Discovering cancer-associated transcripts by RNA sequencing

by

Matthew Kalahasty Iyer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2013

Doctoral Committee:

Professor Arul M. Chinnaiyan, Chair
Assistant Professor James D. Cavalcoli
Assistant Professor John K. Kim
Professor John V. Moran
Professor Gilbert S. Omenn
Assistant Professor Maureen A. Sartor

Dedication

This thesis is dedicated to my brother Kevin Krishnan Iyer (1986-2003) who passed away after a tragic hiking accident. Kevin displayed creativity and innovation in everything he accomplished, including inventing board games, composing and playing music, developing computer programs, and writing stories. Those who appreciated the significance of Kevin's rare genius find it unbearable to imagine the gaping void that his absence leaves, for surely Kevin would have harnessed his brilliance for the betterment of society. Although I do not possess his intellectual gifts, I have overcome my agony by challenging myself to fulfill Kevin's potential. If by some stroke of luck my subsequent devotion to the biomedical sciences leads to new cures or lives saved, then in my eyes Kevin's death will not have been in vain.

Acknowledgements

I am grateful to many individuals and groups for their contributions to my training and personal growth. The University of Michigan Medical Scientist Training Program (MSTP) including Ron Koenig, Ellen Elkin, Hilkka Ketola, and Laurie Koivupalo provided mentorship and financial support that enabled my study. The MSTP staff displayed and enthusiasm and caring that went far beyond their job descriptions, and the support and encouragement of Dr. Koenig has given me added confidence and security. My peers in the MSTP, especially the class beginning in 2007 including Morgan Jones, Jordan Wright, Charlie Kuang, Jason Chua, Joshua Regal, and Mike Mashiba, created a special sense of camaraderie and shared purpose that inspired me as a graduate student. I am grateful to have them as colleagues.

The Chinnaiyan Laboratory and the Michigan Center of Translational Pathology (MCTP) provided me with exceptional training and support. I have appreciated the opportunity to collaborate with biologists in the lab including Dan Robinson, Ram Mani, Irfan Asangani, Saravana "Mohan" Dhanasekaran, Qi Cao, and Chad Brenner. In particular, Dan Robinson consistently dedicated his time to teach me subtleties of high-throughput sequencing and basic concepts of molecular biology. Our long non-coding RNA team, including Wei Chen, Sumin Han, Rohit Malik, Corey Speers, Lalit Patel, Yasuyuki Hosono, Anirban Sahu, John Prensner, and Felix Feng, has been a pleasure to work with. Our bioinformatics team, including Terrence Barrette, Shanker

Kalyanasundaram, Catie Grasso, Mike Quist, Bob Lonigro, Pankaj Vats, Oscar Alejandro Balbin, Brendan Veeneman, Lee Sam, Dan Miller, Scott Begley, Christopher Maher, and Christine Brennan, provided constructive discussions, feedback, and assistance. I especially want to acknowledge Terrence Barrette, who has singlehandedly kept our computational infrastructure running smoothly. Terry has been enthusiastic about helping me with ambitious projects and supported my efforts throughout my time in the lab. Also, Christopher Maher mentored me when I first joined the lab and taught me the fundamentals of high-throughput sequence data analysis. Chris encouraged me to begin studying long non-coding RNAs in prostate cancer and was a key contributor in the early days of the project. Later he gave me the opportunity to work with him to develop and maintain his gene discovery tool, ChimeraScan, for public use. He initially served on my dissertation committee until his faculty appointment at Washington University in St. Louis in 2011. I also want to recognize the support I received from the MCTP staff, including Xiaojun Jing who worked tirelessly to generate microarray datasets for us, Javed Siddiqi who coordinated our use of clinical samples and data, Dianna Banka who bent over backward to help me with my predoctoral grant application which was successfully funded, Karen Giles for coordinating manuscript submissions and requests for external datasets, and Christine Betts for helping me stay organized and for countless other favors. Finally, I want to expressly thank our lab manager Xuhong Cao. She has provided encouragement whenever I faced difficult challenges and has helped me overcome adversity. Apart from the Chinnaiyan Lab, the Department of Defense (BC100238) is funding me to study breast cancer genetics.

I am especially grateful to a few close colleagues that have become lifelong friends. John Prensner, also an MSTP student and member of the Chinnaiyan Lab, rapidly became a close collaborator after I joined the lab. His indelible enthusiasm for science and willingness to take on risky projects enabled my success. John tirelessly worked on our shared projects, contributed to writing manuscripts, shared his exhaustive knowledge of the published literature, and pushed me to achieve more than I possibly could have on my own. Our synergistic collaboration will remain proof of the potential for team science approaches to carry scientific endeavors to new heights. Anirban Sahu, an MSTP student who joined the Chinnaiyan Lab, has also become a great friend and collaborator. Our shared passion for laughing at anything and everything makes life more fun. It is also fun to consistently beat him in tennis and I look forward to our amassing more victories in our upcoming matches. Sameek Rowchowdhury was a fellow in the lab from 2010-2012 and gave me the opportunity to help him design and implement the MI-ONCOSEQ clinical sequencing protocol. Sameek often shared his perspectives on mentorship, leadership, and the importance of a balanced lifestyle. He helped me overcome several challenges and displayed a remarkable degree of unselfishness and caring. Sameek has great leadership potential and I wish him well in his new faculty position at Ohio State University. This work would not be possible without the support of numerous other friends, including Scott Deroo, Nadia Sebastian, Aparna Ghosh, the Auscultations, the Smoker cast of 2012 and 2013, the Georg family, the Rajan family, the Thummalapally and Reddy families, and many more.

I am grateful to the members of my dissertation committee, including Jim Cavalcoli, Maureen Sartor, John Kim, John Moran, and Gil Omenn for their constructive feedback and ideas. They frequently asked difficult questions that I could not answer, but the resulting discussions often led to new ideas and project directions. I thank my committee for their enthusiastic participation and encouragement. Apart from my thesis committee, I received much mentorship and support from Felix Feng. Dr. Feng displayed a passion for studying long non-coding RNAs that led to a number of promising collaborations between our labs. I expect that our continuing collaborations with result in new discoveries together.

Although he leads a large center and the demands on his time are extraordinary, Arul Chinnaiyan has facilitated the growth of our long non-coding RNA project from its fledgling state into a respectably sized team. During this process Arul consistently encouraged me to aim higher and be ambitious. He also provided the ingredients necessary to enable my success, including expensive assays, supercomputing resources, and large cohorts of clinical samples. In particular, Arul's eagerness to adopt new technologies and approaches creates a lab environment with an abundance of data and seemingly endless research opportunities. Although daunting at times, the lab is a computational biologist's personal playground. I will remain forever grateful for the opportunity to work for Arul and hope that our partnership continues.

Lastly, I want to thank my family, including Owen, Scott, Geoffrey, Kristin, and my parents Pam and Hari, for their love and encouragement. My father, a brilliant professor of statistics, assisted me in solving many of the challenging bioinformatics problems that are presented in this thesis. It has been wonderful to work with him and appreciate first-hand his extraordinary mentoring ability.

Preface

The work represented in this dissertation centers around the use of high-throughput sequencing technology to study cancer-associated long ribonucleic acid (RNA) molecules transcribed from the human genome. Therefore, for an introduction I give an overview of the emerging appreciation of the role of RNA in cancer biology, the molecular profiling technologies used to sequence and quantitate RNA levels, and the computational and bioinformatics challenges associated with these technologies. Chapter Two describes the development of ChimeraScan, a software package for detecting cancer-promoting gene fusion events, and its application to studies of breast cancer, solitary fibrous tumors, and clinical sequencing of patients with metastatic cancer. ChimeraScan was published in Bioinformatics in 2011, and the use of ChimeraScan to discover of recurrent families of gene fusions in breast cancers resulted in a mid-author publication in Nature Medicine in 2011. Further, the chapter details the implementation of a personalized clinical sequencing project called MI-ONCOSEQ at the University of Michigan. This resulted in a co-first-author publication in Science Translational Medicine in 2011. Notably, ChimeraScan detected a NAB2-STAT6 gene fusion in a MI-ONCOSEQ patient who suffered from hemangiopericytoma. This gene fusion was found to be highly recurrent in solitary fibrous tumors and the discovery resulted in a mid-author publication in Nature Genetics in 2013.

Chapter Three describes the creation of AssemblyLine, a software package that automatically and accurately annotates the transcriptome from RNA sequencing experiments. The AssemblyLine algorithm established a novel method for filtering sources of contamination from RNA sequencing datasets, as well as a new dynamic programming algorithm for predicting abundant transcript isoforms of a gene. When compared to an existing approach, AssemblyLine produced a more concise and precise transcriptome assembly while still discovering thousands of unannotated long RNAs. This chapter is currently being prepared for submission.

Chapters Four and Five detail the application of AssemblyLine to a prostate cancer RNA sequencing cohort. We identify and characterize thousands of previously undiscovered long RNAs in the human transcriptome, some of which were dysregulated in prostate cancer when compared to benign tissues. In collaboration with experimental biologists John Prensner and Anirban Sahu, we studied the functional roles of two of these RNAs - *PCAT-1* and *SChLAP1* - in greater detail. These studies have resulted in two co-first author publications in Nature Biotechnology in 2011 and Nature Genetics in 2013 (manuscript in press). The dissertation concludes with a discussion of how our findings in prostate cancer can be extended to other diseases, thoughts on the future of RNA sequencing, and a discussion of remaining challenges and opportunities for bioinformatics algorithms to derive useful information from RNA sequencing experiments.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

ASE: allele specific expression
CAGE: Cap Analysis of Gene Expression
cDNA: Complementary DNA
DNA: deoxyribonucleic acid
EST: Expressed Sequence Tag
FDR: False Discovery Rate
HTS: High-throughput Sequencing
lncRNA: long non-coding RNA
miRNA: micro-RNA
mRNA: messenger RNA
ncRNA: non-coding RNA
NGS: Next-Generation Sequencing
PCR: Polymerase Chain Reaction
qPCR: “quantitative” real-time PCR
RACE: Rapid Amplification of Complementary DNA Ends
RNA: ribonucleic acid
RT-PCR: reverse transcription PCR
SNP: single nucleotide polymorphism
SAGE: Serial Analysis of Gene Expression
SNV: single nucleotide variant

Abstract

High-throughput sequencing of poly-adenylated RNA (RNA-Seq) in human cancers shows remarkable potential to identify uncharacterized aspects of tumor biology, including gene fusions with therapeutic significance and novel disease markers such as long non-coding RNA (lncRNA) species. However, the analysis of RNA-Seq data places unprecedented demands upon computational infrastructures and algorithms, requiring novel bioinformatics approaches optimized for accuracy and efficiency. To meet these demands, we present two new open-source software packages - ChimeraScan and AssemblyLine - designed to detect gene fusion events and novel lncRNAs, respectively.

RNA-Seq studies utilizing ChimeraScan, an exquisitely sensitive tool in head-to-head comparisons with similar bioinformatics programs, led to groundbreaking discoveries of new families of recurrent gene fusions in breast cancers and solitary fibrous tumors. Further, ChimeraScan was one of the key components of the repertoire of computational tools utilized in data analysis for MI-ONCOSEQ, a clinical sequencing initiative to identify potentially informative and actionable mutations in cancer patients' tumors in a clinically relevant time frame.

AssemblyLine, by contrast, is a novel algorithm that reassembles RNA sequencing data into full-length transcripts *ab initio*. Head-to-head analyses showed that AssemblyLine compared favorably to existing *ab initio* approaches, and the application of AssemblyLine to human tissues and cell lines unveiled abundant novel lncRNAs,

including antisense and intronic lncRNAs disregarded by previous studies. Moreover, in a first-of-its-kind study, we used AssemblyLine to define the prostate cancer transcriptome from a large patient cohort and discovered myriad lncRNAs, including over a hundred prostate cancer-associated transcripts (PCATs) that could potentially serve as novel disease markers. In-depth functional studies of two PCATs - *PCAT-1* and *SChLAP1* - revealed cancer-promoting roles for these lncRNAs. *PCAT1*, a multi-exonic lncRNA expressed in a 'gene desert' on chromosome 8q24, promotes cell proliferation through transcriptional regulation of target genes and represses the tumor suppressor *BRCA2*. *SChLAP1*, one of several multi-exonic lncRNAs expressed in a chromosome 2q31 'gene desert', independently predicts poor patient outcomes, including metastasis and cancer-specific mortality. Mechanistically, *SChLAP1* antagonizes the genome-wide localization and regulatory functions of the SWI/SNF chromatin-modifying complex.

Collectively, this work demonstrates the utility of ChimeraScan and AssemblyLine as powerful open-source bioinformatics tools. Our applications of ChimeraScan and AssemblyLine led to the discovery of new classes of recurrent and clinically informative gene fusions, and established a prominent role for lncRNAs in coordinating aggressive prostate cancer, respectively. We expect that the methods and findings described herein will establish a precedent for RNA-Seq-based studies in cancer biology and assist the research community at large in making similar discoveries.

Chapter 1: Introduction

The central role of RNA in cellular biology

The transcriptome is defined as the complete set of ribonucleic acid (RNA) transcripts produced by a cell in a given developmental stage or condition[1]. The transcriptome forms a layer of complexity that links the genome (the complete set DNA sequences in a cell) and the proteome (the complete set of proteins present in a cell) through messenger RNAs (mRNAs). The expansive role of RNA in cellular biology far exceeds its role in transferring messages from the nucleus to the ribosome. The many classes of non-coding transcripts carry out myriad biological functions[2], supporting the idea that RNA preceded DNA in the origin of life[3]. Therefore, furthering our understanding of the transcriptome promises to lend unique insights into the mechanisms of development and disease.

Methods for profiling the transcriptome

Technologies for RNA analysis aim to achieve quantitative measurements of RNA levels or underlying sequence information. Microarray technologies assess the amount of hybridization to predefined antisense oligonucleotide probes and provide a high-throughput method for quantitating gene expression levels[4]. Microarray technology can be applied in an unbiased fashion by designing sets of probes that tile large genomic regions[5]. Although it has important limitations, the approach established the pervasive nature of non-coding transcription in areas of the genome originally thought to be devoid of genes[6, 7]. By contrast, capillary sequencing based technologies (also known as Sanger sequencing) were used to sequence thousands of randomly selected complementary DNA (cDNA) clones as expressed sequence tags (EST)[8] and led to the discovery of thousands of new genes[9]. Also, serial analysis of gene expression (SAGE)[10] and cap analysis of gene expression (CAGE)[11] generated sequence information from short (<25bp) tags to quantitate abundance at the 3' and 5' ends of transcripts, respectively. CAGE sequencing efforts led to the identification of transcriptional start sites throughout the genome and constitute a key technology used by the ENCODE project[12, 13].

The emergence of high-throughput sequencing

Protocols for assaying the transcriptome were initially limited by the cost and relatively low throughput of Sanger sequencing[14, 15]. In the 2000s a new breed of technologies known as high-throughput sequencing (HTS) (also referred to as next-generation sequencing (NGS) or massively parallel sequencing (MPS)) emerged with promises of vastly higher throughput[16]. The first of these approaches, the 454 platform, achieved 7.4-fold coverage of a human genome in two months[17], surpassing the Human Genome Project in dramatic fashion. The 454 platform was rapidly followed by competing technologies from Applied Biosystems (ABI)/SOLID, Solexa/Illumina, and Helicos[18, 19]. These technologies offered various advantages and disadvantages that led to significant competition within the industry[19].

High-throughput RNA sequencing offers an unprecedented view of the transcriptome

Initial demonstrations of high-throughput sequencing technologies used human genome sequencing as a benchmark[17, 19], but the technology was rapidly adapted to assess RNA[20, 21]. This method, called RNA-Seq, captured poly-adenylated long RNA, fragmented the RNA to an average length of 200nt by magnesium-catalyzed hydrolysis, and then converted the product to cDNA by random priming and reverse transcription. The cDNA could then by prepared into a library for sequencing using existing protocols[22]. RNA-Seq displayed a number of distinct advantages over previous transcriptome profiling techniques[21]: (1) concurrent sequencing and quantification of abundance levels in a single experiment, (2) extraordinary dynamic range allowing accurate measurement of a wide-range of expression levels, (3) unbiased profiling that captures previously uncharacterized transcripts, and (4) relatively low background noise[1]. Applications of RNA-Seq in mammals established a robust experimental protocol and defined fundamental steps required for data analysis, including the Reads Per Kilobase per Million (RPKM) metric for transcript abundance levels[22]. The approach matured further due to a number of key developments[23]: (1) the ability to sequence both ends of a long DNA fragment (known as paired-end sequencing)[24], (2) the creation of libraries that preserve RNA strandedness (also known as strand-specific RNA-Seq)[25], (3) the capability to produce longer reads up to and exceeding 100nt[26], and (4) continued increases in throughput.

Although RNA-Seq is just a few years old, it has already deepened our understanding of alternative splicing, uncovered functional relationships between DNA and RNA, detected gene fusions in cancer and identified myriad noncoding RNAs[27-30]. Collectively, these studies reveal an enormously complex eukaryotic transcriptional landscape that might give pause to even the most committed supporters of molecular biology's central dogma. Most of these insights were generated using older sequencing instruments that yield orders-of-magnitude fewer reads and substantially shorter read lengths compared with current technology. Because of these technical limitations, the first RNA-Seq analysis methods generally relied on existing Sanger-sequenced reference genomes as a foundation for probing the transcriptome.

RNA-Seq poses computational challenges

The swift increase in sequence data generation has placed escalating demands upon computational platforms. In fact, the rate of sequence data growth continues to exceed Moore's law, leading to a growing worry that resources such as disk storage, processing power, and data transfer time may become the bottleneck for genomics research[31]. To mitigate this worry, increased emphasis is being placed on efficient algorithm design, and a new collections of bioinformatics tools specifically optimized for high-throughput sequencing have appeared[32]. The development of accurate and efficient bioinformatics algorithms for high-throughput sequencing analysis continues to be an area of great interest.

Algorithms for RNA-Seq data analysis

RNA-Seq is an information rich modality capable of interrogating many aspects of genome biology simultaneously, including gene, isoform, and allele-specific expression levels[26, 33-38], changes in transcripts levels across conditions[33, 34, 39-41], gene fusions[24, 30], single nucleotide variants and short indels[42-44], and pathogens such as bacteria and viruses[45]. Perhaps the most promising aspect of RNA sequencing is the ability to delineate the entire set of transcriptional aberrations in a disease, including novel transcripts and long non-coding RNAs (lncRNAs) not measured by conventional analyses[26, 45-51]. To facilitate interpretation of sequence read data, existing computational methods typically process individual samples using either spliced read alignment[52-60] followed by *ab initio* reconstruction[26, 47, 61-66] or *de novo* assembly[46, 50, 67-69] of read sequences followed by genome alignment[70].

The intricate process of computationally assembling sequence reads into full-length transcripts remains an open problem. With improvements in RNA sequencing throughput, algorithm design has become more flexible. In 2010, mammalian transcriptomes were automatically reconstructed for the first time using two software packages, Cufflinks and Scripture, which require reference genomes (referred to as *ab initio* transcriptome assembly)[48]. Several months later, mouse transcripts were assembled without a reference genome using software called Trans-ABySS[50, 61]. More recently, an algorithm called Trinity was introduced that makes it possible to assemble a complete transcriptome in the absence of a reference genome (referred to as *de novo* transcriptome assembly)[46]. The method should prove especially useful for the study of cells with highly rearranged genomes such as cancer cells.

Annotating gene models using RNA-Seq data

Until recently, efforts to annotate gene models used EST sequences as the chief source of data[9]. Millions of ESTs have been assimilated into gene models by a combination of manual and automated efforts such as VEGA, HAVANA, GENCODE, Ensembl, AceView, and RefSeq[71-77]. However, recent reports of thousands of non-coding genes missing from these databases suggest that gene catalogs are far from complete[13, 27, 78-81]. Given the limited bandwidth of manual annotation services, new emphasis has been placed on developing automated transcriptome assembly methods to handle the anticipated growth in RNA-Seq data. Several algorithms are readily available to assist with gene annotation from RNA-Seq datasets[41, 71, 75, 82, 83], and reference databases for model organisms including *Danio rerio* (zebrafish), *Drosophila melanogaster* (fly), and *Caenorhabditis elegans* (worm) have successfully incorporated aspects of RNA-Seq data into gene models[82, 84-86]. Initial efforts to include RNA-Seq in human gene annotations have been encouraging, but discrepancies among the published gene catalogs suggest that the annotation efforts remain incomplete[13, 75, 78, 87, 88].

RNA sequencing to study cancer

Cancers are the second leading cause of death in the United States after heart disease, with over 580,000 deaths related to cancers projected to occur in 2013[89]. Cancer cells arise due to the accumulation of genetic changes that occur in the DNA sequence that drive or permit uncontrolled cell growth[90]. Discovering and characterizing these genetic changes has led to new clinical approaches in the diagnosis and management of the disease[91-97]. 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 and human disease. Unlike modalities confined to monitor predefined molecular lesions, RNA-Seq is uniquely capable of capturing the specific functional products of somatic changes in an unbiased manner and thus promises to further our understanding of cancer biology. To date, numerous cancer genomics studies applied RNA-Seq as a modality for detecting gene fusions (see **Chapter 2**)[30, 98-103], discovering long non-coding RNAs (see **Chapters 3-5**)[27, 81, 104], characterizing expressed somatic variants, and monitoring gene expression changes[105-113].

This thesis describes new methods for RNA-Seq data analysis and their application to study cancer biology. In Chapter 2, we present a software package called ChimeraScan for detecting gene fusions and demonstrate its utility across a variety of cancer profiling studies. Chapter 3 discusses a new bioinformatics tool called AssemblyLine that mitigates the problem of background noise in RNA-Seq data and presents an automated algorithm for annotation of gene models from transcriptome assemblies. Chapter 4 shows the successful application of the AssemblyLine algorithm to discover over 1,800 unannotated lncRNAs in prostate cancers, including 121 Prostate Cancer Associated Transcripts (PCATs) that are aberrantly expressed in the disease. In Chapter 5 we present the extensive characterization of two exemplary PCATs and provide evidence that lncRNAs coordinate the development of aggressive prostate cancers. Finally, we conclude in Chapter 6 and offer an optimistic forecast of discoveries to come.

Chapter 2: Detecting gene fusions in RNA sequencing datasets

AUTHORS: Iyer MK, Roychowdhury S, Robinson DR, Wu YM, Kalyana-Sundaram S, Cao X, Shanker S, Ateeq B, Asangani IA, Barrette T, Grasso CS, Lonigro RJ, Quist M, Sam L, Balbin OA, Siddiqui J, Mehra R, Jing X, Giordano TJ, Sabel MS, Kleer CG, Sung YS, Chen CL, Zhang L, Wang R, Su F, Palanisamy N, Natrajan R, Lambros MB, Reis-Filho JS, Everett J, Kunju LP, Navone N, Araujo JC, Troncoso P, Logothetis CJ, Innis JW, Smith DC, Lao CD, Kim SY, Roberts JS, Mosquera JM, Singer S, Schuetze SM, Antonescu CR, Gruber SB, Pienta KJ, Talpaz M, Kumar-Sinha C, Maher CA, Chinnaiyan AM.

AUTHOR CONTRIBUTIONS

ChimeraScan manuscript: C.A.M. developed preliminary versions of ChimeraScan in Perl. M.K.I. developed the published version of ChimeraScan with mentorship and guidance from C.A.M. C.A.M. and M.K.I. carried out testing and comparison studies. M.K.I., C.A.M., and A.M.C. wrote the manuscript.

Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer: D.R.R., C.K.-S. and A.M.C. conceived of the experiments. D.R.R., C.K.-S., Y.-M.W. and X.C. performed transcriptome sequencing. D.R.R., Y.-M.W. and X.C. performed target capture screening and sequencing. S.K.-S., C.A.M. and M.K.I. performed the bioinformatics analysis of high-throughput sequencing data and the nomination of gene fusions. C.S.G., R.J.L. and M.Q. performed bioinformatics analysis of high-throughput sequencing data for the gene expression profiling. C.K.-S., D.R.R. and Y.-M.W. performed the gene fusion validations. S.S. performed the in vitro experiments of MAST. I.A.A. performed the chorioallantoic membrane assays. B.A. performed the xenograft experiments. D.R.R. and Y.-M.W. performed the in vitro experiments of Notch. X.J. performed the microarray experiments. J.S., M.S.S., C.G.K., T.J.G., N.P., R.N., M.B.L. and J.S.R.-F. provided breast cancer tissue samples and the associated clinical annotation. N.P. performed fluorescence in situ hybridization experiments, and R.M. evaluated the fluorescence in situ hybridization results. D.R.R., C.K.-S. and A.M.C. wrote the manuscript, which was reviewed by all authors.

Personalized oncology through integrative high-throughput sequencing: a pilot study: S.R., M.K.I., K.J.P., S.B.G., M.T., and A.M.C. designed the clinical study; S.R. and M.K.I. accrued patients; J.S. and S.R. processed pathologic specimens; L.P.K. completed pathologic evaluation of tumors; D.R.R. and Y.-M.W. prepared DNA and RNA, prepared sequencing libraries, and completed validations; X.C. and Y.-M.W. performed high-throughput sequencing; T.B., computational systems; M.J.Q. and O.A.B., mutation and indel analysis; R.J.L., L.S., and M.J.Q., copy number analysis; S.K.-S., M.K.I., and L.S., rearrangement and gene fusion analysis; R.J.L., M.J.Q., S.K.-S., and M.K.I., gene expression analysis; O.A.B., S.R., and S.B.G., germline genotyping; S.R., M.K.I., O.A.B., S.B.G., and A.M.C., variant stratification; S.B.G. and J.W.I., clinical genetics; J.S.R. and S.Y.K., bioethics; S.R., S.Y.K., and S.B.G., informed consent; N.N., xenograft samples; M.T., S.R., K.J.P., S.B.G., D.C.S., and C.D.L., clinical oncology; L.P.K. and A.M.C., pathology; S.R., M.K.I., and A.M.C. wrote the manuscript, which was reviewed by all authors. D.R.R., X.C., S.R., M.K.I., and A.M.C. provided overall project management and take responsibility for the integrity of the data and the accuracy of the data analysis.

Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing: D.R.R., C.R.A. and A.M.C. conceived the experiments. D.R.R., Y.-M.W. and X.C. performed exome and transcriptome sequencing. S.K.-S. and M.K.I. carried out bioinformatics analysis of high-throughput sequencing data and nomination of gene fusions. R.J.L. carried out bioinformatics analysis of high-throughput sequencing data for gene expression, copy-number and tumor-content determination. Y.-S.S., C.-L.C., D.R.R., Y.-M.W. and F.S. isolated nucleic acids and performed PCR and Sanger sequencing experiments. Y.-M.W. and F.S. carried out gene fusion validations and gene fusion cloning. Y.-M.W., R.W., F.S. and D.R.R. carried out cell-based in vitro experiments and qPCR assays. L.Z. and C.-L.C. performed immunoblot and immunofluorescence experiments on tissue samples. J.S. collected and prepared tissue samples for next-generation sequencing. L.P.K., J.M.M. and C.R.A. provided pathology review. S.M.S. and S.S. provided the case samples and clinical data. S.R., K.J.P., M.T., S.K.-S., R.J.L., J.S., D.R.R., Y.-M.W., X.C. and A.M.C. developed the integrated clinical sequencing protocol. D.R.R., Y.-M.W., C.R.A. and A.M.C. prepared the manuscript, which was reviewed by all authors.

CITATIONS

Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*. 2011 Oct 15;**27**(20):2903-4.

Roychowdhury S*, Iyer MK*, Robinson DR*, Lonigro RJ, Wu YM, Cao X, Kalyana-Sundaram S, Sam L, Balbin OA, Quist MJ, Barrette T, Everett J, Siddiqui J, Kunju LP, Navone N, Araujo JC, Troncoso P, Logothetis CJ, Innis JW, Smith DC, Lao CD, Kim SY, Roberts JS, Gruber SB, Pienta KJ, Talpaz M, Chinnaiyan AM. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med*. 2011 Nov 30;**3**(111):111ra121.
* These authors made equal contributions

Robinson DR, Wu YM, Kalyana-Sundaram S, Cao X, Lonigro RJ, Sung YS, Chen CL, Zhang L, Wang R, Su F, Iyer MK, Roychowdhury S, Siddiqui J, Pienta KJ, Kunju LP, Talpaz M, Mosquera JM, Singer S, Schuetze SM, Antonescu CR, Chinnaiyan AM. Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing. *Nat Genet*. 2013 Feb;**45**(2):180-5.

Robinson DR, Kalyana-Sundaram S, Wu YM, Shankar S, Cao X, Ateeq B, Asangani IA, Iyer M, Maher CA, Grasso CS, Lonigro RJ, Quist M, Siddiqui J, Mehra R, Jing X, Giordano TJ, Sabel MS, Kleer CG, Palanisamy N, Natrajan R, Lambros MB, Reis-Filho JS, Kumar-Sinha C, Chinnaiyan AM. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat Med*. 2011 Nov 20;**17**(12):1646-51.

Gene fusions as a molecular mechanism for carcinogenesis

Cancers are genetic diseases that arise when accumulated mutations and epigenetic alternations lead to uncontrolled cell growth[90]. Many kinds of mutations are known to promote cancer and range from single nucleotide changes to large-scale alterations in chromosome structure[114]. Structural alterations that result in the colocalization of distant gene elements may allow for the production of new chimeric transcripts, or gene fusions[115]. In 1960, Nowell and Hungerford conducted chromosomal assays in patients with chronic myelogenous leukemia (CML) and described a recurrently abnormal "minute chromosome" known as the Philadelphia chromosome[116]. Their discovery was further characterized by modern cytogenetic assays as a reciprocal translocation causing the first part of the *BCR* gene on chromosome 22 to be positioned immediately upstream from the last part of the *ABL* kinase on chromosome 9[117]. The chimeric *BCR-ABL* transcript had produced a fusion protein with constitutively active kinase activity that promoted neoplasia. The discovery of the *BCR-ABL* chimera established a paradigm for how chromosomal aberrations could promote tumorigenesis and spurred systematic efforts to discover and cataloguing of gene fusions in cancers[115, 118]. It is now commonly understood that gene fusions may be found in most malignancies and could account for up to 20% of cancer morbidity[115].

Transcriptome sequencing to discover gene fusions

Molecular cytogenetics technologies such as fluorescence *in situ* hybridization (FISH) and comparative genomic hybridization (CGH) greatly facilitated the discovery of new gene fusion events, but have important limitations[119, 120]. Standard FISH experiments employ fluorophore-labeled DNA probes designed from bacterial artificial chromosomes

(BACs) approximately 100kb in length, and may lack sufficient resolution to capture focal intrachromosomal aberrations such as inversions. Furthermore, FISH relies on predesigned probes and may be better suited for validation or diagnostic purposes rather than discovery. By contrast, sequencing-based approaches provide an unbiased means for observing chromosomal aberrations and chimeric transcripts at single-nucleotide resolution[121, 122]. In particular, high-throughput RNA sequencing (RNA-Seq) captures the expressed chimeric transcripts emanating from aberrant genomes and helps distinguish 'driver' gene fusion events that potentially contribute to cancer causation from 'passenger' events with irrelevant functional consequences[24, 30]. Thus, RNA-Seq serves as a powerful tool that complements molecular cytogenetics and DNA-based sequencing approaches for gene fusion discovery.

Available algorithms for detecting gene fusions in RNA-Seq data

Interpreting the vast quantities of data from high-throughput RNA-Seq depends upon computational algorithms that can accurately map sequences emanating from chimeric transcripts across the fusion boundary. This has led to the development of a fleet of bioinformatics tools that attempt to predict true gene fusions while discounting sources of artifacts and noise[52, 123-134]. Notably, the FusionSeq package enabled the discovery of novel gene fusions in melanoma[105], the deFuse tool uncovered a recurrent class of gene fusion partners in lymphoid cancers[100], and the unpublished tool GSTRUCT-fusions discovered recurrent R-spondin fusions in colon cancers[99]. Moreover, a preliminary version of ChimeraScan, the tool developed as part of this thesis work, identified recurrent rearrangements of the RAF kinase pathway in multiple cancer types[101]. Although these software tools vary in their underlying architecture, most share a common workflow whereby (1) all sequences are compared to a set of genomic and/or transcriptomic references, (2) sequences that do not support chimeras are designated concordant and not analyzed further, (3) the set of putative chimeric sequences are searched for evidence that they span the junction between disparate genes, (4) multiple sequences supporting the same chimeric transcript are grouped together, (5) the set of chimeric transcripts are further prioritized and/or filtered, and (6) a final set of chimeric transcripts are reported (**Figure 2.1**). Variation within this algorithmic framework typically involves (1) the underlying alignment approach used to compare sequences to a set of references, (2) support for single-end or paired-end sequences, (3) the approach for mapping fusion reads across the chimeric junction, (4) handling of ambiguously mapping sequences, (5) filtering steps employed to reduce sources of error, (6) efficiency and computational performance, and (7) information provided as summary reports.

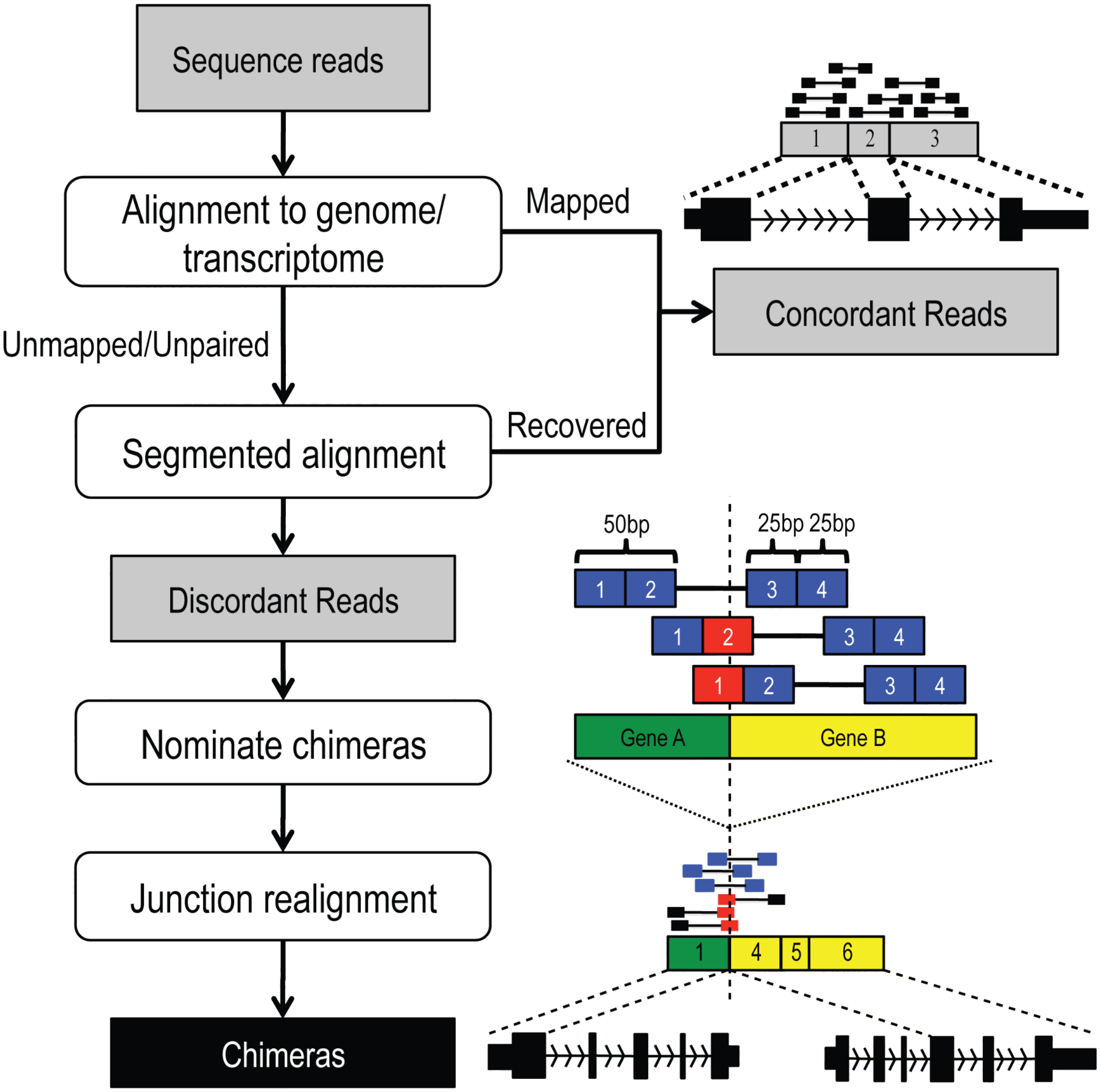


Figure 2.1: Gene fusion discovery workflow.
Paired-end reads failing an initial alignment step are trimmed and realigned to detect discordant reads. Discordant reads that pass filtering criteria are realigned across putative chimeric junctions. Chimeras with encompassing (blue) and spanning (red) fragments may be detected during realignment.

ChimeraScan: a tool for identifying chimeric transcription in sequencing data

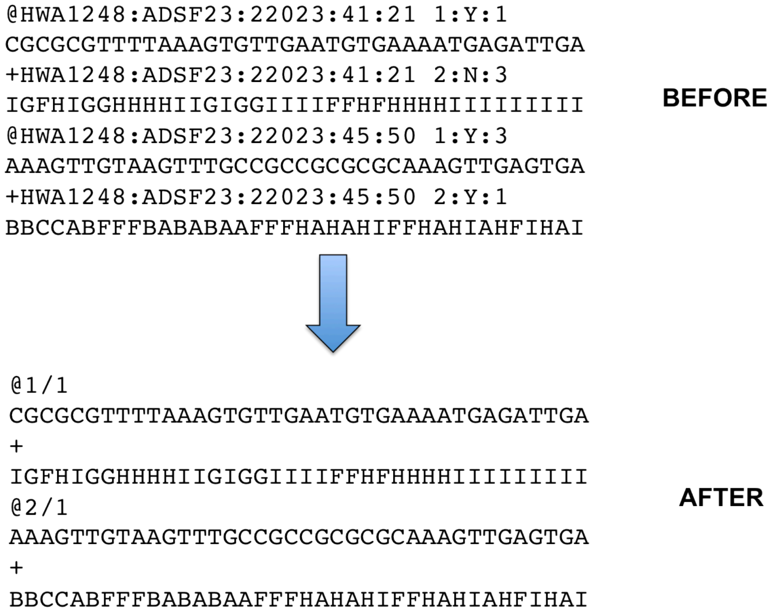
Given the promise of RNA-Seq to uncover recurrent classes of clinically relevant gene fusions, we developed a software tool called ChimeraScan to offer as an open-source package for the community to utilize[124]. The development of ChimeraScan emerged from the initial studies by Maher *et al*. that established the utility of RNA-Seq for gene fusion discovery[24, 30], and proceeded largely in parallel with comparable tools from other labs[52, 125, 128-130, 132, 134]. The software, documentation, and user's guide are hosted online (http://chimerascan.googlecode.com). At the time of this writing the tool was downloaded 436 times and was cited by 10 studies, 8 of which were transcriptome profiling studies that employed ChimeraScan for making novel discoveries, and 2 of which were other bioinformatics methods studies. Furthermore, the commercial software GeneSifter from GeoSpiza incorporates ChimeraScan into its RNA-Seq analysis package. These successful implementations of ChimeraScan are a marker of its impact to the field of cancer genomics, and we believe the tool will continue to be useful in the future.

A guide to the ChimeraScan algorithm

ChimeraScan broadly implements the steps in the generic fusion discovery workflow (**Figure 2.1**). It was engineered for processing paired-end reads and employs the established and markedly efficient Bowtie aligner[135]. Here, we describe ChimeraScan workflow as a series of steps. Prior to running ChimeraScan, users must successfully download the software and construct a set of genomic and transcriptomic references. Step-by-step commands for installing ChimeraScan are described online (https://code.google.com/p/chimerascan/wiki/Installation). The included python script called *chimerascan_index.py* constructs the underlying references files needed to run the Bowtie aligner[135]. It expects genomic sequences (FASTA) as well as a transcriptome annotation file (GTF/GFF or GenePred). The process takes several hours but need only be run once for a particular organism. Detailed instructions can be found online.

Step 1: Pre-process reads before alignment

ChimeraScan processes FASTQ files containing the original read sequences and performs the following: 1) converts non-standard quality scores to Sanger format (Phred + 33), 2) replaces the arbitrarily long read name field with a single unique integer, and adds the suffix "/1" and "/2" to denote read 1 or read 2, respectively (**Figure 2.2**). This pre-processing step dramatically reduces the storage and memory requirements of subsequent steps.



```
@HWA1248:ADSF23:22023:41:21 1:Y:1
CGCGCGTTTTAAAGTGTTGAATGTGAAAATGAGATTGA
+HWA1248:ADSF23:22023:41:21 2:N:3
IGFHIGGHHHHIIGIGGIIIIFFHFHHHHIIIIIIIII
@HWA1248:ADSF23:22023:45:50 1:Y:3
AAAGTTGTAAGTTTGCCGCCGCGCGCAAAGTTGAGTGA
+HWA1248:ADSF23:22023:45:50 2:Y:1
BBCCABFFFBABABAAFFFHAHAHIFFHAHIAHFIHAI
```

```
@1/1
CGCGCGTTTTAAAGTGTTGAATGTGAAAATGAGATTGA
+
IGFHIGGHHHHIIGIGGIIIIFFHFHHHHIIIIIIIII
@2/1
AAAGTTGTAAGTTTGCCGCCGCGCGCAAAGTTGAGTGA
+
BBCCABFFFBABABAAFFFHAHAHIFFHAHIAHFIHAI
```

Figure 2.2: Conversion of quality scores and read identifiers in FASTQ files.
ChimeraScan standardizes quality scores to the common Sanger (Phred + 33) format, and renames read identifiers to single integers to conserve memory in subsequent steps of the algorithm.

Step 2: Align paired-end reads

In this step ChimeraScan uses Bowtie to search for a valid genomic or transcriptomic alignment for each read in the dataset[135]. This initial alignment is performed in paired-end mode where both reads from the fragment must align within a distance range. The default settings use a fragment length range 0-1000bp. This alignment step aims for maximal sensitivity because more correctly mapped paired-end alignments implies fewer potentially false positive chimeras. Users can modify the following parameters to control the behavior of this step:

1. A number of bases may be trimmed from the 5' or 3' end of all reads (*--trim5* and *--trim3*, default 0). This is recommended if sequence quality scores are low at the ends of reads.
2. The number of mismatches tolerated in alignments (*--mismatches*, default 3)
3. The distance range within which alignments are considered valid (*--min-fragment-length* and *--max-fragment-length*, default 0-100bp).

A sorted, indexed BAM file is created from the valid alignments using the pysam library (http://pysam.googlecode.com) and enables fast lookup of aligned reads by subsequent steps of the workflow[44]. Sequence fragments lacking a valid paired-end alignment are saved separately for further analysis (**Figure 2.3)**.

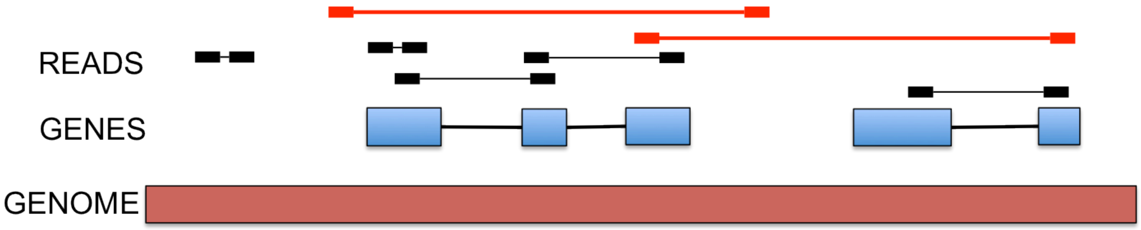


Figure 2.3: Definition of concordant and discordant read pairs.
Concordant reads align as a pair to the genome or a single gene (black). Discordant reads map to independent genes and/or large genomic distances (red).

Step 3: Estimate fragment size distribution

ChimeraScan samples unique paired-end alignments from Step 2 and estimates the empirical distribution of fragment sizes in the library. The fragment size distribution aids in filtering chimeric transcripts in later stages of the workflow (**Figure 2.4**).

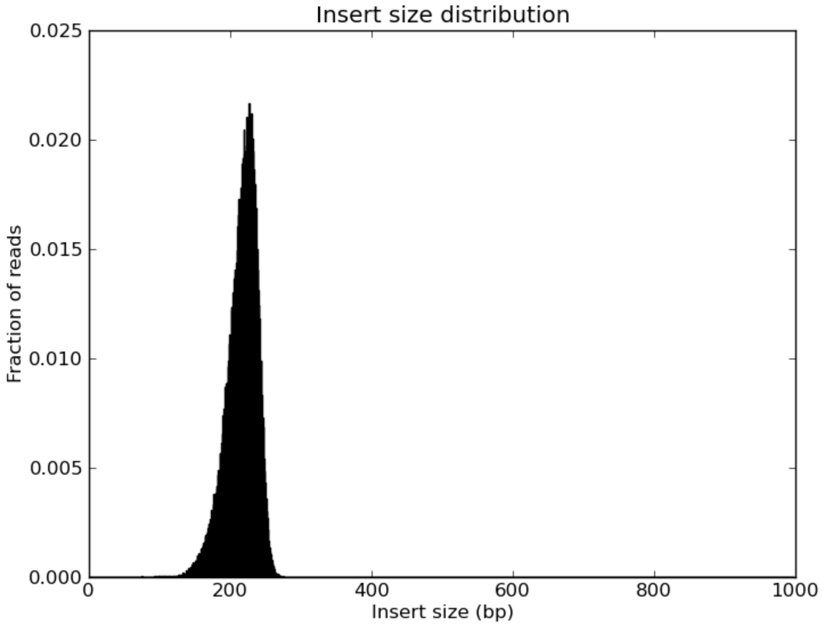


Figure 2.4: Distribution of insert sizes measured by ChimeraScan.
To assess the insert size distribution ChimeraScan considers read pairs that map uniquely to a single transcript isoform. A typical distribution has a mean insert size of 200bp and a standard deviation of 40bp.

Step 4: Realign initially unmapped reads

Read pairs that successfully align are considered concordant reads. They do not support chimeras and need not be processed further. The remaining unmapped fragments could be explained by (1) poor quality sequences with many errors, (2) unannotated collinear transcripts or splicing patterns, (3) foreign sequences such as viruses, (4) "dark matter" missing from the reference genome, or (5) non-collinear chimeric transcripts. To resolve these fragments ChimeraScan realigns them as unpaired reads. Additionally, the reads are trimmed such that only the sequences at the ends of the fragment are aligned. The size of the trimmed segment can be specified by the user using the *--segment-length* option (default 25bp). Setting segment length to a larger number increases the specificity of the mapping process but decreases sensitivity to detect chimeras, since a larger percentage of reads will contain fusion breakpoints and fail alignment. If a library contains relatively small DNA fragments sequenced at long read lengths, then a considerable fraction of the fragments may be sequenced in their entirety with overlapping reads at the center. For example, an average fragment size of 180bp sequenced with 2x100bp reads will generate fragments with an average of 20bp of overlap. To mitigate this overlap it is essential that the segment length be set smaller than half the fragment size (less than 90bp in this example). If not specified by the user, ChimeraScan automatically chooses a segment length that is one-third the size of the fragment length.

Step 5: Discover discordant fragments

A discordant fragment occurs when the two ends of the fragment align to different transcripts. The realigned reads from Step 4 are searched for evidence of discordant pairs. To be considered discordant, both reads in the pair must not align to the same transcript or any of its isoforms. Discordant reads are sorted by reference name and position and stored in an intermediate file format called BEDPE[136].

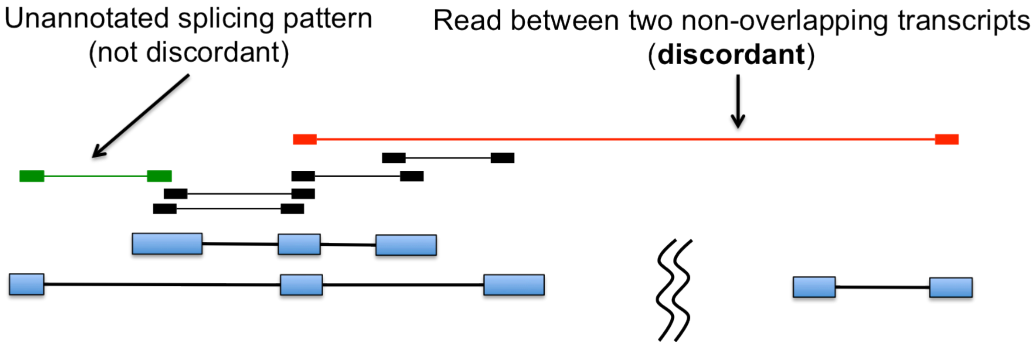

Unannotated splicing pattern
(not discordant)
Read between two non-overlapping transcripts
(discordant)

Figure 2.5: Refinement of discordant fragments
Concordant reads (black) align to a single transcript. Though not strictly concordant, reads that align to different transcript isoforms (green) are not considered discordant. A discordant fragment must align to distinct non-overlapping genes (red).

Step 6: Nominate chimeras

Given that the vast majority of chromosomal aberrations occur in introns or intergenic regions, ChimeraScan only searches for chimeric transcripts that contain fully intact exons. This greatly simplifies the computational complexity of searching for junction spanning reads. In this step the most likely exon boundaries are computed for each discordant fragment using the fragment size distribution information. If a pair of exon boundaries cannot be found within the 99th percentile of fragment sizes, the discordant fragment is discarded as an artifact. Fragments that share the same putative junction boundaries then grouped together.

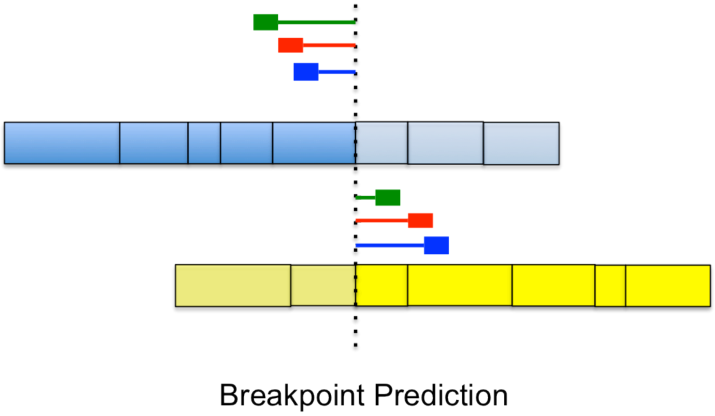


Figure 2.6: Fusion junction prediction.
Discordant read alignments are compared to the fragment size distribution to predict the optimal breakpoint location on each two genes. For each discordant read pair, the most likely pair exon junction is chosen. All putative breakpoints are used in a subsequent realignment phase.

Step 7: Extract chimeric junction sequences

The upstream and downstream sequences surrounding each chimeric junction are extracted from the reference FASTA file, and the *bowtie-build* indexing program is used to create a new alignment index from these junction sequences[135]. In addition, the homology between the 5' and 3' genes at the junction is computed and annotated. Knowledge of the extent of the homology between the two transcripts enables downstream filtering of junction spanning reads.

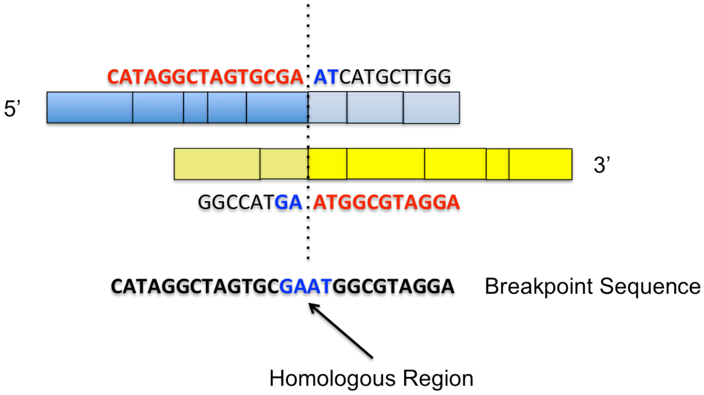


Figure 2.7: Extracting junction spanning sequences.
Combinations of 5' and 3' gene breakpoints are spliced together to form fusion genes *in silico*. The junction sequence is then extracted from the fusion gene. The sequences on either side of the breakpoint are compared to the sequences of the wild-type 5' and 3' genes to determine if homology exists at the junction. Characterizing the upstream and downstream homology aids in filtering junction-spanning reads.

Step 8: Nominate reads that could span chimeric junctions

Two classes of fragments may span chimera junctions: 1) the junction resides in the central portion of the fragment such that both reads in the pair align successfully, and 2) the junction resides on one end of the fragment such that only one of the reads align successfully. Both classes of reads are converted to FASTQ format and aligned to the new junction sequence reference constructed in Step 7.

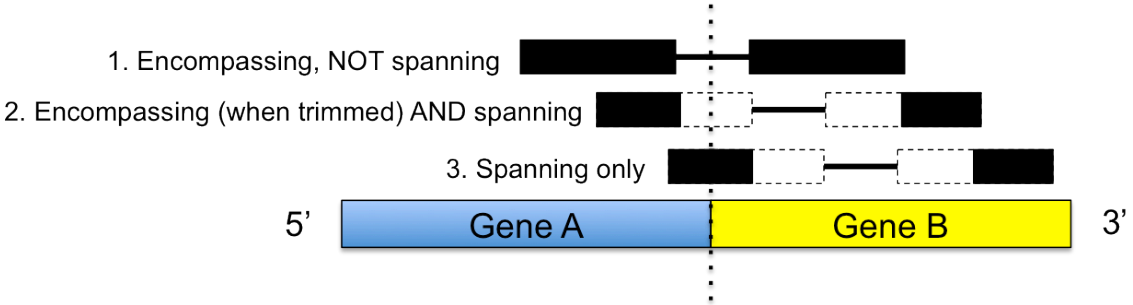

1. Encompassing, NOT spanning
2. Encompassing (when trimmed) AND spanning
3. Spanning only
5'
Gene A
Gene B
3'

Figure 2.8: Classes of junction spanning reads.
Three classes of reads provide support for chimeras: 1) Reads where the unsequenced inner portion of the fragment spans the breakpoint, 2) Reads where the sequenced portion of the fragment spans the junction near to the center of the fragment and 3) Reads where the sequenced portion of the fragment spans the junction near one of the ends of the fragment. Only classes 2 and 3 are nominated for realignment against a breakpoint junction index.

The alignment results are inspected and reads that aligned to the junction reference are considered spanning according to the following criteria:

1. The alignment overlaps the junction by a minimum number of base pairs (specified by –*anchor-min*, default 11bp). The minimum number of anchoring bases must also be greater than the number of bases of homologous sequence at the junction between the two transcripts constituting the chimera.
2. No more than --*anchor-mismatches* mismatches (default: 0) are found within the first --*anchor-min* bases (default: 11bp) of the alignment.

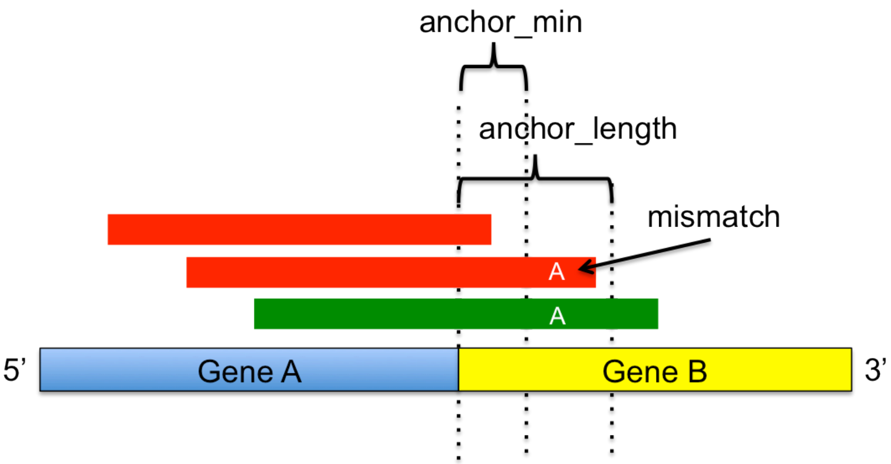


Figure 2.9: Filtering spanning reads in anchor regions.
Reads that align to the junction reference are discarded if the overlap is small (less than *anchor_min* bases) or have larger overlap but contain mismatches (red reads). Reads overlapping the breakpoint by more than *anchor_length* bases are retained (green read).

Step 9: Filter and report chimeras

Junction-spanning alignments from Step 8 are merged with existing chimera information. The chimeras are then categorized and passed through a number of additional filters in order to remove artifacts. These include:

1. Chimeras with low coverage (specified with --*filter-unique-frags*, default 2) that may have arisen from ligation artifacts during library preparation
2. Chimeric transcripts expressed at levels significantly lower than the expression of either of the wild-type alleles in the sample (specified with --*filter-isoform-fraction*, default 0.10).

3. Chimeras that match a list of false positives provided by the user (specified with --*filter-false-pos* as a path to a file containing false positives). A list of false positives was generated from normal human tissue data from the Illumina Bodymap 2.0 project and made available for download on the ChimeraScan website.

Finally, a tab-delimited text file with information about each chimera is generated. Optionally, users may run a companion script (*chimerascan_html_table.py*) to generate an HTML page with results for further investigation.

Evaluation of performance on published data from cell lines

To evaluate the results from ChimeraScan, we applied it to three well characterized cancer cell lines known to harbor multiple chimeric transcripts: VCaP (prostate cancer, 2x53bp)[137], LNCaP (prostate cancer, 2x34bp), and MCF7 (breast cancer, 2x35bp)[138, 139]. Sequence data are deposited in GenBank under the accession number GSE29098. We aligned to human genome version hg19 and the UCSC known transcripts database downloaded in December, 2010, allowing for up to 2 mismatches and no more than 100 alignments per read. The trimmed alignment step was performed with 25bp segments. As our initial benchmark, we confirmed that the current version of ChimeraScan was able to recapitulate the experimentally validated candidates from Maher *et al*., our 'gold standard' (Appendix A)[24]. ChimeraScan detected 9/10, 4/4, and 12/13 chimeras from VCaP, LNCaP, and MCF-7, respectively.

In addition to recapitulating our previous results, we identified additional candidates that demonstrate ChimeraScan's ability to identify and prioritize high-quality chimeras. Overall, we nominated 335 novel chimeras (78 in VCaP, 105 in LNCaP, and

152 in MCF7) from the three cell lines (Appendix B). Interestingly, we detected an inter-chromosomal rearrangement TBL1XR1-RGS17 in the MCF-7 cell line. While our previous version paired-end approach was not able to detect TBL1XR1-RGS17, this fusion event was previously detected by a paired-end diTag approach and experimentally confirmed[140]. Another example of the improved sensitivity was the identification of an intra-chromosomal rearrangement, NDUFAF2-MAST4, in VCaP. Although just two fragments supported the NDUFAF2-MAST4 chimera, one of them was high-quality junction spanning read that uniquely confirmed the fusion junction.

We next compared ChimeraScan with the publicly available tools deFuse[125], ShortFuse[132], and MapSplice[52] using the 10 experimentally validated VCaP chimeras (Appendix C). DeFuse nominated the fewest chimeras, and it only detected 6/10 of the true positives. In comparison, ChimeraScan detection 9/10 of the true positives from 78 predicted chimeras. Of the remaining programs, MapSplice nominated 400 chimeras while detecting 6/10 of the true positives and ShortFuse nominated 245 chimeras while confirming 7/10 of the true positives. Overall, these results suggest that ChimeraScan is both sensitive and relatively specific compared to other recently published methods.

Applications of ChimeraScan

Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancers

Since the discovery of the TMPRSS2-ERG gene fusion in approximately 50% of prostate cancers, emerging evidence has suggested that recurrent gene fusions have a more substantial role in common solid tumors than was previously appreciated[137]. To search for new classes of recurrent gene fusions, we used paired-end transcriptome sequencing on a panel of 89 breast cancer cell lines and tumors and then applied ChimeraScan and another in-house fusion discovery algorithm[124]. Although individual breast cancers harbored a wide variety of gene fusions, we discovered five instances of fusions involving microtubule-associated serine-threonine kinase (MAST) and eight instances of fusions members of the Notch family. Overexpression and knockdown of MAST fusion transcripts in breast cancer cell lines demonstrated that MAST fusions positive and negatively regulated cell proliferation *in vitro*, respectively. Furthermore, cell lines harboring Notch gene rearrangements were uniquely sensitive to treatment with the γ-secretase inhibitor *N*-[(3,5-difluorophenyl)acetyl]-L-al anyl-2-phenyl]glycine-1,1-dimethylethyl ester (DAPT)[141]. These findings indicate that recurrent gene rearrangements have key roles in subsets of breast carcinomas and suggest that transcriptome sequencing could identify individuals with rare but therapeutically targetable gene fusions[102].

Integrative clinical sequencing

Cancers arise from diverse genetic alterations including nucleic acid substitutions, gene fusions and rearrangements, amplifications and deletions, and other aberrations that perturb gene expression[90]. We designed a clinical sequencing strategy called MI-ONCOSEQ that comprehensively identifies informative genomic alterations while remaining cost-effective. We included (i) shallow (5X to 15X) paired-end whole-genome sequencing of the tumor, (ii) targeted exome sequencing of the tumor and matched germline samples (blood or buccal smear), and (iii) paired-end transcriptome sequencing of the tumor (**Figure 2.10**). Whole-genome sequencing can identify copy number alterations (CNAs) and structural rearrangements at relatively shallow depth[142], but accurate point mutation identification requires significantly higher coverage[143]. To fill this niche, we used targeted whole-exome sequencing to capture most human protein-coding exons, including clinically informative and actionable genes in cancer such as *BRAF*, *EGFR*, *JAK2*, *PIK3CA*, and *ALK*[144]. Because tumors are often admixtures with normal tissue or contain multiple tumor clones, the high sequencing depth afforded by exome sequencing was advantageous for the detection of variants. Finally, transcriptome sequencing (RNA-Seq) captured the functional or "expressed" genome of a tumor sample and enabled detection of dysregulated genes and the functional products of genomic alterations[101].

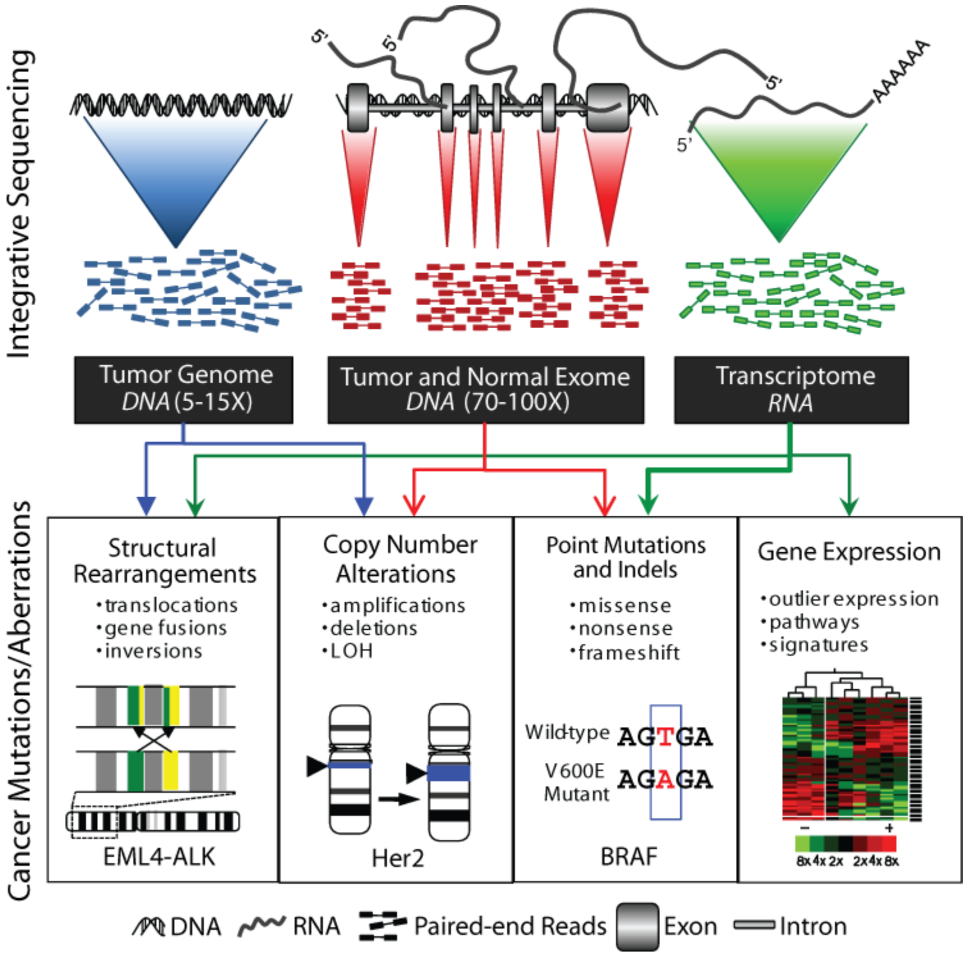


Figure 2.10: Integrative sequencing strategy.
Integration of whole genome sequencing (blue), whole exome capture sequencing for 1-2% of the genome (red), and transcriptome or messenger RNA sequencing (green). Each sequencing strategy can be integrated

(bottom) for analysis of tumor aberrations including structural rearrangements, copy number alteration, point mutations, and gene expression.

We first tested our sequencing strategy on human prostate cancer xenografts, and then applied it to patients with advanced or refractory cancers who were eligible for clinical trials. With this approach, we detected several classes of cancer mutations including copy number alterations, point mutations, and chromosomal aberrations leading to gene fusions[145]. Analysis of the integrative sequencing data relied upon a variety of bioinformatics algorithms (**Figure 2.11**).

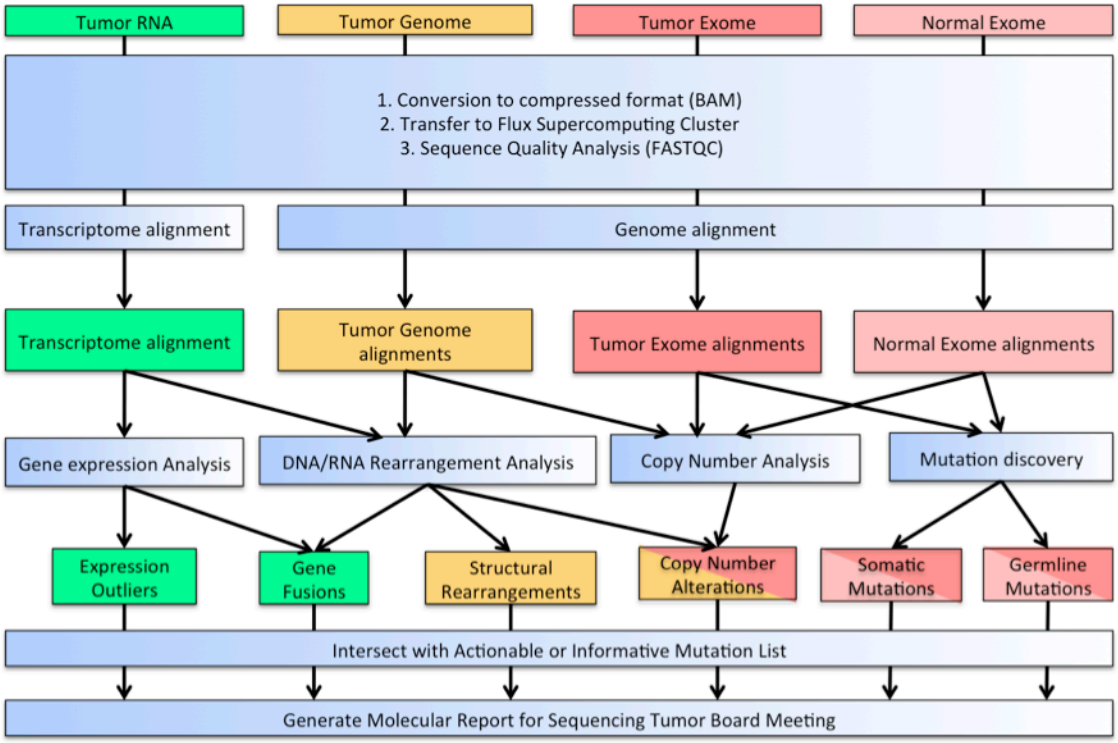


Figure 2.11: Bioinformatics workflow diagram.
DNA and RNA sequences are transferred to a high performance computer cluster and aligned to the human genome using DNA or RNA alignment strategies. Alignment results then feed into multiple analysis pipelines that produce gene expression (RPKM tabulation), gene fusions (ChimeraScan), structural rearrangement (BreakDancer pipeline), copy number alterations (ReadDepth and customized exome copy number assessment), and point mutations (BWA/GATK and Bowtie/in-house pipelines). The results are intersected in a gene-centric manner with a curated list of informative genes. The final results are tabulated for presentation to the Sequencing Tumor Board.

As part of this workflow, ChimeraScan played an essential role in the discovery of gene fusions and uncovered multiple novel fusions in each patient (**Appendix D**).

Patient 1 is a 67-year-old man with castrate-resistant metastatic prostate cancer. A xenograft derived from the patient's tumor harbored the canonical prostate cancer–specific rearrangement of *TMPRSS2* (transmembrane protease, serine 2) and *ERG* (ETS transcription factor) and a novel gene fusion between copine IV (*CPNE4*, a calcium-dependent membrane-binding protein) and *NEK11* (NIMA-related kinase 11) (**Figure 2.12A-C**). The fusion product preserved the full *NEK11* open reading frame and resulted in marked up-regulation of *NEK11* expression (**Figure 2.12D)**. Patient 2 is a 60-year-old man with metastatic prostate cancer not yet treated with hormonal therapies. His xenograft also harbored the *TMPRSS2-ERG* gene fusion, as well as other molecular aberrations **(Figure 2.12E,F)**. Patient 3 is a 46-year-old man diagnosed with colorectal cancer (CRC) in March 2009, who presented with metastatic disease in the liver, bladder perforation, and innumerable polyps upon flexible sigmoidoscopy. ChimeraScan revealed an intrachromosomal gene fusion between acetylserotonin O-methyltransferase–like antisense RNA 1 (*ASMTL-AS1*) and protein phosphatase regulatory subunit 2 (*PPP2R3B*) on chromosome X that abrogated the open reading frame of *PPP2R3B* **(Figure 2.12G)**. Patient 4 is a 48-year-old woman diagnosed with metastatic melanoma who underwent wide local excision for ulcerated spitzoid-type melanoma on her right heel. Her tumor harbored multiple structural aberrations, including a complex interchromosomal rearrangement abolishing the open reading frame of cyclin-dependent kinase inhibitor 2C (*CDKN2C* or *p18INK4C*) **(Figure 2.12H)**.

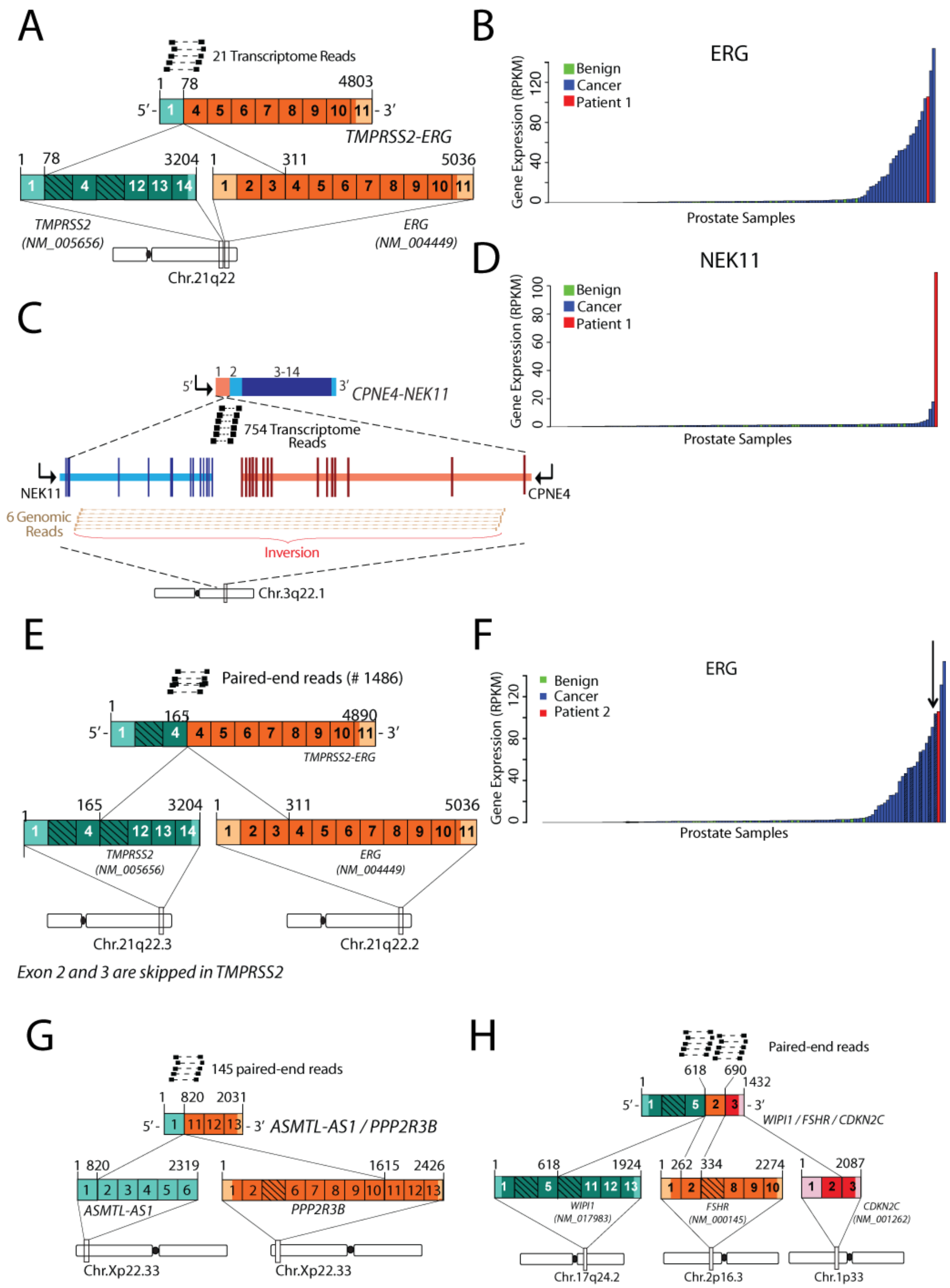


Figure 2.12: Gene fusions nominated in patients from the MI-ONCOSEQ pilot project. (A) Patient 1 harbored the canonical *TMPRSS2-ERG* gene fusion with (B) prominent overexpression of *ERG*. (C) Patient 1 also harbored a novel rearrangement *CPNE4-NEK11* that (D) dramatically upregulated *NEK11* levels. (E-F) Patient 2 also harbored the *TMPRSS2-ERG* fusion. (G) *ASMTL-AS1-PPP2R3B* fusion in Patient 3. (H) Complex rearrangement *WIPI1-FSHR-CDKN2C* in Patient 4.

The MI-ONCOSEQ pilot study used a combination of DNA and RNA sequencing to reveal a broad view of an individual's genetic aberrations. In patient xenografts and tumor samples, ChimeraScan consistently identified gene fusions with high-quality junction spanning reads, including known fusions such as *TMPRSS2-ERG* and novel fusions such as *CPNE4-NEK11*. ChimeraScan served an important role in this integrative bioinformatics setting and should be considered a viable tool for the analysis of RNA-Seq from patient tumor tissues.

Identification of recurrent *NAB2-STAT6* gene fusions in solitary fibrous tumors

In 2011, a 44-year-old woman who had a malignant solitary fibrous tumor (SFT), a rare neoplasm of mesenchymal origin, enrolled in the MI-ONCOSEQ clinical study[145]. Computed tomography (CT)-guided core needle biopsies harvested tissue from a metastatic site in her liver, and whole-exome and transcriptome sequencing were performed. Notably, ChimeraScan identified an intrachromosomal fusion between *NAB2* and *STAT6*. The *NAB2-STAT6* fusion was represented by 1,104 paired-end reads either spanning or encompassing the fusion junction of exon 6 of *NAB2* to exon 18 of *STAT6* (**Figure 2.13**). In the normal genome, *NAB2* and *STAT6* are adjacent genes on chromosome 12q13 that are transcribed in opposite directions. Transcriptome sequencing of 27 additional SFTs was performed, and both ChimeraScan and another in-house algorithm identified the presence of a *NAB2-STAT6* gene fusion in all tumors, indicating high levels of recurrence. Expression of *NAB2-STAT6* fusion proteins was confirmed in SFT, and the predicted fusion products harbor the early growth response (EGR)-binding domain of *NAB2* fused to the activation domain of *STAT6*. Overexpression of the *NAB2-STAT6* gene fusion induced proliferation in cultured cells and activated the expression of

EGR-responsive genes. This study established *NAB2-STAT6* as the defining driver mutation of SFT and provided an example of how neoplasia can be initiated by converting a transcriptional repressor of mitogenic pathways into a transcriptional activator[103].

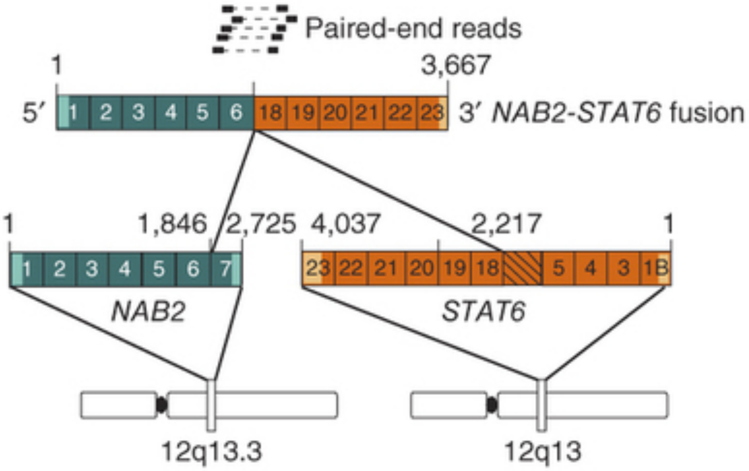


Figure 2.13: NAB2-STAT6 gene fusion detected solitary fibrous tumor patient.
Schematic of the NAB2-STAT6 gene fusion detected in the index case by paired-end sequencing. The has region indicates exons not shown.

Conclusions and future work

In this chapter we described the utility of RNA-Seq for discovering gene fusions that may impact cancer causation and outlined the challenges that must be met by computational methods for gene fusion detection. To address these challenges we developed ChimeraScan[124], a refinement of the original paired-end gene fusion detection methodology developed by Maher *et al.*[24]. Our tests of ChimeraScan suggested that it produces a stringent list of predictions that are enriched with true positives. Our lab and others utilized ChimeraScan to discover recurrent families of gene fusions across multiple cancer types, suggesting that it may continue to serve as a valuable tool for both discovery and clinical RNA-Seq projects[103, 107, 145-148].

Ongoing maintenance and improvement of ChimeraScan will be essential if it is to remain a useful tool for gene fusion discovery. We have identified a number of current issues, feature requests, and enhancements and track them using Google Code issue tracker. An updated release is currently planned that includes migration to the Bowtie 2 alignment tool[149], support for single read datasets, reduced memory requirements for large datasets, outputting a BAM file with fusion alignments, and inclusion of a junction spanning read detection method that supports genomic breakpoints in the middle of exons in addition to introns. Furthermore, we plan to offer new transcriptome references that incorporate new gene annotations provided by Ensembl[75] and GENCODE[87]. Testing will be performed using a variety of read lengths and library types and a robust set of default parameter settings will be established.

The emergence of RNA-Seq as a molecular profiling strategy has fueled the development of a barrage of gene fusion discovery tools, including BreakFusion[123], ChimeraScan[124], deFuse[125], EricScript[126], FusionFinder[127], FusionHunter[128], FusionMap[129], FusionSeq[130], MapSplice[52], ShortFuse[132], SOAPfuse[133], Tophat-Fusion[134], and SnowShoes-FTD[131]. The individual manuscripts describe the creation of a new tool and demonstrate its performance advantages over other tools when tested on a common set of test cases. The performance claims made by each manuscript are likely biased because authors have a chance to optimize their tool on the test datasets while running third-party tools with default parameter settings. Therefore, the field of gene fusion discovery is badly in need of a comprehensive assessment of current tools. This would involve the establishment of a large database of experimentally validated gene fusions from many datasets, as well as a set of simulated gene fusion fragments to enable the measurement of the specificity of algorithms. Each algorithm could then be compared using the large aggregation of test cases and simulated gene fusions. Completing such a study would likely be the most effective way to understand the capabilities of the available tools and inform biologists about how to choose an appropriate tool for a project. A comparative study of gene fusion programs could be modeled after a similar studies done to evaluate ChIP-Seq peak calling programs[150].

Chimeric transcripts detected by RNA-Seq do not necessarily imply chromosomal aberrations; the former may arise due to collinear read-through transcription of adjacent genes, errors in gene annotations, artifacts of the library preparation process, or non-collinear *trans*-splicing of distant genes[151]. Methods that integrate RNA and DNA sequencing data have been shown to accurately discern chromosomal aberrations from other sources of chimeras[152, 153], and such methods may fit nicely into integrative sequencing approaches such as MI-ONCOSEQ.

Despite the fact that RNA sequencing technologies are continuing to mature, the rapid rate of discovery of novel classes of recurrent gene fusions made by existing approaches stakes an irrefutable claim that varieties of RNA sequencing protocols will be a lasting component of cancer genomics studies for years to come. We look forward to future discoveries from this branch of high-throughput sequence data analysis with great anticipation.

Chapter 3: Using RNA sequencing to construct reference gene models

AUTHORS: Iyer MK, Iyer HK, and Chinnaiyan, AM.

AUTHOR CONTRIBUTIONS: MKI and HKI designed the AssemblyLine algorithms and MKI developed the software. MKI generated data for all figures in this chapter and wrote the text.

CITATION: Iyer MK, Iyer HK, Chinnaiyan AM. Large-Scale Meta-Assembly of RNA-Seq Data. Manuscript in preparation.

Abstract

RNA sequencing technologies enable *de novo* reconstruction of full-length transcripts, but establishing a consensus reference transcriptome from collections of RNA-Seq experiments poses numerous challenges. We present AssemblyLine, a meta-assembly algorithm that assimilates transcript predictions from RNA-Seq experiments to produce a merged reference. Innovative features of AssemblyLine include filtering of artifacts arising from incompletely processed RNA and genomic DNA contamination, prioritization of transcript isoforms by abundance level, and inclusion of splicing pattern information into the merging process. In direct comparisons with the Cufflinks meta-assembler AssemblyLine constructed a concise set of transcript predictions with higher precision at a marginal cost in sensitivity. When applied to an existing dataset from

human cell lines and tissues, AssemblyLine nominated an additional 6,397 transcripts that were previously unreported, suggesting that current catalogs of human transcripts are far from complete. Altogether AssemblyLine makes important strides towards the incorporation of RNA-Seq data into mainstream gene databases and the eventual completion of the human transcriptome reference.

Introduction

High-throughput RNA sequencing of eukaryotic organisms has enabled a deeper understanding of the intricate nature of transcription[8, 13, 20, 22, 27, 47, 51, 154]. Intergenic "gene deserts" once thought to be transcriptionally silent express myriad long non-coding RNAs, genic loci once thought to be distinct possess an milieu of overlapping, interleaving, and antisense genes, and catalogs of alternatively spliced protein coding isoforms continue to expand. These discoveries have relied upon transcriptome assembly algorithms that produce full-length transcripts from initial pools of short sequence fragments[26, 46, 47, 61, 63, 64]. The astounding growth rate of RNA-Seq data repositories and continued maturation of assembly strategies forecasts the eventual completion of reference transcriptomes for model organisms, but remaining computational challenges include (1) discerning true transcript expression from sources of background noise and (2) deriving a consensus set of transcript models from independent biological samples. Here we present AssemblyLine, a software package that addresses these challenges and facilitates robust gene model annotation from RNA-Seq data.

Sources of noise in RNA-Seq data

RNA sequencing experiments that isolate poly-adenylated RNA from whole cells inadvertently capture variable amounts of incompletely processed RNA and genomic DNA[13] (**Table 3.1**). These sources of artifacts pervade transcriptome assemblies as lowly expressed, unreliable transcript fragments[78] (Figure: assemblyline_noise_schematic.png).

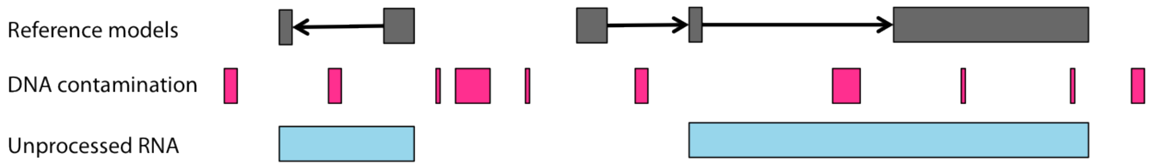


Figure 3.1: Schematic of noise in RNA-Seq data.
Genomic DNA contamination (pink) and incompletely processed RNA (cyan) populate RNA sequencing libraries at variable levels. DNA contamination manifests as spurious reads in both genic and intergenic regions, whereas incompletely processed RNA localized to genic regions only.

Source	Manifestation in assembly
Incompletely processed RNA • Levels vary markedly among libraries for reasons that are incompletely understood	• Intron retention artifacts • Mono-exonic intronic transcripts with sense orientation
Genomic DNA contamination • Relatively higher levels in libraries with low amounts of input RNA	• Mono-exonic transcripts dispersed throughout the genome
Adaptor ligation artifacts • Inefficient A-tailing of blunt ended cDNA	• Chimeras involving highly abundant transcripts

Table 3.1: Source of noise in RNA-Seq libraries.
Chimeric transcripts resulting from inefficiencies in library construction can also produce artifacts but are not relevant to the discovery of collinear transcripts.

To characterize this noise, we partitioned transcripts from *ab initio* assemblies into five categories based on their genomic relationship to reference genes (**Figure 3.2**):

- Well annotated - transcript matches exact splicing pattern of a reference model
- Partially annotated - transcript matches a portion of a reference model but misses some introns
- Intronic - transcript lies within intron of a reference model

- Antisense - transcript overlaps portion of a reference model in the opposite orientation
- Intergenic - transcript has no overlap with exons or introns of reference models

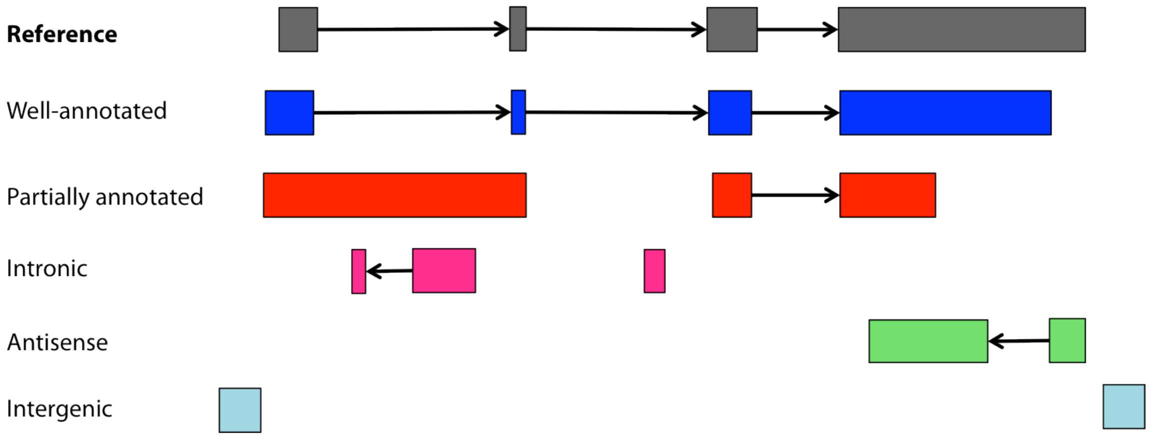


Figure 3.2: Transcript categories based on genomic proximity to known gene models. A well-annotated transcript matches the exact splicing pattern of a reference model. Transcripts that overlap one or more reference models but disagree with splicing patterns are deemed partially annotated. Intronic transcripts lie completely within introns, whereas antisense transcripts overlap reference exons in the opposite orientation. Finally, intergenic transcripts have no overlap with reference models.

We then compared the category fractions of 1,140 RNA-Seq libraries and witnessed striking variability in the fraction of well-annotated transcripts in each library (**Figure 3.3**). As the fraction of well-annotated transcripts decreased, the fractions of partially annotated and intronic transcripts increased dramatically relative to the fractions of intergenic and antisense transcripts. This pattern implicated incompletely processed RNA as the most variable source of noise, followed by genomic DNA contamination.

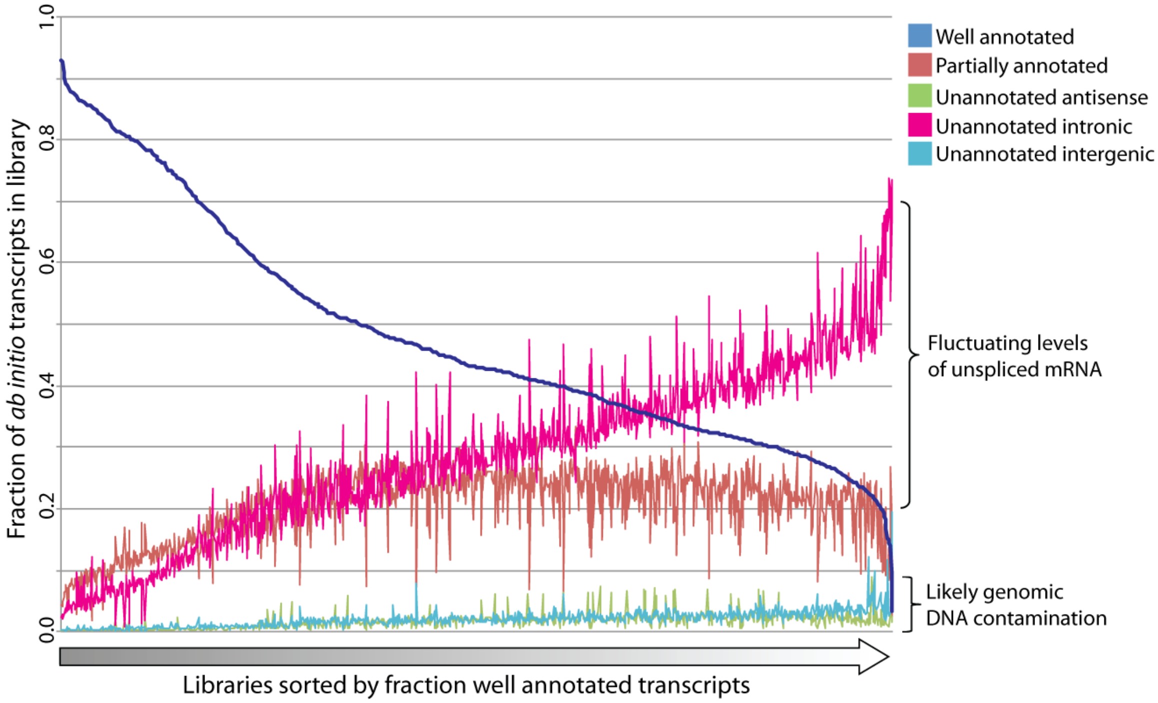


Figure 3.3: Relative fraction of transcript categories across 1,140 RNA-Seq libraries.
Line graph of the relative fraction (y-axis) of transcript categories across 1,140 RNA-Seq libraries. The x-axis contains individual samples sorted by the fraction of well-annotated transcripts in each sample.

Previous attempts at annotating genes using RNA-Seq circumvented background noise by restricting predictions to multi-exonic transcripts or intergenic regions[27, 47, 75, 78, 82]. Given that over 5% of RefSeq transcripts longer than 200nt are mono-exonic[155] and lncRNAs generally have fewer exons than protein-coding genes[156], a significant population of expressed mono-exonic transcripts may be missing from gene catalogs. Furthermore, the high degree of overlapping and interleaving transcription in eukaryotic genomes demands approaches that analyze intronic regions as well.

Apart from excluding areas of the genome, previous studies contended with background noise by designing filtering strategies. Ramskold *et al*. compared the expression levels of exons and intergenic regions to determine an empirical threshold for calling a gene expressed[157]. Similarly, Cabili *et al.* derived empirical detection thresholds by comparing the coverage of full length versus partial length transcripts corresponding to known genes, and further defined a high-confidence set of transcripts that were detected in multiple samples or by independent *ab initio* assembly programs[78]. A recent effort to incorporate zebrafish RNA-Seq data into the Ensembl genebuild discarded exon regions with relatively low coverage[82]. In contrast to empirical filtering methods, Guttman *et al.* developed a statistical approach that models background noise as though read alignments were randomly permuted throughout the genome[47]. Although all of these strategies enrich for expressed genes, they do not account for classes of transcripts that are robustly expressed at relatively low levels[78, 158]. The recent ENCODE study employed a statistic called the non-parametric irreproducible detection rate (*npIDR*)[13, 55]. This statistic embodies the notion that purposeful transcription should be observable by independent experiments. The study filtered novel transcripts that were less than 90% recurrent (*npIDR < 0.1)* between biological replicates of the same sample but still detected an alarming number of novel mono-exonic transcripts. The authors acknowledged the possibility of artifacts due to low levels of DNA contamination but did not compare *npIDR* values between novel and annotated transcripts to credential their chosen detection threshold. Altogether the aforementioned schemes establish the use of noise thresholds based on expressed levels and reproducibility.

In this work we present a machine learning approach for filtering background noise that exploits synergy between transcript expression levels and reproducibility across biological samples. We accommodate variation in noise levels between intergenic and intronic regions by modeling each set of genomic regions independently. Our method weights its predictions by the total amount of noise in each library relative to other libraries such that noisier libraries are handled more stringently.

Meta-assembly

Meta-assembly refers to merging together two or more assemblies to produce a consensus assembly. Establishing a consensus assembly is vital to downstream analysis because it provides a common foundation for comparing transcriptional dynamics[40, 41]. Previously, we developed a merging approach that clustered isoforms into a single set of exon regions per gene[27]. This strategy facilitated the discovery of novel cancer-associated loci but abolished alternative splicing information and relied upon additional assays such as RACE for precise delineation of transcript structure. An earlier generation of algorithms was developed for EST assembly and introduced splicing graphs as an effective representation of the isoform problem[71, 159]. Building on these approaches, Trapnell *et al.* released a meta-assembly utility within the Cufflinks package called Cuffmerge[41]. Cuffmerge converts transcripts from *ab initio* assemblies into faux read alignments and reruns Cufflinks on these alignments in a modified mode. Cufflinks then emits a minimal set of merged transcripts that explains the input transcripts. Alternatively, aggregating the raw sequences from multiple RNA-Seq samples before running standard *ab initio* or *de novo* assembly programs can produce a consensus assembly[46]. However, naively aggregating the raw sequences compounds background noise and forces the choice of a single set of parameters for all samples. Moreover, transcripts specific to a small subset of samples may be unintentionally pruned along with other minor isoforms. We anticipate that the tremendous memory and computational time required to complete large assemblies may also limit the feasibility of this approach.

Here, we present a meta-assembly algorithm that produces isoforms from splicing graphs after pruning sources of noise such as intron retentions and inappropriately long exons. We use a greedy dynamic programming approach that reports the most highly abundant transcripts and optionally discards minor isoforms. Studies of alternative splicing have revealed a tightly controlled system where often only a small number of possible isoforms is observed from loci with innumerable splicing possibilities[28, 160]. Our algorithm incorporates correlative splicing patterns by traversing path graphs built from the original splice graphs.

Methods

AssemblyLine (http://assemblyline.googlecode.com) is a software package written in Python and R that (1) characterizes and filters sources of background noise in RNA-Seq assemblies and (2) performs meta-assembly. We designed the two algorithms to be used together as part of a gene discovery workflow, but they can also be separated or combined with other software tools. Several utility scripts are also included in the package that prepare input data, assess assembly performance, characterize aspects of the consensus assembly, prepare genome browser tracks for visualization, and facilitate downstream analysis. In this section we outline the steps and algorithmic details of the tool.

Aggregating individual transcriptome assemblies

The first stage of AssemblyLine combines a set reference transcripts with distinct transcriptome assemblies together into a single position-sorted GTF file. To run this step, users must provide a list containing sample names, replicate groupings, and paths to individual GTF files as a tab-delimited text file. If the filtering algorithm will be used a set of reference transcripts must also be provided. Users may specify the GTF attribute containing transcript abundance information using the *--gtf-score-attr* parameter (by default this assumes an attribute called "FPKM"). Abundance computation from poor quality RNA-Seq libraries sequenced at shallow depth may be highly inaccurate; therefore, AssemblyLine normalizes abundance values by converting them to percentile ranks to mitigate the impact of outliers. Finally, a transcript length filter can be applied in this step to eliminate artifacts. Certain *ab initio* assemblers are prone to overinflate FPKM values for very short transcripts for reasons that are incompletely understood[161]. Furthermore, transcripts shorter than the average fragment size of the library may correspond to mapping artifacts because such small fragments should have been size-selected away during library preparation. Therefore, users may specify a length cutoff using the *--min-transcript-length* option. The conventional size cutoff used to define long RNAs is 250nt and is the default value for this parameter.

When a reference GTF file is provided a percentage of annotated transcripts are labeled as tests (set using the *--random-test-frac* parameter). In the filtering step of AssemblyLine transcripts that correspond to tests are held out of the training process and are instead used to measure the performance of the classifier. If the user does not specify a list of specific gene identifiers to use as test data using the *--tests* parameter, AssemblyLine will randomly label a fraction of reference transcripts as tests.

Filtering noise from transcriptome assemblies

The AssemblyLine filtering algorithm consists of the following steps: (1) computing transcript annotation status, genomic category, and recurrence (2) modeling noise properties of individual libraries, and (3) predicting whether each transcript is 'expressed' or 'background' noise (**Figure 3.4**).

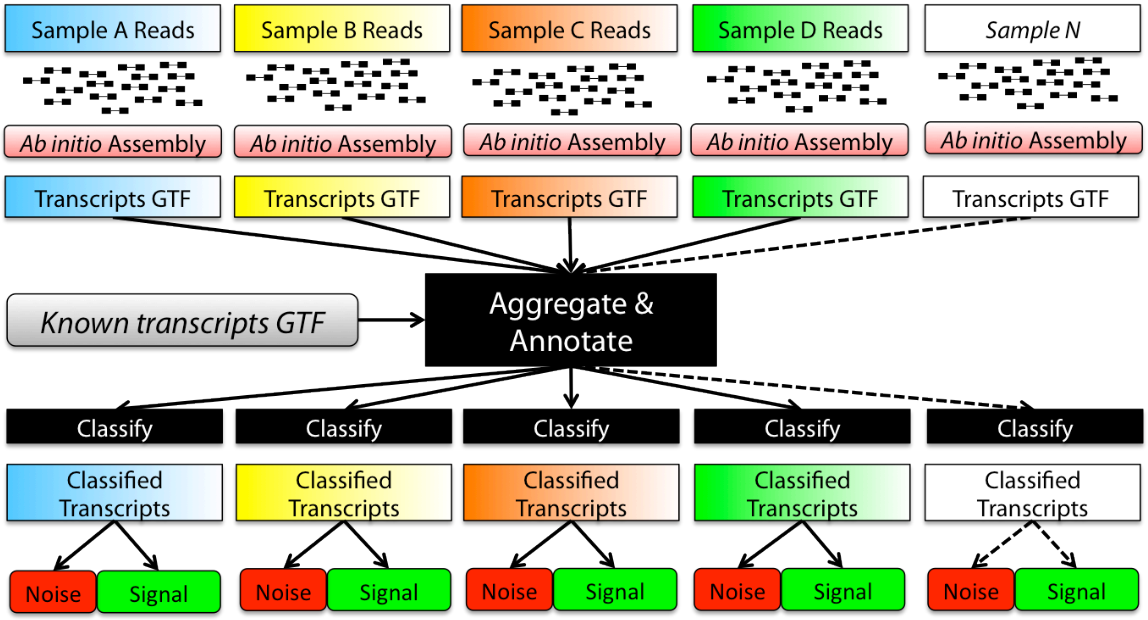


Figure 3.4: AssemblyLine filtering workflow.
Individual RNA-Seq datasets are processed using *ab initio* assembly. The resulting GTF files are aggregated and compared to a set of known transcript models. A machine learning approach classifies each dataset separately, and the robustly expressed transcript "signal" is separated from the "noise".

The filtering algorithm relies on a set of high confidence known transcripts provided by the user. Assembled transcripts that overlap known transcripts in the sense orientation are denoted "annotated". The remaining transcripts are categorized based on their position relative to known transcripts (**Figure 3.2**). Annotated transcripts that were initially marked as tests are treated as unannotated genes and categorized appropriately. We also compute a new measure of recurrence similar to *npIDR* by averaging the number of times each base of a transcript was observed in transcripts from independent biological samples (**Figure 3.5**). The recurrence measure was a powerful differentiator of known and unannotated transcripts in data downloaded from the human lincRNA catalog study[78] (**Figure 3.6**).

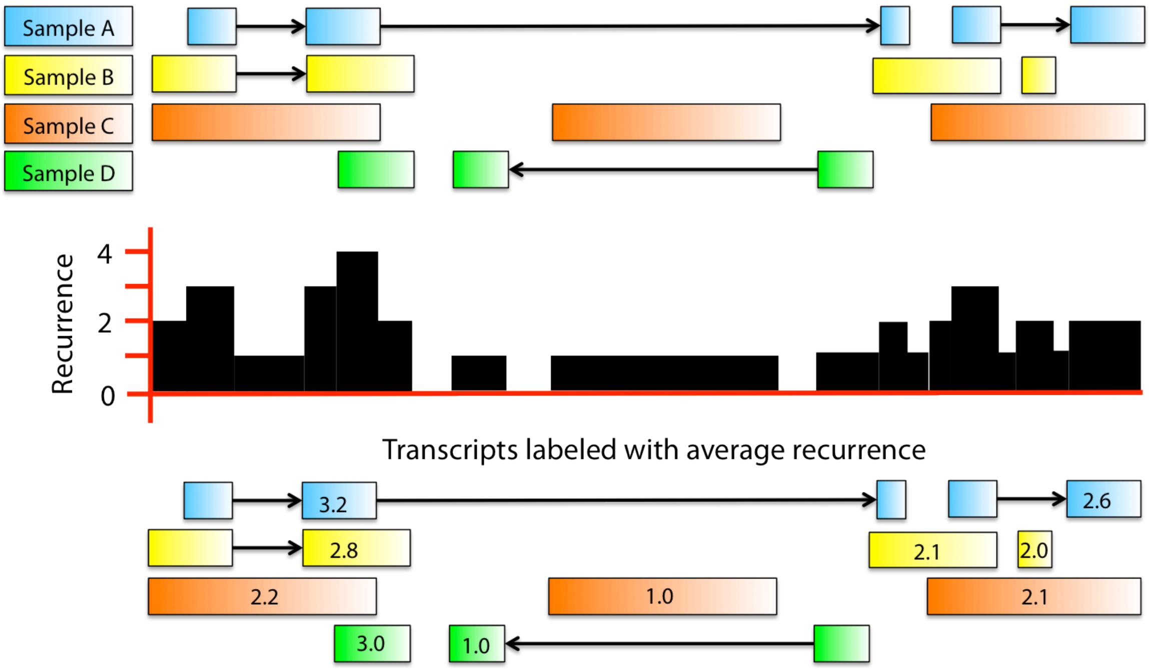


Figure 3.5: Computation of transcript recurrence across multiple independent samples
(top) Transcript models from assemblies of four samples are shown. (middle) A pileup graph is created where the height at each base corresponds to the integer number of samples with transcripts that overlap that base. (bottom) For each transcript the recurrence equals the average recurrence per base. Thus, recurrence scores range from 1.0 <= *R* <= *N*, where *N* is the number of biological samples being analyzed.

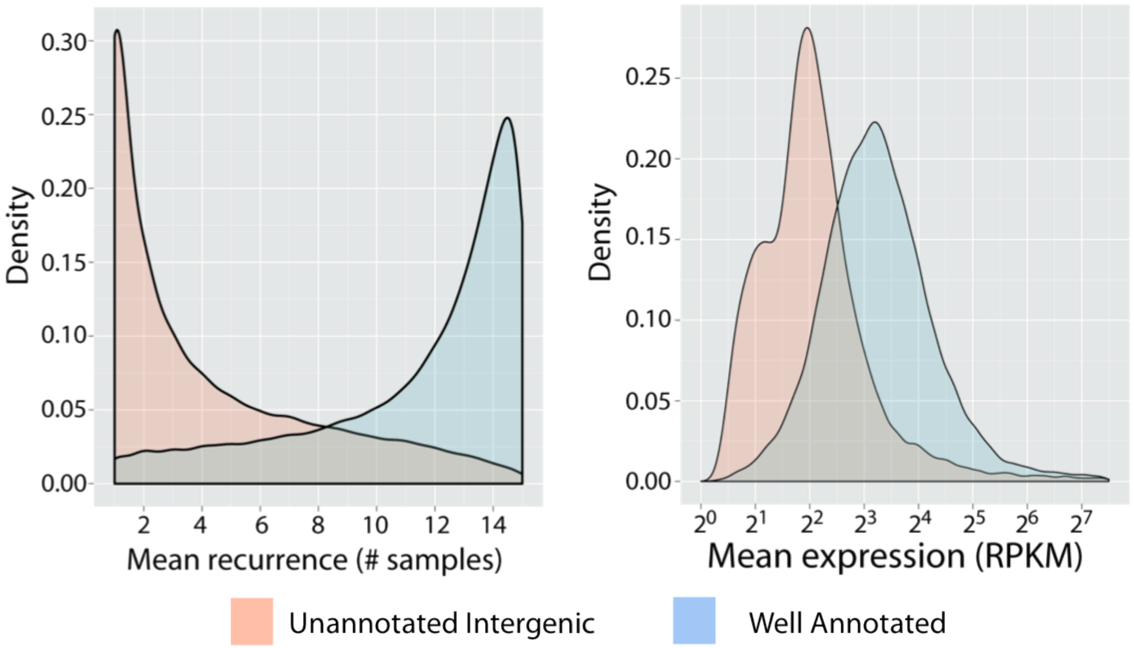


Figure 3.6: Transcript recurrence and abundance distinguish unannotated transcripts
Transcript recurrence (left) and abundance (right) distinguished well-annotated from unannotated transcripts in assemblies from the Human BodyMap 2.0 study.

Discerning novel transcripts from genomic contamination or incompletely processed RNA is problematic because there is no source of true noise with which to train a classifier. To address this concern AssemblyLine treats all unannotated transcripts as pseudo-noise and trains a classifier to predict the likelihood that a transcript is known. Separate classifiers are created for transcripts compatible with incompletely processed RNA and intergenic transcripts. A classifier is trained by computing bivariate kernel density estimates of transcript abundance and recurrence on a square grid for known and unannotated transcripts (**Figure 3.7**).

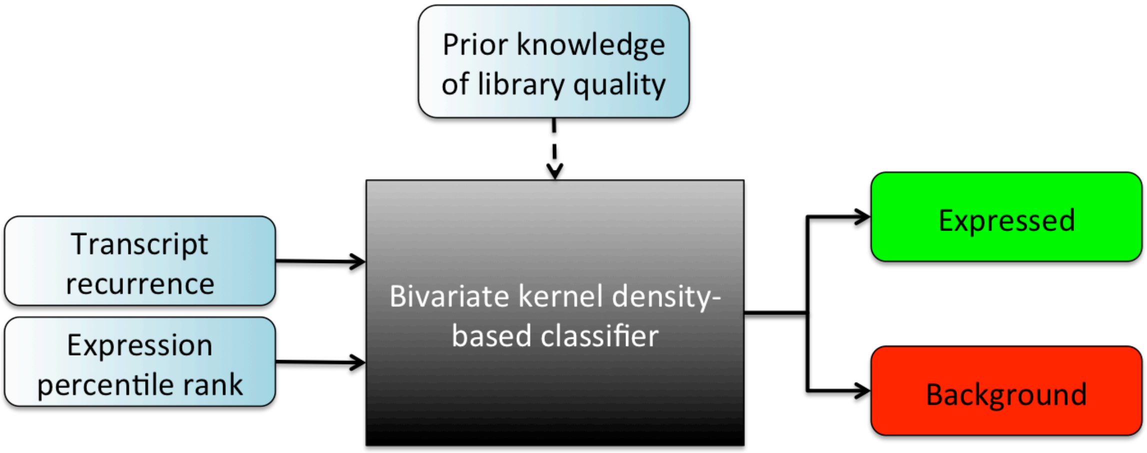

Prior knowledge of library quality
Transcript recurrence
Expression percentile rank
Bivariate kernel density-based classifier
Expressed
Background

Figure 3.7: Schematic of transcript classification approach.
A bivariate kernel density estimation function (*kde2d* in R) models transcript recurrence and relative abundance. Density landscapes are modulated by prior knowledge of library quality. The classifier outputs expressed (green) and background (noise) transcripts as separate files.

We then compute a grid of log-likelihoods by dividing the known density by the unannotated density at each grid point after adding a nominal value to avoid floating point overflow errors:

$$Z_{(x,y)} = log_{10}\left(\frac{known\ density_{(x,y)} + C}{unannotated\ density_{(x,y)} + C}\right)$$

To account for the total noise present in the library, we weight the log-likelihood estimates by a relative measure of total noise in the library when compared to all libraries in the analysis. This weight equals the ratio of the fraction of known to unannotated transcripts in a library divided by the ratio of the medians of these fractions in all libraries:

$$W_L = \frac{\frac{fraction\ known_L}{fraction\ unannotated_L}}{\frac{median(fraction\ known_{all})}{median(fraction\ unannotated_{all})}}$$

$$Z_W = Z \times log_{10}(W_L)$$

For each transcript in the assembly we compute the weighted log-likelihood of the transcript being annotated by linearly interpolating the transcript abundance and recurrence onto the grid. We then determine a log-likelihood threshold by optimizing the balanced accuracy (average of sensitivity and specificity) of the classifier performance on test transcripts. Transcripts with log-likelihood below this threshold are labeled ‘background’ and the remainder ‘expressed’. A number of performance reports and visualizations are also generated during this process. Results from individual libraries are then combined to produce background and expressed files as output. In a typical analysis the transcripts classified as background noise are discarded and meta-assembly is carried out on the expressed fraction.

Meta-assembly

The AssemblyLine meta-assembly program accepts a position-sorted GTF file containing transcript fragments (transfrags) as input. The file is parsed and transfrags are assigned a score according to the *--gtf-score-attr* parameter (default: “FPKM”). Transfrags are then bundled into non-overlapping loci and partitioned by strand. Transfrags lacking strand information are assigned to the strand with the best supporting abundance score. If there are no supporting stranded transcripts the strand is left unknown. Meta-assembly is then carried out on each strand separately (**Figure 3.8**).

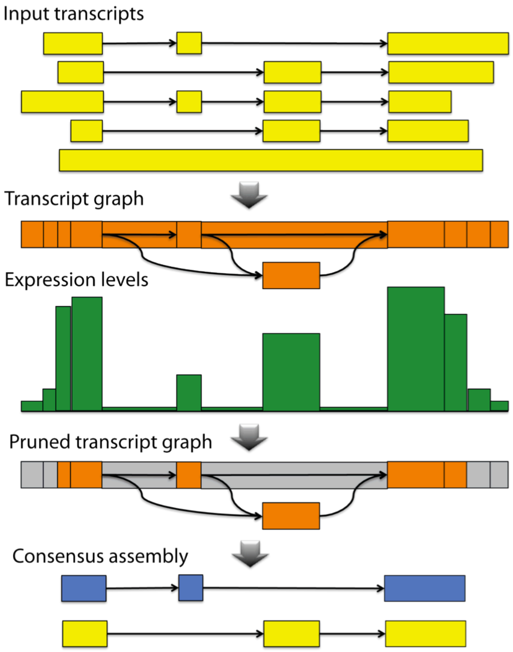


Figure 3.8: Meta-assembly workflow.
A splicing graph (a directed acyclic graph) is produced from the set of input transcripts at each locus. Expression levels are used to apply heuristics that prune or trim lowly expressed intron retentions and transcript ends. A dynamic programming traverses the pruned graph and produces transcript isoforms.

We create directed acyclic splicing graphs where nodes in the graph reflect contiguous exonic regions and edges correspond to alternative splicing possibilities. Nodes in the splicing graph are then pruned according to several criteria. First, we trim low scoring ends in the graph that correspond to extraneously long exons or overhanging exons that extend into introns. The parameter *--trim-utr-fraction* sets the relative score threshold for trimming (default 0.1). Second, nodes within introns are trimmed when their scores are less than a fraction of neighboring exons (set using the *--trim-intron-fraction* parameter, default 0.25). Weakly connected components of the pruned splicing graphs are then extracted and processed independently.

A splicing graph encompasses the milieu of possible isoforms that could be transcribed. Enumerating all possible paths through splicing graphs is impractical; many graphs have millions of paths only minute fractions of which are observed *in vivo*. The initial input transfrags provide paths through the splicing graph and also indicate which parts of the graph are more abundant. Our approach incorporates this path information by building a path graph that subsumes the original splice graph. The path graph is a *De Bruijn* graph[162] where each node represents a contiguous path of length *k* through the splice graph, and edges connect paths with *k-1* nodes in common. As *k* increases so does the amount of correlative path information retained in the graph but at the cost of losing short transfrags with length less than *k*. Thus for each splice graph the partial path length *k* is optimized to maximize the number of nodes in the path graph with the constraint that the summed scores of transfrags with path length greater than or equal to *k* is above a fraction of the total score of all transfrags. This fraction is called *ksensitivity* and is set to 0.90 by default. After the path graph has been constructed, we effectively extend every partial length transfrag into a full-length transcript by transmitting the transfrag's score along incoming and outgoing edges. Scores are allocated proportionally at nodes with multiple incoming or outgoing edges. This smoothing process assures that the sum of incoming node scores and outgoing node scores are equivalent at every node.

Finally, a set of isoforms is predicted from the graph using a greedy algorithm. The algorithm finds and reports the highest abundance transcript by traversing the graph using dynamic programming. The score of the path equals the minimum score of all nodes in the path. The path score is then subtracted from every node in the path and the dynamic programming procedure is repeated. Suboptimal transcripts are enumerated until a path score falls below a fraction of the highest scoring transcript (set by the *--fraction-major-isoform* parameter). The total number of isoforms produced from each gene can also be explicitly constrained using the *--max-paths* parameter. The meta-assembled isoforms are reported in GTF and/or BED format. The genomic landscape of transfrag scores can optionally be reported in BedGraph track format as well.

Results

We assessed the AssemblyLine meta-assembler using transcript assemblies made available as part of the human lincRNA catalog study by Cabili *et al.*[78] (http://www.broadinstitute.org/genome_bio/human_lincrnas). We henceforth refer to this dataset as the Cabili dataset. The transcript assemblies were created by aligning reads to the human genome using TopHat and assembling the alignments *ab initio* with Cufflinks. The dataset consisted of 25 libraries from 21 different human cell and tissue types and contained 4,268,910 transcript fragments (transfrags) in aggregate.

Filtering

A typical filtering procedure was performed on the Cabili assemblies. A large number (2,512,734 out of 4,268,910 or 58.8%) of the transfrags in the assemblies were shorter than 250bp and discarded during the aggregation step. Short transfrags can often occur in abundance when the underlying *ab initio* or *de novo* assembly process performs little or no filtering. The remaining 1,756,176 transcripts were subjected to categorization and filtering steps. Transcripts were compared to the GENCODE v15 reference and 10% of GENCODE genes were randomly selected as tests. The fraction of GENCODE v15 annotated transfrags ranged from 0.27 to 0.71 indicating significant variation in library noise levels (**Figure 3.9**). We noticed that the library with the highest fraction of annotated transfrags, denoted liver_R, was one of two samples eliminated from expression analysis in the Cabili *et al.* study due to its low coverage. The other sample removed, hela_R, also had a high fraction of annotated transcripts.

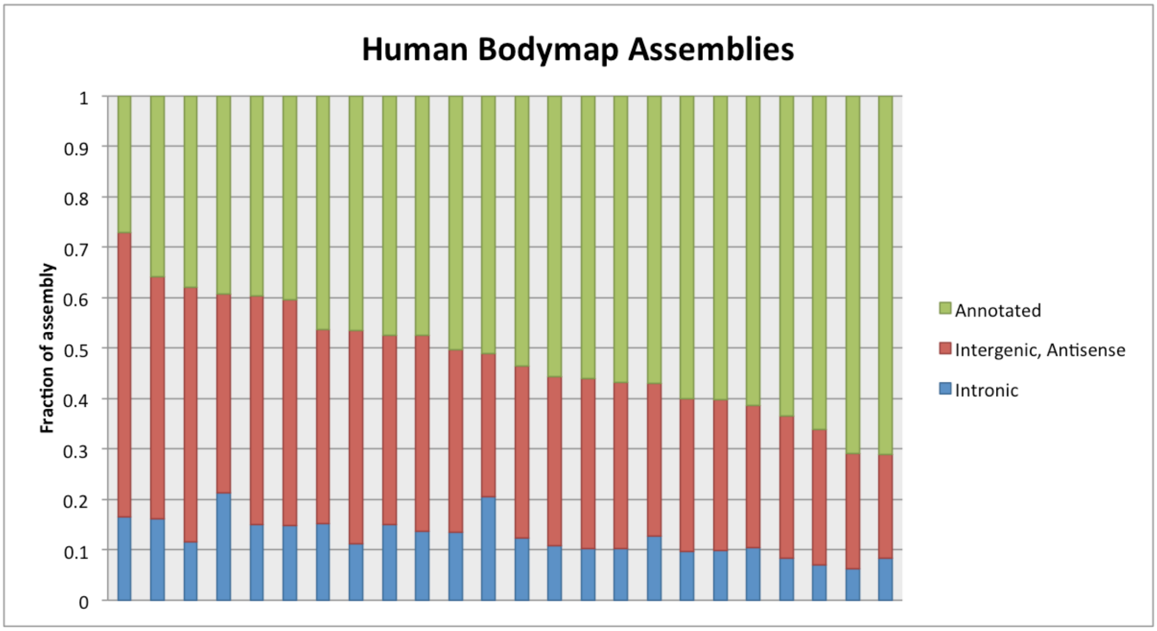


Figure 3.9: Relative fraction of transcript categories in the Cabili assemblies. Transcripts in each assembly were categorized as Annotated (green), Intergenic/Antisense (red), or Intronic (blue). The fraction of annotated transcripts varies from 0.27 to 0.71 across the cohort.

The bivariate kernel density classifier was applied to individual libraries in the Cabili dataset. For each library a series of contour plots were produced indicating the distribution of transcript abundance and recurrence values (**Figure 3.10**). Known transfrags were more highly expressed and recurrent relative to unannotated transfrags (**Figure 3.10 left and middle)**. The resulting log-likelihood landscape reflected the highly disjoint nature of the two distributions (**Figure 3.10 right panel**).

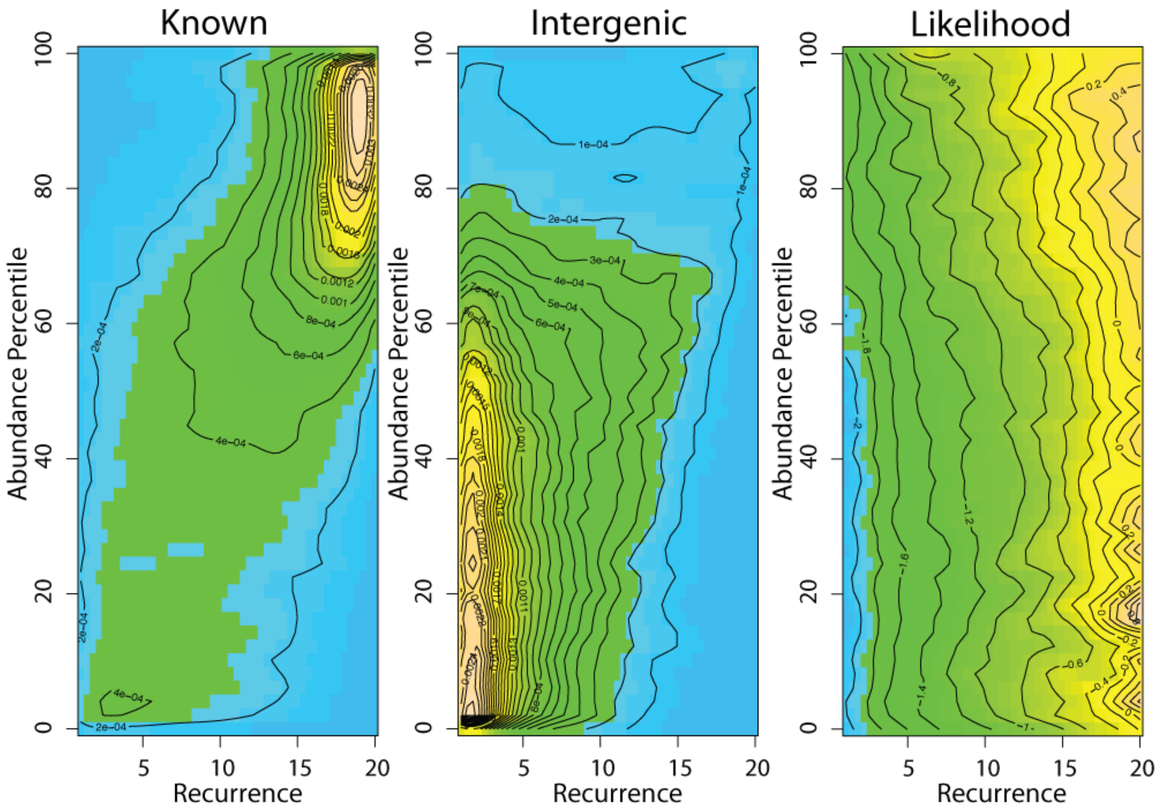


Figure 3.10: Bivariate kernel density classifier example.
(left and middle) Contour plots depicting density of recurrence and abundance (measured as expression percentile rank) for the 'heart' sample from the Cabili dataset across known (left) and unannotated intergenic (middle) transcripts. (right) Contour map depicting the log-likelihood of a transcript being 'Known' created by superimposing the density plots from known and intergenic transcripts. In this plot transcripts with low recurrence and abundance have a low likelihood of being from the pool of known transcripts.

Classifier performance was measured on test transcripts, visualized using receiver-operator characteristic (ROC) curves, and quantified using the area under the curve (AUC) metric (**Figure 3.11a)**. The AUC values ranged from 0.83 to 0.90 (average 0.88) for intergenic classifiers and 0.73 to (average 0.77) 0.87 for intronic classifiers (**Figure 3.11b**). A log-likelihood cutoff was chosen for each classifier by optimizing balanced accuracy (average of sensitivity and specificity) (**Figure 3.11c**). Balanced accuracy values were 0.76-0.83 for intergenic classification and 0.66-0.78 for intronic classification (**Figure 3.11d**). For example, a log-likelihood cutoff of -0.99

(corresponding to a probability < 0.10 of being known) was chosen to partition intergenic transfrags from heart tissue (**Figure 3.12**).

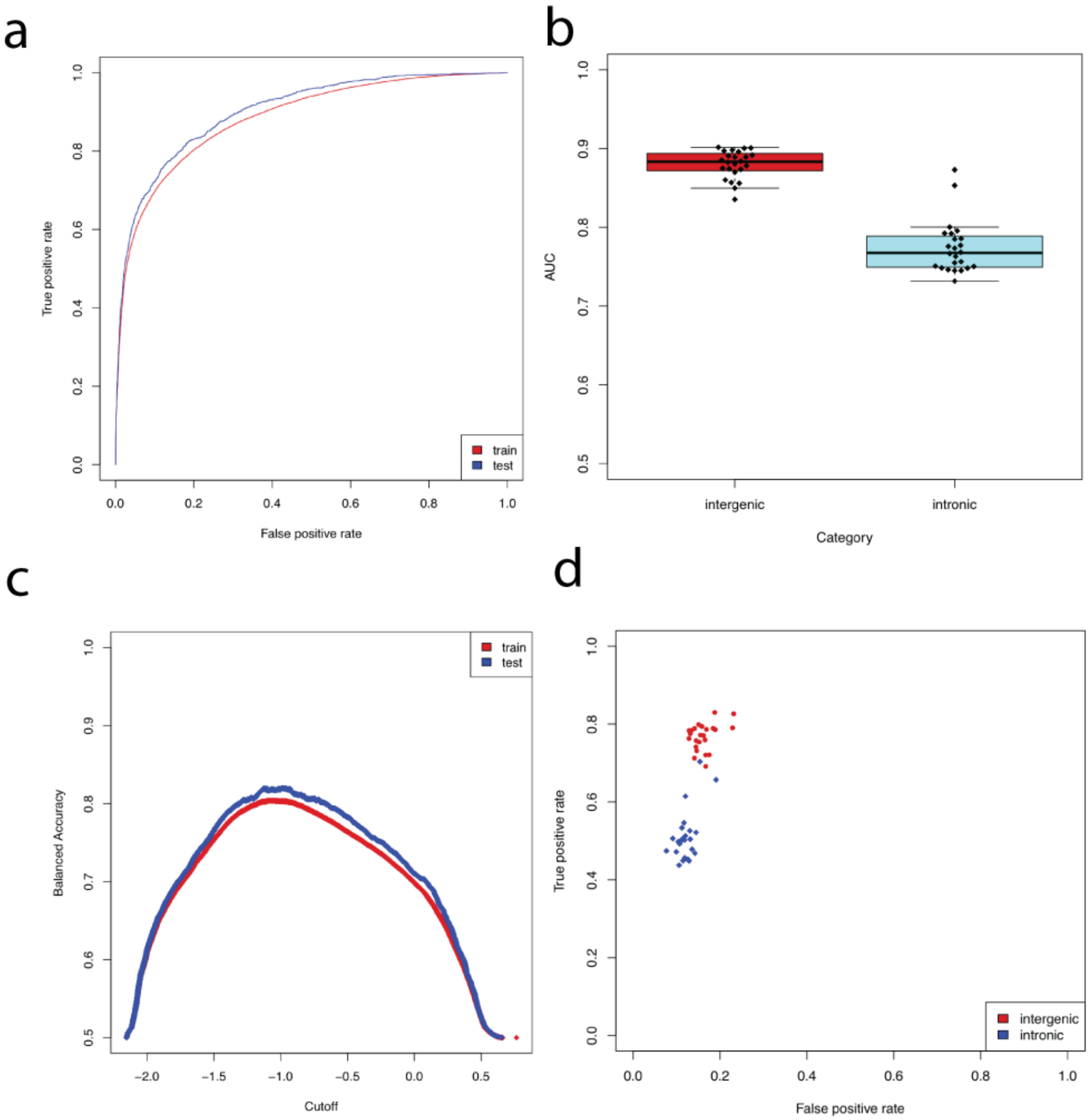


Figure 3.11: Filtering performance on Cabili dataset.
(a) Receiver operating characteristic (ROC) curve showing classifier performance on both training and test datasets for the 'heart' sample from the Cabili dataset. (b) Boxplot showing range of area under the curve (AUC) values for intergenic (red) and intronic (blue) transcripts. (c) Line plot showing the relationship between the log-likelihood cutoff and the balanced accuracy for the 'heart' sample. (d) Scatter plot showing the sensitivity and specificity values at the optimal cutoff point after classification of intergenic and intronic transcripts.

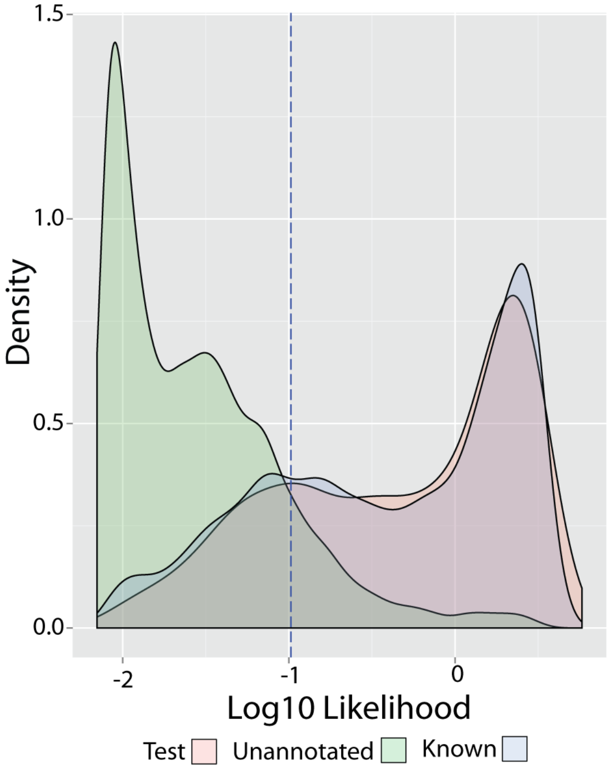


Figure 3.12: Log-likelihood density distribution
Density of the log-likelihood metric across unannotated (green), known (blue) and test (red) transcripts from the Cabili 'heart' sample. The dashed blue line showed the optimal threshold for distinguishing known from unannotated transcripts based on balanced accuracy.

Given that the RNA-Seq libraries were not strand-specific we additionally required intronic transfrags compatible with incompletely processed RNA to be multi-exonic. The filtered transfrags from each library were merged into a set containing 1,017,606 (58%) of the original transfrags, of which 80.8% corresponded to GENCODE v15 transcripts and 19.2% were unannotated (**Table 3.2**).

Category	Filtered transfrags (percentage)
Known	822,446 (80.8%)
Unannotated	195,160 (19.2%)
• Antisense	65,181 (6.4%)
• Intronic same strand (multi-exonic)	6,039 (0.59%)
• Intronic opposite strand (multi-exonic)	8,674 (0.85%)
• Intronic ambiguous strand (multi-exonic)	6,981 (0.69%)
• Interleaving	50,028 (4.9%)
• Intergenic	58,258 (5.7%)

Table 3.2: Filtered transfrag statistics

Meta-assembly performance

We assessed the AssemblyLine and Cuffmerge meta-assemblers on the unfiltered Cabili *et al.* assemblies. We enumerated all isoforms expressed at greater than or equal to one percent of the major isoform for each gene (for AssemblyLine we set *--fraction-major-isoform*=0.01 and for Cuffmerge we set *--min-isoform-fraction*=0.01). AssemblyLine produced a much smaller assembly than Cuffmerge (317,795 versus 854,666 genes, 404,768 versus 978,660 transcripts, and 669,560 versus 1,220,668 exons) with a higher average number of isoform per gene (1.27 versus 1.14) and exons per transcript (1.65 versus 1.25) (**Figure 3.13**).

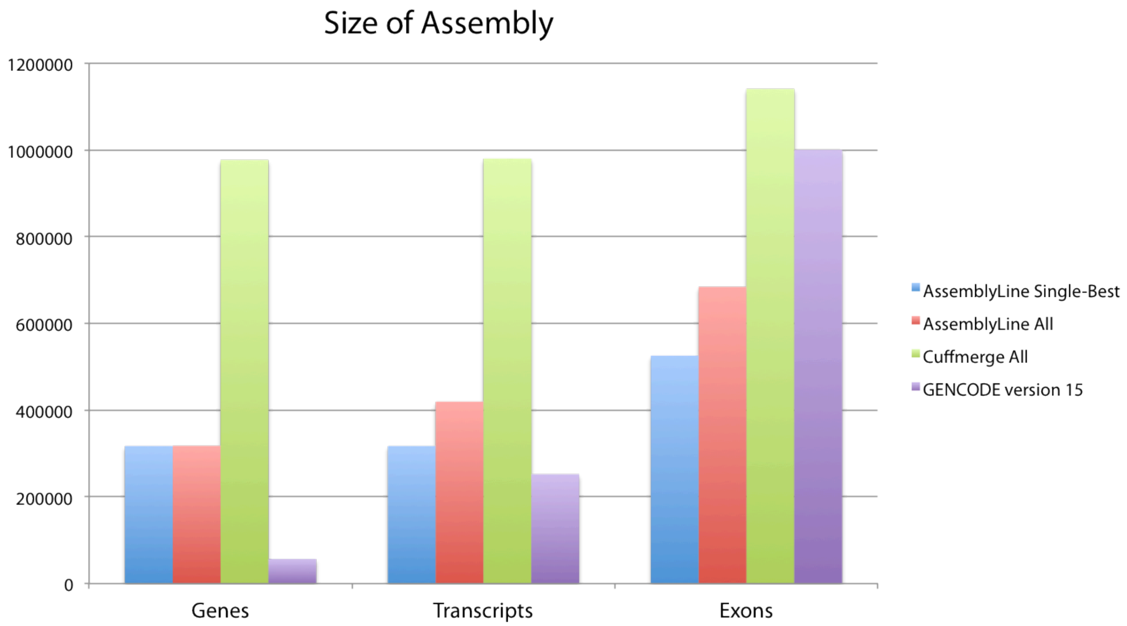


Figure 3.13: Comparison of merged assembly size
Bar plots showing the total assembly size in genes, transcripts, and exons for AssemblyLine (blue and red) and Cuffmerge (green) meta-assemblies. AssemblyLine was run in two modes: 'single-best' (blue) nominated a single isoform for every gene, and 'all' (red) retained all isoforms. The size of the GENCODE v15 reference is provided as a comparison.

Next we compared the ability of the tools to detect reference transcripts from the GENCODE v15 database[87]. We mirrored the approach described by Grabherr *et al.* to predict an empirical upper sensitivity limit called the ‘Oracle Set’[46]. The ‘Oracle Set’ reflects the subset of covered based, introns, and splicing patterns present within the initial unmerged transfrags (it is theoretically possible to outperform the Oracle Set for the detection of splicing patterns because meta-assembly can potentially create new patterns not present in the initial transfrags). AssemblyLine detected slightly fewer reference-covered positions, introns, and splicing patterns than Cuffmerge, but a higher percentage of AssemblyLine predictions matched reference gene models (**Figure 3.14**). AssemblyLine was even more precise when only the major isoform (single-best) at each gene was considered.

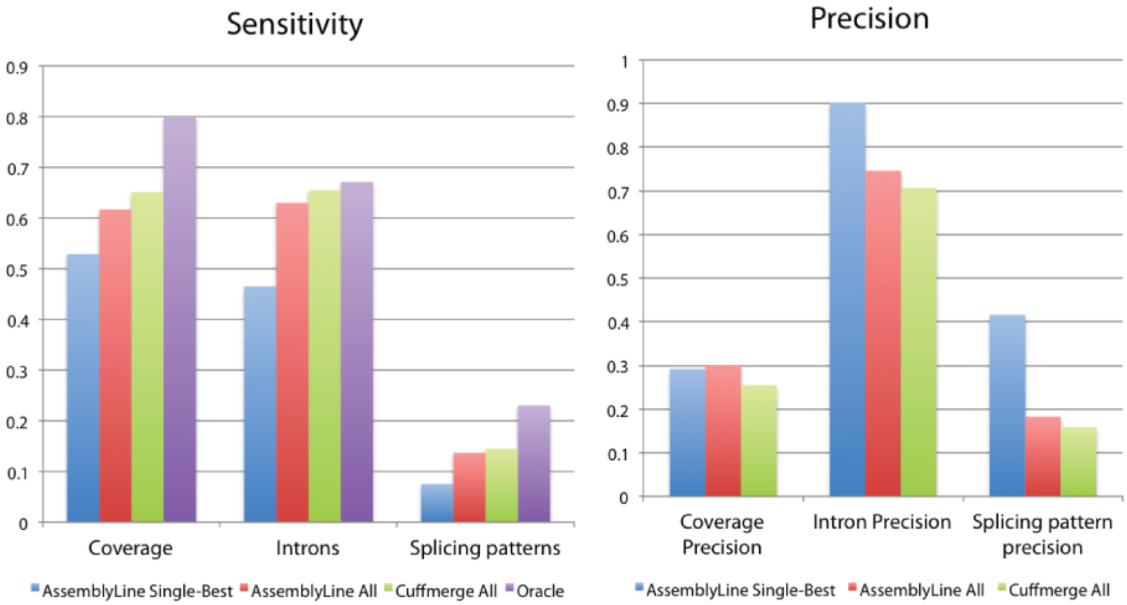
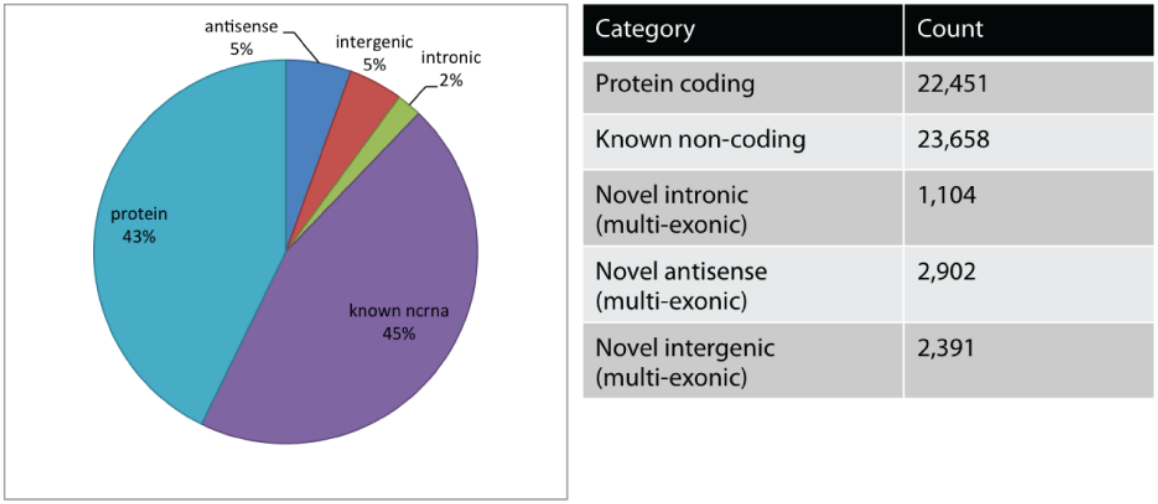


Figure 3.14: Performance of AssemblyLine and Cuffmerge across Cabili dataset
The sensitivity (left) and precision (right) for detection of GENCODE v15 annotated bases, introns, and splicing patterns was assessed. The ‘Oracle’ sensitivity (purple) predicts the maximum attainable performance from the initial transfrags.

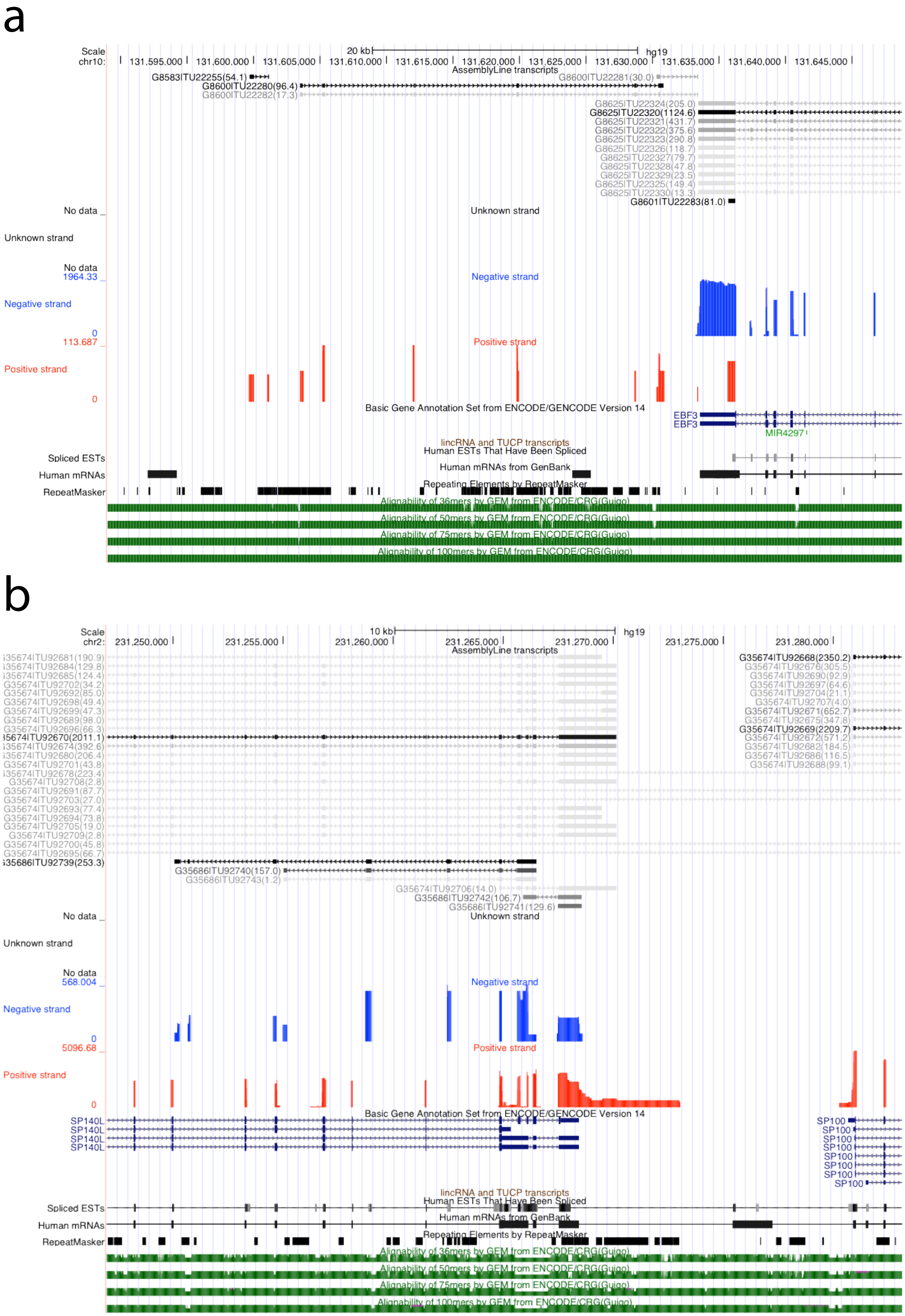
A catalog of human lincRNAs redefined

Despite the fact that the Cabili *et al.* lincRNA BodyMap catalogued 8,195 intergenic long RNAs, including 4,819 that were absent from gene databases, we hypothesized that additional long RNAs remain to be discovered. Therefore, we ran the AssemblyLine workflow on the Cabili dataset to create a consensus assembly. Remarkably, 12% of the 53,506 multi-exonic genes in the consensus assembly were novel, including 1,104 intronic, 2,902 antisense, and 2,391 intergenic loci (**Figure 3.15**). Perusal of the assembly confirmed the presence of many high quality candidates, indicating that the human long RNA transcriptome remains incomplete (**Figure 3.16**).



Category	Count
Protein coding	22,451
Known non-coding	23,658
Novel intronic (multi-exonic)	1,104
Novel antisense (multi-exonic)	2,902
Novel intergenic (multi-exonic)	2,391

Figure 3.15: Discovery of novel transcripts in the Cabili dataset
(left) Pie chart showing the proportions of transcripts corresponding to known protein or ncrna models as well as novel antisense, intergenic, or intronic regions. (right) Table showing the number of transcripts from each category of the meta-assembly.


a
Scale
chr10:
20 kb
hg19
131,595,000
131,600,000
131,605,000
131,610,000
131,615,000
131,620,000
131,625,000
131,630,000
131,635,000
131,640,000
131,645,000
AssemblyLine transcripts
G8583|TU22255(54.1)
G8600|TU22280(96.4)
G8600|TU22282(17.3)
G8600|TU22281(30.0)
G8625|TU22324(205.0)
G8625|TU22320(1124.6)
G8625|TU22321(431.7)
G8625|TU22322(375.6)
G8625|TU22323(290.8)
G8625|TU22326(118.7)
G8625|TU22327(79.7)
G8625|TU22328(47.8)
G8625|TU22329(23.5)
G8625|TU22325(149.4)
G8625|TU22330(13.3)
G8601|TU22283(81.0)
No data
Unknown strand
Unknown strand
No data
1964.33
Negative strand
Negative strand
0
113.687
Positive strand
Positive strand
0
Basic Gene Annotation Set from ENCODE/GENCODE Version 14
EBF3
EBF3
MIR4297
lincRNA and TUCP transcripts
Spliced ESTs
Human ESTs That Have Been Spliced
Human mRNAs
Human mRNAs from GenBank
RepeatMasker
Repeating Elements by RepeatMasker
Alignability of 36mers by GEM from ENCODE/CRG(Guigo)
Alignability of 50mers by GEM from ENCODE/CRG(Guigo)
Alignability of 75mers by GEM from ENCODE/CRG(Guigo)
Alignability of 100mers by GEM from ENCODE/CRG(Guigo)
b
Scale
chr2:
10 kb
hg19
231,250,000
231,255,000
231,260,000
231,265,000
231,270,000
231,275,000
231,280,000
AssemblyLine transcripts
G35674|TU92681(190.9)
G35674|TU92684(129.8)
G35674|TU92685(124.4)
G35674|TU92702(34.2)
G35674|TU92692(85.0)
G35674|TU92698(49.4)
G35674|TU92699(47.3)
G35674|TU92689(98.0)
G35674|TU92696(66.3)
G35674|TU92670(2011.1)
G35674|TU92674(392.6)
G35674|TU92680(206.4)
G35674|TU92701(43.8)
G35674|TU92678(223.4)
G35674|TU92708(2.8)
G35674|TU92691(87.7)
G35674|TU92703(27.0)
G35674|TU92693(77.4)
G35674|TU92694(73.8)
G35674|TU92705(19.0)
G35674|TU92709(2.8)
G35674|TU92700(45.8)
G35674|TU92695(66.7)
G35686|TU92739(253.3)
G35686|TU92740(157.0)
G35686|TU92743(1.2)
G35674|TU92706(14.0)
G35686|TU92742(106.7)
G35686|TU92741(129.6)
G35674|TU92668(2350.2)
G35674|TU92676(305.5)
G35674|TU92690(92.9)
G35674|TU92697(64.6)
G35674|TU92704(21.1)
G35674|TU92707(4.0)
G35674|TU92671(652.7)
G35674|TU92675(347.8)
G35674|TU92669(2209.7)
G35674|TU92672(571.2)
G35674|TU92682(184.5)
G35674|TU92686(116.5)
G35674|TU92688(99.1)
No data
Unknown strand
Unknown strand
No data
568.004
Negative strand
Negative strand
0
5096.68
Positive strand
Positive strand
0
Basic Gene Annotation Set from ENCODE/GENCODE Version 14
SP140L
SP140L
SP140L
SP140L
SP100
SP100
SP100
SP100
SP100
SP100
SP100
SP100
lincRNA and TUCP transcripts
Spliced ESTs
Human ESTs That Have Been Spliced
Human mRNAs
Human mRNAs from GenBank
RepeatMasker
Repeating Elements by RepeatMasker
Alignability of 36mers by GEM from ENCODE/CRG(Guigo)
Alignability of 50mers by GEM from ENCODE/CRG(Guigo)
Alignability of 75mers by GEM from ENCODE/CRG(Guigo)
Alignability of 100mers by GEM from ENCODE/CRG(Guigo)

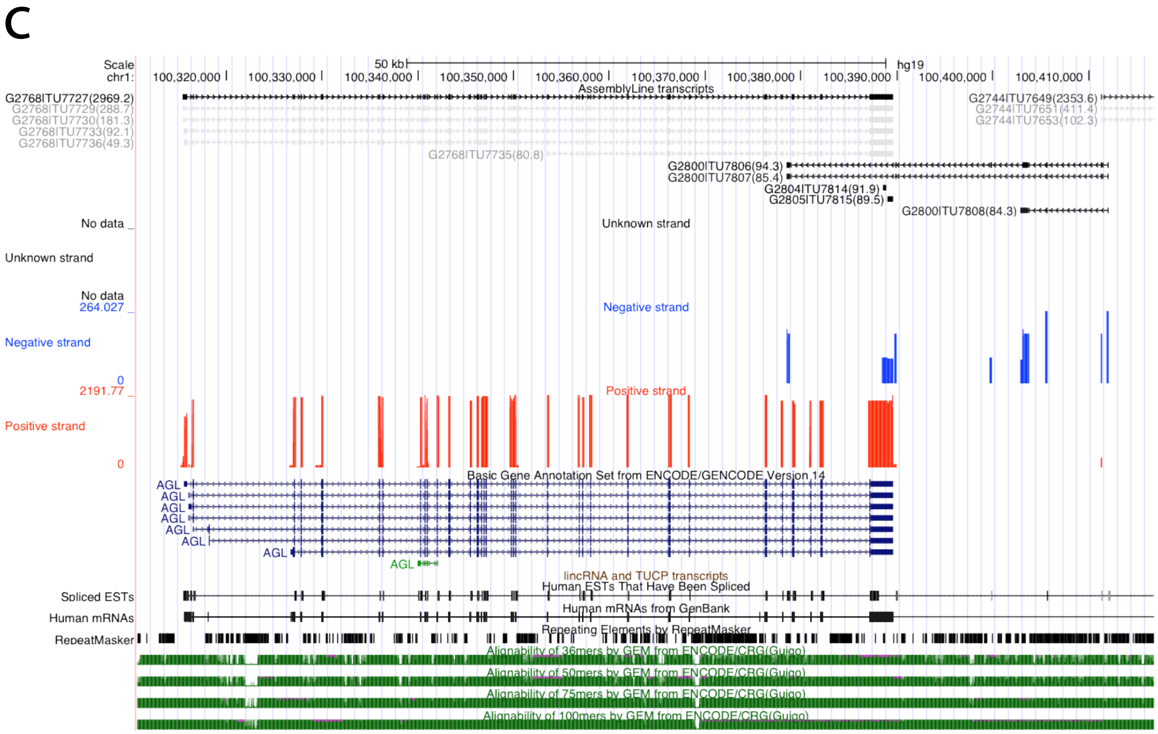


Figure 3.16: Examples of novel transcripts present in the Cabili dataset.
(a) Multi-exonic intergenic transcripts on the negative strand proximal to the *EBF3* gene, (b) Antisense transcription at the *SP140L* locus generating several multi-exonic long RNAs, (c) Interleaving transcription generating multiple isoforms antisense to the 3' end of the *AGL* gene.

Discussion

Assembly of data from high-throughput RNA sequencing experiments has revealed an unanticipated layer of complexity in eukaryotic transcriptomes, but incorporating RNA-Seq transcript models into gene annotation catalogs requires (1) mitigating noise in datasets and (2) condensing the vast amount of data into a concise set of gene models. GENCODE, Ensembl, and other groups are forecasting the incorporation of RNA-Seq data into their catalogs, but the means for doing so remain obscure[75, 82]. Efforts to annotate gene models from collections of RNA-Seq data have had limited success mitigating background DNA contamination and incompletely processed RNA artifacts, and often omitted entire classes of transcripts based on their structure or genomic location in order to simplify the noise problem. Gene discovery pipelines have been created and employed in published works, but few tools have emerged as production-level utilities for RNA-Seq meta-assembly[13, 27, 78]. Thus, the field of RNA-Seq analysis would strongly benefit from a gene discovery software package that could efficiently and confidently delineate a precise set of expressed transcripts in a dataset. AssemblyLine overcomes sources of noise and contamination in the data and employs a meta-assembly strategy that prioritizes the reconstruction of highly abundant isoforms.

The kernel density estimation (KDE) approach used in AssemblyLine has several advantages. It relies on a single bandwidth parameter *h* that controls the degree of smoothing. By contrast, other machine learning methods often require multiple parameters that may be computationally expensive to tune. Currently we use a rule of thumb to select the default bandwidth, but further optimization of bandwidth selection may improve performance further[163]. Additionally, a KDE-based classifier smoothens density estimates and thus may be more robust for datasets with small numbers of transfrags. KDE approaches are certainly appropriate for modeling continuous variables such as transcript abundance and recurrence. Finally, the density landscape lends itself to an intuitive visualization of the patterns of abundance and recurrence in a dataset using contour maps or 3D surface plots.

The balanced accuracy (average of sensitivity and specificity) metric is an intuitive way to select likelihood thresholds for classification, but other measures of accuracy such as the F-measure could be easily substituted. However, the nature of the problem precludes the use of false discovery rate (FDR) as a method to control the error rate. While it may be possible to control the filtering process using simulated noise or by generating matched genomic DNA or nascent RNA sequencing data, we leave these possibilities as future work[164].

The filtering algorithm relies on a robust set of known transcript models. If no reference transcriptome is available it may still be possible to use AssemblyLine. In this case one should run the meta-assembly algorithm first and then employ heuristic filtering to select robust gene models. A second run of AssemblyLine could then use these models as a training set. In this way AssemblyLine could be used iteratively to "bootstrap" the gene discovery process. It is crucial that training models be precisely annotated, and therefore other assays such as RT-PCR-Seq may be useful as complementary assays for validation[88]. If erroneous gene models are used to train the classifier, the results may be unpredictable.

The greedy dynamic programming algorithm used to enumerate paths during meta-assembly assures that the highest scoring path will always be predicted. In the ideal case the high scoring paths match the most abundance isoforms, but highly fragmented or erroneous transfrags can make this difficult. Studies on patterns of alternative splicing lend credence to our greedy approach. It was observed that the number of expressed isoforms plateaus at 10-12 per gene per sample, and the most abundant isoform rarely decreases below 40% of the total gene expression even for genes with large numbers of annotated isoforms[13]. Thus, a greedy approach will recover the most common isoforms at the expense of missing minor isoforms. The problem of splice graph assembly closely resembles the maximum flow problem from optimization theory. Comparison of our dynamic programming solution against other algorithms for computing flow networks remains an area of ongoing study.

AssemblyLine currently utilizes genomic alignment positions and transcript abundance information in its algorithm, but other sources of information may further improve meta-assembly performance. Transcription start site (TSS) information obtained from ChIP-Seq datasets or other sources could be easily incorporated into the meta-assembly program by adding edges between the source node in the splice graph and each known TSS. Similarly, transcription termination sites (TTS) could be incorporated into the splice graph by adding additional edges from poly-adenylation sites to the sink node of the splice graph. Optional incorporation of TSS and TTS data from public databases is considered an important area of future work.

Applying AssemblyLine to the Cabili *et al.* dataset revealed numerous long RNAs absent from GENCODE and the published lincRNA catalog, suggesting that current gene model databases remain incomplete. Indeed, the multifarious nature of cell and types tissues suggests the need for large scale spatiotemporal profiling of transcriptional complexity in both normal and disease states. We believe that after amassing such data AssemblyLine could effectively distill robust transcripts and lead to the eventual completion of reference gene databases.

Chapter 4: Discovery of novel transcripts in prostate cancers

AUTHORS: Iyer MK, Prensner JR, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM.

AUTHOR CONTRIBUTIONS: M.K.I., J.R.P. and A.M.C. designed the project and directed experimental studies. M.K.I., O.A.B., C.S.G. and C.A.M. developed computational platforms and performed sequencing data analysis. M.K.I., O.A.B. and H.K.I. performed statistical analyses. J.R.P., S.M.D., J.C.B., Q.C., N.P., H.D.K., B.L., X.W., I.A.A., X.C., X.J. and D.R. performed experimental studies. J.S. and J.T.W. coordinated biospecimens. M.K.I., J.R.P. and A.M.C. interpreted data and wrote the manuscript.

CITATION:

Prensner JR*, Iyer MK*, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*. 2011 Jul 31;**29**(8):742-9.

* These authors made equal contributions

Abstract

Long non-coding RNAs (lncRNAs) are emerging as key molecules in human cancer, with the potential to serve as novel markers of disease and to reveal uncharacterized aspects of tumor biology. Here we perform poly-A+ long RNA sequencing across a cohort of 102 prostate cancer cell lines and tissues followed by *ab initio* transcriptome assembly and delineate 1,859 unannotated long RNAs throughout the human genome. Among these we define 121 prostate cancer–associated ncRNA transcripts (PCATs) as long non-coding RNAs aberrantly expressed in the disease. These findings establish the utility of RNA-Seq to identify disease-associated ncRNAs that may improve the stratification of cancer subtypes.

Introduction

Recently, RNA-Seq has provided a method to delineate the entire set of transcriptional aberrations in a disease, including novel transcripts not measured by conventional analyses[26, 47, 50, 162, 165]. To facilitate interpretation of sequence read data, existing computational methods typically process individual samples using either short read gapped alignment followed by *ab initio* reconstruction[26, 47] or *de novo* assembly of read sequences followed by sequence alignment[46, 50]. These methods provide a powerful framework to uncover uncharacterized long non-coding RNAs (lncRNAs) >250bp. Although still largely unexplored, long non-coding RNAs have emerged as a new aspect of biology, with evidence suggesting that they are frequently cell-type specific, contribute important functions to numerous systems[166, 167] and interact with known cancer genes such as *EZH2*[168]. Indeed, several well-described examples, such as *HOTAIR*[168, 169] and *ANRIL*[170, 171], indicate that lncRNAs may be essential actors in cancer biology, typically facilitating epigenetic gene repression through chromatin-modifying complexes[172, 173]. Moreover, lncRNA expression may confer clinical information about disease outcomes and have utility as diagnostic tests[169, 174]. The characterization of RNA species, their functions and their clinical applicability is therefore a major area of biological and clinical importance.

Here, we describe a comprehensive analysis of lncRNAs in 102 prostate cancer tissue samples and cell lines by RNA-Seq. We apply *ab initio* computational approaches to delineate the annotated and unannotated transcripts in this disease, and we find 121 lncRNAs, termed PCATs, whose expression patterns distinguish benign, localized cancer and metastatic cancer samples. To our knowledge, our findings describe the first comprehensive study of lncRNAs in prostate cancer, provide a computational framework for large-scale RNA-Seq analyses and highlight prostate cancer lncRNAs associated with disease progression.

Results

RNA-Seq analysis of the prostate cancer transcriptome

Over two decades of research have generated a genetic model of prostate cancer based on numerous neoplastic events, such as loss of the *PTEN*[175] tumor suppressor gene and gain of oncogenic ETS family transcription factor gene fusions[98, 137, 176] in large subsets of prostate cancer patients. As some patients lack these genetic aberrations, we hypothesized that prostate cancer similarly harbored disease-associated lncRNAs that characterized specific molecular subtypes.

To pursue this hypothesis, we applied transcriptome sequencing on a cohort of 102 prostate tissues and cell lines that included 20 benign adjacent prostates (benign), 47 localized prostate cancers (PCA), 14 metastatic tumors (MET) and 21 prostate cell lines. In total, 1.723 billion fragments were generated from 201 lanes of sequencing on the Illumina Genome Analyzer and Illumina Genome Analyzer II. We developed a bioinformatics workflow to define a consensus set of transcript models from this cohort (**Figure 4.1a**).

Ab initio assembly and quantification with TopHat and Cufflinks

Reads were mapped using TopHat v1.0.13[56] to the human genome (hg18)[155] with a maximum of two mismatches. We obtained 1.418 billion unique alignments including 114.4 million that spanned splice junctions, with a median of 14.7 million mapped reads per sample. Ambiguous alignments were discarded in this analysis. Aligned reads from TopHat were assembled with Cufflinks version 0.8.2[26]. Cufflinks assembles exonic and splice-junction reads into transcripts using their alignment coordinates. After assembling transcripts, Cufflinks computes isoform-level abundances by finding a parsimonious allocation of reads to the transcripts within a locus. We filtered transcripts with abundance less than 15% of the major transcript in the locus, and minor isoforms with abundance less than 5% of the major isoform. Default settings were used for the remaining parameters. The Cufflinks assembly stage yielded a set of transcript annotations for each of the sequenced libraries. We partitioned the transcripts by chromosome and used the Cuffcompare utility provided by Cufflinks to merge the transcripts into a combined set of annotations. The Cuffcompare program performs a union of all transcripts by merging transcripts that share all introns and exons. The 5' and 3' exons of transcripts were allowed to vary by up to 100nt during the comparison process. Cuffcompare reported a total of 8.25 million distinct transcripts.

Distinguishing transcripts from background signal

As expected from a large tumor tissue cohort, individual transcript assemblies may have sources of noise, such as artifacts of the sequence alignment process, unspliced intronic pre-mRNA and genomic DNA contamination. To exclude these from our analyses, we trained a decision tree to classify transcripts as expressed versus background (the approach presented here has matured over the course of several years into what is now the AssemblyLine software package described in **Chapter 3**). The approach rests on the premise that a manually curated gene database could represent a reliable set of true positives with which to train a classifier. We used the AceView annotations[72] which we believed had an adequate representation of low abundance lncRNA transcripts that may be cell-type specific. For each transcript predicted by Cufflinks we collected the following statistics: length (bp), number of exons, recurrence (number of samples in which the transcript was predicted), 95th percentile of abundance (measured in Fragments per Kilobase per Million reads (FPKM)) across all samples, and uniqueness of genomic DNA harboring the transcript (measured using the Rosetta uniqueness track from UCSC). Using this information, we used recursive partitioning and regression trees in R (package rpart) to predict, for each transcript, whether its expression patterns and structural properties resembled those of annotated genes. Classification was performed independently for each chromosome in order to incorporate the effect of gene density variability on expression thresholds. Examination of decision trees indicated that expression level and recurrence were most frequently the best predictors of known transcripts (**Figure 4.1b**). Transcripts not classified as background noise were used for further analysis.

Refinement of transcript fragments

The decision tree demonstrated a sensitivity of 70.8% and specificity of 88.3% in our cohort, and 2.88 million (34.9%) unannotated transcript fragments were classified as "expressed". We then developed a program to extend and merge intron-redundant transcripts to produce a minimum set of transcripts that could possibly explain the assemblies produced by Cufflinks. By merging transcripts in this manner we relinquished the ability to detect some types of alternative TSSs, but drastically reduced the total number of independent transcripts in our assembly. We believe merging all intron-redundant transcripts is suitable for qualitative detection of transcriptionally active regions, but more sophisticated methods would be necessary for the study of alternative splicing, alternative TSSs, and alternative poly-adenylation site usage within well characterized regions (the merging approach used here subsequently evolved into the AssemblyLine meta-assembly algorithm presented in **Chapter 3**). The merging step produced a total of 123,554 independent transcripts. We then re-computed transcript abundance levels for these revised transcripts in Reads per Kilobase per Million (RPKM) units. These expression levels were used for the remainder of the study.

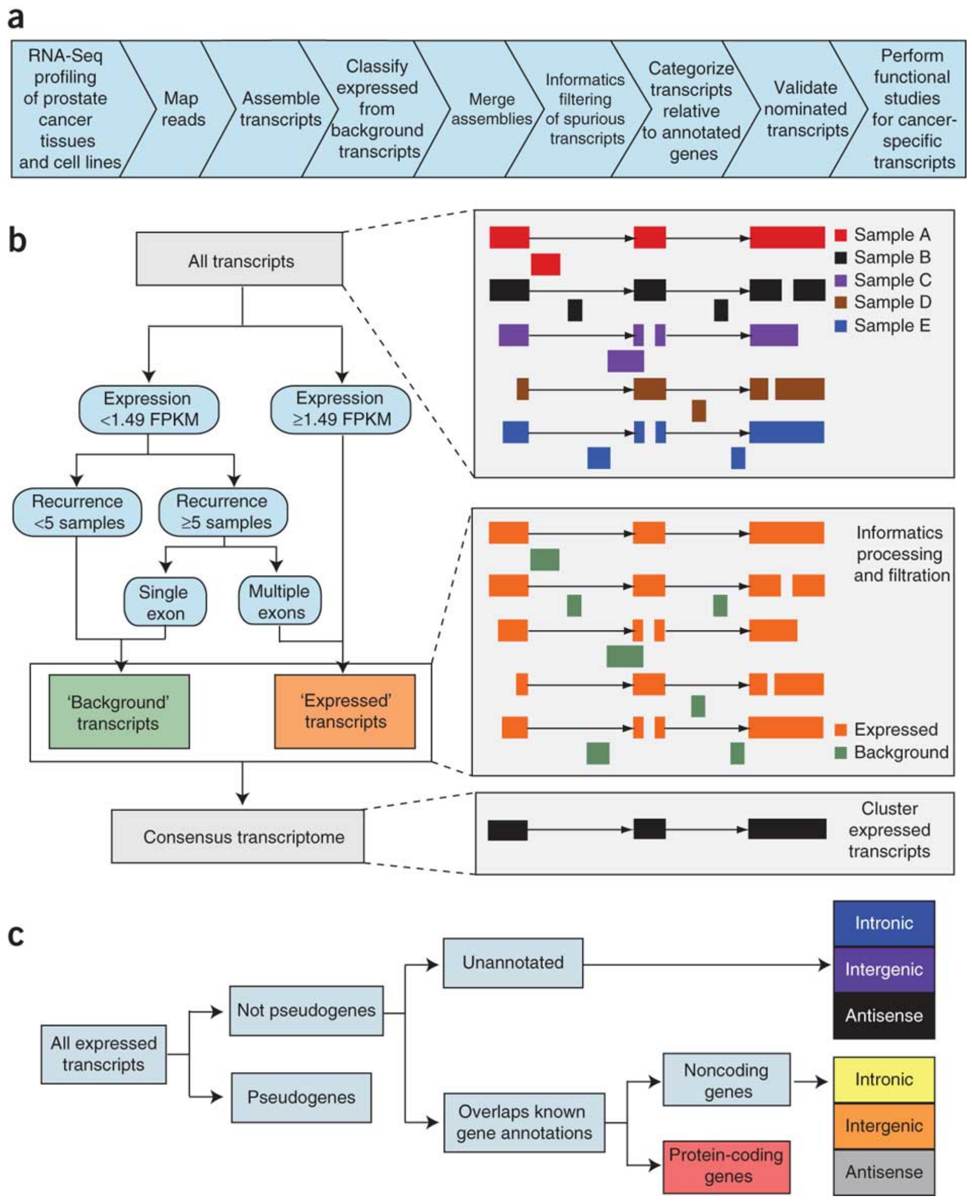


Figure 4.1: Analysis of transcriptome data for the detection of unannotated transcripts
(**a**) Schematic overview of the methodology employed in this study. (**b**) Graphical representation of the bioinformatics filters used to merge individual transcriptome libraries into a single consensus transcriptome. The merged consensus transcriptome was generated by compiling all individual transcriptome libraries and using individual decision tree classifiers for each chromosome to define high-confidence 'expressed' transcripts and low-confidence 'background' transcripts, which were discarded. The example decision tree on the left was trained on transcripts on chromosome 1. The graphics on the right illustrate the application of the informatics filtration pipeline to sample assembled transcripts. (**c**) After informatic processing and filtration of the sequencing data, transcripts were categorized to identify unannotated ncRNAs. Transcribed pseudogenes were isolated, and the remaining transcripts were categorized based on overlap with an aggregated set of known gene annotations into annotated protein coding, noncoding and unannotated. Both annotated and unannotated ncRNA transcripts were then separated into intronic, intergenic and antisense categories based on their relationship to protein-coding genes.

We applied several additional filtering steps to isolate the most robust transcripts. First, we discarded transcripts with a total length less than 200nt. Our size selection protocol isolates RNA molecules larger than this, and small RNA sequencing protocols would likely be needed to quantify smaller molecules with high confidence. Second, we discarded single exon transcripts with greater than 75% overlap to another longer transcript. We believe many of these are produced from unspliced pre-mRNA molecules and do not represent functional RNA products. Third, we removed transcripts that lacked a completely unambiguous genomic DNA stretch of at least 40nt. We measured genomic uniqueness using the Rosetta uniqueness track downloaded from the UCSC genome browser website. We believe transcripts spanning poorly mappable regions are more likely to occur due to mapping artifacts and the availability of longer reads would alleviate the need for this filtering step. Finally, we retained transcripts that were not present in at least 5% of our cohort (>5 samples) at more than 5.0 RPKM. It is possible that certain subtypes of prostate cancer may express highly specific transcripts, and future studies to characterize these transcripts could provide additional insight into the biology of tumor subtypes.

In certain instances we observed transcripts that were interrupted by poorly mappable genomic regions. Additionally, for low abundance genes we observed fragmentation due to the lack of splice junction or paired-end read evidence needed to connect nearby fragments. We reasoned that expression profiles of these fragmented transcripts should be highly correlated. To demonstrate this, we measured the difference in the Pearson correlation between expression of randomly chosen exons on the same transcript versus expression of spatially proximal exons on different transcripts. We found that in our cohort, a Pearson correlation >0.8 had a positive predictive value (PPV) of >95% for distinct exons to be part of the same transcript. Using this criteria, we performed hierarchical agglomerative clustering to extend transcript fragments into larger transcriptional units. Pairs of transcripts further than 100kb apart, transcripts on opposite strands, and overlapping transcripts were not considered for clustering. Groups of correlated transcripts were merged, and introns <40nt in length were removed.

Comparison with gene annotation databases

The 44,534 transcripts produced by the bioinformatics pipeline were classified by comparison with a comprehensive list of "annotated" transcripts from UCSC, RefSeq, ENCODE, Vega, and Ensembl (**Figure 4.1c**). First, transcripts corresponding to processed pseudogenes were separated. This was done to circumvent a known source of bias in older versions of TopHat. Until recently TopHat mapped reads to genomic DNA in its first step, predisposing exon-exon junction reads to align to their spliced retroposed pseudogene homologues. Future improvements to the algorithm eliminated this bias by mapping reads to known transcripts first. Next, transcripts with >1bp of overlap with at least one annotated gene on the correct strand were designated "annotated", and the remainder were deemed "unannotated". Transcripts with no overlap with protein coding genes were subdivided into intronic, intergenic, or partially intronic antisense categories based on their relative genomic locations.

Informatics filtering of unspliced pre-mRNA isoforms

We observed a significant increase in the percentage of intronic transcripts in our assembly relative to known intronic ncRNAs. This led us to observe that in many cases unspliced pre- mRNAs may appear at sufficient levels to escape the filtering steps employed by Cufflinks during the assembly stage. We then removed intronic and antisense transcripts that were correlated (Pearson correlation >0.5) to their overlapping protein coding genes in order to better approximate the true number of intronic or antisense transcripts in the transcriptome. In effect, these steps produced a consensus set of 35,415 transcripts supporting long poly-adenylated RNA molecules in human prostate tissues and cell lines. Overall we detected a similar number of transcripts as present in the either the RefSeq or UCSC databases[177]. To assess the quality of the assembly, we monitored known reference transcripts and noticed that reconstruction quality improves to >90% for transcripts expression levels >10.0 RPKM (**Figure 4.2**). Several examples of accurately reconstructed transcripts are shown, including the known prostate cancer biomarkers *SPINK1* and *PCA3* (**Figure 4.3**).

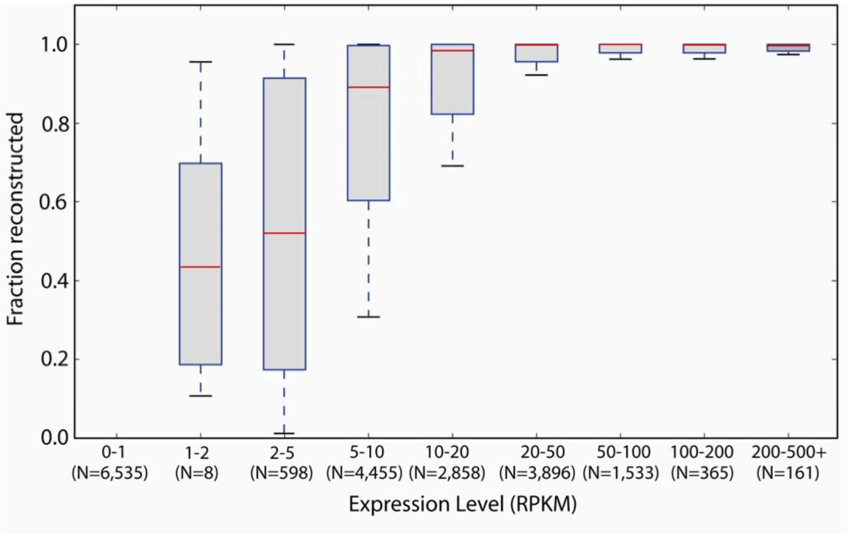


Figure 4.2: Transcriptome reconstruction quality.
We evaluated the quality of our transcriptome reconstruction approach on the set of 20,409 canonical protein-coding genes downloaded from the UCSC Genome Browser version hg18 (knownCanonical). For each UCSC gene, we recorded the transcript with maximal overlap (measured in base pairs) along with the 95th percentile expression level across samples (measured in RPKM). The X axis shows binned expression level windows, and the Y axis shows the fraction of each gene that was encapsulated by our assembly. There were 6,535 genes that were detected at extremely low levels and were not reconstructed. These genes may not be expressed in prostate epithelial cells. The reconstruction quality improves rapidly at expression levels >10.0 RPKM, suggesting that unannotated transcripts detected at levels >10.0 RPKM are likely to be accurately reconstructed.

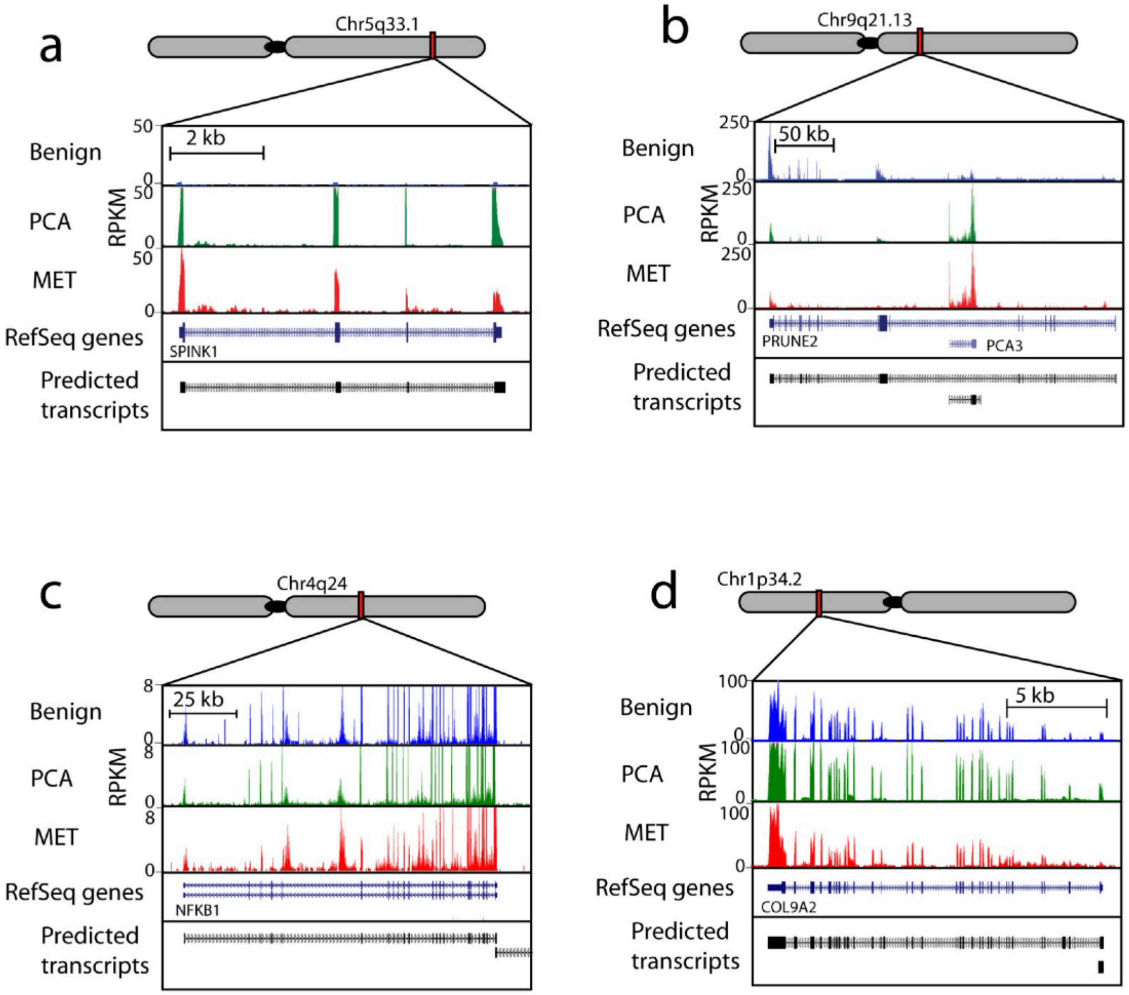


Figure 4.3: Transcript assembly of known genes.
We employed *ab initio* transcript assembly on prostate transcriptome sequencing data to reconstruct the known prostate transcriptome. Four examples of transcriptome reassembly are displayed above. (**a**) *SPINK1*, a biomarker for prostate cancer. (**b)** *PRUNE2* with the *PCA3* non-coding RNA within its intronic regions. Note that *PCA3* is a prostate cancer biomarker while *PRUNE2* is not. The two transcripts remain independent. (**c**) *NFKB1*. (**d**) *COL9A2*.

Discovery of prostate cancer lncRNAs

We compared the assembled prostate cancer transcriptome to the UCSC, Ensembl, RefSeq, Vega and ENCODE gene databases to identify and categorize transcripts. The majority of the transcripts (77.3%) corresponded to annotated protein coding genes (72.1%) and noncoding RNAs (5.2%), but a substantial percentage (19.8%) lacked any overlap and were designated unannotated (**Figure 4.4**). These included partially intronic antisense (2.44%), totally intronic (12.1%) and intergenic transcripts (5.25%), consistent with previous reports of unannotated transcription[5, 178, 179]. Because of the added complexity of characterizing antisense or partially intronic transcripts without strand-specific RNA-Seq libraries, we focused on totally intronic and intergenic transcripts.

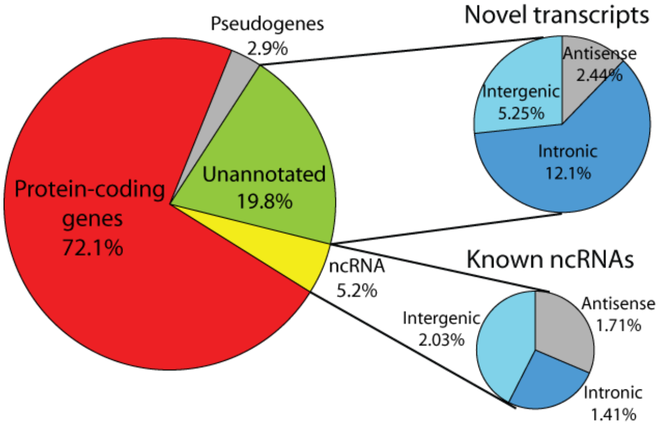


Figure 4.4: Global overview of transcription in prostate cancer.
The pie chart on the left displays transcript distribution in prostate cancer. The pie charts on the right display unannotated (upper) or annotated (lower) ncRNAs categorized as sense transcripts (intergenic and intronic) and antisense transcripts, respectively.

Characterization of unannotated lncRNAs

We extracted the DNA sequences for each transcript and searched for open reading frames (ORFs) using the txCdsPredict program from the UCSC source tool set[180]. This program produces a score corresponding to the protein coding capacity of a given sequence, where scores >800 are ~90% predictive of protein coding genes. We used this threshold to count transcripts with coding potential, and found only 5 of 6,641 unannotated genes with scores >800, compared with 1,669 of 25,414 protein coding transcripts. Additionally, we observed that protein coding genes possess consistently longer ORFs than either unannotated or annotated lncRNA transcripts, suggesting that the vast majority of the unannotated transcripts represent lncRNAs (**Figure 4.5**).

Unannotated transcripts also had greater overlap with expressed sequence tags (ESTs) than randomly permuted controls (**Figure 4.6**).

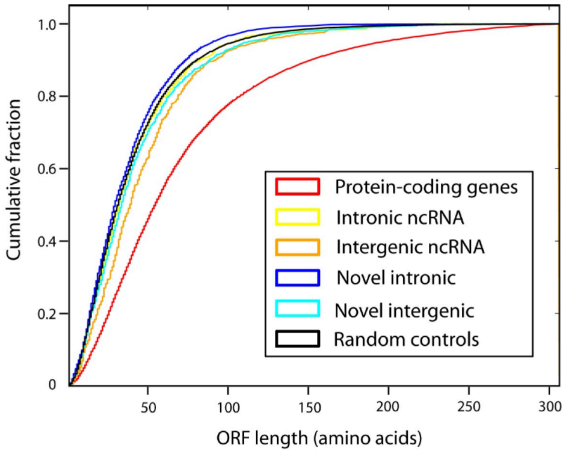
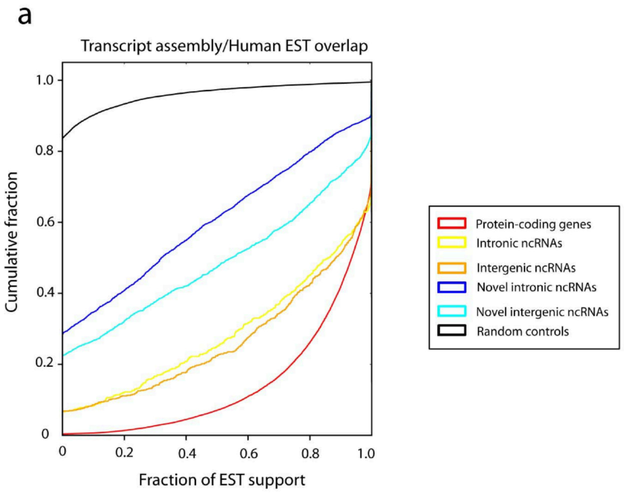


Figure 4.5: Analysis of coding potential of unannotated transcripts.
DNA sequences for each transcript were extracted and searched for open reading frames (ORFs) using the txCdsPredict program from the UCSC source tool set. Using these methods, novel transcripts showed poor protein-coding capacity compared to protein-coding genes, and novel transcripts scored similarly to known ncRNAs, suggesting that the vast majority of unannotated transcripts in prostate cancer represent ncRNAs.



b

Category	Transcripts	EST hits	Percent of all ESTs	ESTs per Transcript
Annotated proteins	25550	4564852	56.43%	178.6634834
Intergenic ncRNA	720	32891	0.41%	45.68194444
Intronic ncRNA	500	57015	0.70%	114.03
Unannotated intergenic ncRNA	1859	17478	0.22%	9.40182894
Unannotated intronic ncRNA	4285	197142	2.44%	46.00746791
Total	**35415**	**4869378**		

Figure 4.6: Analysis of EST support for novel transcripts.
ESTs from the UCSC database table "Human ESTs" were used to evaluate the amount of overlap between ESTs and novel transcripts. (**a**) A line graph showing the fraction of genes whose transcripts are supported by a particular fraction of ESTs. Over 20% of novel transcripts have no support by ESTs. (**b**) A table displaying the number of ESTs supporting each class of transcripts. Percent of all ESTs was calculated using the total number of annotated ESTs (8,089,356), not the total number of observed ESTs in the

transcriptome data (4,869,378).

Characterization of unannotated transcripts demonstrated that they were more highly expressed than randomly permuted controls (**Figure 4.7a**). Further, we used the SiPhy package[181] to estimate the locate rate of variation (ω) of non-repetitive transcript exons across 29 placental mammals (**Figure 4.7b**). Unannotated transcripts displayed a clear but subtle increase in conservation over randomly permuted controls (intergenic transcripts P = 2.7 × 10−4 ± 0.0002 for 0.4 < ω < 0.8; intronic transcripts P = 2.6 × 10−5 ± 0.0017 for 0 < ω < 0.4, Fisher's exact test). By contrast, unannotated transcripts scored lower than protein-coding genes for these metrics, which corroborates data in previous reports2, 24. Notably, a small subset of unannotated intronic transcripts showed a profound degree of conservation (**Figure 4.7b, inset**).

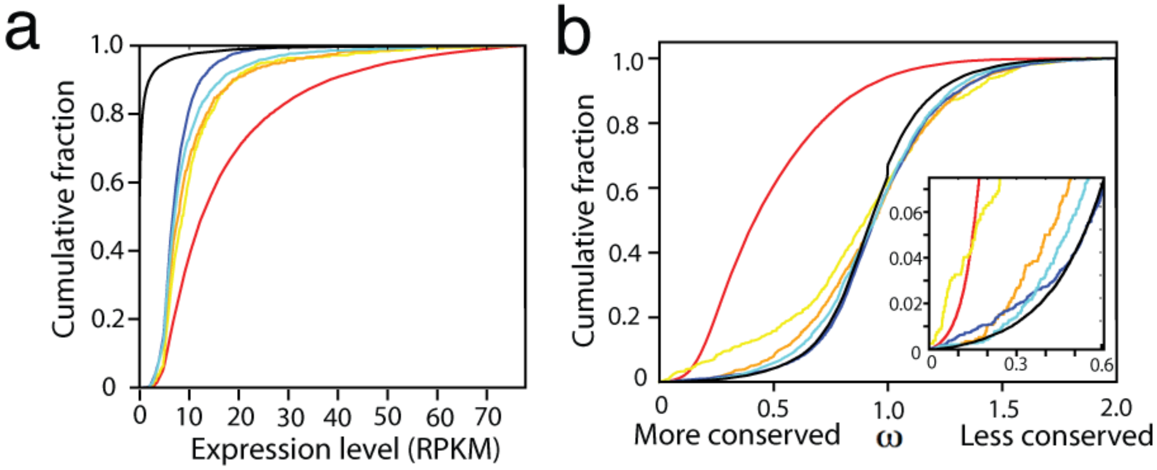


Figure 4.7: Expression and conservation analysis
(a) Line graph showing that unannotated transcripts are more highly expressed (reads per kilobase of transcript per million mapped reads; RPKM) than control regions. Negative control intervals were generated by randomly permuting the genomic positions of the transcripts. (b) Conservation analysis comparing unannotated transcripts to known genes and intronic controls shows a subtle degree of purifying selection among unannotated transcripts. The inset on the right shows an enlarged view.

To determine whether our unannotated transcripts were supported by histone modifications defining active transcriptional units, we used published prostate cancer chromatin immunoprecipitation (ChIP)-Seq data for the prostate cell lines VCaP and LNCaP (GSM353632)[182]. We analyzed the raw ChIP-Seq data (H3K4me2, H3K4me3, Acetylated H3, RNA polymerase II, and Pan-H3) using the MACS peak finder program with default settings[183]. These analyses were performed with the bx-python libraries distributed as part of the Galaxy bioinformatics infrastructure[184]. After filtering our data set for transcribed repetitive elements known to display alternative patterns of histone modifications[185], we observed a strong enrichment for histone modifications characterizing transcriptional start sites (TSSs) and active transcription (**Figure 4.8**). Notably, intergenic lncRNAs showed greater enrichment compared to intronic lncRNAs in these analyses.

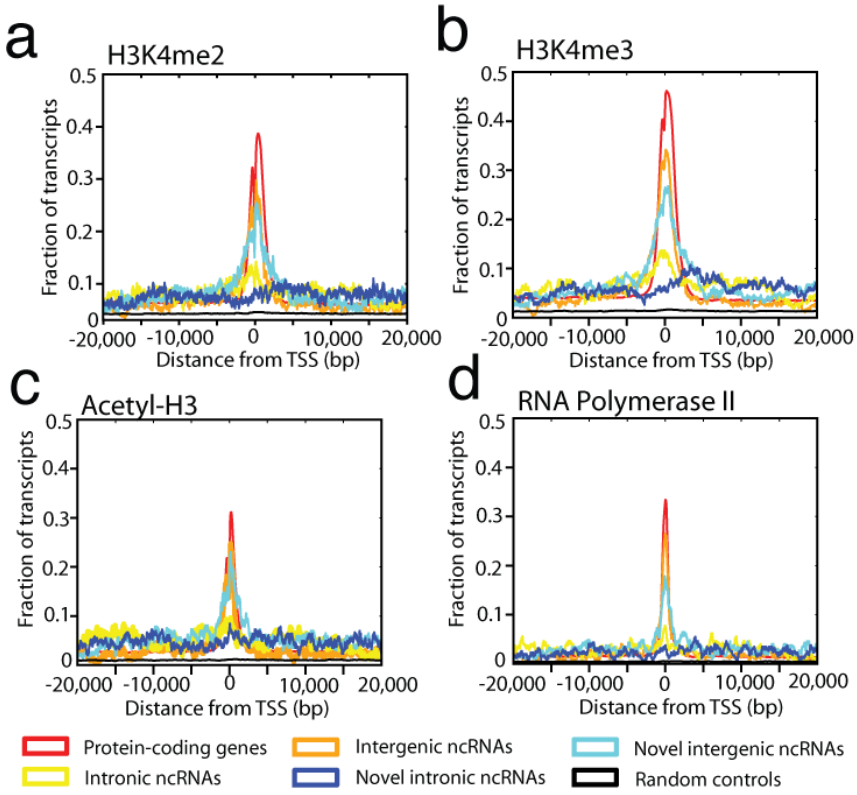


Figure 4.8: ChIP-Seq data supports active transcription of unannotated lncRNAs
Intersection plots displaying the fraction of unannotated transcripts enriched for H3K4me2 (a), H3K4me3 (b), acetyl-H3 (d) or RNA polymerase II (e) at their transcriptional start site (TSS) using ChIP-Seq and RNA-Seq data for the VCaP prostate cancer cell line.

Differential Expression Analysis

To elucidate global changes in transcript abundance in prostate cancer, we analyzed differential expression for all transcripts. We first prepared a matrix of log2-transformed, normalized RPKM expression values after adding a nominal constant 0.1 to all RPKM values. After centering by subtracting the median expression of the benign samples from each transcript, we used the Significance Analysis of Microarrays (SAM) method[186] with 250 permutations of the Tusher *et al*. S0 selection method to predict differentially expressed genes. We chose a delta value corresponding to the 90th percentile FDR desired for individual analyses. The MultiExperiment Viewer application[187] was used to run SAM and generate heatmaps. We found 836 genes differentially expressed between benign samples and localized tumors (false-discovery rate (FDR) < 0.01), with annotated protein-coding and lncRNA genes constituting 82.8% and 7.4% of differentially expressed genes, respectively, including known prostate cancer biomarkers such *AMACR*[188], *HPN*[189] and *PCA3*[174] (**Figure 4.9**). Finally, 9.8% of differentially expressed genes corresponded to unannotated ncRNAs, including 3.2% within gene introns and 6.6% in intergenic regions.

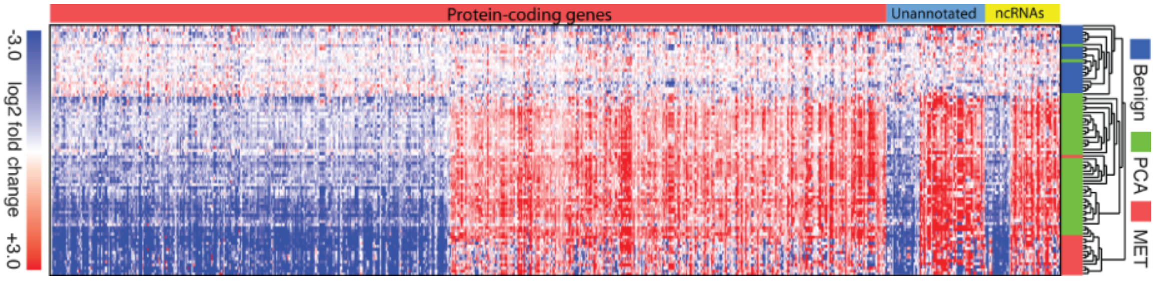


Figure 4.9: Differentially expressed genes in prostate cancers.
A heatmap generated by unsupervised clustering yields 836 differentially expressed transcripts in prostate cancer. Expression is plotted as log2 fold change relative to the median of the benign samples. Transcripts are organized by class (annotated proteins, annotated ncRNAs, unannotated RNAs). Red: upregulated compared to benign; Blue: downregulated compared to benign.

Nomination of Prostate Cancer Associated Transcripts (PCATs)

As lncRNAs may contribute to human disease[166, 169], we identified aberrantly expressed uncharacterized lncRNAs in prostate cancer. We found a total of 1,859 unannotated lncRNAs throughout the human genome. Overall, these intergenic RNAs resided approximately halfway between two protein coding genes (**Figure 4.10**), and over one-third (34.1%) were ≥10 kb from the nearest protein-coding gene, which is consistent with previous reports[190] and supports the independence of intergenic lncRNAs genes. For example, visualizing the Chr15q arm using the Circos program[191] illustrated genomic positions of 89 unannotated intergenic transcripts, including one differentially expressed gene centromeric to *TLE3* (**Figure 4.11**) that we validated by RT-PCR.

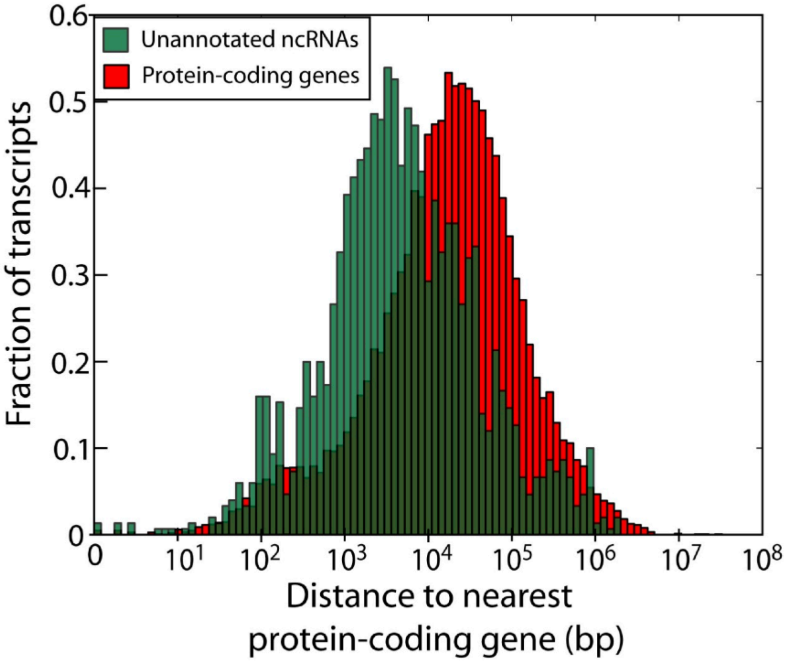


Figure 4.10: Distribution of distances between intergenic unannotated genes and protein-coding genes

The distance between intergenic unannotated ncRNAs to the closest protein-coding gene was calculated, forming a normal distribution around a mean of 4,292 kb. For comparison the distance between protein-coding genes was likewise calculated, forming a normal distribution around a mean of 8,559 kb. These data suggest that, on average, novel intergenic genes are located approximately halfway between protein-coding genes

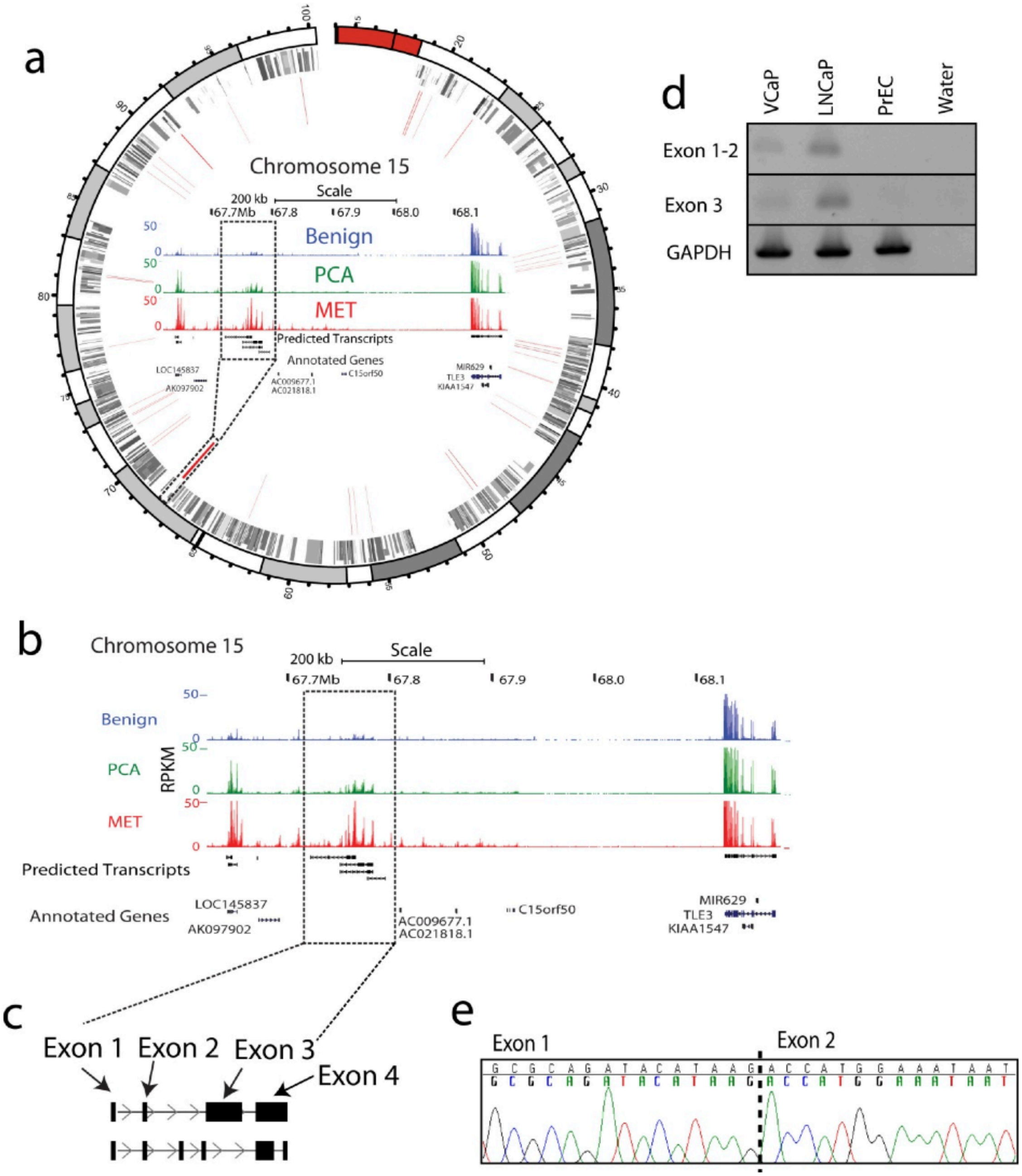


Figure 4.11: Validation of a novel transcript on chromosome 15.
(a) A Circos plot displaying the location of annotated transcripts (grey middle ring) and unannotated transcripts (red inner ring) on Chr15q. Annotated transcripts are represented by 4 individual grey rings representing Ensembl, RefSeq, UCSC, and ENCODE annotations. Unannotated transcripts are widely distributed across the chromosomal arm. Intensity of color indicates relative expression level. (b) Coverage maps showing the average expression levels (RPKM) across the benign, localized tumor, and metastatic samples shows upregulation of a novel transcript. (c) Several predicted isoforms of this transcript were nominated which retained common exons 1 and 2. (d) The exon-exon boundary between exons 1 and 2, as well as an internal portion of exon 3, was validated by RT-PCR in prostate cell line models. (e) Sanger sequencing of the RT-PCR product confirmed the junction of exon 1 and exon 2.

A focused analysis of the 1,859 unannotated intergenic RNAs yielded 106 that were differentially expressed in localized tumors (FDR < 0.05, **Figure 4.12a**). We also applied the Cancer Outlier Profile Analysis (COPA) procedure on the tissue samples and nominated numerous unannotated ncRNA outliers (**Figure 4.12b**) as well as known prostate cancer outliers, such as ERG[137], ETV1[137, 176], SPINK1[192] and CRISP3[193]. Merging these results produced a set of 121 unannotated transcripts that accurately discriminated benign, localized tumor and metastatic prostate samples by unsupervised clustering (**Figure 4.12a**). These transcripts were ranked and named as PCATs according to their fold-change in localized tumor versus benign tissue (**Appendix E**).

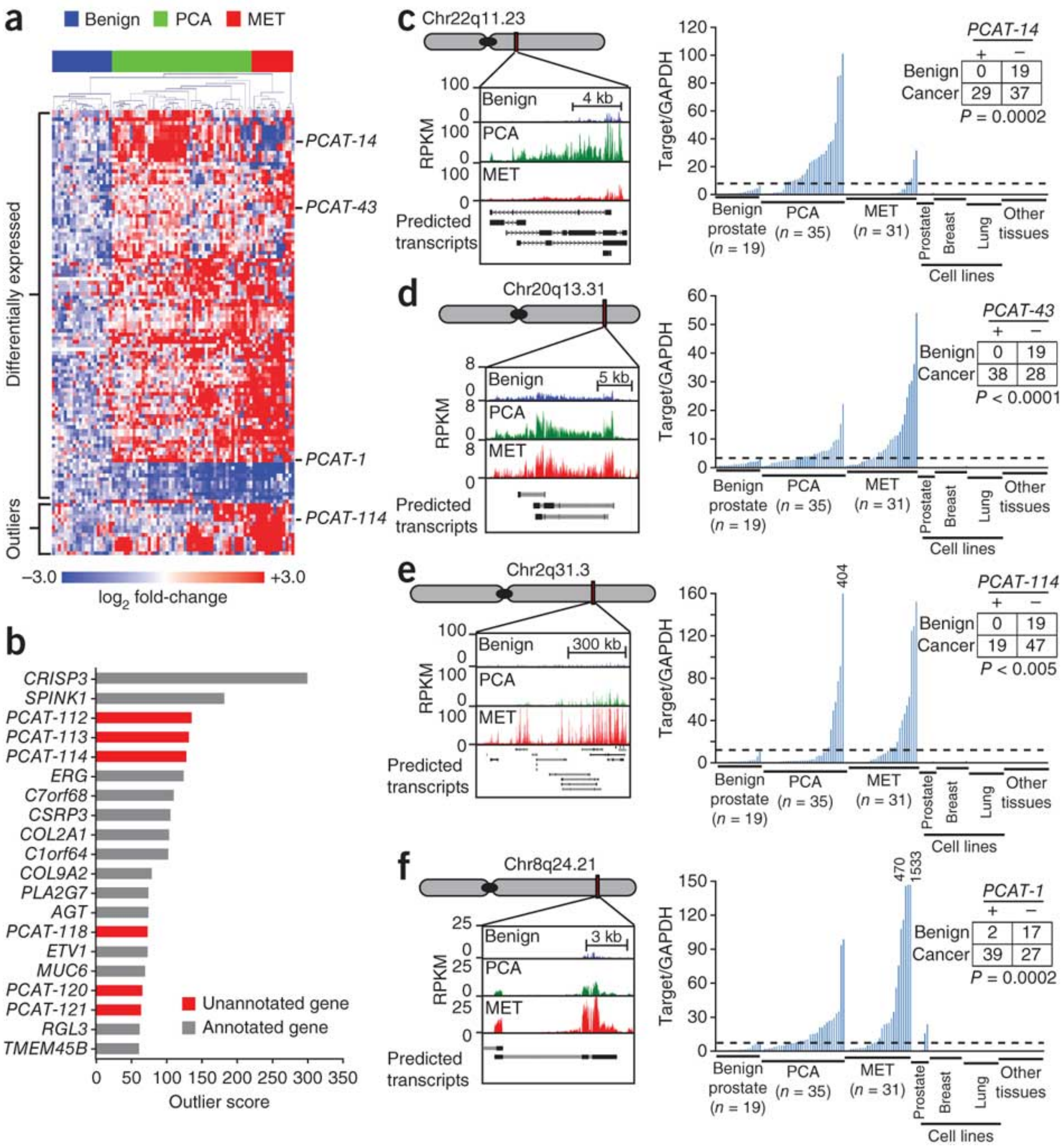


Figure 4.12: Unannotated intergenic transcripts differentiate prostate cancer and benign prostate samples.

(a) Unsupervised clustering analyses of differentially expressed or outlier unannotated intergenic transcripts clusters benign samples, localized tumors and metastatic cancers. Expression is plotted as log2 fold-change relative to the median of the benign samples. The four transcripts detailed in this study are indicated on the side. (b) Cancer outlier expression analysis for the prostate cancer transcriptome ranks unannotated transcripts prominently. (c–f) qPCR on an independent cohort of prostate and nonprostate samples (benign (n = 19), PCA (n = 35), metastatic (MET) (n = 31), prostate cell lines (n = 7), breast cell lines (n = 14), lung cell lines (n = 16), other normal samples (n = 19)) measures expression levels of four nominated ncRNAs—PCAT-14 (c), PCAT-43 (d), PCAT-114 (e), PCAT-1 (f)—and upregulated in prostate cancer. Inset tables on the right quantify 'positive' and 'negative' expressing samples using the cut-off value (shown as a black dashed lines). Statistical significance was determined using a Fisher's exact test. qPCR analysis was performed by normalizing to GAPDH and the median expression of the benign samples.

Validation of novel lncRNAs

To gain confidence in our transcript nominations, we validated multiple unannotated transcripts *in vitro* by reverse transcription PCR (RT-PCR) and quantitative real-time PCR (qPCR). Assays for four transcripts (PCAT-114, PCAT-14, PCAT-43 and PCAT-1) on two independent cohorts of prostate tissues confirmed predicted cancer-specific expression patterns (**Figure 4.12c-f**). Notably, all four are prostate-specific, with minimal expression seen by qPCR in breast (n = 14) or lung cancer (n = 16) cell lines or in 19 normal tissue types. This is further supported by expression analysis of these transcripts in our RNA-Seq compendium of 13 tumor types, representing 325 samples (**Figure 4.13**). This tissue specificity was not necessarily due to regulation by androgen receptor signaling, as only PCAT-14 expression was induced when androgen responsive VCaP and LNCaP cells were treated with the synthetic androgen R1881, consistent with previous data from this locus[176] (**Figure 4.14**). PCAT-1 and PCAT-14 also showed cancer-specific upregulation when tested on a panel of matched tumor-normal pair samples (**Figure 4.15**).

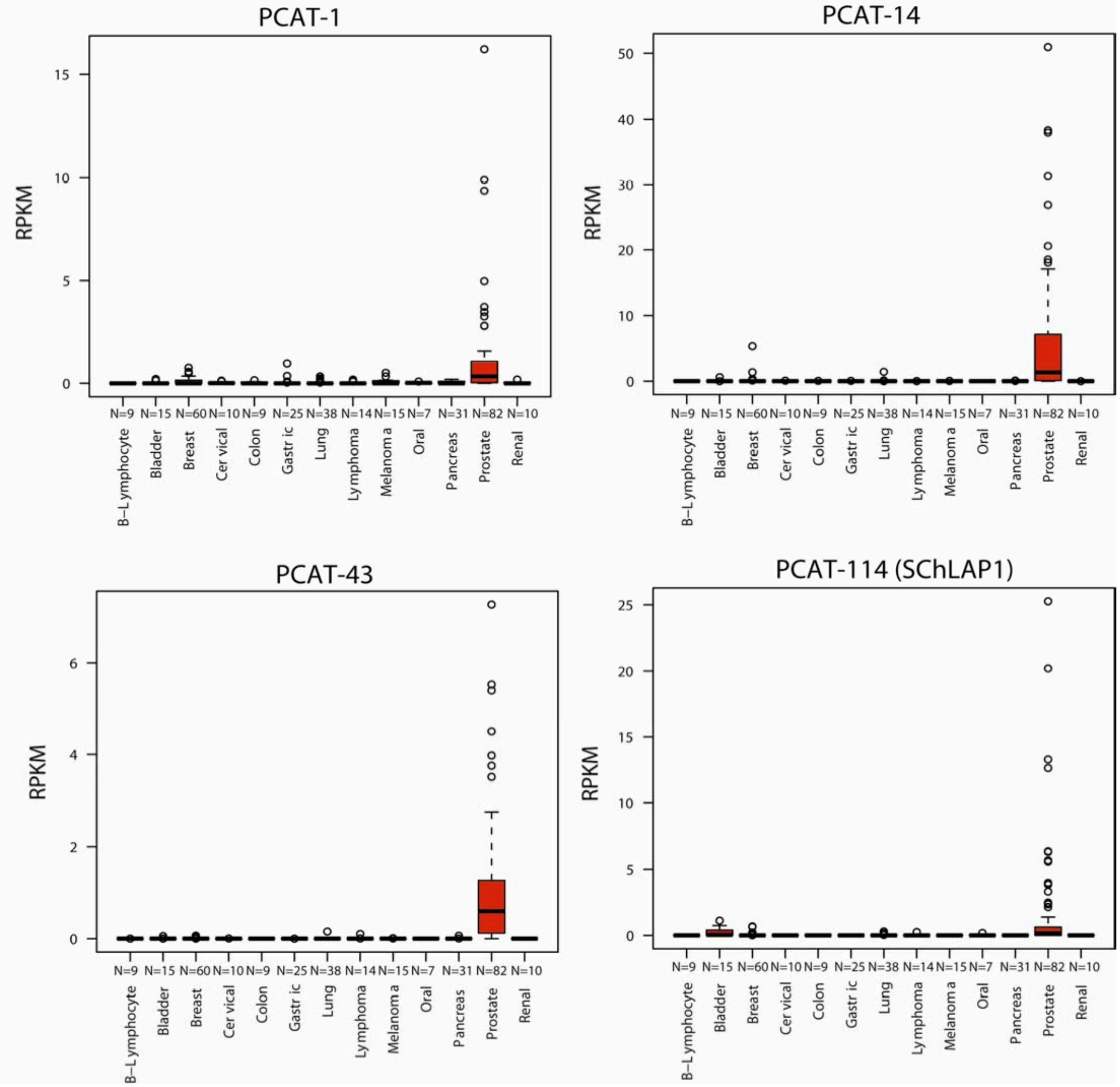


Figure 4.13: Expression of PCATs across tissue types.

RNA- Seq was performed on a compendium of 325 cell lines and tissues, and gene expression was quantified in RPKM. Evaluation of expression levels for PCAT-1, PCAT-14, PCAT-114, or PCAT-43 indicates prostate-specific expression of these transcripts.

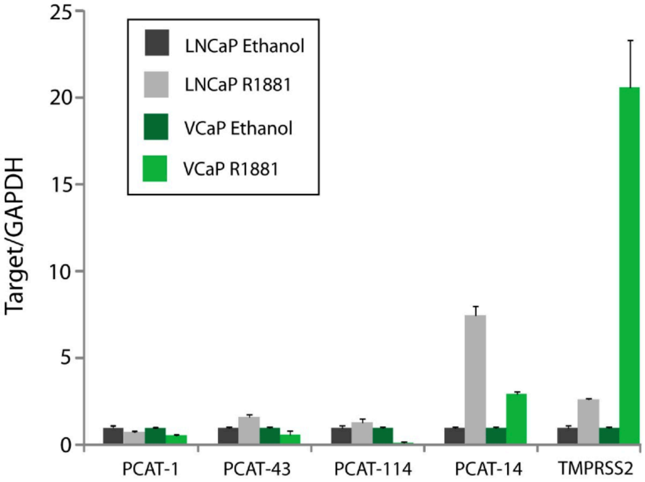


Figure 4.14: PCAT-14 is upregulated by androgen signaling.
VCaP and LNCaP cells were grown in charcoal-stripped serum for 48 hours prior to treatment with 5nM R1881 or vehicle (ethanol) control. 24 hours after treatment, cells were harvested, total RNA and cDNA were generated, and cells were assayed for expression levels of unannotated non- coding RNAs. No consistent change is seen in PCAT-1, PCAT-43 and PCAT-114 expression upon addition of R1881. However, PCAT-14 shows consistent upregulation in both VCaP and LNCaP cells treated with R1881. TMPRSS2 serves as a control androgen-regulated gene. All experiments are normalized to GAPDH and the relative expression of the corresponding ethanol- treated sample. All error bars are mean ± S.E.M.

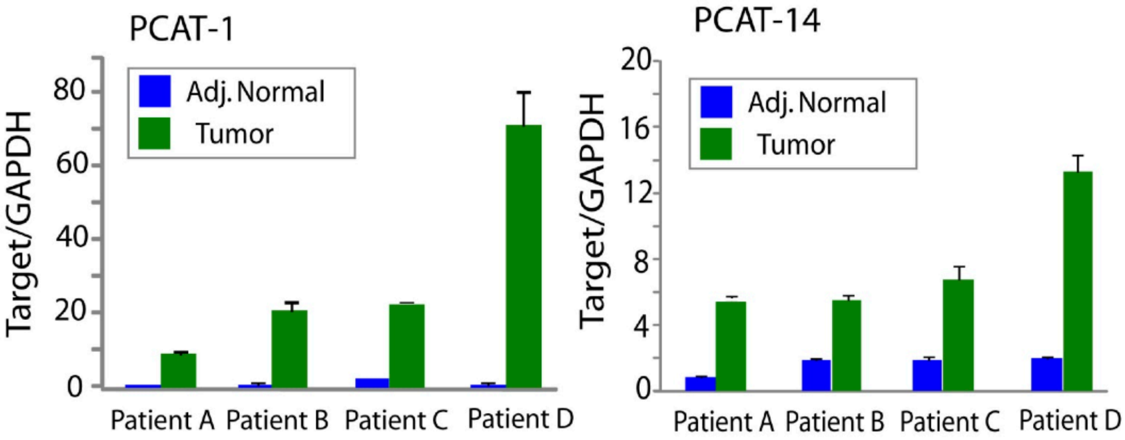


Figure 4.15: PCAT-1 and PCAT-14 are upregulated in matched tumor tissues.
Four matched tumor-normal patient tissue samples were assayed for PCAT-1 and PCAT-14 expression by qPCR. Expression levels were normalized to GAPDH and to the median of the benign samples. All error bars are mean ± S.E.M.

Of note, PCAT-114, which ranks as the fifth best outlier, just ahead of *ERG* (**Figure 4.12b** and **Appendix E**), appears as part of a large, >500 kb locus of expression in a gene desert in Chr2q31. We termed this region 'second chromosome locus associated with prostate (SChLAP) (**Figure 4.16**). Careful analysis of the SChLAP locus revealed both discrete transcripts and intronic transcription, highlighting this region as an intriguing aspect of the prostate cancer transcriptome.

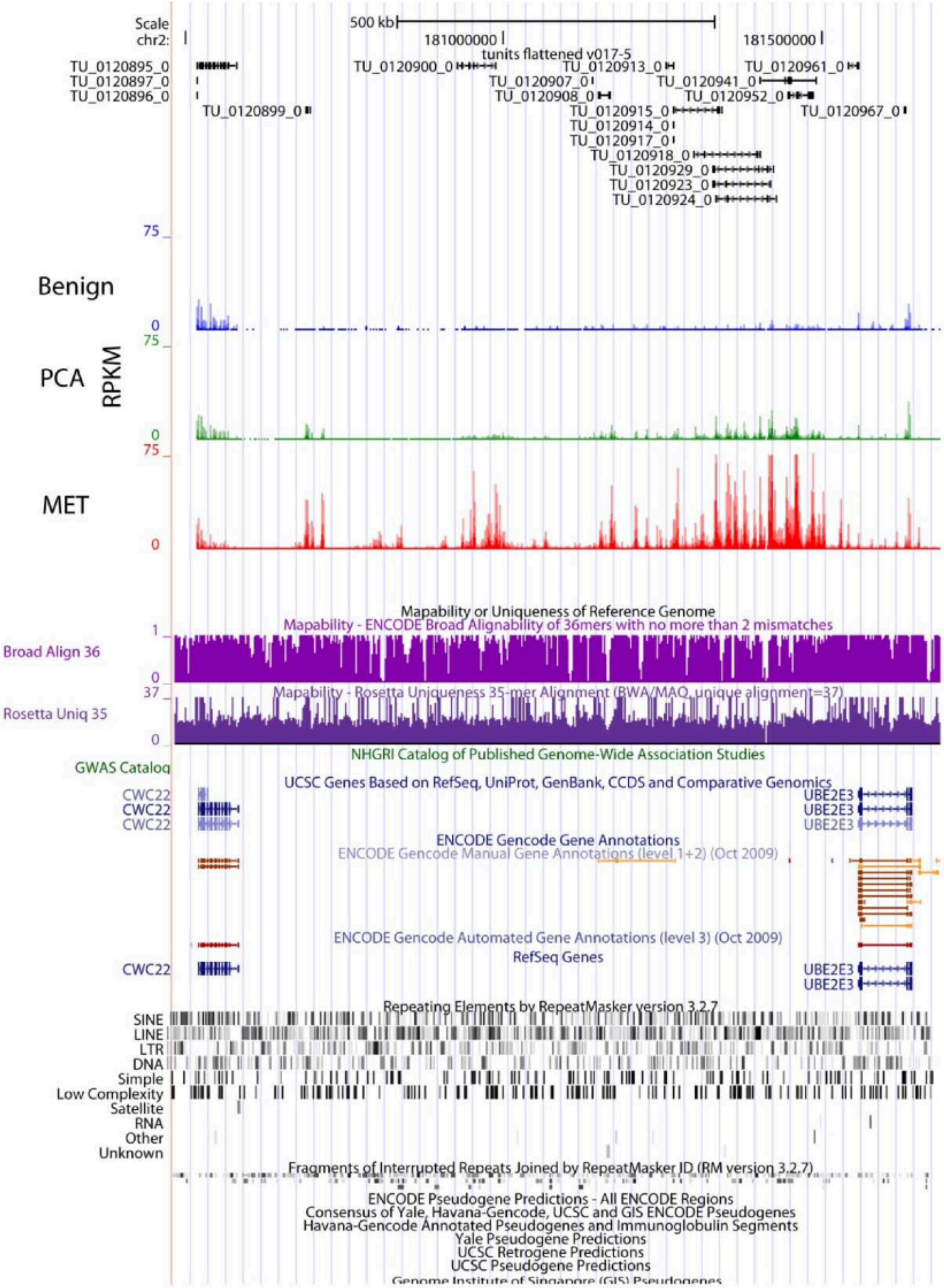

Scale
chr2:
500 kb
181000000
181500000
tunits flattened v017-5
TU_0120895_0
TU_0120897_0
TU_0120896_0
TU_0120899_0
TU_0120900_0
TU_0120907_0
TU_0120908_0
TU_0120913_0
TU_0120915_0
TU_0120914_0
TU_0120917_0
TU_0120918_0
TU_0120929_0
TU_0120923_0
TU_0120924_0
TU_0120941_0
TU_0120952_0
TU_0120961_0
TU_0120967_0
75
Benign
0
75
PCA
RPKM
0
75
MET
0
Mapability or Uniqueness of Reference Genome
Mapability - ENCODE Broad Alignability of 36mers with no more than 2 mismatches
1
Broad Align 36
0
37
Mapability - Rosetta Uniqueness 35-mer Alignment (BWA/MAQ, unique alignment=37)
Rosetta Uniq 35
0
NHGRI Catalog of Published Genome-Wide Association Studies
GWAS Catalog
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CWC22
CWC22
CWC22
UBE2E3
UBE2E3
UBE2E3
ENCODE Gencode Gene Annotations
ENCODE Gencode Manual Gene Annotations (level 1+2) (Oct 2009)
ENCODE Gencode Automated Gene Annotations (level 3) (Oct 2009)
RefSeq Genes
CWC22
UBE2E3
UBE2E3
Repeating Elements by RepeatMasker version 3.2.7
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Fragments of Interrupted Repeats Joined by RepeatMasker ID (RM version 3.2.7)
ENCODE Pseudogene Predictions - All ENCODE Regions
Consensus of Yale, Havana-Gencode, UCSC and GIS ENCODE Pseudogenes
Havana-Gencode Annotated Pseudogenes and Immunoglobulin Segments
Yale Pseudogene Predictions
UCSC Retrogene Predictions
UCSC Pseudogene Predictions
Genome Institute of Singapore (GIS) Pseudogenes

Figure 4.16: The SChLAP locus spans >500 kb.
Visualization of transcriptome sequencing data in the UCSC genome browser indicates that a large, almost 1 Mb section of chromosome 2 is highly activated in cancer, contributing to many individual transcripts

regulated in a coordinated fashion. For reference, the two flanking protein-coding genes, *UBE2E3* and *CWC22*, are shown. Neither *UBE2E3* nor *CWC22* appears differentially regulated in prostate cancer.

Discussion

At the time of publication, this study represented the largest RNA-Seq analysis and the first to comprehensively analyze a common epithelial cancer from a large cohort of human tissue samples. As such, our study has adapted existing computational tools intended for small-scale use[26] and developed new methods to distill large numbers of transcriptome data sets into a single consensus transcriptome assembly that accurately represents disease biology. The bioinformatics methods presented here formed the foundation for the AssemblyLine software package described in Chapter 3 of this thesis.

Among the numerous uncharacterized lncRNA species detected by our study, we have focused on 121 PCATs, which we believe represent a set of uncharacterized lncRNAs that may have important biological functions in this disease. In this regard, these data contribute to a growing body of literature supporting the importance of unannotated lncRNA species in cellular biology and oncogenesis[166, 169-172] and broadly our study confirms the utility of RNA-Seq in defining functionally important elements of the genome[26, 47, 50]. Of particular interest is our discovery of the prostate-specific lncRNA gene *PCAT-1*, which is markedly overexpressed in a subset of prostate cancers, particularly metastases, as well as the *SChLAP1* locus (a region of chromosome 2 containing *PCAT-111*, *PCAT-114*, and *PCAT-118*, among others), which features prominent overexpression in 15-30% of the disease. We describe our extensive follow-up studies of these two genes as Chapter 5 of this thesis.

Recent preclinical efforts to detect prostate cancer noninvasively through the collection of patient urine samples have shown promise for several urine-based prostate cancer biomarkers, including the lncRNA *PCA3*[194, 195]. Although additional studies are needed, our identification of lncRNA biomarkers for prostate cancer suggests that urine-based assays for these lncRNAs may also warrant investigation, particularly for those that may stratify patient molecular subtypes.

Our findings support an important role for tissue-specific lncRNAs in prostate cancer and suggest that cancer-specific functions of these lncRNAs may help to drive tumorigenesis. We further speculate that specific lncRNA signatures may occur universally in all disease states and that applying these methodologies to other diseases may reveal key aspects of disease biology and clinically important biomarkers.

Chapter 5: Long non-coding RNAs coordinate the pathogenesis of prostate cancer

AUTHORS: Prensner JR*, Iyer MK*, Sahu A*, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche RJ, Malik R, Bedenis R, McGregor N, Ma T, Chen W, Han S, Jing X, Cao X, Wang X, Chandler B, Yan W, Siddiqui J, Kunju LP, Dhanasekaran SM, Pienta KJ, Feng FY, Chinnaiyan AM.

* These authors made equal contributions

AUTHOR CONTRIBUTIONS: J.R.P., M.K.I., A.S. and A.M.C. designed the project and directed experimental studies. J.R.P, Q.C., W.C., S.M.D., B.C., S.H., R.M., L.P., T.M. and A.S. performed *in vitro* studies. X.W. performed *in vitro* translation assays. I.A.A. and A.S. performed CAM assays. R.B., N.M. and K.P. performed *in vivo* studies. L.P.K. and W.Y. performed histopathological analyses. M.K.I. performed bioinformatics analysis. X.J. and X.C. performed gene expression microarrays. J.S. and F.Y.F. facilitated biological sample procurement. F.Y.F. performed clinical analyses. For the Mayo Clinic Cohort, R.B.J. provided clinical samples and outcomes data. T.J.T. and E.D. generated and analyzed expression profiles for the Mayo Clinic cohort. E.D., N.E., M.G., and I.A.V. performed statistical analyses of *SChLAP1* expression in the Mayo Clinic cohort. J.R.P., M.K.I., A.S. and A.M.C. interpreted data and wrote the manuscript.

CITATION: Prensner JR*, Iyer MK*, Sahu A*, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M, Jenkins RB, Triche RJ, Malik R, Bedenis R, McGregor N, Ma T, Chen W, Han S, Jing X, Cao X, Wang X, Chandler B, Yan W, Siddiqui J, Kunju LP, Dhanasekaran SM, Pienta KJ, Feng FY, Chinnaiyan AM. The lncRNA SChLAP1 coordinates aggressive prostate cancer and antagonizes the SWI/SNF complex. *Manuscript in press at Nature Genetics*.

* These authors made equal contributions

Abstract

Prostate cancer is a clinically heterogeneous disease in which only a subset of patients has aggressive cancer whereas others have indolent disease[89, 196, 197]. However, the molecular basis for this heterogeneity remains incompletely understood[198, 199]. Previously, we used RNA sequencing and transcriptome assembly methods to define novel transcripts overexpressed in subsets of prostate cancer[27]. Here, we characterize two of these lncRNAs, *PCAT1* and *SChLAP1* (Second Chromosome Locus Associated with Prostate-1). *PCAT1*, a multi-exonic lncRNA expressed in a 'gene desert' on chromosome 8q24, promotes cell proliferation through transcriptional regulation of target genes and represses the tumor suppressor *BRCA2*. *SChLAP1*, one of several multi-exonic lncRNAs expressed in a chromosome 2q31 'gene desert', independently predicts for poor patient outcomes, including metastasis and cancer specific mortality. Mechanistically, *SChLAP1* antagonizes the genome-wide localization and regulatory functions of the SWI/SNF chromatin-modifying complex. These results establish the role of lncRNAs in

coordinating aggressive prostate cancer and as potential prognostic biomarkers for the disease.

Introduction

Prostate cancer is the most common non-cutaneous cancer in U.S. men, with over 200,000 prostate cancer diagnoses per year[89]. However, while 1 in 6 men are diagnosed with prostate cancer, only 1 in 32 men die from this disease[196, 200], and it is estimated that only 20% of prostate cancer patients have a high-risk cancer[197]. Thus, most prostate cancer patients die with their disease, but not from it. While mutational events in key genes characterizes a subset of lethal prostate cancers[198, 199, 201], the molecular basis for aggressive disease remains poorly understood.

Long non-coding RNAs (lncRNAs) are polyadenylated RNA species >200bp in length commonly characterized by splicing of multiple exons, H3K4me3 promoter methylation, and transcription by RNA polymerase II[79, 202]. lncRNA-mediated biology has been implicated in a wide variety of cellular processes, including pluripotency in stem cells[203] and X chromosome inactivation[204]. In cancer, lncRNAs are emerging as a prominent layer of previously underappreciated transcriptional regulation, often by collaborating with epigenetic complexes such as Polycomb Repressive Complex 1[171, 173] (PRC1) and Polycomb Repressive Complex 2 (PRC2)[169, 172, 173, 205], among others. Despite reports showing that upregulation of the lncRNA *HOTAIR* participates in PRC2 function in breast cancer[169], we do not observe strong expression of this lncRNA in prostate (**Figure 5.1**), suggesting that other lncRNAs may be important in this cancer.

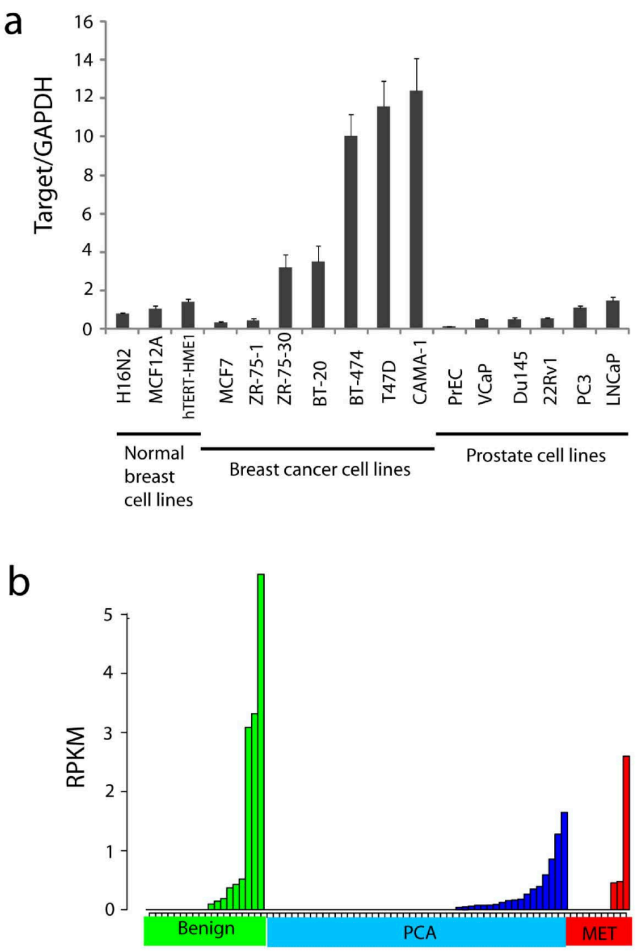


Figure 5.1: HOTAIR is not upregulated in prostate cancer.
(a) qPCR analysis of a panel of breast cell lines and prostate cell lines for HOTAIR expression shows upregulation of *HOTAIR* in numerous breast cancer cell lines but not prostate cancer cell lines. Expression levels are normalized to *GAPDH* and the median expression of benign breast cell lines. (b) RPKM expression levels of *HOTAIR* in the prostate RNA-Seq cohort. Preferential upregulation of *HOTAIR* is not observed in prostate cancer and metastases samples.

We hypothesized that prostate cancer aggressiveness was governed by uncharacterized lncRNAs and sought to discover lncRNAs whose expression characterized the subset of prostate cancer patients with aggressive disease. In Chapter 5, we described our use of RNA-Seq profiling across a prostate cancer cohort to discover 121 lncRNA loci (out of >1,800) that were aberrantly expressed in this disease[27]. Of particular interest was our discovery of the prostate-specific lncRNA gene *PCAT-1*, which was markedly over-expressed in a subset of prostate cancers, particularly metastases (**Figure 4.12f**). *PCAT-1* resides in the well-studied 8q24 genome 'hotspot' associated with development of prostate cancer[206, 207]. Additionally, we observed that a group of PCATs, including *PCAT-109*, *PCAT-114*, and *PCAT-118* showed striking outlier profiles and ranked among the best outliers in prostate cancer when compared to protein-coding genes (**Figure 4.12b,e**). These PCATs were localized in a "gene desert" region on chromosome 2q31.3 with previously unknown ties to prostate cancer that we subsequently named the SChLAP locus (**Figure 4.16**). Here, we explore these loci more closely and discover remarkable roles for lncRNAs in coordinating the pathogenesis of prostate cancer.

PCAT-1 represses distinct target genes and controls cell cycle proliferation

To interrogate *PCAT-1*, we carried out 5' and 3' rapid amplification of cDNA ends (RACE) to define the full-length *PCAT-1* transcript. We cloned the full-length *PCAT-1* transcript and performed *in vitro* translational assays, which were negative as expected (**Figure 5.2**).

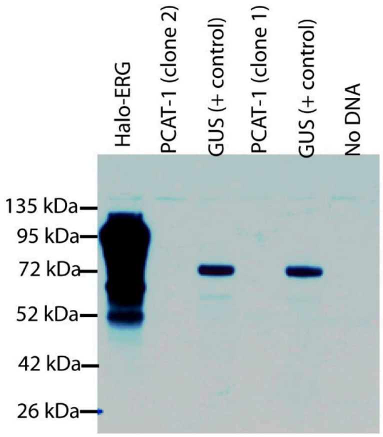


Figure 5.2: ***In vitro*** **translation of PCAT-1 confirms ncRNA status.**
Full length PCAT-1 transcript was cloned into the PCR2.1 vector (Invitrogen) and expressed using the TnT Quick Coupled Transcription/Translation System (Promega). Western blot analysis resolving the proteins by SDS-PAGE indicated that PCAT-1 is a non-coding RNA with no protein-coding capacity. GUS and ERG protein *in vitro* translation served as positive controls.

Interestingly, *PCAT-1* incorporates portions of a mariner family transposase[208, 209], an Alu and a viral long terminal repeat promoter region (**Figure 5.3a**). Because *PCAT-1* was strikingly upregulated in a subset of metastatic and high-grade localized (Gleason score ≥7) cancers (**Figure 4.12f**), we hypothesized that *PCAT-1* may have coordinated expression with the oncoprotein *EZH2*, a core PRC2 protein that is upregulated in solid tumors and contributes to a metastatic phenotype[210, 211]. Surprisingly, we found that *PCAT-1* and *EZH2* expression were nearly mutually exclusive (**Figure 5.3b**), with only one patient showing outlier expression of both. This suggests that outlier *PCAT-1* and *EZH2* expression may define two subsets of high-grade disease. To determine the mechanism for the expression profiles of *PCAT-1* and *EZH2*, we inhibited *EZH2* activity in VCaP cells, which express low-to-moderate levels of *PCAT-1*. Knockdown of *EZH2* by short hairpin (sh)RNA or pharmacologic inhibition of *EZH2* with the inhibitor 3-deazaneplanocin A (DZNep) caused a dramatic upregulation in *PCAT-1* expression levels (**Figure 5.3c,d**), as did treatment of VCaP cells with the demethylating agent 5'-deoxyazacytidine, the histone deacetylase inhibitor SAHA or both (**Figure 5.3e**). ChIP assays also demonstrated that *SUZ12*, a core PRC2 protein, directly binds the *PCAT-1* promoter ~1 kb upstream of the TSS (**Figure 5.3f**). By contrast, LNCaP cells, which express *PCAT-1* at relatively high levels, did not exhibit PRC2-mediated repression of *PCAT-1* (**Figure 5.4a,b**).

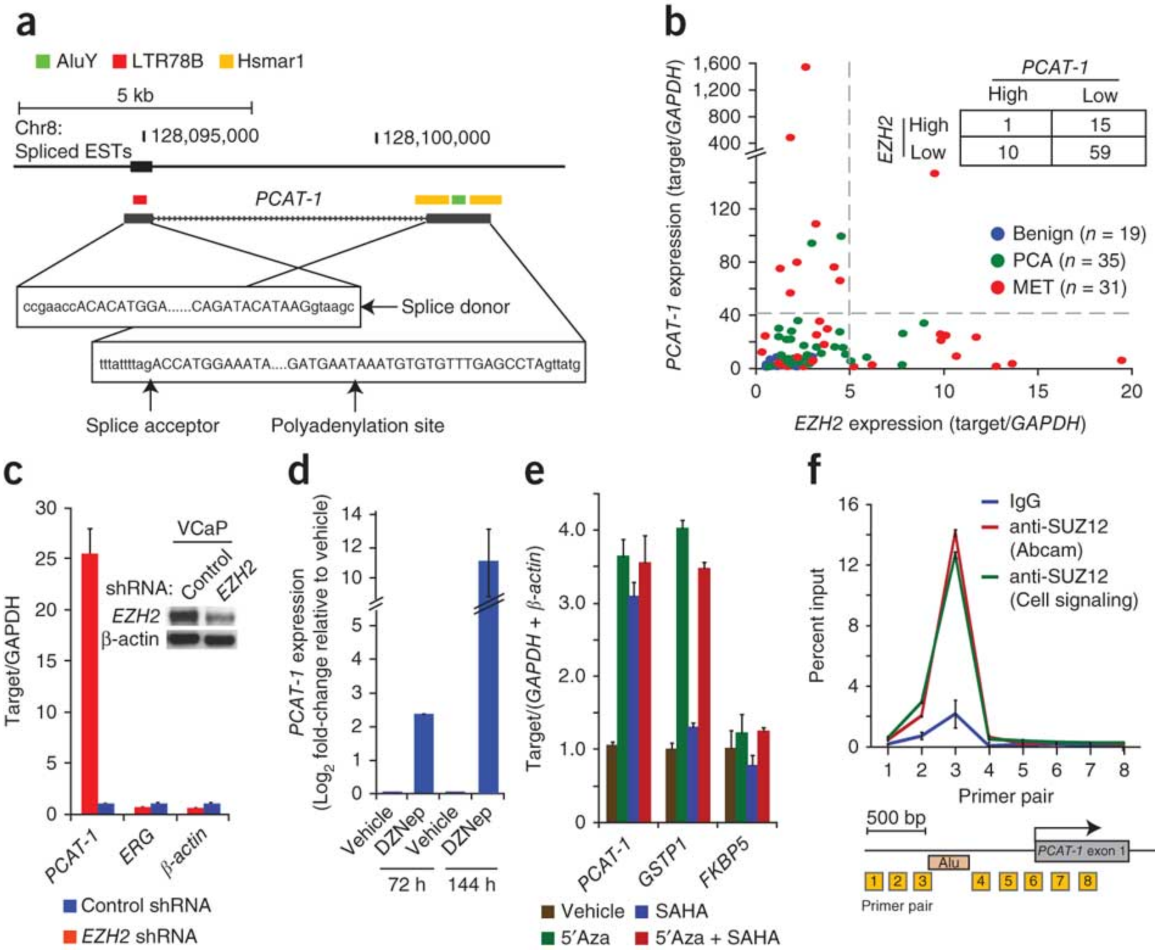


Figure 5.3: PCAT-1 is a marker of aggressive cancer and a PRC2-repressed lncRNA.
(a) The genomic location of PCAT-1 determined by 5' and 3' RACE, with DNA sequence features indicated by the colored boxes. (b) qPCR for PCAT-1 (y axis) and *EZH2* (x axis) on a cohort of benign (n = 19), localized tumor (n = 35) and metastatic cancer (n = 31) samples. The inset table quantifies patient subsets demarcated by the gray dashed lines. (c) Knockdown of *EZH2* in VCaP resulted in upregulation of *PCAT-1*. Data were normalized to *GAPDH* and represented as fold-change. *ERG* and B-actin serve as negative controls. The inset western blot indicates EZH2 knockdown. (d) Treatment of VCaP cells with 0.1 μM of the *EZH2* inhibitor DZNep or vehicle control (DMSO) shows increased expression of *PCAT-1* transcript after *EZH2* inhibition. (e) *PCAT-1* expression is increased upon treatment of VCaP cells with the demethylating agent 5-azacytidine (5-Aza), the histone deacetylase inhibitor SAHA or a combination of both. qPCR data were normalized to the average of (*GAPDH* + *β* -actin) and represented as fold-change. *GSTP1* and *FKBP5* are positive and negative controls, respectively. (f) ChIP assays for *SUZ12* demonstrated direct binding of *SUZ12* to the *PCAT-1* promoter. Primer locations are indicated (boxed numbers) in the *PCAT-1* schematic.

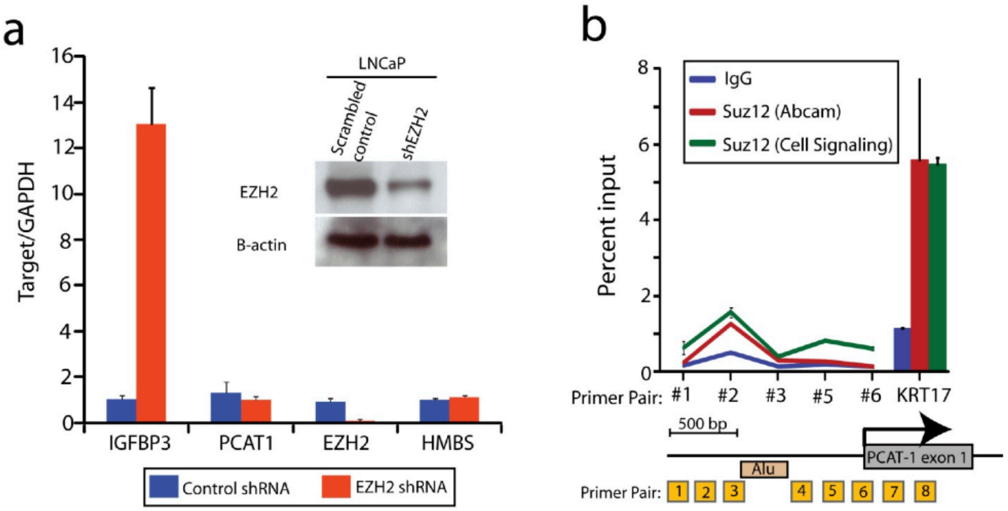


Figure 5.4: PCAT-1 is not a PRC2 target in LNCaP cells.
(a) LNCaP cells were infected with lentivirus for *EZH2* or scrambled control. qPCR showed no change in *PCAT- 1* expression. *IGFBP3* and HMBS serve as positive and negative controls, respectively. (b) ChIP analysis of *SUZ12* in LNCaP cells does not show direct binding of PRC2 proteins upstream of the PCAT-1 transcriptional start site (refer to fig. 3E for comparison). *KRT17* serves as a positive control. (c) RNA-IP analysis of in LNCaP cells does not indicate binding of *PCAT-1* transcript to PRC2 (compare to Supplementary Figure 23). *lincGARS* serves as a positive control. (d) A representative image of *SUZ12* RNA-IP pulldown efficiency. Equal fractions of LNCaP nuclear lysate were treated with either IgG or *SUZ12* antibodies and, following washing, probed for *SUZ12* protein. Treating nuclear lysates with no antibody serves as a negative control.

To explore the functional role of *PCAT-1* in prostate cancer, we stably overexpressed full-length *PCAT-1* or controls in RWPE benign immortalized prostate cells (**Figure 5.4**). Additionally, we designed short interfering (si)RNA oligos to *PCAT-1* and performed knockdown experiments in LNCaP cells (**Figure 5.5**).

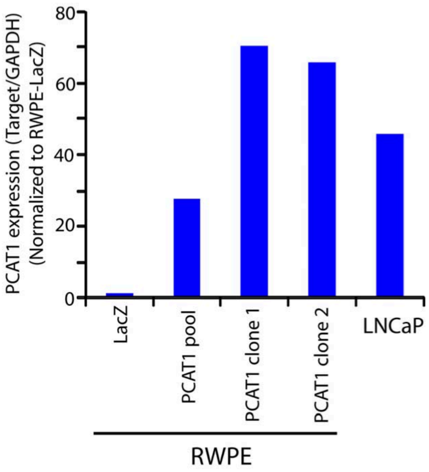


Figure 5.4: *PCAT-1* overexpression in RWPE cells.
Full length *PCAT-1* transcript was cloned into a lentiviral vector, and, following lentivirus production, RWPE benign immortalized prostate cells stably overexpressing *PCAT-1* were generated by selection with blasticidin. *PCAT-1* overexpression compared to the LacZ control cells was confirmed by qPCR. LNCaP serves as a positive control.

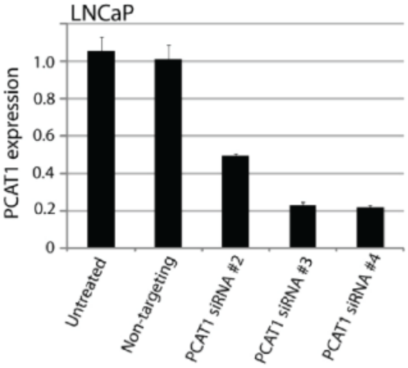


Figure 5.5: qPCR validation of *PCAT-1* knockdown in LNCaP cells

We observed a consistent increase in cell proliferation when *PCAT-1* was overexpressed at physiological levels (**Figure 5.6a**). Supporting our overexpression data, knockdown of *PCAT-1* with three independent siRNA oligos resulted in a 25–50% decrease in cell proliferation in LNCaP cells (**Figure 5.6b**). As expected, knockdown of *PCAT-1* in VCaP cells, in which *PCAT-1* is suppressed by PRC2, did not affect cell proliferation (**Figure 5.7**).

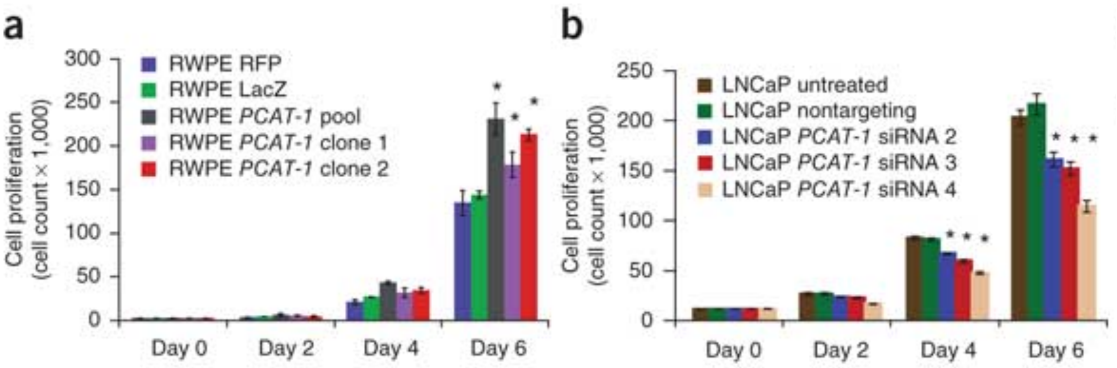


Figure 5.6: PCAT-1 promotes cell proliferation.
(a) Cell proliferation assays for RWPE benign immortalized prostate cells stably infected with *PCAT-1* lentivirus or RFP and LacZ control lentiviruses. An asterisk (*) indicates P ≤ 0.02 by a two-tailed Student's t-test. (b) Cell proliferation assays in LNCaP using *PCAT-1* siRNAs. An asterisk (*) indicates P ≤ 0.005 by a two-tailed Student's t-test. (c) Gene ontology analysis of *PCAT-1* knockdown microarray data using the DAVID program. Blue bars represent the top hits for upregulated genes. Red bars represent the top hits for downregulated genes. DAVID enrichment scores are represented with Benjamini-Hochberg-adjusted P values. All error bars in this figure are mean ± s.e.m.

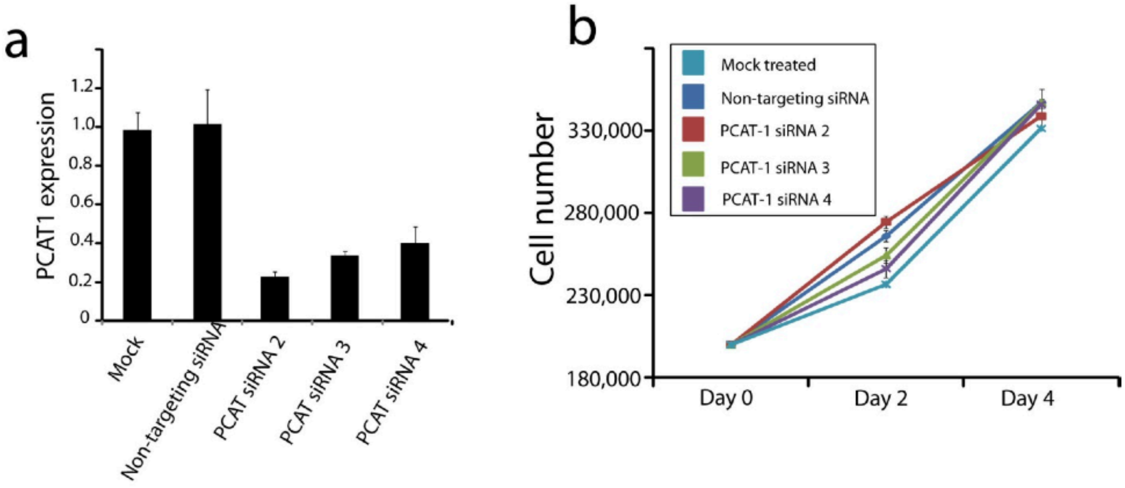


Figure 5.7: *PCAT-1* knockdown in VCAP cells.
VCaP cells were treated with three unique *PCAT-1* siRNA oligos. (a) *PCAT-1* knockdown was confirmed by qPCR. (b) Cell proliferation assays in VCaP cells treated with *PCAT-1* siRNAs show no significant difference from cells treated with controls.

Gene expression profiling of LNCaP knockdown samples on cDNA microarrays indicated that *PCAT-1* modulates the transcriptional regulation of 370 genes (255 upregulated, 115 downregulated; FDR ≤ 0.01) (**Figure 5.8**). Gene ontology analysis of the upregulated genes showed preferential enrichment for gene set concepts such as mitosis and cell cycle, whereas the downregulated genes had no concepts showing statistical significance (**Figure 5.9**). These results suggest that the function of *PCAT-1* is predominantly repressive in nature.

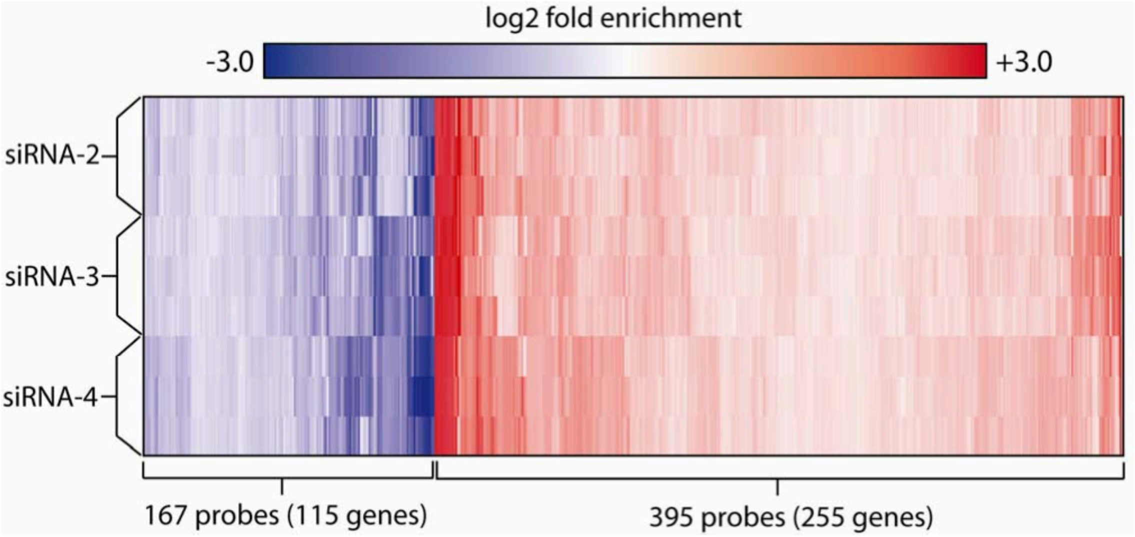


Figure 5.8: Gene expression profiling of PCAT-1 knockdown samples
PCAT-1 knockdown LNCaP samples were run in triplicate against a non-targeting siRNA control sample. Data were analyzed with SAM analysis, producing a signature of 562 differentially-expressed probes (FDR ≤ 0.01; 395 upregulated, 167 downregulated). Concordance between biological and technical replicates was high.

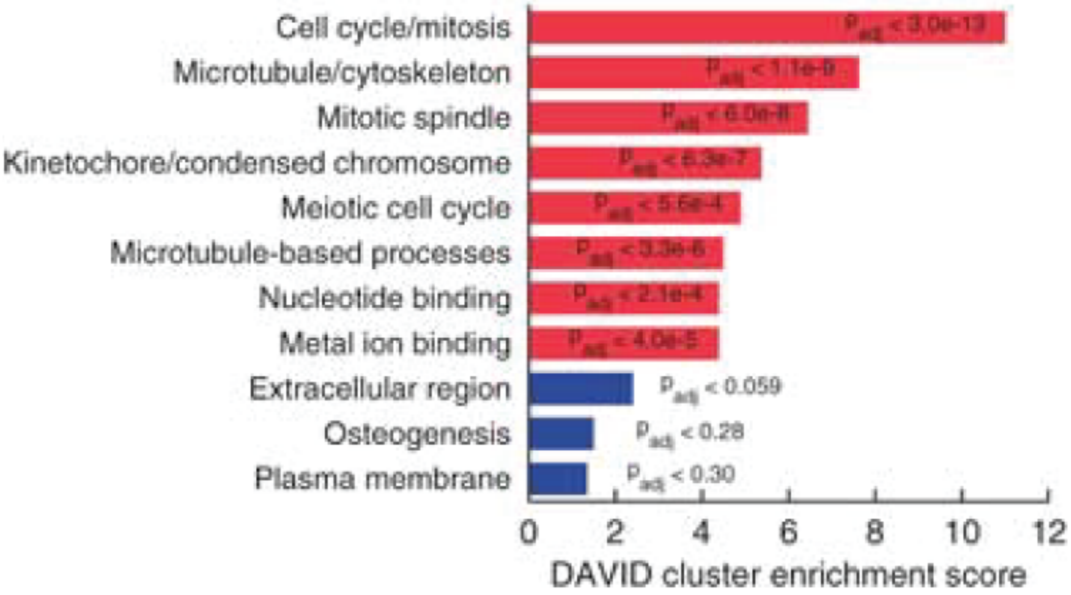


Figure 5.9: Gene ontology analysis of *PCAT-1*.
PCAT-1 knockdown microarray data was analyzed using the DAVID program. Blue bars represent the top hits for upregulated genes. Red bars represent the top hits for downregulated genes. DAVID enrichment

scores are represented with Benjamini-Hochberg-adjusted P values. All error bars in this figure are mean ± s.e.m.

PCAT-1 signatures in prostate cancer

We next validated expression changes in three key *PCAT-1* target genes (*BRCA2*, *CENPE* and *CENPF*) whose expression is upregulated upon *PCAT-1* knockdown (**Figure 5.10a**) in LNCaP and VCaP cells, the latter of which appear less sensitive to *PCAT-1* knockdown likely due to lower overall expression levels of this transcript. Because of the regulation of *PCAT-1* by PRC2 in VCaP cells, we hypothesized that knockdown of *EZH2* would also downregulate *PCAT-1* targets as a secondary phenomenon owing to the subsequent upregulation of *PCAT-1*. Simultaneous knockdown of *PCAT-1* and *EZH2* would thus abrogate expression changes in *PCAT-1* target genes. Carrying out this experiment in VCaP cells demonstrated that *PCAT-1* target genes were indeed downregulated by *EZH2* knockdown, and that this change was either partially or completely reversed using siRNA oligos to *PCAT-1* (**Figure 5.10a**), lending support to the role of *PCAT-1* as a transcriptional repressor. Taken together, these results suggest that *PCAT-1* biology may exhibit two distinct modalities: one in which PRC2 represses *PCAT-1* and a second in which active *PCAT-1* promotes cell proliferation. *PCAT-1* and PRC2 may therefore characterize distinct subsets of prostate cancer.

To examine these findings, we used qPCR to measure expression of *BRCA2*, *CENPE* and *CENPF* in our cohort of tissue samples. Consistent with our model, we found that samples expressing *PCAT-1* tended to have low expression of *PCAT-1* target genes (**Figure 5.10b**). Moreover, comparing *EZH2*-outlier and *PCAT-1*-outlier patients, we found that two distinct phenotypes emerged. Individuals with high *EZH2* tended to have high levels of *PCAT-1* target genes, and those with high expression of *PCAT-1* itself displayed the opposite expression pattern of target genes (**Figure 5.10c**). Network analysis of the top 20 upregulated genes after *PCAT-1* knockdown with the HefaLMP tool[212] further suggested that these genes form a coordinated network (**Figure 5.10d**), corroborating our previous observations. The interplay between PRC2 and *PCAT-1* further suggests that this lncRNA may have an important role in prostate cancer progression (**Figure 5.10e**)

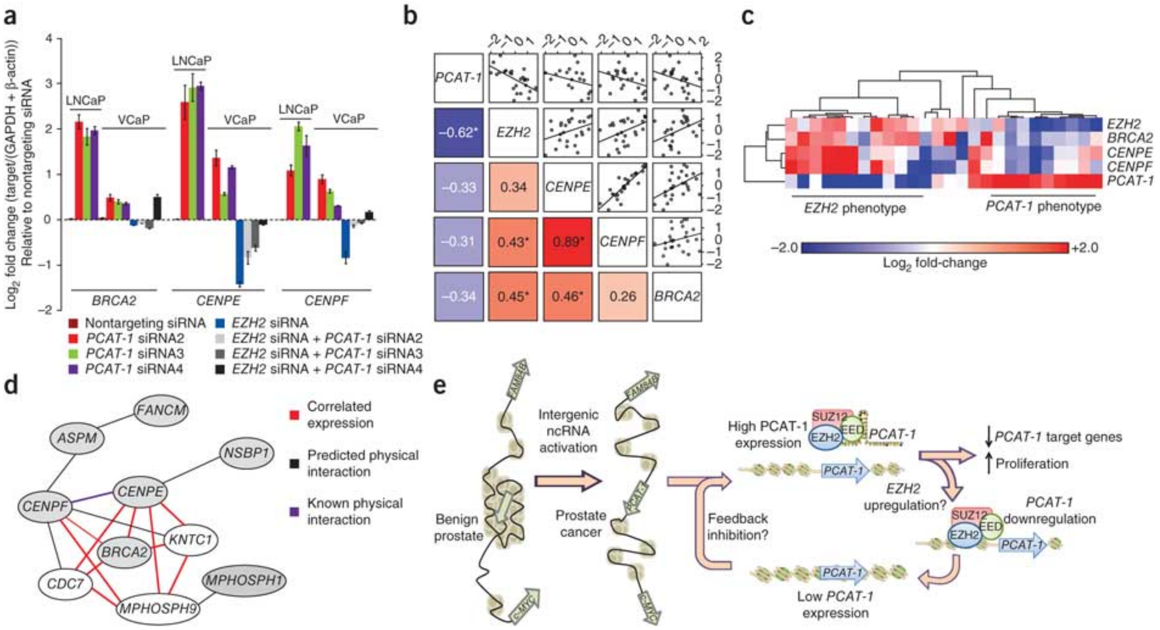


Figure 5.10: Prostate cancer tissues recapitulate PCAT-1 signaling.
(a) qPCR expression of three *PCAT-1* target genes after *PCAT-1* knockdown in VCaP and LNCaP cells, as well as following *EZH2* knockdown or dual *EZH2* and PCAT-1 knockdown in VCaP cells. qPCR data were normalized to the average of (*GAPDH* + β-actin) and represented as fold change. Error bars represent mean ± s.e.m. (b) Standardized log2-transformed qPCR expression of a set of tumors and metastases with outlier expression of either *PCAT-1* or *EZH2*. The shaded squares in the lower left show Spearman correlation values between the indicated genes (* indicates P < 0.05). Blue and red indicate negative or positive correlation, respectively. The upper squares show the scatter plot matrix and fitted trend lines for the same comparisons. (c) A heatmap of *PCAT-1* target genes (*BRCA2*, *CENPF*, *CENPE*) in *EZH2*-outlier and *PCAT-1*-outlier patient samples (see Fig. 4b). Expression was determined by qPCR and normalized as in b. (d) A predicted network generated by the HefaLMP program for 7 of 20 top upregulated genes following *PCAT-1* knockdown in LNCaP cells. Gray nodes are genes found following *PCAT-1* knockdown. Red edges indicate co-expressed genes; black edges indicate predicted protein-protein interactions; and purple edges indicate verified protein-protein interactions. (e) A proposed schematic representing *PCAT-1* upregulation, function and relationship to PRC2.

SChLAP1 contributes to the development of aggressive prostate cancer by antagonizing the SWI/SNF complex

The SChLAP locus harbors several novel cancer-specific transcripts, including *PCAT-109*, *PCAT-113*, *PCAT-114*, *PCAT-115*, and *PCAT-118*. These PCATs were prominently expressed in a subset of disease, and were nominated by COPA analysis to detect highly juxtaposed expression patterns. In fact, PCATs contained in the SChLAP locus scored higher in this analysis than any other novel transcripts (**Figure 5.11a**). We began investigating the SChLAP locus by employing RACE on the PCATs in the region, including *PCAT-109* and *PCAT-114*. Efforts to validate *PCAT-109* by PCR and rapid amplification of cDNA ends (RACE) failed, partly due to the fact that this gene was not robustly expressed in any prostate cell lines. By contrast, in the *PCAT-114* region, PCR experiments and 5' and 3' RACE defined a 1.4 kb, poly-adenylated gene composed of up to seven exons and spanning nearly 200kb on Ch2q31.3 (**Figure 5.11b**). We found that this gene linked together multiple PCATs that were initially assembled independently, and so we renamed the gene Second Chromosome Locus Associated with Prostate-1 (*SChLAP1*).

We employed a published ChIP-Seq dataset of prostate cancer[182] to confirm that the transcriptional start site (TSS) of *SChLAP1* was marked by tri-methylation of H3K4 (H3K4me3) and its gene body harbored tri-methylation of H3K36 (H3K36me3) (**Figure 5.11b**), an epigenetic signature consistent with canonical protein-coding genes and lncRNAs[79]. PCR assays defined numerous splicing isoforms of this gene of which three (termed isoforms #1, #2, and #3, respectively) constituted the vast majority (>90%) of transcripts in the cell (data not shown).

Using quantitative PCR (qPCR), we validated that *SChLAP1* was highly expressed in ~20% of prostate cancers, including metastatic prostate cancer (**Figure 5.11c**), but that *SChLAP1* expression was low in benign prostate tissues. To establish *SChLAP1* as a non-coding gene, we cloned three isoforms (isoforms 1, 2 and 3) and performed *in vitro* translation assays, which were negative (data not shown). Consistent with this, we found that *SChLAP1* expression in prostate cell lines was located in the nucleus (**Figure 5.11d**), whereas protein-coding mRNAs are located in the cytoplasm in order to engage the ribosomal machinery. To confirm the nuclear localization of *SChLAP1* in human samples, we optimized an *in situ* hybridization (ISH) assay to visualize *SChLAP1* expression in formalin-fixed paraffin-embedded (FFPE) prostate cancer samples using a training set and a test set of tissues. We similarly observed that expression of *SChLAP1* was almost exclusively found in the cell nucleus in both localized and metastatic prostate cancers (**Figure 5.11e**).

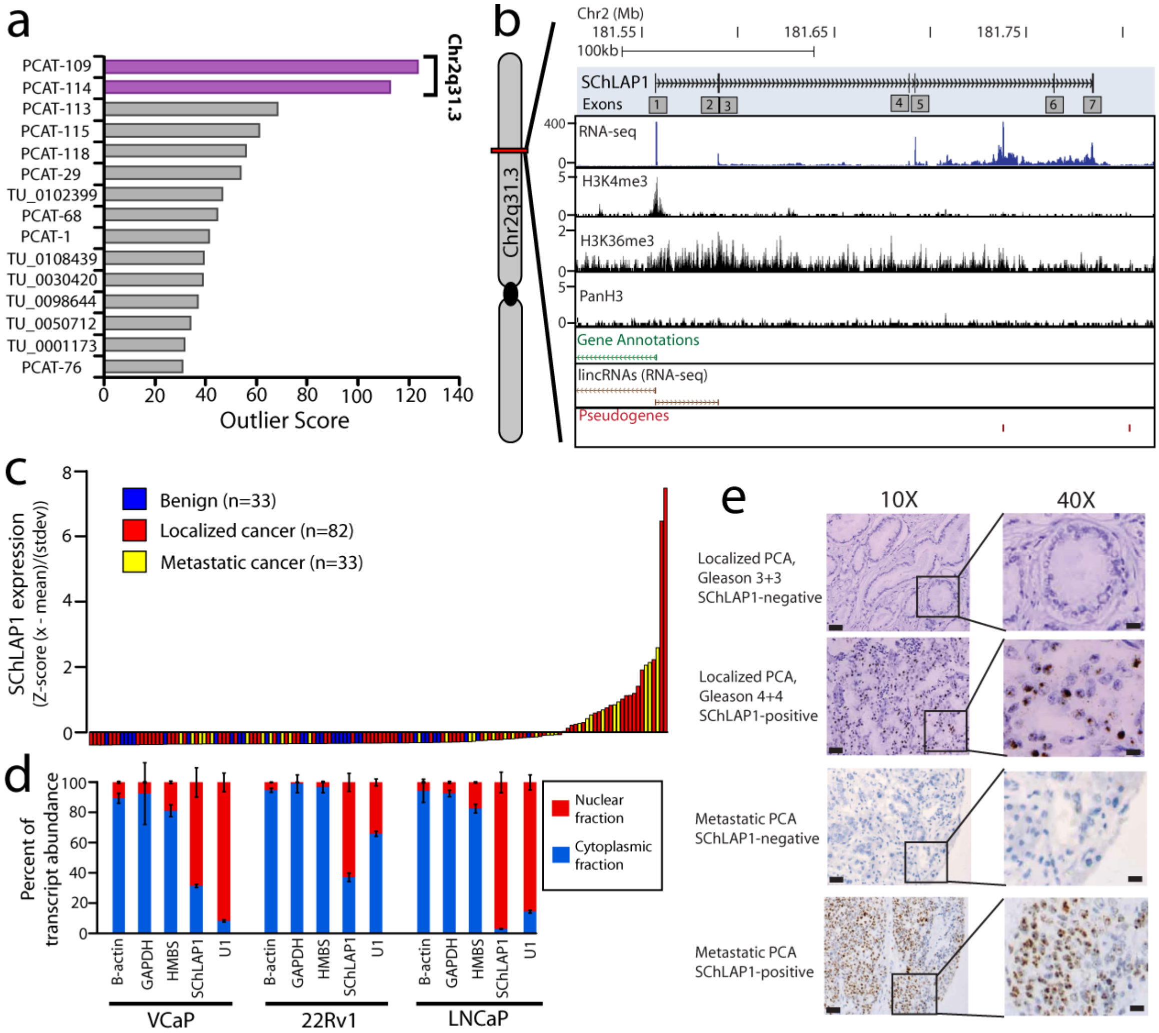


Figure 5.11: Discovery of *SChLAP1* as a prostate cancer lncRNA.
(a) Cancer outlier profile analysis (COPA) for intergenic lncRNAs in prostate cancer nominates two transcripts, *PCAT-109* and *PCAT-114*, as prominent outliers. (b) A representation of the *SChLAP1* gene and its annotations in current databases. *SChLAP1* may consist of up to seven exons on Chr2q31.3. An aggregated representation of current gene annotations for Ensembl, ENCODE, UCSC, Ref-Seq, and Vega shows no annotation for *SChLAP1*. No spliced ESTs represent *SChLAP1*. ChIP-Seq data for H3K4me3, RNA-Pol II, and H3K36me3 show enrichment at the *SChLAP1* gene. Also, RNA-Seq data showing an outlier sample for *SChLAP1* illustrates its expression. (c) qPCR for *SChLAP1* on a panel of benign prostate (n=33), localized prostate cancer (n=82), and metastatic prostate cancer (n=33) samples. qPCR data is normalized to the average of (*GAPDH* + *HMBS*) and represented as standardized expression values. (d) *SChLAP1* expression is predominantly nuclear. Prostate cell lysates were fractionated and nuclear and cytoplasmic fractions were tested for RNA expression. U1 is a positive control for nuclear gene expression. (e) *In situ* hybridization of *SChLAP1* in human prostate cancer. *SChLAP1* staining is shown for both localized and metastatic tissues, and is nuclear in cellular localization.

In silico analysis suggests *SChLAP1* associates with aggressive prostate cancer

Given that *SChLAP1* is not measured by gene expression microarray platforms, we explored a link between *SChLAP1* and aggressive prostate cancer by defining signatures of *SChLAP1*-correlated or anti-correlated genes. To develop these signatures, we augmented our original RNA-Seq cohort with data from 12 primary tumors and 5 benign tissues published in GEO as GSE2226018, and 16 primary tumors and 3 benign tissues released in dbGAP as study phs000310.v1.p119. We reanalyzed the sequencing data using Tophat version 1.3.1[56] against the Ensembl GRCh37 human genome build. Known introns (Ensembl release 63) were provided to Tophat. Gene expression across the Ensembl version 63 genes and the *SChLAP1* transcript was quantified by HT-Seq version 0.5.3p3 using the script htseq-count (www- huber.embl.de/users/anders/HTSeq). Reads were counted without respect to strand to avoid bias between unstranded and strand-specific library preparation methods. This bias results from the inability to resolve reads in regions where two genes on opposite strands overlap in the genome.

Using the count data from this augmented cohort, we developed signatures distinguishing cancers from benign samples, metastatic from primary tumors, and high-grade from low-grade tumors (**Figure 5.12**). Differential expression analysis was performed using R package DESeq version 1.6.1[33]. We called differentially expressed genes by imposing adjusted p-value cutoffs for cancer versus benign (padj < 0.05), metastasis versus primary (padj < 0.05), and Gleason 8+ versus 6 (padj < 0.10).

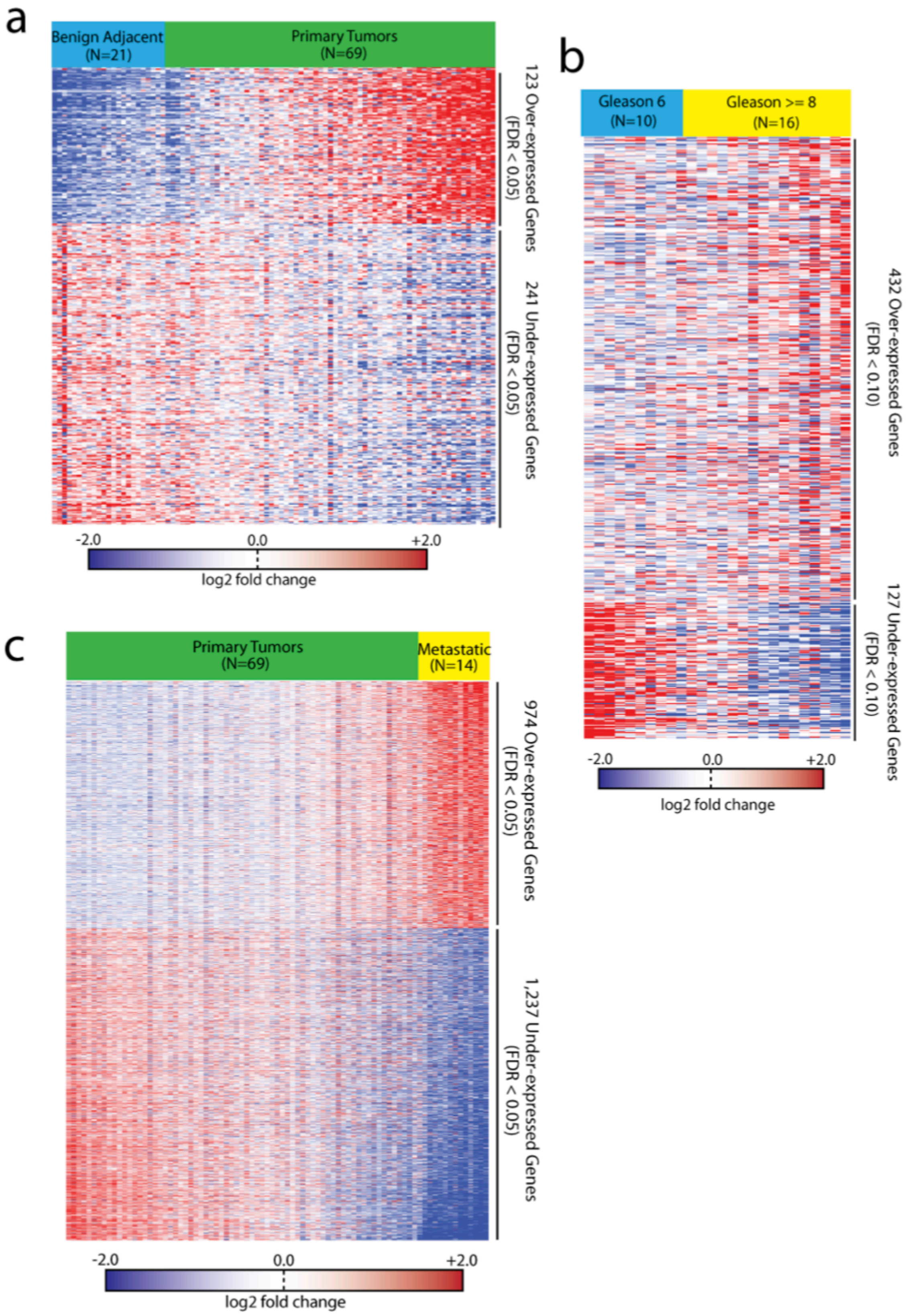


Figure 5.12: Generation of prostate cancer gene signatures from RNA- Seq data.

Heatmap visualizations of RNA-Seq gene signatures distinguishing (a) benign prostate tissues and localized prostate cancer tissues, (b) low grade from high grade localized prostate cancer tissues, and (c) localized prostate cancer tissues from metastatic cancers.

Next we used the count data to derive signatures of correlated and anti-correlated genes to *SChLAP1*. Read count data from HT-Seq were normalized and converted to pseudo-counts using functions from DESeq[33]. Gene expression levels were then mean-centered and standardized using the scale function in R. Pearson correlation coefficients were computed between each gene of interest and all other genes. Statistical significance of Pearson correlations was determined by comparison to correlation coefficients achieved by 1,000 random permutations of the expression data. We controlled for multiple hypothesis testing using the qvalue package in R. A 253-gene *SChLAP1* correlation signature was then determined by imposing a cutoff of $q < 0.05$ on the correlation results.

We interrogated the *SChLAP1* gene signature across published prostate cancer microarray data curated using Oncomine concept analysis[213, 214]. We separated the 253 genes with expression levels significantly correlated to *SChLAP1* into positively and negatively correlated gene lists. We imported these gene lists into Oncomine as custom concepts. We then nominated significantly associated Prostate Cancer concepts with Odds Ratio > 3.0 and p-value < 10-6. We exported these results as nodes and edges of a concept association network, and visualized the network using Cytoscape version 2.8.2[215]. The 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 text labels)

Network analysis representing the significantly enriched concepts (p-value < 1e-6, odds ratio > 3.0) revealed a striking association with concepts related to prostate cancer progression (**Figure 5.13a**). Genes positively correlated with *SChLAP1* were over-expressed in metastatic and high-grade localized tumors. Conversely, genes negatively correlated with *SChLAP1* were under-expressed in metastatic and high-grade localized tumors.

Although Oncomine allows users to import custom gene signatures, it limits the user to select a number of predefined microarray platforms. Therefore, we developed an independent concept association analysis in order to make robust statistical claims based on data from our RNA sequencing cohort. We expanded our correlation analysis to include additional known prostate cancer genes: *EZH2*, a metastasis gene[210, 211], *PCA3*, an over-expressed lncRNA biomarker[174], *AMACR*, a tissue biomarker[188], and B-actin (*ACTB*) as a control gene. For each gene we created signatures from the top 5 percent of positively and negatively correlated genes. We performed a large meta-analysis of these correlation signatures across Oncomine datasets corresponding to disease outcome (Glinsky Prostate, Setlur Prostate), metastatic disease (Holzbeierlein Prostate, Lapointe Prostate, LaTulippe Prostate, Taylor Prostate 3, Vanaja Prostate, Varambally Prostate, and Yu Prostate), advanced gleason score (Bittner Prostate, Glinsky Prostate, Lapointe Prostate, LaTulippe Prostate, Setlur Prostate, Taylor Prostate 3, and Yu Prostate), and localized cancer (Arredouani Prostate, Holzbeierlein Prostate, Lapointe Prostate, LaTulippe Prostate, Taylor Prostate 3, Varambally Prostate, and Yu Prostate). We also incorporated our own concept signatures for metastasis, advanced Gleason score, and localized cancer determined from our RNA-Seq data (**Figure 5.12**). For each concept we downloaded the gene signatures corresponding to the top 5 percent of genes up- and down-regulated from the Oncomine web site. Pairwise signature comparisons were performed using a one-sided Fisher's Exact Test. We controlled for multiple hypothesis testing using the "qvalue" package in R. We considered concept pairs with q < 0.01 and odds ratio > 2.0 as significant.

A heat-map visualization of statistically significant comparisons (q-value < 0.01) confirmed a strong association of SChLAP1-correlated genes with high-grade and metastatic cancers as well as poor clinical outcomes (**Figure 5.13b**). In this respect, *SChLAP1* was highly similar to *EZH2*, a control metastasis gene. By contrast, *PCA3* and *AMACR*, two biomarkers not associated with disease progression, strongly associated with Cancer vs. Normal concepts but not concepts associated with aggressive disease.

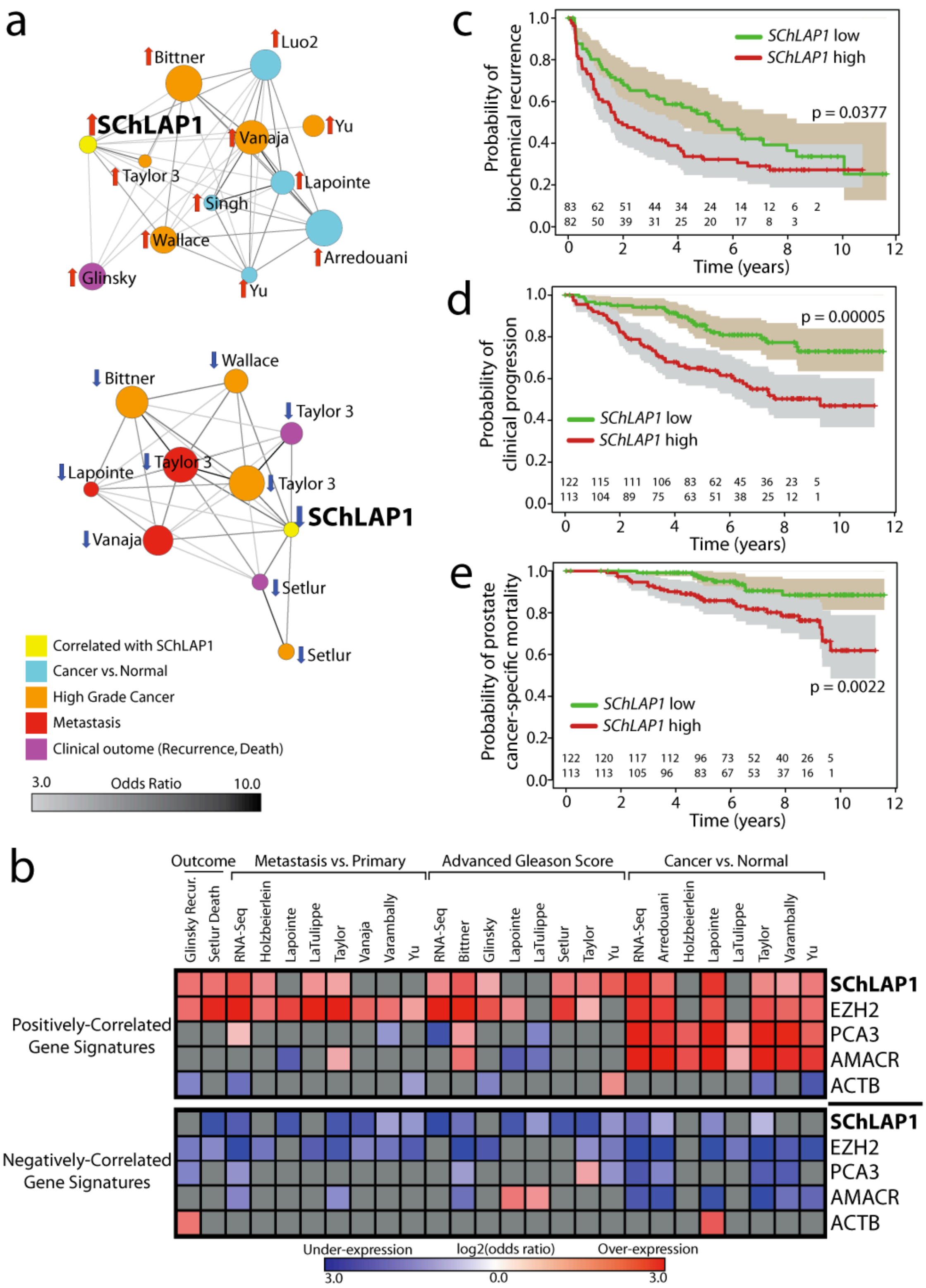


Figure 5.13: *SChLAP1* expression characterizes aggressive prostate cancer.
(a) Network representation of Oncomine concepts analysis of genes positively and negatively correlated with *SChLAP1* expression levels in localized prostate cancers profiled by RNA-Seq. The network was drawn using the Force Directed Layout algorithm in the Cytoscape[215] tool and subtly altered such that text labels could be visualized aesthetically. Node sizes reflect the number of genes that comprise each

molecular concept. The nodes are colored according to the concept category: SChLAP1 (yellow), Cancer vs. Normal (cyan), High Grade Cancer (orange), Metastasis (red), and Clinical Outcome (magenta). Edges are drawn between nodes with statistically significant enrichment (p-value < 1e-6, odds ratio > 3.0) and darker edge shading implies higher odds ratio. (b) Heatmap representation of comparisons between co-expression gene signatures and molecular concepts. Comparisons to positively (top portion) and negatively correlated (bottom portion) gene signatures are shown separately. Comparisons that do not reach statistical significance (q > 0.01 or odds ratio < 2) are shown in grey. In cases where a gene signature associates with both the over- and under-expression gene sets from a single concept, only the most significant result (as determined by odds ratio) is shown. Associations with over-expression concepts are colored red, and under-expression concepts blue. The color shade reflects the base-2 logarithm of the odds ratio. (c-e) Kaplan-Meier analyses of prostate cancer outcomes in the Mayo Clinic cohort. SChLAP1 expression was measured using Affymetrix exon arrays and patients were stratified according to their SChLAP1 expression. Patient outcomes were analyzed for biochemical recurrence (c), clinical progression to systemic disease (d), and prostate cancer-specific mortality (e). The shaded regions represent the 95% confidence interval.

We complemented this analysis, which was based on significantly overlapping gene signatures, with Kaplan-Meier Survival Analysis based on *SChLAP1* gene signature. We downloaded prostate cancer expression profiling data and clinical annotations from GSE8402 published by Setlur *et al*.[216] and found that 80 of the 253 genes in the *SChLAP1* signature had been assayed in the study. We then assigned *SChLAP1* expression scores to each patient sample in the cohort using the un-weighted sum of standardized expression levels across the 80 genes. Given that we observed *SChLAP1* expression in approximately 20% of prostate cancer samples, we used the 80th percentile of *SChLAP1* expression scores as the threshold for “high” versus “low” scores. We then performed 10-year survival analysis using the survival package in R and computed statistical significance using the log-rank test. Additionally, we imported the 253-gene *SChLAP1* signature into Oncomine in order to download the expression data for 167 of the 253 genes profiled by the Glinsky prostate dataset[217]. We assigned *SChLAP1* expression scores in a similar fashion and designated the top 20% of patients as “high” for *SChLAP1*. We performed survival analysis using the time to biochemical PSA recurrence and computed statistical significance as above. Kaplan-Meier analysis of each dataset similarly showed significant associations (log rank test, p < 0.01) between the *SChLAP1* signature and more rapid disease recurrence and decreased survival probability (**Figure 5.14**).

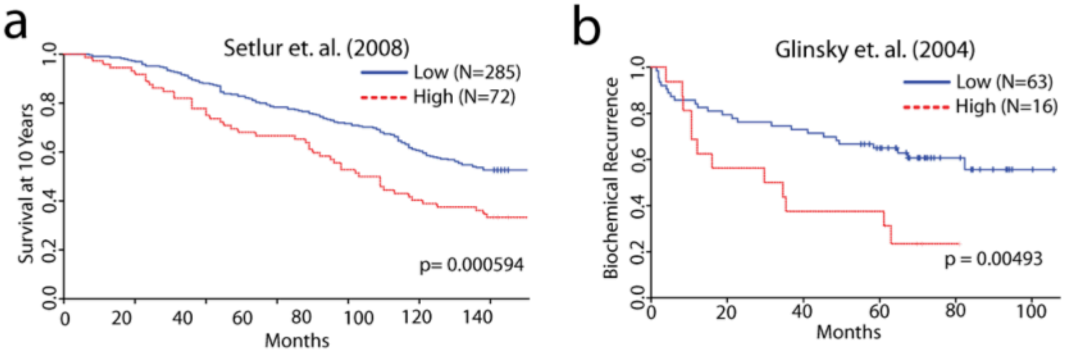


Figure 5.14: SChLAP1 expression stratifies prostate cancer patient outcomes.
(a) Kaplan-Meier analysis of prostate cancer outcomes. Patients were stratified according to their SChLAP1 signature score. Signature scores at or above the 80th percentile were deemed 'High', and the rest 'Low'. Statistical significance was determined by the log rank test. Analysis of the 10-year overall survival probability for prostate cancer patients from the Setlur *et al*. study. (b) As in (a), Analysis of the biochemical recurrence probability for prostate cancer patients from the Glinksy *et al*. study.

SChLAP1 levels associate with aggressive disease in a cohort of high risk prostate cancer patients

To implicate *SChLAP1* expression with clinical outcomes directly, we used Affymetrix exon microarrays, which harbor probes mapping to *SChLAP1* exons (see Methods), to profile its expression in a prospectively-designed study of 235 high-risk prostate cancer patients who underwent radical prostatectomy between 2000-2006 at the Mayo Clinic[218]. Samples were defined as *SChLAP1*-low or *SChLAP1*-high according to unsupervised clustering by the PAM function in R and evaluated for three clinical endpoints: biochemical recurrence (BCR), clinical progression to systemic disease (CP), and prostate cancer-specific mortality (PCSM). At the time of this analysis, patients had a median follow-up of 8.1 years.

We found that *SChLAP1* was a powerful single-gene predictor of aggressive prostate cancer (Fig. 2c-e). *SChLAP1* expression was highly significant when distinguishing CP and PCSM (p = 0.00005 and p = 0.002, respectively) (**Figure 5.13d,e**). For the BCR endpoint, high *SChLAP1* expression in patient primary tumor specimens was associated with a rapid median time-to-progression (1.9 vs 5.5 years for *SChLAP1* high and low patients, respectively) (**Figure 5.13c**). To validate these findings, we confirmed that *SChLAP1*-positive patients are at markedly higher risk for BCR using qPCR on an independent cohort (**Figure 5.15**). Multivariable and univariable regression analyses of the Mayo Clinic data demonstrated that *SChLAP1* expression is an independent predictor of prostate cancer aggressiveness with highly significant hazard ratios for predicting BCR, CP, and PCSM (HR or 3.045, 3.563, and 4.339, respectively, p < 0.01) which were comparable to other clinical factors such as advanced clinical stage and the Gleason histopathological score (**Figure 5.16**). Taken together, our data suggest that *SChLAP1* expression is a powerful indicator of aggressive cancer that either out-performs, or is comparable to, standard clinical parameters for the prediction of CP, PCSM, and BCR.

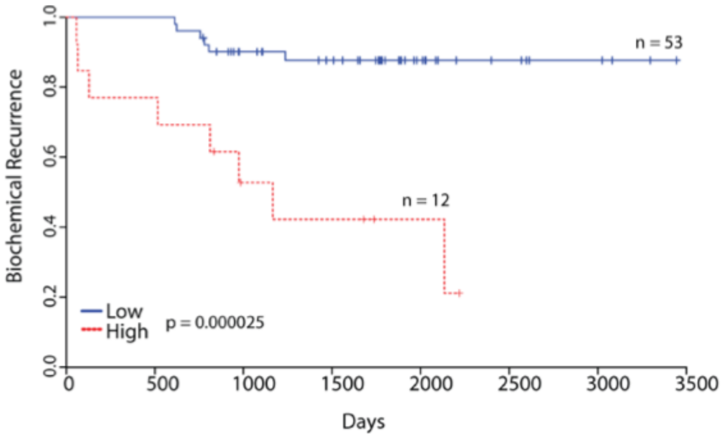


Figure 5.15: *SChLAP1* predicts biochemical recurrence in the University of Michigan cohort. *SChLAP1* expression was measured using qPCR on a cohort of fresh-frozen prostate cancer tissue samples from radical prostatectomy patients for whom follow-up for biochemical recurrence was available. Statistical significance was determined by the log- rank test.

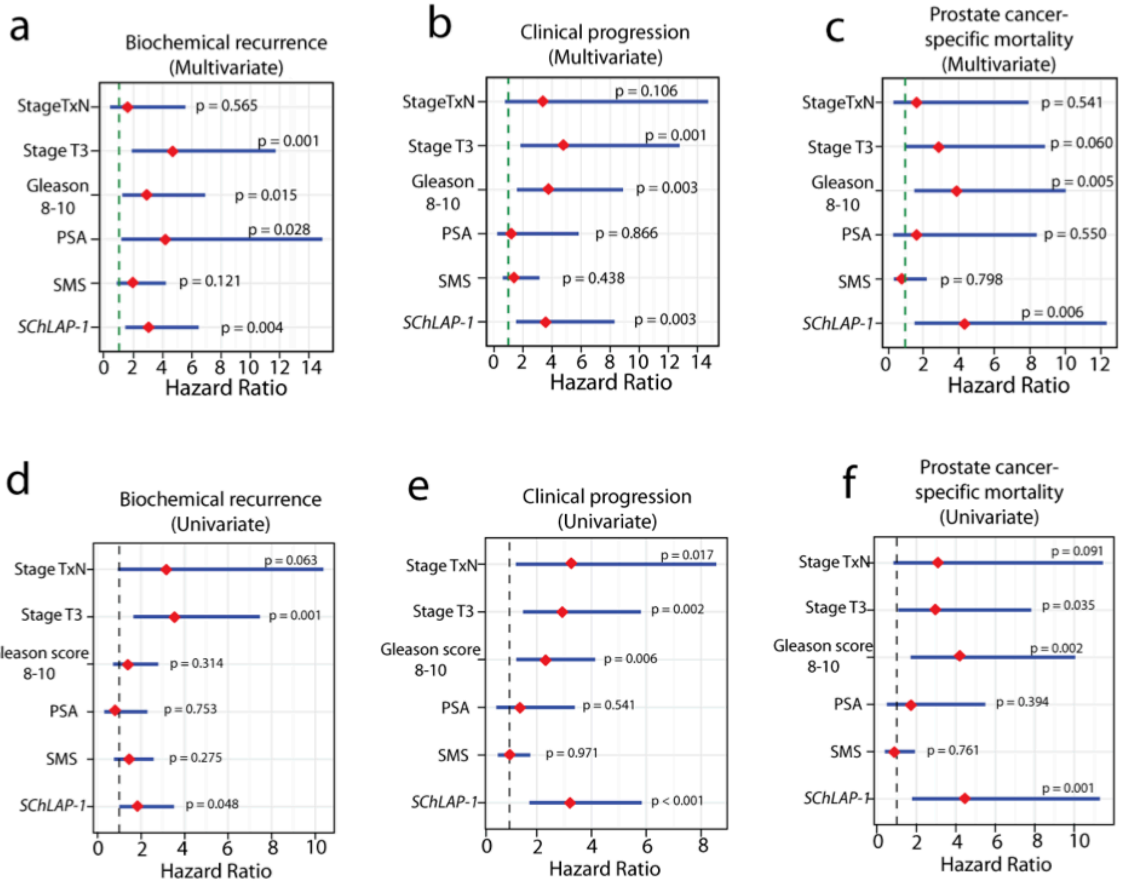


Figure 5.16: *SChLAP1* expression is an independent predictor of patient clinical parameters. (a-f) Multivariate and univariate analyses for *SChLAP1* and disease outcomes. (a-c) Multivariate survival analyses demonstrate that *SChLAP1* is an independent predictor of prostate cancer biochemical recurrence (a), clinical progression (b), and prostate cancer-specific mortality (c) following radical prostatectomy. (d-f) Univariate survival analyses for *SChLAP1* for biochemical recurrence (d), clinical progression (e), and prostate cancer-specific mortality (f) as in (a- c). For these analyses, clinical significance was adjusted for confounding adjuvant treatment, and Gleason score was dichotomized between those samples ≤7 ≥8. Red diamonds indicate the median hazard ratio for each factor and blue lines indicate the 95% confidence interval

SChLAP1 controls cell invasiveness *in vitro* and *in vivo*

To explore a functional role for *SChLAP1*, we performed siRNA knockdowns of this gene using two independent siRNAs as well as siRNA to *EZH2*, a positive control essential for cancer cell invasion[210, 211]. Remarkably, knockdown of *SChLAP1* dramatically impaired cell invasion *in vitro* at a level comparable to *EZH2* (**Figure 5.17a** and **Figure 5.18a**). *SChLAP1* knockdown also impaired cell proliferation (**Figure 5.18b**). Overexpression of *SChLAP1* isoform 2, which lacks the binding site for siRNA-

2, rescued the *in vitro* invasive phenotype of 22Rv1 cells treated with siRNA-2 (**Figure 5.18c,d**), confirming the specificity of our siRNA experiments. Next, we overexpressed the three most abundant *SChLAP1* isoforms in RWPE benign immortalized prostate cells at physiologic levels similar to the LNCaP cell line (**Figure 5.18e**). While *SChLAP1* overexpression did not impact cell proliferation (**Figure 5.18f**), we found that RWPE cells expressing all three *SChLAP1* isoforms, but not control cells, exhibited a dramatic ability to invade through Matrigel model basement membrane matrix *in vitro* (**Figure 5.17b**).

To assess the role of *SChLAP1* on cancer cells *in vivo*, we employed a xenograft model using 22Rv1 cells stably knocking down *SChLAP1* (**Figure 5.19a**) and confirmed that this gene is necessary for appropriate cancer cell metastatic seeding *in vivo*. Specifically, we performed intracardiac injection of tumor cells and monitored luciferase signal from mouse lungs and distant metastases. These experiments showed that 22Rv1 *shSChLAP1* cells displayed impaired metastatic seeding and growth at both proximal (lungs) and distal sites (**Figure 5.17c,d**). Indeed, 22Rv1 shSChLAP1 cells displayed both fewer gross metastatic sites overall (an average 3.66 metastatic sites in shNT mice vs. 2.07 metastatic sites in shSChLAP1 #1 and 1.07 sites in shSChLAP1 #2 mice, $p < 0.05$, Student's t-test) as well as smaller metastatic tumors when they did form (**Figure 5.17d,e**). Histopathological analysis of the metastatic 22Rv1 tumors, regardless of *SChLAP1* knockdown, showed uniformly high-grade epithelial cancer with frequent mitosis noted (**Figure 5.19b**). Interestingly, shSChLAP1 subcutaneous xenografts displayed slower tumor progression *in vivo* (**Figure 5.19c**); however this was due to delayed tumor engraftment rather than decreased tumor growth kinetics and we observed no change in Ki67 staining between shSChLAP1 and shNT cells (**Figure 5.19d-i**).

Next, we used the chick chorioallantoic membrane (CAM) assay[210] to examine the metastatic process more closely. Specifically, this assay measures cancer cell metastasis into a chicken embryo and enables analysis of multiple neoplastic capabilities required for metastasis, including the ability to invade local tissues, intravasate into and extravasate out of blood vessels, seed distant organs, and grow in a foreign microenvironment31. We found that 22Rv1 shSChLAP1 cells demonstrated a greatly reduced ability to invade (**Figure 5.17f**), intravasate (**Figure 5.17g**) and metastasize distant organs (**Figure 5.17h**). Additionally, 22Rv1 shSChLAP1 cells also showed decreased tumor growth in the chick embryo (**Figure 5.17i**). Importantly, RWPE-SChLAP1 overexpression cells partially recapitulated these results, displaying a markedly increased ability to intravasate (**Figure 5.17j**). Together, the murine metastasis and CAM data suggest that the primary function of *SChLAP1* may be to promote invasion and metastasis through cancer cell intravasation, extravasation, and subsequent tumor cell seeding.

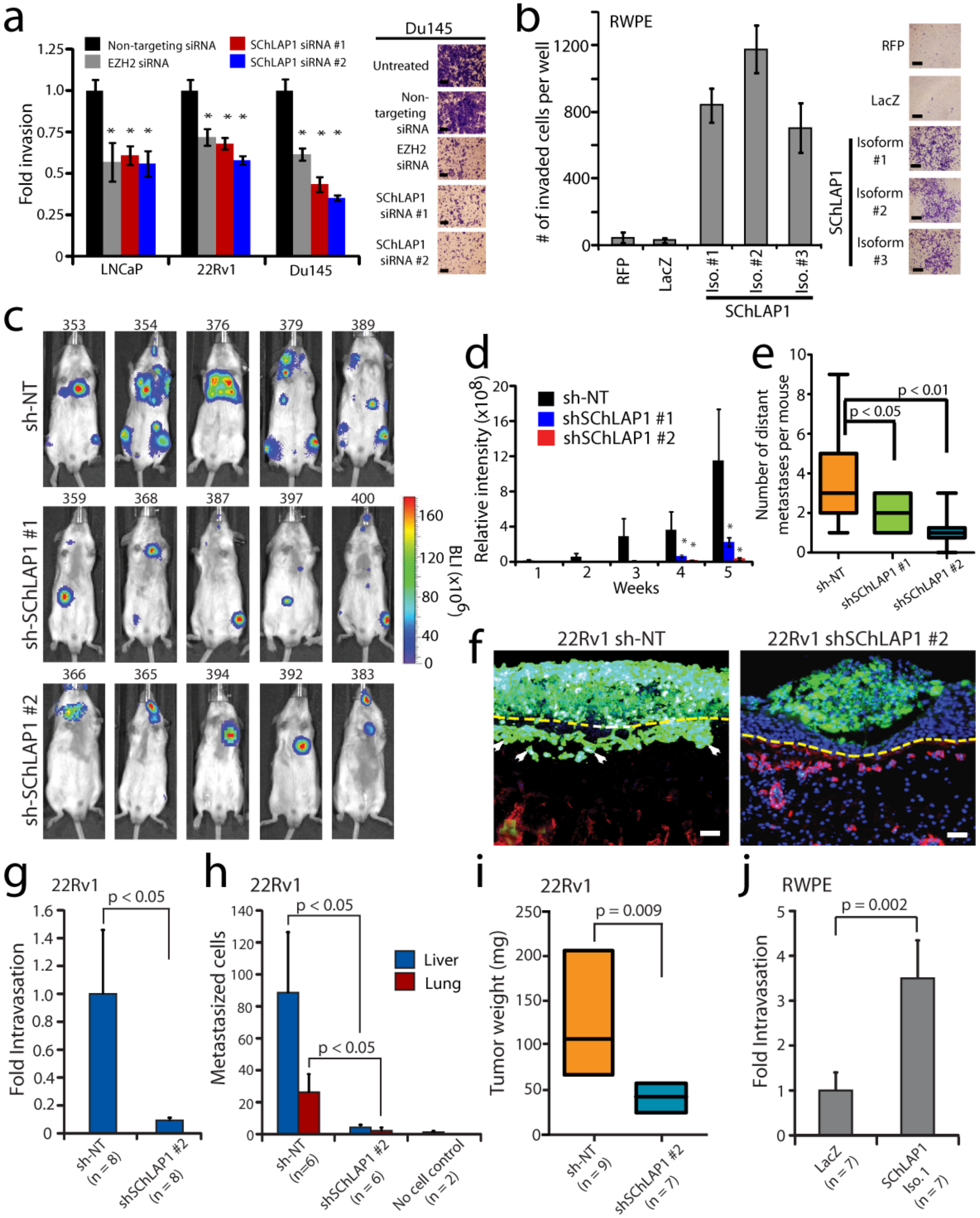


Figure 5.17: SChLAP1 coordinates cancer cell invasion *in vitro* and metastatic seeding *in vivo*. (a) siRNA knockdown of *SChLAP1 in vitro*. Three prostate cell lines (LNCaP, 22Rv1, Du145) were treated with two independent siRNAs for *SChLAP1* and invasion through Matrigel in a Boyden chamber assay was monitored. *EZH2* siRNA serves as a positive control. Data are represented as normalized mean +/- S.E.M. An asterisk (*) indicated $p < 0.05$ by Student's T-test. (b) Overexpression of *SChLAP1* in RWPE cells. Benign RWPE prostate cells overexpressing three isoforms of *SChLAP1*, but not controls, demonstrate increased cellular invasion. Data are represented as normalized mean +/- S.E.M. (c) Intracardiac injection of 22Rv1 cells with stable *SChLAP1* knockdown was performed in severe combined

immunodeficient (SCID) mice, and metastatic seeding and growth of tumor cells was monitored. Example luciferase bioluminescence images from 22Rv1 shNT, shSChLAP1 #1, and shSChLAP1 #2 mice five weeks following intracardiac injection. Mouse IDs are given above each image. (d) The relative intensity of whole-mouse luciferase signal is plotted for 22Rv1 intracardiac injection experiments. Data are represented as mean +/- S.E.M. An asterisk (*) indicates $p < 0.05$ by a two-tailed Student's T-test. (e) The number of gross metastatic sites observed by luciferase signal in 22Rv1 shSChLAP1 cells or shNT controls. Independent foci of luciferase signal were averaged for shNT (n=9), shSChLAP1 #1 (n=14) and shSChLAP1 #2 (n=14) mice. Statistical significance was determined by a two-tailed Student's t-test. (f) Invasion of 22Rv1-shNT and 22Rv1 shSChLAP1 #2 cells across the chorioallantoic membrane in the chick chorioallantoic membrane (CAM) assay. 22Rv1 cells are labeled with GFP. The image is counterstained with chicken collagen IV for vasculature (RFP) and DAPI for nuclei. (g) Quantification of intravasation of 22Rv1-shNT and 22Rv1 shSChLAP1 #2 cells in the CAM assay. (h) Quantification of metastasis to liver and lungs for 22Rv1-shNT and 22Rv1 shSChLAP1 #2 cells in the CAM assay. (i) Quantification of tumor weight of 22Rv1-shNT and 22Rv1 shSChLAP1 #2 cells in the CAM assay. (j) Quantification of intravasation of RWPE-LacZ and RWPE-*SChLAP1* cells in the CAM assay. All data are represented as mean +/- S.E.M. Statistical significance was determined by a two-tailed Student's t-test.

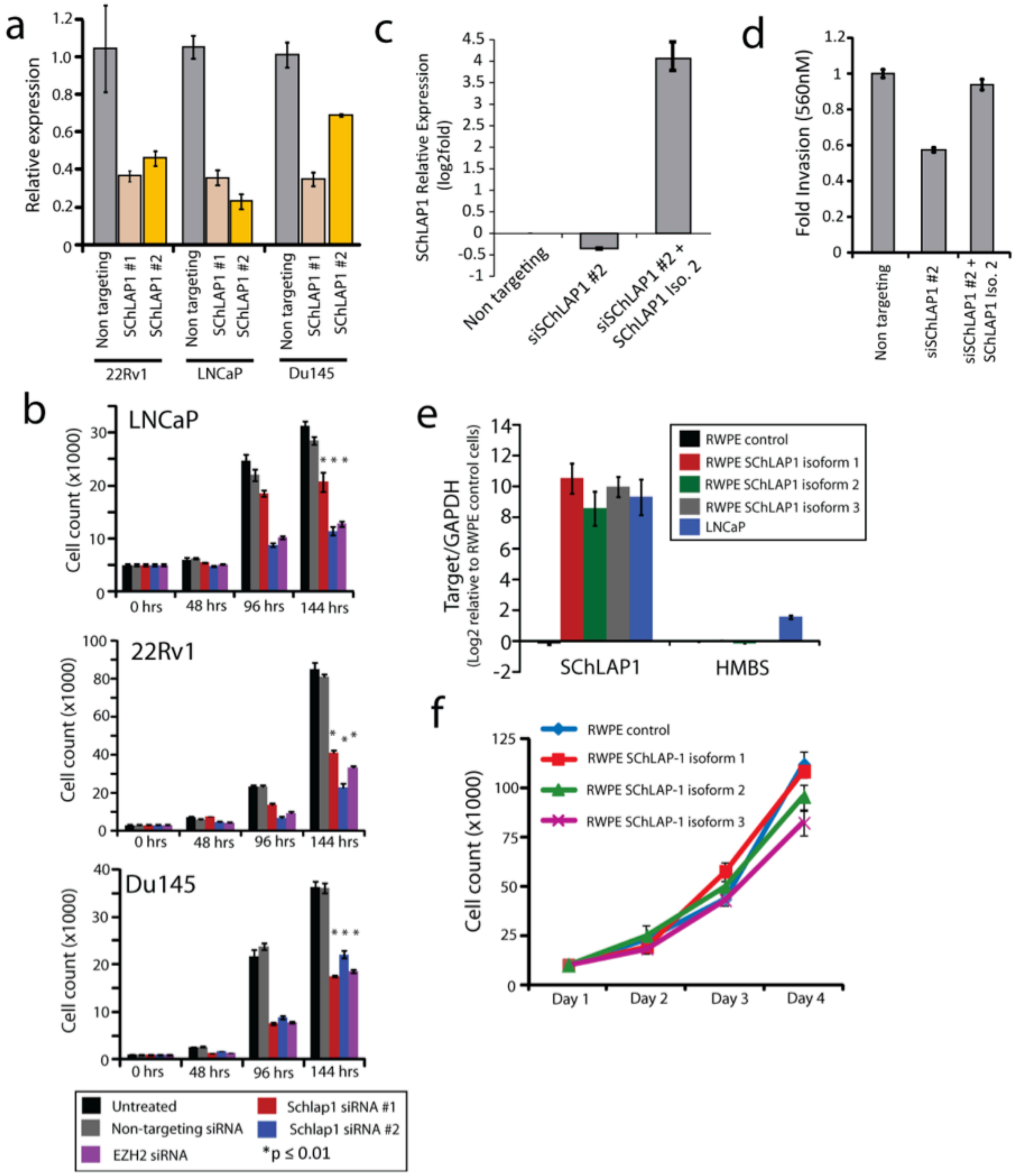


Figure 5.18: *In vitro* knockdown and overexpression of *SChLAP1*.
(a) 22Rv1, LNCaP, and Du145 cells were treated with siRNAs against *SChLAP1*. qPCR indicates relative knockdown efficiency in these cell lines. Error bars represent S.E.M. (b) Expression of *SChLAP1* in 22Rv1 cells treated with non- targeting, siRNA #2 for *SChLAP1*, or siRNA #2 with exogenous overexpression of *SChLAP1* isoform 2. (c) Boyden chamber invasion assay data for 22Rv1 cells treated with non-targeting, siRNA #2 for SChLAP1, or siRNA #2 with exogenous overexpression of *SChLAP1* isoform 2. Data are represented as absorbance at 560nM. Error bars represent S.E.M. (e) Overexpression of *SChLAP1* isoforms 1-3 in RWPE cells was confirmed using qPCR, which demonstrated that the overexpression resulted in comparable levels of SChLAP1 transcript to LNCaP cells that express this gene endogenously. HMBS serves as a negative control. Error bars represent S.E.M. (f) Cell proliferation assays for RWPE cells overexpressing SChLAP1 isoforms. No significant change in cell proliferation is observed. Error bars represent S.E.M.

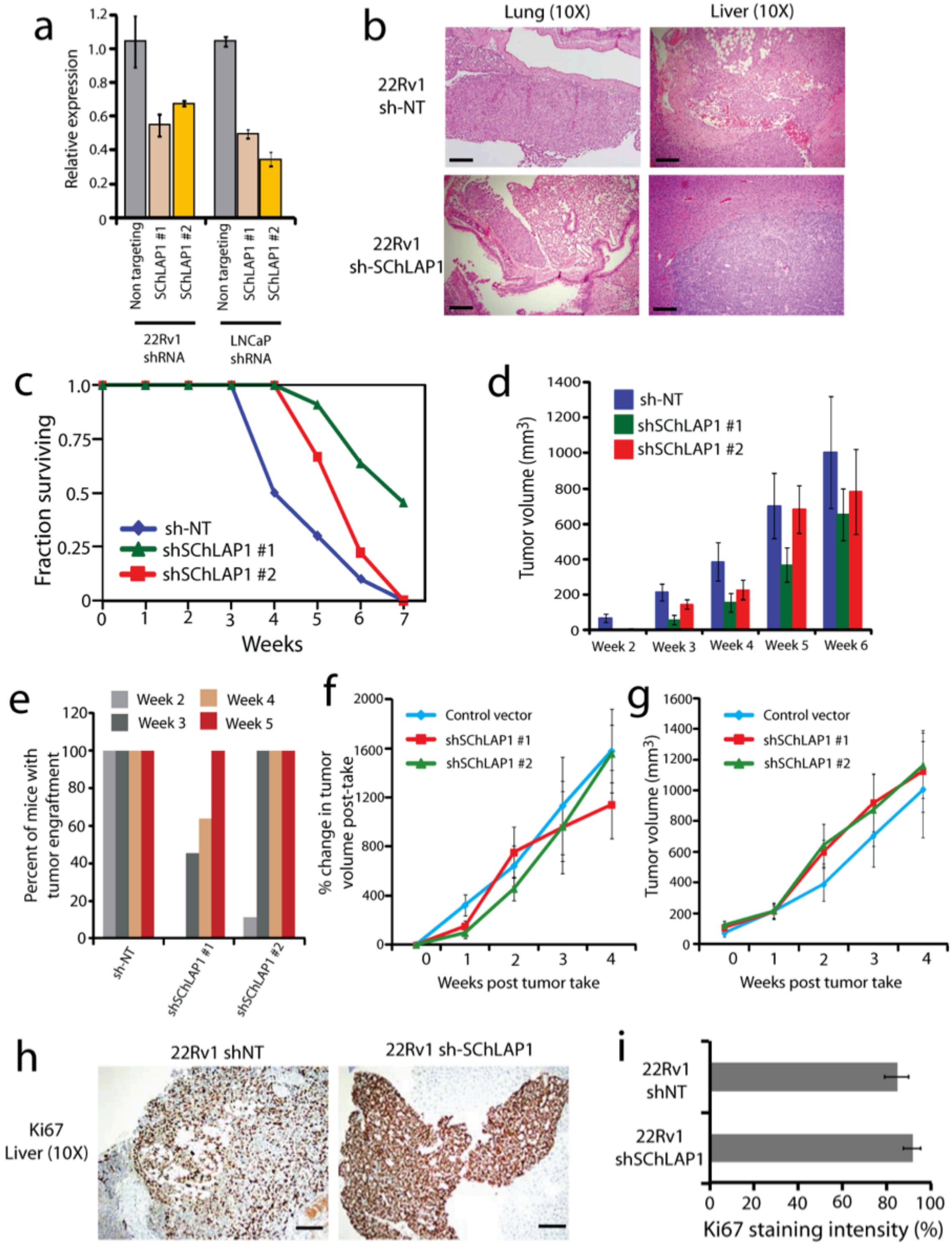


Figure 5.19: Knockdown of *SChLAP1* delays tumor engraftment but not tumor growth kinetics. (a) Knockdown efficiencies for the shRNA knockdown of *SChLAP1* in LNCAP and 22Rv1 cells. Error bars indicate S.E.M. (b) Histolopathology of murine tumors formed by intracardiac injection of 22Rv1 shNT or 22Rv1 sh-*SChLAP1* cells. Images are taken from the lungs and livers or mice with tumors. Slides are stained with H&E. (c) The fraction of mice surviving following subcutaneous injection of the 22Rv1 cell lines. This plot represents tumor-specific death of mice sacrificed when the tumor volume reached the

maximum allowable volume. (d) 22Rv1 cells infected with lentivirus for shNT, sh-*SChLAP1* #1, and sh-*SChLAP1* #2 were injected subcutaneously in mouse flanks and tumor growth was monitored by caliper measurements. N = 10 mice for shNT cells, n = 12 mice for sh-*SChLAP1* #1 cells, n = 9 mice for sh-*SChLAP1* #2 cells. Absolute tumor volume for 22Rv1 shNT, sh-*SChLAP1* #1 and sh-*SChLAP1* #2 cells. Errors bars represent S.E.M. (e) Percent of mice with tumor engraftment over time. Knockdown of *SChLAP1* delays the onset of tumor engraftment. (f) The percent change in tumor volume per cell line normalized to the time of tumor engraftment. Errors bars represent S.E.M. (g) Tumor volume normalized to the time of tumor engraftment. Errors bars represent S.E.M. (h) Immunohistochemistry staining for Ki67 in 22Rv1 shNT and sh-*SChLAP1* liver metastases. (i) Summary of Ki67 tumor staining for 22Rv1 shNT and sh-*SChLAP1* murine tumors show significant difference in Ki67 staining intensity.

SChLAP1 opposes gene expression regulation by the SWI/SNF complex

To interrogate potential mechanisms of *SChLAP1* function, we performed microarray profiling of 22Rv1 and LNCaP prostate cancer cells treated with *SChLAP1* or control siRNAs, which revealed 165 upregulated and 264 downregulated genes in a highly significant manner (q-value < 0.001) (**Figure 5.20a**). After ranking genes according to differential expression by Significance Analysis of Microarrays (SAM)[186], we employed Gene Set Enrichment Analysis (GSEA)[219] to search for enrichment across the Molecular Signatures Database (MSigDB)[220]. Among the highest ranked concepts we noticed genes positively or negatively correlated with the SWI/SNF complex (**Figure 5.20a**)[221], which was independently confirmed using gene signatures generated from our RNA-Seq data (**Figure 5.20c-e**). Importantly, *SChLAP1*-regulated genes were inversely correlated with these datasets, suggesting that *SChLAP1* and SWI/SNF regulate gene transcription in opposing manners, leading to an antagonism of SWI/SNF activity by *SChLAP1*.

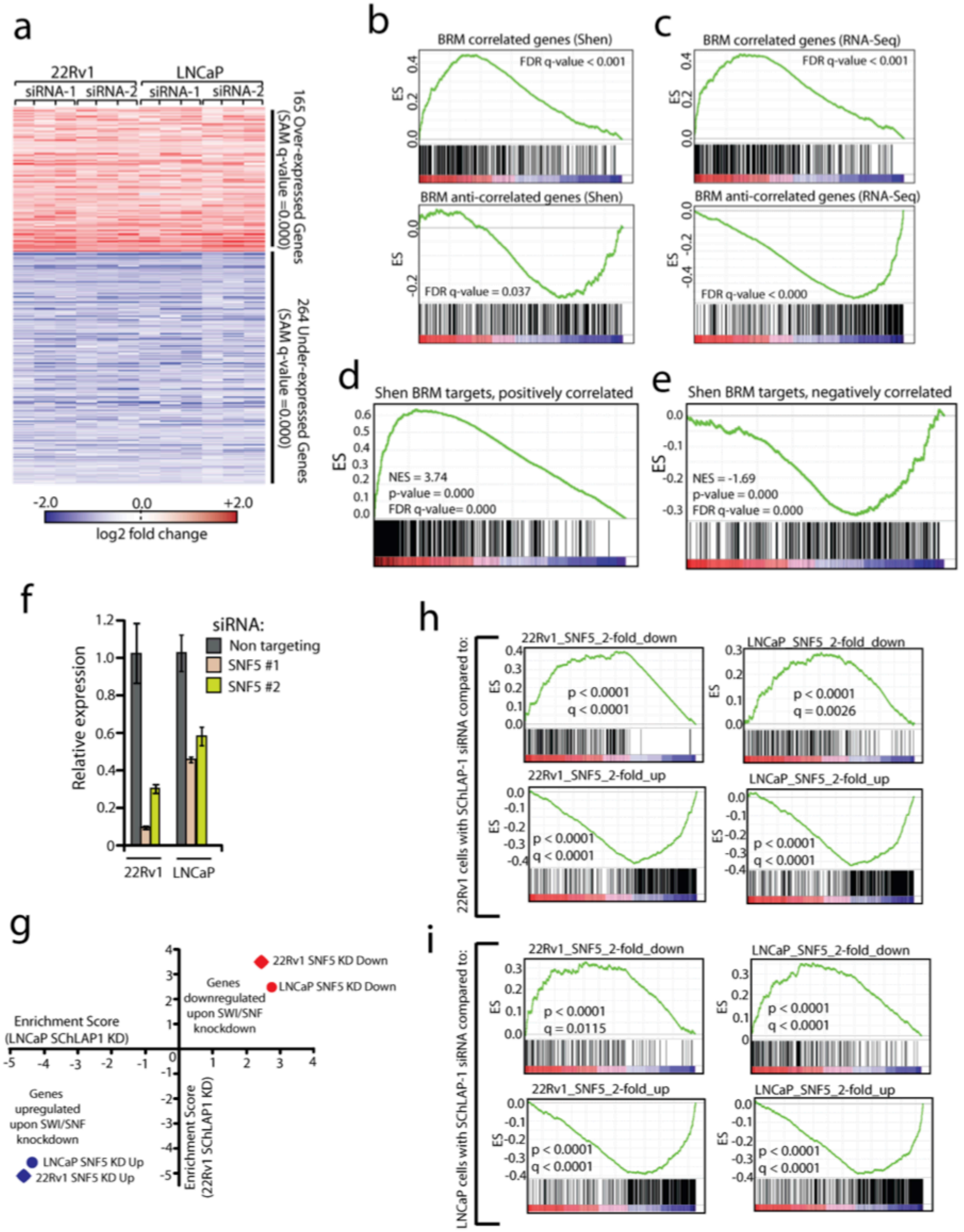


Figure 5.20: *SChLAP1* and the SWI/SNF complex regulate gene expression in an opposing manner. (a) Transcriptome profiling following *SChLAP1* knockdown *in vitro*. Differentially expressed genes were determined by SAM analysis and represented as a heatmap. (b-c) Gene set enrichment analysis (GSEA) of LNCaP and 22Rv1 cells treated with *SChLAP1* siRNAs. GSEA results indicate that *SChLAP1* knockdown results are inversely correlated with SWI/SNF-associated genes using data from Shen *et al*. (b) or using RNA-seq data (c). (d) Comparison of positively correlated *BRM*-associated gene signatures in prostate cancer. The *BRM*-derived signature from RNA-seq samples was compared to the Shen *et al*. signature by GSEA. A highly significant overlap between the signatures is observed. (e) Comparison of negatively

correlated BRM-associated gene signatures in prostate cancer. The *BRM*-derived signature from RNA-seq samples was compared to the Shen *et al*. signature by GSEA. A highly significant overlap between the signatures is observed. (f) Knockdown efficiency of *SNF5* siRNAs in 22Rv1 and LNCaP. Error bars represent S.E.M. (g) GSEA analysis of *SChLAP1* and *SNF5* knockdowns. Across two cell lines (LNCaP and 22Rv1), *SChLAP1* knockdown had the opposite effect on gene expression as knockdown of *SNF5*. Here, a positive GSEA normalized enrichment score (NES) indicates genes up-regulated upon *SChLAP1* knockdown, and a negative GSEA NES indicates genes down-regulated upon *SChLAP1* knockdown. (h) GSEA results from comparisons of *SChLAP1* and *SNF5* knockdown in 22Rv1 cells. *SChLAP1* was knocked-down using siRNAs in 22Rv1 cells. Gene expression changes were compared using GSEA to expression changes observed using *SNF5* siRNAs in LNCaP or 22Rv1 cells. The enrichment plots of these comparisons are shown. (i) GSEA results from comparisons of *SChLAP1* and *SNF5* knockdown in LNCaP cells. *SChLAP1* was knocked-down using siRNAs in LNCaP cells. Gene expression changes were compared using GSEA to expression changes observed using *SNF5* siRNAs in LNCaP or 22Rv1 cells. The enrichment plots of these comparisons are shown.

The SWI/SNF complex operates as a large, multi-protein system that utilizes ATPase enzymatic activity to physically move nucleosomes and, in doing so, regulates gene transcription[222]. Several SWI/SNF complex members are the target of recurrent, inactivating mutations in cancer, including *ARID1A*[223, 224], *PBRM1*[225], and *SNF5*[226], and numerous studies suggest that loss of SWI/SNF functionality promotes cancer progression[222, 227]. SWI/SNF mutations do occur in prostate cancer albeit not commonly[198]. Several reports suggest that down-regulation of SWI/SNF complex members characterizes subsets of prostate cancer[221, 228]. Thus, antagonism of SWI/SNF activity by *SChLAP1* would be consistent with the oncogenic behavior of *SChLAP1* and the tumor suppressive behavior of the SWI-SNF complex.

To directly test whether *SChLAP1* antagonizes SWI/SNF-mediated gene expression regulation, we performed siRNA knockdown of *SNF5* (also known as *SMARCB1*) (**Figure 5.20f**), an essential subunit of the SWI/SNF complex necessary for its function by facilitating histone protein binding[221, 227, 229]. Using two independent cell lines (22Rvl and LNCaP), a comparison of genes regulated by knockdown of *SNF5* to genes regulated by *SChLAP1* demonstrated an antagonistic relationship where *SChLAP1* knockdown affected the same genes as *SNF5* but in the opposing direction.

Quantification of this overlap between *SNF5*- and *SChLAP1*-regulated genes was performed using GSEA, which demonstrated that *SNF5* and *SChLAP1* affect gene expression in a highly significant and opposing manner in LNCaP and 22Rv1 (FDR < 0.05) (**Figure 5.20g-i**, and **Figure 5.22a**). Here, we also found that a shared *SNF5*-*SChLAP1* signature of co-regulated genes was highly enriched for prostate cancer clinical signatures for disease aggressiveness, supporting our observations that SChLAP1 promotes aggressive cancer (**Figure 5.21**).

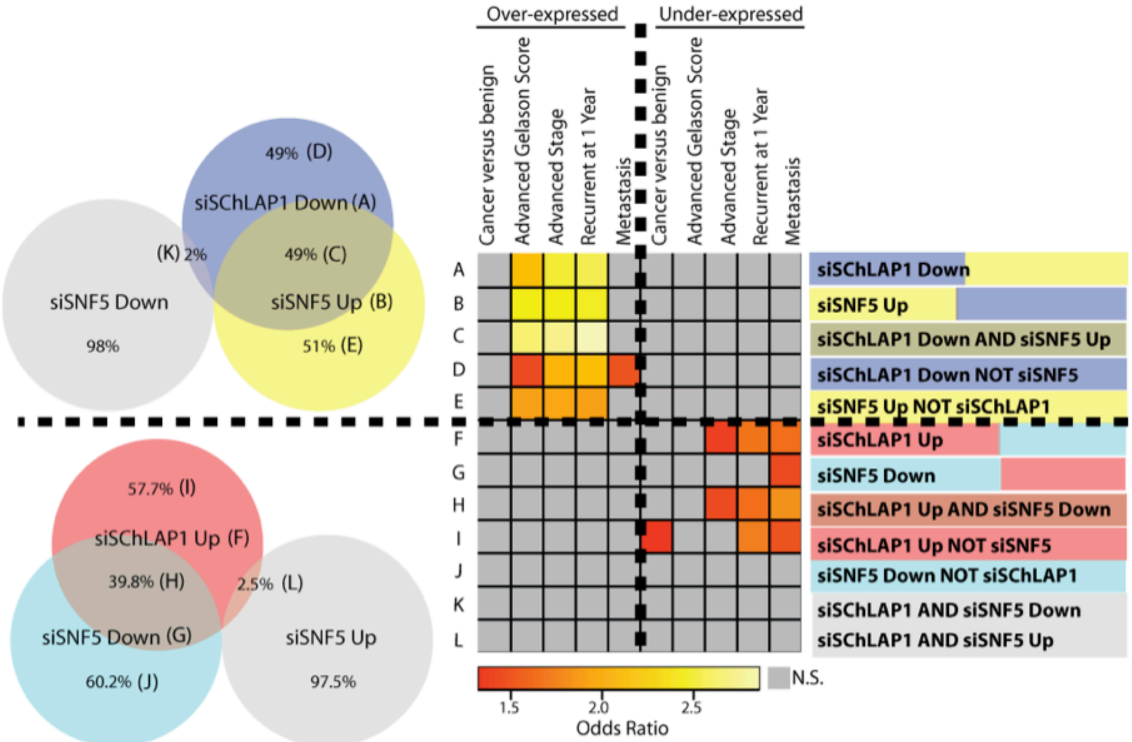


Figure 5.21: *SChLAP1* and *SNF5* co-regulate genes associated with prostate cancer aggressiveness. The top 10% of up- or down-regulated genes for *SNF5*-knockdown or *SChLAP1*-knockdown microarrays in 22Rv1 and LNCaP were intersected to generate an overlapping gene signature for these knockdown experiments. This signature was analyzed for overlap with the Taylor Prostate 3 Oncomine Concept29 for disease aggressiveness. Left, Venn diagrams demonstrating overlap of *SChLAP1* and *SNF5*-knockdown experiments. Right, a heatmap visualization showing statistical (q < 0.05) overlap of gene signatures from the *SNF5* and *SChLAP1* knockdowns with prostate cancer aggressiveness concepts from Oncomine. Odds ratios from the comparisons with q-values <0.05 are shown. One-sided Fisher's exact tests were used for significance.

SChLAP1 co-immunoprecipitates with the SWI/SNF complex

To examine the mechanism of *SChLAP1* regulation of the SWI/SNF complex, we next assessed whether *SChLAP1* regulated *SNF5* itself. Although *SChLAP1* and *SNF5* mRNA levels were comparable in our cohort of human prostate samples (**Figure 5.23a**), we failed to detect any change in *SNF5* protein abundance by Western blot following *SChLAP1* knockdown or overexpression (**Figure 5.23b**), suggesting that *SChLAP1* regulates SWI/SNF activity post-translationally.

As lncRNAs are known to coordinate the function of epigenetic complexes through direct RNA-protein binding, we performed RNA immunoprecipitation assays (RIP) for *SNF5* in 22Rv1 and LNCaP cells. We found that endogenous *SChLAP1*, but not other cytoplasmic or nuclear lncRNAs such as *PCA3*, *PCAT-1*, *MEG3*, *ANRIL*, and *MALAT1*[156, 202], robustly co-immunoprecipitated with *SNF5* protein under both native conditions (**Figure 5.22b**) and UV-crosslinked conditions (**Figure 5.23c**). In addition, we observed that *SChLAP1* did not co-immunoprecipitate with androgen receptor (**Figure 5.22b**), another abundant nuclear protein in prostate cells. Furthermore, we found that both *SChLAP1* isoform #1 and isoform #2 robustly co-immunoprecipitated with *SNF5* in our RWPE overexpression models (**Figure 5.22c**), but not other lncRNAs, including *AK093002* and *LOC145837*, two prostate lncRNAs expressed in RWPE (**Figure 5.23d**). As controls, RIP experiments for *SNRNP70* demonstrated strong binding of this protein to *U1* in all cell lines evaluated (**Figure 5.23e,f**).

SChLAP1 impairs SNF5 genomic binding

Given that *SChLAP1* robustly bound and antagonized the activity of the SWI/SNF complex *in vitro*, we hypothesized that *SChLAP1* may attenuate the ability of SWI/SNF proteins to bind genomic DNA. To investigate this possibility we performed ChIP-Seq of *SNF5* in our RWPE-*LacZ* and RWPE-*SChLAP1* overexpression models. We validated our ChIP pull-down with Western Blots for *SNF5* (**Figure 5.24a**), and sequenced ChIP-Seq libraries using an Illumina Hi-Seq 2000. We aligned the sequence reads to the human genome, called significantly enriched peaks with respect to an IgG control, and aggregated peaks from all samples. This analysis resulted in 6,235 genome-wide binding sites for *SNF5* (FDR < 0.05), which were highly statistically enriched for binding sites near gene promoters (**Figure 5.24b**), supporting other recent genome-wide studies of SWI/SNF complex binding[230-232].

To determine whether *SChLAP1* expression modified *SNF5* genomic binding, we compared the strength of *SNF5* binding across these 6,235 genomic sites in RWPE cells overexpressing *LacZ* or *SChLAP1* isoform #1 or #2. We found a dramatic decrease in *SNF5* genomic binding as a result of *SChLAP1* overexpression (**Figure 5.22d** and **Figure 5.24c**). A summary of the sequence reads surrounding each peak confirmed the considerable attenuation of *SNF5* binding in RWPE cells expressing either *SChLAP1* isoform #1 or #2 (**Figure 5.22e**). Of the 1,299 *SNF5* peaks occurring within 1kb of a gene promoter, 390 of these promoters decreased ≥2-fold in relative *SNF5* binding (**Figure 5.24d**). To verify these findings independently, we next performed ChIP for *SNF5* in 22Rv1-shNT and 22Rv1 sh-*SChLAP1* cells, with the hypothesis that inhibition of *SChLAP1* should increase *SNF5* genomic binding. Using ChIP-PCR, we found that 3 of 4 *SNF5* target genes showed a dramatic increase in *SNF5* binding (**Figure 5.24e**), confirming our predictions.

Finally, we sought to characterize the relationship between *SNF5* genomic binding and *SChLAP1*-mediated gene expression changes. We performed gene expression microarrays of RWPE cells overexpressing *LacZ* or these two *SChLAP1* isoforms, defined genes with highly significant changes in expression, and intersected the microarray data with the ChIP-Seq data. We observed that a significant subset of genes with ≥2-fold relative decrease in *SNF5* genomic binding were dysregulated when *SChLAP1* was overexpressed (**Figure 5.24f**). Decreased *SNF5* binding was primarily associated with downregulation of target gene expression, although a smaller subset of genes was upregulated, consistent with the fact that SWI/SNF binding can regulate gene expression in both directions[227]. We next performed integrative analysis of the microarray data with the *SNF5* ChIP-Seq data using GSEA and observed a significant enrichment for genes that were repressed when *SChLAP1* was overexpressed (q-value = 0.003, **Figure 5.22f**). Overall, these data argue that *SChLAP1* overexpression antagonizes SWI/SNF complex function by attenuating the genomic binding of this complex, thereby impairing its ability to regulate gene expression properly.

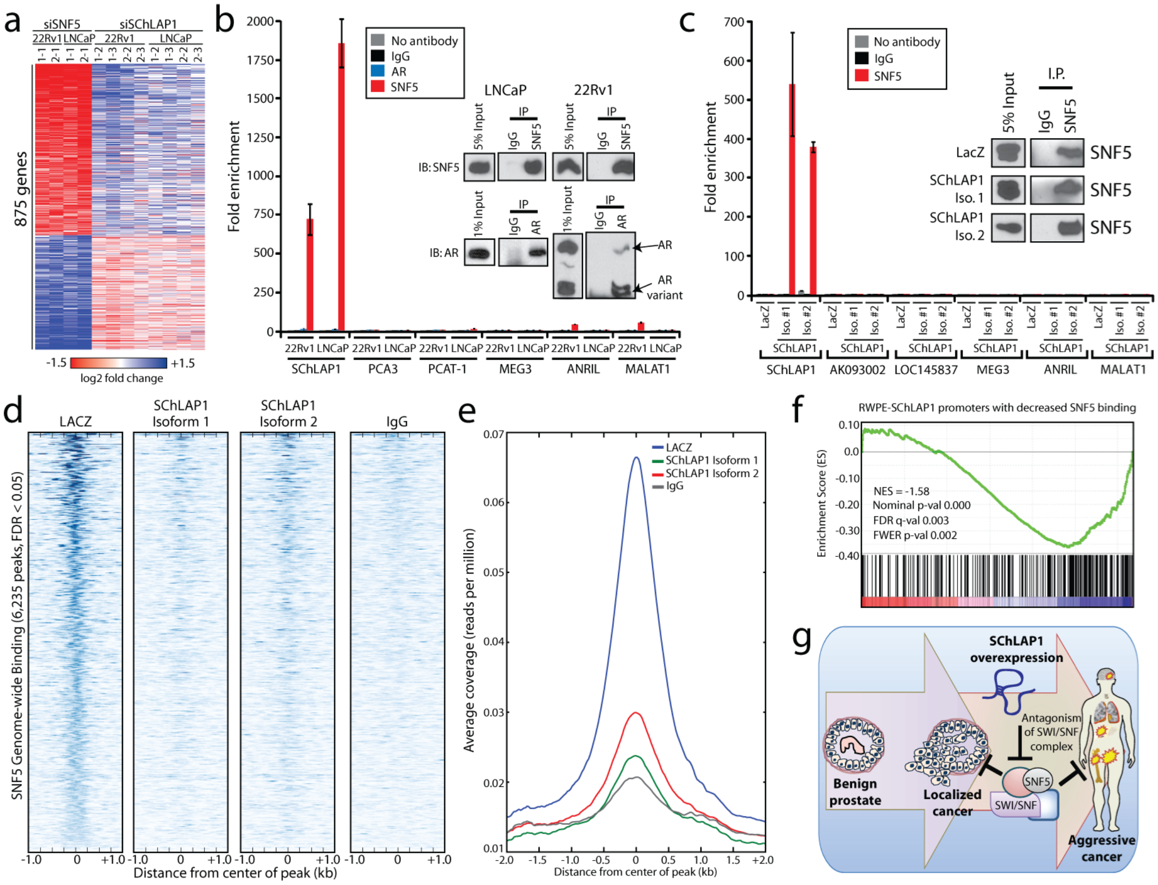


Figure 5.22: *SChLAP1* antagonizes *SNF5* function and attenuates *SNF5* genome-wide localization. (a) Heatmap results for *SChLAP1* or *SNF5* knockdown in LNCaP and 22Rv1 cells demonstrates opposing effects on gene expression regulation by *SNF5* and *SChLAP1*. (b) RNA immunoprecipitation (RIP) of *SNF5* demonstrates *SChLAP1* binding to SNF5 in 22Rv1 and LNCaP cells. Other lncRNAs serve as negative controls. Data are mean +/- S.E.M. (c) RIP analysis of *SNF5* in RWPE cells overexpressing LacZ, *SChLAP1* isoform #1, or *SChLAP1* isoform #2. Other lncRNAs serve as negative controls. Data are mean +/- S.E.M. (d) ChIP-Seq for *SNF5* demonstrates genome-wide loss of *SNF5* binding upon overexpression of *SChLAP1* in RWPE prostate cells. A heatmap represents the interval ±1kb surrounding the called *SNF5* peak. (e) Summary of heatmap data in (d) shown for a ±2kb window surrounding *SNF5* ChIP-Seq peaks. (f) Gene set enrichment analysis results showing significant enrichment of ChIP-Seq promoter peaks with >2-fold loss of *SNF5* binding for underexpressed genes in RWPE-*SChLAP1* cells. (g) A model of *SChLAP1* activity in prostate cancer.

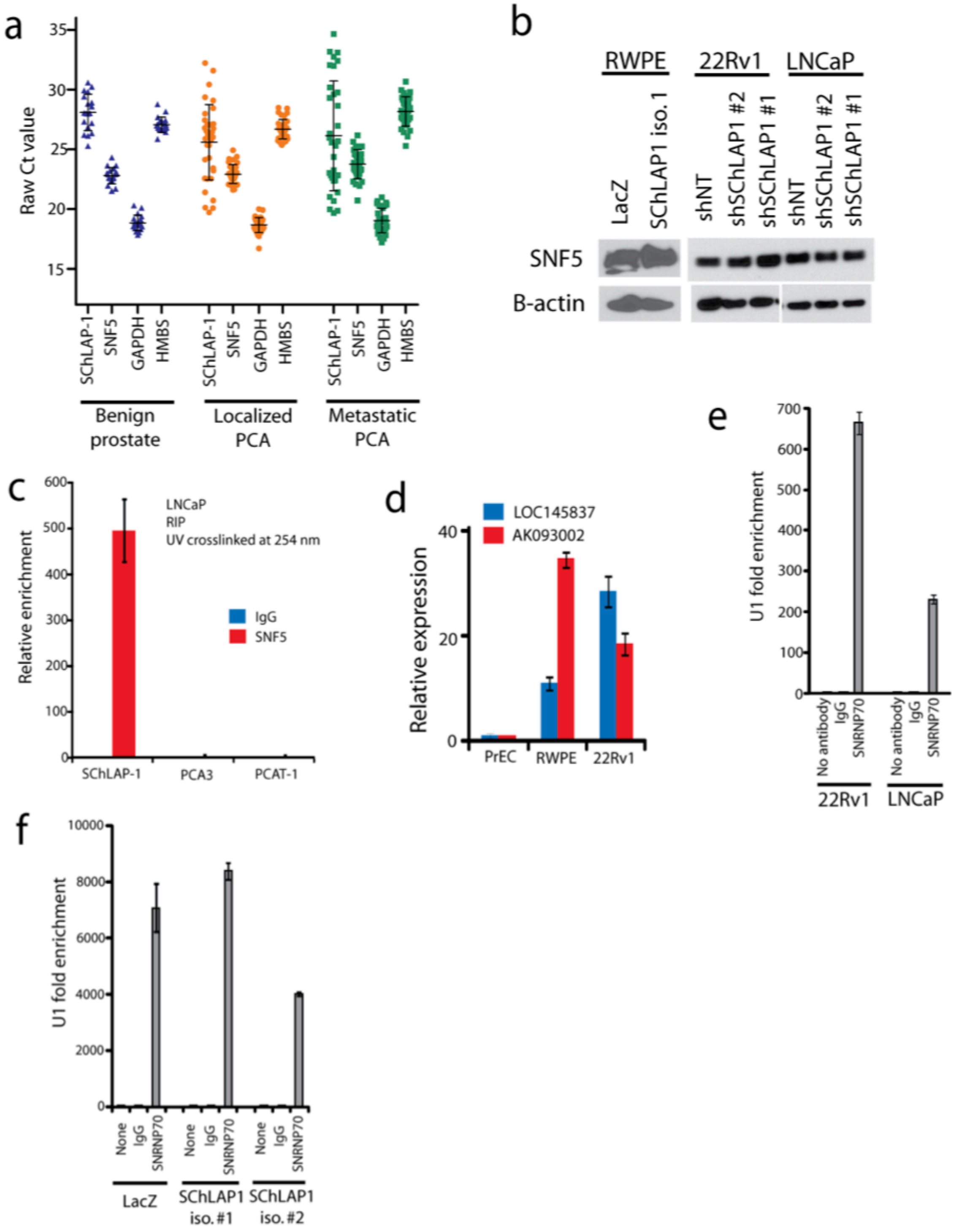


Figure 5.23: *SChLAP1* and *SNF5* expression level and RNA-protein binding of *SChLAP1* with *SNF5*. (a) Relative abundance of *SChLAP1* compared to the SWI/SNF complex in human prostate tissues. qPCR cycle threshold (Ct) values for *SChLAP1*, *SNF5*, *GAPDH*, and *HMBS* are shown. *SChLAP1*-positive samples display Ct values in the low 20s, which is consistent with the abundance of *SNF5*. (b) Western blot analysis of *SNF5* protein abundance in prostate cancer cells either overexpressing *SChLAP1* (RWPE) or with stable knockdown of *SChLAP1* (22Rv1, LNCaP). (c) *SChLAP1* binding to *SNF5* protein by UV-crosslinked RIP assays using UV at 254nM. (d) Expression of *AK093002* and *LOC145837* in prostate cell lines. qPCR data were normalized to the average of *GAPDH* + B-actin and compared to PREC primary

non- immortalized prostate cells. Error bars indicate S.E.M. Expression of these genes in RWPE is comparable to their expression in 22Rv1. (e) RNA-IP experiments for *SNRNP70* in LNCaP and 22Rv1 shows binding of *SNRNP70* to the *U1* ncRNA, indicating specificity of the RNA-IP experiments. Error bars indicate S.E.M. (f) Control *SNRNP70* experiments in the RWPE-*SChLAP1* overexpression models. Enrichment of *U1* is shown as a control for *SNRNP70* IP experiments. Error bars indicate S.E.M.

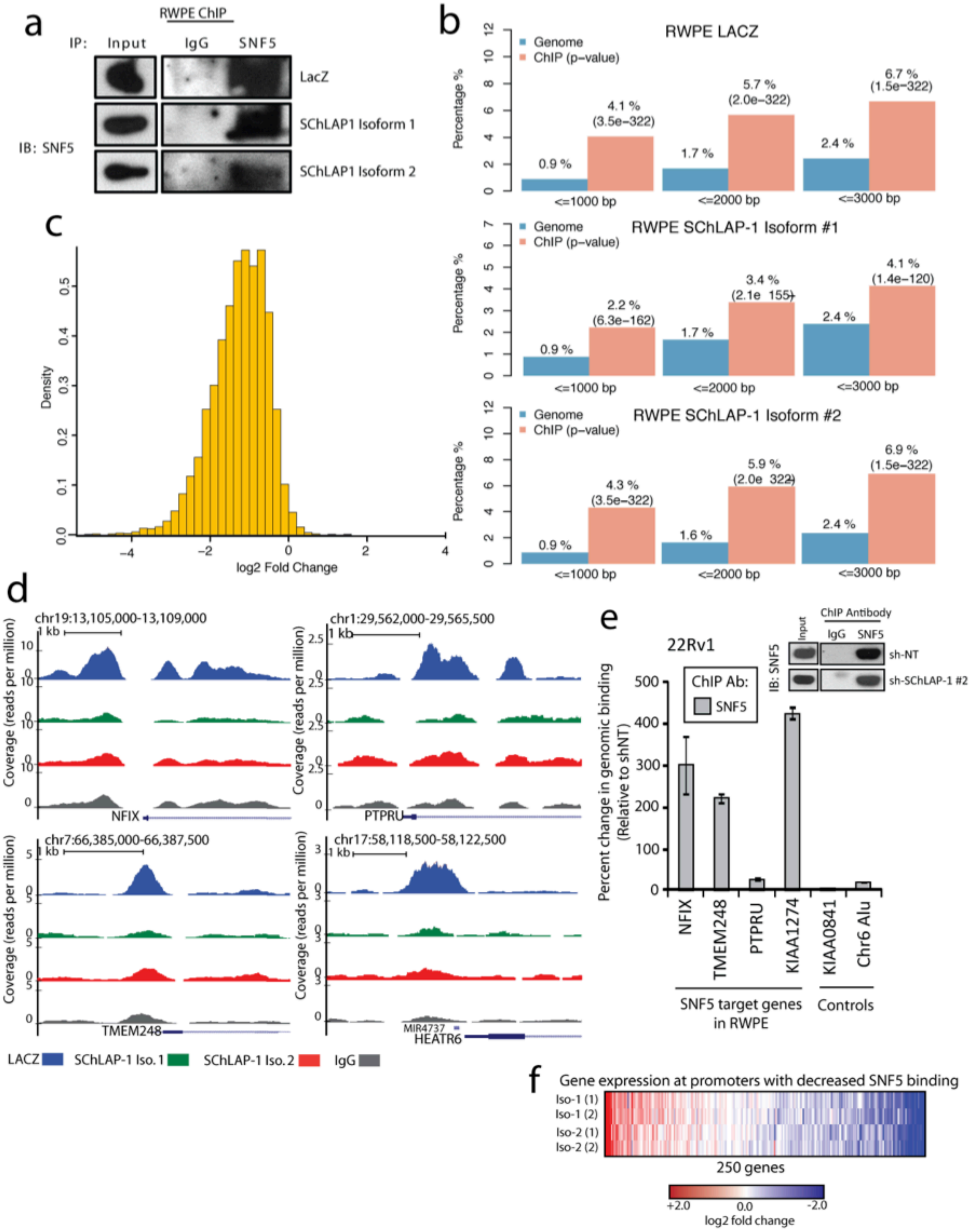


Figure 5.24: *SChLAP1* expression disrupts genomic binding of SNF5.

(a) ChIP for *SNF5* protein followed by Western blot. (b) Bar plots showing enrichment for *SNF5* ChIP-Seq reads at RefSeq gene promoters across the RWPE-*LacZ*, RWPE- *SChLAP1*-Isoform-1 and RWPE-*SChLAP1*-Isoform-2 samples. Blue bars indicate percentage of genomic DNA and red bars indicate percentage of all ChIP-Seq reads in each sample along with the p-value corresponding to the statistical significance of the difference between the blue and red bars. The CEAS software[233] was used to generate these plots and compute the enrichment. (c) Histogram showing the relative log2 fold-change between RWPE-*LacZ* and RWPE-*SChLAP1* (average of both isoforms) coverage across 6,235 genome-wide peaks. (d) Example ChIP-Seq binding sites for SNF5 on gene promoters. SNF5 binding is higher at gene promoters in RWPE-LacZ cells and decreased upon *SChLAP1* overexpression. (e) ChIP for *SNF5* in 22Rv1 shNT and 22Rv1 sh- *SChLAP1* #2. ChIP-PCR for 3 of 4 target genes of *SNF5* in RWPE demonstrates an increase in *SNF5* binding upon *SChLAP1* knockdown. *KIAA0841* and Chr6 Alu serve as negative controls. Data are represented as percent change in genomic binding relative to shNT after being normalized to IgG controls. The inset western blot indicates immunoprecipitation efficiency for SNF5. (f) Heatmap showing the showing the gene expression of RWPE-*SChLAP1* cells (Isoform 1 is labeled as Iso-1 and Isoform 2 is labeled as Iso-2) across 250 genes that exhibited a >2-fold decrease in *SNF5* binding upon *SChLAP1* overexpression. Gene expression is shown as log2 fold-change relative to RWPE-*LacZ*.

Discussion

Here, we characterized two previously undescribed non-coding regions of the human genome that emerged from our *ab initio* assembly and analysis of poly-A+ RNA from a cohort of prostate cancers[27] (**Chapter 4**). *PCAT-1*, which is expressed from the 8q24 'hotspot' implicated by GWAS studies, was markedly overexpressed in primary tumors and metastases. In certain cases *PCAT-1* transcript was repressed by PRC2, and patterns of *PCAT-1* and PRC2 expression stratified patient tissues into molecular subtypes distinguished by expression signatures of *PCAT-1*–repressed target genes. In cases where *PCAT-1* was not repressed by PRC2, it promoted cell proliferation through transcriptional regulation of distinct target genes, including the *BRCA2* tumor suppressor[234](**Figure 5.10e**). Taken together, our findings suggested that *PCAT-1* is a transcriptional repressor implicated in a subset of prostate cancer patients.

Furthermore, we have discovered *SChLAP1*, a highly prognostic lncRNA that is abundantly expressed in 15-30% of prostate cancers and aided the discrimination of aggressive from indolent forms of this disease. Mechanistically, we find that *SChLAP1* coordinates cancer cell invasion *in vitro* and metastatic spread *in vivo*. Moreover, we characterize an antagonistic *SChLAP1*-SWI/SNF axis in which *SChLAP1* impairs *SNF5*-mediated gene expression regulation and genomic binding (**Figure 5.22g**). Thus, while other lncRNAs such as *HOTAIR* and *HOTTIP* are known to assist epigenetic complexes such as PRC2 and MLL by facilitating their genomic binding and enhancing their functions[168, 169, 172, 235], *SChLAP1* is the first lncRNA, to our knowledge, that impairs a major epigenetic complex with well-documented tumor suppressor function[221, 222, 225, 227, 228, 236]. Taken together, our discovery of *SChLAP1* has broad implications for cancer biology and provides evidence for the role of lncRNAs in the progression of aggressive cancers.

Chapter 6: Concluding Remarks

Utilizing RNA-Seq-based discovery algorithms for pan-cancer analyses

As described in Chapters 2-5, applying our novel algorithms ChimeraScan and AssemblyLine to studies of single cancer types to led to the discovery of recurrent classes of gene fusions and lncRNAs that may serve as novel disease markers, respectively. Given these successes, applying these methodologies to pan-cancer studies would likely yield similar insights and may expose overarching patterns of disease biology. However, extending the work presented in this dissertation to large compendia studies places unprecedented demands upon computational infrastructures. In the following sections, we discuss these challenges and propose an automated framework called Oncoseq that addresses the need for a standardized RNA-Seq analysis pipeline that can be deployed on large-scale supercomputing systems.

Exponential increases in RNA-Seq data generation will place unprecedented demands on computational infrastructures and bioinformatics algorithms

Today we are faced with an onslaught of RNA-Seq data large enough to overwhelm existing computational infrastructures. As of November 2012, our lab had compiled 1,140 RNA-Seq libraries from many cancer types (**Figures 6.1** and **6.2**). Analyzing these libraries required 60-500 processor core-hours per sample, which amounted to approximately 136,000 processor core-hours for the entire compendia. We accomplished this analysis at a cost of approximately $3,000 USD on supercomputing resources provided by the University of Michigan Center for Advanced Computing (CAC). The University of Michigan Medical School heavily subsidized the cost of this analysis.

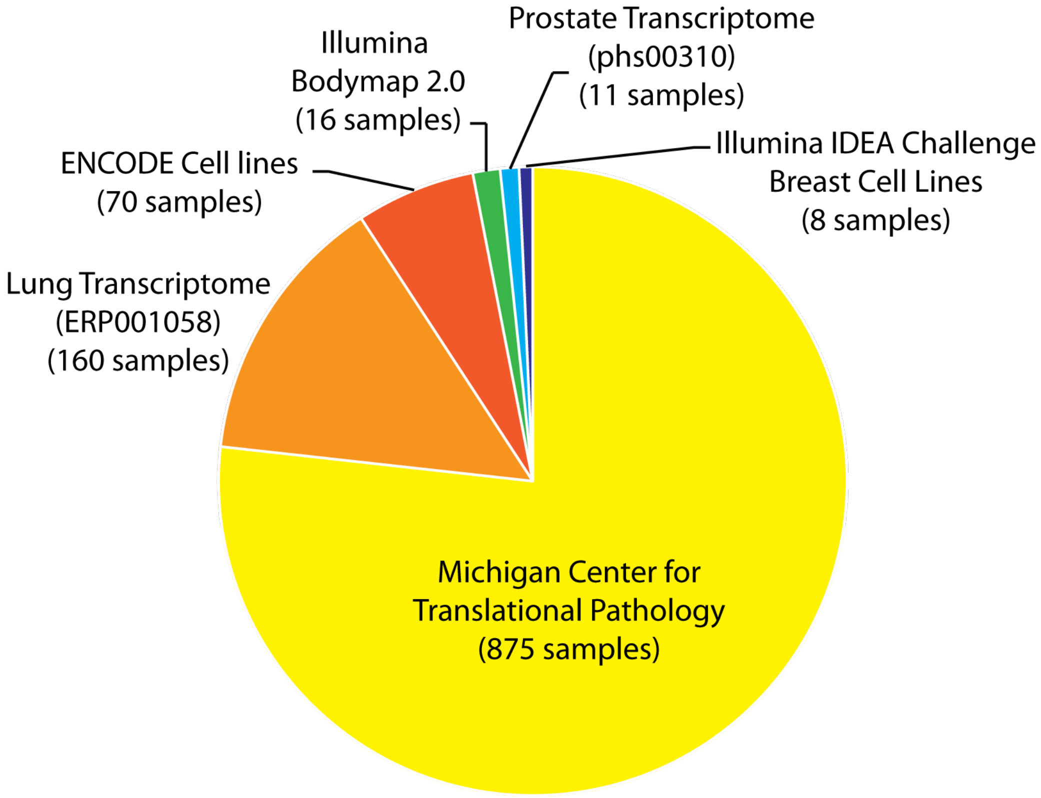


Figure 6.1: Composition of the MCTP compendia as of November, 2012
Pie chart showing the number of samples contributed by MCTP and various public sources

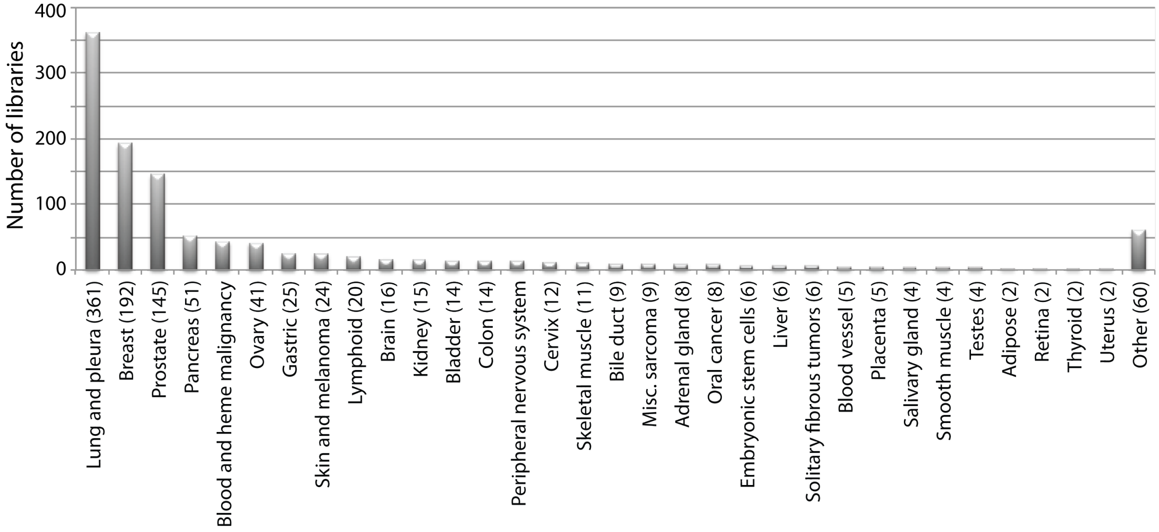


Figure 6.2: Tissue types represented by the MCTP compendia as of November, 2012
Bar graph showing the number of libraries analyzed by tissue type

The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) has recently made RNA-Seq data from over 5,000 samples available for download at the new Cancer Genomics Hub (CGHub) facility (https://cghub.ucsc.edu) with plans to expand to over 10,000 samples in the near future. Additionally, the Cancer Genome Characterization Initiative (CGCI) (http://cgap.nci.nih.gov/cgci.html) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative expect to generate cohorts with hundreds of samples as well. Aside from large consortiums, smaller labs and pharmaceutical companies will likely produce equal if not greater amounts of data with the help of sequencing services such as the Beijing Genomics Institute (BGI). These samples are being sequenced at greater depth than before using instruments such as the Illumina HiSeq 2000. Test analysis of these samples often required more than 300 processor-hours per sample, and we roughly approximate that the full TCGA dataset could require more than 8,300 processor-months. At the subsidized University of Michigan rate this analysis would likely cost well over $100,000 USD in raw computational time, not including the cost of data storage. Such expenses would likely be prohibitive for many labs and requires access to large supercomputing facilities. We note that accomplishing such analysis on pay-as-you-go cloud computing environments would be substantially more expensive as these services charge many times more per processor hour and tack on fees for data transfer.

Given the overwhelming cost of data analysis and storage we envision the need for new federally funded programs that make computational resources available. Instead of burdening small labs, computational loads should be farmed out to large cloud-computing facilities and administrated by specialized groups in order to decrease the technical burden of entry into the field.

Improvements in algorithmic efficiency may mitigate computational demand

In addition to political changes that make data analysis more affordable, the field of bioinformatics must continue to improve algorithm efficiency. For example, the Bowtie[135, 149] and BWA algorithms[237, 238] innovated the use of the novel genome indexing schemes that vastly improved the efficiency of short read alignment. Recently, a new algorithm called STAR appears to produce phenomenal efficiency gains for RNA sequencing datasets, outperforming the Tophat algorithm by a factor of >50[55]. Algorithmic innovations such as this may thus allow computers to keep pace with genomics. We expect that increased involvement by computer engineers could also improve to the performance of genomics algorithms. Graphics processing units (GPUs) based algorithms and field programmable gate array (FPGAs) show remarkable promise for bioinformatics applications that could benefit from extensive parallelization.

Establishing standards for RNA-Seq analysis

Over the past four years, a trend in the field of bioinformatics algorithms has been the parallel and redundant development of software tools. As described in Chapter 2, a growing suite of tools for gene fusion detection now exists, but little is known about the comparative advantages and disadvantages of the tools. With time one expects that ease-of-use, performance, and continued maintenance of the tools will determine their usefulness. To facilitate the evaluation of software tools, we propose the establishment of standardized benchmarks for testing multiple types of algorithms. For gene fusion discovery, such benchmarks would incorporate simulated chimeras in addition to numerous examples of independently validated gene fusion candidates. If such a benchmark dataset existed, existing software tools could be vetted and forthcoming tools could optimize their approaches in a robust and unbiased manner. Creation of such a dataset would involve curating the published literature and performing additional validation experiments where necessary. The idea of standardized benchmarks has been successfully adopted by the field of computer architecture by the Standard Performance Evaluation Corporation (SPEC), a group that manages the selection of benchmarks and the publication of performance results (http://www.spec.org). Mirroring such an approach could greatly benefit high-throughput sequencing data analysis.

The Oncoseq framework for RNA-Seq analysis

In light of the recognized need for a standardized analysis pipeline for processing large number of RNA sequencing experiments, including datasets from TCGA, we prototyped a standardized and comprehensive RNA-Seq analysis workflow called Oncoseq. Oncoseq is unlike frameworks such as Galaxy or Firehose (http://www.broadinstitute.org/cancer/cga/Firehose) that permit extensive customization and selection from a wide array of tools[239]. Instead, Oncoseq represents a curated, tested, and tightly integrated set of analyses that includes third-party tools such as Tophat and Cufflinks as well as tools developed by our lab (**Figure 6.3**). Achieving a standardized analysis framework has a number of important advantages: (1) many nuances of individual software tools can be abstracted away, (2) common steps shared by multiple tools can be shared, (3) data reproducibility can be guaranteed, (4) the approach can be scaled to process large amounts of data.

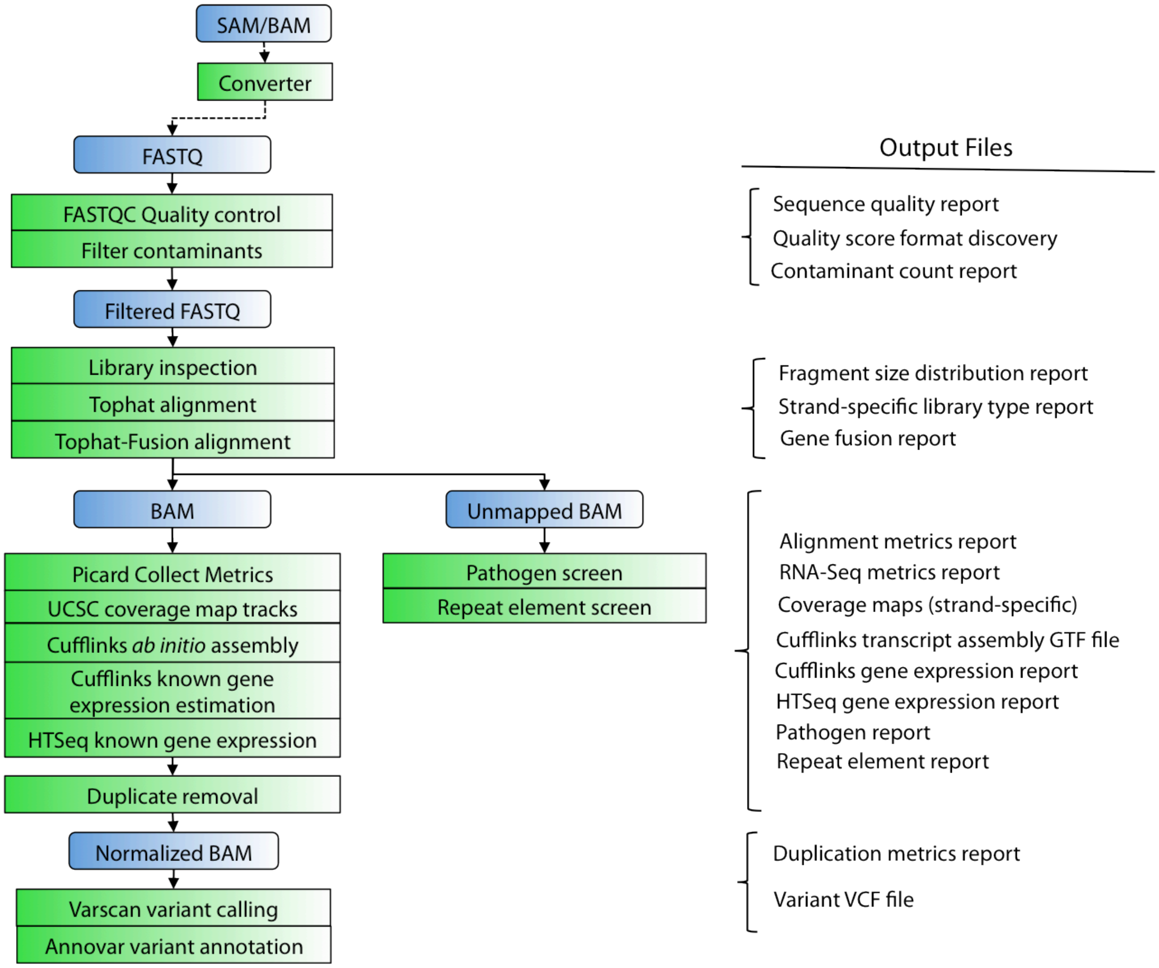


Figure 6.3: Overview of the Oncoseq RNA-Seq analysis framework
Unmapped FASTQ or BAM files feed into the Oncoseq pipeline. (left) Flow chart showing the Oncoseq workflow. (right) Output files produced after each stage of the pipeline.

The primary analyses achieved by Oncoseq are gene fusion detection, *ab initio* transcriptome assembly, gene and transcript abundance assessment, variant calling, pathogen screening, and quantification of repetitive element sequences (**Figure 6.3**). We also expect to accommodate new analysis modules and updates as they become available. A number of additional files provide visualization tracks for the UCSC genome browser[155]. The Oncoseq pipeline is still a work in progress, but to date it has been used to analyze over 1,000 RNA-Seq libraries. In the near future we expect to make the workflow available for other groups to utilize.

Furthering the characterization of cancer-associated RNAs

The establishment of robust computational pipelines such as Oncoseq will facilitate meta-analysis of large cohorts of cancer RNA sequencing datasets. In this section, we discuss unique opportunities for RNA sequencing studies and potential paths forward to making these analyses a reality.

Towards completion of the human transcriptome reference

Given the promise of meta-assembly algorithms such as AssemblyLine (**Chapter 3**) we believe that the full long RNA complement of human cells can now be defined. However, observing the sequence and genomic location of every transcript capable of being produced by an organism requires myriad sequencing datasets that account for gene expression variation across cell lineages, developmental stages, and disease states. Amassing data of this magnitude will likely require years of effort by multiple groups, but poses no additional methodological challenges. In anticipation of large datasets, AssemblyLine has been carefully architected for scalability and each of its steps has been parallelized to take advantage of multi-processor computing systems. A demonstration of the algorithm has now been completed on a set of 1,140 *ab initio* assemblies from a version of our growing RNA-Seq compendia (**Figure 6.2**).

Given the imminent release of thousands of RNA-Seq datasets by the TCGA, we plan to employ AssemblyLine to define the cancer transcriptome at unprecedented depth. Whereas our study of prostate cancers (**Chapter 4**) utilized less than 2 billion mapped reads and the Cabili *et al.* study[78] compiled about 4 billion reads, we anticipate a transcriptome assembly effort with over 250 billion reads. We envision that the reference transcriptomes constructed by AssemblyLine will complement current gene databases. As methods improve and new datasets become available, we hope to continually update these transcript models. We speculate that these efforts will play an important role in shaping our understanding of the human transcriptome.

Incorporation of *de novo* transcriptome assembly

The AssemblyLine algorithm was built upon an *ab initio* assembly approach that relies on reliably aligning sequencing reads to the human genome. By contrast, *de novo* assembly approaches construct full-length transcripts from sequencing reads in the absence of a genome. Results from the recently developed Trinity algorithm suggest that *de novo* assembly will have considerable impact and will be especially useful in the field of cancer genetics[46]. Highly mutated cancer genomes can contain many structural aberrations that deviate from the reference genome. In extreme cases, cancers arise following chromothripsis, a catastrophic shattering of the genome that may generate many fusion genes[240]. Although alignment-based analyses can cope with some of these abnormalities, *de novo* assemblers should be better at deconvoluting aberrant transcriptomes. In principle, *de novo* assembly requires no special considerations to detect events such as cryptic splicing, micro-exons, tiny introns, indels, alternative polyadenylation, tandem duplications, exogenous pathogens or gene fusions. Although the assembly process does not require a reference genome, detecting these events requires the assembled contigs to be long enough to be aligned to an available reference genome. It should be noted that *de novo* assembly requires substantially more computational resources than were previously needed for RNA-Seq analysis. Software optimization will therefore be a key priority in making *de novo* assembly more accessible. Regardless, the initial successes of *de novo* assembly point toward a future where this approach subsumes the need for specialized gene fusion detection or *ab initio* assembly algorithms.

Complementary and alternative RNA sequencing protocols

High-throughput sequencing of size-selected poly-A-selected RNA from whole cells has several important limitations: (1) it neglects classes of non-polyadenylated or bimorphic transcripts such as enhancer RNA[167], (2) it captures steady state RNA levels that cannot infer rates of production and degradation, (3) it does not distinguish protein-coding from non-coding RNA, and (4) RNAs smaller than ~200nt are neglected. Therefore, after defining the human transcriptome by RNA-Seq, we anticipate the need to alternative modalities that provide complementary information. For example, subcellular fractionation followed by ribosome depleted total RNA sequencing has revealed enrichment for transcripts absent from standard RNA-Seq[13, 241, 242]. Also, nascent RNA sequencing can be employed to infer rates of transcription and splicing[164]. Isolation of ribosome-bound mRNA followed by high-throughput sequencing provided evidence for short peptides and non-canonical open reading frames residing in RNAs catalogued as non-coding[243]. Finally, integration of small RNA sequencing data would assist in determining whether long RNAs functions as small RNA precursors. These protocols will serve as complementary tools for RNA interrogation and should be incorporated along with standard RNA-Seq for the global characterization of RNA landscapes in cancer.

The recently appreciated role of lncRNAs in molecular biology suggests the need for techniques to accurately determine global patterns of RNA structure. One such strategy called parallel analysis of RNA structure (PARS) treats RNA with structure-specific enzymes prior to high-throughput sequencing[244]. The results can be used to infer base pairing at nucleotide resolution and can greatly improve the accuracy of RNA structure prediction algorithms. Knowledge of RNA structures is a first step in understanding the interface between lncRNAs and protein complexes. If lncRNAs indeed act as molecular scaffolds, then we expect RNA structural motifs to emerge from these global analyses. The interactions between specific protein complexes and lncRNAs can be complemented by RNA immunopreciptation following by sequencing (RIP-Seq)[245]. We expect that RIP-Seq will provide valuable supporting evidence for the well-established role of lncRNAs as adaptors for chromatin modifying enzymes. In particular, ongoing studies of the interaction between the SChLAP1 lncRNA (presented in **Chapter 5**) and its interaction with the SWI/SNF complex could be greatly facilitated by RIP-Seq.

Underexplored dimensions of RNA-Seq data

The field of cancer genomics should continue to pursue aspects of RNA-Seq data that may lead to the discovery of new disease-specific events. In particular the appreciation of widespread RNA and DNA sequence differences suggested the possibility of cancer-specific RNA editing events[246]. Confirming these suspicions, the recent discovery of RNA editing event in the encoding antizyme inhibitor 1 (*AZIN1*) that predisposes to hepatocellular carcinoma (HCC) provides a glimpse at what could likely be a global phenomenon[247, 248]. RNA editing in cancer could easily be explored using pairs of matched normal and tumor tissue. Such data is available in abundance from the TCGA and other sources, and methods for detecting these events have been proposed[249, 250].

Detecting RNA editing events requires monitoring changes in allele frequencies between conditions. This class of events can be broadly characterized as allele-specific expression (ASE). By expanding the use of the underlying sequencing information provided by RNA-Seq it is possible to detect ASE events on a global scale[251-253]. Although detecting statistically significant ASE requires high coverage depth, we believe that ASE may provide important clues and supporting evidence of intricate isoform-level regulation that may be associated with disease.

A bright future for RNA sequencing

In the past five years RNA-Seq has revolutionized the study of transcriptomes and fostered the emergence of a squadron of new bioinformatics methods. As the size of the data grows and the quality improves, so will the demand for innovative computational solutions. We expect that the algorithms developed for this thesis work will need to evolve considerably to keep pace with improvements in technology, and may one day become obsolete. Nevertheless, we hope that the ideas encapsulated by these methods will continue to be useful to others.

If nothing else, the discoveries of cancer-associated transcripts present in this thesis provide glimpses into the multitudinous roles of RNA in cellular biology. Our observations of highly prognostic lncRNAs motivate the expanded study of this underappreciated layer of biological complexity and the development of novel therapeutic approaches for targeting RNA. We enthusiastically anticipate these innovations and look forward to future roles for RNA sequencing in the diagnosis and treatment of disease.

Appendices

Appendix A: Gold-standard chimeras used to evaluate ChimeraScan

Cell line	Class	Chimera	Detected
VCaP	Intra-chromosomal	TMPRSS2-ERG	Yes
VCaP	Intra-chromosomal	INPP4A-HJURP	Yes
VCaP	Inter-chromosomal	USP10-ZDHHC7	Yes
VCaP	Intra-chromosomal	HJURP-EIF4E2	Yes
VCaP	Intra-chromosomal	RC3H2-RGS3	Yes
VCaP	Read-through	ZNF577-ZNF649	Yes
VCaP	Intra-chromosomal	LMAN2-AP3S1	Yes
VCaP	Intra-chromosomal	SPOCK1-TBC1D9B	No
VCaP	Inter-chromosomal	ZDHHC7-ABCB9	Yes
VCaP	Inter-chromosomal	TIA1-DIRC2	Yes
VCaP	Intra-chromosomal	PIK3C2A-TEAD1	Yes
LNCaP	Inter-chromosomal	MIPOL1-DGKB	Yes
LNCaP	Inter-chromosomal	MRPS10-HPR	Yes
LNCaP	Read-through	C19orf25-APC2	Yes
LNCaP	Read-through	SLC45A3-ELK4	Yes
LNCaP	Intra-chromosomal	BMSUN-PSPC1	Yes
LNCaP	Intra-chromosomal	RERE-PIK3CD	Yes
MCF7	Inter-chromosomal	BCAS4-BCAS3	Yes
MCF7	Intra-chromosomal	ARFGEF2-SULF2	Yes
MCF7	Inter-chromosomal	AHCYL1-RAD51C	Yes
MCF7	Inter-chromosomal	ARHGAP19-DRG1	Yes
MCF7	Intra-chromosomal	MYO9B-FCHO1	Yes
MCF7	Intra-chromosomal	BC017255-TMEM49	Yes
MCF7	Read-through	DEPDC1B-ELOVL7	Yes
MCF7	Read-through	PAPOLA-AK7	Yes
MCF7	Intra-chromosomal	STK11-MIDN	Yes
MCF7	Inter-chromosomal	TEX14-PTPRG	No
MCF7	Inter-chromosomal	SULF2-PRICKLE2	Yes
MCF7	Read-through	RPS6KB1-TMEM49	Yes
MCF7	Read-through	CXorf15-SYAP1	Yes

Appendix B: Novel chimeras discovered by ChimeraScan

sample name	5' gene	3' gene	type	fragments	spanning
VCAP	TMPRSS2	ERG	Intrachromosomal	206	85
VCAP	ZDHHC7	ABCB9	Interchromosomal	18	8
VCAP	MAML3	MED12,TNRC11	Interchromosomal	9	8
VCAP	INPP4A	HJURP	Intrachromosomal_Complex	11	9
VCAP	RC3H2,DKFZp667B165	RGS3	Intrachromosomal_Complex	14	5
VCAP	LENEP	KLK2	Interchromosomal	14	13
VCAP	HJURP	EIF4E2	Intrachromosomal_Complex	11	5
VCAP	PDGFA	WASH1	Interchromosomal	8	4
VCAP	PDGFA	DKFZp434K1323,WASH1	Interchromosomal	8	4
VCAP	PDGFA	WASH1	Interchromosomal	8	4
VCAP	PDGFA	WASH,DKFZp434K1323	Interchromosomal	8	4
VCAP	PIK3C2A	TEAD1	Intrachromosomal_Complex	9	5
VCAP	LMAN2	AP3S1	Intrachromosomal_Complex	4	3
VCAP	ZDHHC7	H3F3B	Interchromosomal	4	3
VCAP	KIAA1267	ARL17P1	Intrachromosomal	4	3
VCAP	KIAA1267	ARL17P1,ARL17	Read_Through	4	3
VCAP	VWA2	PRKCH	Interchromosomal	5	2
VCAP	AK311578	G3BP2	Read_Through	5	2
VCAP	HSF1	RERE,KIAA0458	Interchromosomal	3	2
VCAP	TLK1,TLK2	AX747598	Interchromosomal	7	2
VCAP	TLK1,TLK2	BC006361	Interchromosomal	7	2
VCAP	TLK1,TLK2	AL137655	Interchromosomal	7	2
VCAP	TLK1,TLK2	AL137733	Interchromosomal	7	2
VCAP	TLK1,TLK2	FAM157A	Interchromosomal	7	2
VCAP	ZNF57	LPPR2	Intrachromosomal	6	1
VCAP	EEF1DP3	FRY	Read_Through	5	1
VCAP	TIA1	DIRC2	Interchromosomal	5	1
VCAP	KIAA1592,CNNM4	PARD3B	Intrachromosomal	3	1
VCAP	C16orf70	C16orf48	Intrachromosomal_Complex	3	1
VCAP	NBPF3	NBPF1	Intrachromosomal_Complex	5	1
VCAP	HNRNPK,HNRPK	RPS3	Interchromosomal	3	2
VCAP	PDIA6	PTEN	Interchromosomal	3	1
VCAP	NCKIPSD	CELSR3	Read_Through	2	1
VCAP	NDUFAF2	MAST4	Intrachromosomal	2	1
VCAP	GNAS	RPLP0	Interchromosomal	2	1
VCAP	FZR1	CTBP1	Interchromosomal	2	2
VCAP	TYMP	SCO2	Read_Through	2	1
VCAP	MAP7	APP	Interchromosomal	2	1
VCAP	STIP1	CFL1	Intrachromosomal_Complex	2	1
VCAP	RPL10	FGFRL1	Interchromosomal	2	1
VCAP	COBRA1	C9orf167	Read_Through	2	1
VCAP	NUCKS1	ITPR1	Interchromosomal	2	1
VCAP	SHANK2	SHANK1	Interchromosomal	10	0
VCAP	USP10	ZDHHC7	Intrachromosomal_Complex	8	0
VCAP	BC110060	LRFN1	Read_Through	8	0
VCAP	EEF1A2	HSD11B2	Interchromosomal	15	0
VCAP	PDE4DN2,PDE4D	C5orf47	Intrachromosomal_Complex	7	0
VCAP	SH3D20	ARHGAP27	Read_Through	7	0
VCAP	PTEN	PTENP1	Interchromosomal	7	0
VCAP	FMR1	TM9SF3	Interchromosomal	15	0
VCAP	DKFZp666P032,RANBP17	DOCK2	Intrachromosomal	6	0
VCAP	LOC148189	AK094188	Read_Through	5	0
VCAP	KLK2	KLK3	Read_Through	5	0
VCAP	CR597916	BNIP3	Interchromosomal	7	0
VCAP	BC090058,LOC554248	POM121	Intrachromosomal	8	0
VCAP	OK/SW-cl.16	CR615613	Interchromosomal	5	0
VCAP	TTTY15	USP9Y	Read_Through	4	0
VCAP	KIAA0464,NOS1AP	C1orf226	Read_Through	4	0

VCAP	LOC387647	RAB18	Intrachromosomal	6	0
VCAP	PRELID1	PX19	Interchromosomal	6	0
VCAP	PTMA	LOC441454	Interchromosomal	9	0
VCAP	EEF1A2	ATP5S	Interchromosomal	6	0
VCAP	PRKRIP1,PMS2L3	POM121	Intrachromosomal	7	0
VCAP	HLA-G	hla-b	Intrachromosomal_Complex	5	0
VCAP	SON	PEA15	Interchromosomal	2	0
VCAP	EEF1A2	TECPR1	Interchromosomal	5	0
VCAP	SRP14	FOXP2	Interchromosomal	7	0
VCAP	SRP14	RBMS3	Interchromosomal	7	0
VCAP	HLA-C,hla-b,HLA-B	HLA-A,HLA-G,HLA-F	Intrachromosomal_Complex	6	0
VCAP	HNRNPUL1	WASL	Interchromosomal	5	0
VCAP	PRELID1	PX19	Interchromosomal	5	0
VCAP	UNC13B	UNC13A	Interchromosomal	3	0
VCAP	GAS2L1	PPP1R9B	Interchromosomal	3	0
VCAP	SSSCA1	FAM89B	Read_Through	3	0
VCAP	KLK2	TNK2	Interchromosomal	3	0
VCAP	KLK3	KLK2	Read_Through	3	0
VCAP	hCPE-R,CLDN4	LASS2	Interchromosomal	3	0
VCAP	SF3A2	AMH	Read_Through	3	0

sample name	5' gene	3' gene	type	fragments	spanning
LNCAP	RERE,KIAA0458	PIK3CD	Intrachromosomal_Complex	16	7
LNCAP	NCOR1	SELENBP1	Interchromosomal	6	6
LNCAP	GPS2,KIAA1787	MPP2	Intrachromosomal	18	5
LNCAP	SMA4	NAIP	Intrachromosomal_Complex	11	7
LNCAP	BCL8	NBEA	Interchromosomal	4	4
LNCAP	VMAC	CAPS	Read_Through	6	3
LNCAP	SNX9	CYP2C19	Interchromosomal	6	4
LNCAP	LOC728411	GUSBL2	Interchromosomal	6	5
LNCAP	BC035411	NAIP	Intrachromosomal	8	4
LNCAP	HLA-A,HLA-H	HLA-B,hla-b	Intrachromosomal_Complex	4	3
LNCAP	ZNF92	ZNF680	Intrachromosomal_Complex	3	3
LNCAP	KLK4	KRSP1	Read_Through	5	3
LNCAP	BC039389	GATM	Read_Through	5	2
LNCAP	CHCHD10	VPREB3	Read_Through	3	2
LNCAP	FAM117B	BMPR2	Read_Through	9	1
LNCAP	TFDP1	GRK1	Read_Through	11	1
LNCAP	KIAA1128,FAM190B	CYP2C19	Intrachromosomal	7	1
LNCAP	STX16	GAS5	Interchromosomal	5	4
LNCAP	LQK1	C1orf227	Read_Through	5	1
LNCAP	FAM177B	SOD2	Interchromosomal	4	1
LNCAP	MIPOL1	DGKB	Interchromosomal	3	1
LNCAP	NIPSNAP3A	RPL4	Interchromosomal	3	2
LNCAP	STRF6	ARL2BP	Interchromosomal	3	2
LNCAP	PPP2CA	SKP1	Read_Through	3	1
LNCAP	ATP1A1	EEF1A2	Interchromosomal	2	1
LNCAP	derp10,COPS7A	TVAS5	Interchromosomal	2	1
LNCAP	CCNI	MTND5	Interchromosomal	2	1
LNCAP	TM7SF2	PLEKHB1	Intrachromosomal	2	1
LNCAP	GUK1	N4BP2L2	Interchromosomal	2	1
LNCAP	CHMP5	GAPD,GAPDH	Interchromosomal	2	1
LNCAP	RNF40	DNAJC14	Interchromosomal	2	1
LNCAP	CDH12	AK310013,AK123868	Intrachromosomal	67	0
LNCAP	RLN2	RLN1	Read_Through	31	0
LNCAP	TVAS5	DQ597482	Interchromosomal	65	0
LNCAP	CLDN12	PFTK1	Read_Through	17	0
LNCAP	DQ597482	TVAS5	Interchromosomal	37	0
LNCAP	LOC387647	RAB18	Intrachromosomal	16	0
LNCAP	CR597916	BNIP3	Interchromosomal	14	0
LNCAP	LOC645166	LOC654342	Interchromosomal	18	0
LNCAP	AK311578	G3BP2	Read_Through	9	0
LNCAP	BC110060	LRFN1	Read_Through	9	0
LNCAP	SCNN1A	TNFRSF1A	Read_Through	9	0
LNCAP	BC018860	BC018860	Interchromosomal	11	0

sample name	5' gene	3' gene	type	fragments	spanning
LNCAP	C1QTNF3	AMACR	Read_Through	7	0
LNCAP	EEF1A2	HSD11B2	Interchromosomal	10	0
LNCAP	MRPS10	HPR,HP	Interchromosomal	10	0
LNCAP	BCMSUN	PSPC1	Intrachromosomal	7	0
LNCAP	ATXN3	TRIP11,Trip230	Intrachromosomal	5	0
LNCAP	BC035340	MCF2L	Read_Through	5	0
LNCAP	CR615613	OK/SW-cl.16	Interchromosomal	5	0
LNCAP	LSP1	LOC654342	Interchromosomal	13	0
LNCAP	ANKRD42	RPL32	Interchromosomal	11	0
LNCAP	HSP90Bb	HSP90AB1,HSP90AB3P	Interchromosomal	5	0
LNCAP	GAPD,GAPDH	BC009500	Interchromosomal	11	0
LNCAP	VAMP8	VAMP5	Read_Through	4	0
LNCAP	CIRBP	C19orf24	Read_Through	4	0
LNCAP	KLHL23	SNX22	Interchromosomal	4	0
LNCAP	RPL38	TTYH2	Read_Through	4	0
LNCAP	COPG2	COPG	Interchromosomal	4	0
LNCAP	CR596118	STRF6	Interchromosomal	4	0
LNCAP	LOC100134368	NME4	Read_Through	4	0
LNCAP	DEAF1	SCT	Read_Through	4	0
LNCAP	MKSTYX,STYXL1,DKFZp686O05147	TMEM120A	Read_Through	4	0
LNCAP	OK/SW-cl.16	TVAS5	Intrachromosomal	5	0
LNCAP	AK127238	TNFSF4	Read_Through	5	0
LNCAP	CECR7	AK129567,AK302545	Intrachromosomal_Complex	17	0
LNCAP	LOC643648	MLL3	Interchromosomal	9	0
LNCAP	ADCK4	NUMBL	Read_Through	4	0
LNCAP	UBB	UBB	Interchromosomal	5	0
LNCAP	BAGE5,BAGE2,BAGE	LOC643648	Interchromosomal	7	0
LNCAP	LOC148189	AK094188	Read_Through	4	0
LNCAP	CR615453,AK311167	FRG1B	Interchromosomal	6	0
LNCAP	AK311167	FRG1B	Interchromosomal	6	0
LNCAP	NBPF3	NBPF1	Intrachromosomal_Complex	5	0
LNCAP	RP3-365I19.1-001	SRGAP2	Intrachromosomal	8	0
LNCAP	RPL32	ANKRD42	Interchromosomal	5	0
LNCAP	LOC643648	BAGE1	Interchromosomal	8	0
LNCAP	ZNF83	ZNF765	Intrachromosomal_Complex	6	0
LNCAP	CECR7	AK302545	Intrachromosomal	16	0
LNCAP	NOC2L	LOC401010	Interchromosomal	5	0
LNCAP	PTMS	TVAS5	Interchromosomal	4	0
LNCAP	LOC728855	BC065231	Intrachromosomal_Complex	19	0
LNCAP	LOC728855	BC065231	Intrachromosomal	19	0
LNCAP	BC110832	BC065231	Intrachromosomal_Complex	19	0
LNCAP	CR622584	LOC442028	Interchromosomal	28	0
LNCAP	Z49985	TBL1X	Interchromosomal	3	0
LNCAP	BEX1	BEX2	Intrachromosomal	3	0
LNCAP	TRADD	B3GNT9	Read_Through	3	0
LNCAP	YAF2	RYBP	Interchromosomal	3	0
LNCAP	ZNF264	AURKC	Read_Through	3	0
LNCAP	AX747640	SLC9A7	Interchromosomal	3	0
LNCAP	INCA1	CAMTA2	Read_Through	3	0
LNCAP	AF113016	STRF6	Interchromosomal	3	0
LNCAP	ARNT	CTSK	Read_Through	3	0
LNCAP	HIST1H2BO	OR2B6	Read_Through	3	0
LNCAP	PRKAA1	PPIL1	Interchromosomal	3	0
LNCAP	ITPKC	PPFIA3,KIAA0654	Intrachromosomal	3	0
LNCAP	KIAA1049,TCF25	BC160930	Read_Through	3	0
LNCAP	PIK3CD	TNFRSF8	Intrachromosomal	3	0
LNCAP	CXorf40B	CXorf40A	Intrachromosomal_Complex	3	0
LNCAP	ZMYM2	ZMYM5	Intrachromosomal_Complex	3	0
LNCAP	FAM119B	TSFM	Read_Through	3	0
LNCAP	LOC388789	DTD1	Read_Through	3	0
LNCAP	CATSPER2	BC052612,CATSPER2P1	Intrachromosomal	3	0
LNCAP	IFRD1	C7orf53	Read_Through	3	0

MCF7	BCAS4	BCAS3	Interchromosomal	361	72
MCF7	BC017255	TDC1,TMEM49,DM119428	Intrachromosomal	66	25
MCF7	ARFGEF2	RP5-1049G16.1,SULF2	Intrachromosomal_Complex	173	15
MCF7	CSNK1E	ZNF217	Interchromosomal	18	17
MCF7	STK11	MIDN	Intrachromosomal	31	7
MCF7	PAPOLA	AK7	Read_Through	19	4
MCF7	DEPDC1B	ELOVL7	Read_Through	18	5
MCF7	AHCYL1	RAD51C	Interchromosomal	17	4
MCF7	AK297683	AK297683	Read_Through	14	4
MCF7	CXorf15	SYAP1,DKFZp686K221	Read_Through	14	5
MCF7	RRM2	C2orf48	Read_Through	12	4
MCF7	SMARCA4	CARM1	Intrachromosomal	11	4
MCF7	MFSD1,smap-4	GFM1,EFG	Intrachromosomal	8	4
MCF7	OK/SW-cl.16	TPT1,FLJ44635	Interchromosomal	8	7
MCF7	STRF6	CCDC34	Interchromosomal	8	7
MCF7	OK/SW-cl.16	ACTR6	Interchromosomal	4	3
MCF7	FRG1	FAM122C	Interchromosomal	4	3
MCF7	FRG1B	FAM122C	Interchromosomal	4	3
MCF7	MYO9B	FCHO1	Intrachromosomal	16	2
MCF7	YTHDC1	GNAS	Interchromosomal	11	10
MCF7	ZCCHC7	BC067112	Read_Through	12	2
MCF7	BC041486	CR607999	Interchromosomal	6	5
MCF7	TRIM37	RNFT1	Intrachromosomal	4	2
MCF7	TBL1XR1	RGS17	Interchromosomal	4	2
MCF7	PILRB	STAG3	Intrachromosomal	4	2
MCF7	PARD6G	C18orf1	Intrachromosomal_Complex	3	2
MCF7	X64709	TVAS5	Interchromosomal	4	3
MCF7	X64709	OK/SW-cl.16	Read_Through	9	8
MCF7	KRT79	KRT8	Intrachromosomal	3	2
MCF7	XBP1	SCAND1	Interchromosomal	3	2
MCF7	X64709	OK/SW-cl.5,TPM3	Interchromosomal	9	8
MCF7	EP300	MRFAP1	Interchromosomal	19	1
MCF7	ADAMTS19	SLC27A6	Intrachromosomal	6	1
MCF7	RAB27A	DYX1C1,CCPG1,EKN1	Intrachromosomal	6	2
MCF7	POP1	MATN2	Intrachromosomal	6	1
MCF7	STRF6	DHX40	Interchromosomal	6	5
MCF7	ATXN7L3	FAM171A2	Intrachromosomal	4	1
MCF7	SULF2	PRICKLE2	Interchromosomal	3	1
MCF7	CLDN3	CLDN4	Adjacent_Diverging	3	1
MCF7	ZNF580	CCDC106	Intrachromosomal	3	1
MCF7	ARHGEF7,Nbla10314	C13orf16	Read_Through	3	1
MCF7	SF3B3	EEF1A2	Interchromosomal	3	2
MCF7	CALR	MMS19	Interchromosomal	3	2
MCF7	ADCY3	SLC25A3,OK/SW-cl.48	Interchromosomal	3	2
MCF7	MYC	MYC	Read_Through	3	1
MCF7	GRIK3	EIF1AX	Interchromosomal	3	2
MCF7	X64709	UFSP2	Interchromosomal	9	8
MCF7	STRF6	AB055772	Interchromosomal	3	2
MCF7	X64709	PCBP1	Interchromosomal	8	7
MCF7	RPS16	MAK10	Interchromosomal	2	1
MCF7	RPS9	KIF1C	Interchromosomal	2	1
MCF7	CANX	NUDT7	Interchromosomal	2	1
MCF7	ZNF609	RPL8	Interchromosomal	2	1
MCF7	GRIK3	COPB2	Interchromosomal	2	1
MCF7	CDR2L	EEF2	Interchromosomal	2	1
MCF7	HNRNPU	FBL	Interchromosomal	2	1
MCF7	mccb,MCCC2	SYNE2	Interchromosomal	2	1
MCF7	FOXM1	RNF40	Interchromosomal	2	1
MCF7	SFRS9	MBD3	Interchromosomal	2	1
MCF7	RNF40	AGR2	Interchromosomal	2	1
MCF7	PLEKHG5	RPS16	Interchromosomal	2	1
MCF7	SNX30	DNM2	Interchromosomal	2	1
MCF7	ZBTB41	STRF6	Interchromosomal	2	1
MCF7	PNPLA7	WDR85	Read_Through	2	1

MCF7	IGSF1	EIF4A3	Interchromosomal	2	1
MCF7	DKFZp781M17165,PPP2R4	cytochrome_b	Interchromosomal	2	1
MCF7	hSIPL1A,SHARPIN	ARPC5	Interchromosomal	2	1
MCF7	CSTF3	CAT	Intrachromosomal_Complex	2	1
MCF7	RPL37	PMPCA	Interchromosomal	2	1
MCF7	CSDE1,KIAA0885	SETD1B	Interchromosomal	2	1
MCF7	NQO1	PRNPIP,ERI3	Interchromosomal	2	1
MCF7	KLK10	SLC35A1	Interchromosomal	2	1
MCF7	CLTC	S100A14	Interchromosomal	2	1
MCF7	RPS3	DDX42	Interchromosomal	2	1
MCF7	PRMT2	EEF1A2	Interchromosomal	2	1
MCF7	EDF1	RPL8	Interchromosomal	2	1
MCF7	H2AFJ	H3F3B	Interchromosomal	2	1
MCF7	SLC39A6	PREX1	Interchromosomal	2	1
MCF7	NACC1	AK056267	Interchromosomal	2	1
MCF7	KRT18	EEF1A2	Interchromosomal	2	1
MCF7	FLJ00383,ATP6AP1	STRF6	Interchromosomal	2	1
MCF7	OK/SW-cl.16	CA12	Interchromosomal	2	1
MCF7	PHPT1	PRKCSH	Interchromosomal	2	1
MCF7	EGR3	MTND5	Interchromosomal	2	1
MCF7	UBA52	GNAS	Interchromosomal	2	1
MCF7	BC071809	WDR62,DKFZp686G1024	Read_Through	2	1
MCF7	RPS6KB1	TDC1,TMEM49,DM119428	Read_Through	50	0
MCF7	TVAS5	DQ597482	Interchromosomal	139	0
MCF7	FLJ00194,ARHGAP19	DRG1	Interchromosomal	25	0
MCF7	OK/SW-cl.16	CR615613	Interchromosomal	24	0
MCF7	RSBN1	AK123199	Adjacent_Diverging	14	0
MCF7	TANC2	CA4	Intrachromosomal	12	0
MCF7	TMSB4X	BC113076	Interchromosomal	22	0
MCF7	TMSL3	BC113076	Interchromosomal	22	0
MCF7	TPT1,FLJ44635	FLJ44635	Interchromosomal	14	0
MCF7	RPL31	RPL31P11	Interchromosomal	18	0
MCF7	GATAD2B	DKFZp434E2118,NUP210L	Intrachromosomal	9	0
MCF7	BTCC-1,CD9,5H9	TSPAN18	Interchromosomal	8	0
MCF7	HLA-A,HLA-H,HLA-F	hla-b,HLA-B	Intrachromosomal_Complex	8	0
MCF7	REV3L	RPL28	Interchromosomal	7	0
MCF7	LOC148189	AK094188	Read_Through	7	0
MCF7	EEF1A2	HSD11B2	Interchromosomal	12	0
MCF7	KIAA1049,TCF25	BC160930	Read_Through	6	0
MCF7	TRFP,MED20	CCND3	Read_Through	6	0
MCF7	SNX27	HNRNPA0	Interchromosomal	7	0
MCF7	LOC387647	RAB18	Intrachromosomal	7	0
MCF7	CR610404	PDIA3P	Read_Through	5	0
MCF7	NAV1	GPR37L1	Intrachromosomal	5	0
MCF7	UBC	UBB	Interchromosomal	12	0
MCF7	AK311578	G3BP2	Read_Through	5	0
MCF7	PTMS	ANP32B	Interchromosomal	5	0
MCF7	BC110832	BC065231	Intrachromosomal_Complex	18	0
MCF7	EEF1D	NAPRT1	Read_Through	7	0
MCF7	CDC10L	CR610292	Interchromosomal	10	0
MCF7	RPS18	BC039356	Interchromosomal	13	0
MCF7	SHANK2	SHANK1	Interchromosomal	5	0
MCF7	SHANK2	SHANK3	Interchromosomal	5	0
MCF7	VAX2	ATP6V1B1	Read_Through	4	0
MCF7	HSP90Bb	HSP90AB1,HSP90AB3P	Interchromosomal	4	0
MCF7	RYR1	C16orf35	Interchromosomal	4	0
MCF7	ABCA5	PPP4R1L	Interchromosomal	4	0
MCF7	BC036544	NCRNA00164	Interchromosomal	4	0
MCF7	MTG1	FLJ00268	Read_Through	4	0
MCF7	GGA2	ZFAND5	Interchromosomal	4	0
MCF7	HLA-B	HLA-A	Intrachromosomal_Complex	4	0
MCF7	TAF15	FUS/CHOP	Interchromosomal	4	0
MCF7	LOC389333	MGC29506,PACAP	Read_Through	4	0

MCF7	BC036909	GNAS	Interchromosomal	4	0
MCF7	X64709	X64709	Interchromosomal	4	0
MCF7	LOC728875,BC021732,LOC728855	BC065231	Intrachromosomal_Complex	17	0
MCF7	FLJ39739	BC065231	Intrachromosomal_Complex	17	0
MCF7	BC110832,LOC728855	BC065231	Intrachromosomal	17	0
MCF7	RPS16	ZNF90	Intrachromosomal_Complex	5	0
MCF7	NBPF1	NBPF3	Intrachromosomal_Complex	6	0
MCF7	HLA-B,hla-b	HLA-B	Intrachromosomal_Complex	7	0
MCF7	EP300	MRFAP1L1	Interchromosomal	5	0
MCF7	HLA-B	hla-b	Intrachromosomal_Complex	5	0
MCF7	EEF1G	EEF1G	Interchromosomal	4	0
MCF7	SARS	SNHG8	Interchromosomal	3	0
MCF7	SPATA21	SORT1	Intrachromosomal	3	0
MCF7	DQ786213	TEKT4	Interchromosomal	3	0
MCF7	PTMS	NCL	Interchromosomal	3	0
MCF7	RPL35	RAB7A	Interchromosomal	3	0
MCF7	BC150535	VMO1	Read_Through	3	0
MCF7	CYTH3	pp9943,CYTH2	Interchromosomal	3	0
MCF7	SMG5	PAQR6	Read_Through	3	0
MCF7	NUCKS1	ARCN1	Interchromosomal	3	0
MCF7	TIRAP	DCPS	Read_Through	3	0
MCF7	FAM119B	TSFM	Read_Through	3	0
MCF7	ALDOA	KRT18	Interchromosomal	3	0
MCF7	HSPA1B,HSPA1A	HSPA1B	Read_Through	3	0
MCF7	CSDA	OK/SW-cl.16	Interchromosomal	3	0

Appendix C: Comparison of ChimeraScan with other gene fusion discovery tools

Fusion	ChimeraScan (v0.4.0)*	ShortFuse (v0.1)**	DeFuse (v0.3.5)***	MapSplice (v1.15.2)****
TMPRSS2-ERG	1	1	1	1
ZDHHC7-ABCB9	1	1	1	0
RC3H2-RGS3	1	1	1	1
HJURP-EIF4E2	1	1	1	1
TIA1-DIRC2	1	1	1	0
PIK3C2A-TEAD1	1	0	0	1
USP10-ZDHHC7	1	1	1	0
SPOCK1-TBC1D9B	0	0	0	0
INPP4A-HJURP	1	1	0	1
LMAN2-AP3S1	1	0	0	1
Total positive controls confirmed	9	7	6	6
Total chimeras nominated	78	245	56	400

** ChimeraScan (http://code.google.com/p/chimerascan/) was run with the following parameters (a weighted score >= 3 or a weighted score >= 2 if there is also a spanning read confirming the fusion)*

*** ShortFuse (http://exon.ucsd.edu/ShortFuse) was run with all default parameters*

**** DeFuse was run with Bowtie 0.12.7, max_insert_size = 1000; discord_read_trim = 50, and default values for remaining parameters*

***** MapSplice (http://www.netlab.uky.edu/p/bioinfo/MapSplice) was run using the following parameters: paired_end = yes; segment_length = 25; junction_type = canonical; fusion_junction_type = canonical; full_running = yes; do_fusion = yes; do_cluster = yes. The fusion_remap_junction.unique.chr_seq.extracted.repeat_filtered was processed using a Perl script to determine the HUGO gene symbols overlapping fusion junctions. If the acceptor and donor had the same gene symbol it was removed from further analysis.*

Appendix D: Gene fusions nominated by ChimeraScan in a clinical sequencing study

Patient 1: Xenograft from 67-year old male with prostate cancer

5' transcript	3' transcript	Fusion Genes	Type	Distance	Total Supporting Reads	Breakpoint Spanning Reads
uc003eom.2	uc003enx.2	CPNE4|NEK11	Intrachromosomal_Complex	-359604	754	186
uc003ehn.3	uc003eey.2	TMPRSS2|ERG	Intrachromosomal_Complex	-2347138	120	115
uc003jrh.3	uc001xdz.1	GPBP1|DAAM1	Interchromosomal	NA	78	27
uc002yzj.2	uc002yxa.2	UMPS|CCDC58	Intrachromosomal	-2966050	57	25
uc002ozv.2	uc002pnc.2	PVRL2|CD37	Intrachromosomal	4456230	43	3
uc003vrk.2	uc002hqd.2	EXOC4|SNIP	Interchromosomal	NA	30	21
uc002ilr.3	uc002jdz.2	NPEPPS|ERN1	Intrachromosomal_Complex	16419748	13	9
uc001xiw.2	uc010vwd.1	GPHN|PIGL	Interchromosomal	NA	9	4

Patient 2: Xenograft from 60-year old male with prostate cancer

5' transcript	3' transcript	Fusion Genes	Type	Distance	Total Supporting Reads	Breakpoint Spanning Reads
uc002gij.2	uc001wqm.1	TP53|SCFD1	Interchromosomal	NA	1622	530
uc010gor.2	uc002yxc.3	TMPRSS2|ERG	Intrachromosomal	-2802774	1466	414
uc001vqx.2	uc003tfu.3	COL4A2|TARP	Interchromosomal	NA	496	324
uc002vnj.2	uc002gmp.3	ACSL3|MYH2	Interchromosomal	NA	285	169
uc001wqm.1	uc010cnk.1	SCFD1|TP53	Interchromosomal	NA	130	82
uc003frq.1	uc003csj.1	BCL6|CAMP	Intrachromosomal_Complex	-139172189	52	0
uc002zdv.2	uc002yzj.2	AGPAT3|TMPRSS2	Intrachromosomal_Complex	-2405030	48	5
uc002gkg.3	uc002vni.2	TMEM107|ACSL3	Interchromosomal	NA	34	13
uc003qmj.2	uc003qmq.1	MAP3K7IP2|C6orf72	Read_Through	154779	27	22
uc009zrj.2	uc001sxq.1	FRS2|GLIPR1L2	Intrachromosomal	5811327	24	4
uc002vtg.2	uc002vvs.2	GIGYF2|AGAP1	Intrachromosomal	2718040	22	7
uc002lpp.2	uc001ztj.1	PTBP1|FRMD5	Interchromosomal	NA	16	3
uc001yva.2	uc002mee.1	NIPA2|ACSBG2	Interchromosomal	NA	10	1

Patient 3: 46-year old man with metastatic colon cancer

5' transcript	3' transcript	Fusion Genes	Type	Distance	Total Supporting Reads	Breakpoint Spanning Reads
uc004cpv.2	uc004cpf.2	PP1164|PPP2R3B	Intrachromosomal_Complex	-1216833	145	97
uc003xmr.2	uc003xjw.2	ADAM9|KCNU1	Intrachromosomal	-2063954	11	0

Patient 4: 48-year old woman with metastatic melanoma

5' transcript	3' transcript	Fusion Genes	Type	Distance	Total Supporting Reads	Breakpoint Spanning Reads
uc002jhd.3	uc002rww.2	WIPI1|FSHR*	Interchromosomal	NA	386	232
uc002rww.2	uc001csg.2	FSHR|CDKN2C*	Interchromosomal	NA	55	10
uc003xpe.2	uc002hku.2	SLC20A2|CCL18	Interchromosomal	NA	16	5
uc002jxs.2	uc002rut.2	EIF4A3|PRKCE	Interchromosomal	NA	8	8

*Complex rearrangement involving WIPI1, FSHR, and CDKN2C

Appendix E: Prostate Cancer Associated Transcripts (PCATs)

PCAT ID	Gene	Chromosomal Location	Expected score (dExp)	Observed score(d)	Fold change (PCA vs Benign	q-valu (%)
PCAT-1	TU_0099865_0	chr8:128087842-128095202	-2.2654014	5.444088	6.9071784	0
PCAT-2	TU_0090142_0	chr11:4748677-4760303	-2.4400573	4.6781354	11.39658	0
PCAT-3	TU_0054603_0	chr16:82380933-82394836	-2.1786723	4.4612455	5.8916535	0
PCAT-4	TU_0090140_0	chr11:4748163-4759145	-2.1153426	4.4345	7.1999164	0
PCAT-5	TU_0078288_0	chr12:32393283-32405731	-1.9164219	4.312603	3.5655262	0
PCAT-6	TU_0099864_0	chr8:128094589-128103681	-1.7214081	4.265536	3.8997242	0
PCAT-7	TU_0084308_0	chr5:15938753-15949124	-1.9636476	4.124071	4.747601	0
PCAT-8	TU_0084303_0	chr5:15899476-15955226	-2.0245786	4.0520086	7.1035967	0
PCAT-9	TU_0082746_0	chr12:120197102-120197416	-1.861408	3.7551165	5.1431665	0
PCAT-10	TU_0078296_0	chr12:32394534-32405549	-1.5944241	3.6902914	3.084959	0
PCAT-11	TU_0078290_0	chr12:32394534-32410898	-1.5337954	3.675318	3.1572607	0
PCAT-12	TU_0002597_0	chr6:34335202-34338521	-1.6263148	3.6469774	3.352418	0
PCAT-13	TU_0049368_0	chr4:106772318-106772770	-1.6894134	3.6079373	2.8299346	0
PCAT-14	TU_0106548_0	chr22:22209111-22212055	-1.939075	3.591358	5.962547	0
PCAT-15	TU_0078293_0	chr12:32396393-32414822	-1.5212961	3.5705945	2.9219174	0
PCAT-16	TU_0099884_0	chr8:128301493-128307576	-1.4445064	3.5658643	2.516981	0
PCAT-17	TU_0112014_0	chr15:67722165-67739990	-1.6326295	3.562463	3.6594224	0
PCAT-18	TU_0084306_0	chr5:15896315-15947088	-1.845	3.5603588	5.746707	0
PCAT-19	TU_0114240_0	chr2:1534883-1538193	-1.6970209	3.5233572	4.339947	0
PCAT-20	TU_0008499_0	chr7:24236191-24236455	-1.8302058	3.5071697	6.6821446	0
PCAT-21	TU_0078299_0	chr12:32290896-32292169	-1.7297353	3.506232	3.2923684	0
PCAT-22	TU_0000033_0	chr6:1619606-1668581	-1.7680657	3.494188	2.2470818	0
PCAT-23	TU_0096472_0	chr11:133844590-133862924	-1.8782617	3.410355	5.9854193	0
PCAT-24	TU_0114259_0	chr2:1606782-1607314	-1.6662377	3.3919659	5.060926	0
PCAT-25	TU_0096473_0	chr11:133844590-133862995	-1.8963361	3.3859823	6.1071715	0
PCAT-26	TU_0100361_0	chr8:144914456-144930753	-1.6521469	3.3805158	3.8420231	0
PCAT-27	TU_0040394_0	chr3:133418632-133441282	-1.6208398	3.3201025	2.9724674	0
PCAT-28	TU_0043432_0	chr13:34032994-34050503	-1.6739471	3.2037551	3.2093527	0
PCAT-29	TU_0112020_0	chr15:67764259-67801825	-1.5603316	3.1937351	3.593551	0
PCAT-30	TU_0042717_0	chr13:23149908-23200198	-2.0654948	3.1685438	4.9699407	0
PCAT-31	TU_0078292_0	chr12:32290485-32406307	-1.4503003	3.151379	2.8911364	0
PCAT-32	TU_0084146_0	chr5:14025126-14062770	-1.6452767	3.1257985	2.6190455	0
PCAT-33	TU_0056168_0	chr18:22477042-22477666	-1.5381516	3.0557241	3.1951044	0
PCAT-34	TU_0040383_0	chr3:133360541-133429262	-1.5558791	3.0416508	3.7478442	0
PCAT-35	TU_0112025_0	chr15:67780574-67782345	-1.6815377	3.0412362	3.433415	0

PCAT-36	TU_0041688_0	chr3:186741299-186741933	-1.4749297	3.0062308	2.543468	0
PCAT-37	TU_0103642_0	chr9:109187089-109187455	-1.7387192	2.998956	6.6124363	0
PCAT-38	TU_0040375_0	chr3:133280694-133394609	-1.5469999	2.9753568	3.9068055	0
PCAT-39	TU_0047312_0	chr4:39217669-39222163	-1.6388936	2.9124916	3.6121209	0
PCAT-40	TU_0106545_0	chr22:22218478-22219162	-1.7586497	2.889856	3.7357745	0
PCAT-41	TU_0054541_0	chr16:79408800-79435066	-1.7485934	2.8699164	6.647557	0
PCAT-42	TU_0060446_0	chr1:28438629-28450156	-1.4880521	2.857332	1.9824111	0
PCAT-43	TU_0072907_0	chr20:55759486-55771563	-1.5254781	2.7966201	2.812179	0
PCAT-44	TU_0043403_0	chr13:33844637-33845921	-1.5793877	2.7919009	3.6403422	0
PCAT-45	TU_0038678_0	chr3:53515951-53517078	-1.7047809	2.7858517	3.6908987	0
PCAT-46	TU_0101706_0	chr9:3408690-3415374	-1.4780945	2.7822099	3.3066912	0
PCAT-47	TU_0101709_0	chr9:3411967-3415374	-1.4652373	2.7622206	3.1886175	0
PCAT-48	TU_0106544_0	chr22:22210421-22220506	-1.6153399	2.7578135	3.7418716	0
PCAT-49	TU_0046121_0	chr4:766363-766599	-1.5697786	2.7573307	1.435532	0
PCAT-50	TU_0106542_0	chr22:22211315-22220506	-1.6098742	2.755721	3.3781004	0
PCAT-51	TU_0106541_0	chr22:22209111-22219162	-1.6593723	2.7341027	3.664146	0
PCAT-52	TU_0044453_0	chr13:51505777-51524522	-1.3416	2.732019	2.536953	0
PCAT-53	TU_0104717_0	chr9:130697833-130698832	-1.2938	2.7219732	2.3344588	0
PCAT-54	TU_0089014_0	chr5:176014905-176015351	-1.3967873	2.7047238	1.7803582	0
PCAT-55	TU_0108452_0	chr15:19344745-19362916	-1.5839852	2.6759455	1.8484153	0
PCAT-56	TU_0112003_0	chr15:67645590-67775246	-1.4386703	2.668052	3.045022	0
PCAT-57	TU_0078286_0	chr12:32395588-32405731	-1.3580605	2.6660874	2.6121044	0
PCAT-58	TU_0078303_0	chr12:32274210-32274530	-1.5020599	2.65866	3.3306372	0
PCAT-59	TU_0112004_0	chr15:67644390-67650387	-1.5175762	2.6509888	2.9933636	0
PCAT-60	TU_0071087_0	chr20:21428679-21429454	-1.4916688	2.649109	4.6481714	0
PCAT-61	TU_0072906_0	chr20:55759768-55770657	-1.5059631	2.645004	2.95756	0
PCAT-62	TU_0054240_0	chr16:70155175-70173873	-1.4715649	2.6437716	3.5309577	0
PCAT-63	TU_0047330_0	chr4:39217641-39222163	-1.5139307	2.6277235	3.0695639	0
PCAT-64	TU_0055435_0	chr18:6718938-6719172	-1.6048826	2.6173768	2.9221427	0
PCAT-65	TU_0079791_0	chr12:54971063-54971481	-1.4415668	2.6010823	2.0141602	0
PCAT-66	TU_0043411_0	chr13:33918267-33926769	-1.495064	2.5991623	3.3860362	0
PCAT-67	TU_0056121_0	chr18:20196762-20197522	-1.2526748	2.5938754	1.7191441	0
PCAT-68	TU_0043412_0	chr13:33918267-33935946	-1.5891836	2.590199	4.2804046	0
PCAT-69	TU_0065837_0	chr1:149791252-149795934	-1.3852053	2.5882297	2.9343975	0
PCAT-70	TU_0043401_0	chr13:33825711-33845275	-1.5994886	2.5853698	4.3461533	0
PCAT-71	TU_0006463_0	chr6:144659819-144660143	-1.4985942	2.5744107	2.2007995	0
PCAT-72	TU_0048506_0	chr4:80329017-80348259	-1.5744382	2.5690413	2.8022916	0
PCAT-73	TU_0084140_0	chr5:14003669-14054874	-1.4040573	2.5472755	2.5979335	0
PCAT-74	TU_0082982_0	chr12:121776584-121777370	-1.5293782	2.5458217	2.6197503	0
PCAT-75	TU_0013212_0	chr7:138990883-139001515	-1.2296493	2.544434	1.6879753	0
PCAT-76	TU_0072912_0	chr20:55779532-55780817	-1.4302964	2.5406737	3.8653345	0
PCAT-77	TU_0112281_0	chr15:70586704-70590792	-1.4590155	2.5375097	2.4288568	0

PCAT-78	TU_0048767_0	chr4:88120066-88124880	-1.3735119	2.5323946	2.233308	0
PCAT-79	TU_0108455_0	chr15:19358326-19365341	-1.5651321	2.5261333	1.9462687	0
PCAT-80	TU_0091997_0	chr11:58560356-58573012	-1.3149309	2.5185204	2.1176686	0
PCAT-81	TU_0121655_0	chr2:202985284-202998634	-1.4014161	2.476237	2.2194188	0.859614
PCAT-82	TU_0071798_0	chr20:33775260-33778511	-1.3356665	2.4645917	1.6566333	0.850371
PCAT-83	TU_0049200_0	chr4:102469973-102476087	-1.3222212	2.456723	1.9456172	0.841324
PCAT-84	TU_0121714_0	chr2:203295212-203314868	-1.3457565	2.4496663	1.7624274	0.832468
PCAT-85	TU_0098937_0	chr8:95748751-95751321	-1.4532137	2.42248	2.2326834	0.823797
PCAT-86	TU_0108453_0	chr15:19356996-19364013	-1.8033699	2.4094539	3.839975	0.767811
PCAT-87	TU_0114170_0	chr15:99659312-99669199	-1.4358851	2.4062114	2.1252658	0.767811
PCAT-88	TU_0089906_0	chr11:1042845-1045705	-1.3899238	2.401665	2.6390955	0.767811
PCAT-89	TU_0001559_0	chr6:30283700-30286011	-1.3517065	2.3987799	1.5110766	0.767811
PCAT-90	TU_0050557_0	chr4:159976338-160016453	-1.17525	2.398688	2.0524442	0.767811
PCAT-91	TU_0078294_0	chr12:32395632-32413064	-1.4560982	2.3969867	2.1863208	0.767811
PCAT-92	TU_0044933_0	chr13:94755992-94760688	-1.2905197	2.3965187	2.189938	0.767811
PCAT-93	TU_0017730_0	chr17:52346638-52346880	-1.4169512	2.3874657	1.4708191	0.760428
PCAT-94	TU_0039020_0	chr3:66578329-66607777	-1.2662895	2.3720088	1.7112709	0.712473
PCAT-95	TU_0049213_0	chr4:102461960-102476087	-1.2725139	2.3671806	1.8876821	0.712473
PCAT-96	TU_0093070_0	chr11:64945809-64961189	-1.2954472	2.3645105	1.9128969	0.712473
PCAT-97	TU_0051063_0	chr4:187244297-187244767	1.8922831	-2.8485844	0.50983155	0.732264
PCAT-98	TU_0098190_0	chr8:61704765-61708199	1.9825526	-2.8612607	0.4027831	0.732264
PCAT-99	TU_0038811_0	chr3:57890130-57890834	1.9620296	-2.8837616	0.44431657	0.732264
PCAT-100	TU_0020914_0	chr19:9718612-9721799	1.6433232	-2.9243097	0.50623006	0.732264
PCAT-101	TU_0112056_0	chr15:69658838-69678469	1.837821	-3.0355222	0.46161976	0
PCAT-102	TU_0036396_0	chr14:104617328-104619095	1.849786	-3.1192882	0.45514825	0
PCAT-103	TU_0095765_0	chr11:117640504-117642734	2.1002219	-3.2632742	0.38160667	0
PCAT-104	TU_0050224_0	chr4:147115887-147190781	2.1981242	-3.2975357	0.28569755	0
PCAT-105	TU_0112059_0	chr15:69667695-69691724	1.8148681	-3.3816626	0.43667468	0
PCAT-106	TU_0098382_0	chr8:68494189-68495887	2.5413978	-4.0586042	0.30793378	0

PCAT ID	Gene	Chromosomal Location	Outlier Score	Median Expression (RPKM)	Maximum Expression (RPKM)
PCAT-107	TU_0029004_0	chrX:66691350-66692032	130.7349145	1	90.921
PCAT-108	TU_0054542_0	chr16:79420131-79423590	127.0430957	5.60998	135.85
PCAT-109	TU_0120899_0	chr2:180689090-180696402	123.5416436	1.0525222	94.6932
PCAT-110	TU_0054540_0	chr16:79419351-79423673	119.090847	4.161985	94.4461
PCAT-111	TU_0120918_0	chr2:181297540-181400892	112.710111	1.4533705	92.1795
PCAT-112	TU_0054538_0	chr16:79408946-79450819	98.01851659	1.830343	93.1207
PCAT-113	TU_0059541_0	chr1:20685471-20686432	68.3572507	1.783109	1375.15
PCAT-114	TU_0120924_0	chr2:181331111-181427485	63.95455962	1.3891845	365.202
PCAT-115	TU_0074308_0	chr10:42652247-42653596	60.91841567	1.393607	65.7712
PCAT-116	TU_0049192_0	chr4:102257900-102306678	59.24997694	1.3854525	69.2423
PCAT-117	TU_0054537_0	chr16:79406933-79430041	58.04481977	1.8534395	42.751
PCAT-118	TU_0120900_0	chr2:180926864-180985967	55.8438747	1	67.6582
PCAT-119	TU_0114527_0	chr2:10858318-10858530	54.76455104	1.2969775	35.0059
PCAT-120	TU_0120923_0	chr2:181328093-181419226	52.9793227	1.2821	232.556
PCAT-121	TU_0049231_0	chr4:102257900-102259695	52.77001947	1.34042	67.6276

References

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
2. Mendes Soares, L.M. & Valcarcel, J. The expanding transcriptome: the genome as the 'Book of Sand'. *EMBO J* **25**, 923-931 (2006).
3. Lahav, N. The RNA-world and co-evolution hypotheses and the origin of life: implications, research strategies and perspectives. *Orig Life Evol Biosph* **23**, 329-344 (1993).
4. Xu, W. et al. Human transcriptome array for high-throughput clinical studies. *Proc Natl Acad Sci U S A* **108**, 3707-3712 (2011).
5. 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).
6. Kapranov, P. et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488 (2007).
7. Jacquier, A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**, 833-844 (2009).
8. Adams, M.D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991).
9. Boguski, M.S., Tolstoshev, C.M. & Bassett, D.E., Jr. Gene discovery in dbEST. *Science* **265**, 1993-1994 (1994).
10. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).
11. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* **100**, 15776-15781 (2003).
12. Taylor, M.S. et al. Rapidly evolving human promoter regions. *Nat Genet* **40**, 1262-1263; author reply 1263-1264 (2008).
13. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
14. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).
15. Sanger, F. & Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448 (1975).
16. Rothberg, J.M. & Leamon, J.H. The development and impact of 454 sequencing. *Nat Biotechnol* **26**, 1117-1124 (2008).
17. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).

18. Mardis, E.R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387-402 (2008).
19. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145 (2008).
20. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
21. Wilhelm, B.T. et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).
22. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).
23. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).
24. Maher, C.A. et al. Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* **106**, 12353-12358 (2009).
25. Levin, J.Z. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-715 (2010).
26. 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).
27. 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).
28. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476 (2008).
29. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).
30. Maher, C.A. et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101 (2009).
31. Pennisi, E. Human genome 10th anniversary. Will computers crash genomics? *Science* **331**, 666-668 (2011).
32. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**, S22-32 (2009).
33. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
34. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res* **22**, 2008-2017 (2012).
35. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
36. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493-500 (2010).
37. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**, R22 (2011).

38. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009-1015 (2010).
39. Hu, Y. et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res* **41**, e39 (2013).
40. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31**, 46-53 (2013).
41. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).
42. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
43. Koboldt, D.C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285 (2009).
44. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
45. Kostic, A.D. et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**, 393-396 (2011).
46. Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011).
47. 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).
48. Haas, B.J. & Zody, M.C. Advancing RNA-Seq analysis. *Nat Biotechnol* **28**, 421-423 (2010).
49. Clarke, K., Yang, Y., Marsh, R., Xie, L. & Zhang, K.K. Comparative analysis of de novo transcriptome assembly. *Sci China Life Sci* **56**, 156-162 (2013).
50. Robertson, G. et al. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-912 (2010).
51. Wilhelm, B.T., Marguerat, S., Goodhead, I. & Bahler, J. Defining transcribed regions using RNA-seq. *Nat Protoc* **5**, 255-266 (2010).
52. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178 (2010).
53. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* **38**, 4570-4578 (2010).
54. Huang, S. et al. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Front Genet* **2**, 46 (2011).
55. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
56. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
57. Li, Y. et al. TrueSight: a new algorithm for splice junction detection using RNA-seq. *Nucleic Acids Res* **41**, e51 (2013).
58. Lindner, R. & Friedel, C.C. A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS One* **7**, e52403 (2012).

59. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873-881 (2010).

60. Grant, G.R. et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* **27**, 2518-2528 (2011).

61. Garber, M., Grabherr, M.G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**, 469-477 (2011).

62. Li, J.J., Jiang, C.R., Brown, J.B., Huang, H. & Bickel, P.J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc Natl Acad Sci U S A* **108**, 19867-19872 (2011).

63. Li, W., Feng, J. & Jiang, T. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J Comput Biol* **18**, 1693-1707 (2011).

64. Mezlini, A.M. et al. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res* **23**, 519-529 (2013).

65. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**, 2325-2329 (2011).

66. Denoeud, F. et al. Annotating genomes with massive-scale RNA sequencing. *Genome Biol* **9**, R175 (2008).

67. 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).

68. Martin, J.A. & Wang, Z. Next-generation transcriptome assembly. *Nat Rev Genet* **12**, 671-682 (2011).

69. Martin, J. et al. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics* **11**, 663 (2010).

70. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

71. Haas, B.J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-5666 (2003).

72. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7 Suppl 1**, S12 11-14 (2006).

73. Loveland, J. VEGA, the genome browser with a difference. *Brief Bioinform* **6**, 189-193 (2005).

74. Harrow, J. et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9 (2006).

75. Flicek, P. et al. Ensembl 2013. *Nucleic Acids Res* **41**, D48-55 (2013).

76. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**, D130-135 (2012).

77. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).

78. Cabili, M.N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927 (2011).
79. Guttman, M. et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227 (2009).
80. Khalil, A.M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-11672 (2009).
81. Kalyana-Sundaram, S. et al. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**, 1622-1634 (2012).
82. Collins, J.E., White, S., Searle, S.M. & Stemple, D.L. Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res* **22**, 2067-2078 (2012).
83. Pauli, A. et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**, 577-591 (2012).
84. Graveley, B.R. et al. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473-479 (2011).
85. Daines, B. et al. The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Res* **21**, 315-324 (2011).
86. Gerstein, M.B. et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787 (2010).
87. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774 (2012).
88. 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).
89. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J Clin* **63**, 11-30 (2013).
90. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719-724 (2009).
91. Chapman, P.B. et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* **364**, 2507-2516 (2011).
92. Cohen, Y. et al. BRAF mutation in papillary thyroid carcinoma. *J Natl Cancer Inst* **95**, 625-627 (2003).
93. Hancock, L. The inhibition of anaplastic lymphoma kinase in non-small cell lung tumours with the ALK rearrangement may result in tumour shrinkage. *Thorax* (2011).
94. Antoniu, S.A. Crizotinib for EML4-ALK positive lung adenocarcinoma: a hope for the advanced disease? Evaluation of Kwak EL, Bang YJ, Camidge DR, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010;363(18):1693-703. *Expert Opin Ther Targets* **15**, 351-353 (2011).
95. Kwak, E.L. et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med* **363**, 1693-1703 (2010).
96. Tomlins, S.A. Urine PCA3 and TMPRSS2:ERG Using Cancer-specific Markers to Detect Cancer. *Eur Urol* (2012).
97. Weisberg, E., Manley, P.W., Cowan-Jacob, S.W., Hochhaus, A. & Griffin, J.D. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. *Nat Rev Cancer* **7**, 345-356 (2007).

98. Prensner, J.R. & Chinnaiyan, A.M. Oncogenic gene fusions in epithelial carcinomas. *Curr Opin Genet Dev* **19**, 82-91 (2009).
99. Seshagiri, S. et al. Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660-664 (2012).
100. Steidl, C. et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* **471**, 377-381 (2011).
101. Palanisamy, N. et al. Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. *Nat Med* **16**, 793-798 (2010).
102. 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).
103. Robinson, D.R. et al. Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing. *Nat Genet* **45**, 180-185 (2013).
104. Flockhart, R.J. et al. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res* **22**, 1006-1014 (2012).
105. Berger, M.F. et al. Integrative analysis of the melanoma transcriptome. *Genome Res* **20**, 413-427 (2010).
106. Kalari, K.R. et al. 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. *Front Oncol* **2**, 12 (2012).
107. Liu, J. et al. Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome Res* **22**, 2315-2327 (2012).
108. Ren, S. et al. RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res* **22**, 806-821 (2012).
109. Shah, S.P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395-399 (2012).
110. Sinicropi, D. et al. Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue. *PLoS One* **7**, e40092 (2012).
111. Kaur, H. et al. RNA-Seq of human breast ductal carcinoma in situ models reveals aldehyde dehydrogenase isoform 5A1 as a novel potential target. *PLoS One* **7**, e50249 (2012).
112. Sajnani, M.R. et al. Identification of novel transcripts deregulated in buccal cancer by RNA-seq. *Gene* **507**, 152-158 (2012).
113. Huber-Keener, K.J. et al. Differential gene expression in tamoxifen-resistant breast cancer cells revealed by a new analytical model of RNA-Seq data. *PLoS One* **7**, e41333 (2012).
114. Ku, C.S., Loy, E.Y., Salim, A., Pawitan, Y. & Chia, K.S. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**, 403-415 (2010).
115. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245 (2007).
116. Nowell, P.C. The minute chromosome (Phl) in chronic granulocytic leukemia. *Blut* **8**, 65-66 (1962).

117. Shtivelman, E., Lifshitz, B., Gale, R.P. & Canaani, E. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. *Nature* **315**, 550-554 (1985).
118. Mitelman, F., Johansson, B. & Mertens, F. Catalog of chromosome aberrations in cancer, Edn. 5th. (Wiley-Liss, New York; 1994).
119. Speicher, M.R. & Carter, N.P. The new cytogenetics: blurring the boundaries with molecular biology. *Nat Rev Genet* **6**, 782-792 (2005).
120. Pinkel, D. & Albertson, D.G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11-17 (2005).
121. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-732 (2005).
122. 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).
123. Chen, K. et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics* **28**, 1923-1924 (2012).
124. Iyer, M.K., Chinnaiyan, A.M. & Maher, C.A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903-2904 (2011).
125. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7**, e1001138 (2011).
126. Benelli, M. et al. Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232-3239 (2012).
127. Francis, R.W. et al. FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PLoS One* **7**, e39987 (2012).
128. Li, Y., Chien, J., Smith, D.I. & Ma, J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27**, 1708-1710 (2011).
129. Ge, H. et al. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27**, 1922-1928 (2011).
130. Sboner, A. et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol* **11**, R104 (2010).
131. Asmann, Y.W. et al. A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res* **39**, e100 (2011).
132. Kinsella, M., Harismendy, O., Nakano, M., Frazer, K.A. & Bafna, V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics* **27**, 1068-1075 (2011).
133. Jia, W. et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol* **14**, R12 (2013).
134. Kim, D. & Salzberg, S.L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12**, R72 (2011).
135. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
136. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

137. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648 (2005).

138. Hampton, O.A. et al. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. *Genome Res* **19**, 167-177 (2009).

139. Volik, S. et al. Decoding the fine-scale structure of a breast cancer genome and transcriptome. *Genome Res* **16**, 394-404 (2006).

140. Ruan, Y. et al. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* **17**, 828-838 (2007).

141. Dovey, H.F. et al. Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. *J Neurochem* **76**, 173-181 (2001).

142. Stephens, P.J. et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010 (2009).

143. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685-696 (2010).

144. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-276 (2009).

145. Roychowdhury, S. et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* **3**, 111ra121 (2011).

146. Liang, W.S. et al. Genome-wide characterization of pancreatic adenocarcinoma patients using next generation sequencing. *PLoS One* **7**, e43192 (2012).

147. Plebani, R. et al. Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA. *Neoplasia* **14**, 1087-1096 (2012).

148. Lee, C.S. et al. Transcriptome sequencing in Sezary syndrome identifies Sezary cell and mycosis fungoides-associated lncRNAs and novel transcripts. *Blood* **120**, 3288-3297 (2012).

149. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

150. Wilbanks, E.G. & Facciotti, M.T. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One* **5**, e11471 (2010).

151. Gingeras, T.R. Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206-211 (2009).

152. McPherson, A. et al. Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics* **27**, 1481-1488 (2011).

153. McPherson, A. et al. nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. *Genome Res* **22**, 2250-2261 (2012).

154. Mercer, T.R. et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30**, 99-104 (2012).

155. Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1**, Unit1 4 (2012).

156. 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).

157. Ramskold, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5**, e1000598 (2009).

158. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263 (2009).

159. Heber, S., Alekseyev, M., Sze, S.H., Tang, H. & Pevzner, P.A. Splicing graphs and EST assembly problem. *Bioinformatics* **18 Suppl 1**, S181-188 (2002).

160. Barash, Y. et al. Deciphering the splicing code. *Nature* **465**, 53-59 (2010).

161. Oshlack, A. & Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4**, 14 (2009).

162. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).

163. Silverman, B.W. Density estimation for statistics and data analysis. (Chapman & Hall/CRC, Boca Raton; 1998).

164. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848 (2008).

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

166. 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).

167. Orom, U.A. et al. Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

168. 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).

169. Gupta, R.A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-1076 (2010).

170. Pasmant, E. et al. Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer Res* **67**, 3963-3969 (2007).

171. Yap, K.L. et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell* **38**, 662-674 (2010).

172. Tsai, M.C. et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689-693 (2010).

173. Kotake, Y. et al. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**, 1956-1962 (2011).

174. de Kok, J.B. et al. DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res* **62**, 2695-2698 (2002).

175. Li, J. et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943-1947 (1997).

176. Tomlins, S.A. et al. Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599 (2007).

177. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **41**, D8-D20 (2013).
178. Carninci, P. et al. The transcriptional landscape of the mammalian genome. *Science* **309**, 1559-1563 (2005).
179. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855-1857 (2008).
180. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
181. Garber, M. et al. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62 (2009).
182. Yu, J. et al. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443-454 (2010).
183. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
184. Blankenberg, D. et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**, Unit 19 10 11-21 (2010).
185. Day, D.S., Luquette, L.J., Park, P.J. & Kharchenko, P.V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biol* **11**, R69 (2010).
186. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).
187. Saeed, A.I. et al. TM4 microarray software suite. *Methods Enzymol* **411**, 134-193 (2006).
188. Rubin, M.A. et al. alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA* **287**, 1662-1670 (2002).
189. Dhanasekaran, S.M. et al. Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822-826 (2001).
190. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol* **8**, e1000371 (2010).
191. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).
192. Tomlins, S.A. et al. The role of SPINK1 in ETS rearrangement-negative prostate cancers. *Cancer Cell* **13**, 519-528 (2008).
193. Bjartell, A.S. et al. Association of cysteine-rich secretory protein 3 and beta-microseminoprotein with outcome after radical prostatectomy. *Clin Cancer Res* **13**, 4130-4138 (2007).
194. Hessels, D. et al. DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol* **44**, 8-15; discussion 15-16 (2003).
195. Laxman, B. et al. A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. *Cancer Res* **68**, 645-649 (2008).
196. Etzioni, R., Cha, R., Feuer, E.J. & Davidov, O. Asymptomatic incidence and duration of prostate cancer. *Am J Epidemiol* **148**, 775-785 (1998).
197. Cooperberg, M.R., Moul, J.W. & Carroll, P.R. The changing face of prostate cancer. *J Clin Oncol* **23**, 8146-8151 (2005).

198. Grasso, C.S. et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239-243 (2012).

199. Taylor, B.S. et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11-22 (2010).

200. Prensner, J.R., Rubin, M.A., Wei, J.T. & Chinnaiyan, A.M. Beyond PSA: the next generation of prostate cancer biomarkers. *Sci Transl Med* **4**, 127rv123 (2012).

201. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).

202. Prensner, J.R. & Chinnaiyan, A.M. The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391-407 (2011).

203. Guttman, M. et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300 (2011).

204. Lee, J.T. Lessons from X-chromosome inactivation: long ncRNA as guides and tethers to the epigenome. *Genes Dev* **23**, 1831-1842 (2009).

205. Zhao, J., Sun, B.K., Erwin, J.A., Song, J.J. & Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).

206. Ahmadiyeh, N. et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* **107**, 9742-9746 (2010).

207. Al Olama, A.A. et al. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat Genet* **41**, 1058-1060 (2009).

208. Oosumi, T., Belknap, W.R. & Garlick, B. Mariner transposons in humans. *Nature* **378**, 672 (1995).

209. Robertson, H.M., Zumpano, K.L., Lohe, A.R. & Hartl, D.L. Reconstructing the ancient mariners of humans. *Nat Genet* **12**, 360-361 (1996).

210. 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**, 11606-11611 (2003).

211. Varambally, S. et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629 (2002).

212. Huttenhower, C. et al. Exploring the human genome with functional maps. *Genome Res* **19**, 1093-1106 (2009).

213. Rhodes, D.R. et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166-180 (2007).

214. Rhodes, D.R. et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6 (2004).

215. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).

216. Setlur, S.R. et al. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl Cancer Inst* **100**, 815-825 (2008).

217. Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M. & Gerald, W.L. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* **113**, 913-923 (2004).

218. Nakagawa, T. et al. A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy. *PLoS One* **3**, e2318 (2008).

219. 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).

220. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).

221. Shen, H. et al. The SWI/SNF ATPase Brm is a gatekeeper of proliferative control in prostate cancer. *Cancer Res* **68**, 10154-10162 (2008).

222. Roberts, C.W. & Orkin, S.H. The SWI/SNF complex--chromatin and cancer. *Nat Rev Cancer* **4**, 133-142 (2004).

223. Wiegand, K.C. et al. ARID1A mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med* **363**, 1532-1543 (2010).

224. Jones, S. et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231 (2010).

225. Varela, I. et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539-542 (2011).

226. Versteege, I. et al. Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203-206 (1998).

227. Reisman, D., Glaros, S. & Thompson, E.A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653-1668 (2009).

228. Sun, A. et al. Aberrant expression of SWI/SNF catalytic subunits BRG1/BRM is associated with tumor development and increased invasiveness in prostate cancers. *Prostate* **67**, 203-213 (2007).

229. Dechassa, M.L. et al. Architecture of the SWI/SNF-nucleosome complex. *Mol Cell Biol* **28**, 6010-6021 (2008).

230. De, S. et al. Dynamic BRG1 recruitment during T helper differentiation and activation reveals distal regulatory elements. *Mol Cell Biol* **31**, 1512-1527 (2011).

231. Euskirchen, G.M. et al. Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* **7**, e1002008 (2011).

232. Yen, K., Vinayachandran, V., Batta, K., Koerber, R.T. & Pugh, B.F. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell* **149**, 1461-1473 (2012).

233. Shin, H., Liu, T., Manrai, A.K. & Liu, X.S. CEAS: cis-regulatory element annotation system. *Bioinformatics* **25**, 2605-2606 (2009).

234. Taherian, N. et al. Familial prostate cancer: the damage done and lessons learnt. *Nat Rev Urol* **10**, 116-122 (2013).

235. Wang, K.C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120-124 (2011).

236. Shain, A.H. et al. Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer. *Proc Natl Acad Sci U S A* **109**, E252-259 (2012).

237. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

238. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
239. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**, R86 (2010).
240. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40 (2011).
241. Liu, W. et al. Thousands of Novel Transcripts Identified in Mouse Cerebrum, Testis, and ES Cells Based on ribo-minus RNA Sequencing. *Front Genet* **2**, 93 (2011).
242. Cui, P. et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics* **96**, 259-265 (2010).
243. Ingolia, N.T., Lareau, L.F. & Weissman, J.S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789-802 (2011).
244. Kertesz, M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-107 (2010).
245. Zhao, J. et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939-953 (2010).
246. Li, M. et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**, 53-58 (2011).
247. Chen, L. et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nat Med* **19**, 209-216 (2013).
248. Gallo, A. RNA editing enters the limelight in cancer. *Nat Med* **19**, 130-131 (2013).
249. Bahn, J.H. et al. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res* **22**, 142-150 (2012).
250. Peng, Z. et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol* **30**, 253-260 (2012).
251. Degner, J.F. et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207-3212 (2009).
252. Skelly, D.A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J.M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21**, 1728-1737 (2011).
253. Rozowsky, J. et al. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522 (2011).