# DEVELOPMENT AND APPLICATION OF METHODS TO DISCOVER CANCER-ASSOCIATED TRANSCRIPT VARIANTS

by

Brendan Veeneman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2016

Doctoral Committee:

Professor Arul M. Chinnaiyan, Co-Chair
Associate Professor Alexey Nesvizhskii, Co-Chair
Assistant Professor Hui Jiang
Assistant Professor Ryan Edward Mills
Professor Gilbert S. Omenn
Assistant Professor Scott Arthur Tomlins

# ACKNOWLEDGEMENTS

In completion of this work, I'd like to thank my advisors, Alexey Nesvizhskii and Arul Chinnaiyan, without whose guidance none of this would have been possible and I probably would have gone off on some weird(er) tangents. I'd like to thank the rest of my committee as well, Hui Jiang, Scott Tomlins, Gil Omenn, and Ryan Mills, for their advice and enthusiasm, which helped keep me enthusiastic as well. Maybe most importantly, I want to thank my labmates who helped me and answered stupid questions I had on a daily basis, including Dan Robinson, Marcin Cieślik, and especially Mohan Dhanasekaran, whose tireless humoring and friendship really meant a lot to me. Next, I'd like to thank the people who make running the lab possible, particularly Xuhong Cao, Karen Giles, Christine Betts, and Terry Barrette. Finally, I'd like to thank my lab friends and compatriots in research, past and present, including Sudhanshu Shukla, Pankaj Vats, Rohit Malik, Bhavna Malik, Lisha Wang, Sunita Shankar, Chandan Kumar, George Wang, Yi-Mi Wu, Katayoon Kasaian, Josh Nguyen Vo, Bob Lonigro, Christine Brennan, Yashar Niknafs, Carl Engelke, Yajia Zhang, Lanbo Xiao, Yuanyuan Qiao, Steve Kregel, Ron Siebenaler, Abhijit Parolia, Yasu Hosono, Rohit Mehra, Anton Poliakov, Alejandro Balbin, Matthew Iyer, Lee Sam, Anirban Sahu, John Prensner, Sameek Roychowdhury, Chad Brenner, Ram Mani, Aparna Ghosh, Catie Grasso, Datta Mellacheruvu, Avinash Shanmugam, Chih-chiang Tsou, Andy Kong, and Felipe Leprevost, and anyone else I may have missed (or worse, misspelled)!

In completing my arc as a student, I'd like to thank the many teachers who have had such an immense influence on my life, in Southborough, at the Cambridge School of Weston, and at Oberlin College. In particular, I'd like to thank Mary Garvin, Keith

Tarvin, Yolanda Cruz, John Donaldson, Marilyn Del Donno, Michael Weinstein, Carl LaCombe, Robin Wood, and Mark Santa Maria. I'd like to thank my friends from Febit as well, including Harry Bochner, Lauren Hayes, and Anthony Caruso.

Finally, in reaching this point in my life, I'd like to thank my family and friends. My friends, Talor Walsh, Julia Howells, Gwen Anderson, Jane Sawyer, Ken Wang, Josh Koenig, Kevin McNalley, Tom Newhall, John Bohnert, EmilyKate McDonough, Janet Ackerman, James Webber, Andy Crump, and everyone else (!) for helping to keep life fun and interesting. My family, Liz Henry, Dale Veeneman, Conor Veeneman, and Kate Veeneman for always supporting me. And finally, an extra-special shout out to my life-long best friends, Dan Quirk and David Butler.

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# LIST OF APPENDICES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ACTN1 | Actinin Alpha 1 |
| ALK | Anaplastic Lymphoma Kinase |
| AMACR | Alpha Methylacyl CoA Racemase |
| AR | Androgen Receptor |
| ASCII | American Standard Code for Information Interchange |
| ASO | Anti-Sense Oligonucleotide |
| ATI | Alternative Transcript Initiation |
| ATSS | Alternative Transcription Start Site |
| BLAST | Basic Local Alignment Search Tool |
| BLAT | BLAST-Like Alignment Tool |
| BWA | Burroughs-Wheeler aligner |
| CCLE | Cancer Cell Line Encyclopedia |
| cDNA | Complementary DNA |
| ChIP-Seq | Chromatin Immunoprecipitation Sequencing |
| CNV | Copy Number Variant |
| CPU | Central Processing Unit |
| DDR3 | Double Data Rate type 3 |
| DNA | Deoxyribonucleic acid |
| ENCODE | Encyclopedia of DNA Elements |
| ERG | ETS-Related Gene |
| ESRP* | Epithelial Splicing Regulatory Protein (family) |
| EST | Expressed Sequence Tag |
| ETS | E26 Transformation-Specific |
| ETV1 | ETS Variant 1 |
| ETV4 | ETS Variant 4 |
| EZH2 | Enhance of Zeste Homolog 2 |
| FFPE | Formalin-Fixed Paraffin-Embedded |
| FLI1 | Friend Leukemia Integration 1 |
| FOXA1 | Forkhead Box Protein 1 |
| FPKM | Fragments Per Kilobase per Million mapped fragments |
| GB/MB | Gigabyte/Megabyte |
| GHz/MHz | Gigahertz/Megahertz |
| GNU | GNU's not Unix |
| GPL | GNU public license |

| | |
|---|---|
| GRCh38/hg38 | Genome Reference Consortium Human Build 38 |
| GSEA | Gene Set Enrichment Analysis |
| hg19 | Human Genome revision 19 |
| HISAT | Hierarchical Indexing for Spliced Alignment of Transcripts |
| I/O | Input/Output |
| IDH1 | Isocitrate Dehydrogenase 1 |
| IGV | Integrated Genomics Viewer |
| LASER | Light Amplification by Stimulated Emission of Radiation |
| MAQ | Mapping and Assembly with Quality |
| MATS | Multivariate Analysis of Transcript Splicing |
| MBNL* | Muscleblind-like (family) |
| mCRPC | Metastatic Castration-Resistant Prostate Cancer |
| MET | MET proto-oncogene, receptor tyrosine kinase |
| MiPS | Michigan Prostate Score |
| MISO | Mixture of Isoforms |
| MOV10 | Moloney Leukemia Virus 10 Protein |
| MPRIP | Myosin Phosphatase Rho Interacting Protein |
| MS-MS | Tandem Mass Spectrometry |
| MYLK | Myosin Light Chain Kinase |
| NE | Neuroendocrine |
| NKX3.1 | NK3-homeobox 1 |
| NOVA* | Neuro-Oncological Ventral Antigen 1 |
| NRG1 | Neuregulin 1 |
| OS | Operating System |
| PCA | Principal Components Analysis |
| PCa | Prostate Cancer |
| PCA3 | Prostate Cancer Associated 3 |
| PCR | Polymerase Chain Reaction |
| PDLIM5 | PDZ and LIM Domain 5 |
| PrEC | Prostate Epithelial Cells |
| PSA | Prostate-Specific Antigen |
| PTEN | Phosphatase and Tensin Homolog |
| QKI | Quaking |
| RB1 | Retinoblastoma 1 |
| RBFOX2 | RNA-binding protein Fox-1 homolog 2 |
| RBM* | RNA-Binding Motif protein 1 |
| RIN | RNA Integrity Number |
| RNA | Ribonucleic Acid |
| RNA-Seq | RNA Sequencing |
| SAGE | Serial Analysis of Gene Expression |

| | |
|---|---|
| SAM | Sequence Alignment Map |
| SChLAP1 | Second Chromosome Locus Associated with Prostate-1 |
| SF3B* | Splicing Factor 3B (family) |
| SF3B1 | Splicing Factor 3B subunit 1 |
| SNV | Single Nucleotide Variant |
| SPINK1 | Serine Protease Inhibitor Kazal-type 1 |
| SPOP | Speckle-type POZ protein |
| SRSF* | Serine/arginine-Rich Splicing Factor (family) |
| SRSF2 | Serine/arginine-Rich Splicing Factor 2 |
| STAR | Spliced Transcripts Alignment to a Reference |
| STL | Standard Library |
| SU2C | Stand Up To Cancer |
| TAIR10 | The Arabidopsis Information Resource 10 |
| TC | Tumor Content |
| TCGA | The Cancer Genome Atlas |
| TMPRSS2 | Transmembrane Protease, Serine 2 |
| TP53 | Tumor Protein 53 |
| TPM2 | Tropomyosin 2 (Beta) |
| U2AF1 | U2 Small Nuclear RNA Auxiliary Factor 1 |
| UHRR | Universal Human Reference RNA |
| UTR | Untranslated Region |
| VCL | Vinculin |
| ZRSR2 | Zinc Finger, RNA Binding Motif And Serine/Arginine Rich 2 |

# ABSTRACT

Cancer is and has long been a major threat to human health, and in seeking to better treat cancer, we seek first to better understand cancer. Consequently, the current era of cancer research has aimed to catalog the full range of molecular abnormalities in cancer's genome, epigenome, transcriptome, and proteome. Next-generation short read sequencing has empowered these cataloging efforts, but requires sophisticated algorithms to fully harness, particularly in the case of splicing and transcript variation. The aim of this dissertation was to address this need by establishing and applying novel methods to analyze RNA sequencing data in cancer.

In chapter one, we present Oculus, a software package that attaches to standard aligners and exploits read redundancy by performing streaming compression, alignment, and decompression of input sequences. This nearly lossless process (> 99.9%) led to alignment speedups of up to 270% across a variety of data sets.

In chapter two, we profile performance characteristics of two-pass alignment, which separates splice junction discovery from quantification. Across a variety of transcriptome sequencing datasets, two-pass alignment improved quantification of at least 94% of simulated novel splice junctions, and provided as much as 1.7-fold deeper median read depth over those splice junctions. Two-pass alignment promises to advance quantification and discovery of novel splicing events.

In chapter three, we present a novel bioinformatics pipeline to analyze splicing and transcript variation from cancer transcriptome data, using splice junction read depth, and correlative analysis to circumvent known biases such as tumor content. We demonstrate the value of this pipeline through application to the oncogenes MET and ALK.

Finally, in chapter four, we present the application of our transcript variant calling pipeline to transcriptome data from prostate cancer. We present several recurrently differentially spliced genes which are not attributable to noise or bias and may serve as novel biomarkers, evidence for transcript variants of the androgen receptor, and an apparent genome-wide pattern of alternative transcription start site usage.

# CHAPTER 1
# INTRODUCTION

## 1.1 Cancer

Cancer is a heterogeneous class of diseases which poses a significant threat to human health worldwide. In 2016, in the United States, there are projected to be approximately 1.6 million new cancer cases and six hundred thousand cancer deaths, from a population of about 320 million people.[1,2] To further underscore its significance, cancers are the second-leading cause of death in the United States, and the lifetime probability of being diagnosed with an invasive cancer is 42% for men and 38% for women.[1] As such, cancer has long been and remains an area of intense research interest.

Broadly defined, cancer is the uncontrolled growth of an organism's cells in its own body. The first universal truism of cancer is that cancers follow Darwinian selection, meaning that as cancer cells replicate, mutations arise in single cancer cells and confer selective advantages by which the cancer as a whole adapts over time.[3,4] Precisely, cancers meet the criteria necessary and sufficient for evolution to occur : variability between individuals, heritability of traits, and that traits confer selective advantages.[5] Second, virtually all cancers possess DNA mutations, differences between the genetic code inherited by the organism and shared by the cancer cells, though the burden of mutations varies greatly.[6] The prevailing belief is that each cancer possesses at least one mutation which confers a selective advantage, commonly referred to as a "driver mutation" in contrast to a "passenger mutation," though very rare exceptions involving selective non-DNA traits may exist.[6,7]

Beyond evolution and mutations, cancer begins to defy broad generalization. First, cancers arise from dozens of cell types and tissues in the human body.[8] As an example, The Cancer Genome Atlas, a massive multi-institutional attempt to survey the landscape of mutations in cancer, aggregated samples from at least thirty-three relatively common cancer types, and still excluded many rarer cancer types.[9] Further, even within single cancer types there exists substantial heterogeneity.[9] For instance, invasive breast carcinomas are commonly subtyped into five categories on the basis of their molecular characteristics, with substantial differences in clinical outcome.[10-12] The most enduring attempts to categorize cancers have usually focused on their functional characteristics as in the Hallmarks of Cancer, and on the set of driving mutations common to subtypes of a cancer, as in The Cancer Genome Atlas.[9,13,14] Cancers tend to sustain their own growth signaling, evade suppression of that growth, evade programmed death signals, achieve replicative immortality, recruit blood vessels, and ultimately spread throughout the body, but every cancer is different in its path to achieve these and other functions and not every cancer exhibits every function.[13] In summary, to best understand cancer, it is necessary to first understand that cancer is a heterogeneous class of diseases.

## 1.2 Prostate cancer

The prostate is a walnut-sized organ which contributes alkaline fluids to semen in men and is situated around the urethra.[15] Women do not have prostates in the identical sense, but instead possess a developmentally and functionally homologous organ called the Skene's gland, which develops cancer extremely rarely.[16-19] In contrast, all men develop at least benign prostate tumors as they age.[20-22] In 2016 in the United States there are projected to be about one hundred and eighty thousand new prostate cancer cases diagnosed, which leads amongst all cancer types in men, and about twenty-six thousand deaths, which comes second after lung cancer.[1] A different disease of the prostate, benign prostatic hyperplasia, is also common, but it is completely distinct from prostate cancer, and does not progress into prostate cancer.[15] Prostate cancer incidence and death vary by race, and while environmental links are currently not completely understood, some correlations have been established between prostate cancer and red meat and dairy consumption.[15,23,24]

Histologically, the prostate is comprised of epithelium, including secretory luminal epithelial cells, which produce the alkaline fluid contributed to semen and become cancer, basal epithelial cells situated beneath the luminal cells at a 1:1 ratio, and relatively rarer neuroendocrine epithelial cells; and stroma, including smooth muscle cells and fibroblasts, which provide the mechanical means to expel fluid from the prostate, and other relatively rarer stromal cell types.[25] Of these, the luminal epithelial cells and smooth muscle cells express the androgen receptor (AR), and only the luminal epithelial cells express the prostate specific antigen PSA, whose expression is driven by AR.[25] Extensive paracrine signaling exists between the epithelial cell types, and between the epithelial cells and stroma, which is interesting considering these cell types are displaced over the progression of prostate cancer.[25]

Prostate tissue and cancers are critically linked to the expression and activity of androgen and the androgen receptor, and the most effective therapy for treating progressing prostate cancer works by blocking androgen signaling. Androgen deprivation therapy was pioneered by Charles Brenton Huggins, who was awarded the Nobel prize in physiology or medicine for this work in 1966.[26] While the field previously believed that prostate cancers which progressed after androgen deprivation therapy were beyond or apart from androgen signaling ("androgen-independent"), it is now understood that these cancers have reactivated androgen signaling by circuitous means, and the disease was thus renamed "castration-resistant" prostate cancer to reflect this new understanding.[27,28]

A typical treatment course for prostate cancer is : 1) a patient presents with early symptoms of prostate cancer, including possibly an elevated level of the prostate specific antigen PSA in his blood ; 2) a needle biopsy may then be taken from the prostate and scored for histopathological aggressiveness, called Gleason grade ; 3) if the biopsy appears non-aggressive, a period may elapse wherein the physician and patient wait to monitor if symptoms progress, called active surveillance ; 4) if the biopsy appears aggressive or symptoms progress, then scans may be performed to test the extent of the cancer ; 5) treatment is performed, commonly either surgical removal of the prostate and nearby lymph nodes (to gauge aggressiveness), or localized radiation therapy, or androgen deprivation therapy in advanced cases or if symptoms progress after surgery or radiation therapy ; 6) if the patient's cancer is treated with androgen deprivation therapy and progresses, and they nearly always do progress after androgen deprivation therapy, usually having metastasized to bone, the liver, lymph nodes, or soft tissues, alternative, second-line anti-androgen deprivation therapy or chemotherapy may be applied, but these therapies usually prolong life only by several months.[15,27,29,30] When we study cancer tissue samples, they are taken at the fifth and sixth steps in this treatment course, and are referred to as hormone-naive primary prostate tumors, and metastatic castration-resistant prostate cancers.[30]

As is true with other cancers, prostate cancers are increasingly appreciated as a heterogeneous group of diseases, with multiple different molecular aberrations.[30-33]

Primary molecular aberrations include activating gene fusions of the E26 Transformation-Specific (ETS) family of transcription factors, particularly the TMPRSS2:ERG gene fusion, mutations in the Speckle-type POZ protein (SPOP), mutations in the Forkhead box protein A1 (FOXA1), though aberrations in these genes account for only 74% of prostate cancers, a variety of other rarer possible driver events, and a variety of inactivation variants affecting tumor suppressor proteins common to many cancers, including Tumor protein 53 (TP53), Retinoblastoma protein 1 (RB1), and the Phosphatase and tensin homolog (PTEN), and some apparently specific to prostate cancer such as NK3-homeobox 1 (NKX3.1).[33-40] Beyond these, prostate cancers have recurrent copy-number aberrations, particularly gain of chromosome 8q and loss of chromosome 8p and extreme amplification of the androgen receptor locus after androgen deprivation therapy, recurrent DNA methylation and gene expression changes, and recurrent overexpression of the enhancer of Zeste homolog 2 (EZH2) and serine protease inhibitor Kazal-type 1 (SPINK1), as well as activation of a variety of other developmentally significant signaling pathways.[33,41-44] Finally, prostate cancers are associated with the expression of a number of cancer-specific long non-coding RNA species, most notably the second chromosome locus associated with prostate-1 (SChLAP1).[45-48]

Clinically, PSA is the primary means to detect early prostate cancer, but high false positive rates have undercut the medical field's confidence in the marker.[49,50] PSA is exactly what its name advertises, a prostate-specific antigen, which means that while specific to the prostate, and capable of detecting amplified amounts of prostate signaling in the blood, it can perform poorly in distinguishing prostate cancer from other prostate tissue. Therefore, utmost translational importance is placed on the development of biomarkers which are capable of detecting prostate cancer early, and particularly, capable of prognosticating prostate cancers into those which are likely to metastasize in order to inform decisions to accelerate therapy.[33,36]

## 1.3 RNA splicing

The central dogma of molecular biology was first postulated by Francis Crick in the 1950s, and dictates that genetic sequence information cannot transfer from protein to nucleic acid, or from protein to protein.[51,52] Today, we rephrase this to state that the flow of sequence information in biology proceeds from deoxyribonucleic acid (DNA), to ribonucleic acid (RNA), to protein, with rare exceptions (reverse polymerase in retroviruses, and RNA-dependent RNA replication in RNA viruses). Put simply, DNA is transcribed into RNA, and RNA is translated into protein. In this dynamic, RNA is the messenger that conveys information between the genetic code of the cell to the functional processes the cell carries out.

The understanding of RNA was further advanced in 1977, when two independent research teams led by Phillip Sharp and Richard Roberts discovered the existence of "split genes," a discovery which merited the Nobel Prize in Physiology or Medicine in 1993.[53-55] Briefly, gaps within genes (introns) are removed from neighboring sequence (exons) in RNA by a catalytic molecular process later termed RNA splicing. Splicing is executed by two mechanisms in eukaryotes : catalytic excision by the spliceosome, a complex molecular machine involving many core proteins and context-dependent cofactors (termed "splicing factors"), and self-splicing introns which catalyze their own excision through secondary structural mechanisms. Since the discovery of splicing, we have come to appreciate that nearly all human genes are spliced (about 95%), and further, that genes are frequently spliced in multiple patterns, yielding multiple mature proteins per gene with varying functions.[56] The phenomenon of splicing genes in multiple ways is called alternative splicing, and has been suggested as a possible means of increasing the phenotypic complexity of eukaryotic gene expression.[57-59] Further, we also now appreciate that RNA splicing performs a critical role in regulation, through binding of the exon junction complex and nonsense-mediated decay.[60,61] In short, splicing is equally fundamental to molecular biology as transcription : both are necessary to perform the message conveying function of RNA.

## 1.4 RNA splicing in cancer

Owing to RNA splicing's fundamental role in molecular biology, and the fact that cancers will hijack any means to increase their proliferative potential, there are many examples of both driver and passenger splicing variants in cancer. Examples of cancer-driving splice variants include exon skipping and alternative transcript initiation of the anaplastic lymphoma kinase (ALK), exon skipping variants of the hepatocyte growth factor receptor (MET), truncation of the androgen receptor (AR), and many, many others.[62-66] Similarly, splicing factors themselves are often overexpressed or mutated in cancer, in order to generate driving splice variants or networks of splice variants downstream. These include mutations of splicing factor 3B subunit 1 (SF3B1), U2 small nuclear RNA auxiliary factor 1 (U2AF1), serine/arginine-rich splicing factor 2 (SRSF2), and zinc finger CCCH-type, RNA binding motif and serine/arginine-rich 2 (ZRSR2) in acute myeloid leukemias, and expression modulation of RNA-binding protein Fox-1 homolog 2 (RBFOX2) to drive epithelial to mesenchymal transition in solid tumors, again amongst many, many other examples (NOVA*, ESRP*, SRSF*, MBNL*, QKI, RBM*, SF3B*, and more, where * represents multiple gene family member numbers).[67-72] Next, RNA splicing has specific interest to clinical translation in the form of cancer-specific biomarkers, owing both to the apparent exquisitely-tissue-specific regulation of splicing in many cases, and RNA splicing's inclination to generate novel and specific cell-surface proteins which may be used as neoantigens in immunotherapy.[73-76] Finally, cutting-edge translational research is investigating the use of antisense oligonucleotides (ASOs) to knock-down expression of specific RNA molecules *in vivo*, which is particularly exciting from a treatment standpoint : splice variants at the RNA level may one day be routinely druggable.[77]

## 1.5 Next-generation sequencing and RNA-seq

The current technologies to research nucleotide sequences have a long and storied history, involving contributions from many scientists over many years. Critical highlights in that history include : the development of Sanger sequencing in 1977, which received a Nobel prize ; the isolation of the temperature-resistant Taq polymerase in 1976 ; the invention of the polymerase chain reaction in the 1980s, which received a Nobel prize ; the isolation of reverse transcriptase in 1970, which received a Nobel prize ; the innovation of shotgun sequencing during the Human Genome Project in the late 1990s ; and finally the independent development of "next-generation" technologies, most notably the "sequencing by synthesis" technology of the Solexa corporation in the early 2000s.[57,58,78-84] After these developments, sequential application of reverse transcriptase to RNA to create complementary DNA (cDNA), and high-throughput sequencing of that DNA was a matter of course, and in 2008 high-throughput RNA sequencing or RNA-seq was first described.[85]

To describe a typical RNA-seq experiment : first, a biological sample of interest, possibly cells, tissues, or whole organisms, is disaggregated, lysed, and RNA is extracted; next, the RNA is reverse transcribed to cDNA; the cDNA is fragmented to shorter sequences, and often size-selected, usually to about 350 nucleotides; the cDNA is ligated to sequencing adaptors and amplified by PCR; the resulting cDNA library is sequenced on a high-throughput sequencing instrument (*e.g.*, from the Illumina corporation); and, finally, short (50nt to 150nt) paired-end sequence reads are output as files on a computer. In the end, the result is a list of about 60 million short paired read sequences which require no prior sequence expectation, and which fairly reflect the abundance of RNA molecules in the original sample, owing to random sampling (described at length in section 1.7). Due to these properties, RNA-seq provides a strongly quantitative means to estimate gene expression, and discovery of new sequences, including novel mutations, short insertions and deletions, splice variants, gene fusions, antisense gene expression, long noncoding RNAs, and any other kind of transcript variant.[73,86-92]

## 1.6 Sequence alignment

Bioinformatics analysis begins following high-throughput sequencing. First, a reference copy of the human genome is searched for each sequence read by a process called sequence alignment. The reference human genome most commonly used today ("hg19" or "hg38/GRCh38") has its origin in the Human Genome project, though it continues to be refined over the years.[57,58] Alignment itself has a long history, but key highlights include the development of the Needleman-Wunsch global sequence alignment algorithm in 1970; dynamic programming optimization of Needleman-Wunsch in 1972; the development of the Smith-Waterman local sequence alignment algorithm in 1981, upon which all modern sequence alignment is really based; development of the Basic Local Alignment Search Tool BLAST and its extensions, which serve to computationally optimize Smith-Waterman, in the 1990s; the innovation of searching in Burroughs-Wheeler transformed sequence space using a Ferragina-Manzini index, introduced by the Burroughs-Wheeler Aligner (BWA) and Bowtie in the late 2000s; the development of aligners specifically designed to handle spliced RNA sequences such as Tophat, MapSplice, and SpliceMap; and finally further optimized spliced aligners which use larger memory structures such as STAR and HISAT.[88,93-106] Alignment of RNA sequences is often aided by prior knowledge of gene annotations, and two of the most commonly used gene annotation databases for this purpose are Refseq, and GENCODE (which combines the Havana and Ensembl databases).[107,108] Additional protein-level databases, such as UniProt, can serve to further guide downstream analysis.[109]

The key idea in sequence alignment is that it is fundamentally a process of determining the most likely genomic origin for an observed sequence read. As such, prior expectations of the sample being aligned which affect the parsimony of sequence alignment explanations come into play, such as the likelihood of mismatches (mutations), gaps (insertions and deletions, and splice junctions), and more complex rearrangements (structural variants, and gene fusions), as are possible to call from DNA sequencing data.[110-112] In short, intelligent decisions about parameterization of sequence alignment are critical to any application that uses it, and because sequence alignment is the first step in nearly all bioinformatics analyses of sequence data (the exception being counting approaches, which still implicitly use sequence), these decisions should be weighed carefully.  Much of the work presented in this thesis concerns applications of sequence alignment to specific problems.

### 1.7 Bioinformatics of splicing analysis

Following sequence alignment, three main approaches can be taken to analyze transcript variant expression (*i.e.* isoform expression) levels from RNA-seq, all of which hinge on its "fair sampling" property : that the number of sequence reads is approximately proportional to the number of RNA molecules present in the original sample. The tools and methods listed in this section are meant to be representative ; there many other tools in the area of splicing bioinformatics from RNA-seq.

The first and oldest approach is to estimate and compare expression of individual exons. Computationally this is simple to perform, by counting read depth over exons, but requires prior knowledge of exon boundaries and usually involves sophisticated statistical analysis downstream, such as edgeR or DEseq / DEXseq.[113-117] Exon expression approaches suffer from non-random PCR amplification over the transcriptome, but this PCR bias may be overcome through sophisticated comparison of case and control samples. More importantly, however, prior knowledge of exon boundaries is often infeasible in research projects concerned with novel biology (*e.g.*, cancer research), where actual novel exons or genomic ranges may be expressed, so a non-trivial pre-processing step of identifying such ranges is necessary.

The second and most common approach is to estimate and compare expression of entire gene isoforms. In this approach, reads are counted over exonic regions and splice junctions, and probabilistically or fractionally assigned to multiple annotated isoforms at the gene locus based on unique and shared coordinate ranges. The counting piece of this approach is sometimes coupled with the task of determining the coordinates of the full length transcripts, called sequence assembly. Two popular tools to achieve this dual task are Cufflinks and Trinity, though many others exist, particularly if the steps are handled independently.[118,119] A relatively new variant of this approach is to calculate expression of anticipated transcripts without alignment, by counting observed subsequences termed "k-mers" extremely quickly, as in Sailfish.[120] Once expression of transcript variants has been estimated, relative abundances can be compared between conditions using another range of tools. This approach is robust to PCR bias, but suffers grievously from misalignment owing to rare spurious reads (*e.g.*, ligation artifacts). For instance, an independent review of transcript assemblers found that the best performing tool, of fourteen, recovered merely 21% of full transcripts in *Homo sapiens* transcriptomics data, and expression estimates for the tools correlated between 0.34 and 0.70 with independent expression estimates.[121] Further, in the presence of sample degradation, such as deeper coverage over the 3' end of the gene owing to polyadenylation capture of degraded RNA ("3' bias," which is extremely widespread), probabilistic assignment weighs all the gene's isoforms equally, which makes little sense in terms of parsimony if any of the isoforms are expected to be rare. This is to say, 3' UTR depth provides no actual evidence for the individual presence of all the transcripts that share it, only the set of transcripts sharing that 3' UTR as a group.

Finally, the newest and least common approach to analyze RNA splicing from sequencing data is to estimate and compare expression of splice junctions themselves. This method is the simplest to execute, requiring only spliced alignment to the genome and counting, and uses similar sophisticated downstream statistical analysis to exon expression. Two examples of methods using this approach are MISO and MATS.[122,123] Splice junction expression benefits from prior knowledge of junction boundaries, but novel junctions can easily be discovered by spliced aligners, and through use of methods presented in this dissertation can easily produce expression counts comparable to known junctions. Splice junction expression also suffers from position-specific PCR bias, the same as exon expression, but again through comparison of case and control samples this can be circumvented. Finally, perhaps the largest drawback of junction expression as an approach is that it doesn't make full use of the available data over the length of the transcript, but in cases of degradation this is a boon rather than a liability.

## 1.8 Alternatives to sequencing to study RNA splicing

RNA sequencing is undoubtedly the highest-throughput current means to analyze splicing at the RNA level, but for historical perspective, orthogonal validation, and analysis at the level of the protein, it is useful to be aware of alternative technologies.

High-throughput alternative technologies include expressed sequence tags (ESTs), serial analysis of gene expression (SAGE), and exon microarrays. ESTs were popular in the early 2000s, and work by first reverse transcribing RNA in a sample of interest into complementary DNA (cDNA), inserting that cDNA with a constitutive promoter and without introns and other normal regulatory regions as circular DNA into host cells ("transformation," using bacteria), having the host cells express high levels of the gene, and then performing traditional Sanger sequencing in parallel on many copies of the gene the host cell expresses.[124] ESTs give long sequence reads (>500 nucleotides), and are therefore appropriate for characterizing gene and isoform structure, but are not quantitative with respect to the original sample. SAGE was also popular in the early 2000s, and works by reverse transcribing RNA to cDNA, attaching biotin which serves to anchor cDNA to beads which bind biotin (streptavidin), truncating the cDNA molecules at one end using an enzyme to digest DNA (restriction endonucleases) to short fragments, amplifying the short sequences with PCR, and eventually sequencing the short fragments.[125] SAGE is quantitative to the level of gene expression because it uses samples directly, but loses full transcript structures because of the digestion step, so it cannot be used in analysis of isoform expression. Finally, exon microarrays were a major advancement over these other technologies in the mid 2000s, and work by tiling sequences complementary to known transcribed regions (exons) on a microarray, introducing RNA from a sample directly or after PCR into the microarray, and measuring hybridization intensity as fluorescence, thereby measuring expression of individual exons. Exon microarrays require prior knowledge of exon sequences, and are biased depending on the strength of the complementary oligonucleotide binding, but were a dramatic improvement over previous technologies and still see some use today in validation.  A relatively new variation of the microarray from NanoString Technologies requires no PCR amplification, and could also in principle be used to validate isoform expression, but is more commonly used for gene expression.[126]

There are many targeted validation approaches.  First, the polymerase chain reaction (PCR) can be used to quickly and easily validate the presence of specific, targeted, short sequences in samples, and is highly quantitative.  Next, Sanger sequencing can be used to validate full transcript structures for enriched sequences, but is relatively slow. Similarly, 3' and 5' rapid amplification of cDNA ends (RACE), can provide full transcript structures, given targeted sequences at the 3' or 5' end of the gene. Each of these is well-established, and work well in the context of validating individual targets.

Finally, high-throughput technology such as short read sequencing from Illumina can be complemented, validated, or even replaced with other current high-throughput technology, such as longer read sequencing like IsoSeq from PacBio or Ion Torrent from ThermoFisher.[127,128] These decisions usually weigh expense and throughput, and come down to the individual researcher and project, but it's worth mentioning that many tissue samples have some level of RNA degradation, which limits the utility of long-read sequencing.

## 1.9 Methods to study splicing at the protein level

Although most researchers currently analyze splicing and other transcript variants at the level of RNA to leverage the throughput and sensitivity of NGS, many or most researchers, particularly disease researchers primarily interested in phenotypes, are often more interested in the effect that variation has on expressed mature proteins.

Currently, the best high-throughput means to study splicing at the protein level is through use of mass spectrometry. Briefly, a typical workflow for mass spectrometry is first to isolate proteins from a sample of interest (cells, tissues, or targeted fractions such as organelles or immunoprecipitation pulldowns), digest the proteins into protein fragments called peptides using the digestive enzyme trypsin (produced by one or more animals), ionize and separate the peptides by their mass to charge ratios (using one of several technologies), fragment the peptides and separate again by mass to charge ratio, termed MS-MS owing to this second iteration (again using one of several technologies), and finally search the resulting mass to charge ratio data against a database of expected fragments, usually the non-redundant human transcriptome for human studies.[129-132] In a splicing context, without advance warning of the possible presence of isoforms they may be missed by database searching, so the key idea is to extend the database with novel expected sequences ahead of time ; this field is called proteogenomics.[133-139] This approach has been used successfully to identify breast cancer-specific splice variants at the protein level, and intriguingly further, predict their expected function using annotations and protein folding methods.[140-144]

Splice isoform expression can be easily validated at the protein level using antibodies which bind to each variant, and running the pulled down proteins on Western blots, which separate input proteins by their mass.

# CHAPTER 2

## Oculus: faster sequence alignment by streaming read compression

Citation: **Veeneman BA**, Iyer MK, Chinnaiyan AM. Oculus: faster sequence alignment by streaming read compression. *BMC Bioinformatics*, **13**:297 (2012).[145]

*This manuscript was ranked as "Highly accessed" by BMC Bioinformatics, and has been accessed over 5000 times and cited three times as of September, 2016. The subject of read compression in alignment remains an area of attention for algorithm developers.*[146,147]

### 2.1 Abstract

Despite significant advancement in alignment algorithms, the exponential growth of nucleotide sequencing throughput threatens to outpace bioinformatic analysis. Computation may become the bottleneck of genome analysis if growing alignment costs are not mitigated by further improvement in algorithms. Much gain has been gleaned from indexing and compressing alignment databases, but many widely used alignment tools process input reads sequentially and are oblivious to any underlying redundancy in the reads themselves.

Here we present Oculus, a software package that attaches to standard aligners and exploits read redundancy by performing streaming compression, alignment, and decompression of input sequences. This nearly lossless process (> 99.9%) led to alignment speedups of up to 270% across a variety of data sets, while requiring a

modest amount of memory. We expect that streaming read compressors such as Oculus could become a standard addition to existing RNA-Seq and ChIP-Seq alignment pipelines, and potentially other applications in the future as throughput increases.

Oculus efficiently condenses redundant input reads and wraps existing aligners to provide nearly identical SAM output in a fraction of the aligner runtime. It includes a number of useful features, such as tunable performance and fidelity options, compatibility with FASTA or FASTQ files, and adherence to the SAM format. The platform-independent C++ source code is freely available online, at http://code.google.com/p/oculus-bio.

## 2.2 Background

Nucleic acid sequencing throughput has grown exponentially for the past ten years, and is expected to continue to shatter Moore's law.[148] Though the highly anticipated onslaught of inexpensive sequencing empowers exciting new biological studies, it also presents a critical problem: the skyrocketing computational costs of sequence analysis.[149] Computers may become the bottleneck of genomics research if these growing processing demands are not mitigated by improvements in software algorithms, especially in light of the sequencing demands of personalized medicine.

Much intellectual effort has been invested in minimizing the time required to align a single read against an indexed database. When performed sequentially, each sequence in the input is processed individually, such that the sum of the alignment times of the input sequences is the total running time. Today's fastest and most widely used aligners, such as Bowtie, BWA, MAQ, RazerS, and BLAST, process input reads sequentially.[98,99,104,150,151] These aligners can typically be configured to be consistent and guarantee that identical copies of an input sequence will produce identical alignment results. Therefore, given a set of input reads with ample redundancy, we envisioned that alignment time could be reduced without compromising accuracy by distilling the unique set of sequences and aligning them using a sequential alignment tool.

Harnessing redundancy in sequence alignment input is not a new concept. BLAST + gains a performance benefit by saving alignments within batches.[96] Cloudburst and CloudAligner use MapReduce, and feature a shuffle step wherein seed sequences in the query and database are brought together and combined.[152,153] SEAL also uses MapReduce; it effectively parallelizes BWA, and can remove duplicate reads by comparing alignment position, after aligning all of them.[154] Similarly, SlideSort sorts together sequences with common substrings, and mrsFast uses a sophisticated blocking map to identify unique seeds before performing a direct map-to-map comparison.[155,156] Finally, Fulcrum performs hashing on seed sequences using MapReduce to conserve computation time in genome assembly.[157] While all of these are excellent tools in their own application spaces, sequential aligners such as Bowtie and BWA enjoy extensive support, remain popular for many applications, and can benefit from the same approach. Furthermore, decoupling the process of compressing input reads from the alignment kernel itself could be productive, as improvements to both algorithms can proceed independently. To date, no application exists that performs streaming read compression in a generalized way.

### 2.3 Methods

We explored the nature of read redundancy across thirteen publicly available next-generation nucleotide sequencing datasets. In a series of experiments we measured the contributions of the application (whole genome, targeted exome capture, RNA-Seq, and

ChIP-Seq), read length, and sequencing depth to overall read redundancy, measured in the percentage of unique reads. Using these observations, we wrote the streaming read compression algorithm Oculus and constructed a model to determine the value of streaming read compression for a given dataset. Finally, we benchmarked Oculus on full sequencing datasets.

### 2.3.1 Sequence data profiling

We evaluated thirteen publicly available datasets that were representative of the major applications of high-throughput sequencing, identified here by their NCBI Sequence Read Archive (SRA) accession numbers. There were five RNA-Seq datasets (ERS025093 (pooled), and SRR097790, SRR097792, SRR097786, and SRR097787 from the iDEA challenge), three genome datasets (SRR097850 and SRR097852, also from the iDEA challenge, and ERR000589), three Exome sequencing datasets (SRR098490, SRR098492, and SRR171306), and finally two ChIP-Seq datasets: (SRR227346, and SRR299316 + SRR299313 (pooled)). The ChIP-Seq data was downloaded from the ENCODE Project, hosted on the UCSC genome browser. Illumina, Inc. carried out the IDEA dataset sequencing, first used by Sun et al.[158,159] Additional run metadata can be found in Appendix A.

### 2.3.2 Sequencing type

The sequencing datasets we selected varied widely in their composition. We compared read redundancy between sequencing types by standardizing the number of reads per dataset to 24 million with random subsetting, and read length to 36 bases with 3′ end trimming (both lowest common denominators). RNA-Seq had relatively redundant reads; only 43% to 57% of each single-end dataset was unique (Figure 2.1). In contrast, Exome and Genome sequencing had very little read redundancy. The two ChIP-Seq datasets had disparate content, varying greatly in their %unique reads – without delving into the specifics of those samples, we believe this may reflect the wide variety of ChIP-Seq applications. As expected, paired-end data compressed less well than single-end, since paired-end compression requires identity on both reads.

### 2.3.3 Depth of coverage and read length

Given some fixed input DNA from which fragments are sampled, each incremental read will be more likely to duplicate previous reads. In particular, RNA-Seq reads may disproportionately reflect highly expressed genes, suggesting that higher sequencing coverage could have a nonlinear effect on read redundancy.[160] Therefore, we measured the impact of coverage depth/sequencing run size (number of reads) and read length on the unique read percentage of each dataset, treating reads individually (single-end) or as pairs (paired-end) (Figure 2.2). We fixed the read length for RNA-Seq runs and evaluated %unique reads for a series of random fractions of the original datasets. As predicted, larger sequencing runs corresponded logarithmically to a lower unique fraction of the datasets (Figure 2.2A). The unique read fraction varied between 56-69% for 10 million reads, 32-49% for 25 million reads, and 28% for 385 million reads in RNA-Seq dataset #1. The differences between datasets likely relates to sample biology and preparation. Next, we fixed coverage depth and evaluated the percentage of unique reads for a series of read lengths (trimming from the end) (Figure 2.2B). The impact of read length on uniqueness appeared to be exponential in one case (RNA-Seq #1, for which 100 bp reads were available) and linear in the rest (RNA-Seq #2-5). It's interesting to note that some RNA-Seq algorithms, such as TopHat, dice unmapped reads into segments and align each piece individually.[100] This might entail a ~3-fold alignment speedup for RNA-Seq dataset #1 by use of 25 base segments, if further communication between a streaming read compressor such as Oculus and Tophat's core algorithm could be engineered.

## 2.4 Implementation

The overall architecture of Oculus is shown in Figure 2.3. Oculus reads FASTA or FASTQ input files, processes sequences into a compressed form, and compares them to a map containing all sequences it has seen before; new sequences are passed into the aligner as FASTA, while previously observed sequences increment counts in the map. At the reconstitution step, sequences in the SAM output file are then compared back against the map and re-printed as many times as they appeared in the input, correcting for alignment orientation. Paired-end sequences are handled by concatenating the two sequences to ensure the pair is unique. Oculus can wrap any aligner capable of producing SAM-formatted output.

By design, Oculus sacrifices FASTQ quality scores, read names beyond the first instance of the sequence, and the original order of the reads in the output. Optionally, users can direct Oculus to restore the original read names and quality scores by writing them to an intermediate file, sorting it, and reattaching them during the reconstitution step. This option incurs additional memory overhead, and additional time to sort the intermediate file.

### 2.4.1 Data structures

Oculus uses hashmap data structures to store sequences in memory. Users can either compile in standard library (STL) hashmaps, or Google-SparseHash maps, which are faster and require significantly less memory (2 bits of overhead per entry).[161]

Optionally, users can direct Oculus at runtime to store unique reads in a separate hashset, reducing the burden on the hashmap to only redundant sequences. The effect of this is to reduce lookup times in the reconstitution step and total memory consumption, at the cost of more operations in the compression step. Hashsets are expected to be beneficial for lower redundancy input.

Oculus uses a modified version of MurmurHash2 to hash binary sequence data.[162] It has a low incidence of collision for binary data, and was recommended for use with Google-SparseHash by its developer (C. Silverstein, personal communication). To reduce collisions, the hash algorithm operates only on the sequence field of the compressed sequence objects.

### 2.4.2 Binary compression

Instead of storing sequences in memory as ASCII characters, Oculus uses compressed sequence objects of our own design (cseqs) (Figure 2.4). DNA sequences are dynamically compressed into 2 or 3 bits per base, depending on the presence of N nucleotides. Optionally, a 2-bit encoding can be forced if the user wishes for N's to be evaluated as A's. Each cseq has three fields: a representation bit indicating the nucleotide encoding, its size in memory, and a variable-length compressed sequence. Storing the size is necessary because null-termination is obviated by the possibility of null bytes in the sequence field.

The most obvious benefit of using cseqs is an approximate four-fold reduction in memory use. However, two engineering benefits also arise for cseq string comparison, which help efficiently resolve map collisions. Sequences with different lengths or representations can be differentiated by comparing the first byte in constant time (very quickly). Moreover, by comparing nucleotides in blocks instead of individually, comparison time is reduced four-fold. Memory for sequences is allocated in large chunks (default: 10kB), which reduces overhead greatly.

### 2.4.3 Reverse complements

Lastly, Oculus can be directed to compress together reverse complements in single-end data, or reversed read order in forward-reverse oriented paired-end data, under the presumption that they should align to the same place in the database. This improves compression and therefore reduces aligner runtime. Using reverse complements is optional because BWA and Bowtie both use left-end seed sequences, so the orientation of the read can affect its alignment (though typically in a tiny fraction of sequences).

### 2.4.4 Runtime model

We developed a model to predict the effectiveness of Oculus for any given data set. Given Ni input reads that compress to Nc sequences, and assuming sa and so are the speeds of the aligner and Oculus, in reads/unit time, the following equations give the expected benefit of using Oculus as a fraction of the aligner's run time.

$$\text{Aligner Run Time} = N_i / s_a$$

$$\text{Oculus Run Time} = (N_i / s_o) + (N_c / s_a)$$

$$\text{Run Time Ratio} = \text{Oculus} / \text{Aligner} = (s_a / s_o) + (N_c / N_i)$$

The aligner's run time is simply the total number of input reads divided by the average alignment speed in reads per unit time of the aligner. In the second case, since Oculus passes some fraction $N_c$ of $N_a$ into the aligner, the aligner only has to do $N_c/s_a$ work. However, there's also an overhead for Oculus on the order of the total number of input reads. The fractional benefit of using Oculus is therefore related only to the compression achieved and Oculus's speed relative to the aligner it's wrapping. We therefore derived processing rates in reads per second for Oculus and each aligner, for both single-end and paired-end data, using experimental results for the 50 and 51-mer datasets. Table 2.1 indicates the calculated ratio of the speed of the aligners to Oculus. Based on these parameters we predict that Oculus will have a runtime benefit for sequence data with greater than 10% redundant reads, and that benefits would scale linearly with the unique read fraction. This model discounts non-linear factors such as hash collisions, read length, percent successful alignment, and potentially, alignment location, and disk I/O will produce noise, but it is an effective rule of thumb.

### 2.4.5 Benchmarking

We compared the performance of Oculus with BWA (version 0.5.9-r16) and Bowtie 1 (version 0.12.7 64-bit) by themselves. All alignment was performed against the reference human genome GRCh37/hg19.

Every benchmarking test was run on the Flux supercomputing cluster maintained by the Center for Advanced Computing at the University of Michigan, using single CPU cores of 2.67 GHz Intel X5650 processors, with 64 GB of 1333 MHz DDR3 memory, and distributed access disks. To reduce noise in runtime measurement from disk I/O, each benchmark test was run three times, and the average runtime is presented here. Memory consumption was much less noisy, so similar averaging was unnecessary in reporting memory use. Both aligners ran with entirely default options, and Oculus used only the reverse complement storage option, “--rc”.

To test consistency, we ran Bowtie using “-m 1” to eliminate multi-mapping reads, for which Bowtie reports one random alignment by default. We extracted alignment positions, sorted by read sequence (grouping together forward and reverse orientations), and counted and classified alignment differences. BWA has no such mono-mapping option, so we did not test Oculus’s wrapping of BWA for consistency (BWA was still tested for performance).

## 2.5 Results

### 2.5.1 Compression and performance

Oculus yielded performance benefits that strongly correlated with the unique read fraction of each dataset (Figure 2.5). Notably, the single-end RNA-Seq datasets aligned in 49.7% as much time on average, i.e., they ran 2.0 times as fast in Oculus compared with Bowtie and BWA. The paired-end datasets compressed less well than their single-end counterparts; on average, the paired-end RNA-Seq datasets aligned 1.2x as fast. ChIP-Seq dataset #1 received the greatest performance benefit: its single-end Bowtie alignment ran 3.7x as fast. However, our Genome and Exome datasets, and ChIP-Seq dataset #2, were generally non-redundant and Oculus did not greatly outperform either aligner. This was consistent with our expectations - if reads are not redundant, they cannot be compressed, and the aligner will receive nearly the complete set of input reads. Since compressing and decompressing incurs a small time overhead, it follows that a nearly completely unique dataset might run more slowly.

Though BWA was much slower than Bowtie for single-end data, and somewhat slower for paired-end data, Oculus produced similar fractional speed improvements for the two aligners. Additionally, for the datasets tested, Oculus's hashset option did not yield a significant improvement. For sequencing run information and exact CPU run times, see Appendix A.

### 2.5.2 Consistency

Oculus maintained high fidelity to original alignments for every dataset. Defining accuracy as the percentage of input reads that Oculus mapped to exactly the same location as the aligners, on average Oculus was >99.9% accurate, and in the worst case was 99.874% accurate. For individual dataset accuracy, see Appendix A.

Since they change the seed sequence used in alignment, the vast majority of the differences (inaccuracies) produced were for reads that Oculus either reversed the orientation of (88% of single-end differences), or order of (67% of paired-end differences). Mostly these were previously unaligned reads that aligned and vice versa, but in some cases, an unambiguously mapped read actually changed alignment positions (single-end, 0.09% of differences; paired-end, 10.15% of differences). Though initially surprising, this can be explained by mismatches in seed sequences. Bowtie is less permissive of mismatches in the seed than at the end of a read under the assumption that read quality tends to be better toward 5′ end. Of two closely homologous regions of the genome, one may count as the best hit in the forward orientation, and the other in reverse orientation. For example:

CAGT - read
CATT – genome position 1

CCGT – genome position 2

In this case, if CA is the seed, position 1 would be the optimal alignment and the third base would count as a G-T mismatch. However, if the reverse-complement were aligned, and the seed proceeded from the opposite direction, position 2 would be optimal and the third base would be recorded as a C-A mismatch.

### 2.5.3 Memory use

Oculus very consistently used (sequence length/4) + 20 bytes of memory per map entry. This 20-byte overhead comes from the forward and reverse count integers (4 each), the hash of the sequence (4), a pointer to the sequence (up to 8 on a 64-bit OS), the size field (2), and some heap memory structure overhead. Although these sum to 22 bytes, hash values are not stored multiple times for hash collisions, and pointer memory use varies by OS architecture, often using less than 8. This 20-byte overhead is halved for paired-end map entries, because each pair is stored together. Using the hashset option reduced memory use by about a third, by mitigating some of this overhead for unique reads.

Total memory use is therefore highly dependent on the quantity and redundancy of input sequence, but in a worst-case scenario (perfect non-redundancy), 100 million single-end 80mers will use about 3.7 GB of memory, on top of memory used by the aligner’s database. Redundancy translates linearly to reduction in memory use – if only half of those reads were unique, 1.85 GB would be required instead.

## 2.6 Discussion

Our benchmarking tests suggest Oculus will generally perform very well with RNA-Seq data and on a case-by-case basis in other applications, particularly those with low complexity libraries. The likely source of benefit to RNA-Seq arises from highly expressed genes that are sequenced at great depth and generate multitudes of duplicate reads.

Shorter read length and larger datasets both correlated with higher redundancy in sequencing runs. The hidden variable of actual biological redundancy remains at large (particularly, the effects of PCR and the targeted scope of sequencing), but those two metrics provide good insight into the expected value of streaming read compression for a given sequencing application. We noted the added value Oculus provides for RNA-Seq applications that segment reads (Oculus can significantly benefit the alignment of many 25mers), but Oculus may also yield benefit to customized bioinformatics analyses that take similar approaches. Also of note is that for highly-sensitive but slow aligners such as BLAT and Smith-Waterman, Oculus's relative runtime will be insignificant (i.e., sa/so - > 0), so streaming read alignment will be of greater use to applications that require such sensitivity.[105,163] Perhaps most importantly, as sequencing throughput increases so too will read redundancy and the marginal benefit of compressing input reads, though this will be mitigated by longer read lengths and paired-end reads.

To be effective, Oculus requires read redundancy and an aligner that does not already exploit that redundancy. To be consistent, Oculus requires the aligner to ignore quality score and use parameters that guarantee deterministic behavior. By default, Bowtie will report one alignment at random for ambiguously mapping reads, and Oculus by definition cannot produce multiple alignments for a single read sequence. The exception to this is if the aligner is configured to report multiple alignments per read, either on single or multiple SAM lines, in which case Oculus will reconstitute the reads aligning to each location.

Since both Bowtie and BWA use left-end seeds, it makes sense that Oculus may report different alignments for reverse-complemented single-end reads. However, we were surprised to find alignment differences for paired-end reads with reversed order. Read order shouldn't matter in paired-end alignment: since the read orientation remains the same, so should the seeds. Developers who wish to incorporate streaming read compression into their aligners may be interested in exploring this phenomenon. Another surprising result was that Oculus + Bowtie actually outperformed compression for the second ChIP-Seq data set (it ran in 27.0% of the original time, on 35% of the original data set). Stranger still, the runtime data for that dataset was not noisy – each of the three tests ran in < 28% of the original time. It is possible that Oculus may have compressed a disproportionately large number of slow-aligning reads – reads that take longer to align to the human genome. Better understanding this phenomenon may be a key to further alignment algorithm improvements.

Though Oculus provides immediate benefit to RNA-Seq alignment, further performance gains may be possible by harnessing the idea of streaming read compression. Although implemented here as a customizable "attachment" to a sequential aligner, the streaming compression algorithm could be integrated directly into alignment kernels. One obvious benefit of this would be the ability to store paired-end reads individually (with an extra bit denoting the read number) thereby leveraging additional redundancy (see Figure 2.1). A more nuanced logical continuation of this idea would be for aligners to use cache objects that retain in memory the alignments of the mostly commonly occurring reads. If present, a skew toward very common reads away from reads with few copies could create the perfect conditions for caching. The combinatorics of sequence length suggests an even greater benefit in storing and reusing alignments of common seed sequences, either in a complete object or a cache.

There are three limitations of Oculus's current implementation of streaming read compression: FASTQ quality scores are lost, read names are lost beyond the first instance of the sequence, and the order of the reads in the output will not be consistent with normal aligner output. Quality scores and read names can be restored to the final output at the cost of computation time and memory, which adds value for downstream analyses such as SNP calling. However, the alignment itself is still performed without quality scores, which can alter alignment results. In cases where little faith is placed in the read quality scores this may be acceptable, but to mitigate this loss otherwise, we suggest the use of read filtering or trimming as a preprocessing step.

## 2.7 Conclusion

Oculus provides a demonstrable speed improvement in aligning redundant data, with high fidelity and low memory cost. Further, streaming read compression of redundant reads is generally useful; aligning the unique set of reads is faster than the full set since the overhead of compression is sufficiently low. We expect streaming read compression will play an important role in RNA-Seq alignment and potentially other sequencing applications in the future as data grows and algorithms improve.

### 2.7.1 Availability and requirements

| | |
|---|---|
| Project Name: | Oculus |
| Project Home Page: | http://code.google.com/p/oculus-bio |
| Operating system: | Platform independent |

| | |
|---|---|
| Programming language: | C++ |
| Other requirements: | Perl version 5 or higher (for configuration), g++ version 4.1.2 or higher (lower versions may work but are untested), Bowtie or BWA (versions 0.12.7 or 0.5.9-r16, respectively), or another SAM-compatible alignment algorithm |
| License: | GNU GPL v3 |

### 2.7.2 Author contributions

BAV provided the original idea, wrote the algorithm, performed the benchmarking, modeling, and data interpretation, and drafted the manuscript. MKI contributed critical feedback on the manuscript, suggestions for datasets, and the reverse complement idea. AMC contributed critical feedback on the manuscript and the project, and provided the computational resources necessary to carry out the work. All authors read and approved the final manuscript.

### 2.7.3 Acknowledgements

We would like to thank Terrence Barrette for supporting the lab’s infrastructure, Jeremy Hallum and CAC for supporting Flux, Illumina for referring us to the iDEA data’s reference, Craig Silverstein for developing Google Sparse-hash and his helpful comments on hashing, Austin Appleby for developing MurmurHash, Catherine Grasso and O. Alejandro Balbin for their comments on the manuscript, and Karen Giles and Christine Betts for their assistance with the manuscript submission.

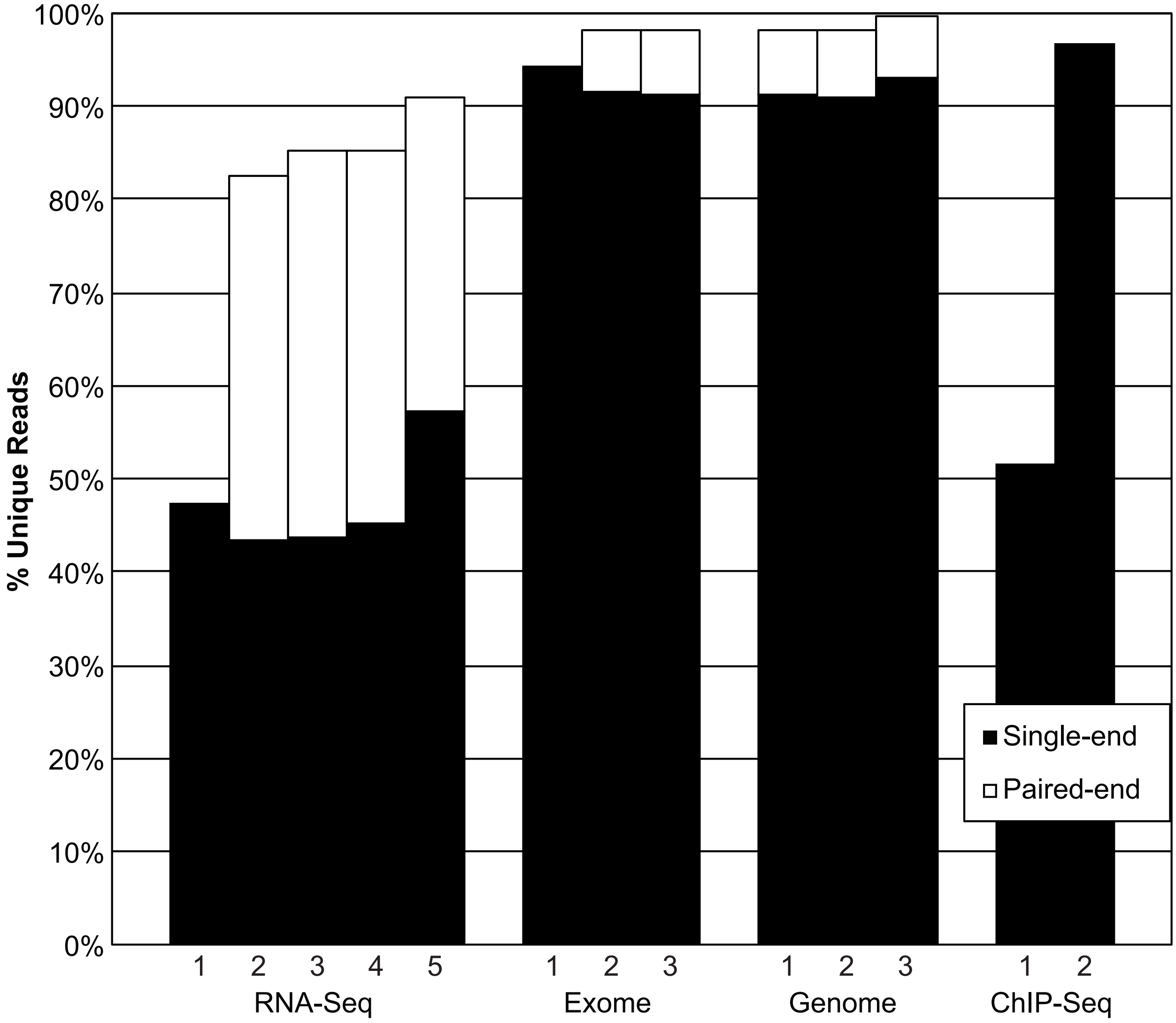


**Figure 2.1 RNA-Seq compresses better than other sequencing platforms.** Each benchmark dataset was randomly subset to the lowest common denominator number of reads (24 million) and read length (36 bases). Subsequently, Oculus computed the unique read fraction for each dataset using the reverse-complement option. For data with paired-ends available, 12 million pairs were used to computer %unique reads. RNA-Seq #1, Exome #1, and ChIP-Seq #1-2 did not have available paired-end data.

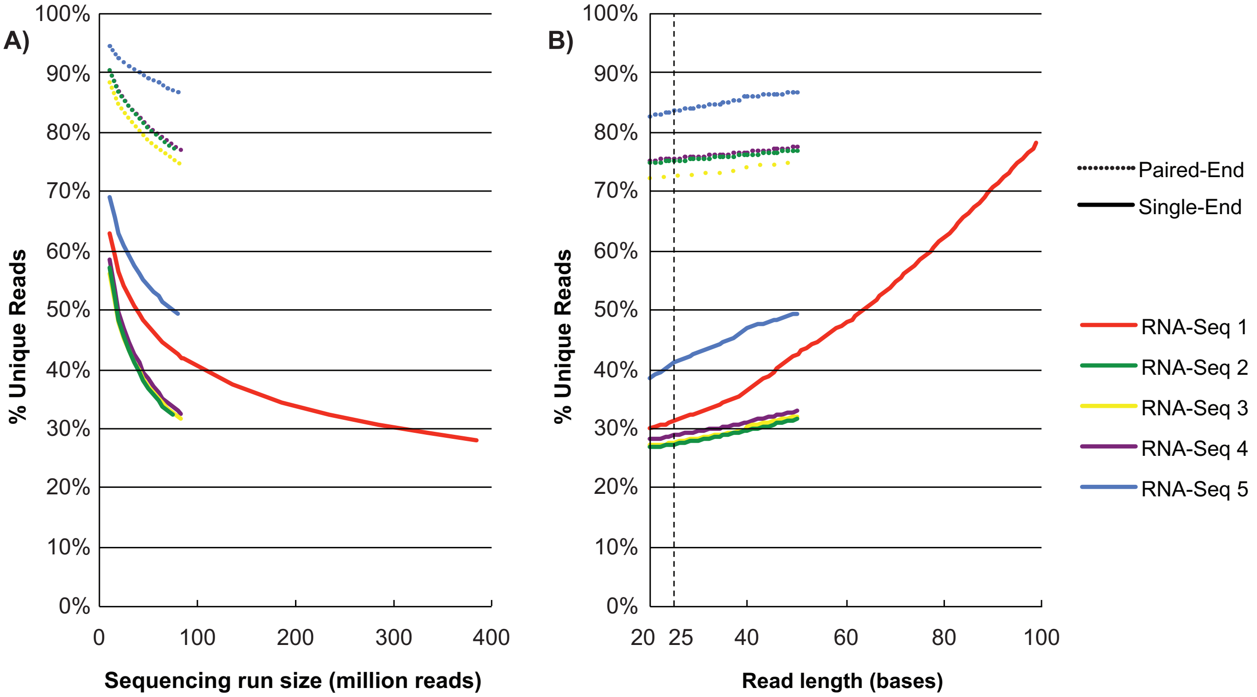


**Figure 2.2 Compression improves for larger sequencing runs and shorter read lengths.** A) Each RNA-Seq dataset was trimmed to 50-base reads, and %unique reads was computed for a series of simulated sequencing run sizes (between 10 million single-end or paired-end reads and their original size). B) Each RNA-Seq dataset was randomly subset to 79 million single-end or paired-end reads, and %unique reads was computed for a series of simulated read lengths by trimming from the end (between 20 bases and their original read size). 25 bases is a typical sequence length that advanced RNA-Seq pipelines such as TopHat may use for segmented alignment.

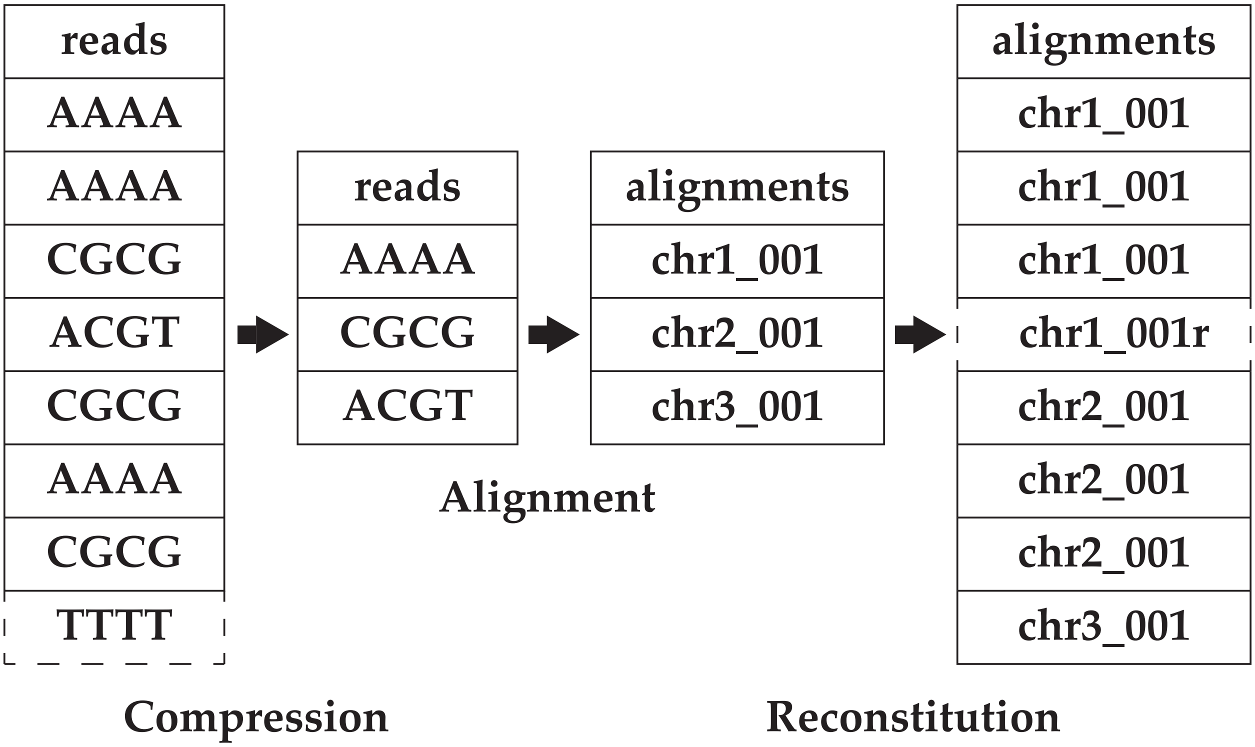


**Figure 2.3 Flowchart depicting Oculus behavior with example sequences.** As input is parsed, new sequences are passed into the aligner in the order they are observed. The aligner then performs normally, mapping each passed read to the database. Downstream of the aligner, Oculus expands the alignment file to reflect the count of each input sequence. Since compression and reconstitution are faster than alignment, there is a net reduction in runtime. In reverse-complement mode (Section 2.4), Oculus would remove the read sequence TTTT, having already seen AAAA, and print an additional alignment: chr-001 with reversed orientation. By default, Oculus treats AAAA and TTTT as distinct sequences – both would be passed into the aligner.

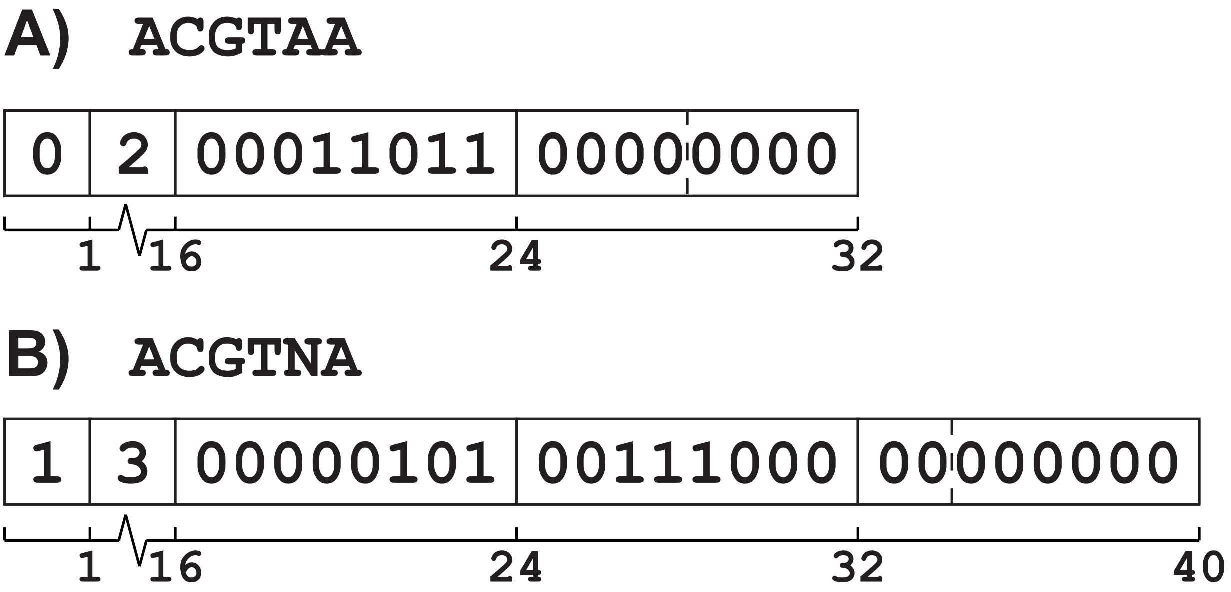


**Figure 2.4 Compressed sequence object (cseq) diagrams.** Numbers below the data fields indicate the 0-based index in bits from the left end. (A) The sequence ACGTAA contains no N's, so its encoding bit is 0, indicating 2 bits per base. By that encoding, two bytes are required to store 6 nucleotides, so the size field is 2. The sequence field is populated by A = 00, C = 01, G = 10, T = 11, etc., with the right-most byte padded on the right by zeros. (B) Compression proceeds as before, until the N nucleotide is encountered, at which point the compression starts over and sets the encoding to 1, indicating 3 bits per base. At that compression, now 3 bytes are required to store 6 nucleotides, and the size field is updated accordingly. The sequence field is populated by A = 000, C = 001, G = 010, T = 011, N = 100, etc., and again the right-end is padded with 0's.

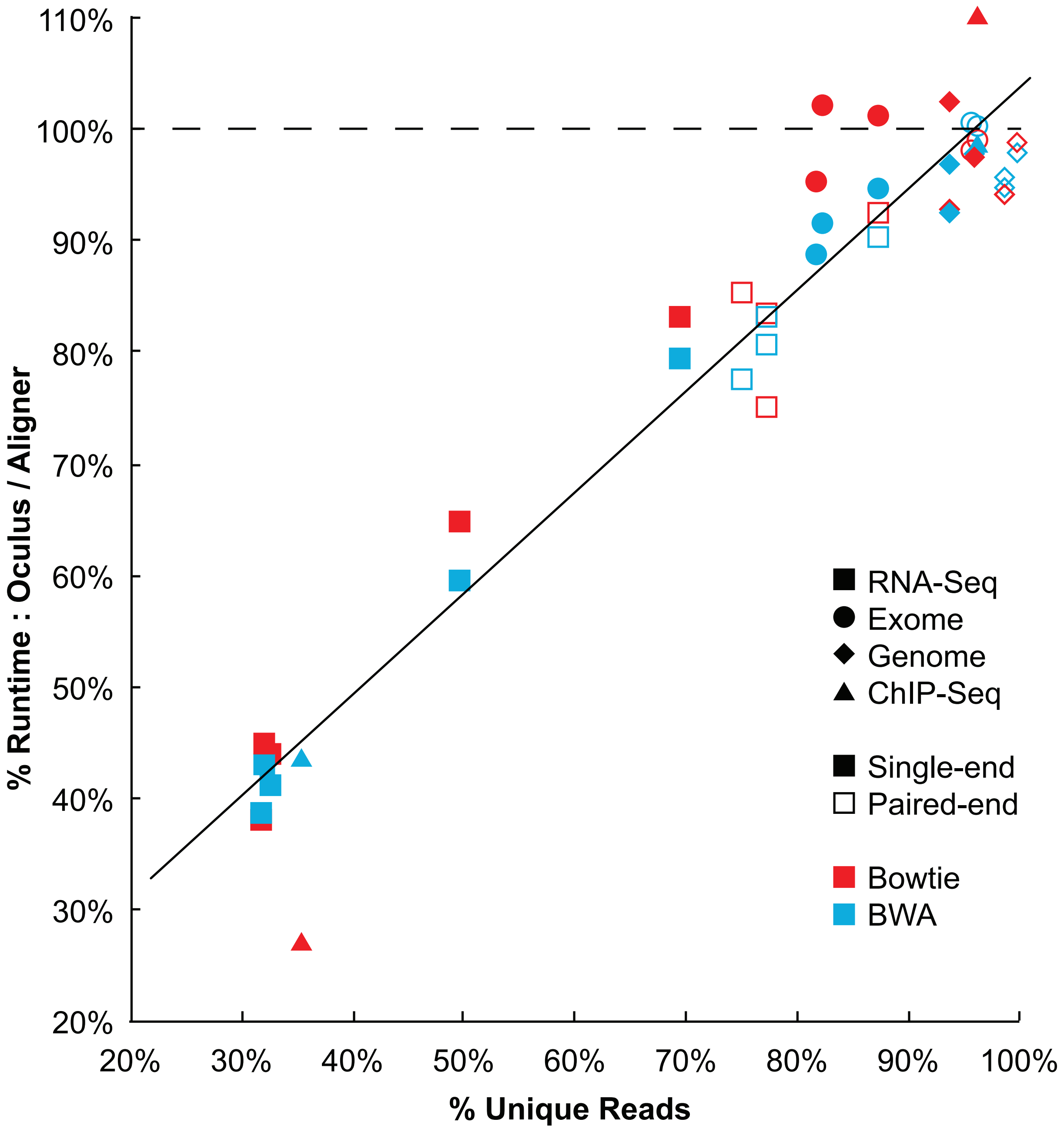


**Figure 2.5 Oculus provides a speedup that correlates linearly with %unique reads.** %Runtime represents the ratio of the runtime of Oculus, wrapping each aligner, to the runtime of the aligner by itself (in CPU time). To best demonstrate fractional benefit, Bowtie and BWA results are combined in this graph – individual run data is available in Appendix A. Oculus provided a speed benefit for points below the dashed line. These datasets span a variety of sequencing types, read number, and read length, which we hypothesized all contribute to the %unique reads for a sequencing run. Filled symbols (rather than black) indicate single-end vs. paired-end. See Appendix A for individual sequencing run characteristics such as read number and read length.

**Table 2.1 Relative processing speeds of Bowtie and BWA to Oculus, for single-end and paired-end data.**

| | | SE | PE |
|---|---|---|---|
| $s_a/s_o$ | Bowtie | 0.079 | 0.023 |
| | BWA | 0.017 | 0.015 |

$s_a$ is the aligner's speed, and $s_o$ is Oculus's speed. Since speeds are measured in reads aligned per second, these values indicate that Oculus runs faster than the aligners, and relatively more fast for paired-end data than single-end data. As expected, Bowtie was measured to be faster than BWA, particularly for single-end data. Both aligners were run with default options.

# CHAPTER 3

# Two-Pass Alignment Improves Novel Splice Junction Quantification

Citation: **Veeneman BA**, Shukla S, Dhanasekaran SM, Chinnaiyan AM, Nesvizhskii AI. Two-pass alignment improves novel splice junction quantification. *Bioinformatics*, **32**:43 (2016).[164]

## 3.1 Abstract

Discovery of novel splicing from RNA sequence data remains a critical and exciting focus of transcriptomics, but reduced alignment power impedes expression quantification of novel splice junctions.

Here, we profile performance characteristics of two-pass alignment, which separates splice junction discovery from quantification. Per sample, across a variety of transcriptome sequencing datasets, two-pass alignment improved quantification of at least 94% of simulated novel splice junctions, and provided as much as 1.7-fold deeper median read depth over those splice junctions. We further demonstrate that two-pass alignment works by increasing alignment of reads to splice junctions by short lengths, and that potential alignment errors are readily identifiable by simple classification. Taken together, two-pass alignment promises to advance quantification and discovery of novel splicing events.

## 3.2 Introduction

Since the first successful application of short read sequencing to cDNA in 2008, broad uptake has proven RNA-seq an indispensable tool in the arsenal of molecular biology.[85] However, for as long as it has existed, analysis of RNA-seq data has been complicated by consequences of the gapped nature of RNA.[87] Briefly, when RNA is transcribed from DNA, putative functional sequences (exons) are interspersed with sequences which are later removed (introns). Because exons originate from noncontiguous genomic contexts, separated by varying distances, the primary challenge in ascribing RNA sequences to their genomic origins is gapped alignment, for which many good tools have been developed.[165] These aligners typically support the use of annotated gene references, which facilitate alignment to known splice junctions, while maintaining the ability to discover novel splice junctions. This approach has the implicit effect of requiring greater evidence for reads spliced over novel splice junctions compared with known splice junctions, and is implemented either by aligning in multiple stages as in Tophat, or by varying alignment scores for different splice junction classes as in STAR (Spliced Transcripts Alignment to a Reference).[102,166] In all such tools, preference is given to known splice junctions, which reduces noise but biases quantification against novel splice junctions.

Two-pass alignment, a framework in which splice junctions are separately discovered and quantified, has recently gained traction owing largely to massive speed enhancements achieved by new aligners, which make aligning twice computationally feasible.[102,165] The rationale behind two-pass alignment is elegant: splice junctions are discovered in a first alignment pass with high stringency, and are used as annotation in a second pass to permit lower stringency alignment, and therefore higher sensitivity. In the absence of annotation, compared to traditional single-pass alignment, an independent analysis demonstrated that two-pass alignment with STAR provides comparable mapping rates (though more multimapping), similar mismatch alignment rates, reduced read truncation, superior read placement accuracy, comparable indel accuracy, improved splice junction recall, and better annotated splice junction detection, with comparable discovery of true novel splice junctions at the cost of more false positive discoveries.[165] While the effects of two-pass alignment on transcript assembly and transcript quantification have also been investigated, our primary interest is in splice junction expression quantification, which is relevant to ascertaining the validity of discovered splice junctions, and has not yet been thoroughly investigated.[121] In light of the evidence that two-pass alignment can improve alignment rate and sensitivity, we investigated what advantages and disadvantages this approach might yield for splice junction quantification.[165]

Here, we describe for the first time several appealing performance characteristics of two-pass alignment.  In an experiment in which known splice junctions are treated as unannotated, two-pass alignment provided excellent quantification accuracy, and significantly more accurate quantification than single-pass alignment. Underscoring the wide applicability of the technique, these quantification benefits were observed across a variety of RNA-seq datasets, including Arabidopsis samples. As a salient takeaway, this corresponded to as much as 1.7-fold median deeper read coverage over novel splice junctions (see Table 3.1 for full per-sample performance statistics).  We go on to demonstrate that two-pass alignment works by permitting alignment of sequence reads by fewer nucleotides to splice junctions.  Finally, while we find that two-pass alignment can introduce alignment errors as previously suspected, we demonstrate that these are relatively simple to detect. In summary, two-pass alignment significantly improves quantification of novel splice junctions, and we recommend its use in studies concerned with novel splice junction discovery and quantification.

## 3.3 Methods

### 3.3.1 Datasets

We acquired twelve publicly-available Illumina paired-end RNA sequencing datasets from five studies, with read lengths ranging between 48 and 101 nucleotides, and library sizes ranging between 34 million and 202 million read pairs. These samples were: two independent pairs of matched tumor-normal lung adenocarcinoma samples from The Cancer Genome Atlas and the study by Seo et al.; two replicates of Agilent's Universal Human Reference RNA (UHRR), sequenced at Illumina; four lung cancer cell lines from the Cancer Cell Line Encyclopedia; and one leaf sample and one flower bud sample from Arabidopsis thaliana (unpublished as of this writing).[167-170] These libraries were selected as high-quality representatives of the breadth of RNA-seq data types typically encountered in biomedical research. Sample descriptions are provided in Table 3.1, and full sample metadata is available in Table B.1.

### 3.3.2 Sequence Alignment

All sequence alignment in this study was performed with STAR (version 2.4.0h1), a fast and sensitive alignment algorithm designed for RNA-seq, which we selected for multiple reasons.[102] First, because STAR was independently reviewed as performing similarly or favorably compared to other methods in splice junction detection and transcript abundance estimation, it reasonably represents modern alignment algorithms in general.[165] Second, STAR provided transparent and fine-grained description and control of critical alignment parameters, which we anticipated would be useful in understanding its behavior. Next, STAR's use in recent publications concerning both broad and sensitive detection of novel transcription suggested it may continue to be used for such purposes, and investigating increased sensitivity using it would be of additional value.[171,172] Finally, STAR's speed made aligning twice in succession more computationally feasible. While aligning in two passes should theoretically affect all single-pass alignment algorithms similarly, here we restricted our analysis to one alignment algorithm for simplicity.

In addition to non-default parameters governing resource management, we followed ENCODE's example as described in the STAR manual in using the following non-default parameters: outFilterType BySJout, for consistency between reported splice junction results and sequence read alignment results; alignIntronMin 20, to set the minimum intron size to 20 nucleotides, for speed and to reduce the likelihood of reporting short indels as introns; alignIntronMax 1000000 and alignMatesGapMax 1000000, to set the maximum intron size to one million nucleotides, longer than the longest known introns, for speed and to reduce the likelihood of mistaking chimeric splice junctions as normal introns; and alignSJoverhangMin 8, to require sequence reads span novel splice junctions by at least eight nucleotides, for specificity. Deviating from ENCODE, we kept: alignSJDBoverhangMin 3, to require sequence reads span known splice junctions by at least three nucleotides (nt), as the suggested 1nt seemed likely to exacerbate alignment errors, and set: scoreGenomicLengthLog2scale 0, to not penalize longer introns compared with shorter introns, which in our experience was more accurate. Full alignment parameters are available in Table B.2.

Human samples were aligned to GRCh38 (full), and Arabidopsis samples were aligned to TAIR10 (all autosomes, plus mitochondrial and chloroplast genomes). We evaluated multiple alternatives for human gene annotation, and selected the GENCODE-Basic gene annotation (v21) as optimal for use in first-pass alignment (when used). It provides a reasonably comprehensive and high-quality gene set, which excludes rarely observed or poorly supported transcript nominations in the complete GENCODE database.  GENCODE-Basic v21 is comprised of 107,529 transcripts, containing a total of 265,193 splice junctions, and is available on the GENCODE website. For Arabidopsis gene annotation, we used TAIR10, acquired from www.arabidopsis.org (127,554 splice junctions across 40,745 transcripts).

To generate data for the quantification accuracy experiments (described below), we performed four types of alignment: single-pass alignment with and without annotation (Annotation 1-pass and De Novo 1-pass), and two-pass alignment with and without annotation (Annotation 2-pass and De Novo 2-pass).  We implemented two-pass alignment as three stages: alignment, reindexing the genome with all discovered splice junctions covered by at least one uniquely mapping read, and alignment to the new genome index. The alignment process is depicted as a flowchart in Figure 3.1.  Higher thresholds for including splice junctions in reindexing may be used, trading off sensitivity for specificity, but we opted for higher sensitivity here. On a related technical note, splice junctions discovered in the second pass, but not the first, are likely artifacts of the alignment process (consistent with reported high  false novel splice junction "discovery" after second-pass alignment cited in the introduction), so we stress that step 4 is for quantification, not discovery.  We also considered an approach in which unannotated alignment is followed by alignment to a pool of discovered splice junctions and the full annotated splice junction list, but it performed similarly to De Novo 2-pass and is uncommon in the field, so we didn't consider it further.

## 3.4 Results and Discussion

### 3.4.1 Quantification Accuracy

To test the splice junction quantification accuracy of two-pass alignment, we designed an experiment as follows, using the sequencing datasets described in the methods and Table B.1. First, we treated the read depth quantification of annotated junctions generated by Annotation 1-pass alignment as correct (a "gold standard"). Annotated 1-pass alignment is very commonly used in projects unconcerned with junction discovery, and should be relatively unaffected by undetected novel junctions, so it is therefore reasonable to believe it provides good quantification of known junctions. Then, treating those splice junctions as if they were novel, we compared the quantification performance of single-pass alignment without annotation (De Novo 1-pass) and two-pass alignment without annotation (De Novo 2-pass), to the "gold standard," essentially testing their ability to recapitulate standard quantification. Because the De Novo alignment approaches had no prior knowledge of the annotated splice junctions, they serve as good proxies for true novel splice junctions. We also performed two-pass alignment with annotation (Annotation 2-pass) out of interest, though that data was not reused in other analyses. Ratios of each alignment approach to Annotation 1-pass are portrayed superimposed for a representative sample, the A549 cell line, in Figure 3.2A, and individually for all samples in Figures B.1-B.12. Extending this analysis, we quantified the extent to which De Novo 2-pass alignment better approximated the gold standard than De Novo 1-pass (i.e., relative quantification accuracy). For each sample, for each splice junction, we calculated quantification improvement as the difference in quantification error between De Novo 1-pass and De Novo 2-pass alignment, as described in Formulae 3.1-3.2, showing x as the quantification level of the given junction in each approach.

Formula 3.1) $\text{error}(x) = \frac{\left|\text{Annotation 1pass} - x\right|}{\text{Annotation 1pass}}$

Formula 3.2) $$\text{improvement} = \text{error(De Novo 1pass)-}\ \text{error(De Novo 2pass)}$$

Tukey boxplots of quantification improvement across splice junctions, per sample, are plotted in Figure 3.2B, and the percentage of splice junctions improved upon are provided in (Table 3.1). Summary statistics per sample, including the median increase in read depth between two De Novo alignment passes, and percentage of splice junctions improved are depicted in (Table 3.1).

From these analyses, we observe that two-pass alignment provides much more accurate quantification of novel splice junctions than single-pass alignment. This is depicted qualitatively for one sample, the A549 cell line, in Figure 3.2A as the blue distribution's deviation from 1.0, compared with the green distribution, and quantitatively as boxplots in Figure 3.2B as deviation from zero. As an example, the median quantification in A549 was approximately 80% of the gold standard (green distribution, Figure 3.2A), and correspondingly, two-pass alignment improved that quantification by about 20% (A549 boxplot center, Figure 3.2B). Across the twelve samples tested, two-pass alignment achieved 1.12x to 1.71x higher coverage over novel splice junctions than single-pass alignment (Table 3.1). Similarly, two-pass alignment improved the quantification of between 94% and 99% of the splice junctions in each sample, over single-pass alignment (Table 3.1).

Next, we ascertained the absolute quantification accuracy of De Novo 2-pass alignment, again in comparison to Annotation 1-pass alignment. For each sample, we counted the number of splice junctions within various accuracy thresholds: "Identical to Standard," meaning De Novo 2-pass alignment produced exactly the same read count as Annotation 1-pass; "Within 1%," meaning De Novo 2-pass produced a count within 1%

of the Annotation 1-pass count (but not identical); "Within 5%," meaning De Novo 2-pass produced a count within 5% of the Annotation 1-pass count (but not within 1%); "Over-quantified," meaning De Novo 2-pass exceeded Annotation 1-pass by more than 5%; "Under-quantified," meaning De Novo 2-pass was less than Annotation 1-pass by more than 5% (but not totally missed); and "Missed," meaning De Novo 2-pass produced zero reads for a splice junction covered by at least one read in Annotation 1-pass. Cumulative barplots for each sample are depicted in Figure 3.2C.

From this analysis, we observe that regardless of its relative improvement over one-pass alignment, two-pass alignment provides accurate novel splice junction quantification. Across the twelve samples, two-pass alignment provided "correct" quantification (identical to Annotation 1-pass) of at least 75% of splice junctions, and provided nearly correct quantification (within 5% accuracy) of at least 88% of splice junctions. We speculate that variability in the percentage of splice junctions quantified identically to the standard, versus those within 5%, was mostly driven by the number of reads per sample - samples with twice as many reads were less likely to produce exactly identical counts (see Table 3.1 for read counts). Instead of normalizing (e.g., read sampling) to eliminate this effect, here we present the accuracy across unadulterated samples.

One interesting (albeit, unfortunate) result was that De Novo two-pass alignment completely missed between 2% and 9% of splice junctions per sample (Figure 3.2C). These splice junctions were also completely missed by De Novo one-pass alignment (A549 example: read depth ratio 0, Figure 3.2A), meaning they were not lost in the second alignment pass, but we were still curious what might introduce difficulty in aligning to these splice junctions. First, we recognized that most missed splice junctions were low expressed, covered by only a few reads in the standard quantification (see

Figures B.1-B.12, B panels), but some missed splice junctions did have high expected quantification. We therefore sorted the splice junctions by their standard quantification in descending order, and found a strong enrichment of AT/AC, GC/AG, and non-canonical splice site motifs at the top of the list (Figure B.13). In particular, annotated AT/AC and GC/AG splice-site containing splice junctions were most likely to be missed, followed by non-canonical splice sites. This result makes qualitative sense, given that STAR penalizes splice junctions with non-canonical splice sites, but the magnitude of the effect was greater than we anticipated. We further note that, in practice, non-canonical annotated splice junctions can still be readily aligned to by use of annotation and aren't damaged by two-pass alignment alone, as evidenced by the Annotation 2-pass distribution in Figure 3.2A, which missed very little (read depth ratio 0).

### 3.4.2 Why Two-Pass Alignment Works

Since we observed quantification differences between one-pass alignment and two-pass alignment, we next investigated what effect might convey those differences. We hypothesized that improved quantification was enabled by improved ability to align reads by shorter over-hanging lengths, and were particularly interested in the effective minimum spanning length for each alignment approach, expecting to see the parameterized values of 3nt and 8nt per read for annotated and unannotated splice junctions (unannotated splice junctions also required a single read span by 12nt). To test this, we extracted splice junction spanning lengths for every spliced read in two representative samples, TCGA-50-5933_N (48nt), and A549 (101nt). Spliced read span length distributions are plotted as histograms for the two samples (Figure 3.3), overlaid for both single-pass and two-pass alignment.

Consistent with parameter selection, in both samples two-pass alignment was capable of aligning reads by at least three nucleotides (to previously discovered splice junctions), and one-pass alignment was capable of aligning reads by at least twelve nucleotides (to novel splice junctions), with some ability to align reads by eight to eleven nucleotides (these reads were present on splice junctions supported by at least one other read spanning by at least twelve nucleotides). We note that in Figure 3.3A, the number of reads spanning splice junctions by the longest amount (24nt) is approximately half other counts because the read length (48nt) is an even number; there are two ways for a read to span by 23nt (23-25 and 25-23), but only one way for a read to span by 24nt, and we did not double count them. The relatively flat distributions demonstrate two-pass alignment possesses little bias for longer or shorter reads.

Critically, while both the 48nt and 101nt libraries demonstrated the same differences in ability to align reads by short spanning lengths, this difference represented a much larger fraction of all spanning lengths in the 48nt library. In other words, the additional ability to align reads by three to eleven nucleotides enables alignment of a greater percentage of reads when the total read length is shorter. As further exploration of this idea, we derived a simple mathematical model to predict how many more reads can be aligned to splice junctions once they are annotated (Formula 3.3).

Formula 3.3) $R = \frac{L - (M_A * 2)}{L - (M_N * 2)}$

Where L is the read length of the sequencing library, MA and MN are the minimum nucleotide spanning lengths required by the aligner for annotated and novel splice junctions, respectively, and R is the expected read depth ratio. Using this formula, the predicted ratio of alignable positions for a 48nt library, with minimum novel and annotated spanning lengths of 12nt and 3nt is therefore: (48 - 3*2) / (48 - 12*2) = 42 / 24 = 1.75, and for a 101nt library using the same lengths is: (101 - 3*2) / (101 - 12*2) = 95 / 77 = 1.23. Across the twelve samples in our analysis, these expected ratios matched the increase in read depth provided by two-pass alignment very well (Table 3.1). We therefore conclude that improved ability to align reads by short spanning lengths is sufficient to explain the quantification benefit of two-pass alignment.

### 3.4.3 Alignment Error Mitigation

While our testing supported two-pass alignment as a sensitive means to quantify novel splice junctions, we carefully considered an anticipated drawback of two-pass alignment. Summarized, this concern is that misaligned reads in the first pass could seed the second pass with false splice junctions, which in turn could distract more reads from their correct contexts, and amplify quantification of these false splice junctions. Because singleton misaligned reads are easily disregarded with cutoffs in downstream analysis, our primary concern was false splice junction quantification, rather than false splice junction discovery. While we appreciated the accuracy and relevance of this concern, even mis-alignment requires stringent sequence matching, and were therefore unclear on exactly how and why these errors might occur.

In place of a read simulation experiment, which would have been difficult to correctly model read distributions for, we instead opted to profile errors within real data, following the rationale that detecting and eliminating these errors was preferable to just knowing they existed. We therefore investigated the mitochondrial genome, which contains 37 known, single-transcript genes, none of which are spliced. Barring population structural variants and relatively rare transcriptional errors, any strongly supported splice junctions on the mitochondrial genome must result from alignment errors.

We began by comparing read depths between the first and second pass alignment, as major differences likely reflect splice junction amplification errors, and paid particular attention to splice junctions where read depth increased five-fold or more between the first and second alignment, as others were likely to be eliminated by minimum read depth thresholds in downstream analysis.  Through manual investigation of read coverage data in the Integrated Genome Browser, we identified three factors which seemed to typify supposed splice junctions with large depth changes. These were: a high sequence read depth of the unspliced context, a high percentage of reads spanning the splice junction by less than the exact sequence identity between the spliced and unspliced contexts, and finally a high percentage of spliced reads spanning the splice junction by very short overhang lengths, typically less than or equal to twelve nucleotides (likely because twelve delineates reads which require and do not require annotation). A genome browser example of a representative alignment error is provided in Figure 3.4A.

We wrote specialized code to extract these three features for every splice junction from the raw data, and plotted per-junction statistics vs. the change in read depth between the two alignments, using "splice junctions" on the mitochondrial genome as true positive errors [Figures B.14-B.17]. While each attribute was positively correlated with erroneously high quantification, unspliced read depth was neither necessary nor sufficient for alignment errors, and sequence identity was sufficient but not necessary. We speculate that the sequence identity check may have failed due either to polymorphisms, or sequence identity between two spliced contexts.  The percentage of reads spanning by twelve nucleotides or less appeared to perform very well in identifying alignment errors, and appeared not to typify annotated splice junctions. Encouraged by this exploratory result, we tested its utility as an alignment error classifier on a representative sample, the A549 cell line.

As a null hypothesis for a 101nt read, on average 12*2 / 101 = 24% of reads should span by twelve nucleotides or less, so we selected 80% as a reasonable cutoff to indicate large deviation from the average. We then calculated sensitivity using known alignment errors, mitochondrial splice junctions which were quantified at least five-fold higher in the second pass than the first pass, and calculated specificity using known true splice junctions, annotated autosomal splice junctions which were not quantified at least five-fold higher in the second pass than the first pass.  Scatterplots and histograms resulting from this analysis are depicted in Figure 3.4B.

This simple classifier performed very well: of 271 mitochondrial splice junctions with five-fold higher coverage in the second pass, 253 had 80% or more of the reads span by less than 12nt (93.4% sensitivity); and of 154,307 annotated splice junctions which had less than five-fold higher coverage in the second pass, only 288 had 80% or more of the reads span by less than 12nt (99.8% specificity). Individual splice junctions are shown as scatterplots in Figure 3.4B, with mitochondrial "splice junctions" depicted in red. Histograms in Figures 3.4B support the scatterplots in demonstrating that more unannotated splice junctions experience alignment errors than annotated splice junctions, and the efficacy of the classifier.

To explain the phenomenon of these alignment artifacts, we speculate that real gapped reads, which we attribute to rare transcriptional events or ligation artifacts of sequence library preparation, provide false positive splice junctions to the second alignment pass. If the normal transcriptional context (unspliced or spliced) has identical sequence to the false splice junction, depending on scoring parameters the aligner could assign reads to the false splice junction with equal likelihood. Worse, if a single-nucleotide polymorphism exists in the normal transcriptional context, i.e., that the individual's genome does not match the human reference genome at one position, potentially all reads could get assigned to the false splice junction. If the transcript is highly expressed (e.g.: mitochondrial genes), many reads may be misaligned, and the expression estimation between the first and second alignment passes increases dramatically. However, a common facet of these misaligned reads is that they all span the splice junction by less than the length of true sequence identity. While we found determining the normal transcriptional context's sequence difficult, measuring the effect of misalignment (short spanning lengths), rather than the cause, proved very effective.

### 3.5 Conclusion

A defining characteristic of RNA-seq is its ability to discover and quantify novel sequences. To maximize this ability in the context of splice junction analysis, we thoroughly investigated two-pass alignment.

Consistent with parameter selection, we found that two-pass alignment enables sequence reads to span novel splice junctions by fewer nucleotides, which confers greater read depth over those splice junctions, and this effect disproportionately benefits samples with shorter reads. The expected read depth benefit from enabling shorter spanning lengths closely matched observed read depth increases across a variety of RNA-seq samples, and affected nearly every splice junction per sample. Further, by aligning significantly more reads to splice junctions, two-pass alignment provides significantly more accurate quantification of novel splice junctions than one-pass alignment, as evidenced by its tight concordance with gene annotation-driven alignment. This quantification is mostly very good, but non-canonical novel splice junctions are likely to be missed using default parameters. Finally, while we observe splice junctions which are likely alignment errors, we demonstrate that these are simple to identify using the distribution of reads spanning the splice junction by short lengths, here less than or equal to twelve nucleotides. In our experience, alignment errors are consistent between samples, underscoring both their sequence-driven nature, and their ease of identification. A similar alignment error classification method is utilized by FineSplice, which also works by modeling splice junction spanning length distributions, and would likely improve on the simple classifier presented here if extended from Tophat results to STAR results.[173]

Beyond these practical benefits, in the context of cancer transcriptomics we anticipate great value in comparing known and novel splice junctions on equal footing, which is enabled only by two-pass alignment. While two-pass alignment particularly benefits shorter read sequences, and technology advances continue to extend read length, much 50nt-100nt read data already exists and stands to benefit from more sensitive reanalysis. In addition to increased sensitivity for rare and low-expressed splice variants, applications include resolving isoform structures of novel non-coding RNAs and genes in non-human organisms, and supplying more confident novel isoforms for proteogenomic database searching. Successful application here to Arabidopsis RNA-seq data bolsters our optimism that the sequence-driven nature of two-pass alignment would benefit analysis of other organisms as well. While we used STAR here, any sequence alignment algorithm which permits scoring differences between annotated and unannotated splice junctions could be run in a two-pass alignment configuration, and should expect to see similar novel splice junction performance improvements.

In conclusion, two-pass alignment significantly improves quantification of novel splice junctions, and we recommend its use in studies concerned with novel splice junction discovery and quantification.

### 3.5.1 Acknowledgements

We thank Marcin Cieślik, Hui Jiang, and Ryan Mills for critical discussion of the research. We thank Xuhong Cao and Terrence Barrette for assistance with informatics resources, and Christine Betts and Karen Giles for assistance with manuscript submission. Finally, we thank The Cancer Genome Atlas (TCGA), for sequencing and hosting the data for some of the tissue samples, and Illumina Inc. for sequencing and hosting the data for the Universal Human Reference RNA.

### 3.5.2 Author Contributions

B.A.V. produced the initial concept, A.I.N. and B.A.V. designed the study, and B.A.V. performed all experiments, analysis and writing. A.I.N., A.M.C., S.M.D., and S.S. contributed significant discussion and direction on experiments and writing. A.M.C. provided necessary computational resources.

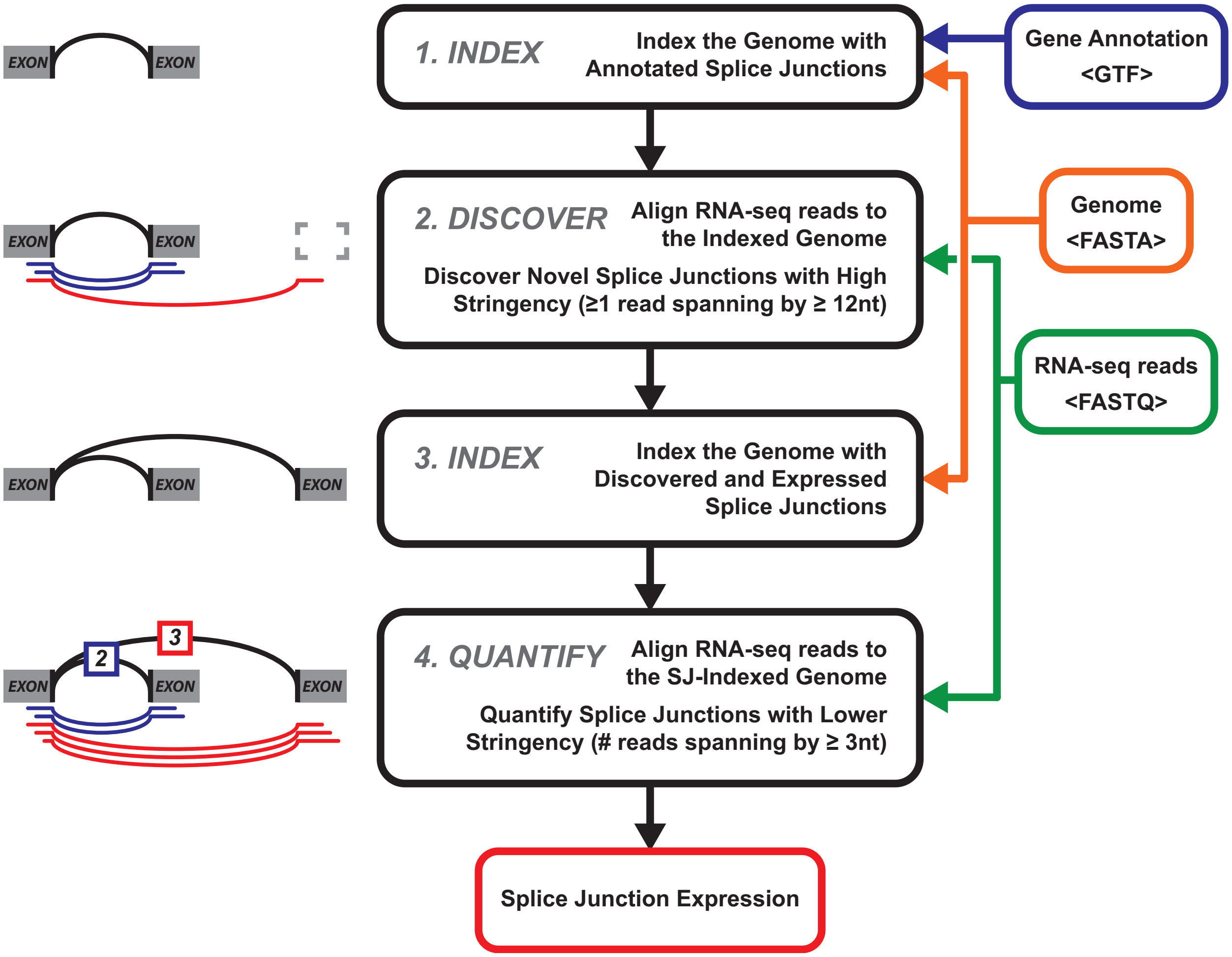

EXON
EXON
1. INDEX
Index the Genome with Annotated Splice Junctions
Gene Annotation <GTF>
2. DISCOVER
Align RNA-seq reads to the Indexed Genome
Discover Novel Splice Junctions with High Stringency (≥1 read spanning by ≥ 12nt)
Genome <FASTA>
RNA-seq reads <FASTQ>
3. INDEX
Index the Genome with Discovered and Expressed Splice Junctions
2
3
4. QUANTIFY
Align RNA-seq reads to the SJ-Indexed Genome
Quantify Splice Junctions with Lower Stringency (# reads spanning by ≥ 3nt)
Splice Junction Expression

**Figure 3.1** **Two-Pass Alignment Flowchart.** Center and right, stepwise progression of two-pass alignment. First, the genome is indexed with gene annotation, here Gencode-Basic. Next, novel splice junctions are discovered from RNA sequencing data at a relatively high stringency (12nt minimum spanning length). Third, these discovered splice junctions, and expressed annotated splice junctions are used to re-index the genome. Finally, alignment is performed a second time, quantifying novel and annotated splice junctions using the same, relatively lower stringency (3nt minimum spanning length), producing splice junction expression. Input files and their associated file formats are shown on the right. Left, pictorial representation of individual steps, for an individual novel splice junction. Exons are illustrated in gray, indexed splice junctions in black, individual sequence reads supporting a known and a novel splice junction in blue and red, and read counts (splice junction quantification) in blue and red boxes. Alignment parameters are provided in the methods, and Table B.2.

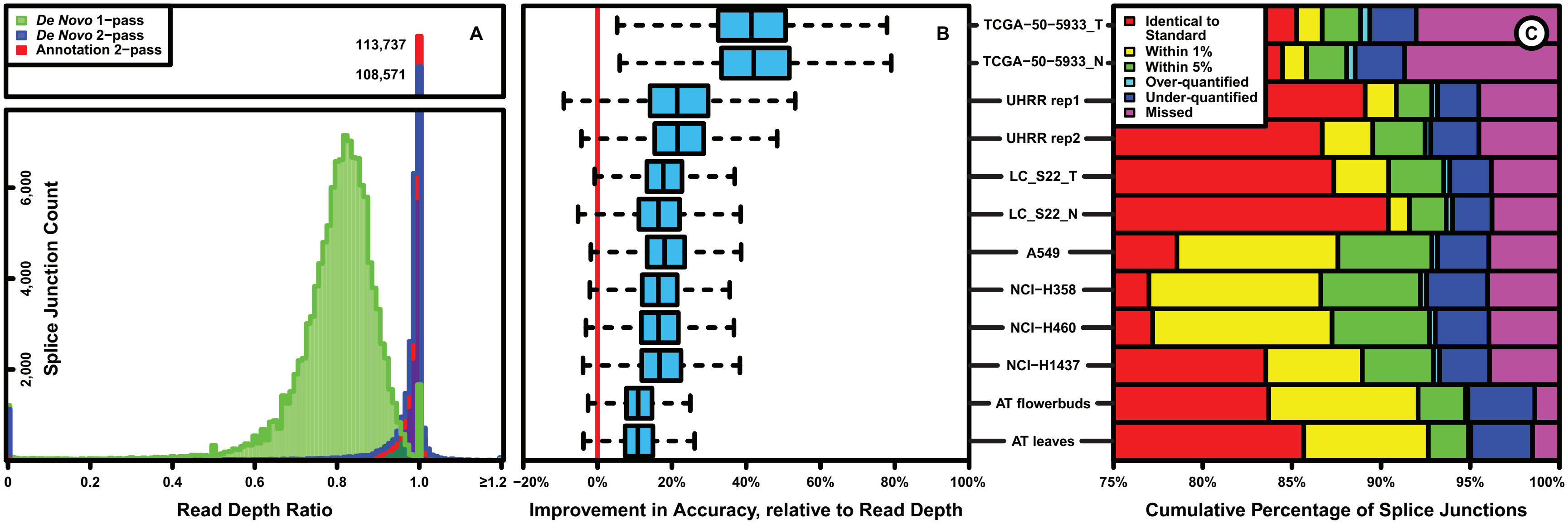


**Figure 3.2 Quantification Accuracy of Two-Pass Alignment.** A) For the A549 cell line, splice junction quantification from three alignment approaches was compared to Annotation 1-pass quantification of annotated splice junctions, testing their ability to recapitulate standard quantification (units: uniquely aligned read counts). Ratios of each approach vs. the standard across all splice junctions are shown as overlaid histograms. B) Across twelve representative RNA-seq samples, across all splice junctions per sample, quantification error was measured for 1-pass and 2-pass De Novo alignment. The extent to which two-pass alignment improved on one-pass alignment is plotted as Tukey boxplots. All samples showed statistically significant deviation from the null hypothesis of zero improvement. C) Absolute quantification accuracy of two-pass alignment was measured by comparing it to one-pass alignment with annotation, and splice junctions within six accuracy thresholds were counted, across twelve representative RNA-seq samples. The samples are described in detail in Table 3.1 and Table B.1. Panels A and B used a cutoff of at least 10 reads in the Annotation 1-pass alignment, and panel C used a cutoff of at least 1 read in the Annotation 1-pass alignment.

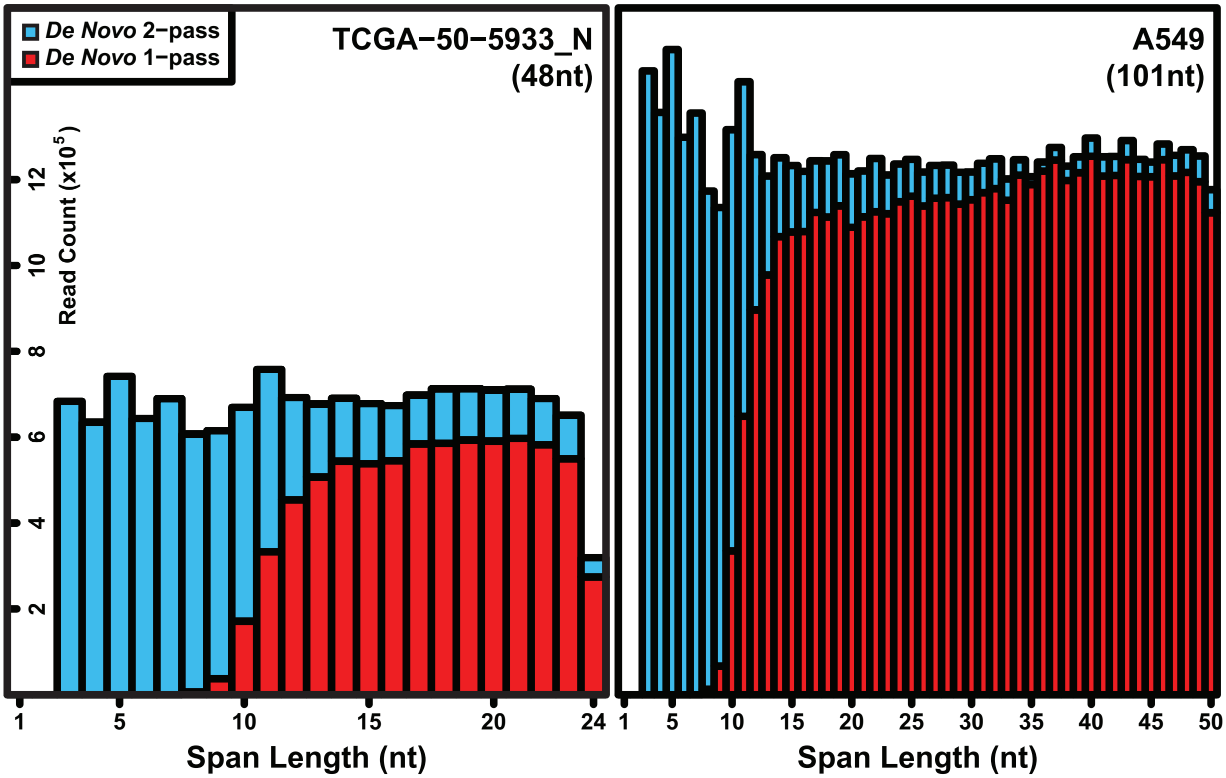


**Figure 3.3 Spliced Read Spanning Length Distributions.** For two samples, TCGA-50-5933_N and A549, all spliced reads were extracted from their one-pass and two-pass De Novo alignment results, and the number of nucleotides those reads spanned splice junctions by were counted. Histograms of the number of reads spanning by each length are depicted overlaid for the two alignment approaches, for the two samples. No cutoffs were used.

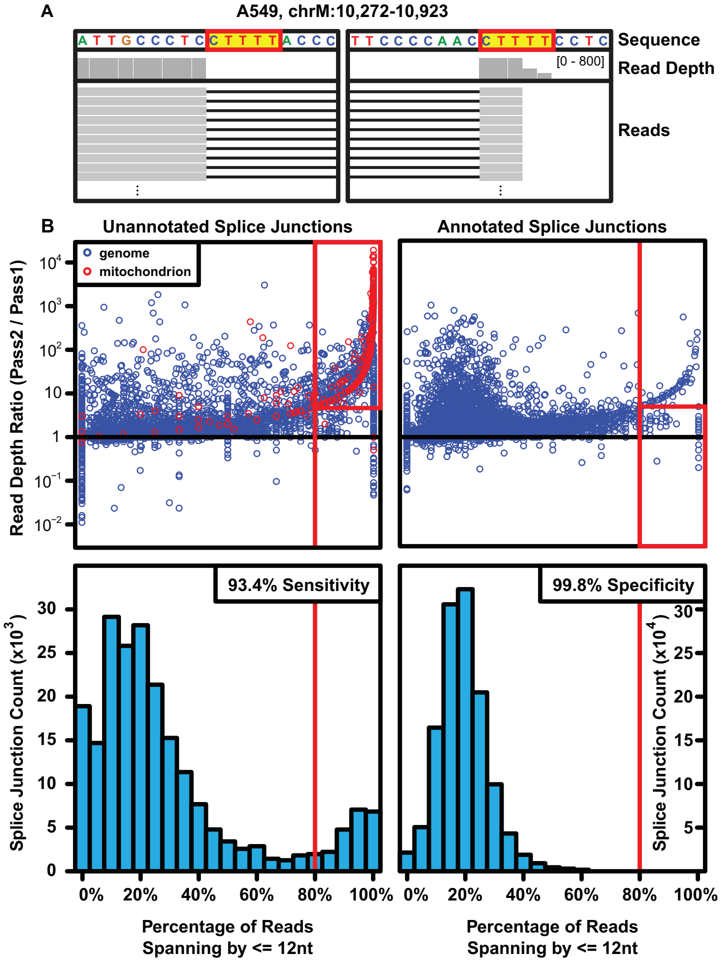


**Figure 3.4 Alignment Error Classification.** A) A representative alignment error from A549 is depicted as an Integrated Genome Viewer screenshot, showing sequence (with identity highlighted in yellow), read depth of coverage, and individual reads. B) Across all unannotated (left) and annotated (right) splice junctions, the percentage of reads spanning by less than or equal to twelve nucleotides was counted. These percentages are plotted vs. the change in read depth between one-pass and two-pass De Novo alignment, which when large indicates possible alignment errors, as scatterplots (top), and as histograms (bottom), with false-positive mitochondrial "splice junctions" identified in red. Using a cutoff of 80% (vertical red lines), 93.4% sensitivity for true-positive alignment errors was found (mitochondrial "splice junctions" with five-fold or higher change in read depth, red boxed area), while only 0.2% of true-negative splice junctions were flagged, yielding 99.8% specificity (annotated splice junctions with less than five-fold change in read depth, red boxed area). Panel B used a cutoff of at least 1 read in De Novo 1-pass alignment for the scatterplots, and at least 10 reads in De Novo 2-pass alignment for the histograms (to eliminate visual distraction at small even ratios, e.g. 1/2, 2/4), while the sensitivity and specificity analysis used no read depth cutoffs.

## Table 3.1 Sample Descriptions and Summary Statistics

| | | | | Splice | | |
|---|---|---|---|---|---|---|
| | | Read Pairs | Read | Junctions | Median Read | Expected Read |
| Sample | Description | (millions) | Length | Improved | Depth Ratio | Depth Ratio |
| TCGA-50-5933_T | Lung Adenocarcinoma Tissue | 48 | 48nt | 99% | 1.68x | 1.75x |
| TCGA-50-5933_N | Lung Normal Tissue | 52 | | 98% | 1.71x | |
| UHRR_rep1 | Reference RNA | 83 | 75nt | 94% | 1.25x | 1.35x |
| UHRR_rep2 | | 85 | | 97% | 1.26x | |
| LC_S22_T | Lung Adenocarcinoma Tissue | 52 | 101nt | 98% | 1.20x | 1.23x |
| LC_S22_N | Lung Normal Tissue | 35 | | 96% | 1.18x | |
| A549 | Lung Cancer Cell Lines | 92 | | 97% | 1.21x | |
| NCI-H358 | | 109 | | 97% | 1.19x | |
| NCI-H460 | | 105 | | 97% | 1.19x | |
| NCI-H1437 | | 76 | | 97% | 1.19x | |
| AT_flowerbuds | Arabidopsis Flower Buds | 192 | | 97% | 1.12x | |
| AT_leaves | Arabidopsis Leaves | 202 | | 95% | 1.12x | |

Twelve publicly-available RNA-seq samples selected to reflect a variety of short read sequencing data types. "Splice Junctions Improved" indicates the percentage of all splice junctions in each sample which were more accurately quantified by two-pass alignment than one-pass alignment. "Median Read Depth Ratio" was calculated as the median across splice junctions, of the fold change in read depth between De Novo 2-pass alignment and De Novo 1-pass alignment. Finally, "Expected Read Depth Ratio" lists the benefit to be expected solely by improved ability to align reads by shorter spanning lengths. No cutoffs were used.

# CHAPTER 4

# Cohort-scale Analysis of Transcript Variation from RNA-seq

*The work presented in this chapter is not currently being pursued as a standalone manuscript.*

## 4.1 Introduction

### 4.1.1 Cancer Research Scopes

The current generation of cancer research exercises an incredibly diverse array of methods to study cancer biology at all levels. Epidemiological approaches study broad trends at the population level, and are best positioned to uncover environmental carcinogens and genetic associations, particularly genome-wide association studies.[24,174] Tissue profiling approaches study molecular aberrations, usually in DNA, RNA, and protein, but sometimes metabolites and other molecules as well, in order to identify common drivers and molecular symptoms of cancer.[30] Patient-derived xenografts, by which human cancer tissue is grown in a host organism (usually mice), are best positioned to study physiological effects of treatment, particularly efficacy and toxicology.[175] On a related note, a relatively new approach, 3D tissue culture ("organoids") can be used to study human cancer tissues in a laboratory setting, but outside of a host organism.[176] Also using model organisms, but very differently, genetically engineered mice are used to study physiological disease progression, through how mice develop cancer themselves.[177] Cancer cells taken from human patients can be immortalized using a variety of approaches, and owing to their availability these cell lines are extremely popular for studies of gene function in a cellular context, be it localization, physiological impact (e.g.: growth, invasion), or molecular impact (e.g.: gene expression).[178] Finally, there is a vast armamentarium of methods to characterize individual genes, by their chemical properties, function, and interaction with other genes which are too numerous to list.

While these approaches all have strengths, the approach that most directly aims to discover molecular drivers of cancer is tissue profiling. A common shortcoming of the model-based approaches is that they do not or cannot address the heterogeneity between human cancers, at least in part because in many cancers this heterogeneity is still incompletely understood. This specific reasoning is a driving force behind large tissue profiling studies, such as The Cancer Genome Atlas, which aim to better understand cancer through its heterogeneity in many cancer patients.[179]

### 4.1.2 Cancer Tissue Profiling Challenges

That said, tissue profiling projects have many issues which must be overcome before they are able to arrive at biologically meaningful results.  First, tissue samples and particularly RNA degrade quickly.  Moreover, to reach the numbers of samples necessary to survey heterogeneity, projects often survey samples from many institutions, whose sample processing pipelines frequently differ. In short, not only must sample degradation be overcome, but *variable* sample degradation must be overcome ("batch effects").

Second, tumor tissue samples are variable in the relative proportions of cell types present, a phenomenon often referred to as "sample mixture," "sample admixture," or simply, "tumor content." This is to say, the cell type of interest is watered down by the presence of other cells in the sample. In the example of prostate cancer, cancerous prostate luminal epithelial cells are the target; everything else, smooth muscle, fibroblasts, non-cancerous epithelium, immune cells, etc., is the background. While signaling from the other cell types is appreciated to contribute to cancer progression, and that appreciation seems likely to grow in the future, currently the primary interest is still molecular characterization of the cancer cells. Similar to sample quality, the analyst must also appreciate that tumor content is variable between tissue samples.

On a related note, multiple different cancer cell lineages may be present in a single tissue sample ("intratumor heterogeneity"), which is interesting and presents its own opportunities and challenges for research, but usually in tissue profiling the researcher focuses on the most abundant lineage and therefore this is less critical to address than the other issues presented here.

Finally, extensive differences exist in genotype and phenotype across cancer samples, this is to say, they are diverse in their driver and passenger aberrations ("intertumor heterogeneity"). Intertumor heterogeneity is really the main reason to study cohorts of tissues (see the introduction to this chapter), but regardless it poses computational challenges, and must be specifically considered in tissue profiling projects.

### 4.1.3 Addressing Cancer Tissue Profiling Challenges

Genomic characterization of cancer tissues, specifically genetic aberrations like point mutations, insertions and deletions, and copy number aberrations, is relatively straightforward to perform around sample challenges. DNA is more stable than RNA, and DNA degradation, while non-random, mostly manifests as loss of coverage rather than mutated bases. Variable DNA degradation can therefore be addressed in analysis as missing values. Next, DNA is present in cells in an integer number of copies, mostly two (and nearly always two in normal tissue, excepting mitotic cells), and most tumor tissue samples are dominated by a single cancer lineage ("clone"), so establishing zygosity is usually tractable. Further, normal cells are mostly not mutated, with the exception of germline polymorphisms which are also present in normal tissue, which is usually also profiled for differential analysis. Therefore, tumor content can usually be circumvented to identify mutations present in the cancer's DNA. Last, intertumor heterogeneity is definitely not non-trivial for genomic variants, but owing to the binary nature of mutations (present or absent), establishing recurrence at the nucleotide, gene, or pathway level is achievable. On the back of the strength of the solutions to these issues, DNA has largely taken center-stage in large tissue profiling studies.

In contrast, RNA degrades quickly, and in samples which have been enriched for messenger RNA by polyadenylation capture (most current samples), this degradation manifests as bias toward the 3' end of the transcript. The extent of this bias varies, and dramatically affects the ability to detect and quantify transcript variants. Also in contrast to DNA, the number of copies of RNA varies widely from cell to cell, so disambiguating which RNA molecules in a tissue sample came from the cancer cells is a serious challenge. It is famously difficult to identify genes which are down-regulated by cancer cells, because the phenomenon of down-regulation is virtually indistinguishable from genes expressed by stromal cells which are displaced in cancer.[180] Finally, because RNA quantity is decidedly non-binary, establishing recurrence at the cohort level requires greater sophistication than for DNA. As a result of these challenges, RNA has largely taken a backseat to DNA in large tissue profiling studies, particularly in the context of integrating DNA and RNA data together in single analyses. For instance, in the TCGA prostate manuscript, RNA is handled completely separately from DNA, with its own clustering analysis for expression.[179] In the most integrative large tissue profiling studies, RNA splicing is not investigated at all, though there are many studies which focus solely on RNA splicing from RNA-seq in cancer tissues.[117,181-183]

### 4.1.4 Aims of this Analysis Pipeline

In this analysis pipeline, we set out to study RNA splicing and transcript variation from heterogeneous cancer tissue cohorts, and address or circumvent the issues raised in the introduction to this chapter. We were specifically interested in identifying differential splicing between cohorts, toward biomarkers, and outlier splicing in individual samples, driven by putative underlying genetic variants, and possibly driver events.

We made a few critical decisions in development of the analysis pipeline. First, we decided to perform tumor / normal comparison at the cohort level rather than the sample level, owing to the paucity of normal tissue RNA-seq data : most studies with tumor transcriptomes do not have per-sample matched normal transcriptomes. Second, we pursued junction expression as the primary driver of the analysis, rather than exon expression or isoform expression. Exon expression's discovery potential suffers from the need to define exon boundaries, as well as variable 3' bias across samples. Isoform expression suffers extremely from 3' bias, to the point where annotated isoforms with the same 3' UTR divide expression equally regardless of how confident the annotation is ; despite how popular isoform expression is, it has serious problems in this regard.[121] More generally, junctions are the lowest level unit which supports transcript variants, and require the fewest assumptions to handle. Third, we decided to agnostically detect differential abundance of transcript variants of any kind, including alternative transcription start sites and end sites, which are not spliceosomally mediated and therefore not "splice variants" in the precisely correct sense of the term. A second example of this is genomic deletion of an exon in cancer : since the exon is not present in the DNA, transcription of that gene faithfully reflects the loss of this exon and is not technically a "splice variant," though exon skipping and exon deletion are phenotypically indistinguishable at the RNA level. In essence, we focused on the effects of transcript variation rather than the causes. Finally, we decided to address detection of differential splicing between cohorts, and outlier splicing in individual samples, as similar but separate analyses.

## 4.2 Methods

### 4.2.1 Junction Quantification

The pipeline begins with input of fastq-formatted sequence data for individual cancer tissue samples, either uncompressed or compressed. Alignment is then performed in two passes as in Chapter 3, briefly described again here. Sequence data is aligned to the GRCh38 / hg38 revision of the human reference genome, which has been indexed with the gene annotation database gencode-basic (version 21).[107] Gencode-basic provides a high-confidence set of transcript annotations, and excludes annotations supported by rare or weak evidence, and in this pipeline serves mostly to guide precise identification of splice junction boundaries. GRCh38 is then re-indexed using junctions discovered in the first alignment pass, and alignment is performed a second time to this newly-indexed genome. Two-pass alignment serves to facilitate alignment to novel junctions, and therefore their quantification - it is analogous to indel realignment in exome analysis or fusion junction realignment in gene fusion calling, and could accurately be renamed "intron realignment." See Chapter 3, and Figure 3.1 for further alignment details.[164] Here, "annotated" and "known" junctions refer to the set of junctions present in the gene annotation database we used, and "unannotated" and "novel" refer to the set of junctions absent from that database. Many novel junctions described here have likely been detected by researchers before, but lacked the significance to be included in gene annotation databases, meaning they have to be "re-discovered" in subsequent analysis. Finally, the novelty we seek is in the form of their cancer-association, which is separate from their novelty with respect to annotation databases.

Splice junction quantification is then supplemented with unspliced read depth over splice junction edges as follows. First, all samples in the cohort under consideration are aggregated, and a unique list of splice junction edges is generated. Next, read alignment data is trimmed by the number of spanning nucleotides desired, in this case three nucleotides to match the minimum spanning nucleotides for a splice junction. Last, read depth is calculated for all samples in the cohort for all splice junction positions detected in any sample. The result is the ability to query the number of reads supporting an unspliced transcript at that position, which may be caused by intron retention, or an alternative splice site. Quantifying unspliced depth is, as far as we are aware, a novel innovation in this project, to extend the potential of splice junction quantification to any arbitrary type of transcript variant. See Figure 4.1 for the full transcript variant "bestiary" detectable using these methods. The combination of splice junctions and "unspliced junctions" is capable of detecting any transcript variant, except for UTR extensions (characterized by a lack of coverage, which is considerably harder to analyze around biases), and internal tandem duplications of exons (which spliced aligners are incapable of aligning to without prior anticipation).

Finally, once splice junctions and "unspliced junctions" have been quantified, the entire cohort of samples is merged together into a [junction] x [sample] matrix of read depths, which is then split by chromosome to facilitate parallel computing in subsequent steps, and sorted in forward and reverse order to facilitate calling of differential splicing in both forward and reverse orientation on the genome without significant memory use.

### 4.2.2 Alternative Splice Junction Usage Analysis

Following the merger of the sample cohort, alternative junction usage analysis is performed. The code is supplied with four lists of samples present in the merged cohort : "control" samples, in most cases normal tissue, which is to be used as the denominator in the analysis ; "test" samples, in most cases tumor tissue, which is to be used at the numerator in the analysis ; "validation" samples, in most cases cell lines, which are not used in calling but are carried through the analysis and presented alongside the samples used in calling ; and "secondary control" samples, so far mostly cultured normal epithelium, which are not used in calling, but have their fractions recorded for subsequent bioinformatic analysis. The beauty of structuring the analysis in this way is that it allows for arbitrary comparison of test and control sample cohorts, where tumor vs. normal is the most obvious, but other nuanced comparisons such as non-aggressive vs. aggressive cancer are equally valid analytically, and permits calling of variants associated with progression or other clinical variables. The gene set enrichment analysis (GSEA) software follows the same case/control framework and served as inspiration for this approach.[184] The validation cohort aggregated for the projects described here is mostly comprised of cell lines, from the Cancer Cell Line Encyclopedia and from Genentech.[167,185]

The calling code works by considering a nucleotide position in the genome, and for all of its detected splice junction partners and the unspliced junction at that position, asks to what extent the test samples deviate from the distribution the control samples establish. More precisely, it considers a specific splice junction and compares that junction's abundance to the sum of other junctions, such that a strong call means a significant change in fractional abundance of a splice junction between the test and control samples (usually framed in the "up in cancer" direction for clarity). Part of the rationale for comparing junctions in this way is that it matches a junction to junctions at the same nucleotide position, which is either exactly the same distance from the polyA tail of the transcript, or an exon's length away, and therefore junction based variant fractions are relatively unaffected by 3' bias. 3' bias in this analysis manifests as a loss of coverage rather than a directional bias, and is therefore considerably easier to handle. Again, all chromosomes, and forward and reverse orientation of calling, are split and processed in parallel for speed.

Most precisely, non-negative integer read depths for each splice junction, and the sum of other junctions that share its left edge in forward orientation (and alternately right edge in reverse orientation), are computed for each of the case and control samples, and are stored as two two-column, [number of samples]-row tables. Then, a Pearson's chi-squared test is performed on the read depths from the control samples, testing the null hypothesis that the joint distribution of cell values (read depths) is the product of the marginal distributions of the rows (samples) and columns (junctions), using the appropriate number of degrees of freedom, and the resultant chi-squared statistic is retained. Simply put, a value is generated, indicating how variable the control samples are around the average variant fraction, given their individual total depth. Next, Pearson's chi-squared tests are performed on individual test samples, against a chi-squared distribution using the average variant fractions in control samples as population probabilities and the chi-squared statistic described above, to generate p-values for individual test samples. Subsequently, the p-values for all the test samples are aggregated using Fisher's method for p-value aggregation to produce a single p-value, signifying the significance of the change in splice variant fraction between the test and control samples. P-values for individual test samples are then corrected for multiple hypothesis testing (Bonferroni correction), and those above an alpha cutoff are retained as outlier calls. See Figure 5.1 for an illustration of this analysis in action.

In addition to this statistical significance, effect size is computed as the average variant fraction in test samples minus the average variant fraction in normals. This effect size is meant to complement the significance of the change with a more absolute value, and is inspired by volcano plots which contrast significance of an expression change against the fold-change to robustify against counts near zero.

Differential and outlier calling is refined by the following cutoffs. Junctions are excluded if they are unannotated and more than 20% of samples with ten or more reads spanning that junction are flagged as alignment errors (to eliminate alignment errors). Junctions are excluded if no test sample has at least ten reads spanning that junction (to eliminate low-expressed transcriptional noise). Junction sets (keyed on a nucleotide position) are excluded unless 20% of the control samples have ten or more total reads across those junctions (to confidently estimate baseline variant fractions). Then, for differential calling, the test samples must average at least one spanning read, and must meet a minimum significance threshold (to reduce low confidence calls at the end), and the significance contribution of a single sample to the aggregated p-value for a junction is capped at 10^-16. For outlier calling, outliers must have at least ten spanning reads, must have a variant fraction of 5% or greater, and must meet a significance threshold (p-value cutoff) of 0.01 (again, to reduce low confidence calls). Variant fractions for a junction are retained for clustering, if at least half of the samples are not "NA," and the junction meets a minimum standard deviation cutoff.

The end result of this analysis is a ranked list of significantly differentially spliced junctions between the test and control cohorts, a ranked list of significant outlier splice junctions in individual samples, and a table of variant fractions used later in clustering analysis.

### 4.2.3 Correlative Bias Analysis

While splicing changes might exist and be called correctly between cohorts of samples, we considered the possibility that some of the splicing changes could be driven by factors other than the difference between cancer cells and their corresponding normal progenitor cell type. Specific examples are loss of normal stroma in the tumor, sample degradation in tumor samples, and any of a number of other biases. Rather than delve into and correct these problems individually, which would be prohibitively time-consuming and difficult to the point of impossibility in some cases (*e.g.*, tumor content), we devised a simple correlative analysis to address them.

Simply, we correlated per-sample, per-junction variant fractions with quantitative metrics of sample quality. The rationale behind this approach is if a variant call is being driven by a dimension of sample bias, that variant should correlate with that bias across samples. As far as the pipeline is concerned, it takes a table with samples as rows, and an arbitrary number of quantitative quality metrics as columns, and runs Pearson correlation on each of the splice variant calls.

In application, the per-sample biases we computed or aggregated were : 3' bias, measured as the log-ratio of last to first splice junction read depths, across genes ; unspliced RNA, intended to reflect incomplete polyadenylation capture, measured as the extent of intron retention across genes ; RNA integrity number (RIN), from the Agilent Bioanalyzer, which uses the measured ratio of ribosomal subunits to estimate the extent of degradation, as a direct metric of RNA quality ; alignment rate, as a stand-in for errors in sequencing library preparation ; FPKM, to test for variants which were passengers of high expression ; tumor content, using estimates from single nucleotide variant fractions in matched exome data ; stromal expression, using a panel of established stromal genes ; androgen receptor activity, using a panel of established AR target genes ; and an additional expression signature for neuroendocrine signaling, which performed poorly and was not considered further.

To walk through an example of this method in more depth, the stromal signature came from a study from our group in which prostate stroma and epithelium were separated by laser capture microdissection and analyzed in parallel.[180] 35 genes which were significantly over-expressed in stroma compared with epithelium were identified, and we took that list from the supplement of the manuscript (Table 4.1), and calculated their expression for all of the samples presented here. Then, per-gene expression was inverse-normal transformed across samples, principal components analysis was performed, and the first principal component was taken as a shared axis of stromal signaling, resulting in a single number for "stromal-ness," per-sample. Alternate methods of aggregating expression performed similarly, including sum of z-scores of FPKM and sum of z-scores of log-FPKM. Figure 4.2 demonstrates the process of generating the signature. We applied this signature to a pair of splice variants strongly suspected to be driven by a smooth muscle contribution (in stroma), alpha actinin-1 (ACTN1) and myosin phosphatase rho-interacting protein (MPRIP), and a clear negative correlation emerges, demonstrating that the cancer-specific splice variant is most enriched in the samples in which the stroma is most absent (Figure 4.3). We can conclude from this analysis that these splice variants are likely specific to epithelium rather than epithelial cancer, or at least *more likely* than variants which do not correlate with stromal expression. Once considering all transcript variants together downstream (at the end of the analysis), the variants most strongly nominated by this correlative analysis are expected to be those most likely to derive from cell lineage differences rather than cancer specificity.  This process is essentially the same for the other biases, though their sources vary as described before.

### 4.2.4 Transcript Variant Annotation

After calling alternative splicing at the level of the splice junction, the junction switch is subsequently annotated.  First, the up-regulated junction is labeled as annotated or unannotated based on its presence in a gene annotation database (gencode version 23), and is contrasted with identical annotation for the most abundant down-regulated junction, such that the call is clearly labeled for instance as "Annotated to Unannotated," such that its novelty is clear.  Second, the edges of the junctions are clearly labeled by the most parsimonious explanation for those junctions within a single gene, in the gene annotation database.  In descending order, these are : exon edge, mid-exon region, intron, 5' of the gene (upstream), 3' of the gene (downstream), and intergenic.  For instance, an alternative transcription start site junction switch might be labeled as "exon edge / exon edge" to "exon edge / 5' of the gene."  Lastly, and most obviously, the gene from which the junction is most likely to have originated is clearly labeled.  It's worth stressing here that the analysis, until this point, uses genomic coordinates and is entirely agnostic of gene definitions, except through use of annotated junctions to assist in resolving ambiguous junction edges.  For instance, the analysis is completely capable of detecting alternative splicing of unannotated intergenic transcripts, though in those cases the impact is more difficult to determine.

Finally, the code appends expression for the gene across samples, "gene of interest" labels for the gene (*e.g.*: "kinase," "splice factor"), and variant fractions in the secondary control samples.  Joined tables are then generated for further analysis, plotting, web visualization, and download.

### 4.2.5 Presentation

Completed analysis results are presented as web browser tables via customized web code, with the analysis and annotation described above, as well as per-call scatter plots which link to genome browser data for visual inspection.  In addition, download links are provided for tabular data sheets of variant fractions in the validation cohort, to facilitate identification of cell line models for a given variant discovered in cancer tissue. See Figure 4.4 for a demonstration of the web and genome browser visualization.

## 4.3 Results

As examples of the effectiveness of this analysis, here we will highlight two examples in which we used identical or similar analysis to identify variants of interest in cancer.

### 4.3.1 MET exon skipping

Citation: Dhanasekaran SM, Balbin OA, Chen G, Nadal E, Kalyana-Sundaram S, Pan J, **Veeneman B**, Cao X, Malik R, Vats P, Wang R, Huang S, Zhong J, Jing X, Iyer M, Wu YM, Harms PW, Lin J, Reddy R, Brennan C, Palanisamy N, Chang AC, Truini A, Truini M, Robinson DR, Beer DG, Chinnaiyan AM. Transcriptome meta-analysis of lung cancer reveals recurrent aberrations in NRG1 and Hippo pathway genes. *Nat Commun.,* **5**:5893 (2014).[65]

In an analysis of 753 lung cancer samples, we detected c-MET exon-14 skipping in 15 samples, 14 of which occurred in driver- unknown samples, a 3.6% (14/386) recurrence rate in this subpopulation (Figure 4.5). Importantly, in 5 out of 15 samples, the skipping of c-MET exon-14 is probably caused by a mutation affecting the splice donor site adjacent to the amino acid position D1010 as previously described.[186] Our RNA-seq data also validated the reported c-MET exon-skipping event in the H596 cell line.

### 4.3.2 ALK alternative transcript initiation

Following exciting research published by another group in the New England Journal of Medicine, we applied an analysis pipeline similar to the pipeline described here to the detection of alternative transcript initiation of the anaplastic lymphoma kinase (ALK), to clinical research samples gathered at the University of Michigan, using targeted exon expression instead because it was more readily available than splice junction expression in the context of that cohort.[63] See Figure 4.6 for the analysis of this variant in an example case.

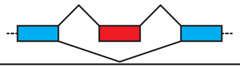
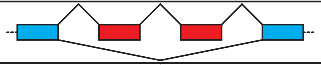
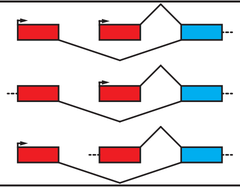
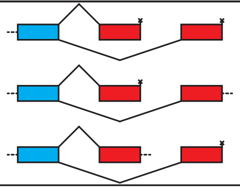
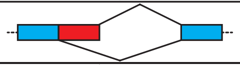
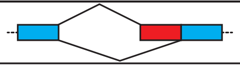
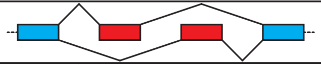
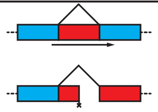
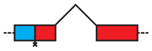
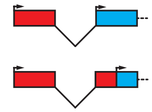
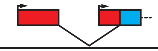
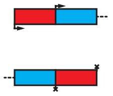
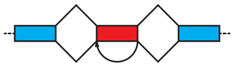
| variant | juncs:1 | juncs:2 | detectable? |
|---|---|---|---|
|  | 1:2 + 3:4 | 1:4 | yes |
|  | 1:2 + 3:4<br>+ 5:6 | 1:6 | yes |
|  | 3:4 | 1:4 | yes |
|  | 1:2 | 1:4 | yes |
|  | 2:3 | 1:3 | yes |
|  | 1:2 | 1:3 | yes |
|  | 1:2 + 3:6 | 1:4 + 5:6 | yes |
|  | 1:2 | retention | yes |
|  | 1:2 | - | no |
|  | - | 1:2 | no |
|  | retention | 1:2 | yes |
|  | - | - | no |
|  | 1:2 + 3:6 | 1:2 + 3:2<br>+ 3:4 | no |

**Figure 4.1 Transcript Variants Detectable by Junction Usage**
Here we list anticipated types of transcript variants. In order, they are : exclusion or inclusion of one or more exons ; alternative transcription start sites ; alternative transcription end sites ; alternative splice donor ; alternative splice acceptor ; cassette exon switching ; intron retention or transcription termination in an intron ; various types of 5' and 3' transcription truncation events ; and internal tandem duplication of an exon. In the diagram blue is constitutive and red is variable. Adjacent to the exon structures are the junctions present in each, numbered by position from left to right, and finally whether that junction is detectable by alternative junction usage.

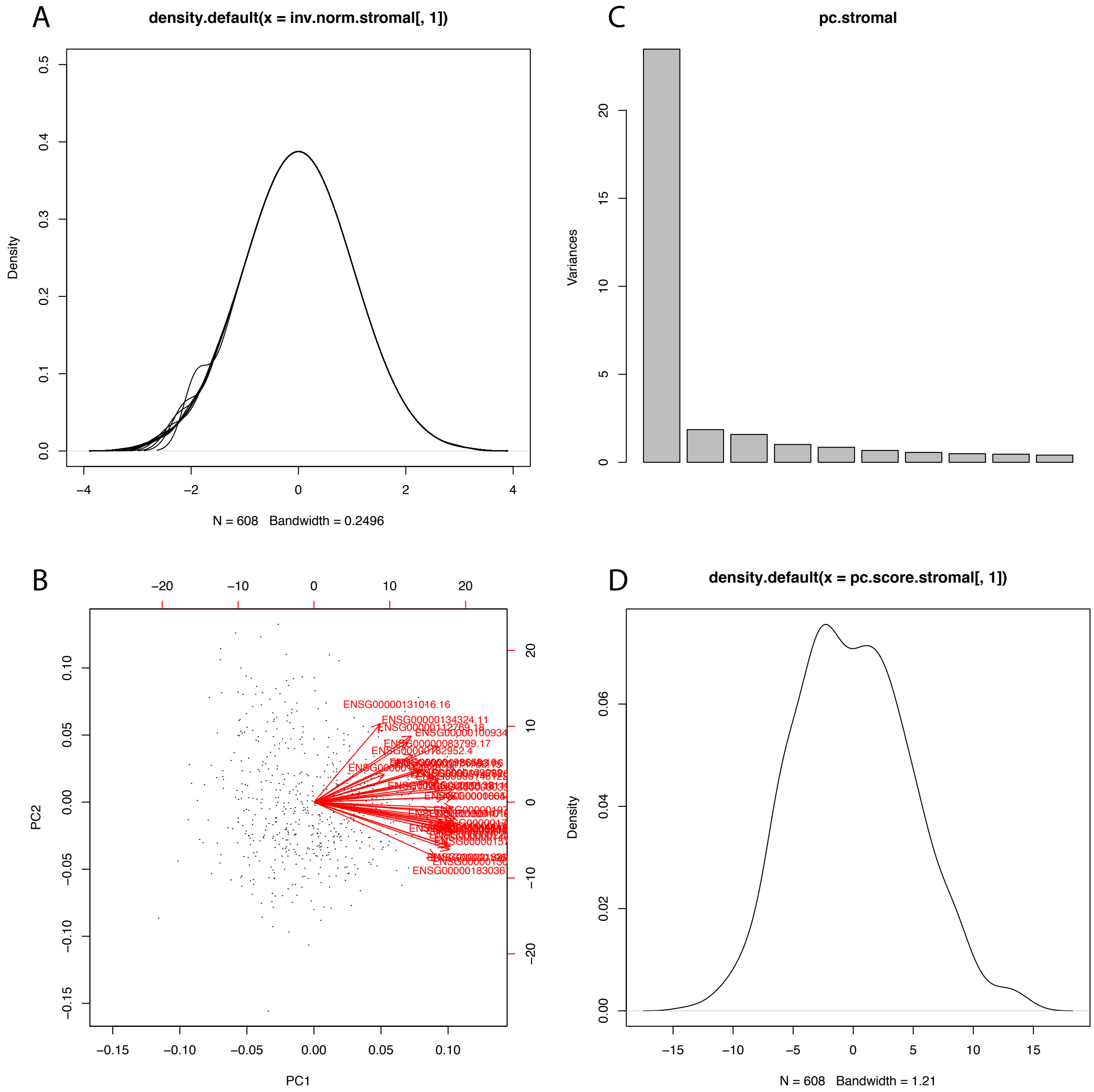


**Figure 4.2 Stromal signature definition**
**A)** Expression of 35 stromal genes in Table 4.1 was estimated for prostate cancer samples described in Chapter 5, was inverse normal transformed, and each genes plotted overlaid as kernel density plots in the upper left. **B, C)** Then, principal components analysis was performed, and the first and second principal components are plotted in the lower left, and the amount of variance explained by the first ten principal components is plotted as a barplot in the upper right. The fact that the first principal component explains most of the variance is good - it means the genes really are on a shared regulatory axis. Individual gene contributions are plotted in red. **D)** Finally, a kernel density plot is displayed of the aggregated stromal signature across samples.

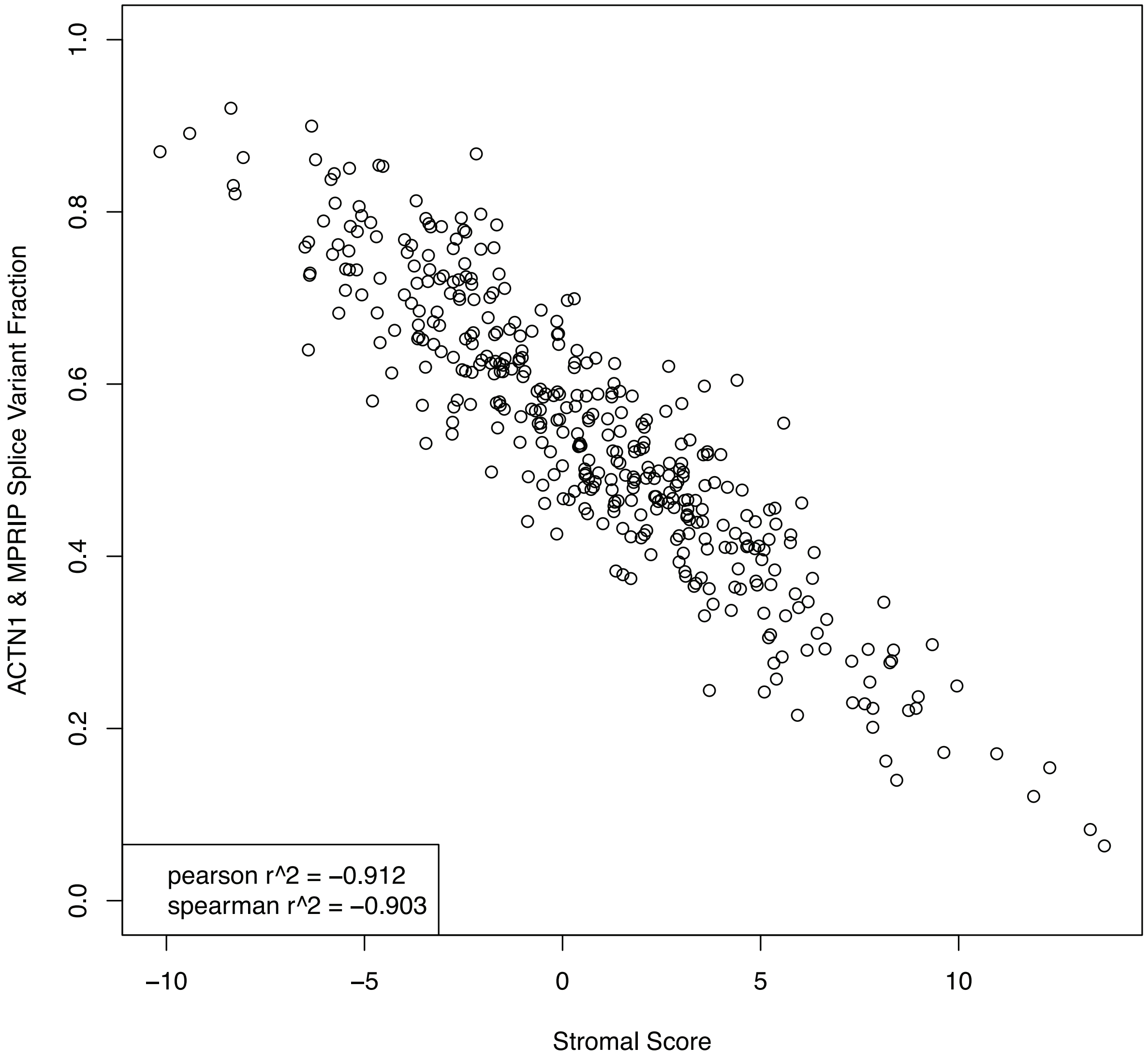
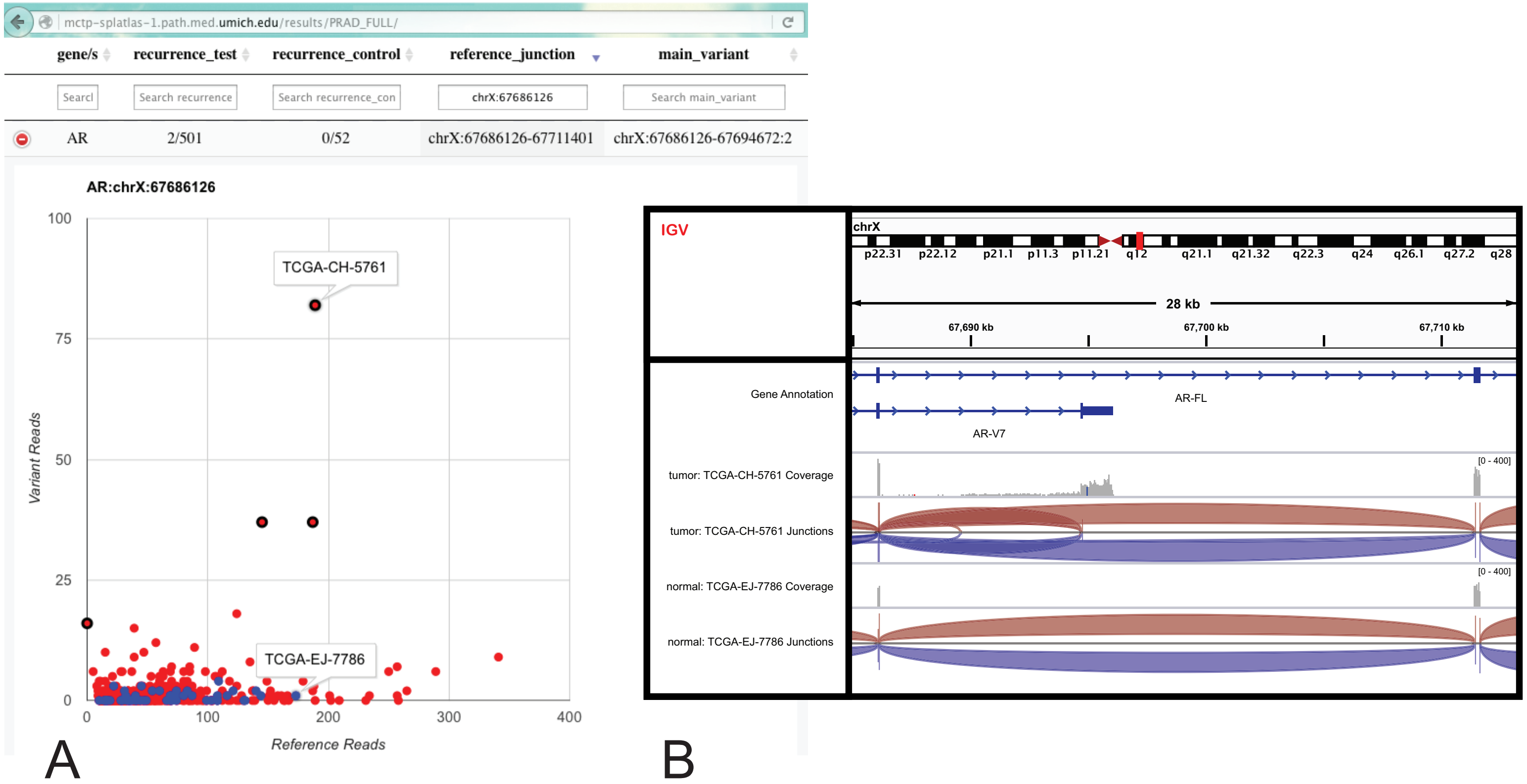


**Figure 4.3 Stromal signature application**
The stromal signature derived in Figure 4.2 is plotted against the average variant fraction from two junctions of two genes (four total), actinin-1 (ACTN1), and myosin phosphatase rho-interacting protein (MPRIP), across the cancer tissue cohort described in Chapter 5. Pearson and Spearman correlations are listed in the lower left. By demonstrating strong negative correlation, we may hypothesize a cause : that these variants are specific to epithelium vs. stroma, rather than cancer epithelium vs. normal epithelium.



mctp-splatlas-1.path.med.umich.edu/results/PRAD_FULL/

| | gene/s | recurrence_test | recurrence_control | reference_junction | main_variant |
|---|---|---|---|---|---|
| | Searcl | Search recurrence | Search recurrence_con | chrX:67686126 | Search main_variant |
| ⊖ | AR | 2/501 | 0/52 | chrX:67686126-67711401 | chrX:67686126-67694672:2 |

**Figure 4.4 Web Portal and Genome Browser Visualization**

**A)** Tabular view of individual splice variant results in a web browser. A variant of the androgen receptor (AR) is shown, plotting reads for the reference splice junction (here, exon 3 to exon 4) against the variant junction ("AR-V7"), across primary tumor and normal samples. Primary tumors are colored red, and normal tissues are colored blue in the scatterplot, outliers are highlighted with black outlines, and two samples have their names highlighted for demonstration purposes. **B)** The two samples highlighted in panel A, one splice-variant-negative normal tissue, and one splice-variant-positive tumor tissue, have their direct read evidence plotted in the integrated genomics viewer (IGV).[187,188] Read depth of coverage is plotted in gray, and junctions are labeled in both red and blue because the sequencing libraries were not strand-specific.

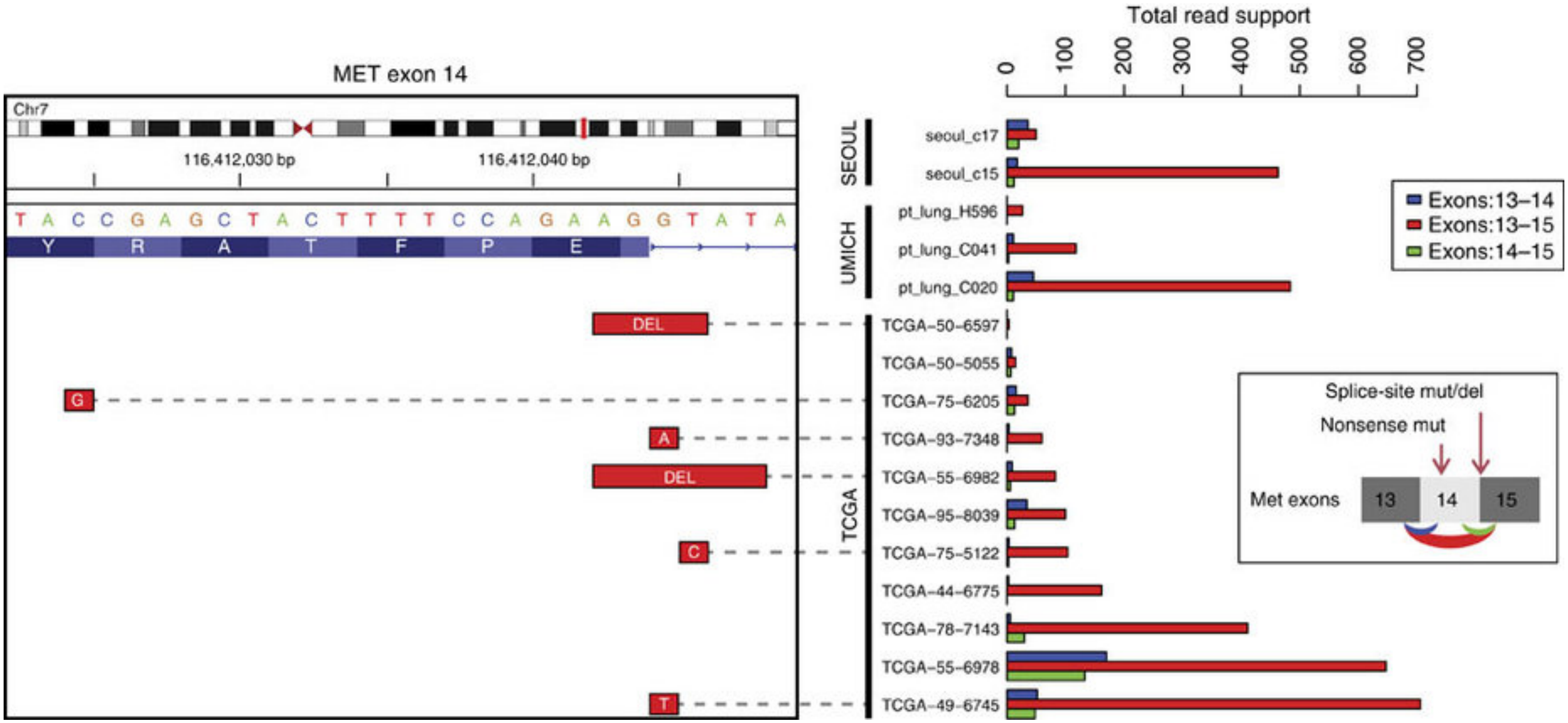


**Figure 4.5 Recurrent activating MET exon-skipping events**
Right panel: an activating MET exon-14 skipping event was observed in a total of 15 tissue samples across all three cohorts. The total reads supporting each splice variant exon13–14 (blue), exon13–15(red) and exon14–15 (green) are represented in the bar plot on the right. In 5 out of 11 TCGA samples where DNA mutation data were available, skipping of MET exon-14 was accompanied by a mutation affecting the splice donor site adjacent to position D1010 (illustrated inset on the right). In addition, one sample harbored a non-sense mutation g.chr7:116412024C>Gp.Y1003*, which accompanied exon-14 skipping. Left panel: IGV browser view of splice site deletions/mutations in the corresponding samples.

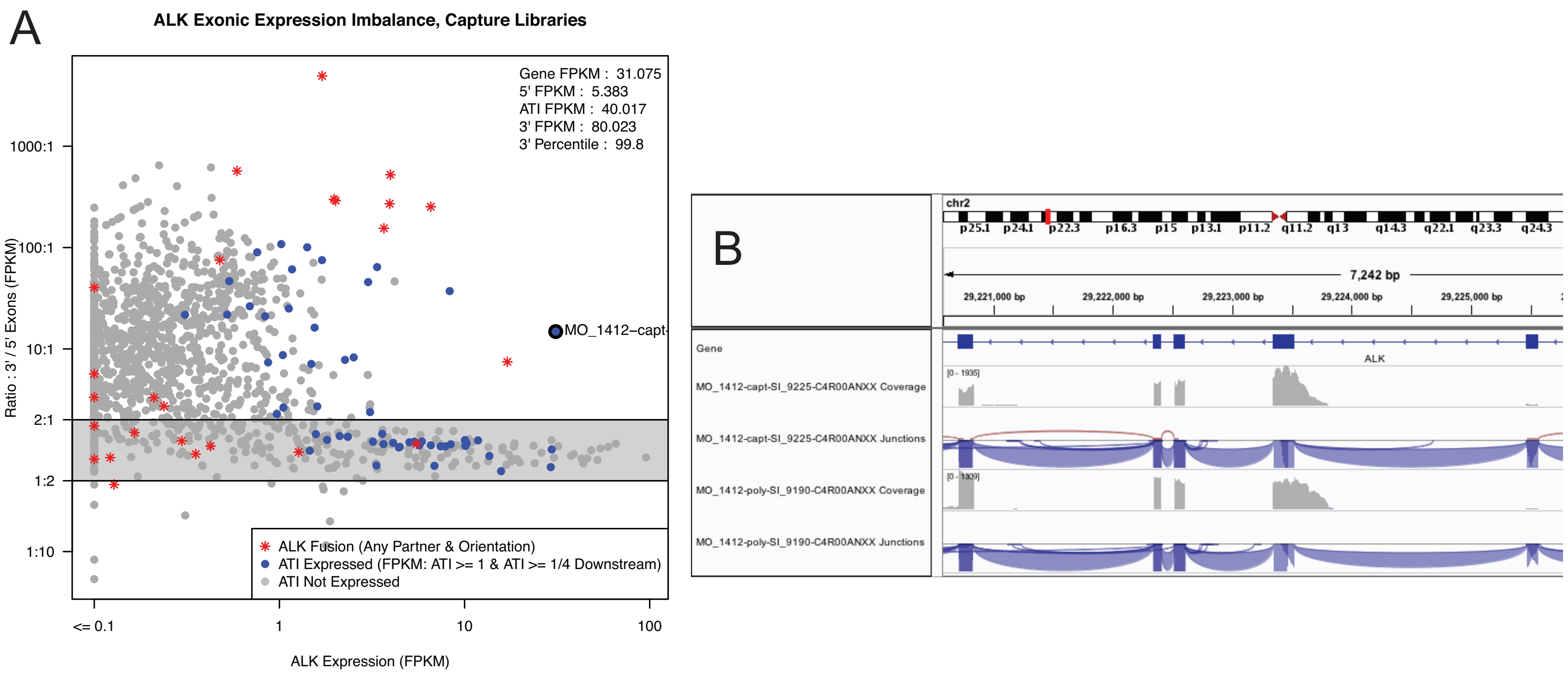


**Figure 4.6 ALK Alternative Transcript Initiation (ALK-ATI) in a Melanoma Patient**
**A)** Example analysis of a melanoma patient for ALK-ATI. Points are individual cancer tissue samples, the X-axis is the expression of ALK, and the Y-axis is the imbalance between the exons 3' (downstream) of the alternative transcript initiation site, and 5' (upstream). The samples are then color-coded by the expression of the intronic region corresponding to the ATI variant, gray for negative, blue for positive, and red for the presence of a gene fusion to ALK instead. This patient is highlighted with a black outline, and has high expression of ALK, exon imbalance approaching that of gene fusion cases, and expression of the ATI region. **B)** Genome browser visualization of the read evidence for this variant, in two variant libraries of RNA-seq, where the gray track is read depth of coverage, and the blue bands show splice junctions on the reverse strand (these libraries were strand-specific). This patient is clearly positive for this transcript variant.

| | Gene | Current | ENSG |
|---|---|---|---|
| Tomlins - Stromal Genes | ATP2B4 | ATP2B4 | ENSG00000058668.14 |
| | FER1L3 | MYOF | ENSG00000138119.16 |
| | CLU | CLU | ENSG00000120885.19 |
| | GSN | GSN | ENSG00000148180.16 |
| | MEIS2 | MEIS2 | ENSG00000134138.19 |
| | SMTN | SMTN | ENSG00000183963.18 |
| | TPM2 | TPM2 | ENSG00000198467.13 |
| | PTRF | PTRF | ENSG00000177469.12 |
| | CNN1 | CNN1 | ENSG00000130176.7 |
| | FHL1 | FHL1 | ENSG00000022267.16 |
| | MYLK | MYLK | ENSG00000065534.18 |
| | PCP4 | PCP4 | ENSG00000183036.10 |
| | ST5 | ST5 | ENSG00000166444.17 |
| | ZNF516 | ZNF516 | ENSG00000101493.10 |
| | TGFB1I1 | TGFB1I1 | ENSG00000140682.18 |
| | PMP22 | PMP22 | ENSG00000109099.13 |
| | SVIL | SVIL | ENSG00000197321.14 |
| | GAS1 | GAS1 | ENSG00000180447.6 |
| | SEC23A | SEC23A | ENSG00000100934.14 |
| | MEIS1 | MEIS1 | ENSG00000143995.19 |
| | RBPMS | RBPMS | ENSG00000157110.15 |
| | TACC1 | TACC1 | ENSG00000147526.19 |
| | PPP1R12B | PPP1R12B | ENSG00000077157.20 |
| | HMGN4 | HMGN4 | ENSG00000182952.4 |
| | CALM1 | CALM1 | ENSG00000198668.10 |
| | GATM | GATM | ENSG00000171766.15 |
| | BTG3 | BTG3 | ENSG00000154640.14 |
| | AKAP12 | AKAP12 | ENSG00000131016.16 |
| | LPIN1 | LPIN1 | ENSG00000134324.11 |
| | LAMA4 | LAMA4 | ENSG00000112769.18 |
| | DAAM2 | DAAM2 | ENSG00000146122.16 |
| | SCRN1 | SCRN1 | ENSG00000136193.16 |
| | VCL | VCL | ENSG00000035403.16 |
| | CYLD | CYLD | ENSG00000083799.17 |
| | C7 | C7 | ENSG00000112936.18 |

**Table 4.1 Genes used in stromal signature**

Genes described in Supplemental Figure 1 of Tomlins *et. al*, which were upregulated in laser-capture-microdissected stroma compared to epithelium.[180] On the left is the gene name used in the original manuscript, the middle column is the currently accepted gene name, and on the right is the ensembl gene id used in expression estimation. "Usual suspect" genes, such as Vinculin (VCL), Myosin light chain kinase (MYLK), and Tropomyosin 2 (TPM2), demonstrate the smooth muscle component of the stroma in this list.

# CHAPTER 5

# The Landscape of Transcript Variation in Prostate Cancer

*The work presented in this chapter is in preparation as a manuscript.*

## 5.1 Introduction

Prostate cancer poses a significant threat to human health. In the United States in 2016, prostate cancer is projected to be the leading cancer type diagnosed in men, and second-leading cause of cancer-related death in men.[1]

Early detection of prostate cancer may significantly inform treatment decisions and improve patient outcomes. However, current methods to detect prostate cancer early, such as measuring serum levels of the prostate-specific antigen (PSA), are famously poor at delineating early aggressive prostate cancer from other benign diseases like benign prostatic hyperplasia and unaggressive prostate tumors which are ubiquitous in aging men.[20-22,49,50] The field therefore recognizes a need for biomarkers which both detect aggressive cancer early *and* are specific versus both normal tissue and benign disease.

Regarding treatment, aggressive prostate cancer is either universally or nearly-universally dependent on signaling of the androgen receptor for growth and survival, and the most common treatment courses are surgical removal of the prostate, localized radiation therapy, chemotherapy, and ultimately androgen deprivation therapy ("castration"), from which the cancer usually recovers, usually by metastasizing to bone, lymph node, liver, or other soft tissue. When we study prostate cancer tissue, it is from relatively untreated "hormone-naive" primary prostate tumors ("PCa"), and metastatic castration-resistant prostate cancer ("mCRPC").

Several molecular landscapes of prostate cancer have already been established. These include, in both PCa and mCRPC : point mutations and short insertions and deletions in exonic regions of known genes ; copy number variants from the level of the gene to whole chromosomes ; DNA methylation at CpG islands ; gene fusions ; and expression of genes and intergenic long non-coding RNA.[30,32,41,45,179,189] These studies have underscored the inter-tumor heterogeneity of prostate cancer, but have uncovered common themes as well, including mutation of AR cofactors, genome-wide shifts in DNA methylation, activation of developmental signaling, cell-cycle deregulation, DNA repair deficiency and knockout of "usual suspect" tumor suppressors, and most strikingly, highly recurrent gene fusions of the ETS transcription factor family.[30,33]

Still, there is a common sentiment amongst researchers that the mechanistic cause of prostate cancer eludes us, and higher-level functional integration of these molecular observations still seems likely to advance our understanding of the disease. Further, critical open questions about prostate cancer's biology remain. While we understand that AR signaling is necessary to prostate cancer, we also know that normal prostate tissue is dependent on AR signaling, which casts some confusion on AR's role. And, while ETS gene fusions are extremely common, their function still seems to evade us. Finally, even considering all of the molecular subtypes the landscaping efforts have characterized, 26% of primary prostate tumors did not have evidence for presence of a main molecular driver, and could be characterized as "known-driver negative," though many of these tumors also exhibited broad-scale copy number aberrations which may drive or help drive cancer progression as well.

Critically, in addition to leaving open broad-scale biology questions, the previous efforts to define molecular landscapes of PCa and mCRPC have not done two things : they have not systematically profiled RNA splicing, and they have not fully addressed the critical need for early detection biomarkers; further biomarkers are still likely to be useful despite the recent development of many diagnostic and prognostic biomarkers. Biomarkers which improve on PSA include detection of TMPRSS2:ERG and PCA3 transcripts in urine using an aggregated prostate score ("MiPS"), detection of other cancer-specific RNA such as AMACR, detection of PSA's alternate form pro-PSA or the ratio of PSA unassociated with serum protease inhibitors in blood (percentage free PSA) or other aggregate PSA measures ("Prostate Health Index"), and an array of protein (ProMark; 4K Score), gene expression (Oncotype DX; Prolaris; Decipher), epigenetic (ConfirmMDx), metabolomic (Prostarix), and even mitochondrial genome assays (Prostate Core Mitomic) from various institutions.[190-192] And, while RNA splicing has been studied in prostate cancer before, using exon microarrays in PCa, whole transcriptome sequencing in mCRPC, polyA RNA sequencing on a Chinese patient population in PCa, and from perspectives of junction detection and pan-cancer analysis in PCa, these efforts neither accounted for tumor content, nor integrated known molecular subtypes, nor integrated together normal tissue, primary tumors, and mCRPC, hampering both the development of splicing biomarkers and their functional contextualization.[193-197]

### 5.1.1 Summary and specific aims

In this study, we aim to survey the landscape of transcript variation over the progression of prostate cancer, with specific aims of characterization of novel diagnostic biomarkers, contextualization of variants against other known molecular aberrations and cancer subtypes, contextualization of variants against AR signaling, investigation of novel driving transcript variants, and finally investigation of transcript variants of the AR itself. In particular, we aim to leverage analytical efforts described in Chapter 4 to address and circumvent technical biases which have hampered previous efforts to study RNA splicing in cancer tissue samples.

## 5.2 Methods and Results

### 5.2.1 Prostate cancer tissue samples

We aggregated a total of 578 polyA RNA-seq datasets, from 78 normal prostate tissue samples, 370 primary prostate tumor tissue samples (PCa), and 130 metastatic castration-resistant prostate cancer tissue samples (mCRPC), from three sources : The Cancer Genome Atlas (TCGA), Stand Up to Cancer (SU2C), and a previously unpublished cohort from the University of Michigan (Michigan).[30,179] To arrive at these numbers, we excluded low-quality TCGA samples that TCGA themselves excluded, and normal tissue samples which were contaminated with tumor RNA (Figure C.1). We split the mCRPC samples by biopsy site, and ran per-cohort pairwise splicing analysis as described in Chapter 4, pairing cohorts as shown in Figure 5.1. Briefly, individual samples were aligned to the genome using a method we optimized to accurately quantify novel splice junctions (see Chapter 3), unspliced coverage over junction edges was also calculated, samples were merged to one large table, and junction switching was determined in aggregate between each paired tumor cohort and normal cohort ("differential splicing"), and for individual tumor samples ("outlier splicing"). See Chapter 4 for a full description of the analytical pipeline.

Significant differential splicing calls were subsequently intersected for the mCRPC biopsy sites, requiring them to be called in each biopsy site independently (albeit, at a very low significance threshold). This approach served to avoid detecting lineage specific splicing changes between prostate and the independent biopsy sites.

### 5.2.2 Application of correlative bias analysis

Additionally, correlative bias analysis was performed as described in Chapter 4, aiming to determine if significant junction switches were explainable by sample variations in 3' bias, total unspliced RNA, RNA integrity, alignment rate, tumor content, aggregate stromal expression, aggregate AR signaling, aggregate neuroendocrine signaling, or expression of the gene the junction came from. Derivations of the per-sample scores for these nine metrics are describe in full in Appendix C (Table C.2, Table C.3, Figures C.2-C.7), and the sample annotation table with their numerical values are presented in Table C.1, excepting per-sample per-gene expression, which is too large to present here. Per-junction correlative values were tracked along with every call and are retained in tabular presentation of the results.

By far the most important of these was the effort to disambiguate tumor content and stromal expression from splicing changes; without having done so, variants specific to the epithelial lineage vs. stromal lineages are impossible to distinguish from variants specific to cancerous epithelium vs. normal epithelium. In an attempt to further refine this approach, we additionally tracked variant fractions for all junctions in a pool of cultured normal prostate epithelium (PrEC cells), and used a combination of the two methods to filter lineage-specific junction switches. Precisely, we flagged junctions as lineage specific rather than cancer specific, per-cohort, if they were in the top 20% of junctions correlating with stromal content and PrEC cells expressed the transcript variant at above 25% variant fraction, then took the union of the cohorts. Figures C.8-C.10 demonstrate this flagging process. Calls which were not identified as lineage-specific were retained in subsequent differential splicing analysis.

### 5.2.3 Differential junction analysis

After filtering cell lineage-specific variants, we further filtered junctions owing to antisense transcription and readthroughs, performed Bonferroni multiple hypothesis correction against the total number of junctions tested to generate q-values, stringently counting at the beginning of the pipeline, and plotted significant differential junction calls for the three cohorts on a shared scatterplot, with significance per-cohort on the X-axis, and effect size as the absolute average change in variant fraction on the Y-axis (Figure 5.2.A). This analysis was inspired by volcano plots, in which statistical significance is supplemented by fold-change to clearly distinguish between variants near zero and those with biologically significant effects.

Next, we took the 25% most significant calls from each cohort with an average variant fraction shift of 10% or greater across samples (dotted lines, Figure 5.2.A), and plotted their count intersections in a Venn diagram (Figure 5.2.B). Considering the genome-wide nature of this analysis, the two primary tumor cohorts showed very strong overlap, but the mCRPC samples had many more unique variants. We interrogated these, and the vast majority were unspliced calls, which upon further inspection appeared to be driven by a pattern of intron retention in a subset of the mCRPC samples. Still, many variants were specific to the primary tumors with respect to the mCRPC samples, and may reflect the actual nature of disease progression. Distributions of bias correlations for the union of the three cohorts' differential splicing calls are plotted in Figure C.11 and without unspliced calls in the SU2C cohort in Figure C.12.

Unspliced calls in the SU2C cohort may reflect either a global pattern of intron retention, a previously unknown dimension of sample degradation, or both. On one hand, many of these calls inversely correlate with RIN (Figure C.11), which strongly implicates sample quality. On the other hand, the nature of the unspliced calls is inconsistent with biases we have previously observed, and frequently involves multiple introns per gene, but critically, not all of them (Figure C.13). Further work is needed to investigate this phenomenon.

Next, we investigated the 11 calls made by all three cohorts. Two of these were caused by expression of overlapping genes on opposite strands (antisense expression), missed because the overlapping region of the gene was unannotated, so we excluded them from further characterization. We additionally noted that a single call made by the two primary tumor cohorts was shared by 3/4s of the mCRPC biopsy site calls, and therefore included it in this characterization. These ten calls (11 - 2 + 1) are presented in Table 5.1. The variants are mixed with respect to prior annotation status (*i.e.*, their presence in the gene annotation database), and whether the variant is driven by a spliced junction or unspliced junction, but intriguingly all ten variants are expected to reflect alternative transcription start sites (ATSS). Furthermore, manual interrogation of AR binding in ChIP-seq data of VCaP at these variants' TSSs uncovered strong AR binding at most of them. We therefore hypothesized that AR mediates a broad pattern of alternative transcription starting in prostate cancer.

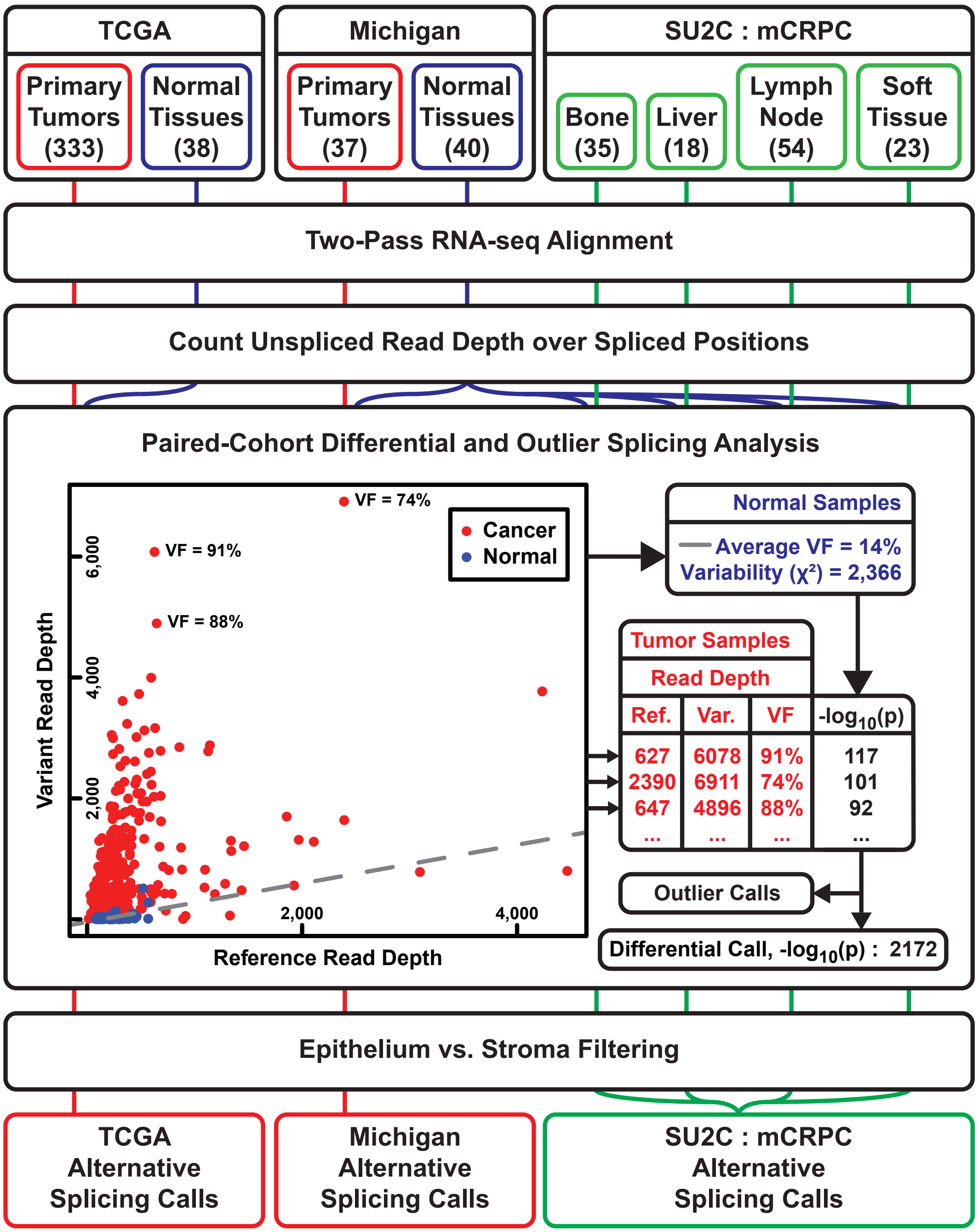
We validated the PDLIM5 ATSS variant in MDA-PCa-2b and VCaP cells, which were respectively expected to express the ATSS (tumor form) and full length (normal form), using the following primers : Total F, R : attctttgcccctgaatgtg, gtagggttcaccatcctcca ; ATSS F, R : ttggttggacattgcataaaa, acagggctcctttctcctct ; Full length F,R : tccacaaacaacatggccta, tcagtgcagatggagactgg. PCR and RNA-seq agreed very well on the variant fractions in these samples, bolstering our confidence in the estimates in tissue data (Figure 5.3).

### 5.2.4 Alternative transcription start site analysis

Owing to the junction-centric nature of our nomination process, the differential calls were not initially labeled by whether they were consistent with a TSS. We therefore mined three gene annotation databases, Gencode (high quality merged annotations), AceView (from cDNA), and MiTranscriptome (from cancer transcriptomes), for previously identified first exons, and labeled our calls as TSSs if the tumor form was associated with a known first exon.[92,107,198] Of the intersection of 316 calls in Figure 5.2 (excluding unspliced junctions in SU2C), 123 (39%) were consistent with known first exons. Deeper characterization of these variants, particularly their possible association with AR, is an area of further research.

### 5.2.5 Androgen receptor transcript variants

Finally, we investigated the tissue samples for the presence of truncating transcripts of the androgen receptor, notably AR-V7, which are of intense clinical interest and may mediate and/or prognosticate disease recurrence and castration resistance.[199] Briefly, we found evidence for both AR-V7 and an unspliced transcript at the locus, which were both expected to be truncating, at a variety of relative (to full-length) and absolute expression levels. See Figure 5.4 for a targeted analysis of AR-V7 expression. Further, while high relative levels of AR-V7 were randomly distributed with respect to AR amplification, they were mutually exclusive with point mutations in the ligand binding domain which are expected to mediate castration resistance (Figure 5.4.C).


TCGA
Primary Tumors (333)
Normal Tissues (38)
Michigan
Primary Tumors (37)
Normal Tissues (40)
SU2C : mCRPC
Bone (35)
Liver (18)
Lymph Node (54)
Soft Tissue (23)
Two-Pass RNA-seq Alignment
Count Unspliced Read Depth over Spliced Positions
Paired-Cohort Differential and Outlier Splicing Analysis
VF = 74%
VF = 91%
VF = 88%
● Cancer
● Normal
Variant Read Depth
6,000
4,000
2,000
2,000
4,000
Reference Read Depth
Normal Samples
Average VF = 14%
Variability ($\chi^2$) = 2,366
Tumor Samples
Read Depth
Outlier Calls
Differential Call, $-\log_{10}(p)$ : 2172
Epithelium vs. Stroma Filtering
TCGA Alternative Splicing Calls
Michigan Alternative Splicing Calls
SU2C : mCRPC Alternative Splicing Calls

| Ref. | Var. | VF | $-\log_{10}(p)$ |
|---|---|---|---|
| 627 | 6078 | 91% | 117 |
| 2390 | 6911 | 74% | 101 |
| 647 | 4896 | 88% | 92 |
| ... | ... | ... | ... |

**Figure 5.1 Transcript variant calling in prostate cancer tissue cohorts**
RNA-seq data from normal prostate tissue, primary prostate tumor tissue, and metastatic castration-resistant prostate cancer was aggregated and analyzed as shown. Full descriptions of the bioinformatics and biostatistics analysis performed here are presented in Chapter 4. Note that mCRPC calls were intersected between the four biopsy sites to eliminate biopsy-site-lineage-specific variants from nomination as cancer-specific variants.

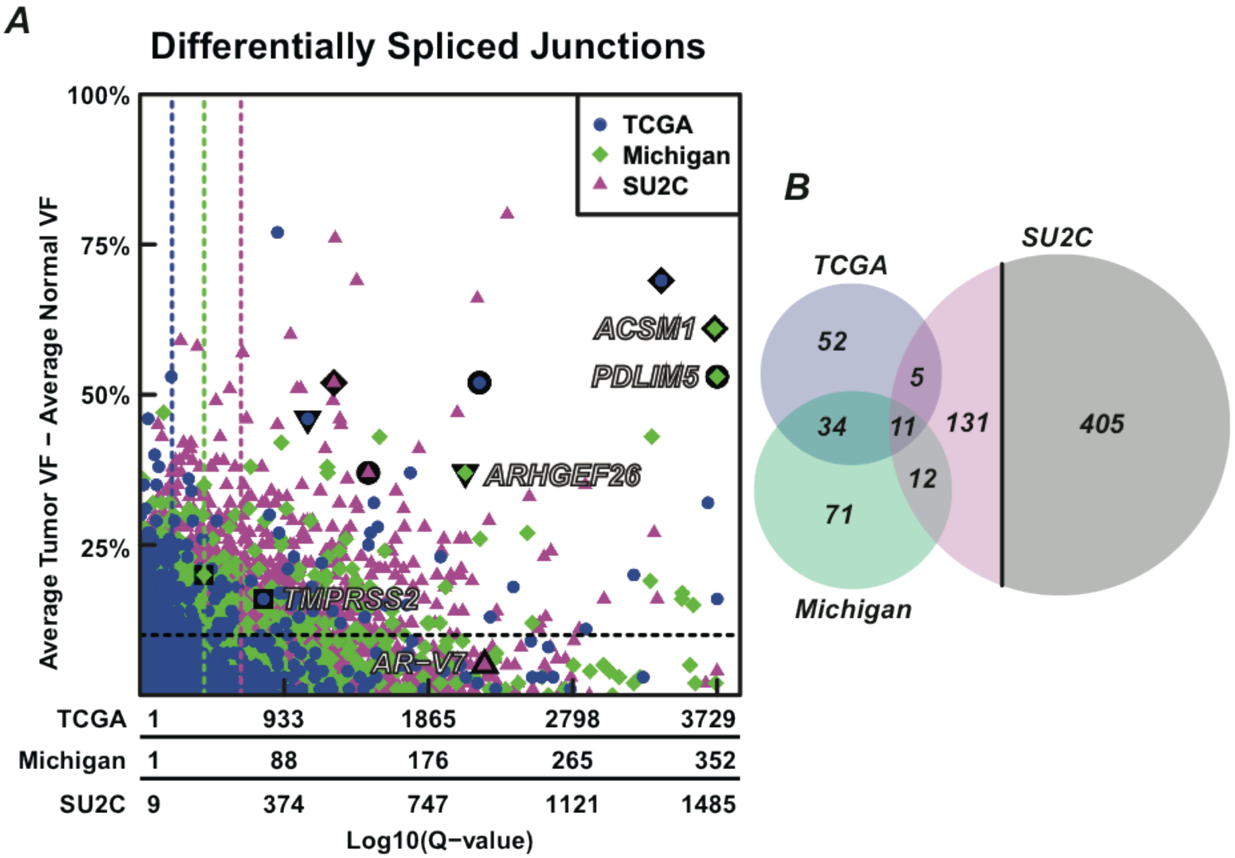


**Figure 5.2 Differential splicing calls in three prostate cancer cohorts**
**A)** Differential splicing analysis was performed on primary tumor tissue samples from TCGA, primary tumor tissue samples from Michigan, and mCRPC samples from SU2C, filtered for variants not likely to be driven by cell lineage differences, antisense expression, or readthrough expression, as is plotted here on a shared axis. Significance of the splicing switch is plotted on the X-axis as log-q-values, with scale labeled per-cohort, and the average variant fraction shift is plotted on the Y-axis. Highlighted variants are labeled and identified by outline symbols, and show each of their occurrences in the cohort calls. Dashed lines indicate a 10% variant fraction shift (horizontal, in black), and the top 25% most significant calls in each cohort (vertical, colored as the cohorts). **B)** Venn diagram of the upper right quadrants defined by the dashed lines in panel A. Calls unique to the SU2C cohort were further split into spliced calls (purple) and unspliced calls (gray) to highlight a broader pattern of unspliced junctions in this cohort. The 11 shared calls in the center are further characterized in Table 5.1.

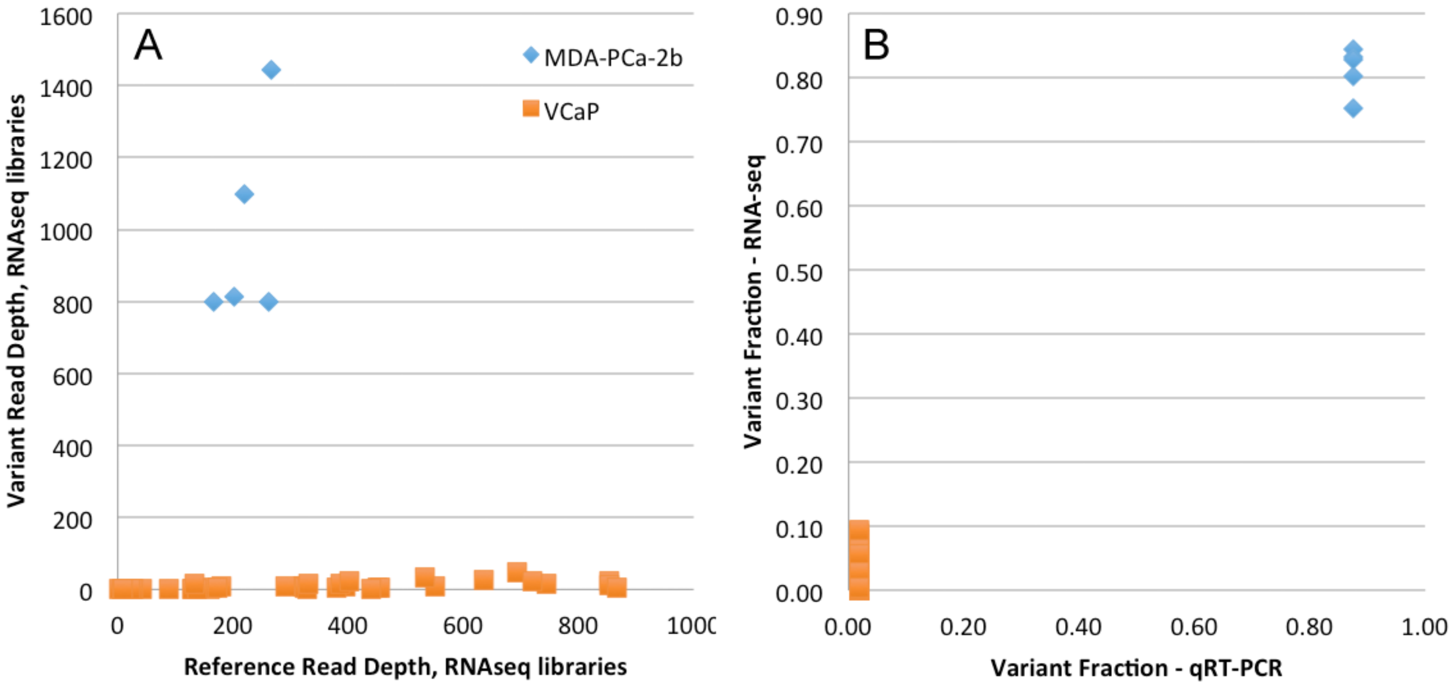


**Figure 5.3 PCR Validation of PDLIM5 ATSS variants in MDA-PCa-2b and VCaP**
**A)** Here we plot read depth spanning an unspliced junction consistent with an ATSS variant of PDLIM5 on the Y-axis, and read depth spanning the normal splice junction at that locus on the X axis. Note that MDA-PCa-2b is positive for the variant, and VCaP is negative. **B)** Variant fractions from panel A are plotted on the Y-axis, along with variant fractions estimated from qRT-PCR, taken by normalizing abundance estimates from PCR to the total transcript, then taking the fraction as ATSS / (ATSS + Full Length).

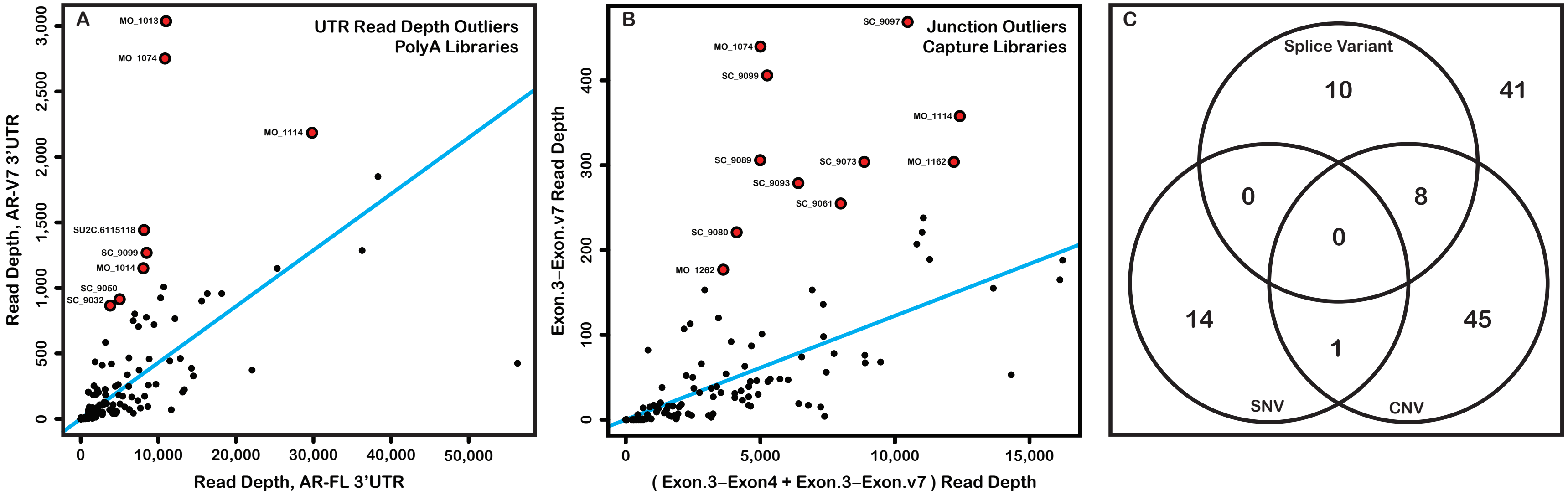
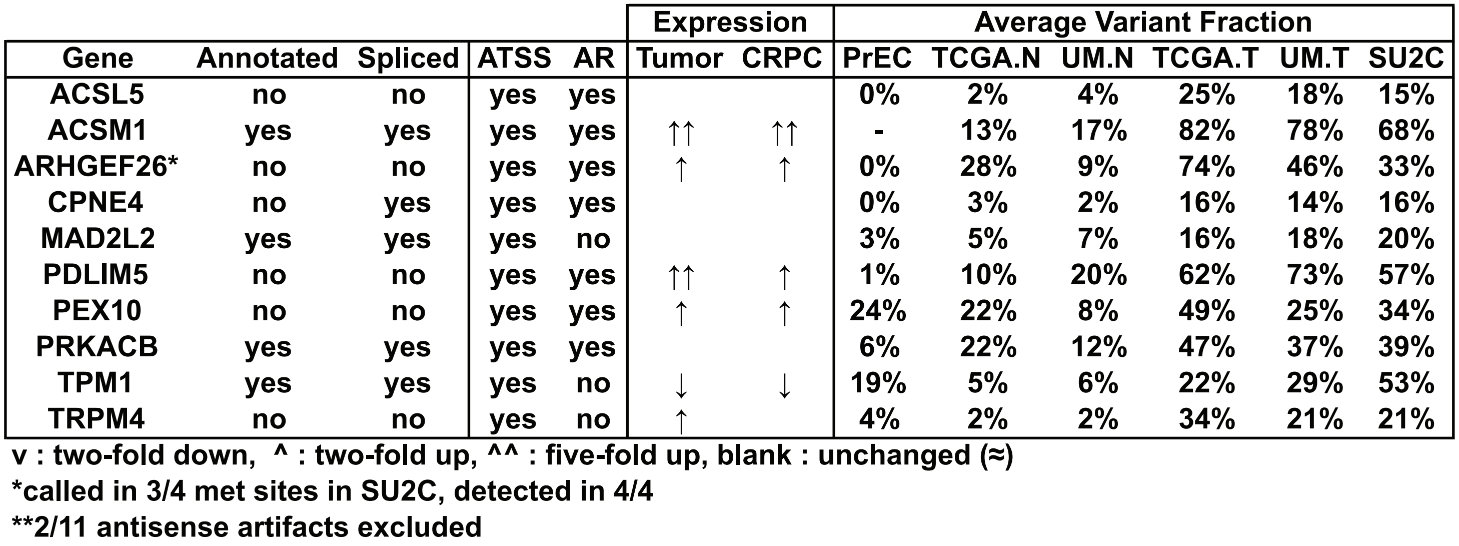


**Figure 5.4 Androgen receptor transcript variants in mCRPC**
**A)** RNA-seq depth of coverage over the last thousand nucleotides of full-length AR's 3' UTR, and the last thousand nucleotides of AR-V7's 3' UTR, was computed and plotted per sample for polyA libraries. **B)** Similarly, RNA-seq depth of coverage over the AR-V7-specific splice junction was computed and compared with read depth over the canonical exon3:exon4 splice junction for capture-RNA-seq libraries. In both panels A and B, a robust linear model is fit to the samples (shown in blue), and samples over three standard deviations away from the fit are labeled as outliers in red, and identified by sample name. **C)** 18 outlier samples identified in panels A and B, which express either a relatively high level of truncated AR (by UTR expression), or AR-V7 (by splice junction expression), are show in a Venn diagram with samples with detected AR SNVs, and detected AR copy-number gain. Overlap of the splice variant samples and CNV samples is statistically insignificant, and negative-overlap of the splice variant samples and SNV samples, while zero (and plausibly biologically significant), is also statistically insignificant.



| | | | | | Expression | | Average Variant Fraction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | Annotated | Spliced | ATSS | AR | Tumor | CRPC | PrEC | TCGA.N | UM.N | TCGA.T | UM.T | SU2C |
| ACSL5 | no | no | yes | yes | | | 0% | 2% | 4% | 25% | 18% | 15% |
| ACSM1 | yes | yes | yes | yes | ↑↑ | ↑↑ | - | 13% | 17% | 82% | 78% | 68% |
| ARHGEF26* | no | no | yes | yes | ↑ | ↑ | 0% | 28% | 9% | 74% | 46% | 33% |
| CPNE4 | no | yes | yes | yes | | | 0% | 3% | 2% | 16% | 14% | 16% |
| MAD2L2 | yes | yes | yes | no | | | 3% | 5% | 7% | 16% | 18% | 20% |
| PDLIM5 | no | no | yes | yes | ↑↑ | ↑ | 1% | 10% | 20% | 62% | 73% | 57% |
| PEX10 | no | no | yes | yes | ↑ | ↑ | 24% | 22% | 8% | 49% | 25% | 34% |
| PRKACB | yes | yes | yes | yes | | | 6% | 22% | 12% | 47% | 37% | 39% |
| TPM1 | yes | yes | yes | no | ↓ | ↓ | 19% | 5% | 6% | 22% | 29% | 53% |
| TRPM4 | no | no | yes | no | ↑ | | 4% | 2% | 2% | 34% | 21% | 21% |

**v : two-fold down, ^ : two-fold up, ^^ : five-fold up, blank : unchanged (≈)**

***called in 3/4 met sites in SU2C, detected in 4/4**

****2/11 antisense artifacts excluded**

**Table 5.1 10 transcript variants called in three prostate cancer cohorts**
The 11 transcript variants identified in the two PCa cohorts and mCRPC cohort in Figure 5.2.B were investigated, two were caused by antisense expression and were excluded, and one variant did not reach statistical significance in a single mCRPC biopsy site and was included here (ARHGEF26). The tumor-specific variant was checked for prior annotation status, whether it was spliced or unspliced, whether the variant was consistent with an ATSS transcript, whether the TSS had demonstrable enrichment of AR in ChIP-seq of VCaP cells, and whether expression was altered in cancer compared to normal tissue. These columns are shown adjacent to the average variant fractions in the tissue cohorts and PrEC samples.

# CHAPTER 6

# Concluding Remarks and Future Directions

## 6.1 Sequence Compression

In chapter 2 of this dissertation, I presented a novel algorithm and accompanying software to accelerate sequence alignment using the unique set of reads from a sequence library.  This work was built on (and succeeded by) two premises : first, that sequence alignment is reproducible, in other words that two instances of the same sequence should align to the same place in the genome, and second, that pre-processing and post-processing sequence reads to eliminate redundancy was computationally cheaper than performing the redundant alignments.  I used exact match hashmaps because it was the most straightforward, both because hashmaps are simple to implement, and because it did not necessitate access to the internal engineering of the Burroughs-Wheeler style sequence aligners (which is complex). The first immediately obvious extension of this work could be to instead perform this kind of "alignment caching" in the internal structure of the aligner, where for instance k-mers used to seed alignments could instead be stored in place of sequence reads, or related variants on this idea. On another related note, other scientists have worked to eliminate redundancy across samples, using I believe exact sequence read matching (identical to the approach presented in chapter 2 here), but instead engineered with the Hadoop / MapReduce framework.[147]

The much more sophisticated extension of this work, that I still strongly suspect would greatly accelerate alignment, is to store input sequence reads in more advanced redundancy-eliminating data structures than a simple hashmap.  Specifically, what I would propose is that, as with the genome, the input sequence reads could be compressed in a Burroughs-Wheeler transformed suffix trie or suffix array, and then the reads and genome could be compared in such a way that whole branches of reads which don't map could be eliminating from consideration (or softclipped) at once.  An even farther-reaching idea would be to collapse multiple samples together into a single massive one of these transformed suffix tries - for instance entire cohorts, or more grandiose, entire sequence repositories like the sequence read archive (though, that repository also already has a search function using BLAST).  By storing sample identifiers along with the sequence it would always be possible to return to the initial per-sample fastq files (except again for the quality scores, which would require special handling). This idea is not so different from how Web-BLAST is set up, and would permit querying sequences against not just the reference genome, but other samples as well.  Constraining this problem to tangible goals, and the engineering, would be the main challenges to its completion.  Storage in particular remains a major obstacle for sequencing centers.

Two noteworthy developments have occurred in this area since our method was published.  First, both personal and cloud computing resources (generally) have access to more memory than before, which enables less compressed storage of the genome, and therefore faster alignment (*e.g.*, STAR).[102] I expect this method of alignment acceleration undermines our method, owing to our more-or-less static compression and decompression steps, but does not undermine the concept of acceleration through input reduction - only the algebra on whether it's worth doing.  Second, the binary alignment map format (BAM), which stores alignment results, has been dramatically improved upon by the compressed alignment map format (CRAM).[200] This format works by storing sequence positions in the genome rather than sequences themselves. I expect the advent of CRAM would have little bearing on accelerating alignment by means of sequence read compression, but it does undermine a possible secondary purpose of compressing reads, which is to reduce storage space. However, index-based storage formats like CRAM may at the same time provide further opportunities for acceleration as well, and should certainly be considered by scientists continuing work in this field.

## 6.2 Longer read sequencing and tissue profiling

Much of the methodological work presented in this dissertation concerns handling of short sequence reads, and using those short sequence reads to interpret splicing and transcript changes at the level of full messenger RNA molecules. However, as new sequencing technologies continue to be developed, the length of sequence reads that can be attained accurately and cheaply will continue to grow. Current technologies such as the Single Molecule, Real-Time sequencing technology from Pacific Biosciences can already attain read lengths longer than ten kilobases (and reportedly, as long as sixty kilobases), and these methods have been used to profile mRNA in addition to DNA.[201] It seems like a matter of historical inevitability that full-transcript, full-transcriptome profiling will eventually be both possible and cost-effective for large sample cohorts.

That said, merely being able to sequence longer reads will have no bearing on the RNA degradation prevalent in patient tissue samples. RNA is unstable and its degradation (which mostly manifests as truncated molecules, rather than nucleotide changes) dramatically complicates analysis of the transcriptome, but the insights that tissue profiling offer force us to tackle the issue head-on rather than study cultured samples instead. Further, one of the necessary steps in RNA sample preparation is to enrich for RNA species other than ribosomal RNA, which is generally regarded as less interesting and would otherwise account for the majority of transcripts. In most tissue studies, poly-adenylation capture is performed, by pulling down molecules with poly-adenine stretches using complementarity to poly-thymine molecules. In the context of RNA degradation though, this manifests as bias toward the polyadenylated end of the transcript (the 3' end). Longer read sequencing will also not help to address this issue. However, I do think that longer read sequencing could further improve other methods of enriching for non-ribosomal-RNA, particularly the method of exome-capture RNA-seq presented by our group, with possibly dramatic effects on the ability to analyze transcript variation from those molecules.[202] In the end, the most critical component to analyzing RNA from tissue samples is simply starting with less-degraded RNA, by means of rapid sample processing.[203]

## 6.3 Tissue Profiling and Cell Lineage Deconvolution

In the work presented in this dissertation, we presented some simple and novel correlative methods to analyze around the "tumor content" problem, and showed some success in terms of eliminating transcript variants specific to epithelium vs. stroma in prostate. With these methods we're really only scratching the surface of this problem though, and there are two big ways to advance the analysis on this issue. First, lineage-specific expression signatures could be generated for more, and more-specific solid tissue cell lineages than we performed here, akin to deconvolution efforts already ongoing with leukocytes.[204] This would push the deconvolution effort more onto the side of bioinformatics, and it remains to be seen how successful these approaches will be. Alternatively, deconvolution could be performed at the level of sample preparation, using laser-capture microdissection, or possibly instead laser ablation of undesired cell lineages.[180] Although current research attention is squarely on the transcriptome of cancer cells, I suspect that soon or eventually researchers will also be interested in the transcriptional profile of other present cell lineages (notably stroma), which is likely to perform an enabling role in cancer by means of paracrine signaling, so laser-capturing individual lineages would be superior to ablating specific lineages (though, obviously it would depend on the specific study). My understanding is that current methods to perform microdissection are both intensely laborious and slow (meaning RNA degradation as well), so further engineering developments in this direction would be welcome. Single-cell sequencing also bears discussion in this context, however, the necessary step of disaggregating single cells from solid tumors involves cleaving cell surface proteins, with effects on the signaling pathways the cells express - it's unclear if and how this could be addressed for effective transcriptome profiling.

### 6.4 Protein-level Analysis

As discussed in the introduction to this dissertation, our main interest in studying splicing is really on the effects those transcript changes have on mature proteins. However, owing to the database-search-centric nature of mass spectrometry, it's difficult to perform *de novo* sequence discovery from the current generation of proteomics tools, so our focus instead has been on RNA. To address this issue, an exciting new field termed proteogenomics has arisen, wherein RNA-seq data is used to predict protein sequences, in order to drive mass spectrometry database searching.[133,134] I expect that routine and paired analysis of RNA and protein in this way could dramatically advance cancer tissue profiling studies, and the National Cancer Institute agrees in having founded the Clinical Proteomic Tumor Analysis Consortium, which aims to perform paired transcriptome and proteome profiling of human cancer tissue samples. It is critical to emphasize here that proteins are more stable than RNA molecules, particularly in archived formalin-fixed paraffin-embedded samples which comprise the vast majority of existing cancer tissues samples, but in biofluids and staining assays as well, and are therefore better positioned as possible biomarkers. Our efforts in studying RNA have possible application toward characterizing their corresponding protein biomarkers, but efforts to study proteins more directly will likely more directly advance this biomarker effort.

As a final comment, if it could be developed, quantitative high-throughput protein sequencing would be similarly significant to biological research as high-throughput DNA sequencing was, and would define decades of future research. For now, directly sequencing proteins in high-throughput, or heretically going against the central dogma of molecular biology and reverse translating protein to nucleotides in order to sequence those, still seems like science fiction. For as well-studied as DNA and RNA have been, proteins remain a vast frontier.

### 6.5 Epitranscriptomics

While we appreciate the significance of covalent modifications to DNA in the form of methylation and other marks, DNA's protein scaffolding (histones) also as methylation and other marks, and other proteins in the form of post-translational modifications, surprisingly little attention has been paid to the significance or landscape of covalent modifications to RNA. Indeed, RNA methylation is a known phenomenon, with apparent physiological function.[205,206] I would predict that further efforts to profile landscapes of aberrations in cancer will likely eventually also profile covalent modifications of RNA, similar to the surge of recent interest in circular RNA molecules in human cancer. I expect such covalent modifications could easily affect sequence recognition and therefore binding of cofactors, in translation, formation of RNA secondary structures, and other pathways which use complementarity like microRNA-mediated silencing.

## 6.6 Further Applications of Junction-Based Splicing Analysis

Briefly, immediate extensions of the work presented here, on alternative splicing analysis at the level of the junction in cancer tissue cohorts and application to prostate cancer, include application of this style of analysis to (all) other cancer types (though, the sample mixture issue remains not completely resolved), retooling the analysis toward application to single cancer tissue samples ("precision medicine"), defining and applying signatures of splicing aberrations associated with dysregulation or mutation of specific splicing factors, and finally extending biological characterization of our results in prostate cancer and ultimately translating those results to clinical tests (particularly, biomarkers).

## 6.7 Concluding Note on Cancer Landscaping

The field of cancer research continues to profile aberrations in larger and larger cancer tissue sample cohorts, in DNA, epigenetics, RNA, RNA Splicing, proteins, metabolites, and probably more dimensions of molecular biology in the future. It seems inevitable to me that researchers will eventually run out of aberrations worth characterizing. At that point, the focus of the field will have to shift to determining how cancer really works, rather than cataloging everything that goes wrong with it. This functional work will be an exciting challenge, because it will necessarily force us to develop deeper understanding of human biology en route, and I expect our understanding of normal human biology and cancer biology to advance lockstep well into the twenty-first century and beyond.

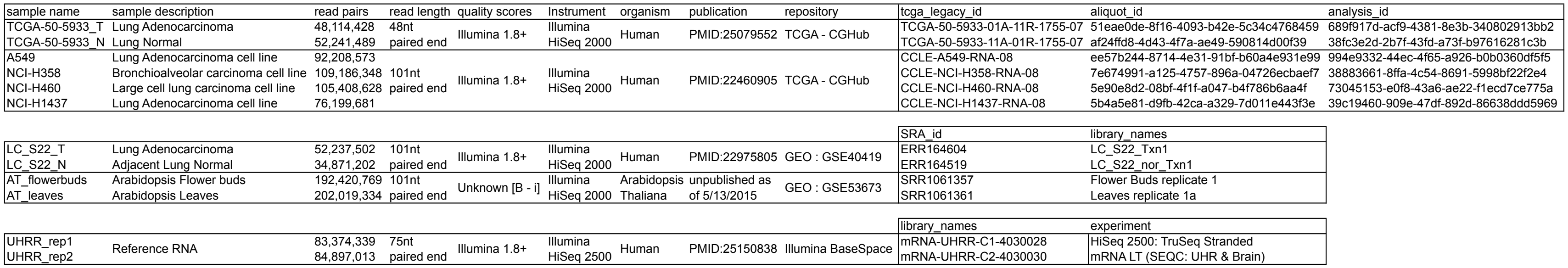
# APPENDIX A
## Supplemental Data for Chapter 2

| | | Genome | | | RNA-Seq | | | | | Chipseq | | Exome | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Oculus performance statistics | #1 | #2 | #3 | #1 | #2 | #3 | #4 | #5 | #1 | #2 | #1 | #2 | #3 |
| | Run | MDA-MB-231 (IDEA) | T-47D (IDEA) | ERR000589 | Bodymap | MDA-MB-231 | T-47D | BT-20 | BT-474 | Broad - H3k4-me3 | Uw - TFBS | SRR098490 | SRR098492 | SRR171306 |
| | Sequence read archive accession id | SRR097850 | SRR097852 | ERR000589 | ERS025093 | SRR097790 | SRR097792 | SRR097786 | SRR097787 | SRR227346 | SRR299316 +<br>SRR299313 | SRR098490 | SRR098492 | SRR171306 |
| | total # of reads (millions) | 25 | 27 | 24 | 385 | 79 | 83 | 84 | 81 | 37 | 66 | 260 | 272 | 154 |
| | read length | 50 | 50 | 51 | 100 | 50 | 50 | 50 | 50 | 36 | 36 | 76 | 76 | 50 |
| SE | % unique | 93.6% | 93.4% | 95.5% | 69.3% | 31.7% | 31.6% | 32.4% | 49.3% | 95.9% | 35.0% | 81.4% | 82.1% | 87.0% |
| | %error | 0.03% | 0.03% | 0.03% | 0.002% | 0.11% | 0.12% | 0.11% | 0.13% | 0.002% | 0.02% | 0.01% | 0.008% | 0.006% |
| | Bowtie CPU runtime (hours) | 0.48 | 0.58 | 0.63 | 29.81 | 1.68 | 1.99 | 1.76 | 2.35 | 1.60 | 1.37 | 13.19 | 13.61 | 8.35 |
| | (Oculus wrapping Bowtie) CPU runtime (hours) | 0.49 | 0.54 | 0.62 | 24.83 | 0.76 | 0.75 | 0.78 | 1.53 | 1.77 | 0.37 | 12.56 | 13.90 | 8.47 |
| | %runtime | 102.7% | 92.8% | 97.7% | 83.3% | 45.1% | 38.0% | 44.0% | 64.9% | 110.4% | 27.0% | 95.3% | 102.2% | 101.5% |
| | BWA CPU runtime (hours) | 2.35 | 2.97 | 2.61 | 146.65 | 7.03 | 7.43 | 8.03 | 9.63 | 2.09 | 2.81 | 39.66 | 41.81 | 26.49 |
| | (Oculus wrapping BWA) CPU runtime (hours) | 2.28 | 2.75 | 2.56 | 116.39 | 3.03 | 2.88 | 3.32 | 5.73 | 2.07 | 1.23 | 35.22 | 38.29 | 25.06 |
| | %runtime | 97.0% | 92.5% | 98.0% | 79.4% | 43.1% | 38.8% | 41.3% | 59.5% | 99.0% | 43.9% | 88.8% | 91.6% | 94.6% |
| PE | % unique | 98.5% | 98.4% | 99.7% | | 77.0% | 74.7% | 77.0% | 87.0% | | | 96.0% | 95.3% | |
| | %error | 0.0004% | 0.001% | 0.001% | | 0.08% | 0.08% | 0.08% | 0.04% | | | 0.004% | 0.004% | |
| | Bowtie CPU runtime (hours) | 4.00 | 4.36 | 1.74 | | 5.25 | 5.11 | 5.16 | 6.28 | | | 20.51 | 22.31 | |
| | (Oculus wrapping Bowtie) CPU runtime (hours) | 3.76 | 4.11 | 1.72 | | 3.94 | 4.37 | 4.31 | 5.82 | | | 20.33 | 21.93 | |
| | %runtime | 94.2% | 94.2% | 98.9% | | 75.1% | 85.4% | 83.5% | 92.7% | | | 99.2% | 98.3% | |
| | BWA CPU runtime (hours) | 2.81 | 3.28 | 3.09 | | 9.42 | 9.15 | 9.25 | 11.38 | | | 42.02 | 44.27 | |
| | (Oculus wrapping BWA) CPU runtime (hours) | 2.69 | 3.11 | 3.02 | | 7.61 | 7.09 | 7.69 | 10.28 | | | 42.17 | 44.51 | |
| | %runtime | 95.8% | 94.8% | 97.9% | | 80.8% | 77.5% | 83.1% | 90.3% | | | 100.4% | 100.6% | |

**Table A.1 Oculus performance statistics**

Detailed benchmarking data used in generating runtime figures.

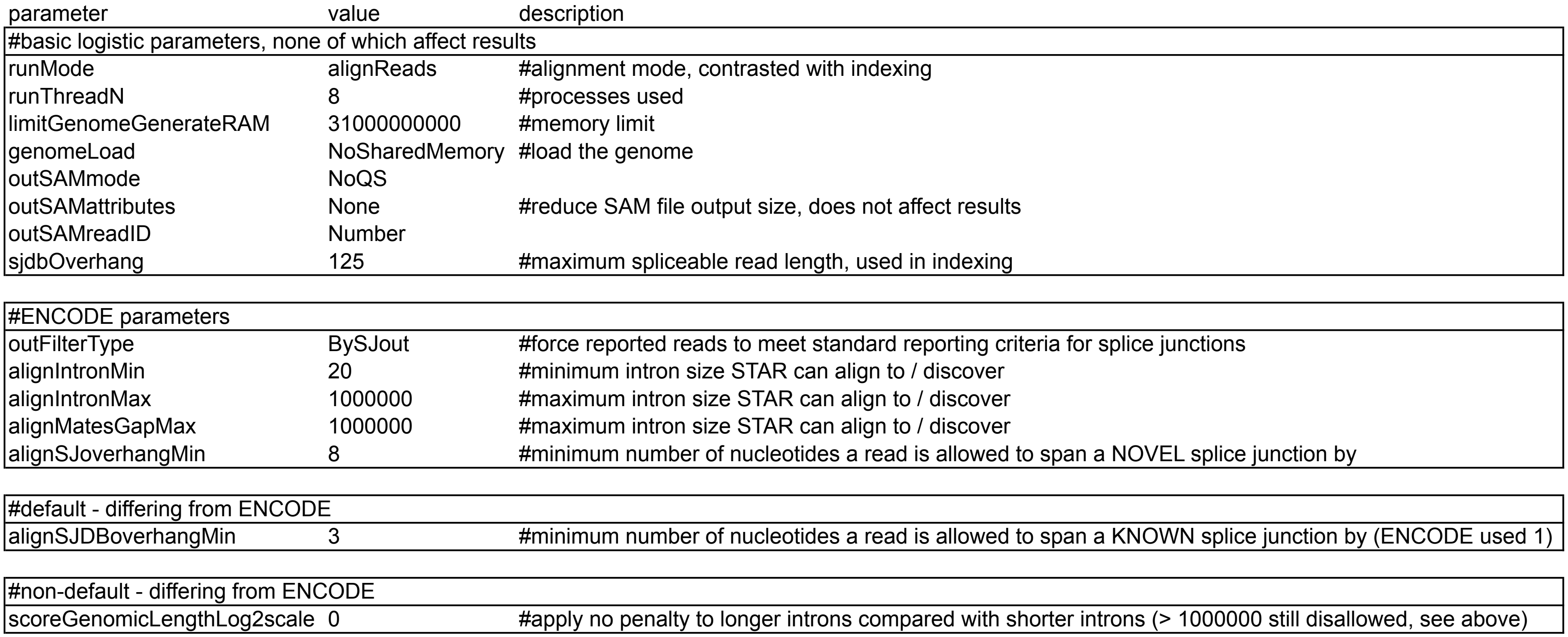
# APPENDIX B
## Supplemental Data for Chapter 3



| sample name | sample description | read pairs | read length | quality scores | Instrument | organism | publication | repository | tcga_legacy_id | aliquot_id | analysis_id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TCGA-50-5933_T | Lung Adenocarcinoma | 48,114,428 | 48nt paired end | Illumina 1.8+ | Illumina HiSeq 2000 | Human | PMID:25079552 | TCGA - CGHub | TCGA-50-5933-01A-11R-1755-07 | 51eae0de-8f16-4093-b42e-5c34c4768459 | 689f917d-acf9-4381-8e3b-340802913bb2 |
| TCGA-50-5933_N | Lung Normal | 52,241,489 | | | | | | | TCGA-50-5933-11A-01R-1755-07 | af24ffd8-4d43-4f7a-ae49-590814d00f39 | 38fc3e2d-2b7f-43fd-a73f-b97616281c3b |
| A549 | Lung Adenocarcinoma cell line | 92,208,573 | 101nt paired end | Illumina 1.8+ | Illumina HiSeq 2000 | Human | PMID:22460905 | TCGA - CGHub | CCLE-A549-RNA-08 | ee57b244-8714-4e31-91bf-b60a4e931e99 | 994e9332-44ec-4f65-a926-b0b0360df5f5 |
| NCI-H358 | Bronchioalveolar carcinoma cell line | 109,186,348 | | | | | | | CCLE-NCI-H358-RNA-08 | 7e674991-a125-4757-896a-04726ecbaef7 | 38883661-8ffa-4c54-8691-5998bf22f2e4 |
| NCI-H460 | Large cell lung carcinoma cell line | 105,408,628 | | | | | | | CCLE-NCI-H460-RNA-08 | 5e90e8d2-08bf-4f1f-a047-b4f786b6aa4f | 73045153-e0f8-43a6-ae22-f1ecd7ce775a |
| NCI-H1437 | Lung Adenocarcinoma cell line | 76,199,681 | | | | | | | CCLE-NCI-H1437-RNA-08 | 5b4a5e81-d9fb-42ca-a329-7d011e443f3e | 39c19460-909e-47df-892d-86638ddd5969 |
| | | | | | | | | | SRA_id | library_names | |
| LC_S22_T | Lung Adenocarcinoma | 52,237,502 | 101nt paired end | Illumina 1.8+ | Illumina HiSeq 2000 | Human | PMID:22975805 | GEO : GSE40419 | ERR164604 | LC_S22_Txn1 | |
| LC_S22_N | Adjacent Lung Normal | 34,871,202 | | | | | | | ERR164519 | LC_S22_nor_Txn1 | |
| AT_flowerbuds | Arabidopsis Flower buds | 192,420,769 | 101nt paired end | Unknown [B - i] | Illumina HiSeq 2000 | Arabidopsis Thaliana | unpublished as of 5/13/2015 | GEO : GSE53673 | SRR1061357 | Flower Buds replicate 1 | |
| AT_leaves | Arabidopsis Leaves | 202,019,334 | | | | | | | SRR1061361 | Leaves replicate 1a | |
| | | | | | | | | | library_names | experiment | |
| UHRR_rep1 | Reference RNA | 83,374,339 | 75nt paired end | Illumina 1.8+ | Illumina HiSeq 2500 | Human | PMID:25150838 | Illumina BaseSpace | mRNA-UHRR-C1-4030028 | HiSeq 2500: TruSeq Stranded mRNA LT (SEQC: UHR & Brain) | |
| UHRR_rep2 | | 84,897,013 | | | | | | | mRNA-UHRR-C2-4030030 | | |

**Table B.1 RNA-seq sample metadata**

Sample metadata for samples used in Chapter 3.



| parameter | value | description |
|---|---|---|
| #basic logistic parameters, none of which affect results | | |
| runMode | alignReads | #alignment mode, contrasted with indexing |
| runThreadN | 8 | #processes used |
| limitGenomeGenerateRAM | 31000000000 | #memory limit |
| genomeLoad | NoSharedMemory | #load the genome |
| outSAMmode | NoQS | |
| outSAMattributes | None | #reduce SAM file output size, does not affect results |
| outSAMreadID | Number | |
| sjdbOverhang | 125 | #maximum spliceable read length, used in indexing |
| #ENCODE parameters | | |
| outFilterType | BySJout | #force reported reads to meet standard reporting criteria for splice junctions |
| alignIntronMin | 20 | #minimum intron size STAR can align to / discover |
| alignIntronMax | 1000000 | #maximum intron size STAR can align to / discover |
| alignMatesGapMax | 1000000 | #maximum intron size STAR can align to / discover |
| alignSJoverhangMin | 8 | #minimum number of nucleotides a read is allowed to span a NOVEL splice junction by |
| #default - differing from ENCODE | | |
| alignSJDBoverhangMin | 3 | #minimum number of nucleotides a read is allowed to span a KNOWN splice junction by (ENCODE used 1) |
| #non-default - differing from ENCODE | | |
| scoreGenomicLengthLog2scale | 0 | #apply no penalty to longer introns compared with shorter introns (> 1000000 still disallowed, see above) |

**Table B.2 Full STAR runtime parameters**
STAR runtime parameters for analysis performed in Chapter 3.

### Figures B.1-B.12 Splice Junction Quantification from Two-Pass Alignment

For each of the 12 samples described in Table 3.1 and Table B.1, we performed alignment with and without annotation, in both one and two alignment passes, yielding splice junction quantification estimates in the form of unique read alignment depth (*De Novo* 1-pass, Gencode 1-pass, *De Novo* 2-pass, and Gencode 2-pass) ("Gencode" is described previously as "Annotation" - they are equivalent). Quantification for splice junctions present in either alignment pass are plotted in log10-scale as scatterplots (SF1-12, panels A-C). The Y=X line, corresponding to equal quantification, is highlighted in red over each scatterplot. The same data presented in the scatterplots is additionally presented as histograms of ratios (Figures B.1-B.12, panels D-F) to convey the plot density. In panels D, the median quantification ratio of 1-pass *De Novo* alignment to 1-pass Gencode alignment is highlighted with a red line, and in text at the top of the plot. Figures B.1-B.12 used a cutoff of at least 10 reads in the Annotation 1-pass alignment for the histograms (to eliminate visual distraction).

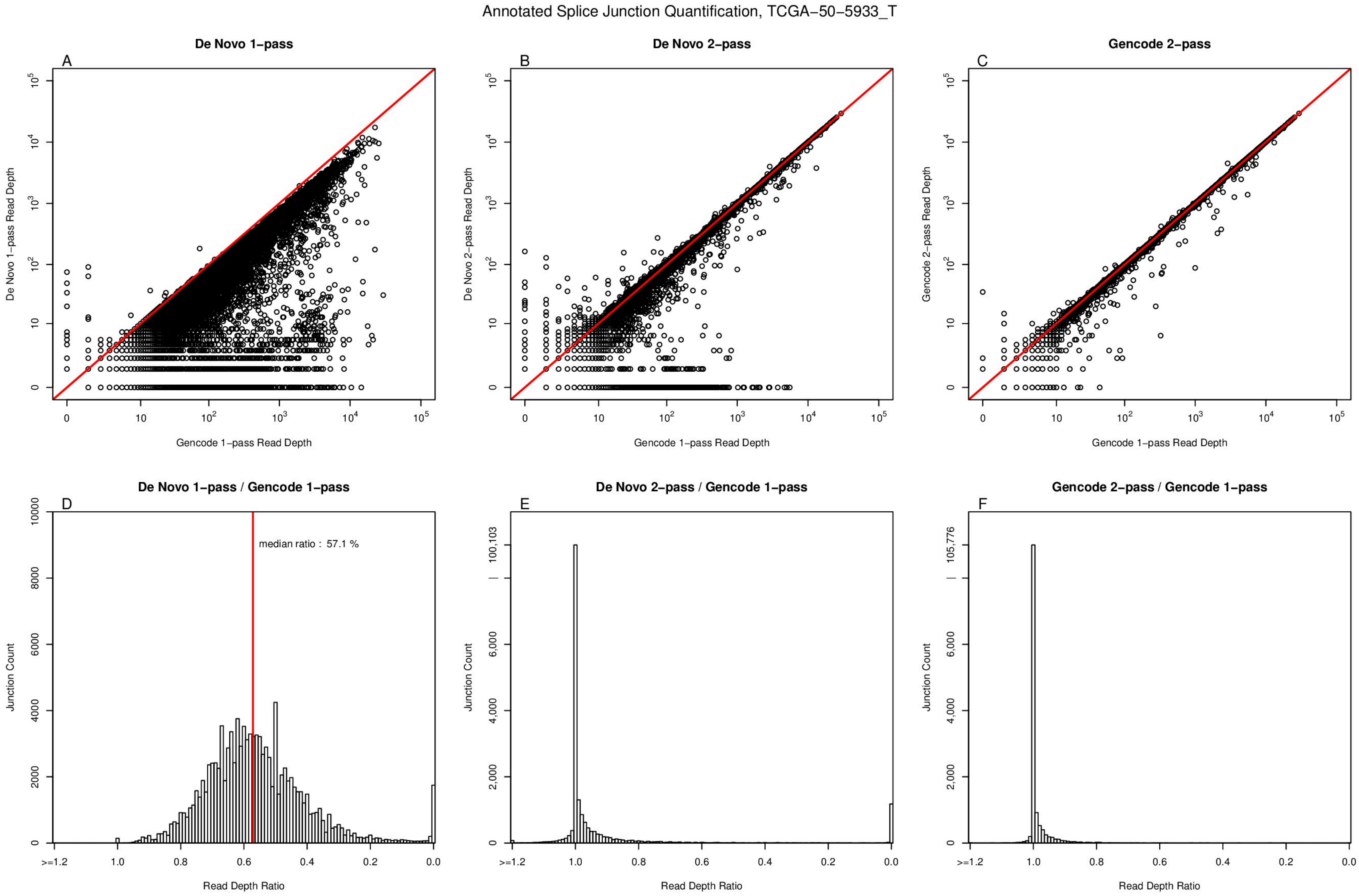


**Figure B.1 Splice Junction Quantification, TCGA-50-5933.T**

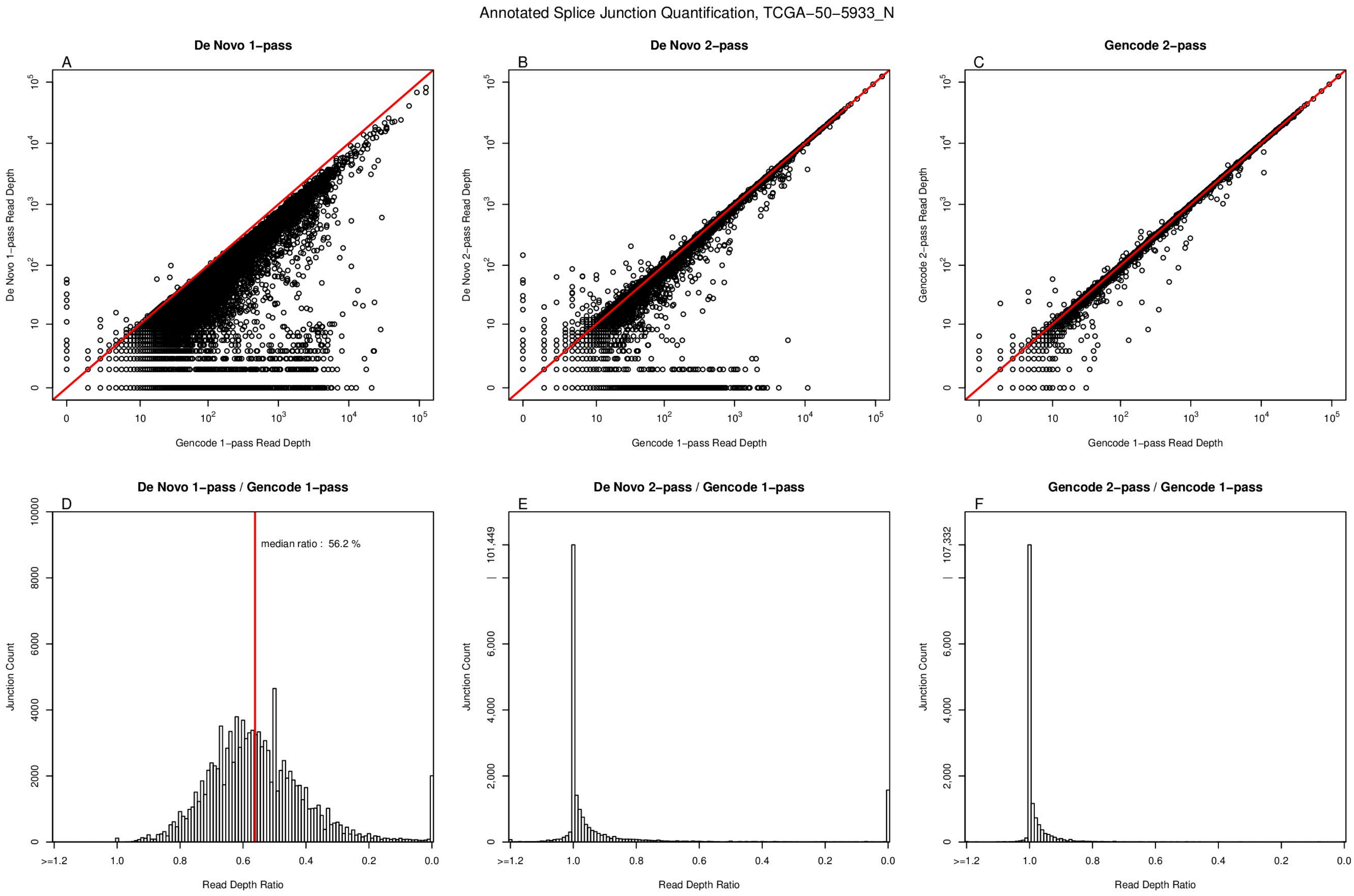


**Figure B.2 Splice Junction Quantification, TCGA-50-5933.N**

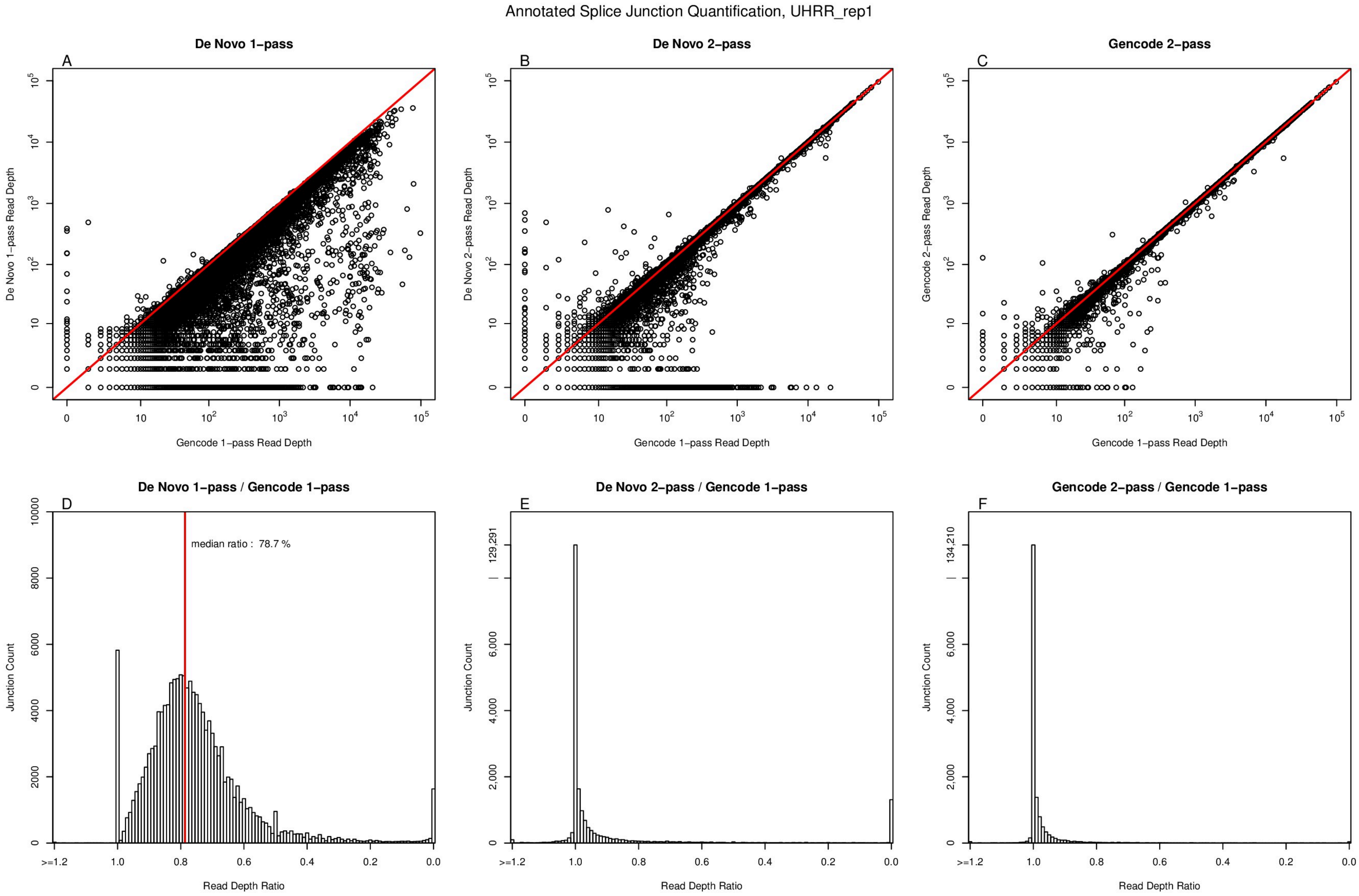
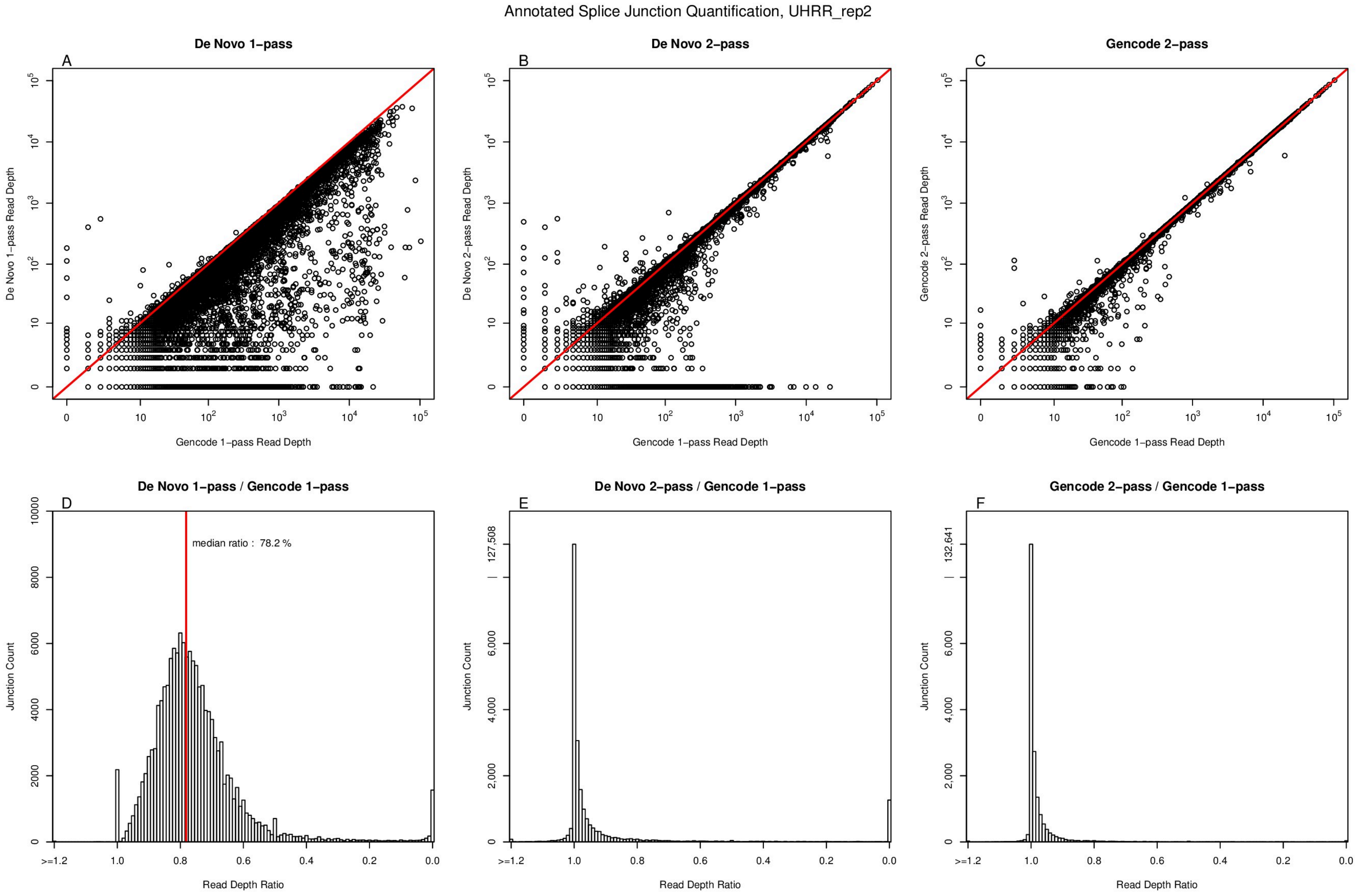


**Figure B.3 Splice Junction Quantification, UHRR_rep1**



## Figure B.4 Splice Junction Quantification, UHRR_rep2

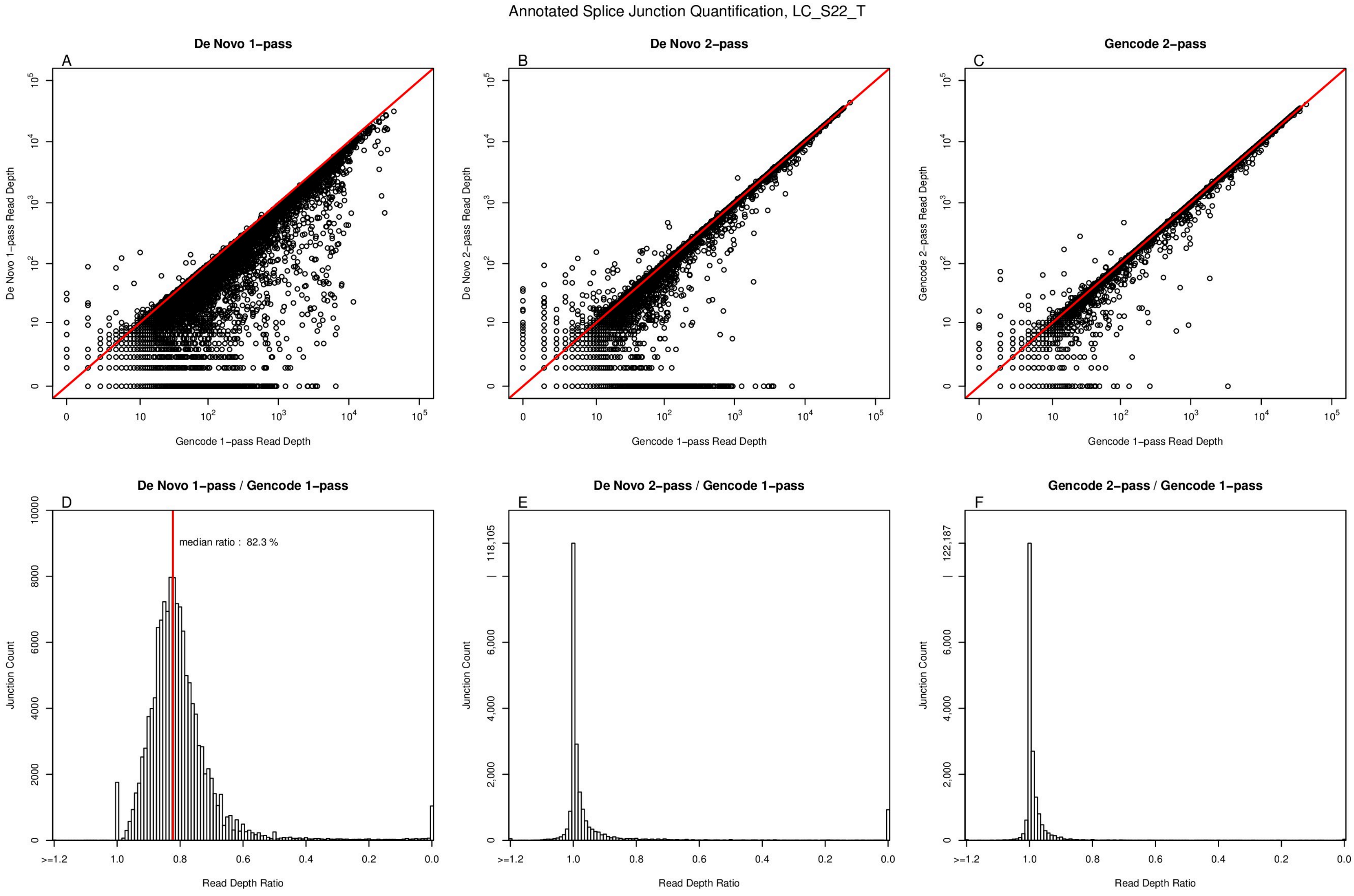


**Figure B.5 Splice Junction Quantification, LC_S22_T**

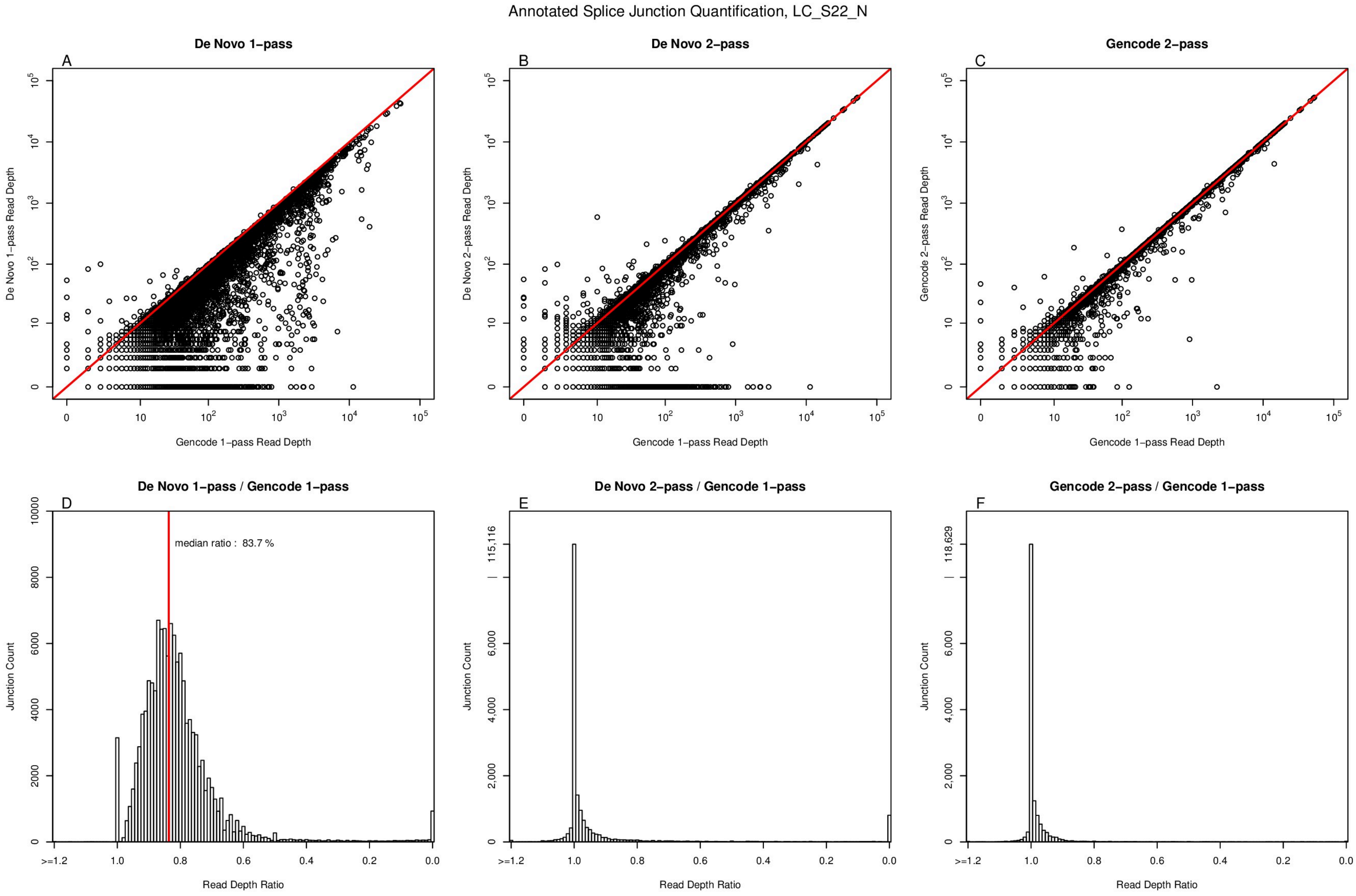


**Figure B.6 Splice Junction Quantification, LC_S22_N**

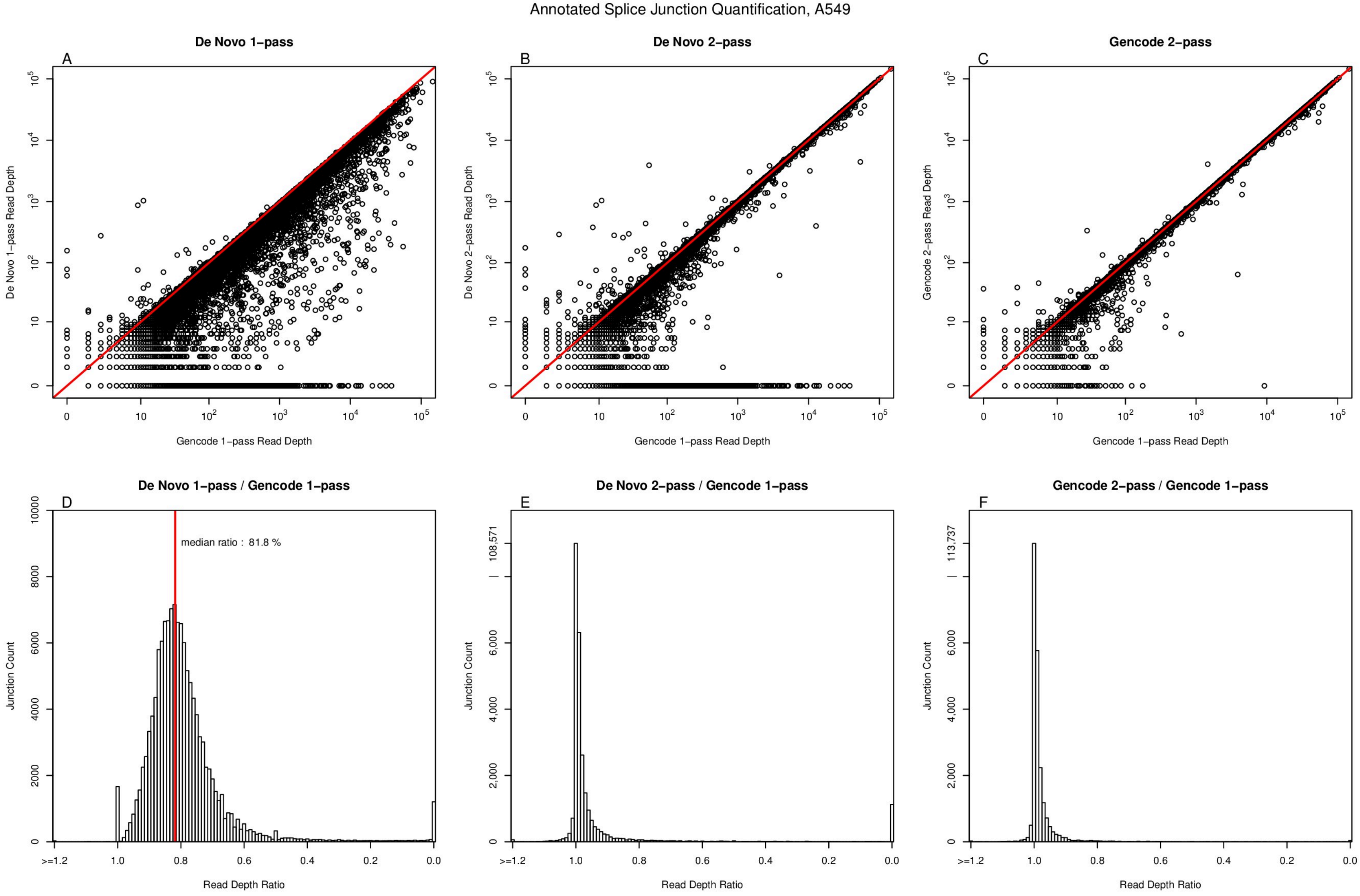


**Figure B.7 Splice Junction Quantification, A549**

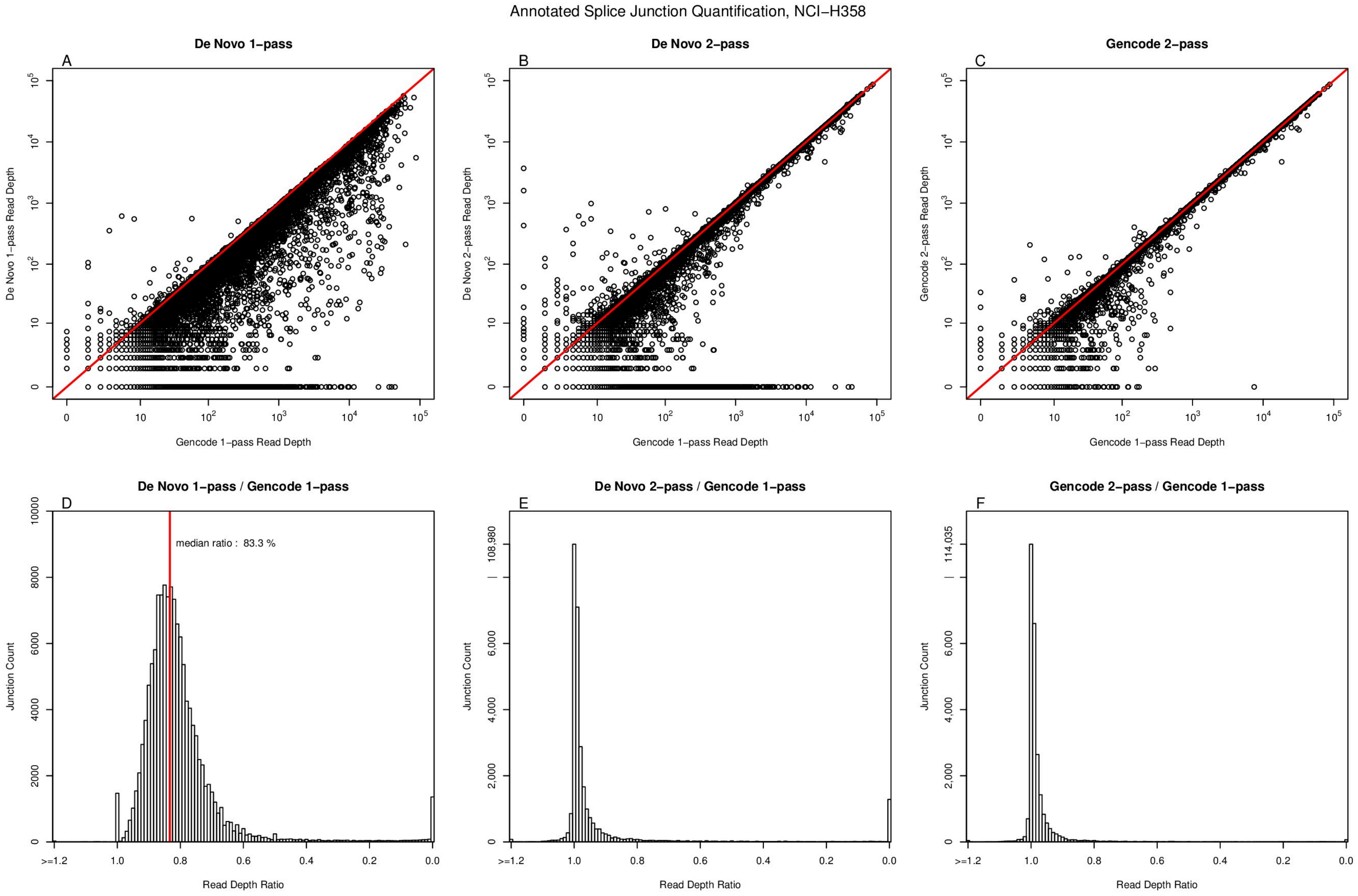


**Figure B.8 Splice Junction Quantification, NCI-H358**

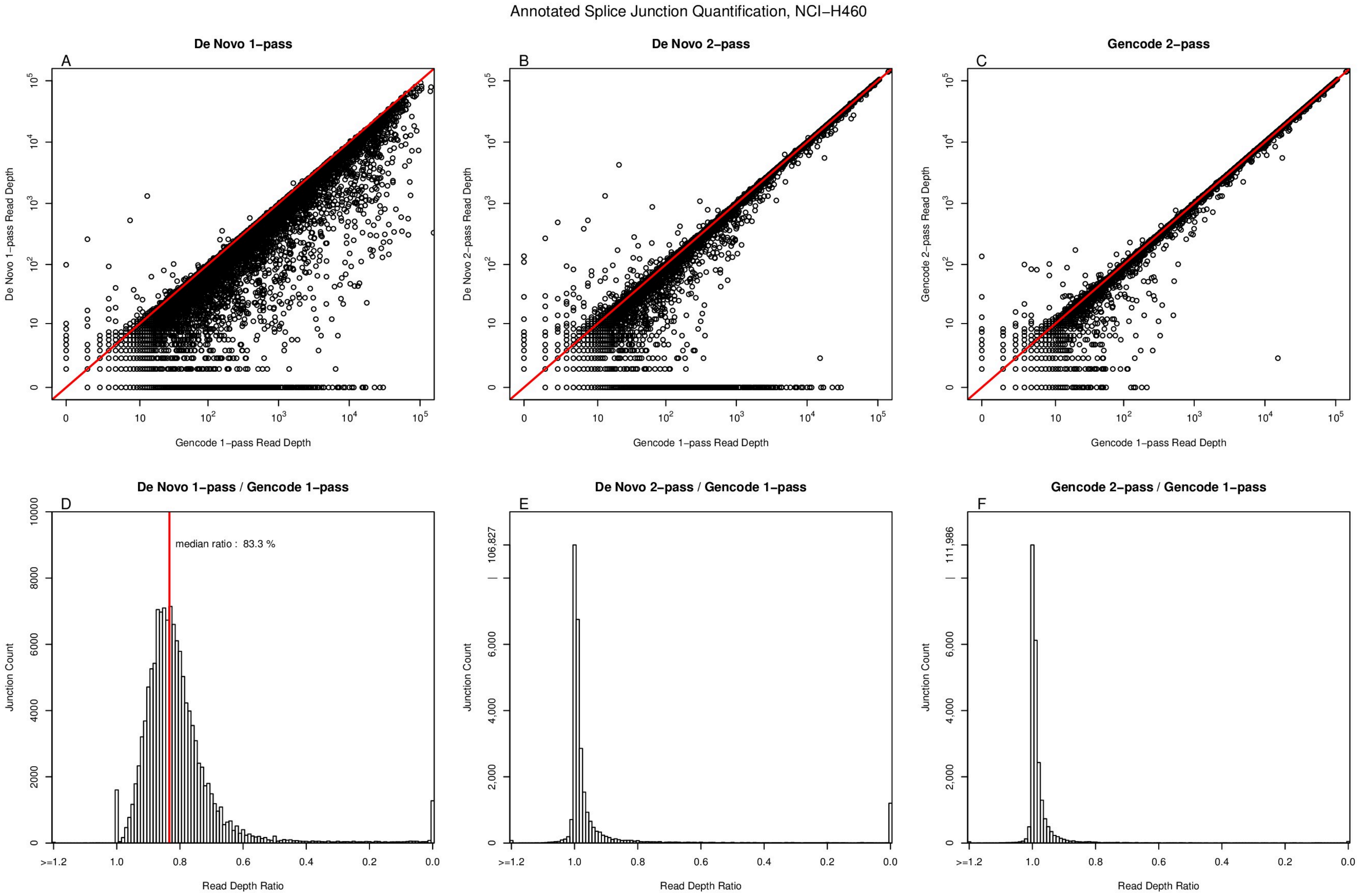


**Figure B.9 Splice Junction Quantification, NCI-H460**

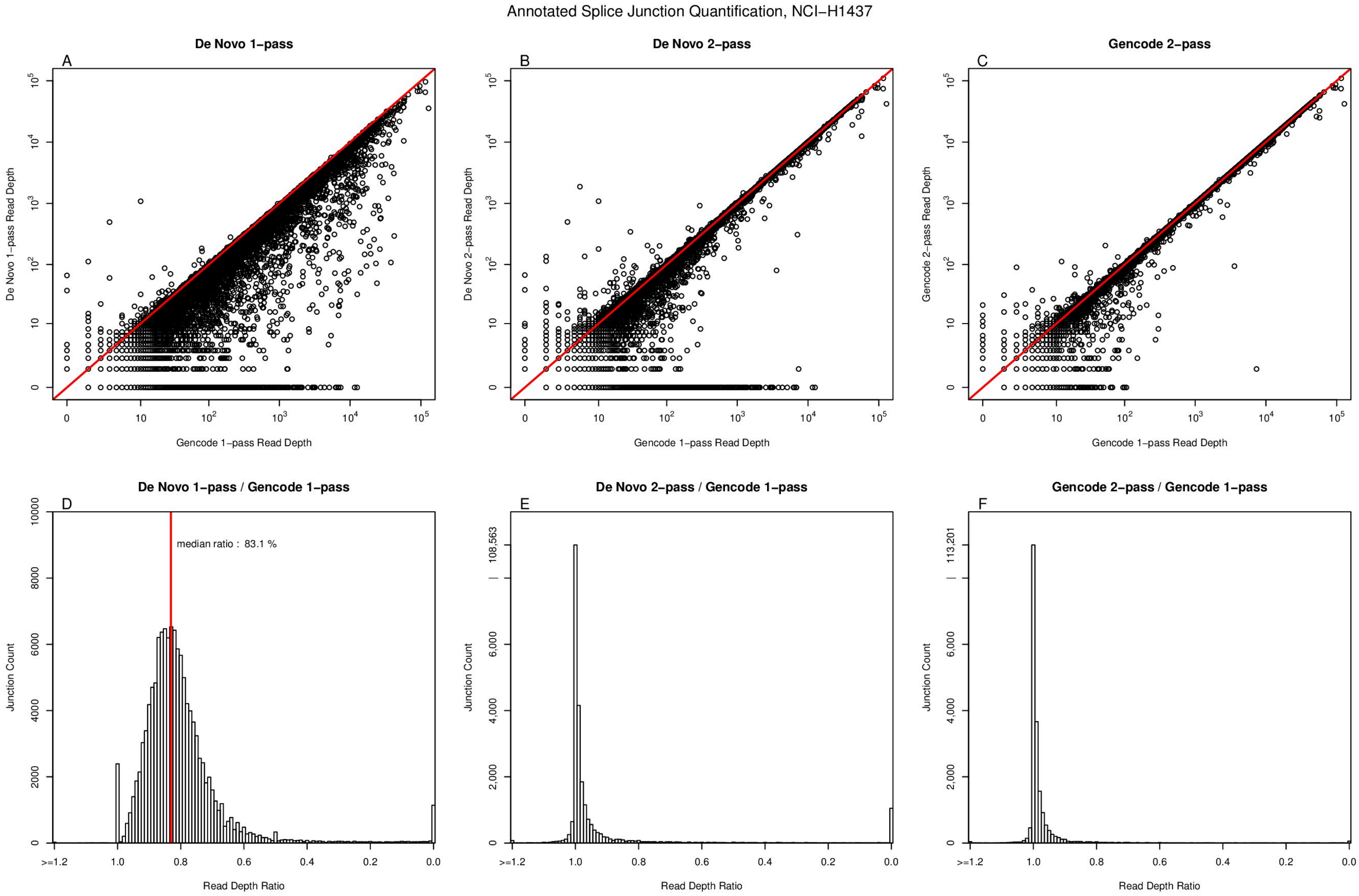
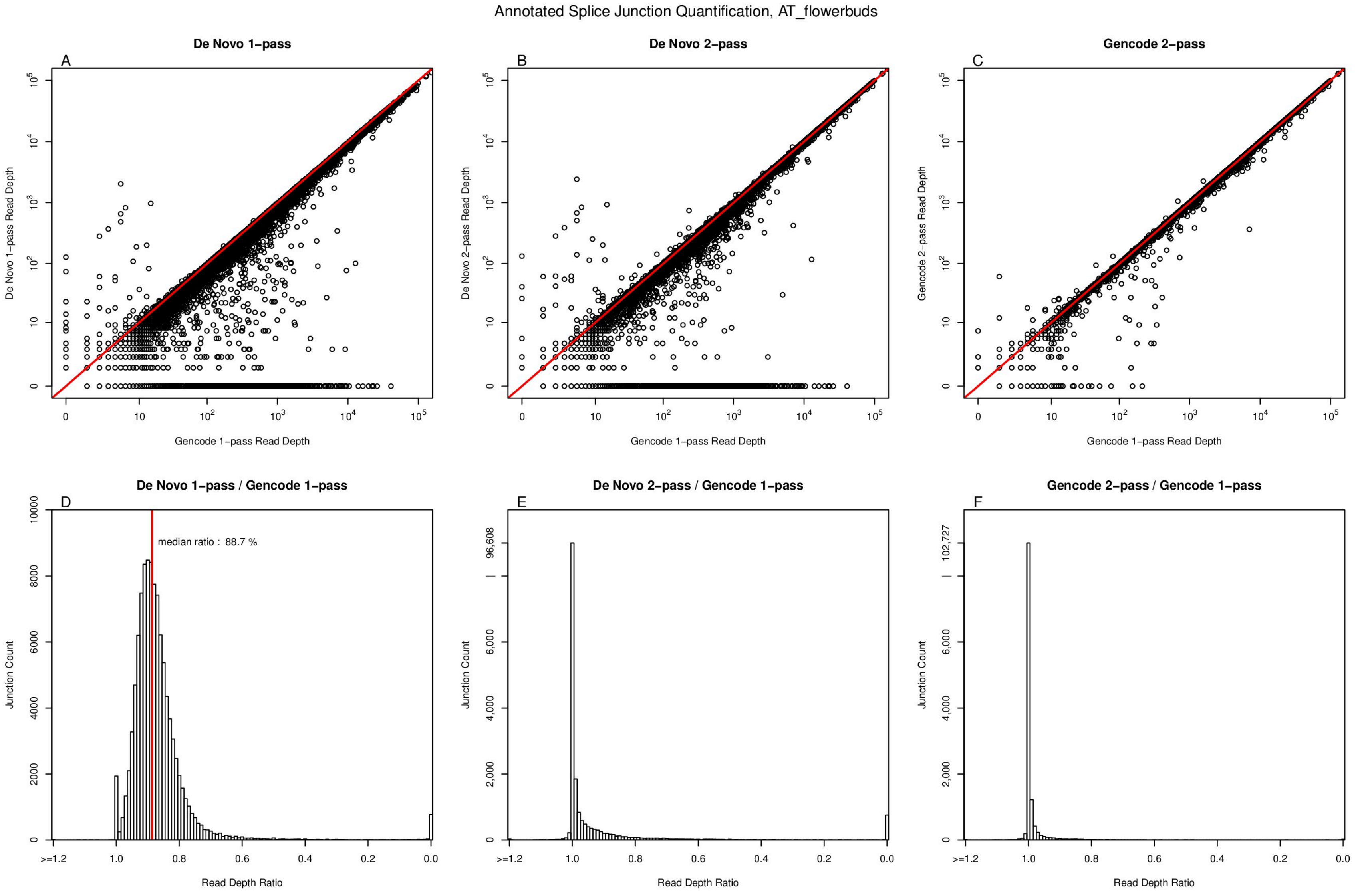


**Figure B.10 Splice Junction Quantification, NCI-H1437**



## Figure B.11 Splice Junction Quantification, AT_flowerbuds

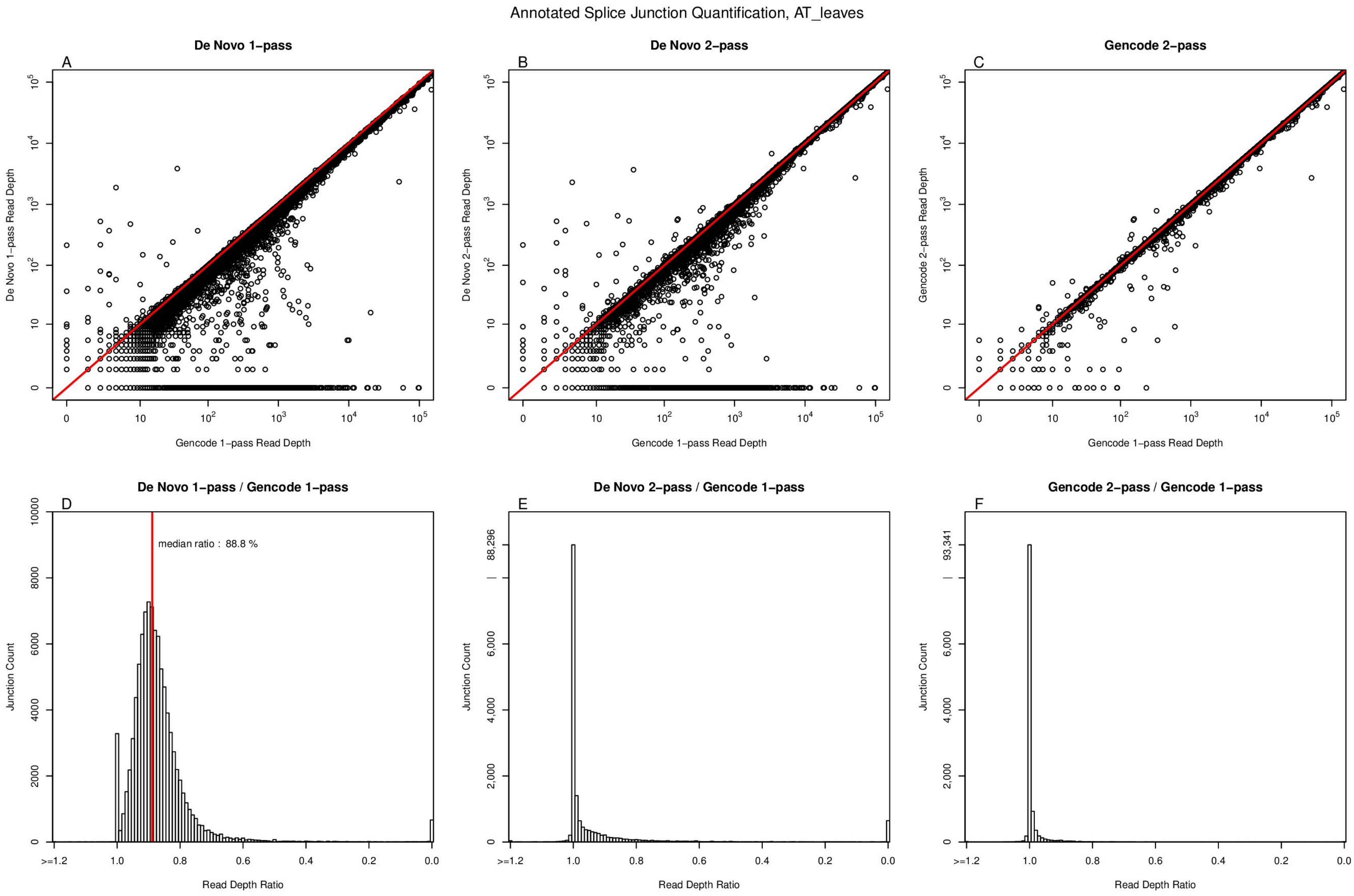


**Figure B.12 Splice Junction Quantification, AT_leaves**

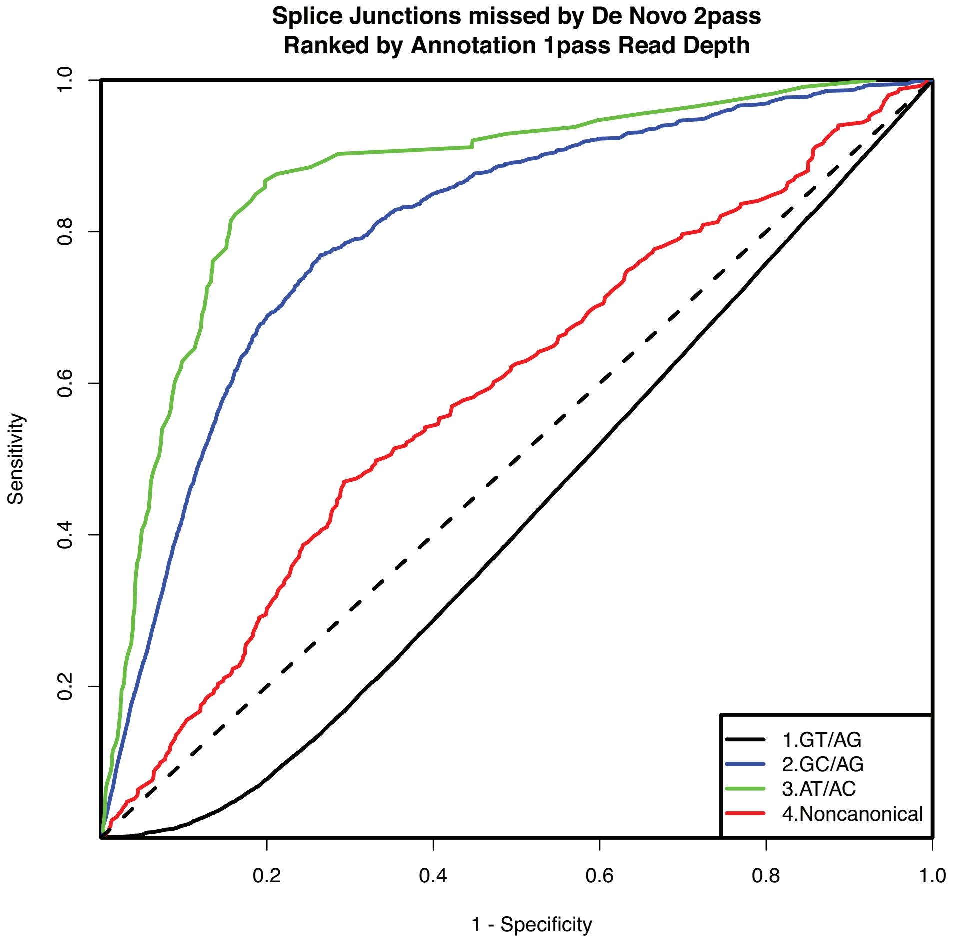


**Figure B.13 Splice Junctions missed by De Novo 2pass Ranked by Annotation 1pass Read Depth**

For splice junctions detected and quantified by 1-pass Annotation alignment of the A549 sample, which were completely missed by 2-pass *De Novo* alignment, we extracted the read depth in the Annotation 1-pass alignment and the internal splice site motif. We then ranked splice junctions in descending order of read depth - splice junctions at the top of the list were the most egregious to miss. Then, we computed ROC-style metrics, where the "sensitivity" for each motif was computed as the running count of observations over the total number, and "specificity" was the percentage of the dataset traversed by that point. For instance, around 90% of the missed AT/AC splice junctions (0.9 sensitivity) were detected in the top 20% of missed splice junctions (0.2 specificity). These statistics were computed for each of the splice site motifs reported - GT/AG, GC/AG, AT/AC, and non-canonical. Y=X is plotted as a dashed black line. Figure B.13 used annotated splice junctions with at least 1 read in Annotation 1-pass, but zero in *De Novo* 2-pass alignment.

### Figures B.14-B.17 Alignment Error Detection

For the A549 and TCGA−50−5933_N samples, we extracted splice junction read depth from *De Novo* 1-pass and *De Novo* 2-pass alignment, the unspliced read depth (calculated as the number of reads unspliced across the splice junction positions by more than ten nucleotides, averaged over the two positions), the percentage of reads spanning the splice junction by less than the length of the exact sequence identity between the unspliced context and the spliced context (in both directions), and finally, the percentage of reads spanning each splice junction by less than 12 nucleotides (calculated as the number of matched bases on either side, from raw SAM data). These data are plotted, split between samples, and split between annotated and unannotated splice junctions, in Figures B.14-B.17 as identified in plot titles. The log10 ratio of read depth is plotted on the Y-axis of the scatterplots (Figures B.14-B.17, panels A-C), and the unspliced read depth, percentage of reads spanning by less than identity, and percentage of reads spanning by less than 12nt are plotted on the X-axes as labeled. Mitochondrial splice junctions, which we considered "true negatives," are colored in red, and a log-ratio of read depth of zero (1:1) is drawn as a black line in each scatterplot. Histograms depicted in (B.14-B.17, panels D-F), are re-illustrations of the same data in panels A-C, to demonstrate density (and share the exact same X-axes). Figures B.14-B.17 used cutoffs of at least one read in *De Novo* 1-pass alignment for the scatterplots, and at least 10 reads in *De Novo* 2-pass alignment for the histograms.

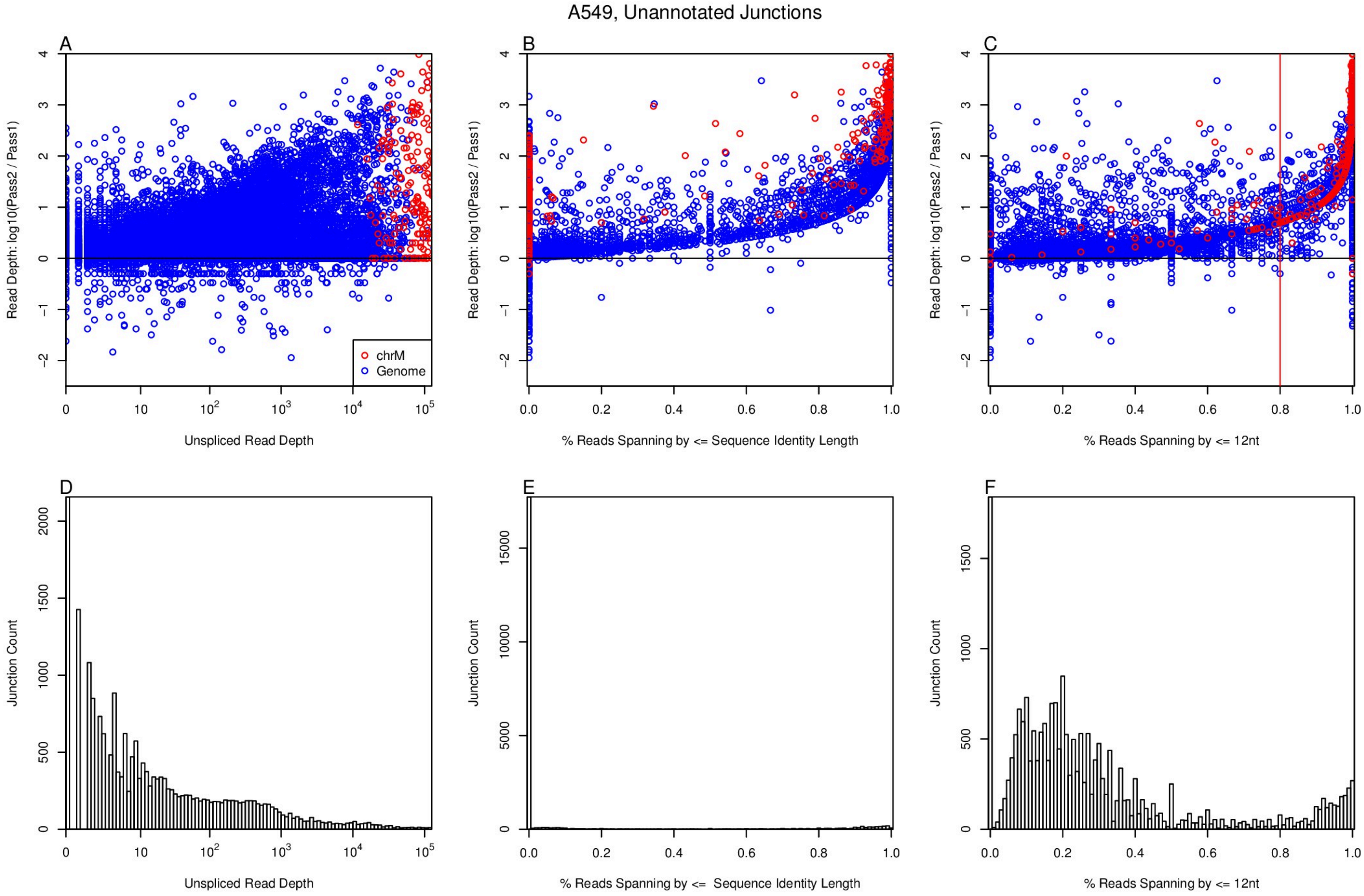


**Figure B.14 Alignment Error Detection, A549 Unannotated Junctions.**

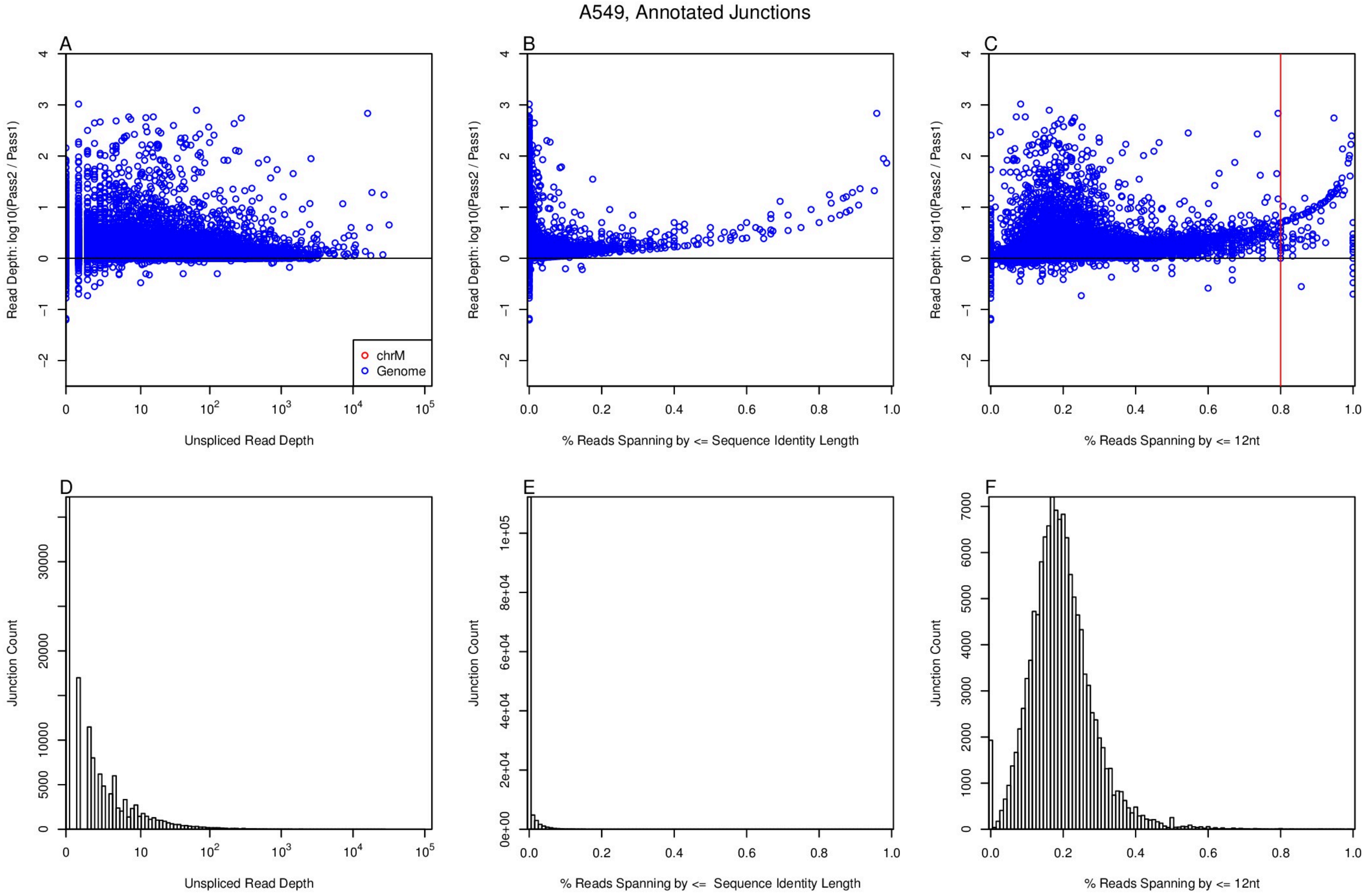


**Figure B.15 Alignment Error Detection, A549 Annotated Junctions.**

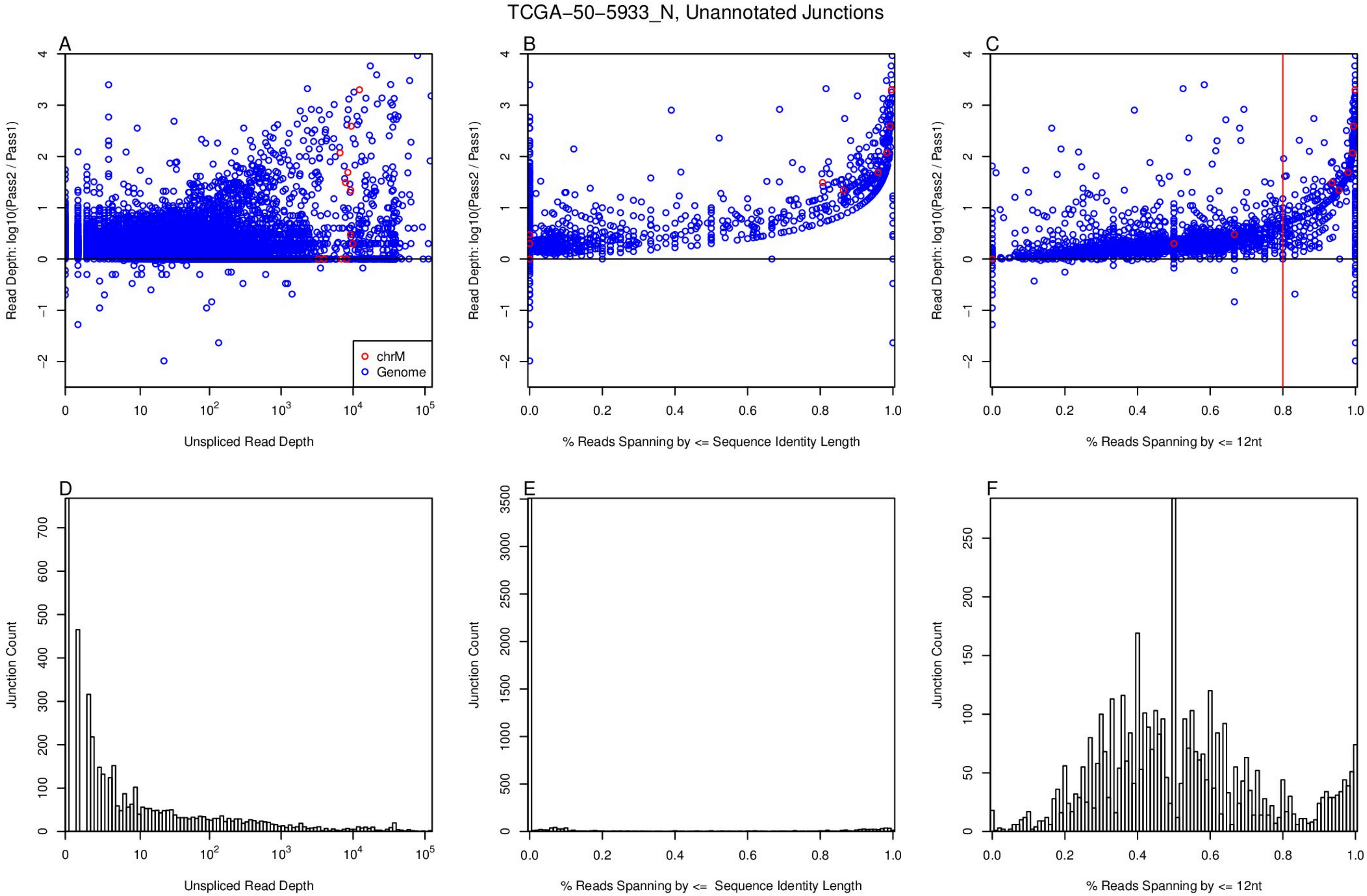


**Figure B.16 Alignment Error Detection, TCGA-50-5933_N Unannotated Junctions.**

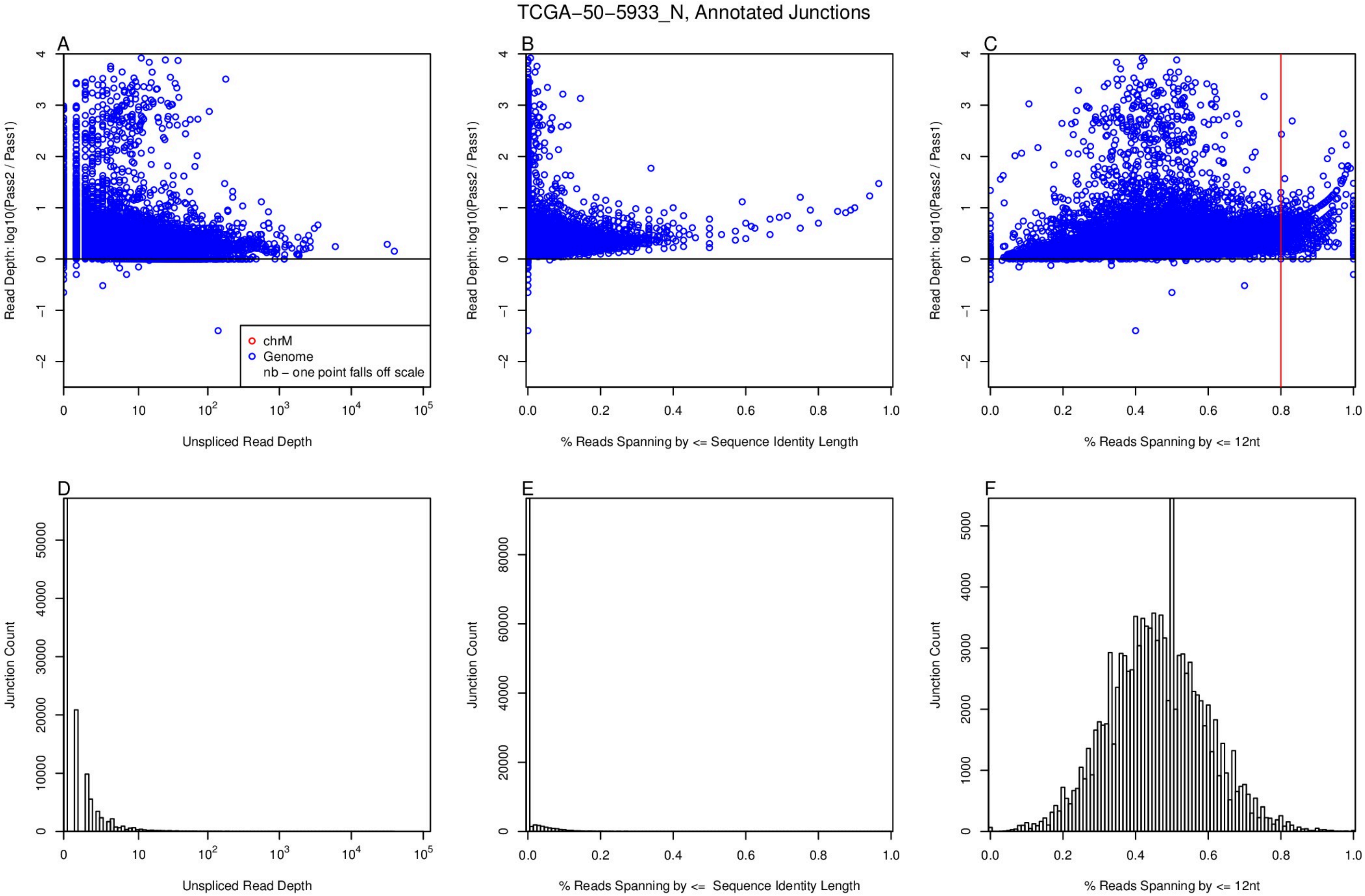


**Figure B.17 Alignment Error Detection, TCGA-50-5933_N Annotated Junctions**

## APPENDIX C
## Supplemental Data for Chapter 5

**Table C.1 Complete Sample Annotation for Prostate RNA-seq Cohorts**

Columns are as follows. Cohort : TCGA.T (primary prostate tumors from TCGA), TCGA.N (normal prostate tissue from TCGA), Michigan.T (primary prostate tumors from Michigan), Michigan.N (normal prostate tissue from Michigan), or SU2C (mCRPC tissue from SU2C). sample : unique sample identifier. RIN : RNA integrity number from the Agilent Bioanalyzer where available, on a scale of 0-10. Gleason : Gleason grade prostate cancer de-differentiation staging as gauged by pathologists, for the largest and section largest tumor sections analyzed ("+"), on a scale of 1 to 5, where available. Biopsy.Site : Where the tissue was biopsied from, one of "Prostate," "Lymph Node," "Liver," "Bone," or "Soft Tissue" for other soft tissue sites. Subtype : Main molecular subtype following TCGA's example, one of "1.ERG," "2.ETV1," "3.ETV4," "4.FLI1," "5.SPOP," "6.FOXA1," "7.IDH1," "8.Other," "9.Normal." For this annotation we aggregated published and internal mutation calls, indel calls, fusion calls, and expression estimation. TC : Tumor Content as estimated by SNV variant fractions from matched exome data, where available. 3' : 3' Bias as estimated by the median log-ratio imbalance between the last and first splice junction of all unambiguous annotated genes. Nascent : Unspliced RNA level as estimated by the median unspliced coverage over junctions from high confidence gene annotations. Aln% : Mapping rate to the genome. AR, Stroma, and NE : Aggregate expression scores for Androgen Receptor signaling, Stromal genes, and Neuroendocrine signaling, as described in Figures 4.2, 4.3, and in Appendix C.

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | 2A-A8VL | 8.7 | 3+3 | Prostate | 1.ERG | 0.51 | 0.93 | 0.00 | 0.91 | 2.18 | -0.21 | 0.04 |
| TCGA.T | 2A-A8VT | 9 | 4+3 | Prostate | 1.ERG | 0.73 | 0.45 | 0.00 | 0.91 | -0.18 | -3.35 | 0.79 |
| TCGA.T | 2A-A8VV | 8.5 | 3+3 | Prostate | 1.ERG | 0.84 | 0.87 | 0.00 | 0.87 | 1.94 | 0.63 | 1.21 |
| TCGA.T | CH-5739 | 8.3 | 4+3 | Prostate | 1.ERG | 0.69 | 0.74 | 0.00 | 0.91 | 1.64 | 0.66 | 0.58 |
| TCGA.T | CH-5740 | 7.2 | 4+4 | Prostate | 1.ERG | 0.69 | 0.85 | 0.00 | 0.80 | 1.41 | -2.56 | -0.66 |
| TCGA.T | CH-5741 | 7.6 | 5+4 | Prostate | 1.ERG | 0.82 | 1.02 | 0.00 | 0.90 | 2.41 | -5.74 | 1.97 |
| TCGA.T | CH-5743 | 7.2 | 4+3 | Prostate | 1.ERG | NA | 0.96 | 0.00 | 0.92 | -2.6 | 4.87 | -3.43 |
| TCGA.T | CH-5744 | 7.4 | 4+4 | Prostate | 1.ERG | 0.60 | 0.74 | 0.00 | 0.91 | 1.04 | -3.53 | 1.41 |
| TCGA.T | CH-5746 | 7.5 | 3+3 | Prostate | 1.ERG | 0.36 | 0.84 | 0.00 | 0.89 | 1.76 | -0.22 | -0.11 |
| TCGA.T | CH-5752 | 9.2 | 5+4 | Prostate | 1.ERG | 0.79 | 0.78 | 0.00 | 0.90 | 0.28 | -2.25 | 0.45 |
| TCGA.T | CH-5754 | 8.4 | 4+4 | Prostate | 1.ERG | 0.65 | 0.74 | 0.00 | 0.90 | -0.31 | -1.57 | 1.11 |
| TCGA.T | CH-5764 | 8.8 | 3+4 | Prostate | 1.ERG | 0.56 | 0.77 | 0.00 | 0.87 | 0.88 | -0.08 | -0.15 |
| TCGA.T | CH-5765 | 8.3 | 3+4 | Prostate | 1.ERG | 0.76 | 0.84 | 0.00 | 0.89 | 1.47 | -2.09 | 1.45 |
| TCGA.T | CH-5766 | 9 | 4+3 | Prostate | 1.ERG | 0.61 | 0.81 | 0.01 | 0.89 | 1.06 | -1.73 | -0.62 |
| TCGA.T | CH-5768 | 8.5 | 3+3 | Prostate | 1.ERG | 0.70 | 0.93 | 0.00 | 0.89 | 1.83 | -2.76 | 1.1 |
| TCGA.T | CH-5769 | 8.2 | 4+5 | Prostate | 1.ERG | 0.63 | 0.84 | 0.00 | 0.88 | -1.13 | 0.33 | 0.94 |
| TCGA.T | CH-5789 | 8.5 | 4+3 | Prostate | 1.ERG | NA | 0.77 | 0.00 | 0.90 | 1.97 | 3.01 | -2.27 |
| TCGA.T | CH-5790 | 8.9 | 3+4 | Prostate | 1.ERG | 0.81 | 0.72 | 0.00 | 0.88 | 2.11 | -3.26 | 1.48 |
| TCGA.T | CH-5791 | 8.1 | 4+3 | Prostate | 1.ERG | 0.67 | 0.76 | 0.00 | 0.88 | 0.47 | -2.76 | 1.19 |
| TCGA.T | CH-5794 | 7.9 | 4+3 | Prostate | 1.ERG | 0.58 | 0.79 | 0.00 | 0.92 | -0.27 | -1.88 | 0.26 |
| TCGA.T | EJ-5495 | 8.8 | 4+5 | Prostate | 1.ERG | 0.41 | 0.70 | 0.00 | 0.90 | -0.64 | 4.85 | -3.49 |
| TCGA.T | EJ-5496 | 8.4 | 3+3 | Prostate | 1.ERG | 0.49 | 0.91 | 0.00 | 0.87 | 3.35 | -1.34 | 0.4 |
| TCGA.T | EJ-5497 | 9 | 3+4 | Prostate | 1.ERG | 0.46 | 0.77 | 0.00 | 0.89 | 3.67 | 1.49 | -1.78 |
| TCGA.T | EJ-5498 | 9 | 3+4 | Prostate | 1.ERG | 0.24 | 0.90 | 0.00 | 0.89 | 1.08 | 6.31 | -4.51 |
| TCGA.T | EJ-5499 | 8.3 | 3+4 | Prostate | 1.ERG | 0.48 | 1.03 | 0.00 | 0.88 | -0.6 | 0.3 | -0.6 |
| TCGA.T | EJ-5502 | 8.1 | 3+4 | Prostate | 1.ERG | NA | 0.69 | 0.00 | 0.92 | 1.42 | 2.61 | -2.29 |
| TCGA.T | EJ-5503 | 7.9 | 4+5 | Prostate | 1.ERG | NA | 0.75 | 0.00 | 0.93 | 2.31 | 4.53 | -3.08 |
| TCGA.T | EJ-5506 | 9.1 | 4+3 | Prostate | 1.ERG | NA | 0.80 | 0.00 | 0.90 | 1.5 | 2.68 | -2.19 |
| TCGA.T | EJ-5507 | 9.4 | 4+5 | Prostate | 1.ERG | 0.68 | 0.85 | 0.00 | 0.91 | -0.46 | -1.72 | 0.2 |
| TCGA.T | EJ-5508 | 8.3 | 3+4 | Prostate | 1.ERG | 0.36 | 0.77 | 0.00 | 0.89 | 2.03 | 1.76 | -1.45 |
| TCGA.T | EJ-5512 | 7.8 | 3+3 | Prostate | 1.ERG | 0.39 | 1.13 | 0.00 | 0.89 | 1.74 | 3.53 | -1.85 |
| TCGA.T | EJ-5516 | 9.1 | 4+3 | Prostate | 1.ERG | 0.39 | 0.81 | 0.00 | 0.91 | 1.98 | 1.74 | -1.64 |
| TCGA.T | EJ-5521 | 8.2 | 4+4 | Prostate | 1.ERG | 0.51 | 0.89 | 0.00 | 0.93 | -0.51 | -0.52 | -0.18 |
| TCGA.T | EJ-5522 | 9.6 | 3+4 | Prostate | 1.ERG | 0.51 | 0.78 | 0.00 | 0.92 | 2.43 | 1.25 | -1.98 |
| TCGA.T | EJ-5524 | 9.6 | 4+3 | Prostate | 1.ERG | 0.54 | 0.82 | 0.00 | 0.92 | 1.71 | 2.68 | -0.96 |
| TCGA.T | EJ-5525 | 9 | 4+5 | Prostate | 1.ERG | 0.87 | 0.91 | 0.00 | 0.90 | 0.08 | -1.07 | 1.57 |
| TCGA.T | EJ-5526 | 9 | 4+5 | Prostate | 1.ERG | 0.51 | 0.81 | 0.00 | 0.91 | 2.12 | 2.67 | -1.15 |
| TCGA.T | EJ-5527 | 9.5 | 4+3 | Prostate | 1.ERG | 0.51 | 0.83 | 0.00 | 0.91 | 1.54 | 2.12 | -2.12 |
| TCGA.T | EJ-5530 | 9.7 | 3+4 | Prostate | 1.ERG | 0.53 | 0.54 | 0.00 | 0.92 | 2.58 | -0.56 | -1.03 |
| TCGA.T | EJ-5542 | 9.8 | 3+4 | Prostate | 1.ERG | 0.58 | 0.92 | 0.00 | 0.92 | 2.7 | 1.32 | -1.74 |
| TCGA.T | EJ-7314 | 8.4 | 4+3 | Prostate | 1.ERG | 0.61 | 0.85 | 0.00 | 0.90 | 1.96 | -2.06 | 0.37 |
| TCGA.T | EJ-7315 | 8.2 | 4+3 | Prostate | 1.ERG | 0.65 | 0.89 | 0.00 | 0.90 | -0.17 | 1.15 | -0.78 |
| TCGA.T | EJ-7321 | 8.7 | 3+3 | Prostate | 1.ERG | 0.72 | 0.82 | 0.00 | 0.90 | 0.93 | -2.63 | 1.49 |
| TCGA.T | EJ-7327 | 9.5 | 4+3 | Prostate | 1.ERG | 0.48 | 0.66 | 0.00 | 0.90 | 0.64 | 2.7 | -2.02 |
| TCGA.T | EJ-7328 | 7.7 | 4+3 | Prostate | 1.ERG | 0.44 | 0.79 | 0.00 | 0.89 | -1.71 | 3.2 | -1.04 |
| TCGA.T | EJ-7783 | 9.3 | 4+4 | Prostate | 1.ERG | 0.41 | 0.72 | 0.00 | 0.90 | 0.75 | 3.08 | -1.98 |
| TCGA.T | EJ-7784 | 9.5 | 4+3 | Prostate | 1.ERG | 0.77 | 0.65 | 0.00 | 0.91 | 2.28 | -2.68 | 0.86 |

**Table C.1 01/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | EJ-7785 | 9.4 | 3+3 | Prostate | 1.ERG | 0.43 | 0.63 | 0.00 | 0.92 | 1.94 | 5.39 | -1.8 |
| TCGA.T | EJ-7793 | 7.7 | 3+4 | Prostate | 1.ERG | 0.33 | 0.82 | 0.00 | 0.92 | 4.05 | 2.42 | -1 |
| TCGA.T | EJ-7797 | 8 | 3+4 | Prostate | 1.ERG | 0.38 | 0.77 | 0.00 | 0.92 | 2.35 | 0.78 | -1.22 |
| TCGA.T | EJ-8469 | 9.3 | 4+5 | Prostate | 1.ERG | 0.79 | 0.68 | 0.00 | 0.92 | 2.01 | -3.1 | -0.31 |
| TCGA.T | EJ-8472 | 8.9 | 4+4 | Prostate | 1.ERG | 0.75 | 1.08 | 0.00 | 0.91 | 0.59 | 0.57 | 1.18 |
| TCGA.T | EJ-A46D | 7.2 | 3+3 | Prostate | 1.ERG | 0.20 | 0.79 | 0.00 | 0.77 | 1.14 | 2.11 | -1.73 |
| TCGA.T | EJ-A65F | 8.1 | 4+3 | Prostate | 1.ERG | 0.67 | 0.80 | 0.00 | 0.91 | 1.32 | -1.45 | -0.57 |
| TCGA.T | EJ-A7NF | 8.6 | 3+4 | Prostate | 1.ERG | 0.76 | 0.90 | 0.00 | 0.89 | 1.73 | -3.66 | 2.1 |
| TCGA.T | EJ-A7NG | 8 | 3+4 | Prostate | 1.ERG | NA | 0.69 | 0.00 | 0.89 | -0.6 | 5.96 | -3.57 |
| TCGA.T | EJ-A7NK | 7.6 | 3+3 | Prostate | 1.ERG | NA | 0.87 | 0.00 | 0.79 | 0.26 | 4.1 | -2.91 |
| TCGA.T | FC-7708 | 9.2 | 3+3 | Prostate | 1.ERG | 0.43 | 0.76 | 0.00 | 0.91 | 0.91 | 3.56 | -2.36 |
| TCGA.T | G9-6329 | 8 | 3+4 | Prostate | 1.ERG | 0.29 | 1.11 | 0.00 | 0.89 | -1.73 | 3.71 | -2.68 |
| TCGA.T | G9-6336 | 7.7 | 3+3 | Prostate | 1.ERG | 0.26 | 1.07 | 0.00 | 0.91 | 1.06 | 0.43 | -0.42 |
| TCGA.T | G9-6342 | 8 | 4+3 | Prostate | 1.ERG | 0.53 | 1.00 | 0.00 | 0.89 | 0.42 | 0.59 | -0.78 |
| TCGA.T | G9-6351 | 8.5 | 3+3 | Prostate | 1.ERG | 0.34 | 1.00 | 0.00 | 0.90 | 3.34 | 1.14 | -0.61 |
| TCGA.T | G9-6353 | 7.9 | 3+3 | Prostate | 1.ERG | NA | 0.99 | 0.00 | 0.90 | 1.39 | 5.08 | -2.55 |
| TCGA.T | G9-6356 | 8.7 | 3+4 | Prostate | 1.ERG | 0.28 | 0.98 | 0.00 | 0.89 | -1.14 | 2.71 | -3.15 |
| TCGA.T | G9-6361 | 7.9 | 3+4 | Prostate | 1.ERG | 0.71 | 1.03 | 0.00 | 0.91 | 0.37 | 3.65 | -0.13 |
| TCGA.T | G9-6363 | 8.4 | 4+3 | Prostate | 1.ERG | 0.51 | 1.07 | 0.00 | 0.89 | -0.19 | -0.43 | 0.5 |
| TCGA.T | G9-6364 | 8.8 | 3+4 | Prostate | 1.ERG | 0.61 | 1.24 | 0.00 | 0.53 | -1.02 | 4.44 | -2.95 |
| TCGA.T | G9-6365 | 7.6 | 3+4 | Prostate | 1.ERG | 0.54 | 1.08 | 0.00 | 0.90 | -1.46 | 4.35 | -2.6 |
| TCGA.T | G9-6377 | 8.2 | 3+4 | Prostate | 1.ERG | 0.66 | 0.98 | 0.00 | 0.90 | 1.58 | -2.45 | 0.88 |
| TCGA.T | G9-6384 | 9 | 3+4 | Prostate | 1.ERG | 0.46 | 0.70 | 0.00 | 0.91 | 2.09 | 2.16 | -1.06 |
| TCGA.T | G9-6385 | 7.4 | 3+4 | Prostate | 1.ERG | NA | 1.04 | 0.00 | 0.91 | 1.49 | 3.09 | -2.32 |
| TCGA.T | G9-7522 | 8.7 | 3+4 | Prostate | 1.ERG | 0.41 | 0.82 | 0.00 | 0.93 | 1.72 | 2.32 | -1.31 |
| TCGA.T | HC-7077 | 8 | 3+4 | Prostate | 1.ERG | 0.78 | 0.71 | 0.00 | 0.90 | 2.58 | -5.76 | 3.68 |
| TCGA.T | HC-7081 | 9.2 | 4+3 | Prostate | 1.ERG | 0.57 | 0.57 | 0.00 | 0.92 | -0.32 | 2.37 | -1.76 |
| TCGA.T | HC-7209 | 9.1 | 3+4 | Prostate | 1.ERG | 0.57 | 0.59 | 0.00 | 0.91 | 1.63 | -1.66 | -0.43 |
| TCGA.T | HC-7211 | 9.5 | 4+5 | Prostate | 1.ERG | 0.69 | 0.67 | 0.00 | 0.91 | 2.33 | -2.29 | 0.61 |
| TCGA.T | HC-7212 | 9.6 | 4+4 | Prostate | 1.ERG | 0.81 | 0.62 | 0.00 | 0.91 | 2.5 | -3.25 | 1.13 |
| TCGA.T | HC-7213 | 9.4 | 4+5 | Prostate | 1.ERG | 0.78 | 0.74 | 0.00 | 0.92 | 0.87 | -1.46 | 1.52 |
| TCGA.T | HC-7230 | 9.1 | 3+3 | Prostate | 1.ERG | 0.85 | 0.78 | 0.00 | 0.92 | 2.99 | -3.45 | 2.04 |
| TCGA.T | HC-7231 | 8.4 | 3+4 | Prostate | 1.ERG | 0.75 | 0.75 | 0.00 | 0.91 | 0.56 | -2.49 | 0.26 |
| TCGA.T | HC-7232 | 9.1 | 3+4 | Prostate | 1.ERG | 0.60 | 0.79 | 0.00 | 0.92 | 0.71 | 1.82 | -1.24 |
| TCGA.T | HC-7744 | 8.8 | 4+5 | Prostate | 1.ERG | 0.82 | 0.75 | 0.00 | 0.91 | 1.49 | -3.38 | 1.54 |
| TCGA.T | HC-7747 | 7.9 | 3+4 | Prostate | 1.ERG | 0.38 | 0.83 | 0.00 | 0.90 | 1.42 | -0.14 | -2.18 |
| TCGA.T | HC-7748 | 9.6 | 3+4 | Prostate | 1.ERG | 0.50 | 0.60 | 0.00 | 0.92 | 2.55 | 2.06 | -1.39 |
| TCGA.T | HC-7818 | 8 | 3+3 | Prostate | 1.ERG | NA | 0.77 | 0.00 | 0.92 | 1.06 | 0.65 | -1.03 |
| TCGA.T | HC-7820 | 7.8 | 3+3 | Prostate | 1.ERG | 0.48 | 0.75 | 0.00 | 0.91 | 3.51 | -1.02 | -0.65 |
| TCGA.T | HC-7821 | 7.5 | 4+5 | Prostate | 1.ERG | 0.85 | 0.71 | 0.00 | 0.91 | 1.63 | -1.81 | -0.72 |
| TCGA.T | HC-8213 | 8.3 | 3+3 | Prostate | 1.ERG | 0.86 | 1.16 | 0.00 | 0.89 | 1.61 | -5.38 | 0.3 |
| TCGA.T | HC-8257 | 7.7 | 3+4 | Prostate | 1.ERG | 0.74 | 0.91 | 0.00 | 0.91 | 0.89 | -3.01 | -0.46 |
| TCGA.T | HC-8260 | 8.1 | 3+3 | Prostate | 1.ERG | NA | 0.93 | 0.01 | 0.87 | 2.47 | 0.89 | -1.23 |
| TCGA.T | HC-8262 | 8.7 | 3+4 | Prostate | 1.ERG | 0.57 | 0.97 | 0.00 | 0.90 | 0.92 | -1.76 | -0.61 |
| TCGA.T | HC-A48F | 9.2 | 4+3 | Prostate | 1.ERG | 0.91 | 0.97 | 0.00 | 0.92 | 0.02 | -4.31 | 1.1 |
| TCGA.T | HC-A632 | 8.4 | 4+5 | Prostate | 1.ERG | 0.71 | 0.88 | 0.00 | 0.89 | 0.41 | -0.61 | 0.64 |
| TCGA.T | HC-A76X | 9.1 | 3+3 | Prostate | 1.ERG | 0.70 | 0.74 | 0.00 | 0.91 | 2.45 | -4.68 | 1.99 |

**Table C.1 02/13**

| | | | | | | | | | | Expression Scores | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | HC-A8D0 | 8.6 | 4+4 | Prostate | 1.ERG | 0.44 | 0.79 | 0.00 | 0.90 | -2.68 | 2.43 | -1.15 |
| TCGA.T | HC-A8D1 | 9.1 | 3+4 | Prostate | 1.ERG | 0.42 | 0.69 | 0.00 | 0.90 | 0.6 | 1.59 | -1.08 |
| TCGA.T | HI-7171 | 9 | 4+4 | Prostate | 1.ERG | 0.88 | 0.97 | 0.00 | 0.90 | 0.19 | -2.78 | 1.94 |
| TCGA.T | J4-8198 | 9.2 | 3+4 | Prostate | 1.ERG | 0.68 | 0.80 | 0.00 | 0.90 | 0.91 | 1.8 | -0.57 |
| TCGA.T | J4-A67T | 7.9 | 3+4 | Prostate | 1.ERG | 0.51 | 0.99 | 0.00 | 0.91 | 1.43 | 0.3 | -0.42 |
| TCGA.T | J4-A6G1 | 8.4 | 4+3 | Prostate | 1.ERG | 0.45 | 1.00 | 0.00 | 0.89 | -1.46 | 2.95 | -1 |
| TCGA.T | J4-A6M7 | 8.9 | 3+3 | Prostate | 1.ERG | 0.81 | 0.79 | 0.00 | 0.91 | 2.59 | -0.64 | 0.06 |
| TCGA.T | J4-A83I | 9.5 | 4+3 | Prostate | 1.ERG | 0.82 | 0.66 | 0.00 | 0.90 | 1.03 | -2.66 | 0.89 |
| TCGA.T | J4-A83K | 9.3 | 3+3 | Prostate | 1.ERG | 0.60 | 0.72 | 0.00 | 0.90 | 0.94 | 3.4 | -0.64 |
| TCGA.T | J4-A83N | 9.2 | 3+3 | Prostate | 1.ERG | 0.62 | 0.94 | 0.00 | 0.86 | 2.36 | -3.74 | 1.6 |
| TCGA.T | J9-A52B | 8.1 | 4+5 | Prostate | 1.ERG | 0.64 | 1.07 | 0.00 | 0.91 | -1.44 | -0.55 | -3.84 |
| TCGA.T | J9-A8CK | 7.3 | 4+3 | Prostate | 1.ERG | 0.56 | 0.85 | 0.00 | 0.87 | -0.75 | 2.81 | -0.94 |
| TCGA.T | J9-A8CM | 8.4 | 3+4 | Prostate | 1.ERG | 0.64 | 0.64 | 0.00 | 0.81 | -2.4 | 3.05 | -0.81 |
| TCGA.T | KC-A4BN | 8.6 | 3+3 | Prostate | 1.ERG | 0.32 | 0.94 | 0.00 | 0.91 | 2.89 | -1.6 | 0.78 |
| TCGA.T | KC-A4BR | 7.1 | 3+4 | Prostate | 1.ERG | NA | 0.56 | 0.00 | 0.88 | -1.19 | 6.63 | -2.14 |
| TCGA.T | KC-A4BV | 8.9 | 4+3 | Prostate | 1.ERG | 0.67 | 0.81 | 0.00 | 0.91 | 0.18 | 0.54 | -0.9 |
| TCGA.T | KC-A7F6 | 8.1 | 3+4 | Prostate | 1.ERG | 0.66 | 0.86 | 0.00 | 0.88 | -0.12 | -0.08 | 0.36 |
| TCGA.T | KK-A59Y | 9.5 | 4+5 | Prostate | 1.ERG | 0.77 | 1.08 | 0.00 | 0.91 | 0.44 | -3.99 | 1.4 |
| TCGA.T | KK-A6DY | 7.8 | 3+4 | Prostate | 1.ERG | 0.47 | 0.87 | 0.01 | 0.91 | -0.52 | -0.54 | 1.21 |
| TCGA.T | KK-A6E1 | 8.8 | 4+5 | Prostate | 1.ERG | 0.74 | 0.82 | 0.00 | 0.91 | 1.31 | -5.47 | 2.2 |
| TCGA.T | KK-A6E2 | 8.5 | 3+4 | Prostate | 1.ERG | 0.86 | 0.71 | 0.00 | 0.90 | 1.63 | -4.7 | 2.4 |
| TCGA.T | KK-A6E6 | 8.3 | 4+3 | Prostate | 1.ERG | 0.46 | 0.97 | 0.00 | 0.91 | -0.63 | 0.3 | -0.15 |
| TCGA.T | KK-A7AU | 8.9 | 4+3 | Prostate | 1.ERG | 0.86 | 0.82 | 0.00 | 0.91 | 1.15 | -8.37 | 3.45 |
| TCGA.T | KK-A7B1 | 9.1 | 4+3 | Prostate | 1.ERG | 0.57 | 0.79 | 0.00 | 0.92 | 0.05 | 1.35 | -0.53 |
| TCGA.T | KK-A7B4 | 9.1 | 4+5 | Prostate | 1.ERG | 0.74 | 0.77 | 0.00 | 0.91 | -0.19 | -3.67 | 1.87 |
| TCGA.T | KK-A8I4 | 7.6 | 4+3 | Prostate | 1.ERG | 0.43 | 0.66 | 0.00 | 0.91 | -1.7 | 2.94 | -0.66 |
| TCGA.T | KK-A8I5 | 8 | 3+3 | Prostate | 1.ERG | 0.85 | 0.82 | 0.00 | 0.88 | 0.77 | 0.37 | 0.22 |
| TCGA.T | KK-A8I6 | 8.1 | 4+3 | Prostate | 1.ERG | 0.68 | 0.74 | 0.00 | 0.88 | 2.26 | -2.6 | 0.74 |
| TCGA.T | KK-A8I8 | 8.4 | 3+4 | Prostate | 1.ERG | 0.74 | 0.83 | 0.00 | 0.88 | -0.62 | -3.06 | 1.4 |
| TCGA.T | KK-A8IA | 8 | 4+4 | Prostate | 1.ERG | 0.87 | 0.82 | 0.00 | 0.91 | 0.4 | -4.24 | 3.08 |
| TCGA.T | KK-A8IC | 8.2 | 4+3 | Prostate | 1.ERG | 0.40 | 0.62 | 0.00 | 0.92 | 0.93 | 2.12 | -0.73 |
| TCGA.T | KK-A8IH | 9.3 | 4+3 | Prostate | 1.ERG | 0.82 | 0.80 | 0.00 | 0.86 | -1.74 | -2.47 | 0.8 |
| TCGA.T | KK-A8II | 10 | 3+4 | Prostate | 1.ERG | 0.93 | 0.74 | 0.00 | 0.92 | 1.35 | -6.33 | 2.41 |
| TCGA.T | M7-A720 | 8.8 | 3+3 | Prostate | 1.ERG | NA | 0.73 | 0.00 | 0.92 | 0.86 | 4.96 | -2.4 |
| TCGA.T | QU-A6IP | 8.1 | 3+3 | Prostate | 1.ERG | 0.58 | 1.07 | 0.00 | 0.90 | 1.43 | 0 | -0.38 |
| TCGA.T | V1-A8WS | 9.1 | 4+3 | Prostate | 1.ERG | 0.78 | 0.70 | 0.00 | 0.90 | -0.59 | -2.83 | 2.16 |
| TCGA.T | V1-A8WW | 7.4 | 4+5 | Prostate | 1.ERG | 0.84 | 0.79 | 0.00 | 0.90 | -1.22 | -1.1 | 0.55 |
| TCGA.T | VN-A88K | 8.3 | 4+4 | Prostate | 1.ERG | 0.82 | 0.73 | 0.00 | 0.85 | 0.58 | -1.79 | 0.52 |
| TCGA.T | VN-A88L | 7.5 | 3+4 | Prostate | 1.ERG | 0.64 | 0.85 | 0.00 | 0.85 | -0.28 | 1.99 | -1.18 |
| TCGA.T | VN-A88Q | 8.4 | 4+3 | Prostate | 1.ERG | 0.68 | 0.74 | 0.00 | 0.87 | 0.82 | -3.81 | 0.65 |
| TCGA.T | VP-A872 | 7.6 | 3+4 | Prostate | 1.ERG | 0.62 | 0.98 | 0.00 | 0.86 | -3.53 | 1.24 | 0.2 |
| TCGA.T | VP-A875 | 9.5 | 4+3 | Prostate | 1.ERG | 0.98 | 0.64 | 0.00 | 0.90 | 3.14 | -6.5 | 2.83 |
| TCGA.T | VP-A876 | 9.1 | 4+3 | Prostate | 1.ERG | 0.89 | 0.90 | 0.00 | 0.90 | 0.42 | -3.62 | 0.85 |
| TCGA.T | VP-A879 | 8.8 | 4+3 | Prostate | 1.ERG | 0.49 | 0.51 | 0.00 | 0.89 | 0.45 | 4.91 | -3.02 |
| TCGA.T | VP-A87C | 8.9 | 4+4 | Prostate | 1.ERG | 0.43 | 0.70 | 0.00 | 0.86 | -1.43 | 3.16 | -2.29 |
| TCGA.T | VP-A87D | 9 | 4+4 | Prostate | 1.ERG | 0.77 | 0.72 | 0.00 | 0.91 | 1.87 | -3.45 | 0.89 |
| TCGA.T | VP-A87K | 8.7 | 3+4 | Prostate | 1.ERG | 0.89 | 0.69 | 0.00 | 0.91 | 2.64 | -5.39 | 1.66 |

**Table C.1 03/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | XJ-A9DI | 7.8 | 5+5 | Prostate | 1.ERG | NA | 0.87 | 0.01 | 0.91 | -2.46 | 4.06 | -2.09 |
| TCGA.T | XJ-A9DK | 8.4 | 4+3 | Prostate | 1.ERG | NA | 0.71 | 0.00 | 0.89 | -0.43 | 3.66 | -1.99 |
| TCGA.T | XQ-A8TB | 7.9 | 3+4 | Prostate | 1.ERG | 0.86 | 0.57 | 0.00 | 0.91 | -0.59 | 0.76 | 0.25 |
| TCGA.T | YL-A8HJ | 7.8 | 4+3 | Prostate | 1.ERG | 0.36 | 0.56 | 0.00 | 0.88 | -0.99 | 5.4 | -2.48 |
| TCGA.T | YL-A8HK | 7.8 | 4+4 | Prostate | 1.ERG | 0.42 | 0.61 | 0.01 | 0.92 | -0.66 | 3.32 | -1.19 |
| TCGA.T | YL-A8HL | 9.1 | 4+4 | Prostate | 1.ERG | 0.85 | 0.64 | 0.00 | 0.89 | 0.38 | -3.64 | 1.45 |
| TCGA.T | YL-A8SA | 9.2 | 4+3 | Prostate | 1.ERG | 0.82 | 0.67 | 0.00 | 0.90 | -0.67 | -5.36 | 2.02 |
| TCGA.T | YL-A8SF | 8.4 | 4+3 | Prostate | 1.ERG | 0.70 | 0.73 | 0.00 | 0.89 | -2.23 | -0.14 | 0.65 |
| TCGA.T | YL-A8SL | 9.4 | 4+4 | Prostate | 1.ERG | 0.77 | 0.69 | 0.00 | 0.89 | -0.5 | -5.38 | 2.56 |
| TCGA.T | YL-A8SP | 9.2 | 4+5 | Prostate | 1.ERG | 0.88 | 0.66 | 0.00 | 0.90 | 1.04 | -4.61 | 1.26 |
| TCGA.T | ZG-A8QZ | 8 | 3+4 | Prostate | 1.ERG | 0.45 | 0.78 | 0.00 | 0.88 | -0.68 | 2.88 | -1.94 |
| TCGA.T | 2A-A8W1 | 8.6 | 4+4 | Prostate | 2.ETV1 | 0.95 | 0.68 | 0.00 | 0.91 | 4.62 | -9.41 | 3.93 |
| TCGA.T | 2A-A8W3 | 8.8 | 5+4 | Prostate | 2.ETV1 | 0.46 | 0.57 | 0.00 | 0.87 | 1.49 | -0.68 | 1.84 |
| TCGA.T | CH-5748 | 8.2 | 3+4 | Prostate | 2.ETV1 | 0.31 | 0.92 | 0.00 | 0.88 | 4.58 | -2.32 | 1.4 |
| TCGA.T | CH-5750 | 8.3 | 3+4 | Prostate | 2.ETV1 | 0.75 | 0.93 | 0.00 | 0.89 | 2.25 | -3.06 | 2.33 |
| TCGA.T | CH-5753 | 8.8 | 4+5 | Prostate | 2.ETV1 | 0.83 | 0.82 | 0.01 | 0.87 | 2.28 | -4.54 | 1 |
| TCGA.T | EJ-5501 | 8.2 | 3+4 | Prostate | 2.ETV1 | 0.40 | 0.75 | 0.00 | 0.92 | 0.05 | 0.47 | -1.31 |
| TCGA.T | EJ-5504 | 8.9 | 3+4 | Prostate | 2.ETV1 | 0.53 | 0.95 | 0.00 | 0.90 | 1.8 | 1.23 | -0.79 |
| TCGA.T | EJ-5510 | 8 | 4+3 | Prostate | 2.ETV1 | 0.40 | 0.82 | 0.00 | 0.92 | 1.29 | 1.25 | -1.88 |
| TCGA.T | EJ-5519 | 9.2 | 4+4 | Prostate | 2.ETV1 | 0.67 | 0.70 | 0.00 | 0.92 | 2.59 | -1.59 | 1.01 |
| TCGA.T | EJ-7318 | 8.3 | 3+3 | Prostate | 2.ETV1 | 0.63 | 1.00 | 0.00 | 0.91 | 2.62 | -0.45 | 1.08 |
| TCGA.T | EJ-7788 | 9.2 | 4+5 | Prostate | 2.ETV1 | 0.71 | 0.73 | 0.00 | 0.91 | 1.27 | 0.56 | -1.01 |
| TCGA.T | EJ-8474 | 8.5 | 3+4 | Prostate | 2.ETV1 | 0.45 | 0.97 | 0.00 | 0.90 | 1.15 | 1.52 | 0.84 |
| TCGA.T | EJ-A7NH | 7.8 | 3+4 | Prostate | 2.ETV1 | 0.41 | 0.66 | 0.01 | 0.90 | 2.01 | 1.3 | -0.4 |
| TCGA.T | EJ-A8FU | 7.8 | 4+4 | Prostate | 2.ETV1 | NA | 0.68 | 0.00 | 0.89 | 0.37 | 6.43 | -3.43 |
| TCGA.T | G9-6348 | 8 | 4+3 | Prostate | 2.ETV1 | 0.39 | 1.07 | 0.00 | 0.91 | -0.24 | 1.37 | -1.25 |
| TCGA.T | G9-6494 | 9.1 | 4+3 | Prostate | 2.ETV1 | 0.60 | 0.81 | 0.00 | 0.91 | 4.57 | -0.15 | 0.37 |
| TCGA.T | HC-A631 | 8.4 | 3+4 | Prostate | 2.ETV1 | 0.92 | 0.89 | 0.00 | 0.89 | 0.56 | -3.39 | 2.66 |
| TCGA.T | HC-A8CY | 9 | 4+5 | Prostate | 2.ETV1 | 0.82 | 0.68 | 0.00 | 0.85 | 3.2 | -5.65 | 2.47 |
| TCGA.T | J4-A83M | 8.6 | 3+4 | Prostate | 2.ETV1 | 0.61 | 0.70 | 0.00 | 0.90 | 1.31 | -2.28 | 0.72 |
| TCGA.T | KK-A7AP | 8.9 | 3+4 | Prostate | 2.ETV1 | 0.90 | 0.54 | 0.00 | 0.90 | 0.17 | -6.42 | 2.46 |
| TCGA.T | KK-A7B3 | 8.3 | 4+3 | Prostate | 2.ETV1 | 0.59 | 0.81 | 0.01 | 0.90 | -0.94 | -0.78 | -0.49 |
| TCGA.T | M7-A71Y | 7.3 | 3+3 | Prostate | 2.ETV1 | 0.17 | 0.90 | 0.00 | 0.91 | -0.37 | 3.14 | -2.16 |
| TCGA.T | SU-A7E7 | 8.7 | 3+4 | Prostate | 2.ETV1 | 0.40 | 0.71 | 0.00 | 0.90 | 0.08 | -0.13 | -0.75 |
| TCGA.T | V1-A8MJ | 7.3 | 3+3 | Prostate | 2.ETV1 | 0.27 | 0.72 | 0.00 | 0.91 | 0.26 | 1.45 | -1.11 |
| TCGA.T | VP-A87J | 8.9 | 4+3 | Prostate | 2.ETV1 | 0.84 | 0.75 | 0.00 | 0.90 | 2.3 | -3.92 | 1.4 |
| TCGA.T | YL-A8SC | 9.2 | 4+4 | Prostate | 2.ETV1 | 0.66 | 0.69 | 0.00 | 0.92 | 0.8 | -2.61 | 1.77 |
| TCGA.T | YL-A8SJ | 8.3 | 4+3 | Prostate | 2.ETV1 | 0.65 | 0.71 | 0.00 | 0.87 | -1.34 | 3.7 | -2.53 |
| TCGA.T | YL-A9WH | 8.8 | 4+3 | Prostate | 2.ETV1 | 0.80 | 0.65 | 0.00 | 0.90 | -1.87 | -3.17 | 2.65 |
| TCGA.T | CH-5762 | 8.3 | 4+3 | Prostate | 3.ETV4 | 0.36 | 0.88 | 0.00 | 0.86 | 0.38 | 3.36 | -0.81 |
| TCGA.T | CH-5763 | 8.3 | 3+4 | Prostate | 3.ETV4 | NA | 0.96 | 0.00 | 0.86 | 1 | 5.75 | -3.05 |
| TCGA.T | CH-5771 | 8.4 | 5+4 | Prostate | 3.ETV4 | 0.42 | 0.81 | 0.00 | 0.92 | 0.13 | 4.49 | -2.23 |
| TCGA.T | EJ-5511 | 9.6 | 4+3 | Prostate | 3.ETV4 | 0.80 | 0.64 | 0.00 | 0.92 | 2.13 | 0.72 | -0.08 |
| TCGA.T | EJ-A7NM | 9.4 | 4+3 | Prostate | 3.ETV4 | 0.66 | 0.75 | 0.00 | 0.90 | -2.1 | -0.95 | -0.79 |
| TCGA.T | G9-6371 | 8.5 | 3+3 | Prostate | 3.ETV4 | 0.27 | 1.12 | 0.00 | 0.88 | 3.39 | -1.51 | 1.12 |
| TCGA.T | HC-7749 | 8.7 | 3+4 | Prostate | 3.ETV4 | 0.69 | 0.86 | 0.00 | 0.91 | 2.32 | -1.69 | 0.04 |
| TCGA.T | HC-A76W | 7.8 | 4+3 | Prostate | 3.ETV4 | 0.63 | 0.67 | 0.00 | 0.89 | -0.95 | 1.78 | -1.73 |

**Table C.1      04/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | HI-7168 | 8.5 | 4+4 | Prostate | 3.ETV4 | 0.58 | 1.06 | 0.00 | 0.89 | -1.02 | 5.09 | -2.85 |
| TCGA.T | KC-A7FD | 9.3 | 3+3 | Prostate | 3.ETV4 | 0.80 | 0.64 | 0.00 | 0.91 | 1.05 | -4.84 | 3.29 |
| TCGA.T | M7-A725 | 8.5 | 4+3 | Prostate | 3.ETV4 | 0.78 | 0.64 | 0.00 | 0.92 | 1.85 | -5.2 | 2.53 |
| TCGA.T | V1-A8WV | 7.8 | 4+5 | Prostate | 3.ETV4 | 0.52 | 1.00 | 0.00 | 0.84 | -3.26 | -0.1 | -0.62 |
| TCGA.T | V1-A8X3 | 8.6 | 3+4 | Prostate | 3.ETV4 | 0.68 | 0.62 | 0.00 | 0.90 | 3.28 | 2.36 | -0.28 |
| TCGA.T | XQ-A8TA | 8.3 | 4+4 | Prostate | 3.ETV4 | 0.96 | 0.77 | 0.01 | 0.88 | -3.24 | -8.06 | 1.51 |
| TCGA.T | CH-5738 | 7.8 | 3+3 | Prostate | 4.FLI1 | 0.44 | 0.76 | 0.00 | 0.91 | 0.44 | 1.45 | -1.57 |
| TCGA.T | H9-7775 | 8.7 | 3+3 | Prostate | 4.FLI1 | 0.80 | 0.82 | 0.00 | 0.91 | 2.72 | 0.44 | -1.82 |
| TCGA.T | HC-7079 | 7.8 | 5+4 | Prostate | 4.FLI1 | NA | 0.88 | 0.00 | 0.85 | -4.31 | 3.58 | -2.59 |
| TCGA.T | V1-A8WN | 9 | 3+3 | Prostate | 4.FLI1 | 0.42 | 0.70 | 0.00 | 0.90 | 2.81 | 3.54 | -1.93 |
| TCGA.T | CH-5788 | 9.4 | 4+5 | Prostate | 5.SPOP | 0.85 | 0.87 | 0.00 | 0.89 | 2.02 | -4.64 | 3.1 |
| TCGA.T | EJ-5505 | 8.7 | 3+4 | Prostate | 5.SPOP | 0.48 | 0.96 | 0.00 | 0.90 | 5.8 | -1.02 | 0.03 |
| TCGA.T | EJ-5509 | 8.8 | 3+3 | Prostate | 5.SPOP | 0.54 | 0.68 | 0.00 | 0.92 | 1.68 | 1.79 | -0.5 |
| TCGA.T | EJ-5531 | 9 | 3+4 | Prostate | 5.SPOP | 0.35 | 0.75 | 0.00 | 0.92 | 2.15 | 2.91 | -1.13 |
| TCGA.T | EJ-7115 | 7.5 | 4+4 | Prostate | 5.SPOP | 0.56 | 0.73 | 0.00 | 0.89 | 3.03 | 1.83 | -0.87 |
| TCGA.T | EJ-7123 | 7.6 | 3+4 | Prostate | 5.SPOP | 0.54 | 0.72 | 0.00 | 0.92 | 4.25 | -2.76 | 2.34 |
| TCGA.T | EJ-7330 | 9.3 | 4+4 | Prostate | 5.SPOP | 0.37 | 0.71 | 0.00 | 0.92 | 2.22 | 3.15 | -2.41 |
| TCGA.T | EJ-7782 | 9 | 4+5 | Prostate | 5.SPOP | 0.76 | 0.64 | 0.00 | 0.92 | 2.61 | 0.01 | -0.19 |
| TCGA.T | EJ-8468 | 8.6 | 4+3 | Prostate | 5.SPOP | 0.58 | 0.67 | 0.00 | 0.93 | 3 | 1.3 | -0.7 |
| TCGA.T | EJ-A65E | 9.1 | 3+3 | Prostate | 5.SPOP | 0.62 | 0.88 | 0.00 | 0.88 | 4.23 | -3.33 | 0.36 |
| TCGA.T | EJ-A8FS | 8.1 | 4+3 | Prostate | 5.SPOP | 0.82 | 0.80 | 0.00 | 0.87 | 2.18 | -5.84 | 3.48 |
| TCGA.T | FC-7961 | 8.6 | 4+5 | Prostate | 5.SPOP | 0.43 | 1.00 | 0.00 | 0.87 | 1.27 | 3.04 | 0.66 |
| TCGA.T | G9-6333 | 8.3 | 4+3 | Prostate | 5.SPOP | 0.47 | 1.06 | 0.00 | 0.89 | 3.6 | 0.6 | 0.4 |
| TCGA.T | G9-7510 | 8.2 | 3+4 | Prostate | 5.SPOP | 0.34 | 0.89 | 0.00 | 0.91 | 1.76 | 2.23 | 0.17 |
| TCGA.T | HC-7080 | 9 | 4+5 | Prostate | 5.SPOP | 0.90 | 0.76 | 0.01 | 0.90 | 3.43 | -6.23 | 4.03 |
| TCGA.T | HC-8258.2 | 7.9 | 3+3 | Prostate | 5.SPOP | NA | 0.97 | 0.00 | 0.90 | 3.03 | 0.36 | -0.79 |
| TCGA.T | HC-8261.2 | 7.7 | 4+3 | Prostate | 5.SPOP | 0.68 | 0.96 | 0.00 | 0.91 | 2.99 | -2.24 | 1.51 |
| TCGA.T | J4-A6G3 | 7.3 | 4+5 | Prostate | 5.SPOP | NA | 0.87 | 0.00 | 0.92 | -0.34 | 2.01 | -1.12 |
| TCGA.T | KK-A59X | 9 | 4+4 | Prostate | 5.SPOP | 0.87 | 1.05 | 0.00 | 0.89 | 1.43 | -2.74 | 3.08 |
| TCGA.T | KK-A59Z | 7.2 | 4+3 | Prostate | 5.SPOP | 0.83 | 1.13 | 0.00 | 0.89 | -0.78 | 1.72 | -0.06 |
| TCGA.T | KK-A6E0 | 8.6 | 4+4 | Prostate | 5.SPOP | 0.64 | 1.04 | 0.00 | 0.92 | 3.39 | -3.69 | 2.61 |
| TCGA.T | KK-A8I9 | 8.1 | 4+3 | Prostate | 5.SPOP | 0.57 | 0.93 | 0.00 | 0.86 | -0.33 | 0.1 | 0.88 |
| TCGA.T | KK-A8IF | 8.5 | 4+4 | Prostate | 5.SPOP | 0.96 | 1.14 | 0.00 | 0.84 | 0.52 | -6.03 | 3.93 |
| TCGA.T | KK-A8IK | 8.7 | 4+4 | Prostate | 5.SPOP | 0.96 | 1.30 | 0.01 | 0.85 | 0.16 | -8.27 | 4.87 |
| TCGA.T | VN-A88O | 7.5 | 3+3 | Prostate | 5.SPOP | 0.36 | 0.82 | 0.00 | 0.86 | 2.83 | -3.1 | 1.05 |
| TCGA.T | VN-A88R | 8.8 | 4+4 | Prostate | 5.SPOP | 0.80 | 1.00 | 0.00 | 0.89 | 3.02 | -3.46 | 2.3 |
| TCGA.T | VP-A878 | 8.5 | 4+5 | Prostate | 5.SPOP | 0.43 | 0.78 | 0.00 | 0.89 | 1.41 | 2.2 | -0.39 |
| TCGA.T | VP-A87B | 8.2 | 4+4 | Prostate | 5.SPOP | 0.91 | 0.79 | 0.00 | 0.80 | 1.43 | -5.48 | 2.28 |
| TCGA.T | VP-A87H | 8.7 | 4+3 | Prostate | 5.SPOP | 0.59 | 0.67 | 0.00 | 0.90 | 2.46 | -2.31 | 1.28 |
| TCGA.T | XJ-A83G | 8.2 | 3+4 | Prostate | 5.SPOP | 0.54 | 0.99 | 0.00 | 0.77 | 3.02 | -4.61 | 3.15 |
| TCGA.T | Y6-A8TL | 8.4 | 3+3 | Prostate | 5.SPOP | 0.61 | 0.89 | 0.00 | 0.87 | 2.27 | 0.66 | -0.14 |
| TCGA.T | YL-A8HM | 9.5 | 4+4 | Prostate | 5.SPOP | 0.87 | 0.81 | 0.00 | 0.90 | 2.95 | -5.81 | 3.91 |
| TCGA.T | YL-A8S8 | 8.1 | 4+4 | Prostate | 5.SPOP | 0.62 | 0.78 | 0.00 | 0.90 | 0.79 | -1.21 | 2.03 |
| TCGA.T | YL-A8SH | 7.1 | 3+4 | Prostate | 5.SPOP | 0.68 | 0.80 | 0.00 | 0.88 | 1.67 | 0.66 | 0.88 |
| TCGA.T | YL-A8SO | 7.9 | 4+3 | Prostate | 5.SPOP | 0.56 | 0.91 | 0.01 | 0.89 | -0.4 | -1.47 | 1.79 |
| TCGA.T | ZG-A8QX | 7.6 | 3+3 | Prostate | 5.SPOP | 0.52 | 0.88 | 0.00 | 0.91 | 1.17 | 1.38 | 0.05 |
| TCGA.T | ZG-A8QY | 7.8 | 4+4 | Prostate | 5.SPOP | 0.57 | 0.69 | 0.00 | 0.89 | 1.96 | 2.07 | -1.02 |

**Table C.1 05/13**

| | | | | | | | | | | Expression Scores | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | CH-5737 | 7 | 3+4 | Prostate | 6.FOXA1 | 0.76 | 0.87 | 0.00 | 0.88 | 3.31 | -3.99 | 1.04 |
| TCGA.T | EJ-5494 | 7.7 | 3+4 | Prostate | 6.FOXA1 | NA | 0.74 | 0.00 | 0.92 | 1.15 | -1.1 | -2.71 |
| TCGA.T | EJ-7789 | 9.2 | 4+5 | Prostate | 6.FOXA1 | 0.96 | 0.72 | 0.00 | 0.90 | 1.35 | -5.13 | 4.06 |
| TCGA.T | HC-7210 | 9.2 | 4+4 | Prostate | 6.FOXA1 | 0.47 | 0.72 | 0.00 | 0.92 | 3.84 | 2.04 | -0.71 |
| TCGA.T | HC-8265.1 | 7.8 | 3+4 | Prostate | 6.FOXA1 | 0.55 | 1.11 | 0.00 | 0.88 | -0.02 | -1.57 | -0.7 |
| TCGA.T | J9-A8CP | 8.6 | 3+4 | Prostate | 6.FOXA1 | 0.60 | 0.86 | 0.00 | 0.85 | 2.42 | -0.49 | 1.19 |
| TCGA.T | KK-A8IB | 7.7 | 4+4 | Prostate | 6.FOXA1 | 0.43 | 0.94 | 0.00 | 0.89 | -1.46 | -1.64 | 2.26 |
| TCGA.T | KK-A8IG | 7.6 | 4+4 | Prostate | 6.FOXA1 | 0.75 | 0.73 | 0.00 | 0.90 | 1.08 | 3.11 | 0.08 |
| TCGA.T | YL-A9WI | 8.6 | 4+4 | Prostate | 6.FOXA1 | 0.89 | 0.80 | 0.00 | 0.90 | 2.86 | -5.18 | 2.55 |
| TCGA.T | CH-5772 | 8.1 | 3+4 | Prostate | 7.IDH1 | 0.84 | 1.08 | 0.00 | 0.87 | 3.98 | -2.04 | -0.15 |
| TCGA.T | EJ-7125 | 9.4 | 3+3 | Prostate | 7.IDH1 | 0.69 | 0.64 | 0.00 | 0.92 | 4.45 | -0.56 | 0.5 |
| TCGA.T | G9-7523 | 7.8 | 3+4 | Prostate | 7.IDH1 | 0.25 | 0.98 | 0.00 | 0.91 | -2.07 | 7.83 | -2.68 |
| TCGA.T | 2A-A8VO | 7.3 | 3+3 | Prostate | 8.Other | 0.56 | 1.03 | 0.00 | 0.85 | 1.19 | -1.07 | 1.57 |
| TCGA.T | CH-5751 | 8.4 | 4+4 | Prostate | 8.Other | 0.38 | 0.79 | 0.00 | 0.88 | -0.1 | -2.44 | 1.99 |
| TCGA.T | CH-5761 | 9.1 | 5+5 | Prostate | 8.Other | 0.71 | 0.70 | 0.00 | 0.91 | -0.7 | -1.49 | 2.54 |
| TCGA.T | CH-5767 | 8.2 | 3+3 | Prostate | 8.Other | 0.84 | 0.95 | 0.00 | 0.75 | 4.55 | -2.54 | 2.79 |
| TCGA.T | CH-5792 | 8 | 3+4 | Prostate | 8.Other | 0.40 | 0.73 | 0.00 | 0.92 | -0.03 | 3.6 | -0.83 |
| TCGA.T | EJ-5514 | 9.2 | 4+3 | Prostate | 8.Other | 0.62 | 0.88 | 0.00 | 0.92 | -2.02 | 0.38 | 0.06 |
| TCGA.T | EJ-5515 | 8.5 | 4+3 | Prostate | 8.Other | NA | 1.02 | 0.00 | 0.90 | 4.32 | 4.69 | -1.96 |
| TCGA.T | EJ-5517 | 8.2 | 3+4 | Prostate | 8.Other | 0.72 | 1.04 | 0.00 | 0.91 | 3.63 | 1.98 | -1.63 |
| TCGA.T | EJ-5518 | 9.2 | 4+5 | Prostate | 8.Other | 0.56 | 0.60 | 0.00 | 0.93 | 2.47 | 1.73 | 0.36 |
| TCGA.T | EJ-5532 | 9.1 | 3+4 | Prostate | 8.Other | 0.68 | 0.83 | 0.00 | 0.93 | 4.81 | -0.5 | -0.52 |
| TCGA.T | EJ-7218 | 8.2 | 3+3 | Prostate | 8.Other | 0.44 | 0.85 | 0.00 | 0.91 | 3 | -3.63 | 1.34 |
| TCGA.T | EJ-7317 | 7.6 | 4+3 | Prostate | 8.Other | 0.71 | 0.93 | 0.00 | 0.91 | 4.68 | -1.67 | 0.54 |
| TCGA.T | EJ-7331 | 9.1 | 4+3 | Prostate | 8.Other | 0.47 | 0.68 | 0.00 | 0.93 | 2.89 | 3.16 | -1.06 |
| TCGA.T | EJ-7781 | 9.4 | 3+4 | Prostate | 8.Other | 0.58 | 0.54 | 0.00 | 0.91 | 2.67 | 1.32 | -1.02 |
| TCGA.T | EJ-7786 | 9.3 | 3+3 | Prostate | 8.Other | 0.62 | 0.77 | 0.00 | 0.91 | 4.26 | 2.06 | -0.83 |
| TCGA.T | EJ-7791 | 8.9 | 4+3 | Prostate | 8.Other | NA | 0.79 | 0.00 | 0.92 | 2.04 | 5.36 | -2.97 |
| TCGA.T | EJ-7792 | 8.8 | 3+3 | Prostate | 8.Other | 0.40 | 0.74 | 0.00 | 0.91 | 0.08 | 6.67 | -3.58 |
| TCGA.T | EJ-7794 | 9.3 | 3+4 | Prostate | 8.Other | 0.36 | 0.77 | 0.00 | 0.91 | 2.39 | 3.22 | -1.81 |
| TCGA.T | EJ-8470 | 8.3 | 3+3 | Prostate | 8.Other | 0.47 | 0.93 | 0.01 | 0.91 | 1.91 | -2.06 | 0.48 |
| TCGA.T | EJ-A46G | 9 | 4+3 | Prostate | 8.Other | 0.74 | 1.01 | 0.00 | 0.91 | 1.81 | -1.66 | 0.07 |
| TCGA.T | EJ-A46I | 8.3 | 3+3 | Prostate | 8.Other | 0.27 | 0.90 | 0.00 | 0.91 | 1.9 | 4.27 | -1.15 |
| TCGA.T | EJ-A65G | 8.2 | 3+3 | Prostate | 8.Other | 0.51 | 0.95 | 0.00 | 0.89 | 4.05 | -2.79 | 1.47 |
| TCGA.T | EJ-A65J | 7.9 | 3+4 | Prostate | 8.Other | 0.85 | 0.88 | 0.00 | 0.91 | 4.64 | -5.07 | 2.77 |
| TCGA.T | EJ-A6RA | 7.6 | 4+3 | Prostate | 8.Other | NA | 0.84 | 0.00 | 0.90 | 1.52 | 0.91 | -0.86 |
| TCGA.T | EJ-A6RC | 7.3 | 3+4 | Prostate | 8.Other | NA | 0.71 | 0.00 | 0.91 | 0.59 | 4.88 | -1.59 |
| TCGA.T | EJ-A7NJ | 9 | 4+5 | Prostate | 8.Other | 0.85 | 0.62 | 0.00 | 0.90 | 3.56 | 0.84 | -1.89 |
| TCGA.T | EJ-A8FN | 8 | 3+3 | Prostate | 8.Other | 0.69 | 0.61 | 0.00 | 0.87 | 1.47 | 1.81 | -0.62 |
| TCGA.T | FC-A4JI | 9.1 | 4+5 | Prostate | 8.Other | 0.69 | 0.80 | 0.00 | 0.91 | 0.92 | -8.31 | 5.21 |
| TCGA.T | FC-A5OB | 8.8 | 4+5 | Prostate | 8.Other | 0.86 | 0.90 | 0.00 | 0.90 | 3.61 | -6.39 | 3.84 |
| TCGA.T | FC-A8O0 | 7.1 | 3+3 | Prostate | 8.Other | NA | 0.87 | 0.00 | 0.87 | -2.06 | 2.78 | -2.53 |
| TCGA.T | G9-6366 | 8.6 | 3+4 | Prostate | 8.Other | 0.64 | 1.05 | 0.00 | 0.91 | 1.57 | -3.54 | 0.71 |
| TCGA.T | G9-6367 | 8.1 | 4+3 | Prostate | 8.Other | NA | 0.90 | 0.00 | 0.90 | 2.34 | 5.63 | -2.45 |
| TCGA.T | G9-6370 | 8 | 3+4 | Prostate | 8.Other | NA | 0.79 | 0.00 | 0.90 | 0.13 | 6.35 | -3.84 |
| TCGA.T | G9-6378 | 7.2 | 3+4 | Prostate | 8.Other | 0.29 | 1.09 | 0.00 | 0.89 | -1.19 | 3.58 | -2.71 |
| TCGA.T | G9-6499 | 7.3 | 3+4 | Prostate | 8.Other | 0.42 | 1.14 | 0.00 | 0.88 | 1.98 | -1 | 0.73 |

**Table C.1 06/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | G9-7519 | 7.1 | 3+3 | Prostate | 8.Other | 0.28 | 1.18 | 0.00 | 0.91 | 2.11 | 1.52 | 0.48 |
| TCGA.T | G9-7521 | 9.2 | 3+4 | Prostate | 8.Other | 0.45 | 0.87 | 0.00 | 0.92 | 2.41 | -1.05 | 0.85 |
| TCGA.T | HC-7075 | 8.4 | 3+4 | Prostate | 8.Other | 0.79 | 0.84 | 0.00 | 0.91 | 4.04 | -5.07 | 2.45 |
| TCGA.T | HC-7078 | 8.1 | 4+4 | Prostate | 8.Other | 0.74 | 0.98 | 0.00 | 0.89 | 1 | -2.45 | 1.05 |
| TCGA.T | HC-7233 | 9.6 | 3+5 | Prostate | 8.Other | 0.56 | 0.62 | 0.00 | 0.91 | 0.88 | 1.95 | -0.89 |
| TCGA.T | HC-7736 | 9.1 | 3+3 | Prostate | 8.Other | 0.60 | 0.55 | 0.00 | 0.92 | 1.35 | 0.82 | 0.12 |
| TCGA.T | HC-7737 | 7.5 | 3+4 | Prostate | 8.Other | 0.33 | 0.75 | 0.00 | 0.90 | 3.03 | 1.27 | 0.66 |
| TCGA.T | HC-7740.1 | 7.9 | 3+4 | Prostate | 8.Other | NA | 0.73 | 0.00 | 0.91 | 0.79 | 5.21 | -3.78 |
| TCGA.T | HC-7742 | 8.4 | 4+4 | Prostate | 8.Other | 0.57 | 0.86 | 0.00 | 0.91 | 1.6 | 3.37 | -0.59 |
| TCGA.T | HC-7750 | 7.7 | 3+3 | Prostate | 8.Other | 0.36 | 0.93 | 0.00 | 0.91 | 0.81 | 5.25 | -1.08 |
| TCGA.T | HC-8216 | 9.4 | 4+3 | Prostate | 8.Other | 0.58 | 0.88 | 0.00 | 0.89 | 2.85 | -1.28 | 0.93 |
| TCGA.T | HC-8256 | 9 | 4+3 | Prostate | 8.Other | 0.47 | 0.87 | 0.00 | 0.90 | 3.03 | -3.4 | 0.63 |
| TCGA.T | HC-8264 | 8.2 | 3+4 | Prostate | 8.Other | 0.72 | 0.89 | 0.00 | 0.90 | 1.68 | 5.2 | -0.62 |
| TCGA.T | HC-8266 | 7.2 | 4+5 | Prostate | 8.Other | 0.31 | 0.90 | 0.00 | 0.90 | -0.12 | 3.05 | -2.52 |
| TCGA.T | HC-A4ZV | 9.3 | 4+5 | Prostate | 8.Other | 0.70 | 0.95 | 0.00 | 0.88 | 0.75 | -2.29 | 2.66 |
| TCGA.T | HI-7170 | 8.3 | 3+3 | Prostate | 8.Other | 0.54 | 0.79 | 0.00 | 0.91 | -1.94 | 5.26 | -3.13 |
| TCGA.T | J4-8200 | 9.3 | 3+4 | Prostate | 8.Other | 0.54 | 0.87 | 0.00 | 0.90 | 2.22 | 2.49 | -0.99 |
| TCGA.T | J9-A8CL | 7.7 | 3+4 | Prostate | 8.Other | 0.76 | 0.64 | 0.00 | 0.83 | 2.49 | -0.55 | 0.93 |
| TCGA.T | J9-A8CN | 7.8 | 3+3 | Prostate | 8.Other | 0.88 | 0.86 | 0.00 | 0.84 | 2.68 | -6.41 | 3.27 |
| TCGA.T | KC-A4BL | 7.6 | 3+3 | Prostate | 8.Other | NA | 0.86 | 0.00 | 0.90 | -1.87 | 3.84 | -2.09 |
| TCGA.T | KC-A7F3 | 8.4 | 3+3 | Prostate | 8.Other | 0.51 | 0.80 | 0.00 | 0.89 | 0.85 | -0.13 | 1.02 |
| TCGA.T | KC-A7FA | 7.7 | 3+4 | Prostate | 8.Other | 0.75 | 0.93 | 0.00 | 0.85 | -0.54 | -1.91 | 0.75 |
| TCGA.T | KC-A7FE | 8.3 | 3+3 | Prostate | 8.Other | NA | 0.77 | 0.00 | 0.90 | -0.2 | 5.93 | -2.56 |
| TCGA.T | KK-A6E5 | 7.3 | 3+3 | Prostate | 8.Other | 0.62 | 1.07 | 0.00 | 0.90 | 2.56 | -1.58 | 0.67 |
| TCGA.T | KK-A7AV | 8 | 3+4 | Prostate | 8.Other | 0.34 | 0.91 | 0.00 | 0.90 | 1.9 | -0.77 | 0.25 |
| TCGA.T | KK-A8ID | 7.9 | 4+4 | Prostate | 8.Other | 0.51 | 0.73 | 0.00 | 0.89 | 2.07 | -6.37 | 3.87 |
| TCGA.T | KK-A8IJ | 8.6 | 3+3 | Prostate | 8.Other | 0.59 | 0.66 | 0.00 | 0.90 | 2.69 | -1.83 | 0.4 |
| TCGA.T | KK-A8IL | 7.5 | 3+4 | Prostate | 8.Other | 0.67 | 0.87 | 0.00 | 0.89 | -0.92 | 0.12 | 1.19 |
| TCGA.T | M7-A721 | 8.5 | 3+4 | Prostate | 8.Other | 0.26 | 0.71 | 0.00 | 0.91 | 2.29 | 0.17 | 1.12 |
| TCGA.T | TK-A8OK | 7 | 5+4 | Prostate | 8.Other | 0.54 | 0.68 | 0.00 | 0.85 | -3.05 | 8.16 | -4.16 |
| TCGA.T | V1-A8MF | 7.8 | 4+4 | Prostate | 8.Other | 0.65 | 0.87 | 0.00 | 0.86 | 0.6 | -2.32 | 0.9 |
| TCGA.T | V1-A8MG | 8.1 | 4+3 | Prostate | 8.Other | 0.71 | 0.82 | 0.00 | 0.89 | 1.11 | 1.3 | 0.14 |
| TCGA.T | V1-A8ML | 8.1 | 3+4 | Prostate | 8.Other | 0.52 | 0.94 | 0.00 | 0.84 | 1.91 | 0.02 | 0.59 |
| TCGA.T | V1-A8MU | 7.4 | 4+3 | Prostate | 8.Other | 0.48 | 0.80 | 0.00 | 0.86 | -0.06 | 4.26 | -2.7 |
| TCGA.T | V1-A8WL | 8.5 | 3+4 | Prostate | 8.Other | 0.71 | 0.89 | 0.00 | 0.76 | 2.51 | 1.02 | -1.97 |
| TCGA.T | VN-A88I | 7.3 | 4+4 | Prostate | 8.Other | 0.84 | 0.85 | 0.00 | 0.88 | -3.99 | 8.74 | -3.63 |
| TCGA.T | VN-A88N | 8 | 4+3 | Prostate | 8.Other | 0.76 | 0.77 | 0.00 | 0.87 | 2.84 | -4.8 | 2.27 |
| TCGA.T | VN-A88P | 8.1 | 3+4 | Prostate | 8.Other | 0.81 | 0.85 | 0.00 | 0.86 | 0.71 | -0.31 | -0.13 |
| TCGA.T | VP-A87E | 7.6 | 3+4 | Prostate | 8.Other | NA | 0.78 | 0.00 | 0.82 | 0.3 | 3.6 | -1.98 |
| TCGA.T | WW-A8ZI | 7.9 | 3+4 | Prostate | 8.Other | 0.93 | 0.80 | 0.00 | 0.80 | 0.13 | -5.66 | 1.16 |
| TCGA.T | XA-A8JR | 7.1 | 3+3 | Prostate | 8.Other | 0.38 | 0.85 | 0.00 | 0.87 | 1.41 | 1.4 | 0 |
| TCGA.T | XJ-A83H | 8.5 | 3+3 | Prostate | 8.Other | 0.53 | 0.82 | 0.00 | 0.81 | 3.01 | -2.44 | 1.52 |
| TCGA.T | YJ-A8SW | 9.5 | 4+4 | Prostate | 8.Other | 0.84 | 0.53 | 0.00 | 0.91 | 3.12 | -3.81 | 1.58 |
| TCGA.T | YL-A8HO | 7.9 | 3+4 | Prostate | 8.Other | 0.57 | 0.67 | 0.00 | 0.85 | -0.14 | 3.5 | -2.46 |
| TCGA.T | YL-A8S9 | 8.6 | 4+3 | Prostate | 8.Other | 0.95 | 0.76 | 0.00 | 0.90 | 2.33 | -10.16 | 5.31 |
| TCGA.T | YL-A8SB | 7.6 | 4+3 | Prostate | 8.Other | 0.53 | 0.82 | 0.00 | 0.91 | 0.91 | 2.33 | -0.89 |
| TCGA.T | YL-A8SK | 7.4 | 3+4 | Prostate | 8.Other | 0.44 | 0.73 | 0.01 | 0.89 | -2.02 | 5.09 | -3.04 |

**Table C.1 07/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.T | YL-A8SQ | 7.8 | 4+3 | Prostate | 8.Other | 0.68 | 0.78 | 0.00 | 0.88 | 3.49 | -1.71 | 0.46 |
| TCGA.T | YL-A8SR | 9 | 4+3 | Prostate | 8.Other | 0.68 | 0.75 | 0.00 | 0.91 | 3.55 | -0.88 | 2.24 |
| TCGA.T | YL-A9WJ | 8.9 | 3+3 | Prostate | 8.Other | 0.42 | 0.61 | 0.00 | 0.88 | 2.11 | 0.56 | -0.62 |
| TCGA.T | ZG-A8QW | 8 | 3+4 | Prostate | 8.Other | 0.41 | 0.87 | 0.00 | 0.88 | -1.15 | 1.45 | -0.29 |
| TCGA.N | CH-5761 | 7.3 | NA | Prostate | 9.Normal | 0.00 | 0.76 | 0.00 | 0.91 | 0.14 | 3.01 | -2.07 |
| TCGA.N | CH-5767 | 8.9 | NA | Prostate | 9.Normal | 0.00 | 0.95 | 0.00 | 0.92 | -1.83 | 5.03 | -2.12 |
| TCGA.N | CH-5768 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 0.89 | 0.00 | 0.88 | 0.91 | 3.66 | -3.35 |
| TCGA.N | CH-5769 | 7.1 | NA | Prostate | 9.Normal | 0.00 | 0.81 | 0.00 | 0.89 | -1.3 | 0.41 | 1.18 |
| TCGA.N | EJ-7115 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.75 | 0.00 | 0.91 | 1.53 | 5.88 | -2.6 |
| TCGA.N | EJ-7123 | 8.7 | NA | Prostate | 9.Normal | 0.00 | 0.70 | 0.01 | 0.91 | -3.69 | 13.32 | -5.27 |
| TCGA.N | EJ-7125 | 8.3 | NA | Prostate | 9.Normal | 0.00 | 0.78 | 0.00 | 0.90 | 0.57 | 3.8 | -2.41 |
| TCGA.N | EJ-7314 | 9 | NA | Prostate | 9.Normal | 0.00 | 0.64 | 0.00 | 0.92 | 0.73 | 8.98 | -4.77 |
| TCGA.N | EJ-7315 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.67 | 0.00 | 0.92 | -2.64 | 9.33 | -5.25 |
| TCGA.N | EJ-7317 | 9.5 | NA | Prostate | 9.Normal | 0.00 | 0.64 | 0.00 | 0.93 | 3.95 | 4.65 | -2.67 |
| TCGA.N | EJ-7321 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.70 | 0.00 | 0.92 | 1.67 | 5.76 | -3.94 |
| TCGA.N | EJ-7327 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.81 | 0.00 | 0.90 | 4.23 | -0.1 | -1.41 |
| TCGA.N | EJ-7328 | 8.2 | NA | Prostate | 9.Normal | 0.00 | 0.68 | 0.00 | 0.91 | -2.64 | 11.89 | -5.66 |
| TCGA.N | EJ-7330 | 8.2 | NA | Prostate | 9.Normal | 0.00 | 0.64 | 0.00 | 0.92 | -3.56 | 9.63 | -5.95 |
| TCGA.N | EJ-7331 | 9.2 | NA | Prostate | 9.Normal | 0.00 | 0.74 | 0.00 | 0.91 | 0.87 | 8.36 | -4.63 |
| TCGA.N | EJ-7781 | 8.5 | NA | Prostate | 9.Normal | 0.00 | 0.75 | 0.00 | 0.92 | -3.11 | 12.3 | -6.35 |
| TCGA.N | EJ-7782 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.67 | 0.00 | 0.90 | 0.93 | 3.19 | -2.22 |
| TCGA.N | EJ-7783 | 8 | NA | Prostate | 9.Normal | 0.00 | 0.77 | 0.00 | 0.92 | -5.65 | 13.68 | -5.15 |
| TCGA.N | EJ-7784 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.70 | 0.00 | 0.92 | 0.77 | 6.04 | -4.97 |
| TCGA.N | EJ-7785 | 8.5 | NA | Prostate | 9.Normal | 0.00 | 0.67 | 0.00 | 0.91 | 0.42 | 8.31 | -4.51 |
| TCGA.N | EJ-7786 | 8.6 | NA | Prostate | 9.Normal | 0.00 | 0.66 | 0.00 | 0.88 | -0.84 | 9.96 | -5.69 |
| TCGA.N | EJ-7789 | 8.6 | NA | Prostate | 9.Normal | 0.00 | 0.69 | 0.00 | 0.90 | 1.39 | 4.65 | -3.2 |
| TCGA.N | EJ-7792 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.76 | 0.00 | 0.91 | -0.52 | 7.72 | -4.96 |
| TCGA.N | EJ-7793 | 8.7 | NA | Prostate | 9.Normal | 0.00 | 0.78 | 0.00 | 0.92 | 1.37 | 8.26 | -4.54 |
| TCGA.N | EJ-7794 | 9.1 | NA | Prostate | 9.Normal | 0.00 | 0.61 | 0.00 | 0.91 | -1.97 | 8.11 | -5.8 |
| TCGA.N | EJ-7797 | 8.5 | NA | Prostate | 9.Normal | 0.00 | 0.84 | 0.00 | 0.92 | 0.73 | 7.76 | -4.6 |
| TCGA.N | EJ-A8FO | 8.4 | NA | Prostate | 9.Normal | 0.00 | 0.63 | 0.00 | 0.90 | -3.53 | 10.96 | -6.08 |
| TCGA.N | G9-6333 | 7.4 | NA | Prostate | 9.Normal | 0.00 | 1.10 | 0.00 | 0.90 | -3.96 | 7.3 | -3.93 |
| TCGA.N | G9-6342 | 7.4 | NA | Prostate | 9.Normal | 0.00 | 1.19 | 0.00 | 0.89 | 1.77 | 2.87 | -1.68 |
| TCGA.N | G9-6348 | 7 | NA | Prostate | 9.Normal | 0.00 | 1.21 | 0.00 | 0.89 | -3 | 8.44 | -4.3 |
| TCGA.N | G9-6351 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 1.05 | 0.00 | 0.90 | 1.53 | 6.17 | -3.36 |
| TCGA.N | G9-6356 | 7.6 | NA | Prostate | 9.Normal | 0.00 | 1.04 | 0.00 | 0.89 | 0.38 | -2.18 | -2.28 |
| TCGA.N | G9-6362 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 1.57 | 0.00 | 0.87 | 0.74 | -0.86 | -0.11 |
| TCGA.N | G9-6363 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 0.90 | 0.00 | 0.89 | -0.58 | 4.16 | -2.7 |
| TCGA.N | G9-6365 | 7.2 | NA | Prostate | 9.Normal | 0.00 | 1.14 | 0.00 | 0.89 | -1.21 | 2.94 | -1.63 |
| TCGA.N | G9-6384 | 7 | NA | Prostate | 9.Normal | 0.00 | 1.02 | 0.00 | 0.90 | 0.16 | 7.32 | -2.63 |
| TCGA.N | G9-6496 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 1.12 | 0.00 | 0.90 | -1.7 | 7.84 | -3.57 |
| TCGA.N | G9-6499 | 7.6 | NA | Prostate | 9.Normal | 0.00 | 1.14 | 0.00 | 0.88 | -0.87 | 5.34 | -0.15 |
| TCGA.N | HC-7211 | 8.2 | NA | Prostate | 9.Normal | 0.00 | 0.86 | 0.00 | 0.91 | -5.95 | 0.31 | -1.04 |
| TCGA.N | HC-7737 | 8.1 | NA | Prostate | 9.Normal | 0.00 | 0.81 | 0.00 | 0.91 | -5.44 | 5.58 | -3.17 |
| TCGA.N | HC-7738 | 7.4 | NA | Prostate | 9.Normal | 0.00 | 0.84 | 0.00 | 0.89 | -6.05 | 6.2 | -3.21 |
| TCGA.N | HC-7740 | 8.1 | NA | Prostate | 9.Normal | 0.00 | 0.88 | 0.00 | 0.91 | -6.07 | 4 | -3.22 |
| TCGA.N | HC-7742 | 7.3 | NA | Prostate | 9.Normal | 0.00 | 0.82 | 0.00 | 0.91 | 1.57 | 5.23 | -2.21 |

**Table C.1 08/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| TCGA.N | HC-7745 | 7.1 | NA | Prostate | 9.Normal | 0.00 | 0.74 | 0.00 | 0.92 | -6.18 | 8.93 | -4.2 |
| TCGA.N | HC-7747 | 7.6 | NA | Prostate | 9.Normal | 0.00 | 0.77 | 0.00 | 0.92 | -6.07 | 3 | -2.3 |
| TCGA.N | HC-7752 | 7.6 | NA | Prostate | 9.Normal | 0.00 | 1.00 | 0.00 | 0.90 | -2.6 | 7.63 | -3.84 |
| TCGA.N | HC-7819 | 7.8 | NA | Prostate | 9.Normal | 0.00 | 0.81 | 0.00 | 0.91 | 1.48 | 5.36 | -3.6 |
| TCGA.N | HC-8258 | 7.5 | NA | Prostate | 9.Normal | 0.00 | 0.82 | 0.00 | 0.92 | -6.11 | 0.61 | -1.57 |
| TCGA.N | HC-8259 | 7.2 | NA | Prostate | 9.Normal | 0.00 | 0.93 | 0.00 | 0.91 | 1.83 | 5.54 | -2.75 |
| TCGA.N | HC-8260 | 8.2 | NA | Prostate | 9.Normal | 0.00 | 0.85 | 0.00 | 0.91 | 2.96 | 4.62 | -3.3 |
| TCGA.N | HC-8262 | 7.9 | NA | Prostate | 9.Normal | 0.00 | 0.78 | 0.00 | 0.91 | -2.28 | 4.41 | -3.85 |
| TCGA.N | J4-A83J | 8.1 | NA | Prostate | 9.Normal | 0.00 | 0.67 | 0.01 | 0.90 | -0.42 | 4.37 | -3.4 |
| Michigan.T | UT_4001 | 8.9 | 3+4 | Prostate | 1.ERG | 0.70 | 0.51 | 0.00 | 0.91 | -0.55 | -3.7 | 1.77 |
| Michigan.T | UT_4003 | 9.4 | 3+3 | Prostate | 1.ERG | 0.47 | 0.38 | 0.00 | 0.92 | 0.79 | 0.67 | -0.68 |
| Michigan.T | UT_4006 | 8.8 | 3+3 | Prostate | 1.ERG | 0.64 | 0.36 | 0.00 | 0.95 | -1.59 | -4.06 | 1.28 |
| Michigan.T | UT_4008 | 9.5 | 3+4 | Prostate | 1.ERG | 0.70 | 0.50 | 0.00 | 0.96 | 2.48 | -2.26 | 2.05 |
| Michigan.T | UT_4016 | 9.7 | 3+3 | Prostate | 1.ERG | 0.21 | 0.52 | 0.00 | 0.96 | 0.26 | 3.9 | -2.73 |
| Michigan.T | UT_4019 | 9.5 | 4+3 | Prostate | 1.ERG | 0.56 | 0.49 | 0.00 | 0.95 | 0.22 | 1.82 | -0.44 |
| Michigan.T | UT_4022 | 9.6 | 3+4 | Prostate | 1.ERG | 0.61 | 0.59 | 0.01 | 0.95 | 0.61 | -1.29 | 0.92 |
| Michigan.T | UT_4023 | 9.6 | 3+3 | Prostate | 1.ERG | 0.40 | 0.57 | 0.00 | 0.95 | 1.6 | 1.74 | 0.06 |
| Michigan.T | UT_4025 | 9.2 | 3+3 | Prostate | 1.ERG | NA | 0.60 | 0.01 | 0.95 | -0.33 | 5.14 | -2.4 |
| Michigan.T | UT_4028 | 9.4 | 3+4 | Prostate | 1.ERG | 0.59 | 0.54 | 0.01 | 0.95 | 3.36 | -1.59 | 1.78 |
| Michigan.T | UT_4028 | 9.2 | 3+4 | Prostate | 1.ERG | 0.48 | 0.64 | 0.01 | 0.94 | 0.86 | 0.53 | -1.06 |
| Michigan.T | UT_4034 | 9.3 | 4+4 | Prostate | 1.ERG | 0.26 | 0.61 | 0.01 | 0.95 | -0.49 | 3.45 | -1.96 |
| Michigan.T | UT_4002 | 9.1 | 3+3 | Prostate | 2.ETV1 | 0.70 | 0.41 | 0.00 | 0.91 | 2.49 | -3.43 | 1.33 |
| Michigan.T | UT_4010 | 9.8 | 3+3 | Prostate | 4.FLI1 | 0.48 | 0.55 | 0.01 | 0.95 | 1.55 | 0.51 | 0.69 |
| Michigan.T | UT_4030 | 9.6 | 4+3 | Prostate | 5.SPOP | 0.82 | 0.46 | 0.00 | 0.95 | 3.33 | -8.2 | 4.39 |
| Michigan.T | UT_4009 | 9.2 | 3+4 | Prostate | 6.FOXA1 | 0.58 | 0.64 | 0.01 | 0.95 | 1.96 | -1.1 | 1.19 |
| Michigan.T | UT_4017 | 9.5 | 3+4 | Prostate | 6.FOXA1 | 0.64 | 0.43 | 0.01 | 0.96 | 2.67 | -3.6 | 1.8 |
| Michigan.T | UT_4018 | 9.2 | 3+3 | Prostate | 6.FOXA1 | 0.32 | 0.67 | 0.01 | 0.96 | 1.29 | 1.91 | 0.32 |
| Michigan.T | UT_4018 | 9.2 | 3+3 | Prostate | 6.FOXA1 | 0.32 | 0.78 | 0.01 | 0.94 | 1.59 | 1.99 | 0.07 |
| Michigan.T | UT_4032 | 7.3 | 4+3 | Prostate | 6.FOXA1 | 0.70 | 0.84 | 0.00 | 0.91 | 1.95 | -4.27 | 2.76 |
| Michigan.T | UT_4005 | 9.3 | 4+3 | Prostate | 8.Other | 0.53 | 0.45 | 0.00 | 0.91 | 1.51 | -3.48 | 1.35 |
| Michigan.T | UT_4006 | 8.8 | NA | Prostate | 8.Other | 0.25 | 0.46 | 0.00 | 0.92 | -2.64 | -2.23 | 0.63 |
| Michigan.T | UT_4007 | 9.4 | NA | Prostate | 8.Other | 0.38 | 0.21 | 0.00 | 0.93 | 4.21 | -0.4 | 1.03 |
| Michigan.T | UT_4011 | 9.8 | 3+3 | Prostate | 8.Other | 0.33 | 0.51 | 0.00 | 0.96 | 2.48 | 2.55 | -0.9 |
| Michigan.T | UT_4012 | 9.5 | 3+3 | Prostate | 8.Other | 0.26 | 0.45 | 0.01 | 0.95 | 2.56 | 0.96 | -1.13 |
| Michigan.T | UT_4013 | 8.9 | 3+3 | Prostate | 8.Other | 0.35 | 0.59 | 0.00 | 0.95 | 3.43 | -4.42 | 2.32 |
| Michigan.T | UT_4014 | 9.1 | NA | Prostate | 8.Other | 0.29 | 0.61 | 0.01 | 0.95 | 2.15 | 2.45 | -2.25 |
| Michigan.T | UT_4015 | 8 | 3+3 | Prostate | 8.Other | 0.46 | 0.58 | 0.01 | 0.94 | -1.63 | 2.89 | -0.88 |
| Michigan.T | UT_4015 | 9.3 | 3+3 | Prostate | 8.Other | 0.51 | 0.54 | 0.00 | 0.95 | -0.9 | 7.09 | -3 |
| Michigan.T | UT_4015 | 9.6 | NA | Prostate | 8.Other | 0.26 | 0.58 | 0.01 | 0.95 | 2.59 | 1.66 | -1.67 |
| Michigan.T | UT_4020 | 9.5 | 3+3 | Prostate | 8.Other | 0.40 | 0.66 | 0.00 | 0.95 | 1.92 | -0.01 | 0.7 |
| Michigan.T | UT_4021 | 9.7 | 3+3 | Prostate | 8.Other | 0.67 | 0.42 | 0.00 | 0.96 | 1.88 | -1.23 | 2.13 |
| Michigan.T | UT_4024 | 7.3 | 3+3 | Prostate | 8.Other | 0.67 | 0.49 | 0.00 | 0.96 | 2.46 | -2.67 | 2.01 |
| Michigan.T | UT_4027 | 9.2 | 3+4 | Prostate | 8.Other | NA | 0.61 | 0.00 | 0.94 | 3.8 | -3.34 | -0.31 |
| Michigan.T | UT_4029 | 8.2 | 4+3 | Prostate | 8.Other | 0.54 | 0.54 | 0.00 | 0.94 | -0.31 | 7.41 | -1.74 |
| Michigan.T | UT_4031 | 8.5 | 3+4 | Prostate | 8.Other | 0.58 | 0.71 | 0.00 | 0.93 | -1.46 | 4.99 | -1.47 |
| Michigan.T | UT_4035 | 6.1 | 3+4 | Prostate | 8.Other | 0.70 | 0.73 | 0.01 | 0.82 | -2.74 | -6.23 | 1.41 |
| Michigan.N | UT_4001 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.37 | 0.00 | 0.92 | -1.19 | 4.66 | -2.76 |

**Table C.1 09/13**

| | | | | | | | | | | Expression Scores | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| Michigan.N | UT_4002 | 8.4 | NA | Prostate | 9.Normal | 0.00 | 0.40 | 0.00 | 0.92 | 0.73 | 7.37 | -4.08 |
| Michigan.N | UT_4003 | 9.4 | NA | Prostate | 9.Normal | 0.00 | 0.37 | 0.00 | 0.92 | 0.33 | 6.7 | -2.62 |
| Michigan.N | UT_4004 | 8.4 | NA | Prostate | 9.Normal | 0.00 | 0.58 | 0.00 | 0.92 | 1.15 | 4.14 | -1.52 |
| Michigan.N | UT_4005 | 9.4 | NA | Prostate | 9.Normal | 0.00 | 0.71 | 0.00 | 0.88 | -3 | 6.85 | -3.72 |
| Michigan.N | UT_4006 | 9.2 | NA | Prostate | 9.Normal | 0.00 | 0.51 | 0.00 | 0.89 | 1.5 | 3.35 | -0.98 |
| Michigan.N | UT_4007 | 9 | NA | Prostate | 9.Normal | 0.00 | 0.47 | 0.00 | 0.93 | 1.28 | 2.15 | -1.37 |
| Michigan.N | UT_4008 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.59 | 0.01 | 0.96 | -2.67 | 8.46 | -3.39 |
| Michigan.N | UT_4008 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.47 | 0.00 | 0.96 | -5.53 | 12.72 | -4.9 |
| Michigan.N | UT_4009 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.61 | 0.01 | 0.96 | -0.84 | 8.41 | -4.6 |
| Michigan.N | UT_4009 | 9.5 | NA | Prostate | 9.Normal | 0.00 | 0.47 | 0.00 | 0.96 | -1.52 | 7.95 | -4.99 |
| Michigan.N | UT_4010 | 9.4 | NA | Prostate | 9.Normal | 0.00 | 0.42 | 0.00 | 0.97 | -4.5 | 13.4 | -4.36 |
| Michigan.N | UT_4010 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.43 | 0.00 | 0.96 | -1.11 | 8.69 | -3.72 |
| Michigan.N | UT_4011 | 8.2 | NA | Prostate | 9.Normal | 0.00 | 0.39 | 0.01 | 0.96 | -0.8 | 7.67 | -4.22 |
| Michigan.N | UT_4011 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.46 | 0.00 | 0.96 | 1.5 | 4.07 | -0.61 |
| Michigan.N | UT_4012 | 8.9 | NA | Prostate | 9.Normal | 0.00 | 0.54 | 0.01 | 0.95 | 1.59 | 4.66 | -1.68 |
| Michigan.N | UT_4013 | 9.7 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.01 | 0.95 | -1.72 | 7.63 | -2.66 |
| Michigan.N | UT_4014 | 8.3 | NA | Prostate | 9.Normal | 0.00 | 0.46 | 0.00 | 0.96 | -6.5 | 13.51 | -5 |
| Michigan.N | UT_4015 | 9.4 | NA | Prostate | 9.Normal | 0.00 | 0.50 | 0.00 | 0.96 | 0.57 | 5.99 | -2.42 |
| Michigan.N | UT_4016 | 9.2 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.00 | 0.96 | -0.4 | 6.75 | -2.49 |
| Michigan.N | UT_4017 | 8 | NA | Prostate | 9.Normal | 0.00 | 0.49 | 0.01 | 0.96 | 0.96 | 6.06 | -3.3 |
| Michigan.N | UT_4018 | 9.2 | NA | Prostate | 9.Normal | 0.00 | 0.56 | 0.00 | 0.96 | 1.32 | 5.97 | -3.41 |
| Michigan.N | UT_4019 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.61 | 0.00 | 0.96 | -2.23 | 6.08 | -2.91 |
| Michigan.N | UT_4019 | 8.4 | NA | Prostate | 9.Normal | 0.00 | 0.46 | 0.01 | 0.96 | -3.71 | 6.63 | -3.07 |
| Michigan.N | UT_4020 | 9.5 | NA | Prostate | 9.Normal | 0.00 | 0.44 | 0.00 | 0.96 | 1.21 | 7.42 | -3.18 |
| Michigan.N | UT_4020 | 9.7 | NA | Prostate | 9.Normal | 0.00 | 0.44 | 0.00 | 0.96 | 3.09 | 4.78 | -2.03 |
| Michigan.N | UT_4021 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.56 | 0.00 | 0.96 | 1.73 | 4.63 | -2.77 |
| Michigan.N | UT_4021 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.45 | 0.00 | 0.95 | 0.84 | 5.57 | -3.59 |
| Michigan.N | UT_4022 | 9.1 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.00 | 0.96 | -1.74 | 8.51 | -3.85 |
| Michigan.N | UT_4022 | 9.7 | NA | Prostate | 9.Normal | 0.00 | 0.55 | 0.00 | 0.96 | 1.89 | -0.05 | 0.1 |
| Michigan.N | UT_4023 | 9.5 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.00 | 0.96 | 0.03 | 7.14 | -3.71 |
| Michigan.N | UT_4023 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.49 | 0.00 | 0.96 | -0.09 | 8.73 | -3.45 |
| Michigan.N | UT_4024 | 9.5 | NA | Prostate | 9.Normal | 0.00 | 0.45 | 0.01 | 0.97 | 2.31 | 4.13 | -2.38 |
| Michigan.N | UT_4024 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.49 | 0.00 | 0.96 | 2.52 | 2.04 | -1.73 |
| Michigan.N | UT_4025 | 9.1 | NA | Prostate | 9.Normal | 0.00 | 0.51 | 0.01 | 0.96 | -0.19 | 8.79 | -3.63 |
| Michigan.N | UT_4025 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.53 | 0.01 | 0.95 | -0.52 | 3.64 | -2.47 |
| Michigan.N | UT_4026 | 8.9 | NA | Prostate | 9.Normal | 0.00 | 0.55 | 0.01 | 0.95 | -1.42 | 9.22 | -4.33 |
| Michigan.N | UT_4026 | 8.6 | NA | Prostate | 9.Normal | 0.00 | 0.51 | 0.00 | 0.96 | -3.22 | 12.3 | -5.09 |
| Michigan.N | UT_4027 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.00 | 0.96 | 1.71 | 5.94 | -3.08 |
| Michigan.N | UT_4027 | 9.2 | NA | Prostate | 9.Normal | 0.00 | 0.54 | 0.00 | 0.95 | 3.23 | 2.71 | -0.92 |
| Michigan.N | UT_4028 | 7 | NA | Prostate | 9.Normal | 0.00 | 0.40 | 0.00 | 0.95 | 3.69 | -2.45 | 2.59 |
| Michigan.N | UT_4028 | 8.8 | NA | Prostate | 9.Normal | 0.00 | 0.46 | 0.00 | 0.95 | -1.66 | 10 | -4.36 |
| Michigan.N | UT_4029 | 7.9 | NA | Prostate | 9.Normal | 0.00 | 0.61 | 0.01 | 0.94 | -2.05 | 3.04 | -0.97 |
| Michigan.N | UT_4029 | 9 | NA | Prostate | 9.Normal | 0.00 | 0.61 | 0.00 | 0.95 | 0 | 4.98 | -2.18 |
| Michigan.N | UT_4030 | 9 | NA | Prostate | 9.Normal | 0.00 | 0.54 | 0.00 | 0.95 | 2.08 | 3.44 | -1.05 |
| Michigan.N | UT_4030 | 9 | NA | Prostate | 9.Normal | 0.00 | 0.52 | 0.00 | 0.95 | -0.83 | 9.17 | -3.85 |
| Michigan.N | UT_4031 | 8.6 | NA | Prostate | 9.Normal | 0.00 | 0.67 | 0.00 | 0.95 | -3.26 | 10.41 | -4.17 |
| Michigan.N | UT_4031 | 9.3 | NA | Prostate | 9.Normal | 0.00 | 0.59 | 0.00 | 0.96 | 0.75 | 6.17 | -1.89 |

**Table C.1 10/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| Michigan.N | UT_4032 | 9.4 | NA | Prostate | 9.Normal | 0.00 | 0.50 | 0.00 | 0.95 | -1.68 | 10.12 | -4.15 |
| Michigan.N | UT_4032 | 8.9 | NA | Prostate | 9.Normal | 0.00 | 0.53 | 0.00 | 0.95 | -1.78 | 8.84 | -3.55 |
| Michigan.N | UT_4033 | 9.6 | NA | Prostate | 9.Normal | 0.00 | 0.46 | 0.00 | 0.95 | 1.2 | 1.21 | -0.06 |
| Michigan.N | UT_4033 | 7.3 | NA | Prostate | 9.Normal | 0.00 | 0.49 | 0.00 | 0.87 | -7.02 | 14.66 | -3.34 |
| Michigan.N | UT_4034 | 9.1 | NA | Prostate | 9.Normal | 0.00 | 0.63 | 0.01 | 0.95 | 0.49 | 6.53 | -2.93 |
| Michigan.N | UT_4034 | 8.9 | NA | Prostate | 9.Normal | 0.00 | 0.51 | 0.00 | 0.94 | -1.99 | 10.67 | -5.73 |
| Michigan.N | UT_4035 | 9.1 | NA | Prostate | 9.Normal | 0.00 | 0.84 | 0.02 | 0.91 | -3.82 | 5.67 | -3.03 |
| Michigan.N | UT_4035 | 7.9 | NA | Prostate | 9.Normal | 0.00 | 0.76 | 0.01 | 0.92 | -3.22 | 4.54 | -2.35 |
| SU2C | MO_1015 | 4.9 | NA | Lymph_Node | 1.ERG | 0.66 | 1.88 | 0.02 | 0.73 | -6.26 | -4.48 | 1.12 |
| SU2C | MO_1040 | 6.8 | NA | Lymph_Node | 1.ERG | 0.70 | 0.79 | 0.01 | 0.89 | 2.85 | -4.48 | 1.73 |
| SU2C | MO_1071 | 7.4 | NA | Prostate | 1.ERG | 0.33 | 0.82 | 0.01 | 0.92 | -4.38 | 3.92 | -2.22 |
| SU2C | MO_1084 | 5.7 | NA | Lymph_Node | 1.ERG | 0.81 | 1.29 | 0.02 | 0.82 | -5.08 | -13.83 | 5.55 |
| SU2C | MO_1095 | 2.5 | NA | Soft_Tissue | 1.ERG | 0.40 | 0.68 | 0.01 | 0.87 | -5.6 | -8.6 | 3.22 |
| SU2C | MO_1114 | 9.8 | NA | Lymph_Node | 1.ERG | 0.88 | 0.80 | 0.00 | 0.91 | 0.43 | -6.72 | 3.86 |
| SU2C | MO_1118 | 10 | NA | Liver | 1.ERG | 0.91 | 0.56 | 0.00 | 0.91 | -6.74 | -4.59 | -4.92 |
| SU2C | MO_1124 | 9.3 | NA | Soft_Tissue | 1.ERG | 0.95 | 0.54 | 0.01 | 0.94 | -6.17 | -5.95 | 1.51 |
| SU2C | MO_1161 | 9.7 | NA | Liver | 1.ERG | 0.82 | 0.73 | 0.01 | 0.92 | -4.09 | -6.18 | -0.02 |
| SU2C | MO_1176 | 8.3 | NA | Lymph_Node | 1.ERG | 0.86 | 0.70 | 0.01 | 0.92 | -1.78 | -6.79 | 3.08 |
| SU2C | MO_1179 | 5.9 | NA | Bone | 1.ERG | 0.30 | 0.88 | 0.01 | 0.91 | -6.64 | -6.73 | 3.38 |
| SU2C | MO_1192 | 9.9 | NA | Lymph_Node | 1.ERG | 0.68 | 0.51 | 0.00 | 0.93 | 3.53 | -4.25 | 2.45 |
| SU2C | MO_1202 | 9.7 | NA | Liver | 1.ERG | 0.47 | 0.59 | 0.01 | 0.92 | -3.37 | -3.02 | 1.95 |
| SU2C | MO_1215 | 9.8 | NA | Soft_Tissue | 1.ERG | 0.88 | 0.52 | 0.00 | 0.93 | -7.38 | -2.4 | -5.37 |
| SU2C | MO_1241 | 9.5 | NA | Liver | 1.ERG | 0.82 | 0.51 | 0.01 | 0.93 | -3.78 | -6.62 | 2.12 |
| SU2C | MO_1244 | 9.3 | NA | Liver | 1.ERG | 0.94 | 0.56 | 0.01 | 0.93 | -2.29 | -6.97 | 5.44 |
| SU2C | MO_1249 | 10 | NA | Lymph_Node | 1.ERG | 0.48 | 0.60 | 0.01 | 0.92 | -3 | -6.97 | 3.46 |
| SU2C | MO_1277 | 9.8 | NA | Bone | 1.ERG | 0.87 | 0.71 | 0.00 | 0.94 | 0.93 | -6.71 | 5.15 |
| SU2C | MO_1316 | NA | NA | Lymph_Node | 1.ERG | 0.85 | 0.63 | 0.01 | 0.90 | 0.09 | -6.83 | 3.36 |
| SU2C | MO_1337 | 9.6 | NA | Liver | 1.ERG | 0.84 | 0.46 | 0.00 | 0.92 | -3.47 | -4.04 | 2.95 |
| SU2C | SC_9009 | 7.3 | NA | Lymph_Node | 1.ERG | 0.86 | 1.26 | 0.03 | 0.89 | -3.22 | -7.66 | 4.11 |
| SU2C | SC_9016 | 9.9 | NA | Lymph_Node | 1.ERG | 0.32 | 1.66 | 0.02 | 0.87 | -2.74 | -6.56 | 3.06 |
| SU2C | SC_9017 | 9.9 | NA | Liver | 1.ERG | 0.74 | 0.82 | 0.01 | 0.90 | -6.1 | -2.78 | 1.64 |
| SU2C | SC_9018 | 8.6 | NA | Bone | 1.ERG | 0.55 | 0.82 | 0.00 | 0.92 | -0.79 | -8.66 | 3.54 |
| SU2C | SC_9022 | 10 | NA | Soft_Tissue | 1.ERG | 0.55 | 0.64 | 0.01 | 0.94 | -3.98 | -1.29 | 0.74 |
| SU2C | SC_9026 | 6.9 | NA | Bone | 1.ERG | 0.39 | 1.18 | 0.00 | 0.82 | -3.57 | -7.23 | 3.03 |
| SU2C | SC_9034 | 4.2 | NA | Bone | 1.ERG | 0.30 | 0.77 | 0.01 | 0.86 | -5.98 | -3.37 | 2.28 |
| SU2C | SC_9035 | 9.6 | NA | Lymph_Node | 1.ERG | 0.63 | 0.43 | 0.01 | 0.94 | -1.55 | -6.37 | 3.53 |
| SU2C | SC_9037 | 9.1 | NA | Liver | 1.ERG | 0.30 | 0.72 | 0.00 | 0.88 | -5.67 | -6.02 | 1.42 |
| SU2C | SC_9043 | 9.3 | NA | Bone | 1.ERG | 0.75 | 0.82 | 0.01 | 0.89 | -0.55 | -5.69 | 3.67 |
| SU2C | SC_9046 | 8 | NA | Liver | 1.ERG | 0.30 | 1.06 | 0.00 | 0.86 | -1.85 | -6.09 | 2.62 |
| SU2C | SC_9049 | 8 | NA | Bone | 1.ERG | 0.30 | 0.65 | 0.00 | 0.89 | -4.5 | -3.53 | 2.72 |
| SU2C | SC_9050 | 8.2 | NA | Lymph_Node | 1.ERG | 0.75 | 1.19 | 0.01 | 0.90 | -4.49 | -9.13 | 3.62 |
| SU2C | SC_9056 | 10 | NA | Bone | 1.ERG | 0.45 | 0.44 | 0.00 | 0.95 | 0.23 | -8.04 | 2.45 |
| SU2C | SC_9059 | 8.7 | NA | Lymph_Node | 1.ERG | 0.82 | 0.60 | 0.00 | 0.93 | -2.01 | -5.44 | -0.84 |
| SU2C | SC_9060 | 9.3 | NA | Liver | 1.ERG | 0.41 | 0.57 | 0.00 | 0.93 | -6.18 | -1.21 | 0.56 |
| SU2C | SC_9061 | 8.9 | NA | Bone | 1.ERG | 0.65 | 0.53 | 0.00 | 0.94 | 0.28 | -4.57 | 4.37 |
| SU2C | SC_9063 | 9.6 | NA | Lymph_Node | 1.ERG | 0.60 | 0.53 | 0.00 | 0.94 | -2.78 | -4.25 | 0.87 |
| SU2C | SC_9068 | 8.9 | NA | Lymph_Node | 1.ERG | 0.58 | 0.77 | 0.01 | 0.89 | -1.79 | -4.3 | 2.29 |

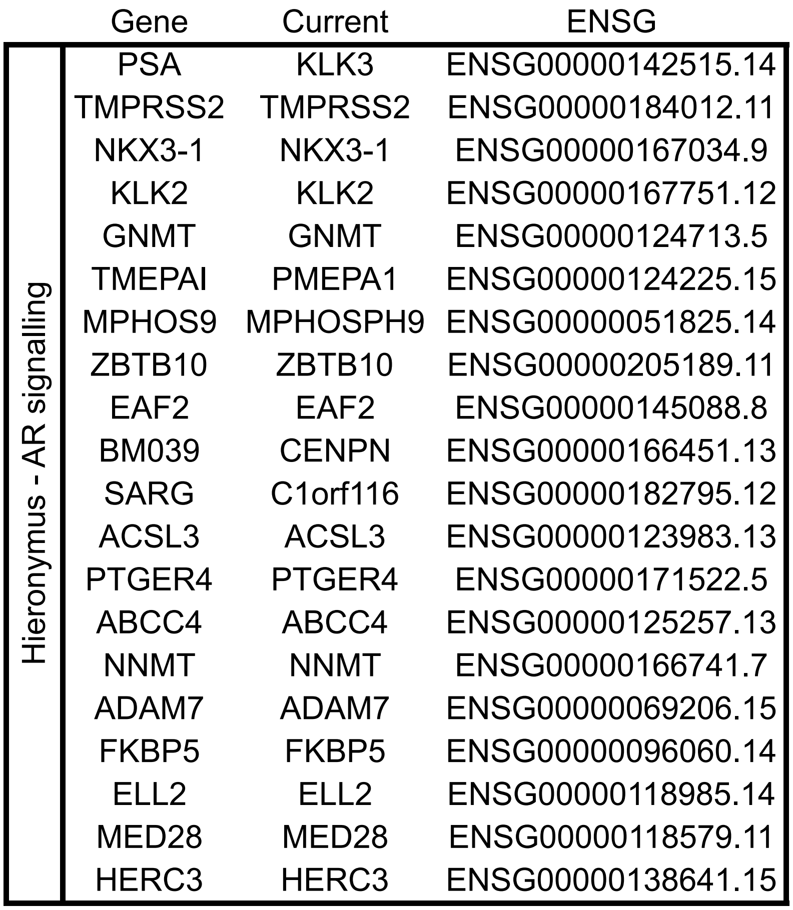
**Table C.1 11/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| SU2C | SC_9071 | 7.6 | NA | Bone | 1.ERG | 0.30 | 1.15 | 0.00 | 0.90 | -6.76 | -5.81 | 3.16 |
| SU2C | SC_9086 | 6.8 | NA | Bone | 1.ERG | 0.42 | 0.79 | 0.01 | 0.91 | -2.11 | -3.48 | 3.02 |
| SU2C | SC_9092 | NA | NA | Bone | 1.ERG | 0.75 | 0.31 | 0.01 | 0.94 | -1.08 | -3.28 | 2.78 |
| SU2C | SC_9097 | 9.6 | NA | Lymph_Node | 1.ERG | 0.82 | 0.28 | 0.00 | 0.95 | -0.42 | -6.02 | 5.45 |
| SU2C | SC_9099 | 9.6 | NA | Lymph_Node | 1.ERG | 0.86 | 0.42 | 0.01 | 0.93 | 1.68 | -6.5 | 3.16 |
| SU2C | SC_9104 | 8.8 | NA | Lymph_Node | 1.ERG | 0.65 | 0.47 | 0.00 | 0.93 | -1.86 | -6.47 | 3.18 |
| SU2C | SC_9107 | 9.7 | NA | Liver | 1.ERG | 0.60 | 0.56 | 0.01 | 0.93 | -3.48 | -5.44 | 3.17 |
| SU2C | SC_9109 | 9.8 | NA | Lymph_Node | 1.ERG | 0.51 | 0.24 | 0.00 | 0.94 | 2.45 | -5.18 | 3.5 |
| SU2C | 1115154 | NA | NA | Lymph_Node | 1.ERG | 0.80 | 0.18 | 0.00 | 0.89 | 0.1 | -4.74 | 4.03 |
| SU2C | 1115156 | NA | NA | Lymph_Node | 1.ERG | 0.69 | 0.32 | 0.00 | 0.89 | 1.03 | -1.95 | 1.03 |
| SU2C | 1115157 | NA | NA | Lymph_Node | 1.ERG | 0.38 | 0.29 | 0.00 | 0.89 | -1.89 | -1.04 | 0.9 |
| SU2C | 1115183 | NA | NA | Bone | 1.ERG | 0.64 | 0.23 | 0.00 | 0.78 | 0.79 | -3.16 | 2.98 |
| SU2C | 1115244 | NA | NA | Bone | 1.ERG | 0.50 | 0.38 | 0.01 | 0.80 | -1.37 | -2.56 | 2.07 |
| SU2C | 6115117 | NA | NA | Soft_Tissue | 1.ERG | 0.52 | 0.67 | 0.01 | 0.87 | -4.61 | -3.31 | 3.41 |
| SU2C | 6115121 | NA | NA | Soft_Tissue | 1.ERG | 0.67 | 1.38 | 0.02 | 0.86 | -4.5 | -9.14 | 4.06 |
| SU2C | 6115122 | NA | NA | Lymph_Node | 1.ERG | 0.56 | 0.56 | 0.00 | 0.85 | -5.42 | 2.24 | 0.55 |
| SU2C | 6115219 | NA | NA | Lymph_Node | 1.ERG | 0.74 | 0.66 | 0.01 | 0.81 | 1.09 | -5.45 | 2.5 |
| SU2C | 6115234 | NA | NA | Lymph_Node | 1.ERG | 0.32 | 0.64 | 0.01 | 0.83 | -1.11 | -4.96 | 4.17 |
| SU2C | 6115247.1 | NA | NA | Lymph_Node | 1.ERG | 0.80 | 0.72 | 0.01 | 0.82 | -0.25 | -4.17 | 2.25 |
| SU2C | TP_2001 | 8.7 | NA | Lymph_Node | 1.ERG | 0.50 | 0.91 | 0.01 | 0.91 | -0.38 | -5.44 | 2.3 |
| SU2C | TP_2034 | 7.6 | NA | Bone | 1.ERG | 0.51 | 0.57 | 0.01 | 0.90 | -6.63 | 6.55 | -3.37 |
| SU2C | TP_2054 | 9 | NA | Lymph_Node | 1.ERG | 0.45 | 0.45 | 0.00 | 0.92 | -3.05 | -1.5 | 2.26 |
| SU2C | MO_1221 | 6.6 | NA | Bone | 2.ETV1 | 0.79 | 0.82 | 0.00 | 0.89 | -2.25 | -4.85 | 3.41 |
| SU2C | SC_9019 | 9.4 | NA | Bone | 2.ETV1 | 0.42 | 0.66 | 0.01 | 0.91 | -2.7 | -2.52 | 2.35 |
| SU2C | SC_9027 | 7.9 | NA | Bone | 2.ETV1 | 0.44 | 0.45 | 0.00 | 0.95 | -2.3 | -7.1 | 3.35 |
| SU2C | SC_9028 | 6.3 | NA | Prostate | 2.ETV1 | 0.30 | 0.85 | 0.01 | 0.90 | -7.59 | -5.47 | 4.03 |
| SU2C | SC_9055 | 9.6 | NA | Lymph_Node | 2.ETV1 | 0.33 | 0.60 | 0.00 | 0.93 | -0.85 | -3.26 | 2.44 |
| SU2C | SC_9057 | 8.5 | NA | Soft_Tissue | 2.ETV1 | 0.72 | 0.56 | 0.00 | 0.94 | -2.85 | -4.96 | 2.38 |
| SU2C | SC_9072 | 10 | NA | Bone | 2.ETV1 | 0.45 | 0.41 | 0.00 | 0.93 | 0.94 | -8.6 | 4.11 |
| SU2C | 6115114 | NA | NA | Soft_Tissue | 2.ETV1 | 0.89 | 0.78 | 0.00 | 0.79 | -2.24 | -10.39 | 5.15 |
| SU2C | 6115118 | NA | NA | Soft_Tissue | 2.ETV1 | 0.54 | 0.47 | 0.00 | 0.87 | -0.68 | -1.83 | 2.81 |
| SU2C | MO_1012 | NA | NA | Soft_Tissue | 3.ETV4 | 0.53 | 2.54 | 0.01 | 0.36 | -8.94 | -3.6 | 1.37 |
| SU2C | MO_1054 | 7.4 | NA | Prostate | 3.ETV4 | 0.30 | 0.63 | 0.01 | 0.91 | -2.97 | -2.27 | 1.23 |
| SU2C | MO_1232 | 7.6 | NA | Soft_Tissue | 3.ETV4 | 0.58 | 0.59 | 0.00 | 0.94 | -5.63 | -0.83 | 1.18 |
| SU2C | MO_1262 | 10 | NA | Lymph_Node | 3.ETV4 | 0.83 | 0.53 | 0.01 | 0.94 | 1.37 | -7.38 | 4.5 |
| SU2C | SC_9001 | 7 | NA | Liver | 3.ETV4 | 0.70 | 1.22 | 0.01 | 0.80 | -6.58 | 1.09 | -4.2 |
| SU2C | SC_9065 | 7 | NA | Bone | 3.ETV4 | 0.63 | 0.69 | 0.00 | 0.84 | -3.65 | -9.87 | 4.79 |
| SU2C | SC_9093 | 9.5 | NA | Soft_Tissue | 3.ETV4 | 0.72 | 0.76 | 0.00 | 0.92 | -2.39 | -8.11 | 5.19 |
| SU2C | 1115153 | NA | NA | Bone | 3.ETV4 | 0.34 | 0.10 | 0.00 | 0.87 | -1.07 | -2.24 | 2.93 |
| SU2C | 6115115 | NA | NA | Lymph_Node | 3.ETV4 | 0.40 | 0.68 | 0.00 | 0.85 | -1.97 | -8.34 | 4.65 |
| SU2C | 6115224 | NA | NA | Soft_Tissue | 3.ETV4 | 0.70 | 0.30 | 0.01 | 0.83 | -1.79 | -6.67 | 4.38 |
| SU2C | 6115237 | NA | NA | Liver | 3.ETV4 | 0.77 | 0.50 | 0.01 | 0.83 | -6.98 | -5.58 | -2.51 |
| SU2C | TP_2009 | 4.8 | NA | Bone | 3.ETV4 | 0.45 | 1.11 | 0.01 | 0.82 | -5.38 | -2.14 | 2.24 |
| SU2C | TP_2020 | 7.6 | NA | Bone | 3.ETV4 | 0.42 | 1.23 | 0.00 | 0.85 | -4.21 | -6.45 | 3.47 |
| SU2C | SC_9007 | 9.2 | NA | Lymph_Node | 4.FLI1 | 0.52 | 0.79 | 0.00 | 0.82 | -1.25 | 1.37 | 0.98 |
| SU2C | MO_1074 | NA | NA | Bone | 5.SPOP | 0.43 | 0.76 | 0.01 | 0.92 | -4.45 | -0.97 | 1.85 |
| SU2C | MO_1128 | 10 | NA | Lymph_Node | 5.SPOP | 0.73 | 0.63 | 0.00 | 0.90 | -4.13 | -1.83 | 3.11 |

**Table C.1 12/13**

|  |  |  |  |  |  |  |  |  |  | Expression Scores |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cohort | sample | RIN | Gleason | Biopsy.Site | Subtype | TC | 3' | Nascent | Aln% | AR | Stroma | NE |
| SU2C | MO_1336 | 4.3 | NA | Bone | 5.SPOP | 0.59 | 0.90 | 0.00 | 0.87 | -1.32 | -5.82 | 3.69 |
| SU2C | SC_9008 | 7.9 | NA | Liver | 5.SPOP | 0.92 | 0.72 | 0.01 | 0.89 | 0.05 | -6.51 | 5.01 |
| SU2C | SC_9094 | 6.6 | NA | Bone | 5.SPOP | 0.55 | 0.41 | 0.00 | 0.89 | -3.73 | -5.91 | 3.78 |
| SU2C | SC_9100 | 7 | NA | Bone | 5.SPOP | 0.20 | 0.55 | 0.00 | 0.87 | -4.43 | -4.68 | 4.01 |
| SU2C | SC_9103 | 9.4 | NA | Bone | 5.SPOP | 0.74 | 0.36 | 0.00 | 0.94 | -3.91 | -6.36 | 4.32 |
| SU2C | TP_2060 | 9.2 | NA | Lymph_Node | 5.SPOP | 0.30 | 1.17 | 0.01 | 0.83 | -3.38 | -3.09 | 2.83 |
| SU2C | SC_9029 | 9.9 | NA | Soft_Tissue | 6.FOXA1 | 0.98 | 0.57 | 0.00 | 0.92 | 0.83 | -12.11 | 6.16 |
| SU2C | SC_9038 | 9.7 | NA | Lymph_Node | 6.FOXA1 | 0.58 | 0.82 | 0.00 | 0.93 | -4.07 | 0.11 | 2.37 |
| SU2C | SC_9047 | 8 | NA | Bone | 6.FOXA1 | 0.65 | 0.80 | 0.00 | 0.87 | -2.24 | -6.76 | 4.81 |
| SU2C | SC_9048 | 8 | NA | Bone | 6.FOXA1 | 0.33 | 0.55 | 0.00 | 0.90 | -3.72 | -3.86 | 2.77 |
| SU2C | SC_9058 | 9.7 | NA | Lymph_Node | 6.FOXA1 | 0.73 | 0.54 | 0.00 | 0.93 | -1.47 | -4.68 | 2.61 |
| SU2C | SC_9091 | 7.9 | NA | Bone | 6.FOXA1 | 0.63 | 0.87 | 0.00 | 0.75 | -3.5 | -10.9 | 5.2 |
| SU2C | 1115161 | NA | NA | Lymph_Node | 6.FOXA1 | 0.87 | 0.33 | 0.00 | 0.81 | 2.15 | -6.29 | 4.94 |
| SU2C | MO_1013 | 6.1 | NA | Lymph_Node | 8.Other | 0.42 | 1.41 | 0.02 | 0.90 | -3.66 | -3.2 | 3.81 |
| SU2C | MO_1014 | NA | NA | Lymph_Node | 8.Other | 0.62 | 1.72 | 0.01 | 0.83 | -4.23 | -5.64 | 4.74 |
| SU2C | MO_1020 | 8 | NA | Liver | 8.Other | 0.36 | 0.90 | 0.01 | 0.88 | -6.39 | 0.33 | -0.9 |
| SU2C | MO_1094 | 9.7 | NA | Bone | 8.Other | 0.85 | 1.64 | 0.02 | 0.91 | -7.15 | -2.33 | 0.61 |
| SU2C | MO_1184 | 9.2 | NA | Liver | 8.Other | 0.56 | 0.87 | 0.00 | 0.90 | -3.36 | -0.95 | 1.12 |
| SU2C | MO_1219 | 8 | NA | Bone | 8.Other | 0.30 | 0.94 | 0.00 | 0.73 | -4.99 | -11.19 | 4.55 |
| SU2C | MO_1339 | 10 | NA | Lymph_Node | 8.Other | 0.79 | 0.37 | 0.00 | 0.93 | -1.24 | -1.98 | -0.44 |
| SU2C | SC_9010 | 7.3 | NA | Lymph_Node | 8.Other | 0.61 | 0.70 | 0.01 | 0.88 | -2.94 | -6.12 | 4.13 |
| SU2C | SC_9012 | 9.6 | NA | Liver | 8.Other | 0.32 | 1.12 | 0.00 | 0.89 | -6.7 | -5.8 | 2.48 |
| SU2C | SC_9023 | 6 | NA | Bone | 8.Other | 0.67 | 0.92 | 0.01 | 0.89 | -5.93 | -6.76 | 2.92 |
| SU2C | SC_9030 | 10 | NA | Soft_Tissue | 8.Other | 0.91 | 0.67 | 0.00 | 0.90 | 1.16 | -9.76 | 5.56 |
| SU2C | SC_9031 | 9.8 | NA | Liver | 8.Other | 0.67 | 0.65 | 0.01 | 0.91 | -6.93 | 1.1 | -6.99 |
| SU2C | SC_9032 | 10 | NA | Lymph_Node | 8.Other | 0.87 | 0.80 | 0.01 | 0.90 | -0.12 | -11.08 | 6.25 |
| SU2C | SC_9036 | 10 | NA | Lymph_Node | 8.Other | 0.46 | 0.49 | 0.00 | 0.90 | 1.63 | -3.91 | 2.96 |
| SU2C | SC_9054 | 9.9 | NA | Lymph_Node | 8.Other | 0.66 | 0.59 | 0.00 | 0.92 | -1.12 | -8.12 | 1.41 |
| SU2C | SC_9062 | 9.5 | NA | Lymph_Node | 8.Other | 0.74 | 0.59 | 0.00 | 0.94 | -5.5 | -2.44 | 3.29 |
| SU2C | SC_9073 | 8.8 | NA | Lymph_Node | 8.Other | 0.38 | 0.52 | 0.00 | 0.93 | 0.86 | -4.93 | 2.75 |
| SU2C | SC_9080 | 9.5 | NA | Lymph_Node | 8.Other | 0.77 | 0.72 | 0.00 | 0.92 | -2.48 | -2.13 | 1.46 |
| SU2C | SC_9081 | 9.8 | NA | Lymph_Node | 8.Other | 0.86 | 0.48 | 0.00 | 0.92 | -0.86 | -3.37 | 2.44 |
| SU2C | SC_9083 | 9 | NA | Bone | 8.Other | 0.62 | 0.63 | 0.01 | 0.89 | -3.1 | -3.49 | 2.8 |
| SU2C | 1115202 | NA | NA | Soft_Tissue | 8.Other | 0.44 | 0.22 | 0.00 | 0.84 | 5.55 | -5.1 | 3.59 |
| SU2C | 6115123 | NA | NA | Soft_Tissue | 8.Other | 0.67 | 1.68 | 0.01 | 0.84 | -5.46 | -3.65 | 2.73 |
| SU2C | 6115227 | NA | NA | Soft_Tissue | 8.Other | 0.87 | 0.39 | 0.01 | 0.83 | -1.33 | -6.75 | 4.51 |
| SU2C | 6115233 | NA | NA | Lymph_Node | 8.Other | 0.84 | 0.41 | 0.01 | 0.84 | -0.9 | -6.03 | 1.94 |
| SU2C | 6115242 | NA | NA | Soft_Tissue | 8.Other | 0.49 | 0.89 | 0.01 | 0.79 | -2.9 | -4.65 | 4.35 |
| SU2C | 6115250.2 | NA | NA | Lymph_Node | 8.Other | 0.95 | 0.62 | 0.01 | 0.84 | 2.93 | -9.32 | 5.03 |
| SU2C | 6115251 | NA | NA | Lymph_Node | 8.Other | 0.73 | 0.46 | 0.01 | 0.83 | -0.22 | -3.58 | 3.03 |
| SU2C | TP_2010 | 7.6 | NA | Lymph_Node | 8.Other | 0.73 | 0.66 | 0.01 | 0.89 | -4.13 | -7.38 | 1.71 |
| SU2C | TP_2032 | 9.4 | NA | Lymph_Node | 8.Other | 0.57 | 1.08 | 0.00 | 0.84 | -3.54 | -2.39 | 3.24 |
| SU2C | TP_2061 | 4.9 | NA | Soft_Tissue | 8.Other | 0.67 | 0.67 | 0.01 | 0.92 | -6.08 | 0.05 | -0.49 |
| SU2C | TP_2064 | 9.9 | NA | Lymph_Node | 8.Other | 0.68 | 0.25 | 0.00 | 0.95 | 4.8 | -4.14 | 4.32 |

**Table C.1 13/13**



| | Gene | Current | ENSG |
|---|---|---|---|
| Hieronymus - AR signalling | PSA | KLK3 | ENSG00000142515.14 |
| | TMPRSS2 | TMPRSS2 | ENSG00000184012.11 |
| | NKX3-1 | NKX3-1 | ENSG00000167034.9 |
| | KLK2 | KLK2 | ENSG00000167751.12 |
| | GNMT | GNMT | ENSG00000124713.5 |
| | TMEPAI | PMEPA1 | ENSG00000124225.15 |
| | MPHOS9 | MPHOSPH9 | ENSG00000051825.14 |
| | ZBTB10 | ZBTB10 | ENSG00000205189.11 |
| | EAF2 | EAF2 | ENSG00000145088.8 |
| | BM039 | CENPN | ENSG00000166451.13 |
| | SARG | C1orf116 | ENSG00000182795.12 |
| | ACSL3 | ACSL3 | ENSG00000123983.13 |
| | PTGER4 | PTGER4 | ENSG00000171522.5 |
| | ABCC4 | ABCC4 | ENSG00000125257.13 |
| | NNMT | NNMT | ENSG00000166741.7 |
| | ADAM7 | ADAM7 | ENSG00000069206.15 |
| | FKBP5 | FKBP5 | ENSG00000096060.14 |
| | ELL2 | ELL2 | ENSG00000118985.14 |
| | MED28 | MED28 | ENSG00000118579.11 |
| | HERC3 | HERC3 | ENSG00000138641.15 |

**Table C.2 Androgen Receptor Target Genes**

A panel of genes whose expression is driven by activity of the Androgen receptor, taken from the manuscript by Hieronymus *et. al*.[207] These genes were used in our derivation of an AR activity signature (Figure C.6), similar to the approach taken by TCGA.[179]

|  | Gene | Current | ENSG | note |
|---|---|---|---|---|
| Beltran - Neuroendocrine expression signature | ASXL3 | ASXL3 | ENSG00000141431.9 |  |
|  | AURKA | AURKA | ENSG00000087586.17 |  |
|  | BRINP1 | BRINP1 | ENSG00000078725.12 |  |
|  | CAND2 | CAND2 | ENSG00000144712.11 |  |
|  | DNMT1 | DNMT1 | ENSG00000130816.14 |  |
|  | ETV5 | ETV5 | ENSG00000244405.7 |  |
|  | EZH2 | EZH2 | ENSG00000106462.10 |  |
|  | GNAO1 | GNAO1 | ENSG00000087258.13 |  |
|  | GPX2 | GPX2 | ENSG00000176153.11 |  |
|  | JAKMIP2 | JAKMIP2 | ENSG00000176049.15 |  |
|  | KCNB2 | KCNB2 | ENSG00000182674.5 |  |
|  | KCND2 | KCND2 | ENSG00000184408.9 |  |
|  | LRRC16B | LRRC16B | ENSG00000186648.14 |  |
|  | MAP10 | MAP10 | ENSG00000212916.4 |  |
|  | MYCN | MYCN | ENSG00000134323.10 |  |
|  | NRSN1 | NRSN1 | ENSG00000152954.11 |  |
|  | PCSK1 | PCSK1 | ENSG00000175426.10 |  |
|  | PROX1 | PROX1 | ENSG00000117707.15 |  |
|  | RGS7 | RGS7 | ENSG00000182901.15 |  |
|  | SCG3 | SCG3 | ENSG00000104112.8 |  |
|  | SEC11C | SEC11C | ENSG00000166562.8 |  |
|  | SEZ6 | SEZ6 | ENSG00000063015.19 |  |
|  | SOGA3 | SOGA3 | ENSG00000255330.8 |  |
|  | ST8SIA3 | ST8SIA3 | ENSG00000177511.5 |  |
|  | SVOP | SVOP | ENSG00000166111.9 |  |
|  | SYT11 | SYT11 | ENSG00000132718.8 |  |
|  | TRIM9 | TRIM9 | ENSG00000100505.13 |  |
|  | C7orf76 | C7orf76 |  | unused, overlaps SHFM1 |
|  | KIAA0408 | KIAA0408 |  | unused, overlaps SOGA3 |
|  | Chromogranin A | CHGA | ENSG00000100604.12 | we added these, previously known markers absent from Beltran list |
|  | Chromogranin B | CHGB | ENSG00000089199.9 |  |
|  | HES6 | HES6 | ENSG00000144485.10 |  |
|  | Synaptophysin | SYP | ENSG00000102003.10 |  |

**Table C.3 Neuroendocrine prostate cancer genes**
A panel of genes whose expression is expected to be up-regulated in the neuroendocrine subtype of prostate cancer, taken from Beltran *et. al*.[208] These genes were used in our derivation of a neuroendocrine signature (Figure C.7).

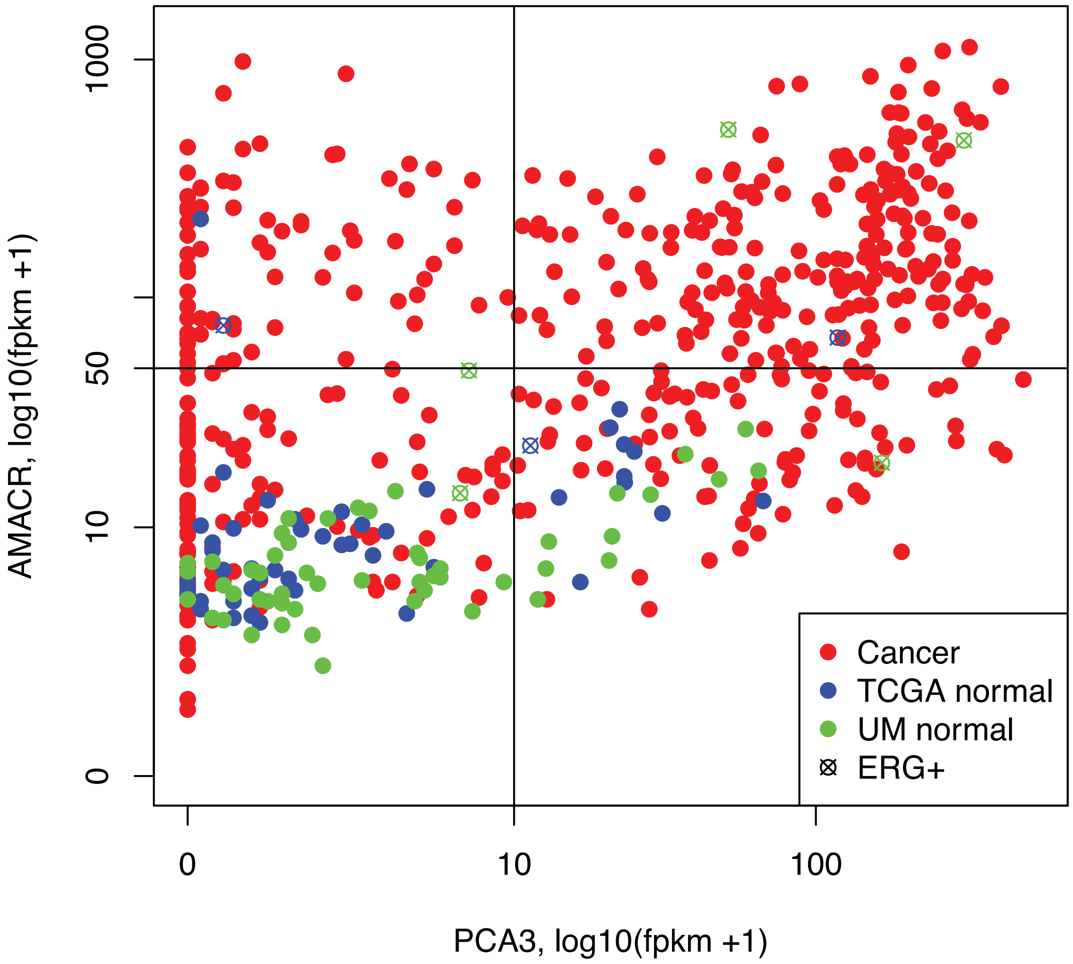


**Figure C.1 PCa-Specific Gene Signature Analysis**
In the course of analyzing normal prostate tissue RNA-seq samples, we became concerned by the presence of known cancer-specific transcripts. We tested for the three most-specific known biomarkers of prostate cancer : ERG, PCA3, and AMACR. Shown here is a scatterplot of log10 expression, of PCA3 and AMACR, with ERG marked if >= 10 fpkm. Normal tissue samples were excluded if : ERG >= 10 fpkm, or PCA3 >= 10 fpkm, or AMACR >= 50 fpkm.

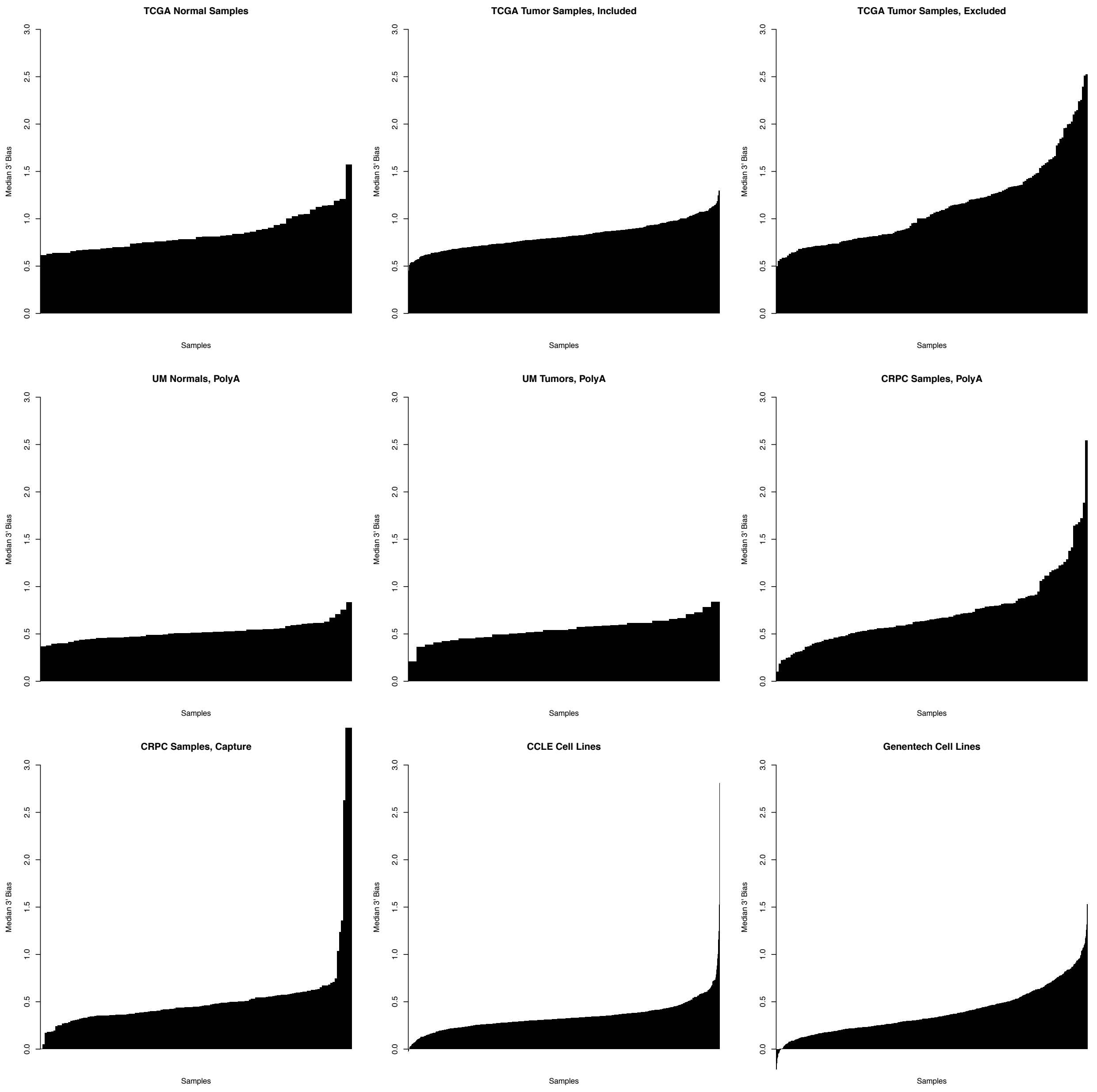


**Figure C.2 Per-cohort Barplots of 3' Bias**

3' Bias was estimated by the median log-ratio imbalance between the last and first splice junction of all unambiguous annotated genes.

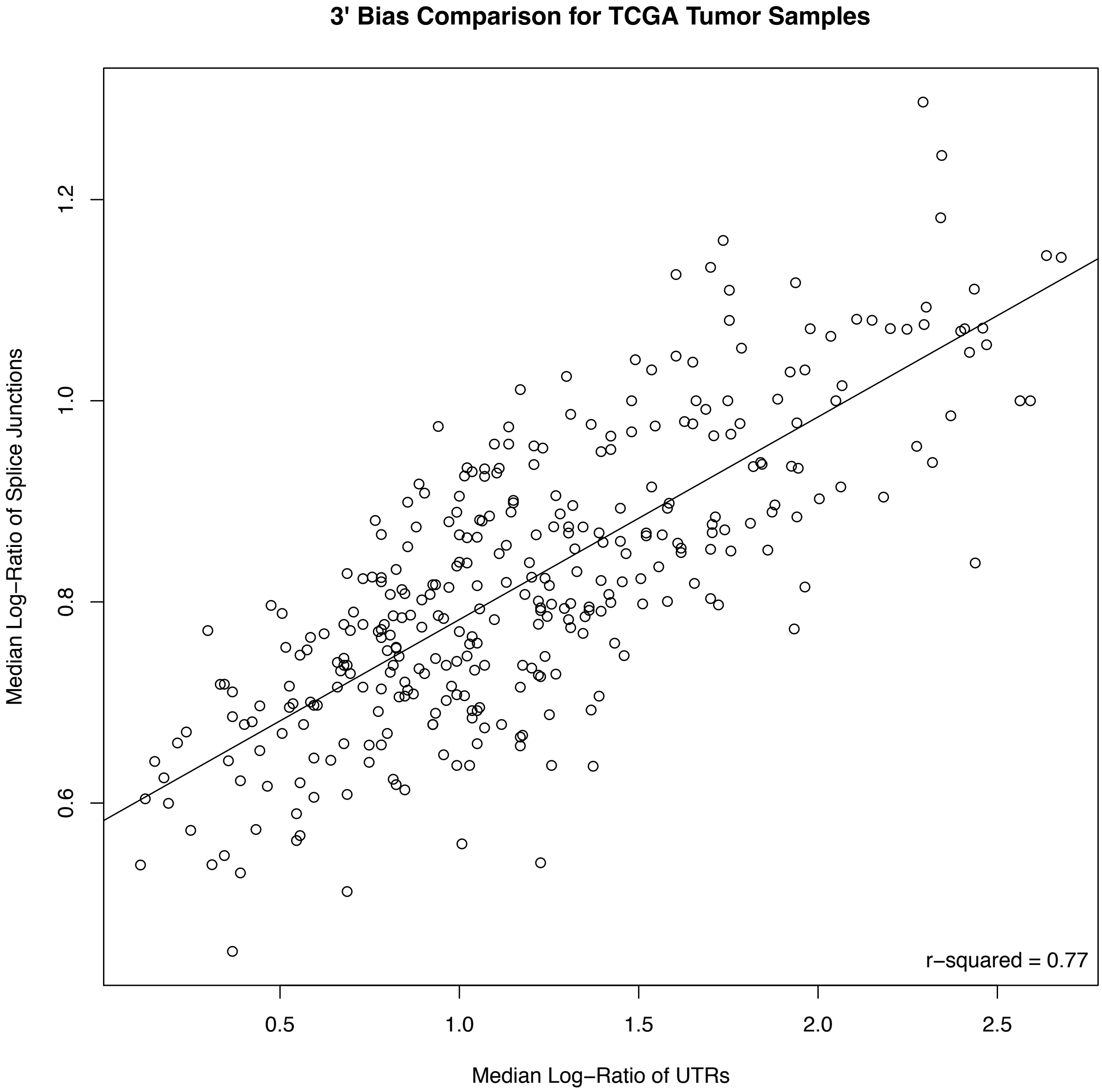


**Figure C.3 Comparison of 3' Bias Estimation with TCGA's 3' Bias Estimation**
Here we compare the 3' bias estimation we generated using junction read depth imbalance, to the 3' bias estimation TCGA generated using 3' UTR average read depth imbalance. The two estimates generally agreed.

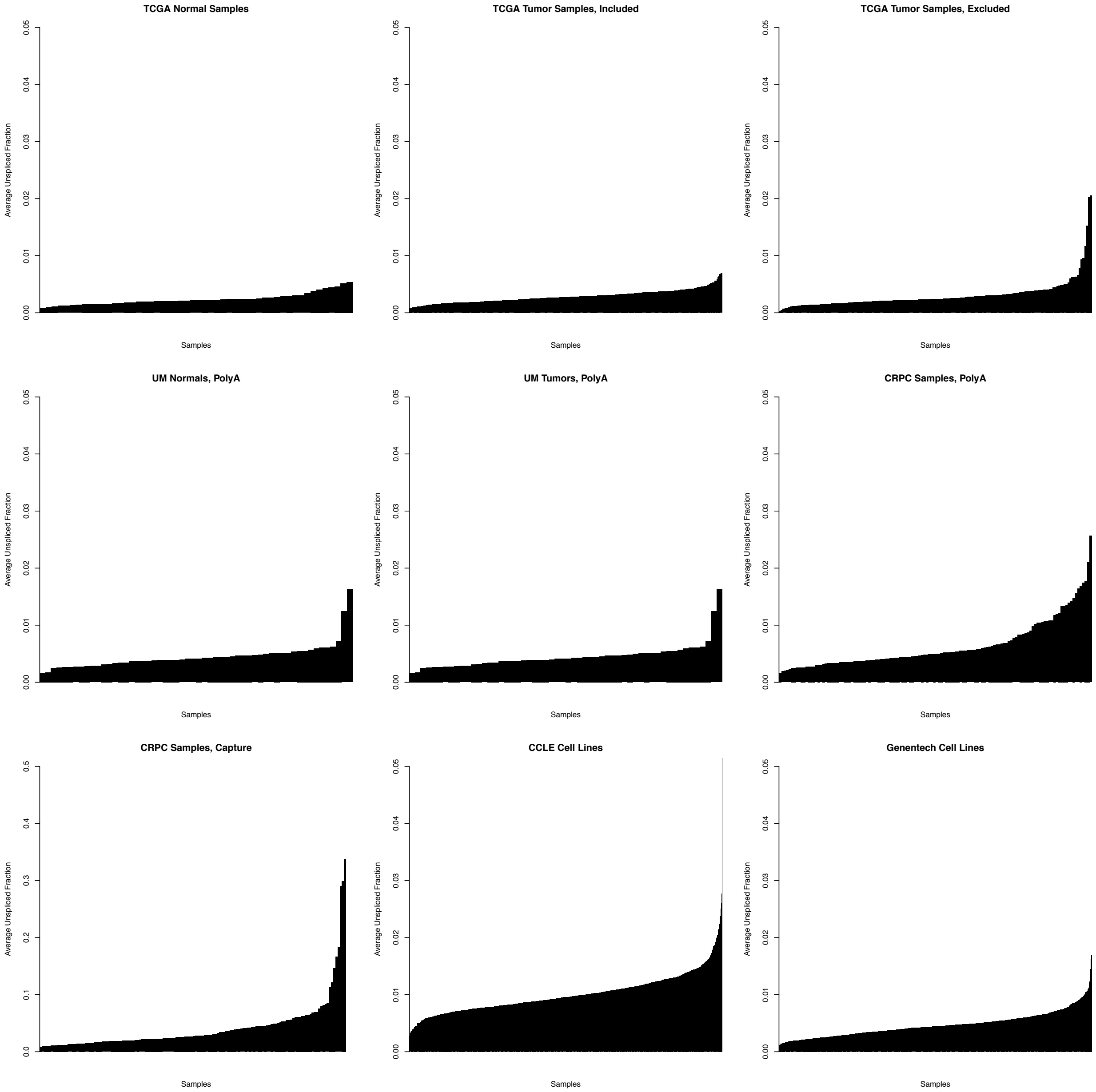


**Figure C.4 Per-cohort barplots of Unspliced RNA content**
Unspliced RNA level was estimated by the median unspliced coverage over junctions from high confidence gene annotations (*i.e.,* across all genes).

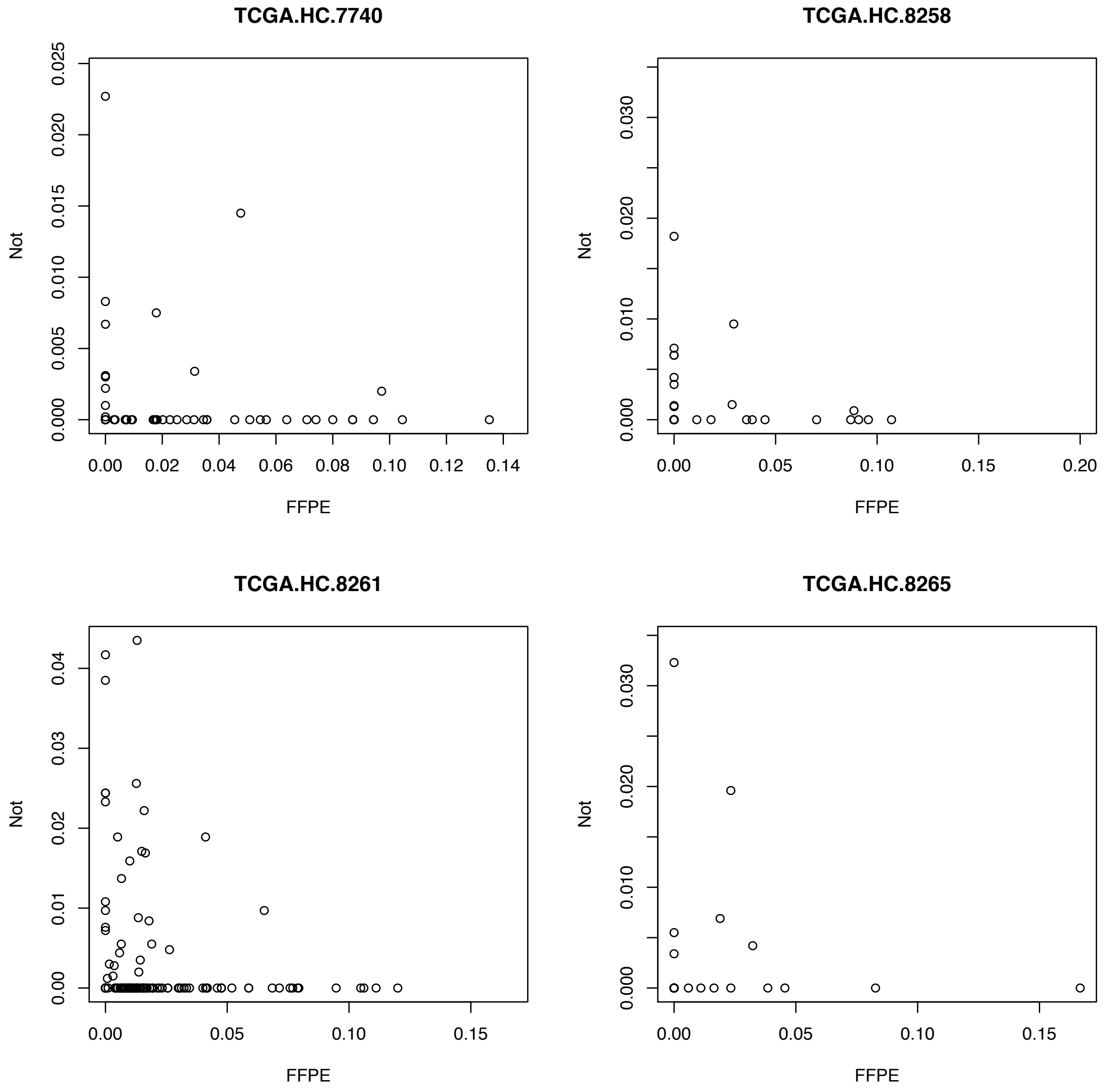


**Figure C.5 Comparison of Unspliced RNA between FFPE and Frozen Samples**
Here we compared the unspliced RNA content as calculated in Figure C.4 between FFPE and Frozen samples for the 4 TCGA primary prostate tumors where both libraries were available, expecting to see that FFPE would have higher levels of unspliced RNA. Points are single unspliced junctions from unambiguous gene annotations. As expected, the unspliced RNA estimate generated here confirmed that FFPE samples have more unspliced RNA (possibly due to quality, or because they were whole transcriptome libraries rather than polyA).

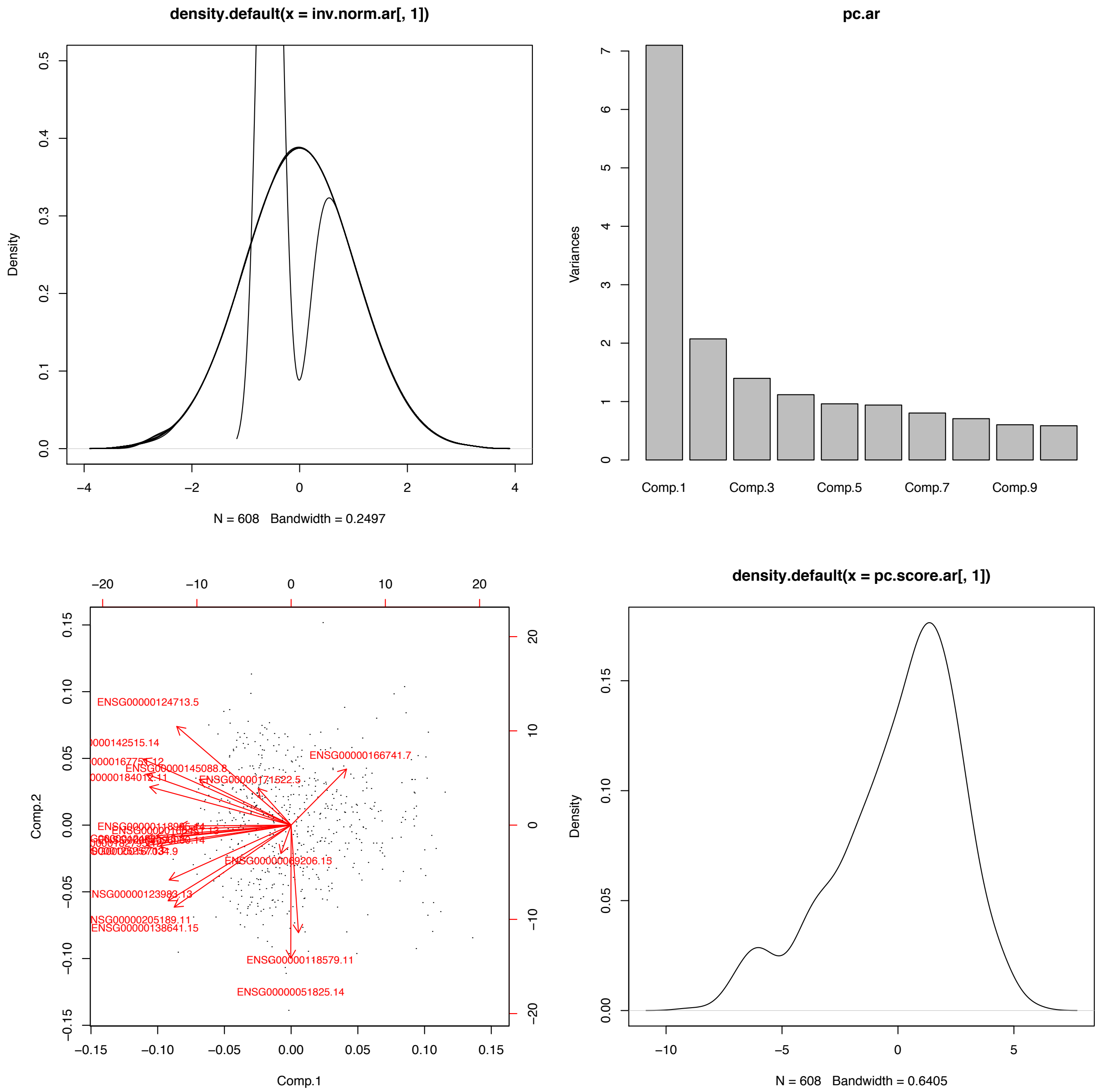


**Figure C.6 Generation of AR activity signature**

The AR activity signature was generated using the first principal component of inverse normal transformed expression as in Figure 4.2, using the genes in Table C.2 from Hieronymus et. al.[207] Individual gene contributions are plotted in red. Samples shown are the full cohort from Chapter 5.

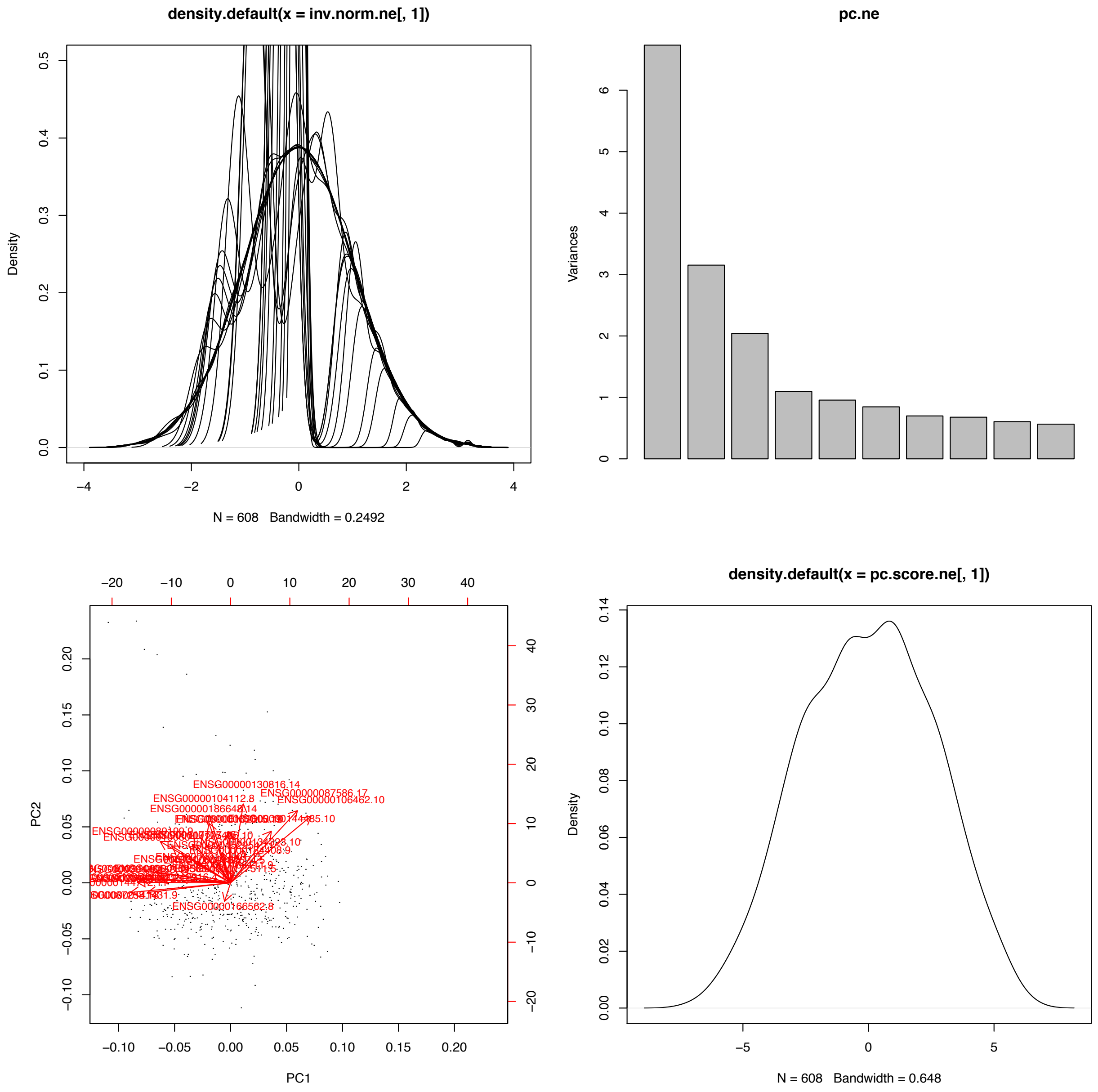


**Figure C.7 Generation of Neuroendocrine expression signature**
The Neuroendocrine signature was generated using the first principal component of inverse normal transformed expression as in Figure 4.2, using the genes in Table C.3 from Beltran et. al.[208] The expression of these genes followed a binary "zero or expressed" pattern, and were originally intended to complement each other rather than reflect a single axis of expression, so our approach here would need to be reworked before drawing significant conclusions from this signature. Individual gene contributions are plotted in red : note also that they don't point in the same direction, again indicating their complementary nature. Samples shown are the full cohort from Chapter 5.

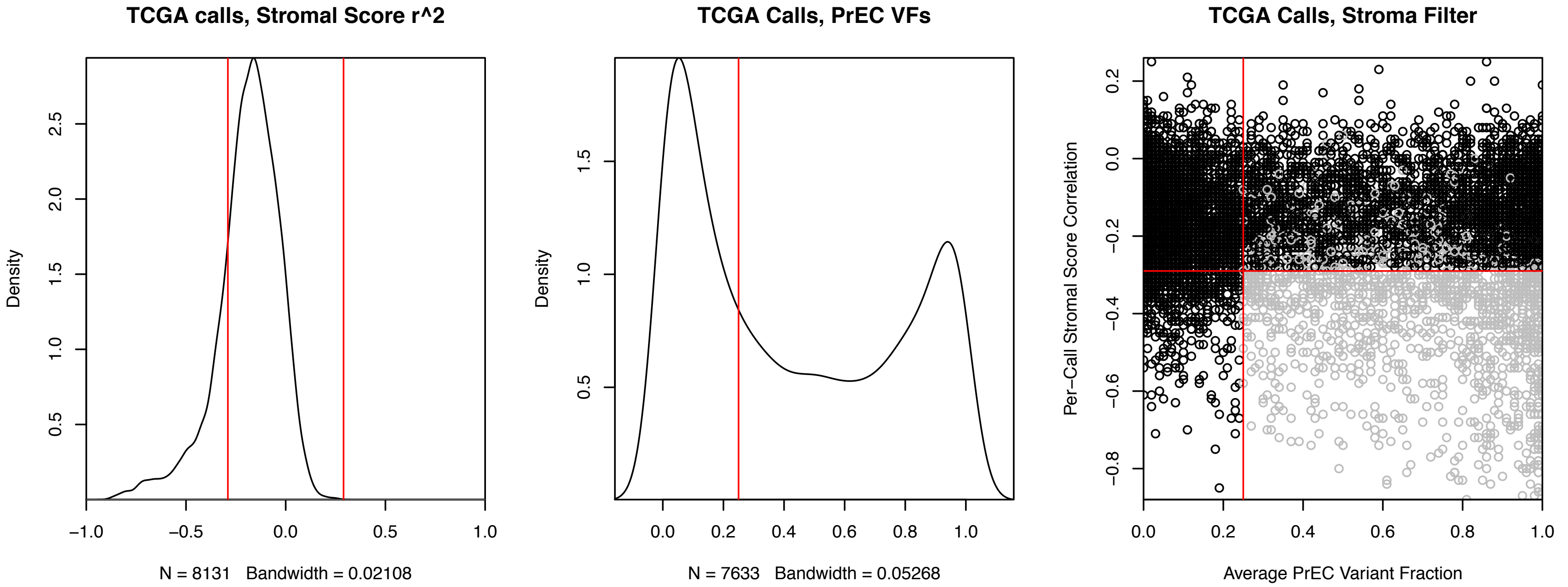


**Figure C.8 Lineage-Specific Variant Filtering, TCGA cohort**
Correlative bias analysis was performed on significant differential variant calls in the TCGA tumor vs. TCGA normal analysis. Left panel) kernel density plot of variant correlation against stromal expression, with red lines indicating the middle 80%. Middle panel) kernel density plot of variant fractions for the same variant in PrEC cells, with a red line indicating the 25% variant fraction marker. Right panel) Scatterplot of the correlations and variant fractions shown in the other two panels, with red lines indicating the same positions - middle 80% of correlation and 25% variant fraction in PrEC. Points plotted in gray were identified as likely lineage-specific variants in this cohort or the other two cohorts.

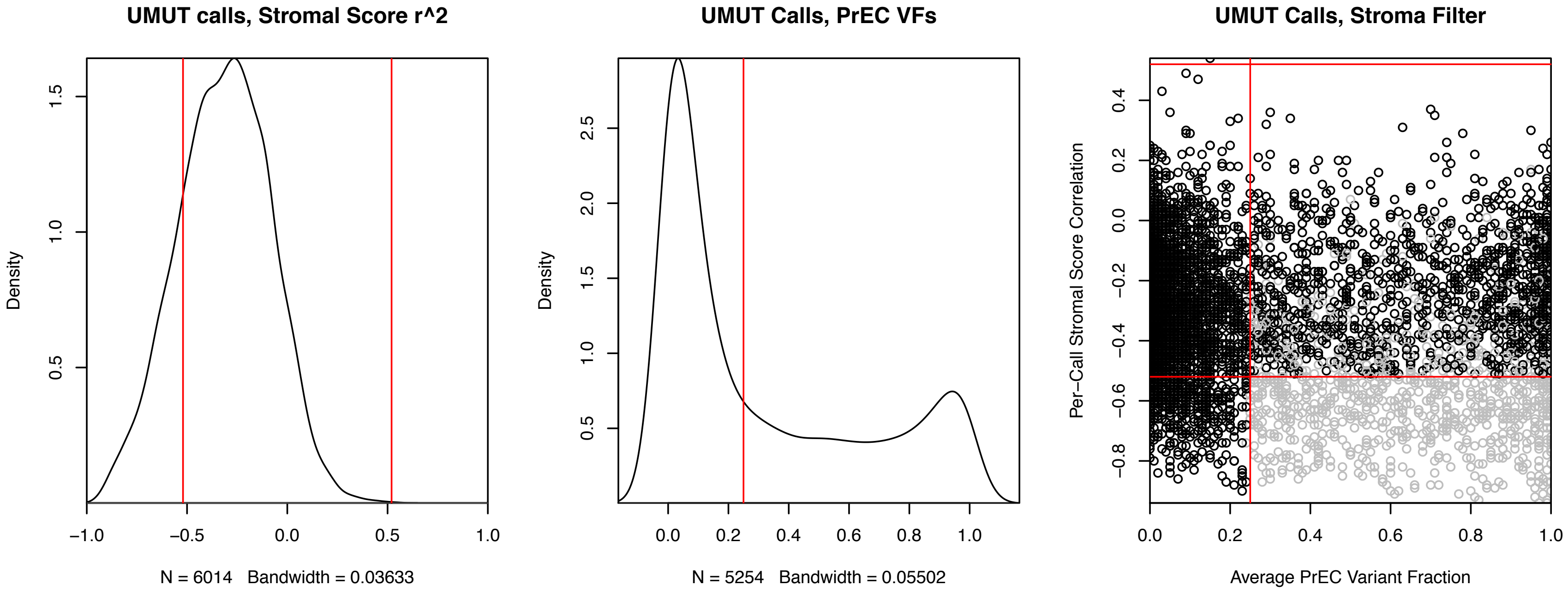


**Figure C.9 Lineage-Specific Variant Filtering, Michigan cohort**
Correlative bias analysis was performed on significant differential variant calls in the Michigan tumor vs. Michigan normal analysis. Left panel) kernel density plot of variant correlation against stromal expression, with red lines indicating the middle 80%. Middle panel) kernel density plot of variant fractions for the same variant in PrEC cells, with a red line indicating the 25% variant fraction marker. Right panel) Scatterplot of the correlations and variant fractions shown in the other two panels, with red lines indicating the same positions - middle 80% of correlation and 25% variant fraction in PrEC. Points plotted in gray were identified as likely lineage-specific variants in this cohort or the other two cohorts.

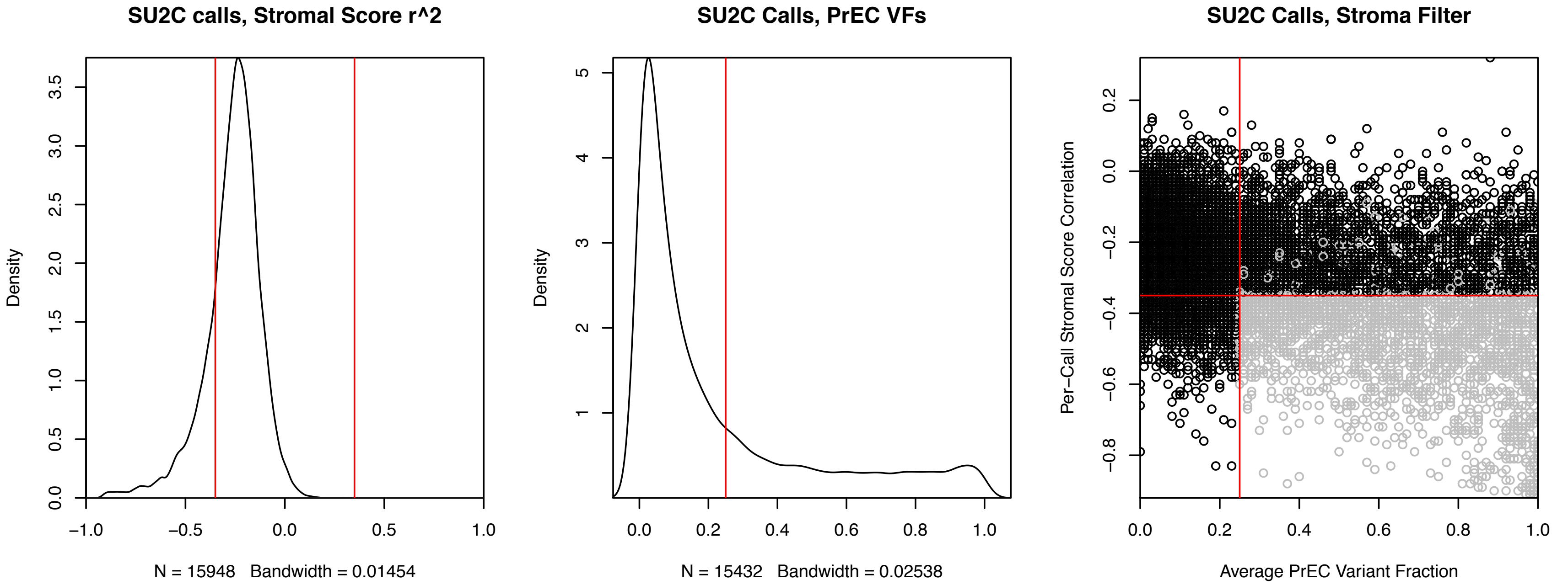


**Figure C.10 Lineage-Specific Variant Filtering, SU2C cohort**
Correlative bias analysis was performed on significant differential variant calls in the SU2C mCRPC vs. Michigan normal analysis. Left panel) kernel density plot of variant correlation against stromal expression, with red lines indicating the middle 80%. Middle panel) kernel density plot of variant fractions for the same variant in PrEC cells, with a red line indicating the 25% variant fraction marker. Right panel) Scatterplot of the correlations and variant fractions shown in the other two panels, with red lines indicating the same positions - middle 80% of correlation and 25% variant fraction in PrEC. Points plotted in gray were identified as likely lineage-specific variants in this cohort or the other two cohorts.

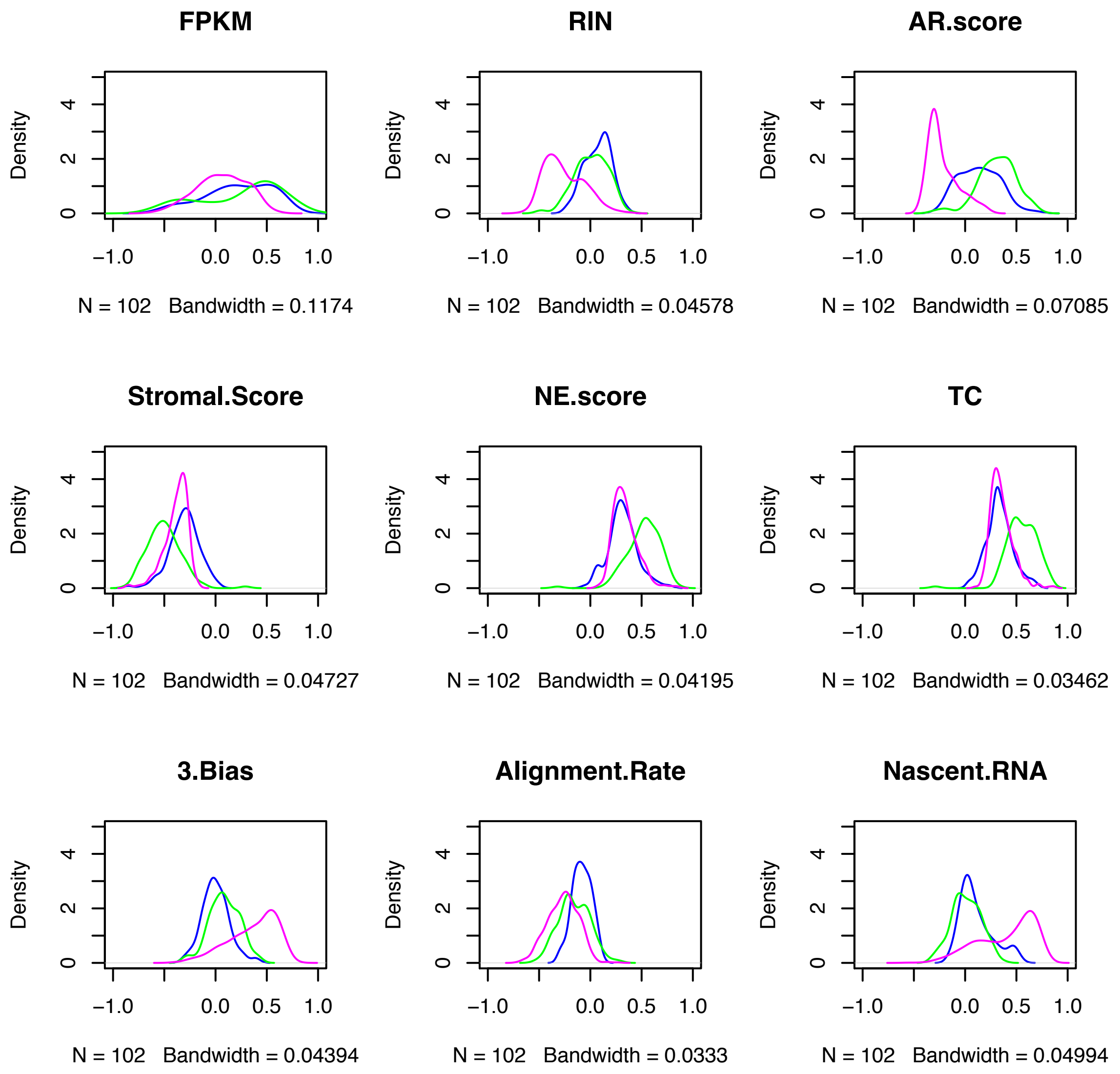


**Figure C.11   Bias Correlation Plots for Significant Differential Splicing Calls** Significant differential splicing was called following the description in Figure 5.1 and filtered for significance and minimum variant fraction shift as in Figure 5.2. Plotted here are the distributions of bias correlations for those calls, where blue is the TCGA cohort, green is the Michigan cohort, and purple is the SU2C cohort.

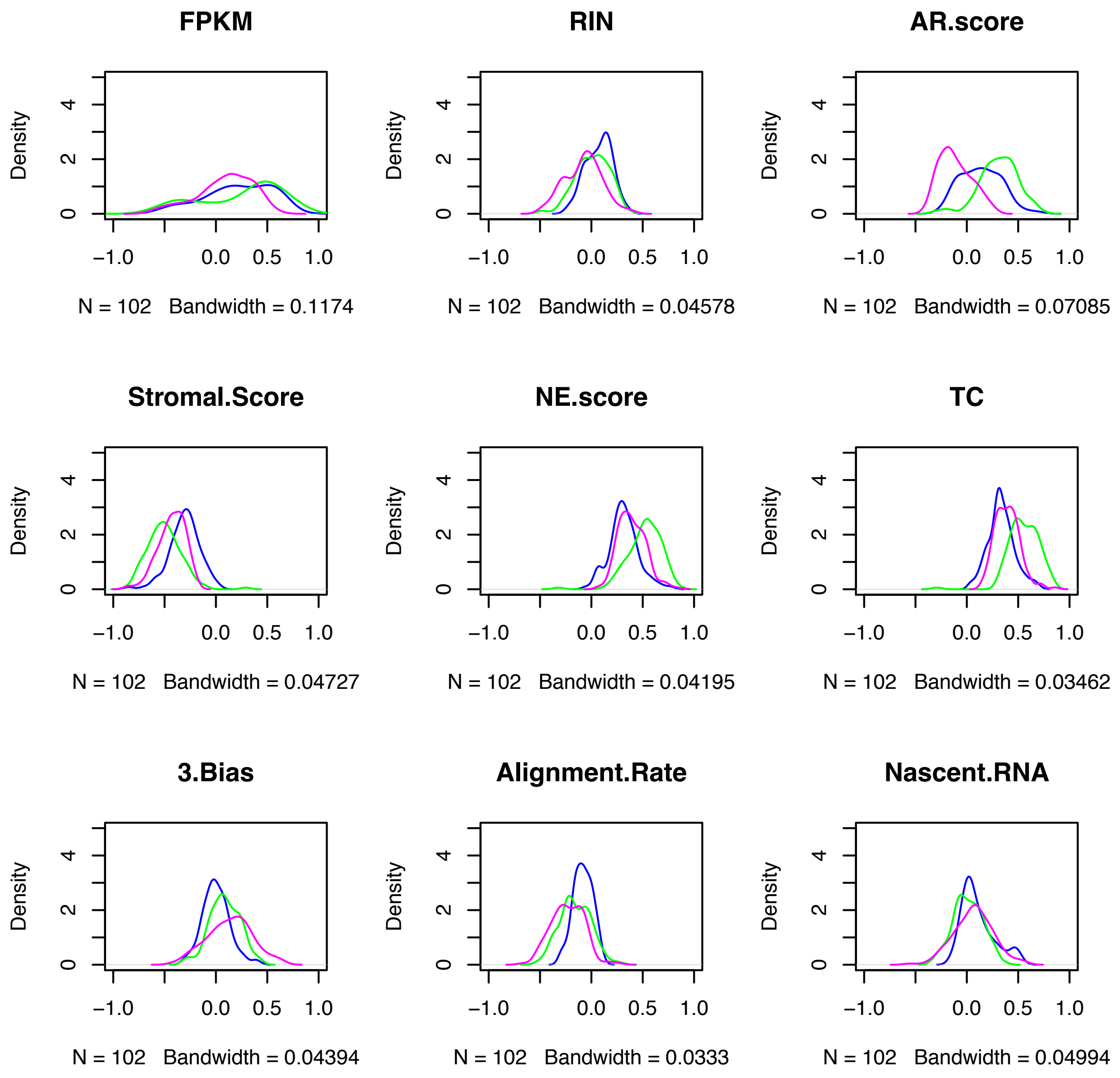


**Figure C.12 Bias Correlation Plots for Significant Differential Splicing Calls, without unspliced calls in SU2C**

Significant differential splicing was called following the description in Figure 5.1 and filtered for significance and minimum variant fraction shift as in Figure 5.2. Additionally, unspliced junction calls were removed from the SU2C cohort. Plotted here are the distributions of bias correlations for those calls, where blue is the TCGA cohort, green is the Michigan cohort, and purple is the SU2C cohort.

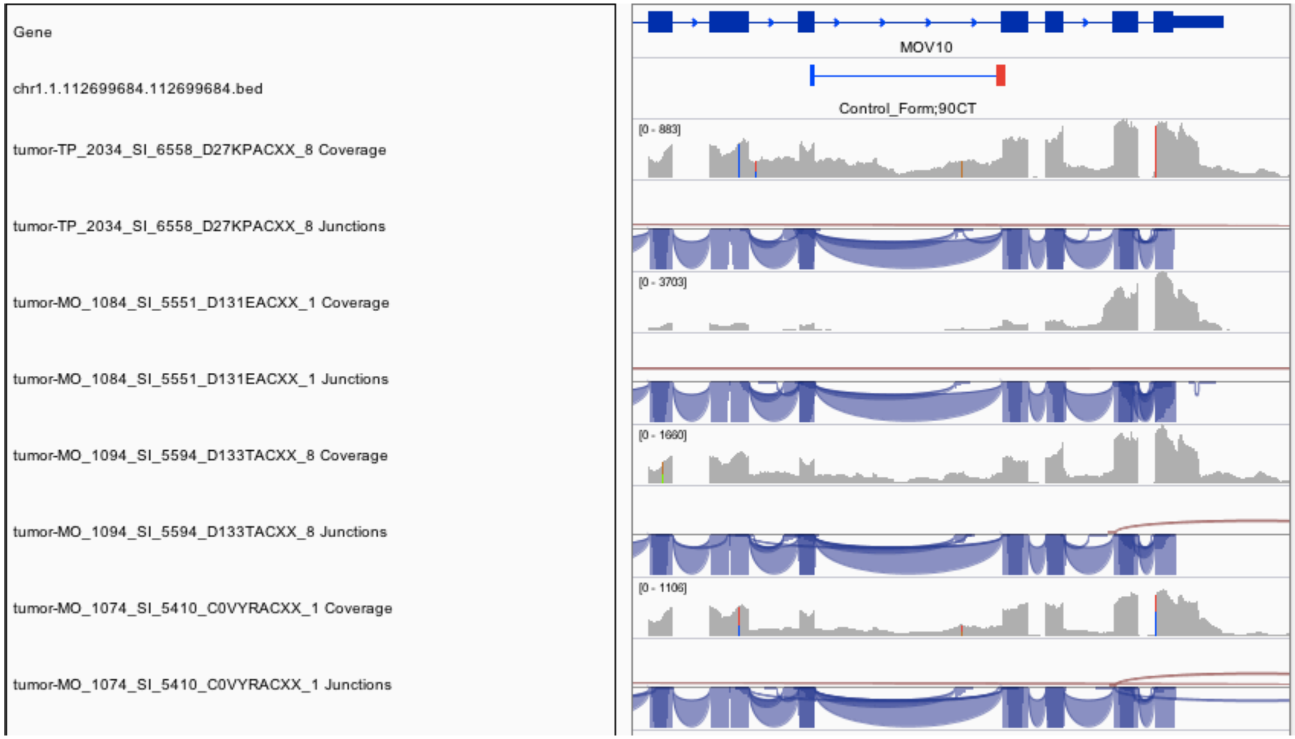

Gene
chr1.1.112699684.112699684.bed
tumor-TP_2034_SI_6558_D27KPACXX_8 Coverage
tumor-TP_2034_SI_6558_D27KPACXX_8 Junctions
tumor-MO_1084_SI_5551_D131EACXX_1 Coverage
tumor-MO_1084_SI_5551_D131EACXX_1 Junctions
tumor-MO_1094_SI_5594_D133TACXX_8 Coverage
tumor-MO_1094_SI_5594_D133TACXX_8 Junctions
tumor-MO_1074_SI_5410_C0VYRACXX_1 Coverage
tumor-MO_1074_SI_5410_C0VYRACXX_1 Junctions
MOV10
Control_Form;90CT
[0 - 883]
[0 - 3703]
[0 - 1660]
[0 - 1106]

**Figure C.13 Intron retention call example from the SU2C cohort**
Plotted are depth of coverage plots in IGV for four mCRPC samples from the SU2C cohort, in which multiple retention variants of the gene MOV10 were called (among many, many other genes). Notably, while some introns are consistently retained, others are consistently spliced at high fidelity, casting doubt on known biases as explanation, including 3' bias, nascent RNA contamination, or genomic DNA contamination. Further work is needed to investigate this phenomenon.

# LITERATURE CITED

1 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA Cancer J Clin* **66**, 7-30, doi:10.3322/caac.21332 (2016).

2 Colby, S. L. & Ortman, J. M. Projections of the Size and Composition of the US Population: 2014 to 2060. *US Census Bureau, Ed*, 25-1143 (2015).

3 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

4 Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat Rev Genet* **13**, 795-806, doi:10.1038/nrg3317 (2012).

5 Darwin, C. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. (John Murray, 1859).

6 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

7 Versteeg, R. Cancer: Tumours outside the mutation box. *Nature* **506**, 438-439, doi:10.1038/nature13061 (2014).

8 Visvader, J. E. Cells of origin in cancer. *Nature* **469**, 314-322, doi:10.1038/nature09781 (2011).

9 The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

10 Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96, doi:10.1186/1471-2164-7-96 (2006).

11 The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

12 Voduc, K. D. *et al.* Breast cancer subtypes and the risk of local and regional relapse. *J Clin Oncol* **28**, 1684-1691, doi:10.1200/JCO.2009.24.9284 (2010).

13 Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

14 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

15 American Cancer Society. *Prostate cancer detailed guide, last revised: 3/11/2016*, <http://www.cancer.org/cancer/prostatecancer/detailedguide/prostate-cancer-references> (2016).

16 Skene, A. J. C. *The anatomy and pathology of two important glands of the female urethra*. (William Wood & Company, 1880).

17 Zaviacic, M. & Ablin, R. J. The female prostate. *J Natl Cancer Inst* **90**, 713-714 (1998).

18 Ostrzenski, A. G-spot anatomy: a new discovery. *J Sex Med* **9**, 1355-1359, doi:10.1111/j.1743-6109.2012.02668.x (2012).

19 Pongtippan, A., Malpica, A., Levenback, C., Deavers, M. T. & Silva, E. G. Skene's gland adenocarcinoma resembling prostatic adenocarcinoma. *Int J Gynecol Pathol* **23**, 71-74, doi:10.1097/01.pgp.0000101144.79462.39 (2004).

20 Franks, L. M. Latent carcinoma of the prostate. *J Pathol Bacteriol* **68**, 603-616 (1954).

21 Rich, A. R. Classics in oncology. On the frequency of occurrence of occult carcinoma of the prostate: Arnold Rice Rich, M.D., Journal of Urology 33:3, 1935. *CA Cancer J Clin* **29**, 115-119 (1979).

22 Martin, R. M. Commentary: prostate cancer is omnipresent, but should we screen for it? *Int J Epidemiol* **36**, 278-281, doi:10.1093/ije/dym049 (2007).

23 Bouvard, V. *et al.* Carcinogenicity of consumption of red and processed meat. *The Lancet Oncology* **16**, 1599-1600, doi:10.1016/s1470-2045(15)00444-1 (2015).

24 Aune, D. *et al.* Dairy products, calcium, and prostate cancer risk: a systematic review and meta-analysis of cohort studies. *Am J Clin Nutr* **101**, 87-117, doi:10.3945/ajcn.113.067157 (2015).

25 Plant, T. M. & Zeleznik, A. J. *Knobil and Neill's Physiology of Reproduction*. (ELSEVIER academic Press, 2006).

26 The Nobel Foundation. *The Nobel Prize in Physiology or Medicine 1966*, <http://www.nobelprize.org/nobel_prizes/medicine/laureates/1966/index.html> (1966).

27 Chandrasekar, T., Yang, J. C., Gao, A. C. & Evans, C. P. Mechanisms of resistance in castration-resistant prostate cancer (CRPC). *Transl Androl Urol* **4**, 365-380, doi:10.3978/j.issn.2223-4683.2015.05.02 (2015).

28 Karantanos, T., Corn, P. G. & Thompson, T. C. Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene* **32**, 5501-5511, doi:10.1038/onc.2013.206 (2013).

29 Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *J Urol* **111**, 58-64 (1974).

30 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228, doi:10.1016/j.cell.2015.05.001 (2015).

31 Shah, R. B. *et al.* Androgen-Independent Prostate Cancer Is a Heterogeneous Group of Diseases. *Lessons from a Rapid Autopsy Program* **64**, 9209-9216, doi:10.1158/0008-5472.can-04-2442 (2004).

32 Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, doi:10.1038/nature11125 (2012).

33 Spratt, D. E., Zumsteg, Z. S., Feng, F. Y. & Tomlins, S. A. Translational and clinical implications of the genetic landscape of prostate cancer. *Nat Rev Clin Oncol*, doi:10.1038/nrclinonc.2016.76 (2016).

34 Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648, doi:10.1126/science.1117679 (2005).

35 Kumar-Sinha, C., Tomlins, S. A. & Chinnaiyan, A. M. Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* **8**, 497-511, doi:10.1038/nrc2402 (2008).

36 Shen, M. M. & Abate-Shen, C. Molecular genetics of prostate cancer: new prospects for old challenges. *Genes Dev* **24**, 1967-2000, doi:10.1101/gad.1965810 (2010).

37 Tomlins, S. A. *et al.* Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. *Nature* **448**, 595-599, doi:10.1038/nature06024 (2007).

38 Mani, R. S. *et al.* Induced chromosomal proximity and gene fusions in prostate cancer. *Science* **326**, 1230, doi:10.1126/science.1178124 (2009).

39 Helgeson, B. E. *et al.* Characterization of TMPRSS2:ETV5 and SLC45A3:ETV5 gene fusions in prostate cancer. *Cancer Res* **68**, 73-80, doi:10.1158/0008-5472.can-07-5352 (2008).

40 Tomlins, S. A. *et al.* TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. *Cancer Res* **66**, 3396-3400, doi:10.1158/0008-5472.can-06-0168 (2006).

41 Kim, J. H. *et al.* Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. *Genome Res* **21**, 1028-1041, doi:10.1101/gr.119347.110 (2011).

42 Varambally, S. *et al.* Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science* **322**, 1695-1699, doi:10.1126/science.1165395 (2008).

43 Varambally, S. *et al.* The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629, doi:10.1038/nature01075 (2002).

44 Tomlins, S. A. *et al.* The role of SPINK1 in ETS rearrangement-negative prostate cancers. *Cancer Cell* **13**, 519-528, doi:10.1016/j.ccr.2008.04.016 (2008).

45 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, doi:10.1038/nbt.1914 (2011).

46 Prensner, J. R. *et al.* The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* **45**, 1392-1398, doi:10.1038/ng.2771 (2013).

47 Prensner, J. R. *et al.* RNA biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of SChLAP1. *Lancet Oncol* **15**, 1469-1480, doi:10.1016/s1470-2045(14)71113-1 (2014).

48 Mehra, R. *et al.* Overexpression of the Long Non-coding RNA SChLAP1 Independently Predicts Lethal Prostate Cancer. *Eur Urol*, doi:10.1016/j.eururo.2015.12.003 (2015).

49 Draisma, G. *et al.* Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. *J Natl Cancer Inst* **101**, 374-383, doi:10.1093/jnci/djp001 (2009).

50 Hayes, J. H. & Barry, M. J. Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence. *Jama* **311**, 1143-1149, doi:10.1001/jama.2014.2085 (2014).

51 Crick, F. H. in *Symp. Soc. Exp. Biol* Vol. 12 138-163 (1958).

52 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

53 Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-3175 (1977).

54 Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8 (1977).

55 The Nobel Foundation. *The Nobel Prize in Physiology or Medicine 1993*, <http://www.nobelprize.org/nobel_prizes/medicine/laureates/1993/index.html> (1993).

56 Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29**, 2850-2859 (2001).

57 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

58 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

59 Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37-47, doi:10.1016/j.cell.2006.06.023 (2006).

60 Zhang, J., Sun, X., Qian, Y. & Maquat, L. E. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *Rna* **4**, 801-815 (1998).

61 Lejeune, F. & Maquat, L. E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol* **17**, 309-315, doi:10.1016/j.ceb.2005.03.002 (2005).

62 Okubo, J. *et al.* Aberrant activation of ALK kinase by a novel truncated form ALK protein in neuroblastoma. *Oncogene* **31**, 4667-4676, doi:10.1038/onc.2011.616 (2012).

63 Wiesner, T. *et al.* Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature* **526**, 453-457, doi:10.1038/nature15258 (2015).

64 Kong-Beltran, M. *et al.* Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res* **66**, 283-289, doi:10.1158/0008-5472.CAN-05-2749 (2006).

65 Dhanasekaran, S. M. *et al.* Transcriptome meta-analysis of lung cancer reveals recurrent aberrations in NRG1 and Hippo pathway genes. *Nat Commun* **5**, 5893, doi:10.1038/ncomms6893 (2014).

66 Hu, R. *et al.* Ligand-independent androgen receptor variants derived from splicing of cryptic exons signify hormone-refractory prostate cancer. *Cancer Res* **69**, 16-22, doi:10.1158/0008-5472.CAN-08-2764 (2009).

67 Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer* **16**, 413-430, doi:10.1038/nrc.2016.51 (2016).

68 Shapiro, I. M. *et al.* An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**, e1002218, doi:10.1371/journal.pgen.1002218 (2011).

69 Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).

70 Braeutigam, C. *et al.* The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. *Oncogene* **33**, 1082-1092, doi:10.1038/onc.2013.50 (2014).

71 Jangi, M., Boutz, P. L., Paul, P. & Sharp, P. A. Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev* **28**, 637-651, doi:10.1101/gad.235770.113 (2014).

72 Wang, E. T. *et al.* Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* **150**, 710-724, doi:10.1016/j.cell.2012.06.041 (2012).

73 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415, doi:10.1038/ng.259 (2008).

74 Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-476, doi:10.1038/nature07509 (2008).

75 Vauchy, C. *et al.* CD20 alternative splicing isoform generates immunogenic CD4 helper T epitopes. *International journal of cancer* **137**, 116-126, doi:10.1002/ijc.29366 (2015).

76 Xu, Q., Modrek, B. & Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30**, 3754-3766 (2002).

77 Dewaele, M. *et al.* Antisense oligonucleotide-mediated MDM4 exon 6 skipping impairs tumor growth. *J Clin Invest* **126**, 68-84, doi:10.1172/jci82534 (2016).

78 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).

79 Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).

80 Chien, A., Edgar, D. B. & Trela, J. M. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J Bacteriol* **127**, 1550-1557 (1976).

81 Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487-491 (1988).

82 Temin, H. M. & Mizutani, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* **226**, 1211-1213 (1970).

83 Baltimore, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* **226**, 1209-1211 (1970).

84 Bennett, S. Solexa Ltd. *Pharmacogenomics* **5**, 433-438, doi:10.1517/14622416.5.4.433 (2004).

85 Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349, doi:10.1126/science.1158441 (2008).

86 Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509-1517, doi:10.1101/gr.079558.108 (2008).

87 Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25**, 1026-1032, doi:10.1093/bioinformatics/btp113 (2009).

88 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, doi:10.1093/nar/gkq211 (2010).

89 Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97-101, doi:10.1038/nature07638 (2009).

90 Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc Natl Acad Sci U S A* **106**, 12353-12358, doi:10.1073/pnas.0904720106 (2009).

91 Balbin, O. A. *et al.* The landscape of antisense gene expression in human cancers. *Genome Res* **25**, 1068-1079, doi:10.1101/gr.180596.114 (2015).

92 Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199-208, doi:10.1038/ng.3192 (2015).

93 Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-453 (1970).

94 Sankoff, D. Matching sequences under deletion-insertion constraints. *Proc Natl Acad Sci U S A* **69**, 4-6 (1972).

95 Smith, T. F., Waterman, M. S. & Fitch, W. M. Comparative biosequence metrics. *J Mol Evol* **18**, 38-46 (1981).

96 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

97 Wang, H., Ooi, B. C., Tan, K. L., Ong, T. H. & Zhou, L. BLAST++: BLASTing queries in batches. *Bioinformatics* **19**, 2323-2324, doi:10.1093/bioinformatics/btg310 (2003).

98 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

99 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, doi:10.1186/gb-2009-10-3-r25 (2009).

100 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

101 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178, doi:10.1093/nar/gkq622 (2010).

102 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

103 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

104 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/s0022-2836(05)80360-2 (1990).

105 Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147**, 195-197 (1981).

106 Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395-2396, doi:10.1093/bioinformatics/btn429 (2008).

107 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

108 Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**, D756-763, doi:10.1093/nar/gkt1114 (2014).

109 The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212, doi:10.1093/nar/gku989 (2015).

110 Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**, 830-839, doi:10.1101/gr.115907.110 (2011).

111 Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65, doi:10.1038/nature09708 (2011).

112 Zhao, X., Emery, S. B., Myers, B., Kidd, J. M. & Mills, R. E. Resolving complex structural genomic rearrangements using a randomized approach. *Genome Biol* **17**, 126, doi:10.1186/s13059-016-0993-1 (2016).

113 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

114 Li, Y., Rao, X., Mattox, W. W., Amos, C. I. & Liu, B. RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. *PLoS One* **10**, e0136653, doi:10.1371/journal.pone.0136653 (2015).

115 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

116 Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765-1786, doi:10.1038/nprot.2013.099 (2013).

117 Risueno, A. *et al.* A robust estimation of exon expression to identify alternative spliced genes applied to human tissues and cancer samples. *BMC Genomics* **15**, 879, doi:10.1186/1471-2164-15-879 (2014).

118 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, doi:10.1038/nbt.1621 (2010).

119 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

120 Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*, doi:10.1038/nbt.2862 (2014).

121 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177-1184, doi:10.1038/nmeth.2714 (2013).

122 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, doi:10.1038/nmeth.1528 (2010).

123 Shen, S. *et al.* MATS: a bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res* **40**, e61, doi:10.1093/nar/gkr1291 (2012).

124 Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform* **8**, 6-21, doi:10.1093/bib/bbl015 (2007).

125 Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).

126 Geiss, G. K. *et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* **26**, 317-325, doi:10.1038/nbt1385 (2008).

127 Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009-1014, doi:10.1038/nbt.2705 (2013).

128 Mangul, S. *et al.* Transcriptome assembly and quantification from Ion Torrent RNA-Seq data. *BMC Genomics* **15 Suppl 5**, S7, doi:10.1186/1471-2164-15-s5-s7 (2014).

129 Yates, J. R., 3rd, Gilchrist, A., Howell, K. E. & Bergeron, J. J. Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol* **6**, 702-714, doi:10.1038/nrm1711 (2005).

130 Walmsley, S. J. *et al.* Comprehensive analysis of protein digestion using six trypsins reveals the origin of trypsin as a significant source of variability in proteomics. *J Proteome Res* **12**, 5666-5680, doi:10.1021/pr400611h (2013).

131 Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* **4**, 787-797, doi:10.1038/nmeth1088 (2007).

132 Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* **10**, 730-736, doi:10.1038/nmeth.2557 (2013).

133 Woo, S. *et al.* Proteogenomic database construction driven from large scale RNA-seq data. *J Proteome Res* **13**, 21-28, doi:10.1021/pr400294c (2014).

134 Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114-1125, doi:10.1038/nmeth.3144 (2014).

135 Shanmugam, A. K., Yocum, A. K. & Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *J Proteome Res* **13**, 4113-4119, doi:10.1021/pr500496p (2014).

136 Ning, K., Fermin, D. & Nesvizhskii, A. I. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J Proteome Res* **11**, 2261-2271, doi:10.1021/pr201052x (2012).

137 Li, H. D., Menon, R., Omenn, G. S. & Guan, Y. Revisiting the identification of canonical splice isoforms through integration of functional genomics and proteomics evidence. *Proteomics* **14**, 2709-2718, doi:10.1002/pmic.201400170 (2014).

138 Omenn, G. S., Menon, R. & Zhang, Y. Innovations in proteomic profiling of cancers: alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology. *J Proteomics* **90**, 28-37, doi:10.1016/j.jprot.2013.04.007 (2013).

139 Li, H. D., Omenn, G. S. & Guan, Y. A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling. *Brief Bioinform*, doi:10.1093/bib/bbv109 (2016).

140 Omenn, G. S., Guan, Y. & Menon, R. A new class of protein cancer biomarker candidates: differentially expressed splice variants of ERBB2 (HER2/neu) and ERBB1 (EGFR) in breast cancer cell lines. *J Proteomics* **107**, 103-112, doi:10.1016/j.jprot.2014.04.012 (2014).

141 Menon, R. *et al.* Distinct splice variants and pathway enrichment in the cell-line models of aggressive human breast cancer subtypes. *J Proteome Res* **13**, 212-227, doi:10.1021/pr400773v (2014).

142 Menon, R. *et al.* Computational Inferences of the Functions of Alternative/Noncanonical Splice Isoforms Specific to HER2+/ER-/PR- Breast Cancers, a Chromosome 17 C-HPP Study. *J Proteome Res* **14**, 3519-3529, doi:10.1021/acs.jproteome.5b00498 (2015).

143 Eksi, R. *et al.* Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput Biol* **9**, e1003314, doi:10.1371/journal.pcbi.1003314 (2013).

144 Menon, R. *et al.* Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. *J Proteome Res* **10**, 5503-5511, doi:10.1021/pr200772w (2011).

145 Veeneman, B. A., Iyer, M. K. & Chinnaiyan, A. M. Oculus: faster sequence alignment by streaming read compression. *BMC Bioinformatics* **13**, 297, doi:10.1186/1471-2105-13-297 (2012).

146 Yorukoglu, D., Yu, Y. W., Peng, J. & Berger, B. Compressive mapping for next-generation sequencing. *Nat Biotechnol* **34**, 374-376, doi:10.1038/nbt.3511 (2016).

147 Nellore, A. *et al.* Rail-RNA: Scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*, doi:10.1093/bioinformatics/btw575 (2016).

148 Wetterstrand, K. A. *DNA sequencing costs: data from the NHGRI large-scale genome sequencing program*, <http://www.genome.gov/sequencingcosts> (2012).

149 Pennisi, E. Human genome 10th anniversary. Will computers crash genomics? *Science* **331**, 666-668, doi:10.1126/science.331.6018.666 (2011).

150 Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858, doi:10.1101/gr.078212.108 (2008).

151 Weese, D., Emde, A. K., Rausch, T., Doring, A. & Reinert, K. RazerS--fast read mapping with sensitivity control. *Genome Res* **19**, 1646-1654, doi:10.1101/gr.088823.108 (2009).

152 Schatz, M. C. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25**, 1363-1369, doi:10.1093/bioinformatics/btp236 (2009).

153 Nguyen, T., Shi, W. & Ruden, D. CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. *BMC Res Notes* **4**, 171, doi:10.1186/1756-0500-4-171 (2011).

154 Pireddu, L., Leo, S. & Zanetti, G. SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics* **27**, 2159-2160, doi:10.1093/bioinformatics/btr325 (2011).

155 Shimizu, K. & Tsuda, K. SlideSort: all pairs similarity search for short reads. *Bioinformatics* **27**, 464-470, doi:10.1093/bioinformatics/btq677 (2011).

156 Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-577, doi:10.1038/nmeth0810-576 (2010).

157 Burriesci, M. S., Lehnert, E. M. & Pringle, J. R. Fulcrum: condensing redundant reads from high-throughput sequencing studies. *Bioinformatics*, doi:10.1093/bioinformatics/bts123 (2012).

158 The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640, doi:10.1126/science.1105136 (2004).

159 Sun, Z. *et al.* Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One* **6**, e17490, doi:10.1371/journal.pone.0017490 (2011).

160 Labaj, P. P. *et al.* Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27**, i383-391, doi:10.1093/bioinformatics/btr247 (2011).

161 Silverstein, C. *sparsehash: An extremely memory-efficient hash_map implementation*, <http://code.google.com/p/sparsehash/> (2005).

162 Appleby, A. *MurmurHash*, <http://sites.google.com/site/murmurhash> (2008).

163 Kent Informatics Inc. *BLAT and other fine software*, <http://www.kentinformatics.com> (2016).

164 Veeneman, B. A., Shukla, S., Dhanasekaran, S. M., Chinnaiyan, A. M. & Nesvizhskii, A. I. Two-pass alignment improves novel splice junction quantification. *Bioinformatics* **32**, 43-49, doi:10.1093/bioinformatics/btv642 (2016).

165 Engstrom, P. G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* **10**, 1185-1191, doi:10.1038/nmeth.2722 (2013).

166 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

167 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

168 Seo, J. S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res* **22**, 2109-2119, doi:10.1101/gr.145144.112 (2012).

169 SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32**, 903-914, doi:10.1038/nbt.2957 (2014).

170 The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).

171 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).

172 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods*, doi:10.1038/nmeth.2639 (2013).

173 Gatto, A. *et al.* FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. *Nucleic Acids Res* **42**, e71, doi:10.1093/nar/gku166 (2014).

174 Eeles, R. A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41**, 1116-1121, doi:10.1038/ng.450 (2009).

175 Brenner, J. C. *et al.* PARP-1 inhibition as a targeted strategy to treat Ewing's sarcoma. *Cancer Res* **72**, 1608-1613, doi:10.1158/0008-5472.can-11-3648 (2012).

176 Drost, J. *et al.* Organoid culture systems for prostate epithelial and cancer tissue. *Nat Protoc* **11**, 347-358, doi:10.1038/nprot.2016.006 (2016).

177 Abate-Shen, C. & Pandolfi, P. P. Effective utilization and appropriate selection of genetically engineered mouse models for translational integration of mouse and human trials. *Cold Spring Harb Protoc* **2013**, doi:10.1101/pdb.top078774 (2013).

178 Mani, R. S. *et al.* TMPRSS2-ERG-mediated feed-forward regulation of wild-type ERG in human prostate cancers. *Cancer Res* **71**, 5387-5392, doi:10.1158/0008-5472.CAN-11-0876 (2011).

179 The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011-1025, doi:10.1016/j.cell.2015.10.025 (2015).

180 Tomlins, S. A. *et al.* Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* **39**, 41-51, doi:10.1038/ng1935 (2007).

181 Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**, 1242-1248, doi:10.1038/ng.3414 (2015).

182 Hsu, T. Y. *et al.* The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature* **525**, 384-388, doi:10.1038/nature14985 (2015).

183 Danan-Gotthold, M. *et al.* Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Res*, doi:10.1093/nar/gkv210 (2015).

184 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, doi:10.1073/pnas.0506580102 (2005).

185 Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol*, doi:10.1038/nbt.3080 (2014).

186 Onozato, R. *et al.* Activation of MET by gene amplification or by splice mutations deleting the juxtamembrane domain in primary resected lung cancers. *J Thorac Oncol* **4**, 5-11, doi:10.1097/JTO.0b013e3181913e0e (2009).

187 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

188 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

189 Friedlander, T. W. *et al.* Common structural and epigenetic changes in the genome of castration-resistant prostate cancer. *Cancer Res* **72**, 616-625, doi:10.1158/0008-5472.can-11-2079 (2012).

190 Tomlins, S. A. *et al.* Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. *Eur Urol* **70**, 45-53, doi:10.1016/j.eururo.2015.04.039 (2016).

191 Laxman, B. *et al.* A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. *Cancer Res* **68**, 645-649, doi:10.1158/0008-5472.can-07-3224 (2008).

192 Saini, S. PSA and beyond: alternative prostate cancer biomarkers. *Cell Oncol (Dordr)* **39**, 97-106, doi:10.1007/s13402-016-0268-6 (2016).

193 Thorsen, K. *et al.* Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol Cell Proteomics* **7**, 1214-1224, doi:10.1074/mcp.M700590-MCP200 (2008).

194 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, doi:10.1038/cr.2012.30 (2012).

195 Sveen, A., Johannessen, B., Teixeira, M. R., Lothe, R. A. & Skotheim, R. I. Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. *BMC Genomics* **15**, 672, doi:10.1186/1471-2164-15-672 (2014).

196 Sowalsky, A. G. *et al.* Whole transcriptome sequencing reveals extensive unspliced mRNA in metastatic castration-resistant prostate cancer. *Mol Cancer Res* **13**, 98-106, doi:10.1158/1541-7786.MCR-14-0273 (2015).

197 Rezaeian, I., Tavakoli, A., Cavallo-Medved, D., Porter, L. A. & Rueda, L. A Novel Model Used to Detect Differential Splice Junctions as Biomarkers in Prostate Cancer from RNA-Seq Data. *J Biomed Inform*, doi:10.1016/j.jbi.2016.03.010 (2016).

198 Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* **7 Suppl 1**, S12 11-14, doi:10.1186/gb-2006-7-s1-s12 (2006).

199 Antonarakis, E. S. *et al.* AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. *N Engl J Med* **371**, 1028-1038, doi:10.1056/NEJMoa1315815 (2014).

200 Cochrane, G. *et al.* Facing growth in the European Nucleotide Archive. *Nucleic Acids Res* **41**, D30-35, doi:10.1093/nar/gks1175 (2013).

201 Treutlein, B., Gokce, O., Quake, S. R. & Sudhof, T. C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. *Proc Natl Acad Sci U S A* **111**, E1291-1299, doi:10.1073/pnas.1403244111 (2014).

202 Cieslik, M. *et al.* The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res* **25**, 1372-1381, doi:10.1101/gr.189621.115 (2015).

203 Rubin, M. A. *et al.* Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res* **6**, 1038-1045 (2000).

204 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).

205 Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635-1646, doi:10.1016/j.cell.2012.05.003 (2012).

206 Saletore, Y. *et al.* The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol* **13**, 175, doi:10.1186/gb-2012-13-10-175 (2012).

207 Hieronymus, H. *et al.* Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* **10**, 321-330, doi:10.1016/j.ccr.2006.09.005 (2006).

208 Beltran, H. *et al.* Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* **22**, 298-305, doi:10.1038/nm.4045 (2016).