Pan-Cancer Identification and Prioritization of Cancer-Associated Alternatively Spliced and Differentially Expressed Genes: A Biomarker Discovery Application

by

Daryanaz Dargahi

B.Sc., University of Tehran, 2009
M.Sc., Simon Fraser University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY

Fall 2016

Approval

Name: **Daryanaz Dargahi**

Degree: **Doctor of Philosophy**

Title: ***Pan-Cancer Identification and Prioritization of Cancer-Associated Alternatively Spliced and Differentially Expressed Genes: A Biomarker Discovery Application***

Examining Committee: **Chair:** Sharon Gorski
Professor

Steven J.M. Jones
Senior Supervisor
Professor

David L. Baillie
Supervisor
Professor

Robert A. Holt
Supervisor
Professor

Angela Brooks-Wilson
Supervisor
Professor

Martin Hirst
Supervisor
Associate Professor

Fiona Brinkman
Internal Examiner
Professor

Denise Clark
External Examiner
Professor
Biology
University of New Brunswick

Date Defended/Approved: November 21, 2016

Abstract

Tumour cells arise through aberrant expression of genes and the proteins they encode. This may result from a direct change to DNA sequence or perturbations in the machinery responsible for production or activity of proteins, such as gene splicing. With the advent of massively parallel RNA-sequencing (RNA-seq), large-scale exploration of changes at the stage of transcription and posttranscriptional splicing has the potential to unravel the landscape of gene expression changes across human cancers. Aberrantly expressed genes in cancer can serve as molecular biomarkers for discrimination of tumour and normal cells if localized to the cell surface and therefore can be used as targets for targeted antibody-based cancer therapy. In the current study, I devised an analysis pipeline to identify and rank such events from human cancer RNA-seq datasets. Using my pipeline, I conducted a pan-cancer analysis in the RNA-sequencing data of more than 7,000 patients from 24 different cancer types generated by the cancer genome atlas (TCGA). I identified abnormally expressed and alternatively spliced genes, which seemed to be cancer-associated in comparison to a large compendium of transcriptomes from non-diseased tissues gathered from Genotype-Tissue Expression (GTEx) and TCGA. My analysis revealed 1,503 putative tumor-associated abnormally expressed genes and 1,142 novel cancer-associated splice variants occurring in 694 genes. In order to rank identified candidate genes, I performed an extensive literature search and studied known therapeutic antibody targets to collect the characteristics of an ideal antibody target in cancer. I developed an R package, Prize, based on the Analytic Hierarchy Process (AHP) algorithm. AHP is a multiple-criteria decision making solution that allows a user to prioritize a list of elements based of a set of user-define criteria and numerical score that express the importance of each criterion to achieving the goal. I built an AHP model to depict cancer biomarker target properties for ranking and prioritizing the genes. Using this model, Prize was able to successfully recognize and rank known tumour biomarker targets among the top 25 ranked list along with other novel candidates.

Keywords: RNA-sequencing; Alternative splicing; Gene expression; Biomarker target; Prioritization; Analytic Hierarchy Process

Preface

Portions of section 3.1 and 3.3 is in preparation for submission as Daryanaz Dargahi, Christopher Bond, Ryan Dercho, Richard Swayze, Leanna Yee, Peter Bergqvist, Alireza Heravi-Moussavi, Bradley Hedberg, Jianghong An, Edie Dullaghan, Ismael Samudio, John Babcook, and Steven Jones. (2016). Pan-cancer Identification and Prioritization of Cancer-Associated Abnormally Expressed Genes: A Biomarker Discovery Application. I am the lead researcher and author of this publication. I performed data analysis, generated figures, performed literature search, designed and implemented the R package, and am writing the manuscript. Myself, SJ and JB conceived and designed the study. Myself, SJ, JB, CB, IS, RS, LY, PB, BH, ED, RD, JA, AHM designed the problem hierarchy, chose decision criteria and rating categories for prioritization, and generated consensus pairwise comparison matrices via multiple discussions and literature search. JB, CB, and IS are leading experts in antibody-drug conjugate development.

The Prize R package described in section 3.3.1 is currently available to public on Bioconductor at https://www.bioconductor.org/packages/release/bioc/html/Prize.html. The package has been downloaded more than 1,300 time since the date of publication (October 2015).

Portions of section 3.2 has been published as Daryanaz Dargahi, Richard Swayze, Leanna Yee, Peter Bergqvist, Bradley Hedberg, Alireza Heravi-Moussavi, Edie Dullaghan, Ryan Dercho, Jianghong An, John Babcook, and Steven Jones. (2014). A Pan-Cancer Analysis of Alternative Splicing Events Reveals Novel Tumor-Associated Splice Variants of Matriptase. Cancer Informatics. 2014 Dec; 13: 167–177. doi: 10.4137/CIN.S19435. I was the lead researcher and author of this publication. I performed data analysis, generated figures, performed literature search, wrote the manuscript, and was involved in designing validation experiments. Myself, SJ and JB conceived and designed the study. RS, LY, PB, BH, ED, and RD designed and ran validation experiments. JA and AHM provided technical assistance. All the authors made critical revisions and approved the final version of the manuscript.

In addition, novel matriptase splice variants described in section 3.2.3 have been filled as a PCT international patent application No. PCT/CA2014/000875 entitled: MATRIPTASE VARIANTS ASSOCIATED WITH TUMORS, filed December 9, 2014. Inventors: Dargahi, D., Babcook, JS. and Jones SJM. Applicant: British Columbia Cancer Agency Branch and The Centre for Drug Research and Development.

Dedication

To my parents whose unconditional love and support has made this possible for me to be here today...

Acknowledgements

I would like to greatly thank my senior supervisor, Dr. Steven J.M. Jones for giving me the opportunity to pursue my PhD and for providing me with exceptional mentorship, consistent support, and endless scientific expertise over the past 5 years.

I would also like to thank my committee members, Dr. David L. Baillie, Dr. Robert Holt, Dr. Angela Brooks-Wilson, and Dr. Martin Hirst for their support over the past 5 years as well as their advice and guidance not only scientifically, but also in relation to my professional and personal development. In addition, I would like to acknowledge Dr. Fiona Brinkman and Dr. Denise Clark for being my internal and external examiners, respectively.

This work would not have been made possible without the financial support of several funding agencies. I am deeply grateful for a PhD fellowship from the Mitacs Accelerate program, and grants from Genome British Columbia strategic opportunities fund, and the Terry Fox Research Institue (TFRI) new frontiers program. I would like to also thank Dr. John Babcook, The Centre for Drug Research and Development (CDRD), and CDRD ventures Inc. for the three years internship opportunity through Mitacs Accelerate program. In addition, I would like to thank Simon Fraser University and the Molecular Biology and Biochemistry Department.

The results published in this thesis are in whole or part based upon data generated by Genotype-Tissue Expression (GTEx) and The Cancer Genome Atlas (TCGA) pilot projects established by national cancer institute (NCI) and national human genome research institute (NHGRI). I would like to thank GTEx and TCGA groups for making these data publically available. Information about TCGA can be found at http://cancergenome.nih.gov. Additional information about GTEx project is also available at http://www.gtexportal.org/.

Finally, I am extremely grateful for the support of my friends and family. Specifically my mother Nahid Mojaverian, and my father Mohammad Ali Dargahi, who have always encouraged me to embrace and develop my individuality and have

unconditionally supported the pursuit of my passions. Their endless love, support, and respect have made me the person I am today.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ADC	Antibody Drug Conjugate
AHP	Analytic Hierarchy Process
AIJ	Aggregated Individual Judgement
AIP	Aggregated Individual Priority
AS	Alternative Splicing
ASTD	Alternative Splicing and Transcript Discovery Database
BAM	Binary Alignment/Map
CA9	Carbohydrase 9
CADE	Cancer Associated Differentially Expressed
CAM	Category Assignment Matrix
cDNA	Complementary DNA
CDRD	Center for Drug Research and Development
CHO	Chinese Hamster Ovary
CI	Consistency Index
COX-2	Cyclooxygenase-2
CR	Consistency Ratio
CUB	Complement C1r/C1s, Uegf, Bmp1
DAC	Data Access Committee
dbGAP	Database of Genotypes and Phenotypes
DCC	Data Coordinating Center
DM	Decision Making
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EGFR	Epidermal Growth Factor Receptor
EGFRvIII	Epidermal Growth Factor Receptor variant III
EM	Expectation-Maximization
EMT	Epithelial-to-Mesenchymal Transition
ERK	Extracellular-signal Regulated Kinase
ESE	Exonic Splicing Enhancers
ESS	Exonic Splicing Silencers
FDA	Food and Drug Administration

FDA	Food and Drug Administration
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase Million
GABA	γ-amino butyric acid
GC	Guanine-Cytosine
GEA	Gene Expression Analysis
GO	Gene Ontology
GSC	Genome Sciences Centre
GTEx	Genotype-Tissue Expression
HAI-1	Hepatocyte growth factor Activator Inhibitor-1
HGF	Hepatocyte Growth Factor
hnRNP	Heterogeneous Nuclear Ribonucleoproteins
ICR	Individual Consistency Ratio
IGV	Integrated Genome Viewer
ISE	Intronic Splicing Enhancers
ISS	Intronic Splicing Silencers
LDLRA	Low-Density-Lipoprotein Receptor class A
mAb	Monoclonal Antibody
MAPK	Mitogen Activated Protein Kinase
MDS	Multidimensional Scaling
MITF	Microphtalmia-associated Transcription Factor
ML	Maximum Likelihood
mRNA	Messenger Ribonucleic Acid
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
NIH	National Institute of Heath
NPM	Nucleotides Per Million
ORF	Open reading frame
PCM	Pairwise Comparison Matrix
PCR	Polymerase Chain Reaction
PKA	Protein Kinase A
PSA	Prostate Specific Antigen
RI	Random Index
RNA	Ribonucleic Acid

RNA-seq	RNA Sequencing
RPKM	Reads Per Kilobase Million
rRNA	Ribosomal RNA
RSEM	RNA-Seq by Expectation Maximization
SAGE	Serial Analysis of Gene Expression
SAM	Sequence Alignment/Map
SEA	Sea urchin stem region, Enteropeptidase, and Argin
SF	Splicing Factor
snoRNA	Small Nucleolar RNA
snRNP	Small Nuclear Ribonucleoprotein
SPINT1	Serine Peptidase Inhibitor encoded by Kunitz type 1
SR	Serine-Rich
SRE	Splicing Regulatory Element
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of M-values
TNF-α	Tumour Necrosis Factor-α
TPM	Transcripts Per Million
uPA	Urokinase Plasminogen Activator
UQUA	Upper Quartile

Chapter 1. Introduction

1.1. Gene expression and splicing

Gene expression is a fundamental process in the cell during which the deoxyribonucleic acid (DNA) is transcribed to the corresponding ribonucleic acid (RNA) and the RNA is translated to the corresponding protein. Gene expression can change from one cell to another, between tissues and at different points in time (Alberts et al., 2007). Measuring gene expression by the quantification of the transcript levels is an invaluable tool in biomedical sciences to study a disease diagnosis, prognosis and search for drug targets (Schulze & Downward, 2001). For instance, the study of gene expression in cancer, alzheimer's disease, schizophrenia and HIV infection have revealed much about the biology and potential treatment of these diseases (Minagar et al., 2004). Therefore, measuring gene expression is of high scientific interest, and many methods have been developed for measuring gene expression.

The splicing of messenger RNA (mRNA) transcripts is a highly regulated process during gene expression that can results in a single gene coding for multiple distinct protein sequences (Roy, Haupt et al., 2013). The human genome contains approximately 22,000 protein-coding gene loci (Pruitt, Tatusova et al., 2012). However, the number of unique protein isoforms is greater than can be explained by the number of genes alone. In order to understand this disparity, we must study the pathway that leads to the formation of proteins. In this process a region of DNA that encodes at least one gene is transcribed into an RNA molecule. If the transcribed gene encodes a protein, the resultant mRNA will serve as a template for the protein's synthesis through translation. In order for an RNA molecule to become mRNA and translate into a protein peptide, it must undergo a series of modifications (Roy et al., 2013). In eukaryotes, splicing is a pre-

mRNA processing mechanism that commonly occurs and this process serves to remove non-protein coding introns, joining the resultant exons to form a complete in-fame coding transcript.

Alternative splicing (AS) is the process by which a single primary transcript yields different mature RNAs leading to the production of protein isoforms with possibly diverse and even antagonistic functions. Studies of human genome have estimated that 94% of genes produce alternatively spliced transcripts (Wang, Sandberg et al., 2008). There are several different types of AS (Figure 1-1). In rare cases, a whole intron can be retained during the splicing process. Alternative 5′ splice sites or 3′ splice sites can result in exons of different sizes. Exclusion or skipping of one or more exons is a common form of AS. Similar to other cellular processes that are modified during cellular growth, differentiation and tissue development, AS is also affected. Recently, several mRNA isoforms specific to stages of cellular development and disease, including cancer, have been described (Oltean & Bates, 2014). With the recognition of the importance of splicing defects in human disease has come a realization that constitutive splicing events are potential therapeutic targets. Many different approaches such as conventional small-molecule drugs and antibody-based therapeutics have been proposed to target alternative splice variants.

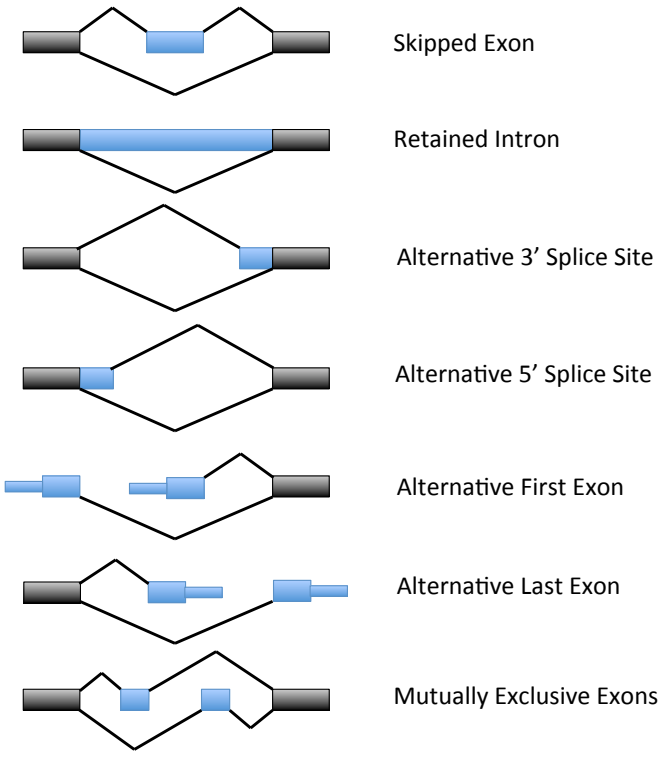


Figure 1-1. Alternative splicing (AS) event types. Constitutive exonic regions are solid black. Regions that may be differentially included are blue. Thin black lines represent introns.

1.1.1. The regulation of gene expression

Gene expression is believed to be one of the most tightly controlled processes in the body (Alberts et al., 2013). This process needs to be strictly regulated to ensure that the required amounts of RNA/proteins are being generated within the right cells at the right time. Disruption of gene expression regulation may lead to disease, including cancer (Hanahan & Weinberg, 2011)

Gene expression is regulated according to the needs of the cells. Regulation of gene expression encompasses a wide range of mechanisms that are used by cells to increase or decrease the production of a specific gene product including RNA and protein (Alberts et al., 2013). Also, cells can produce or block specific gene products in response to external signals or cellular damage (Alberts et al., 2013). Although the different cell types within a multicellular organism contain the same genome, different cell types can respond differently to the same signal. This can be explained in great part by the difference in the gene expression profile, which helps establish cell types. Cells have the ability to change which genes they express and how much without altering the nucleotide sequence of their DNA (Alberts et al., 2013). Therefore, gene expression regulation determines the cell's overall structure and function. It also governs cell differentiation, cell morphology and adaptability to the environment.

Gene expression regulation can occur at many stages in the pathway from DNA to RNA to protein. A cell can control the amount of produced proteins by (Figure 1-2);

1) Regulating the amount of transcription,

2) Regulating the processing of RNA molecules, including AS to produce more than one protein product from a single gene,

3) Selecting which mRNAs are exported from the nucleus to the cytosol,

4) Selectively degrading certain mRNA molecules,

5) Regulating the rate of translation.

Although every step mentioned above can participates in regulating gene expression, the control of transcription is paramount for most genes (Alberts et al., 2013). The reason is that only transcriptional control can ensure no unnecessary intermediates are synthesized. Transcriptional regulation is capable of turning the process of transcription on or off for individual genes in cells. Many different transcriptional regulators such as transcription factors, epigenomic features and promoters typically control the expression of eukaryotic genes (Alberts et al., 2013). For example, in order for transcription to take place, the enzyme that synthesizes RNA, known as RNA polymerase, must attach to the DNA near a gene. Promoters contain specific DNA sequences that provide a secure initial binding site for RNA polymerase and for transcription factors that recruit RNA polymerase. These transcription factors have specific activator or repressor sequences of corresponding nucleotides that attach to specific promoters and regulate gene expression (Alberts et al., 2013). Although we have good tools to quantitate changes in transcript expression, we lack the molecular biology tools to easily determine the precise reason for a change in gene expression. A major reason is simply the vast complexity of the regulatory network inside and outside of the cell (Alberts et al., 2013).

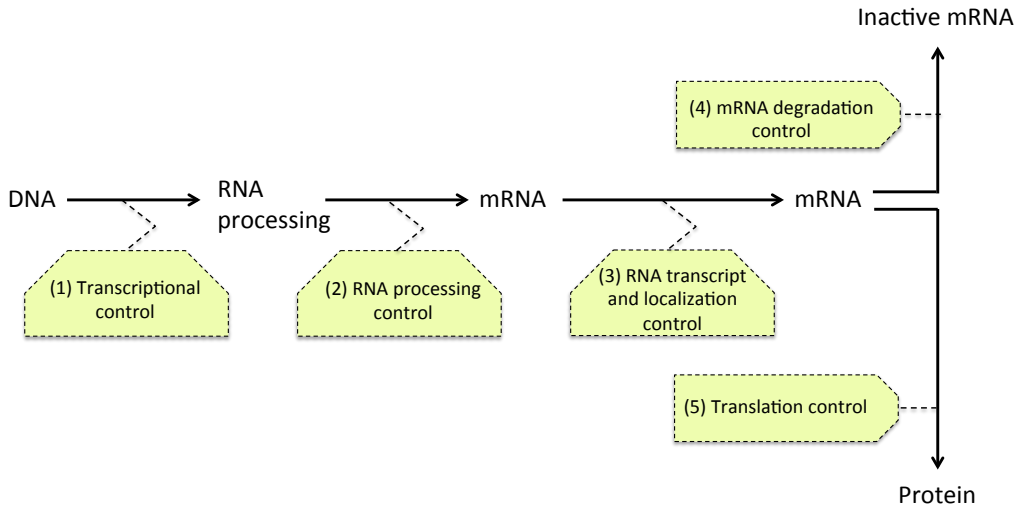


Figure 1-2. Gene expression can be controlled at several different steps. Examples of regulation at each of the steps are known, although for most genes the main site of control is step 1: transcription of a DNA sequence into RNA.

1.1.2. The regulation of splicing

Splicing is an editing of the nascent pre-mRNA transcript through which intronic sequence is systematically excised and flanking exons are ligated. It is one of several transcriptional processing steps. For splicing to take place, the involvement of many distinct proteins and ribonucleoprotein particles is required (Chen & Manley, 2009). During the splicing process, a subset of splicing factors (SFs) assemble onto the mRNA precursor around exon junctions to form a spliceosome complex. The spliceosome then cleaves the RNA molecule, removes the non-coding intron segment, and ligates the remaining exons together. Recognition and precise definition of exon boundaries involves several cis- and trans-acting elements that can either promote or inhibit splicing at a candidate exon junction (Chen & Manley, 2009).

The spliceosome is a dynamic, macromolecular complex that is systematically assembled at splice sites to catalyse the splicing reaction. It is composed of five small nuclear ribonucleoprotein particles (snRNPs: U1, U2, U3, U4, U5, and U6), in conjunction with many auxiliary proteins (Will & Luhrmann, 2011). The snRNPs form the core of the spliceosome. They are directly involved in the recognition of splice sites and branch-point sequences, as well as the catalysis of the splicing reaction. Assembly and activity of the spliceosome complex occurs during transcription of the pre-mRNA. The assembly of spliceosome complex occurs in a step-wise fashion, forming several intermediate complexes before forming the final complex (Matlin, Clark et al., 2005). The first pre-spliceosomal complex is called the E complex. It forms when the U1 snRNP binds to the 5’ splice site of an intron, followed by binding the splicing factor 1 (SF1) to the intron branch point, and the U2 auxiliary factors, U2AF1 and U2AF1, to the 3′ splice site and the polypyrimidine tract, respectively. The E complex can be converted to the A complex (pre-spliceosome complex) if the U2 snRNP displaces SF1 and binds to the intron branch point sequence. Recruitment of the U5/U4/U6 tri-snRNP to the A complex generates the B complex (pre-catalytic spliceosome complex) with the binding of U5 snRNP to exons at the 5' site and U6 to U2. Extensive rearrangements are required to produce the C complex (catalytic spliceosome complex). The C complex catalyzes the next step in the splicing process before disassociating.

The splice site choice is regulated through cis-acting splicing regulatory elements (SREs, enhancers and silencers) and trans-acting SFs (repressors or activators) (Matlin et al., 2005). On the basis of their locations and activities, SREs are categorized into four groups; exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs) and intronic splicing silencers (ISSs). These SREs specifically recruit SFs to assist in the placement of the spliceosome on the appropriate splice sites, and to consequently promote or reduce the usage of a particular splice site. Common splicing factors include Serine-Rich (SR) proteins, which recognize ESEs to promote splicing, as well as various heterogeneous nuclear ribonucleoproteins (hnRNPs), which typically recognize ESSs to inhibit splicing. Both SR proteins and hnRNPs often affect the function of U2 and U1 snRNPs during spliceosomal assembly (Figure 1-3).

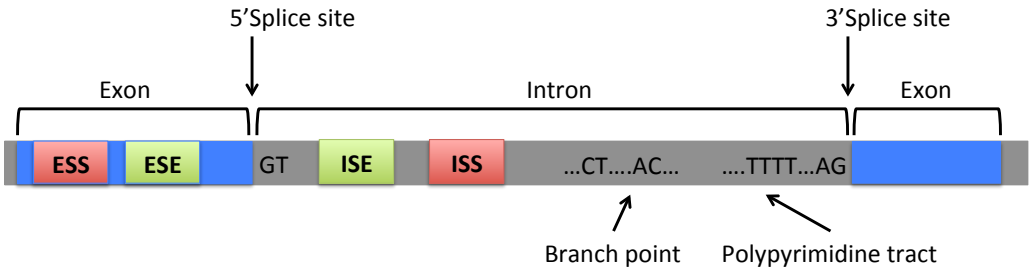


Figure 1-3. **The regulation of splicing.** The cis-acting sequences involved in the regulation of intron removal are shown. In addition to the core splicing signals (i.e. 5' splice site, branch-point and 3' splice site), several regulatory sequences influence the splicing decision by recruiting trans-acting SFs. Common SFs include SR proteins and hnRNPs, which typically promote and inhibit splicing, respectively. ESE: Exonic Splicing Enhancers. ISE: Intronic Splicing Enhancers. ESS: Exonic Splicing Silencers. ISS: Intronic Splicing Silencers.

It has been shown that the relative concentrations and activities of SFs can affect the ability of the spliceosome to determine the precise location of a splice site and therefore assemble on exon junctions (Chen & Manley, 2009). Hence, altering SFs' expression, localization, or functional efficacy can modulate splicing. For example, disrupting the phosphorylation of SR proteins could negatively impact splicing regulatory programmes. The secondary structure of the pre-mRNA transcript, chromatin structure and nucleosome positioning also play a role in regulating splicing by influencing the accessibility of splice sites or cis-acting SREs (Brown, Stoilov, & Xing, 2012). Moreover, splicing is also affected by factors that control transcription initiation and elongation. This is because the splicing of most introns happens before transcription termination, a phenomenon known as co-transcriptional splicing. For example, the rate of transcription elongation can affect splicing events; slow elongation rates generally promote the inclusion of weak exons.

Changes in the set of selected splice sites will impact the structural composition of the final RNA molecule. Given the potential differences in biological function between the resulting alternative transcripts, AS can result in the generation of proteins with different biological functions, structure, localization and interaction capabilities. Therefore, AS may occur in a tissue- or disease-specific manner (Oltean & Bates, 2014; Wang, Sandberg et al., 2008). In addition, it likely plays a role in dynamic processes such as development and cellular differentiation (Kalsotra & Cooper, 2011; Trapnell et al., 2010). It has also been suggested that a considerable amount of the detected AS products result simply from noisy splicing, reflecting an inherent error rate, and will have no specific function at all (Melamud & Moult, 2009). AS of pre-mRNAs can also contribute to the regulation of resultant protein product levels, through the formation of transcripts that will be targeted by the nonsense-mediated decay pathway, as well as producing transcripts incapable of producing functional proteins, for example through intron retention events or exon loss (McGlincy & Smith, 2008; Yap, Lim et al., 2012)

1.2. Next-generation sequencing

A transcriptome is the complete set of transcripts and their relative abundance within a cell, for a specific developmental stage or physiological condition. Understanding the transcriptome is an essential step towards interpreting the functional elements of the genome, revealing the molecular constituents of cells and tissues, and understanding development and disease. Therefore the key goals of transcriptome studies are: to discover and catalogue all species of transcripts; to determine the transcriptional repertoire of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions.

DNA microarray was the first technology developed for the high throughput comparison of expression levels across different cell types and environmental conditions (Malone & Oliver, 2011). Nonetheless, it had several limitations. For example, background hybridization limits the accuracy of expression measurements, particularly for transcripts present in low abundance. Furthermore, probes differ considerably in their hybridization properties, and arrays are limited to interrogating only those genes for which probes are designed. Therefore, in the past few years, RNA sequencing (RNA-seq) - the direct sequencing of transcripts by high-throughput sequencing technologies - has become the method of choice for the study of transcriptome composition (Wang, Gerstein, & Snyder, 2009). RNA-seq offers a much bigger dynamic range to study gene expression patterns compared to array technologies, and enables a much broader set of analyses. For example, besides standard differential gene expression analysis, RNA-seq allows for the identification of novel transcribed regions, including rearranged and fused genes, the study of allele specific expression, and the possibility to estimate transcript expression levels and to study differential splicing across conditions. However, RNA-seq poses novel algorithmic and logistical challenges for data analysis and storage. Many computational methods have been developed for alignment of reads, quantification of gene and/or transcripts, and identification of differentially expressed genes from RNA-seq data (Conesa et al., 2016).

The first next generation sequencing machine was released by 454 Life Sciences in 2005, followed by Solexa Genome Analyzer and SOLiD (Supported Oligo Ligation Detection) by Agencourt in 2006 (Mardis, 2013). In 2006 Agencourt was purchased by Applied Biosystems, and in 2007, 454 was purchased by Roche, while Illumina purchased Solexa. These are the best known next generation sequencing systems due to their competitive cost, accuracy, and performance. However, currently Illumina's platforms are the most commonly used for sequencing RNA. The reason behind such wide adoption of Illumina's systems is likely due to the large volume of information obtained from a typical sequencing run (i.e. sequencing depth) and good sequence accuracy compared to other competitors (Mardis, 2013).

The Illumina sequencing platform generates short-read (up to 150 bases) RNA-seq data. The major limitation of short-read RNA-seq is the difficulty in accurately reconstructing expressed full-length transcripts from the assembly of reads. This is particularly complicated in complex transcriptomes, where different but highly similar isoforms of the same gene are expressed. Therefore, the size of the final sequencing fragments is crucial for proper subsequent analysis. With improvement in RNA sequencing protocols, Pacific Biosciences recently introduced long-read (up to several kilobases) PacBio RNA-seq technology, which is capable of sequencing a single transcript to its full length in a single read. Nevertheless, long-read sequencing has its own set of limitations, such as a high error rate and low accuracy. If PacBio technology reaches a throughput that is comparable to the next-generation technologies, then the need for transcriptome assembly will probably be eliminated (Conesa et al., 2016).

In the current thesis, I focus on RNA-seq data generated by the Illumina technology.

1.2.1. RNA Sequencing experiment work-flow

Library preparation

Library preparation is the first step in sequencing RNA (van Dijk, Jaszczyszyn, & Thermes, 2014). It consists of obtaining the starting material, and converting it into a cDNA library that can be loaded into the sequencing machine. Once RNA is extracted from a sample, it is typically subjected to ribosomal RNA (rRNAs) reduction, i.e. the most abundant RNA species in the cell. This can be done through either polyA selection or ribodepletion. PolyA selection approach uses oligo-dT beads, which enable the specific extraction of polyAdenylated RNAs, hence ensuring a good representation of mRNAs. Ribodepletion approach relies on the use of ribonucleases to specifically digest rRNAs. Therefore, it has the advantage of not restricting the analyses to a specific type of RNA. Datasets produced with the polyA selection protocol are known as polyA-selected, and those obtained with ribodepletion are referred to as total RNA. Due to the simpler protocol and its lower price, polyA selection emerges as the most popular choice amongst the currently available RNA-seq datasets. However, studies that aim at characterising non-coding RNA species, which typically lack a polyA tail would be an exception (Figure 1-4, step 1).

The RNA is then fragmented via hydrolysis with divalent cations and retro-transcribed into double stranded cDNA by using random hexamer primers. The reason to use random primers is due to the unknown sequence of the obtained fragments (Figure 1-4 - step 2). Next, adapter sequences are ligated at both ends of each cDNA fragment. These adaptors enable the hybridisation of RNA fragments into the flow cell, where the sequencing takes place (Figure 1-4 - step 3). In addition, they serve as primer binding sites for the sequencing reaction. Using gel electrophoresis resulting cDNA fragments are size-selected to fit within the range required by the sequencing machine (typically 300-500 bp), and fragments outside this range will be missed. Finally, the resulting cDNA library is amplified by Polymerase Chain Reaction (PCR).

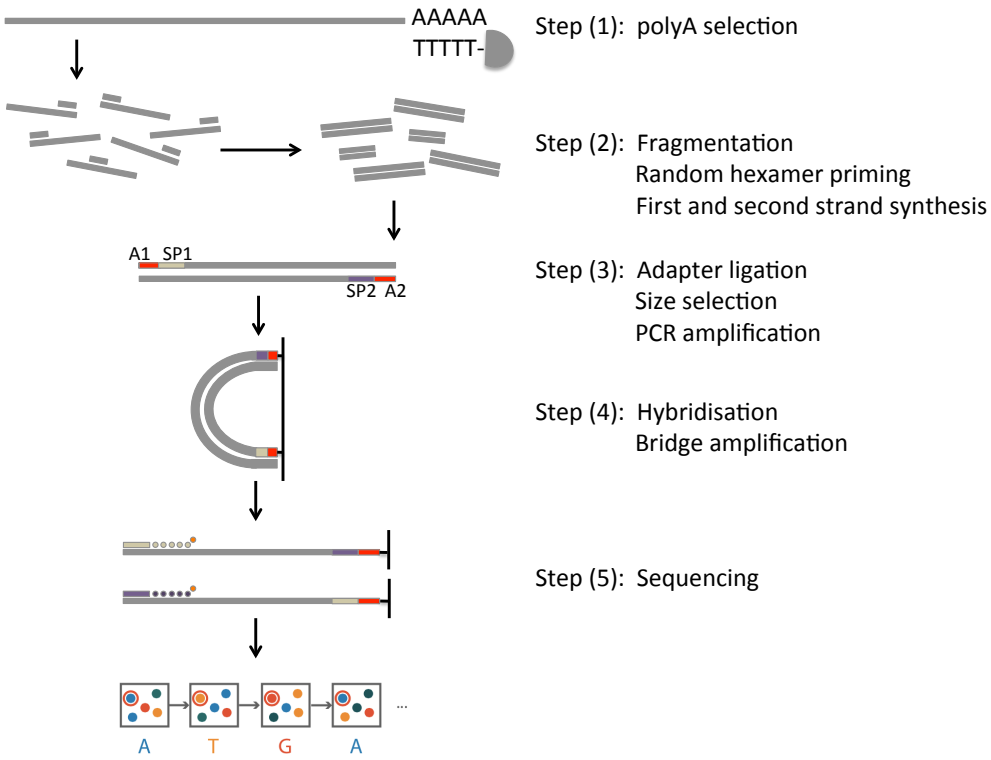


Figure 1-4. Overview of paired-end library preparation and sequencing steps in an Illumina platform. A workflow consists of ligating different adaptors at each end of the initial cDNA molecule, which enables sequencing each cDNA fragment from both ends, in two separate reactions. Paired-end sequencing has advantages for the downstream bioinformatic analyses compared to single-end sequencing.

Sequencing RNA

Once the RNA-seq libraries are produced, they can be loaded onto a flow cell for sequencing. A sequencing flow cell is saturated with complementary adapters to the ones ligated at both ends of the cDNA fragments, therefore they promote the hybridisation of the denatured double strand molecules (Mardis, 2013). In order to increase the signal from the sequencing reaction, the starting material is amplified once again through bridge amplification (Figure 1-4 - step 4). This amplification process allows clonal amplification of a large number of DNA fragments simultaneously and includes the synthesis of fragments that are complementary to the hybridised cDNA molecules. The cDNA fragments then bend over and hybridise with adjacent adapters thus enabling subsequent rounds of synthesis. At the end of each round the double stranded DNA is denatured so that each strand can separately attach to an oligonucleotide sequence anchored to the flow cell. At the end of bridge amplification, all of the reverse strands are washed off the flow cell, leaving only forward strands. As a result, a large number of clusters with identical sequences will be formed. The sample is now ready to undergo sequencing.

The Illumina platform relies on sequencing by synthesis to read the base pair composition of each cDNA cluster (Bentley et al., 2008). It uses modified versions of the four nucleotides (bases), which incorporate a reversible terminator, as well as a fluorescent dye (Figure 1-4 - step 5). Hence, only one base can be added during each sequencing cycle. The reason is that the reversible terminator on every nucleotide blocks elongation after a successful base incorporation. The identity of the incorporated base then can be recorded by measuring its fluorescent signal. Repetition of this process will lead to a set of images, which will be converted into a set of sequences or reads using a base calling software. The recorded reads represent the set of molecules expressed in the initial sample. The length of reads corresponds to the number of cycles performed during the sequencing reaction. The sequencing machine stores the obtained sequence information, together with the probability of a wrong base call at each given position of the read (i.e. Phred score) in a plain text file in FASTQ format (Cock, Fields et al., 2010).

Despite of its many advantages, effective RNA-seq utilization still faces some challenges. For example, the PCR amplification step can lead to differential amplification of fragments with higher or lower GC content (Benjamini & Speed, 2012). In addition, the failure to block the elongation reaction or to remove the fluorescent dye during the sequencing step can lead to incorrect base calls (Metzker, 2010). Alternative protocols or analysis methods have been introduced to overcome such biases. For example, alternative library preparation methods using random barcodes (i.e. molecular identifiers) to quantify the absolute number of molecules have been proposed to account for PCR bias (Shiroguchi, Jia et al, 2012). Alternative library preparation strategies also can add further information to the sequencing experiment. This is the case of strand-specific protocols, which are able to provide information on the DNA strand from which the specific transcript originates (Levin et al., 2010). The paired-end sequencing protocol (as opposed to the single-end) is a common strategy to overcome limitations on the read length. It allows sequencing of each cDNA fragment from both ends by ligating different adaptors at each end of the initial cDNA molecule, and sequencing each cDNA fragment from both ends, in two separate reactions (Mardis, 2013). Paired-end RNA-Seq facilitates discovery applications such as detecting gene fusions in cancer and characterizing novel splice isoforms.

Quality assessment of RNA-seq data

RNA-seq is a complicated, multistep process involving reverse transcription, amplification, fragmentation, purification, adaptor ligation, and sequencing. A disruption at any of these steps could lead to biased or even unusable data. Hence, comprehensive quality assessment is a critical step for all downstream analyses and results interpretation. RNA-seq quality control metrics include but is not limited to: base quality, sequence quality, nucleotide composition bias, guanine-cytosine (GC) bias, reads duplication rates (clonal reads), overrepresented sequences and sequencing adaptor contamination (Li, Nair et al., 2015).

A base quality analysis can be done using the Phred score provided in the FASTQ files by the sequencing machine for each sequenced nucleotide (Ewing, Hillier et al., 1998). The Phred score is defined as $Q=-10\times\log_{10}(P)$, where P is the probability of erroneous base calling. For example, a Phred quality score of 30 means the chance that

this base is called incorrectly is 1 in 1,000. Although there is no guideline to determine if the quality of a particular base is good or bad, in general, scores over 30 indicate very good quality, 20-30 indicate reasonable good and less than 20 indicate poor quality. Phred quality scores can be visualized in parallel boxplots illustrating per base quality score for all reads at each position (Andrews, 2016). In addition, one can also calculate the average quality score per read (per sequence quality score) and check the quality score distribution of all sequences. This analysis allows identification of subset of sequences that may have universally low quality values (Andrews, 2016). It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged (e.g. on the edge of the field of view), however these should represent only a small percentage of the total sequences.

Assuming that RNA-seq reads were randomly sampled from expressed transcripts, one would expect to see little to no differences between the nucleotide composition (percentage of A, C, G, and T) at each position. Where, random fluctuations are cancelled out because of the large sample size. GC content is the percentage of bases in a sequence that are either guanine or cytosine. Measuring the GC content is a simple way to evaluate the nucleotide composition of DNA or RNA. Per sequence GC content can be roughly used to measure the randomness of sequencing library as GC content of reads from random sequence library follows normal distribution with the mean equals to the overall GC content of the transcriptome. While, a poorly prepared or contaminated library will exhibit a skewed distribution. The dependence between read coverage and the GC content of reference genome in high-throughput sequence data has been shown previously (Benjamini & Speed, 2012). A serious bias suggests the existence of overrepresented sequences in a sample, and such bias will influence coverage uniformity as well as transcripts abundance estimation. Therefore, evaluating GC content bias in RNA-seq data is of great importance to both transcript detection and abundance quantification. The reason to use GC rather than AT (or AU in RNA) is that GC content carries more direct biologic meaning. GC pairs are more stable than AT (3 vs. 2 hydrogen bonds). Therefore, it has implications in PCR experiments, since the GC content of primers predicts their annealing temperature. Furthermore, exons have on average a higher GC content than introns and intergenic regions.

Read duplication rate can be affected by read length, sequencing depth, transcript abundance and PCR amplification. A read is duplicated if there is an exact sequence match over the whole length of the read. Therefore, supposing the sequencing library is purely random the chance to get a duplicated read is very slim even if the sequencing depth reaches hundreds of millions. A low level of duplication may indicate a very high level of coverage of the target sequence, however a high level of duplication is more likely to indicate an enrichment bias. The majority of duplicated reads are artifactually generated from PCR amplification (Andrews, 2016). And because of this, duplication rate analysis mostly only includes checking for PCR amplification bias. In general, if there are more than 50% of duplicated sequences in total in an RNA-seq sample, the sample will be considered as seriously biased and not randomly sampling the target sequence.

A high-throughput library with good quality contains a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a single sequence is overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated. It can also be an indication that the sequenced sample is not as diverse as expected. Overrepresented sequences may also be detected due to high duplication rate. One of the common sources of overrepresented sequences is the read-through adapter sequences that are built up on the end of sequences (Andrews, 2016).

The trimming process, which removes N nucleotides from the beginning or the end of a sequencing read, can improve the quality of a sequenced sample by removing low quality bases as well as adaptor sequences (Babraham Bioinformatics, 2015). One may also improve the sample quality by removing overrepresented and duplicated reads. Many tools have been developed that preform RNA-seq quality assessment and trimming (Andrews, 2016; Babraham Bioinformatics, 2015; Li et al., 2015).

Read mapping strategies

The next step in an RNA-seq analysis pipeline is to identify the genomic region that each read has originated from. This task for an RNA-seq sample is equivalent to

discovering the loci that are expressed in a given sample. There are two strategies to perform this task: in the first approach, reads are directly aligned to the reference genome or transcriptome (Li & Homer, 2010). Therefore, using this approach depends on the availability of a reference, which may not always be the case. In the second approach, reads can be directly assembled into contigs (i.e. contiguously expressed regions) with the aim of reconstructing the set of expressed transcripts (Martin & Wang, 2011). In general, the first strategy constitutes a much simpler approach, and it is typically the method of choice when working with model organisms.

Read mapping is usually the bottleneck of an RNA-seq analysis workflow. Therefore, available mapping tools make use of heuristic parameters such as the maximum number of allowed mismatches per read to speed up this task. While, this process can lead to information loss due to the lower sequence quality at the 3' end of the read. The quality difference commonly occurs when working with Illumina platforms, since interpreting the fluorescent signal as sequencing cycles accumulate becomes more difficult (Minoche, Dohm et al., 2011). Therefore, the sequence quality assessment and trimming, as explained in the previous section, helps with identifying and removing such sequences in order to speed up the subsequent mapping process. The trimming process either shortens the read by cutting off the low quality sequence, or removes the entire low quality read.

When a reference genome is available, the commonly used approach is to align the reads directly to the genome sequence. Similarly, reads can be aligned to a transcriptome reference if a good annotation exists. The advantage of the second strategy is that due to the lack of intronic sequences in a transcriptome reference the alignment process will be simplified. However, this approach limits the downstream analysis that can be performed (Martin & Wang, 2011). For instance, alignment to the transcriptome is neither compatible with the identification of novel expressed regions nor the study of intronic expression levels. Some RNA-seq read mapping tools use a hybrid approach (e.g. TopHat) (Trapnell, Pachter, & Salzberg, 2009). Such tools have the advantage of using a reference genome along with the available exon-exon junction annotation. In addition, there are some short read mapping tools (such as Bowtie) that are able to detect exon-exon junctions without the need for any priori knowledge on the

annotation (Langmead, Trapnell et al., 2009). Such aligners usually report a splice junction whenever a read appears to span multiple exons. The identified splice sites and their flanking sequences are then concatenated into a novel transcriptome, which is then used to re-align the set of unmapped reads. If the RNA-seq data is paired-end, each read is usually processed separately. Once the potential alignments are obtained, they are evaluated by taking into account additional information such as fragment length and orientation of the reads. All the information gathered during the mapping process is reported in SAM/BAM format. SAM stands for Sequence Alignment/Map. Similarly, BAM stands for Binary Alignment/Map.

When the species of interest lack a reference genome, *de novo* assembly emerges as an advantageous strategy. It also can be used in situations where the genome composition of a given sample is expected to differ largely from that of the reference assembly (e.g. cancer samples). *De novo* assembly relies largely on the overlap among the reads to assemble them into contigs. (Martin & Wang, 2011). Although the short read length makes the task of *de novo* assembly difficult, the use of paired-end data can slightly simplify this process. The assembly of lowly expressed genes will still be a challenging task to do. There are several *de novo* assembly tools including Trans-ABySS (Robertson et al., 2010) and Trinity (Grabherr et al., 2011) that are commonly being used by the bioinformatics community. The *de novo* transcriptome assembly allows identification of novel splice junctions and AS events.

1.2.2. Detecting and measuring expression differences in the transcriptome

Once the reads are mapped to the reference genome or transcriptome, the next step of an RNA-seq analysis pipeline is to estimate the level of expression for genes and transcripts. Similar to the read mapping strategies, the quantification of expression levels can be achieved by relying on existing information (i.e. gene and isoform annotation), or it can be done through *de novo* identification of transcribed regions and independent of any annotation information.

Gene expression levels

The abundance of gene transcription products is an important measure to infer the endogenous state or response of a cell under various conditions, and identifying differentially expressed genes is a powerful approach to help determine their functions. When a complete gene annotation exists, the abundance estimation can be easily achieved by counting the number of reads that overlap with each gene locus. Once the raw read counts are estimated for the entire genes, many downstream analysis can be performed including differential gene expression analysis (Love, Huber et al., 2014; Robinson, McCarthy, & Smyth, 2010; Tarazona, Garcia-Alcalde et al., 2011). Despite the simplicity of the coverage analysis, there are some challenges that need to be considered while performing this analysis. First, reads that map to multiple locations in the genome, and those that arise from repetitive or duplicated loci need to handled carefully to avoid over-estimating the expression levels. In this case, coverage analysis tools often discard such reads. However, they can also be handled by uniformly distributing them to all the mapped positions or probabilistically assigning them depending on the coverage at each mapping locus in order to avoid information loss (Trapnell et al., 2010). The second challenge arises from the overlapping features. In most cases, such reads remain ambiguously assigned (Figure 1-5).

Alternatively, gene expression levels can be calculated after estimation of transcript expression levels by aggregating the corresponding individual transcript abundances.

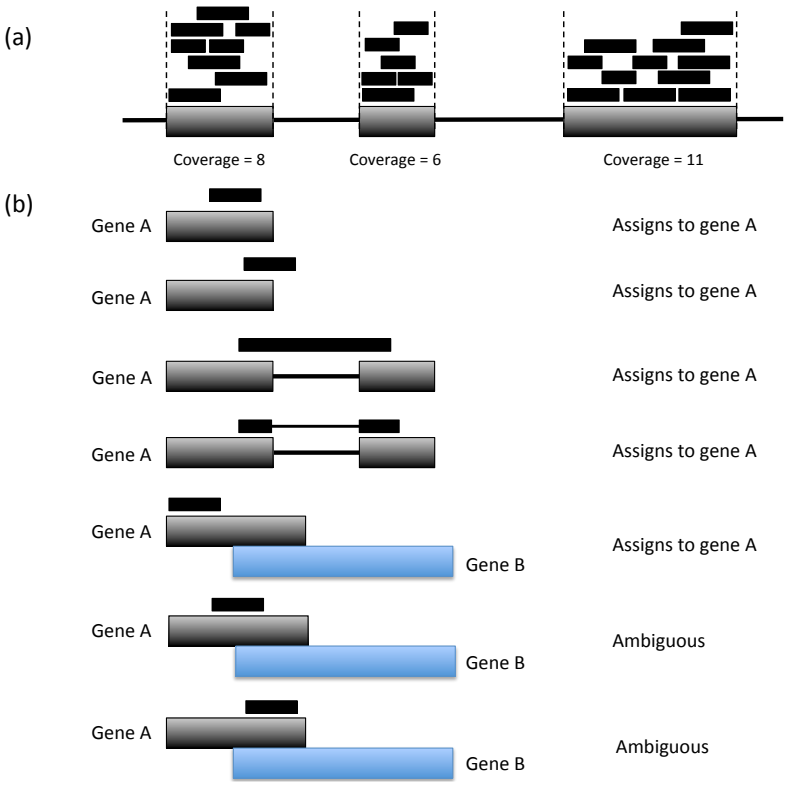


Figure 1-5. **Counting reads.** (a) An illustration of the read counting concept. (b) Examples of challenges of counting reads. When a read overlap with multiple locations, it is not always clear where it should be aligned. Different methods take different approaches. A simple process is shown above.

Transcript expression levels

The estimation of expression level becomes more complicated when the focus is on individual transcripts. The reason is that many reads overlap with exons that are shared by multiple isoforms of the same gene. Currently available algorithms rely on those reads that map uniquely to one of the annotated transcripts within the loci. In addition, split reads (i.e. those that span two different exons) and the paired-end information becomes especially informative (Li & Dewey, 2011). Similarly, the fragment length distribution can be used to deconvolute ambiguous assignments by attributing a lower likelihood to those that would require extreme distances between the paired reads.

De novo transcript identification

One of the main advantages of RNA-seq over other gene expression analysis techniques such as microarray is the possibility to gather information on novel expressed loci in a more high throughput manner. The *de novo* assembly of RNA-seq data allows for the identification of novel genes and alternate splice isoforms independent of the knowledge of the reference genome. Detection of AS events usually involves assessing part of a gene associated with the transcript isoform of interest. For example, in the case of a gene with a cassette exon (i.e. the inclusion or skipping of a single exon) and two transcript isoforms, the presence or absence of one or more transcript isoforms will be indicated by the relative expression of this exon. This can be assessed by a test between experimental groups for the normalised expression of that exon (Martin & Wang, 2011; Robertson et al., 2010)

The *de novo* transcriptome assembly strategy does not use a reference genome; instead it leverages the redundancy of short-read sequencing to find overlaps between the reads and assembles them into transcripts. The *de novo* assemblers usually assemble the data set multiple times using a De Bruijn graph-based approach to reconstruct transcripts from a broad range of expression levels and then post-process the assembly to merge contigs and remove redundancy (Martin & Wang, 2011). Most of the currently available *de novo* assemblers are developed and optimized using short-read data sets, while longer second-generation reads, such as 454 reads, can also be

integrated into *de novo* transcriptome assemblies, which may even improve the ability to resolve alternative isoforms.

The *de novo* transcriptome assembly has several advantages over the reference-based strategy (Martin & Wang, 2011). First, it does not depend on a reference genome. It can recover transcripts that are transcribed from segments of the genome that are missing from the genome assembly, or detect transcripts from an unknown exogenous source. Second, the *de novo* assembly does not depend on the correct alignment of reads to known splice sites. Similarly, it is independent of the accuracy of prediction of novel splicing sites, as required by reference-based methods. Finally, trans- and alternatively spliced transcripts and similar transcripts originating from chromosomal rearrangements can be assembled using the *de novo* approach.

Read count normalization

The result of an RNA-seq quantification approach is an estimate on the number of reads that can be attributed to a certain feature, which is referred to as counts. Although the counts are proportional to the levels of expression for the certain feature of interest, they depend on the total number of sequenced reads (sequencing depth) as well as the length of the feature. The counts may also be impacted by further experimental biases (Hansen, Irizarry, & Wu, 2012; Lee et al., 2011; Oshlack & Wakefield, 2009; Roberts, Trapnell et al., 2011). Therefore a normalization method is needed in order to enable the comparison of read counts across different samples and features. One of the commonly used measures to report the level of expression derived from an RNA-seq experiment is the Reads per Kilobase per Million mapped reads (RPKMs) in the case of single-end data. While the Fragments per Kilobase per Million mapped reads (FPKMs) has been recommended for paired-end RNA-seq data (Mortazavi, Williams et al., 2008)

$$U_{ij=}{}^{K_{ij}}\Big/_{N_j L_i} . 10^9$$

Where:

U_{ij} = Normalized expression of gene *i* in sample *j*

K_{ij} = Observed counts (reads/fragments) for gene *i* in sample *j*

N_{ij} = Total number of reads in sample *j* (sequencing depth)

L_{ij} = Length of gene *i*

The RPKMs and FPKMs are currently being used as established intuitive measure of expression levels in RNA-seq. However, they work based on the assumption that the overall RNA levels are similar across samples, which may not always be the case. Therefore, they may fail to properly estimate the normalisation factors in cases where the compared libraries differ in their composition (Robinson & Oshlack, 2010). This caveat can be illustrated by comparing the expression of genes in two RNA-seq samples; one expressing an extra small set of highly expressed genes (sample A), while the other one does not (sample B). The sample A is more likely to detect reads from genes with high expression levels. This is due to the sampling nature of the RNA-seq. Therefore, even if the two samples are sequenced at similar depth, the signal from commonly expressed genes will be lower in the sample A. In such cases, if one uses the above mentioned normalisation method, it leads to the identification of most genes undergo expression differences between the two samples. Whilst the observed differences could be better explained by the isolated differential expression of the few non-overlapping genes. This example illustrates the need for more robust normalisation methods than the RPKM/FPKM for RNA-seq, especially when the goal is to compare across libraries. An example of those methods is developed within the DESeq2 Bioconductor package, where it calculates a geometric mean for each gene in order to capture the variability of the observed measurements across all the libraries (Love et al., 2014). This approach is similar to obtaining a reference sample for the expression analysis. These values are then used to normalize the read counts. Lastly, the library-specific normalisation factors are obtained from the median of the calculated ratios;

$$S_j = \underset{i:\, K_i^R \neq 0}{\text{Median}} \Big/ K_i^R$$

Where:

S_j = size factor for sample *j*

K_{ij} = observed counts for gene *i* in sample *j*

K_i^R = geometric mean for gene *i* across the *m* sample, where geometric mean is $(\prod_{v=1}^{m} K_{iv})^{1/m}$.

Differential expression analysis

The assessment of differences in expression levels is one of the most common uses of RNA-seq data. Once the coverage analysis is performed and the corresponding counts are obtained, differential expression analysis can be performed at both gene and transcript levels. Many tools including DESeq2 Bioconductor package have been developed for such analysis (Love et al., 2014). In order to address the significance of the detected expression changes, the majority of these methods rely on the use of Generalised Linear Models (GLMs) of the Negative Binomial (NB) family. A differential expression analysis workflow would consist of normalising the observed counts in order to enable their comparison across libraries. Next, using the replicate samples, for each gene, an estimate on the amount of variability is calculated. Replicates may either be biological or technical replicates. Finally, the differential expression test is performed.

1.3. Disruption of RNA processing in human cancer

Cancer cells have two intrinsic properties that make them pathological for living organisms: They reproduce in defiance of the natural limitations on cell growth and division, and invade and colonize areas normally occupied by other cells (Alberts et al., 2007). A cell that grows and proliferate abnormally and uncontrollably into a mass will result in a neoplasm i.e. a tumour. A neoplasm is considered benign when its cells do not invade nearby tissue or spread to other parts of the body. Such tumours are usually easy to treat by surgically removing the tumour mass. However, if tumour acquires an

ability to invade into the surrounding tissues, it is considered malignant or cancerous. Cancer cells may invade to the surrounding tissues and spread to form secondary tumours called metastases. It is usually the metastases that result in the death of the cancer patient (Alberts et al., 2007; Hanahan & Weinberg, 2000).

Cancer is typically caused by genetic changes effecting protein coding genes and impacting the role of their protein products. These changes can be mutations, deletions, and insertions that change the amino acid sequence of the translated peptide. In addition, synonymous changes, copy number variations (CNVs), as well as changes occurring in intronic regions can lead to gene dysfunction and cancer (Stratton, Campbell, & Futreal, 2009).

In general, studies of human genetic diseases have shown that up to 50% of mutations contributing to disease affect RNA splicing, where 10% directly disrupt splice sites (Krawczak et al., 2007; Lopez-Bigas, Audit et al., 2005). Mutations affecting RNA splicing have also been implicated in cancer formation and progression. For example, the splicing factor SF3B1 is mutated in approximately 20% of patients with myelodysplastic syndromes (Malcovati et al., 2015). Similarly, a mutation creates an ESE in the KLF6 gene in prostate cancer, where it promotes expression of an isoform that accelerates tumour progression (Narla et al., 2008). Also the up regulation of SR proteins in ovarian and colon cancer regulates splicing of a number of oncogenes (Ward & Cooper, 2010).

Currently there are ten known hallmarks of cancer, including self-sufficiency in growth signals, insensitivity to anti-growth signals, evading programmed cell death (apoptosis), limitless replicative potential, developing blood vessels (angiogenesis), tissue invasion and metastasis, deregulated metabolism, evading the immune system, unstable DNA, and Inflammation (Hanahan & Weinberg, 2000; Hanahan & Weinberg, 2011). Each of these widely accepted hallmarks could be affected by aberrant splicing (Oltean & Bates, 2014). In particular, apoptosis and metastasis are affected by AS in a number of genes. For example, the overexpression of the anti-apoptotic transcript variants of BCL2L1 (BCLXL) confers resistance to apoptosis in cancer (Oltean & Bates, 2014). In addition, abnormal expression of TP53 splicing isoforms is involved both in

apoptosis and cell proliferation (Oltean & Bates, 2014). A splicing switch between pro- and anti-angiogenic isoforms of VEGFA is also observed between cancer and healthy samples in several tissue types (Oltean & Bates, 2014).

AS is known as a process contributing to structural transcript variation and proteome diversity. It also regulates gene expression by generation of premature termination codons, and subsequent targeting by nonsense-mediated mRNA decay. Although numerous normal and disease related AS events have been identified and characterized in recent years, the function of the majority of observed splicing events is unknown. In addition, in some cases, AS appears to results in non-functional end-products. It has to be noted that the splicing pathway can also be considered error-prone which introduces noise and stochastic variation in the transcriptome, resulting in generation of mis-spliced and non-canonical transcripts at low abundance in most genes. Regardless, aberrant splicing commonly denotes splicing events that are associated with disease, and differs from the splicing patterns found in healthy tissues.

1.4. Cancer therapeutics

Treatment of cancer is currently a double-edged sword. It needs to be aggressive enough to destroy tumour cells completely. However, it is this aggressiveness that causes severe side effects through deleterious effects on normal cells. One way in which the efficacy of systemic therapeutics can be improved would be to locally enhance their concentration at the tumour site. One approach to accumulate therapeutic agents at the tumour site, while minimizing their presence at other sites in the body, is to conjugate/fuse them with tumour-specific monoclonal antibodies (Zhang, Chen et al., 2007).

Antibodies represent a natural response by the immune system to the presence of foreign proteins within the body. An antibody is a protein that identifies and binds to a specific protein called an antigen (Figure 1-6). They circulate throughout the body until they find and attach to their antigen. Once attached, they can recruit other parts of the immune system to destroy the cells presenting the antigen. Monoclonal antibodies

(mAbs) are identical antibodies that are generated from a cell population derived from a single isolated immune cell to specifically target a certain antigen (Scott, Wolchok, & Old, 2012). Therefore, in order to make mAbs, it is critical to identify the right antigen to attack. There are three types of mAbs: naked, conjugated, and bispecific mAbs. Naked mAbs are the most common type of mAbs used in cancer treatment. They can block and kill tumour cells in different ways (Scott et al., 2012); naked mAbs can boost a person's immune response against cancer cells by attaching to them and acting as a marker for the body's immune system to destroy them, or they can boost the immune response by targeting immune system checkpoints. Other naked mAbs work mainly by attaching to and blocking antigens on the surface of cancer cells that help cancer cells grow or spread. For example, trastuzumab (Herceptin®) is an antibody against the HER2 protein (Albanell & Baselga, 1999). Most patients with ovarian and breast tumours express high levels of this gene. When HER2 is activated, it helps tumour cells to grow and proliferate. The binding of trastuzumab to these proteins sterically hinders the oncogenic function of HER2 (Albanell & Baselga, 1999).

Although mAbs targeting certain surface receptors may possess sufficient anti-tumour activity to be viable therapeutics themselves, e.g. by hindering the function of the bound protein as is the case with anti-HER2 trastuzumab, the concept of coupling highly potent cytotoxic molecules to antibodies via linkers expands significantly the potential for antibody based approaches (Figure 1-6). Conjugated mAbs are mAbs joined to a chemotherapy drug or to a radioactive particle (Polakis, 2016). In this case, the antibody is being used as a homing device to deliver the conjugated drug directly to the cancer cells. These antibodies circulate throughout the body until they can find and bind onto their specific target protein. Then, they can deliver their toxic payload to the cancer cells. This approach minimizes the damage to normal cells in other parts of the body. The key to achieve this goal is to identify a target protein that is specific to the tumour cells and it is expressed at a low level or is absent in healthy normal tissues. Chemolabeled antibodies, also known as antibody-drug conjugates (ADCs), usually carry a drug that is often too powerful to be used systematically on its own. An example of these antibodies is TDM-1 (Kadcyla®), an antibody that targets the HER2 protein, attached to a chemotherapeutic drug called DM1 (Verma et al., 2012). This drug is suitable for the treatment of breast cancer patients whose cancer cells express HER2 at a high level

(Verma et al., 2012). Interestingly, in case of conjugated mAbs, the target protein need not even necessarily be driving proteins of oncogenesis as long as they have tumour specific or enriched profiles compared to normal tissues – although one might presume that proteins involved in oncogenesis would be preferred targets for therapeutic development.

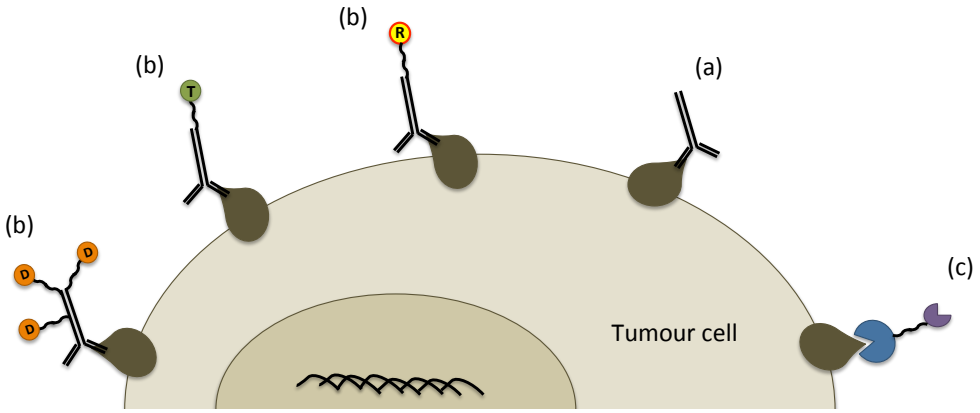


Figure 1-6. **Targeted antibody-based therapeutics.** (a) Targeting mAbs to the tumour can result in destruction of tumour cells by antibody-dependent cellular cytotoxicity or complement-dependent cytotoxicity. (b) A direct approach to kill tumour cells is the conjugation of cytotoxic drugs (D), toxins (T) or radionucleotides (R) to mAbs. (c) Bispecific antibodies can modulate immune response against tumour cells. They are capable of targeting two proteins on the surface of tumour cells simultaneously. In addition, they can bring immune cells to the tumour site by binding to a target on the surface of a tumour cell and the other target on the surface an immune cell.

The bispecific mAbs are made up of parts of two different mAbs (Figure 1-6). This means that they can attach to two different proteins at the same time(Chames & Baty, 2009)An example is blinatumomab (Blincyto), which is used to treat some types of acute lymphocytic leukemia (Sanford, 2015). One part of blinatumomab attaches to the CD19 protein, which is found on some leukemia and lymphoma cells. Another part attaches to CD3, a protein found on immune cells called T cells. Therefore, blinatumomab brings the cancer cells and immune cells together by binding to both of these proteins. This process is thought to cause the immune system to attack the cancer cells.

Complex diseases such as cancer are often multifactorial in nature and involve redundant or synergistic action of disease mediators or up regulation of different receptors. Therefore, blockade of multiple different pathological factors and pathways simultaneously may improve the therapeutic efficacy. This goal can be achieved by using the dual targeting strategies applying bispecific antibodies. Bispecific antibodies offer more binding specificity and improved efficacy than mAbs, since they can bind to two target proteins on the surface of tumour cells simultaneously. An example of such antibodies is the bispecific antibody that targets EGFR and IGFR proteins on the surface of tumour cells that express both of them (Kontermann, 2012).

Over the past couple of decades, the US Food and Drug Administration (FDA) has approved more than a dozen antibodies including all three types to treat certain cancers (www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm279174.htm). Although mAbs represent a powerful way to target cancer cells, a critical and key step in the process is the identification of appropriate proteins to target. While, cell surface proteins highly expressed in cancers have been shown as attractive and successful targets in the clinic, protein changes that occurs in the extracellular region of surface proteins and accessible by antibodies, e.g. splice variants, may also serve as attractive targets for mAbs.

1.4.1. Differentially expressed genes as therapeutic targets

The safety and efficacy of therapeutic mAbs in oncology vary depending on the nature of the target (Papkoff, 2007). An ideal antibody target should be abundant and accessible and should be expressed homogeneously, consistently and exclusively on the surface of cancer cells. Aside from tumour overexpression, a second key aspect is normal tissue expression. Ideally, there should be no or very little expression in any normal human tissues. Normal expression of the target in tissues such as lung, heart or kidney that cannot sustain damage from a targeted therapeutic represents a major concern and consequently can disqualify a promising target (Papkoff, 2007). Conversely, some normal tissues such as uterus, thyroid or prostate can be subjected to toxic therapies without life-threatening consequences. Therefore, some level of normal expression may be tolerable. In such cases the potential toxicities that could ensue depend on the relative ratio between tumour and normal expression levels. Examples of overexpressed proteins that have been identified as suitable targets for antibody therapy include EGF receptor and HER2/neu. Even though they are expressed on a variety of normal tissues the therapeutic antibodies apparently have minimal toxicities due to this normal expression within their therapeutic windows (Deckert, 2009). Other examples include antibodies targeting CD20 and VEGF, which have all shown significant benefit for the treatment of patients in both liquid and solid tumours (Deckert, 2009).

1.4.2. Alternatively spliced genes as therapeutic target

Within human cancer alternatively spliced forms of proteins on the cell surface are obvious targets for antibody-based therapies - particularly if the splice variant is tumour specific. Even in the cases where the splice variant is rare but not tumour specific a comprehensive understanding of the normal tissues where it is expressed and its expression levels can allow the potential toxicity to essential organs and side effects to be predicted. Certainly, targeting a therapeutic antibody to the tumour is fundamentally more appealing than systemic untargeted application of chemotherapeutics. An example of successful mAbs targeting a splice variant is AMG595, which specifically bound to a constitutively-activated, oncogenic form of

epidermal growth factor receptor variant III (EGFRvIII) (Carlsson, Brothers, & Wahlestedt, 2014). EGFRvIII has restricted tumour-specific expression, including glioma, breast, non-small cell lung, ovarian, head and neck, and prostate cancers (Carlsson et al., 2014). This splice variant of EGFR has a deletion of exons 2-7 creating a novel epitope unique to the tumour. Genomic deletion of exons 2-7 has also been detected in some but not all of the EGFRvIII expressing tumours (Wheeler et al., 2015). AMG595 is an immunoconjugate, which consists of a human mAb directed against the deletion-mutant of EGFR. This mAb is conjugated via a non-cleavable linker to the cytotoxic agent maytansinoid DM1, with potential antineoplastic activity (Carlsson, Brothers, & Wahlestedt, 2014).

Hence, somatically generated cancer-specific protein isoforms may represent attractive candidates for mAb development in oncology, particularly if such protein isoforms are highly recurrent either within or across tumour types. Even though many of these variants may have lower expression than the canonical isoforms the ability to conjugate highly potent cytotoxic compounds to the binding mAbs can mitigate this problem.

1.4.3. Identifying optimal therapeutic targets

The availability of large datasets of cancer transcriptome data now provide an unprecedented opportunity to comprehensively identify cancer associated differentially overexpressed genes as well as alternative spliced forms or other transcriptomic re-arrangements that are involved in oncogenesis or arise in tumours as a consequence of therapy (Dargahi et al., 2014; Sebestyen et al., 2016; Tsai, Dominguez et al., 2015). Bioinformatic analyses of these datasets has the potential to identify novel biomarkers that can discriminate between tumour and normal tissues, interrogating the tumours at both the DNA and RNA levels. The advent of next-generation sequencing approaches allows identification of large classes of genetic defects and differentially expressed transcripts within and across cancer types. The candidate therapeutic targets identified by these technologies can be further narrowed down through three key characteristics representative of an ideal therapeutic antibody target, including target localization, expression pattern, and function (Papkoff, 2007):

(1) A desirable tumour biomarker should be located on the cell surface. Currently, more than 4,000 human cell-surface proteins have been annotated. Proteins localized to the surface of human cells are potential diagnostic and therapeutic targets.

(2) An ideal tumour biomarker should be highly expressed or uniquely expressed on the majority of tumour cells with no or limited normal tissue expression. Therefore, the expression profile of an ideal candidate should be abundant on the surface of tumour cells at all stages of cancer development to provide a broader window of opportunities for treating patients, while its expression is restricted or absent from vital normal tissues to minimize the risk of toxicities. An exception to overexpression would be proteins expressed by both normal and cancerous cells at a similar level, while a unique form is expressed within the cancer, including novel splice variants and fusion proteins. Tumour-associated aberrant proteins are highly-attractive targets, since mAbs can be directed towards a protein domain uniquely expressed on the tumour cell surface but absent from normal tissue. For example, AS may add or delete functional domains from protein coding sequences, causing a completely different physiological activity and structural conformation of splice variant compared to the wild-type protein. Also, treatment resistance-associated splice variants, which may arise as response to traditional chemotherapy through survival adaption mechanisms can be specifically targeted by mAbs. As they are specific responses to a compound's action, these changes may be more likely to be recurrent and specific to cancer cells and thus, make attractive targets for mAb therapy.

(3) It is favourable that a tumour biomarker plays a defined role in malignant transformation, however it is not required. Tumour biomarkers with a role in malignant transformation may be essential for cancer cell survival. Therefore resistance to a therapeutic mAb through gene loss or mutation might be less likely to arise. Since the conjugated mAbs are very effective and strong agents, even if the function of target is unknown or is not driving oncogenesis, it still can be used as a tumour cell surface marker for these antibodies to specifically deliver toxins to the tumour site. In this case, it is required that the target expresses at low levels or be completely absent from critical normal tissues.

1.5. Experimental design and Aims

The work presented in this thesis focuses on the use of RNA sequencing for the high throughput study of differentially expressed genes and alternative transcript products in human cancer samples. Overall, the goal is to identify such cases ideal for antibody-based therapeutics in cancer. Hence, the following chapters will discuss methods and algorithms of studying gene expression and splicing using RNA-seq data (chapter 2), a pan-cancer analysis of The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) transcriptome data in order to identify cancer-associated differentially expressed genes and the affected pathways (chapter 3.1), a pan-cancer analysis of TCGA and GTEx transcriptome data in order to identify cancer-associated splicing variants (chapter 3.2), introducing an R package based on analytic hierarchy process approach (AHP) to prioritize and rank identified candidate genes according to their tumour target biomarker potential (chapter 3.3), and finally the conclusion and discussions (chapter 4).

Chapter 2. Methods and Materials

2.1. Datasets

The Cancer Genome Atlas project (TCGA) began in 2005 under the supervision of the National Cancer Institute's Center for Cancer Genomics (NCI) and the National Human Genome Research Institute (NHGRI) with the goal of cataloguing cancer causing events including genomic mutations using the latest sequencing technologies and bioinformatics approaches. TCGA is currently characterizing 33 cancer types including 10 rare cancer (http://cancergenome.nih.gov/). In order to achieve these goals, TCGA primarily performs genome, transcriptome, and exome sequencing. All of the TCGA data are available to researchers at the Data Coordinating Center (DCC) established to provide data access (https://tcga-data.nci.nih.gov/). Most of the TCGA data is completely open access, except for data that could potentially identify specific patients (e.g. rare germline variants). This Clinically Controlled-Access data can be accessed through application to the Data Access Committee (DAC).

The availability of large datasets of genetic information such as TCGA introduces an unprecedented opportunity to exploit these data and generate novel hypothesis and approaches for the treatment of cancer. In order to achieve the goals within this thesis, I used TCGA as my primary discovery dataset. Raw RNA-seq and clinicopathological data were downloaded from the TCGA data portal (http://cancergenome.nih.gov). Permission to access TCGA data was obtained from the DAC of the National Center for Biotechnology Information's Genotypes and Phenotypes Database (dbGAP) at the National Institute of Heath (NIH). Sample collection, library preparation, and sequencing RNA methodologies have been described by TCGA previously (Cancer Genome Atlas

Research Network, 2011). Predominantly, TCGA data is generated from primary tumours that have not received any systemic treatment.

Similar to TCGA, The Genotype-Tissue Expression (GTEx) is a large project that was begun in 2010 under the supervision of NIH with the goal of studying the relationship between genetic variation and gene expression in human tissues (http://www.gtexportal.org/). GTEx characterises more than 30 tissue types collected from deceased donors and organ/tissue transplant patients. These tissues are collected from individuals free of major disease processes. Hence, combining the two TCGA and GTEx datasets offers a unique opportunity to identify cancer-associated events in each available TCGA cancer type by comparing each cancer against the entire available repertoire of GTEx normal tissues. Permission to access GTEx data was obtained from the dbGAP at the NIH. Sample collection, library preparation, and sequencing RNA were described by GTEx previously (GTEx Consortium, 2013).

Illumina BodyMap 2.0 project is also consists of 19 normal transcriptomes from 16 different tissue types, making it an invaluable source for studying tissue-specific transcript models (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/). The raw RNS-seq data is publically available to download through the above-mentioned link.

2.2. RNA-seq quality control and trimming

As mentioned in chapter 1, RNA-seq quality control metrics include but are not limited to: base quality, sequence quality, nucleotide composition bias, guanine-cytosine (GC) bias, reads duplication rates (clonal reads), over-represented sequences and sequencing adaptor contamination. There are number of tools developed to perform quality control and trimming (if necessary) on RNA-seq data including Trim Galore (Babraham Bioinformatics, 2015).

Trim Galore is a wrapper script around Cutadapt (Martin, 2014) and FastQC (Andrews, 2016) tools to consistently apply quality control checks as well as quality and adapter trimming to FASTQ files. The Cutadapt tool finds and removes adapter

sequences, primers, poly-A tails and other types of unwanted sequence (e.g. low quality) from the reads of an RNA-seq libaray. It performs these trimming tasks by finding the adapter or primer sequences in an error-tolerant way. Trim Galore leverages Cutadapt to perform adaptor sequence detection and trimming by using the first 13 bp of Illumina standard adapters ('AGATCGGAAGAGC'). If a sequencing read becomes too short (shorter than 25 bases) after trimming, Trim Galore can either remove the read from the dataset, or write it to a separate file so the information is not entirely lost and can be used with caution (e.g. using strict parameters for alignment of such short reads). In addition, using the Phred quality of base calls, Trim Galore can identify bases with lower quality than 20 and trim the reads. Trimming the adaptor sequences as well as low quality bases will improve the alignment quality of an RNA-seq data. Then, Trim Galore uses FasyQC to perform quality control by measuring the metrics listed in section 1.3.1 (quality assessment of RNA-seq data).

2.3. RNA-seq alignment

RSEM (RNA-Seq by Expectation Maximization) is a software package for both gene and isoform quantification (Li & Dewey, 2011). It is designed to work with reads aligned to transcript sequences, instead of a reference genome. Hence, it consists of two major steps: First, it requires generation of a set of reference transcript sequences. Then, it aligns a set of RNA-Seq reads to the reference transcripts and uses these alignments to estimate abundances.

RSEM uses the Bowtie alignment program to align the RNA-seq short reads (Langmead et al., 2009). It uses parameters specifically chosen for RNA-Seq quantification, which later will help with better quantification of the data. For example, RSEM runs Bowtie to find all alignments of a read with at most two mismatches in its first 25 bases. This will allow RSEM to determine which alignments are most likely to be correct, rather than assigning the aligner this responsibility. This approach leads into a more accurate estimation, due to the more detailed model used by RSEM for the RNA-seq read generation process compared to the read aligner programs. The alignment step can also be performed using any alignment program other than Bowtie.

2.3.1. Gene and isoform quantification guided by a transcriptome

RSEM aligns short RNA-seq reads to a set of transcript sequences. There are several advantages to this approach: The alignment of RNA-seq reads directly to a reference genome is complicated due to the splicing and polyadenylation events. For example, reads that span splice junctions or extend into poly(A) tails are challenging to align at the genome level. Therefore, this approach allows for a faster alignment at the transcript-level, since the total length of all possible transcripts is often much smaller than the length of the reference genome. Lastly, using transcript-level alignments allows for analyses of samples from species without reference genome, since a decently characterized transcriptome can be achieved by methods such as RNA-Seq transcriptome assembly (Martin & Wang, 2011).

By itself, RNA-seq data allows the estimation of the relative expression level of isoforms within a sample. There are two natural measures of relative expression: *the fraction of transcripts* and *the fraction of nucleotides* of the transcriptome made up by a given gene or isoform (Li & Dewey, 2011). These quantities can be referred to as τ_i and ϑ_i for the transcript *i,* respectively. Therefore, at the isoform level these quantities are related by the following equations:

$$\vartheta_i = \frac{\tau_i l_i}{\sum_j \tau_j l_j}$$

$$\tau_i = \frac{\vartheta_i / l_i}{\sum_j \vartheta_j / l_j}$$

Where, l_i is the length of isoform *i* in nucleotides. At the gene level, expression is simply the sum of the expression of possible isoforms. The expression levels estimated from these quantities are called nucleotides per million (NPM) and transcripts per million (TPM), which are obtained by multiplying ϑ and τ by 10^6, respectively.

RSEM is capable of estimating the level of expression at both the gene and transcript levels. For isoform expression-level estimation, it infers the values of the

model parameters θ = [θ_0, θ_1,..., θ_M], where *M* is the number of isoforms. Under the assumption that reads are uniformly sampled from the transcriptome, these parameters correspond to relative expression levels. Therefore, ϑ_i can be estimated by $\frac{\theta_i}{1-\theta_0}$, where θ_i represents the probability that a fragment is derived from transcript *i*, and θ_0 represents the noise-transcript from which reads that have no alignments may be derived. RSEM then computes maximum likelihood (ML) abundance estimates using the Expectation-Maximization (EM) algorithm. This can be achieved by computing the ML values of the parameter θ. When the values of θ is estimated, they can be converted into τ_i:

$$\tau_i = \frac{\theta_i / l_i'}{\sum_{j \neq 0} \theta_j / l_j'}$$

Where, l_i' is the length of isoform *i* in nucleotides. The effective length can be thought of as the mean number of positions from which a fragment may start within the sequence of transcript *i*.

The TCGA bioinformatics group use RSEM in their pipeline for gene and transcript expression level estimation, therefore in order to minimize the variability and avoid software biases in expression analysis, I also used this tool for the expression analysis.

2.4. Differential expression analysis

Several methods have been developed for the differential expression analysis, including DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010), and NOISeq (Tarazona et al., 2011). Each of these methods is described below.

2.4.1. DESeq2

DESeq2 is a parametric statistical method based on a negative binomial model. The implemented DESeq2 analysis workflow consists of several steps: it begins with

normalization of the observed read counts in order to enable their comparison across samples (this step is explained in section 1.3.2). Then for each gene, DESeq2 estimates the amount of variability that can be expected on the measurements from biological replicates. As with any counting process, one would not expect the detected counts for a given gene to be exactly the same across all observations from a single condition. Therefore, the key question in differential expression analysis is whether the observed counts across the two evaluated conditions are similar enough to be derived from the same distribution (null hypothesis), or whether they are better explained by two separate ones (alternative hypothesis). Hence, based on the nature of RNA-seq data, the Poisson distribution was first proposed to model noise intrinsic to the counting process. However, later it was shown that although this method works well for technical replicates, it underestimates the variability across biological replicates. As a result, the negative binomial distribution was introduced:

$$K_{ij} = NB(\mu_{ij}, \delta_{ij}^2)$$
$$\mu_{ij} = s_j q_{ij}$$

Where:

K_{ij} = observed counts for gene *i* in sample *j*

μ_{ij} = distribution mean for gene *i* in sample *j*

δ_{ij}^2 = dispersion for gene *i* in sample *j*

s_j = size factor for sample *j*

q_{ij} = quantity proportional to the concentration of cDNA fragments for gene *i* in sample *j*

In order to evaluate the significance of detected changes in the data, it is important to correctly identify the amount of variation across biological replicates.

However, due to the low number of replicates that are typically available for RNA-seq experiments, such variation cannot be directly calculated. Instead, it needs to be estimated from the data. DESeq2 uses the assumption that genes with similar level of expression have similar sample-to-sample variance, and therefore obtains gene-specific variance estimates by taking into account not only the observed dispersion for each given gene, but also that of all other genes. This goal is achieved by fitting a regression curve to the data, which is average normalised counts vs. observed dispersion. Then, the outcome is used to modify the observed dispersion values. DESeq2 further decomposes the mean into a function of independent variables (covariates), and therefore take all known sources of variation into account:

$$\log_2(\mu_{ij}) = \sum_r x_{jr}\,\beta_{ir}$$

where:

μ_{ij} = mean for gene *i* in sample *j*

x_{jr} = independent variable *r* in sample *j*

β_{ir} = coefficient for gene *i* and variable *r*

Therefore, in summary, the DESeq2 algorithmic approach fits the model mentioned in the above equations for both the null and alternative hypotheses, and evaluates the significance of coefficient of interest.

DESeq2 can be used for the study of differential expression at both gene and transcript level. DESeq2 is available as a bioconductor R package (Love et al., 2014).

2.4.2. EdgeR

Similar to DESeq2, edgeR is a parametric statistics method that works based on a negative binomial model. edgeR is the first statistical method developed for the differential expression analysis. It was originally developed for the analysis of Serial Analysis of Gene Expression (SAGE) data. The model formulation is very similar to that of DESeq2. When estimating variances, DESeq2 and edgeR both borrow information between genes but in different ways. edgeR uses conditional maximum likelihood conditioning on the total count for a gene to estimate the gene-wise variance or dispersion. Then using an empirical Bayes procedure, it shrinks the dispersions towards a consensus value. edgeR assesses differential expression using an exact test similar to that of DESeq2 with modification for over-dispersed data for each gene. It requires each condition to have at least one replicate for input data. One of the main differences between DESeq2 and edgeR is the way they estimate variance. edgeR estimates a single common dispersion parameter for all genes, whereas DESeq2 estimates the variance using a more flexible, mean-dependent local regression. edgeR is available as a bioconductor R package (Robinson et al., 2010).

2.4.3. NOISeq

In contrast to DESeq2 and edgeR, NOISeq is a non-parametric statistical method. NOISeq offers several normalization methods for the raw read counts, including: RPKM, FPKM, Trimmed Mean of M-values (TMM), and upper quartile (UQUA). TMM calculates a normalization factor based on a weighted average expression ratio of all genes after removing extremely high and low counts data. UQUA calculates scaling factors based on per-lane upper-quartile (75th percentile) of all the gene counts excluding those that have zero counts for all lanes. Once the data is normalized, NOISeq calculates the log-ratio (*M*) and the absolute value (*D*) of difference. If x_g^i is the mean or median of gene *i* expression for all available replicates in experimental condition *g* (where *g* can be one of two experimental conditions), the *M* and *D* values for gene *i* are:

$$M_i = \log_2 {x_1^i}/{x_2^i}$$

$$D_i = |x_1^i - x_2^i|$$

The value *M* collects fold-change information, while the value *D* collects the absolute difference that compensates for the unstable behaviour of *M* at low expression values. NOISeq then calculates the probability of a gene being differentially expressed which is the probability that |*M*| and *D* are greater that the noise |M^{noise}| and D^{noise}.

$$P(|M^{noise}| < |M^i|, D^{noise} < D^i)$$

NOISeq empirically computes the probability distributions of M^{noise} and D^{noise} by comparing gene expression counts between each pair of replicates within the same condition. Therefore, the odds of gene *i* being differentially expressed to non-differentially expressed is calculated as:

$$P_{noiseq} = {P(|M^{noise}| < |M^i|, D^{noise} < D^i)}/{1 - P(|M^{noise}| < |M^i|, D^{noise} < D^i)}$$

In the case that no replicates are available, NOISeq simulates replicates based on a multinomial distribution for read counts with the following parameters:

N = the number of replicates to be simulated

Pnr = the number of the total reads for each replicate to be simulated expressed in a percentage of the total reads of the available sample

V = the variability in the total read numbers of the simulated samples

NOISeq is available as a bioconductor R package (Tarazona et al., 2011).

2.5. *De novo* transcriptome assembly

A comprehensive study of the transcriptome includes identifying novel transcripts from unannotated genes, splicing isoforms and gene-fusion transcripts. Recent advances in the sequencing of the whole transcriptomes using next-generation sequencing technologies enable the study of the complex and dynamic landscape of the human transcriptome at an unprecedented level of sensitivity and accuracy. RNA-seq technology is capable of capturing RNA at base-pair-level resolution and a much higher dynamic range of expression levels than previous hybridisation methods. It is also capable of *de novo* annotation. *De novo* transcriptome assembly is a method of creating a transcriptome without the aid of a reference genome. Therefore, reconstructing the full-length transcripts from billions of short RNA-seq reads (35–500 bp) by transcriptome assembly poses a significant informatics challenge (Martin & Wang, 2011). One of the challenges of transcriptome assembly is the presence of multiple transcript variants from the same gene, which can share exons and are difficult to resolve unambiguously with short reads. Second, unlike genomic sequencing, in which both strands are sequenced, RNA-seq can be strand-specific. Therefore, assemblers should be designed to use strand information to resolve overlapping sense and antisense transcripts. Lastly, the sequencing depth of RNA-seq can vary by several orders of magnitude. This is opposite to the DNA sequencing, where the sampling depth is expected to be similar across the genome. Although high sequence coverage for a genome may indicate the presence of repetitive sequences and thus be masked, in RNA-seq experiments it typically represents abundant genes. Several transcriptome assemblers (e.g. Trans-ABySS) have been developed in the past few years (Robertson et al., 2010). Depending on whether a reference genome assembly is available, current transcriptome assembly strategies fall into one of the following categories: a reference-based strategy, a *de novo* strategy or a combined strategy that merges the two *de novo* and reference based strategies (Martin & Wang, 2011).

When a reference genome is available, the transcriptome assembly can be built upon it by following three steps: (1) short RNA-seq reads are aligned to a reference genome using a splice-aware aligner, (2) overlapping reads from each locus are clustered to build a graph representing all possible isoforms, and (3) in order to resolve

individual isoforms, the assembler traverses the graph built on step 2. Therefore, this strategy is known as reference-based or *ab-initio* assembly (Martin & Wang, 2011). This approach has several advantages. It allows parallel processing of the data since it breaks a large assembly problem into many smaller assembly problems i.e. independent assemblies across each locus. In addition, contamination or sequencing artefacts are not confounding since they are not expected to align to the reference genome and therefore will be filtered out from further analysis. The most important advantage of this approach is the ability to assemble transcripts of low abundance. Because the underlying genome sequence is known, small gaps within the transcript that have been caused by a lack of read coverage can be filled in using the reference sequence. Therefore, the reference-based approach allows for discovery of novel transcripts as in general such transcripts have lower expression levels. However, this approach does not allow identification of trans-spliced and fusion genes, and relies on correct identification and alignment of the reads to splice junctions by the aligner.

The *de novo* transcriptome assembly strategy does not use a reference genome. Instead, it leverages the redundancy of short-read sequencing to find overlaps between the reads and assembles them into transcripts (Martin & Wang, 2011). The *de novo* assemblers generally assemble the data set multiple times using a De Bruijn graph-based approach. Then, they post-process the assembly to merge contigs and remove redundancy. *De novo* assemblers then traverse the De Bruijn graph by applying paired-end read information to assemble isoforms at each locus. Compared to the reference-based strategy, *de novo* transcriptome assembly has several advantages. First, it does not depend on a reference genome. Second, it does not depend on the correct alignment of reads to known splice sites or the prediction of novel splicing sites, as required by reference-based assemblers. Finally, a *de novo* approach is able to assemble and identify trans-spliced transcripts and similar transcripts originating from chromosomal rearrangements. Disadvantages of this approach include the sensitivity of *de novo* assemblers to sequencing errors and to the presence of chimeric molecules. Even though, algorithms have been developed to correct error containing reads from abundant transcripts, this distinction is more difficult to make for reads that are sequenced from low-abundance transcripts. Furthermore, *de novo* assemblers are likely to assemble highly similar transcripts (e.g. different alleles or paralogues) into a single

transcript. Therefore, the outcome will require additional post-assembly steps to resolve this.

The two approaches mentioned above can be combined to create a more comprehensive method for transcriptome analysis. The combined transcriptome assembly approach therefore brings together the advantages of the two previous approaches allowing for detection of novel and trans-spliced transcripts by *de novo* assembly while leveraging the high sensitivity of reference-based assemblers. The combined method can be carried out by either first aligning the reads to the reference genome or by *de novo* assembling the reads. Trans-ABySS, which is being used in this thesis, is an example of combined transcriptome assembly method (Robertson et al., 2010). It assembles the RNA-seq data set using the reference genome, and then performs *de novo* assembly on the reads that failed to align to the genome. In addition, transcripts that result from the reference-based assembly could also serve as input to the *de novo* assembly. The combined approach therefore requires less computational time and resources in comparison to the *de novo* approach as with a reference, most of the reads will be assembled, leaving only a small fraction of the reads to be *de novo* assembled. The main advantage of this approach is the ability to merge incomplete transcripts by aligning both the assembled transcripts and the unassembled reads to the reference genome.

2.5.1. Trans-ABySS *de novo* assembly package

The Trans-ABySS package includes ABySS (Simpson et al., 2009), a genome assembler tool, and Trans-ABySS itself for the post-processing of assembly outcomes. ABySS algorithm works based on de Bruijn di-graph representation of sequence neighbourhoods. In this graph, a sequence read is decomposed into tiled sub-reads of length k (also called k-mers) and sequences sharing k−1 bases are connected by directed edges. Therefore, it captures the adjacency information between sequences that overlap the last and the first k−1 characters. Once ABySS establishes adjacency information by cataloging k-mers in a given set of reads, the resulting graph is inspected to identify potential sequencing errors and small-scale sequence variation. If a sequence has a read error, it alters the k-mers that span it. Therefore, it results in the formation of

branches in the graph. However, since such errors are stochastic in nature, their rate of observation is substantially lower than that of correct sequences. Therefore, using the coverage information, such errors can be discerned in order to improve the quality and contiguity of an assembly. However, this fact is especially true for genomic sequences. In the case of RNA-seq, where sequence coverage depth is a function of the transcript expression level, the removal of low coverage branches needs to be performed with care. ABySS trims such low coverage branches in RNA-seq data, when the absolute coverage levels falls below a threshold of 2-fold.

After removing false branches, the unambiguously linear paths along the de Bruijn graph are connected to form contigs. Contigs are contiguous sequences that are used to indicate a contiguous piece of DNA/RNA assembled from shorter overlapping sequence reads. Therefore, contigs formed in this stage consist of uniquely occurring k-mers. Next a streamlined read-to-assembly alignment routine is performed by Trans-ABySS. In this step, the aligned read pairs are used to infer read distance distributions between pairs in the RNA-seq sample, and identify contigs that are in a certain neighbourhood defined by these distributions. The adjacency and the neighbourhood information are used to further merge contigs unambiguously connected by read pairs.

In practice, ABySS assembles the short read RNA-seq data set multiple times using a De Bruijn graph-based approach for different values of K-mer. This approach reconstructs transcripts from a broad range of expression levels. Trans-ABySS post-processes the assembly to merge contigs and remove redundancy.

The quality of a *de novo* assembly can be evaluated by the length of the shortest as well as the longest contigs and the N50 value - which is the size at which half of all assembled bases reside in contigs of this size or longer. A larger N50 is more desirable.

2.6. Downstream analysis

2.6.1. Pathway and enrichment analysis

The analysis of high-throughput sequencing data typically yields a long list of differentially expressed genes or proteins. Such lists are useful in identifying genes that may play a role in a given phenomenon or phenotype. However, extracting the meaning of a long list of differentially expressed genes and proteins and understanding the underlying biology of the condition presents a challenge in bioinformatics. One approach to overcome this challenge is to group long lists of individual genes into smaller sets of related genes or proteins, which can significantly reduce the complexity of analysis. Knowledge base approaches can help with this task. They can describe biological processes and components in which individual genes and proteins are known to be involved in, as well as how and where gene products interact with each other. One example of this idea is to identify groups of genes that function in the same pathways. Pathway analysis allows researchers to determine up and down regulated genes/proteins, expression changes in a network overall, infer upstream regulators, downstream molecules, and associated diseases. Ingenuity Pathway Analysis (IPA) is a commercial software package that performs such analysis(Kramer, Green et al., 2014)

Data analysis and interpretation with IPA is built on the comprehensive, manually curated content of the Ingenuity Knowledge Base database. Given a gene-expression dataset, IPA elucidates the upstream biological causes and probable downstream effects on cellular and organismal biology. In addition, it performs functional enrichment and pathways/networks analysis (Kramer et al., 2014). IPA follows a network-based approach based on the knowledge derived from the Ingenuity Knowledge Base, where the nodes are genes, chemicals, protein families, complexes, microRNA species and biological processes, and edges are observed cause-effect relationships. Each edge is associated with a set of underlying findings obtained from the literature, and marked with the regulation direction of the effect. IPA constructs many possible networks from the data serving as hypotheses for the biological mechanism underlying the data. It constructs networks that optimize for both interconnectivity and number of focus genes

(uploaded genes that pass initial filters such as fold change, expression level, or significance level) under the constraint of a maximal network size. IPA then scores these networks by their statistical significance. It uses two scores that address two independent aspects of the inference problem. An enrichment score, which is Fisher's exact test P-value, that measures overlap of observed and predicted regulated gene sets. In addition, it uses Z-score to assess the match of observed and predicted up/down regulation patterns. The Z-score is suited for this kind of problem since it serves as both a significance measure and a predictor for the activation state of the regulator. Therefore, IPA calculates significance scores based on the number of genes/molecules that map to a biological function, pathway, or network. Once the IPA network is created, it can determine information such as over represented canonical pathways, as well as up-stream and down-stream regulators. IPA is available at http://www.ingenuity.com/products/ipa.

2.7. Statistical analysis

Survival analysis was performed using the Kaplan-Meier survival curve approach, and differences in overall survival rates were determined by the log-ranked test (Goel, Khanna, & Kishore, 2010). Overall survival time was defined as the period between initial pathologic diagnosis and the time of death. Survival time of patients who were still alive was noted with the data of the most recent follow-up appointment.

The Fisher's exact test was used to compare two categorical variables. The Mann-Whitney t-test was used to determine significant differences in gene expression between groups. A statistically significant P-value was defined as P-value ≤ 0.05. I have used R for statistical and survival analysis.

Chapter 3. Results

3.1. Pan-cancer identification of cancer-associated differentially expressed genes

Cancer is fundamentally a disease of disordered gene expression (Pelengaris & Khan, 2013). There are number of mechanisms that can alter gene expression patterns in cancer cells (Holliday & Jeggo, 1985). These mechanisms may occur via a direct change to DNA sequence, such as mutations within genes or closely linked DNA that regulates activity of those genes, deletions that remove various genes and gene regulatory sequences, amplification of genomic regions containing various genes, and fusions of two genes by recombination between DNA sequences. Furthermore, gene expression can be affected by perturbations in the machinery responsible for production or activity of proteins, such as gene splicing. Splicing is a regulatory mechanism by which variations in the incorporation of exons, or coding regions, into mRNA leads to the production of alternate proteins, or isoforms.

Overall, aberrant gene expression ultimately results in an imbalance of cell replication and cell death in a cell population that leads to formation and expansion of tumour tissue. Differentially expressed genes are particularly relevant in oncology since they may contribute to the etiology of cancer, may provide selective drug targets and can serve as a marker set for cancer diagnosis (Goodison, Sun, & Urquidi, 2010). Therefore, measuring gene expression is of great interest to scientists and many gene expression measurement methods have been developed for biomedical studies. EGFR is one of the differentially expressed genes that is currently in clinic as a drug target. Cetuximab is a monoclonal antibody that targets EGFR and is indicated for the treatment of patients with colorectal- and head and neck- EGFR expressing cancers (Wong, 2005).

Diagnostic markers such as PSA (prostate-specific antigen) and CA125 (also known as MUC16) are also being clinically used for the detection of prostate and ovarian cancers, respectively (Felder et al., 2014; Schroder, 2009). However, the effectiveness of detection is compromised with a high false positive rate.

The availability of large datasets such as TCGA provides an unprecedented opportunity to comprehensively mine such datasets for genes that are being differentially expressed within one cancer or amongst several different cancer types. In this chapter, I describe the identification of tumour marker genes, along with their associated pathways that are either common to multiple types of cancer or specific to individual cancer types. I studied RNA-seq data from 24 cancer types available from the TCGA as well as RNA-seq from a number of non-cancerous normal tissues generated by the GTEx project. I further examined these genes to identify those that may serve as suitable tumour biomarkers for targeting with antibody therapeutics in cancer.

3.1.1. Gene expression analysis pipeline

I developed a count-based gene expression analysis (GEA) pipeline for the analysis of TCGA RNA-seq data. This pipeline, shown in Figure 3-1, is built from a set of tools that are available to public and demonstrated better performance compared to their counterparts (Conesa et al., 2016; Seyednasrollah, Laiho, & Elo, 2015; Yang & Smith, 2013). The GEA pipeline consists of four major steps including: a sequence quality check, alignment of short RNA-seq reads to reference and coverage analysis, differential expression analysis, pathway analysis and subsequent post-processing.

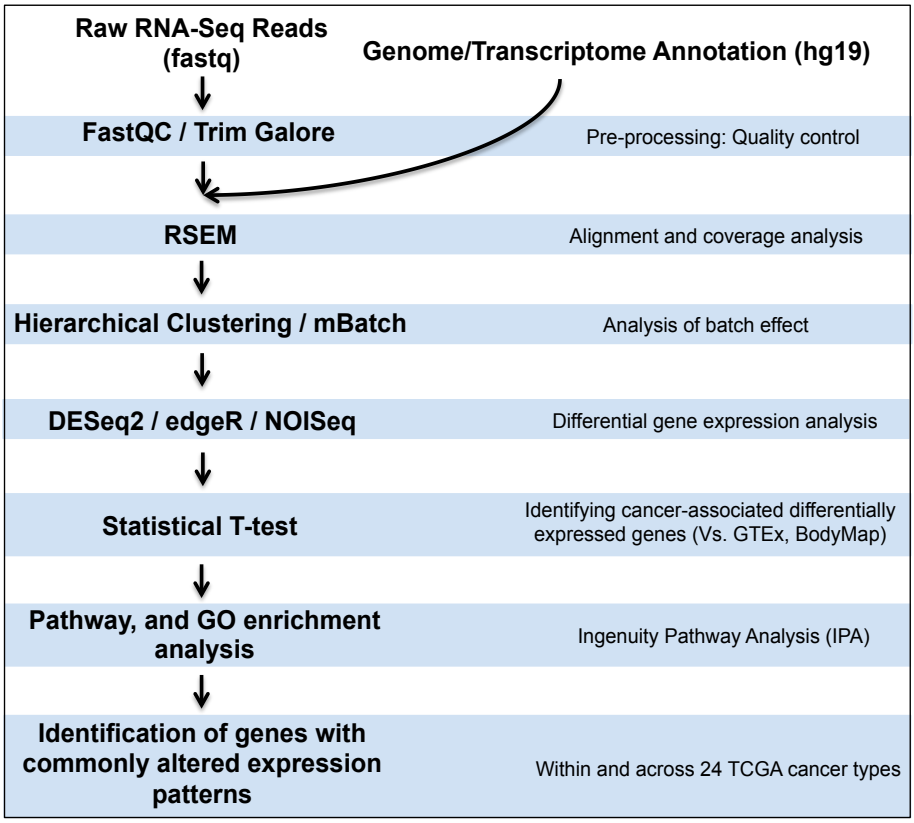


Figure 3-1. Gene Expression Analysis (GEA) pipeline

Quality Control

The GEA pipeline accepts raw RNA-seq reads in FASTQ format, a reference genome, and the corresponding transcriptome annotation as input. In the first step, Trim Galore is used to perform RNA-seq quality assessment and improvement by trimming the adaptor and over-represented sequences as well as low quality bases (Babraham Bioinformatics, 2015). This step improves the mappability of the RNA-seq reads to the reference and therefore allows for more accurate coverage analysis. Using FastQC (Andrews, 2016), Trim Galore estimates the sample's duplication rate. Samples with over 50% duplication rate are discarded from the rest of analysis.

Read alignment and coverage analysis

Raw RNA-seq reads are then mapped to the reference transcriptome using RSEM (version 1.2.20). RSEM uses the genomic aligner Bowtie to perform read mapping to the reference transcriptome (Langmead et al., 2009). The default parameters suggested by RSEM is used for this step. Then, the quality of the alignments is assessed by examining the percentage of the reads that are successfully mapped to the reference. This is done using samtools flagstst (H. Li et al., 2009). A threshold of greater than or equal to 70% is used to select samples that were successfully mapped to the reference transcriptome. Those with less than 70% mapped reads were discarded from the rest of the analysis. Once the read mapping is complete, RSEM reports the gene coverage in form of raw read counts as well as FPKM and TPM measures. The raw read counts can further be used for the differential gene expression analysis.

Batch effect and hierarchical clustering

A large project such as TCGA collects and sequences tumour samples in batches at different times. Since the samples are processed in batches, rather than all at once, the data can be vulnerable to systematic noise such as batch effects (unwanted variation between batches) and trend effects (unwanted variation over time), which can lead to misleading analysis results. Therefore, Hierarchical clustering was performed through the mBatch online tool (http://bioinformatics.mdanderson.org/tcgambatch/), which helps to assess, diagnose and correct for any batch effects in TCGA data. For

example, Figure 3-2 shows the hierarchical clustering of 479 lung squamous cell carcinoma samples. No major batch effect was observed in this dataset.

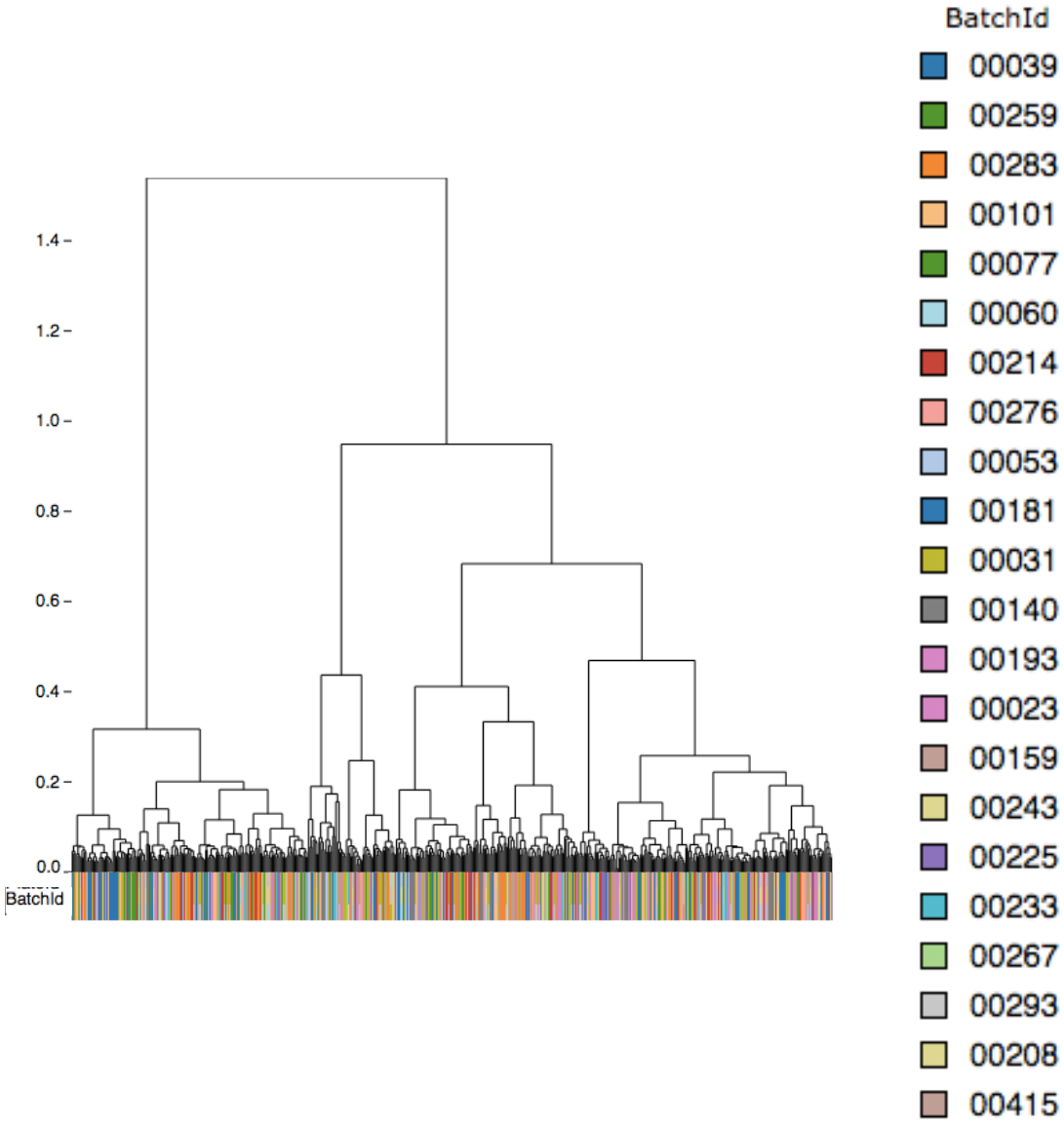


Figure 3-2. Hierarchial clustering of Lung squamous cell carcinoma (LUSC) RNA-seq data using mBatch version 1.2 (http://bioinformatics.mdanderson.org/tcgambatch/)

Differential expression analysis

The GEA pipeline incorporates three differential expression analysis methods, including DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010), and NOISeq (Tarazona et al., 2011). These methods use different normalization and statistical approaches, which leads to the identification of different sets of genes as being differentially expressed. Although the majority of differentially expressed genes are found by all three approaches, there are small subsets of genes that are only found by an individual method. In order to prevent information loss, the union of differentially expressed genes found by the three approaches are being reported by GEA pipeline. Although genes found by more than one method are being marked as more confident. A gene is considered to be differentially expressed if the P-value and false discovery rate (FDR) reported by DESeq2 and/or edgeR are less than or equal to 0.05. Similarly, a gene is found as differentially expressed by NOISeq if the differential expression probability is less than or equal to 0.8. In addition, a fold change of greater than 1.5 for overexpressed genes and smaller than -1.5 for under-expressed ones must be observed. The reported list of differentially expressed genes is further trimmed using the level of gene expression in tumour and matched-normal samples. A gene is considered to be overexpressed in tumour compared to the matched normal libraries, if it is expressed higher that 0.1 FPKM in at least 25% of tumour samples. Similarly, if a gene is under-expressed in tumour samples, it must be expressed in at least 25% of the available matched-normal libraries with a FPKM value of equal or greater than 0.1.

The described approach above works well when tumours have paired control samples from the same group of patients. However, there are several TCGA tumour types that have no matched-normal. In these cases, non-cancerous normal tissues were obtained from GTEx dataset (Table 3-1). Therefore, for datasets with unpaired cancer and control samples, Mann-Whitney test was applied on the normalized expression values (FPKM) to identify genes that are differentially expressed in cancer *versus* control samples.

Table 3-1. GTEx tissue types

Tissue type	Number of samples
Adipose Tissue	10
Adrenal Gland	10
Bladder	11
Blood	11
Blood Vessel	10
Bone Marrow	10
Brain	25
Breast	25
Cervix Uteri	20
Colon	25
Esophagus	10
Fallopian Tube	10
Heart	11
Kidney	12
Liver	12
Lung	25
Muscle	10
Nerve	10
Ovary	11
Pancreas	11
Pituitary	10
Prostate	16
Salivary Gland	10
Skin	10
Small Intestine	15
Stomach	10
Testis	10
Thyroid	20
Uterus	10
Vagina	10

Downstream analysis

The downstream analysis assesses the list of differentially expressed genes for each cancer type individually and as a whole to identify frequently occurring events. This analysis includes identifying commonly enriched pathways across different malignancies, potential transcription factors regulating the gene expression and their target genes, commonly over- and under expressed genes across multiple cancer types and potential tumour markers. Such analyses are performed using Ingenuity software (Kramer et al., 2014).

3.1.2. Identification of differentially expressed genes within and across multiple cancer types

The majority of cancers undergo a common set of alterations during oncogenesis, such as self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of apoptosis, and tissue invasion and metastasis (Hanahan & Weinberg, 2011). Since the same group of proteins may execute some of these biological processes during the formation and progression of different cancers, I hypothesize that it is possible to find common genes with disrupted expression patterns across different cancer types. In addition, the availability of large datasets such as TCGA allow for identification of genes with commonly altered expression patterns across different cancer types. Therefore, in this section, I use the GEA pipeline introduced in section 3.1.1 to analyze large batches of RNA-seq data from TCGA and GTEx to find genes differentially expressed in cancers. Next, I present an exploratory analysis on these genes to identify commonly enriched pathways across multiple malignancies as well as transcription factors that may play a role in regulating the observed gene expression patterns. Such analysis brings the gene differential expression information together to explain the underlying mechanisms of the disease, and hence would help to highlight the key players of those mechanisms. Such genes could be attractive therapeutic targets. Then, I bring the differentially expressed genes in each tumour type together to find those that are commonly found in cancers. Such genes, if expressed on the cell surface could potentially be used as global tumour cell markers for targeted therapeutic avenues such as antibodies (mAbs, ADCs, and bi-specific). Finally, I use the GTEx dataset to find cancer-associated events. Differentially expressed genes with high tumour expression and zero to low level of expression in non-cancerous tissues are the most attractive candidate genes for antibody-based therapeutics. GTEx includes RNA-seq samples from 30 tissue types of human body, which are obtained from individuals free of major diseases including cancer. Hence, GTEx is an important resource for studying gene expression patterns in normal human tissues.

As of December 2014, there are RNA-seq from 24 types of malignancies available at TCGA data repository (Table 3-2), of which only 17 cancer types have available matched-normal samples. Raw RNA-seq data from 24 cancer types and 30

normal non-cancerous tissue types were downloaded from TCGA and GTEx data repository, respectively. At least 10 samples were downloaded for each GTEx tissue type (Table 3-1). The raw RNA-seq reads were run trough the GEA pipeline for data quality assessment and gene coverage analysis with RSEM.

Table 3-2. Cancer RNAseq datasets used for pan-cancer identification of differentially expressed genes

ID	Type	Tumour sample	Matched normal
ACC	Adrenocortical carcinoma	79	N/A
AML	Acute Myeloid Leukemia	158	N/A
BLCA	Bladder Urothelial Carcinoma	220	19
BRCA	Invasive Breast carcinoma	1010	96
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	167	3
COAD	Colon adenocarcinoma	451	41
GBM	Glioblastoma multiforme	145	N/A
HNSC	Head and Neck squamous cell carcinoma	422	42
KICH	Kidney Chromophobe	66	25
KIRC	Kidney renal clear cell carcinoma	506	72
KIRP	Kidney renal papillary cell carcinoma	171	30
LGG	Brain Lower Grade Glioma	452	N/A
LIHC	Liver hepatocellular carcinoma	189	50
LUAD	Lung adenocarcinoma	485	58
LUSC	Lung squamous cell carcinoma	479	50
OV	Ovarian serous cystadenocarcinoma	253	N/A
PAAD	Pancreatic adenocarcinoma	75	4
PRAD	Prostate adenocarcinoma	272	49

ID	Type	Tumour sample	Matched normal
READ	Rectum adenocarcinoma	160	9
SARC	Sarcoma	102	2
SKCM	Skin Cutaneous Melanoma	82	N/A
THCA	Thyroid carcinoma	493	57
UCEC	Uterine Corpus Endometrial Carcinoma	512	35
UCS	Uterine Carcinosarcoma	57	N/A

Pathway enrichment analysis of differentially expressed genes

Pathway enrichment analysis has been applied to cancer data sets to find driver genes and pathways, to identify cancer mechanisms and biomarkers, and to identify key regulators of the disease (Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium, 2015). Since some cancer genes cannot be targeted directly or it is different components of a pathway that alter during the disease progression, studying the pathways would reveal other potential key players as well as mechanism that can be targeted for cancer therapy.

Following the GEA pipeline, differential gene expression analysis was performed on TCGA tumours *versus* either the available TCGA paired matched-normal tissue samples or unpaired normal samples from GTEx. Genes with at least 1.5-fold changes and FDR of equal or smaller than 0.05 were identified as differentially expressed. Pathway enrichment analysis on genes that are differentially expressed in any of the 17 TCGA cancer types with matched normal samples revealed a number of signalling pathways that are consistently and highly enriched across all TCGA tumour types (P-value and FDR <= 0.05). These pathways are shown in Table 3-3. Estrogen-mediated S-phase entry, coagulation system, MIF-mediated glucocorticoid regulation, acute phase response signalling, GABA receptor signalling, Wnt/B-catenin signalling, p38 MAPK signalling, cAMP mediated signalling, and chemokine signalling (in addition to the general cellular processes such as cell cycle, DNA replication and repair, and apoptosis) are among the most commonly enriched signalling pathways within different cancer types. Some of these enriched pathways may also arise from the presence of non-cancerous cells within the sampled tumour microenvironment. This analysis was performed using IPA (Kramer et al., 2014).

Table 3-3. Top 50 commonly enriched pathways across TCGA cancer types with matched-normal tissue. The analysis is performed on differentially expressed genes in each cancer type separately using the IPA software.

Pathway	#Tumor	Tumor*
Eicosanoid Signaling	17	CESC,BLCA,PAAD,SARC,BRCA,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
Bladder Cancer Signaling	17	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,SARC,THCA,UCEC
Coagulation System	17	PAAD,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,SARC,THCA,UCEC
FXR/RXR Activation	17	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,SARC,THCA,UCEC
Agranulocyte Adhesion and Diapedesis	17	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,SARC,THCA,UCEC
Granulocyte Adhesion and Diapedesis	17	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,SARC,THCA,UCEC
LXR/RXR Activation	17	BLCA,BRCA,COAD,HNSC,KIRC,KIRP,LICH,LUAD,PAAD,PRAD,READ,THCA,UCEC,CESC,KICH,LUSC,SARC
Glutamate Receptor Signaling	16	KIRC,PAAD,SARC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRP,LICH,LUAD,LUSC,READ,THCA,UCEC
nNOS Signaling	16	LICH,PAAD,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,PRAD,READ,THCA,UCEC
cAMP-mediated signaling	16	KIRC,PAAD,BRCA,LICH,LUSC,BLCA,CESC,COAD,HNSC,KICH,KIRP,LUAD,PRAD,READ,UCEC,THCA
VDR/RXR Activation	16	BRCA,CESC,UCEC,KICH,SARC,PAAD,PRAD,BLCA,COAD,HNSC,KIRC,KIRP,LUAD,LUSC,READ,THCA
Atherosclerosis Signaling	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,THCA,UCEC
Axonal Guidance Signaling	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,THCA,UCEC
Transcriptional Regulatory Network in Embryonic Stem Cells	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,PRAD,READ,THCA,UCEC
MIF-mediated Glucocorticoid Regulation	15	LICH,PAAD,SARC,BRCA,CESC,COAD,KICH,KIRC,KIRP,LUAD,LUSC,PRAD,READ,THCA,UCEC
Basal Cell Carcinoma Signaling	15	LUSC,KICH,BLCA,BRCA,CESC,COAD,HNSC,KIRC,KIRP,LICH,LUAD,PRAD,READ,THCA,UCEC
Acute Phase Response Signaling	15	COAD,KIRC,KIRP,BLCA,BRCA,HNSC,KICH,LICH,LUAD,LUSC,PRAD,READ,SARC,THCA,UCEC
Leukocyte Extravasation Signaling	15	KIRC,KIRP,THCA,BLCA,PAAD,BRCA,CESC,COAD,HNSC,KICH,LUAD,LUSC,READ,SARC,UCEC
Amyotrophic Lateral	15	BLCA,BRCA,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUA

Pathway	#Tumor	Tumor*
Sclerosis Signaling		D,LUSC,PAAD,PRAD,READ,THCA,UCEC
Differential Regulation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F	15	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PAAD,READ,THCA,UCEC
Hepatic Fibrosis / Hepatic Stellate Cell Activation	15	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
Role of Cytokines in Mediating Communication between Immune Cells	15	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
Dopamine-DARPP32 Feedback in cAMP Signaling	15	COAD,LUAD,READ,UCEC,PAAD,BLCA,BRCA,HNSC,KICH,KIRC,KIRP,LICH,LUSC,PRAD,THCA
Estrogen-mediated S-phase Entry	14	BLCA,BRCA,KIRC,KIRP,LICH,LUAD,LUSC,UCEC,THCA,CESC,COAD,KICH,READ,SARC
Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL-17A and IL-17F	14	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,PAAD,READ,THCA,UCEC
Human Embryonic Stem Cell Pluripotency	14	BLCA,BRCA,CESC,COAD,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
MIF Regulation of Innate Immunity	14	PAAD,SARC,BRCA,CESC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
Role of IL-17A in Psoriasis	14	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,READ,THCA,UCEC
TR/RXR Activation	14	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,PRAD,READ,THCA,UCEC
eNOS Signaling	14	BLCA,COAD,READ,PAAD,SARC,CESC,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,PRAD,UCEC
Corticotropin Releasing Hormone Signaling	13	LICH,BLCA,LUAD,READ,UCEC,BRCA,CESC,COAD,KICH,KIRP,LUSC,PRAD,THCA
p38 MAPK Signaling	13	THCA,UCEC,PAAD,BLCA,BRCA,CESC,COAD,HNSC,LICH,LUAD,LUSC,PRAD,READ
GABA Receptor Signaling	13	LUAD,LUSC,THCA,UCEC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH
Wnt/catenin Signaling	13	BLCA,BRCA,CESC,COAD,KICH,KIRP,LICH,LUAD,LUSC,PRAD,READ,THCA,UCEC
Protein Kinase A Signaling	13	BLCA,COAD,READ,UCEC,BRCA,CESC,HNSC,KIRC,KIRP,LUAD,LUSC,PAAD,PRAD
Endothelin-1 Signaling	13	BLCA,CESC,COAD,KICH,KIRC,KIRP,LICH,LUSC,PAAD,PRAD,READ,THCA,UCEC
PCP pathway	12	LUSC,BLCA,BRCA,CESC,COAD,KIRP,LICH,LUAD,PRAD,READ,THCA,UCEC

Pathway	#Tumor	Tumor*
Complement System	12	KIRC,LUAD,THCA,UCEC,PAAD,BLCA,BRCA,KICH,KIRP,LICH,LUSC,READ
Antioxidant Action of Vitamin C	12	COAD,CESC,SARC,KICH,KIRC,KIRP,LUSC,PAAD,PRAD,READ,THCA,UCEC
Altered T Cell and B Cell Signaling	12	BLCA,BRCA,CESC,COAD,HNSC,KIRC,LUAD,LUSC,PRAD,READ,THCA,UCEC
Embryonic Stem Cell Differentiation into Cardiac Lineages	12	BLCA,BRCA,COAD,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,THCA,UCEC
Hematopoiesis from Multipotent Stem Cells	12	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUSC,READ,THCA
MSP-RON Signaling Pathway	12	BLCA,BRCA,CESC,HNSC,KICH,KIRC,KIRP,LICH,LUSC,PAAD,PRAD,SARC
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	12	BLCA,UCEC,PRAD,BRCA,CESC,COAD,KICH,KIRP,LICH,LUAD,LUSC,SARC
Colorectal Cancer Metastasis Signaling	11	LICH,PAAD,BLCA,BRCA,CESC,COAD,LUAD,LUSC,READ,THCA,UCEC
Cell Cycle Control of Chromosomal Replication	10	BLCA,BRCA,CESC,HNSC,KIRP,LICH,LUAD,LUSC,SARC,UCEC
Chemokine Signaling	9	PRAD,PAAD,BLCA,BRCA,COAD,HNSC,KICH,KIRP,THCA
CREB Signaling	8	COAD,READ,PAAD,BLCA,CESC,KICH,KIRP,UCEC
ERK/MAPK Signaling	7	BLCA,BRCA,KIRP,PAAD,PRAD,READ,UCEC
FGF Signaling	7	BLCA,COAD,KIRP,PRAD,READ,THCA,UCEC

* The expanded form of each tumour type abbreviation is available in Table 3-2.

The pathway analysis revealed common mechanisms shared by different types of cancer. For example, the GABA receptor signalling is enriched in 13 cancer types including BLCA, BRCA, CESC, COAD, HNSC, KICH, KIRC, KIRP, LICH, LUAD, LUSC, THCA, and UCEC. The γ-amino butyric acid (GABA) has been shown to control secretion in peripheral organs and acts as a developmental signal in both embryonic and adult developing or regenerating tissues (Young & Bordey, 2009). Through $GABA_A$ receptors, GABA affects every stage of cell development including proliferation, migration, and differentiation. In particular, it has been shown to control the proliferation of many different cell types including stem cells (Young & Bordey, 2009). Since the levels of $GABA_A$ receptors are frequently up-regulated in cancer cells, there is a possibility that manipulating $GABA_A$ receptor activity reduces tumour growth. With growing evidence implicating the existence and role of cancer stem cells in tumour generation and progression, genes involved in such pathways provide attractive therapeutic targets for manipulating the proliferation of cancer cells and perhaps cancer stem cells.

Acute phase response signaling is another highly enriched pathway among 15 cancer types (BLCA, BRCA, HNSC, KICH, LICH, LUAD, LUSC, PRAD, READ, SARC, THCA, UCEC, COAD, KIRC, and KIRP). The acute phase response is a rapid inflammatory response that provides protection against microorganisms using non-specific defense mechanisms (Davalieva et al., 2015). This pathway is associated with cancer, since inflammation is often observed in tumours and appears to play a role in the pathogenesis of various cancer types. Interestingly, majority of tumour types with enriched acute phase response pathway, also present enrichment in P38 Mitogen activated protein kinase (P38 MAPK), and extracellular-signal-regulated kinase (ERK)/MAPK pathways. P38 MAPK and ERK1/MAPK are members of the mitogen activated protein kinase super family and are involved in the production of inflammatory mediators, including tumour necrosis factor-α (TNF-α) and cyclooxygenase-2 (COX-2) (Dhillon, Hagan et al., 2007; Gui, Sun et al., 2012). Abnormal regulation of the MAPK pathways has also been reported in cancers (Dhillon et al., 2007).

The gene expression profile of TCGA cancers also revealed the enrichment of cAMP-mediated signalling pathway in 16 cancer types. The cAMP-mediated pathway,

also known as the adenylyl cyclase pathway, is a G protein-coupled receptor-triggered signalling cascade, which is used in cell communication. There are conflicting reports in literature that cAMP signalling can either activate or inhibit tumour growth. In thyroid papillary carcinoma cell lines, it has been reported that cAMP signalling acts in an inhibitory manner on the proliferation, even though the cAMP pathway physiologically promotes the proliferation of normal follicular cells as well as hormonogenesis (Matsumoto et al., 2008). One of the suggested mechanisms of cAMP growth inhibiting function is through the inhibition of the mitogen activated protein kinase (MAPK) pathway, a pathway important for both proliferation and differentiation. The MAPK pathway is usually initiated by tyrosine kinase receptor stimulation of small G-proteins (e.g, RAS) followed by the activation of several downstream protein kinases (RAF, MEK and ERK). Each RAF protein (ARAF, BRAF and CRAF) performs a different function, both physiologically and in cancer. For example, in normal melanocytes, BRAF but not CRAF transduces the signal from RAS to MEK because CRAF is inhibited by cAMP–dependent protein kinase A (PKA). This inhibitory cAMP pathway is often hijacked in melanoma. In RAS-mutated melanoma, CRAF rather than BRAF is utilized to activate MEK/ERK and this switch in RAF utilization is due to a disruption of cAMP signaling, most likely from the activation of PDE4. In other words, a loss of cAMP promotes melanoma growth in RAS-mutated melanoma. In contrast, multiple investigators have demonstrated that some melanomas favour elevated cAMP signalling. cAMP may play a role in promoting melanoma drug resistance. Activation of the AC-cAMP-PKA axis may confer resistance to MAPK inhibitors in melanoma; where the expression of transcription factors downstream of the MAPK and cAMP pathways (e.g. microphtalmia-associated transcription factor (MITF)) resulted in resistance. Treatment with a combination of MAPK-pathway and histone deacetylase inhibitors suppressed cAMP-mediated resistance and MITF expression. The reduction of MITF activity sensitized melanoma cells to chemotherapeutic agents. Therefore, the importance of cAMP signalling in promoting tumour growth suggests this pathway may be a therapeutic target for cancer (Nardin, Fitzpatrick, & Zippin, 2014).

The presence of common pathways among different cancer types supports the assumption that different types of cancer share similar process. Such processes can be summarized into the hallmarks of cancers, which are six biological capabilities acquired

during the multistep development of human tumours (Hanahan & Weinberg, 2011). They include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. The fact that these pathways are being enriched in multiple cancer types suggests that there is a selection for these processes as they benefit the formation and progression of the disease. Therefore, such common pathways may present attractive targets for development of therapeutic intervention.

Of the list of identified abnormally expressed genes, this step identified those that may play a role in the development and progression of cancer by identifying genes that are a member of well-studied cancer-promoting pathways. Here, I refer to them as potential cancer-related genes. This information was subsequently used to prioritize potential targets for therapeutic utility.

Identification of transcription factors and their target genes common to multiple cancer types

Analysis of upstream gene expression regulators (i.e. transcription factors) based on the list of differentially expressed genes for each TCGA cancer type revealed commonly occurring transcription factors across multiple human malignancies. Studying transcription factors involved in cancer help to illuminate the biological activities occurring during the tumour formation and progression. With the limited list of transcription factors activated in most human cancers and their effect on many target genes and processes, transcription factors are logical targets for the development of anticancer drugs. For example in glioblastoma, the transcriptional regulator Id-1 plays a critical role in modulating the invasiveness of glioblastoma cell lines and primary glioblastoma cells by regulating multiple tumor-promoting pathways (Soroceanu et al., 2013). The down-regualtion of Id-1 gene expression showed significant association with decrease in glioma cell invasiveness and self-renewal. Therefore, it is currently being pursued as a novel and promising target for improving the therapy and outcome of patients with glioblastoma.

Transcription factors potentially involved in TCGA tumours were predicted using the Ingenuity® platform (Kramer et al., 2014). The discovery analysis works based on

the available prior knowledge of expected effects between transcriptional regulators and their target genes in the literature. In other words, it compares the number of known targets of each transcription factor present in the list of differentially expressed genes for a cancer type, as well as their direction of change to what is expected from the literature in order to predict likely relevant transcriptional regulators. In addition, if the direction of change in the samples is consistent with a particular transcriptional regulator's activation state (activation or inhibition), then a prediction is made about the activation state. The P-value calls likely upstream regulators based on significant overlap between genes in the study (differentially expressed genes) and known targets regulated by a transcriptional regulator. The activation Z-score infers likely activation states of upstream regulators based on comparison with a model that assigns random regulation directions.

This analysis identified many transcription factors with known involvement in cancer such as FOXO1, MITF, STAT3 and 5, NOTCH, TP53, and FOXM1. The top 15 putatively activated transcription factors in TCGA cancers are shown in Table 3-4.

Among the identified transcription factors, MITF and FOXM1 are the most significantly activated occurring in 15 and 12 cancer types, respectively. MITF (Microphthalmia-Associated Transcription Factor) is a transcription factor that regulates the expression of genes with essential roles in cell differentiation, proliferation, and survival by binding to symmetrical DNA sequences (E-boxes) (5-CACGTG-3) found in the promoters of target genes, such as BCL2 and tyrosinase (TYR). As mentioned in the last section, MITF is highly associated with melanoma progression, invasiveness and metastasis (Vachtenheim & Ondrusova, 2015). Current studies are examining potential avenues to target this transcription factor mechanism for cancer prevention, as MITF itself is not a druggable target. Therefore, MITF-targeting approaches are based on the modulation of its upstream regulatory pathways (Hartman & Czyz, 2015). Similarly, FOXM1 (Forkhead Box M1) is a transcription factor required for a wide spectrum of essential biological functions, including DNA damage repair, cell proliferation, cell cycle progression, cell renewal, cell differentiation and tissue homeostasis. It is also known as a master regulator for a broad array of genes required for Cancer stem cells (Bao et al., 2011; Bergamaschi et al., 2014). The expression of FOXM1 is frequently up-regulated in many malignancies, where it is an early event during cancer development. Accordingly,

genome-wide profiling studies of gene expression in cancers have identified FOXM1 as one of the most frequently up-regulated genes in human malignancies (Zona, Bella et al., 2014)These findings suggest that FOXM1 has a key role in cancer initiation and promotes cancer progression by facilitating cancer angiogenesis, invasion and metastasis. Suppression of FOXM1 has been shown to sensitize human cancer cells to apoptosis induced by DNA-damaging agents or oxidative stress (Gartel, 2014). In addition, *in-vivo* studies showed FOXM1 inhibition leads to inhibition of human xenograft tumor growth in nude mice (Halasi et al., 2013).

Table 3-4. Top 15 putatively activated transcription factors in TCGA cancer types with available matched-normal samples

Transcription Factor	#Tumours	Tumours*
MITF	15	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,SARC,UCEC
FOXM1	12	BLCA,BRCA,CESC,HNSC,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,SARC,UCEC
E2F3	10	BLCA,BRCA,CESC,HNSC,LICH,LUAD,LUSC,READ,SARC,UCEC
ETS1	9	BLCA,BRCA,CESC,COAD,HNSC,KIRP,LUSC,READ,THCA
ESR1	8	BRCA,CESC,COAD,LICH,LUAD,LUSC,SARC,UCEC
RARA	8	BLCA,BRCA,CESC,LICH,LUAD,LUSC,SARC,UCEC
FOXO1	7	BRCA,CESC,HNSC,LUAD,LUSC,PRAD,SARC
CCND1	6	BRCA,CESC,LICH,LUAD,LUSC,UCEC
FOSL1	5	BRCA,CESC,HNSC,LUAD,THCA
JUN	5	BRCA,CESC,COAD,HNSC,KIRC
ATF6	4	CESC,LUAD,PRAD,SARC
DNMT3B	4	CESC,COAD,READ,UCEC
EZH2	4	CESC,COAD,READ,THCA
PAX8	4	CESC,LICH,LUSC,UCEC
HIF1A	3	KIRC,LUSC,UCEC

* The expanded form of each tumour type abbreviation is available in Table 3-2.

Conversely, TP53 (Tumor Protein P53) is the most inhibited transcription factor identified in 11 TCGA cancer types including BLCA, BRCA, CESC, HNSC, KICH, LICH, LUAD, LUSC, PRAD, SARC, and UCEC. This tumour suppressor and transcription factor is a key modulator of cellular stress responses to regulate expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism (Riley, Sontag et al., 2008). Activation of p53 leads to inhibition of cell cycle progression, induction of senescence, differentiation or apoptosis, therefore it is favourable for the tumour cells to inhibit P53 and its downstream associated processes via different approaches such as mutations and deletions. Re-expression of p53 in lymphomas and sarcomas cells lacking functional p53 caused significant, sometimes complete, regression of the tumour by inducing apoptotic cell death in the lymphomas, and cell-cycle arrest with signs of cellular senescence in sarcomas (Kastan, 2007).

Deregulation of transcription factors is a pervasive theme across many, if not all, forms of human cancer (Bhagwat & Vakoc, 2015). Some cancers may alter the function of transcription factors to implement favourable expression changes in the downstream transcription factor target genes to drive oncogenic cell transformation (Bhagwat & Vakoc, 2015). Hence, transcription factor target genes may also offer interesting therapeutic targets in cancer. For example, transmembrane glycoprotein NMB (GPNMB) is one of the known MITF target genes (Gutknecht et al., 2015), where its overexpression is associated with the ability of cancer cells to invade and metastasize (Roth et al., 2016; Zhou et al., 2012). Antibody drug conjugates targeting GPNMB have shown promising results in cancer treatment (Roth et al., 2016). Therefore, transcription factor target genes that are highly expressed in TCGA tumours were also identified and marked as potential cancer-related genes. Such target genes that are commonly perturbed within or across multiple cancer types may play a favourable role in the development and progression of the disease, and therefore present interesting therapeutic targets in cancer. This information was subsequently used for ranking and prioritization of potential targets for therapeutic utility.

Survival Analysis of differentially expressed genes

Survival analysis of differentially over- and under-expressed genes revealed a number of significant associations with survival after correction for multiple testing in each cancer type studied. Kaplan-Meier method was used to assess survival outcomes. This analysis identified both known and novel associations. As shown in Figure 3-3, the overexpression of WNT2 and IL8 in colorectal adenocarcinoma is found to associate with shorter survival time in patients. These associations have also been previously shown by other groups (Jiang et al., 2014; Ning et al., 2011). WNT2 encodes a secreted signalling protein involved in the Wnt signalling pathway and is frequently overexpressed in malignant tissues including colorectal cancer (Park et al., 2009). The overexpression of WNT2 has also been associated with poor clinical outcome of pancreatic patients (Jiang et al., 2014). IL8, a pro-inflammatory chemokine, is known to possess tumorigenic and proangiogenic properties. The overexpression of IL8 has been detected in many tumours and, including colorectal cancer, and is associated with poor prognosis (Ning et al., 2011).

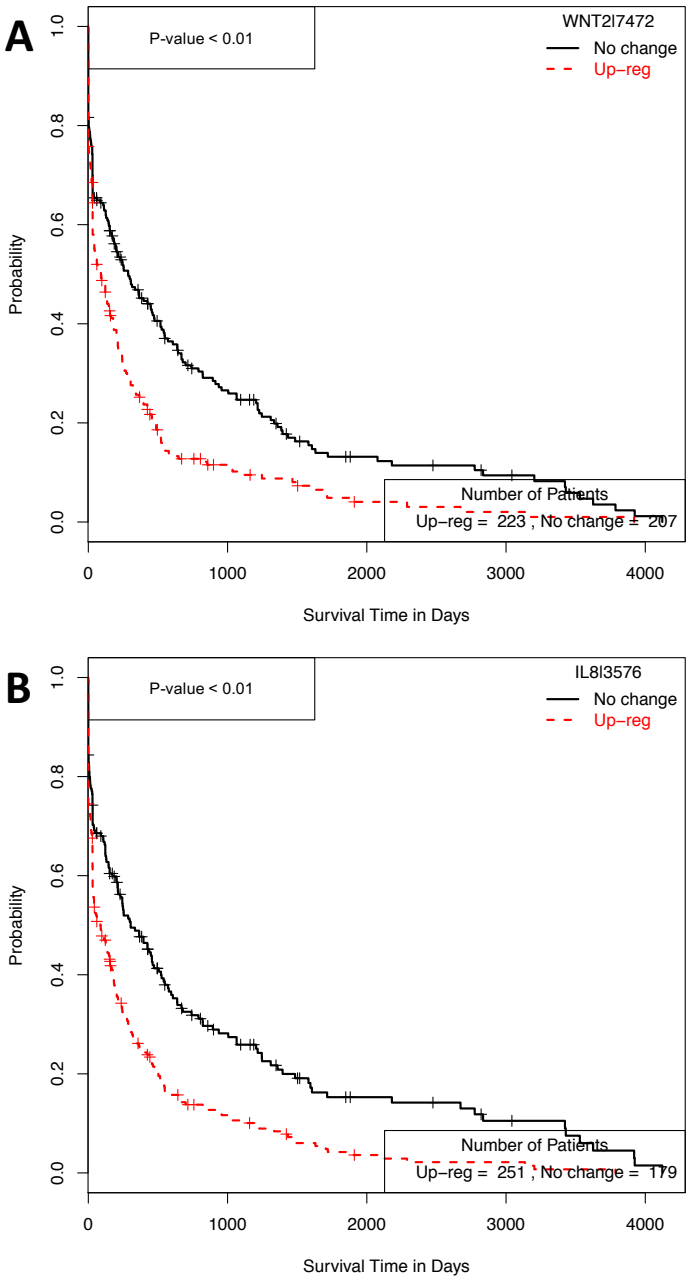


Figure 3-3. Kaplan-Meier survival analysis revealed significantly lower overall survival in Colon Adenocarcinoma (COAD) patients with overexpression of (A) WNT2 and (B) IL8. Up-regulated samples demonstrate a greater than or equal to 2 log fold difference compared to the normal colon tissue. No significant expression difference was observed between the tumour and normal tissues for samples marked as no change.

Survival analysis of differentially expressed genes also revealed novel associations with patient outcome. In lung squamous cell carcinoma, the overexpression of PIF1 is found to be significantly associated with poor patient outcome (Figure 3-4). PIF1 encodes a highly conserved DNA helicase, which is implicated in the maintenance of telomeres and genome stability. It has been suggested that PIF1 plays a role in S-phase entry and progression that are essential to protect human tumour cells from apoptosis (Gagou et al., 2011). Therefore, depletion of PIF1 resulted in reduction of the survival of tumour cells by triggering cell death, while non-malignant cells are unaffected by PIF1 depletion (Gagou et al., 2011).

Similarly, the overexpression of SCARNA12 (Small Cajal Body-Specific RNA 12) significantly correlates with poor outcome in patients with lung squamous cell carcinoma. SCARNA12 gene produces a small nucleolar RNA (snoRNA), which acts as a guide to direct posttranscriptional modification of RNAs (omim.org/entry/625642). In recent years, a number of studies have emerged that indicated a role for snoRNAs in cancer (Su et al., 2014; Williams & Farzaneh, 2012)For example, overexpression of SNORA42, a snoRNA, is frequently found in non-small-cell lung cancer (NSCLC). The down-regulation of SNORA42 in lung cancer cell lines is shown to induce apoptosis and reduce colony-forming ability *in vitro*, and also inhibited tumour formation in a mouse model (Williams & Farzaneh, 2012). On the other hand, ectopic expression of this gene resulted in enhanced proliferation of NSCLC cells (Williams & Farzaneh, 2012). High SNORA42 expression in clinical lung cancer samples showed a significant correlation with poor survival (Williams & Farzaneh, 2012).

Hence, Highly expressed genes in TCGA tumours with significant associations with survival were identified and marked as potential cancer-related genes. This information was used later for prioritization of the potential candidate target genes.

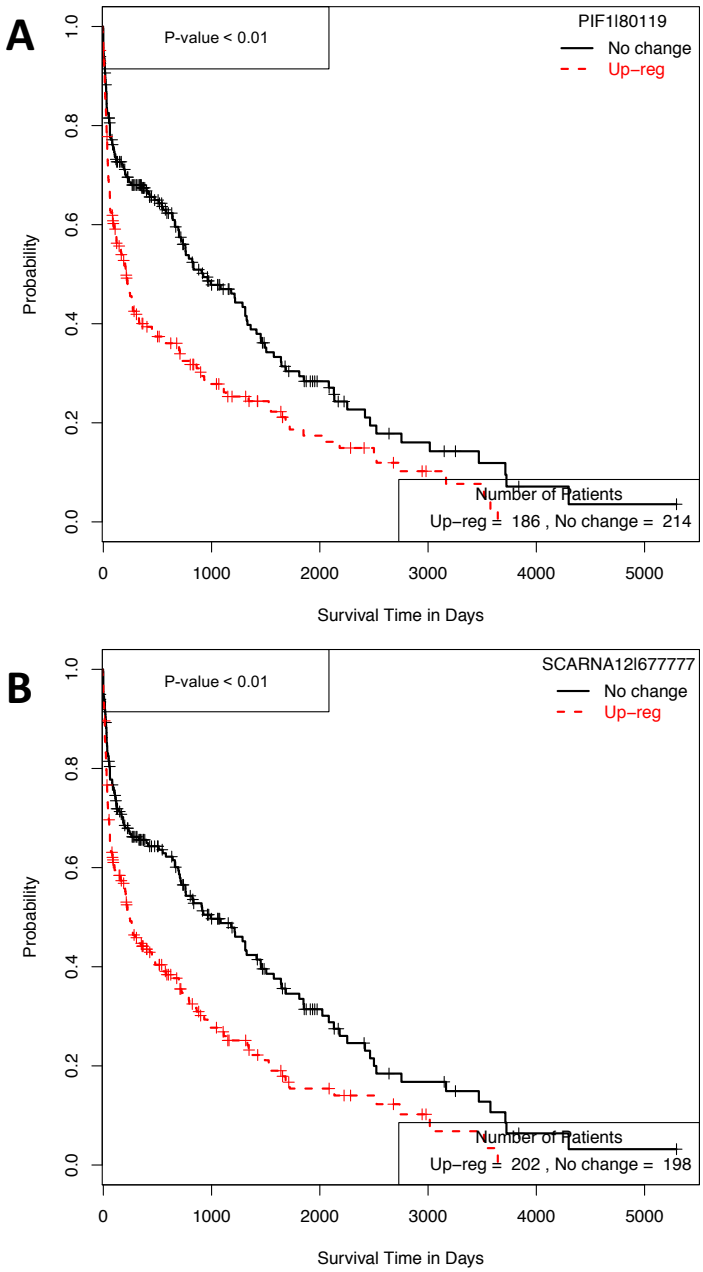


Figure 3-4. Kaplan-Meier survival analysis revealed significantly lower overall survival in Lung squamous cell carcinoma (LUSC) patients with overexpression of (A) PIF1 and (B) SCARNA12. Up-regulated samples demonstrate a greater than or equal to 2 log fold difference compared to the normal lung tissue. No significant expression difference was observed between the tumour and normal tissues for samples marked as no change.

Although many of the identified differentially expressed genes in TCGA tumours do not directly associate with the disease outcome and/or play a defined role in malignant transformation, those that present high tumour expression on the surface of cancer cells compared to the normal tissue may present interesting biomarker targets for antibody-based targeting of tumour cells. Especially those that are found in a number of different cancer types, suggesting their potential favourable role in cancer.

Identification of genes differentially expressed across multiple cancer types

Identified differentially over- and under-expressed genes in TCGA cancers were merged to find those that commonly undergo expression changes. Tables 3-5 and 3-7 show the top 25 most commonly over- and under-expressed genes, respectively. The top most overexpressed genes across TCGA cancers are UBE2C, MYBL2, IQGAP3, and CDKN2A. They are found in 21 out of 24 examined cancer types, and have been previously shown in literature to be involved in cancer development and progression.

The protein encoded by UBE2C (Ubiquitin-Conjugating Enzyme E2C) gene is required for cell cycle progression and checkpoint control by targeted degradation of short-lived proteins. It also plays an important role in mitotic spindle checkpoint control (Hao, Zhang, & Cowell, 2012)Cells that overexpress UBE2C ignore the mitotic spindle checkpoint signals and lose genomic stability, which is a hallmark of cancer. Upon malignant transformation, the expression of UBE2C increases, and this overexpression correlates with the aggressiveness of the tumour. The high UBE2C expression is predictive of poor survival and likely a high risk for relapse (Hao et al., 2012). Also the inhibition of UBE2C reduces proliferation and sensitizes breast cancer cells to radiation, doxorubicin, tamoxifen and letrozole.

MYBL2 (V-Myb Avian Myeloblastosis Viral Oncogene Homolog-Like 2) is a member of the v-myb family of transcription factors and is involved in the regulation of cell survival, proliferation, and differentiation (Papetti & Augenlicht, 2011)More interestingly, there are several lines of evidence that link this gene to a stem cell-like phenotype, which potentially allows for self-renewal, a hallmark of cancer. First, MYBL2 is one of 39 critical transcription factors that are commonly expressed in several different

types of pluripotent stem cells (Muller et al., 2008). Second, it maintains embryonic stem cells in an undifferentiated state. It may also be involved in early steps of differentiation by transcriptionally activating pluripotency-associated genes (Tarasov, Tarasova et al., 2008; Tarasov, Testa et al., 2008). Lastly the absence of functional MYBL2 is embryonic lethal in mice. It is likely because of the inability in these embryos to form an inner cell mass, the source of embryonic stem cells (Tanaka, Patestos et al., 1999). Therefore, developing and maintaining a stem cell phenotype that may play an important role in proliferation and differentiation of several cancer types.

IQGAP3 (IQ Motif Containing GTPase Activating Protein 3) is a member of IQGAP family, which display complicated and often contradictory activities in tumorigenesis. Other members of this family, IQGAP1 and IQGAP2 have oncogenic potential and putative tumour-suppressive function, respectively (White et al., 2010). Similar to IQGAP1, the overexpression of IQGAP3 promote tumour cell growth, migration and invasion. While, its knockdown exhibits opposite effects (Yang et al., 2014). Suppression of this gene in a lung cancer cell line caused a reduction in the tumorigenicity of the cancer cells in lung tissue (Yang et al., 2014).

CDKN2A (Cyclin-Dependent Kinase Inhibitor 2A), also known as P16, plays an important role in cell cycle regulation by decelerating cells progression from G1 phase to S phase. CDKN2A is mainly known to act as a tumour suppressor (Romagosa et al., 2011). However, the overexpression of this gene has also been reported in multiple different cancer types (Dong et al., 1997; Milde-Langosch et al., 2001; Romagosa et al., 2011). In breast cancer, it is associated with a more malignant phenotype (Milde-Langosch et al., 2001). Similarly, in prostate cancer, the overexpression of P16 is associated with tumour recurrence (Lee et al., 1999).

Table 3-5. Top 25 commonly differentially overexpressed genes across TCGA cancer types. This observation suggests a common underlying disease mechanism shared by different cancer types.

Gene	Entrez ID	#Tumour	Tumours*
UBE2C	11065	21	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LGG,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
MYBL2	4605	21	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LGG,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
IQGAP3	128239	21	ACC,BLCA,BRCA,CESC,COAD,GBM,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PAAD,PRAD,READ,SARC,SKCM,THCA,UCEC,UCS
CDKN2A	1029	21	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LGG,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,THCA,UCEC,UCS
UHRF1	29128	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LGG,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCEC,UCS
TROAP	10024	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
TPX2	22974	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LGG,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCEC,UCS
KIF18B	146909	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
HJURP	55355	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
CDC45	8318	20	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
SKA3	221150	19	ACC,BLCA,BRCA,CESC,COAD,GBM,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
PLK1	5347	19	ACC,BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SARC,SKCM,UCEC,UCS
PKMYT1	9088	19	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCEC,UCS
KIF4A	24137	19	ACC,BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP,

Gene	Entrez ID	#Tumour	Tumours*
			LGG,LICH,LUAD,LUSC,OV,PRAD,SARC,SKCM,UCEC, UCS
KIF14	9928	19	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIR C,KIRP,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCE C,UCS
IBSP	3381	19	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIR C,KIRP,LICH,LUAD,LUSC,OV,READ,SKCM,THCA,UCE C,UCS
FOXM1	2305	19	ACC,BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP, LGG,LICH,LUAD,LUSC,OV,PRAD,SARC,SKCM,UCEC, UCS
E2F7	144455	19	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIR C,KIRP,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCE C,UCS
E2F1	1869	19	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KICH,KIR C,KIRP,LICH,LUSC,OV,READ,SARC,SKCM,THCA,UCE C,UCS
CENPF	1063	19	ACC,AML,BLCA,BRCA,CESC,GBM,KICH,KIRC,KIRP,L GG,LICH,LUAD,LUSC,OV,PRAD,SARC,SKCM,UCEC,U CS
TUBB3	10381	18	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRP,LICH ,LUAD,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UC S
TRIP13	9319	18	ACC,BLCA,BRCA,CESC,COAD,GBM,HNSC,KIRC,KIRP ,LICH,LUAD,LUSC,OV,READ,SARC,SKCM,UCEC,UCS
TOP2A	7153	18	ACC,AML,BLCA,BRCA,CESC,GBM,KICH,KIRC,KIRP,L GG,LICH,LUAD,LUSC,OV,SARC,SKCM,UCEC,UCS
TERT	7015	18	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIR P,LICH,LUAD,LUSC,OV,PRAD,READ,SKCM,UCEC,UC S
SOX11	6664	18	ACC,BLCA,BRCA,CESC,GBM,HNSC,KICH,KIRC,KIRP, LGG,LICH,LUAD,LUSC,OV,PRAD,THCA,UCEC,UCS

Identified differentially overexpressed genes in each cancer types are significant with P-value less than or equal to 0.05 and are overexpressed with fold change greater than or equal to 1.5. * The expanded form of each tumour type abbreviation is available in Table 3-2.

Genes differentially overexpressed in cancers are commonly considered for therapeutic and diagnostic purposes especially if they are involved in critical mechanisms in favour of the disease. The gene ontology (GO) enrichment analysis using IPA, shown in Table 3-6, reveals significant association between the identified commonly overexpressed genes and cancer. This observation supports the assumption that cancers benefit from some core processes, which are shared by different cancer types during the oncogenesis. Such genes may serve as therapeutic and/or diagnostic biomarker targets since they represent essential activities and are frequently overexpressed in cancer.

Table 3-6. GO enrichment analysis reveals significant association between the identified commonly overexpressed genes and cancer

Category	Pvalue	Number of Molecules
Diseases and Disorders		
Cancer	7.59E-21-3.13E-04	3800
Endocrine System Disorders	7.59E-21-2.93E-04	1481
Organismal Injury and Abnormalities	7.59E-21-3.13E-04	3946
Reproductive System Disease	7.59E-21-2.93E-04	1372
Infectious Diseases	5.05E-16-3.14E-04	255
Immunological Disease	3.09E-15-3.13E-04	870
Inflammatory Disease	3.09E-15-2.95E-04	799
Connective Tissue Disorders	2.14E-14-2.93E-04	537
Skeletal and Muscular Disorders	2.14E-14-1.3E-04	1144
Inflammatory Response	2.13E-13-2.97E-04	632
Developmental Disorder	6.9E-11-2.93E-04	218
Neurological Disease	1.48E-10-2.95E-04	677
Gastrointestinal Disease	1.87E-10-1.62E-04	630
Respiratory Disease	3.35E-10-2.93E-04	653
Hereditary Disorder	2.83E-08-2.93E-04	569
Renal and Urological Disease	3.42E-08-2.95E-04	760
Metabolic Disease	2.56E-07-1.66E-04	520
Hematological Disease	3.62E-07-3.13E-04	450
Tumor Morphology	6.9E-07-3.13E-04	125
Molecular and Cellular Functions		
Cellular Movement	1.41E-17-3.13E-04	706
Cellular Development	2.23E-15-2.61E-04	971
Cellular Growth and Proliferation	2.23E-15-3.15E-04	1206
Cell-To-Cell Signaling and Interaction	3.36E-12-3.15E-04	734
Cell Signaling	1.24E-11-2.4E-04	278
Molecular Transport	1.24E-11-1.62E-04	576
Vitamin and Mineral Metabolism	1.24E-11-2.58E-04	299
Cell Death and Survival	6.62E-11-3.13E-04	971
Cellular Function and Maintenance	1.31E-07-2.28E-04	377
Cell Morphology	8.41E-07-9.23E-05	221
Nucleic Acid Metabolism	8.43E-07-3.2E-05	127
Small Molecule Biochemistry	8.43E-07-2.58E-04	343

Cellular Compromise	1.82E-06-1.43E-04	43
Cellular Assembly and Organization	2.63E-06-1.81E-04	54
DNA Replication, Recombination, and Repair	2.63E-06-1.81E-04	231
Cell Cycle	7.43E-06-2.09E-04	117
Lipid Metabolism	1.22E-05-2.58E-04	204
Free Radical Scavenging	1.45E-05-1.55E-04	120
Post-Translational Modification	1.78E-05-1.59E-04	111
Protein Synthesis	1.78E-05-1.59E-04	102
Carbohydrate Metabolism	2.84E-05-2.05E-04	25
Protein Degradation	4.23E-05-1.59E-04	83
Amino Acid Metabolism	1.07E-04-1.07E-04	30
Physiological System Development and Function		
Embryonic Development	5.4E-20-3.15E-04	241
Hair and Skin Development and Function	5.4E-20-7E-05	103
Organ Development	5.4E-20-3.15E-04	150
Organismal Development	5.4E-20-3.15E-04	390
Tissue Development	5.4E-20-3.15E-04	512
Immune Cell Trafficking	7.13E-17-2.97E-04	292
Hematological System Development and Function	9.91E-17-3.15E-04	461
Cell-mediated Immune Response	4.07E-07-1.83E-05	54
Tissue Morphology	4.38E-07-4.38E-07	138
Digestive System Development and Function	4.01E-06-1.17E-04	48
Connective Tissue Development and Function	2.84E-05-5.71E-05	31
Skeletal and Muscular System Development and Function	2.84E-05-3.15E-04	71
Organismal Survival	5.68E-05-5.68E-05	78
Cardiovascular System Development and Function	1.04E-04-3.15E-04	228
Hematopoiesis	1.45E-04-2.52E-04	132
Renal and Urological System Development and Function	1.68E-04-1.68E-04	52
Reproductive System Development and Function	2.02E-04-2.02E-04	17
Lymphoid Tissue Structure and Development	2.52E-04-2.52E-04	55

Similar to commonly up-regulated genes among TCGA cancer, there are genes that are commonly down-regulated in cancers compared to their matched-normal tissue. The most common genes include TCEAL2 and SCARA5, which are found to under-express in 17 out of 24 cancer types. TCEAL2 (Transcription Elongation Factor A (SII)-Like 2), nuclear phosphoprotein, is a member of TCEAL family that modulates transcription in a promoter context-dependent manner. It has been recognized as an important nuclear target for intracellular signal transduction. Although the role of TCEAL2 is not clear in cancer, the down regulation of other members of this family including TCEAL7 and TCEAL4 has been reported in different cancer types (Akaishi et al., 2006; Chien et al., 2008). TCEAL7 is a tumour suppressor gene, while the down-regulation of TCEAL4 has been associated with development of anaplastic thyroid cancer from differentiated thyroid cancer. SCARA5 (Scavenger Receptor Class A, Member 5) is a member of class A scavenger receptors that has been proposed recently as a novel candidate tumour suppressor gene in human hepatocellular carcinoma (Huang et al., 2010). SCARA5 down-regulation is essential for epithelial-to-mesenchymal transition (EMT)-induced migration (Liu et al., 2013). Therefore, EMT-regulator Snail1 suppresses the expression of SCARA5 to promote cancer progression. In addition, SCARA5 down-regulation has been reported in several types of human malignancy, and interestingly its up-regulation inhibits tumour growth and metastasis via inactivating signal transducer and activator of transcription 3, as well as downstream signaling including cyclinB1, cyclinD1, AKT, survivin, matrix metalloproteinase-9 and vascular endothelial growth factor-A (Yan et al., 2012).

The exploratory analysis identified over-represented pathways and putative transcription factors regulating the observed gene expressions, as well as commonly up-regulated genes across multiple types of malignancies. All together such knowledge builds a platform that allows for identification of optimal therapeutic targets such as tumour-associated taregts.

Table 3-7. Top 25 commonly down regulated genes across TCGA cancers

Gene	Entrez ID	#Tumours	Tumour*
TCEAL2	140597	18	ACC,AML,BLCA,CESC,COAD,GBM,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,OV,READ,SKCM,THCA,UCEC,UCS
ADH1B	125	18	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UCS
SCARA5	286133	17	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,PRAD,READ,SKCM,THCA,UCEC
MAMDC2	256691	17	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,LUAD,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UCS
FXYD1	5348	17	AML,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRP,LICH,LUAD,LUSC,OV,READ,SKCM,THCA,UCEC,UCS
TMEM132C	92293	16	ACC,BLCA,BRCA,CESC,HNSC,KIRC,KIRP,LICH,LUAD,LUSC,OV,PRAD,SKCM,THCA,UCEC,UCS
PI16	221476	16	ACC,AML,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,PRAD,READ,THCA,UCEC
LMOD1	25802	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRP,LUAD,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UCS
GSTM5	2949	16	ACC,BLCA,BRCA,CESC,COAD,HNSC,KIRP,LICH,LUAD,LUSC,OV,READ,SKCM,THCA,UCEC,UCS
DPT	1805	16	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUSC,READ,SKCM,THCA,UCEC,UCS
CHRDL1	91851	16	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,LUAD,LUSC,PRAD,READ,SKCM,THCA,UCEC,UCS
CDO1	1036	16	AML,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,LUAD,LUSC,OV,PRAD,READ,SKCM,UCEC,UCS
C7	730	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,READ,SKCM,UCEC,UCS
C1QTNF7	114905	16	BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,OV,READ,SKCM,THCA,UCEC,UCS
BAI3	577	16	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,LICH,LUAD,LUSC,OV,READ,SKCM,UCEC,UCS
AOX1	316	16	AML,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRP,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UCS
ANGPTL1	9068	16	BLCA,BRCA,CESC,COAD,HNSC,KIRC,KIRP,LUAD,LUSC,OV,PRAD,READ,SKCM,THCA,UCEC,UCS
ADRA1A	148	16	ACC,AML,BLCA,BRCA,COAD,HNSC,KICH,LICH,LUAD,LUSC,OV,PRAD,READ,SKCM,UCEC,UCS
TGFBR3	7049	15	BLCA,BRCA,CESC,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,OV,PRAD,READ,SKCM,UCEC,UCS
TCF21	6943	15	AML,BLCA,BRCA,CESC,COAD,KICH,KIRC,KIRP,LICH,LUAD,LUSC,OV,READ,UCEC,UCS

Gene	Entrez ID	#Tumours	Tumour*
SVEP1	79987	15	ACC,AML,BLCA,BRCA,CESC,COAD,KIRC,KIRP,LUAD,LUSC,READ,SKCM,THCA,UCEC,UCS
SCN7A	6332	15	ACC,BLCA,BRCA,CESC,COAD,HNSC,KIRC,KIRP,LUAD,LUSC,OV,READ,SKCM,UCEC,UCS
SCN2B	6327	15	AML,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,LUAD,LUSC,PRAD,READ,SKCM,UCEC,UCS
RSPO1	284654	15	ACC,BLCA,BRCA,CESC,COAD,HNSC,KICH,KIRC,KIRP,LUAD,LUSC,READ,SKCM,UCEC,UCS
PGM5P2	595135	15	ACC,BLCA,BRCA,CESC,COAD,HNSC,KIRC,KIRP,LUSC,OV,PRAD,READ,SKCM,UCEC,UCS

Identified differentially overexpressed genes in each cancer types are significant with P-value less than or equal to 0.05 and are overexpressed with fold change greater than or equal to 1.5. * The expanded form of each tumour type abbreviation is available in Table 3-2.

3.1.3. Identifying optimal tumour biomarker targets

Antibodies (mAbs, ADCs, bi-specific) are a rapidly growing class of drugs used for treatment of human cancers and other diseases. Currently there are several antibody drugs approved by FDA for various indications including cancer treatment and diagnosis. Some of them are shown in Table 3-8 (http://www.fda.gov/). In addition, many more antibodies are in preclinical and clinical development (Scott et al., 2012). Antibodies can be used effectively to target tumour-associated molecules and thereby modulate key signaling pathways that play a role in tumour growth, survival and metastasis. Using the Fc region, these proteins can recruit the host immune system to fight against cancer by mediating antibody-dependent cellular cytotoxicity and complement-dependent cytotoxicity, which can result from an antibody binding to a target on tumour cells (Scott et al., 2012). In addition, ADCs can also be used to deliver a payload such as a cytokine, chemotherapeutic small molecule or radionuclide by binding to tumour cells (Polakis, 2016). Thus, the high specificity, long half-life and relative safety of antibodies compared with other cancer therapeutics together with their ability to bind to and modulate key players in pathways that drive malignant transformation and enhance antitumor immune functions make them highly desirable therapeutic agents. In addition to mAbs and ADCs that bind to one target, the bi-specific antibodies can attach to two different proteins at the same time (Chames & Baty, 2009)In this case, both targets are either expressed on the surface of a tumour cell simultaneously, or one is expressed on the surface of a cancerous cell and the other one is present on the surface of an immune cell. Therefore, in the latter, the antibody brings the cancer cells and immune cells together to facilitate a tumour cell killing process by the immune system. In the case of the former scenario, bi-specific antibodies offer more specificity and improved efficacy than mAbs and ADCs in identifying and binding target cancer cells. They also can be paired with toxins to specifically deliver them to the tumour site.

Table 3-8. Currently approved antibody-based diagnostic and therapeutic agents

Trade name	Other names	Company	Target	Therapeutic indication(s)
Verluma® (Diagnostic)	Nofetumomab	Boehringer Ingelheim, NeoRx	Carcinoma-associated antigen	Diagnostic imaging of small-celllung cancer (non-therapeutic)
ProstaScint® (Diagnostic)	Capromab	Cytogen	PSMA	Detection of prostate adenocarcinoma (non-therapeutic)
CEA-scan® (Diagnostic)	Arcitumomab	Immunomedics	CEA	Detection of colorectal cancer (non-therapeutic)
Rituxan® MabThera®	Rituximab	Biogen Idec, Genentech (Roche)	CD20	Non-Hodgkin's lymphoma Chronic lymphocytic leukemia Rheumatoid arthritis
Herceptin®	Trastuzumab	Genentech(Roche)	HER-2	Breast cancer Metastatic gastric or gastroesophageal junction adenocarcinoma
Mylotarg®	Gemtuzumab ozogamicin	Wyeth	CD33 (ADC)	Acute myeloic leucemia (AML)
Campath®	Alemtuzumab	Millennium Pharmaceuticals and Genzyme	CD52	B-cell chronic lymphocytic leukemia
Zevalin®	Ibritumomab tiuxetan	Biogen Idec	CD20	Non-Hodgkin's lymphoma
Bexxar®	Tositumomab and iodine 131 tositumomab	Corixa and GSK	CD20	Non-Hodgkin's lymphoma
Avastin®	Bevacizumab	Genentech (Roche)	VEGF	Metastatic colorectal cance, rNon-small cell lung cancer, Metastatic breast cancer, Glioblastoma multiforme, Metastatic renal cell carcinoma
Erbitux®	Cetuximab	ImClone (Eli Lilly), Merck Serono and BMS	EGFR	Head and neck cancerColorectal cancer
Vectibix®	Panitumumab	Amgen	EGFR	Metastatic colorectal carcinoma
Arzerra®	Ofatumumab	Genmab and GSK	CD20	Chronic lymphocytic leukemia
Adcetris®	Brentuximab	Seattle Genetics	CD30 (ADC)	Hodgkin lymphoma (HL),systemic anaplastic large cell lymphoma (ALCL)

Trade name	Other names	Company	Target	Therapeutic indication(s)
Xgeva®	Denosumab	Amgen	RANKL	Prevention of SREs in patients with bone metastases from solid tumours
Vervoy®	Ipilimumab	BMS	CTLA-4	Melanoma
Perjeta®	Pertuzumab	Roche	HER2	Breast cancer
Kadcyla®	Trastuzumab emtansine	Roche	HER2	Breast cancer

Although the unique properties of antibodies themselves are key components of a successful antibody-based therapeutic approach, the target proteins recognized by these antibodies play an equally important role. Cancer is caused by genetic and epigenetic changes that regulate cell proliferation, apoptosis, migration, angiogenesis and other biologic properties that underlie cell growth, survival and interaction with the extracellular environment. These genetic and epigenetic changes may lead to cancer-specific expression of genes. These changes in gene expression can be identified in tumour cells or the host environment such as tumour stroma or components of the adaptive and innate immune system.

With the availability of datasets such as TCGA and GTEx, a bioinformatic approach can be used to identify novel targets that can discriminate between tumour and normal tissues. Here, I postulate that the list of differentially expressed genes within and across multiple cancer types can be further narrowed down through three key characteristics representative of an ideal tumour target for targeting with a therapeutic antibody, including target localization, expression pattern, and function:

(1) A desirable tumour target is located on the surface of tumour cells. In addition, in case of ADCs it is favourable that the target is capable of internalizing into the cells. Proteins localized to the surface of human cells are potential diagnostic and therapeutic targets. Cell surface proteins of interest with respect to antibody-based drug targets include: integral membrane, phospho-lipid-linked, or surface associated proteins by other means such as those expressed by tumour epithelium, angiogenic endothelium, stroma, or immune cells (Papkoff, 2007).

(2) An ideal tumour target should be overexpressed or uniquely expressed on the majority of tumour cells with no or limited normal tissue expression. The expression of an ideal tumour target is to be abundant on the surface of tumour cells at all stages of cancer development to provide a broader window of opportunities for treating patients, and is restricted or absent from vital normal tissue to minimize the risk of antibody-dependent toxicities (Carter et al., 2004). An exception to overexpression would be proteins expressed by both normal and cancerous cells at a similar level, while

a unique form is expressed within the cancer, including novel splice variants and fusion proteins.

(3) Conceptually, an ideal target is preferred to play a defined role in malignant transformation, however this is not necessary for a target to become successful. Tumour targets with a role in malignant transformation may therefore be essential for cancer cell survival and thus resistance to a therapeutic antibody through gene loss might be less likely to arise (Papkoff, 2007). GO and pathway analysis are some of the approaches to elucidate the target's role in the biology of the disease.

Identification of cell surface proteins

A catalog of human cell-surface associated proteins was compiled through an extensive search of literature (Da Cunha et al., 2009; Diaz-Ramos, Engel, & Bastos, 2011; Fagerberg, Jonasson et al., 2010) and databases such as human protein atlas (proteinatlas.org), UniProt (uniprot.org), cancer vaccine center (bio.dfci.harvard.edu), and available gene ontology (geneontology.org). Cell-surface proteins could be integral membrane, GPI-linked, expressed by tumor epithelium, angiogenic endothelium, stroma or immune cells. It has to be noted that since some of the localizations are predicted based on sequence information and bioinformatic tools, they may not in fact be localized as expected or may localize to membranes that are inside the cell such as mitochondria, endoplasmic reticulum, golgi or nucleus and, therefore, would not be available to a therapeutic antibody. In addition, localization of proteins may differ between tumor and normal cells. In total, more than 4,000 cell-surface proteins have been collected in this analysis. Where available, the extracellular region of the proteins was also annotated using Uniprot protein annotation.

The differential expression analysis, described in 3.1.2, revealed 14,217 genes differentially overexpressed in at least one of the 24 different types of malignancies available from TCGA; of which 10,923 were found in more than one type of cancer. Comparison of the differentially expressed genes with the compiled list of surface proteins revealed 2,824 genes that could code for cell surface proteins, hence their protein product my localize to the surface of tumour cells.

Identification of cancer-associated differentially expressed genes

The GTEx project characterises more than 30 non-cancerous tissue types collected from deceased donors and organ/tissue transplant patients with the goal of studying the relationship between genetic variation and gene expression in human tissues. Therefore, it offers a unique opportunity to study the expression of identified differentially expressed genes in TCGA cancers across normal tissues to identify cancer-correlated expression. Therefore, 400 RNA-seq samples were downloaded from the GTEx data repository, where at least 10 samples were downloaded for each tissue type. The raw RNA-seq reads were run through the first steps of the GEA pipeline for data quality assessment and gene coverage analysis with RSEM. A compendium matrix of FPKM values was created from the expression of genes of interest across all 30 GTEx tissue types. Similarly the expression of target genes in each cancer type that was found to be differentially expressed were collected. A Mann-Whitney test was applied on the normalized expression values (FPKM) to identify genes that show significant difference in tumour samples in comparison to the compendium of normal tissues. Gene showing significant difference (p-value and FDR <= 0.05) in their expression pattern between tumour and normal conditions will be referred to as cancer-associated differentially expressed (CADE) genes. This analysis revealed 1,503 genes (out of 2,824) with higher level of expression in cancer in comparison to the GTEx database. This list is available as appendix A.

Identification of optimal targets for antibody targeting

Studying a list of targets that are FDA approved or are currently in clinical trial for antibody-based therapeutics (shown Table 3-8) revealed that an optimal tumour target follows one of the following three expression patterns in normal tissues:

(1) The most desirable tumour taregts are those that are only expressed on the surface of tumour cells at a high level, while their expression in normal tissues are either very low or it is completely absent. The normal tissue can be further broken down into regenerative tissues where damage to them is not life threatening, and critical tissues that may cause severe side effects if damaged. Non-critical tissues may include the

reproductive system, breast, and thyroid tissues. Examples of critical tissues are heart, lung, kidney, small intestine, and skin.

(2) Tumour targets that are expressed in multiple normal tissues while the tumour expression is much higher than normal tissue expression.

(3) Tumour targets that are expressed at similar level in both tumour and normal tissues, but play a major role in tumour survival and progression while their normal function is not critical. In addition, protein variants (that are products of alternative splicing, mutation, and etc.) that are specifically expressed on the tumour cell surface fall into this group of targets.

Tumour-specific biomarker targets are the most favourable targets. However, the number of such targets with no expression in normal tissues is very limited. The majority of 1,503 cell surface localized cancer-associated genes identified in previous section that show higher expression in tumour cells than normal follow the second class of targets described above. Of those putative candidate genes, 28 present no to low expression (<=20 FPKM) across all normal tissues, while 54 genes have low to no expression in critical normal tissues. Such targets are a favourable target for naked antibodies if they play a significant role in the disease, a target for ADCs to deliver a load of toxins to the tumour site if they internalize, or a desirable target for bi-specific antibodies that use a combination of a marker on the surface of tumour cells with a marker expressed on the surface of immune cells to bring them together in order to initiate natural tumour cell killing by the immune system. The mRNA expression profile of some candidates is shown in Figures 3-5 to 3-7. This analysis successfully identified known cancer targets suggesting that this method may identify putative novel targets as well.

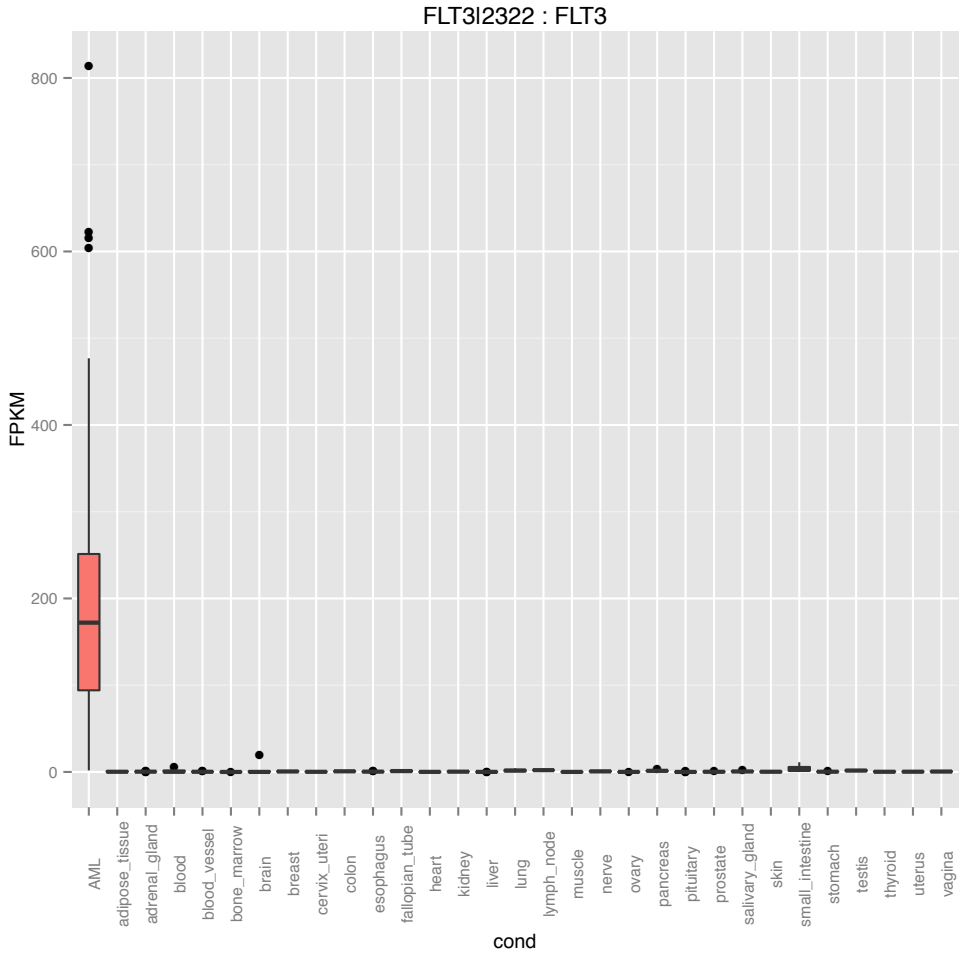


Figure 3-5. Putative tumour biomarker target FLT3 demonstrates high expression in AML samples while has no to little expression across normal tissues tested. The expanded form of each tumour type abbreviation is available in Table 3-2.

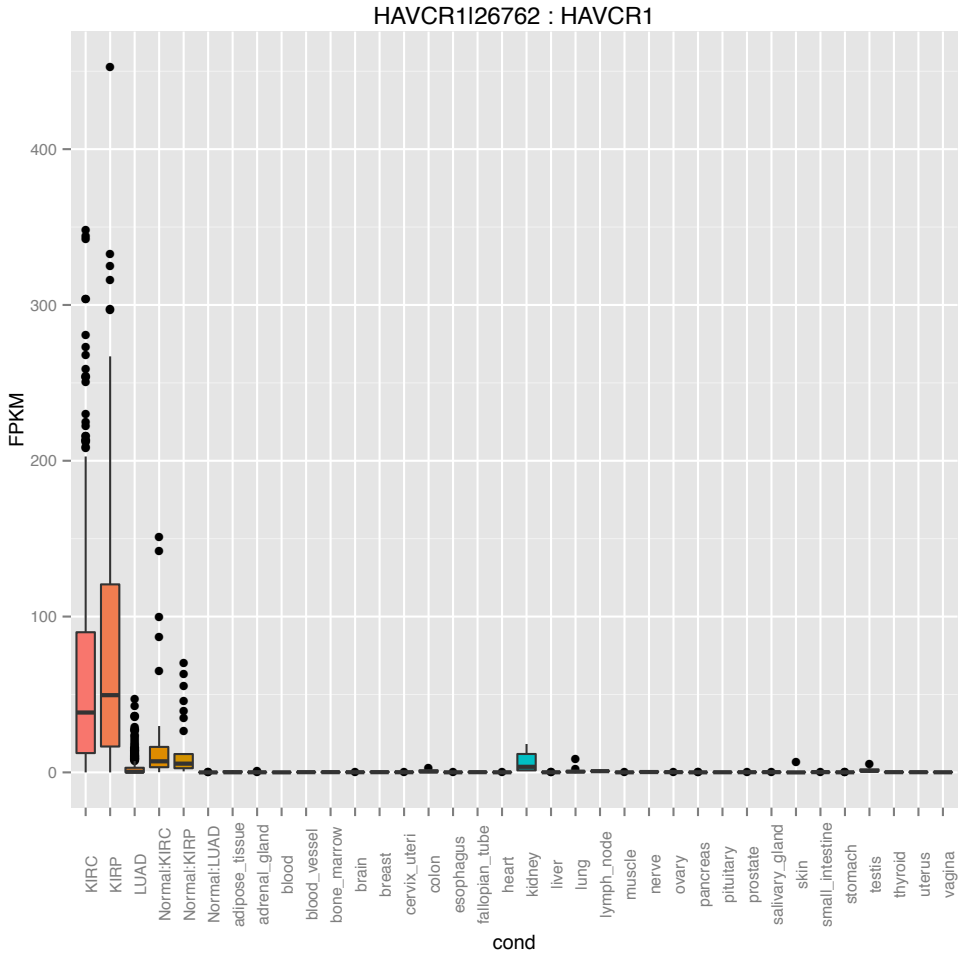


Figure 3-6. Putative tumour biomarker target HAVCR1 demonstrates high expression in kidney and lung cancer samples while has low expression in matched normal tissue. The expanded form of each tumour type abbreviation is available in Table 3-2.

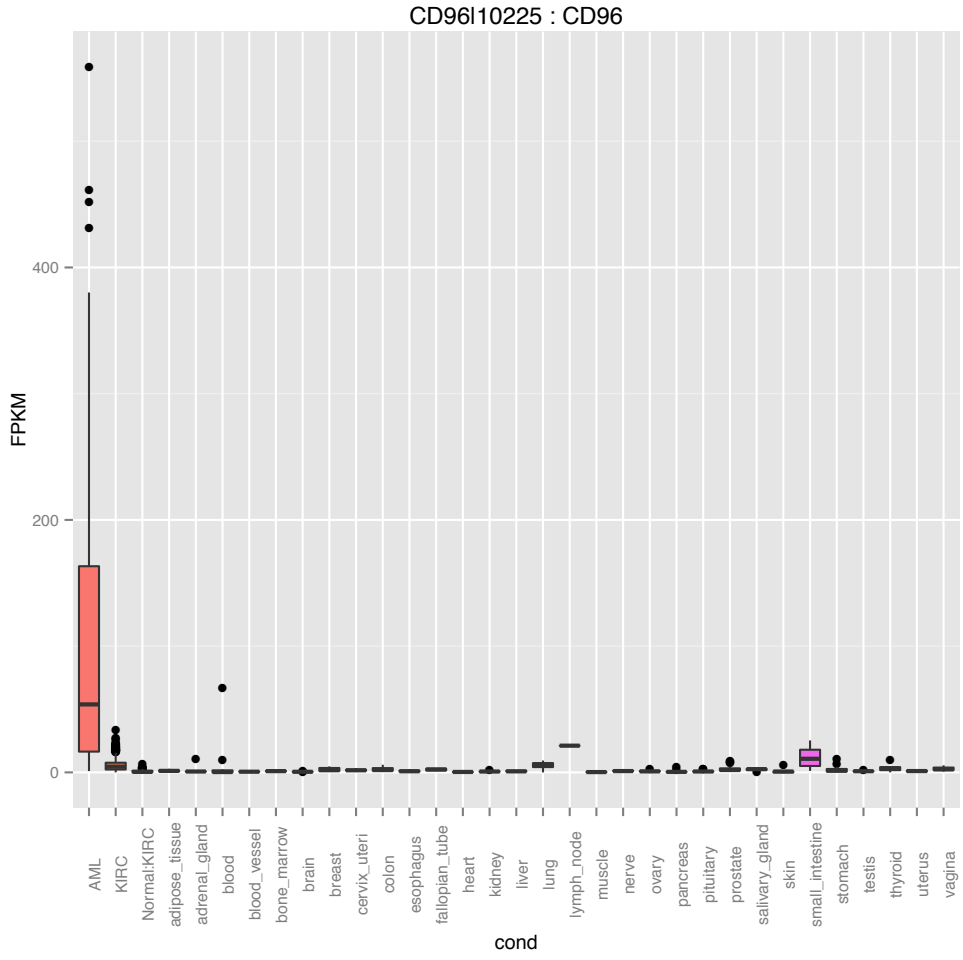


Figure 3-7. Putative tumour biomarker target CD96 demonstrates high expression in AML samples while has lower expression in critical normal tissue including small intestine, blood, lung, lymph node and adrenal gland. The expanded form of each tumour type abbreviation is available in Table 3-2.

In addition to the targets with low or no expression in every normal tissue, there are cases that are expressed only in a limited number of normal tissues and still can play a role as an attractive target. For example, carbohydrase 9 (also known as CA9) is highly expressed in normal stomach tissue (Figure 3-8). However, antibodies targeting CA9 are currently in clinical trial and are showing promising results (McDonald, Winum et al., 2012; Zatovicova et al., 2010). Therefore, identified candidates must each be evaluated based on their level of expression and the type of the normal tissue that they are expressed in. Because, the large number of identified candidates makes it challenging to evaluate each target individually, a method is required to rank and prioritize these candidates. In addition to the normal expression profile, the expression profile in tumour tissues is another key criterion in the success of a tumour target. The higher the target is expressed, the chance that antibodies find and bind to it. In addition, higher tumour expression compared to lower expression in normal tissues decreases the chance of antibodies binding to the target expressed on the surface of healthy normal cells. Considering all the above, and other characteristics of a tumour target, I developed an R package, Prize, based on the analytic hierarchy process algorithm to perform ranking and prioritization of identified putative tumour markers based on a set of user-defined criteria. In addition, I developed an AHP model to depict the characteristics of tumour targets to perform this ranking. This method is described in section 3.3.

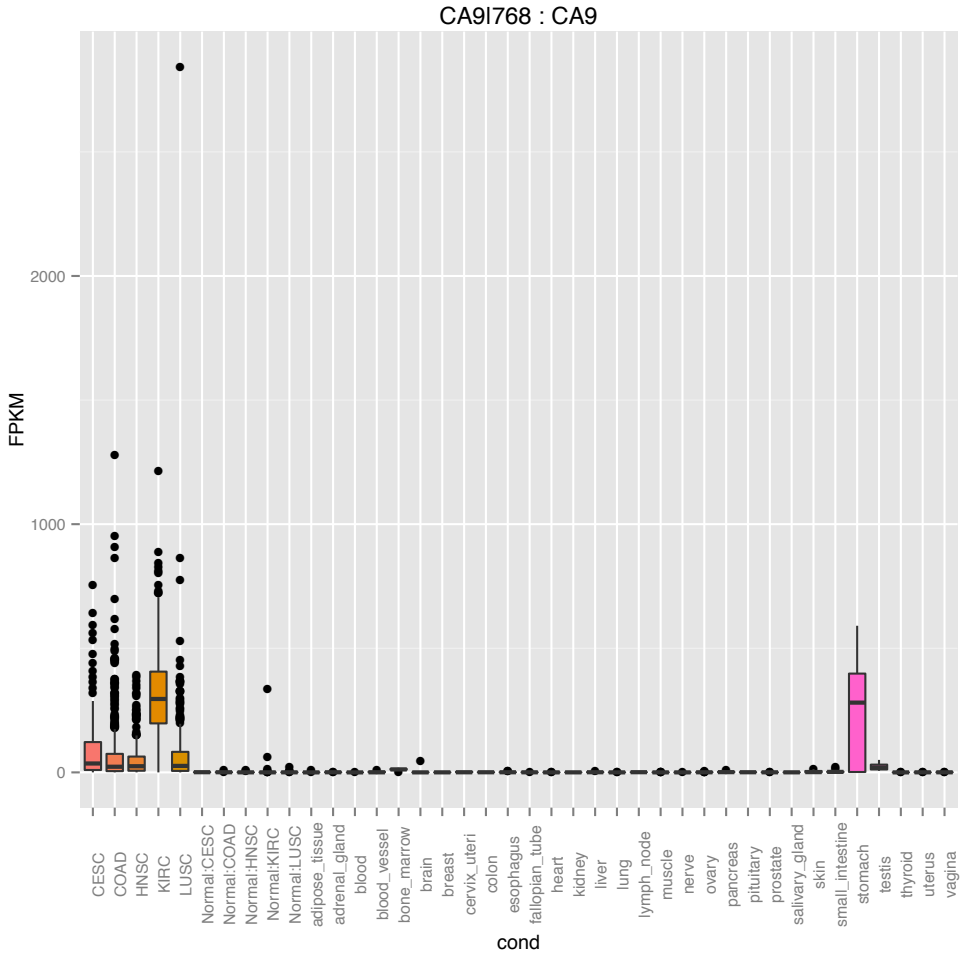


Figure 3-8. The expression profile of putative tumour biomarker target CA9. Even though CA9 demonstrates high expression in normal stomach tissue, it has been shown as an effective tumour target in tumour cell killing with no severe side effects (McDonald et al., 2012; Zatovicova et al., 2010). The expanded form of each tumour type abbreviation is available in Table 3-2.

Identification of potential targets for bi-specific anbodies

Bi-specific antibodies are capable of targeting two targets on the surface of tumour cells simultaneously. The fact that these antibodies bind to two targets significantly improves their specificity compared to mAbs. In addition, binding to different cell surface proteins, bi-specific antibodies allow for blocking more than one pathway component, or simultaneously hitting complementing pathways, which may limit potential escape mechanisms of cancer cells. Similar to mAbs, they may also be used as vehicles to deliver immune effector cells and/or cytokines to tumours. Therefore, an optimal pair of targets for bi-specific antibodies is a pair of genes that are both expressed on the surface of tumour cells (preferably at high levels) while their normal tissue expression is limited and mutually exclusive. The mutual exclusive expression defines as; there are no normal tissues that express the two targets simultaneously except the matched normal tissue of the tumour of interest.

In order to identify such pair of genes with mutually exclusive expression pattern across normal tissues, I studied the RNA-seq data available from GTEx. Since damage to critical normal tissue (including tissues from adipose, adrenal gland, blood and blood vessel, bone marrow, brain, colon, esophagus, heart, kidney, liver, lung, lymph node, muscle, nerve, pancreas, pituitary, salivary gland, skin, small intestine, and stomach) is mainly the cause of severe side effects in patients, only critical tissues were included in this analysis. To identify pairs with mutually exclusive expression in critical tissues, first a 0-1 matrix was generated from the expression of every gene present in the human genome (total of 26,761 genes) according to the expression profile across the 21 critical tissues. An entry is equal to 0 when a gene is not expressed (FPKM < 10), while it is equal to 1 when it is expressed in the tissue of interest with FPKM greater than or equal to 10. Then, the generated profile for each gene was multiplied into the 0-1 matrix (Figure 3-9). The outcome is equal to zero if a pair of genes has mutually exclusive expression pattern in the critical normal tissues, while it is greater than or equal to 1 if they are not. If the outcome is greater than 0, the value represents the number of normal tissues that the pair of genes is expressed in simultaneously.

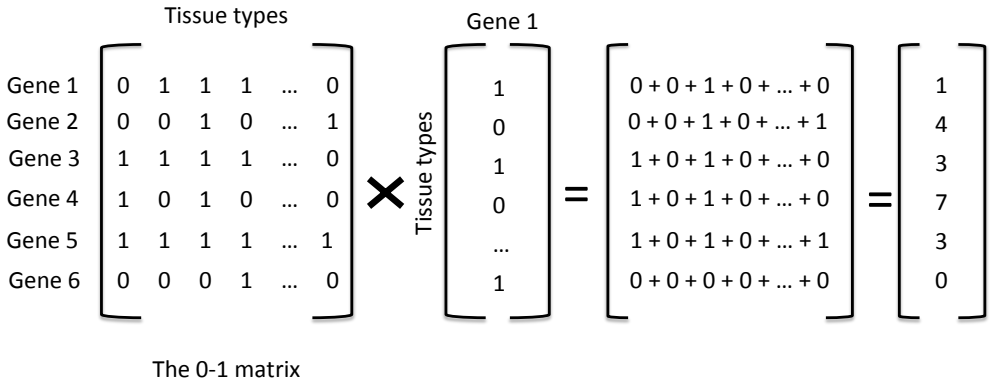


Figure 3-9. A 0-1 matrix was generated from the expression of every gene present in the human genome in any of the 21 critical tissue types available from GTEx. Genes were multiplied one by one to the 0-1 matrix. The outcome is zero if the pair are mutually exclusive across critical normal tissues. Here gene 1 is mutually exclusive with gene 6. This means that there is no critical tissue that expresses both genes at the same time. While gene 1 is expressed in 1, 4, 3, 7, and 3 tissues as genes 1 to 5 also do.

An ideal pair of targets for bi-specific antibodies can be considered the one where both genes are highly expressed on the surface of tumour cells, while their normal expression is limited and mutually exclusive across critical normal tissues. Therefore, for each TCGA cancer type, the list of cell-surface associated differentially expressed genes where compared with the list of identified mutually exclusive pairs to identify such candidate pairs. In total 1,280 pairs were identified. This list is available as Appendix B. An example of genes with mutual exclusive expression pattern is shown in Figure 3-10.

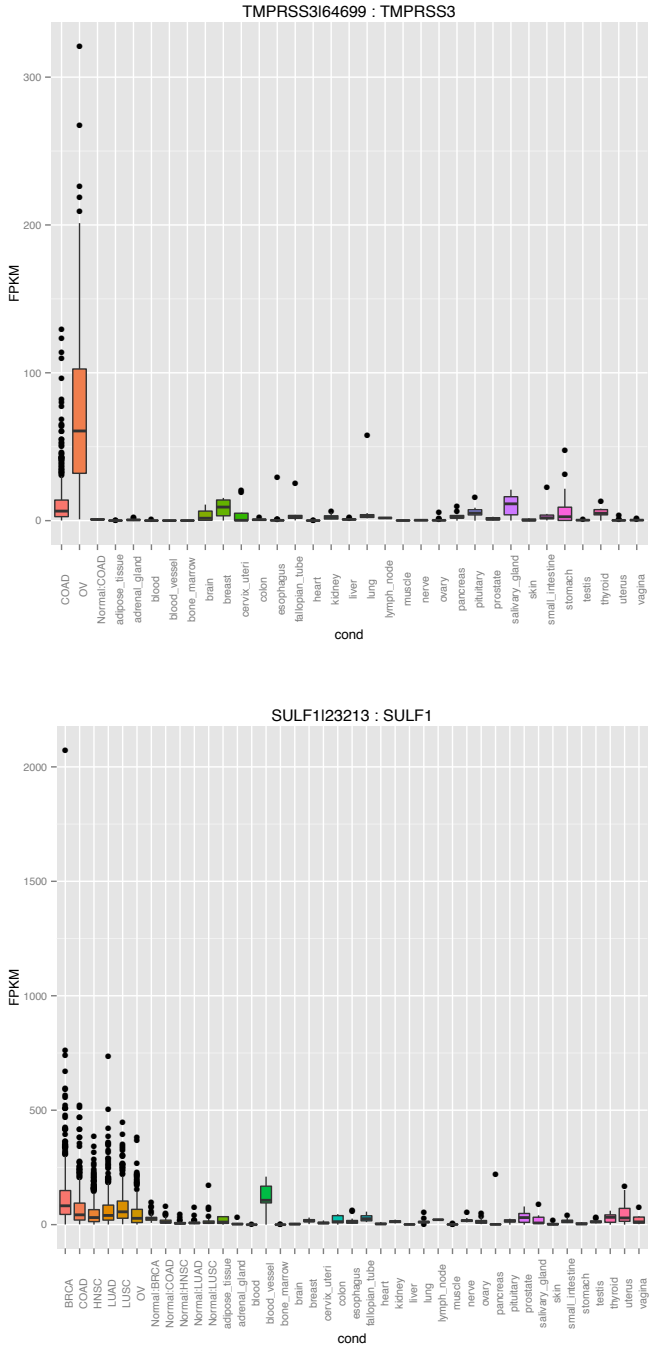


Figure 3-10. TMPRSS3 and SULF1 demonstrate mutually exclusive expression pattern in normal critical tissues, while both are differentially overexpressed in colon and ovarian cancers. The expanded form of each tumour type abbreviation is available in Table 3-2.

3.2. Pan-cancer identification of cancer-associated alternatively spliced genes

One of the mechanisms by which oncogenic events can occur is through the modification of the transcriptome. The AS of pre-mRNA transcripts is common in eukaryotic cells and provides a mechanism for a normal cell to generate a number of diverse protein products from a single gene locus. AS is thus thought to increase the functional diversity of the encoded genome. Some transcript variants may only be generated during certain times of development and only in certain tissues. In cancer, cells are able to recapitulate variants that are involved in developmental and proliferative stages, while those variants are normally absent in differentiated tissues. Tumour-associated alternatively spliced variants represent attractive biomarker targets especially if the presence of these variants is otherwise low or absent in normal patient tissues. Since alternatively spliced transcripts possess new exon-exon boundaries and can involve the loss or gain of a number of exons they can lead to relatively large changes in the primary and three-dimensional structure of a protein. This in turn can provide a relatively large and specific target for mAb generation. These splice variant specific mAbs have the potential to be used both prognostically and therapeutically.

A number of cancer associated alternate splicing events have been identified that confirm this process contributes to multiple facets of oncogenesis and tumour establishment. Some aberrant splice variant transcripts are involved in aspects of embryonic development while others appear to be aberrant novel forms only arising within cancer cells (He, Zhou et al., 2009). For example, VEGF is typically secreted by hypoxic cancer cells where it ultimately binds to the VEGF2 receptor present on the surrounding endothelial cells, there it stimulates the growth of endothelial tissue and the formation of new capillaries (Potente, Gerhardt, & Carmeliet, 2011). The VEGF ligand is also known to undergo extensive alternative splicing, producing both pro-angiogenic and anti-angiogenic isoforms. In cancer, the AS of the VEGF ligand is skewed toward the pro-angiogenic form compared to the ratio observed in normal tissue (Qiu, Hoareau-Aveilla et al., 2009). Similarly, hypoxia induces AS of the CD44 gene. In CD44 where the

overall function of the protein product is poorly understood, numerous cancer associated spliced variants have been identified (Orian-Rousseau, 2010; Ponta, Sherman, & Herrlich, 2003)In particular, the presence of the spliced variants CD44v6 and CD44v8 are associated with poorer outcome and more rapid progression in a number of tumour types (Kopp, Fichter et al., 2009; Saito et al., 2013).

AS has also been found to play a key role in the process of epithelial-to-mesenchymal transition (EMT) whereby cells undergo de-differentiation and lose their tight cell-cell junctions, ultimately allowing the cells to disperse to other sites in the body giving rise to metastasis. The Ron proto-oncogene (MST1R) was the first gene involved in EMT determined to be regulated through alternative splicing. In this case a constitutively active isoform produced through the loss of exon 11 confers pro-motility properties to the cancer cell (Ghigna et al., 2005; Zhou, He, Chen et al., 2003). Subsequently, numerous other genes involved in EMT have been found to undergo tumour associated alternative splicing, including Rac1 (Jordan, Brazao et al., 1999), KLF6 (Narla et al., 2008), FAM3B (Li et al., 2013), Cortactin (Van Rossum et al., 2003), MENA (Di Modugno et al., 2007) and L1CAM (Hauser et al., 2011). Apoptosis is also influenced through the tumour-associated AS of CASP8 (Mohr et al., 2005), CASP9 (Shultz & Chalfant, 2011), and BCL-X (Boise et al., 1993). Other oncogenic processes are also influenced by AS such as increased telomerase activity and altered centrosome function through the AS of TERT (Wong et al., 2013) and TACC1 (Line, Slucka et al., 2002) respectively.

Within human cancer alternatively spliced forms of proteins on the cell surface are obvious targets for antibody based-therapies - particularly if the spliced variant is tumour-specific. Even in the cases where the splice variant is not tumour-specific a comprehensive understanding of the normal tissues where it is expressed and its expression levels can allow the potential toxicity to essential organs and side-effects to be predicted. Certainly, targeting a therapeutic antibody to the tumour is fundamentally more appealing than systemic untargeted application of chemotherapeutics.

The availability of large datasets such as TCGA and GTEx provides the opportunity of studying the landscape of AS in human malignancies as well as normal

healthy tissues. Therefore, in this section, I introduce an AS variant detection pipeline from RNA-seq data. Using this pipeline, I examine the TCGA and GTEx data in order to identify cancer-associated events. Identified variants were then further examined to identify putative tumor markers for antibody therapeutics.

3.2.1. AS detection pipeline

Cancer cells can usurp the cells splicing mechanism to produce functional transcripts that favour the malignant state. Novel splice variants have been identified in a variety of cancers, suggesting that widespread aberrant and AS may be a common consequence or even a cause of cancer (Venables, 2004). Even though the biological activity of the majority of AS isoforms, and in particular, their contribution to cancer biology, has yet to be elucidated. A number of studies have demonstrated that cancer-associated splice variants can serve as diagnostic or prognostic markers, or predict sensitivity to certain drugs (Griffith et al., 2012; Pajares et al., 2007; Venables et al., 2008). RNA-seq allows the exploration of cancer-related changes at the level of transcription and splicing. Here, I devised an AS-detection pipeline based on a *de novo* assembly approach.

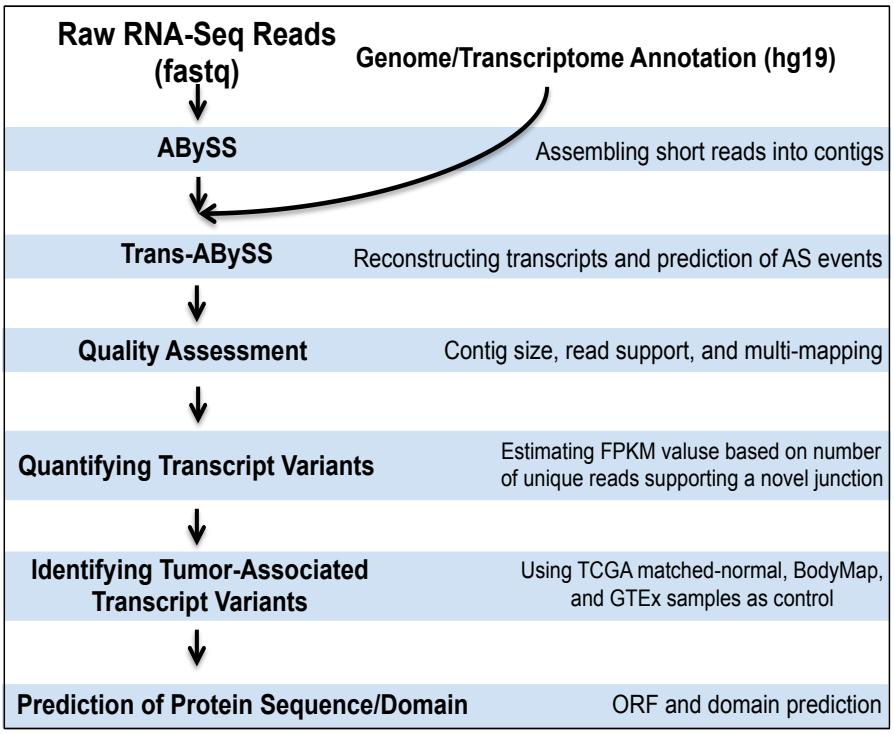


Figure 3-11. Alternative Splicing (AS) detection pipeline

The AS-detection pipeline starts with raw RNA-seq data (FASTQ files). The pipeline core step includes *de novo* transcriptome assembly using ABySS and Trans-ABySS software package. ABySS is a *de novo*, parallel, and paired-end sequence assembler designed for short reads. It assembles a data set multiple times using a De Bruijn graph-based approach. Trans-ABySS post-processes ABySS assemblies to merge contigs and remove redundancy. This approach reconstructs transcripts from a broad range of expression levels, including those expressed at low levels. The pipeline also consists of the following steps: assessing the quality of assembled transcripts, identifying tumour-associated events, quantifying predicted transcripts, and prediction of protein sequence and domains (Figure 3-11). These steps are described below:

De novo transcriptome construction

The *de novo* transcriptome assembly leverages the redundancy of short-read sequencing to find overlaps between the reads and assembles them into transcripts. We assembled short RNA-seq reads into contigs using ABySS version 1.3.4 for multiple K-mer values. A K-mer is all the possible subsequences (of length K) from a read obtained through sequencing of RNA. TCGA RNA-seq libraries are paired-end and the read length is 48 bp. We assembled each library for 13 different values of K-mer from 24 to 48 in increments of two. This approach captures transcripts from a broad range of expression levels, thus allowing lowly expressed transcripts to be constructed. Trans-ABySS (version 1.4.4) was then used to merge ABySS assemblies, removing redundancy and reconstructing transcripts. The *de novo* transcriptome construction therefore captures major splice rearrangements and novel variations that occur in the transcriptome, including exon-skipping, novel exons, retained introns and AS at 3'-acceptor and 5'-donor sites. Since this approach does not rely on a reference genome, it can assemble novel AS as well as trans-spliced transcripts. Constructed transcripts were then annotated by mapping them to the human reference genome (hg19).

Transcript quality assessment

Predicted AS transcripts were evaluated by their contig size, number of reads supporting predicted novel junction, and their alignment quality. Transcripts with contigs smaller than 200 bp and less than 4 reads supporting a predicted novel junction were

removed from further analysis. The mis-assembly of transcriptome reads may occur as a result of mutation, low quality and low complexity of the reads, as well as presence of repeats. This could lead to the prediction of false splice junctions. In order to identify such cases, we aligned predicted AS transcripts back to the human genome (hg19) using BLAT from UCSC (http://hgdownload.cse.ucsc.edu/admin/exe/) and evaluated the alignment quality of sequences that span predicted novel junctions. BLAT was run using default parameters. If sequences that span a novel junction were also aligned to a different part of genome with similarity greater than 70%, we labelled such transcripts as unreliable and removed them from further analysis. Transcripts that passed initial quality assessment were visualized by UCSC genome browser (https://genome.ucsc.edu/) or Integrative Genome Viewer (IGV, http://www.broadinstitute.org/igv/).

Quantifying predicted transcripts

Only the reads that align to a novel junction are isoform informative. Trans-ABySS estimates the number of these reads, which allows the quantification of the novel AS isoform abundance. Assuming each unique read spanning a novel junction is generated from a transcript uniformly, each exon in a AS isoform was assigned an equal number of reads as the number of spanning reads, and estimated FPKM values. For example, gene A with 4 exons is shown in Figure 3-12. There are five reads (shown in red) that suggest the skipping of exon two in this gene. The five reads that align to the novel junction suggest that there are at least five transcripts that support the novel splice variant. Therefore, five reads is assigned to each remaining exon to estimate the total number of reads supporting this novel AS isoform. This value then is used toward estimation of FPKM.

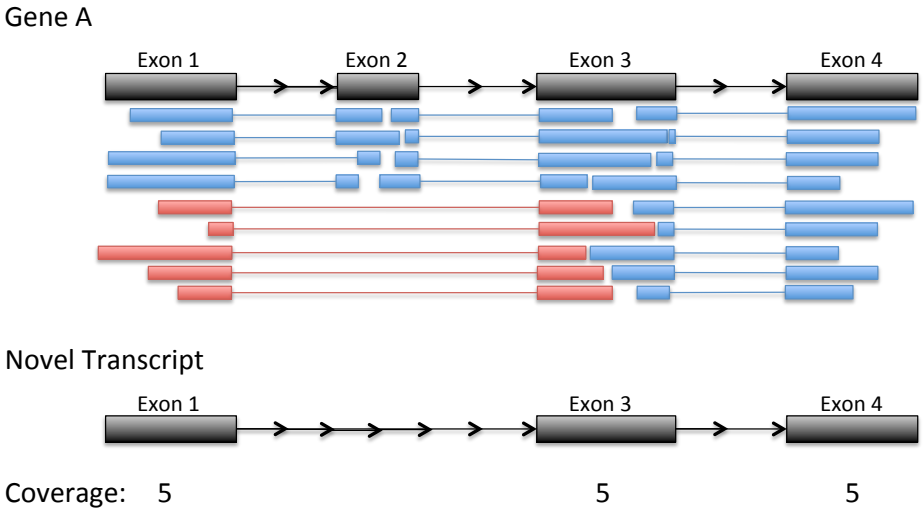


Figure 3-12. Estimation of total number of reads supporting a novel splice variant. Assuming each unique read spanning a novel junction is generated from a transcript uniformly (shown in red here), each exon in a novel splice variant was assigned an equal number of reads as the number of spanning reads. This value was then used towards estimation of values.

Identification of tumor-associated transcripts

In order to identify and remove tissue-specific splicing variants, we compared predicted transcripts from tumour libraries with the ones present in available corresponding normal data from TCGA as well as GTEx and Illumina BodyMap 2.0 project. BodyMap consists of 19 normal transcriptomes from 16 different tissue types, making it an invaluable source for studying tissue-specific transcript models (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/). Similarly GTEx offers a large RNA-seq dataset including samples from 30 non-cancerous tissue types. Tissue-specific AS events were also predicted using ABySS/Trans-ABySS software package as described above. Transcript variants not detected by the *de novo* transcriptome assembly approach are considered as not being expressed.

Prediction of protein sequence and domain

Open reading frame (ORF) prediction is performed using NCBI ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) to identify the longest open reading frame in each transcript. Protein domains are predicted by RPS-BLAST at NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi).

3.2.2. Identification of alternatively spliced genes within and across multiple cancer types

The raw RNA-seq reads for 20 TCGA cancer types (Table 3-9) were obtained from the TCGA data repository. The *de novo* reconstruction of transcripts was performed for both tumour and matching normal samples from TCGA using the AS-detection pipeline. The pipeline identifies 5 types of events including skipped exon, retained intron, AS at 3' acceptor site, AS at 5' donor site, and novel exon. Each predicted AS event is required to be supported by at least 4 reads mapped to the novel junction. In addition, in case of novel exon and retained intron a minimum of 10 read is required to support the novel insertion.

The AS events for the adjacent non-cancerous normal tissue from TCGA, if available, were also predicted using the AS-pipeline. If an AS event predicted in the TCGA cancer samples is also found in the matched normal tissues, then it is marked as a non-somatic event and is removed from the further analysis.

Table 3-9. Tumour and corresponding adjacent non-cancerous tissue sample from TCGA investigated to identify novel cancer-associates splice variants

ID	Type	Tumour sample	Matched normal	Platform
ACC	Adrenocortical carcinoma	79	N/A	RNA-seq
AML	Acute Myeloid Leukemia	161	N/A	RNA-seq
BLCA	Bladder Urothelial Carcinoma	116	14	RNA-seq
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	104	3	RNA-seq
ESCA	Esophageal carcinoma	186	23	RNA-seq
GBM	Glioblastoma multiforme	73	N/A	RNA-seq
HNSC	Head and Neck squamous cell carcinoma	177	25	RNA-seq
KICH	Kidney Chromophobe	66	33	RNA-seq
KIRC	Kidney renal clear cell carcinoma	398	63	RNA-seq
KIRP	Kidney renal papillary cell carcinoma	141	30	RNA-seq
LIHC	Liver hepatocellular carcinoma	161	50	RNA-seq
LUAD	Lung adenocarcinoma	183	57	RNA-seq
LUSC	Lung squamous cell carcinoma	303	41	RNA-seq
OV	Ovarian serous cystadenocarcinoma	429	N/A	RNA-seq
PAAD	Pancreatic adenocarcinoma	55	4	RNA-seq
PRAD	Prostate adenocarcinoma	166	38	RNA-seq
SKCM	Skin Cutaneous Melanoma	256	N/A	RNA-seq
STAD	Stomach	430	33	RNA-seq

ID	Type	Tumour sample	Matched normal	Platform
	adenocarcinoma			
TNBC	Triple negative breast cancer	109	10	RNA-seq
UCS	Uterine Carcinosarcoma	57	N/A	RNA-seq

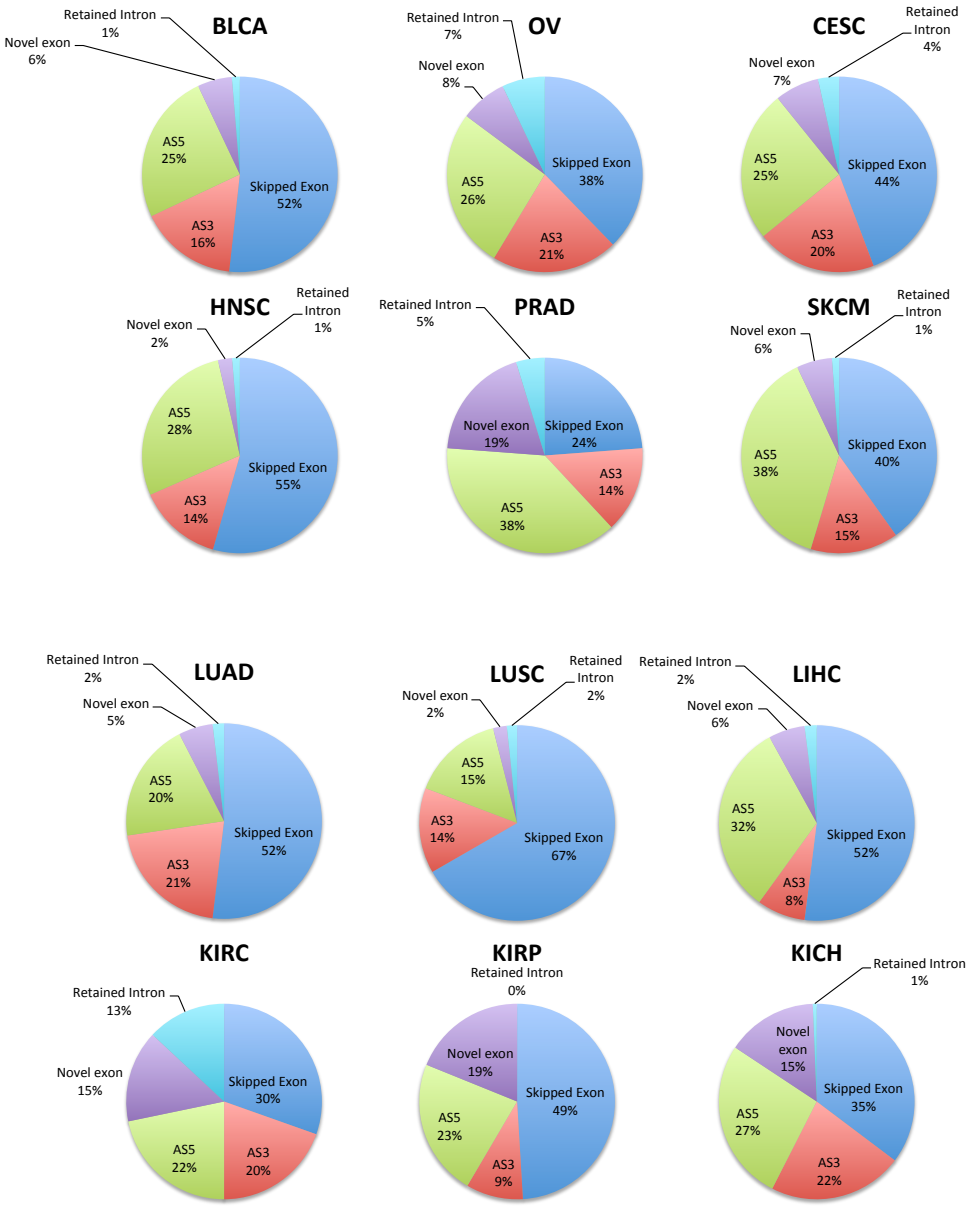
The AS-detection pipeline was successfully able to predict both known and novel splice variants in human cancers. One example is the prediction of the epidermal growth factor receptor variant III, also known as EGFRvIII (Sampson et al., 2008). This variant has a deletion of exons 2-7 which creates a novel epitope unique to the tumour-associated form of the receptor. As reported previously, EGFRvIII has restricted tumour specific expression, including glioblastoma (GBM) tumours (Sampson et al., 2008). The AS-detection pipeline was able to identify this variant in 8% of the GBM tumours available from TCGA.

The prediction of the human cancers AS landscape also revealed skipped exon as the most common type of AS in cancer (Figure 3-13). During an exon-skipping event, exons are included or excluded from the final gene transcript leading to extended or shortened mRNA variants. As exons represent the coding regions of a gene and are responsible for producing proteins that are utilized in various cell types for a number of functions. Skipped exon events may therefore result in formation of protein isoforms that display functional diversity. Therefore, tumours could use this mechanism to form protein isoforms that favour their malignant state. Similar observation has also been made by (Tsai et al., 2015).

Interestingly, a lower number of splicing variants was observed in prostate adenocarcinoma (PRAD) in comparison to the other cancer types in this study. PRAD is also the only cancer type that skipping exon is not the dominant from of AS events. This observation may be consistent with the lower mutation rate in prostate cancer (Taylor et al., 2010). It also should be noted that the modest number of samples tested here limits this analysis.

Tumour-associated AS variants represent attractive taregts for mAb development in oncology, especially if the presence of these variants is otherwise low or absent in normal patient tissues. For instance, the EGFRvIII is currently being investigated by several research groups to be used as a target for antibody-based cancer therapeutics in oncology (Padfield, Ellis, & Kurian, 2015) and mAbs targeting EGFRvIII coupled to cytotoxic molecules to form an ADC have demonstrated very potent anti-tumour activity.

Since alternatively spliced transcripts possess new exon-exon boundaries and can involve the loss or gain of a number of exons, they can lead to relatively large changes in the primary and three-dimensional structure of a protein. This in turn can provide a relatively large and specific target for mAb-based agents. However, a challenge of targeting AS events might be the lower expression of these variants compared to the canonical isoforms, as was the case for many of the identified AS variants in my analysis. Although advances in antibody engineering technologies allow effective targeting of these splice variants even with low expression for both potential prognostic and therapeutic purposes.


BLCA
Retained Intron 1%
Novel exon 6%
AS5 25%
Skipped Exon 52%
AS3 16%
OV
Retained Intron 7%
Novel exon 8%
Skipped Exon 38%
AS5 26%
AS3 21%
CESC
Retained Intron 4%
Novel exon 7%
AS5 25%
Skipped Exon 44%
AS3 20%
HNSC
Novel exon 2%
Retained Intron 1%
AS5 28%
Skipped Exon 55%
AS3 14%
PRAD
Retained Intron 5%
Novel exon 19%
Skipped Exon 24%
AS3 14%
AS5 38%
SKCM
Novel exon 6%
Retained Intron 1%
AS5 38%
Skipped Exon 40%
AS3 15%
LUAD
Retained Intron 2%
Novel exon 5%
AS5 20%
Skipped Exon 52%
AS3 21%
LUSC
Novel exon 2%
Retained Intron 2%
AS5 15%
AS3 14%
Skipped Exon 67%
LIHC
Retained Intron 2%
Novel exon 6%
AS5 32%
Skipped Exon 52%
AS3 8%
KIRC
Retained Intron 13%
Novel exon 15%
Skipped Exon 30%
AS5 22%
AS3 20%
KIRP
Retained Intron 0%
Novel exon 19%
Skipped Exon 49%
AS5 23%
AS3 9%
KICH
Retained Intron 1%
Novel exon 15%
Skipped Exon 35%
AS5 27%
AS3 22%

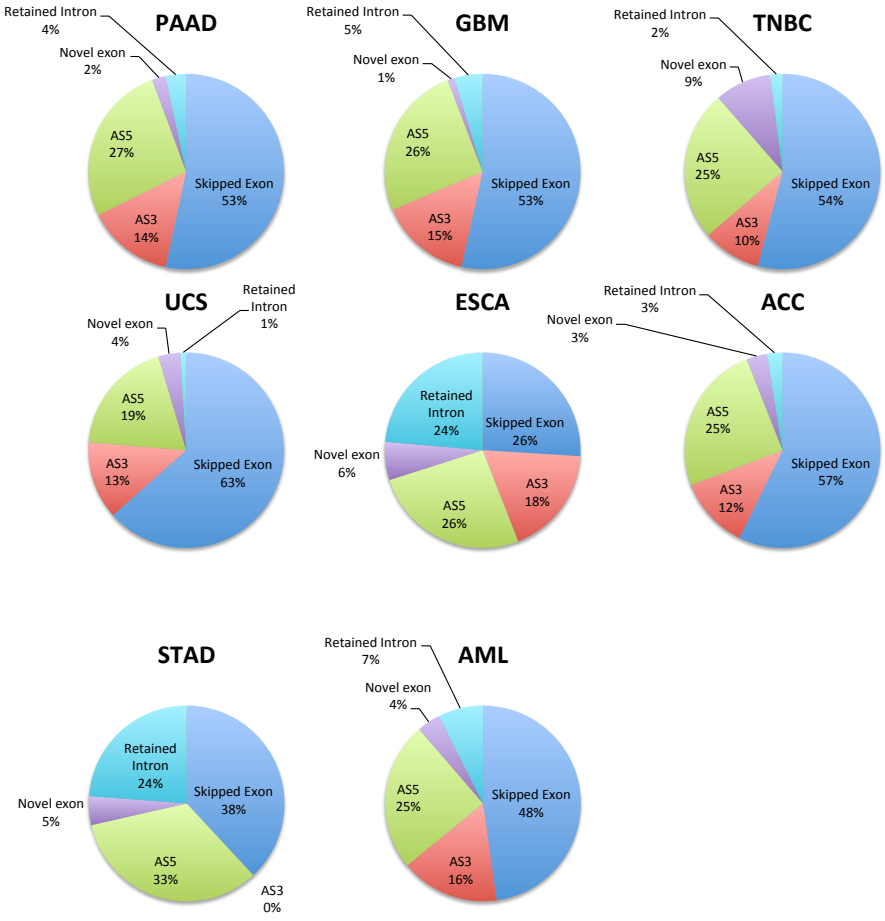


Figure 3-13. Skipped exons are the most common type of splicing variants in human cancers. AS3: Alternative 3' splice site (also known as acceptor). AS5: Alternative 5' splice site (also known as donor). The expanded form of each tumour type abbreviation is available in Table 3-9.

Identification of optimal AS variants for antibody-based cancer therapy

While highly expressed surface proteins in cancers represent excellent targets for antibody-based therapeutics (section 3.1.3), there also exist many splice variant isoforms that appear to be unique to cancer cells and these represent a significant potential for clinical development. In order to identify such events, raw RNA-seq reads from 30 non-cancerous tissue types were downloaded from the GTEx and Body Map data repositories and the AS landscape was predicted using the AS-detection pipeline (over 400 samples). TCGA tumour samples were then compared against this dataset to identify the cancer-associated events. Those AS variants that occur in normal non-cancerous tissue were identified and removed from future analysis. In total 1,142 cancer-associated splice variants occurring in 694 genes were identified. This list is available as appendix C. Next, the cell surface-associated genes were identified using the compiled dataset in section 3.1.3. In total 180 cancer-associated cell surface AS variants were identified across the TCGA cancer types.

I observed many of these AS variants demonstrating lower expression than their respective canonical isoforms. The ability to conjugate highly potent cytotoxic compounds to the binding antibodies could potentially mitigate this problem. Therefore, somatically cancer-specific protein isoforms represent attractive candidates for mAb development in oncology, particularly if such protein isoforms are recurrent either within or across tumour types at clinically relevant frequencies. From my analysis, one of the most commonly occurring AS variant among TCGA cancer types, are two skipping exon events of a known cancer-associated gene named matriptase (also known as ST14). The following sections will describe the bioinformatic analysis and further validation of these variants across independent tumour tissues and cell lines. This work is published in the journal of cancer informatics (Dargahi et al., 2014). It has been done as collaboration between Genome Sciences Centre (GSC) and the Centre for Drug Research and Development (CDRD) in Vancouver. CDRD is a non-profit company focused on identification of genetic alterations in human cancer for diagnostic and therapeutic purposes. The splice variants were identified through bioinformatic analysis at GSC by myself and were validated by CDRD in orthogonal samples by performing qRT-PCR and flow cytometry analysis.

3.2.3. Epithelial-derived tumours express novel splicing variants of matriptase

Matriptase (MT-SP1/TADG-15/ST14) is a type II transmembrane serine protease (TTSP) encoded by a gene located at human chromosome 11q24-25, and is localized to the cell surface (Lin et al., 1997). It has a multi-domain structure common for the TTSP family. The intracellular domain at its amino terminal contains a consensus phosphorylation site for protein kinase C, followed by a signal anchor transmembrane domain. At the extracellular region, matriptase contains a single SEA domain (sea urchin stem region, enteropeptidase, and argin), two CUB repeats (complement C1r/C1s, Uegf, Bmp1), and four tandem repeats of a LDLRA domain (ligand binding repeats of the low-density-lipoprotein receptor class A) (Tanimoto et al., 2001). It is synthesized as an inactive, single chain zymogen and catalyzes its own auto-activation (Lee et al., 2007). Once activated, matriptase cleaves and activates the hepatocyte growth factor/scattering factor (HGF/SF), and urokinase plasminogen activator (pro-uPA) (Lee, Dickson, & Lin, 2000; Takeuchi et al., 2000; Unterholzner et al., 2010) suggesting that this protease functions as an epithelial membrane activator for other proteases and latent growth factors. Matriptase substrate proteins are known to play important roles in tumour development. Activated HGF/SF binds to its receptor, Met proto-oncogene (Met), and stimulates multiple downstream pathways including Rat sarcoma viral oncogene-Mitogen Activated Protein Kinase (Ras-MAPK), Phosphoinositide-3-Kinase (PI3K), Schmidt-ruppin A-2 oncogene (Src), and Signal transducer and activator of transcription 3 (Stat3). In turn, this leads to the activation of gene products required for invasive growth (Kang et al., 2003; K. Matsumoto & Nakamura, 1996; Trusolino & Comoglio, 2002) uPA regulates cell/extracellular matrix (ECM) interactions as an adhesion receptor for vitronectin, and cell migration as a signal transduction molecule and by its intrinsic chemotactic activity, thereby promoting tumour invasion and metastasis (Sidenius & Blasi, 2003). By controlling the activity of uPA and HGF/SF, matriptase is a prime constituent in the activation cascade for invasive growth and metastasis.

Matriptase activity is tightly regulated via antagonism from hepatocyte growth factor activator inhibitor-1 (HAI-1). HAI-1 is a serine peptidase inhibitor encoded by Kunitz type 1 gene (SPINT1) (Shimomura et al., 1997). HAI-1 has not only an inhibitory

function, but is also required for matriptase activation, and regulates the proper expression and intracellular trafficking of matriptase (Oberst, Williams et al., 2003; Oberst et al., 2005). It has been shown that in the absence of HAI-1, matriptase biosynthesis is significantly lower due to auto-proteolytic activation in the Golgi-endoplasmic reticulum apparatus. This event has a detrimental effect upon the trafficking of the matriptase protease, and the cessation of further matriptase translation (Oberst et al., 2005). The role of HAI-1 as both inhibitor and activator of matriptase provides a means to prevent unwanted proteolysis and the subsequent harmful effects of matriptase on cells.

Matriptase is widely expressed by the epithelia of almost all organs examined so far (Oberst et al., 2003). Studies of matriptase-deficient mice have shown that matriptase is essential for postnatal survival, epidermal barrier function, hair follicle development, and thymic homeostatsis (List et al., 2002). Matriptase has also been shown to overexpress in a variety of human cancers. In many cases, high matriptase expression levels are correlated with poor clinical outcome (List et al., 2005; Oberst et al., 2002). In addition to matriptase overexpression, an imbalance in the ratio of matriptase to HAI-1 has been reported in late stage tumours leading to the proposal that uninhibited matriptase activity may contribute to the development of advanced disease (Oberst et al., 2002).

Although many studies present matriptase as a promising potential therapeutic target in oncology (Oberst et al., 2002; Wu et al., 2010), its therapeutic use is limited by its widespread expression and essential function in normal epithelial tissues. However, a unique form of matriptase within tumour cells could potentially overcome this limitation. Using the AS-detection pipeline, I identified two novel tumour-associated spliced isoforms of matriptase in the transcriptome of primary ovarian, breast, prostate, head and neck, lung, stomach, and bladder carcinoma that were not in normal transcriptomes from the adjacent non-tumour tissue. This finding is confirmed by quantitative analysis of mRNA expression of matriptase splice variants using qRT-PCR on cDNA panels obtained from an orthogonal set of tumour tissues and cell lines. Then using flow cytometry, the presence of matriptase splice variants on the surface of transfected CHO cells with cDNA encoding these variants were demonstrated. Tumour association and

the high frequency of matriptase splice variants within and across epithelial tumours suggest that these mutant matriptase transcripts may be of potential therapeutic value. This is the first study reporting tumour-associated transcripts of matriptase in human cancers.

Identification of two novel splice variants of matriptase

De novo assembly of matriptase transcripts revealed two novel splice variants in epithelial-derived tumours. As depicted in Figure 3-14, these variants contain an in-frame exon skipping of the LDLRA1 or LDLRA3 domain, respectively. The novel transcripts were therefore denoted A1 (skipping LDLRA1), and A3 (skipping LDLRA3). Similar analysis for transcriptomes derived from melanoma, leukemia, and glioblastoma tumors did not identify A1 and A3 variants. This is consistent with the observation that matriptase is predominantly expressed by the epithelial tissue ($p=0.006$ and 0.0242, respectively).

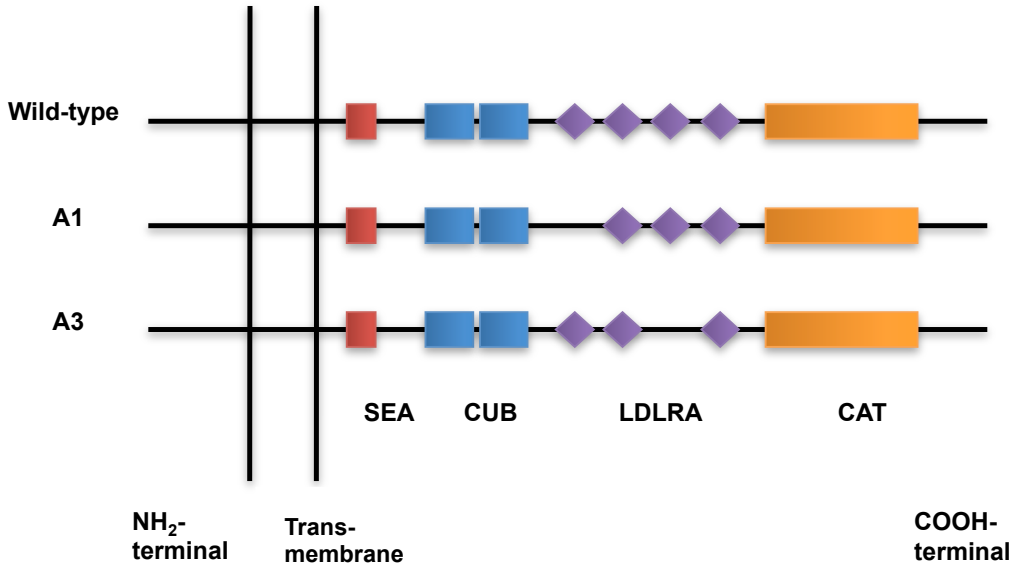


Figure 3-14. Schematic representation of novel matriptase AS transcripts. Four LDL receptor class A domains are found in matriptase, including: LDLRA1: residues 452–486, LDLRA2: residues 487–523, LDLRA3: residues 524–561, and LDLRA4: residues 566–604. A1 and A3 are produced by skipping exon 12 (encoding LDLRA1) and exon 14 (encoding LDLRA3), resulting in in-frame deletion of 105 and 114 bp, respectively. CAT: serine protease catalytic domain.

An estimation of A1 and A3 transcript abundances using the number of reads supporting the novel exon-exon junction from Trans-ABySS indicated higher expression for A1 compared to the A3 transcript in all tumours studied (Figures 3-15 and 3-16). We observed a wide range in the frequency of epithelial tumours displaying these matriptase splice variants, from 3% in prostate adenocarcinoma (PRAD) to 69% in lung squamous cell carcinoma (LUSC) (Figure 3-17). Matriptase variant A1 was found more frequent than A3 across all tumours studied ($p=0.01$). In addition, A3 variant was not detected in the transcriptomes from the prostate adenocarcinoma (PRAD). Among samples with matriptase splice variant-positive cancer, we observed cases that either express one or both splice variants of matriptase (Figure 3-17).

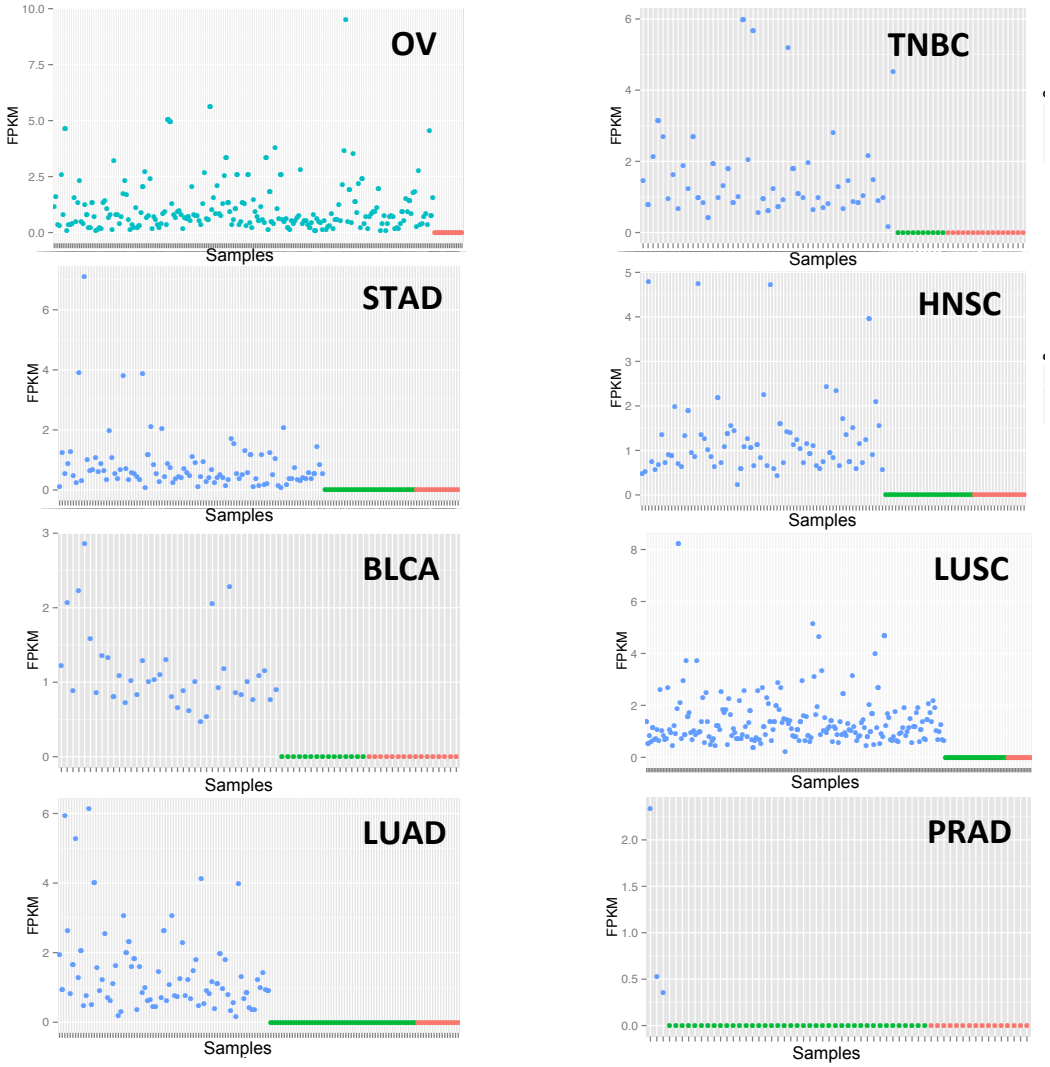


Figure 3-15. Estimated level of expression for matriptase variant A1. The x-axis represent samples that express matriptase variant A1 (Skipping exon 12). The expression in tumour samples is shown in blue. There is no evidence for matriptase novel transcript A1 in adjacent non-cancerous tissue from TCGA (shown in green with FPKM equal to zero) nor in the transcriptome data available from the GTEx and BodyMap 2.0 project (shown in red with FPKM equal to zero). The expanded form of each tumour type abbreviation is available in Table 3-9.

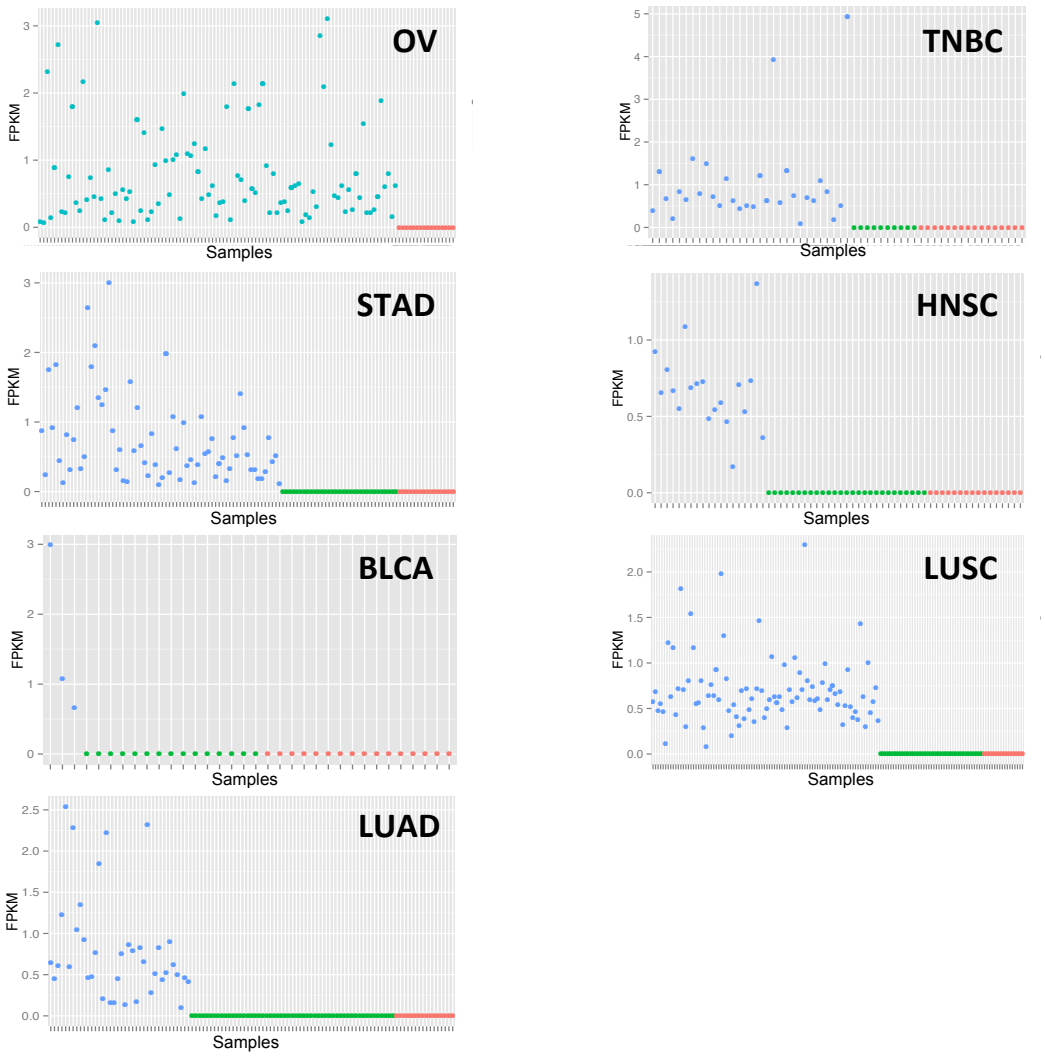


Figure 3-16. Estimated level of expression for matriptase variant A3. The x-axis represent samples that express matriptase variant A3 (Skipping exon 14). The expression in tumour samples is shown in blue. There is no evidence for matriptase novel transcript A3 in adjacent non-cancerous tissue from TCGA (shown in green with FPKM equal to zero) nor in the transcriptome data available from the GTEx and BodyMap 2.0 project (shown in red with FPKM equal to zero). The expanded form of each tumour type abbreviation is available in Table 3-9.

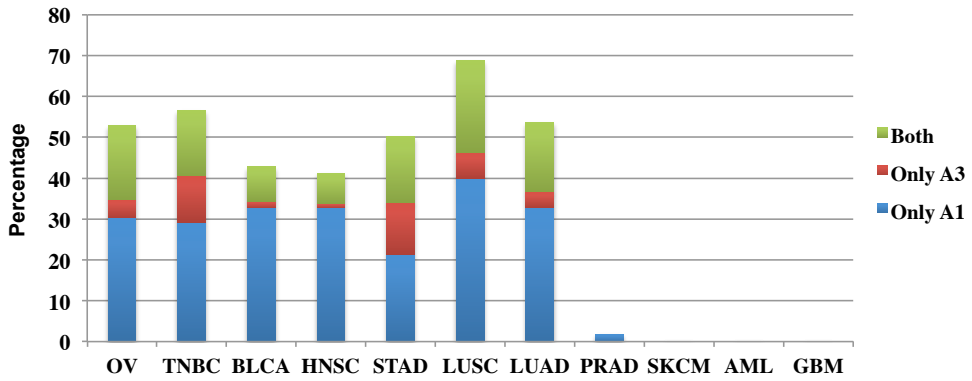


Figure 3-17. Frequency of novel matriptase novel AS transcripts. Samples expressing matriptase novel transcripts were divided into three groups: (1) expressing transcript A1, (2) expressing transcript A3, and (3) expressing both A1 and A3 transcripts. Transcript A3 was not detected in prostate cancer samples. The expanded form of each tumour type abbreviation is available in Table 3-9.

The human matriptase gene is located on chromosome 11 (q24-25), spanning a genomic region of 50 kilobases. It is comprised of 19 exons (NCBI reference sequence GeneBank: NM_021978), and codes for a protein containing 855 amino acids. The nucleotide sequence analysis revealed that A1 was produced as a result of skipping exon 12. Similarly the A3 deletion occurred by skipping exon 14. Analysis of predicted protein sequences revealed both matriptase variants contain fully functional open reading frames, suggesting the possibility of expressing two novel proteins (Figures 3-18 and 3-19). Protein domain prediction further demonstrated that matriptase variants A1 and A3 lack LDLRA1 and LDLRA3 domains, respectively. Pairwise protein sequence alignment versus wild-type matriptase showed that the predicted protein for A1 transcript skips amino acids 452 to 487 followed by occurrence of an amino acid arginine (R) through the resultant of a novel exon-exon junction (Figure 3-18). The protein product of A1 transcript contains 820 amino acids. The A3 transcript encodes a protein of 817 amino acids, which is the result of skipping amino acids 524 to 562 followed by substitution of a methionine (M) due to the formation of a novel exon-exon junction (Figure 3-19).

>Blastp_Wildtype_ST14_vs_A1
Range 1: 1 to 855

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
1657 bits(4290)	0.0	Compositional matrix adjust.	819/855(96%)	819/855(95%)	35/855(4%)

```
Query  1    MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA  60
            MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA
Sbjct  1    MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA  60

Query  61   VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV  120
            VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV
Sbjct  61   VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV  120

Query  121  KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM  180
            KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM
Sbjct  121  KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM  180

Query  181  LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA  240
            LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA
Sbjct  181  LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA  240

Query  241  HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS  300
            HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS
Sbjct  241  HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS  300

Query  301  YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY  360
            YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY
Sbjct  301  YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY  360

Query  361  PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS  420
            PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS
Sbjct  361  PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS  420

Query  421  NSNKITVRFHSDQSYTDTGFLAEYLSYDSSD----------------------------  451
            NSNKITVRFHSDQSYTDTGFLAEYLSYDSSD
Sbjct  421  NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH  480

Query  452  ------RCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK  505
                   CDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK
Sbjct  481  SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK  540

Query  506  SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK  565
            SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK
Sbjct  541  SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK  600

Query  566  DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID  625
            DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID
Sbjct  601  DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID  660

Query  626  DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP  685
            DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP
Sbjct  661  DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP  720

Query  686  AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL  745
            AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL
Sbjct  721  AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL  780

Query  746  PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT  805
            PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT
Sbjct  781  PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT  840

Query  806  RLPLFRDWIKENTGV  820
            RLPLFRDWIKENTGV
Sbjct  841  RLPLFRDWIKENTGV  855
```

Figure 3-18. Pairwise sequence alignment of wild-type and A3 matriptase transcripts

>Blastp_Wildtype_ST14_vs_A3
Range 1: 1 to 855

Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
1654 bits(4284)	0.0	Compositional matrix adjust.	816/855(95%)	817/855(95%)	38/855(4%)

```
Query  1    MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA  60
            MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA
Sbjct  1    MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA  60

Query  61   VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV  120
            VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV
Sbjct  61   VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV  120

Query  121  KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM  180
            KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM
Sbjct  121  KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM  180

Query  181  LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA  240
            LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA
Sbjct  181  LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA  240

Query  241  HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS  300
            HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS
Sbjct  241  HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS  300

Query  301  YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY  360
            YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY
Sbjct  301  YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY  360

Query  361  PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS  420
            PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS
Sbjct  361  PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS  420

Query  421  NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH  480
            NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH
Sbjct  421  NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH  480

Query  481  SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGC-----------------  523
            SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGC
Sbjct  481  SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK  540

Query  524  ----------------------MNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK  562
                                 +NVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK
Sbjct  541  SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK  600

Query  563  DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID  622
            DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID
Sbjct  601  DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID  660

Query  623  DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP  682
            DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP
Sbjct  661  DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP  720

Query  683  AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL  742
            AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL
Sbjct  721  AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL  780

Query  743  PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT  802
            PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT
Sbjct  781  PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT  840

Query  803  RLPLFRDWIKENTGV  817
            RLPLFRDWIKENTGV
Sbjct  841  RLPLFRDWIKENTGV  855
```

Figure 3-19. Pairwise sequence alignment of wild-type and A3 matriptase transcripts

Matriptase splice variants are novel and tumor-associated

To search for AS information for matriptase, I performed literature searches using PubMed, OMIM, and other databases of AS including the AS and Transcript Discovery database (ASTD) (Koscielny et al., 2009). In addition, I searched publicly available EST and mRNA databases including GeneBank, Ensembl, dbEST, and Unigene. My search did not find these novel matriptase variants. I only found three AS transcripts of matriptase, which are formed as result of an intron retention event (Ensembl ID: ENST00000530532, ENST00000524718, and ENST00000530376). Furthermore, I did not detect the novel transcripts of matriptase in adjacent non-cancerous tissue from TCGA nor in the transcriptome data available from GTEx and BodyMap 2.0 project, thus suggesting these variants are tumour-associated.

qRT-PCR analysis confirms differential expression of novel matriptase transcripts in epithelial-derived tumours

To validate the expression of matriptase splice variants in epithelial tumours, a matriptase wild-type or splice variant-specific probes was designed to perform qRT-PCR (supporting methods, section 3.2.4). qRT-PCR was carried out on orthogonal panels of cell lines and human primary and metastatic tumour tissue from ovarian, breast, lung and bladder cancer and a panel of normal tissue. The normal panel includes 48 healthy tissues (Supporting methods, section 3.2.4) and normal ovary, lung, bladder and breast. We measured changes in the gene expression by comparing the threshold cycle (Ct) of PCR product detection normalized against a reference gene transcript. The expression levels detected by qRT-PCR for wild-type matriptase and its splice variants showed that wild-type matriptase was the predominant transcript in both tumour and normal tissues (p-value < 0.0001). A1 transcript was overexpressed in tumour samples compared to normal tissues for ovarian (p-value < 0.0001) and lung panels (p-value = 0.0082). However, this did not apply to the bladder (p-value = 0.6414) and breast (p-value = 0.6466) panels. We also investigated the expression level of A3 splice variant in a panel of ovarian tissues and cell lines. A3 was overexpressed in ovarian tumours compared to normal samples (p-value = 0.0004). However, we observed lower expression of A3 transcript compared to A1 in ovarian tumours (p-value = 0.0004).

We further tested the expression of matriptase splice variants in a panel of normal tissue samples including 48 normal tissues from across the human body. Both matriptase splice variants A1 and A3 showed higher expression in tumour samples compared to the normal tissue panel (p-value < 0.0001). In fact, the majority of tissues in the normal tissue panel did not express matriptase A1 and A3 transcript variants at all, while a small number showed a much lower expression compared to tumour samples (Figure 3-20). That is, the A1 and A3 transcripts were detected only in 16 and 17 out of the 48 normal tissues in the normal tissue panel, respectively.

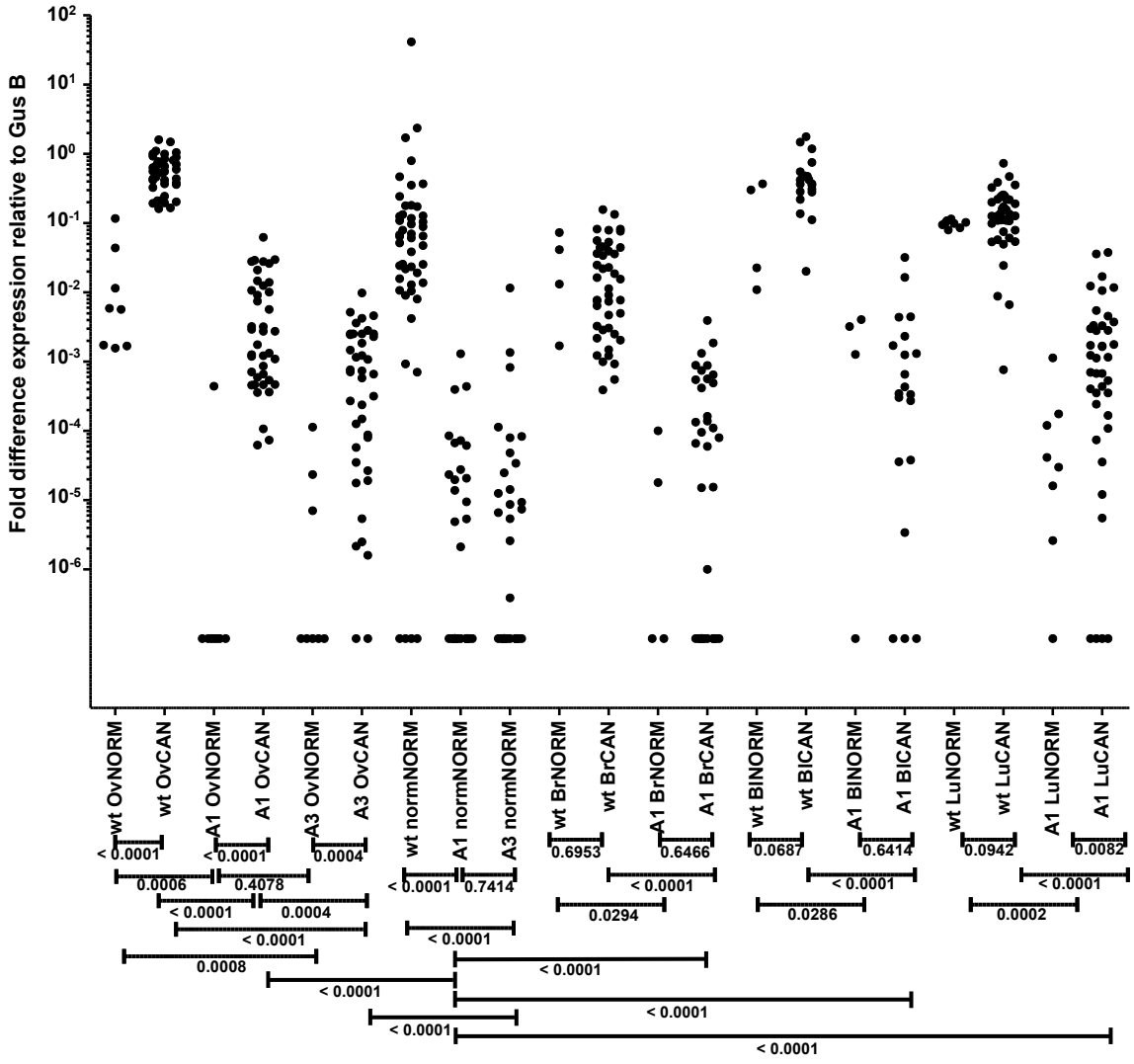
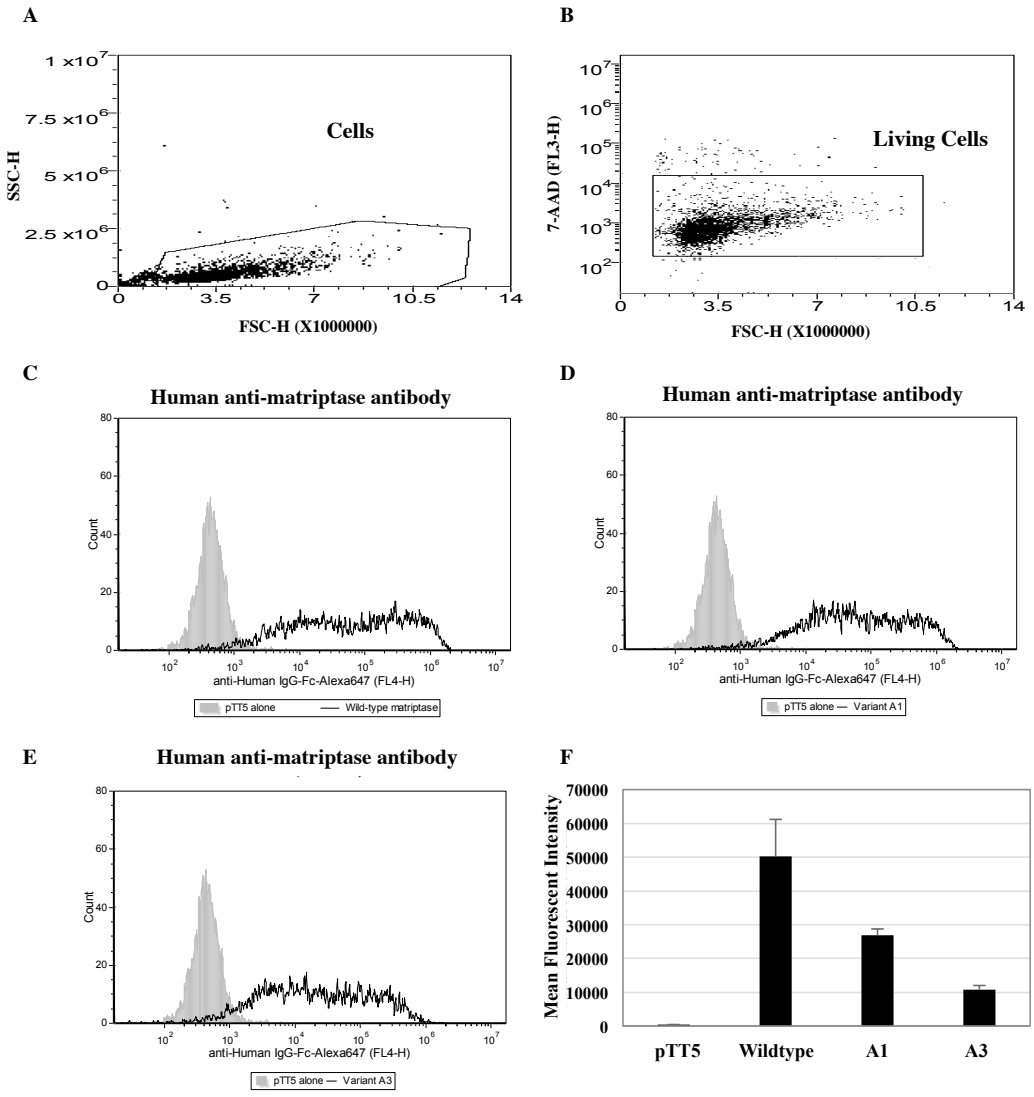


Figure 3-20. qRT-PCR validation. qRT-PCR was carried out on orthogonal panels of cell lines and human primary and metastatic tumor tissues from ovarian, breast, lung, and bladder cancer and a panel of normal tissues. Mann–Whitney t-test was used to determine significant differences in gene expression between groups. The resulting P-values are summarized below the x-axis. The x-axis labels from left to right are (1) wild type in normal ovary, (2) wild type in ovarian cancer, (3) A1 in normal ovary, (4) A1 in ovarian cancer, (5) A3 in normal ovary, (6) A3 in ovarian cancer, (7) wild type in normal tissue panel, (8) A1 in normal tissue panel, (9) A3 in normal tissue panel, (10) wild type in normal breast, (11) wild type in breast cancer, (12) A1 in normal breast, (13) A1 in breast cancer, (14) wild type in normal bladder, (15) wild type in bladder cancer, (16) A1 in normal bladder, (17) A1 in bladder cancer, (18) wild type in normal lung, (19) wild type in lung cancer, (20) A1 in normal lung, and (21) A1 in lung cancer. The y-axis is log scaled.

Matriptase splice variants can be translocated to the surface of transfected CHO cells

To address the question of whether matriptase A1 and A3 transcripts yield protein variants that are capable of being translocated to the cell surface, transiently transfected CHO cells with cDNA encoding these genes were developed, followed by flow cytometric analysis of surface matriptase proteins (wild-type, variant A1 and variant A3) (Supporting methods, section 3.2.4). For this experiment, a human anti-matriptase antibody was used that binds to the catalytic domain of all three matriptase variants and is not variant specific. Co-expression of the matriptase variants with HAI-1 resulted in a significant increase in the mean fluorescent intensity for wild-type, variant A1 and variant A3 (p-value < 0.05; Figure 3-21 sections C-F), whereas expression of matriptase variants alone showed modest increases in surface expression (data not shown). So to verify that the recombinant proteins detected by flow cytometry were the expected molecular weight for each variant, matriptase variants were immunoprecipitated from transfected CHO cells using the same human anti-matriptase antibody and analysed by Western blot (Figure 3-21 section G) (Supporting methods, section 3.2.4). As observed in the flow cytometry experiment, endogenous matriptase was not detected in the elution from CHO cells transfected with the empty vector alone. In contrast, bands corresponding to the expected molecular weight for each variant were detected in the respective elutions. These results support the assertion that proteins corresponding to the expected molecular weight of matriptase variant A1 and A3 are trafficked to the cell surface of transiently transfected cells despite the deletion of the LDLRA domains.


A
Cells
SSC-H
FSC-H (X1000000)
B
Living Cells
7-AAD (FL3-H)
FSC-H (X1000000)
C
Human anti-matriptase antibody
Count
anti-Human IgG-Fc-Alexa647 (FL4-H)
pTT5 alone
Wild-type matriptase
D
Human anti-matriptase antibody
Count
anti-Human IgG-Fc-Alexa647 (FL4-H)
pTT5 alone
Variant A1
E
Human anti-matriptase antibody
Count
anti-Human IgG-Fc-Alexa647 (FL4-H)
pTT5 alone
Variant A3
F
Mean Fluorescent Intensity
pTT5
Wildtype
A1
A3

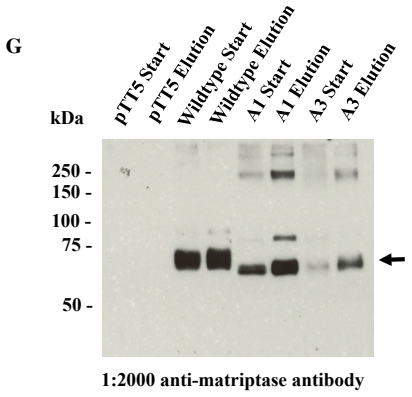

G
pTT5 Start
pTT5 Elution
Wildtype Start
Wildtype Elution
A1 Start
A1 Elution
A3 Start
A3 Elution
kDa
250 -
150 -
100 -
75 -
50 -
1:2000 anti-matriptase antibody

Figure 3-21. Flow cytometric analysis reveals surface expression of matriptase splice variants. Cells were transfected with 10 µg of empty vector alone (pTT5) or 5µg of each matriptase variant plus 5µg of HAI-1 (A-G). The next day, duplicate wells containing 100,000 cells/well were stained with either human anti-matriptase or mouse anti-SPINT1 (HAI-1) antibodies (data not shown) followed by species specific secondary Alexa Fluor® 647 Goat anti-IgG-Fc antibodies plus the live/dead cell discriminator 7-AAD followed by flow cytometric analysis. The gating tree is as follows: (A) SSC vs. FSC depicts the distribution of cells as opposed to the debris that was excluded; to (B) living cells not stained with 7-AAD. (C) wildtype matriptase, (D) matriptase variant A1, and (E) matriptase variant A3 (F) graph depicting the mean fluorescent intensity plus/minus the standard error of mean of matriptase expressed on the surface of CHO cells. This data is representative of 3 independent experiments analyzed with a student's t-test (p-value < 0.05). Flow cytometry data was acquired on an Intellicyte® HTFC, which uses an Accuri® C6 Flow Cytometer® (BD Biosciences) with the sip time set at 3 seconds. Laser lines for this instrument are 488nm and 640nm. FL3 emission detection for 7-AAD is >670nm, and FL4 emission detection for Alexa Fluor® 647 is 675/25nm. (G) Recombinant wildtype, A1 and A3 variants were immunoprecipitated with 1.5µg of human anti-matriptase antibody, followed by Western blot analysis on the clarified start lysates (20µg each) and elutions (15µl each). The arrow shows the bands corresponding to the expected size of each matriptase variant.

In this study, I introduced an AS-detection pipeline, and used it toward identification of novel AS variants in TCGA tumours. This analysis revealed two novel tumour-associated splice variants of matriptase, which were confirmed in an orthogonal set of tumour tissues and cell lines. Matriptase variants are highly frequent (up to 69% in lung cancer) among patients with epithelial-derived tumours with low or no occurrence in normal tissue. In addition to gene expression data, the flow cytometric analysis confirmed protein expression of both matriptase variants on the surface of CHO cells, suggesting matriptase variants as potential biomarkers of tumour cells. Clinical validation would prove valuable in confirming the utility of matriptase variants for therapeutic use.

No splice-sites mutation associated with skipping exons 12 and 14 of matriptase was identified in TCGA mutation analysis data derived from matching whole-exome sequencing dataset. This analysis was done online through cBioPortal website (http://www.cbioportal.org/), which allows visualization and analysis of available TCGA datasets. Furthermore, no correlation ($p>0.05$) between expression of matriptase variants and patient's survival time, age, tumour size, tumour clinical stage and histological grade were identified. Table 3-10 shows this analysis in ovarian cancer.

Table 3-10. Relationship between matriptase splice variants and clinicopathological data in ovarian serous cystadenocarcinoma. Clinicopathological data was downloaded from the TCGA data portal (http://cancergenome.nih.gov).

	Number of A1 positive (percentage)	Number of A1 negative (percentage)	p-value	Number of A3 positive (percentage)	Number of A3 negative (percentage)	P-value
Age						
> 50	164 (75)	178 (77)		75 (74)	267 (77)	
<= 50	55 (25)	54 (23)	0.6613	27 (26)	82 (23)	0.599
Clinical Stage						
IV	31 (14)	41 (18)		15 (15)	57 (16)	
IIA-IIC	15 (7)	7 (3)		9 (9)	13 (4)	
IIIA-IIIC	171 (79)	183 (79)	0.1224	78 (76)	276 (80)	0.1278
Histological Grade						
G1-G2	28 (13)	26 (11)		11 (11)	43 (12)	
G3-G4	186 (86)	202 (87)		89 (87)	299 (86)	
GB, GX	3 (1)	4 (2)	0.8152	2 (2)	5 (2)	0.7847
Tumour Size						
1-10 mm	103 (64)	109 (67)		55 (76)	157 (63)	
11-20 mm	13 (8)	17 (11)		6 (9)	24 (10)	
> 20 mm	44 (28)	36 (22)	0.4747	11 (15)	69 (27)	0.0783

3.2.4. Supporting methods

The qRT-PCR validation of matriptase splice variants

Reverse transcription reaction was performed using commercially available sets of human normal tissue or ovarian, breast, lung and bladder cancer cDNA (OriGene Technologies), as well as cDNA synthesized from RNA isolated from ovarian cell lines including OvCAR3, CaOV3, UACC-1598, Ov-90 and triple negative breast cell lines including MDA-MB-231, MDA-MB-468 and HCC 1937. All cells lines were cultured under ATCC recommended culture conditions.

Primer and probes in designed splice variant assays were all tested at varying concentrations and with the use of positive controls to ensure efficiency of 90-110%, and to determine assay range of detection. Positive amplification controls were also run against their counterpart reaction to ensure that amplification was specific to the designed splice variant or wild-type assay. All designed assays were run as duplex reactions using GusB as a reference gene. PCR amplification was performed for 40 cycles. Matriptase-Exon12 (A1) RT-PCR experiments utilized 300nM final concentration of Forward (GAC ACC GGC TTC TTA GCT GAA T) and Reverse (GAA GAG GGG CTT GCA GAA CTT G) primers, 100nM of either Exon 12 wildtype (/56-FAM/TCC AGT GAC /ZEN/CCA TGC CCG GG/3IABkFQ) or Exon 12 splice variant mutant /56-FAM/CAG TGA CCG /ZEN/TTG CGA CGC CG/3IABkFQ probes, and human commercial available assay Hs00939627_m1 (Applied Biosystems, USA) for GusB reference gene in 1X master mix reaction. Matriptase-Exon14 (A3) RT-PCR were performed identically, but utilizing different sequences for Forward (GAA CGA CTG CGG AGA CAA CA) and Reverse (TGC TCA AGC AGA GCC CAT T) primers and wild-type (/56-FAM/TCC GGC CCA /ZEN/GAC CTT CAG GTG TT/3IABkFQ/) or Exon 14 variant (/56-FAM/AGT GAC GAC /ZEN/GTT CAT GCA CCC CTG /3IABkFQ/) probes. The GusB reference gene was used to normalize data for RT-PCR analysis and to estimate the relative fold change for each sample, similar to the approach taken by Beillard et al (Beillard et al., 2003). Since the reference gene is exposed to the same preparation steps as the gene of interest, this approach adjusts expression of target gene for differences in experimental condition, which are not a result of experimental design and allows for an exact comparison of

mRNA transcription level between different samples. To assemble each reaction, 3ng cDNA or control RNA in 5µL volume was used as input template sample. 15µL reaction master mix was added to each well for a final 20µL reaction mixture in each well of the RT-PCR plates, and read on real-time PCR platform 7900HT FAST RT-PCR with SDS 2.3 software (Applied Biosystems, USA). Each master mix reaction was made to final 1X target gene (FAM) reaction, 1X GusB (VIC) in 1X master mix reaction. RT-PCR cycle conditions were set to ΔΔCt plate setting in 96-well FAST format, using standard RT-PCR cycle settings: 2mins at 50°C, 20secs at 95°C and a minimum of 40 cycles of 1sec at 95°Cand 20secs at 60°C. ΔΔCt plate setting data files were loaded into ABI RQ Manager software, and Ct values that were automatically generated were used to calculate ΔCt values for each reaction. Samples that did not have amplification as detected by ABI RT-PCR platform software were indicated as having levels of transcript below the limit of detection. Samples were grouped by cell lines, cancer subtypes, or normal tissue, and graphed using GraphPad Prism software version 5.0 (GraphPad Software Inc).

Transfection constructs

Total RNA was isolated with the RNeasy Mini Kit (Qiagen) from MDA-MB-468 and HCC 1937 cells to generate cDNA encoding HAI-1 and wild-type matriptase, respectively. cDNA was generated as per manufacturer's instructions using SuperScript® III Reverse Transcriptase (Life Technologies) and Oligo$(dT)_{18}$ primer (Thermo Fisher Scientific). HAI-1 and wild-type matriptase were amplified from the above cDNA using Q5® Hot Start High-fidelity DNA Polymerase (New England Biolabs). *SPINT1* (Accession# GeneBank: NM_181642.2) encoding HAI-1 was cloned into the pTT5 vector (National Research Council of Canada, Biotechnology Research Institute) using Gibson Assembly® (New England Biolabs) as per manufacturer's instructions. The HAI-1 forward primer was 5'-aaacggatctctagcgaattcgccaccATGGCCCCTGCGAGGACG-3' and the reverse primer was 5'-aggtcgaggtcgggggatccTCAGAGGGGCCGCGTGGT-3'. Lower case letters correspond to the pTT5 vector, upper case to HAI-1, and the start/stop codons are underlined. Wild-type matriptase (aka *ST14*; Accession# GeneBank: NM_021978) was cloned into the pTT5 vector using 5' EcoRI and 3' BamHI restriction sites (in bold) incorporated into the *ST14* forward (5'-GATC**GAATTC**GCCACCATGGGGAGCGATCGGGCCCGCAA-3') and *ST14* reverse

primer (5'-GATC**GGATCC**CTATACCCCAGTGTTCTCTTTGATCCAGTCCC-3'). The exon 12 deletion (variant A1) was introduced by amplifying the regions 5' and 3' to the exon deletion from the wild-type matriptase cDNA using the *ST14* forward and reverse primers. The *ST14* forward primer was paired with the variant A1 reverse primer (5'-CCGGCGTCGCAACGGTCACTGGAGTCGTAGGAGAG-3') to amplify the region 5' to the exon deletion, and the *ST14* reverse primer was paired with the variant A1 forward primer (5'-ACTCCAGTGACCGTTGCGACGCCGGCCACCAGTT-3') to amplify the region 3' to the exon deletion. The variant A1 primers introduced an overhang depicted by the underlined sequence. Equimolar amounts of the above 5' and 3' PCR products were added to a PCR reaction as the template, along with the *ST14* forward and reverse primers to produce a full-length construct with the region corresponding to exon 12 deleted. The exon 14 deletion (variant A3) was constructed the same way using the variant A3 forward primer (5'-AGCAGGGGTGCATGAACGTCGTCACTTGTACCAA-3') and the variant A3 reverse primer (5'-TGACGACGTTCATGCACCCCTGCTCGTCGCTGTT-3'). All constructs were verified by DNA sequencing.

Cell culture conditions, and transfection

CHO-K1 cells (ATCC) were maintained in Ham's F-12 media (Life Technologies) supplemented with 10% Fetal Bovine Serum (FBS; Life Technologies) at 37°C and 5% CO_2. The day before transfection $2.5x10^6$ cells per a plate were seeded in the above media on four 10cm plates for each transfection. The four transfections consisted of empty pTT5 vector alone, wild-type plus HAI-1, variant A1 plus HAI-1 and variant A3 plus HAI-1. Twenty-four hours later, each transfection was performed by mixing a total of 10µg of cDNA into 500µl of Opti-MEM® (Life Technologies), and 30ug of Polyethylenimine (PEI) Max (Polysciences, Inc.) into another tube with 500µl of Opti-MEM®. The two tubes were incubated at room temp for 5 minutes, and then the PEI Max solution was added to the cDNA solution followed by a 25 minute incubation. PEI/cDNA complexes were added drop-wise to the 10cm plate while swirling/rocking to mix, and the cells were returned to the incubator.

Flow Cytometry

Twenty-four hours after transfection, the plates were washed once with PBS (Life Technologies), and the cells were dissociated from the plate with non-enzymatic cell dissociation solution (Sigma-Aldrich). After 15 minutes at 37°C, the cells were collected by pipetting up and down in PBS plus 1% FBS (PBS/FBS), counted on a ViCell™, and resuspended in PBS/FBS. Cells were added to a 96 well plate, spun at 400xg for 5 minutes, and resuspended in 5 µg/ml of human anti-matriptase or 10µg/ml of mouse anti-*SPINT1* (OriGene Technologies). Isotype controls were also prepared for each transfection with 5µg/ml of human IgG1 Kappa (Sigma-Aldrich) or 10µg/ml of mouse IgG1 Kappa (eBioscience). After a 1 hour incubation on ice, cells were washed 2 times in ice-cold PBS/FBS and resuspended in PBS/FBS containing 2.5µg/ml of 7-Aminoactinomycin D (7-AAD; Sigma-Aldrich) plus 2µg/ml of either Alexa Fluor® 647 Goat anti-human IgG-Fc (Jackson ImmunoResearch Labs, Inc.) or Alexa Fluor® 647 Goat anti-mouse IgG-Fc (Jackson ImmunoResearch Labs, Inc.). Cells were incubated for 30 minutes on ice in the dark, then washed 2 times in PBS/FBS and resuspended in PBS/FBS. Data was acquired with an Intellicyte® High Throughput Flow Cytometer (HTFC) that consisted of an Accuri® C6 Flow Cytometer® (BD Biosciences), CFlow® Software (version 1.0.227.4), HyperCyt® CFlow Automator (version 3.4.0.0) and HyperView iDM® Client Edition 4.0 (R2 version 4.0.4395). Analysis was carried out using the CFlow® Software (version 1.0.227.4) and FCS Express 4 Professional Standalone Research Edition with histogram smoothing set to 1 (*De Novo* Software™, version 4.07.0014).

Immunoprecipitation and Western Blot Analysis

The immunoprecipitation was performed as described by Swayze et al. with the following modifications (Swayze & Braun, 2001). Unless otherwise stated, all reagents were purchased from Sigma. As outlined above for the flow cytometry experiment, HAI-1 plus either wildtype, A1 or A3 transfected CHO-K1 cells were dissociated from 10 cm plates with non-enzymatic dissociation solution, and collected by pipetting up and down in PBS alone. Cells were spun for 5 minutes at 400xg and the supernatant was aspirated. Pellets were resuspended in 0.5-1ml of ice cold lysis buffer [50mM Tris-HCl, pH 7.4, 150mM NaCl, 1% Triton X-100, 0.1% sodium dodecyl sulfate (SDS), 1mM

$CaCl_2$, 1mM $MgCl_2$ and one Complete mini EDTA-free protease inhibitor cocktail tablet (Roche) per 10ml of buffer]. While on ice, the cells were broken open with 10 strokes of the pestle using a pestle and microtube set (VWR), and then the lysate was passed through a 26 gauge syringe 10 times to shear the DNA. DNase was added to 10µg/ml and the lysates were gently rotated at 4°C for 30 minutes. Lysates were clarified by centrifugation at 20,000xg for 10 minutes at 4°C and supernatant was subjected to a BCA protein concentration assay (Pierce). Clarified lysates were adjusted to 1mg/ml in 1ml (Figure 3-21 5G "start"). 40µl of a 50% slurry of Protein G Sepharose Fast Flow beads (GE Healthcare) pre-equilibrated in lysis buffer was added followed by rotation at 4°C for 1-2 hours to pre-clear the lysate. The beads were removed by centrifugation at 2500xg for 2.5 minutes at 4°C, and the pre-cleared lysate was transferred to a new 1.7ml tube. 1.5ug of human anti-matriptase antibody was added followed by rotation for 14-16 hours at 4°C. Matriptase-antibody complexes were then rotated with 40µl of the above Sepharose bead preparation for another 2 hours at 4°C. The beads were washed three times in 1ml of ice cold lysis buffer by centrifuging at 2500xg for 2.5 minutes at 4°C followed by supernatant aspiration. The beads were resuspended in non-reducing Laemmli sample buffer (Laemmli, 1970), and heated at 95°C for 5 minutes to dissociate the matriptase-antibody-bead complex. The beads were removed by centrifugation using a custom-made spin column, and the proteins (Figure 3-21 section G "elution") were separated by SDS-polyacrylamide gel electrophoresis in 1X Tris/Glycine/SDS buffer (Bio-Rad). The resolved proteins were electrotransferred to 0.45µm nitrocellulose membrane (Bio-Rad) at 100 volts for 90 minutes in 1X Tris/Glycine buffer with 20% Methanol (Towbin, Staehelin, & Gordon, 1979) (Bio-Rad). The nitrocellulose was air dried to fix the proteins, and then subjected to Western blot analysis as described (Swayze & Braun, 2001)The primary rabbit anti-matriptase antibody was used at 1:2000 (Millipore) and the secondary anti-rabbit conjugated horseradish peroxidase was used at 1:50000 (GE Healthcare). Proteins were detected with SuperSignal West Dura Chemiluminescent substrate (Pierce) and exposed to Amersham Hyperfilm (GE Healthcare).

3.3. Identification and prioritization of optimal therapeutic targets

With high-throughput studies often producing long lists of genes and proteins of interest, an approach is needed to narrow down such lists by ranking and prioritizing the candidates. Analytic hierarchy process (AHP), developed by T. Saaty, is one of the best known multiple criteria decision-making (DM) techniques (Saaty, 1977) and has been widely used around the world in a variety of decision situations (Liberatore & Nydick, 2008; Subramanian & Ramanathan, 2012; Vaidya & Kumar, 2006). It offers an objective way to reproducibly narrow down a long list of candidates thorough prioritization using a series of user-specified preferences.

The AHP algorithm (Saaty, 1977) provides a rational framework to decompose a problem into a hierarchy of sub-problems, which can be more easily comprehended and evaluated. This hierarchical structure may include the goal, objectives (criteria and sub-criteria), and alternatives (candidates to be ranked) (Saaty, 1980). Once the hierarchy is built, decision elements can be evaluated to obtain their relative importance to achieve the final goal. Then, these evaluations are converted into numerical values and processed to rank each candidate on a numerical scale. The AHP approach is described below using a simple step-by-step example:

An example of a simple decision: determining a thesis topic

Assume a scenario that a graduate student is looking for a topic for her thesis project. She is planning to use AHP to make her decision. The methodology of the AHP can be explained in following steps:

Step 1. Defining the problem and determining the kind of knowledge sought. Here, user defines the problem as selecting a topic for her thesis project amongst three topics A, B, and C (alternatives). She considers (1) research cost, (2)

level of attractiveness, and (3) how fast it is possible to finish the project as the criteria to make her decision. This information is summarized in Figure 3-22 A.

Step 2. Decomposing the problem into a hierarchy of goal, objectives, and alternatives. Structuring the decision problem as a hierarchy is fundamental to the process of the AHP. Hierarchy indicates the relationship between decision elements in one level of hierarchy with those of the level immediately below. Figure 3-22 B illustrates a decision hierarchy, where the first level includes the goal, second level illustrates decision objectives (i.e. criteria), and the last level (leaf nodes) are the alternatives to be ranked (i.e. three thesis topics).

Step 3. Pairwise evaluation of decision elements. AHP uses pairwise comparisons to determine the relative importance of decision elements. Each element in an upper level is used to compare the elements in the level immediately below with respect to it. For example, criteria are evaluated in terms of their importance to achieve the goal. While, alternatives are evaluated with respect to their immediate upper criterion in the problem hierarchy. Therefore, in the current example, thesis topics are required to be pairwise evaluated once per each criterion including research cost (Figure 3-22 D), attractiveness (Figure 3-22 E), and time to finish (Figure 3-22 F). Similarly, the criteria are required to be evaluated based on their importance to achieve the goal, which is choosing a thesis topic (Figure 3-22 C). To make pairwise comparisons, AHP method offers a numeric scale that indicates how many times more important or dominant one element is over another element. Table 3-12 exhibits this scale. For example, here the user has determined that the research cost is three times more important than the required time to finish the project (Figure 3-22 C), while the topic's attractiveness is three times more important than the research cost (Figure 3-22 C).

Step 4. Constructing pairwise comparison matrices. The pairwise comparisons obtained in previous step are organized into a square matrix named pairwise comparison matrix (PCM). In this matrix, the diagonal elements are equal to 1. If the decision element in the row *i* is better than decision element in the column *j,* the value of *(i, j)* entry in the matrix is more than 1; otherwise the decision element in the column *j* is better than the one in the row *i*. In addition, the *(j, i)* entry of PCM is the

reciprocal of the *(i, j)* entry. PCMs of choosing a thesis topic are shown in Figures 3-22 C to F.

Step 5. Estimating local priorities. The principal eigenvalue and the corresponding normalised right eigenvector of a PCM give the relative importance of the decision elements being compared (Saaty, 1977). The elements of the normalised eigenvector are known as weights. For the current example, the weights - which are also known as local priority - are shown in Figure 3-22 C to F.

Step 6. Estimating the consistency of pairwise comparisons. The consistency of PCMs can be examined through the estimation of consistency index (CI).

$$\text{CI} = {}^{(\lambda_{max} - n)}/_{n-1}$$

Where;

λ_{max} is the maximum eigenvalue of the comparison matrix.

This value is then used to compute Consistency Ratio (CR), which indicates the amount of allowed inconsistency in a decision matrix.

$$\text{CR} = {}^{CI}/_{RI}$$

Where;

Random Index (RI) is the average CI value of randomly-generated comparison matrices (PCMs) using Saaty's preference scale.

Saaty suggests the value of CR should be less than 0.1 (Saaty, 1977). Although AHP tolerates some inconsistency due the amount of redundancy in the approach, pairwise comparison may be re-examined if the CR fails (greater than 0.1).

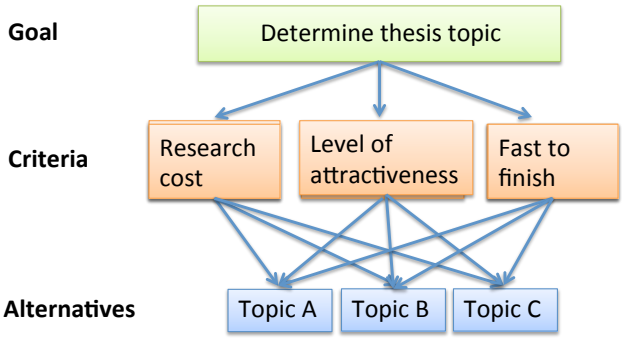
The CR of the PCMs shown in Figures 3-22 is as following: (C) 0.036, (D) 0.067, (E) 0.00, and (F) 0.0041.

Step 7. Prioritization. In order to compute final ranking, the local priority of each alternative is multiplied by the weight of the immediate upper level criterion to get global priorities. Once the global priorities in each level of hierarchy are determined, for each alternative the weighted values can be add up to obtain the overall priority. The calculated overall priority demonstrates how an alternative contributes to the goal. The final ranking of thesis topics is shown in Figure 3-22 G.

A

Goal	Criteria	Alternatives
Determine thesis topic	1. Research cost 2. Level of attractiveness 3. Fast to finish	List of thesis topics

B



C Criteria PCM

	Research cost	Level of attractiveness	Fast to Finish	Local priority
Research cost	1	1 / 3	3	0.258
Level of attractiveness	3	1	5	0.637
Fast to Finish	1 / 3	1 / 5	1	0.105

D Research cost PCM

	Topic A	Topic B	Topic C	Local priority	Global priority
Topic A	1	3	7	0.65	0.168
Topic B	1 / 3	1	5	0.28	0.072
Topic C	1 / 7	1 / 5	1	0.07	0.018

E Level of attractiveness PCM

	Topic A	Topic B	Topic C	Local priority	Global priority
Topic A	1	1	3	0.4285	0.273
Topic B	1	1	3	0.4285	0.273
Topic C	1 / 3	1 / 3	1	0.143	0.091

F Fast to finish PCM

	Topic A	Topic B	Topic C	Local priority	Global priority
Topic A	1	1 / 2	3	0.31	0.032
Topic B	2	1	5	0.58	0.061
Topic C	1 / 3	1 / 5	1	0.11	0.012

G Final AHP prioritization

	Research cost	Level of attractiveness	Fast to finish	Total priority
Topic A	0.168	0.273	0.032	**0.473**
Topic B	0.072	0.273	0.061	0.406
Topic C	0.018	0.091	0.011	0.12
Totals	0.258	0.637	0.105	1

Figure 3-22. A step-by-step example of AHP relative model. (A) Determining the problem goal, objectives and alternatives. (B) Building the problem hierarchy. (C) Constructing PCM for decision criteria with respect to the goal. (D-F) Constructing alternative PCMs with respect to their associated criteria. Table C illustrates the PCM of criteria and their local priorities. Tables D - F demonstrate the PCMs of alternatives with respect to (D) research cost, (E) level of attractiveness, and (F) fast to finish, respectively. In addition computed local and global priorities are shown in the last two columns. An alternative global priority is computed by multiplying the alternatives' local priority to the priority of its associated criterion. (G) Total priority values showing Topic A with a score of 0.473 is the alternative that contributes most to the goal than Topics B and C. The consistency ratio of PCMs C-F is as following; (C) 0.036, (D) 0.067, (E) 0.00, (F) 0.0041, respectively.

In the given example above, topic A with a final priority of 0.473 is the alternative that contributes the most to the goal of choosing a thesis topic with respect to the three criteria including cost, attractiveness, and time to finish. Topic B is a close second with a priority of 0.406.

In a decision problem with n alternatives, $n(n\text{-}1)/2$ comparisons are required to build a PCM. Hence, when the number of alternatives is large or if the possibility of adding or deleting alternatives exists, using pairwise comparisons (AHP relative) is not practical. In this case an AHP rating approach is often used (Saaty, 2008). This approach requires a series of categories/intensities to be established for each criterion. For instance, in the above example, the attractiveness criterion can be broken down into the following categories: very interesting, interesting, and not interesting. Next, these categories are pairwise compared and their priorities with respect to their associated criterion (e.g. attractiveness) are obtained (instead of pairwise comparing the alternatives). Then, for each criterion, alternatives are evaluated and weighted by selecting the appropriate category that they fall into.

As shown in an example above, AHP offers a simple yet powerful technique in which to rank alternatives and express preference. Using this approach, a user is required to only provide two sets of information: (1) The problem hierarchy - i.e. breaking a problem into smaller sub-problems, each of which may be easier to solve, and (2) PCMs – i.e. expressing her preference of decision elements in a pairwise manner. Then, AHP computes a ranking score for each alternative using this information.

Prioritization with the AHP method depends on the available knowledge about the decision alternatives. Similarly, the design of the problem hierarchy - the choice of the alternatives, criteria, and sub-criteria, as well as their weights – can affect the final ranking (Saaty, 2008). On the other hand, AHP allows decision makers to select and define criteria/sub-criteria as it fits best to their research question. In addition, it is flexible enough to allow adding and/or removing decision elements. Other benefits of the AHP includes; 1) incorporating data and judgments of experts, 2) it is a valuable tool for solving problems with both quantitative and qualitative factors (Vaidya & Kumar, 2006). AHP has been successfully used in different fields and disciplines such as business,

industry, healthcare, and education (Liberatore & Nydick, 2008; Subramanian & Ramanathan, 2012; Vaidya & Kumar, 2006).

Most interestingly, the National Cancer Institute (NCI) highlighted the application of AHP in translational research by using it for prioritization of cancer antigens in order to provide a basis for deciding which antigens are most likely to generate successful cancer vaccine candidates for testing in later-stage clinical trials (Cheever et al., 2009).

As of 2014 that this analysis was in progress, there was no comprehensive AHP R package available. Therefore, in order to leverage AHP for bioinformatics applications, I have implemented the AHP technique as an R package along with multiple visualization tools for further analysis of the prioritization. Prize supports both AHP relative and rating models. Since November 2016 a second R implementation of the AHP method is available on CRAN at https://cran.r-project.org/web/packages/ahp/index.html. However, unlike Prize package this implementation does not offer group decision aggregation and does not support AHP rating model.

Prioritization using AHP approach offers unique advantages compared to other weight-based methods. This may include: (1) AHP uses a hierarchical structure which enables decision makers to define high level strategic objectives and specific metrics for a better assessment of alternatives, (1) It measures the level of inconsistency in pairwise comparisons and weightings, (3) It integrates quantitative and qualitative considerations and transforms them into numerical value, (4) AHP enables decision makers to measure the relative importance of alternatives, and (5) allows for group decision making where communication among team members is impeded by their different specializations.

3.3.1. Implementation

The purpose of Prize is to allow users to simplify complex problems into elementary hierarchy system and calculate alternatives' prior probabilities. Prize is an R implementation of the AHP algorithm, which allows users to evaluate the information

quantitatively and qualitatively using both subjective and objective ranking scales. Using Prize, the user is only required to decompose the decision problem into a hierarchy and evaluate its various elements by comparing them to each other in a pairwise manner, with respect to their impact on an element above them in the hierarchy (building PCMs). Prize uses this information to compute the final priorities and allows visualization of the ranking. Prize can be run on any platform with an existing R and Bioconductor installation. The package includes 10 functions (Table 3-11), which allow for simple prioritization and visualization of final rankings. Prioritization with Prize consists of three main steps:

Decomposing the problem into a hierarchy

A problem may define as a related set of sub-problems, which indicates the relationship among decision elements. Once the user breaks down the problem into a hierarchy of goal, objectives, and alternatives, the hierarchy can be visualized using ahplot(). This function takes a matrix that consists of two columns. The first column consists of the level of elements in the hierarchy and the second column consists of the name of the decision elements. Figure 3-23 shows a hierarchic structure.

Table 3-11. Prize Functions

Function	Description
Analysis tools	
gaggregate()	Aggregating individual judgements
rating()	Estimating alternative's rating value in AHP rating model
pipeline()	AHP analysis pipeline
ahp()	Computing AHP weights and CR
ahmatrix()	Converting a triangular matrix into a square PCM
Visualization tools	
crplot()	Plotting CR of individual judgements
dplot()	Illustrating the distance among individual judgements and aggregated group judgement
ahplot()	Plotting the problem hierarchy, showing the relationship among goal, objectives, and alternatives
wplot()	Plotting AHP weights in a bar/pie chart
rainbowplot()	Plotting prioritized alternatives in a color coded stacked bar plot

Building PCMs from individual and/or group judgements

In AHP methodology, each element in an upper level of problem hierarchy is used to compare the elements in the level immediately below with respect to it. Once the user performs the pairwise comparison of decision elements using the AHP scale (Table 3-12), these values can be organized into a square PCM. For an immediate evaluation, Prize offers an ahp() function that takes a PCM and reports the weight of decision elements and CR. In addition, prize offers a pipeline() function that takes in all the PCMs and performs the overall prioritization. This function is introduced in the next step.

AHP is an individual and group DM technique. In case of group DM, group members can either engage in discussion to achieve a consensus PCM or express their own preferences in form of individual PCMs. In case of latter, individual judgments can be aggregated in different ways to achieve a group PCM. Two of the methods that have been found to be most useful are the aggregation of individual judgments (AIJ) and the aggregation of individual priorities (AIP) (Forman & Peniwati, 1998). These methods perform the aggregation using geometric and arithmetic mean, respectively. In addition, the decision-makers' expertise and background can be reflected on the group judgment by weighting the individuals. The gaggregate() function computes group PCM/priority, CR of individual judgments (ICR), CR of aggregated group judgment, and CI measuring the consensus degree between individual judgments and the aggregated group judgment. Although AHP tolerates some degree of inconsistency, a severe inconsistency might cause the decision-making results to become invalid. Therefore, it is recommended to evaluate the CR of PCMs before it can be used to make decisions. Prize offers crplot(), which allows visualization of CR of individual judgments. The distance between individuals and group judgements can also be computed and visualized using dplot() function. dplot() uses the classical multidimensional scaling (MDS) approach to compute the distance (Gower, 1966).

If n is the number of elements in a level of hierarchy, $n(n\text{-}1)/2$ comparisons are required to build a PCM. Hence, with increasing the number of alternatives, the amount of pairwise comparisons becomes large. In this case, user can establish a rating category (e.g. excellent, good, fair, and poor) with respect to the corresponding criterion

for the evaluation of alternatives (AHP rating model). Prize offers a rating() function that computes the weight of alternatives according to the category that they fall into. This function takes two matrices as input, including a PCM of rating categories and a category assignment matrix (CAM), which states what category an alternative belongs to. rating() returns alternatives idealised priorities, weight of rating categories, and CR of category PCM. To obtain idealised priorities, weights of categories are divided by the largest weight. In case of AHP rating model, idealised priorities are used as the weight of alternatives in further steps of prioritization process.

Prioritization estimation

Prize offers an ahp() function, which can be called by a PCM matrix to compute weights and CR. In an actual analysis, ahp() must be called for each decision element to compute their weights. As a problem gets more complicated and the number of elements increases, it becomes complicated to perform this analysis manually. Therefore, in order to facilitate AHP analysis, I developed a pipeline() function, which can simply be called by a matrix including the problem hierarchy and PCMs built for each element. The pipeline() function returns the overall prioritization as well as CR of all input PCMs in a convenient format that facilitates further processing and visualization. Prize offers rainbowplot() and wplot() functions to visualize the final prioritization results and the weights of decision criteria, respectively. An example is shown in Figures 3-24 and 3-25.

Table 3-12. Saaty's fundamental scale for pairwise comparison

Intensity of importance	Definition	Explanation
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgement slightly favor one element over an other
5	Strong importance	Experience and judgement strongly favor one element over an other
7	Very strong importance	One element is favored very strongly over an other, its dominance is demonstrated in practice
9	Extreme importance	The evidence favoring one element over another is of the highest possible order of affirmation

* Intensities of 2,4,6, and 8 can be used to express intermediate values. Intensitise 1.1, 1.2, 1.3, etc. can be used for elements that are very close in importance

3.3.2. Prioritizing putative cancer-associated targets

Prize can efficiently rank and prioritize a list of alternatives according to a series of user-defined criteria. In order to demonstrate Prize application in translational bioinformatics research, here I rank and prioritize the putative tumour targets identified in section 3.1. The goal of this analysis is to identify and prioritize candidate genes that are most likely to generate successful cancer targets for antibody treatment.

The key step in decision-making is to gather and organize the critical information and data required to make a decision. Therefore, through an extensive literature search and systematic review, a list of criteria that are the most indicative of a tumour target were identified. The criteria include cancer expression profile, tumour-specificity, expression fold change in tumour compared to a compendium of normal tissues, target heterogeneity, role in cancer, therapeutic need, and annotation of extracellular region. They are also summarized in Table 3-13. The identified putative tumour targets, described in section 3.1, were chosen as alternatives to prioritize. Using collected information, a problem hierarchy were build as shown in Figure 3-23.

Table 3-13. Decision elements and their weights

Criteria, subcriteria, and rating scale categories	Definition	Weight*
Specificity		15.7% (0.157)
Low or no expression in normal tissues	No or little expression in normal tissues (< 20 FPKM)	100% (1.0)
Low expression in critical normal tissues**	Little expression in critical normal tissues (< 20 FPKM)	38.1% (0.381)
Medium expression in critical normal tissues**	Medium expression in critical normal tissues (>= 20 FPKM and < 50 FPKM)	14.5% (0.145)
Other		0.0% (0.0)
Expression level in cancer tissue		37.6% (0.376)
High	Differentially expressed in cancer with high level of expression (>= 100 FPKM)	100% (1.0)
Medium	Differentially expressed in cancer with medium level of expression (>= 50 TPM and < 100 FPKM)	38.1% (0.381)
Low	Differentially expressed in cancer with low level of expression (< 50 FPKM)	14.5% (0.145)
Fold Difference		25.5% (0.255)
FD High	Fold difference >= 4	100% (1.0)
FD Medium	Fold difference >= 2 and < 4	53.1% (0.531)
FD Low	Fold difference < 2	18.8 % (0.188)
Other		0.0% (0.0)
Target heterogeneity		6.8% (0.068)
Many patients, with high level of expression	High level of expression in many patients (>= 20%)	100% (1.0)
Few patients, with high level of expression	High level of expression in a small subset of patients (< 20 %)	31.4% (0.314)
Many patients, with lower level of expression	Lower level of expression in many patients (>= 20%)	19.8% (0.198)
Other		0.0% (0.0)

Criteria, subcriteria, and rating scale categories	Definition	Weight*
Accessibility		2.6% (0.026)
Annotated extracellular region		100% (1.0)
Predicted		0.0% (0.0)
Cancer gene/Function		4.0% (0.040)
Candidate is a known cancer gene	Putative cancer-genes identified through discovery analysis and literature search ***	100% (1.0)
Not Available		0.0% (0.0)
Therapeutic need		7.7% (0.077)
High interest	Cancers with high interest to develop novel therapy include; PAAD, LUSC, LUAD, LICH, LGG, GBM, AML	100% (1.0)
Medium interest	Cancers with medium interest to develop novel therapy include; OV, HNSC, COAD, KIRP, KIRC, KICH, BLCA, CESC	55.0% (0.550)
Low interest	Cancers with low interest to develop novel therapy include; BRCA, UCS, UCEC, PRAD, THCA, SKCM	30.2% (0.302)
Other		0.0% (0.0)

*Pairwise comparisons were performed via multiple discussions with a panel of antibody drug conjugate (ADC) development experts from CDRD and bioinformatic experts including myself from GSC to achieve a consensus PCM for each criterion and their rating categories. Final criteria and category PCMs were tested for inconsistency through measuring the CR value. All PCMs satisfied a CR smaller than 0.1. ** Critical normal tissue include tissues from adipose, adrenal gland, blood and blood vessel, bone marrow, brain, colon, esophagus, heart, kidney, liver, lung, lymph node, muscle, nerve, pancreas, pituitary, salivary gland, skin, small intestine, and stomach. *** In addition to the information obtained through pathway analysis, analysis of transcription factor target genes, and survival analysis (section 3.1.2), a list of cancer-associated genes was compiled using literature search and publically available databases including COSMIC (Bamford et al., 2004), allOnco (www.bushmanlab.org). The expanded form of each tumour type abbreviation is available in Table 3-2.

Since the number of alternatives (i.e. genes) is large (n = 1,503), I chose to use AHP rating model to perform prioritization. Therefore, each criterion was broken down into smaller categories that better represent the alternatives' characteristics. These categories are shown in Figure 3-23 as well as Table 3-13. Categories were then pairwise compared using AHP scale (Table 3-12) with respect to their associated criterion and their weights were obtained. Literature search and multiple discussions with a panel of experts at the Centre for Drug Research and Development (CDRD) including myself were used as the source to obtain the relative importance of categories to each other. These weights are listed in the last column of Table 3-13. For example, the level of expression in cancer tissue was broken down into three categories: low, medium, and high. The PCM and computed weights are shown in Table 3-14 A and B, respectively. Prize's rating() function was used to obtain alternatives' idealised priorities.

Similarly using a panel discussion and literature search, twenty-one pairwise comparisons were performed to assess the relative priority of the seven criteria. The obtained weight for each criterion is shown in Table 3-13 and Figure 3-24.

Table 3-14. (A) Category PCM for cancer expression criterion. (B) Computed AHP weights and idealised priorities for each category is shown. Idealised priories are computed by dividing AHP weights by the largest weight. Alternatives were then assigned a score (i.e. the value of idealised priority) with respect to the category that they fall into. If an alternative fulfilled more than one category within a criterion, the category with the highest value was selected.

(A)

Cancer Expression	High	Medium	Low
High	1	3	6
Medium	1/3	1	3
Low	1/6	1/3	1

(B)

Cancer Expression	Weight	Idealised priority
High	0.654	1
Medium	0.249	0.381
Low	0.0952	0.145

The Prize's pipeline() function was then used to perform final prioritization of candidate genes. This function takes in the problem hierarchy in form of a matrix and the associated PCMs and reports a final score for each alternative. The higher this score is, the better the performance of the alternative is with respect to the goal. The final prioritization is visualized in Figure 3-25 using the rainbowplot() function and is available as appendix D. The rainbow plot illustrates how the final scores are built from the user-defined criteria. In this plot, alternatives are placed on the y-axis, while the x-axis shows the final score. For instance, the color red represents the expression specificity of candidate genes to tumour tissues by evaluating the level of gene expression across a compendium of normal tissues. The larger block of red means that a gene is assigned a higher score due to its favourable expression pattern (i.e. low to no expression) across the compendium of normal tissue samples. Similarly, the color purple illustrates if a candidate gene is known to play a role in cancer. If this color is missing for a gene, it means that the gene is not classified as a cancer-associated based on the pathway analysis performed in section 3.1.2 and literature search.

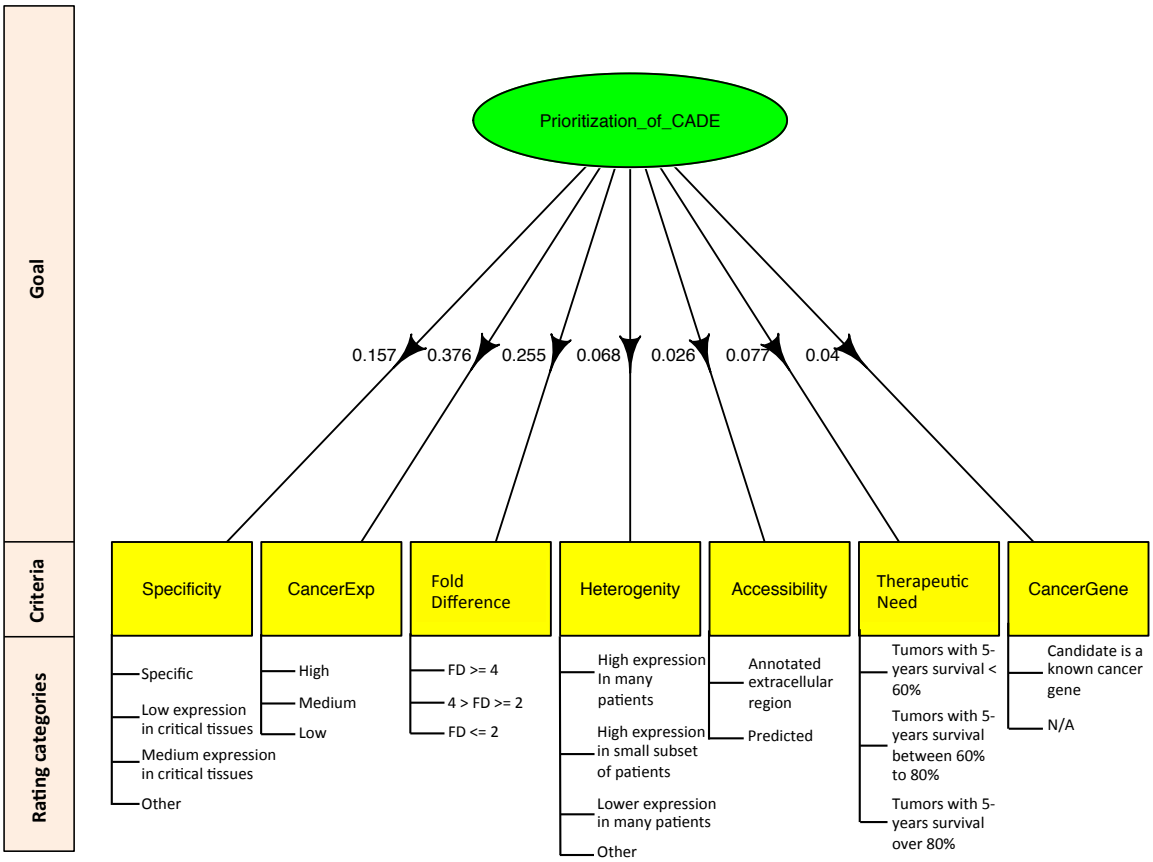


Figure 3-23. The problem hierarchy. Since the number of alternatives (i.e. genes) is large, AHP rating model is selected to perform the ranking. Therefore, each criterion is broken down into smaller categories that better represent the characteristics of alternatives with respect to the associated criterion. The weigh of each criterion with respect to the goal is shown on the edges of the hierarchy structure.

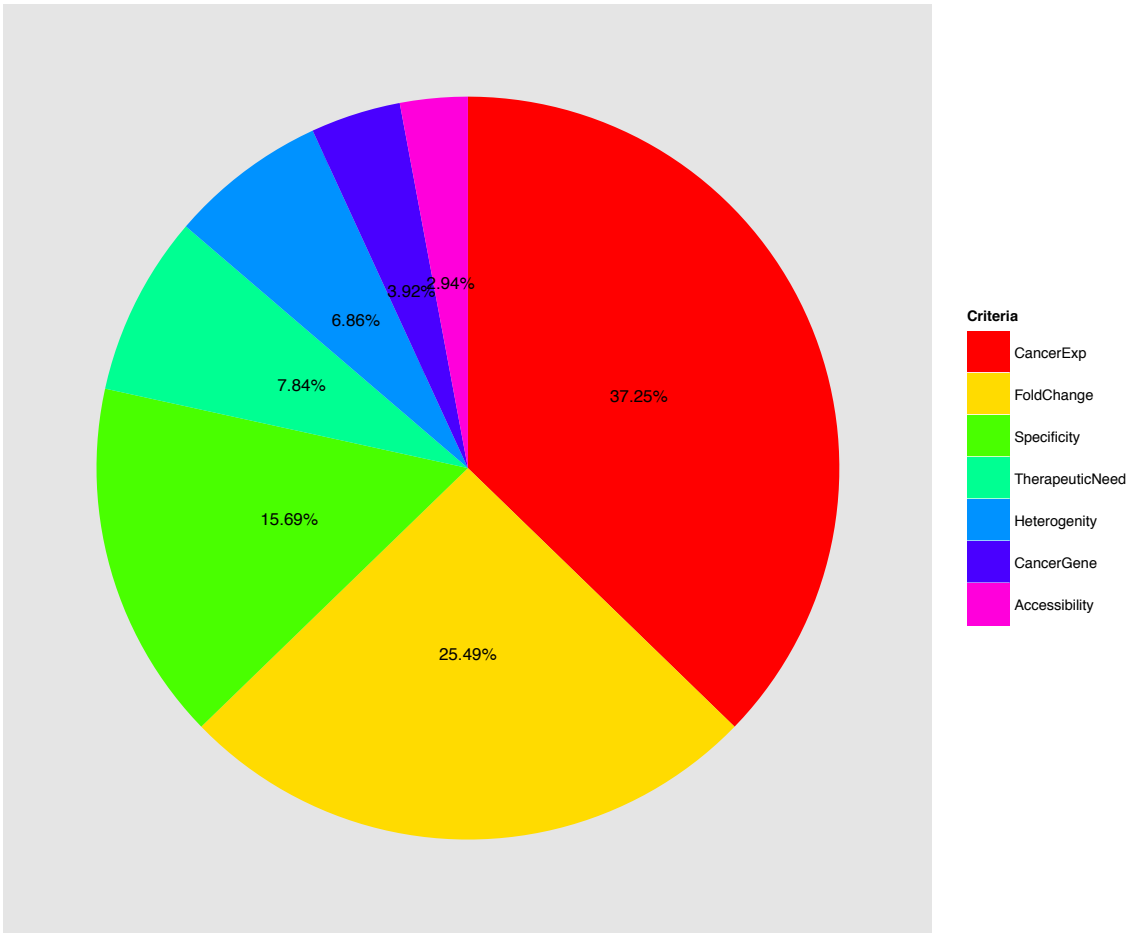


Figure 3-24. The pie chart represents the weight of each criterion with respect to the goal. The weights are obtained through twenty-one pairwise comparisons organized into a PCM. Prize computes the weight of each criterion using this PCM. The higher the weight, the more important the criterion is to achieve the final goal of prioritization.

Prize generates priority rankings of 1,503 putative cancer targets based on criteria pre-identified and weighted according to literature and expertise of a panel of antibody experts at genome sciences centre (GSC) and CDRD. This ranking is dynamic, given that the initial priorities could change as knowledge accrues from new studies. In total, such ranking provides a basis for rapidly deciding which target should advance to further validation and study.

Prize offers a simple approach to perform ranking and prioritization according to a user-specified list of criteria. The user is only responsible to provide problem hierarchy and PCMs of decision elements. The package then applies AHP method to obtain final ranking. Prize is simple to use and does not require an extensive knowledge of programming language R to work with. A detailed and simple manual is available on the Prize webpage at: https://www.bioconductor.org/packages/devel/bioc/vignettes/Prize/inst/doc/Prize.pdf

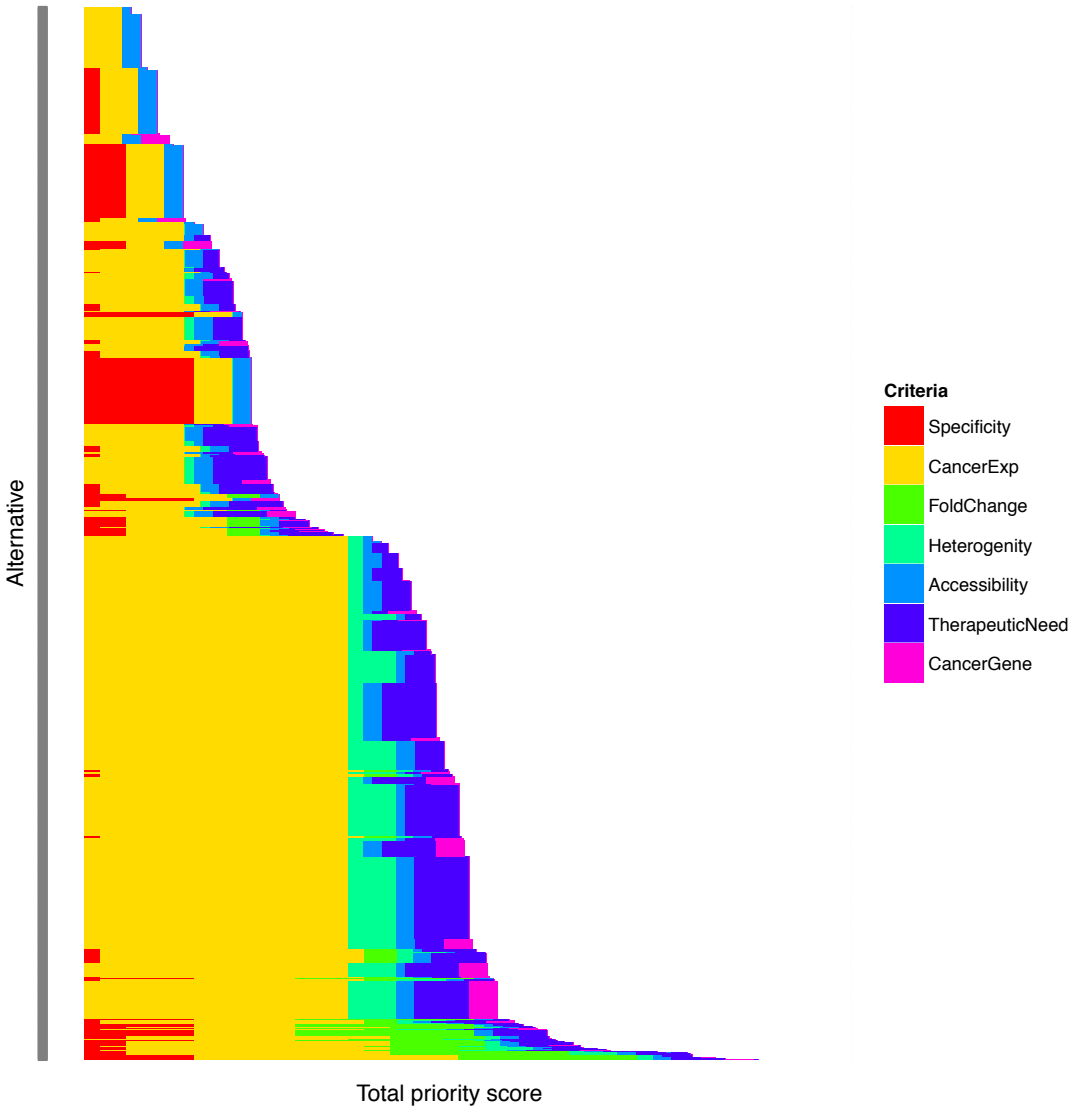


Figure 3-25. Prioritized candidates shown in a color-coded format (rainbow plot). In addition to the prioritization order, this plot illustrates how the final score for each gene is built as a combination of the user-defined criteria. The x-axis shows the final prioritization score, while alternatives are placed on the y-axis.

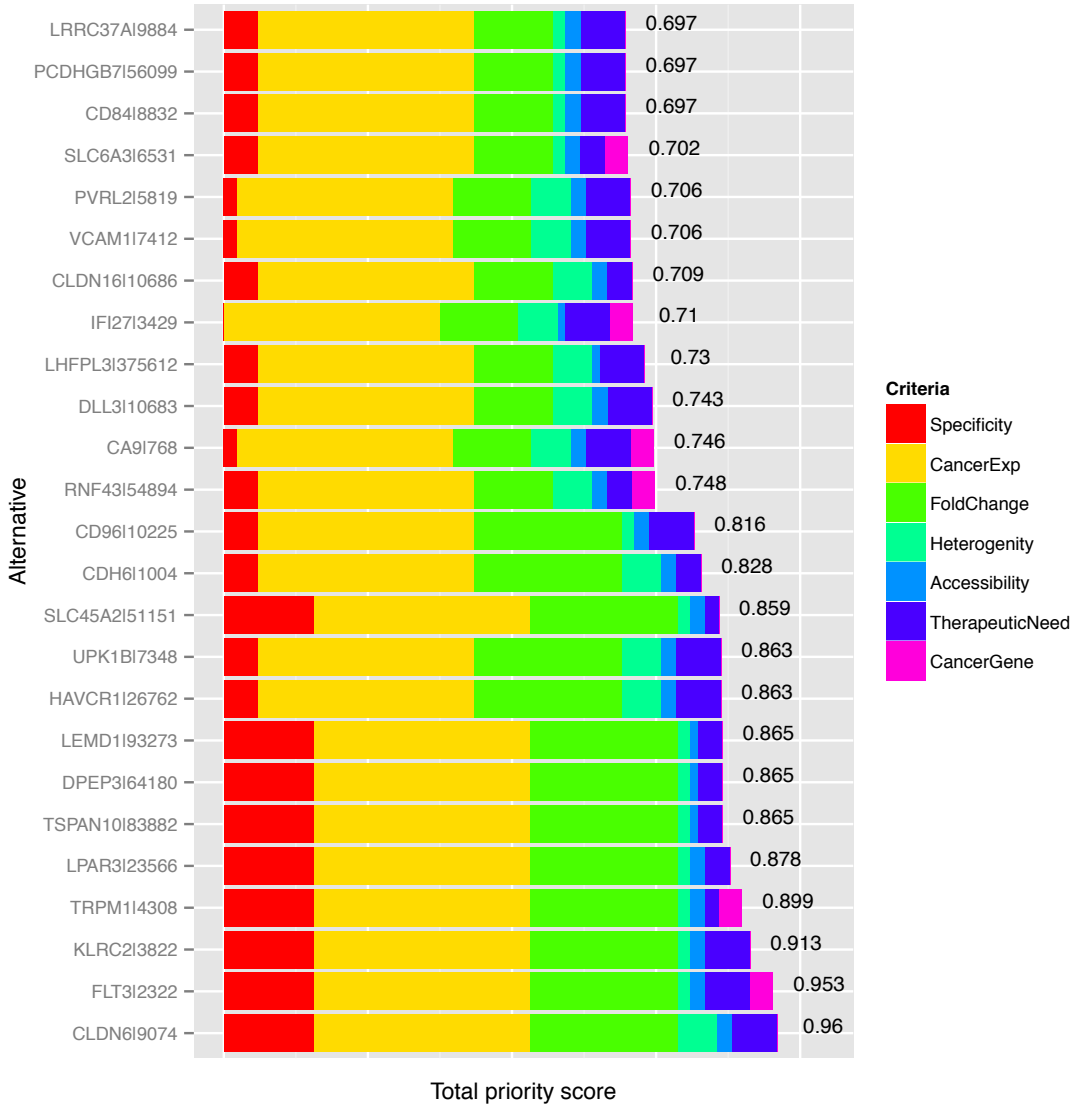


Figure 3-26. The top 25 prioritized candidates shown in a rainbow plot

Prize demonstrated successful prioritization and ranking of known tumour targets within the top 25 prioritised candidates. As shown in Figure 3-26, well-characterized biomarker targets in human malignancies including CLDN6 (Micke et al., 2014), FLT3 (Konig & Levis, 2015), HAVCR1/TDM-1 (Rees & Kain, 2008), CDH6 (Sancisi et al., 2013), CD96 (Hosen et al., 2007), CA9 (Tafreshi et al., 2014), DLL3 (Saunders et al., 2015), VCAM-1 (Chen & Massague, 2012), PVRL2 (Oshima et al., 2013), and CD84 (Binsky-Ehrenreich et al., 2014) are ranked among the top 25 candidates. These candidates are currently being investigated in pre-clinical studies and clinical trials. For example, CLDN6 (Figure 3-27), a cell surface protein and a member of claudin family, is often found to be abnormally expressed in cancer. In addition, strong CLDN6 expression has been associated with higher mortality rate in some cancer types. On the other hand, CLDN6 is absent from majority of healthy adult tissue. MAB027, developed by Ganymed (http://www.ganymed-pharmaceuticals.com/pipeline/imab027.html), is a monoclonal antibody that selectively binds to CLDN6, and is being tested in phase I/II clinical trial. The tumour cell specificity of CLDN6 makes IMAB027 a cancer cell selective drug allowing it to efficiently kill tumor cells without harming healthy non-cancerous cells. Similarly, DLL3 (Figure 3-28), a member of delta protein ligand family, functions as a Notch ligand that is characterized by a DSL domain, EGF repeats, and a transmembrane domain. It inhibits primary neurogenesis, and may be required to divert neurons along a specific differentiation pathway. DLL3 has been shown to express at high levels in multiple cancer types. Rova-T, developed by Stemcentrx (http://www.stemcentrx.com/ct-small-cell-lung-cancer.html), is an ADC that is made to target DLL3, enter the tumour cells, and release a potent drug to kill these cells. The antibody has been shown to successfully eradicate DLL3-expressing tumour cells *in vivo.*

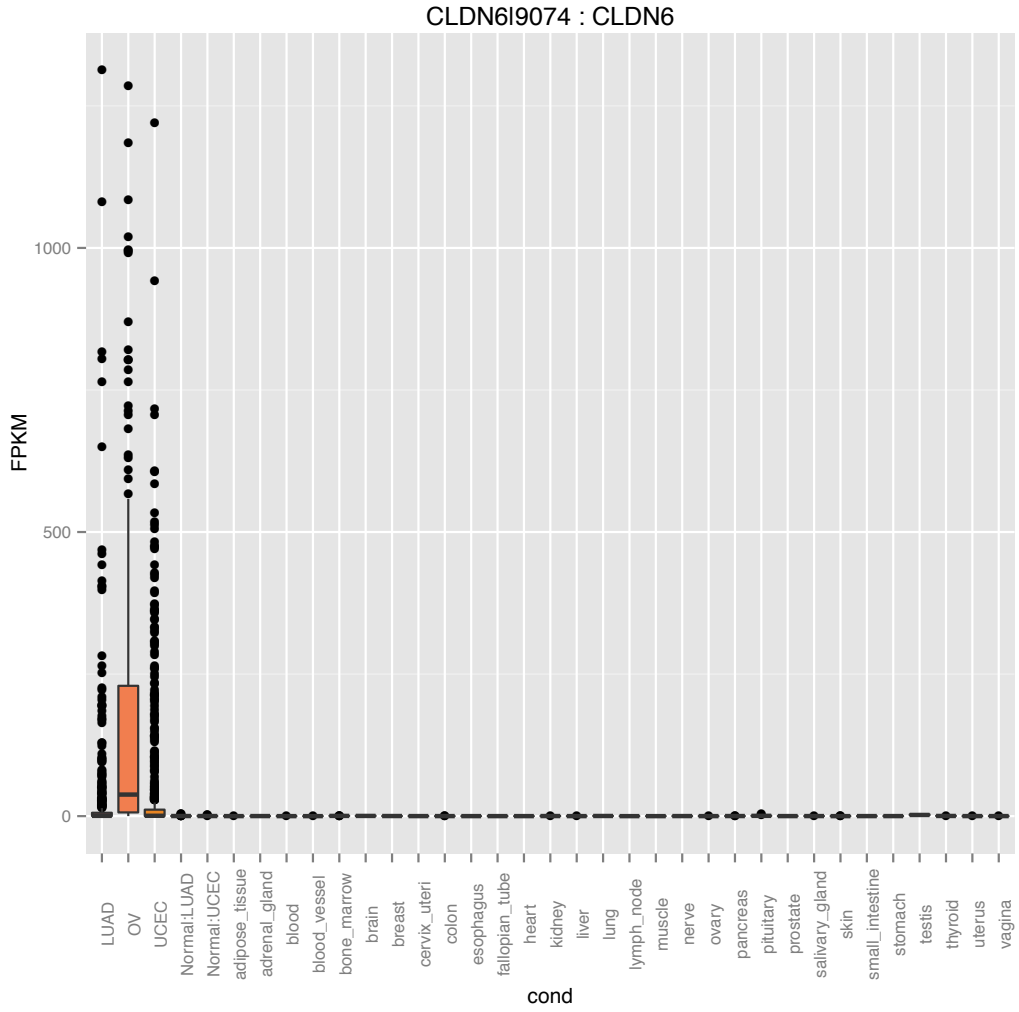


Figure 3-27. The expression profile of CLDN6. It is found to be overexpressed in lung, ovarian, and uterus tumours while it's expression is absent from matched normal TCGA and available normal tissues from GTEx.

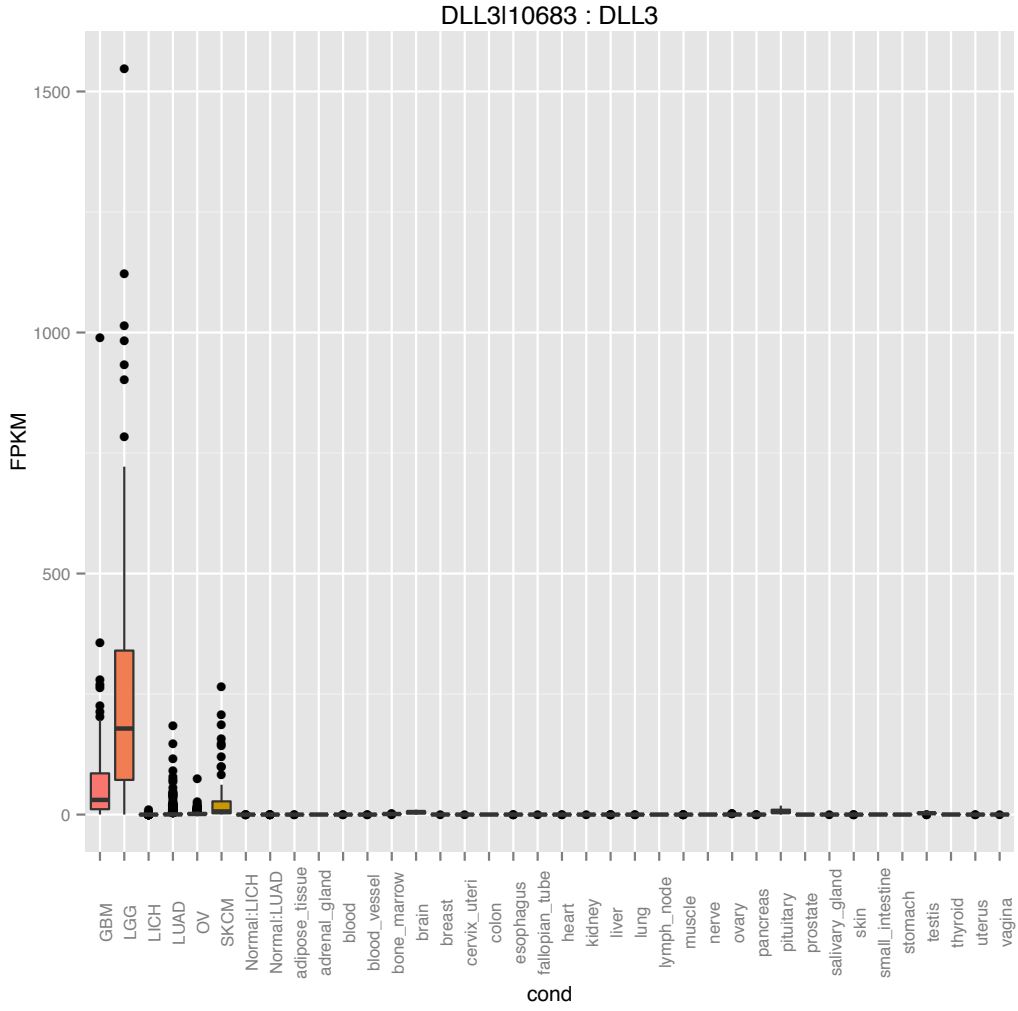


Figure 3-28. The expression profile of DLL3. It is found to be overexpressed in several TCGA tumors while it's expression is absent from matched-normal TCGA and available normal tissues from GTEx.

Among the identified well-characterized tumour targets are also novel candidates that may present potential therapeutic targets. For example, UPK1B, a member of the transmembrane 4 superfamily and a cell surface protein, demonstrates an ideal expression profile across the tumour and compendium of non-cancerous healthy tissues (Figure 3-29) (Olsburgh et al., 2003). The protein mediates signal transduction events that play a role in the regulation of cell development, activation, growth and motility. Even though it demonstrates a tissue specific expression in normal bladder and may play a role in normal bladder epithelial physiology, the much higher expression in tumour compared to the normal bladder makes it an attractive candidate to investigate further (Figure 3-29).

LPAR3, also known as LPA3, is a G protein-coupled receptor and functions as a cellular receptor for lysophosphatidic acid and mediates lysophosphatidic acid-evoked calcium mobilization. The aberrant expression of LPAR3 in ovarian cancer cells has been reported previously, and it is hypothesized that LPA3 overexpression during ovarian carcinogenesis contributes to ovarian cancer aggressiveness (Yu et al., 2008)The expression profile of LPAR3 makes it an interesting candidate for further validation and analysis as a potential target for cancer therapeutics (Figure 3-30).

The prioritization of the putative tumour targets illustrates Prize ability to efficiently rank a list of candidates according to a set of user-defined decision criteria. The use of Prize is not limited to the medical and biological decision making, it has a great potential to be used in variety of studies involving multiple-criteria DM toward ranking and prioritization of decision alternatives. Prize is currently available to public through Bioconductor (the R package repository) at: https://www.bioconductor.org/packages/release/bioc/html/Prize.html.

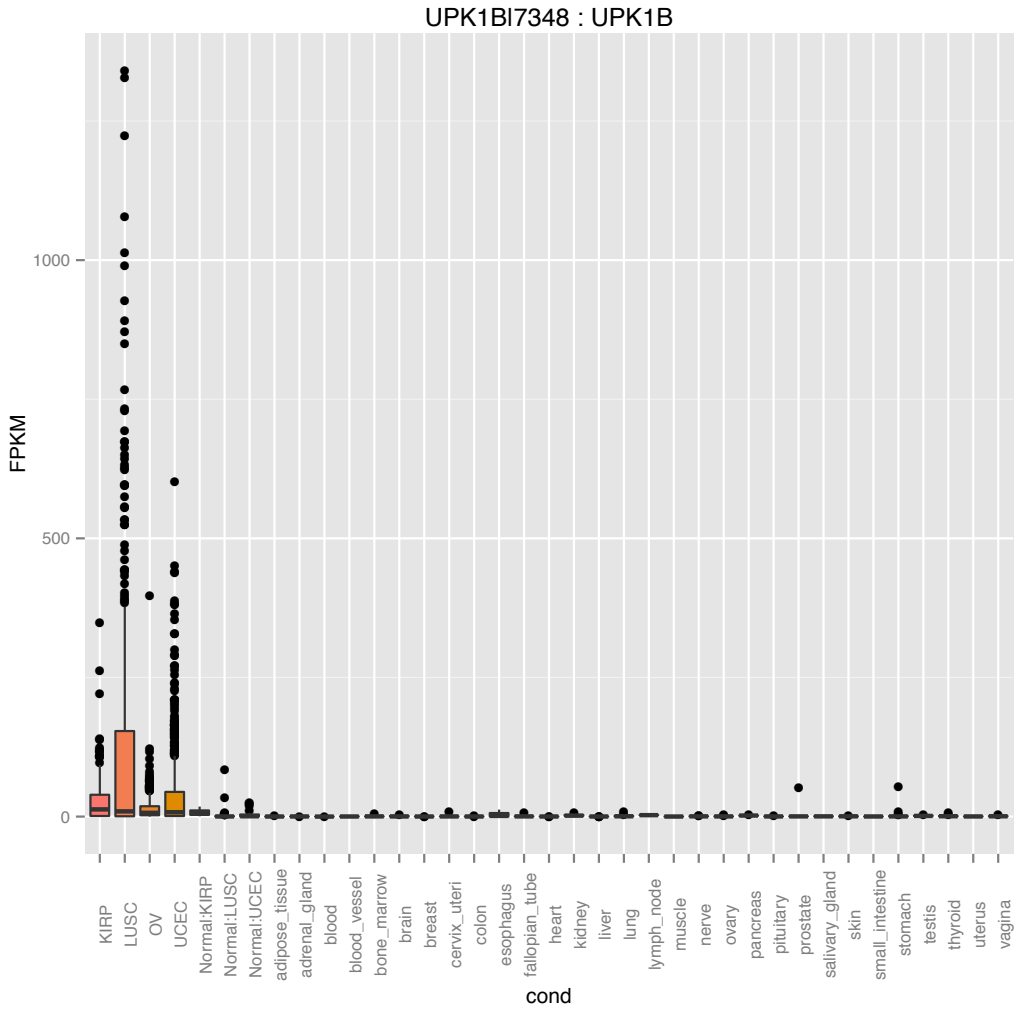


Figure 3-29. The expression profile of UPK1B across tumour and normal samples

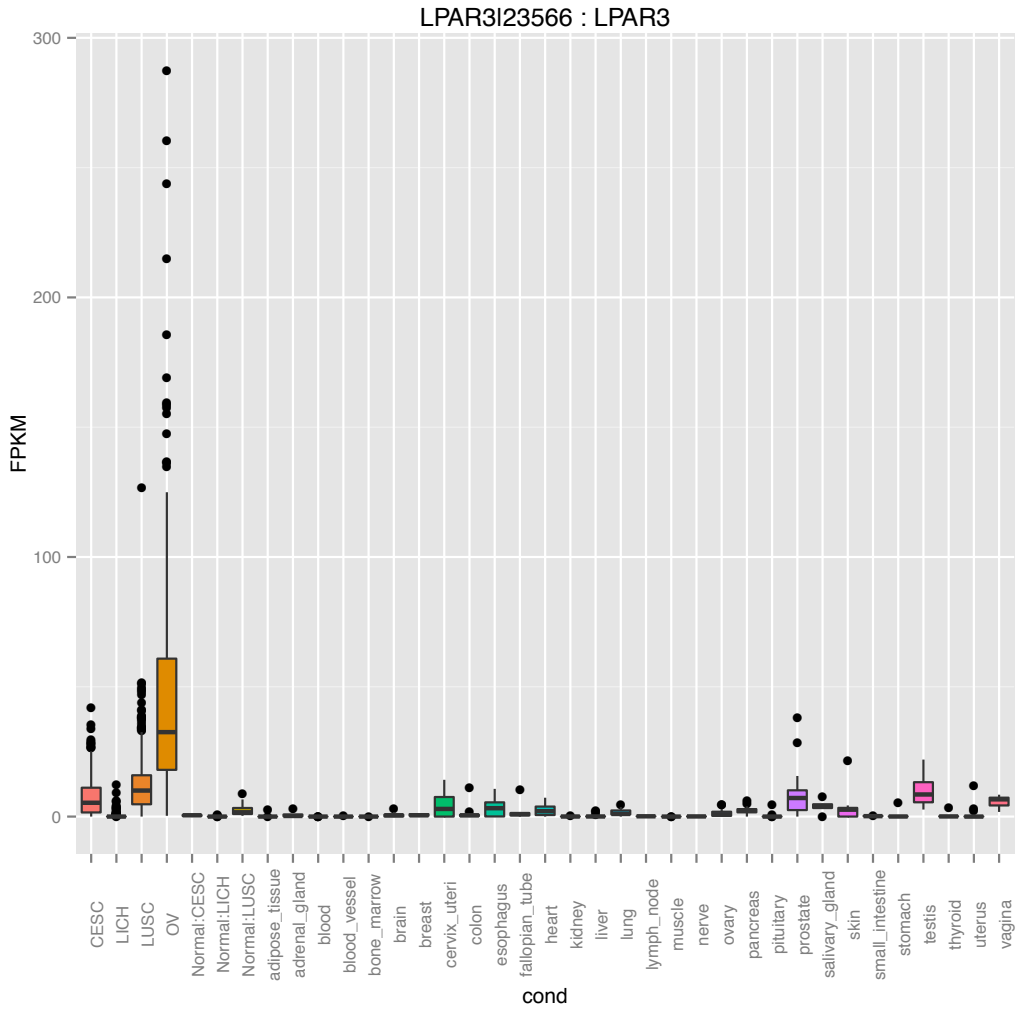


Figure 3-30. The expression profile of LPAR3 across tumour and normal samples

Chapter 4. Discussion

The traditional cancer therapy that is given to the patients with malignant tumors mainly includes chemotherapy, radiation therapy and removal of the tumor mass with surgery. However these methods either result in a non-optimal treatment or cause severe side effects and regardless may not be curative (De Angelis, 2008). For instance, standard chemotherapy often results in collateral damage to healthy tissue, causing unwanted side effects that impair the circulatory system, the immune system, the digestive system, and others (De Angelis, 2008). The reason is that chemotherapeutics usually affect processes that occur in all rapidly dividing cells, hence many normal cells throughout the body that are undergoing active growth and cell division can also be damaged. Although radiation therapy is more focused compared to chemotherapeutics, the high doses of radiation used to kill cancer cells can also damage healthy cells in the treatment area (De Angelis, 2008). In some cases a secondary malignancies and heart disease have been reported (De Angelis, 2008; Vega-Stromberg, 2003). Curative or primary surgery on the other hand is usually done when cancer is found in only one part of the body and presents less critical side effect compare to the two previous methods (Urruticoechea et al., 2010). However, after surgery radiation or chemotherapy may be given to the patients to eliminate any remaining cancer cells.

An ideal cancer therapy is one that specifically targets cancer cells while sparing normal tissues (Scott et al., 2012; Zhang et al., 2007). Targeted cancer therapies were introduced in 1990s and since then many have been approved by the Food and Drug Administration (FDA) to treat specific types of cancer (Baudino, 2015). Among the earliest targeted therapies are trastuzumab (Herceptin), gefitinib (Iressa), imatinib (Gleevec), and cetuximab (Erbitux) (Bou-Assaly & Mukherji, 2010; Harries & Smith, 2002; Hernandez-Boluda & Cervantes, 2002; Li et al., 2004). Targeted cancer therapies are drugs or other substances that block the growth and spread of cancer by interfering

with specific molecules (targets) that are involved in the growth, progression, and spread of cancer (Baudino, 2015). Most targeted therapies are either antibodies or small-molecule drugs (Baudino, 2015). Antibodies can be raised against a specific target protein (monoclonal antibodies). If this target is specific to cancer, then the antibody is able to distinguish between the cancerous and healthy cells and specifically bind to the cells that express the target protein (Scott et al., 2012; Zhang et al., 2007). Some monoclonal antibodies can be conjugated with toxins, chemotherapy drugs or radioactive isotopes to specifically deliver them to the cancer cells that express a specific target on their surface (Carter & Senter, 2008). Whilst, cells that do not express the target protein (e.g. normal tissue) will not be targeted. Therefore, targeted antibody-based therapeutics have the potential to significantly decrease the treatment side effects in patients as well as effectively targeting and destroying cancer cells by accumulating at the tumour site and increasing the effective dose to the tumour.

Although the unique properties of antibodies themselves are key components of a successful antibody-based therapeutic approach, the target proteins recognized by these antibodies play an equally important role. An ideal antibody target in cancer is a protein that is expressed on the surface of the tumour cells and is absent from healthy tissues (Papkoff, 2007). In addition, it is highly favourable if the target is necessary for cancer cells to grow and survive (Papkoff, 2007). But it is not a necessary factor for a target to become successful. The next best group of antibody targets are the ones that are highly expressed on the surface of tumour cells with much lower expression detected in healthy tissues (Papkoff, 2007). In this case, the fold change difference between tumour and normal expression and the type of the normal tissues that express the target are important to consider. Lastly, a tumour cell marker may be expressed on the surface of both tumour and normal tissues, while a unique protein form is expressed in tumour cells (Papkoff, 2007).

The work presented in this thesis focuses on the use of RNA sequencing for the high throughput study of differentially overexpressed and alternatively spliced genes in human cancers. Overall, the goal is to identify such cases ideal for targeting with antibody-based therapeutics. I collected a list of characteristics that are desirable for an ideal tumour biomarker target (to be targeted with therapeutic antibodies) by conducting

an extensive literature review (Baudino, 2015; Carter et al., 2004; Carter & Senter, 2008; Cheever et al., 2009; Papkoff, 2007; Scott et al., 2012) and studying those targets that are currently either used in the clinic or clinical trials (Table 3-8). I used this information to explore and mine TCGA transcriptome data to identify abnormally expressed genes and alternatively spliced variants that may serve as tumour cell markers. I devised a gene expression analysis pipeline (GEA) that resulted in identification of 1,503 differentially overexpressed genes across 24 different cancer types that may code for cell surface proteins. In addition, my analysis revealed a number of transcription factors that appear to commonly play a role in regulating the gene expression patterns across different cancer types. The pathway analysis also revealed similar mechanisms interrupted as a result of cancer. This observation suggests that majority of cancers undergo a common set of alterations during oncogenesis and it may be the same group of proteins that execute some of these biological processes across different cancers. Such proteins may offer interesting targets for therapeutic antibodies.

With my study as well as many other high-throughput studies often producing long lists of genes and proteins of interest, an approach is needed to narrow down such lists by ranking and prioritizing the candidates. I developed an R package, Prize, based on the analytic hierarchy process (AHP) approach to perform ranking and prioritization of the putative cancer biomarker targets. Prize allows prioritization based on a set of user-define criteria and numerical score to express the importance of each criterion to achieving the goal.

In addition, I developed an AHP model that depicts the characteristics of an ideal tumour cell target to perform this ranking. In summary, the key properties of an ideal biomarker target include: 1) abundant tumour cell surface localization; 2) significant overexpression in human cancer with little or no normal tissue expression, and 3) a function in promoting tumour growth and spread is favourable. Even though these characteristics appear to be simple, an ideal target is difficult to find and the selection of suitable targets can be complex. In chapter 3.3, the list of 1,503 putative tumour biomarker tagets were ranked and prioritized using Prize package according to their target potentials.

Prize successfully ranked known tumour markers within the top 25 prioritized genes (Figure 3-22), including CLDN6 (Micke et al., 2014), FLT3 (Konig & Levis, 2015), HAVCR1/TDM-1 (Rees & Kain, 2008), CDH6 (Sancisi et al., 2013), CD96 (Hosen et al., 2007), CA9 (Tafreshi et al., 2014), DLL3 (Saunders et al., 2015), VCAM-1(Q. Chen & Massague, 2012), PVRL2 (Oshima et al., 2013), and CD84 (Binsky-Ehrenreich et al., 2014) that are well-characterized targets in several human malignancies. Drugs targeting above genes are currently in pre-clinical and clinical studies. For example, MAB027, developed by Ganymed, is a monoclonal antibody that selectively binds to CLDN6, and is being tested in phase I/II clinical trial (http://www.ganymed-pharmaceuticals.com/pipeline/imab027.html). Similarly, Rova-T, developed by Stemcentrx, is an ADC that is made to target DLL3, enter the tumour cells, and release a potent drug to kill these cells (http://www.stemcentrx.com/ct-small-cell-lung-cancer.html). The antibody has been shown to successfully eradicate DLL3-expressing tumour cells *in-vivo.* IMC-EB10 is a novel antibody directed against FLT3 developed by ImClone Systems Coorporation (Youssoufian, Rowinsky et al., 2010). The binding of IMC-EB10 to FLT3 results in anti-proliferative effects *in-vitro* and in mouse models engrafted with human leukemia cells that harbour wild-type or constitutively activated FLT3. Yeda Research and Development Co. also is targeting CD84 with a monoclonal antibody for the treatment of Chronic lymphocytic leukemia as well as other B cell-related cancers including gastric cancer and renal cell carcinoma. They showed that inhibition of CD84 activity with a blocking antibody down-regulates the expression of another protein, which controls B-CLL survival, thus inducing cell death (http://www.yedarnd.com/technologies/cd84-novel-regulator-b-cll-survival).

In addition to the well-characterized tumour targets, GEA pipeline and Prize identified and prioritized genes that may present novel putative therapeutic targets. For example, UPK1B and LPAR3 both demonstrate an ideal expression profile across the tumour and compendium of non-cancerous healthy tissues. UPK1B codes for a protein that mediates signal transduction events that play a role in the regulation of cell development, activation, growth and motility (Olsburgh et al., 2003). While, the high expression of LPAR3 is believed to contribute to ovarian cancer aggressiveness (Yu et al., 2008)

Pan-Cancer identification and prioritization of both known and novel tumour targets demonstrate the ability of GEA pipeline and Prize to efficiently identify and rank putative tumour cell markers. The use of Prize is not limited to the medical and biological decision-making. Given a list of alternatives and a series of user-defined criteria and numerical score, Prize is able to perform ranking and prioritization. Therefore, Prize R package has a great potential to be used in variety of studies involving multiple-criteria decision-making.

Bispecific antibodies are a group of targeted antibody-based therapeutics that are capable of recognizing two different epitopes simultaneously. This dual specificity opens up a wide range of applications, including redirecting immune cells to tumour cells, blocking two different signalling pathways simultaneously, dual targeting of different disease mediators, and delivering payloads to targeted sites (Fan, Wang et al., 2015). Since bispecific antibodies are able to bind to two targets on the surface of tumour cells simultaneously, they may have higher specificity compared to monoclonal antibodies. Therefore, identified differentially overexpressed genes in each cancer type (section 3.1.2) were further examined to find pairs of surface-localized genes with mutually exclusive expression profile in critical normal tissues. This analysis revealed 1,200 candidate pairs across 24 different cancer types with some pairs being shared across multiple types of cancer.

In addition to differentially overexpressed genes, alternative splice variants are another class of tumour targets that can be targeted with antibodies. They may express in both tumour and normal healthy tissues, however a unique form may present on the surface of cancerous cells. Alternative splicing (AS) is a widespread mechanism for the generation of diverse protein products and regulation of protein expression. Tumour cells exploit this mechanism to favour the malignant state (Ghigna, Valacca, & Biamonti, 2008; Venables, 2004)In the past decade, cancer-associated splice variants of genes that control mechanisms such as DNA damage and proliferation (EGFR, Fibroblast Growth Factor Receptor 3 (FGFR3), Breast Cancer 1 (BRCA1)), adhesion and invasion (CD44, Macrophage Stimulating 1 Receptor (MST1R)), angiogenesis (Vascular Endothelial Growth Factor (VEGF)) and apoptosis (B-Cell Lymphoma/Leukemia 10 (BCL10), Caspase 2 (CASP2)) have been reported (Brinkman, 2004). Among these,

alternatively spliced transcripts with altered protein structure localized to the cell surface are of particular interest since they represent potential targets for discrimination between healthy and cancerous cells. That is, monoclonal antibodies can be produced to selectively target cancerous cells expressing such protein isoforms. An antibody against a tumour-associated surface-localized variant of EGFR (EGFRvIII) with exons 2-7 deleted, has shown effective anti-tumour activity in pre-clinical studies (Sampson et al., 2008) and is now is in phase I clinical trials. With the advent of massively parallel RNA sequencing, the large-scale exploration of cancer-related changes at the stage of transcription and post-transcriptional splicing has the potential to determine many more tumour-associated or enriched alternatively spliced targets.

In order to identify splicing variants that may play a role as tumour cell markers, I devised an AS-detection pipeline from high throughput RNA-seq data. The AS-detection pipeline allowed me to mine large sets of tumour transcriptomes to identify novel tumour-associated alternatively spliced variants. Most notably, I identified two novel tumour-associated splicing variants of matriptase. The variant designated as A1 has an in-frame skipping of exon 12, and variant A3 is generated as result of skipping exon 14. This analysis revealed a high frequency of these variants across epithelial-derived tumours, which were absent or expressed at extremely low levels in transcriptomes derived from normal tissues. Novel matriptase isoforms appear to form 2 to 8% of the overall matriptase gene expression in studied TCGA tumour samples, with wild-type being the dominantly expressed form. The qRT-PCR experiment confirmed the mRNA expression of matriptase variants in an independent set of tissues and cell lines, and revealed differential higher expression of variant A1 in ovarian and lung tumour tissues and cell lines compared to low or no expression in normal samples. Similarly, the A3 transcript was overexpressed in ovarian tumour tissues and cells. The variants A1 and A3 expression also were investigated in cDNA panels derived from 48 healthy tissue types from across the human body, such as brain, heart, kidney, and lung. Two third of normal samples demonstrated no mRNA expression of matriptase variants and a low level of expression in the remainder was observed.

Sequence analysis of novel matriptase variants indicated that the transcript variants could produce two fully functional open reading frames. The

immunoprecipitation results showed that these two novel proteins are being produced in CHO cells transiently transfected with cDNA encoding matriptase splice variants. With matriptase localized to the cell surface, we hypothesized there is a possibility that these novel isoforms of matriptase are also present on the cell surface. This hypothesis was tested by performing flow cytometry on CHO cells expressing these recombinant proteins. This analysis demonstrated the presence of these novel proteins on the surface of CHO cells, where wild-type matriptase surface expression predominated followed by variant A1 and then variant A3. Thus, protein expression of matriptase splice variants on the surface of CHO cells supports the notion that A1 and A3 protein products can localize on the surface of tumour cells as well.

The LDL receptor class A domain is an ~40-amino acid-long structure. The prototype structure of the LDLRA domain is found in the LDL receptor itself, which contains seven such domains. The crystal structure of the fifth LDLRA domain in the LDL receptor revealed that this domain contains six amino acids that bind calcium in an octahedral arrangement (calcium cage) (Fass, Blacklow et al., 1997). It has been shown that point mutations at critical residues in this calcium cage potently inhibit the LDLRA ligand binding (Esser, Limbird et al., 1988). Oberst et al. showed that mutations in the Ca2+-binding motifs of any or all of the four LDLRA domains of matriptase prevent its activation (Oberst et al., 2003). Interestingly, however, the complete deletion of all four LDLRA domains allowed constitutive activation of this enzyme. Additional experiments are required to demonstrate the impact of deleting LDLRA1 and LDLRA3 domains as observed in the A1 and A3 variants. Although these two deletions may have variable effects on matriptase activity, the results demonstrated here show that they do not affect the ability of the protein products to form a complex with HAI-1 and traffic to the cell surface. Hence, they may serve as potential tumour biomarker targets for targeting with therapeutic antibodies.

Cancer is characterized by uncontrolled cell proliferation and an absence of cell death that result in formation of an abnormal cell mass or tumour. The primary tumour can grow, acquire metastatic potential, and spreads to other body sites. Currently, local and non-metastatic cancers are treated by surgery and radiotherapy, while anti-cancer drugs (e.g. chemotherapy) are being used in metastatic cancers. Chemotherapeutic

drugs target rapidly growing cells, which is a characteristic of the cancerous cells, but it also affects normal cells with fast proliferation rates, such as the hair follicles, bone marrow and gastrointestinal tract cells, generating severe side effects in patients. The indiscriminate destruction of normal cells as well as the toxicity of chemotherapeutic drugs support the need to find new effective targeted treatments based on the changes in the molecular biology of the tumour cells. Targeted therapies either block biologic transduction pathways and/or specific cancer proteins to induce the death of cancer cells or specifically deliver chemotherapeutic agents to cancer cells, minimizing the undesirable side effects. One approach to specifically deliver therapeutic agents to the tumour cells, while minimizing their presence at other sites in the body, is to conjugate them with tumour-specific monoclonal antibodies. Although the unique properties of antibodies themselves are key components of a successful antibody-based therapeutic approach, the target proteins recognized by these antibodies play an equally important role. The current thesis provides a comprehensive list of putative cancer-associated biomarker targets that may serve as targets for therapeutic antibody development in cancer. Further clinical validation would prove valuable in the utility of identified putative biomarker targets for therapeutic use.

References

Akaishi, J., Onda, M., Okamoto, J., Miyamoto, S., Nagahama, M., Ito, K., et al. (2006). Down-regulation of transcription elogation factor A (SII) like 4 (TCEAL4) in anaplastic thyroid cancer. *BMC Cancer, 6*, 260. doi:1471-2407-6-260 [pii]

Albanell, J., & Baselga, J. (1999). Trastuzumab, a humanized anti-HER2 monoclonal antibody, for the treatment of breast cancer. *Drugs of Today (Barcelona, Spain : 1998), 35*(12), 931-946. doi:564040 [pii]

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., et al. (2013). ***Essential cell biology*** (4th ed.) Garland Science.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2007). ***Molecular biology of the cell
*** (5th ed.) Garland Science.

Andrews, S. (2016). *Fastqc: A quality control tool for high throughput sequence data.*http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Babraham Bioinformatics. (2015). ***Trim galore!.***http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., et al. (2004). The COSMIC (catalogue of somatic mutations in cancer) database and website. *British Journal of Cancer, 91*(2), 355-358. doi:10.1038/sj.bjc.6601894 [doi]

Bao, B., Wang, Z., Ali, S., Kong, D., Banerjee, S., Ahmad, A., et al. (2011). Over-expression of FoxM1 leads to epithelial-mesenchymal transition and cancer stem cell phenotype in pancreatic cancer cells. *Journal of Cellular Biochemistry, 112*(9), 2296-2306. doi:10.1002/jcb.23150 [doi]

Baudino, T. A. (2015). Targeted cancer therapy: The next generation of cancer treatment. *Current Drug Discovery Technologies, 12*(1), 3-20. doi:CDDT-EPUB-67825 [pii]

Beillard, E., Pallisgaard, N., van der Velden, V. H., Bi, W., Dee, R., van der Schoot, E., et al. (2003). Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients using 'real-time' quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR) - a europe against cancer program. *Leukemia, 17*(12), 2474-2486. doi:10.1038/sj.leu.2403136

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research, 40*(10), e72. doi:10.1093/nar/gks001 [doi]

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59. doi:10.1038/nature07517 [doi]

Bergamaschi, A., Madak-Erdogan, Z., Kim, Y. J., Choi, Y. L., Lu, H., & Katzenellenbogen, B. S. (2014). The forkhead transcription factor FOXM1 promotes endocrine resistance and invasiveness in estrogen receptor-positive breast cancer by expansion of stem-like cancer cells. *Breast Cancer Research : BCR, 16*(5), 436-014-0436-4. doi:10.1186/s13058-014-0436-4 [doi]

Bhagwat, A. S., & Vakoc, C. R. (2015). Targeting transcription factors in cancer. *Trends in Cancer, 1*(1), 53-65. doi:10.1016/j.trecan.2015.07.001 [doi]

Binsky-Ehrenreich, I., Marom, A., Sobotta, M. C., Shvidel, L., Berrebi, A., Hazan-Halevy, I., et al. (2014). CD84 is a survival receptor for CLL cells. *Oncogene, 33*(8), 1006-1016. doi:10.1038/onc.2013.31 [doi]

Boise, L. H., Gonzalez-Garcia, M., Postema, C. E., Ding, L., Lindsten, T., Turka, L. A., et al. (1993). Bcl-X, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell, 74*(4), 597-608. doi:0092-8674(93)90508-N [pii]

Bou-Assaly, W., & Mukherji, S. (2010). Cetuximab (erbitux). *AJNR.American Journal of Neuroradiology, 31*(4), 626-627. doi:10.3174/ajnr.A2054 [doi]

Brinkman, B. M. (2004). Splice variants as cancer biomarkers. *Clinical Biochemistry, 37*(7), 584-594. doi:10.1016/j.clinbiochem.2004.05.015

Brown, S. J., Stoilov, P., & Xing, Y. (2012). Chromatin and epigenetic regulation of pre-mRNA processing. *Human Molecular Genetics, 21*(R1), R90-6. doi:dds353 [pii]

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*(7353), 609-615. doi:10.1038/nature10166 [doi]

Carlsson, S. K., Brothers, S. P., & Wahlestedt, C. (2014). Emerging treatment strategies for glioblastoma multiforme. *EMBO Molecular Medicine, 6*(11), 1359-1370. doi:10.15252/emmm.201302627 [doi]

Carter, P., Smith, L., & Ryan, M. (2004). Identification and validation of cell surface antigens for antibody targeting in oncology. *Endocrine-Related Cancer, 11*(4), 659-687. doi:11/4/659 [pii]

Carter, P. J., & Senter, P. D. (2008). Antibody-drug conjugates for cancer therapy. *Cancer Journal (Sudbury, Mass.), 14*(3), 154-169. doi:10.1097/PPO.0b013e318172d704

Chames, P., & Baty, D. (2009). Bispecific antibodies for cancer therapy. *Current Opinion in Drug Discovery & Development, 12*(2), 276-283.

Cheever, M. A., Allison, J. P., Ferris, A. S., Finn, O. J., Hastings, B. M., Hecht, T. T., et al. (2009). The prioritization of cancer antigens: A national cancer institute pilot project for the acceleration of translational research. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 15*(17), 5323-5337. doi:10.1158/1078-0432.CCR-09-0737; 10.1158/1078-0432.CCR-09-0737

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nature Reviews.Molecular Cell Biology, 10*(11), 741-754. doi:10.1038/nrm2777 [doi]

Chen, Q., & Massague, J. (2012). Molecular pathways: VCAM-1 as a potential therapeutic target in metastasis. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 18*(20), 5520-5525. doi:10.1158/1078-0432.CCR-11-2904 [doi]

Chien, J., Narita, K., Rattan, R., Giri, S., Shridhar, R., Staub, J., et al. (2008). A role for candidate tumor-suppressor gene TCEAL7 in the regulation of c-myc activity, cyclin D1 levels and cellular transformation. *Oncogene, 27*(58), 7223-7234. doi:10.1038/onc.2008.360 [doi]

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The sanger FASTQ file format for sequences with quality scores, and the solexa/illumina FASTQ variants. *Nucleic Acids Research, 38*(6), 1767-1771. doi:10.1093/nar/gkp1137 [doi]

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 13-016-0881-8. doi:10.1186/s13059-016-0881-8 [doi]

Da Cunha, J. P., Galante, P. A., de Souza, J. E., de Souza, R. F., Carvalho, P. M., Ohara, D. T., et al. (2009). Bioinformatics construction of the human cell surfaceome. *Proceedings of the National Academy of Sciences of the United States of America, 106*(39), 16752-16757. doi:10.1073/pnas.0907939106

Dargahi, D., Swayze, R. D., Yee, L., Bergqvist, P. J., Hedberg, B. J., Heravi-Moussavi, A., et al. (2014). A pan-cancer analysis of alternative splicing events reveals novel tumor-associated splice variants of matriptase. *Cancer Informatics, 13*, 167-177. doi:10.4137/CIN.S19435 [doi]

Davalieva, K., Kiprijanovska, S., Komina, S., Petrusevska, G., Zografska, N. C., & Polenakovic, M. (2015). Proteomics analysis of urine reveals acute phase response proteins as candidate diagnostic biomarkers for prostate cancer. *Proteome Science, 13*(1), 2-014-0059-9. eCollection 2015. doi:10.1186/s12953-014-0059-9 [doi]

De Angelis, C. (2008). Side effects related to systemic cancer treatment: Are we changing the promethean experience with molecularly targeted therapies? *Current Oncology (Toronto, Ont.), 15*(4), 198-199.

Deckert, P. M. (2009). Current constructs and targets in clinical development for antibody-based cancer therapy. *Current Drug Targets, 10*(2), 158-175.

Dhillon, A. S., Hagan, S., Rath, O., & Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene, 26*(22), 3279-3290. doi:1210421 [pii]

Di Modugno, F., DeMonte, L., Balsamo, M., Bronzi, G., Nicotra, M. R., Alessio, M., et al. (2007). Molecular cloning of hMena (ENAH) and its splice variant hMena+11a: Epidermal growth factor increases their expression and stimulates hMena+11a phosphorylation in breast cancer cell lines. *Cancer Research, 67*(6), 2657-2665. doi:67/6/2657 [pii]

Diaz-Ramos, M. C., Engel, P., & Bastos, R. (2011). Towards a comprehensive human cell-surface immunome database. *Immunology Letters, 134*(2), 183-187. doi:10.1016/j.imlet.2010.09.016

Dong, Y., Walsh, M. D., McGuckin, M. A., Gabrielli, B. G., Cummings, M. C., Wright, R. G., et al. (1997). Increased expression of cyclin-dependent kinase inhibitor 2 (CDKN2A) gene product P16INK4A in ovarian cancer is associated with progression and unfavourable prognosis. *International Journal of Cancer, 74*(1), 57-63. doi:10.1002/(SICI)1097-0215(19970220)74:1<57::AID-IJC10>3.0.CO;2-F [pii]

Esser, V., Limbird, L. E., Brown, M. S., Goldstein, J. L., & Russell, D. W. (1988). Mutational analysis of the ligand binding domain of the low density lipoprotein receptor. *The Journal of Biological Chemistry, 263*(26), 13282-13290.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research, 8*(3), 186-194.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Research, 8*(3), 175-185.

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlen, M., & Berglund, L. (2010). Prediction of the human membrane proteome. *Proteomics, 10*(6), 1141-1149. doi:10.1002/pmic.200900258

Fass, D., Blacklow, S., Kim, P. S., & Berger, J. M. (1997). Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. *Nature, 388*(6643), 691-693. doi:10.1038/41798

Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., et al. (2014). MUC16 (CA125): Tumor biomarker to cancer therapy, a work in progress. *Molecular Cancer, 13*, 129-4598-13-129. doi:10.1186/1476-4598-13-129 [doi]

Gagou, M. E., Ganesh, A., Thompson, R., Phear, G., Sanders, C., & Meuth, M. (2011). Suppression of apoptosis by PIF1 helicase in human tumor cells. *Cancer Research, 71*(14), 4998-5008. doi:10.1158/0008-5472.CAN-10-4404 [doi]

Gartel, A. L. (2014). Suppression of the oncogenic transcription factor FOXM1 by proteasome inhibitors. *Scientifica, 2014*, 596528. doi:10.1155/2014/596528 [doi]

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P. M., et al. (2005). Cell motility is controlled by SF2/ASF through alternative splicing of the ron protooncogene. *Molecular Cell, 20*(6), 881-890. doi:S1097-2765(05)01721-1 [pii]

Ghigna, C., Valacca, C., & Biamonti, G. (2008). Alternative splicing and tumor progression. *Current Genomics, 9*(8), 556-570. doi:10.2174/138920208786847971; 10.2174/138920208786847971

Goel, M. K., Khanna, P., & Kishore, J. (2010). Understanding survival analysis: Kaplan-meier estimate. *International Journal of Ayurveda Research, 1*(4), 274-278. doi:10.4103/0974-7788.76794 [doi]

Goodison, S., Sun, Y., & Urquidi, V. (2010). Derivation of cancer diagnostic and prognostic signatures from gene expression data. *Bioanalysis, 2*(5), 855-862. doi:10.4155/bio.10.35 [doi]

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis.*53*(3/4), 325-338.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology, 29*(7), 644-652. doi:10.1038/nbt.1883 [doi]

Griffith, M., Mwenifumbo, J. C., Cheung, P. Y., Paul, J. E., Pugh, T. J., Tang, M. J., et al. (2012). Novel mRNA isoforms and mutations of uridine monophosphate synthetase and 5-fluorouracil resistance in colorectal cancer. *The Pharmacogenomics Journal,* doi:10.1038/tpj.2011.65; 10.1038/tpj.2011.65

GTEx Consortium. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics, 45*(6), 580-585. doi:10.1038/ng.2653 [doi]

Gui, T., Sun, Y., Shimokado, A., & Muragaki, Y. (2012). **The roles of mitogen-activated protein kinase pathways in TGF-β-induced epithelial-mesenchymal transition**. *Journal of Signal Transduction, 2012*

Halasi, M., Pandit, B., Wang, M., Nogueira, V., Hay, N., & Gartel, A. L. (2013). Combination of oxidative stress and FOXM1 inhibitors induces apoptosis in cancer cells and inhibits xenograft tumor growth. *The American Journal of Pathology, 183*(1), 257-265. doi:10.1016/j.ajpath.2013.03.012 [doi]

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70. doi:S0092-8674(00)81683-9 [pii]

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674. doi:10.1016/j.cell.2011.02.013 [doi]

Hansen, K. D., Irizarry, R. A., & Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics (Oxford, England), 13*(2), 204-216. doi:10.1093/biostatistics/kxr054 [doi]

Hao, Z., Zhang, H., & Cowell, J. (2012). Ubiquitin-conjugating enzyme UBE2C: Molecular biology, role in tumorigenesis, and potential as a biomarker. *Tumour Biology : The Journal of the International Society for Oncodevelopmental Biology and Medicine, 33*(3), 723-730. doi:10.1007/s13277-011-0291-1 [doi]

Harries, M., & Smith, I. (2002). The development and clinical use of trastuzumab (herceptin). *Endocrine-Related Cancer, 9*(2), 75-85.

Hartman, M. L., & Czyz, M. (2015). MITF in melanoma: Mechanisms behind its expression and activity. *Cellular and Molecular Life Sciences : CMLS, 72*(7), 1249-1260. doi:10.1007/s00018-014-1791-0 [doi]

Hauser, S., Bickel, L., Weinspach, D., Gerg, M., Schafer, M. K., Pfeifer, M., et al. (2011). Full-length L1CAM and not its Delta2Delta27 splice variant promotes metastasis through induction of gelatinase expression. *PloS One, 6*(4), e18989. doi:10.1371/journal.pone.0018989 [doi]

He, C., Zhou, F., Zuo, Z., Cheng, H., & Zhou, R. (2009). A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PloS One, 4*(3), e4732. doi:10.1371/journal.pone.0004732 [doi]

Hernandez-Boluda, J. C., & Cervantes, F. (2002). Imatinib mesylate (gleevec, glivec): A new therapy for chronic myeloid leukemia and other malignancies. *Drugs of Today (Barcelona, Spain : 1998), 38*(9), 601-613. doi:696536 [pii]

Holliday, R., & Jeggo, P. A. (1985). Mechanisms for changing gene expression and their possible relationship to carcinogenesis. *Cancer Surveys, 4*(3), 557-581.

Hosen, N., Park, C. Y., Tatsumi, N., Oji, Y., Sugiyama, H., Gramatzki, M., et al. (2007). CD96 is a leukemic stem cell-specific marker in human acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America, 104*(26), 11008-11013. doi:10.1073/pnas.0704271104

Huang, J., Zheng, D. L., Qin, F. S., Cheng, N., Chen, H., Wan, B. B., et al. (2010). Genetic and epigenetic silencing of SCARA5 may contribute to human hepatocellular carcinoma by activating FAK signaling. *The Journal of Clinical Investigation, 120*(1), 223-241. doi:10.1172/JCI38012 [doi]

Jiang, H., Li, Q., He, C., Li, F., Sheng, H., Shen, X., et al. (2014). Activation of the wnt pathway through Wnt2 promotes metastasis in pancreatic cancer. *American Journal of Cancer Research, 4*(5), 537-544.

Jordan, P., Brazao, R., Boavida, M. G., Gespach, C., & Chastre, E. (1999). Cloning of a novel human Rac1b splice variant with increased expression in colorectal tumors. *Oncogene, 18*(48), 6835-6839. doi:10.1038/sj.onc.1203233 [doi]

Kalsotra, A., & Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nature Reviews.Genetics, 12*(10), 715-729. doi:10.1038/nrg3052 [doi]

Kang, J. Y., Dolled-Filhart, M., Ocal, I. T., Singh, B., Lin, C. Y., Dickson, R. B., et al. (2003). Tissue microarray analysis of hepatocyte growth factor/met pathway components reveals a role for met, matriptase, and hepatocyte growth factor activator inhibitor 1 in the progression of node-negative breast cancer. *Cancer Research, 63*(5), 1101-1105.

Kastan, M. B. (2007). Wild-type p53: Tumors can't stand it. *Cell, 128*(5), 837-840. doi:S0092-8674(07)00246-2 [pii]

Konig, H., & Levis, M. (2015). Targeting FLT3 to treat leukemia. *Expert Opinion on Therapeutic Targets, 19*(1), 37-54. doi:10.1517/14728222.2014.960843 [doi]

Kontermann, R. E. (2012). Dual targeting strategies with bispecific antibodies. *mAbs, 4*(2), 182-197. doi:10.4161/mabs.4.2.19000 [doi]

Kopp, R., Fichter, M., Schalhorn, G., Danescu, J., & Classen, S. (2009). Frequent expression of the high molecular, 673-bp CD44v3,v8-10 variant in colorectal adenomas and carcinomas. *International Journal of Molecular Medicine, 24*(5), 677-683.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J. J., Nardone, F., et al. (2009). ASTD: The alternative splicing and transcript diversity database. *Genomics, 93*(3), 213-220. doi:10.1016/j.ygeno.2008.11.003; 10.1016/j.ygeno.2008.11.003

Kramer, A., Green, J., Pollard, J.,Jr, & Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics (Oxford, England), 30*(4), 523-530. doi:10.1093/bioinformatics/btt703 [doi]

Krawczak, M., Thomas, N. S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., et al. (2007). Single base-pair substitutions in exon-intron junctions of human genes: Nature, distribution, and consequences for mRNA splicing. *Human Mutation, 28*(2), 150-158. doi:10.1002/humu.20400 [doi]

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature, 227*(5259), 680-685.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10*(3), R25-2009-10-3-r25. Epub 2009 Mar 4. doi:10.1186/gb-2009-10-3-r25 [doi]

Lee, C. T., Capodieci, P., Osman, I., Fazzari, M., Ferrara, J., Scher, H. I., et al. (1999). Overexpression of the cyclin-dependent kinase inhibitor p16 is associated with tumor recurrence in human prostate cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 5*(5), 977-983.

Lee, M. S., Tseng, I. C., Wang, Y., Kiyomiya, K., Johnson, M. D., Dickson, R. B., et al. (2007). Autoactivation of matriptase in vitro: Requirement for biomembrane and LDL receptor domain. *American Journal of Physiology.Cell Physiology, 293*(1), C95-105. doi:10.1152/ajpcell.00611.2006

Lee, S., Seo, C. H., Lim, B., Yang, J. O., Oh, J., Kim, M., et al. (2011). Accurate quantification of transcriptome from RNA-seq data by effective length normalization. *Nucleic Acids Research, 39*(2), e9. doi:10.1093/nar/gkq1015 [doi]

Lee, S. L., Dickson, R. B., & Lin, C. Y. (2000). Activation of hepatocyte growth factor and urokinase/plasminogen activator by matriptase, an epithelial membrane serine protease. *The Journal of Biological Chemistry, 275*(47), 36720-36725. doi:10.1074/jbc.M007802200

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods, 7*(9), 709-715. doi:10.1038/nmeth.1491 [doi]

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323-2105-12-323. doi:10.1186/1471-2105-12-323 [doi]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics, 11*(5), 473-483. doi:10.1093/bib/bbq015 [doi]

Li, J., Kleeff, J., Giese, N., Buchler, M. W., Korc, M., & Friess, H. (2004). Gefitinib ('iressa', ZD1839), a selective epidermal growth factor receptor tyrosine kinase inhibitor, inhibits pancreatic cancer cell growth, invasion, and colony formation. *International Journal of Oncology, 25*(1), 203-210.

Li, X., Nair, A., Wang, S., & Wang, L. (2015). Quality control of RNA-seq experiments. *Methods in Molecular Biology (Clifton, N.J.), 1269*, 137-146. doi:10.1007/978-1-4939-2291-8_8 [doi]

Li, Z., Mou, H., Wang, T., Xue, J., Deng, B., Qian, L., et al. (2013). A non-secretory form of FAM3B promotes invasion and metastasis of human colon cancer cells by upregulating slug expression. *Cancer Letters, 328*(2), 278-284. doi:10.1016/j.canlet.2012.09.026 [doi]

Liberatore, M. J., & Nydick, R. L. (2008). The analytic hierarchy process in medical and health care decision making: A literature review. *European Journal of Operational Research, 189*(1), 194. doi:http://dx.doi.org/10.1016/j.ejor.2007.05.001"

Lin, C. Y., Wang, J. K., Torri, J., Dou, L., Sang, Q. A., & Dickson, R. B. (1997). Characterization of a novel, membrane-bound, 80-kDa matrix-degrading protease from human breast cancer cells. monoclonal antibody production, isolation, and localization. *The Journal of Biological Chemistry, 272*(14), 9147-9152.

Line, A., Slucka, Z., Stengrevics, A., Li, G., & Rees, R. C. (2002). Altered splicing pattern of TACC1 mRNA in gastric cancer. *Cancer Genetics and Cytogenetics, 139*(1), 78-83. doi:S0165460802006076 [pii]

List, K., Haudenschild, C. C., Szabo, R., Chen, W., Wahl, S. M., Swaim, W., et al. (2002). Matriptase/MT-SP1 is required for postnatal survival, epidermal barrier function, hair follicle development, and thymic homeostasis. *Oncogene, 21*(23), 3765-3779. doi:10.1038/sj.onc.1205502

List, K., Szabo, R., Molinolo, A., Sriuranpong, V., Redeye, V., Murdock, T., et al. (2005). Deregulated matriptase causes ras-independent multistage carcinogenesis and promotes ras-mediated malignant transformation. *Genes & Development, 19*(16), 1934-1950. doi:10.1101/gad.1300705

Liu, J., Hu, G., Chen, D., Gong, A. Y., Soori, G. S., Dobleman, T. J., et al. (2013). Suppression of SCARA5 by Snail1 is essential for EMT-associated cell migration of A549 cells. *Oncogenesis, 2*, e73. doi:10.1038/oncsis.2013.37 [doi]

Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., & Guigo, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters, 579*(9), 1900-1903. doi:S0014-5793(05)00253-X [pii]

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. doi:s13059-014-0550-8 [pii]

Malcovati, L., Karimi, M., Papaemmanuil, E., Ambaglio, I., Jadersten, M., Jansson, M., et al. (2015). SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts. *Blood, 126*(2), 233-241. doi:10.1182/blood-2015-03-633537 [doi]

Malone, J. H., & Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology, 9*, 34-7007-9-34. doi:10.1186/1741-7007-9-34 [doi]

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry (Palo Alto, Calif.), 6*, 287-303. doi:10.1146/annurev-anchem-062012-092628 [doi]

Martin, M. (2014). *Cutadapt removes adapter sequences from high-throughput sequencing reads.*http://cutadapt.readthedocs.io/en/stable/index.html

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews.Genetics, 12*(10), 671-682. doi:10.1038/nrg3068 [doi]

Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: Towards a cellular code. *Nature Reviews.Molecular Cell Biology, 6*(5), 386-398. doi:nrm1645 [pii]

Matsumoto, H., Sakamoto, A., Fujiwara, M., Yano, Y., Shishido-Hara, Y., Fujioka, Y., et al. (2008). Cyclic AMP-mediated growth suppression and MAPK phosphorylation in thyroid papillary carcinoma cells. *Molecular Medicine Reports, 1*(2), 245-249.

Matsumoto, K., & Nakamura, T. (1996). Emerging multipotent aspects of hepatocyte growth factor. *Journal of Biochemistry, 119*(4), 591-600.

McDonald, P. C., Winum, J. Y., Supuran, C. T., & Dedhar, S. (2012). Recent developments in targeting carbonic anhydrase IX for cancer therapeutics. *Oncotarget, 3*(1), 84-97. doi:422 [pii]

McGlincy, N. J., & Smith, C. W. (2008). Alternative splicing resulting in nonsense-mediated mRNA decay: What is the meaning of nonsense? *Trends in Biochemical Sciences, 33*(8), 385-393. doi:10.1016/j.tibs.2008.06.001 [doi]

Melamud, E., & Moult, J. (2009). Stochastic noise in splicing machinery. *Nucleic Acids Research, 37*(14), 4873-4886. doi:10.1093/nar/gkp471 [doi]

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews.Genetics, 11*(1), 31-46. doi:10.1038/nrg2626 [doi]

Micke, P., Mattsson, J. S., Edlund, K., Lohr, M., Jirstrom, K., Berglund, A., et al. (2014). Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *International Journal of Cancer, 135*(9), 2206-2214. doi:10.1002/ijc.28857 [doi]

Milde-Langosch, K., Bamberger, A. M., Rieck, G., Kelp, B., & Loning, T. (2001). Overexpression of the p16 cell cycle inhibitor in breast cancer is associated with a more malignant phenotype. *Breast Cancer Research and Treatment, 67*(1), 61-70.

Minagar, A., Shapshak, P., Duran, E. M., Kablinger, A. S., Alexander, J. S., Kelley, R. E., et al. (2004). HIV-associated dementia, alzheimer's disease, multiple sclerosis, and schizophrenia: Gene expression review. *Journal of the Neurological Sciences, 224*(1-2), 3-17. doi:10.1016/j.jns.2004.06.007 [doi]

Minoche, A. E., Dohm, J. C., & Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biology, 12*(11), R112-2011-12-11-r112. doi:10.1186/gb-2011-12-11-r112 [doi]

Mohr, A., Zwacka, R. M., Jarmy, G., Buneker, C., Schrezenmeier, H., Dohner, K., et al. (2005). Caspase-8L expression protects CD34+ hematopoietic progenitor cells and leukemic cells from CD95-mediated apoptosis. *Oncogene, 24*(14), 2421-2429. doi:1208432 [pii]

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods, 5*(7), 621-628. doi:10.1038/nmeth.1226 [doi]

Muller, F. J., Laurent, L. C., Kostka, D., Ulitsky, I., Williams, R., Lu, C., et al. (2008). Regulatory networks define phenotypic classes of human stem cell lines. *Nature, 455*(7211), 401-405. doi:10.1038/nature07213 [doi]

Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium. (2015). Pathway and network analysis of cancer genomes. *Nature Methods, 12*(7), 615-621. doi:10.1038/nmeth.3440 [doi]

Nardin, C., Fitzpatrick, L., & Zippin, J. (2014). **Diverse effects of cAMP signaling in melanoma support the role of distinct cAMP microdomains in melanomagenesis, metastasis, and resistance to therapy.**. *OA Dermatology, 2*(1)

Narla, G., DiFeo, A., Fernandez, Y., Dhanasekaran, S., Huang, F., Sangodkar, J., et al. (2008). KLF6-SV1 overexpression accelerates human and mouse prostate cancer progression and metastasis. *The Journal of Clinical Investigation, 118*(8), 2711-2721. doi:10.1172/JCI34780 [doi]

Ning, Y., Manegold, P. C., Hong, Y. K., Zhang, W., Pohl, A., Lurje, G., et al. (2011). Interleukin-8 is associated with proliferation, migration, angiogenesis and chemosensitivity in vitro and in vivo in colon cancer cell line models. *International Journal of Cancer, 128*(9), 2038-2049. doi:10.1002/ijc.25562 [doi]

Oberst, M. D., Chen, L. Y., Kiyomiya, K., Williams, C. A., Lee, M. S., Johnson, M. D., et al. (2005). HAI-1 regulates activation and expression of matriptase, a membrane-bound serine protease. *American Journal of Physiology.Cell Physiology, 289*(2), C462-70. doi:10.1152/ajpcell.00076.2005

Oberst, M. D., Johnson, M. D., Dickson, R. B., Lin, C. Y., Singh, B., Stewart, M., et al. (2002). Expression of the serine protease matriptase and its inhibitor HAI-1 in epithelial ovarian cancer: Correlation with clinical outcome and tumor clinicopathological parameters. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research, 8*(4), 1101-1107.

Oberst, M. D., Singh, B., Ozdemirli, M., Dickson, R. B., Johnson, M. D., & Lin, C. Y. (2003). Characterization of matriptase expression in normal human tissues. *The Journal of Histochemistry and Cytochemistry : Official Journal of the Histochemistry Society, 51*(8), 1017-1025.

Oberst, M. D., Williams, C. A., Dickson, R. B., Johnson, M. D., & Lin, C. Y. (2003). The activation of matriptase requires its noncatalytic domains, serine protease domain, and its cognate inhibitor. *The Journal of Biological Chemistry, 278*(29), 26773-26779. doi:10.1074/jbc.M304282200

Olsburgh, J., Harnden, P., Weeks, R., Smith, B., Joyce, A., Hall, G., et al. (2003). Uroplakin gene expression in normal human tissues and locally advanced bladder cancer. *The Journal of Pathology, 199*(1), 41-49. doi:10.1002/path.1252 [doi]

Oltean, S., & Bates, D. O. (2014). Hallmarks of alternative splicing in cancer. *Oncogene, 33*(46), 5311-5318. doi:10.1038/onc.2013.533 [doi]

Orian-Rousseau, V. (2010). CD44, a therapeutic target for metastasising tumours. *European Journal of Cancer (Oxford, England : 1990), 46*(7), 1271-1277. doi:10.1016/j.ejca.2010.02.024 [doi]

Oshima, T., Sato, S., Kato, J., Ito, Y., Watanabe, T., Tsuji, I., et al. (2013). Nectin-2 is a potential target for antibody therapy of breast and ovarian cancers. *Molecular Cancer, 12*, 60-4598-12-60. doi:10.1186/1476-4598-12-60 [doi]

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct, 4*, 14-6150-4-14. doi:10.1186/1745-6150-4-14 [doi]

Padfield, E., Ellis, H. P., & Kurian, K. M. (2015). Current therapeutic advances targeting EGFR and EGFRvIII in glioblastoma. *Frontiers in Oncology, 5*, 5. doi:10.3389/fonc.2015.00005 [doi]

Pajares, M. J., Ezponda, T., Catena, R., Calvo, A., Pio, R., & Montuenga, L. M. (2007). Alternative splicing: An emerging topic in molecular and clinical oncology. *The Lancet Oncology, 8*(4), 349-357. doi:10.1016/S1470-2045(07)70104-3

Papetti, M., & Augenlicht, L. H. (2011). MYBL2, a link between proliferation and differentiation in maturing colon epithelial cells. *Journal of Cellular Physiology, 226*(3), 785-791. doi:10.1002/jcp.22399 [doi]

Papkoff, J. (2007). New solid tumor targets for therapeutic monoclonal antibodies. *Expert Opinion on Therapeutic Targets, 11*(5), 585-588. doi:10.1517/14728222.11.5.585 [doi]

Park, J. K., Song, J. H., He, T. C., Nam, S. W., Lee, J. Y., & Park, W. S. (2009). Overexpression of wnt-2 in colorectal cancers. *Neoplasma, 56*(2), 119-123.

Pelengaris, S., & Khan, M. (2013). *The molecular biology of cancer: A bridge from bench to bedside* (2nd ed.) Wiley-Blackwell.

Polakis, P. (2016). Antibody drug conjugates for cancer therapy. *Pharmacological Reviews, 68*(1), 3-19. doi:10.1124/pr.114.009373 [doi]

Ponta, H., Sherman, L., & Herrlich, P. A. (2003). CD44: From adhesion molecules to signalling regulators. *Nature Reviews.Molecular Cell Biology, 4*(1), 33-45. doi:10.1038/nrm1004 [doi]

Potente, M., Gerhardt, H., & Carmeliet, P. (2011). Basic and therapeutic aspects of angiogenesis. *Cell, 146*(6), 873-887. doi:10.1016/j.cell.2011.08.039 [doi]

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Research, 40*(Database issue), D130-5. doi:10.1093/nar/gkr1079 [doi]

Qiu, Y., Hoareau-Aveilla, C., Oltean, S., Harper, S. J., & Bates, D. O. (2009). The anti-angiogenic isoforms of VEGF in health and disease. *Biochemical Society Transactions, 37*(Pt 6), 1207-1213. doi:10.1042/BST0371207 [doi]

Rees, A. J., & Kain, R. (2008). Kim-1/tim-1: From biomarker to therapeutic target? *Nephrology, Dialysis, Transplantation : Official Publication of the European Dialysis and Transplant Association - European Renal Association, 23*(11), 3394-3396. doi:10.1093/ndt/gfn480 [doi]

Riley, T., Sontag, E., Chen, P., & Levine, A. (2008). Transcriptional control of human p53-regulated genes. *Nature Reviews.Molecular Cell Biology, 9*(5), 402-412. doi:10.1038/nrm2395 [doi]

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biology, 12*(3), R22-2011-12-3-r22. Epub 2011 Mar 16. doi:10.1186/gb-2011-12-3-r22 [doi]

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods, 7*(11), 909-912. doi:10.1038/nmeth.1517 [doi]

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England), 26*(1), 139-140. doi:10.1093/bioinformatics/btp616 [doi]

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology, 11*(3), R25-2010-11-3-r25. Epub 2010 Mar 2. doi:10.1186/gb-2010-11-3-r25 [doi]

Romagosa, C., Simonetti, S., Lopez-Vicente, L., Mazo, A., Lleonart, M. E., Castellvi, J., et al. (2011). p16(Ink4a) overexpression in cancer: A tumor suppressor gene associated with senescence and high-grade tumors. *Oncogene, 30*(18), 2087-2097. doi:10.1038/onc.2010.614 [doi]

Roy, B., Haupt, L. M., & Griffiths, L. R. (2013). Review: Alternative splicing (AS) of genes as an approach for generating protein complexity. *Current Genomics, 14*(3), 182-194. doi:10.2174/1389202911314030004 [doi]

Saaty, T. L. (1980). *The analytic hierarchy process, planning, piority setting, resource allocation*. New york: McGraw-Hill.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15*(3), 234. doi:http://dx.doi.org/10.1016/0022-2496(77)90033-5"

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences, 1*(1), 83-98.

Saito, S., Okabe, H., Watanabe, M., Ishimoto, T., Iwatsuki, M., Baba, Y., et al. (2013). CD44v6 expression is related to mesenchymal phenotype and poor prognosis in patients with colorectal cancer. *Oncology Reports, 29*(4), 1570-1578. doi:10.3892/or.2013.2273 [doi]

Sampson, J. H., Archer, G. E., Mitchell, D. A., Heimberger, A. B., & Bigner, D. D. (2008). Tumor-specific immunotherapy targeting the EGFRvIII mutation in patients with malignant glioma. *Seminars in Immunology, 20*(5), 267-275. doi:10.1016/j.smim.2008.04.001; 10.1016/j.smim.2008.04.001

Sancisi, V., Gandolfi, G., Ragazzi, M., Nicoli, D., Tamagnini, I., Piana, S., et al. (2013). Cadherin 6 is a new RUNX2 target in TGF-beta signalling pathway. *PloS One, 8*(9), e75489. doi:10.1371/journal.pone.0075489 [doi]

Sanford, M. (2015). Blinatumomab: First global approval. *Drugs, 75*(3), 321-327. doi:10.1007/s40265-015-0356-3 [doi]

Saunders, L. R., Bankovich, A. J., Anderson, W. C., Aujay, M. A., Bheddah, S., Black, K., et al. (2015). A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. *Science Translational Medicine, 7*(302), 302ra136. doi:10.1126/scitranslmed.aac9459 [doi]

Schroder, F. H. (2009). Review of diagnostic markers for prostate cancer. *Recent Results in Cancer Research.Fortschritte Der Krebsforschung.Progres Dans Les Recherches Sur Le Cancer, 181*, 173-182.

Schulze, A., & Downward, J. (2001). Navigating gene expression using microarrays--a technology review. *Nature Cell Biology, 3*(8), E190-5. doi:10.1038/35087138 [doi]

Scott, A. M., Wolchok, J. D., & Old, L. J. (2012). Antibody therapy of cancer. *Nature Reviews.Cancer, 12*(4), 278-287. doi:10.1038/nrc3236 [doi]

Sebestyen, E., Singh, B., Minana, B., Pages, A., Mateo, F., Pujana, M. A., et al. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Research, 26*(6), 732-744. doi:10.1101/gr.199935.115 [doi]

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics, 16*(1), 59-70. doi:10.1093/bib/bbt086 [doi]

Shimomura, T., Denda, K., Kitamura, A., Kawaguchi, T., Kito, M., Kondo, J., et al. (1997). Hepatocyte growth factor activator inhibitor, a novel kunitz-type serine protease inhibitor. *The Journal of Biological Chemistry, 272*(10), 6370-6376.

Shiroguchi, K., Jia, T. Z., Sims, P. A., & Xie, X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America, 109*(4), 1347-1352. doi:10.1073/pnas.1118018109 [doi]

Shultz, J. C., & Chalfant, C. E. (2011). Caspase 9b: A new target for therapy in non-small-cell lung cancer. *Expert Review of Anticancer Therapy, 11*(4), 499-502. doi:10.1586/era.11.23 [doi]

Sidenius, N., & Blasi, F. (2003). The urokinase plasminogen activator system in cancer: Recent advances and implication for prognosis and therapy. *Cancer Metastasis Reviews, 22*(2-3), 205-222.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research, 19*(6), 1117-1123. doi:10.1101/gr.089532.108 [doi]

Soroceanu, L., Murase, R., Limbad, C., Singer, E., Allison, J., Adrados, I., et al. (2013). Id-1 is a key transcriptional regulator of glioblastoma aggressiveness and a novel therapeutic target. *Cancer Research, 73*(5), 1559-1569. doi:10.1158/0008-5472.CAN-12-1943 [doi]

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*(7239), 719-724. doi:10.1038/nature07943 [doi]

Su, H., Xu, T., Ganapathy, S., Shadfan, M., Long, M., Huang, T. H., et al. (2014). Elevated snoRNA biogenesis is essential in breast cancer. *Oncogene, 33*(11), 1348-1358. doi:10.1038/onc.2013.89 [doi]

Subramanian, N., & Ramanathan, R. (2012). A review of applications of analytic hierarchy process in operations management. *International Journal of Production Economics, 138*(2), 215. doi:http://dx.doi.org/10.1016/j.ijpe.2012.03.036"

Swayze, R. D., & Braun, A. P. (2001). A catalytically inactive mutant of type I cGMP-dependent protein kinase prevents enhancement of large conductance, calcium-sensitive K+ channels by sodium nitroprusside and cGMP. *The Journal of Biological Chemistry, 276*(23), 19729-19737. doi:10.1074/jbc.M005711200 [doi]

Tafreshi, N. K., Lloyd, M. C., Bui, M. M., Gillies, R. J., & Morse, D. L. (2014). Carbonic anhydrase IX as an imaging and therapeutic target for tumors and metastases. *Sub-Cellular Biochemistry, 75*, 221-254. doi:10.1007/978-94-007-7359-2_12; 10.1007/978-94-007-7359-2_12

Takeuchi, T., Harris, J. L., Huang, W., Yan, K. W., Coughlin, S. R., & Craik, C. S. (2000). Cellular localization of membrane-type serine protease 1 and identification of protease-activated receptor-2 and single-chain urokinase-type plasminogen activator as substrates. *The Journal of Biological Chemistry, 275*(34), 26333-26342. doi:10.1074/jbc.M002941200

Tanaka, Y., Patestos, N. P., Maekawa, T., & Ishii, S. (1999). B-myb is required for inner cell mass formation at an early stage of development. *The Journal of Biological Chemistry, 274*(40), 28067-28070.

Tanimoto, H., Underwood, L. J., Wang, Y., Shigemasa, K., Parmley, T. H., & O'Brien, T. J. (2001). Ovarian tumor cells express a transmembrane serine protease: A potential candidate for early diagnosis and therapeutic intervention. *Tumour Biology : The Journal of the International Society for Oncodevelopmental Biology and Medicine, 22*(2), 104-114. doi:50604

Tarasov, K. V., Tarasova, Y. S., Tam, W. L., Riordon, D. R., Elliott, S. T., Kania, G., et al. (2008). B-MYB is essential for normal cell cycle progression and chromosomal stability of embryonic stem cells. *PloS One, 3*(6), e2478. doi:10.1371/journal.pone.0002478 [doi]

Tarasov, K. V., Testa, G., Tarasova, Y. S., Kania, G., Riordon, D. R., Volkova, M., et al. (2008). Linkage of pluripotent stem cell-associated transcripts to regulatory gene networks. *Cells, Tissues, Organs, 188*(1-2), 31-45. doi:10.1159/000118787 [doi]

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research, 21*(12), 2213-2223. doi:10.1101/gr.124321.111 [doi]

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., et al. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell, 18*(1), 11-22. doi:10.1016/j.ccr.2010.05.026 [doi]

Towbin, H., Staehelin, T., & Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proceedings of the National Academy of Sciences of the United States of America, 76*(9), 4350-4354.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-seq. *Bioinformatics (Oxford, England), 25*(9), 1105-1111. doi:10.1093/bioinformatics/btp120 [doi]

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology, 28*(5), 511-515. doi:10.1038/nbt.1621 [doi]

Trusolino, L., & Comoglio, P. M. (2002). Scatter-factor and semaphorin receptors: Cell signalling for invasive growth. *Nature Reviews.Cancer, 2*(4), 289-300. doi:10.1038/nrc779

Tsai, Y. S., Dominguez, D., Gomez, S. M., & Wang, Z. (2015). Transcriptome-wide identification and study of cancer-specific splicing events across multiple tumors. *Oncotarget, 6*(9), 6825-6839. doi:3145 [pii]

Unterholzner, L., Keating, S. E., Baran, M., Horan, K. A., Jensen, S. B., Sharma, S., et al. (2010). IFI16 is an innate immune sensor for intracellular DNA. *Nature Immunology, 11*(11), 997-1004. doi:10.1038/ni.1932

Urruticoechea, A., Alemany, R., Balart, J., Villanueva, A., Vinals, F., & Capella, G. (2010). Recent advances in cancer therapy: An overview. *Current Pharmaceutical Design, 16*(1), 3-10.

Vachtenheim, J., & Ondrusova, L. (2015). Microphthalmia-associated transcription factor expression levels in melanoma cells contribute to cell invasion and proliferation. *Experimental Dermatology, 24*(7), 481-484. doi:10.1111/exd.12724 [doi]

Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research, 169*(1), 1. doi:http://dx.doi.org/10.1016/j.ejor.2004.04.028"

Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research, 322*(1), 12-20. doi:10.1016/j.yexcr.2014.01.008 [doi]

Van Rossum, A. G., de Graaf, J. H., Schuuring-Scholtes, E., Kluin, P. M., Fan, Y. X., Zhan, X., et al. (2003). Alternative splicing of the actin binding domain of human cortactin affects cell migration. *The Journal of Biological Chemistry, 278*(46), 45672-45679. doi:10.1074/jbc.M306688200 [doi]

Vega-Stromberg, T. (2003). Chemotherapy-induced secondary malignancies. *Journal of Infusion Nursing : The Official Publication of the Infusion Nurses Society, 26*(6), 353-361. doi:00129804-200311000-00004 [pii]

Venables, J. P. (2004). Aberrant and alternative splicing in cancer. *Cancer Research, 64*(21), 7647-7654. doi:10.1158/0008-5472.CAN-04-1910

Venables, J. P., Klinck, R., Bramard, A., Inkel, L., Dufresne-Martin, G., Koh, C., et al. (2008). Identification of alternative splicing markers for breast cancer. *Cancer Research, 68*(22), 9525-9531. doi:10.1158/0008-5472.CAN-08-1769; 10.1158/0008-5472.CAN-08-1769

Verma, S., Miles, D., Gianni, L., Krop, I. E., Welslau, M., Baselga, J., et al. (2012). Trastuzumab emtansine for HER2-positive advanced breast cancer. *The New England Journal of Medicine, 367*(19), 1783-1791. doi:10.1056/NEJMoa1209124 [doi]

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature, 456*(7221), 470-476. doi:10.1038/nature07509 [doi]

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. *Nature Reviews.Genetics, 10*(1), 57-63. doi:10.1038/nrg2484 [doi]

Ward, A. J., & Cooper, T. A. (2010). The pathobiology of splicing. *The Journal of Pathology, 220*(2), 152-163. doi:10.1002/path.2649 [doi]

Wheeler, S. E., Egloff, A. M., Wang, L., James, C. D., Hammerman, P. S., & Grandis, J. R. (2015). Challenges in EGFRvIII detection in head and neck squamous cell carcinoma. *PloS One, 10*(2), e0117781. doi:10.1371/journal.pone.0117781 [doi]

White, C. D., Khurana, H., Gnatenko, D. V., Li, Z., Odze, R. D., Sacks, D. B., et al. (2010). IQGAP1 and IQGAP2 are reciprocally altered in hepatocellular carcinoma. *BMC Gastroenterology, 10*, 125-230X-10-125. doi:10.1186/1471-230X-10-125 [doi]

Will, C. L., & Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology, 3*(7), 10.1101/cshperspect.a003707. doi:10.1101/cshperspect.a003707 [doi]

Williams, G. T., & Farzaneh, F. (2012). Are snoRNAs and snoRNA host genes new players in cancer? *Nature Reviews.Cancer, 12*(2), 84-88. doi:10.1038/nrc3195 [doi]

Wong, M. S., Chen, L., Foster, C., Kainthla, R., Shay, J. W., & Wright, W. E. (2013). Regulation of telomerase alternative splicing: A target for chemotherapy. *Cell Reports, 3*(4), 1028-1035. doi:10.1016/j.celrep.2013.03.011 [doi]

Wong, S. F. (2005). Cetuximab: An epidermal growth factor receptor monoclonal antibody for the treatment of colorectal cancer. *Clinical Therapeutics, 27*(6), 684-694. doi:S0149-2918(05)00096-2 [pii]

Wu, S. R., Cheng, T. S., Chen, W. C., Shyu, H. Y., Ko, C. J., Huang, H. P., et al. (2010). Matriptase is involved in ErbB-2-induced prostate cancer cell invasion. *The American Journal of Pathology, 177*(6), 3145-3158. doi:10.2353/ajpath.2010.100228; 10.2353/ajpath.2010.100228

Yan, N., Zhang, S., Yang, Y., Cheng, L., Li, C., Dai, L., et al. (2012). Therapeutic upregulation of class A scavenger receptor member 5 inhibits tumor growth and metastasis. *Cancer Science, 103*(9), 1631-1639. doi:10.1111/j.1349-7006.2012.02350.x [doi]

Yang, Y., & Smith, S. A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics, 14*, 328-2164-14-328. doi:10.1186/1471-2164-14-328 [doi]

Yang, Y., Zhao, W., Xu, Q. W., Wang, X. S., Zhang, Y., & Zhang, J. (2014). IQGAP3 promotes EGFR-ERK signaling and the growth and metastasis of lung cancer cells. *PloS One, 9*(5), e97578. doi:10.1371/journal.pone.0097578 [doi]

Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B., & Makeyev, E. V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes & Development, 26*(11), 1209-1223. doi:10.1101/gad.188037.112 [doi]

Young, S. Z., & Bordey, A. (2009). GABA's control of stem and cancer cell proliferation in adult neural and peripheral niches. *Physiology (Bethesda, Md.), 24*, 171-185. doi:10.1152/physiol.00002.2009 [doi]

Youssoufian, H., Rowinsky, E. K., Tonra, J., & Li, Y. (2010). Targeting FMS-related tyrosine kinase receptor 3 with the human immunoglobulin G1 monoclonal antibody IMC-EB10. *Cancer, 116*(4 Suppl), 1013-1017. doi:10.1002/cncr.24787 [doi]

Yu, S., Murph, M. M., Lu, Y., Liu, S., Hall, H. S., Liu, J., et al. (2008). Lysophosphatidic acid receptors determine tumorigenicity and aggressiveness of ovarian cancer cells. *Journal of the National Cancer Institute, 100*(22), 1630-1642. doi:10.1093/jnci/djn378 [doi]

Zatovicova, M., Jelenska, L., Hulikova, A., Csaderova, L., Ditte, Z., Ditte, P., et al. (2010). Carbonic anhydrase IX as an anticancer therapy target: Preclinical evaluation of internalizing monoclonal antibody directed to catalytic domain. *Current Pharmaceutical Design, 16*(29), 3255-3263. doi:BSP/CPD/E-Pub/000218 [pii]

Zhang, Q., Chen, G., Liu, X., & Qian, Q. (2007). Monoclonal antibodies as therapeutic agents in oncology and antibody gene therapy. *Cell Research, 17*(2), 89-99. doi:7310143 [pii]

Zhou, Y. Q., He, C., Chen, Y. Q., Wang, D., & Wang, M. H. (2003). Altered expression of the RON receptor tyrosine kinase in primary human colorectal adenocarcinomas: Generation of different splicing RON variants and their oncogenic potential. *Oncogene, 22*(2), 186-197. doi:10.1038/sj.onc.1206075 [doi]

Zona, S., Bella, L., Burton, M. J., Nestal de Moraes, G., & Lam, E. W. (2014). FOXM1: An emerging master regulator of DNA damage response and genotoxic agent resistance. *Biochimica Et Biophysica Acta, 1839*(11), 1316-1322. doi:10.1016/j.bbagrm.2014.09.016 [doi]

Appendix A.

Cell surface cancer-associcated abnormally expressed genes across TCGA cancers

This table is attached as an excel file.

Appendix B.

Putative biomarker target pairs for therapeutic bispecific antibodies

This table is attached as an excel file.

Appendix C.

Cell surface cancer-specific spliced variants across TCGA cancers

This table is attached as an excel file.

Appendix D.

Final prioritization of putative biomarker genes by Prize R package

This table is attached as an excel file.