade in a Boyden chamber invasion assay (**Figure 4.3e**), and substantially abolished the ability of these cells to form colonies in soft agar (**Figure 4.3f**). While ER regulates levels of *DSCAM-AS1*, *ER* expression and protein levels are not dependent on level of *DSCAM-AS1* (**Figure C.4a**), ruling out the possibility that the phenotype observed could be explained through changes in the level of *ER*. Additionally, knockdown of *DSCAM-AS1* exhibited no affect on RNA or protein levels of the *DSCAM* gene, in which *DSCAM-AS1* resides antisense and intronic (**Figure C.4b**). To further demonstrate the impact of *DSCAM-AS1* on aggressive cancer phenotypes, we overexpressed *DSCAM-AS1* in T47D (**Figure C.4c**) and ZR75-1 (**Figure C.4d**), two ER-positive breast cancer cell lines with moderate *DSCAM-AS1* expression (**Figure 4.2c**), and observed an increase in the invasion phenotype (**Figure 4.3g** and **Figure C.4e**). MCF7 cells were not included in the overexpression studies as *DSCAM-AS1* is already expressed at a very high level in these cells (**Figure 4.2c**). Overexpression was also tested in MDA-MB-231 cells (**Figure C.4f**), a common ER-negative cell line, however, exogenous *DSCAM-AS1* was unable to confer oncogenicity via proliferation (**Figure C.4g**) and invasion (**Figure C.4h**). This phenomenon may be explained by a requisite genetic and epigenetic milieu provided by ER-positive cells in order for *DSCAM-AS1* to confer its cancer phenotype, and more

investigation into the precise mechanisms through which it acts will shed light on this finding. Furthermore, the simple presence of DSCAM-AS1 alone is not sufficient to make cells highly aggressive, as evidenced by its high expression in ER-positive cell lines that are moderately invasive (e.g., MCF7). To further characterize the impact of *DSCAM-AS1* on cacner phenotype, we performed a mouse xenograft tumor growth assay, showing that loss of *DSCAM-AS1* reduces the growth of implanted T47D cells *in vivo* (**Figure 4.3h**). The metastatic potential of these implanted cells were also reduced with *DSCAM-AS1* knockdown, as evidenced through decreased liver metastasis following xenograft (**Figure 4.3i**).

4.3.4 Role of hnRNPL in *DSCAM-AS1* mechanism

LncRNAs have been shown to be functional through their binding interactions with other RNAs, DNA, and with proteins[38]. Thus, identifying protein binding partners for *DSCAM-AS1* is a crucial step in determining the mechanism through which it confers oncogenicity. To identify *DSCAM-AS1* binding partners, we performed pull-down of *DSCAM-AS1* and performed mass spectrometry on the pull-down product to identify proteins bound to *DSCAM-AS1* (**Figure 4.4a**). The protein hnRNPL was observed to have the highest spectral counts for the sense form of *DSCAM-AS1* with zero spectral counts in the antisense pull-down (**Figure 4.4b**). Additionally, PCBP2, a protein known to complex with hnRNPL[165], was also among the top proteins bound to *DSCAM-AS1*. We thus investigated the interaction between *DSCAM-AS1* and hnRNPL further. HnRNPL is a protein widely expressed in many tissue types (**Figure C.5a**) and has been implicated in regulating RNA stability and processing with subsequent effects on gene expression[166–169]. The binding of hnRNPL to *DSCAM-AS1* was confirmed by RNA pulldown followed by Western blot, with no binding of hnRNPL to the negative control antisense transcript (**Figure 4.4c**). Other RNA-binding proteins did not bind *DSCAM-*

AS1, however, suggesting that *DSCAM-AS1* does not promiscuously bind to RNA-binding proteins in general (**Figure 4.4c**). To further confirm this binding interaction and its specificity, RNA-immunoprecipitation (RIP) was performed with using antibodies directed against hnRNPL. *DSCAM-AS1* was highly enriched by anti-hnRNPL RIP in both MCF7 and T47D cells, while control coding and non-coding genes exhibited modest binding (**Figure 4.4d**). Additionally, anti-snRNP70 and anti-HuR RIP failed to pull down *DSCAM-AS1*, further suggesting the specificity of the *DSCAM-AS1*-hnRNPL interaction (**Figure C.5b**).

To more specifically investigate the functional relationship of *DSCAM-AS1* and hnRNPL, we performed rescue studies assessing the impact of *hnRNPL* knockdown on the invasive advantage conferred by *DSCAM-AS1* overexpression, observing that reduction of *hnRNPL* levels entirely reversed the increase in invasion observed upon *DSCAM-AS1* overexpression (**Figure 4.4e** and **Figure C.6a**). Although the reduction in invasiveness was not significant in the control cells with *hnRNPL* knockdown, there appears to be a slight reduction in invasive capacity upon knockdown of *hnRNPL*. Thus, the dramatic reduction in invasion observed in the DSCAM-AS1 overexpressing cells with *hnRNPL* knockdown may be the result of *hnRNPL* affecting invasion in a mechanism exclusive to *DSCAM-AS1*. So, in order to further characterize the functional relationship between *DSCAM-AS1* and hnRNPL, we set out to localize the binding site of hnRNPL within the *DSCAM-AS1* lncRNA. Using *in silico* prediction drawing from prior studies of hnRNPL crosslinking-immunoprecipitation sequencing (CLIP-seq)[170], a single strong predicted binding peak was identified near the 3' end of *DSCAM-AS1* (**Figure 4.4f**). HnRNPL has been shown to bind CACA-rich RNA sites[167], and the predicted binding region possessed a 10 base pair CACA stretch. To identify if this predicted region does in fact account for the hnRNPL binding, multiple mutant forms of *DSCAM-AS1* were created with or without the binding site. DSCAM-AS1-5 and

DSCAM-AS1-3 are large deletion mutants containing only the 5′ and 3′ end, respectively, with only DSCAM-AS1-3 possessing the predicted binding site, and DSCAM-AS1-D is a mutant form with the 27 nucleotides comprising the predicted binding site deleted (**Figure 4.4f,g**, red). The various mutant forms of *DSCAM-AS1* were expressed in HEK293, a cell line which lacks endogenous *DSCAM-AS1* expression while still expressing hnRNPL (**Figure C.6b**). While both the full length and DSCAM-AS1-3 mutant retained hnRNPL binding, loss of the predicted binding region was effective in abrogating hnRNPL binding via both Western blot following RNA pull-down (**Figure 4.4g**) and by qPCR following hnRNPL RIP (**Figure C.6c**). All deletion mutants were expressed at comparable levels, ruling out the possibility of falsely diminished binding due to failed expression of the mutant construct (**Figure C.6d**). RNA-secondary structure is a crucial component of RNA functionality and is a key player in RNA-protein interactions. While the 27 nucleotide deletion in the DSCAM-AS1-D mutant is a small fraction of the total number of bases in the transcript, to ensure that this deletion was not causing a dramatic RNA secondary structure change, we investigated the impact of this deletion on RNA-secondary structure via the RNAfold structure prediction tool[171]. Evidenced by a minimal free energy prediction, the posited secondary structure of *DSCAM-AS1* is largely similar to that of DSCAM-AS1-D (**Figure C.6e**), suggesting that the loss of hnRNPL binding observed with the DSCAM-AS1-D mutant is not due to a dramatic secondary structure rearrangement. Quite interestingly, overexpression of the DSCAM-AS1-D mutant in T47D cells failed to recapitulate the increased invasion observed when overexpressing full length *DSCAM-AS1* (**Figure 4.4h**). This finding, in combination with the rescue studies following *hnRNPL* knockdown (**Figure 4.4e**), strongly suggest that *DSCAM-AS1* promotes oncogenicity via its interaction with hnRNPL in these ER-positive breast cancer cells.

4.3.5 Role of *DSCAM-AS1* in tamoxifen resistance

A substantial number of patients with ER-positive breast cancer eventually develop resistance to endocrine therapy and present with clinical recurrence and metastasis[136,172,173]. Thus, as *DSCAM-AS1* is implicated in poor-prognosis ER-positive breast cancer (**Figure 4.3a-c** and **Figure C.3**), we set out to investigate its potential role in subverting estrogen dependence and promoting resistance to anti-estrogen therapies. We continuously passaged MCF7 cells in 1uM tamoxifen for 6 months until we attained a subpopulation of MCF7 cells that were able to grow in in tamoxifen and termed these tamoxifen resistant MCF7 cells (TamR-MCF7). Interestingly, although expression of canonical ER targets (*GREB1* and *PGR*) was decreased compared to the parental MCF7 cells, *DSCAM-AS1* expression was significantly upregulated despite already being expressed at very high levels in MCF7 cells (**Figure 4.5a**). The levels of *ER* were also increased, which is likely a compensatory upregulation in response to the continual anti-estrogen effects of tamoxifen. Additionally, short-term tamoxifen treatment of parental MCF7 cells transiently reduced *DSCAM-AS1* levels at 8hrs following tamoxifen treatment, with a rise back pre-treatment levels after 24 hours (**Figure C.7a**). In contrast, canonical ER target, *GREB1*, exhibited pronounced expression reduction at both the short- and long-term timescale (**Figure C.7b**). To interrogate whether this upregulation of *DSCAM-AS1* in the TamR-MCF7 cells is functionally significant, we assessed the proliferative capacity of these cells following *DSCAM-AS1* knockdown. With knockdown levels of *DSCAM-AS1* comparable to the endogenous levels in parental MCF7 cells (**Figure C.7c**), knockdown of *DSCAM-AS1* in TamR-MCF7 cells led to a loss of their baseline proliferative advantage when cultured in tamoxifen, exhibiting a proliferation profile nearly identical to that of the parental MCF7 cells (**Figure 4.5b**). Additionally, knockdown of *hnRNPL* in these cells produced a similar loss of proliferative capacity in the TamR-MCF7 cells (**Figure C. 7d,e**), suggesting that both

DSCAM-AS1 and hnRNPL may be playing a role in promotion of the tamoxifen resistance developed by these cells.

We also interrogated the ability of *DSCAM-AS1* to confer tamoxifen resistance in native T47D cells via overexpression of *DSCAM-AS1*. *DSCAM-AS1* overexpression was also associated with tamoxifen-resistant growth in a dose-dependent manner, with a striking increase in cell viability at levels of tamoxifen as low was 100nM (**Figure 4.5c**). Additionally, in line with the ability of *DSCAM-AS1* to provide estrogen-independent growth advantage, cells overexpressing *DSCAM-AS1* also exhibited a proliferative advantage when grown in estrogen-deprived medium compared to normal serum (**Figure C.7f**). This growth advantage was abolished with the addition of estrogen, and returned with the subsequent addition of tamoxifen (**Figure C.7f**). Conversely corroborating the relationship of *DSCAM-AS1* on estrogen dependence in these cells, we witnessed an increased estrogen dependence of T47D cells following *DSCAM-AS1* knockdown (**Figure C.7g**).

To corroborate our *in vitro* findings in a tissue model, we obtained data previously generated performing ChIP-seq for ER in primary and metastatic breast tumor tissue[138]. These tumors were grouped into the following categories as previously described[138]: primary ER-negative (n=2), primary ER-positive, tamoxifen-responder (n=8), primary ER-positive, tamoxifen-non-responder (n=7), metastatic ER-positive (n=3). Strikingly, investigation of the *DSCAM-AS1* promoter revealed that ER preferentially binds to the *DSCAM-AS1* promoter in tumors with clinical aggression (i.e., metastatic and tamoxifen non-responders) (**Figure 4.5d**), while a canonical ER target, *GREB1*, exhibits ER-biding to its promoter in nearly all ER-positive tumors, lacking preference for the more clinically aggressive tumors. All together, these data suggest that the association

between *DSCAM-AS1* expression with clinical aggressiveness in ER-positive breast cancer samples may be explained, in part, by the ability of *DSCAM-AS1* to facilitate estrogen-independent oncogenicity, thus potentially promoting resistance to endocrine therapy with tamoxifen.

4.4 Discussion

Further investigation and study of the mechanisms through which ER-dominant breast cancers become aggressive and eventually evade traditional clinical therapies is of intense clinical interest. In this study, we identify a myriad of potentially ER-associated lncRNAs, and functionally and mechanistically characterize one of the most intriguing candidates. Nevertheless, further investigation of some of these other lncRNAs may also contribute to our understanding of ER biology and ER-driven oncogenesis. LncRNAs have been shown to function through multiple mechanisms, and the study of the interaction of *DSCAM-AS1* with hnRNPL is a promising step towards understanding the ways through which this molecule executes its oncogenic function. While we show that the binding of hnRNPL to *DSCAM-AS1* is responsible for at least some of its oncogenicity, a further understanding of how the interaction between hnRNPL and *DSCAM-AS1* is mediating this phenotype is necessary.

Novel mediators of tumor aggression, such as *DSCAM-AS1*, can provide insight into the mechanism of endocrine therapy resistance. This increased understanding may in turn lead to more effective strategies to overcome this resistance, which is one of the last, great clinical challenges in treatment of ER-positive breast cancer. Additionally, there is little known regarding the role of non-coding RNAs in developing resistance to anti-estrogen therapy, with a small number of studies implicating some of the more prominent, well characterized breast cancer lncRNAs[174,175]. *DSCAM-AS1* is just one of many potentially relevant ER-regulated lncRNAs in breast cancer, and further

investigation of the other candidates is likely to yield a greater understanding of ER-mediated cancer biology. Ultimately, this study provides key insight into the role of lncRNAs in ER breast cancer biology, and is an important step in better understanding this common disease.

4.5 Methods

4.5.1 Cell lines and cell culture

All cell lines were obtained from the American Type Culture Collection (Manassas, VA). Cell lines were maintained using standard media and conditions. Specifically, T47D cells were maintained in Roswell Park Memorial Institute (RPMI) 1640 medium supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin and 5 mg/ml insulin. ZR75-1 cells were maintained in RPMI 1640 medium supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin. HEK293 cells were maintained in DMEM plus 10% fetal bovine serum (FBS) plus 1% penicillin-streptomycin. MCF7 cells were cultured in Dulbecco's modified Eagle's medium plus GlutaMAX (DMEM, Invitrogen) containing 10% fetal bovine serum (Hyclone) and 1% penicillin-streptomycin. To establish the tamoxifen resistant cell line, MCF7 cells were grown in IMEM phenol-red free medium with10% Charcoal-stripped FBS in the presence of 1 uM (Z)-4-hydroxytamoxifen (Sigma) for 6 months. All cell lines were grown at 37°C in a 5% CO_2 cell culture incubator, and were genotyped for identity at the University of Michigan Sequencing Core and tested routinely for *Mycoplasma* contamination.

4.5.2 Cell proliferation assay

Cells were seeded in a 48-well plate at $3x10^4$ cells per well. Plates were added to Incucyte machine (Essen Bioscience) 16-20 hours following seeding. Growth curves were constructed by imaging plates using the Incucyte system, where the growth curves are generated from confluence measurements acquired during continuous kinetic imaging. Four wells were measured per condition. For tamoxifen treatment, 16-20 hours after seeding, the medium was changed to RPMI phenol-red free medium containing 10% charcoal-treated FBS in the presence of 1 uM tamoxifen or ethanol. Growth curves were obtained using Incucyte system as described above.

4.5.3 Cell viability assay

Cells were seeded in 96-well plates at 5000 cells per well in a total volume of 100 μl media containing 10% FBS. Serially diluted tamoxifen in 100 μl of media was added to the cells 12 hours after seeding. Medium containing tamoxifen was replenished every 2-3 days. Following 10 days of incubation, cell viability was assessed by WST assay (WST-8, Dojindo). All assays were performed in triplicate and repeated at least three times. The relative cell viability was expressed as a percentage of the control that was treated with vehicle solutions.

4.5.4 Soft agar colony formation assay

10000 cells were suspended in medium containing 0.3 % agar, 10 % FBS, and layered on medium containing 0.6% agar and 10% FBS in 6-well plate. Colonies were stained for 18-24 hours with iodonitrotetrozolium chloride (Sigma #18377) following 3 weeks of incubation. Colonies from three replicate wells were quantified.

4.5.5 Quantitative RT-PCR assay

RNA was isolated from cell lysates by the miRNeasy mini kit and cDNA was synthesized from 1 μg RNA using SuperScript III (Invitrogen) and Random Primers (Invitrogen), per the manufacturer's protocol. qRT-PCR was carried out on the ABI7900 HT Fast Real time system (Applied Biosystems) using gene-specific primers designed with the Primer3 software and synthesized by IDT Technologies. qRT-PCR data were analyzed using a relative quantification method and plotted as average fold change compared to the control. GAPDH and actin were used as an internal reference. All primers used for qPCR are detailed in **Table C.2**. Three technical replicates were utilized in each assay, and all data shown was performed with at least three biological replicates.

4.5.6 Estrogen and tamoxifen treatment

To evaluate the effect of estrogen stimulation, cells were first hormone depleted via growth in phenol-red free medium containing 10% charcoal-treated FBS for 72 hours and then treated with ethanol vehicle, 10nM β-estradiol, or 10nM β-estradiol plus 1uM tamoxifen. After 10 hours, RNA was isolated as described above and qPCR was performed as described above using Power SYBR Green Mastermix (Applied Biosystems).

4.5.7 Subcellular fractionation

Cellular fractionation was performed using a RiboTrap Kit (MBL International), according to manufacturer's instructions. RNA was isolated and qRT-PCR was performed as described above.

4.5.8 Knockdown and overexpression studies

Knockdown of *DSCAM-AS1* and *hnRNPL* in T47D and MCF7 cells was accomplished by siRNA from Dharmacon. Transfections were performed with OptiMEM (Invitrogen) and RNAi Max (Invitrogen) per manufacturer instruction. Target sequences used for shRNA or siRNA knock-down are listed in **Table C.3**. For stable knockdown of *DSCAM-AS1*, MCF7 and T47D cells were transfected with lentiviral constructs containing 2 different *DSCAM-AS1* shRNAs or no targeting shRNAs in the presence of polybrene (8 μg/mL; **Table C.3**). After 48 hours, transduced cells were grown in culture media containing 2 μg/mL puromycin.

For *DSCAM-AS1* overexpression, the predominant isoform (isoform 2, **Table C.1**) was cloned into the pLenti6.3 vector (Invitrogen) using PCR8 non-directional Gateway cloning (Invitrogen) as an initial cloning vector and shuttling was then done to pLenti6.3 using LR clonase II (Invitrogen) according to the manufacturer's instructions. As control, *LacZ* was also cloned into the same vector system. The primers for making *DSCAM-AS1* mutation and truncations are listed in **Table C.2**. Lentiviral particles were made and T47D and ZR75.1 cells were transduced as described above. Stable cell lines were generated by selection with 3μg/ml blasticidin. Transient transfection of *DSCAM-AS1* and its derivative mutants was done in HEK293 cells was performed with Lipofectamine LTX (Invitrogen) following manufacturer's instruction. Cells were collected at 48 hours post-transfection.

4.5.9 *In vitro* RNA binding assay

The RNA binding assay was performed according to the protocol of the RiboTrap Kit (MBL International). In brief, 5-bromo-UTP (BrU) was randomly incorporated into sense *DSCAM-AS1*, antisense *DSCAM-AS1*, and *LacZ* control via PCR-based transcription. The primers are shown in **Table C.2**. The the BrU labelled RNA

transcripts were bound to beads conjugated with anti-BrdU antibodies. Then, the cytoplasmic or nuclear fractions from MCF7 or T47D cells were mixed for 2 hours. Samples were washed four times with Wash Buffer II prior to elution. The samples were sent to the Michigan Center for Translational Pathology (MCTP) proteomic core facility for mass spectrometry.

4.5.10 Mass spectrometry

The samples were treated with SDS-PAGE loading buffer supplied with 10 mM DTT for 5 min at 85ºC. The proteins were alkylated by the addition of iodoacetamide to the final concentration of 15 mM. The samples were subjected to SDS-PAGE and the whole lanes were cut out and digested with trypsin in-gel for 2 hours. The resulting peptides were extracted, dried and resuspended in 0.1% formic acid with 5% acetonitrile prior to loading onto a 2 cm EASY-column (Thermo Scientific) coupled to an in-house made nano HPLC column (20 cm x 75 um) packed with LUNA C18 media (Phenomenex). Analysis was performed on a Velos Pro mass spectrometer (Thermo Scientific) operated in data-dependent mode using 120-min gradients in EASY-LC system (Proxeon) with 95% water, 5% acetonitrile (ACN), 0.1% formic acid (FA) (solvent A), and 95% ACN, 5% water, 0.1% FA (solvent B) at a flow rate of 220 nl/min. The acquisition cycle consisted of a survey MS scan in the normal mode followed by twelve data-dependent MS/MS scans acquired in the rapid mode. Charge state was not recorded. Dynamic exclusion was used with the following parameters: exclusion size 500, repeat count 1, repeat duration 10 s, exclusion time 45 s. Target value was set at 10^4 for tandem MS scan. The precursor isolation window was set at 2 m/z. The complete analysis comprised two independent biological replicates.

4.5.11 Mass spectrometry data analysis

The resulting spectrum files were transformed into MGF format by MSConvert software and interrogated by MASCOT 2.4 search engine using human UniProt database version 15 concatenated with reverse sequences for estimation of false discovery rate (FDR) and with a list of common contaminants (40729 entries in total). The search parameters were as follows: full tryptic search, 2 allowed missed cleavages, peptide charges +2 and +3 only, MS tolerance 1 Da, MS/MS tolerance 0.5 Da. Permanent post-translational modification was: cysteine carbamidomethylation. Variable post-translational modifications were: protein N-terminal acetylation, Met oxidation and N-terminal Glutamine to pyro-Glutamate conversion. The remaining analysis was performed as previously described[176]. To summarize, the minimal ion score threshold was chosen such that a peptide FDR below 1% was achieved. The peptide FDR was calculated as: 2 × (decoy_hits)/(target + decoy hits). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium[177] via the PRIDE partner repository with the dataset identifier PXD002421 and 10.6019/PXD002421. Spectral counts for all detected proteins were assembled using an in-house written Python script. The adjustment of spectral counts was done as previously described[176].

4.5.12 RNA immunoprecipitation

RIP assays were performed using a Millipore EZ-Magna RIP RNA-Binding Protein Immunoprecipitation kit (Millipore, #17-700) according to the manufacturer's instructions. RIP-PCR was performed as qPCR, as described above, using total RNA as input controls. 1:150th of RIP RNA product was used per PCR reaction. Antibodies used for RIP are listed in **Table C.4**. All RIP assays were performed in biological duplicate.

4.5.13 Invasion assay

3×10^5 cells were seeded in a 24 well corning FluoroBlok chamber pre-coated with Matrigel (BD Biosciences). Medium containing 10% FBS in the lower chamber served as chemoattractant. After 48 hours, cells remaining on the lower side of the membrane were stained with calcein AM (C34852 invitrogen). The invasive cells adhering to the bottom surface of the filter were quantified under a fluorescent microscope (×2).

4.5.14 Antibodies and immunoblot analyses

Cell lysates were separated on 4–12% SDS polyacrylamide gels (Novex) and transferred onto nitrocellulose membrane (Novex) using wet transfer (30 V overnight). The membrane was incubated for 1 hour in blocking buffer (Tris-buffered saline, 0.1% Tween (TBS-T), 5% nonfat dry milk) and incubated overnight at 4 °C with indicated antibodies. Blots were developed using enhanced chemiluminescence (ECL Prime) per the manufacturer's protocol. All the antibodies used in this study are described in **Table C.4**. Representative full blot images are shown in **Figure C.8**.

4.5.15 Chromatin immunoprecipitation (ChIP)

The ChIP assays were performed using HighCell# ChIP kit (Diagenode) according to the manufacturer's protocol. Briefly, MCF7 cells were grown in charcoal-stripped serum media (described above) for 72 hours followed by stimulation with 10nM estradiol for 12 hours. Next, cells were cross-linked for 10 min with 1% formaldehyde. Cross-linking was terminated by the addition of a one-tenth volume of 1.25 M glycine for 5 min at room temperature, followed by cell lysis and sonication (Bioruptor, Diagenode), resulting in an average chromatin fragment size of 300bp. Chromatin equivalent of 5 ×

10^6 cells was done for ChIP using different antibodies. ChIP DNA was isolated (IPure Kit, Diagenode) from samples by incubation with the antibody at 4°C overnight followed by wash and reversal of cross-linking.

4.5.16 ChIP-seq library construction and sequencing analysis

DNA was purified for library preparation using the IPure Kit (Diagenode). The ChIP-seq sample preparation for sequencing was performed according to the manufacturer's instructions (Illumina). ChIP-enriched DNA samples (1–10ng) were converted into blunt-ended fragments using T4 DNA polymerase, *E. coli* DNA polymerase I large fragment (Klenow polymerase) and T4 polynucleotide kinase (New England BioLabs (NEB)). A single adenine base was added to fragment ends by Klenow fragment (3′ to 5′ exo$^-$; NEB) followed by ligation of Illumina adaptors (Quick ligase, NEB). The adaptor-modified DNA fragments were enriched by PCR using the Illumina Barcode primers and Phusion DNA polymerase (NEB). PCR products were size selected using 3% NuSieve agarose gels (Lonza) followed by gel extraction using QIAEX II reagents (QIAGEN). Libraries were quantified with the Bioanalyzer 2100 (Agilent) and sequenced on the Illumina HiSeq 2000 Sequencer (100 nucleotide read length). ChIP-seq data were mapped to human genome version hg19 using BWA[178]. The MACS program[179] was used to generate coverage map files to visualize the raw signal on the UCSC genome browser[74]. Hpeak[180], a hidden Markov model (HMM)-based peak-calling software program designed for the identification of protein-interactive genomic regions, was used for ChIP-seq peak determination.

4.5.17 ChIP-seq peak promoter overlap

Overlap of ChIP-seq peaks with gene promoters was performed using the BEDTools "coverage" tool. Intervals of +/- 5-10 kilobases surrounding unique transcriptional starts were used to assess promoter overlap.

4.5.18 Coding potential scoring

Coding potential for all lncRNA transcripts was determined as described previously[43]. The alignment-free Coding Potential Assessment Tool (CPAT)[75] was used to determine coding probability for each transcript. CPAT determines the coding probability of transcript sequences using a logistic regression model built from ORF size, Fickett TESTCODE statistic, and hexamer usage bias.

4.5.19 Xenograft analysis

All experimental procedures were approved by the University of Michigan Committee for the Use and Care of Animals (UCUCA) and conform to all regulatory standards. . A total of 5×10^6 cells of T47D control or T47D shM41 cells were suspended in 100 uL of PBS/Matrigel (1:1) were injected subcutaneously in 5-week-old pathogen-free female CB-17 severe combine immunodefiecient mice (CB-17 SCID) which simultaneously received a 60-day slow release pellet containing 0.18 mg of 17b-estradiol (Innovative Research of America). Tumors were measured weekly using a digital caliper. Growth in tumor volume was recorded using digital calipers and tumor volumes were estimated using the formula $(\pi/6)$ $(L \times W^2)$, where L = length of tumor and W = width. Additionally, mouse livers were harvested to determine spontaneous metastasis by measuring human-Alu sequence. Briefly, genomic DNA from livers were prepared using Puregene DNA purification system (Qiagen), followed by quantification of human Alu sequence by human Alu specific Fluorogenic Taqman qPCR probes.

4.5.20 RNA-seq data processing

Sequence quality control was done using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Next, reads mapping to mitochondrial DNA, ribosomal RNA, poly-A, poly-C, Illumina sequencing adaptors, and the spiked-in phiX174 viral genome were filtered. Sequences were downloaded from the Illumina iGenomes server (2012, March 9). Mapping was performed using bowtie2 (2.0.2). Reads were mapped using TopHat2 (2.0.6 and 2.0.8) using default parameters. A human genome reference was constructed from UCSC version Feb 2009 (GRCh37/hg19) chromosomes 1-22, X, Y, and mitochondrial DNA, and references from alternate haplotype alleles were omitted. Bowtie-build and bowtie2-build were used to build genome reference for Bowtie versions 0.12.8 and 2.0.2 were, respectively. The Ensembl version 69 transcriptome was used as a reference gene set. Using the --transcriptome-index option in TopHat version 2.0.6[181], alignment index files were prepared from this reference for Bowtie versions 0.12.8 and 2.0.2.

4.5.21 RNA-seq transcript expression estimation

Cufflinks version 2.1.1[51] was used with the following parameters to estimate transcript abundance from RNA-seq data: '--max-frag-multihits=1', '--no-effective-length-correction', '--max-bundle-length 5000000', '--max-bundle-frags 20000000'. To convert FPKM abundance estimates (generated by Cufflinks) to approximate fragment count values we multiplied each FPKM by the transcript length (in kilobases) and by the "Map Mass" value (divided by 1.0e6) found in the Cufflinks log files.

4.5.22 Breast cancer tissue expression heatmap generation

The "gplots" R-package was used to generate heatmaps using the *heatmap.2* function. For the cancer versus normal heatmap, expression was normalized as log2 of the fold-change over the median of the normal samples for each transcript. For the ER-positive versus ER-negative heatmap, expression was normalized to the median of the ER-negative samples. Unsupervised heirarchical clustering was performed with the *hclust* function, using Pearson correlation as the clustering distance, using the "ward" agglomeration method.

4.5.23 RNA-seq differential expression testing

Differential expression testing was performed using the Sample Set Enrichment Analysis (SSEA) tool described previously[68]. Briefly, following count data normalization, SSEA performs the weighted KS-test procedure described in GSEA[164]. The resulting enrichment score (ES) statistic describes the enrichment of the sample set among all samples being tested. To test for significance, SSEA enrichment tests are performed following random shuffling of the sample labels. These shuffled enrichment tests are used to derive a set of null enrichment scores (1000 null enrichment scores computed). The nominal p value reported is the relative rank of the observed enrichment score within the null enrichment scores. Multiple hypothesis testing is performed by comparing the enrichment score of the test to the null normalized enrichment score (NES) distributions for all transcripts in a sample set. This null NES distribution is used to compute FDR q values in the same manner used by GSEA[164].

4.5.24 Associations with Oncomine clinical signatures

We identified the top 150 positively and negatively correlated genes (Spearman's

correlation) to *DSCAM-AS1* among the ER-positive breast cancer samples. These gene lists were imported into Oncomine[149] as custom concepts. We then nominated significantly associated breast cancer concepts with odds ratio > 4.0 for negatively associated concepts and > 6.0 for positively associated concepts and $p\text{-}value < 1 \times 10^{-6}$. Nodes and edges of these associations were exported and a concept association network was generated using Cytoscape version 3.2.1. Node positions were computed using the Force-Directed Layout algorithm in Cytoscape using the odds ratio as the edge weight. Node positions were subtly altered manually to enable better visualization of node labels.

4.5.25 Association of correlation signatures with Oncomine concepts

Correlation analysis described above was performed for *DSCAM-AS1*, *EZH2*, *HOTAIR*, *MALAT1*, and *NEAT1*. For each gene, we created a signature of the top 150 most positively and top 150 most negatively correlated genes. We performed a Fisher's exact test of overlap for each of the above gene signatures with Oncomine clinical signatures for cancer versus normal, clinical recurrence, clinical survival, metastasis, and high clinical grade. The following studies were utilized: Curtis Breast[94], Ma Breas[182], TCGA Breast[150], Zhao Breast[151], Bittner Breast[7], Desmedt Breast[152], Ivshina Breast[153], Loi Breast[154], Lu Breast[155], Perou Breast[146], Schmidt Breast[156], Sorlie Breast[147], vantVeer Breast[183], Wang Breast[158], Boersma Breast[159], Kao Breast[160], Symmans Breast[161], vandeVijver Breast[162]. For each Oncomine concept, overlap was tested for the top 1%, 5%, and 10% of genes up- and downregulated, and the gene signature with the greatest odds ratio was selected for each study. Signature comparisons were performed using a one-sided Fisher's exact test.

4.5.26 Survival analysis with TCGA breast data

Association of DSCAM-AS1 levels on clinical outcomes was assessed using the TCGA breast cohort. Survival data was obtained from the TCGA data portal. ER-positive samples were used for survival analysis as indicated by the TCGA clinical metadata via IHC status. Samples with DSCAM-AS1 expression >10FPKM were grouped into the "DSCAM-AS1 high" category and samples with expression <1FPKM were grouped into the "DSCAM-AS1 low" category. Kaplan-Meier analysis was performed, and log-rank test was performed to assess statistical significance.

4.5.27 Tissue expression level percentile metric

To generate a metric to summarize the expression of each lncRNA in breast cancer tissues, we identified the expression level of the 95th percentile sample among all breast RNA-seq samples including cancers tissue, normal tissue, and cell lines.

4.5.28 RNA-sequencing library preparation

Total RNA was obtained from cancer cell lines, and RNA quality was determined using an Agilent Bioanalyzer. Poly-A transcriptome libraries from the mRNA fractions were generated following the Illumina RNA-seq protocol. Each sample was sequenced in a single lane with the Illumina HiSeq 2000 (100-nucleotide read length) as previously described[30,184]. The dUTP method of second-strand marking was used for strand-specific library preparation as described previously[185].

4.5.29 Gene Set Enrichment Analysis

Expression levels of *DSCAM-AS1* were correlated (Spearman) to the expression of all protein-coding genes across all ER-positive breast cancers. The protein-coding genes

were then ranked by the Spearman Rho value, and used in a weighted, preranked GSEA analysis against MSigDB gene sets V5.0[115].

4.5.30 *In silico* binding prediction

To obtain potential HNRNP-L binding sites on DSCAM-AS1, we utilized GraphProt[186] to learn a predictive model from genome-wide HNRNP-L binding sites identified by iCLIP-seq[170]. For training data generation, we extracted the genomic binding positions (GSE37560) with BED table scores >= 10, followed by an extension of -/+ 20 nt resulting in 41 nt long binding sites. After mapping the sites to annotated RefSeq genes obtained from UCSC, an equally-sized set of negative sites was selected such that the sites were on the same RefSeq genes and did not overlap with any of the identified positive sites from the initial table. The GraphProt sequence model trained on these data was then used to identify high-scoring sites in the *DSCAM-AS1* sequence (NCBI GenBank NR_038899.1). The highest-scoring site centered at RNA position 923 contains a CA-repeat motif known for its affinity towards HNRNP-L and was thus used for subsequent analysis.

4.5.31 Rapid Amplification of cDNA Ends (RACE)

5′ and 3′ RACE was performed using the GeneRacer RLM-RACE kit (Invitrogen) according to the manufacturer's protocols. RACE PCR products obtained using Platinum Taq high-fidelity polymerase (Invitrogen), were resolved on a 1.5% agarose gel. Individual bands were gel purified using a Gel Extraction kit (Qiagen), and cloned into PCR4 TOPO vector, and sequenced using M13 primers.

4.5.32 Single-molecule fluorescence *in situ* hybridization (smFISH)

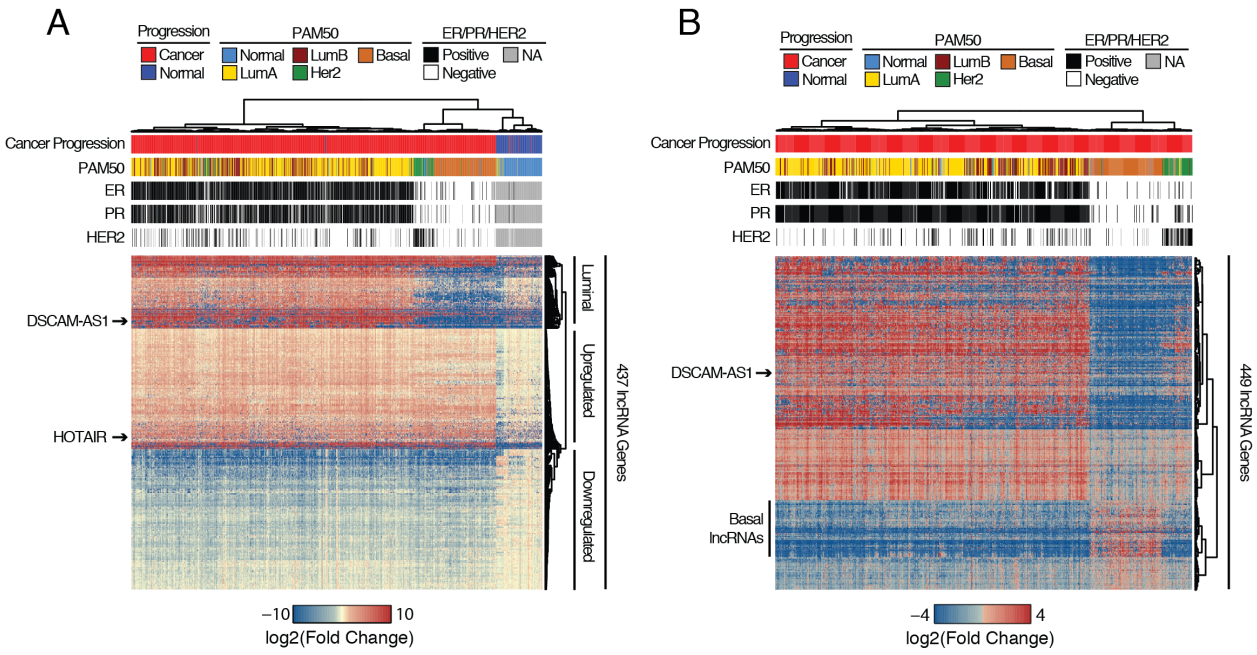
smFISH was performed as described[187], with some minor modifications. Cells were grown on 8-well chambered coverglasses, formaldehyde fixed and permeablized overnight at 4°C using 70% ethanol. Cells were rehydrated in a solution containing 10% formamide and 2x SSC for 5 minutes and then treated with 10nM FISH probes for 16 h in 2x SSC containing 10% dextran sulfate, 2 mM vanadyl-ribonucleoside complex, 0.02% RNAse-free BSA, 1 μg/μL E.coli tRNA and 10% formamide at 37°C. After hybridization, cells were washed twice for 30 minutes at 37°C using a wash buffer (10% formamide in 2x SSC). Cells were then mounted in solution containing 10 mM Tris/HCl pH 7.5, 2x SSC, 2 mM trolox, 50 μM protocatechiuc acid (PCA) and 50nM protocatechuate dehydrogenase (PCD). FISH samples were imaged in 3 dimensions using HILO illumination as described[188]. Images were processed using custom-written macros in ImageJ. Analysis routines comprised of 3 major steps: background subtraction, Laplacian of Gaussian (LoG) filtering and thresholding. Spots with intensity above set threshold are represented in images. Probes were designed to target all isoforms of the DSCAM-AS1 transcript. Probe sequences targeting DSCAM-AS1 (21 probes per transcript) are as follows: cctatccctttctctaagaa, acttctgcaaaaacgtgctg, ggttccactccattttaatt, ctatagcgtcttatcagctg, catgtgtccggatatcattt, tcagtgagtggataactggt, aattctagtggaggcaccta, ctaagtagcttcatctttcc, caactgcgtgtttcctagtc, agcattctctgttttaacca, ttagcaactgccttgctctg, gctgtccagttttagtaaca, cgttgtgagcctgagagatc, agaacttccctagaggagtg, atggggagtgagaccaaaca, tggaggagggacagagaagg, tgtgggtgattggtactttt, atggatgagtatgtcatgcc, tattgccatggttagcatga, aatgcatgcttgatggagct.

4.6 Acknowledgements

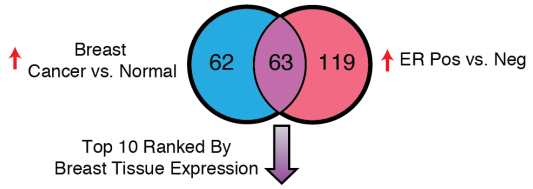
We thank K. Wilder-Romans for help with the mouse Xenograft work, the University of Michigan DNA Sequencing Core for Sanger sequencing and K. Giles for critically reading the manuscript and for the submission of documents. This work was supported in part by US National Institutes of Health Prostate Specialized Program of Research Excellence grant P50CA186786, Early Detection Research Network grant UO1 CA111275, US National Institutes of Health grants R01 CA132874 and RO1 CA154365 (A.M.C.), and US Department of Defense grant PC100171 (A.M.C.). A.M.C. is supported by the Prostate Cancer Foundation and the Howard Hughes Medical Institute. A.M.C. is an American Cancer Society Research Professor and a Taubman Scholar of the University of Michigan. R.M. was supported by a Prostate Cancer Foundation Young Investigator Award and by US Department of Defense Post-Doctoral Fellowship W81XWH- 13-1-0284. Y.S.N. is supported by a University of Michigan Cellular and Molecular Biology National Research Service Award Institutional Predoctoral Training Grant.

4.7 Author Contributions

Y.S.N., S.H., A.M.C. and F.Y.F. conceived the study and analyses. Y.S.N. performed all bioinformatics analyses and guided study design with the assistance of M.K.I. and S.Z. S.H. performed cellular and molecular biology experiments with the assistance T.M., R.M., C.Z., J.R.P., Y.H., S.R., S.K., R.F.S., L.X., and U.S. K.W.R. performed mouse xenograft work. X.C. generated RNA-seq libraries and performed the sequencing. A.P. performed the mass spectrometry. M.U., A.G., R.B., and C.S.S. performed the in silico binding predictions. J.M.R. produced the tamoxifen resistance cell line. D.F.H. and L.J.P. provided guidance and analysis for the clinical work. Y.S.N., A.M.C., and F.Y.F. wrote the manuscript. All authors discussed results and commented on the manuscript.



C



MiTranscriptome Name	Genomic Coordinates	Breast Ca vs Norm SSEA %ile	ER+ vs ER- SSEA %ile	Breast Expression 95th %ile	ER Promoter Binding	ER Response
DYNLL1-AS1	chr12:120899317-120933832(-)	99.03	99.89	231.53	T47D	↑
CERS6-AS1	chr2:169627826-169678566(-)	99.71	99.93	115.62	⊘	↑
DSCAM-AS1	chr21:41754824-41757280(+)	98.02	97.51	45.29	BOTH	↑↑↑
BRCAT1	chr10:8768317-9084377(+)	99.71	99.53	42.65	BOTH	—
CAT1222	chr10:5500354-5543763(-)	98.71	99.06	35.10	MCF7	↑
BRCAT54	chr5:44742417-44809099(-)	97.97	99.82	30.10	MCF7	—
BRCAT32	chr6:151913411-151937977(-)	98.27	99.86	29.33	⊘	—
CAT146	chr1:197122504-197127797(+)	98.16	97.91	26.80	⊘	↑
CAT659	chr5:44826244-44828803(-)	97.68	98.87	25.55	⊘	—
CAT505	chr4:780400-792957(+)	97.54	99.73	15.71	BOTH	—

Figure 4.1 Identification of ER and breast cancer associated lncRNAs. (**a**) heatmap depiction of the top cancer versus normal differentially expressed lncRNAs among the TCGA breast RNA-seq cohort (n=946). 437 lncRNAs were differentially expressed with an SSEA FDR < 1e-5 and an SSEA percentile cutoff of 0.975. Expression values are depicted as log2 of the fold-change over the median of the normal samples (n=104). Unsupervised hierarchical clustering was done on both lncRNAs and patients. Cancer progression, PAM50 classification, and ER, PR, and HER2 status are shown above heatmap. LncRNAs clustered into 3 distinct categories, "Luminal", "Upregulated", and "Downregulated". Two representative lncRNAs are highlighted. (**b**) heatmap depiction of the top ER-positive versus ER-negative lncRNAs. 449 lncRNAs met the SSEA criteria described in (**a**). Unsupervised clustering was performed for samples and lncRNAs.

Expression values depicted as log2 of the fold-change over the median of the ER-negative samples (n=538). Cancer progression, PAM50 classification, and ER, PR, and HER2 status are shown above heatmap. One representative lncRNA is highlighted along with a group of lncRNAs with basal specific expression. (**c**) Venn diagram of the intersection of the breast cancer versus normal and ER-positive versus ER-negative analyses. Intersection is shown for the overexpressed lncRNAs in both categories. The top 10 lncRNAs based on expression level in breast cancer tissues (expression value of 95th percentile sample) are depicted in table. ER promoter binding determined via ChIP-seq is depicted along with expression response from RNA-seq following 3 hours of estrogen stimulation in MCF7 cells (one arrow represents > 1.5 fold increase, three arrows represents > 2.5 fold increase).

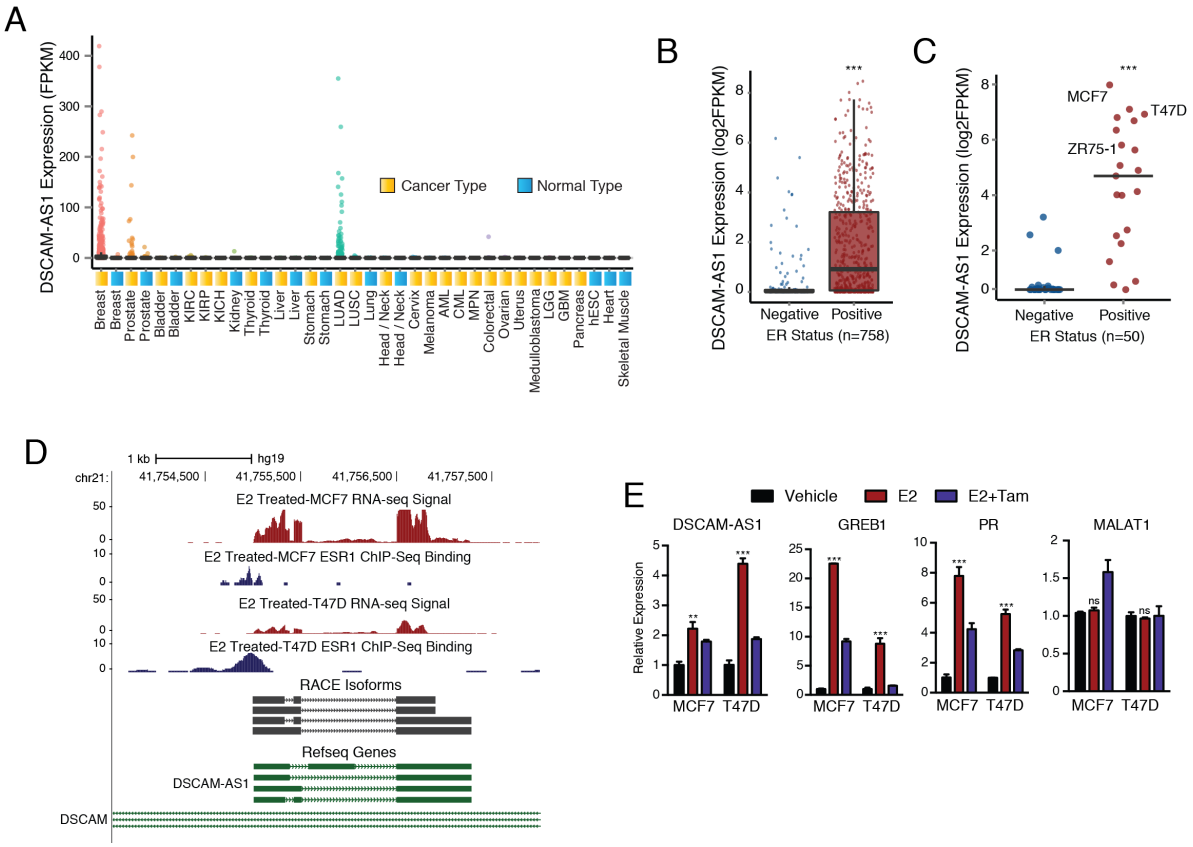


Figure 4.2 Characterization of *DSCAM-AS1*. (**a**) Plot highlighting the expression in FPKM of *DSCAM-AS1* in the 6,503 sample MiTranscriptome RNA-seq compendium[68] categorized by the different cancer/tissue types. Each point represents one RNA-seq tissue sample. (**b**) Expression of *DSCAM-AS1* is significantly higher in ER-positive breast cancer tissue samples (n=584) compared to ER-negative samples (n=174). Expression was analyzed in samples for which ER IHC was performed. Each point represents one RNA-seq sample. ***: $p < 0.0001$, comparing ER-positive to -negative. (**c**) Expression of *DSCAM-AS1* by RNA-seq in breast cancer cell lines categorized by ER status. *DSCAM-AS1* expression is significantly higher in ER-positive cell lines (n=21) versus ER-negative cell lines (n=29). Each point represents one cell line. ***: $p < 0.0001$, comparing ER-positive to –negative via Student's t-test. (**d**) UCSC genome browser depiction of *DSCAM-AS1* region on chromosome 21. RNA-seq expression track shown in red, and ESR1 ChIP-seq shown in blue. Refseq transcripts shown in green. RACE verified transcript structure shown in black. (**e**) qPCR expression of *DSCAM-AS1*, *GREB1*, *PGR*, and *MALAT1* 8 hours following addition of DMSO vehicle (black), 10nM estrogen (red), and 10nM estrogen and 1uM tamoxifen (blue) in MCF7 and T47D cell lines. Error bars represent standard error of the mean (s.e.m.) for three biological replicates. **: $p < 0.001$, ***: $p < 0.0001$, ns: $p > 0.01$ comparing to vehicle for each condition via Student's t-test.

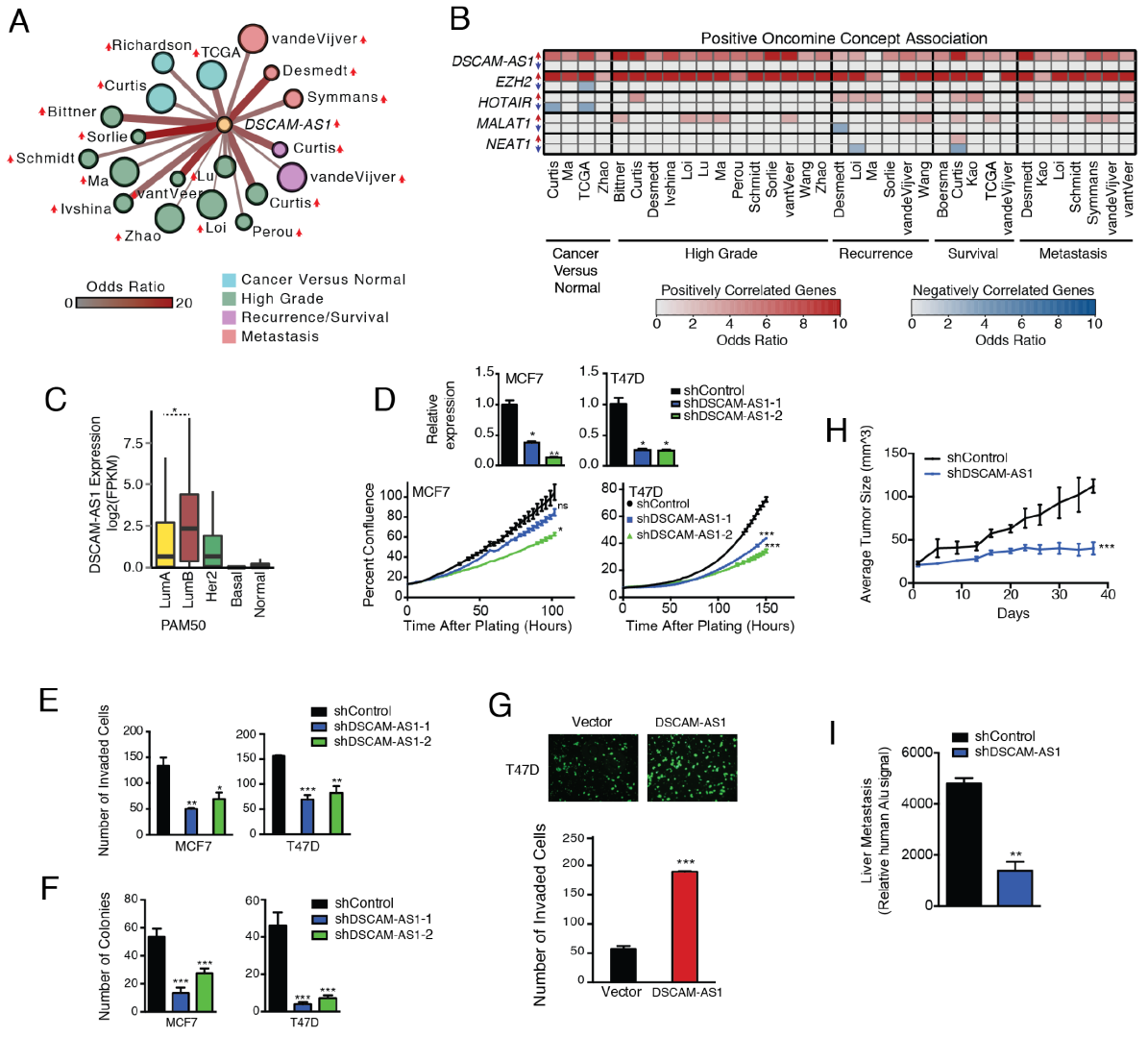
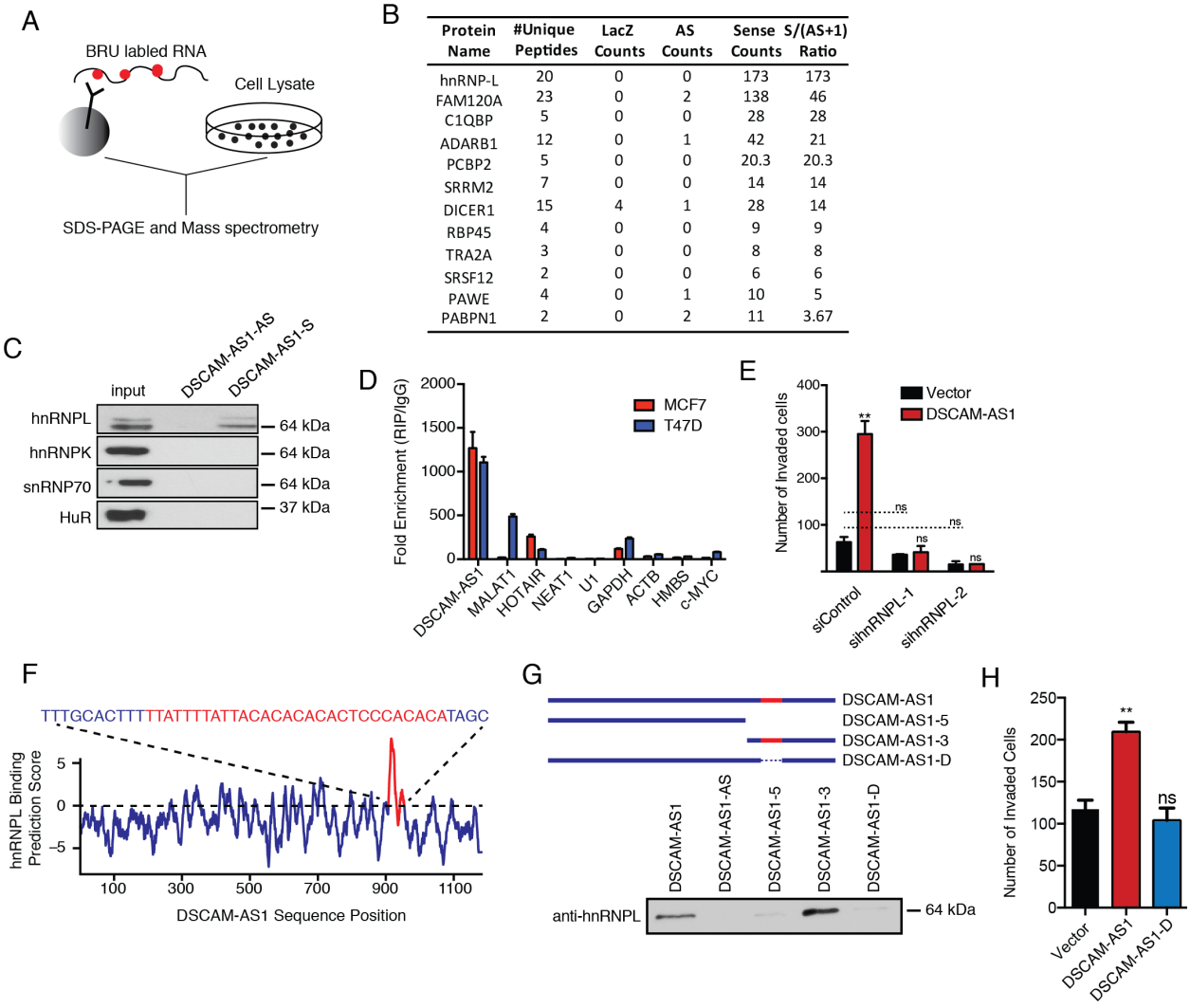


Figure 4.3 ***DSCAM-AS1* is implicated in cancer aggression clinically and in cell lines.** (**a**) Cytoscape depiction of the overlap between the 150 genes most positively correlated with *DSCAM-AS1* and clinical signatures from Oncomine[149] for breast cancer clinical outcomes (i.e., recurrence, survival, and metastasis), high cancer grade, and cancer versus normal. All significant associations with an odds ratio > 6 are shown (Fisher's p-value < 1e-4). Size of node reflects the size of the gene signature, and the thickness/redness of the line represents the magnitude of the odds ratio. (**b**) Heatmap displaying the overlap between the top 150 genes correlated to *DSCAM-AS1*, *EZH2*, *HOTAIR*, *MALAT1*, and *NEAT1* and the genes positively associated with various breast cancer clinical signatures (see above). For each gene, the top row depicts the odds ratio for the positively correlated genes (red), and the bottom row represents the odds ratio for the negatively correlated genes (blue). The first name of the author for each clinical study is listed. (**c**) Expression of *DSCAM-AS1* from breast cancer RNA-seq by PAM50 classification (n=946). Luminal B expression is significantly greater than Luminal A

(Student's t-test, $p = 0.006$) (**d**) Incucyte proliferation assay performed following knockdown of *DSCAM-AS1* using two independent shRNAs. Degree of knockdown determined by qPCR shown above. Error bars represent the s.e.m. for three biological replicates. *: $p < 0.01$, **: $p < 0.001$, ***: $p < 0.0001$, ns: $p > 0.01$ comparing to shControl for each condition via Student's t-test. (**e**) Invasion assay following shRNA knockdown of *DSCAM-AS1* using Matrigel coated Boyden chamber assay. Error bars represent the s.e.m. for three biological replicates. *: $p < 0.01$, **: $p < 0.001$, ***: $p < 0.0001$ comparing to shControl for each condition via Student's t-test. (**f**) Soft agar colony formation assay following shRNA knockdown of *DSCAM-AS1*. Error bars represent the s.e.m. for three biological replicates. ***: $p < 0.0001$comparing to shControl for each condition via Student's t-test. (**g**) Invasion assay following overexpression of *DSCAM-AS1* and *LacZ* control. Error bars represent the s.e.m. for three biological replicates. Representative invasion images shown above. ***: $p < 0.0001$, comparing to vector overexpression via Student's t-test. (**h**) Mouse xenograft study of tumor growth for T47D cells with shRNA knockdown of *DSCAM-AS1*. Xenografts with shRNA knockdown of *DSCAM-AS1* (n=24) exhibited reduced growth when compared to control shRNA knockdown (n=26). Error bars represent the s.e.m. for all xenografts used. ***: $p < 0.0001$, comparing to shControl via Student's t-test. (**i**) Assessment of xenograft metastasis to liver by human Alu PCR, which detects the human cancer cells in the mouse liver. Error bars represent the s.e.m. for three biological replicates. **: $p < 0.001$ comparing to shControl via Student's t-test.



Protein Name	#Unique Peptides	LacZ Counts	AS Counts	Sense Counts	S/(AS+1) Ratio
hnRNP-L	20	0	0	173	173
FAM120A	23	0	2	138	46
C1QBP	5	0	0	28	28
ADARB1	12	0	1	42	21
PCBP2	5	0	0	20.3	20.3
SRRM2	7	0	0	14	14
DICER1	15	4	1	28	14
RBP45	4	0	0	9	9
TRA2A	3	0	0	8	8
SRSF12	2	0	0	6	6
PAWE	4	0	1	10	5
PABPN1	2	0	2	11	3.67

Figure 4.4 Physical and functional relationship of *DSCAM-AS1* with hnRNPL. (**a**) Schematic representation of the RNA pull-down technique used to identify protein binding partners of *DSCAM-AS1*. The BRU labeled RNA transcripts are incubated with cell lysate from T47D cells and the eluted protein is resolved by SDS-PAGE. RNA-bound protein product is then processed by mass spectrometry. (**b**) Top protein binding partners for *DSCAM-AS1*. Pulldown of *LacZ* and antisense *DSCAM-AS1* used as control. S/AS ratio determined as sense counts divided by 1 + antisense counts. (**c**) Western blot of hnRNPL, hnRNPK, snRNP70, and HuR following pulldown of BRU labeled *DSCAM-AS1* and antisense *DSCAM-AS1*. (**d**) qPCR following RIP for hnRNPL performed in MCF and T47D cells. Data represented as fold-enrichment over IgG RIP. Error bars represent the s.e.m. for three biological replicates. (**e**) Invasion assay for T47D cells overexpressing *LacZ* control or *DSCAM-AS1* following siRNA-mediated knockdown of *hnRNPL*. Error bars represent the s.e.m. for two biological replicates. **: p

< 0.001, ns: p > 0.01 comparing to siControl (unless otherwise specified with dotted line) for each condition via Student's t-test. (**f**) Per base i*n silico* prediction for binding of hnRNPL to *DSCAM-AS1.* One strong predicted binding peak exists in the 3′ region of *DSCAM-AS1* shown in red. (**g**) Schematic depicting the mutant forms of *DSCAM-AS1* generated with or without the predicted binding site (top). Western blot for hnRNPL shown following pulldown of each mutant form of *DSCAM-AS1* in HEK293 cells (bottom). (**h**) Invasion assay in T47D cells overexpressing *LacZ* control, full-length *DSCAM-AS1,* and the *DSCAM-AS1-D* mutant. Error bars represent the s.e.m. for three biological replicates. **: p < 0.001, ns: p > 0.01 comparing to vector overexpression for each condition via Student's t-test.

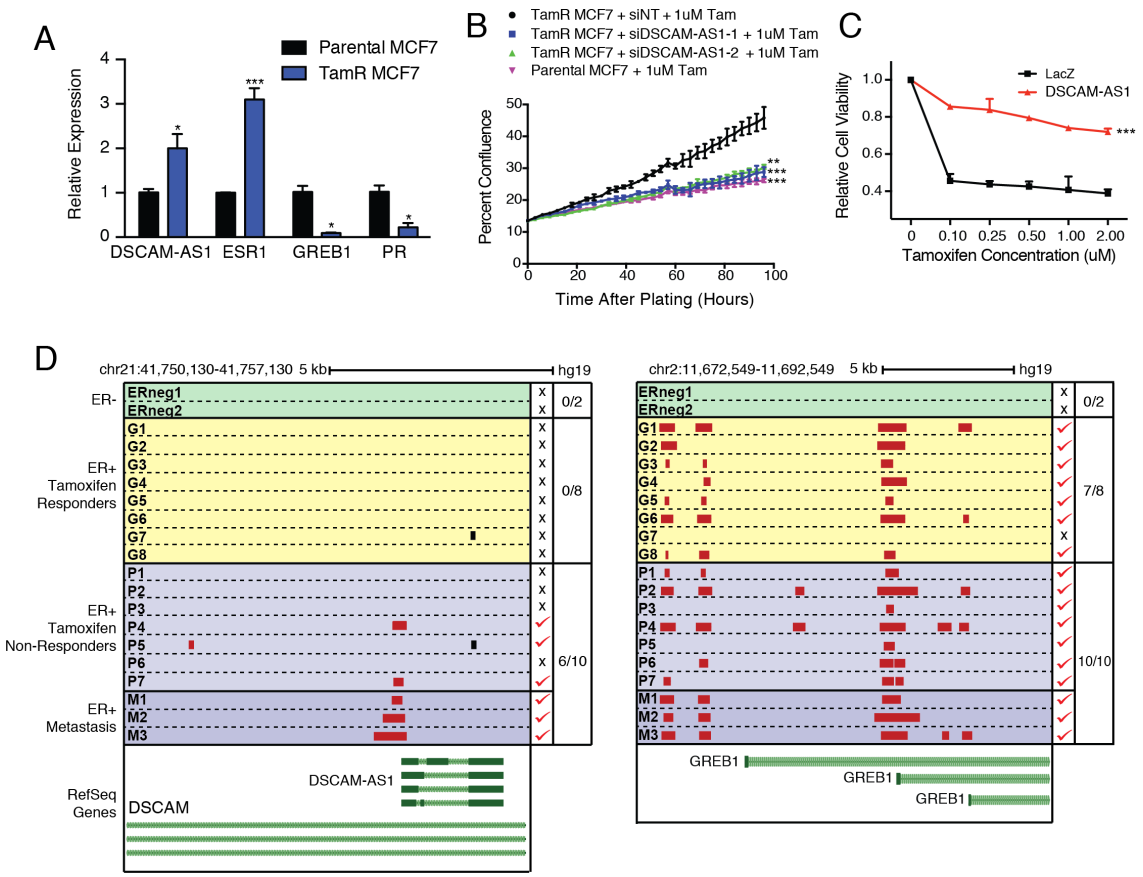


Figure 4.5 ***DSCAM-AS1* is implicated in tamoxifen resistance.** (**a**) qPCR expression of *DSCAM-AS1*, *ESR1*, *GREB1*, and *PGR* in tamoxifen resistant MCF7 cells relative to parental MCF7. Error bars represent the s.e.m. for three biological replicates. *: $p < 0.01$, ***: $p < 0.0001$, comparing to parental MCF7 for each condition via Student's t-test. (**b**) Proliferation assay in parental MCF7 cells and in TamR MCF7 cells following siRNA-mediated knockdown of *DSCAM-AS1* via two independent siRNAs. Error bars represent the s.e.m. for three biological replicates. **: $p < 0.001$, ***: $p < 0.0001$, comparing to parental TamR siNT for each condition via Student's t-test. (**c**) WST viability assay following 10 days of culture in varying levels of tamoxifen performed for T47D cells overexpressing LacZ control and *DSCAM-AS1*. Error bars represent the s.e.m. for three biological replicates. ***: $p < 0.0001$, comparing to LacZ overexpression via Student's t-test. (**d**) Depiction of estrogen receptor binding to the *DSCAM-AS1* (left) and *GREB1* (right) promoters via ChIP-seq performed in primary and metastatic breast cancer tumor tissues. ER-status and response to tamoxifen treatment detailed to left. ER binding peaks (determined using MACS software) are depicted in red (for promoter binding) and black (for non-promoter binding). Promoter defined as 5KB upstream of any transcriptional start site. ER promoter-binding indicated by red check or "x" to the right. Genomic coordinates in hg19 listed above.

CHAPTER 5

Oncogenic Role of *THOR*, a Conserved Cancer/Testis Long Noncoding RNA

The work presented in this chapter has been written into a manuscript currently under in review

5.1 Abstract

Large scale transcriptome sequencing efforts have vastly expanded the catalog of long non-coding RNAs (lncRNAs) with varying evolutionary conservation, lineage expression, and cancer specificity. Here we functionally characterize a novel ultraconserved lncRNA, *THOR*, which exhibits expression exclusively in testis and a broad range of human cancers. *THOR* knockdown and overexpression in multiple cell lines and animal models alters cell or tumor growth supporting an oncogenic role. We discovered a conserved interaction of *THOR* with IGF2BP1 and show that *THOR* contributes to the mRNA stabilization activities of IGF2BP1. Notably, transgenic *THOR* knockout produced fertilization defects in zebrafish and also conferred a resistance to melanoma onset. Likewise, ectopic expression of human *THOR* in zebrafish accelerated the onset of melanoma. *THOR* represents a novel class of functionally important cancer/testis lncRNAs whose structure and function have undergone positive evolutionary selection.

5.2 Introduction

Over the past decade, there has been a paradigm shift in our understanding of molecular biology sparked by the discovery of non-coding RNAs (ncRNAs) challenging the central dogma of molecular biology [189]. It has become apparent that the transcriptome is far more intricate than previously appreciated [190], with various types of non-coding RNAs implicated in key physiological roles in cells [191]. Long non-coding RNAs (lncRNAs) have emerged as an abundant and functionally diverse species of ncRNA [38,192]. Despite their striking prevalence in the transcriptome and countless efforts to interrogate their function, our understanding of the function of the vast majority of lncRNAs remains anecdotal, making their classification particularly challenging [193]. Novel classes of lncRNAs continue to be identified with categorization criteria including their degree of conservation [41], association with various DNA elements [194,195], ability to bind miRNAs [196], regulation of chromatin remodeling [43,46,197], induction of aneuploidy [57], influence on RNA stability machinery [198], and their potential to produce small peptides [199,200].

Recently, we discovered that the human genome possesses the potential to transcribe tens of thousands of lncRNAs [192]. With the overwhelmingly large number of potentially functional elements to investigate, features such as degree of evolutionary conservation and expression pattern are logical criteria that can be employed to discover functionally important lncRNAs in cells. Although the general level of conservation of lncRNAs has been contentious [28,192], there is a clear subclass of lncRNAs that are highly conserved, many of which possess "ultraconserved" regions (i.e., at least 200 base-pairs (bps) of nearly perfect vertebrate conservation) [41,201,202]. While conservation is a trait highly suggestive of functional relevance in cells, it also permits the characterization and mechanistic investigation of lncRNAs in model organisms [41,203], a particularly exciting avenue given the recent popularization of genome editing techniques [204].

In searching for highly conserved lncRNAs with intriguing expression patterns, we defined a novel class of lncRNA, with normal tissue expression limited to the testis and widespread expression in multiple cancer types. This cancer/testis expression pattern is characteristic of cancer/testis antigens, a well-defined class of proteins that has been suggested as targets for cancer therapy [205–207]. While many cancer-associated lncRNAs have been identified [66,208], none previously characterized have exhibited cancer/testis expression. Here we discover the first cancer/testis lncRNA, *THOR* (*T*estis-associated *H*ighly-conserved *O*ncogenic long non-coding *R*NA), and investigate its role in oncogenesis and testis physiology, identifying an evolutionarily conserved functional interaction with IGF2 mRNA-binding proteins (IGF2BPs). Moreover, we leverage the sequence conservation of *THOR* to generate transgenic zebrafish models that shed light on its function in the testis and in cancers.

5.3 Results

5.3.1 Discovery of *THOR*, a conserved lncRNA expressed in the testis

In a recent large-scale RNA-sequencing analysis, we comprehensively profiled the human transcriptome, discovering tens of thousands of novel lncRNAs [192]. Further interrogation of the degree of conservation of these lncRNAs revealed that, while lncRNAs tend to be less conserved than protein-coding genes (**Figure D.1a**) and most do not exhibit marked sequence conservation, a subset of conserved lncRNAs does exist (**Figure 5.1a**). We measured both the average base-wise conservation of the entire transcript (**Figure 5.1a**, *x*-axis) and the level of conservation of the best 200bp window (**Figure 5.1a**, *y*-axis), a metric previously utilized to determine "ultraconserved" elements [192,201]. We identified 82 intergenic ultraconserved lncRNAs with expression of at least 1 FPKM in the top 1% of samples in our tissue RNA-seq compendium (**Figure**

5.1a,b). Despite possessing a 200bp ultraconserved segment, these lncRNAs possessed varying degrees of base-wise conservation (**Figure 5.1b**, top; range, 0.1% - 55.4% conserved bases), with *THOR* exhibiting the second highest degree of base-wise conservation. Interestingly, when interrogating the GTEX benign tissue RNA-seq dataset [120,209], two of these ultraconserved lncRNAs, *THOR* and *CRNDE*, displayed substantial expression in the testes (**Figure 5.1b**, bottom). While both *THOR* and *CRNDE* were most highly expressed in the testis in comparison to a panel of 30 other benign tissue types, the expression pattern of *THOR* was almost exclusively in the testis (**Figure 5.1c**), while *CRNDE* exhibited more distributed expression in other tissues (**Figure D.1b**). Thus, drawing from its strikingly testis specific expression pattern and its substantial base-wise conservation, we identified *THOR* as a prime candidate for further investigation.

5.3.2 Transcriptional *THOR* homologues exist in the mouse and zebrafish

Utilizing structure prediction from our prior RNA-seq assembly [192] and from corroboration of 5′ and 3′ rapid amplification of cDNA ends (RACE), we identified two isoforms of *THOR*, comprised of either 2 or 3 exons on chromosome 2 (**Figure D.2a and Figure D.3a**). Given its substantial sequence conservation we set out to identify *THOR* homologues in other species. Utilizing the BLAT tool[210] we identified predicted regions in the mouse and zebrafish genome homologous to the human *THOR* (h-*THOR*) (**Figure 5.1d**). In order to confirm transcription of these homologous regions, we performed 5′ and 3′ RACE on mouse and zebrafish testis RNA. In the mouse, we identified a 2 exon homologue of *THOR* (m-*THOR*), that was partially annotated by Refseq (gm29359) (**Figure D.2b** and **Figure D.3b**), and in the zebrafish we identified a large, previously unidentified, 2056bp monoexonic orthologue (z-*THOR*) (**Figure D.2c and Figure D.3c**). The conservation of h-*THOR* extends to exons 2 and 3, which are both represented in m-

THOR and z-*THOR* (**Figure 5.1d**). Additionally, the CPAT tool[75] was utilized to confirm the identity of *THOR* as a non-coding transcript (**Figure 5.3d**).

5.3.3 *THOR* exhibits an evolutionarily conserved expression pattern in normal tissues

To obtain an independent validation of testis specific *THOR* expression observed in the GTEX RNA-seq data, we performed quantitative real-time PCR (qRT-PCR) with cDNA derived from various normal human tissues, observing a similarly testis-specific expression pattern (**Figure 5.1e**). Moreover, RNA *in situ* hybridization (ISH) of human testis tissue using h-*THOR* specific probes revealed an enrichment of *THOR* testis expression in the spermatocyte and spermatid (**Figure 5.1f**), which is interestingly a similar expression pattern reported for cancer/testis antigens not found on the X-chromosome [205,211]. Querying additional RNA from tissue panels in the mouse and zebrafish identified testis-specific expression for both m-*THOR* (**Figure 5.1g**) and z-*THOR* (**Figure 5.1h**). Additionally, we also observed elevated *THOR* expression during the early development of both the mouse and zebrafish (**Figure 5.1g,h**, right).

5.3.4 Expression and functional implication of *THOR* in human cancers

Further interrogation of the expression of *THOR* in cancer tissue RNA-seq samples from the TCGA revealed that, despite bearing a benign expression pattern restricted to the testis, *THOR* is widely expressed in a number of tumor types and is expressed significantly higher in cancers compared to both the GTEX normal and the TCGA normal samples (**Figure 5.2a**). Additionally, we also found *THOR* to be expressed in a multitude of the cell lines in the Cancer Cell Line Encyclopedia [123] with the highest expression in lung cancer cell lines (**Figure 5.2b**). Interestingly, in both subtypes of lung

cancer represented in the TCGA (lung adenocarcinoma, LUAD; lung squamous carcinoma, LUSC), *THOR* expression was significantly higher in cancer in comparison to matched benign adjacent normal tissue (**Figure 5.2c**). We confirmed this cancer specific non-small cell lung cancer (NSCLC) expression by performing h-*THOR* qRT-PCR on RNA from an independent cohort of tumor and normal lung tissue obtained from the University of Michigan (**Methods**) (**Figure 5.2d**). This cohort also contained both melanoma and benign melanocytes, in which *THOR* similarly exhibited cancer specific expression (**Figure 5.2d**).

In order to interrogate the potential role of *THOR* in cancer processes, we investigated its function in NSCLC and melanoma cell lines. Knockdown of *THOR* via siRNA in H1299 and MM603, NSCLC and melanoma cell lines with high levels of *THOR* (**Figure 5.2e**) resulted in a dramatic reduction in the proliferative capacity of these cells (**Figure 5.2f** and **Figure D.4a**). *THOR* knockdown also exhibited a reduction in the ability of H1299 to form colonies in a soft agar colony formation assay (**Figure 5.2g**). To rule out potential of off-target effects, we also performed antisense oligo (ASO)-mediated knockdown of *THOR* in H1299 cells, and observed a significant reduction in proliferation (**Figure 5.2h**). Additionally, siRNA and ASO-mediated knockdown of *THOR* in H1437, a cell line without *THOR* expression (**Figure 5.2e**), failed to have any effect on proliferation in these cells (**Figure D.4b,c**), further corroborating the on-target effects of the RNA-interference techniques. Cells with stable lentiviral *THOR* overexpression in H1437 and SKMEL5, cell lines with little to no expression of *THOR* (**Figure 5.2e**), also exhibited significant increases in proliferative capacity (**Figure 5.2i** and **Figure D.4d**) and in soft agar colony formation (**Figure 5.2j**). This proliferative advantage conferred by *THOR* was confirmed in a mouse xenograft experiment wherein we observed that tumor xenografts derived from cells stably overexpressing *THOR* exhibited a significant proliferative advantage when compared to cells stably

overexpressing LacZ control (**Figure 5.2k**). These findings demonstrate a profound growth advantage conferred by *THOR* expression in tumors, suggesting an oncogenic role for this lncRNA, which in combination with its testis-specific expression establish *THOR* as a functional cancer/testis lncRNA.

5.3.5 Conservation of the *THOR*-IGF2BP1 interaction

To begin characterization of the potential mechanism through which *THOR* is functioning in cells, we investigated the cellular localization of *THOR*, observing it to be present in both the cytoplasm and nucleus via qRT-PCR following cellular fractionation (**Figure D.5a**) and via single molecule fluorescence *in situ* hybridization (**Figure D.5b**). With this pan-cellular expression pattern, we hypothesized that *THOR* may be executing its function through a protein partner and set out to identify potential *THOR* protein interactors via RNA-pulldown followed by mass-spectrometry. Building on the sequence conservation of *THOR*, we aimed to identify protein-binding interactions that were also conserved. We accomplished this goal by utilizing four conditions for RNA-pulldown mass spectrometry: 1) pulldown of h-*THOR* added to human H1299 cancer cell lysate, 2) pulldown of h-*THOR* added to zebrafish embryo lysate, 3) pulldown of z-*THOR* added to human H1299 cancer cell lysate, and 4) pulldown of z-*THOR* added to zebrafish embryo lysate. In all conditions pulldown of antisense *THOR* was utilized as a negative control. While varying numbers of proteins were pulled down in the various conditions, two proteins, IGF2BP1 and IGF2BP3, were pulled down in all four conditions (**Figure 5.3a**). Interestingly, the IGF2BP proteins were also the only proteins pulled down by h-*THOR* in both the nuclear and cytoplasmic fractions of H1299 cells (**Figure D.5c**). Both IGF2BP1 and IGF2BP3 have been reported to exhibit a cancer/testis expression pattern similar to that of *THOR* [212] (**Figure D.6a-d**).

IGF2BP1 and IGF2BP3 have been implicated in mediating RNA-stability and translation through their binding to a number of well-defined mRNA targets [212,213]. Additionally, this function has been documented to occur in concert with a number of other proteins comprising a messenger ribonucleoprotein (mRNP) complex [214]. We observe that many members of this complex are also pulled down by *THOR* in the various conditions (**Figure 5.3a** and **Figure D.5d**), and confirm that IGF2BP1, IGF2BP2, IGF2BP3, STAU1, and YBX1 are present in the complex via immunoprecipitation and Western blotting (**Figure 5.5e**). To further validate the specificity of the interaction between h-*THOR* and mRNP complex proteins, we performed RNA immunoprecipitation (RIP) assays with antibodies against IGF2BP1-3, YBX1, STAU1 and HuR with subsequent qRT-PCR for *THOR* and a panel of additional control RNAs (**Figure 5.3b**). As expected, *THOR* and *IGF2* RNAs are pulled down by the various proteins of the mRNP complex, while negative control lncRNAs *NEAT1*, *TINCR*, and *HOTAIR* exhibit a substantially reduced extent of pulldown (**Figure 5.3b**). We also observed that HuR, a protein not in the mRNP complex (**Figure 5.5e**), does not to pull down any of the RNAs used in the study (**Figure 5.3b**). We further characterized the relationship of *THOR* and IGF2BP1 as a direct binding interaction, showing that *in vitro* transcribed h-*THOR* and purified myc-tagged IGF2BP1 exhibit substantial binding compared to negative controls (**Figure 5.3c**).

In order to further implicate that the conserved regions of *THOR* are responsible for its binding to IGF2BP1, suggesting that this binding interaction was selected for evolutionarily, we added multiple deletion isoforms of both h-*THOR* and z-*THOR* to human H1299 cancer cell lysate and to zebrafish embryo lysate, respectively. Pulldown of the various isoforms of h-*THOR* followed by Western blot for IGF2BP1 revealed that the region of *THOR* responsible for IGF2BP1 binding is in exon 2 and 3, the conserved region of h-*THOR* (**Figure 5.3d**). Additionally, this observation was also observed for z-

THOR, wherein the 5′ conserved portion of z-*THOR* was sufficient to result in pulldown of the zebrafish igf2bp1 protein (**Figure 5.3e**).

5.3.6 *THOR* regulates IGF2BP1's target mRNAs levels

Having identified a specific and evolutionarily conserved interaction between *THOR* and IGF2BP1, we set out to further characterize this functional relationship. As IGF2BP1 has been reported to regulate the mRNA stability of a set of well-described target RNAs[212,215], we hypothesized that *THOR* may be playing a role in IGF2BP1-facilitated mRNA stabilization. To test this hypothesis, we monitored the levels of canonical IGF2BP1 targets (*IGF2*, *CD44, KRAS, ACTB, PABPC1, GLI1, MYC, MAPT, CTNNB1, PPP1R9B, BTRC, PTEN* and *H19*)[212] in various conditions. Remarkably, the levels of nearly all IGF2BP1 targets were decreased upon knockdown of *THOR* in both H1299 and MM603 cells, and conversely increased with stable overexpression of *THOR* in H1437 and SKMEL5 cells (**Figure 5.4a**). As expected, knockdown of *IGF2BP1* produced a similar reduction in its targets, while altering levels of *IGF2* and *CD44*, two of the canonical IGF2BP1 targets, failed to show a trend in the expression of IGF2BP1 targets (**Figure 5.4a**). IGF2BP1 binds to target mRNAs, and has been shown to mostly increase their stability[212,214], although there has been a report of an IGF2BP1-mediated destabilization effect [215]. Thus, we hypothesized that the effects of *THOR* levels on IGF2BP1 targets (**Figure 5.4a**) may be explained by *THOR* stabilizing the interaction of IGF2BP1 with its targets. To examine this function, we performed qRT-PCR for IGF2BP1 targets, *IGF2* and *CD44*, following IGF2BP1 RIP. Interestingly, knockdown of *THOR* reduced binding of both *IGF2* and *CD44* to IGF2BP1, while overexpression of *THOR* increased *IGF2* and *CD44* binding. However, knockdown of *IGF2* did not result in an altered binding of *CD44* to IGF2BP1, while *CD44* knockdown failed to cause an effect of *IGF2* binding to IGF2BP1 (**Figure 5.4b** and **Figure D.7a**). Further corroborating

this hypothesis, *THOR* overexpression substantially increased the mRNA stability of IGF2BP1 targets *IGF2* and *CD44* following Actinomycin D treatment (**Figure 5.4c**), while having no effect on the stability of the *GAPDH* and *UBC,* two control mRNAs that do not interact with IGF2BP1 (**Figure D.7b**). The *THOR* mRNA has a half-life of 14 hours (**Figure D.7c**), which is longer than the dynamic range observed for the stabilization effects on *IGF2* and *CD44* (**Figure 5.4c** and **Figure D.7b**), confirming *THOR* is present in cells long enough to exert these effects.

IGF2, one of the canonical IGF2BP1 targets from which it derives its name[216], is a secreted protein that has been shown to contribute to oncogenesis via mitogenic activation through receptor tyrosine kinase-MEK-ERK signaling (**Figure D.7d**)[217,218]. Interestingly, we observed that in addition to altering levels of *IGF2* (**Figure 5.4a**) and modulating *IGF2*-IGF2BP1 binding (**Figure 5.4b**), *THOR* is also successful in regulating the downstream signaling pathway of *IGF2* (**Figure D.7e**), suggesting that the *THOR*-mediated changes in IGF2BP1 target expression are sufficient to result in functional downstream signaling. With the hypothesis that *THOR* is mediating oncogenic phenotypes in cancer cells (**Figure 5.2**) through a functional relationship with IGF2BP1 (**Figure 5.3a-e**), we assessed the proliferative capacity of cells with *THOR* overexpression and *IGF2BP1* knockdown. In both H1437 and SKMEL5 cells, the proliferative advantage conferred by *THOR* overexpression was abrogated by reduction of *IGF2BP1* levels in those cells, with a particularly striking phenotype in the SKMEL5 cell line (**Figure 5.4d**). These data suggest that the binding interaction between *THOR* and IGF2BP1 leads to functionally relevant consequences in cells that have implications for the oncogenicity of *THOR*.

To more broadly assess the transcriptional phenotype of *THOR*-knockdown in comparison to *IGF2BP1*-knockdown, differential expression was assessed for RNA-seq

performed with two independent siRNAs targeted to *THOR*, *IGF2BP1*, and *HUR*. Providing further evidence for a shared functional role for *THOR* and *IGF2BP1*, we observed significant overlap between the genes with significant differential expression upon *THOR* and *IGF2BP1* knockdown (**Figure 5.5a,b**). As a negative control, we assessed the transcriptional changes upon knockdown of *HUR* that failed to exhibit a similar differential expression profile to that of *THOR* and *IGF2BP1* knockdown. The gene signatures most altered upon knockdown of *THOR* and *IGF2BP1* were also highly concordant when comparing GSEA [164] analyses upon gene expression changes following knockdown (Pearson r = 0.50; **Figure 5.5c**). The gene signature changes following HUR knockdown, however, were not correlated to those following either *THOR* (**Figure D.8a**) or *IGF2BP1* (**Figure D.8b**) knockdown. Interestingly, two independent gene signatures associated with metastasis and relapse of melanoma[219,220] were among the top gene signatures altered upon knockdown of both *IGF2BP1* and *THOR* (**Figure 5.5c**), further implicating the *THOR-IGF2BP1* relationship in cancer progression, particularly in melanoma.

5.3.7 *THOR* knockout zebrafish exhibit fertilization defects and resistance to melanoma formation

Given that we observed sequence conservation of *THOR* (**Figure 5.1a,b,d**, and **Figure D.2**), a conserved tissue expression pattern (**Figure 5.1e,g,h**), and conservation of its binding interaction to IGF2BP1 (**Figure 5.3a,c,d**), we set out to interrogate its potential function in a different animal model, extending the implications of its functionality beyond human cancer cell lines (**Figure 5.2** and **Figure 5.3**). The zebrafish animal model has become a commonly used system for investigation of development [41,221], and more recently it has become a relevant model system for cancer investigation [222,223]. With the advent of CRISPR-Cas genome editing technology [204], the genomes of model organisms

can be molded to interrogate specific genomic questions with ease [224,225]. Thus, utilizing the CRISPR-Cas9 genome editing system, we produced a transgenic *THOR* knockout zebrafish line. Firstly, two guide RNAs (gRNAs) targeting the conserved region of z-*THOR* (**Figure D.9**) and Cas9 mRNA were injected into zebrafish embryos, producing a mosaic F0 generation that was subsequently mated to wildtype zebrafish to generate heterozygous F1 offspring ($THOR^{+/-}$). The heterozygotes were then mated with one another to generate a population of homozygous *THOR* knockout zebrafish ($THOR^{-/-}$) in the F2 generation (**Figure 5.6a**). Knockout of z-*THOR* was validated at the DNA and RNA level (**Figure D.9**). As many interesting mammalian lncRNAs lack the robust evolutionary conservation of *THOR*, the generation of a transgenic $THOR^{-/-}$ zebrafish provides a unique and powerful tool to investigate the function of this lncRNA.

Upon generation of $THOR^{-/-}$ zebrafish, we observed a striking phenotypic effect on the fertility of $THOR^{-/-}$ zebrafish in comparison to wildtype zebrafish, with 55% of embryos from mating of $THOR^{-/-}$ zebrafish either dead or unfertilized 6 hours post fertilization (hpf), compared to only 11% from wildtype mating (**Figure 5.6b**). This fertilization effect is an intriguing phenotype given the cancer/testis expression pattern observed of h-*THOR* and the testis specific expression for z-*THOR* (**Figure 5.1h**). Moreover, when mating wildtype males to female $THOR^{-/-}$ zebrafish, the fertilization defect was substantially diminished, while mating wildtype females to male $THOR^{-/-}$ zebrafish produced a significant fertilization defect in the zebrafish offspring (**Figure 5.6c**). This finding supports the role of *THOR* in the testis, and suggests a primary functional role in fertility for *THOR* in the testis. Interestingly, when assessing the subpopulations of sperm at various stages of development, expression of *THOR* was found to be isolated to spermatocytes in meiosis II at much higher levels than sperm at any other stage of development (**Figure 5.6d** and **Figure D.10a,b**).

Having shown a shared role transcriptional modulation for *THOR* and *IGF2BP1* in human cancer cell lines, we set out to corroborate shared gene expression changes in the zebrafish model, observing a significant decrease in the expression of *IGF2BP1* targets in *THOR*$^{-/-}$ zebrafish embryos compared to wildtype embryos (**Figure 5.6e**). Additionally, given that the transcriptional changes resulting from *THOR* and *IGF2BP1* knockdown exhibited strong association with previously identified melanoma gene signatures, we aimed to investigate the oncogenicity of *THOR* using a zebrafish melanoma model wherein driver genes are injected into zebrafish to elicit various cancer phenotypes [226]. Specifically, we utilized a previously described zebrafish melanoma model that employs embryonic injection of human NRAS-K61 driven by the mitfa promoter (a zebrafish gene expressed in melanocytes) resulting in the formation of easily visible zebrafish melanoma (**Figure 5.4a**) [227]. Using this NRAS melanoma system, we observed a striking resistance to melanoma development in the *THOR*$^{-/-}$ zebrafish (**Figure 5.6f**). The fertilization defect and resistance to melanoma mediated by *THOR* present compelling evidence for a conserved role of *THOR* in vertebrate physiology and pathophysiology.

5.3.8 Human *THOR* enhances the onset of melanoma in zebrafish

In order to further investigate the role of *THOR* in zebrafish melanoma development, we assessed the function of the addition of human *THOR* to zebrafish embryos utilizing the mitfa-promoter driven injection method with injection of mCherry used as a negative control (**Figure 5.6g**)[226]. Injection was performed on p53 knockout (p53$^{-/-}$) zebrafish to enhance the melanoma phenotype observed, as previously described [227]. The overexpression of h-*THOR* in zebrafish was sufficient for a significant increase in not only the onset of melanoma development (**Figure 5.6h**), but also on the size of the melanoma tumors that developed (**Figure 5.6i,j**). NRAS 61K induced melanomas

(**Figure 5.6k**) were positive for Melan-A staining, a mitf target gene and marker for melanoma in human specimens (**Figure 5.6l**). Additionally, the zebrafish melanomas were also positive for p-ERK, a downstream molecule of the IGF2 signaling cascade (**Figure 5.6m**) [228], a finding mirroring the IGF2 signaling regulation we observed with modulating levels of *THOR* in human cancer cell lines (**Figure D.7d,e**). Thus, the striking ability of the human isoform of *THOR* to promote melanoma in zebrafish argues for an evolutionarily conserved role of *THOR* in mediating cellular processes that are potentially involved in tumor development.

5.4 Discussion

We have identified the first cancer/testis lncRNA, *THOR*, as a lncRNA with a cancer/testis expression pattern that exhibits a conserved interaction with IGF2BP1, potentially promoting oncogenesis (**Figure 5.7**). In doing so we have defined a new class of lncRNA that may provide insight into our understanding of cancer biology, and in line with other cancer/testis antigens, this new class of lncRNA may provide potential for development of future cancer therapeutics [66,205]. LncRNAs are highly abundant in the human transcriptome, but little is known regarding the ways in which most of these lncRNAs are functioning in cells. Thus, in this study, we aimed to define a new class of lncRNA, and perform a thorough investigation of its functional relevance in cells. We therefore enforced significant evolutionary conservation as a selection criterion for nomination of *THOR* (**Figure 5.1a,b**) in order to leverage an animal model to robustly interrogate its role in cells. With a testis-specific tissue expression pattern observed in humans, mouse, and zebrafish, *THOR* presented the possibility of bearing an evolutionarily conserved function, strengthening its likelihood of performing an important role in many vertebrate species (**Figure 5.1e,g,h**). Other interesting cancer/testis lncRNAs that are not highly conserved likely exist, and present an exciting

avenue for future studies. In line with our goal of identifying a conserved function for *THOR*, we identified protein-binding interactions for both human *THOR* and zebrafish *THOR* in human and zebrafish cells (**Figure 5.3a**). This method provided a powerful way to identify the *THOR*-IGF2BP1 interaction as one conserved across multiple vertebrate species.

We subsequently implicated *THOR* in assisting IGF2BP1 in its regulation of multiple mRNA targets (**Figure 5.4a**), potentially via modulating the ability of IGF2BP1 to bind to these targets (**Figure 5.4b**), and show that this *THOR*-IGF2BP1 relationship is associated with the proliferation cancer phenotype observed with *THOR* overexpression (**Figure 5.4d**). Post-transcriptional regulation of mRNAs is controlled tightly by their interaction with various RNA-binding proteins [212,229,230]. IGF2BPs are almost exclusively observed in the peri-nuclear region of cytoplasm where they associate with target mRNAs in cytoplasmic ribonucleoprotein complexes (mRNPs) and play a central role in mRNAs regulation [212]. We have shown in this study that *THOR* binds to IGF2BP1 and regulates broad target mRNAs stabilization. In light of the surprisingly long half-life of IGF2BP-RNA complexes *in vitro* [231], *THOR* may have a potential role in promoting the formation of 'stable' protein–RNA complexes, although further experiments are required to determine the precise mechanism.

While animal models have been employed to investigate lncRNAs [41,203], we present the first investigation of an oncogenic lncRNA employing a zebrafish cancer assay (**Figure 5.6a**). Moreover, many of these prior studies were limited to investigation of lncRNAs with little tissue expression in humans. *THOR*, however, is expressed in multiple species and is also widely expressed in a myriad of human cancers (**Figure 5.2A**), elevating its relevance to our understanding of normal and disease processes in humans. We show that loss of zebrafish *THOR* results in aberrant fertilization (**Figure**

5.6b,c), targeting this function to the male testis, and we also show that loss of *THOR* yields a resistance to melanoma formation (**Figure 5.6f**), implicating a function for z-*THOR* in the zebrafish. Interestingly, we also show that the human *THOR* is capable of utilizing zebrafish cellular machinery to also produce a striking cancer phenotype (**Figure 5.6h,i**). This trans-species function of *THOR* provokes interest, suggesting that its sequence and function have both been evolutionarily selected for. Investigating highly conserved lncRNAs like *THOR* elucidate key mechanisms through which lncRNAs function, and as the scientific community continues to investigate the multitude of unstudied novel lncRNAs, the zebrafish cancer model system presented here provides a powerful platform that can be used for further investigation of lncRNA function.

5.5 Methods

5.5.1 RNA-seq data processing

TCGA prostate FASTQ files were obtained from the CGhub. Reads were aligned using STAR version 2.4.2[124] and read abundance was calculated using FeatureCounts version 1.4.6[232], providing the MiTranscriptome gene annotation GTF (mitranscriptome.org) [192]. GTEX (accession: phs000424.v6.p1) and TCGA (accession: phs000178.v8.p7) data were downloaded from dbGAP.

5.5.2 Conservation analysis

Evolutionary conservation of transcripts was assessed via the fraction of significantly conserved bases ($P \leq 0.01$, phyloP algorithm), and the most conserved 200nt sliding window (phastCons scores averaged within each window). For contiguous sliding

window conservation an average PhastCons probability of 0.9986 was used to identify ultraconserved elements as previously described[192].

5.5.3 Cell lines

All cell lines were obtained from the American Type Culture Collection (Manassas, VA). Cell lines were maintained using standard media and conditions. Specifically, NCI-H1299 and NCI-H1437 were maintained in RPMI 1640 (Invitrogen) plus 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. MM603 and SK-MEL-5 were maintained in DMEM (Invitrogen) plus 10% FBS and 1% penicillin-streptomycin. All cell lines were grown at 37°C in a 5% CO_2 cell culture incubator and genotyped for identity at the University of Michigan Sequencing Core and tested routinely for Mycoplasma contamination. *THOR* or control-expressing cell lines were generated by cloning *THOR* or control into the pLenti6 vector (Invitrogen) using pCR8 non-directional Gateway cloning (Invitrogen) as an initial cloning vector and shuttling to plenti6 using LR clonase II (Invitrogen) according to the manufacturer's instructions. Stably-transfected NCI-H1437 and SK-MEL-5 cells were selected using blasticidin (Invitrogen). All lentiviruses were generated by the University of Michigan Vector Core.

5.5.4 Tissue Samples

All human subject studies were approved by the Institutional Review Board of the University of Michigan. The lung cancer and paired non-tumoral lung tissues were obtained from patients undergoing curative cancer surgery during the period from 1991 to 2012 at the University of Michigan Health System. None of the patients included in this study received any preoperative radiation or chemotherapy. All melanoma tissues were procured from the University of Michigan Hospitals Cutaneous Surgery and

Oncology Program with appropriate informed consent. Resected specimens were frozen in liquid nitrogen and then stored at −80°C until use. Total RNA panels from human and mouse normal tissues were purchased from Clontech and Zyagen. Mouse embryos were obtained from the University of Michigan Transgenic Core. Zebrafish tissues and embryos were obtained from AB strain wild type zebrafish.

5.5.5 RNA isolation and cDNA synthesis

Total RNA from human and mouse normal tissues were purchased from Clontech and Zyagen. Mouse embryos were obtained from the University of Michigan Transgenic Core. Zebrafish tissues or embryos were obtained from AB strain wild type fish. Nuclear and cytoplasmic fractions were separated using NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo) according to the manufacturer's instructions. Total RNA was isolated using miRNeasy Mini Kit (Qiagen) with DNase I (Qiagen) digestion according to the manufacturer's instructions. RNA integrity was verified on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). cDNAs weresynthesized from total RNA using Superscript III (Invitrogen) and random primers (Invitrogen).

5.5.6 Quantitative Real-time PCR

Quantitative Real-time PCR (qRT-PCR) was performed using Power SYBR Green Mastermix (Applied Biosystems, Foster City, CA) on an Applied Biosystems 7900HT Real-Time PCR System. All oligonucleotide primers were obtained from Integrated DNA Technologies (Coralville, IA). The housekeeping genes, *GAPDH*, *HMBS* and *UBC*, were amplified as controls. Fold changes were calculated relative to housekeeping genes and normalized to the median value of the lung benign samples.

5.5.7 RNA-ligase-mediated Rapid Amplification of cDNA Ends (RACE)

5′ and 3′ RACE was performed using the GeneRacer RLM-RACE kit (Invitrogen) according to the manufacturer's instructions. RACE PCR products were obtained using Platinum Taq High Fidelity polymerase (Invitrogen). RACE-PCR products were separated on a 2% agarose gels, bands excised and the extracted DNA (Gel Extraction kit, Qiagen) were cloned into pCR4-TOPO vector (Invitrogen), and sequenced bidirectionally using M13 forward and reverse primers at the University of Michigan Sequencing Core. At least four colonies were sequenced for every gel product that was purified and the data was analyzed using Sequencher software (GeneCodes).

5.5.7 siRNA Knockdown Experiments

Knockdown experiments were carried out in approximately 1 - 2 x 10^5 cells plated in 100 mm dishes. While *THOR* knockdown in MM603 cells was achieved with two sequential transfections (at 24 hr and 48 hr post-plating) with 50 achieved with two sequential transfections (at 24 hr and 48*THOR* knockdown in H1299 cells only one siRNA transfection (24hr post plating) was done. Only one transfection was performed for all protein coding gene knockdowns. Knockdowns were performed with RNAiMAX (Invitrogen) in OptiMEM media and its efficiency was determined by qRT-PCR 96 hr post-plating. All siRNAs were purchased from Dharmacon.

5.5.8 Overexpression Studies

The *THOR* expression construct were generated by amplifying the full-length transcript from NCI-H1299 cells and subcloning into the pLenti6 expression vector (Invitrogen),

LacZ constructs were used as controls. Following Sanger sequencing (University of Michigan Sequencing Core) confirmation of the inserts, lentiviruses were generated at the University of Michigan Vector Core. NCI-H1437 and SK-MEL-5 cells were infected with lentiviruses expressing *THOR* or LacZ and stable pools and clones were generated by blasticidin selection (Invitrogen).

5.5.9 Cell Proliferation Assays

Proliferation experiments were carried out by plating 1-2 x 10^4 cells or 0.15-0.3 x 10^4 cells in 24-well or 96-well plates respectively and grown in regular media. Growth rate was monitored by IncuCyte live-cell imaging system (Essen Biosciences) for the specified durations.

5.5.10 Murine Subcutaneous *In Vivo* Models

All experimental procedures were approved by the University of Michigan Committee for the Use and Care of Animals (UCUCA). Mice (CB-17 SCID) aged 5-7 weeks were injected with 0.25 x 10^6 cells with a Matrigel scaffold (BD Matrigel Matrix, BD Biosciences) in the posterior dorsal flank region (n = 10 per cell line). Tumors were measured weekly using a digital caliper, and endpoint was determined as a tumor volume of 1000 mm^3. Upon reaching endpoint, or if the animal became fatally ill, the mouse was euthanized and the primary tumor resected. The resected specimen was divided in two halves: one preserved in 10% buffer formalin and the other snap frozen.

5.5.11 RNA *In Situ H*ybridization in Testis

THOR ISH was performed on thin (approximately 4 μm thick) tissue sections (Advanced Cell Diagnostics, Inc., Hayward, CA), as described previously [233]. Appropriate batch positive and negative controls demonstrated expected staining patterns (data not shown). Slides were examined for THOR ISH signals in morphologically intact cells and scored manually by a study pathologist (R.M.). Specific THOR ISH signal was identified as brown, punctate dots.

5.5.12 Single-Molecule Fluorescence *in situ* Hybridization in Cell Lines

smFISH was performed as described [187], with some minor modifications. Cells were grown on 8-well chambered coverglasses, formaldehyde fixed and permeablized overnight at 4°C using 70% ethanol. Cells were rehydrated in a solution containing 10% formamide and 2x SSC for 5 minutes and then treated with 10nM FISH probes for 16 h in 2x SSC containing 10% dextran sulfate, 2 mM vanadyl-ribonucleoside complex, 0.02% RNAse-free BSA, 1 μg/μL E.coli tRNA and 10% formamide at 37°C. After hybridization the cells were washed twice for 30 minutes at 37°C using a wash buffer (10% formamide in 2x SSC). Cells were then mounted in solution containing 10 mM Tris/HCl pH 7.5, 2x SSC, 2 mM trolox, 50 μM protocatechiuc acid (PCA) and 50nM protocatechuate dehydrogenase (PCD). FISH samples were imaged in 3 dimensions using HILO illumination as described [188]. Images were processed using custom-written macros in ImageJ. Analysis routines comprised of 3 major steps: background subtraction, Laplacian of Gaussian (LoG) filtering and thresholding. Spots with intensity above set threshold are represented in images. All probes were obtained from Biosearch.

5.5.13 Western Blot Analysis

Western blot analysis was performed according to standard procedures using Immobilon-P filters (Millipore) and an Enhanced Chemiluminescence detection system (GE Healthcare).

5.5.14 RNA Immunoprecipitation (RIP) Assay

RIP assays were performed using a Millipore EZ-Magna RIP RNA-Binding Protein Immunoprecipitation kit (Millipore, #17-701) according to the manufacturer's instructions. RIP-PCR was performed using total RNA as input controls and 1:150 of RIP RNA product was used per PCR reaction. The antibodies were used at 3 - 5 ug of antibody per RIP reaction.

5.5.15 RNA Pulldown Assay

RNA-pull down assays were performed using a RiboTrap Kit (MBL, RN1011/RN1012) according to the manufacturer's instructions. Briefly, 5-bromo-UTP (BrU) was randomly incorporated into the *THOR* RNA upon transcription using *THOR* full-length or deleted fragments PCR products as templates. Next Anti-BrdU antibodies conjugated with protein G beads (Invitrogen), were bound to the *in vitro* synthesized RNA before incubating with NCI-H1299 cell lysates for 4 hr. Finally, the samples were washed, eluted, and subjected to Mass spectrometry analysis.

5.5.16 Mass Spectrometry

The samples were treated with SDS-PAGE loading buffer supplied with 10 mM DTT for 5 min at 85ºC. The proteins were alkylated by the addition of iodoacetamide to the final concentration of 15 mM. The samples were subjected to SDS-PAGE and the whole lanes

were cut out and digested with trypsin in-gel for 2 hours. The resulting peptides were extracted, dried and resuspended in 0.1% formic acid with 5% acetonitrile prior to loading onto a trap EASY-column (Thermo Scientific) coupled to an in-house made nano HPLC column (20 cm x 75 um) packed with LUNA C18 media. Analysis was performed on Velos Pro mass spectrometer (Thermo Scientific) operated in data-dependent mode using 90-min gradients in EASY-LC system (Proxeon) with 95% water, 5% acetonitrile (ACN), 0.1% formic acid (FA) (solvent A), and 95% ACN, 5% water, 0.1% FA (solvent B) at a flow rate of 220 nl/min. The acquisition cycle consisted of a survey MS scan in the normal mode followed by twelve data-dependent MS/MS scans acquired in the rapid mode. Dynamic exclusion was used with the following parameters: exclusion size 500, repeat count 1, repeat duration 10 s, exclusion time 45 s. Target value was set at 104 for tandem MS scan. The precursor isolation window was set at 2 m/z. The complete analysis comprised two independent biological replicates.

5.5.17 MS Data Analysis

The resulting spectrum files were transformed into MGF format by MSConvert software and interrogated by MASCOT 2.4 search engine using human UniProt database version 15 concatenated with reverse sequences for estimation of false discovery rate (FDR) and with a list of common contaminants (40729 entries in total). The search parameters were as follows: full tryptic search, 2 allowed missed cleavages, peptide charges +2 and +3 only, MS tolerance 1 Da, MS/MS tolerance 0.5 Da. Permanent post-translational modifications was: cysteine carbamidomethylation. Variable post-translational modifications were: protein N-terminal acetylation, Met oxidation and N-terminal Glutamine to pyro-Glutamate conversion. The remaining analysis was performed as previously described [176]. To summarize, the minimal ion score threshold was chosen such that a peptide false discovery rate (FDR) below 1% was achieved. The

peptide FDR was calculated as: 2 × (decoy_hits)/(target + decoy hits). Spectral counts for all detected proteins were assembled using an in-house written Python script. The adjustment of spectral counts was done by the same script as in[176].

5.5.18 RNA-protein Interaction Assay

The *in vitro* transcribed BrU labeled RNA were heated at 92°C for 2 min (to remove secondary structure), and incubated with recombinant myc-tagged proteins in RIP buffer (150 mM KCl, 25 mM Tris pH 7.4, 0.5 mM DTT, 0.5% NP40, 1 mM PMSF and protease inhibitor (Roche Complete Protease Inhibitor Cocktail Tablets) for 3 hr at 4°C. RNA–protein complexes of interest were then partially purified with anti-myc magnetic beads (Thermo) and the products were treated with proteinase K, to remove the protein components leaving the RNAs intact. The recovered RNAs were extracted using miRNeasy Mini Kit as described above.

5.5.19 RNA Endogenous Degradation Assay

Cells were treated with 5 μg/ml Actinomycin D (Sigma) and collected in Quiazol at the indicated time points after treatment. Purified RNA was subjected cDNA synthesis and qRT-PCR as described above. The slopes for decay plots were determined by simple linear regression, and transcript half-life was calculated as the x intercept at y = 0.5, using GraphPad Prism.

5.5.20 Anchorage-independent soft agar colony formation Assay

For H1437 and H1299 soft-agar colony formation assay, 1-3x10^3 cells were suspended in DMEM containing 0.3 % agar, 10 % fetal bovine serum, and layered on DMEM

containing 0.6% agar, 10% FBS in 6-well plate. After 2 weeks incubation, colonies were stained with iodonitrotetrozolium chloride (Sigma) for overnight. Visible colonies were enumerated from two replicate wells.

5.5.21 RNA-seq Data Processing

RNA-sequencing reads were quantified to the human transcriptome (GENCODEv24) using Kallisto (v0.43.0) [128]. GENCODEv24 GTF was obtained from GENCODE [48], and transcriptome fasta file was produced using the *rsem-prepare-reference* function of RSEM (version 1.2.26) [234]. Kallisto index was generated using the *kallisto index* function. Transcript level quantification obtained using the *kallisto quant* function. Gene level expression obtained by summing the TPM values for all transcripts within each gene.

5.5.22 RNA-Seq Differential Expression Testing

Differentially expressed genes were obtained by comparing non-targeting shRNA control to each of the two replicates for the three genes tested (i.e., *THOR*, *IGF2BP1,* and *HUR*) using DESeq2 [113]. Significantly differentially expressed genes were defined as genes with a greater than 2^0.75 log-fold-change with a q-value < 0.05.

5.5.23 Gene signature GSEA analysis

For each gene a rank list was generated by ordering each gene in the differential expression analysis by the DESeq log-fold-change value (*log2foldchange)* by the q-value (*padj)*. These rank lists were used in a weighted, pre-ranked GSEA analysis against MSigDBv5. Significant associations were determined for any gene set having an FWER p-value below 0.01.

5.5.24 Zebrafish Maintenance

Zebrafish (*Danio rerio*) were maintained and reared as described previously [235]. Approval for zebrafish research was obtained from the University Committee on the Use and Care of Animals (UCUCA) of the University of Michigan.

5.5.25 Cas9 Target Site Design, Vector Construction and *in vitro* RNA Synthesis

The plasmids MLM3613 bacterial Cas9 expression vector ((Addgene plasmid # 42251), [236]) and DR274 sgRNA expression vector ((Addgene plasmid # 42251), [236]) were purchased from Addgene (Cambridge, MA). We selected two sgRNA targets with ZIFIT Targeter (http://zifit.partners.org/zifit/Introduction.aspx) to generate a deletion of the conserved portion of THOR in zebrafish. For each target we annealed the oligonucleotide pairs and ligated into Bsa I-linearized DR274. sgRNAs were transcribed from Dra I-linearized templates using the MEGAscript® T7 Kit (Ambion). Cas9 mRNA was transcribed *in vitro* with the mMESSAGE mMACHINE T7 ULTRA kit (Ambion). RNAs were purified by RNA Clean &Amp; Concentrator (Zymo Research) and re-dissolved in RNase-free water.

5.5.26 Microinjection of Zebrafish Embryos

One-cell stage embryos were microinjected with 250 ng/ul Cas9 mRNA and 150 ng/ul of each sgRNAs by using a pneumatic pico-pump (PV-820, World Precision Instrument).

5.5.27 DNA Isolation and PCR Analysis for Identifying Deletion

For embryonic gDNA extraction, 20 pooled embryos were lysed in 20 µL lysis buffer (10mM Tris HCl pH8.0, 2mM EDTA, and 0.2% Triton) containing proteinase K (10 [illegible]or embryonic gDNA extraction, 20 pooled embryos were lysed in 20 [illegible]

D[illegible] [illegible]rose gel.

5.5.28 Germ cell sorting

Germ cell sorting was performed as described previously [237]. Briefly, zebrafish testes were placed in 6 ml Collagenase I/Dnase I solution (200 U/ml Collagenase type I (Sigma-Aldrich) and 5 µg/ml DNAse I (Invitrogen) in Gey's Balanced Salt Solution (GBSS) (Sigma-Aldrich)) and shaken at 150 rpm for 10 min at 35°C. The temperature and agitation speed were the same for all subsequent incubation steps. The testes were gently pipetted halfway into the 10 minute incubation. Tubules were settled for 2 min at room temperature (RT), then the supernatant, enriched in interstitial testicular cells (somatic cells), was harvested. 6 ml Collagenase I/Dnase I/Trypsin solution (200 U/ml Collagenase type I, 5 µg/ml DNAse I and 0.025% Trypsin (Gibco) in GBSS) was added to the pellet and the tubules were gently pipetted. Halfway into the 25 minute digestion period, 60 µl of 2.5% Trypsin was added, and the tubules were pipetted again. At the end of the incubation time, pipetting was repeated. The resulting cell-dense suspension was passed through a Nylon cell strainer (Falcon). To the resulting filtered cell suspension 10 µl of 1 mg/ml DNAse I and 10 µl of 10 mg/ml Hoechst 33342 (Life Technologies) were added and incubated for 20 min. Halfway into the 20 minute period, the suspension was pipetted. At the end of incubation, 600 µl of FBS was added to inactivate the trypsin. After determining the cell number, the suspension was spiked

with 10 μl of 1 mg/ml DNAse I, and stained with Hoechst dye for the final 6 μg Hoechst/million cells. The suspension was incubated for 25 min. The cells were then stained with 10 μl of PI (Sigma-Aldrich) at RT. Cell sorting and data analysis were performed by University of Michigan Flow Cytometry Core using MoFlo Astrios cell sorter (Beckman Coulter). Hoechst was excited using 375 nm laser, and the dye's wide emission spectrum detected in two distinct channels: the "Ho Blue" (450/40 nm band-pass filter) and the "Ho Red" (670 nm long pass filter). Cells from each subpopulation were sorted and subjected to qRT-PCR.

5.5.29 Generation of THOR Knockout Zebrafish

F0 zebrafish was crossed to wild-type AB* to generate F1 embryos that were screened for THOR deletion. F0 zebrafish that were able to produce germ-line deletion of THOR were crossed to produce F1 heterozygotes, which were subsequently genotyped and crossed to generate THOR homozygous. THOR homozygotes and matched wild-type fish were used for phenotypic analyses.

5.5.30 Zebrafish Mosaic Melanoma Model

All transgenic constructs were made using the Tol2/Gateway kit (a gift from Dr. Kristen Kwan)[238]. Full-length human NRAS 61K was amplified from pBabe NRAS 61K construct (a gift from Channing Der (Addgene plasmid # 12543),[239]) subcloned into BglII and BamHI restriction enzyme sites. GFP was amplified and subcloned into SalI and BamHI sites of pME entry vector. *THOR* transcript was amplified from *THOR* expression plasmid and cloned into SalI and BamHI sites of pME entry vector. The mitfa promoter was amplified using gDNA extracted from embryos as a template, and cloned into p5'E entry vector; and polyA tail was cloned into p3'E entry vector. These

were assembled into the Tol2 destination vector using MultiSite Gateway Technology system (Invitrogen). 2.5 ng/uL of mitfa:NRAS 61K was co-injected into one-cell stage of p53-/- embryos[240] with mitfa: *THOR* or mitfa:mCherry (25 ng/ul each) with 2.5 ng/ul of Tol2 mRNA. For injections into *THOR* -/- embryos and their corresponding wild type embryos, 5 ng/uL of mitfa:NRAS 61K was injected into one-cell embryos with 5 ng/ul of Tol2 mRNA. Zebrafish were inspected weekly, then euthanized when 17 weeks old and fixed in 4% paraformaldehyde overnight. After taking photos, they were then decalcified in 0.5 M ethylenediaminetetraacetic acid before paraffin embedding and sectioning. Staining and immunohistochemistry were done using the standard techniques by the University of Michigan URAM Core. Percentage of melanoma area per body was calculated using ImageJ software.

5.6 Acknowledgements

We thank S. Shukla, L. Xiao, and Y. Zhang for technical assistance; J. Shavit for discussion; T. Saunders for mice embryos. This work was supported in part by the National Cancer Institute Early Detection Research Network (UO1CA113913) and US National Institutes of Health grants RO1 CA154365 (D.G.B. and A.M.C.). A.M.C. is an American Cancer Society Research Professor, a Howard Hughes Medical Institute Investigator, and a Taubman Scholar of the University of Michigan. Y.H. was supported by a Genentech Fellowship. R.M. is supported by a Prostate Cancer Foundation Young Investigator Award and by US Department of Defense Post-Doctoral Fellowship W81XWH- 13-1-0284. Y.S.N. was supported by a University of Michigan Cellular and Molecular Biology National Research Service Award Institutional Predoctoral Training Grant. W.Z. was supported by grants from the National Institutes of Health (NIH) (DK091405, DK081943 sub), the Janette Ferrantino Investigator Award and the

American Society of Nephrology Carl W. Gottschalk Scholar Research Grant. Y.S.N. is supported by an NIH F30 grant (F30-CA-200328).

5.7 Author Contributions

Y.H., Y.S.N., W.Z. and A.M.C. designed the project and directed experimental studies. Y.H., J.R.P., S.P. and K.S. performed in vitro studies. Y.H., J.T., J.E.W. and W.Z. performed in vivo studies. R.M. performed histopathological analyses. Y.S.N and M.K.I. performed bioinformatics analysis. A.P. performed the proteomics analysis. F.S. and X.C. performed library construction. S.G., S.M.F. and H.H.B registered the transcript and designed the ASOs. T.M.J. and D.G.B. facilitated biological sample procurement. Y.H., Y.S.N., S.M.D., R.M., F.Y.F, W.Z. and A.M.C. interpreted data and wrote the manuscript.

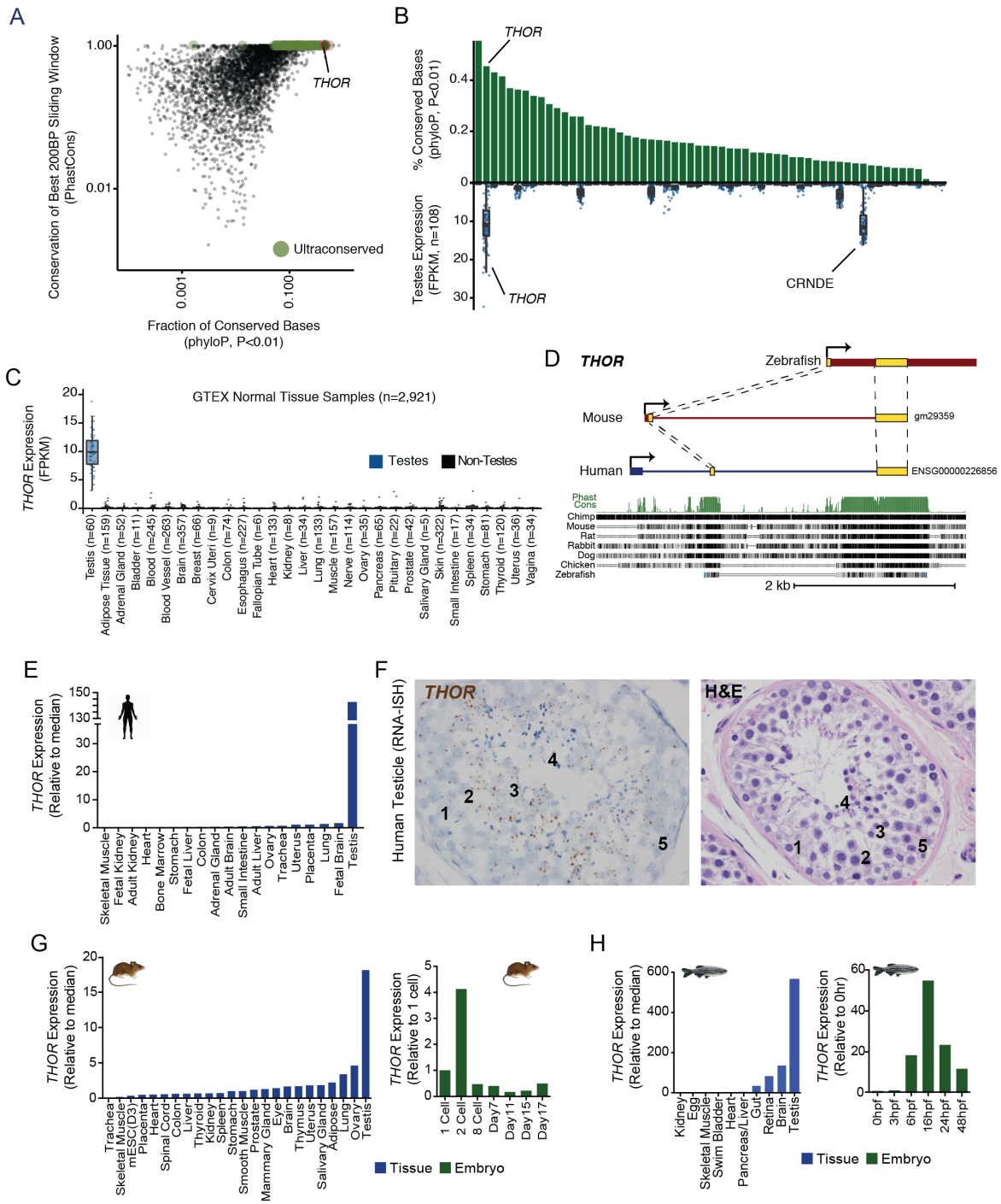


Figure 5.1 ***THOR* is a conserved testis specific lncRNA**

(a) Scatter plot depicting the distribution of basewise transcript conservation levels (x axis) and the average conservation for the best 200bp window (y axis) for all intergenic transcripts expressed at 1 FPKM or more in the top 1% of TCGA samples. Full transcript conservation levels were measured using the fraction of conserved bases (PhyloP $p <$ 0.01). Sliding window conservation levels were measured using the average PhastCons

score across 200bp regions along the transcript. Green points indicate transcripts with 200bp windows that meet the criteria for 'ultraconserved' regions. **(b)** Dual plot depicting the fraction of conserved bases (top, green, **(a)**, *x* axis) and the expression across testes RNA-seq samples (bottom, blue) for all ultraconserved lncRNAs identified in **(a-c)** Expression in FPKM of *THOR* amongst the GTEX normal tissue RNA-seq dataset, spanning a myriad of different normal tissue types. **(d)** Genome browser depiction of *THOR* and its conserved analogues in mouse and zebrafish. *THOR* is annotated in the mouse as gm29359. Multiz alignment of multiple vertebrate species depicted as well as the per base PhastCons conservation score. **(e)** Estimation of *THOR* mRNA expression by qRT-PCR in human adult normal tissue panel. **(f)** H&E stain of human testis at high magnification (400x) (right), and RNA-ISH of *THOR* in human testis (left). Various cells of the testis are labelled as follows: (1) spermatogonia, (2) spermatocytes, (3) spermatids, (4) mature spermatozoa, and (5) scattered Sertoli cells with a single central prominent nucleolus. *THOR* expression is observed in the spermatid and spermatocyte. **(g)** Measurement of mouse *THOR* expression by qRT-PCR on an adult murine tissue panel (left) and embryos (right). **(h)** Quantification of zebrafish *THOR* expression by qRT-PCR on a piscine tissue panel (left) and embryos (right).

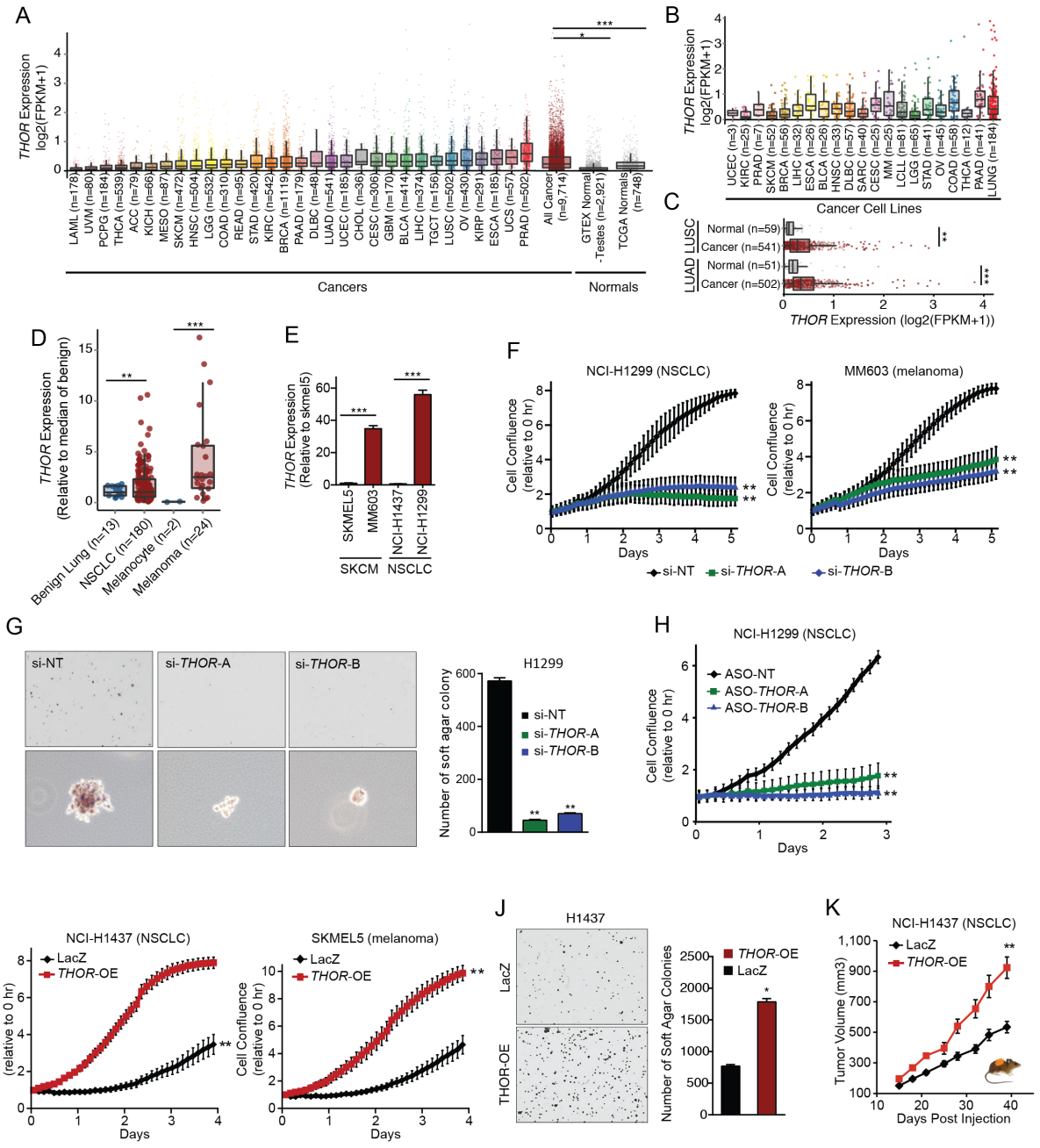


Figure 5.2 ***THOR* is expressed in cancers and potentiates tumorigenesis**

(a) Expression of *THOR* across a panel of 9,714 TCGA tumors from a myriad of different tissues, and 2,921 normal tissue samples from GTEX and 748 normal tissue samples from the TCGA. Expression represented as log2 (FPKM+1). **(b)** Expression of *THOR* in the CCLE cell line panel. Expression represented as log2 (FPKM+1). **(c)** Expression of *THOR* in the TCGA lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) samples represented alongside each tissue's matched normal samples. *THOR* is significantly overexpressed in both LUAD and LUSC. **(d)** qRT-PCR validation in an

independent tissue cohort reveals *THOR* upregulation in cancer compared to benign samples both in lung adenocarcinoma (benign, n=13; cancer, n=180), and melanoma tissues (benign, n = 2; cancer, n=24). **(e)** Expression levels of *THOR* in two melanoma (SKCM) and two non-small cell lung cancer (NSCLC) cell lines. Data show mean ± S.D. **(f)** Cell proliferation assays for NCI-H1299 and MM603 cells treated with 2 independent *THOR* siRNAs. **(g)** Anchorage-independent growth of H1299 cells transfected with non-targeting siRNA (si-NT) or two *THOR* siRNAs (siTHOR-A, siTHOR-B). Left, representative image of surviving colonies and individual colony. Right, quantification of number of colonies. **(h)** Cell proliferation assays for NCI-H1299 cells treated with 2 independent *THOR* ASOs. Data show mean ± S.E. from one of the two independent experiments. **(i)** Cell proliferation assays in NCI-H1437 and SK-MEL-5 cells stably transfected with *THOR* overexpression or LacZ control lentivirus . Data show mean ± S.E. from one of the two independent experiments. **(j)** Anchorage-independent growth of LacZ or *THOR* overexpressing H1437 cells. Left, representative images of surviving soft agar colonies. Right, quantification of number of colonies. **(k)** *THOR* overexpressing NCI-H1437 cell line xenografts (N=10) demonstrate increased tumor growth relative to control LacZ samples (N=10). Tumor volumes at each time point by caliper measurement are shown. Asterisk (*) indicates $P \leq 0.001$ by a two-tailed Student's t-test. Data show mean ± S.E.M. from one of the two independent experiments. For all panels, asterisk (*) indicates $P \leq 0.01$, (**) indicates $P \leq 0.001$, (**) indicates $P \leq 0.0001$ by a two-tailed Student's t-test.

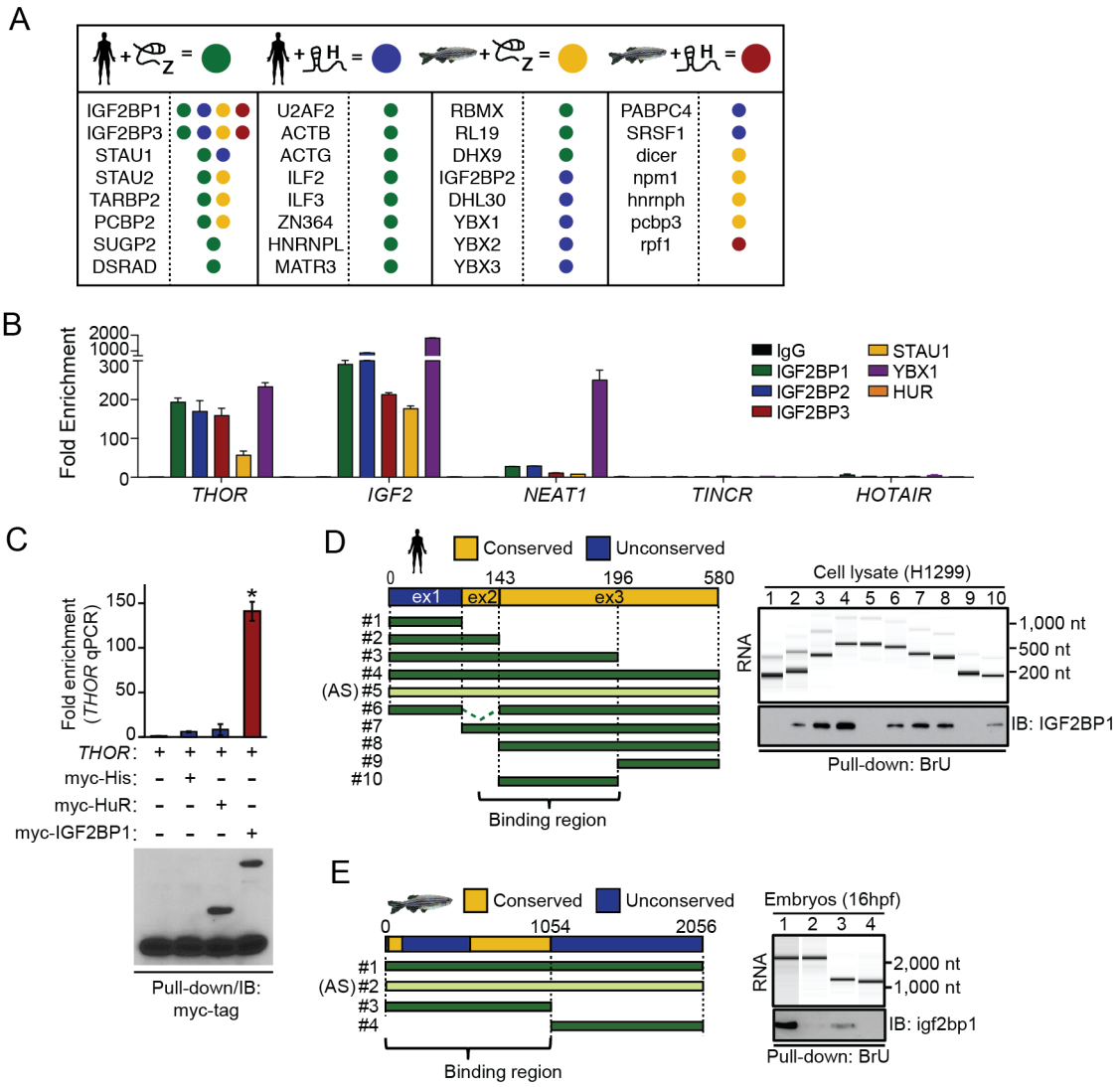


Figure 5.3 Conserved interaction of *THOR* and IGF2BP1

(a) Table reporting the protein binding partners for *THOR* in four different experimental conditions of RNA pull-down analysis: zebrafish *THOR* added to human H1299 cell lysate (green), human *THOR* added to human H1299 cell lysate (blue), zebrafish *THOR* added to zebrafish embryo lysate (yellow), and human *THOR* added to zebrafish embryo lysate (red). All proteins bound in any condition are displayed in the table, and each dot represents binding in the respective condition. **(b)** qRT-PCR following RIP of IGF2BP1, IGF2BP2, IGF2BP3, STAU1, YBX1, HUR, and IgG in H1299 cells. Data show mean ± S.D. from one of the two independent experiments. **(c)** *In vitro* RNA-protein binding assay. In vitro transcribed *THOR* added to purified myc-tagged proteins. *THOR* qRT-PCR was then performed following anti-myc pull-down. Asterisk (*) indicates $P \leq 0.01$ by a two-tailed Student's t-test. Data show mean ± S.D. from one of the two independent experiments. **(d)** Schematic representation of human *THOR*,

antisense-*THOR* (AS), and various deletion constructs generated to interrogate IGF2BP1 binding (left). Fragment sizes confirmed by PCR (right, top), and binding of each fragment to IGF2BP1 determined via pulldown of BRU-labelled RNA fragments (right, bottom) in H1299 cells. **(e)** Schematic representation of zebrafish *THOR* constructs generated to study IGF2BP1 binding. Fragment sizes confirmed by PCR (right, top), and binding of each fragment to zebrafish igf2bp1 determined via pulldown of BRU-labelled RNA fragments (right, bottom) in 16 hpf embryos.

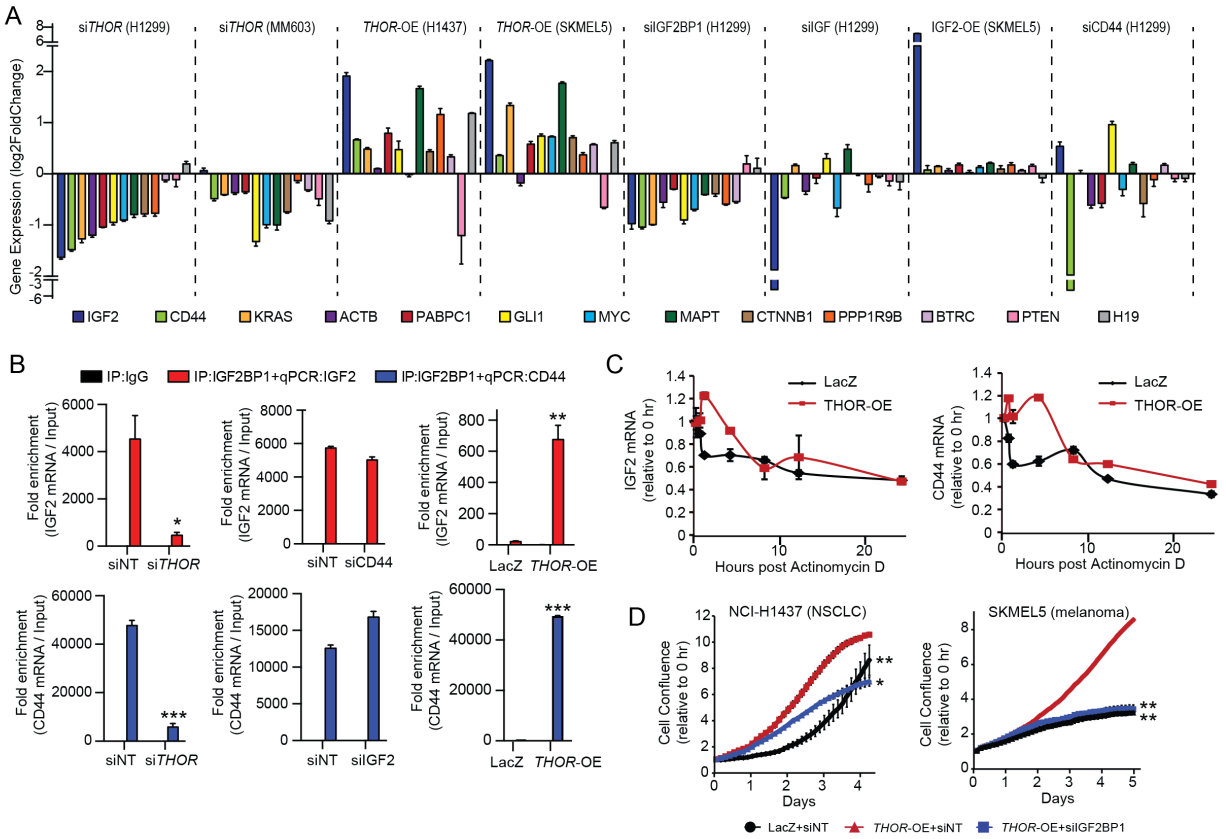


Figure 5.4 Interrogation of the functional relationship of *THOR* and IGF2BP1
(a) Bar plot depiction of the expression levels of 13 canonical IGF2BP1 target genes by qRT-PCR in various conditions: *THOR* knockdown, *THOR* overexpression, *IGF2BP1* knockdown, *IGF2* knockdown, *IGF2* overexpression, and *CD44* knockdown. Data show mean ± S.D. from one of the two independent experiments. **(b)** qRT-PCR expression levels for *IGF2* (red, black) and *CD44* (blue, black) following RIP of IGF2BP1 or IgG as negative control. RIPs performed in H1299 cells under various experimental conditions: *THOR* siRNA knockdown, *IGF2* siRNA knockdown, *CD44* siRNA knockdown, and *THOR* overexpression. Asterisk indicate *P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001 by two-tailed Student's t-test. Data show mean ± S.D. from one of the two independent experiments. **(c)** qRT-PCR expression levels for *IGF2* (top) and *CD44* (bottom) following Actinomycin D treatment in *THOR* or *LacZ* overexpressing H1347 cells. Data show mean ± S.D. from one of the two independent experiments. **(d)** Cell proliferation assays for H1437 and SKMEL5 cells overexpressing *LacZ* control and *THOR* in addition to siRNA knockdown of *IGF2BP1*. Asterisk indicate *P ≤ 0.05; **P ≤ 0.001 by two-tailed Student's t-test. Data show mean ± S.E. from one of the two independent experiments.

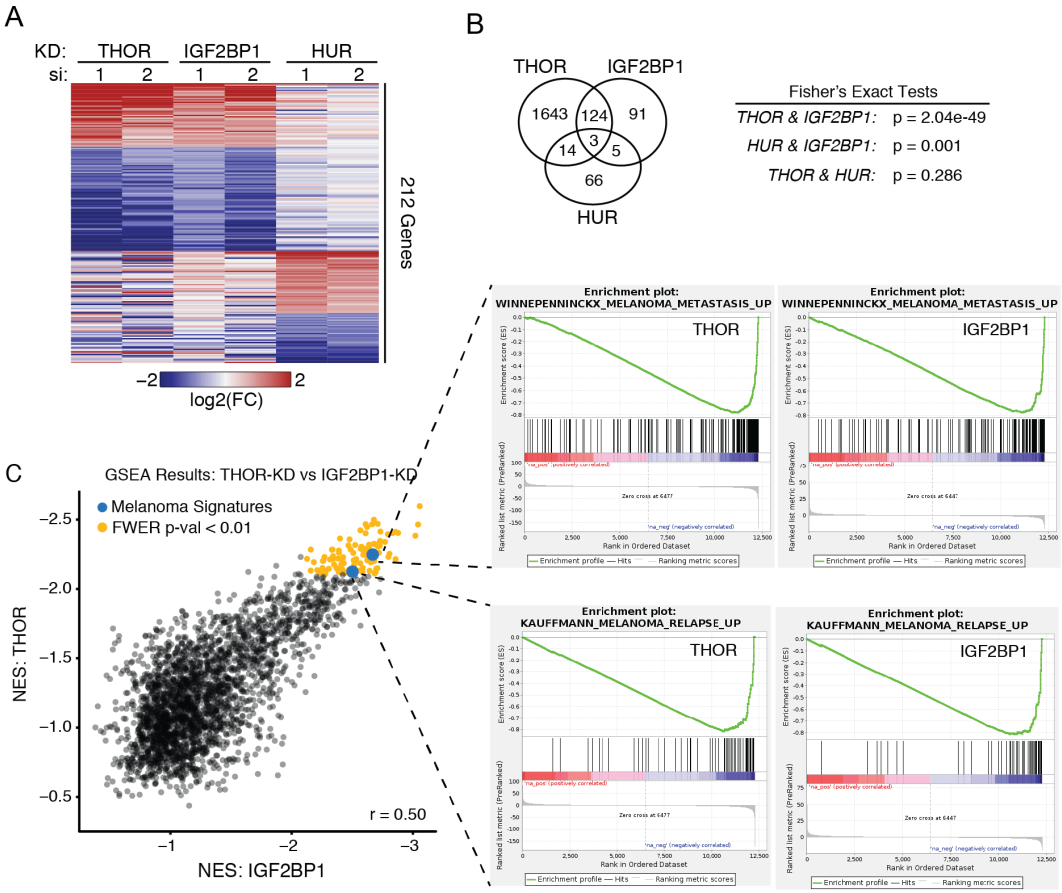


Figure 5.5 Shared transcriptional regulation by *THOR* and *IGF2BP1*
(a) Heatmap depicting the expression of the genes significantly differentially expressed (DESeq FDR < 0.05) in knockown of *THOR* and *IGF2BP1* in addition to those genes with significant differential expression in *HUR* knockdown. Expression depicted as the log2(fold-change) for each siRNA compared to the non-targeting siRNA control. **(b)** Venn diagram depiction of the overlap for the significant differentially expressed genes in *THOR, IGF2BP1,* and *HUR* knockdown. Fisher's exact statistics shown on the right. **(c)** Scatterplot depicting the GSEA performance for MSigDBv5.0 gene signatures with NES < 0 for both *THOR* and *IGF2BP1* knockdown (left). Signatures significant upon knockdown of both genes (FWER p-value < 0.01) depicted in gold. Two melanoma gene signatures depicted in blue. Pearson correlation coefficient shown in bottom right of scatterpot. GSEA plots for two significant melanoma signatures depicted for knockdown of *THOR* and *IGF2BP1* (right).

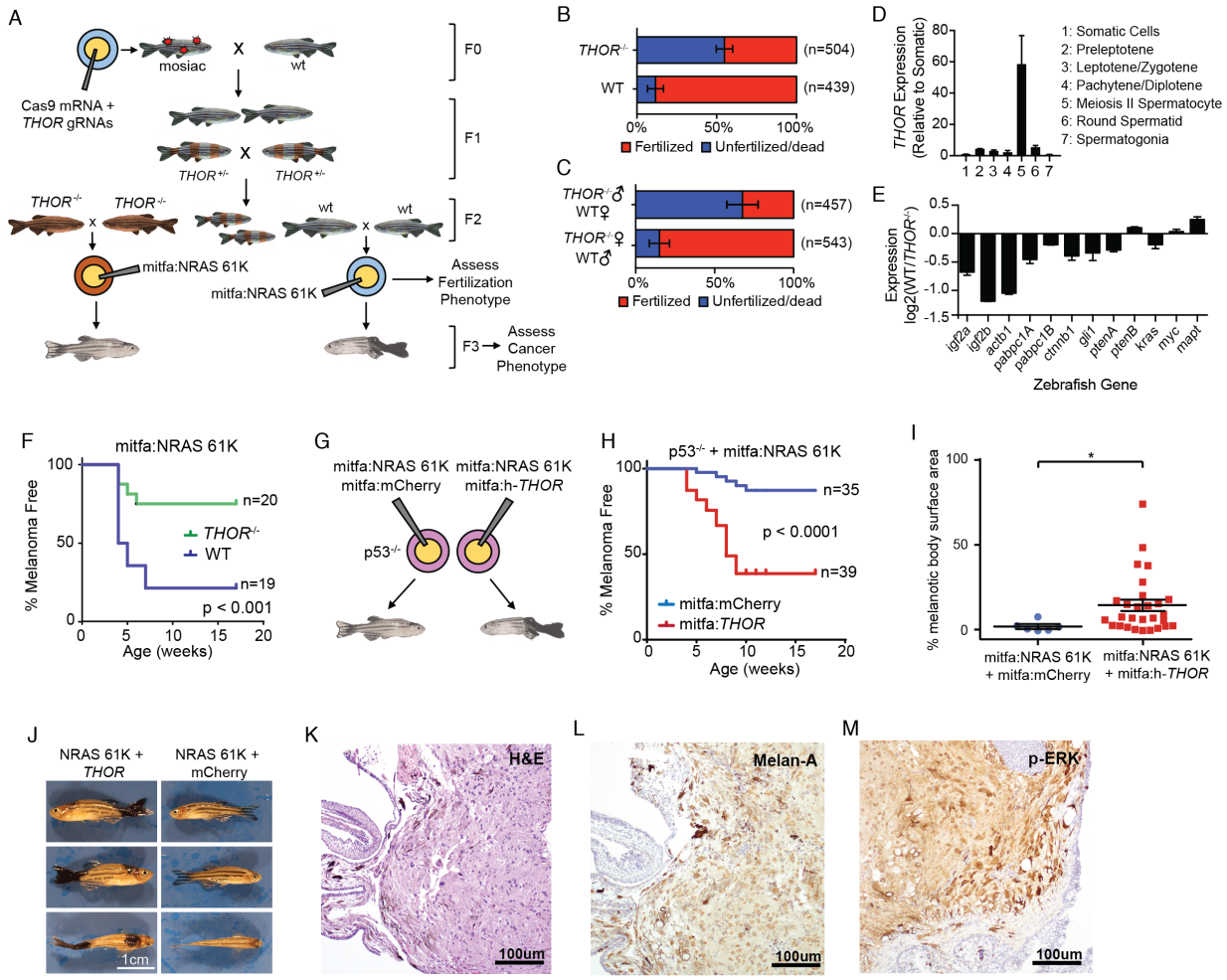


Figure 5.6 ***THOR* regulates melanoma onset in zebrafish**

(a) Schematic depicting creation of *THOR* knockout zebrafish model. **(b)** Fraction of fertilized zebrafish embryos derived from wild-type, or *THOR* knockout zebrafish (-/-) 6 hours following mating. Asterisk (*) indicates P ≤ 0.01 by a $\chi 2$ test. Data show mean ± S.D. from two independent experiments. **(c)** Fraction of fertilized 6hpf zebrafish embryos derived from wild-type female crossed with *THOR* knockout male, or *THOR* knockout female crossed with wild-type male. Asterisk (*) indicates P ≤ 0.01 by a $\chi 2$ test. Data show mean ± S.D. from two independent experiments. **(d)** Expression levels of z*THOR* in testicular somatic cells as well as in six Hoechst profiled subpopulations of testicular germ cells. Data show mean ± S.D. **(e)** Bar plot demonstrates expression of 12 zebrafish orthologs of the canonical IGF2BP1 target genes by qRT-PCR in zebrafish embryos. Expression represented as log2 of the fold change of *THOR*$^{-/-}$ compared to wild-type embryos. Data show mean ± S.D. from one of the two independent experiments. **(f)** Kaplan-Meier curve of melanoma free period for mitfa promoter driven NRAS 61K zebrafish in either *THOR*$^{-/-}$ background or wildtype background. P values were determined using a log-rank test. **(g)** Schematic describing the generation of the h-

THOR overexpression melanoma zebrafish model. **(h)** Kaplan-Meier curve of melanoma free period for p53-/- zebrafish co-injected with either mitfa promoter driven NRAS 61K + mitfa promoter driven human *THOR* or mitfa promoter driven NRAS 61K + mCherry. P values were determined using a log-rank test. **(i)** Percentage of body area covered in melanoma for mCherry and h-*THOR* injected zebrafish also containing mitfa promoter driven NRAS 61K. Asterisk (*) indicates $P \leq 0.01$ obtained by a two-tailed Student's t-test. **(j)** Specimen example of NRAS 61K driven melanomas in zebrafish co-injected with mitfa promoter driven h-*THOR* or mitfa promoter driven mCherry in a $p53^{-/-}$ background. **(k)** H and E staining (100x) of melanoma from an NRAS 61K + *THOR* zebrafish. **(l)** Immunohistochemistry staining (100x) for Melan-A of melanoma from an NRAS 61K + *THOR* zebrafish. Melan-A stained in brown. **(m)** Immunohistochemistry staining for phospho-extracellular signal–regulated kinase (ERK) at 100x magnification of NRAS 61K driven zebrafish melanomas co-injected with mitfa promoter driven *THOR*. p-ERK stained in brown.

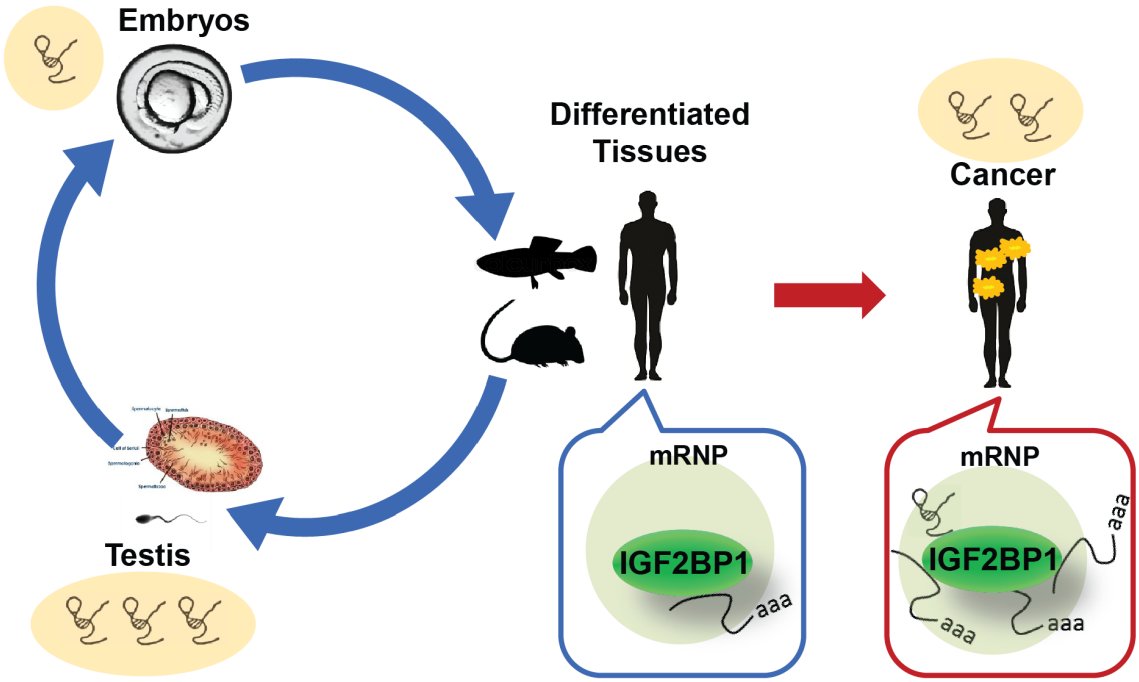


Figure 5.7 Summary Schematic
Summary schematic depicting *THOR* as a conserved cancer-testis lncRNA that functions as a regulator of IGF2BP1 in cancers.

CHAPTER 6

Concluding Remarks and Future Directions

6.1 RNA-seq driven transcript discovery

The work detailed within this dissertation flows from leveraging RNA-seq data for the discovery of lncRNAs to a methods based algorithmic approach to improving the lncRNA discovery workflow. From there, two individual lncRNAs are functionally interrogated, one implicated in ER-positive breast cancers, and the other a highly conserved lncRNA with implications in cancer and fertility.

In Chapter 2 of this dissertation, I presented a large-scale effort of bringing together over 6,500 RNA-seq samples from a myriad of different cancer and tissue types for the purpose of generating the most comprehensive lncRNA annotation to date. The scale of this study is largely responsible for the discovery of so many novel transcripts. Prior studies of lncRNAs using human RNA-seq data had been limited to a small number of samples from a small number of normal human tissue or from cell lines. While these studies did, in fact, identify a myriad of novel transcripts, the tissue and context dependent nature of transcription (of lncRNAs in particular[28]) necessitates inclusion of a wide array and large number of samples in the transcriptome assembly process. If a transcript exists that is exclusively expressed in the amygdala, for example, without RNA-seq data from amygdala tissue this transcript can never be identified using *ab initio* assembly from RNA-seq. Thus, with a RNA-seq compendium orders of

magnitude larger than any previous study, we were able to discover tens-of-thousands of novel transcripts.

As sequencing efforts continue to generate a myriad of RNA-seq data from more and more tissues and diseases, there will continue to be areas in which the transcriptome meta-assembly can be improved. Since the MiTranscriptome study, the GTEX consortium has contributed nearly 10,000 RNA-seq samples from normal human tissue to the scientific community, and other consortia continue to make samples available. Thus, in order to continue to expand our understanding of the human transcriptome, being sensitive to those transcripts that are tissue and/or disease specific, continual updating of the meta-assembly with newer samples will be necessary.

6.2 Transcriptome meta-assembly

In Chapter 3 of this dissertation, I detail a novel algorithm I developed for the purpose of generating high-fidelity transcriptome meta-assemblies from large quantities of individual transcriptome assemblies. The tool, TACO, was shown to dramatically outperform its competitors, especially in its ability to accurately predict splicing patterns and its ability to identify accurate 3′ and 5′ ends of transcripts. While the MiTranscriptome has become a widely accessed and cited resource in the community, it was generated prior to the development of TACO. At that time, the algorithm utilized to generate the meta-assembly was victim to noise aggregation and inaccurate 3′ and 5′ ends. Thus, there is now a compelling need to utilize the expanded RNA-seq compendium, and leverage TACO to produce an updated and more robust version of the MiTranscriptome. With access to over 3 times as many RNA-seq samples now, and with a transcriptome meta-assembly tool that has been shown to generate robust and accurate transcript structure predictions from large numbers of samples we are well

poised to expand our pool of known transcripts, and do so in a clean way that is amenable to facile downstream analyses.

It is important to note that the myriad of available RNA-sequencing data is in the form of short-read RNA-seq. Thus, any meta-assembly approach will be bottlenecked by the accuracy of the tools used to generate the individual transcriptome assemblies that are used for meta-assembly. The algorithms outlined in TACO could be dramatically improve the performance of individual sample transcriptome assemblers, and thus should be considered to expand the capacity of current transcriptome assembly tools.

6.3 Functional interrogation of lncRNAs

In Chapters 4 and 5 of this dissertation, I outline two studies that functionally interrogate a specific lncRNA that was discovered in the MiTranscriptome. The MiTranscriptome contains over 50,000 lncRNAs. This number is simply too great to functionally interrogate individually. However, without an affordable and viable approach to do high-throughput functional interrogation, the field is being advanced mostly through anecdotal studies performing in depth investigation of single lncRNAs. The advent of CRISPR-screening technology provides an opportunity for large scale investigation of lncRNAs and will prove an interesting avenue moving forward[241]. Screening tens-of-thousands of novel transcripts, however, is currently cost prohibitive.

Through dissemination of our MiTranscriptome findings to the scientific community via the mitranscriptome.org web portal, it is my hope that the scientific community can come together to chip away at the mountain of potentially functional RNAs in the human genome. Regardless of whether large-scale screens become easily available, the in depth studies, such as those performed in Chapter 4 and 5 will be necessary to fully

understand the role of these lncRNAs in cells. As our understanding of the function of lncRNAs continues to expand, we will be able to begin to identify functional classes of lncRNAs. Despite belief that many lncRNAs are non-functional and transcriptional noise, there is unequivocal evidence that many do perform function in cells. Thus, we are tasked with teasing this apart in order to continue to further our understanding of the mechanistic underpinnings of cell and molecular biology.

APPENDICES

APPENDIX A
Supplemental Data for Chapter 2

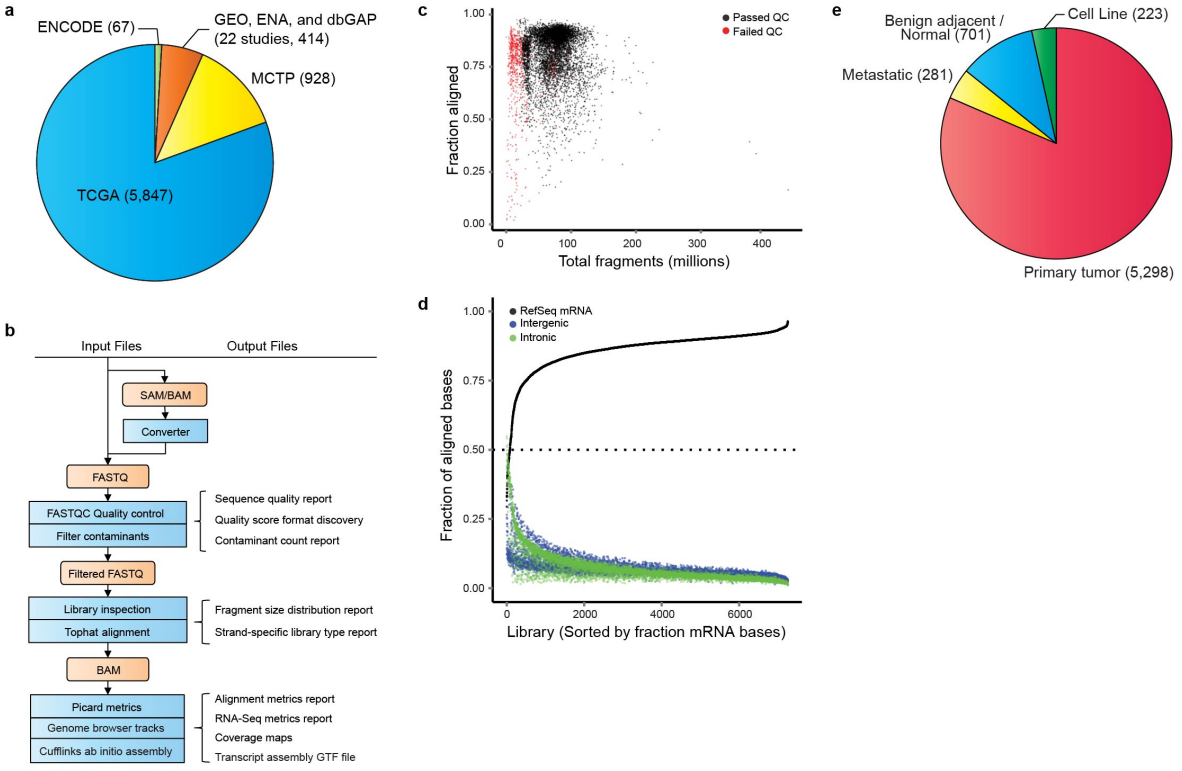


Figure A.1 Curation and processing of samples in the MiTranscriptome compendium

A) Pie chart showing the number of studies curated from TCGA, ENCODE, MCTP, and other publicly available datasets. B) Workflow for bioinformatics processing of individual RNA-Seq libraries. Datasets downloaded as BAM files were first converted to FASTQ format. Quality assessment of FASTQ files was performed using FASTQC. Reads mapping to mitochrondria, ribosomal RNA, poly-A sequence, poly-C sequence, or phiX virus (a spiked-in control) were filtered. Fragment length distribution and orientation were determined by mapping a subset of the input reads to a set of large human exons (>500bp). Reads were aligned using TopHat (v2.0.6) with Bowtie2 (v2.1.0).

Gene fusion calling was performed using Tophat-Fusion (v2.0.6) with bowtie1 (v0.12.9). Read alignment metrics were computed using Picard Tools, and genome track information was generated using BEDTools and UCSC binary utilities. Finally, ab initio transcriptome assembly was performed using Cufflinks version 2.0.2. C) Scatter plot showing total fragments (x axis) and the fraction of aligned fragments (y axis) for each RNA-Seq library. Coarse quality control filters used to remove libraries with fewer than 20 million total fragments or 20 million alignments (red point). D) Dot plot showing for each library the fraction of aligned bases corresponding to Refseq mRNA (black points), intronic regions (green points), or intergenic regions (blue points) on the y axis. Libraries with fewer than 50% of aligned bases corresponding to RefSeq mRNA were filtered (dotted line). E) Pie chart showing numbers of primary tumors (red), metastatic tumors (yellow), benign adjacent tissues or tissues from healthy individuals (blue), or cell lines (green) for 6,503 RNA-Seq libraries that passed coarse quality control filters.

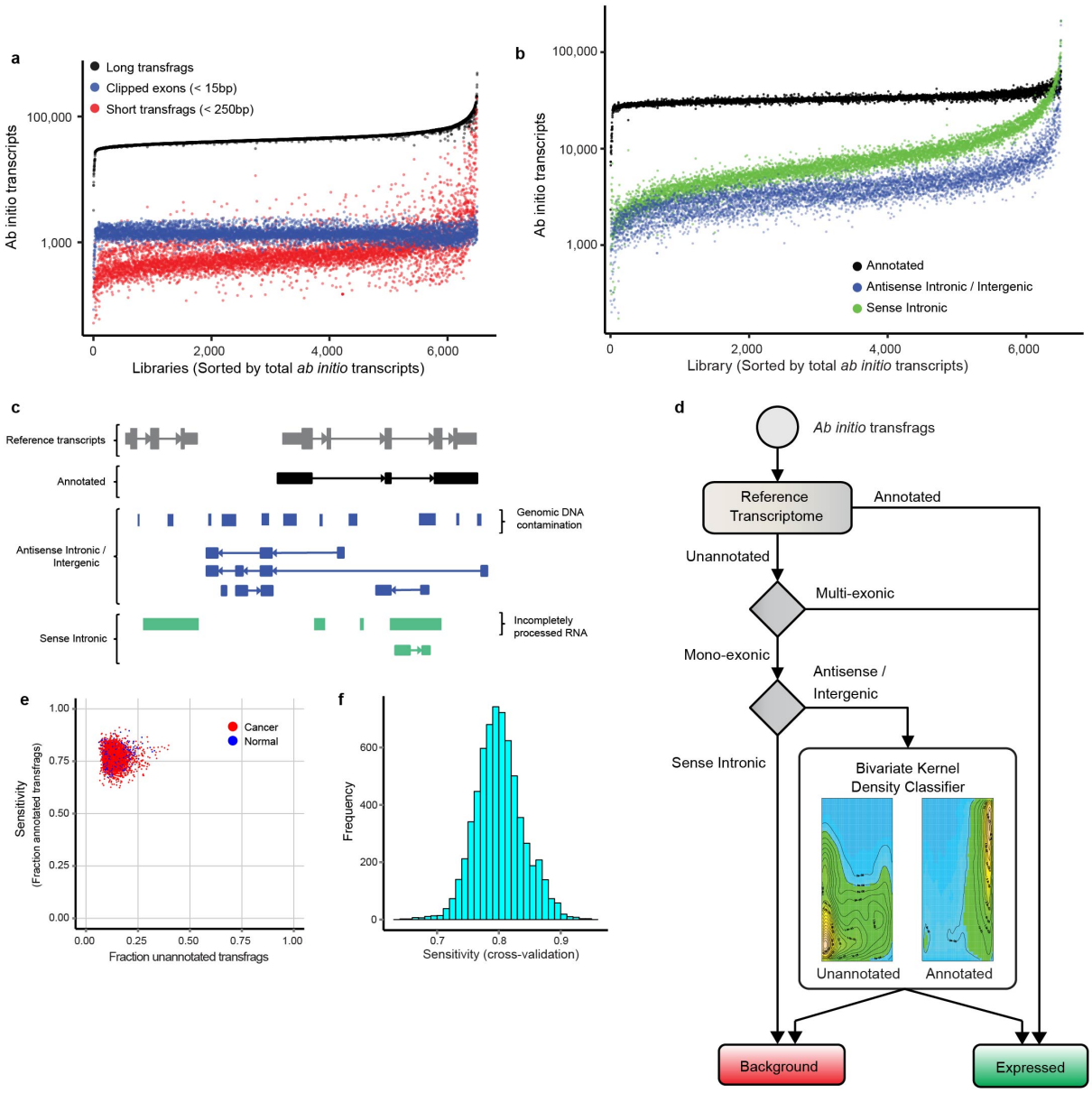


Figure A.2 Transfrag filtering.

A) Dot plot shows the numbers of short transfrags (red), short clipped exons (blue), and long transfrags (black) for each library. B) Dot plot shows the numbers of unannotated intergenic or antisense transfrags (blue), sense intronic transfrags (green), and annotated transfrags (black) for each library. C) Example transcript models illustrating categories of ab initio transcripts and sources of background noise. Annotated transfrags (black) overlap reference transcripts on the same strand. Unannotated antisense intronic or intergenic transfrags (blue) may be confounded by genomic DNA contamination. Unannotated sense intronic transfrags (green) may be confounded by both genomic DNA and incompletely processed RNA contamination. D) Decision tree depicting transfrag filtering steps for a single library. First, transfrags were labeled

'Annotated' or 'Unannotated' based on overlap with a reference transcriptome catalog. Annotated transfrags and unannotated multi-exonic transfrags were considered expressed. Unannotated mono-exonic transfrags within introns in the sense orientation of an overlapping transcript were discarded as incompletely processed RNA artifacts. Unannotated antisense or intergenic mono-exonic transfrags were subjected to a bivariate kernel density classification method to discriminate recurrent, reliable transcription from genomic DNA contamination artifacts. Transfrags predicted as 'expressed' were incorporated into meta-assemblies. E) Scatter plot comparing the sensitivity of the mono-exonic transfrag classifier for correctly detecting annotated transcripts (y axis) and the fraction of unannotated transfrags predicted to be expressed (x axis). F) Histogram demonstrating sensitivity for correctly detecting annotated test transcripts held out of the classifier training process.

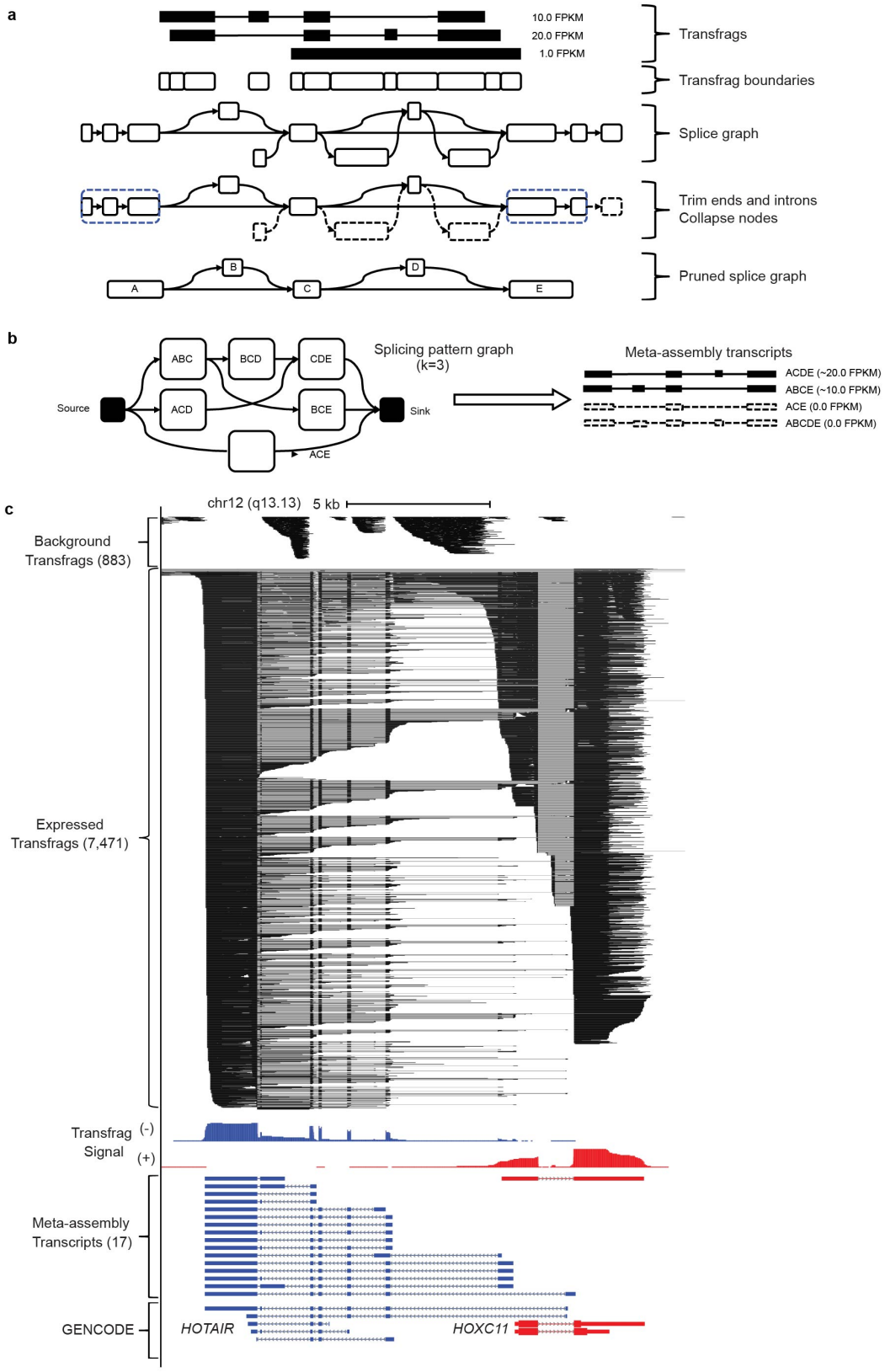


Figure A.3 Meta-assembly.

A) Schematic of transcriptome meta-assembly algorithm using a simplified example with three transfrags transcribed from left-to-right. The input to the meta-assembly is a list of weighted transfrags (in this case, the weights correspond to FPKM expression values). First, a splice graph is constructed using the transfrag exon boundaries. The splice graph is a directed acyclic graph (DAG) with nodes (rounded rectangular boxes) representing contiguously transcribed genomic bases and edges (arrows) corresponding to possible alternative splicing and promoter usage. The splice graph is then trimmed to remove lowly expressed starting/ending nodes, and adjacent nodes with a degree of one are collapsed. B) The pruned splice graph from panel a is subjected to meta-assembly. To encapsulate splicing pattern information present in the original transfrags the pruned splice graph is converted into a splicing pattern graph. A splicing pattern graph is a *De Bruijn* graph where each node represents a group of *k* consecutive connected nodes from the splice graph (in this example *k*=3), and edges connect adjacent node groups. In real cases *k* is automatically chosen to optimize the number of nodes in the splicing pattern graph. Finally, the splicing pattern graph is repeatedly traversed using a greedy dynamic programming algorithm to determine the set of most highly abundant isoforms from the graph. In this example, isoforms ACDE and ABCE recapitulate input transfrags with nearly identical FPKM values, and invalid isoform combinations ACE and ABCDE are discarded. C) Genome view showing an example of the meta-assembly procedure for breast cohort transfrags in a chromosome 12q13.3 locus containing the lncRNA *HOTAIR* and the protein-coding gene *HOXC11* on opposite strands (chr12:54,349,995-54,377,376, hg19). 883 transfrags were considered background noise and not used for meta-assembly. A dense cluster of 7,471 expressed transfrags from 1,076 breast RNA-Seq libraries was used as input. The aggregated transfrag signal on the positive (+) and negative (-) strands is shown below. Meta-assembly produced 17 transcripts from the transfrags, including transcripts that matched GENCODE *HOTAIR* and *HOXC11* splicing patterns as well as *HOTAIR* transcripts with unannotated splice sites.

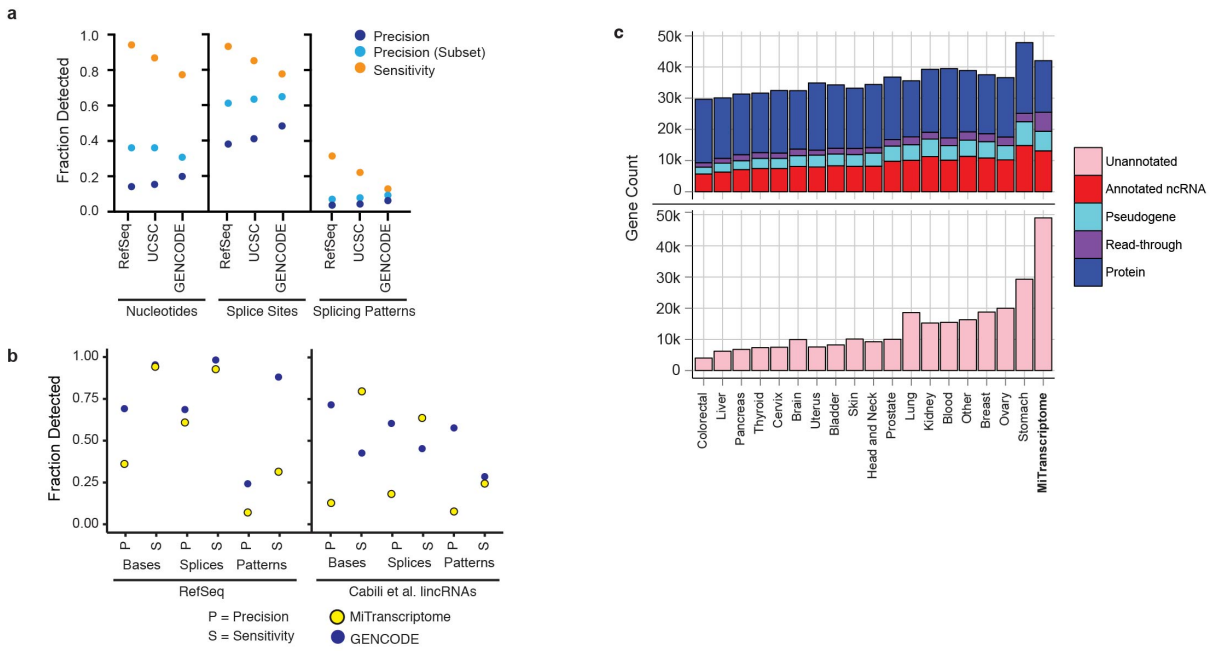


Figure A.4 Characterization of unannotated transcripts.
A) Dot plots depicting comparison of MiTranscriptome with reference transcripts from RefSeq, UCSC, or GENCODE. Precision (blue), precision for the subset of transcripts overlapping annotated transcripts (light blue), and sensitivity (orange) are plotted for each comparison. B) Dot plots comparing the basewise, splice site, and splicing pattern precision and sensitivity of MiTranscriptome and GENCODE using RefSeq (left) or Cabili et al. lncRNAs (right). C) Bar plots comparing numbers of unannotated versus different classes of annotated transcripts for each of the 18 cohorts. (top) Stacked bar plot showing annotated ncRNAs (red), pseudogenes (cyan), read-throughs (purple), and protein-coding genes (blue). (bottom) Bar plot showing unannotated transcripts (pink).

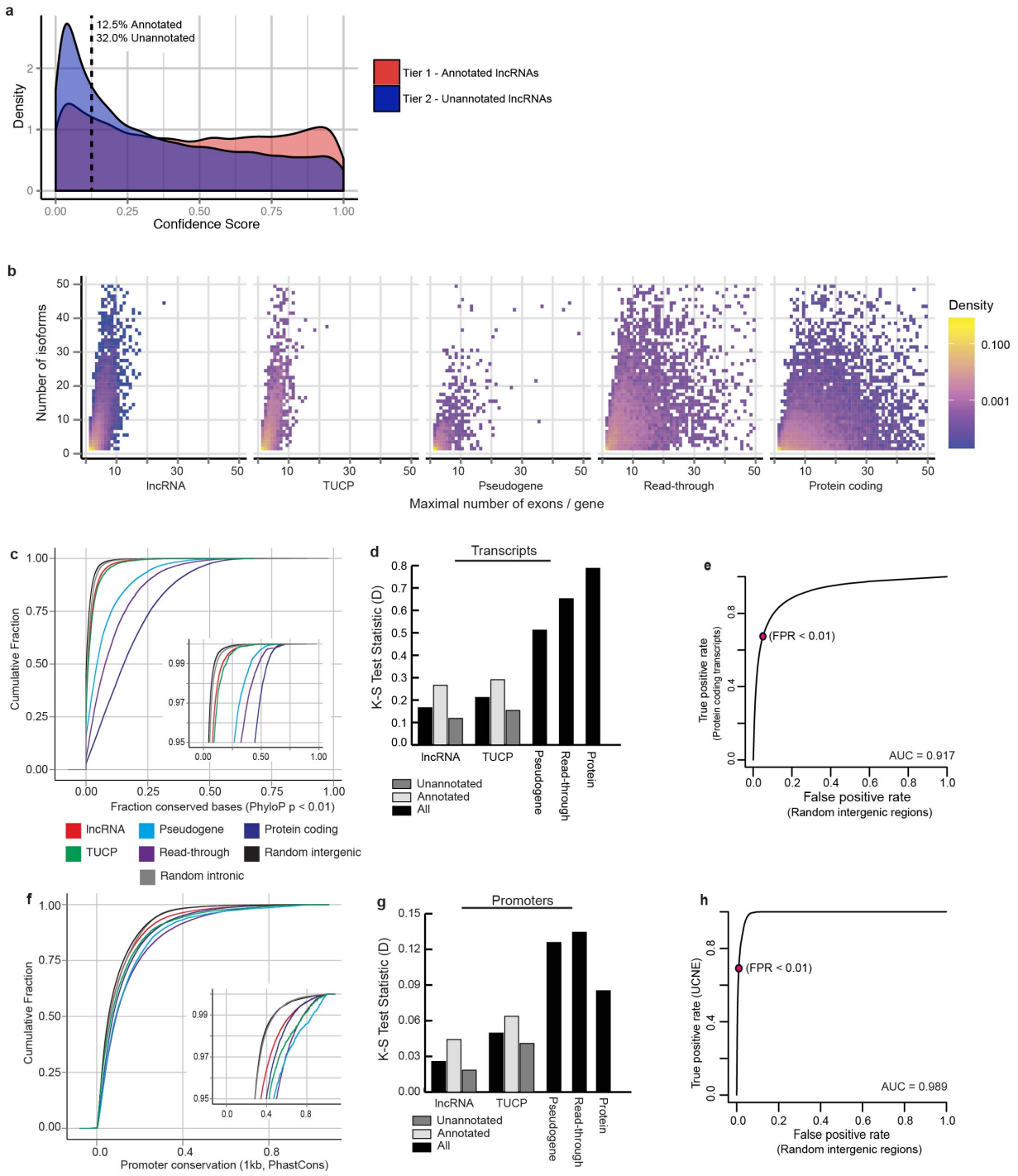


Figure A.5 MiTranscriptome characterization.

A) Density histogram depicting the confidence scores for annotated and unannotated lncRNAs. B) Comparison of the relationship of maximum number of exons per gene to the number of isoforms per gene. LncRNAs tend to have fewer exons than protein-coding genes, but they have complex splicing patterns that yield multiple transcript isoforms. C) Cumulative distribution plot for basewise conservation fraction of (blue)

proteins, (purple) read-throughs, (cyan) pseudogenes, (green) TUCPs, (red) lncRNAs. Random intergenic (black) and intronic (grey) regions are plotted as controls. Inset plot highlights upper 5th percentile of distribution. D) Bar plot showing KS test statistics for classes of transcripts versus random intergenic controls. E) ROC curve for predicting conservation of protein-coding genes versus random intergenic controls. The cutoff (pink point) chosen for calling highly conserved transcripts is plotted. F) Cumulative distribution plot for promoter conservation (legend shared with C). Inset plot highlights upper 5th percentile of distribution. G) Bar plot showing KS tests for promoter conservation versus random intergenic regions. H) ROC curve for predicting ultra-conserved non-coding elements versus random intergenic regions. Cutoff (pink point) chosen for nominating ultraconserved lncRNAs is plotted.

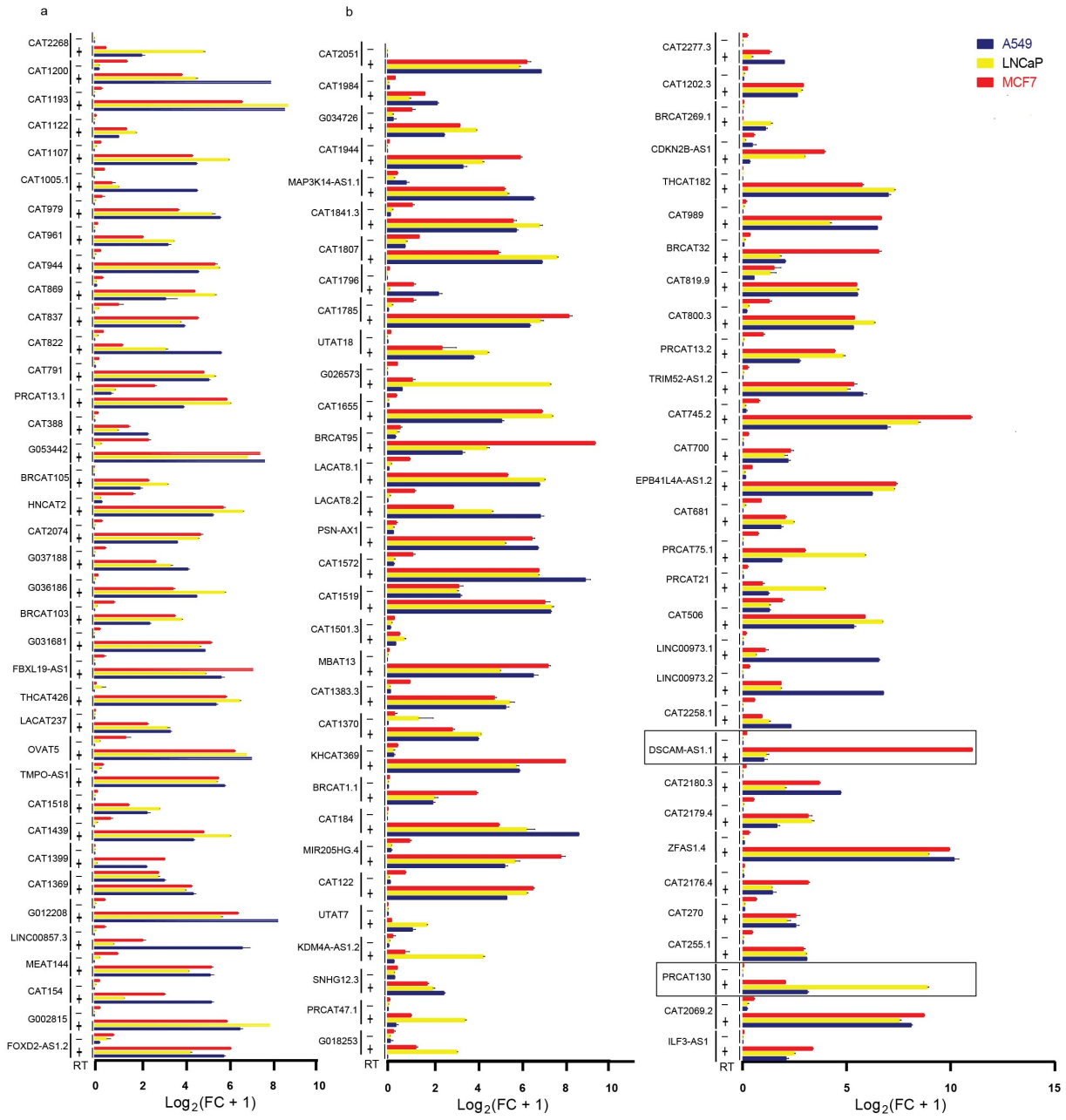


Figure A.6 Validation of lncRNA transcripts.

One hundred lncRNA transcripts were validated by qRT-PCR across A549, LNCaP, and MCF7 cell lines using a +/- RT approach. Ct values were first normalized to housekeeping genes (*CHMP2A, EMC7, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3*) and then to the median value of all samples using the delta-delta Ct method. Here, data is plotted as a logirithmic of fold change over median with standard error of the mean. Validation was performed on **a**, thirty-eight mono-exonic transcripts and **b**, sixty-two multi-exonic transcripts with associated MiTranscriptome name or Gene ID. Boxed

transcripts are two representative examples of lncRNAs with lineage/cancer specficity in breast, or prostate according to SSEA analysis whose cell line expression profile (by qRT-PCR) reflects what is expected from tissue analysis.

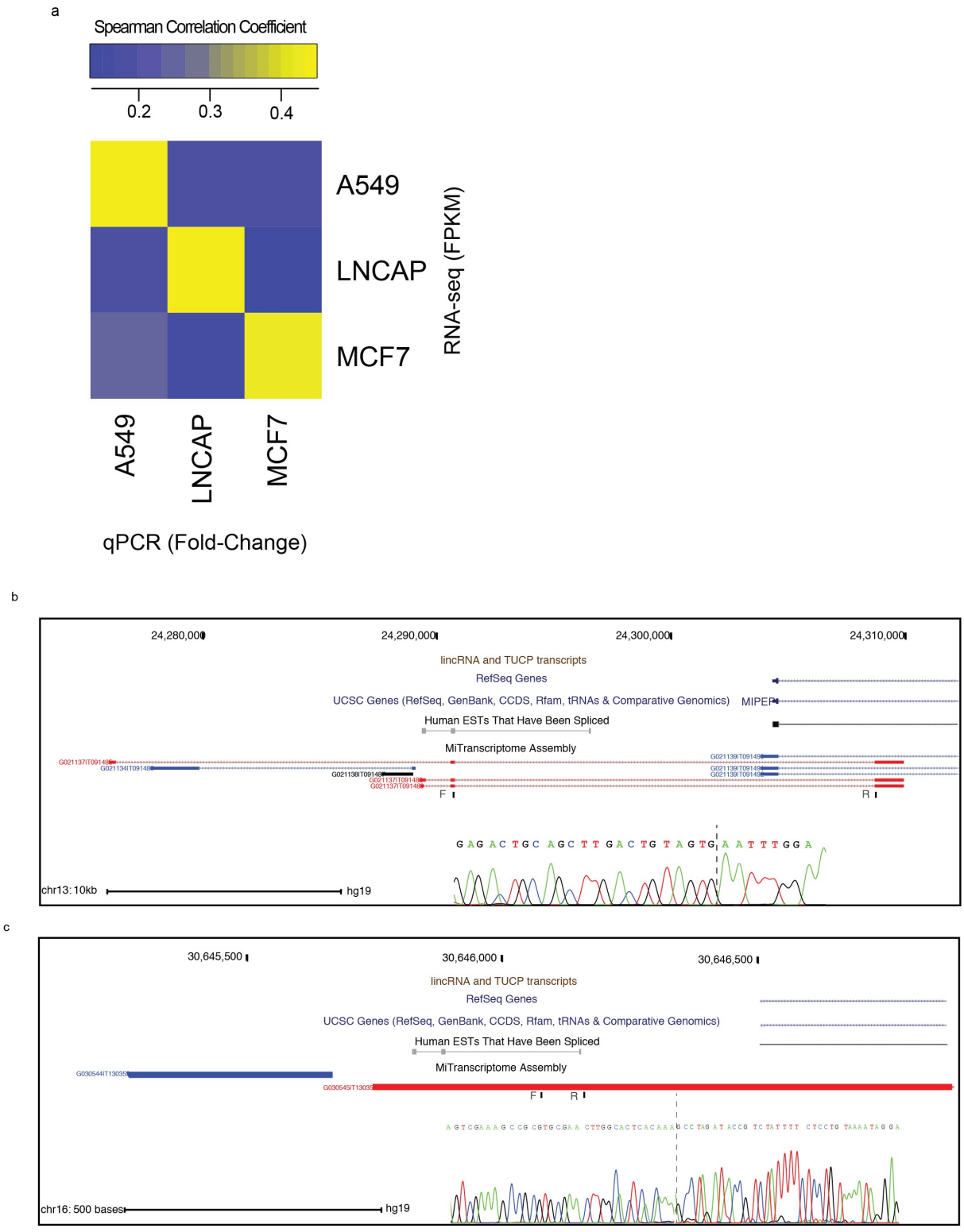


Figure A.7 Validation of lncRNA transcripts, continued.

A) Heatmap representation of the correlation between qPCR (fold change over median) with RNA-seq (FPKM) of 100 selected transcripts in cell lines A549, LNCaP, and MCF7. B) and C) Representative example of two of twenty previously unannotated lncRNA transcripts that were analyzed by Sanger sequencing to ensure primer specificity with their associated chromatograms. As seen in the UCSC Genome Browser View a B) multi-exonic lncRNA (Gene ID: G021137) and C) mono-exonic lncRNA (Gene ID: G030545).

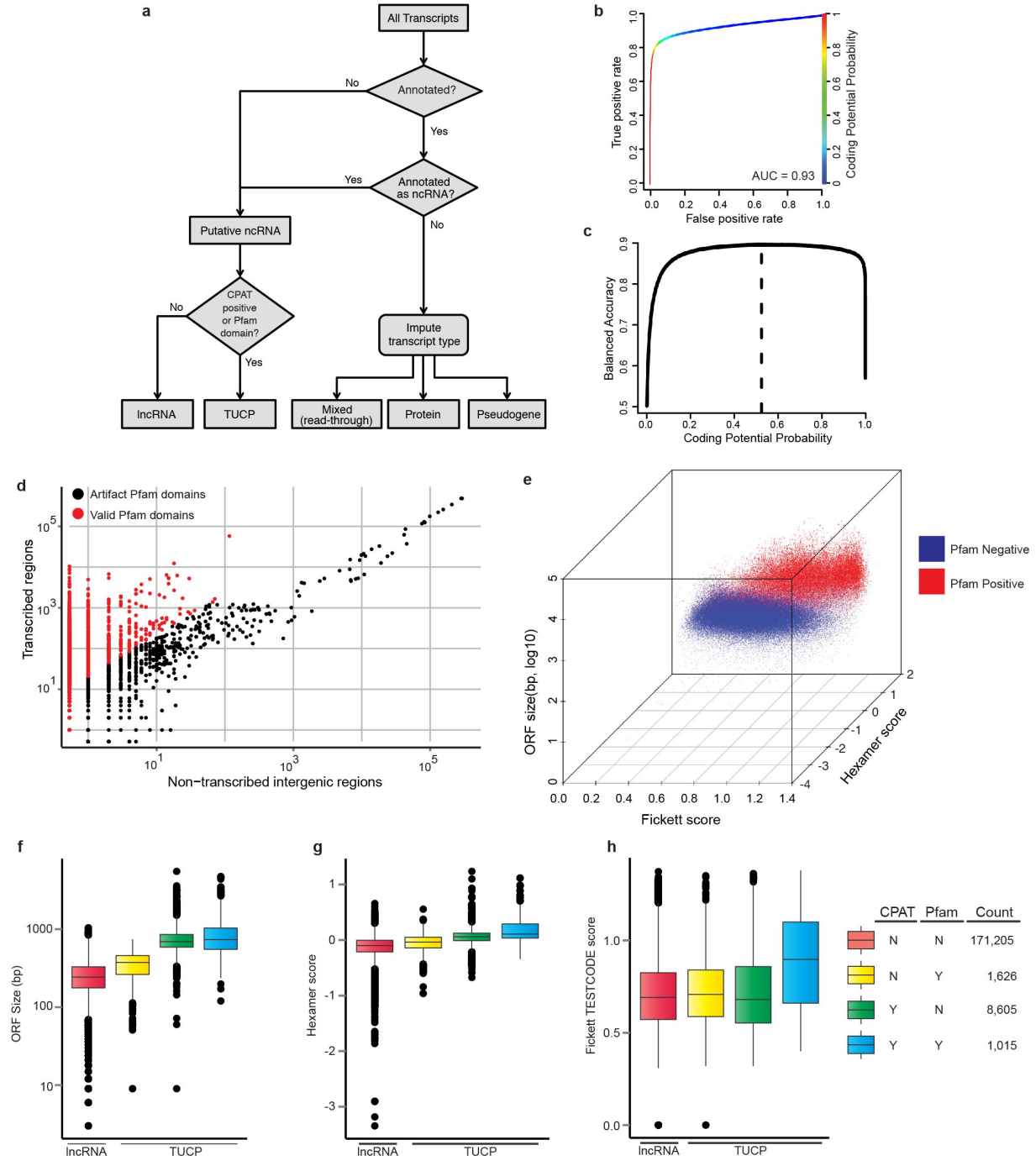


Figure A.8 Classification of transcripts of unknown coding potential.
A) Decision tree showing categorization of *ab initio* transcripts. Unannotated transcripts and annotated ncRNAs were classified as either lncRNA or TUCP. Transcript categories for protein-coding genes, pseudogenes, and read-throughs were imputed from overlapping reference annotations. B) ROC curve comparing false positive rate (x axis) with true positive rate (y axis) for CPAT coding potential predictions of ncRNAs versus

protein-coding genes. C) Curve comparing probability cutoff (*x* axis) with balanced accuracy (*y* axis). Dotted line show cutoff used to call TUCP transcripts. D) Scatter plot comparing frequencies of Pfam domain occurrences in non-transcribed intergenic space versus transcribed regions. Points in red were considered valid Pfam domain hits and points in black were considered artifacts. E) Three-dimensional scatter plot comparing Fickett score (*x* axis), ORF size (*y* axis), and Hexamer score (*z* axis) for all transcripts. Red points contain valid Pfam domains and blue points do not. F,G,H) Boxplots comparing ORF size (F), Hexamer score (G), and Fickett score (H) for lncRNAs (red), TUCPs predicted by Pfam only (yellow), TUCPs predicted by CPAT (green), and TUCPs predicted by both Pfam and CPAT (blue).

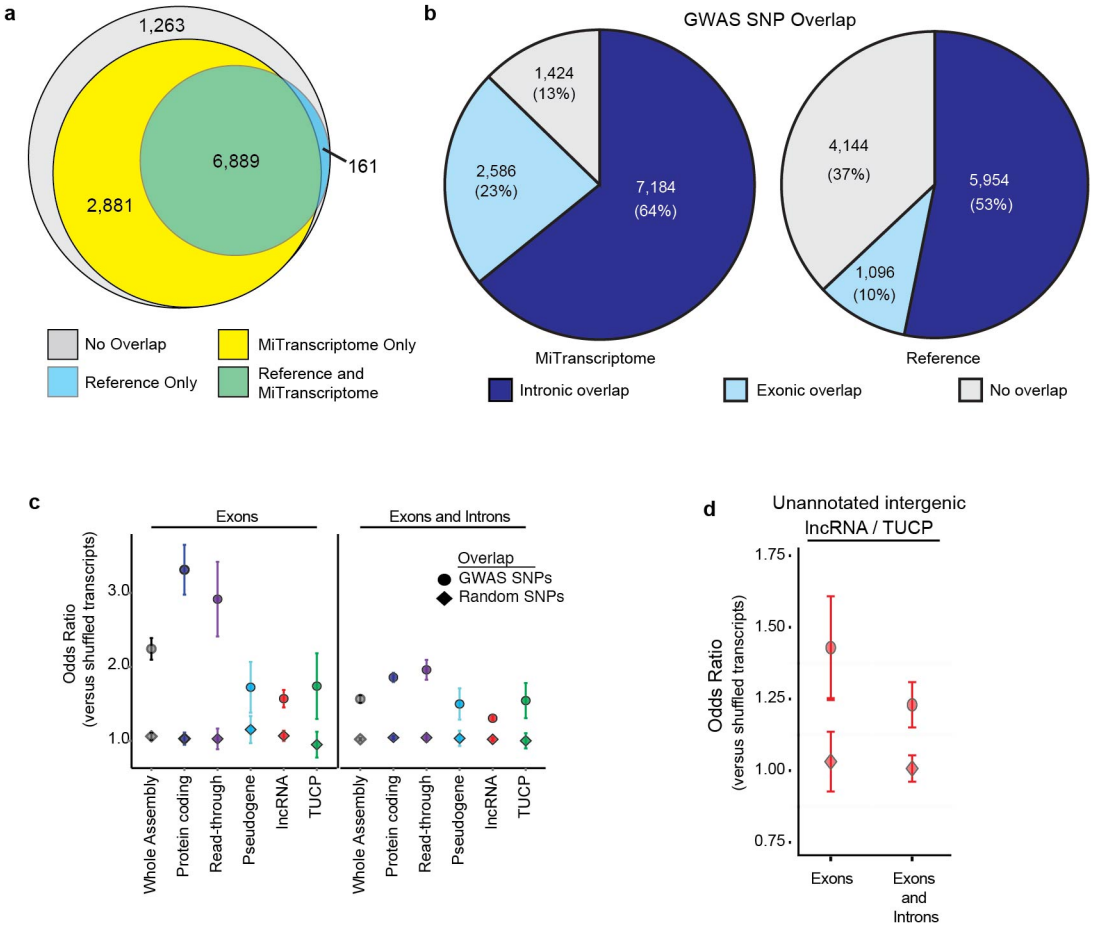


Figure A.9 Enrichment of MiTranscriptome assembly for disease-associated regions.

A) Venn diagram comparing coverage of disease- or trait-associated genomic regions (i.e. GWAS SNPs) for the MiTranscriptome assembly (yellow) in comparison to reference catalogs (blue), with the area of intersection is shaded green. B) Pie charts comparing distributions of intronic and exonic GWAS SNP coverage of the MiTranscriptome assembly (left) and reference catalogs (right). C) Dot plot displaying enrichment of GWAS SNPs versus random SNPs for different transcript categories. Enrichment odds ratios (transcripts-SNP overlaps versus shuffled transcript-SNP overlaps) are plotted on the *y* axis. Points indicate the mean of 100 permutations for tests of enrichment with GWAS SNPs (circle) or random SNPs (diamond), and error bars depict +/- 2 standard deviations of the distribution of odds ratios. Both exonic and whole-transcript enrichment is reported. D) Dot plot showing enrichment of GWAS SNPs (circle) versus random SNPs (diamond) for novel intergenic lncRNAs and TUCPs. Enrichment odds ratios (transcripts-SNP overlaps versus shuffled transcript-SNP

overlaps) are plotted on the *y* axis. Points indicate the mean of 100 shuffles for comparisons with GWAS SNPs (circle) or random SNPs (diamond), and error bars depict +/- 2 standard deviations of the distribution of odds ratios. Both exonic and whole-transcript enrichment is reported.

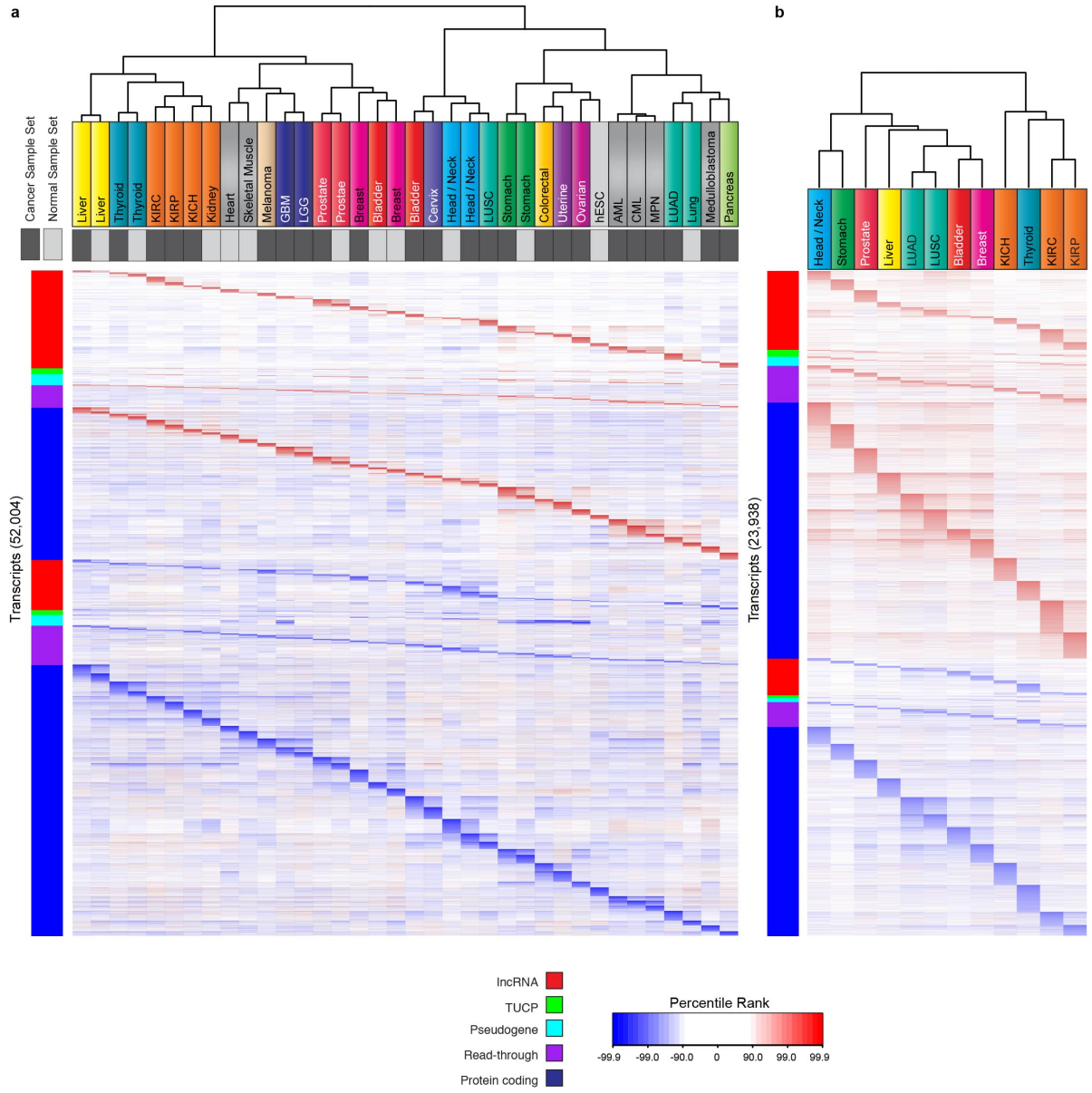


Figure A.10 Discovery of lineage-associated and cancer-associated transcripts. A) Heatmap of lineage-specific transcripts (LATs) nominated by SSEA. Each column represents a sample set from one of 25 cancer (dark grey) and 13 normal (light grey) lineages and each row represents an individual transcript. Colored labels above columns reflect organ system cohorts used in assembly. Row side colors correspond to lncRNAs (red), TUCPs (green), pseudogenes (cyan), read-throughs (purple), and protein coding transcripts (blue). All transcripts were statistically significant (FDR < $1e^{-7}$) and ranked in the top 1% most positively or negatively enriched transcripts within at least one sample set. The heatmap color spectrum corresponds to percentile ranks, with under-expressed transcripts colored blue and over-expressed transcripts colored red.

The column dendrogram shows unsupervised hierarchical clustering of sample sets. B) Heatmap of cancer-specific transcripts (CATs) nominated by SSEA. Columns represent 12 cancer types, and colored column labels reflect organ system cohorts used in assembly. All transcripts were statistically significant (FDR < $1e^{-3}$) and ranked in the top 1% most positively or negatively enriched transcripts within at least one sample set. Column dendrogram shows unsupervised clustering results. Row side color and heatmap color schemes are identical to (A).

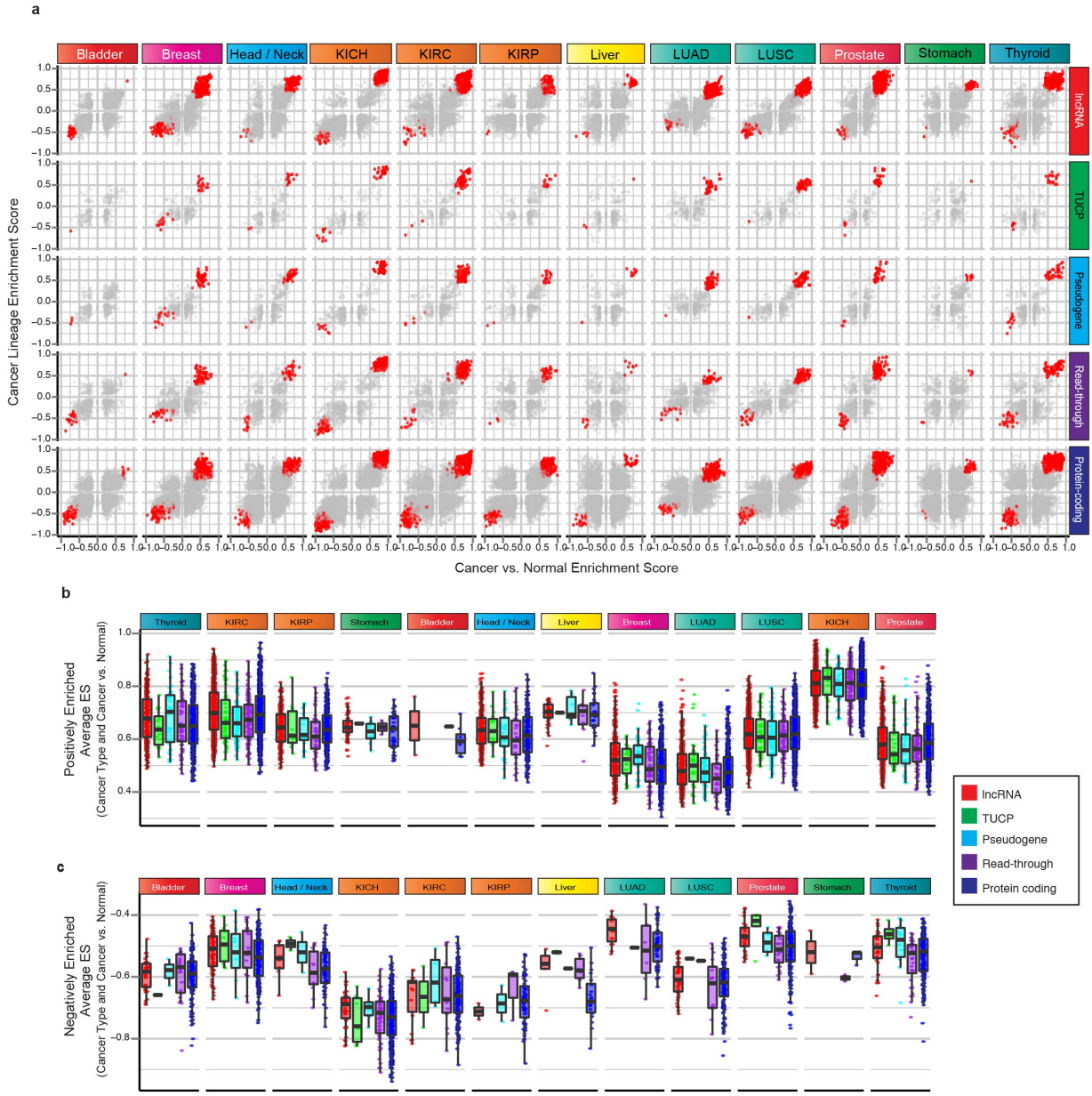


Figure A.11 Lineage-specific and cancer-specific transcripts.
A) Scatter plot grid showing lineage-specific and cancer-specific transcripts (CLATs) nominated by SSEA. A row of scatter plots for transcript category is plotted across 12 cancer types. Each plot shows Cancer vs. Normal enrichment score (x axis) and the Cancer Lineage enrichment score (y axis). Red points indicate CLATs within the respective cancer types, and grey points indicate CLATs for other cancer types. B and C) Boxplots comparing the performance of (B) positively enriched CLATs and (C) negatively enriched CLATs for each transcript category across 12 cancer types. The average of the lineage and cancer versus normal ES is plotted on the y axis.

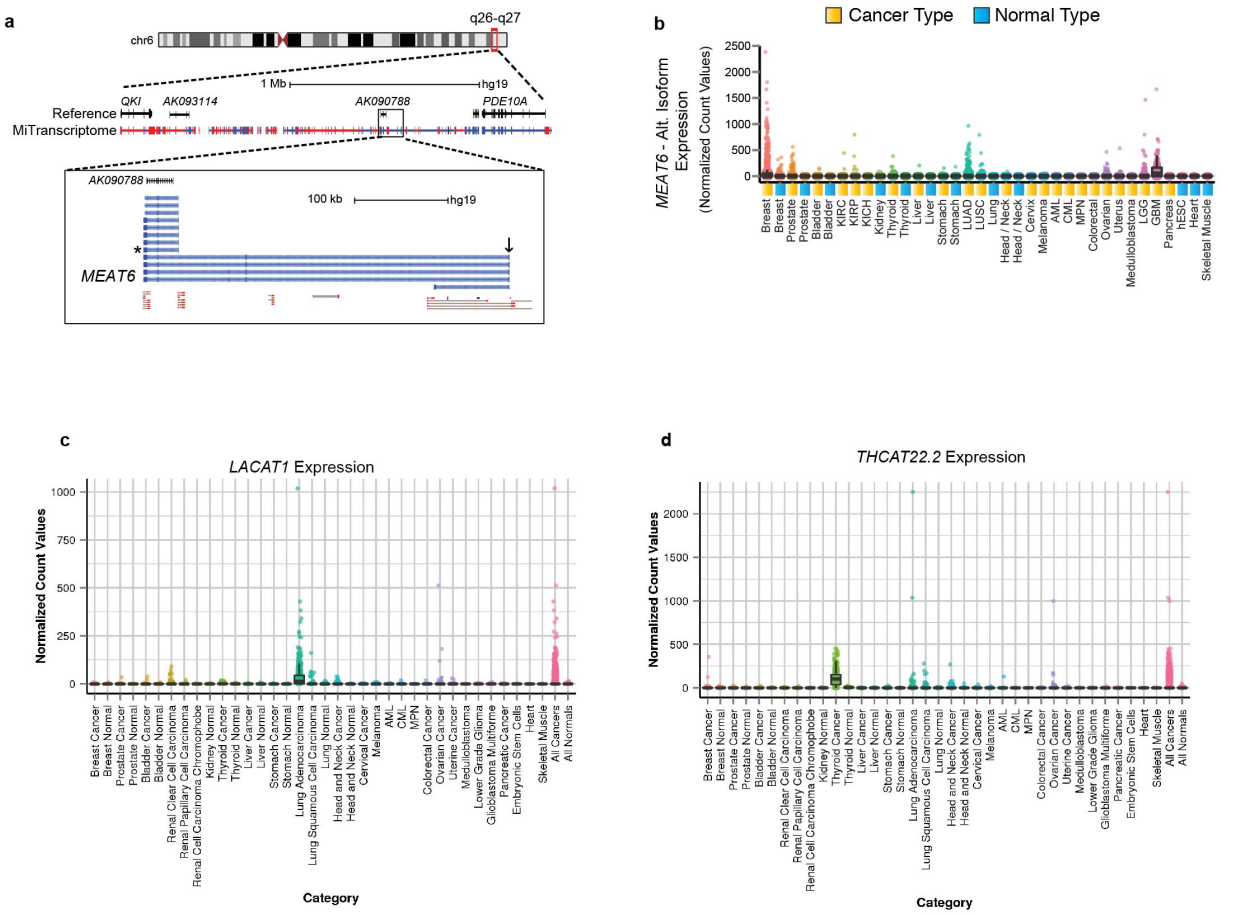


Figure A.12 Examples of cancer and/or lineage associated transcripts.
A) Genomic view of chromosome 6q26-q27 locus. Protein coding genes *QKI* and *PDE10A* flank an intergenic region with two annotated lncRNAs, *AK093114* and *AK090788*. MiTranscriptome transcripts are shown in a dense view populating this intergenic space. Most zoomed view (bottom) depicts *MEAT6*, a melanoma associated lncRNA. *AK090788* overlaps a portion of *MEAT6*, but the full *MEAT6* transcript uses an alternate start site (black arrow). B) Expression data for *MEAT6* (demarcated by asterisk in A). This isoform variant does not use the alternate start site used by *MEAT6*, and closely resembles *AK090788*. Expression profile for cancer and lineage associated transcripts across all MiTranscriptome tissue cohorts are shown for (C) lung adenocarcinoma and (D) thyroid cancer.

Table A.1 RNA-seq datasets.

Supplementary Table 1: RNA-Seq Datasets

Study ID	Sources	Study Name(s)	Study URL(s)	Total Samples Processed	Samples Used For Assembly
encode	ENCODE	ENCODE RNA-Seq (CSHL and Caltech)	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/	67	67
erp000546	ENA	Illumina Bodymap 2.0 Transcriptome	http://www.ebi.ac.uk/ena/data/view/ERP000546	35	35
erp000550	ENA	Complete transcriptomic landscape of prostate cancer in the Chinese population using RNA-seq	http://www.ebi.ac.uk/ena/data/view/ERP000550	23	22
erp001058	ENA	The transcriptional landscape and mutational profile of lung adenocarcinoma	http://www.ebi.ac.uk/ena/data/view/ERP001058	163	160
gse20156	GEO	RNA-Seq of melanoma short-term cultures and cell lines	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20156	14	0
gse22260	GEO	Comparative transcriptomic analysis of prostate cancer and matched normal tissue using RNA-seq	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22260	30	0
gse22765	GEO	SplicePL: de novo splice junction detection from paired-end RNA-seq reads	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22765	1	0
gse23776	GEO	Alternative expression analysis by RNA sequencing (comparison of 5-FU sensitive and resistant colorectal cancer cell lines)	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23776	4	0
gse26109	GEO	A human transcriptome array for high-throughput clinical studies (with a comparison to RNA-seq technology) [RNA-Seq]	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26109	7	6
gse29006	GEO	mRNA-seq of Human Airway Epithelial Cells	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29006	4	3
gse30352	GEO	The evolution of gene expression levels in mammalian organs	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30352	18	2
gse30554	GEO	Integrative Annotation of Human Large Intergenic Non-Coding RNAs Reveals Global Properties and Specific Subclasses	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30554	9	7
gse31999	GEO	RNA-sequencing of left and right atrial	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31999	8	8
gse32424	GEO	Identification of a Novel Angiogenesis and Tumor Suppressor Gene Rab25 in Esophageal Squamous Cell Carcinoma	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32424	12	0
gse32625	GEO	RNA-Seq of human neurons derived from iPS cells Reveals candidate long non-coding RNAs Involved in neurogenesis and neuropsychiatric disorders.	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32625	2	0
gse33092	GEO	Oncogenic BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration [HT-seq]	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33092	4	2
gse33328	GEO	Transcriptomic profiling of a glioblastoma multiforme patient with matched control brain tissue	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33328	2	0
gse34914	GEO	Deep Sequence Analysis of non-small cell lung cancer: Integrated analysis of gene expression, alternative splicing, and single nucleotide variations in lung adenocarcinomas with and without oncogenic KRAS mutations	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34914	16	6
gse35178	GEO	Transcriptome of normal, paracancerouse and cancerous bladder tissues based on RNA-Seq method	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35178	3	3
gse41009	GEO	Divergent transcription of lncRNA/mRNA gene pairs in embryonic stem cells	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41009	4	4
gse42146	GEO	RNA-Seq analysis of atypical chronic myeloid leukemia samples	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42146	13	9
idea	Illumina	Breast Cancer Cell Line RNA-Seq	http://www.illumina.com/science/data_library.ilmn	16	7
mctp	dbGAP and previously unpublished data	Molecular profling of cancer, MAST Kinase and Notch Gene Families in Breast Cancer, NAB2-STAT6 Gene Fusions in Solitary Fibrous Tumor by Integrative Sequencing, Expressed Pseudogenes in the Transcriptional Landscape of Human Cancers, and previously unpublished data	http://www.ncbi.nlm.nih.gov/gap	928	513
phs000178	dbGAP	The Cancer Genome Atlas	https://cghub.ucsc.edu/	5847	5648
phs000310	dbGAP	Discovery of Non-ETS Gene Fusions in Human Prostate Cancer using Next Generation RNA Sequencing	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000310.v1.p1	26	19

Table A.2 Software packages.

Software	Purpose	Version(s)	Website
Modules	Linux environment management	3.2.8	modules.sourceforge.net
Python	Scripting language	EPD 2.7.2	www.enthought.com/products/epd.php
Java	Runtime environment	1.7.0_1	www.oracle.com/technetwork/java/javase/downloads/index.html
R	Statistical language	3.0.1	www.r-project.org
bx-python	Bioinformatics utilities	0.7.0	bitbucket.org/james_taylor/bx-python
pysam	SAM/BAM processing	0.6	code.google.com/p/pysam
FastQC	Diagnostics	0.10.1	www.bioinformatics.babraham.ac.uk/projects/fastqc
Picard	Alignment quality metrics	1.8	picard.sourceforge.net
samtools	SAM utilities	0.1.18	samtools.sourceforge.net
bowtie	Sequence alignment	0.12.8	bowtie-bio.sourceforge.net
bowtie2	Sequence alignment	2.0.9	bowtie-bio.sourceforge.net/bowtie2
tophat	RNA sequence alignment	2.0.6 and 2.0.7	tophat.cbcb.umd.edu
cufflinks	Transcript assembly and gene expression estimation	2.0.2 and 2.1.1	cufflinks.cbcb.umd.edu
BEDTools	Utilities for working with genomic data	2.17.0	code.google.com/p/bedtools
UCSC	Utilities for working with bigWig/bigBed formats		hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64

Table A.3 Confidence tiers.

Tier	Description
Tier 1a	Multi-exonic transcripts that overlap existing annotations on the same strand and match splicing (allowing for 5' and 3' end variation)
Tier 1b	Multi-exonic transcripts that overlap existing annotations on the same strand and have at least one shared splice junction
Tier 1c	Transcripts that overlap existing annotations on the same strand but do not share splice junctions
Tier 2a	Multi-exonic unannotated transcripts
Tier 2b	Single exon unannotated transcripts

Table A.4 Validation information.

Transcript Name	MiTranscriptome Gene ID	Forward Primer	Reverse Primer	Validated by Sanger Sequencing?
SNHG12.3	G001404	TCCCAGAGACAACTCCAGAGG	AGTAAGTCGACACCGGGAAT	not tested
FOXD2-AS1.1	G002192	CCACGCTTCAAAATCCCTGC	AGTGTGGCCTGAGAATGAGC	not tested
NA	G002815	CTCACTGCCTCAGTGCTTGT	AGGGTTTTACGTAGGGGAGAGA	not tested
NA	G002815	TGCAACTTGTGCCAGTCAGA	AGTGGCCTTCTACATTTGATCAGT	not tested
CAT122	G004900	TGCTCCTGAAAAGAGACACCC	GCTGTCTAACCTCCATCCTTCC	not tested
CAT154	G006409	GCGCAGAGAACCAGTTTGAC	TCCCCTTTTCCGTTCAGTGG	not tested
CAT184	G007371	GGGGCAAAGGATTCGTCAGA	CCTGGGTGCTCAGGAATGTT	not tested
MEAT144	G007561	GACCACATGATGTTGCCTGC	CACTGAACCTGTCAGGACCC	not tested
LINC00857.2	G010412	TAGCAGGGCAAGGCCATTTT	CCCAGGATGCCTGTTGTTCA	no
NA	G012208	TGCCTGTCTGGTCTTCGTTC	AAAGAGCATAGAGTGCCGGG	not tested
CAT1369	G014085	TCACCGTCTGAACAACCAGG	GGAGCTCCATCTTGGCGTG	not tested
CAT1370	G014086	AGGGGAGTGGACATCGTG	AGCAGTGGATTTGAAGACTGT	not tested
NA	G014665	TCCAGCCAGCTGACGAAAAA	CTCGGTGCTAGGCTACTCTG	not tested
CAT1399	G014757	TGACATGCGTGACAAGTCGG	CACCTCCTTGTTCCCACAGAA	not tested
CAT1439	G016152	GCTGTCTCCCACACTGTACC	CCTAAGCGTGAGTTGGCTCT	not tested
CAT1518	G018248	TCAGTGGCTCTCACCTGTACG	CGGAGAAGATCCCTGGCTTTG	not tested
TMPO-AS1	G019668	GGGGAGTTACCCCTCTCCTC	CCCGATTGTAGAGGTCTGGC	not tested
CAT1572	G020156	CAGCAACTGCTAACAGCGAG	TTCAGCAGTCAGGGAGGTCT	yes
LACAT8.1	G020776	ATTCAAGGCTGTCTGCTCCG	CACATCTTCCTTGGTCCCCC	not tested
LACAT8.2	G020776	ACATGGGCGATCTCACTGTC	ACATCTCTGCCTCACTTGGC	not tested
OVAT5	G026110	GGCCTCGCGAGACTTCC	GATCGATTGGCCGAGAGGAG	no
NA	G026573	GAACCGGAACTCCCCTTGC	TGAAGAGGCGCTTGGGTAGA	not tested
LACAT237	G026779	TCCAAAGTCTCTTTCCCGCC	TTGACCCGTTCGGTCAACAA	not tested
CAT1796	G029106	TGGATGTGTGCAAAGAATCTGC	TCTTTGGGGATCCAAATGATGA	not tested
THCAT426	G030545	ACAATGTAGCCACGGAGTCG	GGGGGTCCATCCCTCCTATT	yes
FBXL19-AS1	G030577	CAGTCTTGTCGAACAGGGGG	ACTGATTCCCGGAGGTCTGA	yes
NA	G031681	GACGGTCACGGACACTTTCT	TTGTGTCTCCCAGGACAACG	not tested
MAP3K14-AS1.1	G034565	GGTTCTTCCTGCACAGTCGG	AGTCCTCTGGAGCTGATTCTGT	not tested
CAT1944	G034663	ACATGCTCCTAGGCCTGAATG	CTCGCACCCTCTGATGTAGG	not tested
NA	G034726	AGAGAGCTGGGGAGTCTTCTT	CTGGTCCCTCTTCCAACCAAA	not tested
NA	G036186	CCTCGTAAGGCAAGAACGTG	CTCCAACAGCAAAACCAGGTA	not tested
NA	G037188	GTGGTGAGGGGTTCTCTCAA	GGGGAAAGAGGAGCTAGTGC	not tested
ILF3-AS1	G039106	GCATAGTCACATGATACGTCAAA	CAAGAATATGATTATGCACCTCTGA	not tested
CAT2069.2	G039292	CTGCTGCTGCGTCGTTTG	GGCTGCACCGGAGTCTG	yes
CAT2074	G039503	ACCTCTCTCTCGGCTCTGAAG	ACAGTTTTTAAAGCATTTTTCTGGC	not tested
HNCAT2	G040445	ACAGATAAGCCGGTGCTGG	ATTCACGCCCGAACAAAGAG	not tested
BRCAT105	G041270	TTGAGCCTCCATTAACGCCA	AGGGCGCTTGCTTTTGTTTC	not tested
PCAT130	G041952	CAGGGCACCCCTAAATCCAG	GAGGGAGCAATCTCTCATCGG	yes
CAT255.1	G043944	AAGAGGAAGAAAGCCCTCTGT	GTTTCGGCATGAATCCCCAG	not tested
ZFAS1.3	G050732	TGCAGTCAGGCTTCATACGC	GTAGATGTCTGCACGTGGCT	yes
CAT2179.2	G050801	CCCTTTGGCCGAAATTCCAC	TTGTGAGCTTCTCAAGGGCA	not tested
CAT2180.1	G050799	TTAACCCTCACAGCACCCTTC	TGCAGCTCAACTCCAAAGATG	not tested
NA	G053442	TCTGGCAGAGAGAGGCAGAT	CCTGGTGAGACATGTGGCTT	not tested
CAT2258.3	G054111	CTCGGAGCAGAAAGAGGAAACT	GTGTGGCTGCATTCCATTTG	not tested
CAT388	G054791	TCTGGATTGCATTCAGAGACTGT	AGCTCAAAATAAAACTACGCCTGT	not tested
LINC00973.2	G056743	CCTCACTGTCAACCCTGATCTC	TCCAGCTGTGTAATCATCTCCC	not tested
LINC00973.1	G056743	CCCTCACTGTCAACCCTGAT	CCCTCTGTGACCTCGGACAA	not tested
CAT506	G059675	GGTTGGGTTCTCCTCTTCCA	CACCATGAAAAGGGCTTGTGG	not tested
PCAT21	G059684	CAGAACATGACCACCCGTCA	CGGACGCGACTTTTCATTGG	not tested
PCAT75.2	G062751	GGCCTCCTTTCCTAAGTGGG	TGGGAGGACACAGACCTGAA	not tested
CAT681	G066452	GCATTTTAGCCGGACTTGGT	TCTGGGTCGGCCTTCCTTTA	not tested
CAT700	G067117	AGTCAATCTGGCTGGAAGCA	TGTCTTGGTGATGCTAAAGATCG	not tested
PCAT13.1	G069333	AGTGGCTCACCTTGACTTGG	AGTATTGGGTGGGTTGTGCC	not tested
PCAT13.1	G069333	CTCCCCAGGCTTCTTAACCC	TGCGTCAGTCTACACTCCAG	not tested
CAT791	G070500	AACCCTCATCAAAGTAGAATACCA	CAGCACTGATTGTTGCTGTGG	not tested
CAT800.1	G070842	AGAGAAAACCACGTGCACCA	TTTTCAGATCTGCCGCCTCA	not tested
CAT819.8	G071445	TGAAGTTGAGGGTCTGTGGC	CAATGAGCATGGGTTGACGC	not tested
CAT822	G071584	GATGACGCTGACCAGAGAGG	TGGCAAGTGCCCATACAGTC	not tested
CAT837	G071978	TGATACCTGCACAACACTGC	TGGCCAATTTGGGGTATGGT	not tested
CAT869	G073201	GTCAGACTACTCCTTCGGCG	GGGTAGCGGTAGAAGCATCC	not tested
CAT944	G076482	CAGCGACGGCCATCAAAAA	CGTTCTGCTTCCACGTGTCT	not tested
CAT979	G078253	GGGGAGTGCTCTGGGTTATC	ACGTATGTTGAGCCTGCGTT	not tested
CAT989	G078524	GGATTTCTCAGCCTGCGACT	TTGGAGGGACTGGACACTCA	yes
CAT1005.3	G079390	CCCCCAGACCAGTTTTTCCA	TGATGCCATCTGTGCCTCAAT	not tested
CAT1107	G083084	CTGAGACCTCTTGCCTGACG	TGCAGTGAACACCTTGACCT	not tested
CAT1122	G083486	ACTTCCTGTCCTGGTGAACTG	TCCCCGGAGACAGAAGTAGA	not tested
CAT1193	G086614	GGGGCCAATGAAGCCAATTC	GCTTCTGGCCCCATAGACTG	yes
CAT2268	G090676	AGTTTCTTTTGCCTGCACAGC	GGGGACTCCTAGGCTGCTAA	not tested
CAT1501.3	G017919	TGCCAGCATAAGAAATCAAGGG	GCCGAGAAAAAGAAACATCCC	not tested
DSCAM-AS1.3	G052296	TGGAACCCACTGATGGTATGC	CCGATCCGTCGTCCATCTC	yes
CAT2051	G038822	AAGCATGTGAGACACAGAACA	CTCTCCGCAGGCTCAATGAA	not tested
CAT2277.3	G090914	AGCCCACTTTTAGGCTTAGGAC	CCACACCTTTTATTGACACAGAA	not tested
CAT1984	G036124	CTTTCCCTGTCCGTTGGAGT	GGGTGCTCCTCATTCCTTCA	not tested
CAT1785	G028840	GACAGCAGCATACAGAGCAC	TCCTAGGAGTCTGAAGGCCT	yes
CAT270	G044540	CCTCTTCAGACTGGACTCCA	CAGCCAGTTCAATCCTCTTCAT	not tested
CAT1519	G018251	TTCAGGCCAGGAAAGCTGAG	ACTGTACAGAAGAGCCACCA	not tested
CAT1383.2	G014466	GAATGCTCACTCAGTCCCAG	CATTCCACAGGTCCCCTCTC	not tested
CAT1655	G024299	GAGACAGCGAGTGAAGAAGC	ACACATCTTCCTGGGGTCAT	not tested
BRCAT95	G021137	CTCACCTTGCTGCAGTTGGT	TGATTGCTGTGAGTTGGGTG	yes
CAT1841.3	G030453	GAGGAGACGGGACCAGTGAG	CTGTGGAGAGCAAGTTCCAG	not tested
CAT745.1	G068660	GATGGAGACGGGGAAGTTGA	ATGCTGGAGGAAGGGTGTTT	yes
EPB41L4A-AS1.3	G067083	TTACAGCTGTTTCCACGTGC	AGCAGTCTACACTTCAGGCA	yes
CAT1202.3	G086826	CATCCTCCTCGTCGTCTTCC	TGATGGGGAGTGCAGGTATC	not tested
BRCAT32	G073548	AGCATGCTCTCACCAGACAT	TGACAAGAGCAGAACCCCAT	yes
CAT2176.3	G050667	GGAGTTGCCTTTTCTGGACG	GACCCAACTTTCAGCTTGCA	not tested
PXN-AS1	G020360	AAGAGGAAGCACGCAAATCC	GGGAAAGAGGCTCTAGGACA	not tested
MIR205HG.10	G006597	GTGATTCCGGGATGCCAAAG	CCTCATGGTTGTCAGCTCCA	yes
BRCAT1.6	G008385	AGTACACGATCACTGCCAGT	CAAGACCCAGGAAAGCGAAT	not tested
TRIM52-AS1.1	G069076	CCCATTCCCAAGCCCCTTCT	CCTCCCCAAACCAATGACCA	not tested
UTAT7	G004700	GGATCAGGGAGGAAGGAACA	CGCCTTCCTTTGGTTAGAATCT	not tested
CAT1807	G029343	CTTCCCTCCCCTGCAGTTCA	ATGAAGTGAACCGCAAGCTG	yes
THCAT182	G081983	GGGCGGTGACTTCTAAAATCA	ATTTCAGGCACCCACCAGAT	yes
UTAT18	G027086	TTTCCAGTCCCTCTCCCAAC	CCTCAAGTCTAGCCTCAGGAG	not tested
CDKN2B-AS1	G084101	TGTGTGTTTCCTTGTGAGCT	CAGCTTTGATCTATATGCTTGGG	not tested
KHCAT369	G010197	TGCGCCTCTTTCATCTTCCC	GGTGGAATGGCAGAACTGAT	yes
KDM4A-AS1.3	G002047	GAAGGAGGGTGAAAGGAACG	TACTTTGCCAGGTCCCATGG	not tested
BRCAT269.1	G084941	AACCTGGGAGTGTGTGAGAG	GATTGTAGAGCTCACAAGTGGT	not tested
CAT961	G077310	TGGGAGAAGGACAGAGCTTG	CCCTCCCTGTGGTATCTGAC	not tested
NA	G018253	TAGCCAAAGTCGCCCAAGTT	TGGTGCTAGGTCCCAAGAAG	not tested
PCAT51	G032033	GTTCCATGGCTCTTCCTGAG	CGCAAAGGCACTATTTCCAG	not tested

Table A.5 SSEA Statistics.

Sample Set Abbreviated Name	Sample Set Name	Sample Set Description	Set Type	Set Family	Total Samples	Hits	Misses	FDR Threshold	# positively enriched	# negatively enriched	# enriched (both directions)	# not significant
cancer_type_all	Cancer Type: All	All tumor tissues and cell lines	pancancer	cancer_type	6475	5726	749	1.00E-07	29670	45617	75287	306534
cancer_type_aml	Cancer Type: AML	Acute myelogenous leukemia tissues and cell lines	aml	cancer_type	6475	30	6445	1.00E-07	43298	9010	52308	329513
cancer_type_bladder	Cancer Type: Bladder	Bladder cancer tissues and cell lines	bladder	cancer_type	6475	124	6351	1.00E-07	6343	12880	19223	362598
cancer_type_breast_carcinoma	Cancer Type: Breast Carcinoma	Breast cancer tissues and cell lines	breast	cancer_type	6475	967	5508	1.00E-07	28519	72481	101000	280821
cancer_type_cervical_carcinoma	Cancer Type: Cervical Carcinoma	Cervical cancer tissues and cell lines	cervical	cancer_type	6475	112	6363	1.00E-07	6875	16290	22965	358856
cancer_type_cml	Cancer Type: CML	Chronic myelogenous leukemias	cml	cancer_type	6475	17	6458	1.00E-07	6531	2689	9220	372601
cancer_type_colorectal_carcinoma	Cancer Type: Colorectal Carcinoma	Colorectal carcinoma tissues and cell lines	colorectal	cancer_type	6475	186	6289	1.00E-07	21097	39559	60656	321165
cancer_type_glioblastoma_multiforme_gbm	Cancer Type: Glioblastoma Multiforme (GBM)	Glioblastoma multiforme cancers	gbm	cancer_type	6475	159	6316	1.00E-07	33414	31964	65378	316443
cancer_type_head_and_neck_carcinoma	Cancer Type: Head and Neck Carcinoma	Head and neck cancer tissues	head_neck	cancer_type	6475	302	6173	1.00E-07	19622	72803	92425	289396
cancer_type_hepatocellular_carcinoma	Cancer Type: Hepatocellular Carcinoma	Liver cancer tissues and cell lines	liver	cancer_type	6475	73	6402	1.00E-07	10966	13947	24913	356908
cancer_type_lower_grade_glioma_lgg	Cancer Type: Lower Grade Glioma (LGG)	Lower grade gliomas	lgg	cancer_type	6475	219	6256	1.00E-07	49432	50525	99957	281864
cancer_type_lung_adenocarcinoma	Cancer Type: Lung Adenocarcinoma	Lung adenocarcinoma tissues and cell lines	luad	cancer_type	6475	520	5955	1.00E-07	35346	22509	57855	323966
cancer_type_lung_squamous_cell_carcinoma	Cancer Type: Lung Squamous Cell Carcinoma	Lung squamous cell carcinoma tissues and cell lines	lusc	cancer_type	6475	378	6097	1.00E-07	24486	43456	67942	313879
cancer_type_medulloblastoma	Cancer Type: Medulloblastoma	Medulloblastoma tissues and cell lines	medulloblastoma	cancer_type	6475	14	6461	1.00E-07	5206	1544	6750	375071
cancer_type_melanoma	Cancer Type: Melanoma	Melanoma tissues and cell lines	melanoma	cancer_type	6475	263	6212	1.00E-07	17369	52973	70342	311479
cancer_type_mpn	Cancer Type: MPN	Myeloproliferative neoplasia tissues	mpn	cancer_type	6475	10	6465	1.00E-07	4878	2386	7264	374557
cancer_type_ovarian	Cancer Type: Ovarian	Ovarian cancer tissues and cell lines	ovarian	cancer_type	6475	423	6052	1.00E-07	53659	43536	97195	284626
cancer_type_pancreatic_carcinoma	Cancer Type: Pancreatic Carcinoma	Pancreatic cancer tissues and cell lines	pancreatic	cancer_type	6475	59	6416	1.00E-07	4825	870	5695	376326
cancer_type_prostate_carcinoma	Cancer Type: Prostate Carcinoma	Prostate carcinoma tissues and cell lines	prostate	cancer_type	6475	243	6232	1.00E-07	21529	26912	48441	333380
cancer_type_renal_cell_carcinoma_chromophobe	Cancer Type: Renal Cell Carcinoma (Chromophobe)	Chromophobe renal cell carcinomas	kich	cancer_type	6475	64	6411	1.00E-07	14190	13678	27868	353953
cancer_type_renal_clear_cell_carcinoma	Cancer Type: Renal Clear Cell Carcinoma	Renal clear cell carcinomas	kirc	cancer_type	6475	463	6012	1.00E-07	46282	42838	89120	290701
cancer_type_renal_papillary_cell_carcinoma	Cancer Type: Renal Papillary Cell Carcinoma	Renal papillary cell carcinoma	kirp	cancer_type	6475	76	6399	1.00E-07	8035	4010	12045	369776
cancer_type_stomach_adenocarcinoma	Cancer Type: Stomach Adenocarcinoma	Stomach adenocarcinoma tissues	stomach	cancer_type	6475	164	6311	1.00E-07	48853	18143	66996	314825
cancer_type_thyroid	Cancer Type: Thyroid	Thyroid carcinoma tissues	thyroid	cancer_type	6475	464	6011	1.00E-07	33654	79276	112930	268891
cancer_type_uterine_endometrial_carcinoma	Cancer Type: Uterine Endometrial Carcinoma	Endometrial carcinoma tissues	uterine	cancer_type	6475	276	6199	1.00E-07	38910	42238	81148	300673
cancer_versus_normal_bladder	Cancer versus Normal: Bladder	Cancer versus Normal: Bladder	bladder	cancer_vs_normal	142	124	18	0.001	327	3932	4259	377902
cancer_versus_normal_breast	Cancer versus Normal: Breast	Cancer versus Normal: Breast	breast	cancer_vs_normal	1076	967	109	0.001	23335	46000	69335	312486
cancer_versus_normal_head_and_neck	Cancer versus Normal: Head and Neck	Cancer versus Normal: Head and Neck	head_neck	cancer_vs_normal	337	302	35	0.001	6804	7228	14032	367789
cancer_versus_normal_kidney_chromophobe	Cancer versus Normal: Kidney Chromophobe	Cancer versus Normal: Kidney Chromophobe	kich	cancer_vs_normal	89	64	25	0.001	11472	17963	29435	352386
cancer_versus_normal_kidney_clear_cell	Cancer versus Normal: Kidney Clear Cell	Cancer versus Normal: Kidney Clear Cell	kirc	cancer_vs_normal	533	463	70	0.001	54330	28649	82979	298842
cancer_versus_normal_kidney_papillary_cell	Cancer versus Normal: Kidney Papillary Cell	Cancer versus Normal: Kidney Papillary Cell	kirp	cancer_vs_normal	101	76	25	0.001	4466	10332	14798	367023
cancer_versus_normal_liver	Cancer versus Normal: Liver	Cancer versus Normal: Liver	liver	cancer_vs_normal	109	73	36	0.001	2923	4716	7639	374182
cancer_versus_normal_lung_adenocarcinoma	Cancer versus Normal: Lung Adenocarcinoma	Cancer versus Normal: Lung Adenocarcinoma	luad	cancer_vs_normal	656	520	136	0.001	27589	56264	83853	297968
cancer_versus_normal_lung_squamous	Cancer versus Normal: Lung Squamous	Cancer versus Normal: Lung Squamous	lusc	cancer_vs_normal	424	378	46	0.001	21728	34610	56338	325483
cancer_versus_normal_prostate	Cancer versus Normal: Prostate	Cancer versus Normal: Prostate	prostate	cancer_vs_normal	304	243	61	0.001	4551	7403	11954	369867
cancer_versus_normal_stomach	Cancer versus Normal: Stomach	Cancer versus Normal: Stomach	stomach	cancer_vs_normal	188	164	24	0.001	2161	2317	4478	377343
cancer_versus_normal_thyroid	Cancer versus Normal: Thyroid	Cancer versus Normal: Thyroid	thyroid	cancer_vs_normal	524	464	60	0.001	11987	26841	38828	342993
normal_type_all	Normal Type: All	All normal tissues and cell lines	pancancer	normal_type	6475	749	5726	1.00E-07	41179	26054	67233	314588
normal_type_bladder	Normal Type: Bladder	Normal or benign adjacent bladder tissue	bladder	normal_type	6475	18	6457	1.00E-07	214	26	240	381681
normal_type_embryonic_stem_cells	Normal Type: Embryonic Stem Cells	Human embryonic stem cell samples	hesc	normal_type	6475	10	6465	1.00E-07	2627	3115	5742	376079
normal_type_head_and_neck	Normal Type: Head and Neck	Benign adjacent tissues from head and neck cancer patients	head_neck	normal_type	6475	35	6440	1.00E-07	3360	5597	8957	372864
normal_type_heart	Normal Type: Heart	Tissues derived from heart	heart	normal_type	6475	10	6465	1.00E-07	1900	933	2833	378988
normal_type_kidney	Normal Type: Kidney	Kidney tissue adjacent to tumor and tissue from healthy patients	kidney	normal_type	6475	122	6353	1.00E-07	25599	29219	54818	327003
normal_type_liver	Normal Type: Liver	Normal or benign adjacent liver tissues	liver	normal_type	6475	36	6439	1.00E-07	10343	8373	18716	363105
normal_type_lung	Normal Type: Lung	Benign lung tissue and lung tissue from cancer free individuals	lung	normal_type	6475	184	6291	1.00E-07	45224	20744	65968	315853
normal_type_mammary_gland	Normal Type: Mammary Gland	Normal tissues and cell lines arising from mammary gland	breast	normal_type	6475	109	6366	1.00E-07	20120	12244	32364	349457
normal_type_prostate	Normal Type: Prostate	Benign adjacent tissues from prostate and benign cell lines	prostate	normal_type	6475	61	6414	1.00E-07	6202	3343	9545	372276
normal_type_skeletal_muscle	Normal Type: Skeletal Muscle	Skeletal muscle tissues and cell lines	skeletal_muscle	normal_type	6475	12	6463	1.00E-07	609	1356	1965	379856
normal_type_stomach	Normal Type: Stomach	Normal tissue arising from stomach	stomach	normal_type	6475	24	6451	1.00E-07	3456	2236	5692	376129
normal_type_thyroid	Normal Type: Thyroid	Benign adjacent tissues from thyroid cancer patients	thyroid	normal_type	6475	60	6415	1.00E-07	14595	9827	24422	357399

Table A.6 MiTranscriptome Overlap With Intergenic Cancer GWAS SNPs.

GWAS SNP ID	SNP Disease Association	Pubmed	Transcript Name	MiTranscriptome Gene ID	Genomic Location	SSEA Association	Distance (Bases)
rs965513	Thyroid cancer	19198613	CAT1172	G085855	chr9:1004997	thyroid,lusc	48866
rs965513	Thyroid cancer	19198613	THCAT577	G085858	chr9:1005485	thyroid	0
rs965513	Thyroid cancer	19198613	THCAT557	G085861	chr9:1005654	thyroid	9336
rs944289	Thyroid cancer	19198613	LINC00609.3	G024258	chr14:363667	thyroid	0
rs944289	Thyroid cancer	19198613	PTCSC3.4	G024264	chr14:365987	thyroid,kich	15343
rs944289	Thyroid cancer	19198613	PTCSC3.3	G024264	chr14:365987	thyroid,kich	3116
rs944289	Thyroid cancer	19198613	PTCSC3.2	G024264	chr14:365987	thyroid	1891
rs944289	Thyroid cancer	19198613	PTCSC3.1	G024264	chr14:365987	thyroid	0
rs944289	Thyroid cancer	19198613	THCAT174	G024265	chr14:366055	thyroid	37385
rs944289	Thyroid cancer	19198613	PTCSC3.5	G024264	chr14:366170	thyroid	1417
rs944289	Thyroid cancer	19198613	CAT1654	G024266	chr14:366220	thyroid,kich	23370
rs944289	Thyroid cancer	19198613	LINC00609.4	G024258	chr14:366260	thyroid,kich	0
rs944289	Thyroid cancer	19198613	CAT1653	G024258	chr14:366831	thyroid,lusc	33898
rs9383938	Breast cancer	22976474	BRCAT32	G073548	chr6:1519134	breast	49381
rs817826	Prostate cancer	23023329	PCAT37.2	G086117	chr9:1101804	prostate	24184
rs817826	Prostate cancer	23023329	PCAT37.5	G086117	chr9:1101804	prostate	24184
rs817826	Prostate cancer	23023329	PCAT37.4	G086117	chr9:1101804	prostate	24184
rs817826	Prostate cancer	23023329	PCAT35	G086118	chr9:1101870	prostate	30772
rs817826	Prostate cancer	23023329	PCAT37.1	G086117	chr9:1101903	prostate	34011
rs817826	Prostate cancer	23023329	PCAT37.3	G086117	chr9:1101970	prostate	40737
rs7837688	Prostate cancer	20676098	CAT1099.2	G082926	chr8:1282997	prostate	19609
rs7837688	Prostate cancer	20676098	CAT1099.1	G082926	chr8:1285036	prostate	0
rs7716600	Breast cancer	20872241	CAT659	G065609	chr5:4482624	kirc,breast	46203
rs7716600	Breast cancer	20872241	BRCAT160	G065610	chr5:4483197	breast	41984
rs720475	Breast cancer	23535729	OR2A1-AS1.2	G078570	chr7:1440042	breast	22447
rs6983267	Colorectal cancer,	23266556, 21	CAT1099.2	G082926	chr8:1282997	prostate	0
rs614367	Breast cancer	23535729, 20	CAT1383.2	G014466	chr11:692903	kirc,breast	34098
rs614367	Breast cancer	23535729, 20	CAT1383.3	G014466	chr11:692903	kirc,breast	32708
rs614367	Breast cancer	23535729, 20	CAT1382.2	G014467	chr11:692903	kirc,breast	36801
rs614367	Breast cancer	23535729, 20	CAT1382.1	G014467	chr11:692903	kirc,breast	36663
rs614367	Breast cancer	23535729, 20	BRCAT8.7	G014468	chr11:692964	breast	14961
rs614367	Breast cancer	23535729, 20	BRCAT8.3	G014468	chr11:692964	breast	14961
rs614367	Breast cancer	23535729, 20	BRCAT8.6	G014468	chr11:692965	breast	14961
rs614367	Breast cancer	23535729, 20	BRCAT8.5	G014468	chr11:692965	breast	14961
rs614367	Breast cancer	23535729, 20	BRCAT8.4	G014468	chr11:692975	breast	14961
rs614367	Breast cancer	23535729, 20	BRCAT8.1	G014468	chr11:692987	breast	20846
rs614367	Breast cancer	23535729, 20	BRCAT8.2	G014468	chr11:693005	breast	20846
rs614367	Breast cancer	23535729, 20	BRCAT3.1	G014469	chr11:693057	breast	8044
rs614367	Breast cancer	23535729, 20	BRCAT3.2	G014469	chr11:693057	breast	0
rs614367	Breast cancer	23535729, 20	BRCAT9.1	G014470	chr11:693221	breast	0
rs614367	Breast cancer	23535729, 20	BRCAT9.2	G014470	chr11:693221	breast	0
rs614367	Breast cancer	23535729, 20	BRCAT9.3	G014470	chr11:693458	breast	17076
rs4784227	Breast cancer	20585626	CASC16.2	G031065	chr16:525858	breast	0
rs445114	Prostate cancer	21743057, 19	CAT1099.2	G082926	chr8:1282997	prostate	0
rs4415084	Breast cancer	21263130	BRCAT107.3	G065601	chr5:4449467	breast	0
rs4415084	Breast cancer	21263130	BRCAT107.2	G065601	chr5:4449467	breast	0
rs4242384	Prostate cancer	21743057, 19	CAT1099.2	G082926	chr8:1282997	prostate	0
rs4242384	Prostate cancer	21743057, 19	CAT1099.1	G082926	chr8:1285036	prostate	0
rs4242382	Prostate cancer	18264096	CAT1099.2	G082926	chr8:1282997	prostate	0
rs4242382	Prostate cancer	18264096	CAT1099.1	G082926	chr8:1285036	prostate	0
rs3903072	Breast cancer	23535729	CAT1373	G014216	chr11:655628	skeletal_muscle,lusc,gbm,lgg,medulloblastoma,breast,luad,kirc	11391
rs3803662	Breast cancer	23535733, 23	CASC16.2	G031065	chr16:525858	breast	0
rs3757318	Breast cancer	23535729, 20	BRCAT32	G073548	chr6:1519134	breast	0
rs3734805	Breast cancer	21263130	BRCAT32	G073548	chr6:1519134	breast	1374
rs3112612	Breast cancer	21263130	CASC16.2	G031065	chr16:525858	breast	0
rs2735839	Prostate-specific a	23269536, 18	CAT2124.4	G041354	chr19:513382	thyroid,prostate,kich	8404
rs2735839	Prostate-specific a	23269536, 18	CAT2124.1	G041354	chr19:513446	prostate	8404
rs2735839	Prostate-specific a	23269536, 18	CAT2124.3	G041354	chr19:513446	prostate	8404
rs2735839	Prostate-specific a	23269536, 18	CAT2124.2	G041354	chr19:513446	prostate	8404
rs2242652	Prostate cancer	21743467	CAT635	G064388	chr5:1312499	prostate,breast,luad	32471
rs2072590	Ovarian cancer	23535730, 23	HOXD-AS2.1	G047016	chr2:1769980	ovarian	38841
rs2046210	Breast cancer	23535733, 23	BRCAT32	G073548	chr6:1519134	breast	10390
rs1983891	Prostate cancer	20676098	CAT800.2	G070842	chr6:4149162	prostate	22375
rs1983891	Prostate cancer	20676098	CAT800.1	G070842	chr6:4149162	prostate	22375
rs1805007	Hair color, Non-m	23548203, 23	CAT1889	G032462	chr16:899862	head_neck,kirp,melanoma,breast,luad	117
rs16902094	Prostate cancer	19767754	CAT1099.2	G082926	chr8:1282997	prostate	0
rs16867225	Prostate cancer (g	22219177	PCAT122	G047129	chr2:1809779	prostate	34804
rs1512268	Prostate cancer	22923026, 20	PCAT85.2	G079889	chr8:2351152	prostate	0
rs1512268	Prostate cancer	22923026, 20	PCAT85.1	G079889	chr8:2354052	prostate	14062
rs1447295	Prostate cancer	19767754, 17	CAT1099.2	G082926	chr8:1282997	prostate	0
rs1447295	Prostate cancer	19767754, 17	CAT1099.1	G082926	chr8:1285036	prostate	18658
rs1354774	Prostate-specific a	23359319	CAT2124.4	G041354	chr19:513382	thyroid,prostate,kich	36899
rs1354774	Prostate-specific a	23359319	CAT2124.1	G041354	chr19:513446	prostate	36899
rs1354774	Prostate-specific a	23359319	CAT2124.3	G041354	chr19:513446	prostate	36899
rs1354774	Prostate-specific a	23359319	CAT2124.2	G041354	chr19:513446	prostate	36899
rs13387042	Breast cancer	23535729, 21	BRCAT49	G048017	chr2:2179558	breast	49983
rs12710696	Breast cancer	23535733	CAT215.1	G042577	chr2:1916558	breast	0
rs12409639	Prostate-specific a	23269536	CAT156	G006465	chr1:2056010	thyroid,prostate	33850
rs12203592	Hair color, Non-m	23548203, 23	MEAT9	G069094	chr6:389614-3	melanoma	3398
rs116909374	Thyroid cancer	22267200	LINC00609.3	G024258	chr14:363667	thyroid	32908
rs116909374	Thyroid cancer	22267200	LINC00609.4	G024258	chr14:366260	thyroid,kich	32908
rs116909374	Thyroid cancer	22267200	CAT1653	G024258	chr14:366831	thyroid,lusc	32908
rs11672691	Prostate cancer	23065704, 23	CAT2106.4	G040735	chr19:419597	prostate	144
rs10941679	Breast cancer	23535729	BRCAT107.3	G065601	chr5:4449467	breast	0
rs10941679	Breast cancer	23535729	BRCAT107.2	G065601	chr5:4449467	breast	0
rs10941679	Breast cancer	23535729	BRCAT54.3	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.8	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.7	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.9	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.4	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.1	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT54.2	G065605	chr5:4474241	breast	35919
rs10941679	Breast cancer	23535729	BRCAT77.1	G065606	chr5:4474427	breast	37772
rs10941679	Breast cancer	23535729	BRCAT77.2	G065606	chr5:4475291	breast	46418
rs10503733	Prostate cancer	22923026	PCAT85.2	G079889	chr8:2351152	prostate	5566
rs10503733	Prostate cancer	22923026	PCAT85.1	G079889	chr8:2354052	prostate	6507
rs10090154	Prostate cancer	22923026	CAT1099.2	G082926	chr8:1282997	prostate	12386
rs10090154	Prostate cancer	22923026	CAT1099.1	G082926	chr8:1285036	prostate	0
rs10069690	Breast cancer	23535733, 23	CAT635	G064388	chr5:1312499	prostate,breast,luad	32709

Table A.7 Reported Functional lncRNAs.

lncRNA	Reported Cancer Association	Reported Mechanism of action	SSEA cancer/lineage association	MiTranscriptome GID	References	
HOTAIR	Breast, hepatocellular	Chromatin modification, PRC2 association	Breast	G018377	Gupta, R. A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071-1076, doi:10.1038/nature08975 (2010).	
SCHLaP1	Prostate	Chromatin modification, SWI/SNF association	Prostate	G047139	Prensner, J. R. et al. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Nature genetics 45, 1392-1398, doi:10.1038/ng.2771 (2013).	
KCNQ1OT1	Kidney	Imprinting and chromatin modification	Papillary Renal Cell Carcinoma	G012319	Martens-Uzunova, E. S. et al. Long noncoding RNA in prostate, bladder, and kidney cancer. European urology 65, 1140-1151, doi:10.1016/j.eururo.2013.12.003 (2014).	Pandey, R. R. et al. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Molecular cell 32, 232-246, doi:10.1016/j.molcel.2008.08.022 (2008).
PRNCR1	Prostate	Regulation of AR activity	Prostate	G082910	Chung, S. et al. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. Cancer science 102, 245-252, doi:10.1111/j.1349-7006.2010.01737.x (2011).	
MEG3	Breast, hepatocellular	P53 regulation	Breast, bladder	G026143	Zhou, Y. et al. Activation of p53 by MEG3 non-coding RNA. The Journal of biological chemistry 282, 24731-24742, doi:10.1074/jbc.M702029200 (2007).	
PCAT1	Prostate	Genotoxic stress response	Prostate	G082921	Prensner, J. R. et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nature biotechnology 29, 742-749, doi:10.1038/nbt.1914 (2011).	Prensner, J. R. et al. PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. Cancer research 74, 1651-1660, doi:10.1158/0008-5472.CAN-13-3159 (2014).
PCA3	Prostate	Regulation of androgen signaling	Prostate	G085157	Ferreira, L. B. et al. PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling. BMC cancer 12, 507, doi:10.1186/1471-2407-12-507 (2012).	
PTCSC3	Thyroid	miRNA sponge	Thyroid, low grade glioma, kidney chromophobe, squamous cell lung cancer	G024258	Jendrzejewski, J. et al. The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. Proceedings of the National Academy of Sciences of the United States of America 109, 8646-8651, doi:10.1073/pnas.1205654109 (2012).	
PCAT29	Prostate	Tumor suppressor	Prostate, melanoma, papillary renal cell carcinoma, lung adenocarcinoma, cervical	G028045	Malik, R. et al. The lncRNA PCAT29 inhibits oncogenic phenotypes in prostate cancer. Molecular cancer research : MCR 12, 1081-1087, doi:10.1158/1541-7786.MCR-14-0257 (2014).	

APPENDIX B

Supplemental Data for Chapter 3

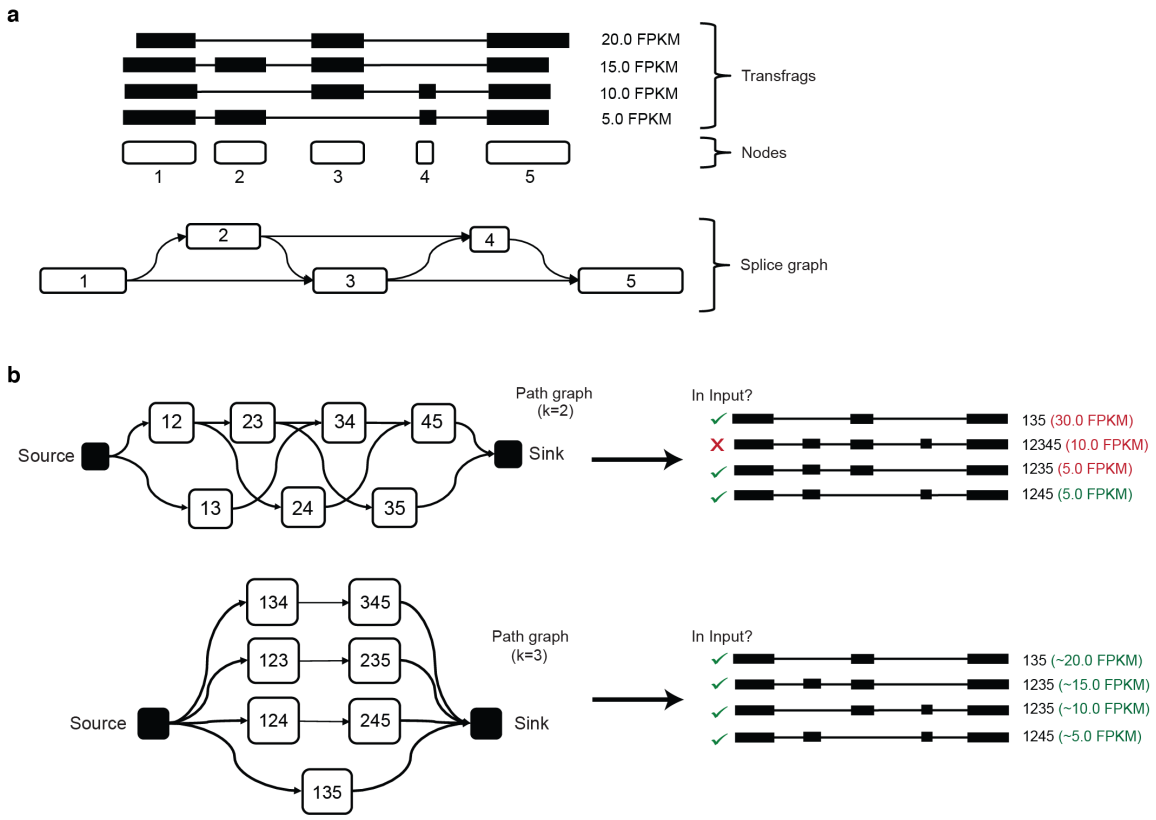


Figure B.1 Schematic example of TACO splice and path graph generation. **a**. Depiction of splice graph generation from 4 input transfrags for a given gene. Expression of input transfrags listed to right of transfrags. Individual nodes indicated by numerical identifier. **b**. Path graphs for k=2 and k=3 from the splice graph represented in (**a**). At k=2, traversing the path graph generates 4 isoforms, one of which is not represented in the input transfrags, additionally the expression distribution of the transfrags does not accurately reflect the input. At k=3, path graph generates 4 transcripts, which are accurate representations of the input data, manifesting the power of utilizing path graphs instead of splice graphs.

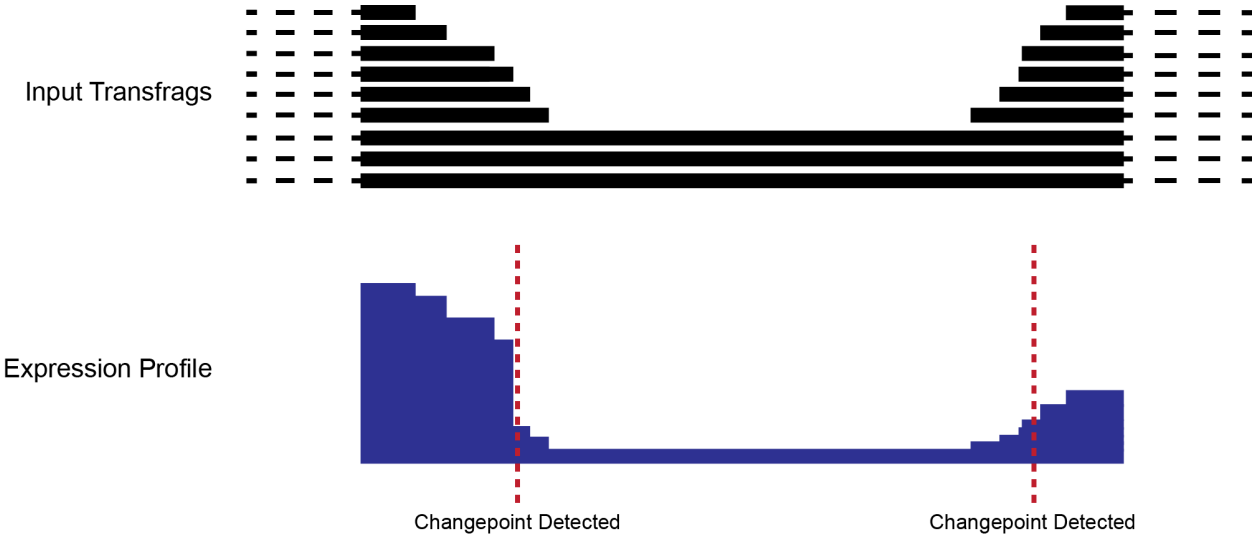


Figure B.2 Schematic depicting the change point detection method.
Multiple transfrags with different expression are used as input. The expression profile is then determined for each base in each node of the splice graph. Changes in the expression profile are then identified as change points.

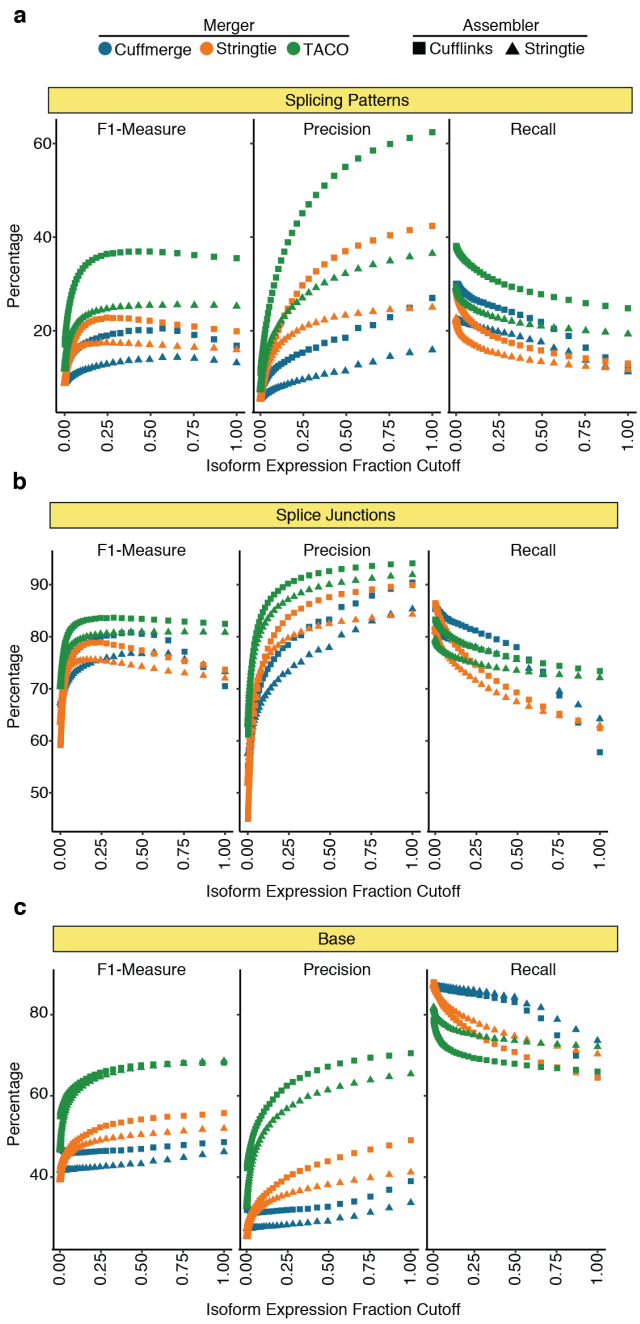


Figure B.3 Assessment of TACO performance at different isoform fraction cutoffs. Performance metrics (recall, precision, and f-measure) describe the merging tool performance on the 55 CCLE breast cell lines at multiple isoform fraction cutoffs (n=50, ranging from 0.001-0.999) for Cuffmerge (blue), Stringtie (organge), and TACO (green) utilizing transcriptome assembly input Cufflinks (squares) and Stringtie (triangles). Performance measured for (**a**) splicing patterns, (**b**) splice junctions, and (**c**) bases.

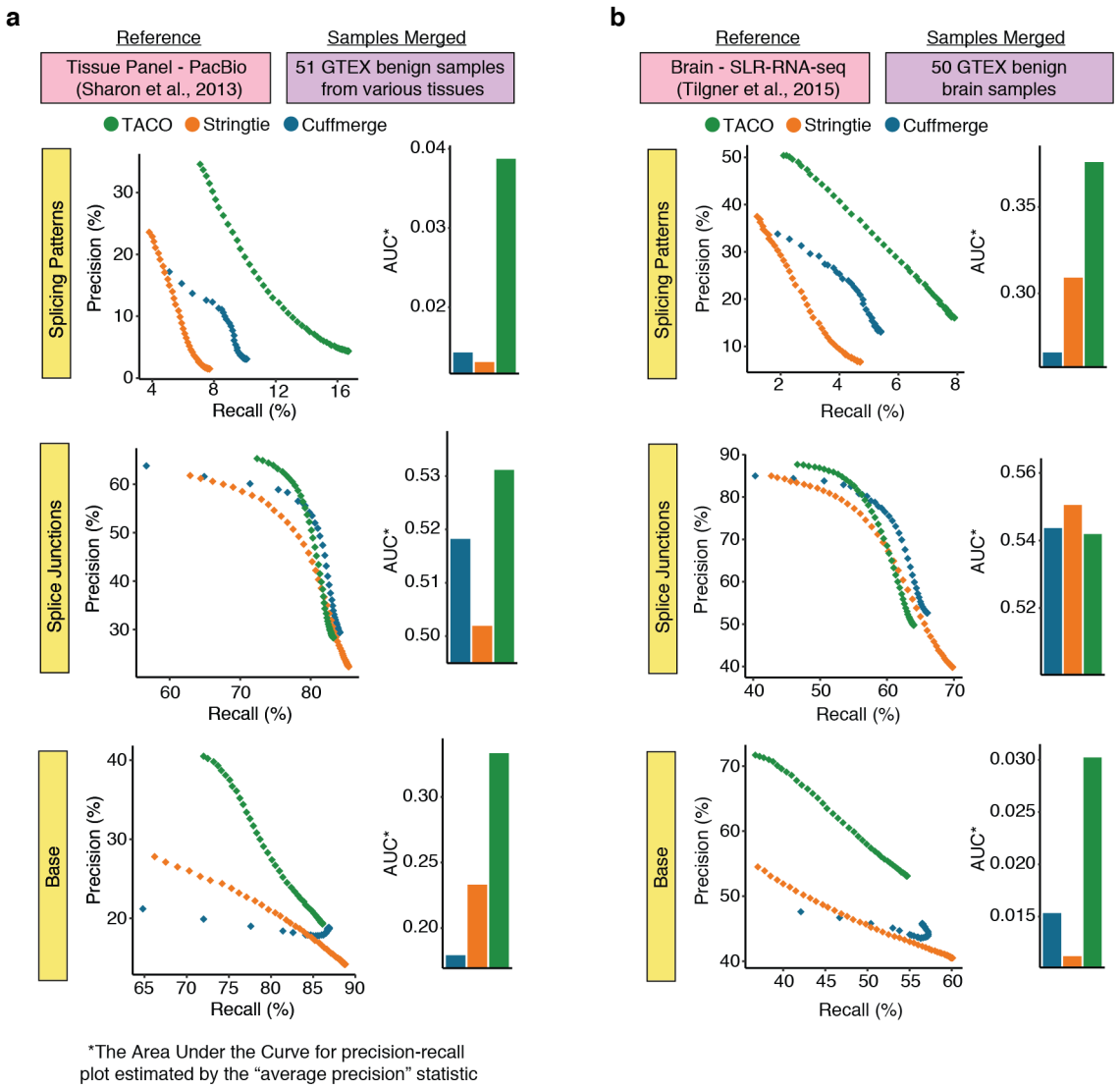


Figure B.4 Assessment of TACO performance using long-read sequencing reference.

a. Precision-recall plots highlighting the performance for the three merging tools at 50 different isoform fraction cutoffs ranging from 0.001-0.999. Statistics describe the merger of the 51 GTEX RNA-seq samples from 17 different tissues types. PacBio long-read sequencing data from a multi-tissue panel was used as a reference for comparison statistics. **b.** Precision-recall plots highlighting the performance for the three merging tools at 50 different isoform fraction cutoffs ranging from 0.001-0.999. Statistics describe the merger of the 50 randomly selected GTEX benign brain tissue RNA-seq sample. Long-read sequencing data of multiple brain samples was used as a reference for comparison statistics.

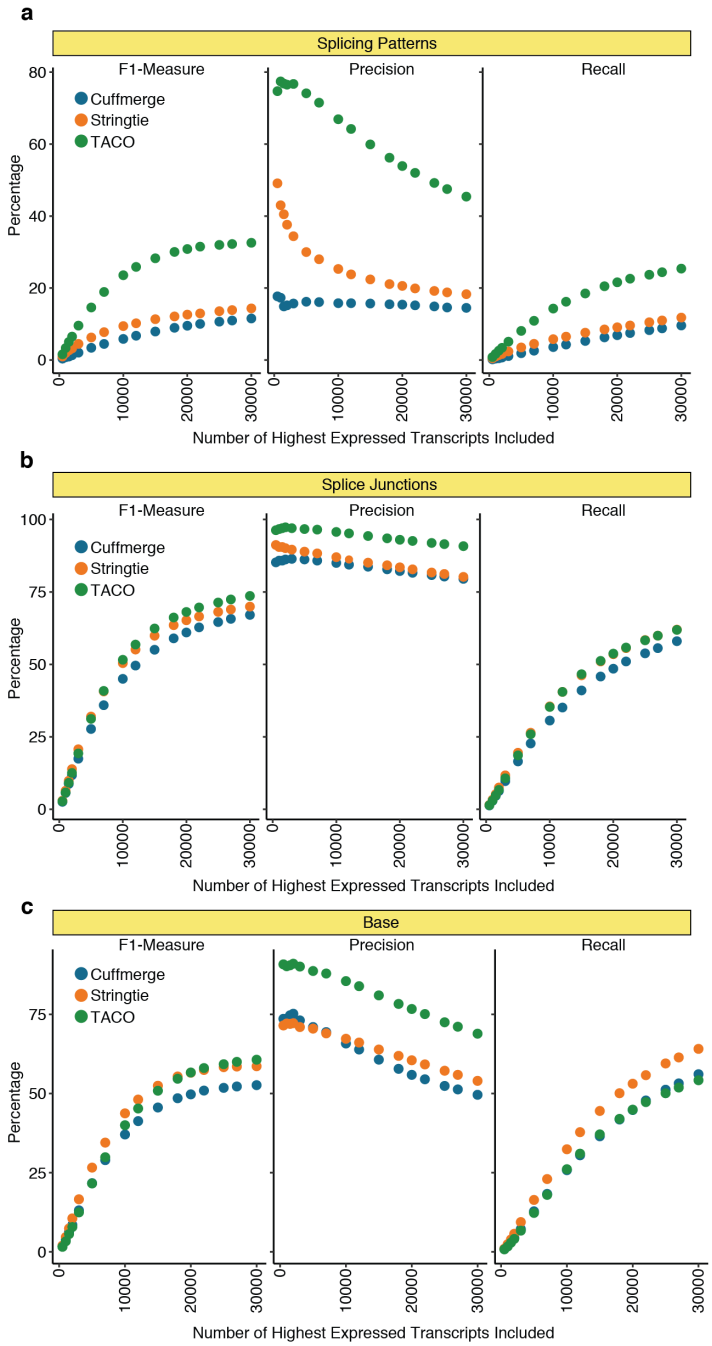


Figure B.5 Assessment of TACO performance for the highest expressed transcripts. Performance metrics (recall, precision, and f-measure) describe the merging tool performance on the 55 CCLE breast cell lines at multiple numbers of the highest expressed transcripts in each meta-assembly generated by Cuffmerge (blue), Stringtie (organge), and TACO (green). Performance measured for (**a**) splicing patterns, (**b**) splice junctions, and (**c**) bases.

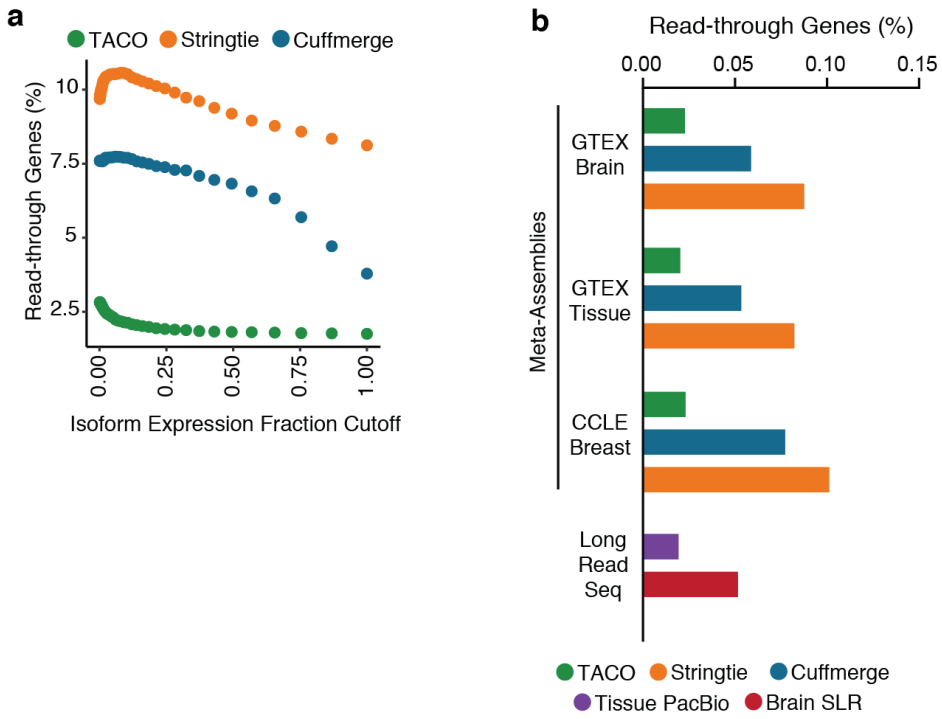


Figure B.6 Assessment of read-through transcription.
a. Scatterplot depicting the fraction of read-through genes produced in the meta-assembly from the three merge tools. Gene designated as being a read-through if it contains at least one transcript isoform bearing exonic overlap with two separate protein-coding genes on the same strand. Fraction of read-throughs was determined for assemblies at 50 different isoform fraction cutoffs ranging from 0.001-0.999. Meta-assembly was performed on the 55 CCLE breast cancer cell lines. **b**. Barplot depicting the fraction of read-through genes produced by meta-assembly of the CCLE breast cohort, the GTEX tissue cohort, and the GTEX brain cohort by all three merging tools. Additionally, the fraction of read-through genes in the long-read sequencing data is also shown. Statistics for meta-assemblers are shown at the default isoform fraction for each tool.

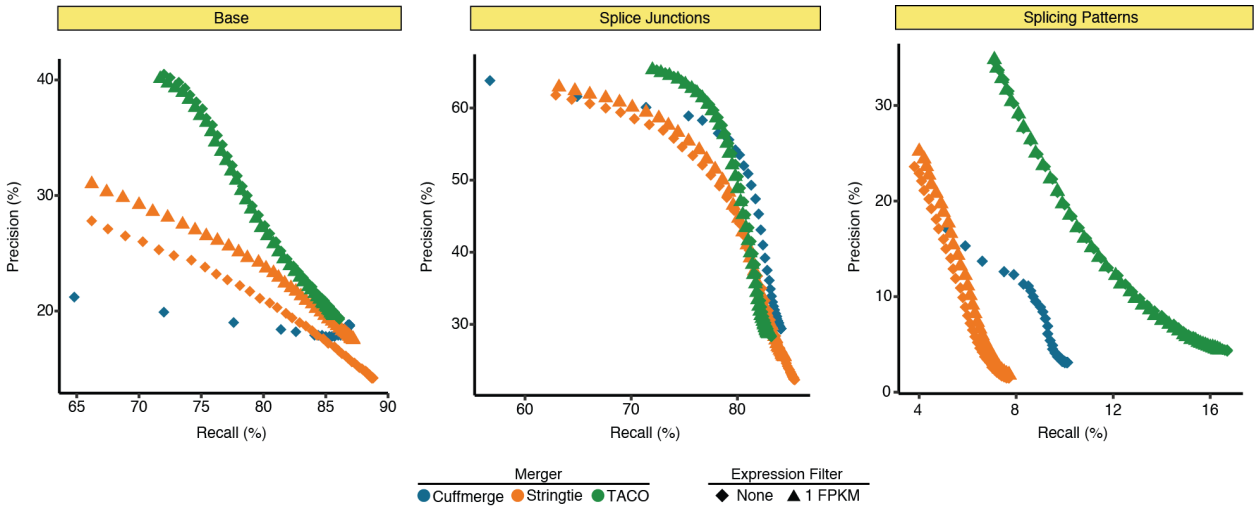


Figure B.7 Assessment of expression filtering on meta-assembly performance. Precision-recall plots are shown highlighting the performance for the three merging tools at 50 different isoform fraction cutoffs ranging from 0.001-0.999. Statistics describe the merger of the 51 GTEX RNA-seq samples from 17 different tissues types. PacBio long-read sequencing data from a multi-tissue panel was used as a reference for comparison statistics. TACO and StringTie-merge were run with and without a 1 FPKM expression filter, delineated by the shape of the point.

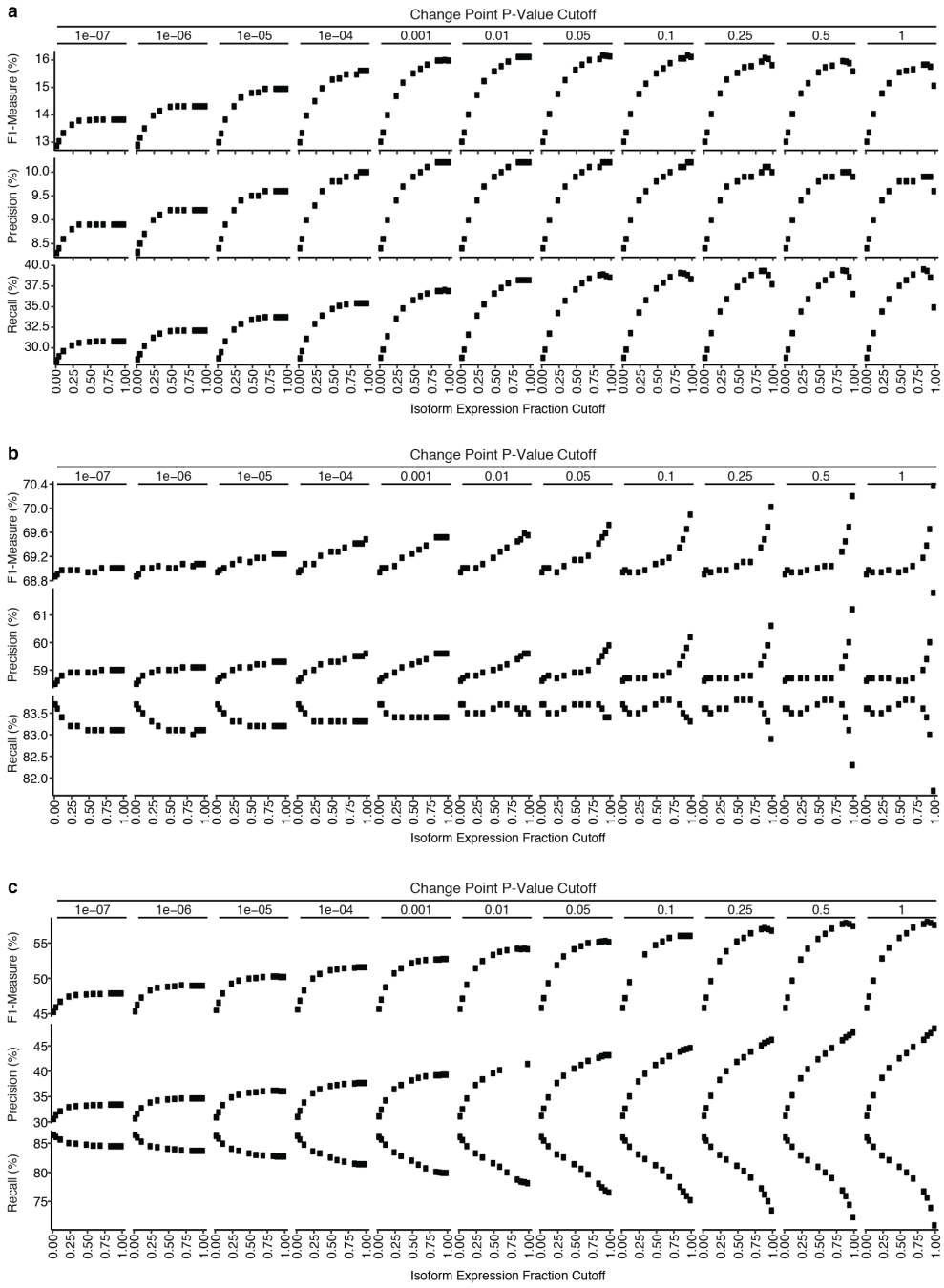


Figure B.8 Change point detection parameters.

Precision, recall, and f-measure statistics for the performance of TACO on merging the 55 CCLE breast samples. Different p-values for the Mann-Whitney U measure performed during change point detection were tested. Additionally, different expression fold-change cutoffs were tested for the two groups produced following change point detection. Statistics reported for (**a**) splicing patterns, (**b**) splice junctions, and (**c**) bases.

APPENDIX C
Supplemental Data for Chapter 4

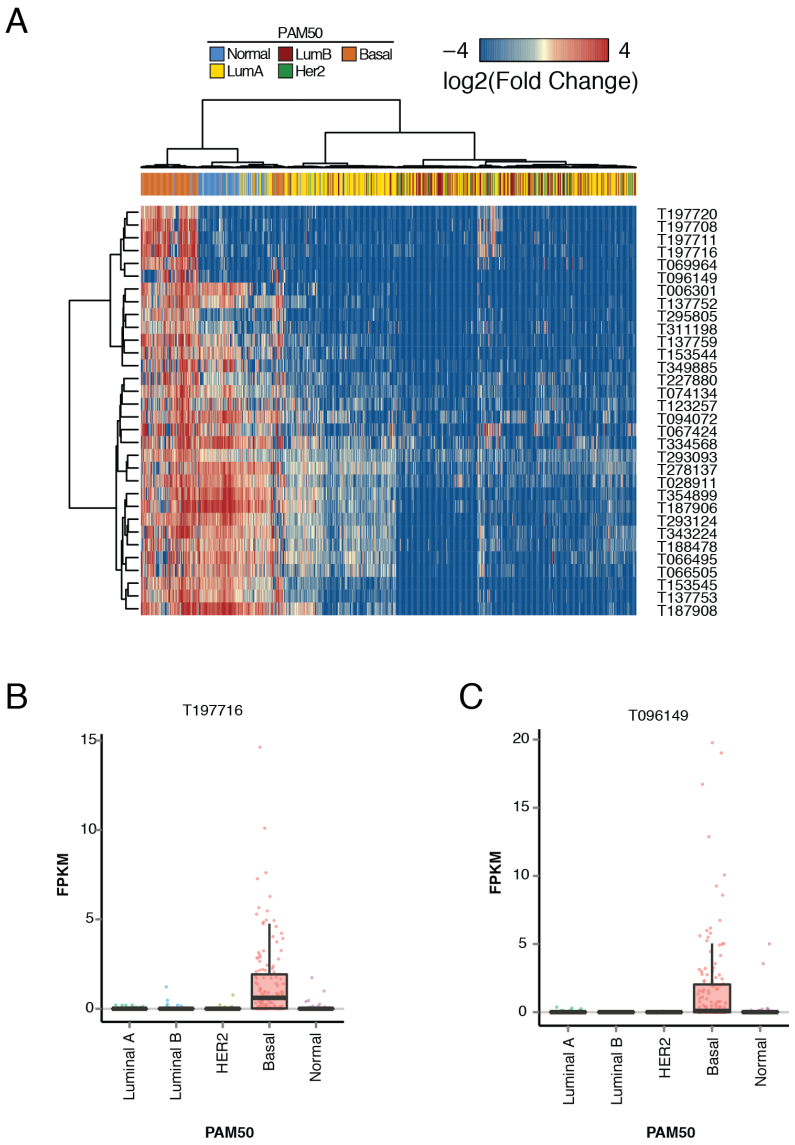


Figure C.1 Discovery of basal specific lncRNAs.
(**a**) heatmap depiction of basal specific lncRNAs delineated in **Figure 4.1b**. 32 lncRNAs were identified as being basal specific. Expression values are depicted as log2 of the fold-change over the median of the ER-negative samples. Unsupervised hierarchical clustering was done on both lncRNAs and patients (n=946). PAM50 classification is shown above heatmap. (**b**) Representative example of expression pattern for

MiTranscriptome transcript T197716 across the PAM50 subtypes. (**c**) Representative example of expression pattern for MiTranscriptome transcript T096149 across the PAM50 subtypes.

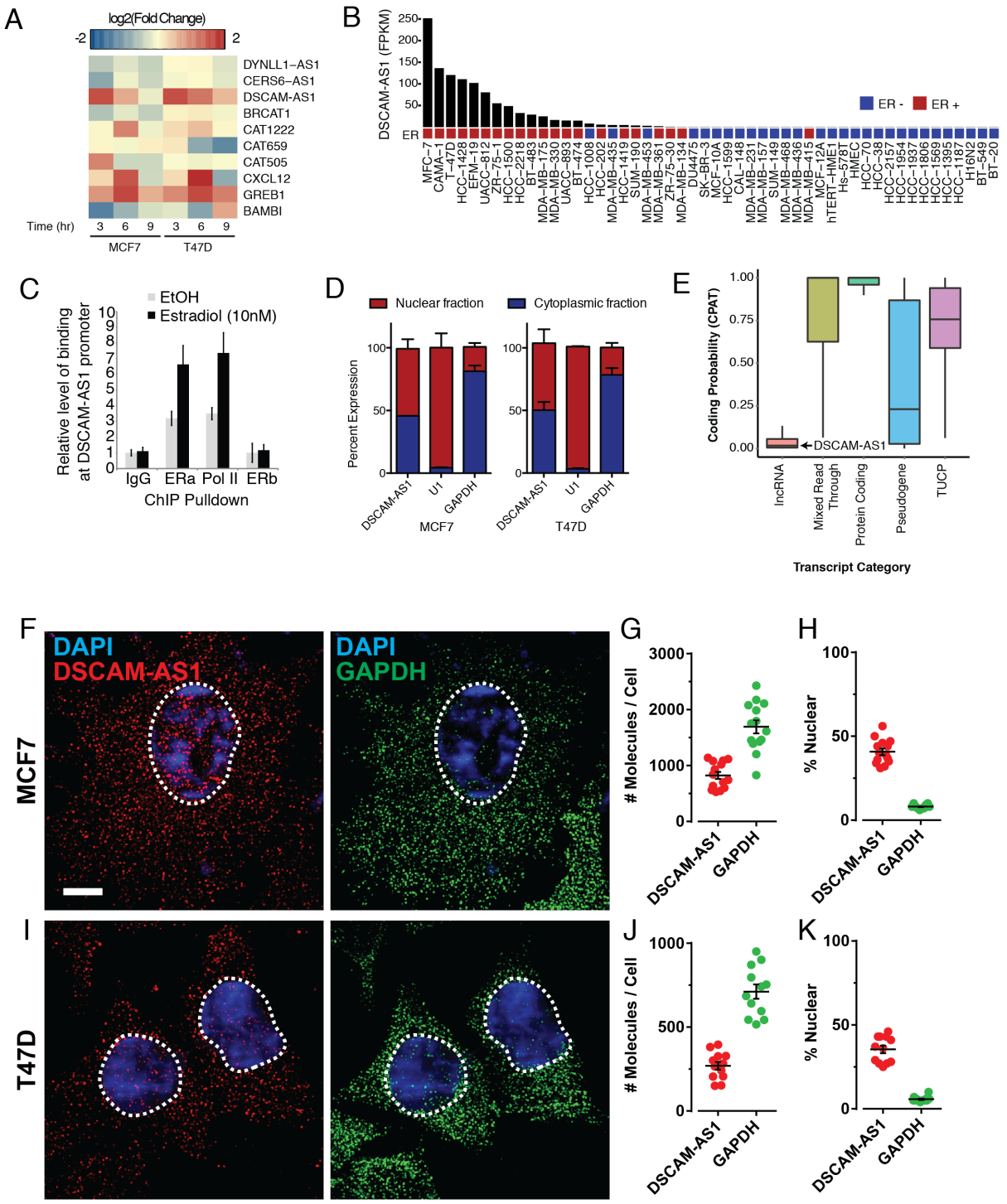


Figure C.2 Characterization of *DSCAM-AS1*.

(**a**) Heatmap representation of qPCR validation of estrogen responsiveness in MCF7 and T47D cell lines. Heatmap values are the log2 of the fold change over time zero for each cell line. The most highly expressed cancer and ER-positive associated lncRNAs were tested. *GREB1* and *CXCL12* were used as a positive control. And *BAMBI* used as a

negative control. (**b**) Expression of *DSCAM-AS1* in 50 breast cancer cell lines from RNA-seq data. ER status of cell line shown in blue/red below expression bar. (**c**) qPCR of the DSCAM-AS1 promoter following ChIP for ERa, ERb, and RNA Pol II following 12hr estradiol or DMSO vehicle stimulation. Expression normalized to IgG pulldown. Error bars represent the s.e.m. for three biological replicates. (**d**) Cellular localization of *DSCAM-AS1* in MCF7 and T47D cells. qPCR performed following cellular fractionation. U1 and GAPDH were used as controls for the nuclear and cytoplasmic fractionation, respectively. Error bars represent the s.e.m. of three biological replicates. (**e**) Boxplot depiction of the CPAT coding potential scores for all MiTranscriptome transcripts by transcript category. CPAT coding potential score for *DSCAM-AS1* (4 isoforms with CPAT score ranging from 0.015-0.016) is highlighted. (**f**) Representative pseudocolored image of an MCF7 cell probed for *DSCAM-AS1* (red, left) and *GAPDH* (green, right). Nucleus is stained with DAPI (blue). White dotted line represents nuclear boundary. Scale bar, 10 μm. B and C, Scatter plots representing the number of DSCAM-AS1 or GAPDH molecules per MCF7 cell. (**g**) and the percentage of transcripts that localize in the nucleus (**h**). Black line and error bars depict the mean and s.e.m. respectively (n =14 cells). (**i**) Representative pseudocolored image of a T47D cell, probed and stained as in (**f**). (**j** and **k**) Scatter plots are similar to (**g**) and (**h**) respectively (n =12 cells).

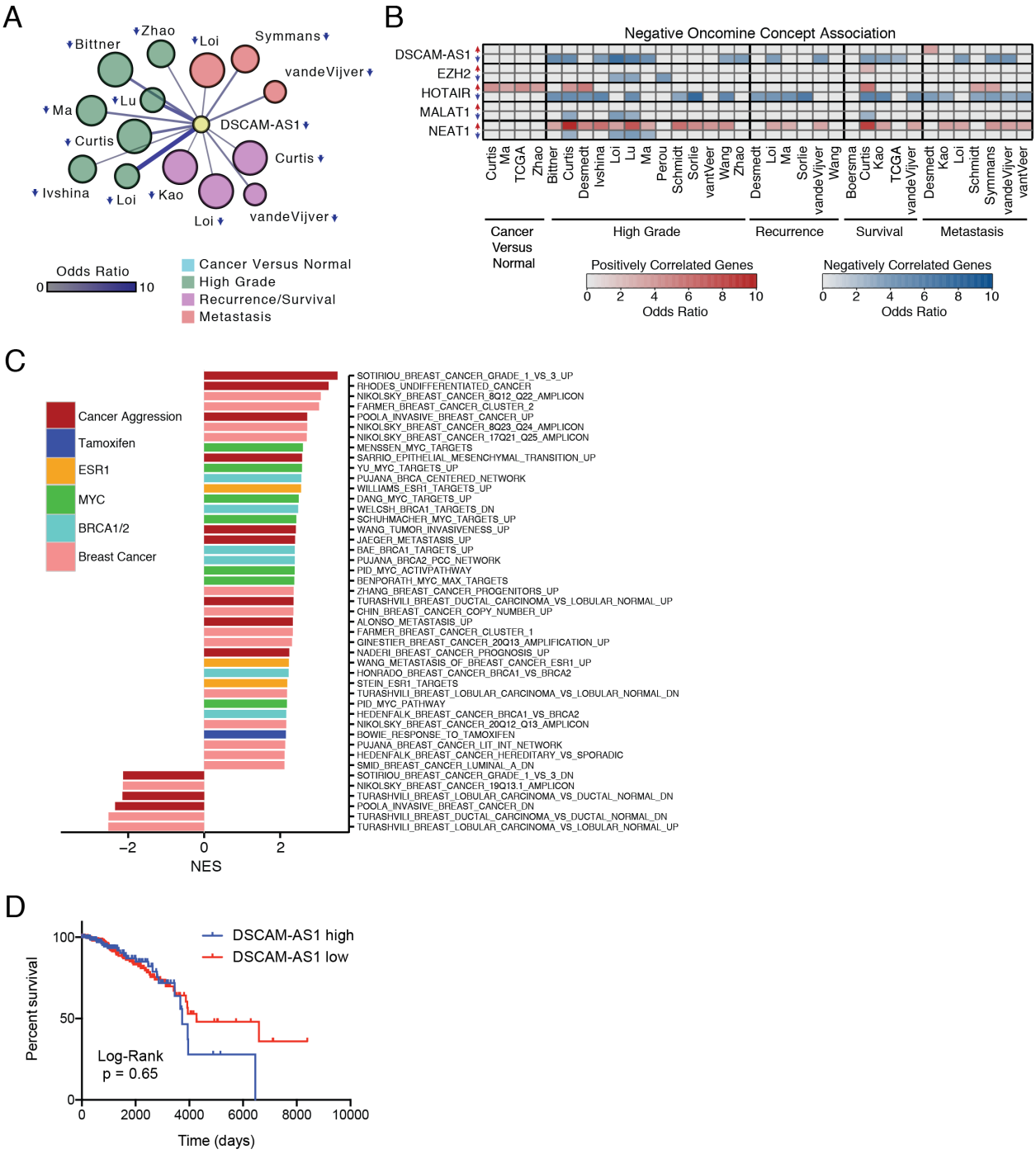


Figure C.3 ***DSCAM-AS1* clinical associations.**

(**a**) Cytoscape depiction of the overlap between the 150 genes most negatively correlated with *DSCAM-AS1* and clinical signatures from Oncomine[149] for breast cancer clinical outcomes (i.e., recurrence, survival, and metastasis), high cancer grade, and cancer versus normal. All significant associations with an odds ratio > 4 are shown (Fisher's p-value < 1e-4). Size of node reflects the size of the gene signature, and the thickness/redness of the line represents the magnitude of the odds ratio. (**b**) Heatmap

displaying the overlap between the top 150 genes correlated to *DSCAM-AS1*, *EZH2*, *HOTAIR*, *MALAT1*, and *NEAT1* and the genes negatively associated with various breast cancer Oncomine clinical signatures for cancer versus normal, high clinical grade, recurrence, survival, and metastasis. For each gene, the top row depicts the odds ratio for the positively correlated genes (red), and the bottom row represents the odds ratio for the negatively correlated genes (blue). The first name of the author for each clinical study is listed. (**c**) Preranked GSEA performed for the genes correlated to *DSCAM-AS1* using the ER-positive breast cancer samples. NES values for all significant (GSEA FDR < 1e-5) MSigDB[115] signatures related to cancer aggression and breast cancer are plotted. (**d**) Kaplan-Meier analysis of survival in the TCGA breast cohort for ER-positive samples. Samples were divided based on DSCAM-AS1 expression (> 10 FPKM = DSCAM-AS1 high; < 1 FPKM = DSCAM-AS1 low). Statistical significance determined by Log-Rank test.

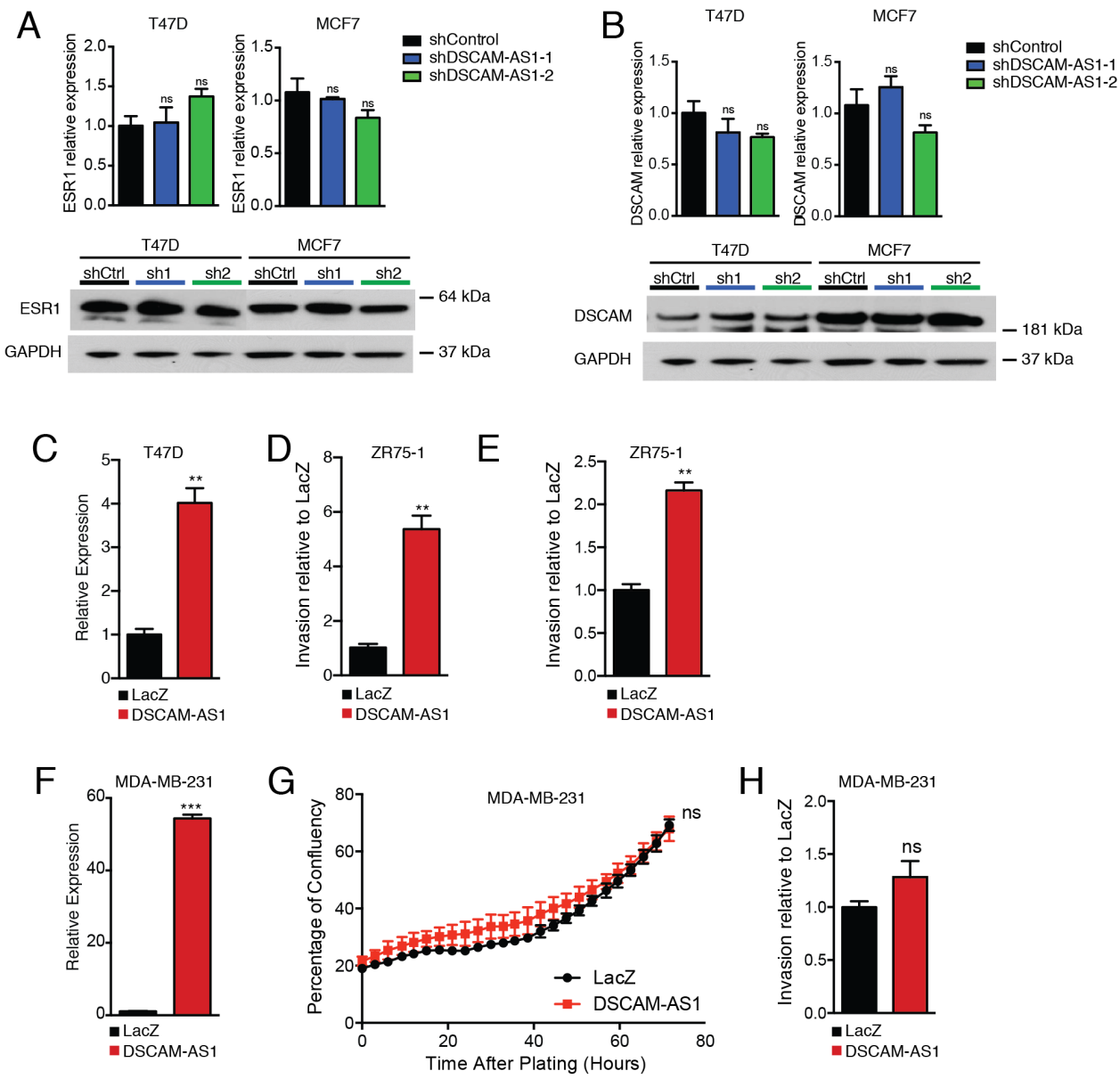


Figure C.4 ***DSCAM-AS1* knockdown and overexpression mediate cancer phenotype.**

(**a**) RNA (top) and protein (bottom) levels of *ESR1* following shRNA knockdown of *DSCAM-AS1* in MCF7 and T47D cells. qPCR expression normalized to shControl. Error bars represent the s.e.m. for three biological replicates. Blot is representative of three independent experiments. (**b**) RNA (top) and protein (bottom) levels of *DSCAM* following shRNA knockdown of *DSCAM-AS1* in MCF7 and T47D cells. qPCR expression normalized to shControl. Error bars represent the s.e.m. for three biological replicates. Blot is representative of three independent experiments. ns: $p > 0.01$, comparing to shControl for each condition via Student's t-test. (**c** and **d**) qPCR expression of *DSCAM-AS1* following overexpression of *LacZ* control and *DSCAM-AS1* in (**c**) T47D cells and (**d**) ZR75-1 cells. **: $p < 0.001$, comparing to LacZ overexpression

via Student's t-test. (**e**) Invasion assay following overexpression of *LacZ* control and *DSCAM-AS1* in ZR75-1 cells. Error bars represent the s.e.m. for three biological replicates. **: $p < 0.001$, comparing to LacZ overexpression via Student's t-test. (**f**) qPCR expression of *DSCAM-AS1* following overexpression of *LacZ* control and *DSCAM-AS1* in MDA-MB-231 cells. ***: $p < 0.0001$, comparing to LacZ overexpression via Student's t-test. (**g**) Incucyte proliferation assay performed following overexpression of *DSCAM-AS1* in MDA-MB-231 cells. Error bars represent the s.e.m. for three biological replicates. ns: $p > 0.01$ comparing to LacZ overexpression via Student's t-test. (**h**) Invasion assay following overexpression of *LacZ* control and *DSCAM-AS1* in MDA-MB-231 cells. Error bars represent the s.e.m. for three biological replicates. ns: $p > 0.01$, comparing to LacZ overexpression via Student's t-test.

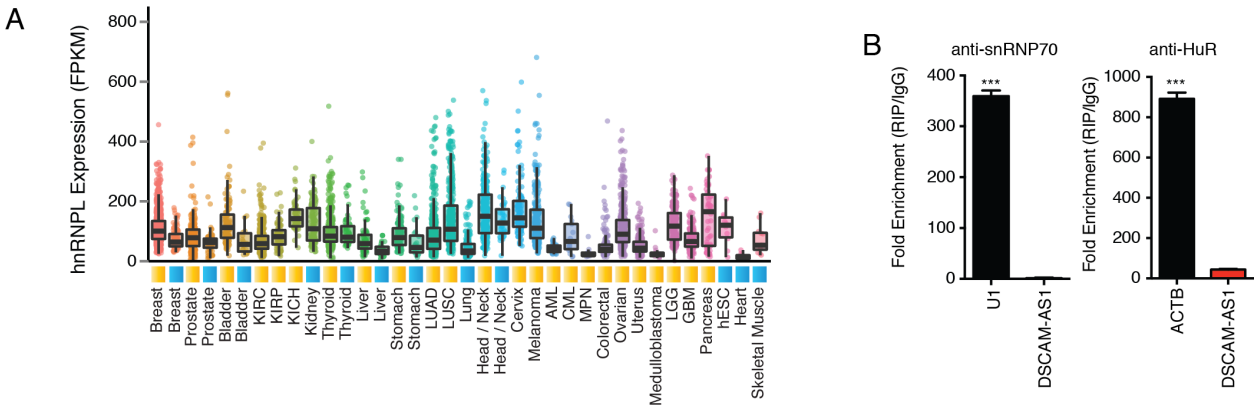


Figure C.5 HnNRPL binds to and phenocopies *DSCAM-AS1*.
(**a**) Expression of hnRNPL in the 6,503 sample MiTranscriptome RNA-seq compendium[192] categorized by the different cancer/tissue types. Each point represents one RNA-seq tissue sample. (**b**) RIP-qPCR for pulldown of snRNPL70 and HuR. *U1* is used as a positive control for snRNP70 pulldown, and *ACTB* is used as a positive control for HuR pulldown. Error bars represent s.e.m. for three biological replicates. ***: $p < 0.0001$, comparing to DSCAM-AS1 fold enrichment via Student's t-test.

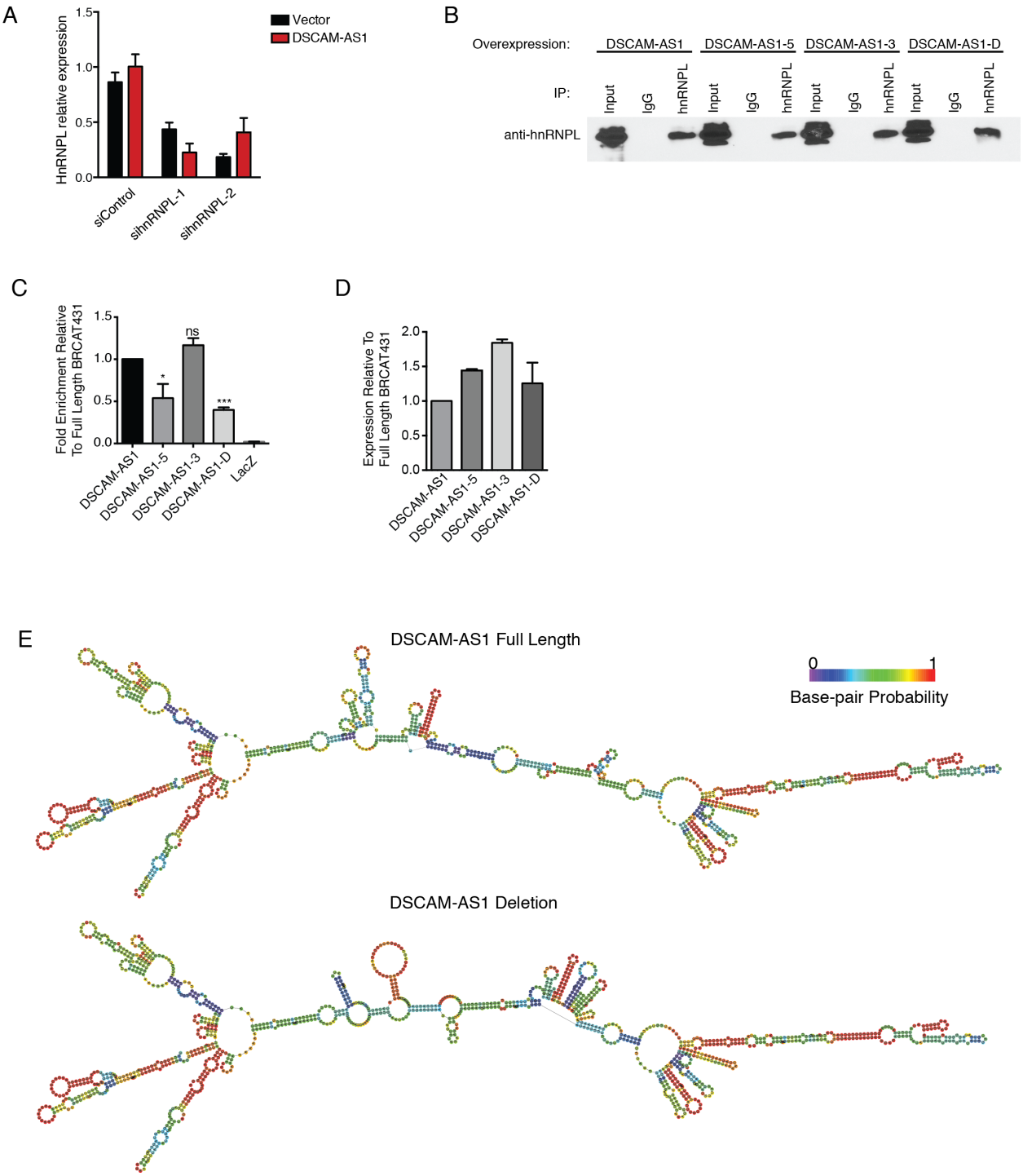


Figure C.6 Binding of hnRNPL localized to region at 3′ end of *DSCAM-AS1*.
(**a**) qPCR expression levels of *hnRNPL* following siRNA knockdown of *hnRNPL* in T47D cells overexpressing *DSCAM-AS1* using two independent siRNAs. (**b**) Western blot for hnRNPL following pulldown of hnRNPL in HEK293 cells expressing the wildtype and mutant forms of DSCAM-AS1, confirming hnRNPL expression in these cells. (**c**) qPCR for the wildtype and mutant forms of *DSCAM-AS1* following hnRNPL RIP in HEK293

cells. Expression normalized to fold-enrichment (hnRNPL RIP/ IgG RIP) of full length DSCAM-AS1. *: $p < 0.01$, ***: $p < 0.0001$, ns: $p > 0.01$, comparing to DSCAM-AS1 full length fold enrichment via Student's t-test (**d**) Expression of wildtype and mutant *DSCAM-AS1* by qPCR in HEK293 cells. (**e**) RNA-fold minimum free energy secondary structure prediction for DSCAM-AS1 full length and the DSCAM-AS1-D deletion construct. Color represents base-pair probability.

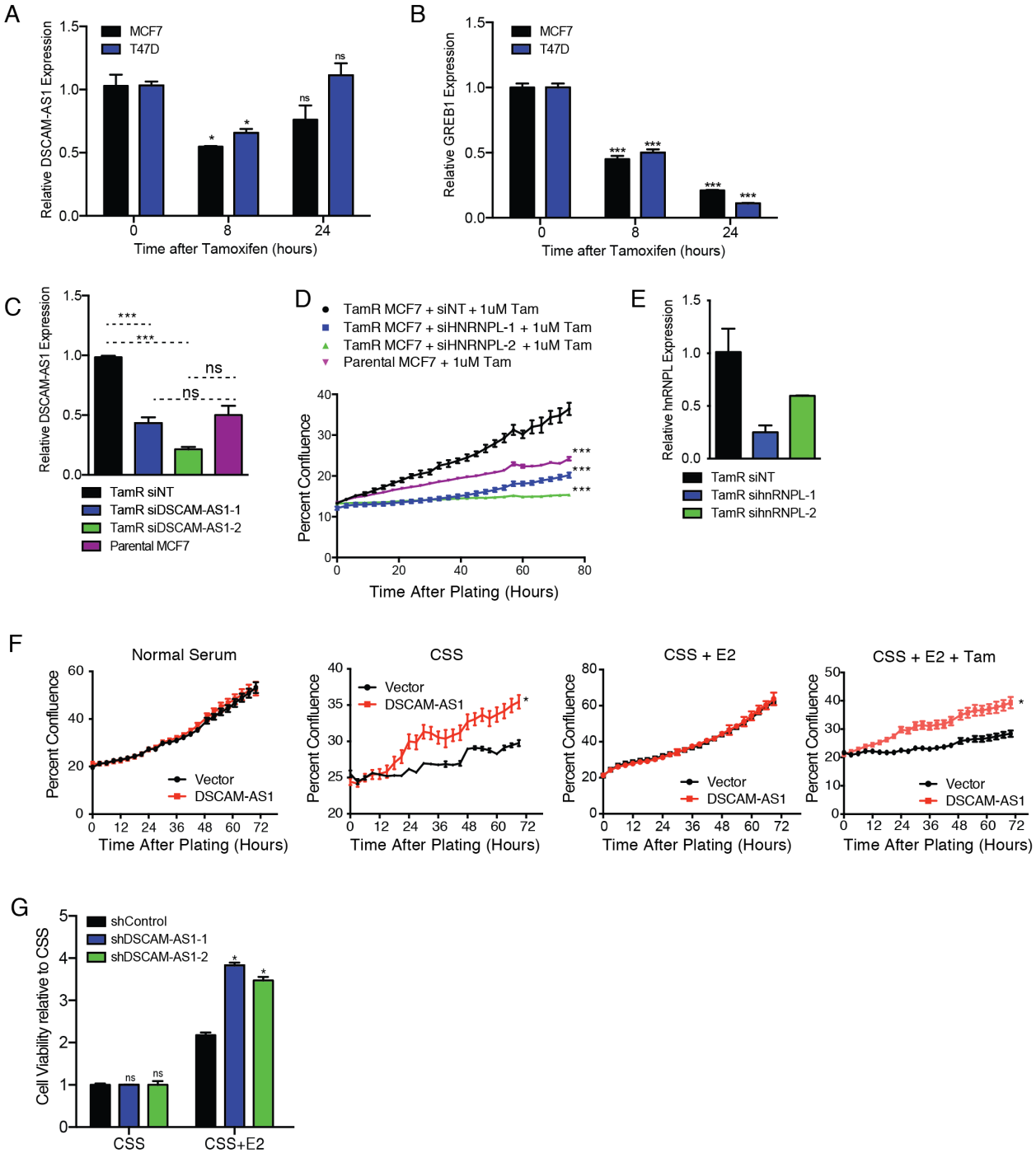
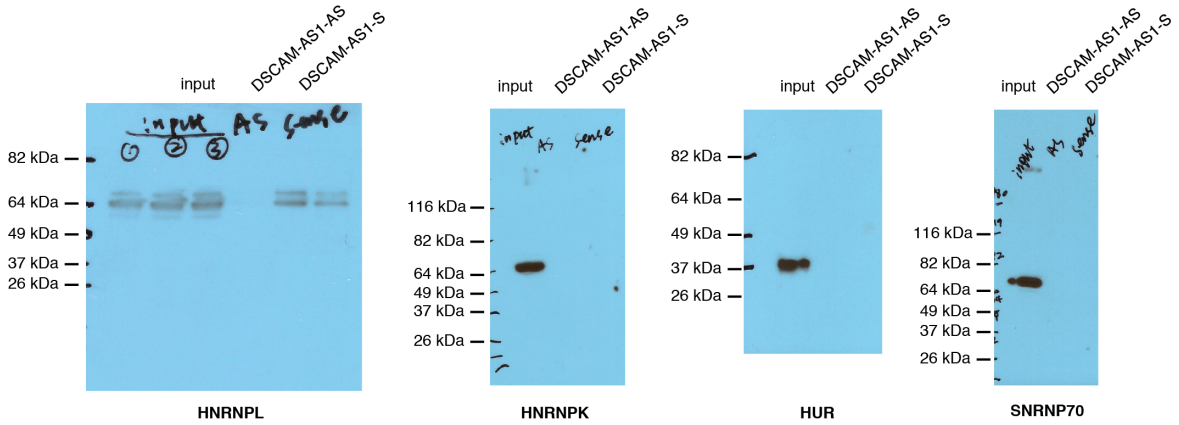


Figure C.7 *DSCAM-AS1* and hnRNPL mediate estrogen independent growth in breast cancer cells.

(**a** and **b**) qPCR expression of (**a**) *DSCAM-AS1* and (**b**) *GREB1* following addition of 1.25 uM tamoxifen in parental MCF7 cells. *: $p < 0.01$, ***: $p < 0.0001$, ns: $p > 0.01$, comparing to 0hr in each condition via Student's t-test (**c**) qPCR expression of *DSCAM-AS1* in TamR MCF7 cells following siRNA knockdown of *DSCAM-AS1*. Expression plotted

relative to non-targeting siRNA control. (**d**) Proliferation assay in parental MCF7 cells and in TamR MCF7 cells following siRNA-mediated knockdown of *hnRNPL* via two independent siRNAs. (**e**) qPCR expression of *hnRNPL* in TamR MCF7 cells following siRNA knockdown of *hnRNPL*. Expression plotted relative to non-targeting siRNA control. (**f**) Proliferation assay comparing T47D cells overexpressing *DSCAM-AS1* versus *LacZ* control in the presence of normal estrogen-containing serum, charcoal stripped serum (CSS), charcoal stripped serum with estrogen (CSS + E2), and charocoal stripped serum with estrogen and 1.25uM tamoxifen (CSS + E2 + Tam). *: $p < 0.01$, ns: $p > 0.01$, comparing to vector overexpression in each condition via Student's t-test. (**g**) Crystal violet cell viability assay following 10 days of culture in charcoal stripped serum (CSS), or charcoal stripped serum with estrogen (CSS + E2) with shRNA knockdown of DSCAM-AS1. Error bars represent the s.e.m. for three biological replicates. *: $p < 0.01$, ns: $p > 0.01$, comparing to shControl in each condition via Student's t-test.

Full Blots for Figure 4.4c



Full Blot for Figure 4.4g

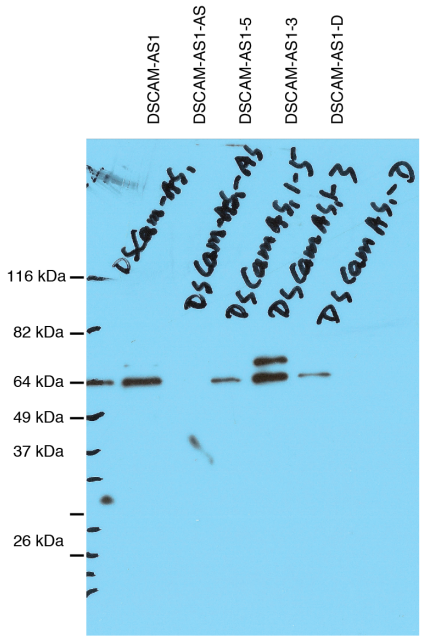


Figure C.8 Representative full Western blot images presented in Figure 4.4c,g.

Table C.1 RACE-determined isofrom sequences for *DSCAM-AS1.*

Isoform	Sequence
1	ctttgggaggctgaggcaggcagatcacctaaggccaggaattcgacaccagcctggccaacgtggcaaaacccgtctctactaaaaatacaaaaattagccgggcgtggtggtgtgcgcctggaatcccagctacccaggaggctgaggcaggagaaatgctggaacccgggaggcagaggctgcagtgagctgagatcatgccactactgcactccagcctgggtgacacagcaagactccctctaaaaaaagaaaaaaagaaaagaaaagaaaagaaaatgatatatccatgatgaattaaaatggagtggaacccactgatggtatgcagctgataagacgctatagagaaatgatatccggacacatggtgagacttttctcctctgtaaatctcaattcttgtgattctttcaaagggaaagccacagaaggtaccagttatccactcactgacttaggtgcctccactagaattctcagcacgtttttgcagaacctgggcaacaagagcgaaaccccatctcaaaaccacaacaacaacaacaggacaacagagatggacgacggatcgggaaagccaaccagacagcgtgaggccaggacggaaagaggcacagggagctctgctcagtgtcgctacaggggatctctcaggctcacaacgggccactcctctagggaagttctggtctcatcatgatccttgtttggtctcactccccatgtccttctctgtccctcctccaactgccatttatttatttaactgaaaaagtaccaatcacccacataggcatgacatactcatccatgtacccatttcttaaaattgatcattgttaacatttggtgtaatttgctttatttatttttaatgaaataaataaaactttacagaaatgctttatttttctctttgttccctccccatcctatatttttctcctaaaaaaccctattatcagaaatattagtgtgtattcccagtttcgacttttattttattacacacacactcccacacatagctgtgacaatgaacttcacatagtatggttctgtattttctctttgtttttcaaacttacataagtattttactatttctcatggaaaaactcacttttcccatccaacgttatgtttccttaagatctctctatgttgatataaagaaatctagctcattctttttaatgaaaataaagtatatttttatgaatgtaactcatgctaaccatggcaataaaagctccatcaagcatgcatttagt
2	ctttgggaggctgaggcaggcagatcacctaaggccaggaattcgacaccagcctggccaacgtggcaaaacccgtctctactaaaaatacaaaaattagccgggcgtggtggtgtgcgcctggaatcccagctacccaggaggctgaggcaggagaaatgctggaacccgggaggcagaggctgcagtgagctgagatcatgccactactgcactccagcctgggtgacacagcaagactccctctaaaaaaagaaaaaaagaaaagaaaagaaaagaaaatgatatatccatgatgaattaaaatggagtggaacccactgatgggaaagccacagaaggtaccagttatccactcactgacttaggtgcctccactagaattctcagcacgtttttgcagaacctgggcaacaagagcgaaaccccatctcaaaaccacaacaacaacaacaggacaacagagatggacgacggatcgggaaagccaaccagacagcgtgaggccaggacggaaagaggcacagggagctctgctcagtgtcgctacaggggatctctcaggctcacaacgggccactcctctagggaagttctggtctcatcatgatccttgtttggtctcactccccatgtccttctctgtccctcctccaactgccatttatttatttaactgaaaaagtaccaatcacccacataggcatgacatactcatccatgtacccatttcttaaaattgatcattgttaacatttggtgtaatttgctttatttatttttaatgaaataaataaaactttacagaaatgctttatttttctctttgttccctccccatcctatatttttctcctaaaaaaccctattatcagaaatattagtgtgtattcccagtttcgacttttattttattacacacacactcccacacatagctgtgacaatgaacttcacatagtatggttctgtattttctctttgtttttcaaacttacataagtattttactatttctcatggaaaaactcacttttcccatccaacgttatgtttccttaagatctctctatgttgatataaagaaatctagctcattctttttaatgaaaataaagtatatttttatgaatgtaactcatgctaaccatggcaataaaagctccatcaagcatgcatttagt
3	ctttgggaggctgaggcaggcagatcacctaaggccaggaattcgacaccagcctggccaacgtggcaaaacccgtctctactaaaaatacaaaaattagccgggcgtggtggtgtgcgcctggaatcccagctacccaggaggctgaggcaggagaaatgctggaacccgggaggcagaggctgcagtgagctgagatcatgccactactgcactccagcctgggtgacacagcaagactccctctaaaaaaagaaaaaaagaaaagaaaagaaaagaaaatgatatatccatgatgaattaaaatggagtggaacccactgatggtatgcagctgataagacgctatagagaaatgatatccggacacatggtgagacttttctcctctgtaaatctcaattcttgtgattctttcaaagggaaagccacagaaggtaccagttatccactcactgacttaggtgcctccactagaattctcagcacgtttttgcagaacctgggcaacaagagcgaaaccccatctcaaaaccacaacaacaacaacaggacaacagagatggacgacggatcgggaaagccaaccagacagcgtgaggccaggacggaaagaggcacagggagctctgctcagtgtcgctacaggggatctctcaggctcacaacgggccactcctctagggaagttctggtctcatcatgatccttgtttggtctcactccccatgtccttctctgtccctcctccaactgccatttatttatttaactgaaaaagtaccaatcacccacataggcatgacatactcatccatgtacccatttcttaaaattgatcattgttaacatttggtgtaatttgctttatttatttttaatgaaataaataaaactttacagaa
4	ctttgggaggctgaggcaggcagatcacctaaggccaggaattcgacaccagcctggccaacgtggcaaaacccgtctctactaaaaatacaaaaattagccgggcgtggtggtgtgcgcctggaatcccagctacccaggaggctgaggcaggagaaatgctggaacccgggaggcagaggctgcagtgagctgagatcatgccactactgcactccagcctgggtgacacagcaagactccctctaaaaaaagaaaaaaagaaaagaaaagaaaatgatatatccatgatgaattaaaatggagtggaacccactgatgggaaagccacagaaggtaccagttatccactcactgacttaggtgcctccactagaattctcagcacgtttttgcagaacctgggcaacaagagcgaaaccccatctcaaaaccacaacaacaacaacaggacaacagagatggacgacggatcgggaaagccaaccagacagcgtgaggccaggacggaaagaggcacagggagctctgctcagtgtcgctacaggggatctctcaggctcacaacgggccactcctctagggaagttctggtctcatcatgatccttgtttggtctcactccccatgtccttctctgtccctcctccaactgccatttatttatttaactgaaaaagtaccaatcacccacataggcatgacatactcatccatgtacccatttcttaaaattgatcattgttaacatttggtgtaatttgctttatttatttttaatgaaataaataaaactttacagaa

Table C.2 Oligonucleotide Primers.

Assay	Gene/Region	Primer (Forward)	Primer(Reverse)
Expression qPCR	DSCAM-AS1	CTAAGGCCAGGAATTCGACAC	CTGGGTAGCTGGGATTC
Expression qPCR	DSCAM-AS1	CTCACTTTTCCCATCCAACG	TGCCATGGTTAGCATGAGTT
Expression qPCR	hnRNPL	GTGTGGTGGAAGCAGACCTTGT	CAAACTCCACCAGTGCTTGTCTC
Expression qPCR	MALAT1	GACGGAGGTTGAGATGAAGC	ATTCGGGGCTCTGTAGTCCT
Expression qPCR	NEAT1	GCTGGACCTTTCATGTAACGGG	TGAACTCTGCCGGTACAGGGAA
Expression qPCR	GAPDH	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG
Expression qPCR	VEGFA	TTGCCTTGCTGCTCTACCTCCA	GATGGCAGTAGCTGCGCTGATA
Expression qPCR	HOTAIR	CCAGAGAACGCTGGAAAAACCTG	GGAGATGATAAGAAGAGCAAGGAA
Expression qPCR	PGR	GTCGCCTTAGAAAGTGCTGTCAG	GCTTGGCTTTCATTTGGAACGCC
Expression qPCR	GREB1	GGTCTGCCTTGCATCCTGATCT	TCCTGCTCCAAGGCTGTTCTCA
Expression qPCR	ACTIN	AAGGCCAA-CCGCGAGAAG	ACAGCCTGGATAGCAACGTACA
Expression qPCR	MYC	CCTGGTGCTCCATGAGGAGAC	CAGACTCTGACCTTTTGCCAGG
Expression qPCR	HMBS	TGGGGCCCTCGTGGAATGTTA	GATGGGCAACTGTACCTGACTGGA
Expression qPCR	U1	GGGAGATACCATGATCACGAAGGT	CCACAAATTATGCAGTCGAGTTTCCC
ChIP-PCR	DSCAM-AS1 - Promoter	AGGCAAACAGTGTCAGTTGG	TGGGATTACAGGCATGAGTGAC
Cloning	DSCAM-AS1 - Full Length	CTTTGGGAGGCTGAGGCAGGCAGATC	ACTAAATGCATGCTTGATGGAGC
Cloning	DSCAM-AS1 5	CTTTGGGAGGCTGAGGCAGGCAGATC	AGGATGGGGAGGGAACAAAG
Cloning	DSCAM-AS1-3	CTTTGTTCCCTCCCCATCCT	ACTAAATGCATGCTTGATGGAGC
Cloning	DSCAM-AS1-D	TTTCGACTTTTTATTTAGCTGTGACAATGA	TCATTGTCACAGCTAAATAAAAAGTCGAAA
BrdU labelling	DSCAM-AS1 WT or D	TAATACGACTCACTATAGGGAGACCTTTGGGAGGCTGAGGCAGGCAGATC	ACTAAATGCATGCTTGATGGAGC
BrdU labelling	DSCAM-AS1-5	TAATACGACTCACTATAGGGAGACCTTTGGGAGGCTGAGGCAGGCAGATC	AGGATGGGGAGGGAACAAAG
BrdU labelling	DSCAM-AS1-3	TAATACGACTCACTATAGGGAGACCCTTTGTTCCCTCCCCATCCT	ACTAAATGCATGCTTGATGGAGC
BrdU labelling	DSCAM-AS1-AS	CTTTGGGAGGCTGAGGCAGGCAGATC	TAATACGACTCACTATAGGGAGACACTAAATGCATGCTTGATGGAGC
BrdU labelling	LacZ	GATCACTAATACGACTCACTATAGGGAGAGGAGAGTCGTTTTACAACGTCGTGACTG	GTATTCGCAAAGGATCAGCGGGC

Table C.3 sh/siRNA Sequence Used.

Gene	Sequence
DSCAM-AS1-1	GAAUUAAAAUGGAGUGGAA
DSCAM-AS1-2	GGACAACAGAGAUGGACGA
hnRNPL-1	GAAUGGAGUUCAGGCGAUG
hnRNPL-2	CUACGAUGACCCGCACAAA

Table C.4 Antibodies Used.

Protein	Species	Vendor	Catalogue number	Purpose
hnRNPL	Mouse	Abcam	ab1606	RIP/WB
hnRNPK	Mouse	Santa Cruz	sc28380	WB
HUR	Rabbit	Millipore	07-468	RIP/WB
SNRNP70	Rabbit	Millipore	03-103	RIP/WB
GAPDH	Mouse	Abcam	ab9484	WB
Estrogen Receptor Alpha	Rabbit	Abcam	ab32063	ChIP
Estrogen Receptor Beta	Rabbit	Abcam	ab3577	ChIP
IgG	Rabbit	Millipore	PP64B	RIP
IgG	Mouse	Millipore	cs200621	RIP
DSCAM	Rabbit	Abcam	ab85362	WB
Anti-Rabbit IgG	Mouse	Jackson Immunoresearch	211-032-171	WB
Anti-Mouse IgG	Goat	Jackson Immunoresearch	115-035-174	WB

APPENDIX D Supplemental Data for Chapter 5

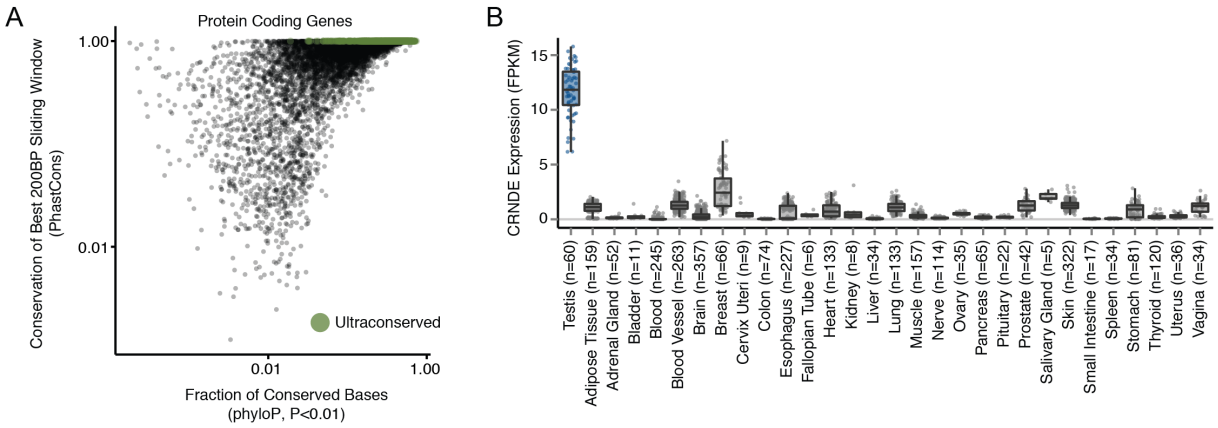


Figure D.1 Conservation of protein-coding genes and expression of CRNDE in normal tissue.

(a) Scatter plot depicting the distribution of basewise transcript conservation levels (x axis) and the average conservation for the best 200bp window (y axis) for all protein coding genes expressed at 1 FPKM or more in the top 1% of TCGA samples. Full transcript conservation levels were measured using the fraction of conserved bases (PhyloP $p < 0.01$). Sliding window conservation levels were measured using the average PhastCons score across 200bp regions along the transcript. Green points indicate transcripts with 200bp windows that meet the criteria for 'ultraconserved' regions. **(b)** Expression of lncRNA CRNDE amongst the GTEX normal tissue RNA-seq dataset, spanning a myriad of different normal tissue types.

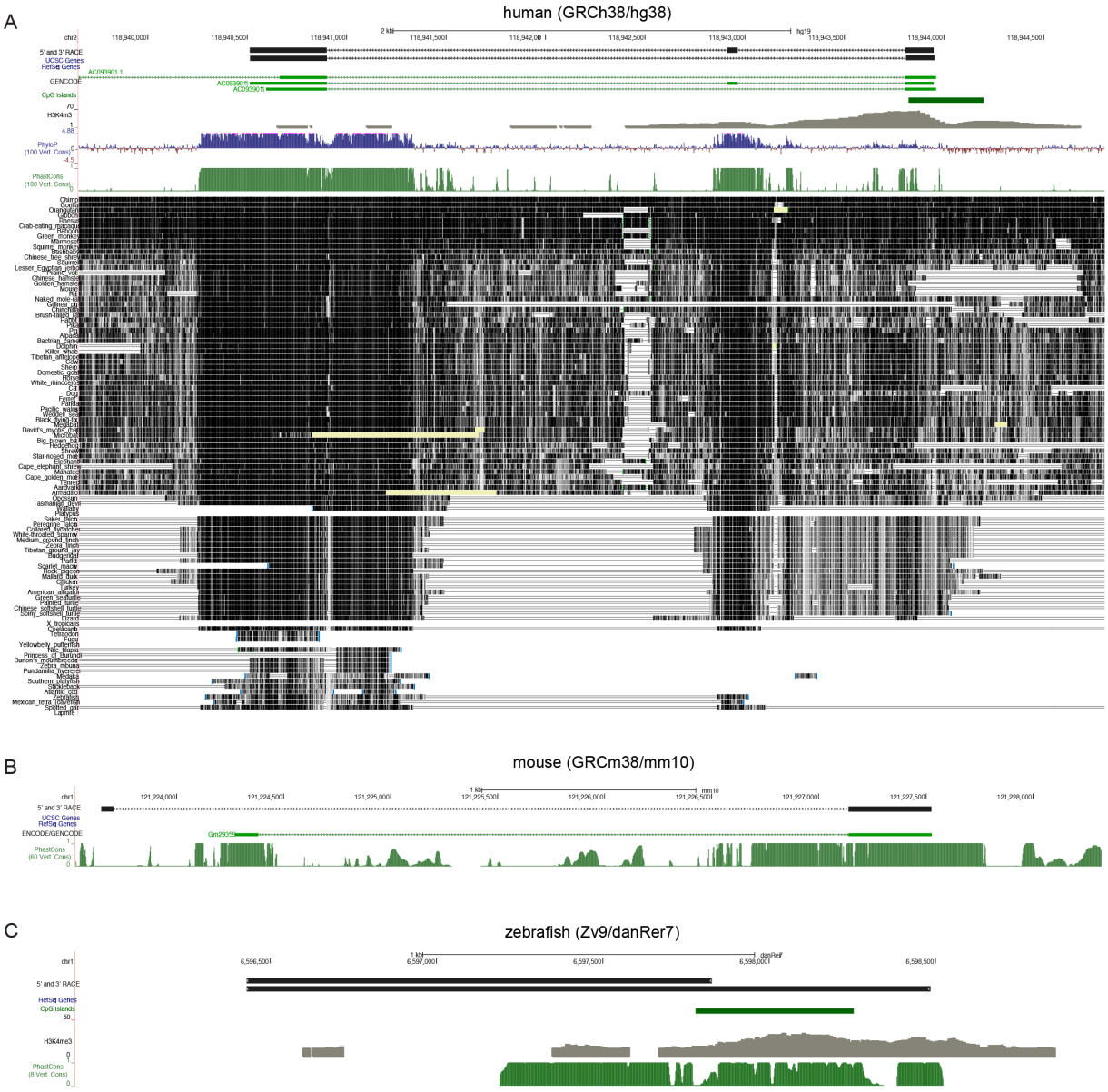
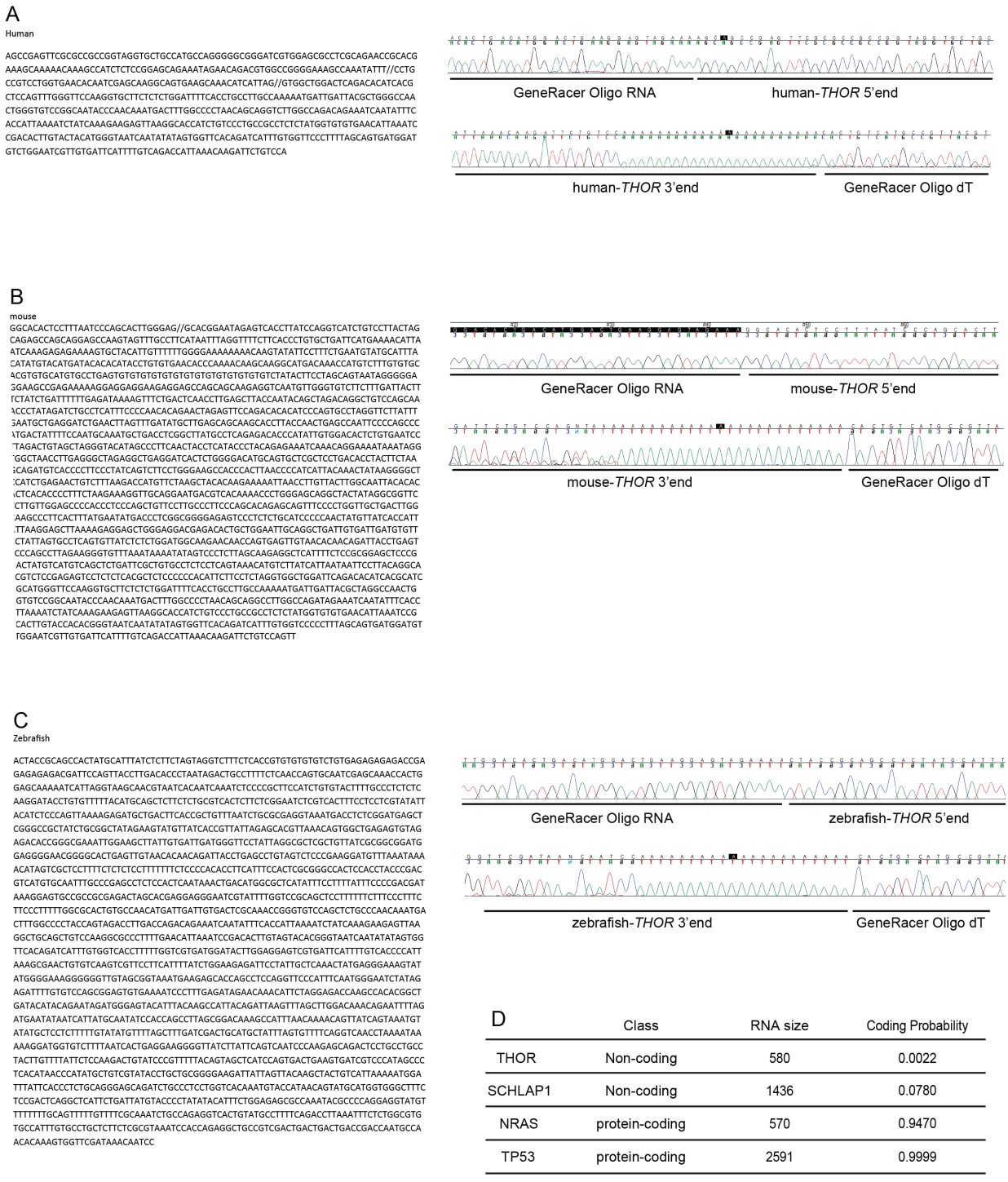


Figure D.2 Genomic organization of the long non-coding RNA *THOR*.
UCSC genomic browser view of *THOR* represented in the UCSC browser for (**a**) human GRCh37, (**b**) mouse GRCm38, and (**c**) zebrafish Zv9. *THOR* structure depicted along with H3K4me3 histone marks (ENCODE), conservation (Phylop and PhastCons) and Multiz 100 vertebrate alignment.



D

Class	RNA size	Coding Probability
Non-coding	580	0.0022
Non-coding	1436	0.0780
protein-coding	570	0.9470
protein-coding	2591	0.9999

Figure D.3 RACE verified *THOR* sequence.
(a-c) Sequence information (left) and 5′ and 3′ end RACE sequencing chromatogram verification (right) of *THOR* in (**a**) human, (**b**) mouse, and (**c**) zebrafish. **(d)** Coding probability scores for the transcripts were assessed by Coding Potential Assessment

Tool (CPAT). *NRAS* and *TP53* used as positive control, and *SCHLAP1* as a negative control.

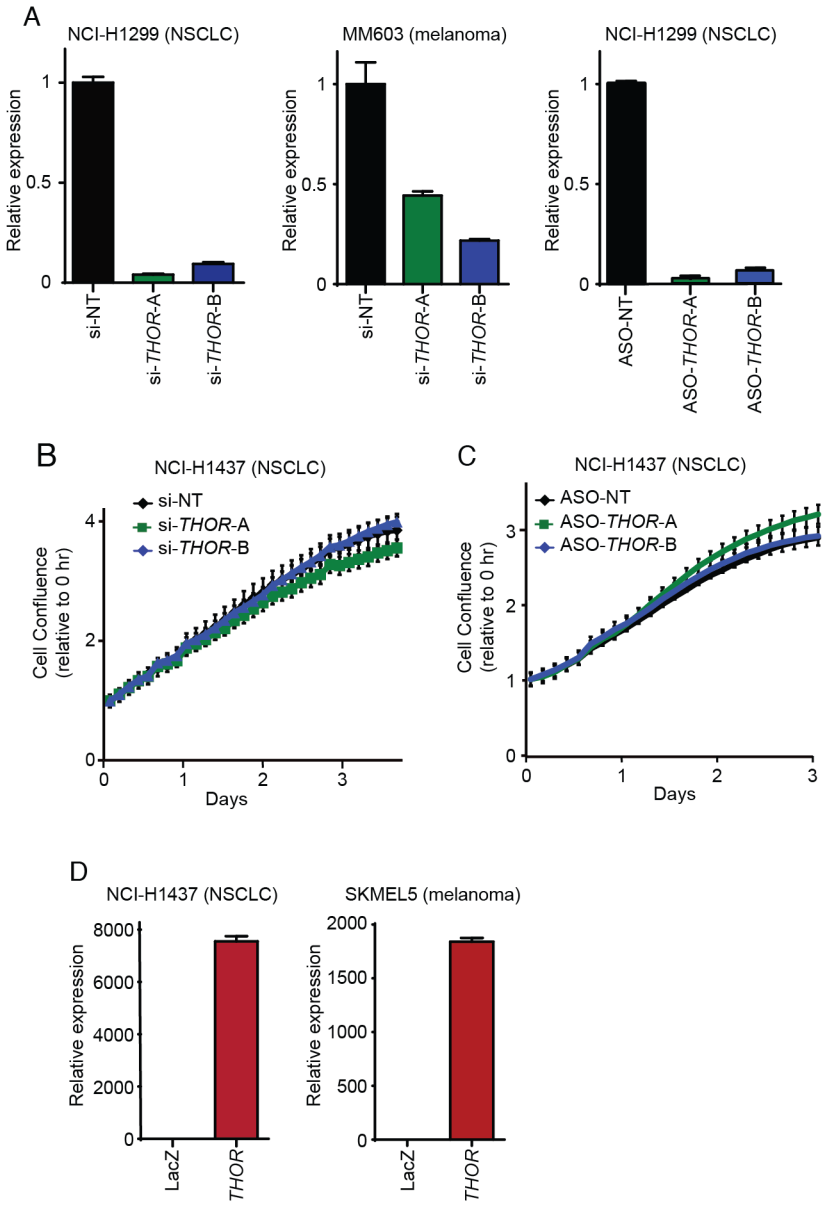


Figure D.4 ***THOR*** **knockdown efficiency and proliferation effects.**
(a) Knockdown efficiency of two independent siRNAs against *THOR* in NCI-H1299 and MM603 cells determined by qRT-PCR. Data show mean ± S.D. **(b)** Cell proliferation assays for NCI-H1437 cells treated with two independent *THOR* siRNAs. Data show mean ± S.E. from one of the two independent experiments. **(c)** Cell proliferation assays for SK-MEL-5 cells treated with two independent *THOR* ASOs. Data show mean ± S.E. from one of the two independent experiments. **(d)** Overexpression efficiency of *THOR* in NCI-H1299 and SK-MEL-5 cells. Data show mean ± S.D.

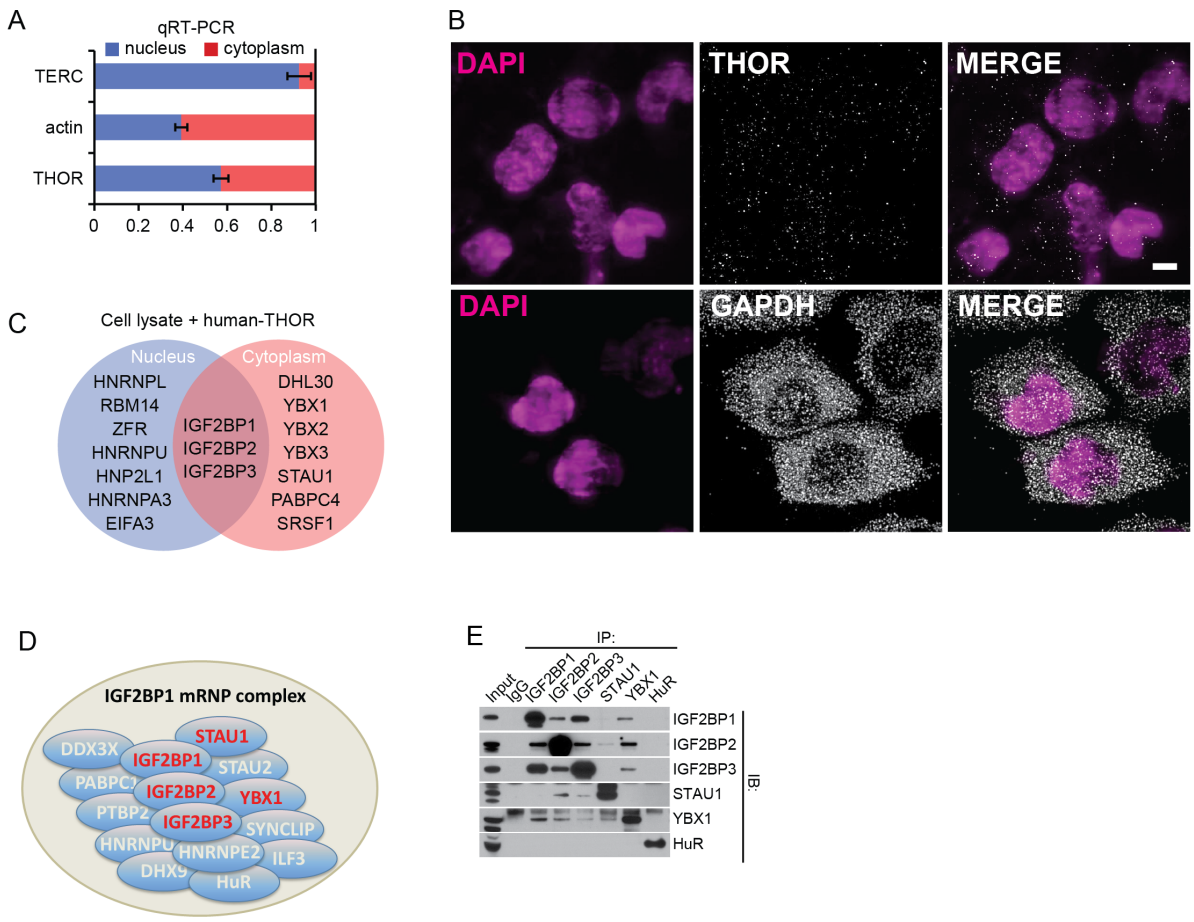


Figure D.5 ***THOR* cellular localization.**
(a) qRT-PCR for *TERC*, *ACTB*, and *THOR* following nuclear and cytoplasmic fractionation of NCI-H1299 cell lysates demonstrates both nuclear and cytoplasmic expression of *THOR*. *TERC* serves as a control for nuclear gene expression and *ATCB* serves as a control for cytoplasmic expression. Error bars represent the standard deviation (s.d.). **(b)** Single molecule RNA in situ hybridization in NCI-H1299 cells. Staining performed for DAPI, *THOR*, and *GAPDH*. **(c)** Venn diagram depiction of the proteins preferentially bound to sense *THOR* (compared to antisense) from nuclear or cytoplasmic lysate from H1299 cells. Protein binding was identified via mass spectrometry following pull-down of BrU-labelled RNA. **(d)** Schematic diagram of the IGF2BP1 binding proteins described in Weidensdorfer et. al.[214]. The proteins which were detected as binding to THOR are highlighted in red. **(e)** Immunoprecipitation western blotting analysis (IP-WB) for various components of the IGFBP complex which contains IGF2BP1, IGF2BP2, IGF2BP3, STAU1 and YBX1. HuR (ELAV1) was used as a negative control.

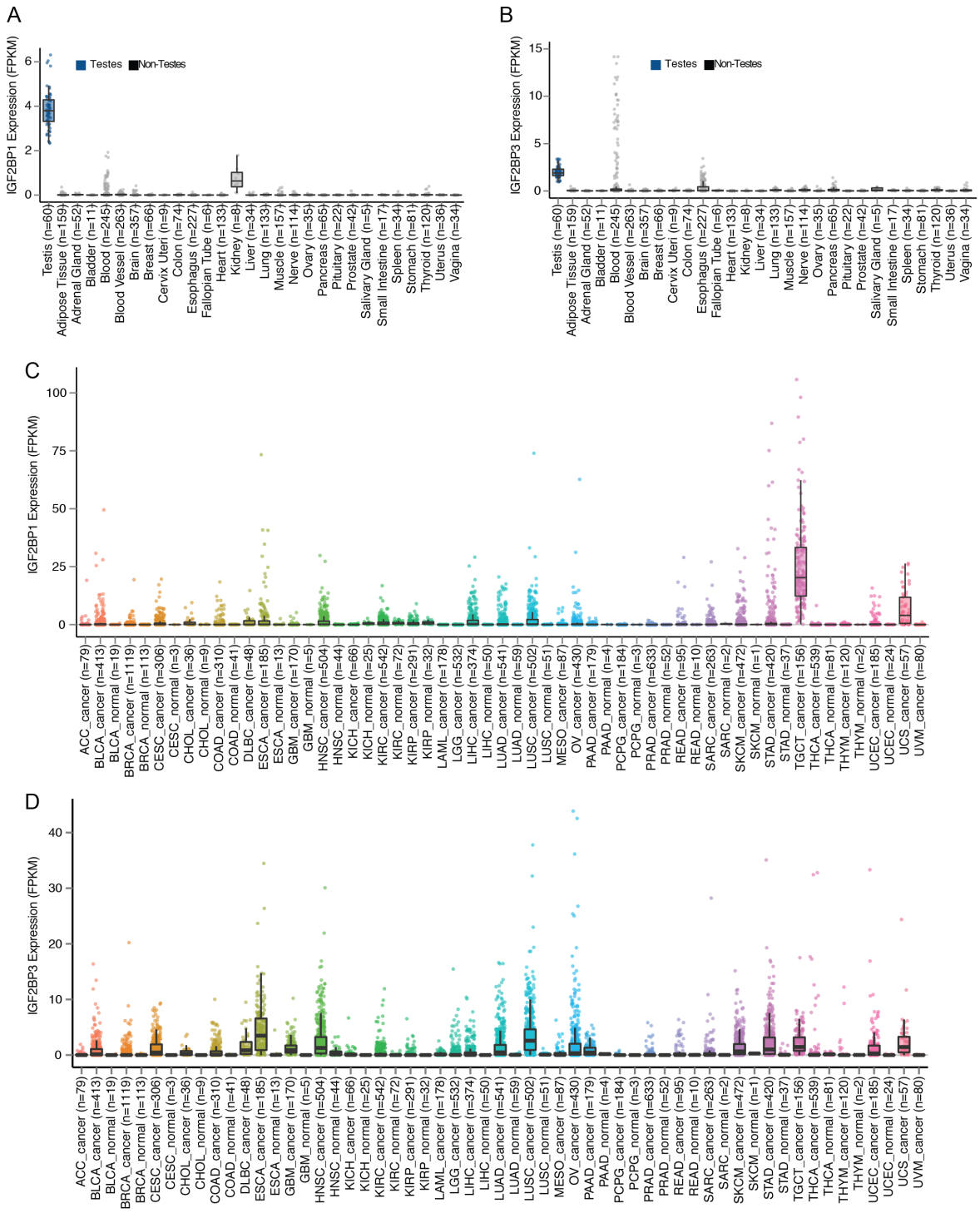


Figure D.6 Expression of *IGF2BP1* and *IGF2BP3* in cancer and normal tissue. **(a,b)** Expression of (**a**) *IGF2BP1* and (**b**) *IGF2BP3* in the GTEX normal tissue RNA-seq dataset, spanning a myriad of different normal tissue types. **(c,d)** Expression of (**c**) *IGF2BP1* and (**d**) *IGF2BP3* across a panel of 9,714 TCGA tumor RNA-seq samples from a myriad of different tissues.

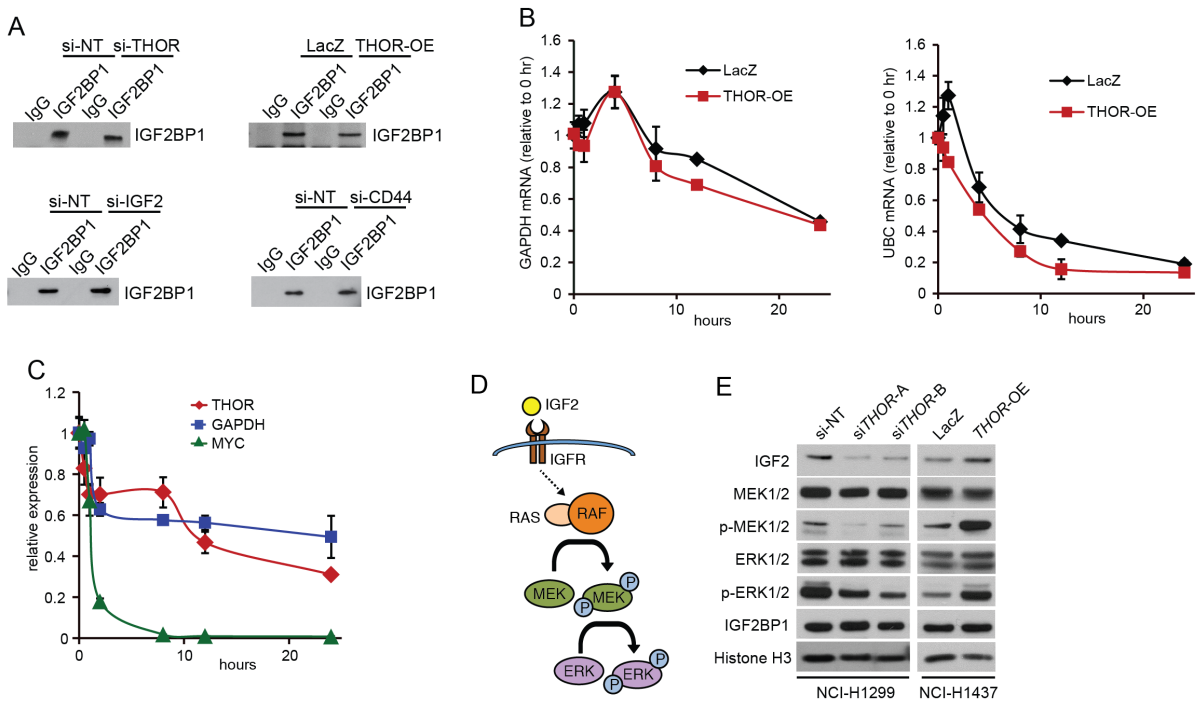


Figure D.7 ***THOR* interaction with *IGF2BP* proteins.**
(a) Western blot confirming IGF2BP1 pull-down utilized for the RIP experiments depicted in **Figure 5.4b**. **(b)** qRT-PCR expression levels for *GAPDH* (left) and *UBC* (right) following Actinomycin D treatment in *THOR* or *LacZ* overexpressing H1347 cells. Data show mean ± S.D. from one of the two independent experiments. **(c)** qRT-PCR expression levels for *THOR* (red), *GAPDH* (blue) and *MYC* (green) following Actinomycin D treatment in H1299 cells. Data show mean ± S.D. from one of the two independent experiments. **(d)** Schematic diagram of the IGF2-MEK-ERK signaling cascade. **(e)** The expression of IGF2-MEK-ERK axis proteins and their corresponding phosphorylated forms in siTHOR treated H1299 cells (left) and *THOR* overexpressing H1437 cells (right).

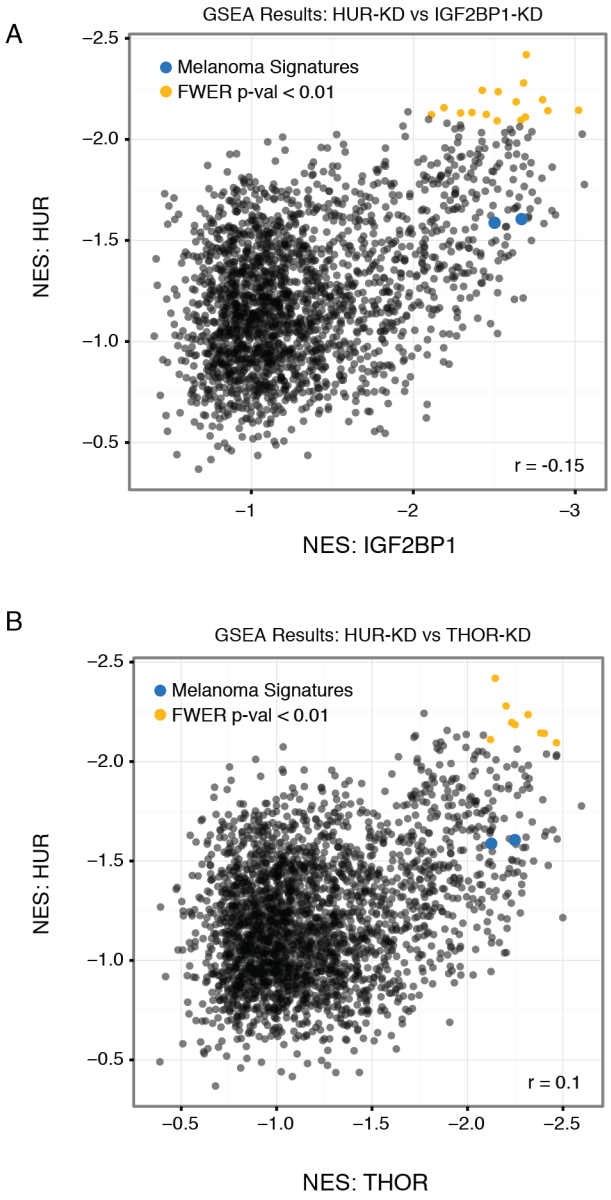


Figure D.8 Comparing GSEA performance for knockdown of *THOR*, *IGF2BP1*, and *HUR*.

(a,b) Scatterplot depicting the GSEA performance for MSigDBv5.0 gene signatures with NES < 0 for (**a**) *HUR* and *IGF2BP1* and (**b**) *HUR* and *THOR*. Signatures significant upon knockdown of both genes (FWER p-value < 0.01) depicted in gold.

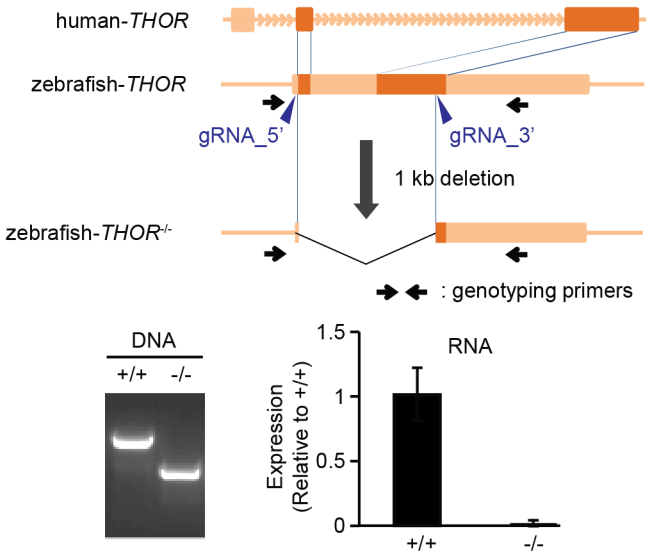


Figure D.9 *THOR* genetic model in zebrafish contributes to melanoma development.

Schematic representation of human and zebrafish *THOR* transcript structure and the guide RNA (gRNA) design used to delete the conserved transcript region of zebrafish *THOR* via CRISPR/Cas9 genome editing (top). Validation of *THOR* knockout at both DNA (genotyping) and RNA (qRT-PCR) level (bottom). Data show mean ± S.D.

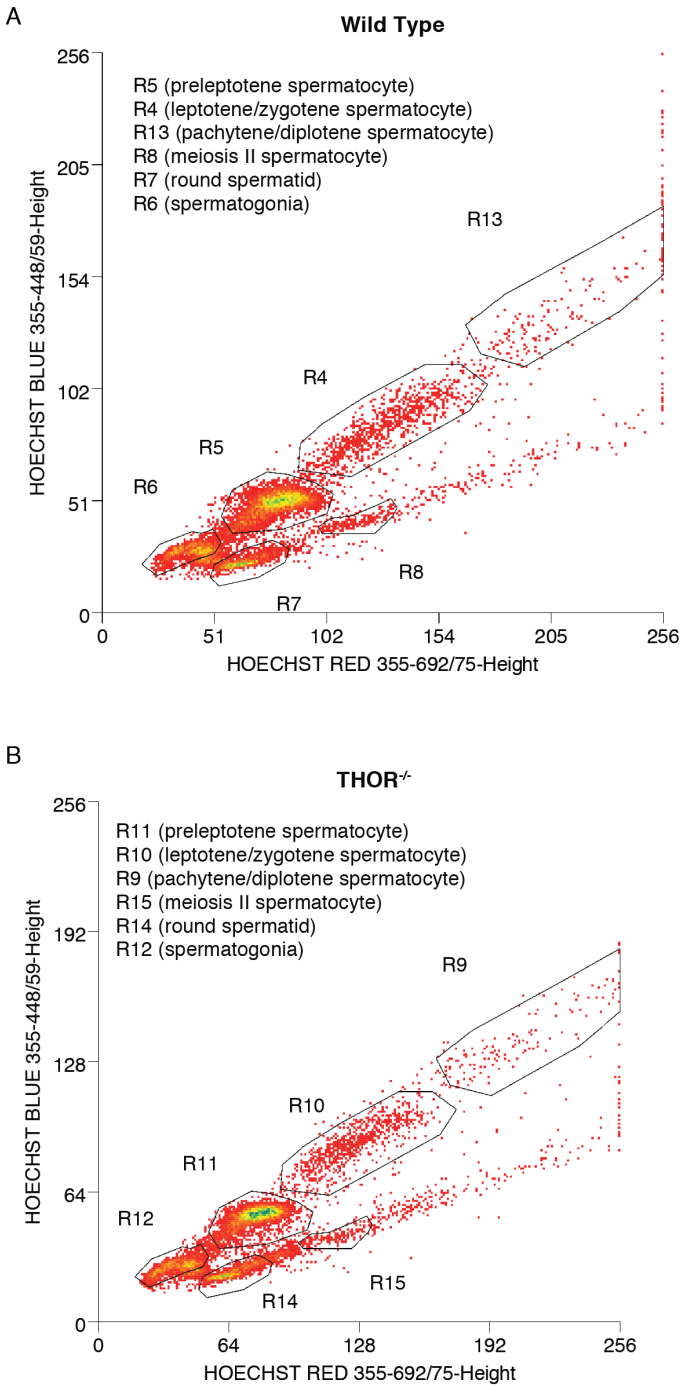


Figure D.10 Interrogation of developmental sperm stages.
Cell selection is visualized in a "Hoechst Blue"/"Hoechst Red" contour plot, in which the density of the cells is displayed as contour lines that form circular contours upon high cell density. Contour plots shown for sorted zebrafish spermatocytes from (**A**) wildtype zebrafish and (**B**) *THOR* knockout zebrafish.

LITERATURE CITED

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66,** 7–30 (2016).
2. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).
3. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153,** 17–37 (2013).
4. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
5. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
6. (Chairperson), T. J. H. *et al.* International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
7. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).
8. Network, T. C. G. A. R. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
9. Network, T. C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
10. Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).
11. Stransky, N. *et al.* The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science (80-.).* **333,** 1157–1160 (2011).
12. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
13. Wang, L. *et al.* SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *N. Engl. J. Med.* **365,** 2497–2506 (2011).
14. Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476,** 298–303 (2011).
15. Jones, S. *et al.* Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science (80-.).* **321,** 1801–1806 (2008).
16. Sjöblom, T. *et al.* The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science (80-.).* **314,** 268–274 (2006).
17. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
18. Network, T. C. G. A. R. Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).
19. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid

Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
20. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).
21. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
22. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
23. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).
24. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45,** 1134–1140 (2013).
25. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6,** 99–103 (2009).
26. Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.* **40,** 722–729 (2008).
27. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
28. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25,** 1915–27 (2011).
29. Iyer, M. K. *et al.* The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nat Genet* **In Press,** (2014).
30. Prensner, J. R. *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.* **29,** 742–749 (2011).
31. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17,** 669–681 (2007).
32. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven symposia in biology* **23,** 366–370 (1972).
33. Services, U. S. D. of H. and H. Understanding Our Genetic Inheritance. *US Hum. genome Proj. first five years. NIH Publ.* (1990).
34. Venter, J. C. *et al.* The sequence of the human genome. *Science (80-.).* **291,** 1304–1351 (2001).
35. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
36. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: Past, present, and future. *Genetics* **193,** 651–669 (2013).
37. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).

38. Ulitsky, I. & Bartel, D. P. LincRNAs: Genomics, evolution, and mechanisms. *Cell* **154,** 26–46 (2013).
39. Prensner, J. R. & Chinnaiyan, A. M. The Emergence of lncRNAs in Cancer Biology. *Cancer Discov.* **1,** 391–407 (2011).
40. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477,** 295–300 (2011).
41. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–50 (2011).
42. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci.* **106,** 11667–11672 (2009).
43. Prensner, J. R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.* **45,** 1392–8 (2013).
44. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419 (2010).
45. Wapinski, O. & Chang, H. Y. Long noncoding RNAs and human disease. *Trends Cell Biol.* **21,** 354–361 (2011).
46. Gupta, R. a *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076 (2010).
47. Pugacheva, E. M. *et al.* Familial cases of point mutations in the XIST promoter reveal a correlation between CTCF binding and pre-emptive choices of X chromosome inactivation. *Hum. Mol. Genet.* **14,** 953–965 (2005).
48. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22,** 1760–1774 (2012).
49. Pruitt, K. D. *et al.* RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res.* **42,** (2014).
50. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2015).
51. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).
52. Mihaela Pertea, J. T. M. S. L. S. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** 290–295 (2015).
53. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).
54. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28,** 1086–1092 (2012).
55. Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10,** 1177–1184 (2013).

56. Hu, X. *et al.* A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell* **26,** 344–57 (2014).
57. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* (2015). doi:10.1016/j.cell.2015.12.017
58. Schmitt, A. M. *et al.* An inducible long noncoding RNA amplifies DNA damage signaling. *Nat. Genet.* **48,** 1370–1376 (2016).
59. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2013,** 1–24 (2013).
60. Niknafs, Y. S. *et al.* The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nat. Commun.* **7,** 12791 (2016).
61. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol. Cell* **44,** 667–678 (2011).
62. Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341,** 1237973 (2013).
63. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159,** 188–199 (2014).
64. McHugh, C. a. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* (2015). doi:10.1038/nature14443
65. Chu, C. *et al.* Systematic Discovery of Xist RNA Binding Proteins. *Cell* 404–416 (2015). doi:10.1016/j.cell.2015.03.025
66. Sahu, A., Singhal, U. & Chinnaiyan, A. M. Long Noncoding RNAs in Cancer: From Function to Translation. *Trends in Cancer* **1,** 93–109 (2016).
67. Bertrand, J.-R. *et al.* Comparison of antisense oligonucleotides and siRNAs in cell culture and in vivo. *Biochem. Biophys. Res. Commun.* **296,** 1000–1004 (2002).
68. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47,** 199–208 (2015).
69. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136,** E359–E386 (2015).
70. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
71. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012).
72. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31,** 46–53 (2013).
73. Pruitt, K. D. *et al.* RefSeq: An update on mammalian reference sequences. *Nucleic Acids Res.* **42,** 756–763 (2014).
74. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** 764–770 (2014).

75. Wang, L. *et al.* CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41,** 1–7 (2013).
76. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42,** D222-30 (2014).
77. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
78. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22,** 1775–1789 (2012).
79. Prensner, J. R. *et al.* PCAT-1, a long noncoding RNA, regulates BRCA2 and controls homologous recombination in cancer. *Cancer Res.* **74,** 1651–1660 (2014).
80. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28,** 503–510 (2010).
81. Consortium, E. P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
82. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
83. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–640 (2014).
84. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science (80-.).* **304,** 1321–1325 (2004).
85. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42,** D1001-6 (2014).
86. 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 (2005).
87. Grasso, C. S. *et al.* The Mutational Landscape of Lethal Castrate Resistant Prostate Cancer. *Nature* **487,** 239–243 (2012).
88. Yu, Y. P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J Clin Oncol* **22,** 2790–2799 (2004).
89. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18,** 11–22 (2010).
90. Gluck, S. *et al.* TP53 genomics predict higher clinical and pathologic tumor response in operable early-stage breast cancer treated with docetaxel-capecitabine +/- trastuzumab. *Breast Cancer Res Treat* **132,** 781–791 (2012).
91. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
92. Rhodes, D. R. *et al.* Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia* **9,** 166 (2007).

93. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Res.* **43,** D1079–D1085 (2015).
94. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* (2012). doi:10.1038/nature10983
95. Chen, D. *et al.* LIFR is a breast cancer metastasis suppressor upstream of the Hippo-YAP pathway and a prognostic marker. *Nat Med* **18,** 1511–1517 (2012).
96. Thomas, G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41,** 579–584 (2009).
97. Stacey, S. N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39,** 865–869 (2007).
98. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45,** 353–61, 361–2 (2013).
99. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42,** 504–507 (2010).
100. Li, J. *et al.* A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res Treat* **126,** 717–727 (2011).
101. Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E. & Mattick, J. S. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* **39,** D146-51 (2011).
102. Volders, P. J. *et al.* LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res* **41,** D246-51 (2013).
103. Park, C., Yu, N., Choi, I., Kim, W. & Lee, S. lncRNAtor: a comprehensive resource for functional investigation of long noncoding RNAs. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu325
104. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet.* **9,** (2013).
105. Zhou, Y. *et al.* Activation of p53 by MEG3 non-coding RNA. *J Biol Chem* **282,** 24731–24742 (2007).
106. Tomlins, S. A. *et al.* Urine TMPRSS2:ERG fusion transcript stratifies prostate cancer risk in men with elevated serum PSA. *Sci Transl Med* **3,** 94ra72 (2011).
107. Fickett, J. W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res* **10,** 5303–5318 (1982).
108. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30,** 918–920 (2012).
109. Ye, J. *et al.* Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13,** 134 (2012).
110. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing

genomic features. *Bioinformatics* **26,** 841–842 (2010).
111. Dimitrieva, S. & Bucher, P. UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res* **41,** D101-9 (2013).
112. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
113. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
114. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18,** 1509–1517 (2008).
115. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).
116. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Meth* **14,** 68–70 (2017).
117. Mercer, T. R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30,** 99–104 (2011).
118. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).
119. Consortium, I. C. G. International network of cancer genome projects. *Nature* **464,** 993–998 (2010).
120. Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
121. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.* **22,** 1698–1710 (2012).
122. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* **11,** 1650–1667 (2016).
123. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–307 (2012).
124. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).
125. Zhu, M. Recall, precision and average precision. *Dep. Stat. Actuar. Sci. …* 1–11 (2004).
126. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31,** 1009–14 (2013).
127. Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33,** 736–742 (2015).

128. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech* **34,** 525–527 (2016).
129. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
130. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5,** 557–572 (2004).
131. Chikhi, R. & Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **30,** 31–37 (2014).
132. Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142,** 409–419 (2010).
133. Cancer Genome Atlas Research, N. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2015).
134. Sun, M., Gadad, S. S., Kim, D. S. & Kraus, W. L. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol. Cell* **59,** 698–711 (2015).
135. Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 10393–10398 (2003).
136. Haque, R. *et al.* Impact of breast cancer subtypes and treatment on survival: An analysis spanning two decades. *Cancer Epidemiol. Biomarkers Prev.* **21,** 1848–1855 (2012).
137. Carroll, J. S. *et al.* Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38,** 1289–1297 (2006).
138. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481,** 389–393 (2012).
139. Robinson, D. R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nat. Genet.* **45,** 1446–51 (2013).
140. Toy, W. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat. Genet.* **45,** 1439–45 (2013).
141. Osborne, C. K. & Schiff, R. Mechanisms of endocrine resistance in breast cancer. *Annu. Rev. Med.* **62,** 233–47 (2011).
142. Qiu, J.-J. *et al.* Expression and clinical significance of estrogen-regulated long non-coding RNAs in estrogen receptor α-positive ovarian cancer progression. *Oncol. Rep.* **31,** 1613–22 (2014).
143. Chakravarty, D. *et al.* The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat Commun* **5,** (2014).
144. Bhan, A. *et al.* Antisense transcript long noncoding RNA (lncRNA) HOTAIR is transcriptionally induced by estradiol. *J. Mol. Biol.* **425,** 3707–3722 (2013).

145. Miano, V. *et al.* Luminal long non-coding RNAs regulated by estrogen receptor alpha in a ligand-independent manner show functional roles in breast cancer. **7,**
146. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
147. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98,** 10869–10874 (2001).
148. Robinson, D. R. *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat. Med.* **17,** 1646–1651 (2011).
149. Rhodes, D. R. *et al.* ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform1. *Neoplasia* **6,** 1–6 (2004).
150. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
151. Zhao, H. *et al.* Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell* **15,** 2523–36 (2004).
152. Desmedt, C. *et al.* Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.* **14,** 5158–5165 (2008).
153. Ivshina, A. V. *et al.* Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* **66,** 10292–10301 (2006).
154. Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.* **25,** 1239–1246 (2007).
155. Lu, X. *et al.* Predicting features of breast cancer with gene expression patterns. *Breast Cancer Res. Treat.* **108,** 191–201 (2008).
156. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68,** 5405–5413 (2008).
157. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–6 (2002).
158. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365,** 671–9 (2005).
159. Boersma, B. J. *et al.* A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer* **122,** 1324–1332 (2008).
160. Kao, K.-J., Chang, K.-M., Hsu, H.-C. & Huang, A. T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11,** 143 (2011).
161. Symmans, W. F. *et al.* Genomic index of sensitivity to endocrine therapy for breast cancer. *J. Clin. Oncol.* **28,** 4111–4119 (2010).
162. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347,** 1999–2009 (2002).

163. Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11606–11 (2003).
164. 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,** 15545–50 (2005).
165. Kim, J. H., Hahm, B., Kim, Y. K., Choi, M. & Jang, S. K. Protein-protein interaction among hnRNPs shuttling between nucleus and cytoplasm. *J. Mol. Biol.* **298,** 395–405 (2000).
166. Liu, X. & Mertz, J. E. HnRNP L binds a cis-acting RNA sequence element that enables intron- independent gene expression. *Genes Dev.* **9,** 1766–1780 (1995).
167. Hui, J., Stangl, K., Lane, W. S. & Bindereif, A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* **10,** 33–37 (2003).
168. Rothrock, C. R., House, A. E. & Lynch, K. W. HnRNP L represses exon splicing via a regulated exonic splicing silencer. *EMBO J.* **24,** 2792–2802 (2005).
169. Hung, L.-H. *et al.* Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis. *RNA* **14,** 284–296 (2008).
170. Rossbach, O. *et al.* Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biol.* **11,** 146–55 (2014).
171. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).
172. Mouridsen, H. *et al.* Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. *N. Engl. J. Med.* **361,** 766–776 (2009).
173. Winer, E. P. *et al.* American Society of Clinical Oncology technology assessment on the use of aromatase inhibitors as adjuvant therapy for postmenopausal women with hormone receptor-positive breast cancer: status report 2004. *J. Clin. Oncol.* **23,** 619–29 (2005).
174. Xue, X. *et al.* LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. *Oncogene* (2015).
175. Godinho, M., Meijer, D., Setyono-Han, B., Dorssers, L. C. J. & Van Agthoven, T. Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells. *J. Cell. Physiol.* **226,** 1741–1749 (2011).
176. Poliakov, A. *et al.* Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Mol. Cell. Proteomics* **10,** M110.007039 (2011).
177. Vizcaíno, J., Deutsch, E. & Wang, R. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. …* **32,** 223–226 (2014).
178. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

179. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).
180. Qin, Z. S. *et al.* HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* **11,** 369 (2010).
181. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–11 (2009).
182. Ma, X.-J. *et al.* Gene expression profiles of human breast cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 5974–5979 (2003).
183. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).
184. Maher, C. a *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12353–8 (2009).
185. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7,** 709–15 (2010).
186. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15,** R17 (2014).
187. Raj, A., van den Bogaard, P., Rifkin, S. a, van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5,** 877–879 (2008).
188. Pitchiaya, S., Androsavich, J. R. & Walter, N. G. Intracellular single molecule microscopy reveals two kinetically distinct pathways for microRNA assembly. *EMBO Rep.* **13,** 709–715 (2012).
189. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15 Spec No,** 17–29 (2006).
190. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
191. Morris, K. V & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15,** 423–37 (2014).
192. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. doi:10.1038/ng.3192
193. St. Laurent, G., Wahlestedt, C. & Kapranov, P. The Landscape of long noncoding RNA classification. *Trends Genet.* **31,** 239–251 (2015).
194. Luke, B. & Lingner, J. TERRA: telomeric repeat-containing RNA. *EMBO J.* **28,** 2503–10 (2009).
195. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–7 (2010).
196. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–8 (2011).
197. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472,** 120–4 (2011).

198. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493,** 231–5 (2013).
199. Pauli, A. *et al.* Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343,** 1248636 (2014).
200. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science (80-.).* **351,** 271–275 (2016).
201. Hudson, R. S. *et al.* Transcription signatures encoded by ultraconserved genomic regions in human prostate cancer. *Mol. Cancer* **12,** 13 (2013).
202. Calin, G. a *et al.* Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12,** 215–29 (2007).
203. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2013,** 1–24 (2013).
204. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **339,** 819–823 (2013).
205. Simpson, A. J. G., Caballero, O. L., Jungbluth, A., Chen, Y.-T. & Old, L. J. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5,** 615–625 (2005).
206. Fratta, E. *et al.* The biology of cancer testis antigens: Putative function, regulation and therapeutic potential. *Mol. Oncol.* **5,** 164–182 (2011).
207. Scanlan, M. J., Gure, A. O., Jungbluth, A. a, Old, L. J. & Chen, Y.-T. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol. Rev.* **188,** 22–32 (2002).
208. Prensner, J. R. & Chinnaiyan, A. M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1,** 391–407 (2011).
209. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Sci.* **348,** 660–665 (2015).
210. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
211. Tapparel, C. *et al.* The TPTE gene family: Cellular expression, subcellular localization and alternative splicing. *Gene* **323,** 189–199 (2003).
212. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci.* **70,** 2657–75 (2013).
213. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–41 (2010).
214. Weidensdorfer, D. *et al.* Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. *RNA* **15,** 104–15 (2009).
215. Hämmerle, M. *et al.* Posttranscriptional destabilization of the liver-specific long noncoding RNA HULC by the IGF2 mRNA-binding protein 1 (IGF2BP1). *Hepatology* **58,** 1703–1712 (2013).

216. Nielsen, J. *et al.* A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol. Cell. Biol.* **19,** 1262–1270 (1999).
217. Livingstone, C. IGF2 and cancer. *Endocr. Relat. Cancer* **20,** 321–339 (2013).
218. El-Shewy, H. M., Lee, M. H., Obeid, L. M., Jaffa, A. a. & Luttrell, L. M. The insulin-like growth factor type 1 and insulin-like growth factor type 2/mannose-6-phosphate receptors independently regulate ERK1/2 activity in HEK293 cells. *J. Biol. Chem.* **282,** 26150–26157 (2007).
219. Kauffmann, A. *et al.* High expression of DNA repair pathways is associated with metastasis in melanoma patients. *Oncogene* **27,** 565–573 (2007).
220. Winnepenninckx, V. *et al.* Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. *JNCI J. Natl. Cancer Inst.* **98,** 472–482 (2006).
221. Giraldez, A. J. *et al.* MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308,** 833–838 (2005).
222. White, R., Rose, K. & Zon, L. Zebrafish cancer: the state of the art and the path forward. *Nature reviews. Cancer* **13,** 624–36 (2013).
223. Lieschke, G. J. & Currie, P. D. Animal models of human disease: zebrafish swim into view. *Nat. Rev. Genet.* **8,** 353–367 (2007).
224. Sánchez-Rivera, F. J. & Jacks, T. Applications of the CRISPR-Cas9 system in cancer biology. *Nat. Rev. Cancer* **15,** 387–395 (2015).
225. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31,** 227–9 (2013).
226. Langenau, D. M. *et al.* Co-injection strategies to modify radiation sensitivity and tumor initiation in transgenic Zebrafish. *Oncogene* **27,** 4242–4248 (2008).
227. Dovey, M., White, R. M. & Zon, L. I. Oncogenic NRAS cooperates with p53 loss to generate melanoma in zebrafish. *Zebrafish* **6,** 397–404 (2009).
228. Lister, J. a *et al.* A conditional zebrafish MITF mutation reveals MITF levels are critical for melanoma promotion vs. regression in vivo. *J. Invest. Dermatol.* **134,** 133–40 (2014).
229. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582,** 1977–1986 (2008).
230. Decker, C. J. & Parker, R. P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. *Cold Spring Harb. Perspect. Biol.* **4,** 1–16 (2012).
231. Nielsen, J., Kristensen, M. A., Willemoës, M., Nielsen, F. C. & Christiansen, J. Sequential dimerization of human zipcode-binding protein IMP1 on RNA: A cooperative mechanism providing RNP stability. *Nucleic Acids Res.* **32,** 4368–4376 (2004).
232. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

233. Mehra, R. *et al.* A novel RNA in situ hybridization assay for the long noncoding RNA SChLAP1 predicts poor clinical outcome after radical prostatectomy in clinically localized prostate cancer. *Neoplasia* **16,** 1121–7 (2014).
234. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
235. Westerfield, M. The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio), 5th Edition. *Univ. Oregon Press. Eugene* (2007).
236. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–6 (2013).
237. Gaysinskaya, V., Soh, I. Y., van der Heijden, G. W. & Bortvin, A. Optimized flow cytometry isolation of murine spermatocytes. *Cytom. Part A J. Int. Soc. Anal. Cytol.* **85,** 556–565 (2014).
238. Kwan, K. M. *et al.* The Tol2kit: a multisite gateway-based construction kit for Tol2 transposon transgenesis constructs. *Dev. Dyn.* **236,** 3088–99 (2007).
239. Khosravi-Far, R. *et al.* Oncogenic Ras activation of Raf/mitogen-activated protein kinase-independent pathways is sufficient to cause tumorigenic transformation. *Mol. Cell. Biol.* **16,** 3923–3933 (1996).
240. Berghmans, S. *et al.* Tp53 Mutant Zebrafish Develop Malignant Peripheral Nerve Sheath Tumors. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 407–412 (2005).
241. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–4 (2014).