IDENTIFYING NOVEL TARGETABLE GENES AND PATHWAYS IN CANCER BY INTEGRATING DIVERSE OMICS DATA

by

Oscar Alejandro Balbin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2014

Doctoral Committee:

Professor Arul. M. Chinnaiyan, Co-Chair
Associate Professor Alexey I. Nesvizhskii, Co-Chair
Assistant Professor James Cavalcoli
Associate Professor Jun Li
Professor Gilbert S. Omenn

This thesis is dedicated to my family

To my parents, they inculcated in me the importance of education

To **Jesús William**, my father, he taught me how to read and later in his life, he taught me how to fight cancer and how to demystify this disease

To **Matilde**, my mother, she taught me the strength of tenderness

To **Diego**, my brother and his beautiful daughter **Ana Sofía**, they form a dynamic duo and have filled my life with happiness

To **Vilma**, for her support

To **Liliana**, my wife, lover and friend, she infuses my life with love every new sunrise

Esta tesis es dedicada a mi familia

A **mis padres**, quienes juntos me enseñaron la importancia de la educación

A **Jesús William**, mi padre, él me enseño a leer y luego, al final de su vida, me enseñó como pelear contra el cáncer y como desmitificar esa enfermedad

A **Matilde**, mi madre, ella me enseñó la fuerza de la ternura y me regalo su amor de madre

A **Diego**, mi hermano y a su hermosa hija **Ana Sofía**, quienes juntos son un dúo dinámico y han llenado mi vida de alegría

A **Vilma**, por su apoyo

A **Liliana**, mi esposa, amante y amiga, ella inunda de amor mi vida cada nuevo amanecer

Acknowledgements

This journey started seven years ago, when I decided to move to the United States pursing my academic dreams and, as in all good stories, the love of a wonderful woman. It has been a fantastic expedition, full of surprises and challenges: a new culture, a new language, seasons and the -20 Celsius, new ways of understanding the world and of course new and exciting science. Through this voyage I have had the fortune of meeting many friends, colleagues and mentors who have given me words of advice, encouragement and vast amounts of support. This thesis is a living document because I was able to count on all of you along this journey. To all of you, I want to express my deepest gratitude.

I would like to thank my two advisors Dr. Arul M. Chinnaiyan and Dr. Alexey I. Nesvizhskii. I was very fortunate to have you both as mentors; your complementary perspectives contributed to the improvement of different aspects of my work and science. Arul, thank you for giving me the freedom and the challenge to explore, search and come up with new projects and new ideas for this thesis. Your thoughtful insights and perspectives have helped me understanding the value of translational research. Alexey, thank you for providing me with rigorous methodological foundations for analyzing complicated datasets. Your ability and speed to identify the correct and misleading trends in figures and other data representations have always challenged me to improve my analysis and to come up with better ways to understand the hidden patterns in the data. I have such high regards and respect for both of you as scientists and team leaders and I hope that we are collaborators for many years to come.

I would like to express heartfelt thanks to my thesis committee – Dr. Gilbert Omenn, Dr. Jun Li, and Dr. James Cavalcoli. Thank you for all your suggestions and comments on all my three projects. I truly believe that each one of them have contributed to improve the quality of my research, my presentation skills and overall the science formation. Gil, I deeply appreciate your support on all my fellowship applications and your kind disposition for finding time in your busy calendar to meet with me and discuss about my research. Jim, your insights about the job market gave me a clear perspective that was essential on deciding the next stage on my career. Jun, your very intelligent comments on each of my projects always motivated me to develop new ways of analyzing the data and displaying the results.

I would also like to thank the Department of Computational Medicine and Bioinformatics. I was privileged to have been part of such an inter-disciplinary department that gave me the opportunity to grow as a data scientist and cancer researcher. Julia, thank you for making my life easier by always being on top of everything and more importantly for having a smile ready every time that I visited your office. To my friends Dattatreya Mellacheruvu, Michelle Wynn, Claudia McDonald, Rohit Malik and John Prensner thanks so much for all the laughter and support through this years.

I also wish to thank all of my colleagues in both the Chinnayian and Nesvizhskii labs. The results presented in this thesis were made possible because of the great collaboration and support of a large team of people in both laboratories. John Prensner, Mohan Dhanasekaran, Anastasia Yocum, Anirban Sahu, Rohit Malik, Damian Fermin, Ernest Nadal, Sunita Shankar, David Beer, Matthew Iyer, Chad Brenner, Brendan Veeneman, Dattatreya Mellacheruvu, Terrence Barrette, Dan Robinson, Yi-Mi Wu, Shanker Kalyana-Sundaram and

many others. Xuhong Cao thanks so much for organizing everything in the lab and making sure that all runs smoothly from the sequencing machine to the MCTP annual party! Diana Banka and Jyoti Athanikar thanks for your great help on fellowship and grant writing. Karen Giles thanks for making our life easier by taking care of all the overwhelming administrative responsibilities, from securing funding to submitting articles. Christine Betts, running Arul's schedule deserves a PhD in itself thanks for finding slots of time for me in his very packed schedule.

I would like to thank my undergraduate and master's thesis advisor and friend professor Eugenio Andrade. Thank you for encouraging me to pursue a scientific career since my early beginnings at the Universidad Nacional de Colombia. I hope I can contribute to the advancement and development of science in Colombia, my beautiful, diverse, and amazingly complex home country.

To my family, thank you for all your unconditional love, constant encouragement and support. Thank you for believing in me, for being there every time that I need you, for listening, for giving me reasons to keep going, in essence for being the most fundamental reason of happiness.

To Liliana, my wife, lover, friend and accomplice, thank you for sharing and helping me, hand to hand, to make my dreams reality. Thank you for your deep love, for believing in me, for reminding me with your simple presence how beautiful life is and how small things can make us tremendously happy, thank you for chasing sunrises and sunsets with me … in every language Te amo.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AKT1 v-akt murine thymoma viral oncogene homolog 1

ALK anaplastic lymphoma receptor tyrosine kinase

BDNF brain-derived neurotrophic factor

BRAF v-raf murine sarcoma viral oncogene homolog B1

CD74 CD74 molecule, major histocompatibility complex, class II invariant chain

CDKN2A cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)

DDR2 discoidin domain receptor tyrosine kinase 2

DHPS deoxyhypusine synthase

EML4 echinoderm microtubule associated protein like 4

ERBB3 v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)

FGFR1 fibroblast growth factor receptor 1

FGFR3 fibroblast growth factor receptor 3

HIF1A hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)

HOTAIRM1 HOX antisense intergenic RNA myeloid 1

HOXD1 homeobox D1

HOXD3 homeobox D3

HRAS v-Ha-ras Harvey rat sarcoma viral oncogene homolog

KRAS v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog

LCK lymphocyte-specific protein tyrosine kinase

MLL2 myeloid/lymphoid or mixed-lineage leukemia 2

NF1 neurofibromin 1

NKX2 NK2 Homeobox 1

NOTCH1 Notch homolog 1, translocation-associated (Drosophila)

NRAS neuroblastoma RAS viral (v-ras) oncogene homolog

NRG1 neuregulin 1

PIK3CA phosphoinositide-3-kinase, catalytic, alpha polypeptide

PTEN phosphatase and tensin homolog; phosphatase and tensin homolog pseudogene 1

RET ret proto-oncogene

ROS1 c-ros oncogene 1 , receptor tyrosine kinase

SOX2 SRY (sex determining region Y)-box 2

STK11 serine/threonine kinase 11

TP53 tumor protein p53

TPPP tubulin polymerization promoting protein

WDR83 mitogen-activated protein kinase organizer 1

WRAP53 WD repeat containing, antisense to TP53

WT1 Wilms tumor 1 NSCLC, Non-small cell lung cancer

WT-AS1 Wilms tumor 1 antisense 1

CDKN2A-AS cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4), antisense

BDNF-AS brain-derived neurotrophic factor, antisense

NKX2-1-AS NK2 Homeobox 1, antisense

HOXD-AS1 homeobox D antisense1

NSCLC Non-small cell lung cancer

LUAD Lung adenocarcinoma

LUSC Lung squamous carcinoma

RNASeq RNA sequencing

ssRNASeq Strand specific RNA sequencing

TCGA The Cancer Genome Atlas

ICGC International Cancer Genome Consurtium

PCST Prize Collecting Steiner Tree

KRAS-Dep KRAS dependent

KRAS-Ind KRAS independent

S-AS Sense and antisense transcripts pairs

HTH Head-to-head, transcripts that overlap by the 5'regions

TTT Tail-to-tail, transcripts that overlap by the 3'regions

EMD Overlapping transcripts in which one of the transcripts is fully contained within the other.

cis-NAT All overlapping pairs.

ASloci Loci with significant expression of the antisense strand

OPSratio Reverse read counts/(Forward read counts + Reverse read counts))

Abstract

The rapid evolution of omics technologies for profiling the human genome, transcriptome and proteome is revolutionizing cancer research and driving a paradigm shift in clinical care. Omics have forever changed our view of cancer, from a uniform disease to a highly heterogeneous ecosystem of diseases driven by different genetic events. Standard care is, as well, evolving from “one size” fits all treatments towards more precise and molecularly informed therapies. The success of this precision medicine paradigm will depend on our ability to integrate diverse omics measurements to distill clinically relevant information that can be act upon. This thesis developed bioinformatics approaches to integrate multi-omics datasets and applied these approaches in three distinct studies that identified novel actionable genes and pathways in cancers.

In the first study, we aim at finding alternative target proteins in cancer samples that share activating mutations in *KRAS* a well-known, but undruggable, oncogene. We profile the transcriptome, proteome and phosphoproteome in a panel of non-small cell lung cancer (NSCLC) cell lines in order to reconstruct targetable networks associated with KRAS dependency. A bioinformatics strategy was developed addressing the challenge of integrating these disparate datasets and the Prize Collecting Steiner Tree algorithm was used to identify functional sub-networks. Three modules centered on KRAS and MET, LCK and PAK1 and β-Catenin were identified. We validated activation of these proteins in KRAS-dependent cells and performed functional studies defining *LCK* as a critical gene for cell proliferation in KRAS-dependent but not KRAS-independent NSCLCs. These results are the first evidence that suggest LCK as a potential druggable target protein in KRAS-dependent lung cancers.

In the second study, the landscape of fusions in lung adenocarcinoma and lung squamous carcinoma tissue types was described in order to identify potentially oncogenic gene fusions in driver negative patients. The landscape was found to be highly heterogeneous and gene fusions incidence was discovered to be an independent prognostic factor for poor outcome. By integrating gene mutation status, the lung cohort was divided into driver positive and driver negative patients (who do not have mutations in known cancer genes). Focusing in driver negative patients we identify NRG1 as a novel low recurrence 3' fusion partner present exclusively in this subset; resembling previously reported kinase fusions. The documented success of targeted therapies against low recurrence oncogenic fusions in lung cancer and the high heterogeneity of the fusions' landscape, shown in this study, reinforce the demand for more personalized and tailored drug therapies.

Finally in the third study, the landscape of antisense expression in human cancers was characterized in order to identify sense-antisense gene pairs involving tumor suppressors and oncogenes, which could be suitable for emerging antisense-targeted therapies. More than 60% of DNA loci were found to have measurable antisense transcription. Expression of sense and antisense transcript pairs is in general positively correlated and directed by bidirectional promoters in cases of overlapping divergent genes. By comparing with known sense-antisense pairs, our results raise the possibility that antisense transcripts could be regulating the expression of well-known tumor suppressors and oncogenes. This study provides a resource, oncoNATdb, a catalogue of cancer related genes with significant antisense transcription, which will allow cancer researchers to investigate the mechanisms of sense-antisense regulation and further advance our understanding of their role in cancer.

We anticipate that the computational methods developed and the results found in this thesis would assist others with similar tasks and warrant further studies of the therapeutic opportunities provided by these novel targets.

Chapter 1
Multi-omics data integration

1.1 Background and significance

The collective characterization and quantification of pools of biological molecules such as genes, transcripts and proteins have emerged as the complete new fields of genomics, transcriptomics and proteomics. Collectively, these and others high-throughput fields are known as Omics.

Omics technologies for high-throughput profiling of the human genome, transcriptome and proteome are revolutionizing cancer research and driving a paradigm shift on clinical care. Omics have forever changed our view of cancer, from a uniform disease to a highly heterogeneous ecosystem of diseases driven by different genetic events. Standard care is, as well, evolving from “one size” fits all treatments towards more precise and molecularly informed therapies. This precision medicine paradigm depends on our ability for integrating diverse omics measurements to distill clinically relevant information that can be act upon. This dissertation focuses on developing bioinformatics approaches to integrate multi-omics datasets to identify novel actionable genes and pathways in cancer. In three independent studies we integrate multi-omics cancer data in order to reconstruct novel targetable pathways in KRAS dependent lung cancer, search for novel oncogenic fusions in lung cancer patients with no known driver genes and study sense/antisense gene regulation in cancer. Our results warrant further studies of the therapeutic opportunities provided by these novel targets.

1.2 Molecular characterization of lung cancers

Lung cancer is the leading cause of cancer mortality in the world with more than one million deaths a year[1]. Non-small cell lung cancer (NSCLC) is the most predominant type of this malignancy, and it can be subdivided into lung adenocarcinoma (LUAD) and lung squamous carcinoma (LUSC).

Recent genomic analyses have deepened our understanding of the genetic alterations characterizing both LUAD and LUSC, and have revealed very different mutation landscapes. LUAD are mutated in several well-characterized oncogenes and tumor suppressor genes including KRAS (~30%), EGFR (~14%), BRAF (~10%), *TP53* (~46%), and *STK11* (~17%) among others[2]. Importantly, activating mutations in KRAS are mutually exclusive with activating mutations in EGFR. On the other hand, LUSC is characterized by mutations in *TP53* (~81%)*, CDKN2A (15%), PTEN (8%), PIK3CA (16%), DDR2, AKT1, MLL2, NOTCH1,* and *RB1* as well as several recurrent gene copy number alterations of *FGFR1, SOX2* and *TP63*[3]. This molecular heterogeneity underlies the difficulties in effectively treating patients with this disease.

Remarkably, despite this deep molecular characterization of lung cancer, there is still above 30% of patients with no known driver genes. This driver negative subpopulation has been recently subject to intense study and additional driver events such as oncogenic gene fusions have been discovered.

Several important gene fusions occur in lung cancer including the *EML4-ALK* fusion gene identified in approximately 4% of adenocarcinomas[4]. This fusion protein links the N-terminal portion of echinoderm microtubule-associated protein-like 4 (EML4) with the intracellular signaling portion of the anaplastic lymphoma kinase (ALK) tyrosine kinase receptor. The *EML4-ALK* translocation is

mutually exclusive with *EGFR* and *KRAS*[5]. Additional gene fusions have now been identified in LUAD involving *ROS1* [6], as well as *RET* [7,8] kinases as 3' partner genes.

1.2.1 KRAS mutations lung cancers

As mentioned above, mutations in the Ras oncogenes characterize 30-40% of all NSCLC, with KRAS, NRAS, and HRAS being somatically mutated in 30%, 1-5%, and 1% of the cases respectively [9]. Recent studies suggest that a subset of KRAS mutant tumors are dependent on KRAS for survival[10], implying that targeting KRAS or other genes downstream in this signaling cascade could yield potential drug targets to treat NSCLC.

Ras is a GTP binding protein that communicates signaling information through five major cancer related pathways: Akt/PI3K, Raf –MAPK, RalGDS, phospholipase-Ce, and Rac [11]. Mutations on residues 12, 13 and 61 in the GTPase pocket disrupt Ras GTPase activity generating constitutively active Ras proteins, which in turn affect transcription of numerous genes promoting cell proliferation and survival. Microarray profiling has been extensively used for defining gene expression signatures characterizing Ras activity in cell lines and tissues[10,12,13] but results are inconsistent across studies. Complicating this matter, it has been shown that NSCLC can be subdivided into KRAS-dependent (KRAS-Dep) and KRAS-independent (KRAS-Ind) according to their requirement of KRAS for survival[10,14]; and more importantly KRAS pathway activity predicts KRAS dependency and drug resistance better than mutation status[12].

The active KRAS signaling pathway transmits information in the form of post-translational modification such as phosphorylation. Although previous studies have used semi or quantitative phospho-proteomics experiments to characterize these pathways[15-17], none of those studies profiled simultaneously

gene expression, protein abundance and protein phosphorylation status[6,15-19], as we have accomplished in chapter 2, missing the key opportunity of synthesizing all these levels of information.

1.2.2 Targeted therapies

Lung cancer patients whose tumors harbor *EGFR* activating mutations show responsiveness to drugs inhibitors such as Erlotinib and Gefitinib that target these particular alterations[20]. More importantly, the presence of the *ALK* fusions is an indicator of therapeutic responsiveness to *ALK* inhibitors such as Crizotinib[4]. These results have accelerated the development of new drug inhibitors targeting additional genetic aberrations such as *ROS1* and *RET* fusions and mutations in *FGFR* and *PI3KCA* among others.

Remarkably, however, it is the absence of therapeutic options for treating the two biggest groups of lung cancers: patients with mutations in *KRAS* and driver negative patients.

1.2.3 Drug therapies targeting Ras pathways

The development of drug therapies aimed at disrupting Ras activity or blocking Ras pathways has proved challenging[9]. However, Ras driven tumors could harbor vulnerabilities in other pathways due to proteins which are not oncogenic themselves but are required for Ras dependency[21]. Therefore, inhibitors targeting various Ras effectors could be indirectly effective in treating tumors driven by Ras activity[22]. RNAi profiling studies aiming at identifying genes whose inhibition constitutes synthetic lethality with KRAS have identified vulnerable points in networks as diverse as the mitotic[21], the epithelial differentiation[10] and NF-kB pathways[14]. Each study reported a different but not overlapping set of vulnerable genes. Remarkably, no strong susceptibility points

were detected in the Akt/PI3K and Raf –MAPK pathways neither in NSCLC cell lines[10,14] nor in the KRAS dependent DLD-1 colorectal cell line[21], indicating that other poorly characterized pathways may be contributing to the KRAS-induced oncogenic state.

1.3 Omics Technologies

Second generation sequencing, or next generation sequencing (NGS) as was initially named, allows researchers to sequence billions of DNA strands in parallel generating substantially more high throughput than conventional Sanger sequencing. Although recently developed, NGS technologies are being applied in a variety of fields ranging from gene mutation profiling, gene fusions detection, novel transcripts discovery, transcript expression, ribosome profiling and nascent RNA characterization to mention just a few. Through this dissertation we have primarily used DNA exome sequencing (ExomeSeq), RNA sequencing (RNASeq) and strand-specific RNA sequencing (ssRNASeq).

1.3.1 DNA Sequencing

In 2008 the first whole cancer genome was sequenced using NGS and since then several more genomes have been sequenced as part of The Cancer Genome Atlas project (TCGA). Although very useful for detecting large chromosomal rearrangements and somatic mutations of non-coding regions including promoters, enhancers and un-annotated regions the cost of sequencing the whole genome is still high enough to prevent its implementation on a routine basis. Exome sequencing, or targeted NGS of the coding regions of the genome is a more cost effective approach to reliably detect somatic mutations in the regions of interest due to increased sequence coverage gained by concentrating all the sequencing efforts to a small region of the genome. In contrast to

maximum sequence coverage of about 30x seen for whole genome sequencing, ExomeSeq can typically deliver greater than 100x over the targeted regions.

1.3.2 RNA sequencing

NGS of the transcriptome, or RNASeq, has been used to profile mRNA, total RNA and small RNAs in cancer and normal samples. RNASeq allows transcript quantification, transcript discovery and detection of in-frame oncogenic gene fusions, as well as alternative splice variants. RNASeq can also be used to detect somatic mutations; however, determining this without matched normal is challenging because normal tissues should not express the same gene repertoire as cancers. Moreover, gene expression level and infrequent mechanisms, such as RNA editing, need to be taken into account when using RNASeq for somatic mutation calling. Despite these challenges, several studies have properly used RNASeq to determine somatic mutations, cleverly restricting the analysis to well-known mutations in cancer genes[23].

1.3.3 Strand specific RNA sequencing

Standard RNASeq does not preserve information about which DNA strand was originally transcribed. In this method double stranded cDNA is randomly primed followed by addition of adaptors for NGS. In this process the information about what strand was present in the original mRNA template is lost. Strand information can improve the value of RNASeq experiments by providing accurate information about antisense transcripts, helping to clearly delimit gene boundaries of adjacent genes and to correctly resolve the expression levels of overlapping transcripts.

Although many methods have been developed for generating strand specific RNASeq, they rely on one of three approaches: ligation of adaptors in

predetermined orientation to ends of the RNAs molecules, direct sequencing of the first-strand cDNA products, or selective marking of the second-strand cDNA and subsequent degradation of the first-strand cDNA. Selective labeling is accomplished by using dUTP during cDNA synthesis or by bisulfite conversion of cytosine to uracil in RNA. Levin *et al.,* 2010[24] compared the performance of seven ssRNASeq protocols and observed differences with respect to their level of strand specificity, evenness of coverage, agreement with known annotations, library complexity and the ability to generate quantitative expression profiles. They identified the dUTP labeling methods and Illumina RNA adaptor-ligation methods as the leading protocols, with dUTP libraries providing the additional advantage of conducting pair-end sequencing.

1.3.4 Proteomics

Proteomics high-throughput methods, for profiling the abundance and post-translational modifications of proteins, is providing deep insights about the proteome organization of normal and cancer tissues[25].

In particular quantitative phospho-proteomics, label or label-free, allows researchers to characterize signaling pathways[15-17]. A general pipeline for label-free phospho-proteome quantification is summarized in the following steps[26-41]: tryptic sample preparation, phospho-peptide enrichment, label-free quantitative tandem mass spectrometry, peptide identification through database search, and quantification. It is important to note that phospho-peptide enrichment is necessary because phospho-peptides correspond to a small fraction of all peptides obtained after tryptic digestion. Several enrichment methods have been proposed [36], such as immobilized metal affinity chromatography (IMAC), Titanium or Zirconium dioxide (TiO2, ZrO2) and phosphoamidate chemistry (PAC), which is the one most commonly used [36,41,42]. In addition, phospho-tyrosine peptides

are typically under-represented with respect to phospho-Ser/Thr peptides, but they play an important role upstream and downstream of many signaling cascades. Finally, peptides and phospho-peptides are quantified by label or label-free liquid chromatography-tandem mass spectrometry (LC-MS/MS).

Improvements in LC-MS/MS mass spectrometers are increasing the coverage of the proteome that is achieved in a single experiment to above 10,000 proteins [43]. Continuous improvements on the mass spectrometers resolution will close the gap between the number of transcripts identified by RNASeq experiments and the number of proteins identified by LC-MS/MS, allowing for better comparisons between the transcriptome and proteome of matched samples.

1.4 Software used for data processing

In order to organize the following sections, the data levels schema proposed by TCGA consortium[1] was adopted. There are four data levels: Level 1 (for Raw Data), Level 2 (for Processed Data), Level 3 (for Segmented or Interpreted Data) and Level 4 (for Summary or Region of Interest Data). Table 1.1-1 below describes the data levels with examples. The aims of this dissertation focus primarily on integration of level 3 and 4 datasets; however, a significant amount of work was devoted to generate bioinformatics pipelines for processing raw data from level 1 to level 3 for hundreds of samples. A great deal of expertise on the computational tools for cancer genomics was gained through this process. A list of computational tools for processing next generation sequencing data and proteomics is given in Table 1.1-2.

[1] https://wiki.nci.nih.gov/display/TCGA/Data+Classification

Table 1.1-1. Data levels as adopted from the Cancer Genome Atlas project.

Data level	Level type	Description	Example
1	Raw	Low-level data for single sample Not normalized	Raw sequences Raw spectra
2	Processed	Normalized single sample data Interpreted for presence or absence of specific molecular abnormalities	Germline and Somatic mutations
3	Segmented/Interpreted	Aggregate of processed data from single sample	Gene expression data for all genes across sample and cohort
4	Summary	Quantified association across classes of samples	Integrative analysis and association between molecular variables and clinical parameters

Table 1.1-2. Computational tools for cancer genomics and proteomics.

Category	Method	Comments	Refs
Alignment	BWA	Genome and exome alignment	44
	Bowtie/Bowtie2	Transcriptome alignment	45, 46
	ELAND	Transcriptome and genome alignment	
	TopHat2	Transcriptome alignment	47
Mutation Calling	VarScan2	Germline and somatic mutation calling	48
	GATK	Germline and somatic mutation calling	49
	Samtools	SNV mutation calling	50
Gene Expression	Cufflinks	Gene expression quantification, FPKM	51
	Samtools	Gene expression quantification, read counts	50
	HTSeq-count	Gene expression quantification, read counts	
Differential Expression Analysis	DESeq	Negative binomial and variance estimation	52
	Cufflinks	Differential expression analysis	53
	EdgeR	Negative binomial and variance estimation	54
Fusion Calling	TopHat-Fusion	Fusion discovery from pair end sequencing	55
	ChimeraScan	Fusion discovery from pair end sequencing	56
	Defuse	Fusion discovery from pair end sequencing	57
Proteomics	X!Tandem	mzXML search	
	PeptideProphet and ProteinProphet	Post-processing of X!Tandem Searches	58, 59

1.5 Integration of omics datasets

The term data integration is used in different contexts and does not always have the same meaning. The term is often employed to describe tools, methods and software used to interrogate different data sources such as databases or federate data repositories. The term is also utilized when combining related studies in order to obtain stronger conclusions or to increase the power of previous studies by collecting more data of the same type. Finally, the term is used when combining diverse and heterogeneous data types, measured in the same individual, in order to improve our understanding of a biological process or to uncover previously unappreciated relationships or measurements. Throughout this dissertation we will use the term data integration referring to the second and third examples described above.

Genomics, transcriptomics, proteomics and epigenomics, to mention just a few of the omics platforms, each provides a one dimensional view of the cell components; integrative analysis promises a global and systemic view of these levels and their interactions. However, the huge amount of information obtained from each of these omics technologies and diversity in the platforms discussed above pose multiple bioinformatics challenges for data processing and combination.

Despite these challenges collaborative projects such as the Encode project (ENCODE) and The Cancer Genome Atlas project (TCGA) are generating vast muti-omics datasets. The Encode project has deepened our understanding of gene elements and gene regulation, while TCGA is providing a full characterization of more than 25 different cancer types. These efforts highlight the importance and challenges of multi-omics data integration[60].

There are numerous methods for integrating omics datasets. This diversity is explained by the fact that the types of data integration used in a particular situation depend on the scientific question motivating the analyses. In general,

however, algorithms for multi-omics data integration usually belong to one of few categories: data reduction supervised algorithms and unsupervised algorithms.

1.5.1 Data reduction

In omics projects data reduction happens at several levels. For example, in RNASeq and ChipSeq the first step of data processing is to reduce the million of reads produced by the sequencing machine to a few hundreds or thousands of points (integers) representing the amount of those reads over a genomic interval. Then, a popular method of data integration is to perform intersection analysis between those genomic intervals and other regions from the same or different experiment, such as ChIPSeq peaks indicating specific chromatin marks.

In a second level of data integration, data reduction statistical methods such as principal component analysis, multiple factor analysis and non-negative matrix factorization[61,62] aim at reducing or transforming the variable space into one that represents the stronger tendencies in the data. Importantly, these methods are individually applied to each data set and then the results are combined. When applied to the full multi-omics datasets these methods usually depend on a strong correlation between the genomic levels, which are being combined. For example the positive correlation observed between the number of copies of a gene and its transcription level has been exploited for several methods to integrate copy number changes profiles (CNVs) with gene expression profiles[63,64].

1.5.2 Unsupervised data integration

In unsupervised learning the goal is to summarize a large dataset into smaller groups that can be easily understood. The methods in this category answer the question, “What are the more frequent patterns present in the

dataset?" An assumption to have in mind when applying these methods is that the patterns that are discovered are usually the ones that appear more frequently, and therefore outlier events would not be identified despite their potential importance. An important caveat of these methods is that they will always find a pattern in the data regardless of its biological significance.

Similarly to data reduction methods a standard approach with these unsupervised, or clustering, methods is to first find clusters in each data set and then map clusters between data types. The mapping procedure is not unsupervised and usually constitutes the most demanding part, for example clusters found at the gene expression level, representing co-expressed genes, are analyzed in light of clusters found in the chromatin level, representing co-regulated loci. Because clustering methods are agnostic, as long as "pattern frequency" represents genes/proteins activity, they can be used, in principle, for summarizing multiple omics datasets at once as long as the data are properly normalized in advance and a strong correlation exists between data types[65].

1.5.3 Limitations of data reduction and unsupervised methods

As described above statistical data reduction techniques and unsupervised clustering algorithms depended on strong correlation between data types in order to be useful in multi-omics integration tasks. This degree of correlation has been shown for example, between gene copy number and gene expression and activating chromatin marks and gene expression.

However, this degree of correlation between omics datasets is not the norm and it is, indeed, not expected when integrating somatic mutations and gene expression, mRNA and protein abundance, or mRNA levels and phosphorylation status. In these cases the degree of correlation observed is usually low and in some cases it is determined by true biological factors.

Mutations abolishing gene expression, unstable mRNAs that are not correlated with protein abundance level, and proteins with constant abundance, but differential levels of phosphorylation, are some examples. The lack of correlation in those contexts is therefore not only due to noise but also due to the biological phenomena that are being studied.

Another challenge when integrating heterogeneous omics datasets is the wide level of coverage between different technologies. For example, a single RNASeq experiment can identify above 20,000 expressed transcripts, while a very good proteomics experiment will identify at the most 10,000 proteins. Similarly, the number of genes harboring somatic non-synonymous mutations in a sample is typically less than a 1000, and usually closer to 100. The number of phospho-proteins detected varies greatly depending on the enrichment protocol and LC-MS/MS mass spectrometer instrument used. This wide range of coverage generated by different omics platforms creates a high number of missing values and sparsity. The high number of missing values impedes the use of data reduction methods such as standard principal component analysis.

1.6 Network biology approaches to data integration

Another approach to multi-omics-based data integration is network biology. This approach leverages our current knowledge about the systemic relationships between the different components; such as, genes, proteins, and their interactions, and at the same time find new relationships in the data. Molecular pathways and protein-to-protein interactions are typical representations of our current knowledge about the molecular interactions in the cell.

Network biology approaches address the challenges imposed by the low correlations between omics-based measurements and the diverse coverage

range of different omics platforms, by using the pathways or networks as common frameworks over which the information obtained from different omics-based measurements is combined. Overlaying different omics measurements on top biological networks brings functional information in order to make sense of the information gained through a multi-omics-based experiment. These approaches, are, therefore, becoming a common strategy for multi-omics data analysis.

1.6.1 Network reconstruction

Identification of pathways, modules, or functional sub-networks is a central theme in understanding oncogenesis from an integrative perspective, as well as a very challenging computational problem[66]. However, multi-omics data integration has been successful in building more complete models of cancer molecular networks[29,67-80].

Numerous computational methods are being proposed to identify functional and/or differential expressed modules[68,77-79,81-94]. Those methods can be sub-divided into two main different approaches according to their use of *a priori* information regarding the network of interactions. Inference methods that do not use a priori information require vast amounts of data in order to estimate their model's parameters, which make them inappropriate for small datasets with few conditions. Methods that use *a priori* information depend on the quality and extent of this information. Fortunately, our knowledge of biological pathways and interactions is increasing constantly and pathway models can be refined and updated as needed. The focus of this dissertation is on the second type of methods, but a detailed comparison of both can be found elsewhere[95]. Within methods that use *a priori* information, methods proposed so far aim at finding a dense connected sub-network, based on a pre-specified protein-to-protein interaction network (PPI) and gene expression data. Gene expression profiles

are treated as a snapshot of the dynamic behavior of the system, while the PPI, although incomplete, represents the universe of potential interactions.

From a computational point of view, extracting functional sub-modules from high throughput omics-based data can be formulated as an optimization problem whose objective function is defined according to specific requirements. There are two types of module extraction methods heuristic and exact[90]. Heuristic methods were used to find cancer modules that distinguish breast cancer subtypes[78] and to organize the Reactome database into pathways[81]. Heuristic approaches cannot guarantee the optimality of their solutions, whereas 'exact' methods do so. Exact methods commonly employ integer or mixed-integer linear programming techniques in order to find optimal solutions to the network extraction problem[84,91,96]. Among exact approaches, the Prize Collecting Steiner Tree (PCST) formulation has been successfully applied to find functional sub-networks in yeast and cancer[77,842]. In the second chapter of this dissertation the problem of integrating NSCLC transcriptome, proteome and phospho-proteome datasets will be formulated as Prize Collecting Steiner Tree Problem, which solutions allow us to reconstruct active networks in KRAS dependent cells.

1.7 Aims and structure of this thesis

This dissertation focuses on developing bioinformatics approaches to integrate multi-omics datasets. As emphasized through this first chapter, data integration itself is designed to generate novel hypothesis that can be experimentally or computationally tested in order to answer specific scientific questions. The bioinformatics approaches developed in this thesis are all aimed at identifying novel actionable genes and pathways in cancer. These approaches were applied to find novel targets in three distinct scenarios, representing different cancer patient populations with unmet therapeutic needs.

[2] Notably, Zhao, et al 2008 algorithms can be formulated as the PCST problem used for Dittrich, et al., 2008.

In the first study, we aim at finding alternative target proteins in cancer samples sharing activating mutations in *KRAS* a well known, but undruggable, oncogene. We profile the transcriptome, proteome and phosphoproteome in a panel of non-small cell lung cancer (NSCLC) cell lines in order to reconstruct targetable networks associated with KRAS dependency. We develop a bioinformatics strategy addressing the challenge of integrating these disparate datasets and use the Prize Collecting Steiner Tree algorithm to identify functional sub-networks. We identify three modules centered on KRAS and MET, LCK and PAK1 and β-Catenin. We validate activation of these proteins in KRAS-dependent cells and perform functional studies defining *LCK* as a critical gene for cell proliferation in KRAS-dependent but not KRAS-independent NSCLCs. These results are the first evidence to suggest LCK as a potential druggable target protein in KRAS-dependent lung cancers.

In the second study, we describe the fusions landscape of lung adenocarcinoma and lung squamous carcinoma tissue types in order to identify potentially oncogenic gene fusions in driver negative patients. We show the high heterogeneity of this landscape and discover that gene fusions incidence is an independent prognostic factor for poor outcome. By integrating gene mutation status, we divide the cohort into driver positive and driver negative patients, who do not have mutations in known cancer genes. Focusing in driver negative patients we identify NRG1 as a novel low recurrence 3' fusion partner present exclusively in this subset; resembling previously reported receptor kinase fusions. The documented success of targeted therapies against low recurrence oncogenic fusions in lung cancer and the high heterogeneity of the fusions' landscape, shown in this study, reinforce the demand for more personalized and tailored drug therapies.

Finally in the third study, we characterize the landscape of antisense expression in human cancers in order to identify sense-antisense gene pairs involving cancer related genes, which could be suitable for emerging antisense targeted therapies. We show that > 60% of DNA loci have measurable antisense

transcription and that the expression of sense and antisense transcript pairs is in general positively correlated and directed by bidirectional promoters in cases of overlapping divergent genes. By comparing with known sense-antisense pairs, our results raise the possibility that antisense transcripts could be regulating the expression of well-known tumor suppressors and oncogenes. This study provides a resource, oncoNATdb, a catalogue of cancer related genes with significant antisense transcription, which will allow cancer researchers to investigate the mechanisms of sense-antisense regulation and further advance our understanding of their role in cancer.

These studies are presented consecutively in chapters 2, 3, and 4, followed for general conclusions and the future directions of this work.

Chapter 2
Reconstructing targetable pathways in KRAS dependent lung cancers

The content of this chapter was previously published by the author as an original article in Nature Communications [70].

2.1 Background

Activating mutations in the Ras oncogenes characterize 20-40% of all non-small cell lung cancer (NSCLC)[9,97,98], the leading cause of cancer mortality in the United States[99], which establishes Ras genes as the most commonly mutated oncogenes in this malignancy. KRAS, NRAS, and HRAS, the main members of this family of GTPase proteins, are activated by somatic mutations in 20-30%, 1-5%, and 1% of the NSCLC cases respectively[9]. Mutated Ras has been implicated in activating numerous pathways that control cell proliferation and survival; however, development of drug therapies aimed at disrupting Ras activity has proved challenging[9]. Consequently, recent efforts have focused on identifying indirect mechanisms to disrupt Ras signaling by targeting either upstream activators or downstream effectors[13,14,22,100,101]. To this end, microarray gene expression profiling has been extensively used to define expression signatures characterizing Ras mutations in cell lines and tumors[12,79,82], but gene signatures vary considerably across these studies.

Complicating these initial studies, recent work has shown that NSCLCs with activating KRAS mutations can be stratified into KRAS-dependent (KRAS-Dep) or KRAS-independent (KRAS-Ind) groups according to their requirement for

mutant KRAS signaling to sustain growth and proliferation[10,13,101,102]. Therefore, after shRNA knock down of KRAS, KRAS-Ind cells would grow at rates resembling cells treated with control shRNAs, while KRAS-Dep grow at slower rates. Here, gene expression profiles of NSCLC cell lines found that KRAS dependency correlated with a differentiated phenotype, whereas KRAS independency was associated with the epithelial mesenchymal transformation phenotype[10,102]. Moreover, recent work associated KRAS dependency with activation of the Wnt signaling pathway in colorectal cancers[102]. Taken together, these results suggest that specific pathways are activated in KRAS-Dep cell lines but not in KRAS-Ind cells, and that those pathways play a role in the varying disease phenotypes found in these cancers.

While such expression profiling studies are useful for the analysis of KRAS signaling, it is well established that KRAS frequently exerts oncogenic functions through changes in protein abundance or post-translational modifications of proteins, specifically kinases that in turn induce a signaling cascade of downstream effectors[15,17,40,103]. Consequently, global transcriptome, proteome and phosphophospho-proteome profiling methods should be applied in order to identify causative pathways in KRAS-Dep and KRAS-Ind NSCLC cells in an unbiased fashion. However, to date no study has comprehensively integrated these diverse sets of data[14,15,18,40,79,82,103], leading to potential biases and inadequacies in our understanding of the mechanistic basis for KRAS function in NSCLC.

One reason why such studies are lacking is because integration of such diverse datasets is a major challenge with existing integrative methods. Yet when employed, integrative methods have been successful in building more comprehensive models of molecular signaling networks in cancer[67,68].

2.2 Aims of this study

In this study we generate a matched dataset of KRAS-mutated NSCLC cell lines with global and unbiased transcriptome, proteome and phosphoproteome profiles. We develop a bioinformatics approach to integrate these disparate omics datasets and nominate biologically informative signaling modules using network analysis. We find that KRAS-dependent cell lines harbor an active and targetable sub-network composed of lymphocyte-specific tyrosine kinase (*LCK*), *cMET*, *KRAS* and the p21 serine/threonine activated kinase (*PAK1*). We characterize a KRAS-LCK-PAK1 pathway and show that KRAS-Dep, but not KRAS-Ind cell lines require LCK for proliferation. This KRAS-LCK-PAK1 network further coordinates anti-apoptotic pathways both through inhibition of pro-apoptotic proteins such as BAD and/or activation of anti-apoptotic proteins in KRAS-Dep cell lines. In summary this study identifies active networks associated with the KRAS-dependent phenotype in NSCLC and nominates a novel KRAS-LCK-PAK1 pathway in KRAS-Dep cells that may serve as a druggable pathway for treating KRAS-dependent lung cancers.

2.3 Bioinformatics Methods

2.3.1 Protein quantification by label free LC-MS/MS

The mass spectrometry proteomics and phosphoproteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000439. The general workflow used for label-free phosphoproteome quantification is summarized in the following steps[26-41]: sample preparation, phospho-peptides enrichment, label-free quantitative tandem mass spectrometry, peptide identification through database search, and quantification by the spectral count method. Cell lines were grown on vendors recommended media until they were 70% confluent and then protein extraction

and sample preparation was performed as previously reported[40] in the presence of proteases and phosphatases inhibitors.

For mass Spectrometry eluted proteins were separated by 1D SDS-PAGE (4-12% Bis-Tris Novex-Invitrogen, Carlsbad, CA). 24 equal-sized gel bands were excised and subjected to in-gel tryptic digestion. Because phospho-peptides correspond to a small fraction of all peptides of after tryptic digestion, phospho-peptide enrichment was performed using immobilized metal affinity chromatography (IMAC). Tryptic peptides were then divided into two fractions: phospho-enriched and flow-through or unmodified peptides. Both fractions of extracted peptides were independently reconstituted with mobile phase A prior to on-line reverse phase nanoLC-MS/MS (LTQ-Velos with Proxeon nanoHPLC, ThermoFinnigan). Peptides were eluted on-line to the mass spectrometer with a reverse phase linear gradient from 97% A (0.1 % formic acid in water) to 45 % B (0.1 % formic acid in acetonitrile) over 60-minutes. Peptides were detected and fragmented in the mass spectrometer in a data-dependent manner sending the top 12 precursor ions that exceeded a threshold of 500 ion counts, excluding singly charged ions, for collisional-induced dissociation. Dynamic mass exclusion was enabled with a repeat count of 2 for 1.5 minutes for a list size of 500 *m/z*.

For the database search raw spectra files were converted to mzXML using ReadAW. The mzXML files were searched using X!Tandem with the k-score plug-in[104]. The proteomic searches were performed using the following options: allow up to 2 missed tryptic cleavages, a parent ion tolerance window of -1 to +4 Daltons, and a fragment ion tolerance of 0.8 Da. The following variable modifications were allowed: phosphorylation of Serine, Threonine, and Tyrosine (+79.966331@[STY]), oxidation of Methionine (+15.994920@M), and carbamidomethylation of Cysteine (+57.021460@C). All protein searches were performed using the Human Refseq protein database (release 47). Appended to this database were common proteomic contaminants and reversed protein

sequences to serve as decoys[105,106]. The X!Tandem results were then post processed with PeptideProphet and ProteinProphet[58,59]. Spectral counts were then obtained for all of the proteins identified in our cohort of 13 cell lines using the Abacus software tool[107]. For Abacus, the following parameters were used: count only peptide-to-spectrum-matches (PSMs) with a PeptideProphet score above 0.5 (iniProbTH=0.50), retain only proteins with at least one peptide with a PeptideProphet score above 0.99 (maxIniProbTH=0.99) and a ProteinProphet probability greater than 0.9 in the COMBINED file (minCombinedFilePw=0.90). For the phosphorylated fraction, peptides were required to have at least one phosphorylated Serine, Threonine or Tyrosine (reqAAmods=+S[167];+T[181];+Y[243]). Proteins and phospho-proteins identified with at least one spectral count in two independent cell lines were kept for downstream analysis (Balbin *et al.,* 2013[70], Supplementary Data 1, 3), while those identified in one cell line only were filter out (Balbin *et al.,* 2013[70] Supplementary Data 2, 4).

The spectrum counts for each protein were normalized with respect to the total number of spectrum counts within each sample. This normalization was applied independently for unmodified and modified proteins. Common contaminants and "Deja vu"[108] proteins were filter out before quantification of differentially abundant proteins. For both, unmodified and phosphorylated proteins, the fold change was calculated with respect to the comparison KRAS-Dep vs. KRAS-Ind cell lines. This fold change was then log transformed and z-score normalized. Finally, the p-value was calculated using the standard normal distribution. The final master tables with the normalized spectrum counts for phosphorylated and flow through fraction for each cell line are provided as Supplementary Data 1 and 2 in Balbin *et al* 2013.

Phospho-enrichment was calculated as the ratio between the number of phospho-peptides identified and the total number of peptides (phosphorylated and unphosphorylated) at a particular PeptideProphet score for the best peptide

match (bestInitProbability). All enrichment calculations were made using only peptides that have Ser, Thr or Tyr in them. Peptides without any of those amino acids were excluded from the calculation. Finally, the phospho-enrichment value is taken for a PeptideProphet score above 0.94 (bestInitProbability=0.9413), which produces a 0.01 FDR. The calculated phospho-peptide enrichment, for all samples, ranges from 26 to 38%.

2.3.2 Gene Expression Data

Gene expression data used in this study are publicly available at ArrayExpress with accession number E-MTAB-783. Gene expression was scaled and log2 normalized previous to additional downstream analysis.

2.3.3 Integration of Datasets

Because different protein functional groups (e.g. transcription factors, kinases or secretory proteins) have distinct gene expression dynamic range, the gene expression dataset was split into two different categories: “informative” genes and “all other” genes and subsequently analysis were performed independently on each one of them. “Informative” refer to genes that are well known to drive a carcinogenic process such as *KRAS*, *TP53*, *ERBB2* and *CDKN2A*, etc., as well as to genes that could have the potential to drive oncogenesis as kinases, phosphatases among others. A list of “Informative” genes was compiled by combining the Sanger’s cancer census genes, all kinases and phosphatases as well as additional and recently reported genes important for carcinogenesis (Balbin *et al.,* 2013[70] Supplementary Data 8).

Raw data was preprocessed as described in the experimental methods section. Phosphoproteome, proteome and transcriptome datasets were log transformed and the log fold change (LFC) was taken with respect to the

comparison KRAS-Dep vs KRAS-Ind cell lines. The LFC was z-score normalized and a p-value was calculated using the standard normal distribution.

In order to synthesize for each protein the information obtained from gene expression, protein and phospho-protein abundance, we calculated a combined abundance S score as $S = \frac{\sum_i^k w_k z_i}{\sqrt{\sum w_k^2}}$, where z is the z-transformed LFC of protein i in the dataset k, while w corresponds to the weight of each dataset $w_k = 1/\sqrt{N_k}$. N_k represents the size of dataset k.

Finally a p-value for the combined score was calculated using the standard normal distribution and then adjusted using the Hochberg procedure in order to correct for multiple hypothesis testing.

2.3.4 Network Analysis

We use the Signaling Pathway Impact Analysis Algorithm (SPIA[109]) in order to perform network enrichment analysis. The source code for this algorithm is available as an R package from http://bioconductor.org/biocLite.R. SPIA calculates the significance of a pathway according to both the over-representation evidence (e.g. any commonly used enrichment test) and perturbation's based evidence using the topology of the network. The KEGG database (http://www.genome.jp/kegg/kegg1.html) was used as the main source for pathway's definition and we used the set of differential expressed genes as defined by the combined abundance score with adjusted p-value <=0.05 as the seed genelist. Significant pathways with FDR <=0.05 are reported (Table A-1).

For the Network reconstruction methodology, we built a focused undirected and weighted protein-to-protein interaction network (G) using significant (FDR<=0.05) pathways identified by the SPIA algorithm[109]. Those

pathways were downloaded from the KEGG database[65] and then merged into a unified meta-pathway (G) using the bioconductor KEGGgraph library[66]. This meta-pathway (G) is provided for the interested reader as Supplementary Data 9 in Balbin *et al.,* 2013[70].

We assigned weights to both nodes (V) and edges (E). Node weights correspond to the combined score (S) for differential abundance between KRAS-Dep and KRAS-Ind phenotypes, while the edge weights correspond to the experimental confidence on that interaction as derived from the STRING database. For each edge in the meta-pathway, we obtained from STRING the experimental and physical interaction scores and then combined them into a single score using a naïve Bayes approach. In addition, in order to decrease redundancy, multiple gene family members with the same interaction partners were summarized into a "consensus gene" defined as the gene with highest scoring interaction neighborhood. This step is advised due to the node redundancy introduced within the KEGG database and the fact that the interactions for many gene family members are annotated by similarity to other members in the family and not by direct experimental validation.

Finally, we used the Prize Collecting Steiner Tree (PCST) algorithm to find sub-networks, T, in the meta-pathway (G) that represent the most differentially abundant proteins connected through the most reliable interactions. Formally, the PCST is formulated as follows:

$$T = \min_{\substack{E' \subseteq E;\, V' \subseteq V \\ (E', V')\, connected}} \left(\sum_{e \in E'} c_e - \lambda \sum_{v \in V'} b_v \right) \text{[1]}$$

where $b_v = -\log p(S)$ with $p(S)$ as the p-value for the S score of each protein, and $c_e = 1 - \prod_i^k R_i$ with R_i for the string score for the edge's physical and experimental evidence. This choice of b_v and c_e assigns high values to the most differentially abundant proteins in the pathway, and low values to the high confidence interactions in the network. Finally, the constant λ controls the trade off of adding new proteins into the reconstructed network, by balancing the cost

of new edges and the prize gained by bringing in a new protein. λ indirectly controls the size of the final sub-networks. All results presented here were obtained with λ=0.3. In order to choose λ, we solved the prize collecting steiner problem, varying λ between 0.01-1 in increments of 0.01, and choose the value of λ at which 60% of the essential nodes of simulated network of similar size were recovered. In order to solve the PCST, we used the implementation based on information message passaging described by [110], for which the source code availability is annotated in the Table A-2.

The PCST has been used in similar settings before[75,84,110] because it identifies sub-networks that represent cross talk between pathways, as well as *"connecting proteins"* that are not directly measured in the experiment but that are relevant to link other measured proteins with high weight in the network.

2.3.5 Analysis of LCK knock-down experiments

We used the Signaling Pathway Impact Analysis Algorithm (SPIA) as described above to identify pathways specifically activated or inhibited after LCK knockdown (Table A-5), confirming the involvement of a lung cancer pathway but more importantly several pathways controlling apoptosis induction such as the natural killer cell-mediated cytotoxicity, Toll-like receptor signaling and the NOD-like receptor signaling pathway. This is in agreement with the fact that Module M1 containing LCK and PAK1 were enriched for proteins belonging to the apoptosis pathways (Figure A.7A). Therefore, we focused the additional analysis of the microarray data on identifying altered proteins belonging to the apoptosis pathways.

To perform *BCL2A1* nomination we first collect apoptosis gene concepts from KEGG, gene ontology and Reactome and generate a meta-apoptosis gene concept with all unique genes found. We reasoned that proteins specifically activated by LCK should simultaneously satisfy the following three characteristics: to be overexpressed when comparing KRAS-Dep vs KRAS-Ind cells, to be under-expressed when comparing the LCK knock down vs. the non-targeting control in H441 and H358 cell lines and to be unaffected after knocking down any other gene in different cell lines. Characteristic 3 is included to control for changes in gene expression induced by any knockdown treatment irrespective of the gene of interest.

Representing conditions 1, 2, and 3 in Cartesian plot results in a plot shown in Figure A.8A. The x-axis shows the differential expression of those genes when comparing KRAS-Dep vs KRAS-Ind cell lines. The y-axis shows the average differential expression of the same genes when comparing a siRNA knockdown of LCK in H441 and H358 cell lines with respect to the targeting control (red dots), or the average differential expression when comparing the knockdown of a “random” gene compared to its respective control (black dots) in three unrelated prostate cell lines. Genes affected by the overall siRNA treatment would by overlapping or very close in this plot, while genes specifically affected by LCK would be located far apart in the y-axis. We measure this effect by taking the Euclidean distance between red and black dots representing the same gene in the above representation.

Genes that are specifically affected by LCK would have positive or negative Euclidean distances according to the magnitude of their perturbation, while genes nonspecifically affected by the siRNA treatment would have Euclidean distances close to 0 (Figure A.8).

2.4 Experimental Methods

2.4.1 Cell lines

All cell lines were obtained from ATCC and maintained using standard procedures. Specifically, H441, H358, H2009, H1734, H727, H460, H2122, H1792, H23, H1155 cells were maintained in RPMI 1640 (Gibco) plus 10% FBS and 1% penicillin-streptomycin. A549 cells were maintained in DMEM (Gibco) plus 10% FBS and 1% penicillin-streptomycin. SKLU1 cells were maintained in DMEM/F12 plus 10% FBS and 1% penicillin-streptomycin. SW900 cells were maintained in L15 plus 10% FBS and 1% penicillin-streptomycin. Cell lines were grown at 37°C in a 5% CO_2 cell culture incubator. All cell lines were genotyped for identity at the University of Michigan Sequencing Core.

2.4.2 shRNA knock down studies

For LCK and KRAS knockdowns all cells were plated at 100000 cells/ml in 6 well plates and let them attached overnight. Cells were infected next day with the lentivirus RNA and 24 hours after infection old media was replaced with new cell media. Cells were allowed to grow for 96 hours in this fresh media. At this point cells were treated with 1mg/ml puromycin for 5 days to eliminate uninfected cells. Media was replaced and proliferation assays set up with the stable selected clones. Knockdown efficiency was confirmed by Western blot. shRNA sequences are provided in the supplementary methods.

2.4.3 siRNA knockdown studies

Cells were plated in 100mM plates at 30% confluency and transfected twice at 12 hours and 24 hours post-plating. Knockdowns were performed using 20uM siRNA oligos or non-targeting controls (Dharmacon) with Oligofectamine

(Invitrogen) in Opti-MEM media (Gibco). Knockdown efficiency was confirmed by Western blot. siRNA used are listed in the supplementary methods. 72 hours post-transfection, cells were rinsed twice with 10mL PBS, harvested with a rubber policeman in 1mL PBS and centrifuged for 5 min at 2,500x g. The supernatant was discarded and the cells were prepared for Western blot analysis.

2.4.4 Western Blots

Cell pellets were lysed in RIPA lysis buffer (Sigma) supplemented with HALT protease inhibitor and phosphatase inhibitor (Fisher). Western blotting was performed using standard protocols. Briefly, protein lysates were boiled in sample buffer for 5 min at 98C and 10ug of protein was separated by SDS-PAGE gel electrophoresis. Proteins were transferred onto a PVDF membrane (GE Healthcare) and blocked for 30 minutes in blocking buffer (5% milk in 1x TBS supplemented with 0.1% Tween (TBS-T)). Membranes were incubated with primary antibody overnight at 4C and then with secondary antibody for 2 hours at room temperature. Signals were visualized by enhanced chemiluminescence system (GE Healthcare). The primary antibodies used are listed in the supplementary methods and full blots can be found in Supplementary Fig S9-S15 in Balbin *et al.,* 2013.

2.4.5 Proliferation Assays

Proliferation assays were performed with stable clones of the scramble RNA, and two independent constructs against LCK or KRAS for each cell line. Cells were plated at 30000 cells/ml in 24 well plates and cell counts were taken with a Beckman coulter Z2 particle count instrument every 48 hours for 8 days. Three independent replicates of each experiment were performed.

2.4.6 WST Drug Assays

Cells were plated in a 96-well plate 12 hours prior to drug treatment at a density of 3500 cells per well in a 100ul of growth media. Desired concentrations of LCK Inhibitor (Santa Cruz, sc-204052,CAS 213743-31-8) and LCK Inhibitor II (Millipore, Lck Inhibitor II,CAS 918870-43-6) were prepared using growth media and 100ul of the drug solution was added directly to the wells. After 72 hours of incubation at 37C, 20ul of WST Cell proliferation reagent (Roche) was added to each well. Following 2 hours of incubation at 37C, the absorbance of the wells was measured at 450nm.

2.4.7 Confocal microscopy

H460 and H441 cells were fixed with 3.7% paraformaldehyde, and then permeabilized with 0.1% (w/v) saponin for 15 min. Cells were co-incubated with primary antibodies against phosphor □-catenin and total beta catenin for 12hr at 4 °C, followed by incubating with appropriate Alexa-Fluor-conjugated secondary antibodies for 30 min at 37 °C. Cells were washed and mounted onto glass slides using Vectashield mounting medium containing DAPI. Samples were analyzed using a Nikon A1 laser-scanning confocal microscope equipped with a Plan-Apo ×63/1.4 numerical aperture oil lens objective. Acquired images were then analyzed using ImageJ software (version 1.41o).

2.4.8 KRAS Genotyping

Genomic DNA from resected lung cancer tissue samples was prepared using a Qiagen Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions. KRAS mutations were determined using standard RT-PCR and Sanger sequencing protocols for *KRAS* exon 1, which harbors codons 12 and 13, and exon 2, which harbors codon 61. RT-PCR was performed with 5 ng genomic

DNA with 38 cycles of PCR according to the following conditions: 94C for 30 seconds, 56C for 30 seconds, 68C for 45 seconds. PCR products were subsequently purified using ExoSAP-IT PCR purification product (USB/Affymetrix) according to the manufacturer's instructions. PCR products were then unidirectionally sequenced using the M13 forward primer at the University of Michigan Sequencing Core. Sequence data was analyzed for the presence of canonical activating *KRAS* mutations at codons 12, 13, and 61. Primers used for the PCR reactions are listed in the supplementary methods.

2.4.9 Immunohistochemistry

Immunohistochemical (IHC) analyses on paraffin-embedded formalin-fixed (FFPE) tumor tissue sections were carried out using the automated DiscoveryXT staining platform from Ventana Medical Systems. All FFPE sections were represented in triplicate on the tissue microarray. The primary rabbit monoclonal LCK antibody was obtained from Cell Signaling (#2984). Antigen recovery was conducted using heat retrieval and CC1 standard, a high pH Tris/borate/EDTA buffer (VMSI, catalogue no. 950-124). Slides were incubated with 1:50 of the LCK antibody (Cell Signaling) overnight at room temperature. Primary antibody was detected using the ChromoMap DAB detection kit (VMSI, catalogue no. 760-159) and UltraMap anti-Rb HRP (VMSI, catalo no. 760-4315). The anti-Rb HRP secondary antibody was applied for 30 minutes at room temperature. Slides were counterstained with Hematoxylin for 10 minutes followed by Bluing Reagent for 5 minutes at 37C. Staining was scored (DG Beer) as negative (score = 0), minimal (score = 1), weak (score = 2), moderate (score = 3), or high (score = 4).

2.5 Results

2.5.1 Omics integration improves the nomination of actionable protein

To study KRAS function in lung cancer, we generated matched global transcriptome, proteome and phosphoproteome datasets for a panel of KRAS-Dep and KRAS-Ind NSCLC cell lines, as well as a bioinformatics methodology to integrate all those data types (Figure 2.1A). Transcript, protein and phospho-protein abundance were measured by microarrays and label free LC-MS/MS respectively (Methods). We identified 3213 proteins in the unmodified state and 1044 proteins in the phosphorylated state, with at least 1 spectrum count in two independent cell lines. The number of unique peptides and phospho-peptides for each cell line are shown in the Figure A.1A, Figure A.1B, and the full proteome and phosphoproteome datasets for all cell lines are given in Balbin *et al* 2013[70] Supplementary Data 1, 2 and Supplementary Data 3, 4 respectively.

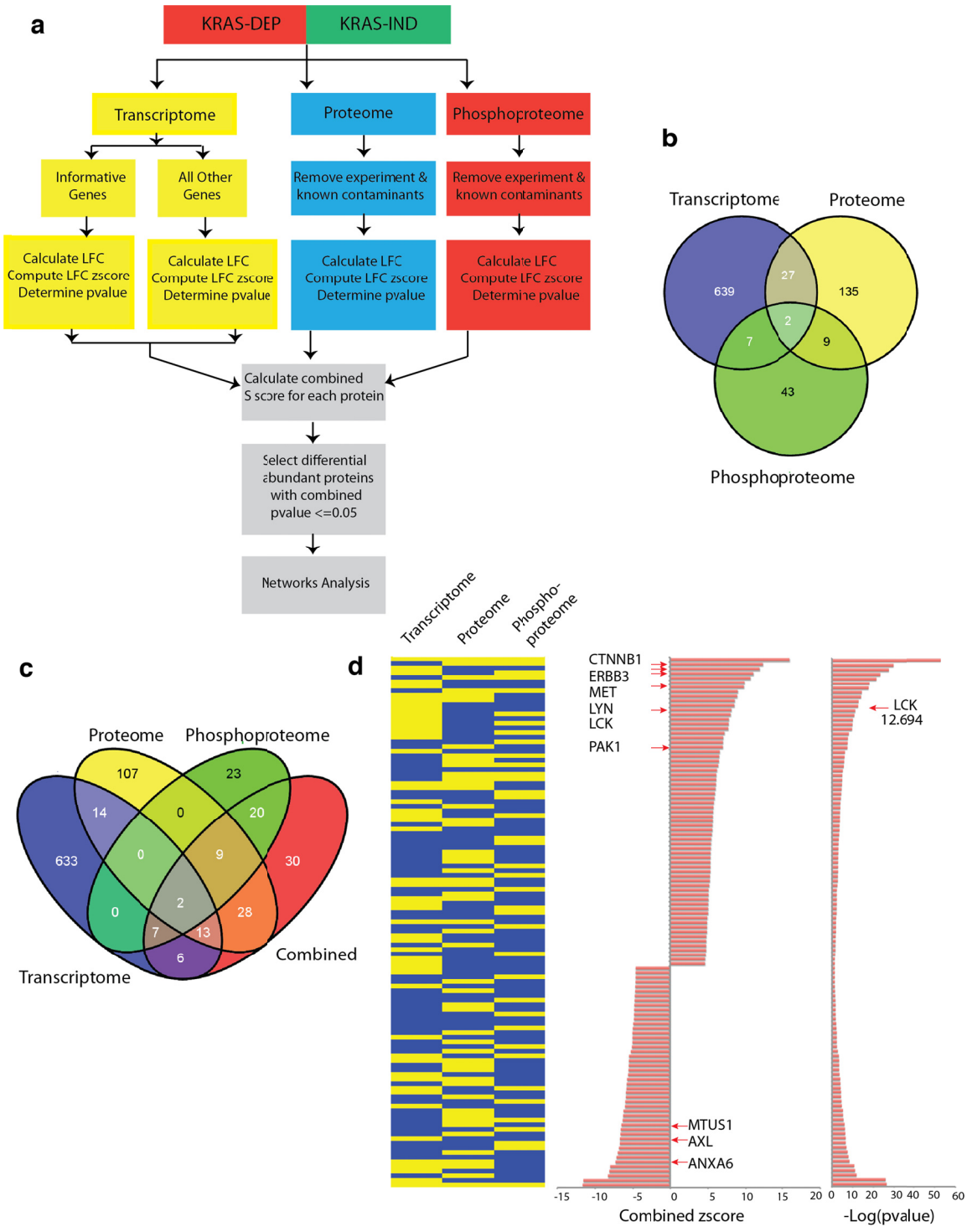


Figure 2.1 Integrative analysis of omics data reveals targetable kinases in NSCLC *KRAS* dependent cell lines.

A) A panel of KRAS-dependent and -independent cell lines was interrogated by transcriptomics, proteomics and phospho-proteomics techniques. Transcripts were split in two different categories: “informative” genes and “all other” genes. Proteome and phopho-proteome datasets were normalized with respect to the total number spectral counts in each library, and common contaminants and “Deja vu” proteins were filtered out before quantification of differentially abundance. All datasets were log transformed and the log fold change (LFC) was taken with respect to the comparison KRAS-Dep vs KRAS-Ind cell lines. The LFC was z-score normalized and a p-value was calculated using the standard normal distribution. The combined S score was used to integrate all three datasets (methods) and select differentially expressed proteins. Network and enrichment analysis were performed using the Signaling Pathway Analysis Algorithm (SPIA) and the Prize Collecting Steiner Tree Algorithm (PCST). **B)** Naïve integration of datasets. Only ~ 5.2 % of the proteins are shared among two of the datasets (adjusted p-value <= 0.05 was used as a threshold to select differentially expressed proteins). A major drawback of this method is the absence of an objective criterion to include proteins differentially expressed in only one dataset. **C)** A meta-integration of the independent signatures using the combined S score (S). The S score integration improves by five-fold the percentage of shared proteins among datasets (~ 26 %), and defines an objective rule for including proteins differentially expressed in one, two or all datasets. **D)** Integrative analysis of transcriptome, proteome and phospho-proteome nominates receptor tyrosine kinases *MET* and *ERBB3*, Src family members *LCK* and *LYN*, *PAK1*, and *CTNNB1*, *CTNNA1*, and *CDH1* among others as differentially “activated” proteins in KRAS-Dep cell lines. Left) Presence/absence heatmap. Proteins that are differentially abundant in a particular dataset are represented in yellow and unaffected proteins are represented in blue. Middle) Combined S score (S) for all differentially abundant proteins in KRAS-Dep vs KRAS-Ind cell lines. Right) Combined statistical significance each differentially abundant protein. –log of the Hochberg adjusted p-value, -log(0.05)= 1.30.

Integration of transcriptome, proteome and phosphoproteome data is challenging due to differences in technological methods and detection power. Hence, we first calculated the log fold change (LFC) in transcript, protein and phospho-protein abundance between KRAS-Dep and KRAS-Ind cell lines. We then correlated LFC mRNA abundance with LFC protein abundance as well as LFC protein abundance with LFC phospho-protein abundance. We found generally low to intermediate correlations, which is consistent with previous studies describing intermediate correlations between mRNA and protein abundance [111-113] (Figure A.2A, B) Correlation between LFC transcript and LFC protein 95 confidence interval (CI) = 0.29 - 0.36, p-value <= 2×10^{-16}; correlation between LFC unmodified protein and LFC phospho-protein 95 CI=0.29 - 0.43, p-value <=2×10^{-16}).

A naïve method of integrating those diverse sets of data is either to look for genes that are differentially abundant at the transcript, protein and phospho-protein level or to look for genes differentially abundant in at least one of these datasets. In this study, naïve integration called 675 differentially abundant

transcripts, 173 differentially abundant proteins in the unmodified state and 61 differentially abundant proteins in the phosphorylated state (Figure 2.1B and Supplementary Data 5 provided in Balbin *et al.,* 2013). However, naïve integration commonly produces a limited number of proteins that are differentially abundant across all signatures. Out of the 862 unique proteins called as differentially abundant, only 2 proteins are shared across all signatures and 45 by two independent datasets (Hochberg-adjusted p-value <=0.05, Figure 2.1B) resulting in only a ~ 5.2 % overlap among signatures. Furthermore, naïve integration typically produces a final list of differentially abundant proteins that is dominated by proteins identified only in the largest dataset, the transcriptome in this case (Figure A.2C). Moreover, this list is enriched in genes that appear not to be causative cancer genes but which have a high dynamic range of expression (Figure A.2D, A2E, A2F).

In order to address these issues, we developed a bioinformatics methodology to integrate transcriptomics, proteomics and phosphoproteomics datasets that aims at identifying differentially abundant proteins that are nominated as such by any combination of these datasets. This methodology focuses on identifying proteins that change consistently across transcript, protein and phospho-protein levels as they constitute candidates that can be uniformly assessed, and therefore potentially used for interrogating tissue samples at either the protein, phospho-protein or transcript level with similar results.

We first distinguish between “informative” and “all other” genes and assign weights to each dataset in proportion to that dataset’s size (Figure 2.1A, and Methods) in order to control for differences in the dynamic range of different proteins and the coverage of each “omics” dataset. We then calculate the combined “abundance score”, S, to measure the overall differential abundance of a protein across all datasets as $S = \frac{\sum_i^k w_k z_i}{\sqrt{\sum w_k^2}}$, where z is the z-transformed LFC of protein i in the dataset k, while w corresponds to the weight of each dataset

$w_k = 1/\sqrt{N_k}$. N_k represents the size of dataset k. Our score is inspired by the Stouffer's score that is used for meta-analysis[114]. Variations of the Stouffer score have been previously used to aggregate multiple studies involving only one type of "omics" datasets, such as microarrays[115].

Moreover, although other integration methods such as the combined Fisher p-value or the scores proposed by Ramasay *et al*., and Huang *et al*.,[75,115] could be used for nominating differentially abundant proteins, when compared to those methods the S score demonstrates several key advantages for discriminating informative genes. First, because the S score normalizes the original data into z-scores, the combined distribution is also normal, allowing for simple statistics (Figure A.3A). Second, the weight for each dataset is flexibly defined, i.e. according to the size of the dataset. Third, the S score can identify consistently changing proteins that would be missed otherwise (Figure A.3B). Fourth, because the S score is based on the average of z_i and the fisher method on the average of -log(p-value), these scores follow a close linear relationship for most values of S. Deviations of this linear relationship are observed for extreme values of S and instances where the transcript, protein and phospho-protein abundances change in discordant directions (Figure A.3C). Therefore, the combined used of the Fisher and S scores could identify proteins with discordant changes in abundance. In summary, by using the S score we defined a metric for selecting transcripts, proteins and phospho-proteins that are differentially abundant uniquely or consistently across different datasets, overcoming the drawbacks of naïve integration.

Our S-score analysis of the phosphoproteome, proteome and transcriptome nominated 115 differentially abundant proteins at a Hochberg-adjusted p-value <= 0.05. Out of the 115 proteins, 30 were nominated uniquely by our method and were missed using naïve integration of the datasets (Figure 2.1C). The S score also helps with prioritizing, as 20 proteins in phosphorylated

state, 28 proteins in un-phosphorylated state and 6 transcripts that were differentially expressed would have been unattended by a naïve approach (Figure 2.1C). By using the S score, the percentage of overlap among datasets in the list of differentially expressed proteins is ~26 %, which represents an increase of five-fold with respect to the naïve integration approach. Moreover, genes identified by our method show higher correlation between the LFC abundance of the transcript and protein in unmodified state as well as the protein in unmodified and phosphorylated state (Figure A.2A, Figure A.2B). We also note that the list of differentially expressed genes nominated by the S score is enriched for proteins with functions such as kinase, phospho-transferase activity and alternative splicing, and localized both in the cytoplasm and nucleus (Figure A.2G). These functions are expected for proteins in signaling cascades, such as the ones downstream of KRAS, but these functions were completely missed on the proteins nominated by the naïve integration approach.

Finally, comparison of NSCLC KRAS-Dep cell lines against KRAS-Ind cell lines showed that of 115 proteins nominated by our integrative analysis, 68 also demonstrated increased mRNA, unmodified protein or phosphorylated protein abundance in KRAS-Dep cells, whereas 47 were found to be decreased (Figure 2.1D, Supplementary Data 6 provided in Balbin *et al.,* 2013). Of the 68 that were increased, 57 proteins are classified as phospho-proteins, 14 as kinases, 8 as proto-oncogenes and 9 as involved in lymphocyte activation among other functions. Similarly, out of the 47 genes that were decreased, 37 are classified as phospho-proteins, 8 as kinases and 5 as proto-oncogenes among other functions. These results demonstrate that our analysis is able to identify functionally relevant proteins by integrating the transcriptome, proteome and phosphoproteome datasets.

2.5.2 Validation in NSCLC cell lines

To confirm our computational predictions, we employed a panel of 13 NSCLC cell lines for experimental studies, for which profiles of somatic mutations is provided in Table A-3. Of these, 8 have been defined as KRAS-Ind and 5 have been defined as KRAS-Dep based on previous studies [10,102] and confirmed in our hands. We selected highly ranking proteins predicted to be up-regulated in KRAS-Dep but not KRAS-Ind cells for further experimental validation. Of the top 20 nominated proteins, we included several proteins known to be associated with KRAS dependency in colorectal cancers (CTNNB1, PAK1) [102,116] and others that have not been implicated to date (LCK and cMET) with the KRAS-dependent phenotype in any cancer (Figure 2.2). Western blot analyses of these proteins and their phosphorylated forms validated that cMET, LCK, PAK1, and □-catenin were enriched in expression in KRAS-Dep cell lines. Furthermore, phosphorylated forms of these proteins were also specific, suggesting that these proteins are activated in KRAS-Dep cells. These experiments validate our computational method and suggest that the S score accurately identifies proteins that are highly activated in KRAS-Dep cell lines.

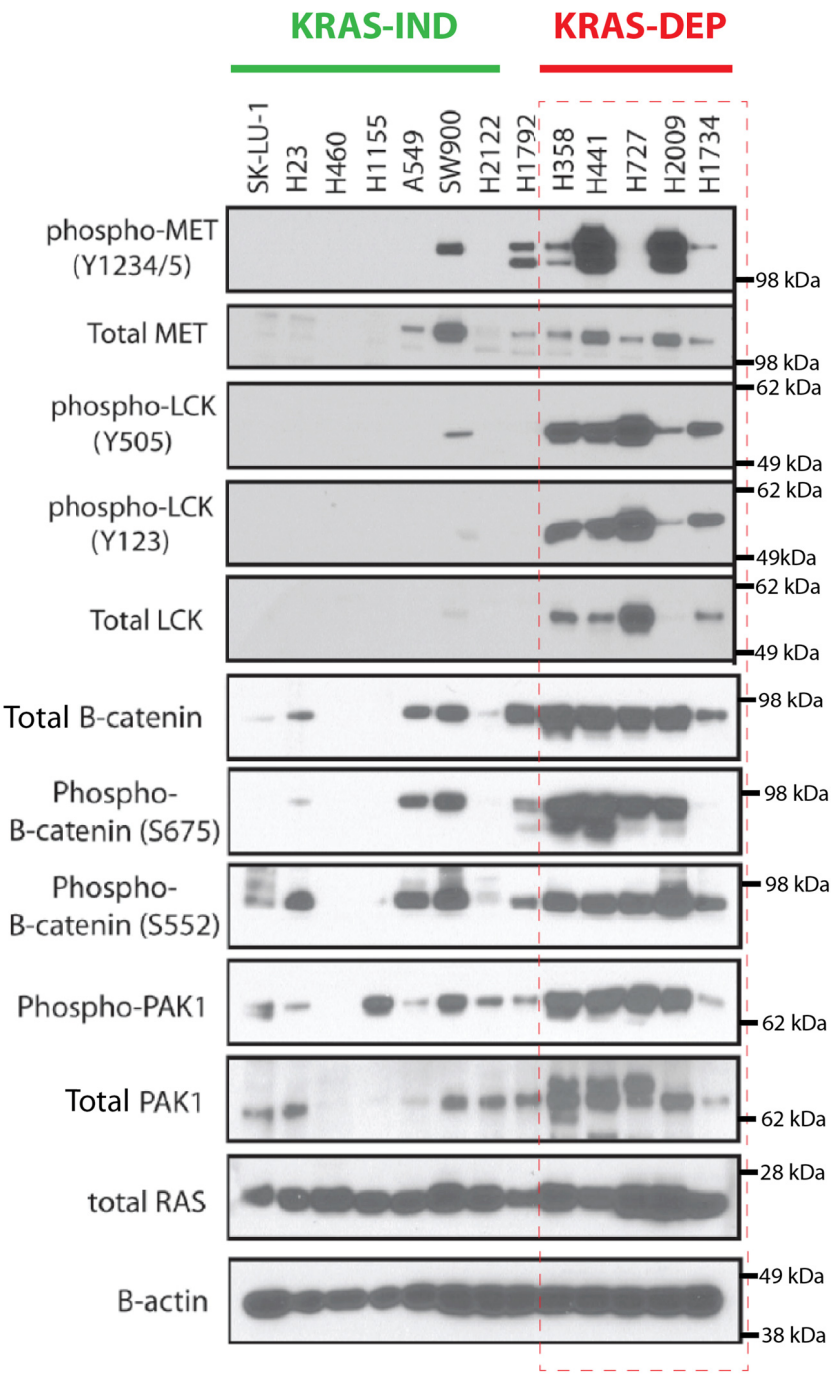


Figure 2.2. Activation of proteins nominated by the S score was confirmed by an orthogonal and low throughput method.

The western blot demonstrates high levels of total and phosphorylated protein for *MET*, *LCK*, *PAK1*, and β-catenin in KRAS-Dep. This pattern confirms the utility of our integrative analysis in nominating differentially activate proteins. It also suggests signaling modules that are differentially active in KRAS-dependent but not in KRAS-independent cell lines. Total RAS, and β-actin were used as controls.

2.5.3 Network analysis identifies active modules in KRAS-Dep cells

We next developed a three-step methodology for reconstruction of biological modules associated with KRAS status (Figure 2.3A). In the first step, we identified differential expressed pathways using the Signaling Pathway Impact Analysis algorithm (SPIA [109]). We then build a focused undirected and weighted protein-to-protein interaction network (G). Finally, in the third step, we used the Prize Collecting Steiner Tree algorithm to find sub-networks, T, in the weighted protein-protein interaction network (G) that maximized the number of differential expressed proteins recovered as well as the confidence in their interaction (Methods).

Specifically, in the first step we performed pathway enrichment analysis using SPIA in order to identify pathways with overall increased or decreased activity in KRAS-Dep cell lines (Figure A.4A). SPIA calculates the significance of a pathway according to both a gene set over-representation index and a network's perturbation index that takes into consideration the topology of and interactions within the pathway (Methods). This analysis revealed activation of main signaling programs in KRAS-Dep NSCLC cell lines when compared to KRAS-Ind, such as the ERBB signaling pathway, cancer specific associated pathways and tight junctions/cell adhesion pathways (Figure A.4B). Interestingly, immune-related signaling modules such as the T cell receptor, natural killer cell mediated cytotoxicity and Fc epsilon RI pathways were present, which suggested a relationship to LCK as immune-predominant kinase aberrantly up-regulated in KRAS-Dep cells. Moreover, although cancer associated-pathways are expected to appear enriched in our analysis of cancer cell lines, it is remarkable that the cancer pathways enriched in KRAS-Dep cell lines correspond to cancers types driven by activating Ras oncogene mutations (Figure A.4C), suggesting that certain molecular features are common to KRAS dependency across different cancers types.

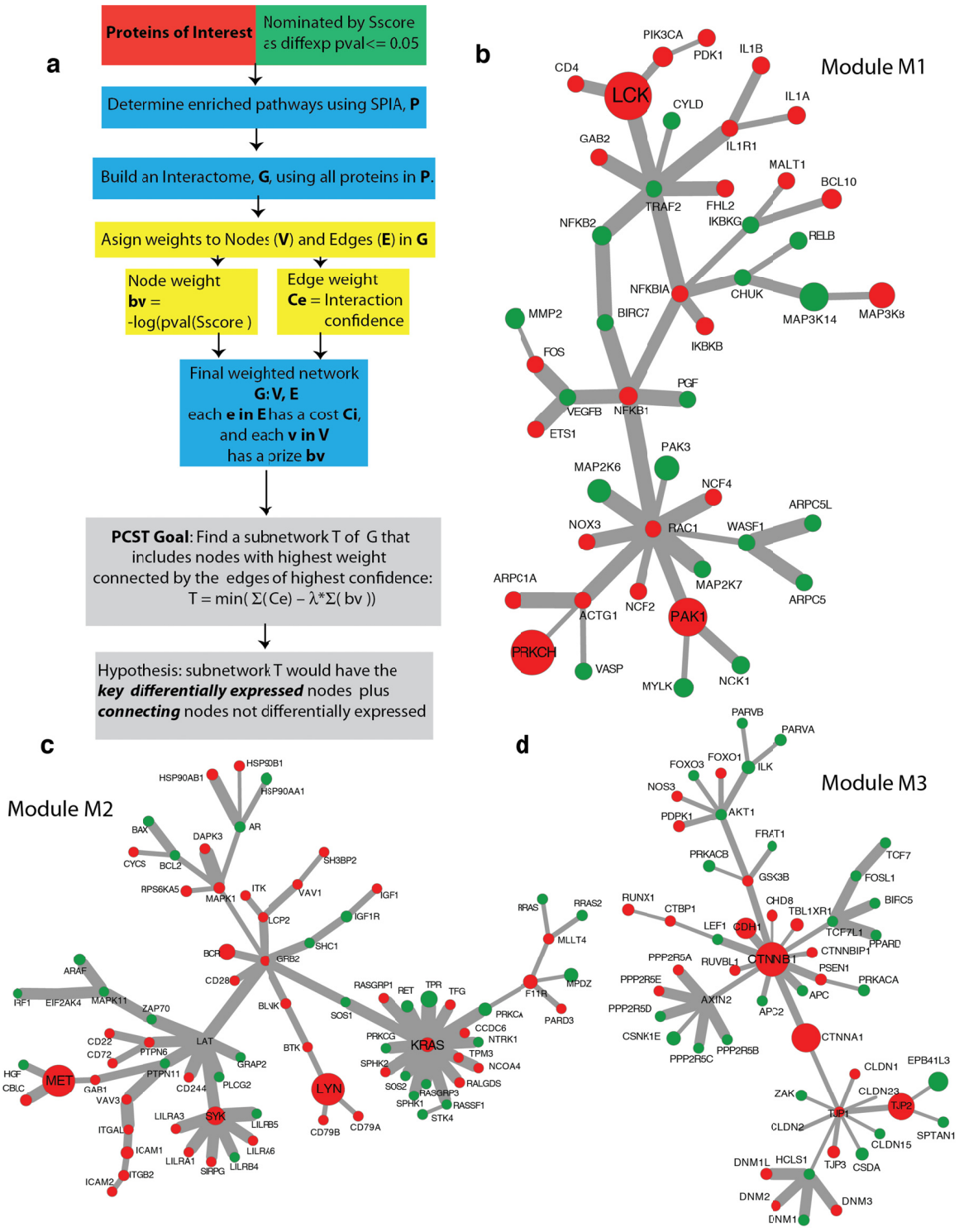


Figure 2.3. PCST-based network reconstruction method identifies active sub-modules in KRAS dependent cell lines.

A) Network reconstruction methodology. We built a focused undirected and weighted protein-to-protein interaction network (G) using differential expressed pathways identified by the SPIA algorithm [109]. We assign weights to both nodes (V) and edges (E). Node weights (bv) correspond to –log(pvalue(Sscore)) for

differential abundance between KRAS-Dep and KRAS-Ind phenotypes, while the edge's (Ce) weight correspond to the experimental confidence of that interaction as reported for the STRING database. Finally, we used the Prize Collecting Steiner Tree algorithm to find sub-networks, T, in G that maximized the number of differential expressed proteins recovered as well as the confidence in their interaction. **B)** Module M1. This module, identified by the PCST, connects *LCK* and *PAK1* in KRAS-dependent cell lines. The module joins *LCK* and *PAK1* with other proteins that belong to the NF-Kappa B and apoptosis pathways such as *NFKBIA*, NFKBs, TRAFs, and BIRCs. Node size is proportional to the absolute value of the combined S score. Node color represents over-expressed (red) or under-expressed (green) in KRAS-Dep cells. Edge thickness corresponds to edge's confidence as calculated from STRING database (methods). **C)** Module M2. This module, identified by the PCST, involves *KRAS* and *MET* in KRAS-dependent cell lines. Additional targetable proteins such as *SYK* and *LYN* are also part in this module. Described as in b. **D)** Module M3. This module, identified by the PCST, connects *CTNNB1* (β-catenin), *CTNNA1*, *CDH1*, *TJP2* and other proteins associated cell adhesion complexes and the tight junction pathways. Described as in b.

Furthermore, in the second step we built a focused undirected and weighted protein-to-protein interaction network (G) using all proteins that belong to those pathways identified by SPIA and we assigned weights to both nodes (V) and edges (E). The weight of each Node (Bi) corresponds to the combined score (S) for differential abundance between KRAS-Dep and KRAS-Ind phenotypes, while the weight of each edge (Ce) corresponds to the experimental confidence on that interaction. The edge weight is derived from the STRING database [117], by combining STRING's experimental and physical interaction scores using a naïve Bayesian approach.

Finally in the third broad step of this methodology, in order to identify specific network sub-modules that are active in KRAS-Dep cell lines, we formulated this network reconstruction task as a Prize Collecting Steiner Tree (PCST) problem[75,84,96,110] (Methods). The PCST allowed us to synthesize transcriptome, proteome and phosphoproteome signatures in the context of the weighted protein-to-protein interaction network mentioned above. This formulation facilitated the identification of crosstalk between pathways nominated by SPIA, as well as identification of relevant proteins that were not directly measured in our experiments. We identified three modules –referred to as M1, M2, M3 – using the PCST formulation.

M1 contains LCK, PAK1, and PRKCH as well as proteins involved in regulation of inflammation, antiviral responses and apoptosis proteins such as several TRAFs, BIRCs and NFKBs (Figure 2.3B). M2 contains KRAS as well as the kinases MET, LYN, SYK, and MAPK1 among others (Figure 2.3C). M3 contains CTNNB1 (β-catenin), CDH1, CTNNA1 (α-catenin), TJP2 and other proteins associated with the adhesion complex (Figure 2.3D). M3 is consistent with our observation that □-catenin is mainly localized in the cellular membrane of KRAS-Dep cells (Figure A.4D), supporting a role in cellular adhesion in NSCLC cell lines.

2.5.4 KRAS-LCK-PAK1 signaling axis in KRAS-Dep lung cancer

Intriguingly, module M1 suggests a link between LCK and PAK1 that has not been reported previously in solid tumors despite the fact that PAK1 overexpression has been already implicated in lung and breast cancers[118]. LCK is a tissue-specific kinase normally expressed in T-lymphocytes. It is commonly overexpressed in myeloid and lymphocytic leukemia, as well as Burkitt and non-Hodgkin's B-cell lymphoma [119] and acts as a proto-oncogene, inducing cellular transformation through regulation of cell proliferation and survival[119,120]. A role for LCK is not known in solid tumors. Therefore, we hypothesized that the aberrant overexpression of LCK in KRAS-Dep lung cancers could also play a role in this disease.

To confirm our network reconstruction approach and further dissect the functional connections among KRAS, MET and LCK, we performed knockdown experiments using independent siRNAs in the H441 and H358 cell lines that display KRAS dependency[10]. Immunoblot analysis showed that knockdown of KRAS decreased the abundance of MET, phospho-MET, LCK, phospho-LCK, phospho-PAK1/2 and phospho-BAD (Figure 2.4A, Figure A.5A, A.5B). These results demonstrate that MET, LCK, PAK1/2 and BAD are downstream of

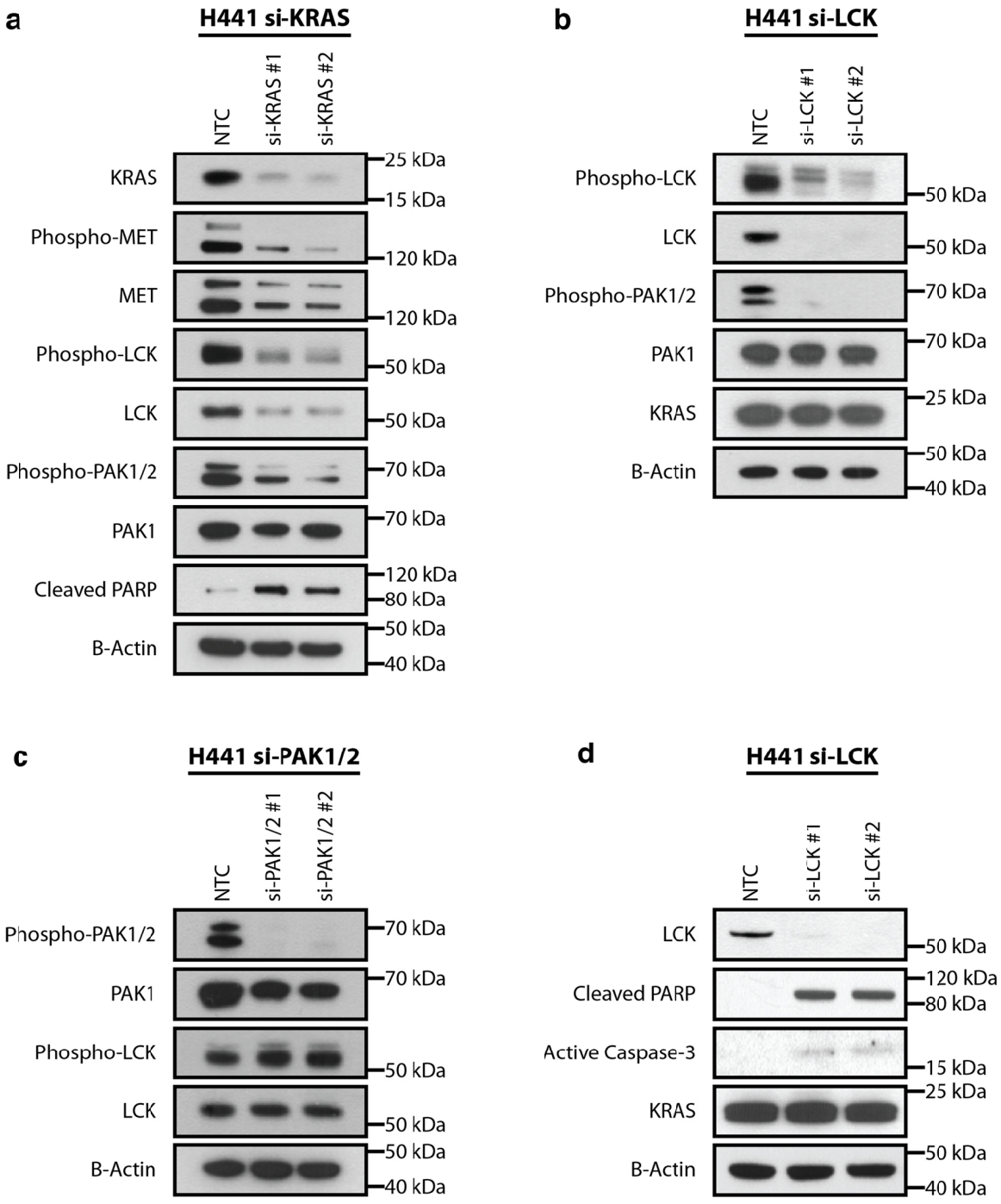


Figure 2.4 Experimental validation of protein modules in KRAS-Dep cells.

A) *KRAS* influences total and phosphorylated protein level of potential druggable kinases *LCK* and *MET* in KRAS-Dep cell lines. Knock down of *KRAS* with two independent siRNAs reduces phosphorylation levels of *LCK*, *MET*, *PAK1/2* in H441 cell line. KRAS-KD also reduced total protein levels of *LCK* and *MET*, but not *PAK1/2*. **B)** *LCK* influences *PAK1/PAK2* activation in KRAS-dependent cell lines. Knockdown of *LCK* using two independent siRNAs reduces phosphorylation levels of PAK1/2 but not their protein level in H441 KRAS-Dep cell line. **C)** *PAK1/2* are downstream of *LKC* in KRAS-dependent cell lines. *PAK1/2* knockdown

does not affect phosphorylation or protein level of *LCK* in H441-Dep cell line. **D)** *LCK* knockdown increases the level of cleaved *PARP* and caspase-3, markers of apoptosis in H441 KRAS-Dep cell line.

KRAS and regulated by KRAS *in vitro*. In contrast, knockdown of LCK did not reduce KRAS levels indicating that LCK does not regulate KRAS protein abundance (Figure 2.4B, Figure A.5C), although previous reports have suggested a role for LCK in KRAS activation[121]. Knockdown of LCK did however reduce phospho-PAK1/2 levels, but not total PAK1/2 protein, defining PAK1/2 as targets for LCK-mediated phosphorylation (Figure 2.4B, Figure A.5C). Figure 2.3B indicates that this effect is potentially mediated through a small network of interacting proteins. Moreover, knockdown of PAK1/2 did not change the phosphorylation or protein levels of LCK, confirming that PAK1 and PAK2 are downstream of LCK (Figure 2.4C). Taken together, our bioinformatics and experimental results suggest an active KRAS-LCK-PAK1/2 network in KRAS-Dep cell lines (Figure A.5D). Our results also present evidence that KRAS can influence both the phosphorylation and protein levels of LCK and MET kinases, which complements previous reports suggesting that those kinases could be upstream of the RAS-MEK pathways[121,122], and suggests the possibility of a feedback loop among these proteins in KRAS-dependent cells (Figure A.5D).

2.5.5 KRAS-Dep cells are also dependent on LCK for proliferation

In order to extend our results and investigate potentially aberrant expression of LCK in other cell lines, we performed a gene outlier expression analysis on an extended panel of 122 lung cancer cell lines (11 KRAS-Dep, 18 KRAS-Ind and 93 KRAS-WT) (Methods). We evaluated informative genes observed as outliers in KRAS-Dep but not in KRAS-Ind cell lines (Figure 2.5A).

This analysis revealed LCK, MET, ERBB3, MST1R and LYN are kinases that frequently exhibit outlier expression in KRAS-Dep cell lines, with expression levels in the top 80 percentile in over 60% of cell lines in this group (Figure 2.5B).

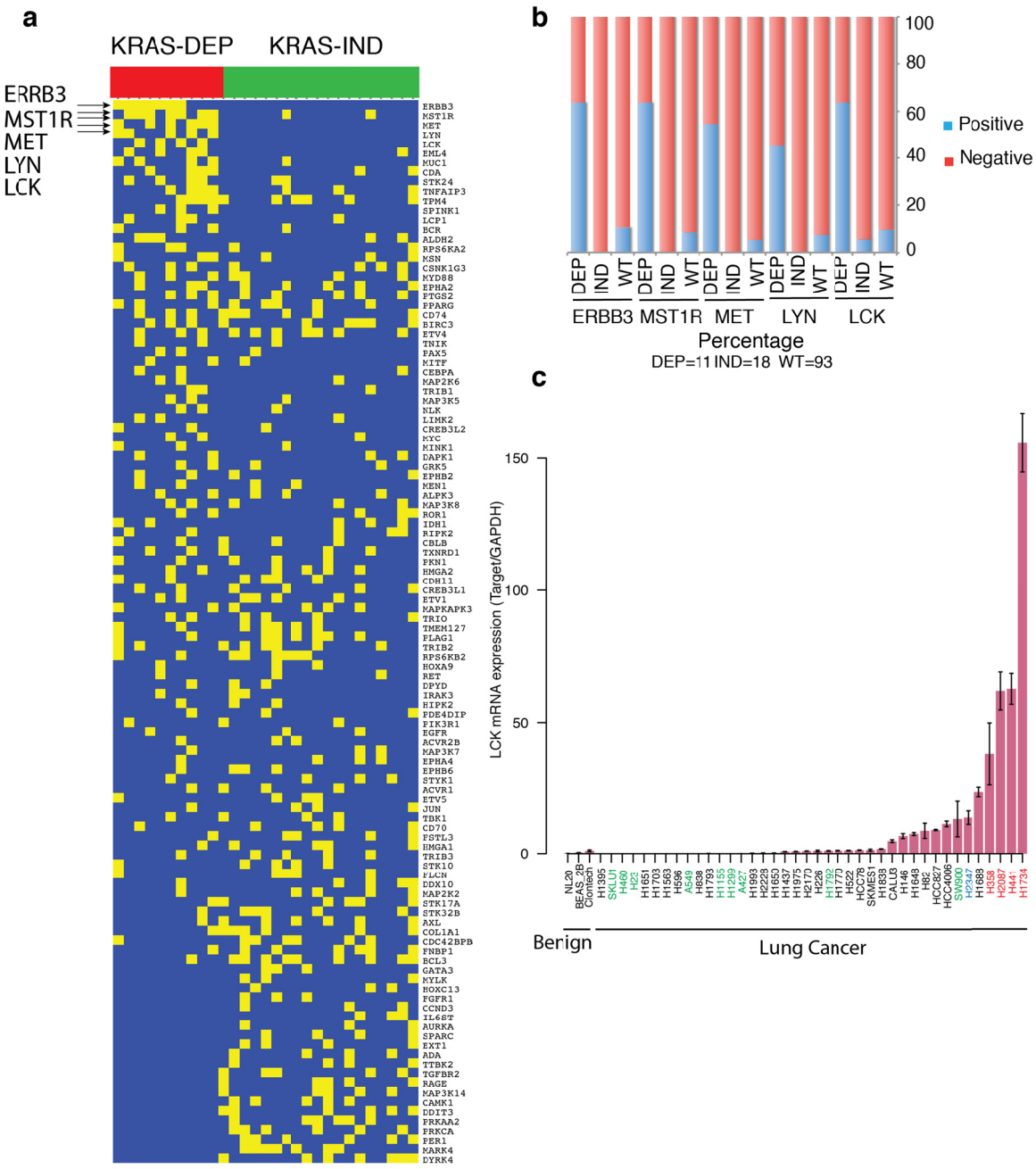


Figure 2.5. Outlying kinases in KRAS-Dep cell lines.

A) Cancer outlier profile analysis (COPA) of "informative" genes on an extended gene expression dataset of KRAS-mutated cell lines (KRAS-Dep=10, KRAS-Ind=11) confirms *LCK*, *MET*, *LYN* and *ERBB3* as differentially abundant proteins in KRAS-dependent but not in KRAS-independent cell lines. 11 KRAS-Dep and 18 KRAS-Ind were analyzed. **B)** Overexpressed *LCK* is present in at least 60% but in less than 10% of the either wild type or KRAS-Ind cell lines. *MET*, *ERRB3*, *MST1R* and *LYN* show a similar pattern. 11 KRAS-Dep, 18 KRAS-Ind and 93 KRAS-WT cell lines were analyzed. **C)** *LCK* expression measured by RTQ-PCR in a panel of KRAS-Dep (red label), KRAS-Ind (green label) and KRAS-WT (black label) cell lines

confirms high levels of *LCK* in *KRAS* dependent cell lines and none or negligible expression in KRAS-Ind or WT cell lines. Cell line H2347 (blue label) harbors *NRAS* Q61K mutation, but its dependency status could not be established. Bar height corresponds to the average over three independent replicates and error bars are defined as s.e.m.

By contrast, the kinases DIRK4 and MARK4 showed outlier expression in KRAS-Ind cell lines (Figure 2.5A). To validate our approach, we experimentally confirmed that LCK is overexpressed in KRAS-Dep cells using quantitative PCR on a panel of 43 lung cell lines (Figure 2.5C).

Given that LCK is a known lineage-specific proliferation factor in B-lymphocytes, we hypothesized that KRAS-Dep NSCLC overexpressing LCK also require this kinase for cell growth and survival. We performed shRNA knockdown experiments for LCK and determined whether ablation of LCK activity with independent shRNAs could selectively impair cell proliferation on KRAS-Dep cells (Methods). Figure 2.6A shows that knockdown of LCK dramatically impairs cell proliferation in KRAS-Dep cells but not KRAS-Ind cells, validating our predictions (shRNA1 t-test p-value=0.0001822, shRNA3 t-test p-value = 4.14 exp -6). We further confirmed that independent knockdown of KRAS also produced similar results (Figure A.6A).

Moreover, as a kinase, LCK is also an attractive candidate for strategies of targeted therapy. While specific LCK inhibitors are still in development, we tested whether prototype small molecule inhibitors of LCK would selectively affect the viability of NSCLC KRAS-Dep cells. We treated a panel of 3 KRAS-Dep cell lines and 2 KRAS-Ind cell lines with increasing doses of LCK inhibitor (CAS 213743-31-8) and measured cell viability at different drug concentrations. All three KRAS-Dep cell lines tested in this experiment were sensitive to LCK inhibition while the KRAS-Ind cell lines were insensitive to LCK inhibition, as expected from our hypothesis (Figure 2.6B). We further confirmed these results

using a second LCK inhibitor (CAS 918870-43-6) that showed similar results (Figure A.6B).

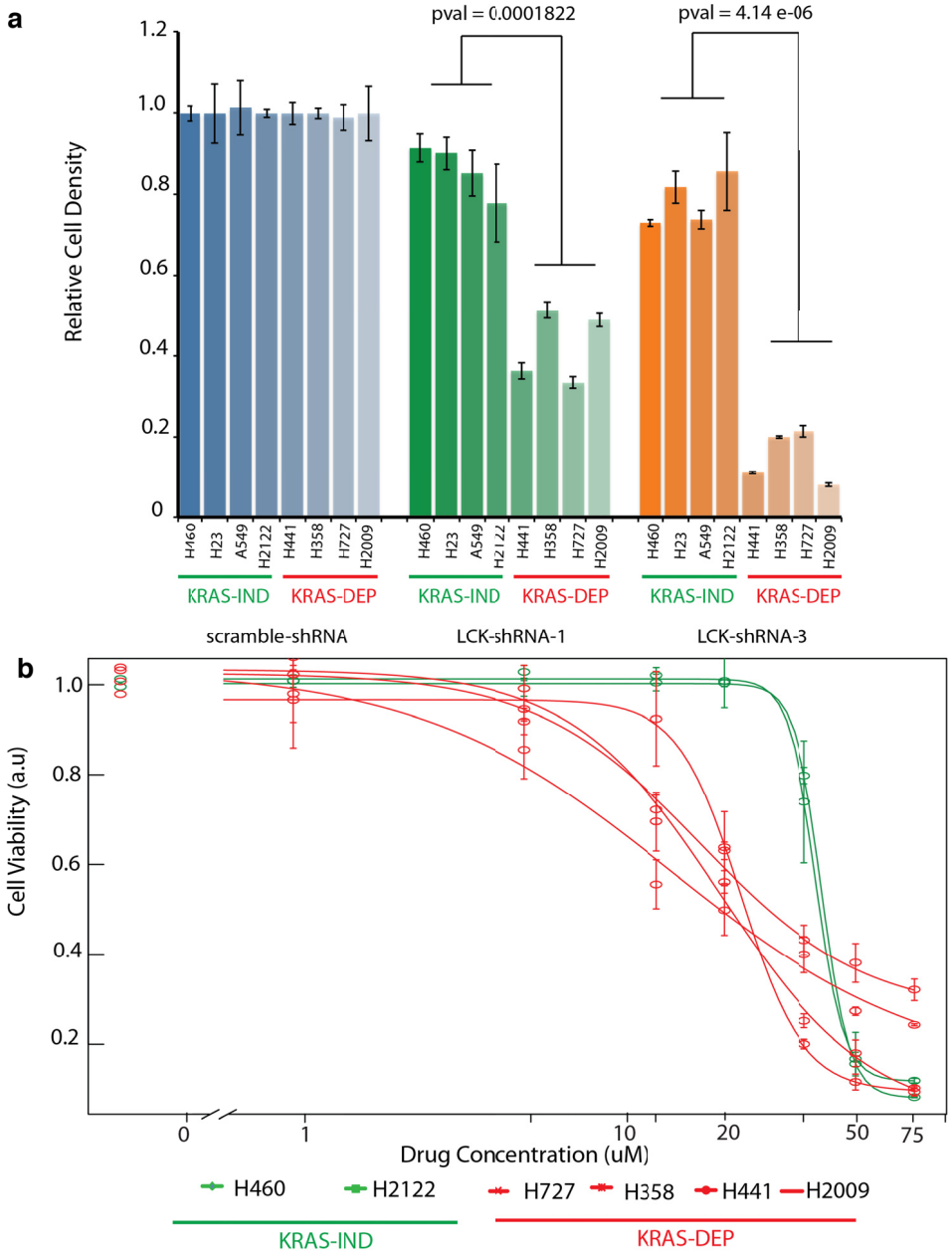


Figure 2.6. *LCK* constitutes a potential novel drug target in NSCL KRAS-Dep cell lines.

A) *LCK* knockdown impairs cell proliferation specifically in KRAS-dependent cell lines. *LCK* knockdown with two independent shRNAs showed statistically significant reduction in cell proliferation in KRAS-Dep but not KRAS-Ind cell lines (*LCK* shRNA-1 t-test p-value = 0.0001822 and *LCK* shRNA-3 t-test p-value 4.14exp-06). Bars correspond to the average of three independent experiments and error bars are defined as s.e.m. **B)** Inhibition of *LCK* using small molecule inhibitor preferentially impaired cell proliferation in KRAD-Dep but not

in KRAS-Ind cell lines. Points represent the average over four independent experiments and error bars are defined as s.e.m.

These results demonstrate that KRAS-Dep lung cancer cell lines have aberrant overexpression and activity of LCK. Similarly, we observed that MET shRNA knockdown as well as MET inhibition with small molecule inhibitors selectively impaired cell growth of KRAS-Dep cell lines (Figure A.6C, Figure A.6D), further supporting the biological relevance of our computational network reconstructions and predictions of targetable proteins in KRAS-Dep cells.

To evaluate whether LCK expression can be used to stratify the KRAS dependency status of human lung cancers, we assessed LCK expression in a panel of 29 lung adenocarcinoma tissue samples with mutations in KRAS. To confirm the KRAS mutations, we genotyped canonical positions in codons 12, 13 and 61, known to produce a constitutively active KRAS when mutated (Table A-4). As there is currently no clinical biomarker to identify the KRAS dependency status of NSCLCs, we sought to evaluate LCK expression in these samples as a potential biomarker for KRAS dependency. Because LCK is normally highly expressed in lymphocytes, LCK mRNA expression from surgical samples is not an accurate method to assess LCK expression in epithelial-derived lung cancer cells, as the infiltrating lymphocytes in these samples would distort the analysis. Thus, a previous study that detected LCK in lung cancer tissues by gene expression microarrays is likely confounded by the lack of cell-type specificity [123].

We therefore used immunohistochemistry (IHC) to determine the abundance of phosphorylated LCK in epithelial lung cancer cells in our 29 clinical samples. We first validated our IHC assay using a panel of normal tissues and cell lines that demonstrated high levels of LCK expression in the spleen where lymphocytes are abundant, but not in other tissue types. Next, a TMA of KRAS-Dep cell lines H441 and H358 also showed high levels of phosphorylated LCK

expression, while a TMA of H460 and H23 KRAS-Ind cell lines did not showed any staining. Finally, applying this method to our 29 lung tumor samples harboring KRAS mutation, we found that 58.6% (17/29) of tumors showed high levels of phosphorylated LCK staining, whereas 41.4% (12/29) tumors showed low levels of phosphorylated LCK (Supplementary Table S1). These results are consistent with *in vitro* data demonstrating that KRAS-mutant lung cancer tissues can be subdivided in two groups according to their levels of phosphorylated LCK, similar to NSCLC cell lines. Although, it is not possible currently to determine the dependency status of a tissue through direct experimentation, this subdivision of tumor samples is suggestive of the correlation described here between KRAS dependency and LCK activation in cell lines. However, a larger cohort of tissues with matched profiles of KRAS mutation, gene expression as well as immunohistochemistry of phosphorylated LCK would be required to further determine the prognostic value and the extent of this association between KRAS dependency and LCK activation in tissue specimens. A proof of principle analysis in this direction is shown in Figure A.6E.

2.5.6 KRAS and LCK could regulate anti-apoptosis pathways

To explore potential functional roles of the KRAS-LCK-PAK1/2 pathway, we evaluated our computational predictions of modules M1, M2, and M3 in lung cancer. We were struck by the enrichment for apoptosis-related proteins in module M1 that included LCK and PAK1 (Figure A.7A), suggesting a potential connection between LCK and apoptosis. Indeed knockdown of LCK in H441 cells was correlated with increased levels of cleaved PARP and caspase-3, markers of apoptosis, which further supports the association between LCK and apoptosis (Figure 2.4D).

To further explore this association, we used microarrays to profile gene expression changes following knockdown of LCK in the H441 and H358 KRAS-

Dep cell lines, and we evaluated the microarray data for pathways specifically inhibited or activated by LCK (Table A-5 and methods for specific details on this analysis of these microarray data). We assumed that pathways activated specifically by LCK in the context of KRAS dependency would be inhibited after knockdown of this kinase. Interestingly, we observed a module comprised of TRAF1, BIRC3 and BCL2L1, three proteins that regulate apoptosis (Figure A.7B). These proteins were part of a canonical KEGG pathway for lung small cell cancer, a pathway specifically inhibited after LCK knockdown (Table A-5).

Moreover, we reasoned that causative genes should be both overexpressed in KRAS-Dep compared to KRAS-Ind cell lines and also down regulated upon LCK knockdown in H441 and H358 (Methods). Performing this analysis yielded BCL2A1, a BCL2-related protein A1 (Figure A.8A, A.8B). BCL2A1 can bind to and inhibit or neutralize pro-apoptotic multi-domain proteins such as BAK and BAX as well as pro-apoptotic BH3-only proteins such as tBID, BIM, PUMA, BIK, HRK and NOXA but not BAD [124]. Pro-apoptotic protein BAD is inhibited when phosphorylated [125,126]. Indeed, knockdown of KRAS in H441 decreased phosphorylation levels of BAD (p112, p136) (Figure 2.5A), which is consistent with increased levels of cleaved PARP observed in the knockdown samples (Figure 2.4A) and supports a role for KRAS in preventing apoptosis *via* BAD. The effect on BAD phosphorylation was observed downstream of KRAS but not downstream of LCK or PAK1/2. Knockdown of LCK or PAK1/2 did not decrease phosphorylation levels of BAD suggesting independent mechanisms.

Taken together, these computational and experimental data suggest a potential regulatory network in KRAS-Dep cells that both “directly” inhibits apoptosis by inducing phosphorylation of BAD and “indirectly” by modulating the apoptotic response through the LCK module.

2.6 Discussion

The advent of high-throughput technologies has greatly advanced the study of cancer biology. However to date, most studies employ only an individual technology and studies that do include multiple profiling technologies frequently analyze them separately without integrating across modalities. While these approaches are effective for identifying single events in cancer (i.e. a new point mutation or an overexpressed gene), they do not uncover integrated biological modules that coordinate higher-level biological processes (i.e. apoptosis, RNA splicing, etc).

Here we developed a novel method to integrate disparate profiling modalities to explore novel functional networks differentiating KRAS-dependent from KRAS-independent NSCLCs. We used transcriptome, proteome, and phosphoproteome profiling to comprehensively analyze gene expression at the RNA and/or protein level, as well as signaling proteins activated or inactivated by post-translational modification. Using this approach on 13 KRAS-mutant NSCLC cell lines known to be KRAS-Dep or KRAS-Ind, our integrative analysis nominated 115 proteins that were differentially abundant between these two groups (Hochberg-adjusted p-value <= 0.05). Specifically, our method identified a set of proteins with highly correlated changes between transcript and protein levels or unmodified protein and phosphorylated protein levels, and then enriched these results for specific functions associated with KRAS. Of these, we validated four proteins (LCK, MET, PAK1 and β-catenin) selected from the top 20 nominated genes. LCK, MET, and PAK1 have not previously been studied in the context of KRAS-dependent lung cancer.

Of particular interest to this study was LCK, a lymphocyte-specific kinase well studied in B-lymphocyte development [119,120] but uncharacterized in solid tumors. We define a KRAS-LCK-PAK1/2 pathway in KRAS-Dep lung cancers

that has not previously been described. We find that KRAS regulates LCK protein and phospho-protein levels, and LCK in turn regulates PAK1/2 phosphorylation but not total protein levels. Previous studies have identified a role for PAK1/2 in the phosphorylation of β-catenin in KRAS-mutated colorectal cancer [102,116]; however, we did not observe β-catenin as a direct target of the KRAS-LCK-PAK1/2 pathway in lung cancer. Knockdown of KRAS and LCK did not impact β-catenin phosphorylation or cellular localization. Indeed β-catenin localized to the cell membrane in our experiments (Figure A.4D), not the cell nucleus where □-catenin is known to be active in the stimulation of the Wnt signaling pathway [102,116]. In addition, our work finds that □-catenin associates with the M3 reconstructed network module that also contains cell surface adhesion proteins such as CDH1, CTNNA1 (α-catenin), and TJP2. Thus □-catenin in NSCLC cell lines may operate through cell adhesion pathways as opposed to a role in regulating transcription as reported in colorectal cancer [102]. This further helps to explain earlier observations that associate KRAS-Dep lung cancer cell lines with differentiated phenotypes [10].

To explore the function of LCK in lung cancer, we performed knockdown experiments and observed that depletion of LCK impaired cellular proliferation and phenocopied knockdown of KRAS in KRAS-Dep cell lines. In addition, small molecule inhibition of LCK resulted in preferential decrease in cell viability in KRAS-Dep cells. Using the Prize Collecting Steiner Tree formulation, we also found that LCK was associated with a reconstructed Module M1 containing several proteins involved in regulation of apoptosis in addition to PAK1. Indeed, we observed that knockdown of LCK or KRAS induces an increase in cleaved PARP levels indicating an increase in apoptosis. KRAS-Dep cells may then modulate apoptosis through two complementary mechanisms. KRAS may regulate the apoptotic response by regulating phosphorylation of BAD, while LCK may regulate BCL2-related anti-apoptotic proteins. Previous studies in T cells and CLL cells support this role of LCK as a guardian against apoptosis, as well

as LCK inhibition through small molecule inhibitors as an effective mean to sensitize those cells to apoptosis [119]. Finally, we evaluated LCK expression in KRAS-mutant NSCLC tumors. We observed that almost 60% (17/29) of the KRAS-mutated tumors showed high staining levels of phosphorylated LCK by IHC, suggesting they are likely KRAS-dependent. As projects such as The Cancer Genome Atlas (TCGA) approach their goal of enrolling thousands of patients with matched -omics datasets such as exome/genome and RNA sequencing and reverse phase protein arrays (among others), as well as detailed clinical follow ups, we will be able to assess the prognostic value of the LCK-KRAS-PAK1/2 pathway in the context of KRAS dependency. A proof of principle analysis in this direction is presented in Figure A.6E.

Taken together, this study establishes a potentially actionable pathway in KRAS-Dep NSCLCs comprised of KRAS, LCK, and PAK1/2. We find that KRAS induces LCK activation, leading to a signaling cascade specific to KRAS-Dep cells that promotes cell proliferation and could reinforce a positive feedback loop with KRAS activity (Figure A.5D). Furthermore, our study develops a method to integrate multiple proteomic and transcriptomic datasets for the identification of biologically relevant modules in cancer. We thus provide a framework for the complex analysis of multiple cancer datasets to make biologically-informed computational predictions for uncharacterized signaling pathways in cancer.

2.7 Contributions

Science is a collective enterprise and it is much more fun when done with friends and good collaborators. The results presented in this chapter were made possible because of the great collaboration and support of a team of people in the Chinnayian and Nesvizhskii labs.

Author contributions: O.Alejandro Balbin, Alexey Nesvizhskii, and Arul M. Chinnayian designed the study; O. Alejandro Balbin developed all bioinformatics methods and computational analysis, designed functional assays, and performed proliferation assays; John Prensner., Benjamin Chandler and Anirban Sahu performed knock down functional assays and western blots; Anirban Sahu performed the drug assays; Anastasia Yocum completed Mass Spectrometry; Damian Fermin helped with proteomics analyses; Rohit Malik performed β-catenin immunoflorescence assays; Sunita Shankar help with Preliminary drug assays; David Beer and Dafydd Thomas provide tumor microarrays and performed LCK staining scoring. I am deeply grateful to all of them.

2.8 Dedication

I want to dedicate this research to my father, Jesus William. Jesus William fought a very aggressive KRAS mutated colorectal cancer. While he was giving the fight for his life, I was, paradoxically, trying to find alternative ways to treat KRAS dependent cancers. Although, the results of this research were not on time to help saving his life, I hope they can contribute to save the life of some other fathers.

Chapter 3
Identifying driver fusions in lung cancers without known driver mutations

3.1 Background

3.1.1 Lung cancer

Lung cancer is the leading cause of cancer-related deaths worldwide, generating more than a million deaths each year[127,128]. Lung cancer is histologically classified as either non–small cell lung cancer (NSCLC) or small cell lung cancer (SCLC). NSCLC accounts for 80% of all lung cancers and includes lung adenocarcinoma (LUAD), squamous cell carcinoma (LUSC) and large cell carcinoma (LULC). Adenocarcinoma is the most prevalent subtype and most often observed in non-smokers, however tobacco smoking is associated with the majority of lung cancers[129]. The overall 5-year survival rate for lung cancer remains poor ~15%, due primarily to late diagnosis when tumor removal is no longer an option[128].

Genomic analyses of LUAD have revealed mutations in several well-characterized tumor suppressor and oncogenes including *TP53*, *STK11*, *KRAS*, *EGFR* and *BRAF* among others[2]. These tumors also demonstrate copy number alterations with most occurring at relatively low frequency with some having therapeutic implications such as *ERBB2* amplification[130]. Remarkably mutations in KRAS are mutually exclusive with mutations in EGFR. Recent analyses by

TCGA for lung squamous cell carcinomas indicate that these tumors undergo *TP53, FGFR1, DDR2, AKT1, PIK3CA, CDKN2A, MLL2, NOTCH1,* and *RB1* gene mutations as well as several recurrent gene copy number alterations in genes such as *FGFR1, SOX2 and TP63*[3]. The heterogeneity observed in lung cancer both histologically and molecularly, underlie the difficulties in effectively treating patients with this disease.

Similar to these known "driver" somatic gene mutations, several important gene fusions, formed by the breakage and re-joining of two different genes, occur in lung cancer including the *EML4-ALK* gene fusions identified in approximately 4% of adenocarcinomas[4,8]. This fusion protein links the N-terminal portion of echinoderm microtubule-associated protein-like 4 (EML4) with the intracellular signaling portion of the anaplastic lymphoma kinase (ALK) tyrosine kinase receptor. The *EML4-ALK* translocation is mutually exclusive with *EGFR* and *KRAS* mutations and tumors with *EML4-ALK* translocation have fewer *TP53* gene mutations[5]. *EML4-ALK* gene fusions occur in LUAD and in never- or light smokers. Additional gene fusion events have now been identified in LUAD including *ALK* [131], *ROS1*[6], as well as *RET*[7,8] kinases. These chromosomal rearrangements have been strongly associated with a history of never or light smoking.

Most importantly, patients with tumors containing *EGFR* mutations show at least initial responsiveness to drugs that target these alterations[20] and the presence of the *EML4-ALK* gene fusion is an indicator of therapeutic responsiveness to ALK inhibitors[4]. Under this logic, lung tumors with *KIF5B-ALK* fusions also have the potential for sensitivity to ALK inhibitors and RET fusions may be treated using drugs that target this kinase[132]. These chromosomal rearrangements have been strongly associated with a history of never or light smoking. Our group also identified the *NFE2-R3HDM2* and FGFR3 gene fusions present in a small percent of lung cancers[133,134].

3.1.2 Gene fusion detection and control of the false positive rate

The increase on RNASeq experiments to study the cancer transcriptome has propelled the development of numerous algorithms for fusions discovery[55-57,135,136]. A major task in fusion discovery is handling false positive. False positives can be generated during the sequencing and bioinformatics steps. In the sequencing step chimeric cDNA artifacts are generated by template switching during reverse transcription and amplification[137]. Template switching occurs when the nascent cDNA that is being synthesized dissociates from the template RNA and re-anneals to a different stretch of RNA with a similar sequence to the initial template, generating artifactual gene fusions[137]. This behavior is observed even more for highly abundant transcripts such as ribosomal RNA. In the bioinformatics step false positives are generated because all fusion discovery tools are error prone; they all identify fusion genes that are not present in control synthetic datasets[138]. Strikingly, the number of false positive fusions increased with read length for all tools, and all tools detected less fusion reads than were expected.

A recent comparison of eight fusion calling algorithms showed that overall algorithms have a maximum sensitivity of about 80%. However, in order to recover the higher number of true positives, the most sensitive algorithms pay an extremely high prize increasing by several orders of magnitude the number of false positives. For example, TopHat-fusions (THF) and ChimeraScan the most sensitive algorithms produced well above 13000 false positive fusions in order to detect 27 true fusions[139]. False positives can be generated by the software or during the sequencing step.

Taken together those results demonstrate that reducing the number of false positives would be the biggest challenge on a high throughput assessment of gene fusions across large cohort of patients, such as the TCGA cohorts and the one assembled for this study.

Previous studies have relied on applying hard thresholds to filter out potential false positives. For example establishing a minimum number of reads that support the gene fusion product. According to the identification of fusion boundary, the nucleotide coordinates defining the breakpoint of both genes involved in the fusion, we can define three types of reads: read spanning, read encompassing and mate pair encompassing reads (for pair-end sequencing). Encompassing reads harbor a fusion boundary and each read maps on a different gene of the fused gene couple; in spanning reads one mate overlaps with a fusion event, while the corresponding paired-end mate matches with one of the two genes involved in the chimera; lastly in mate pair encompassing reads one read maps to one side of the fusion boundary while its mate maps to the other side, but none of them harbor the fusion breakpoint.

Although effective for blindly eliminating most false chimeric events, two issues appear with applying hard thresholds using the minimum number of reads. First, there are real and functional fusions such as EML4-ALK, which have a low read support despite being driver fusions. On the other hand, highly expressed, such as ribosomal proteins, tend to form a great variety of chimeric fusions due to template switching during the sequencing step. In consequence, thresholding approaches would eliminate real and functional fusions while keeping clear false positives.

In this study, we address that challenge of controlling the false positive rate on fusion detection from a completely different perspective. Here, we developed a gene fusion classifier to distinguish between true and false positives. We used structural properties such as: 3’ and 5’ partner genes, cohort, 3’ breaking exon, 5’ breaking exons, median alignment quality of reads that support 3’ gene, median alignment quality of reads that support 5’ gene, number of spanning reads, encompassing reads, spanning mate pairs, expression of the 5’ and 3’ gene in the sample, among other characteristics. We employed a

random forest algorithm for classification, because random forest algorithms do not make assumptions about the distribution of the data. Moreover, random forest algorithms determine the importance of each feature in the classification process and therefore they allow for feature selection. This approach showed to be very efficient while at the same time it recovered the main structural properties that are frequently used for thresholding.

3.1.3 Aims of this study

Given the importance of gene fusions, yet their apparent low frequency in lung cancer, in the present study we examine the landscape of gene fusions in the largest RNASeq cohort of NSCLC assembled so far. We have performed comprehensive RNA sequencing of our cohort of primary NSCLC with a history of heavy smoking and integrated the results with the available data from TCGA and another public available dataset, the “Seoul” cohort.

We characterize gene fusions in lung tumors both with and without known driver mutated genes, and we examine the relationship between fusions incidence, tumor and clinical characteristics including patient survival. This is the first study to show that fusions incidence is an independent factor associated with poor prognosis. Moreover, we identified recurrent Neuregulin 1 (NRG1) gene fusion events exclusively in driver negative patients, resembling known kinase fusions, which may provide future therapeutic opportunities for patients harboring NRG1 rearrangements. This study also generates a database of lung cancer fusions that can be used by other researches looking for low recurrent fusions in this disease.

3.2 Bioinformatics methods

3.2.1 Sequence alignment

Sequence alignment was performed using the Tuxedo pipeline: Bowtie2 (Bowtie2/2.0.2) and Tophat2 (TopHat/2.0.6)[47]. We supplied TopHat with the set of transcript models annotated in the Homo sapiens ensemble database version 69. The flag fr-firststrand was used for the strand specific RNASeq libraries while fr-unstranded was used for the unstranded libraries. All other parameters were used with default values.

3.2.2 Fusion calling

Fusion calling was performed with TopHat-fusion[47] (THF) on the UMICH, TCGA and Seoul cohorts. ChimeraScan[56] was applied to the UMICH cohort to increase sensitivity in our discovery cohort. TopHat-fusion was run with the following arguments: bowtie, fusion-search, keep-fasta-order, no-coverage-search, fusion-min-dist=0, fusion-anchor-length=13, fusion-ignore-chromosomes=chrM. TopHat post-processing was run with the arguments: skip-blast, num-fusion-reads=1, num-fusion-pairs=1, num-fusion-both=3. Chimerascan was run with the following options: trim5=1, trim3=1, frag_size_percentile=1.0, arg= -v, keep-tmp.

3.2.3 Fusion annotation and lung cancers fusions database

A database of fusions in lung cancers was developed, and for each fusion structural and functional annotation was recorded. The structural information correspond to chromosomes of 3' and 5' partner genes, cohort, 3' chromosome, 5' chromosome, 3' breaking exon, 5' breaking exons, median alignment quality of reads that support 3' gene, median alignment quality of reads that support 5'

gene, number of spanning reads, spanning mate pairs and encompassing reads, 3' and 5' partner recurrence across the cohort and fusion type (Inter-chromosomal, Intra-chromosomal, Tandem-duplication).

The functional annotation corresponds to kinase status, oncogene status, tumor suppressor status and targetable status (TRUE/FALSE) of both 3' and 5' partner genes. Other functional annotations include the gene family of both fusion genes, as well as the their gene biotype (protein-coding, ncRNA, rRNA, etc.). Moreover, the gene expression of each fusion gene was calculated in fragments per-kilobase per million (FPKM) using Cufflinks[140] and stored in the database. In addition, an outlier score was calculated for the expression of both 5' and 3' partners in order to identify cases in which the 3' partner is highly expressed as consequence of the fusion event.

This database was created using pytables and hd5 format for fast access and storage and includes the following tables: patient table, patient clinical information table, fusions structural information table and expression table. In addition to these tables corresponding to fusion events, we create an additional table to store the mutation status for each patient, mutation table. The mutation table allows us to classify each patient as "driver positive" or "driver negative" according to mutation status of well-known cancer related genes (see below).

3.2.4 Fusions classifier

As described in section 3.1.2 all fusion calling algorithms produce a significant number of false positive fusions when applied on RNASeq data. Many of these spurious fusions are due to diverse and difficult to model bioinformatics, sequencing and biological factors such as: template switching and random chimeric events associated with amplicon regions among others.

Therefore, in order to separate potentially genuine fusions from spurious ones, we developed a classifier to predict potentially true fusions based on the structural and functional features collected for each fusion, which were described above and stored in our fusion's database.

THF called 31,304 fusions across the combined cohort. The task of separating false positive fusions from potentially true ones in this dataset is then far from trivial. We first reason that functional fusion proteins have open reading frames (ORFs); therefore fusions in which the exon of one gene is fused to the intron of another or two introns are fused together would not produce fusion products with ORFs. This first level filtering reduced to 6,465 the number of fusions to classify. Then, we reason that fusions found in normal samples; fusions involving pseudogenes, lincRNAS, or antisense transcripts and fusions for which the median alignment quality of reads supporting any of the gene partners was equal to zero (indicating multi-mapping) are potentially false positives, and there were excluded from downstream analysis. This second level filtering reduced to 4,990 the number of fusions called by THF. Assessing the quality of each one of those fusions manually is impossible in practice. Therefore, we build a random forest classifier to determine the potentially true and false positives out those 4,990 gene fusions.

For the classification step, we train a random forest classifier with 10,000 trees using the following features: chromosomes of 3 and 5' genes, 3' gene, 5' gene, 3' breaking exon, 5' breaking exons, median alignment quality of reads that support 3' gene, median alignment quality of reads that support 5' gene, number of spanning reads, spanning mate pairs and encompassing reads, 3' and 5' partner recurrence, fusion type, gene biotype of both 3' and 5' genes, FPKM expression of both 3' and 5' genes, and FPKM expression of both 3' and 5' genes normalized across the combined cohort.

True positives examples were selected from the TCGA, Seoul and UMICH cohorts. On one hand, the examples chosen from the TCGA and Seoul Cohorts correspond to well known fusions involving ALK, RET and ROS1 kinases. On the other, the examples chosen from the UMICH cohort correspond to fusions called by at least two independent algorithms, carefully curated manually and validated by PCR (Table B-1). False positive examples were selected representing different types of spurious fusions: e.g. overlapping genes, fusions involving highly expressed genes such as ribosomal proteins among others.

An additional advantage of using a classifier to determine the potential true fusions, as opposed to hard filters defined a priori, is that we can learn those features or rules from the data itself. In our dataset, the top five features that contributed the most for the random forest classifier were, in decreasing order of importance, fusion type (Inter-chromosomal, Intra-chromosomal, Tandem-duplication), sum of the median alignment quality of both gene partners, number of reads spanning and encompassing reads across the fusion junction and the cohort normalized expression value of the 3' gene (Figure 3.4).

Two additional sets of true fusions were left out of the training dataset to calculate the recovery rate. First, a set of 11 fusions called in the Seoul cohort[23] and validated by PCR by the same authors, and a second set of 15 fusions called in the UMICH cohort by THF, ChimeraScan, manually curated and validated by PCR. In the first of these datasets, our classifier recovered 10 out 11 fusions for a 90.1% recovery rate (Table 3.3-1). In the second set, the classifier recovered 14 out 15 validated fusions for a 93.3% recovery rate (Table 3.3-2).

Table 3.3-1. Fusions recovered by our classifier in the Seoul cohort.

All these fusions were called and validated independently of our study. Recovered fusions are label with 1, while missed fusions with 0.

Patient	5'Chr	3'Chr	3'Gene	5'Gene	Cohort	5'exon	3'exon	5'AlignQ	3'AlignQ	AlignQS	SpanReads	MatePairs	Encompasing	FusionType	Recovered
Seoul_lc_c25	3	6	ASCC3	UBE2E1	Seoul	1	5	9	9	18	16	2	9	InterC	1
Seoul_lc_s11	11	11	C11orf93	HYOU1	Seoul	13	3	9	9	18	155	7	49	IntraC	1
Seoul_lc_c25	1	6	CGA	ZFYVE9	Seoul	15	4	9	9	18	16	3	12	InterC	1
Seoul_lc_s18	19	19	DNM2	NMRK2	Seoul	8	2	3	3	6	3	1	1	IntraC	0
Seoul_lc_s18	11	11	FGF3	RBM14-RBM4	Seoul	1	2	9	9	18	20	1	8	IntraC	1
Seoul_lc_c17	12	12	GPR133	TXNRD1	Seoul	16	14	9	9	18	14	2	10	IntraC	1
Seoul_lc_s23	19	14	MBIP	AXL	Seoul	19	6	9	9	18	1566	153	1210	InterC	1
Seoul_lc_s38	12	4	PDGFRA	SCAF11	Seoul	15	2	9	9	18	132	13	110	InterC	1
Seoul_lc_s9	10	6	ROS1	CCDC6	Seoul	5	9	9	9	18	77	6	53	InterC	1
Seoul_lc_c36	12	12	SLC16A7	MUCL1	Seoul	2	3	9	9	18	82	7	16	IntraC	1
Seoul_lc_c15	11	13	TNFSF11	APLP2	Seoul	7	5	9	9	18	150	12	83	InterC	1

Table 3.3-2. Fusions recovered by our classifier in the UMICH cohort.

All these fusions were called and validated independently of our study. Recovered fusions are label with 1, while missed fusions with 0.

Patient	5'Chr	3'Chr	5'Gene	3'Gene	Cohort	5'exon	3'exon	5'AlignQ	3'AlignQ	AlignQS	SpanReads	MatePairs	Encompassing	FusionType	Recovered
A28	9	9	PTCH1	FAM120AOS	umich	23	2	7	7	14	8	2	6	IntraC	1
C028	1	1	WASF2	FGR	umich	9	12	9	9	18	21	48	2	TD	1
C074	8	8	GTF2E2	GSR	umich	5	1	9	9	18	87	33	0	TD	1
C004	3	3	HLTF	HPS3	umich	18	11	9	9	18	40	7	37	TD	1
A49	11	11	CPT1A	HRASLS2	umich	6	3	9	9	18	105	13	115	IntraC	1
C040	11	11	AHNAK	KAT5	umich	6	6	6	6	12	9	43	6	IntraC	1
H1838	6	6	PCMT1	LATS1	umich	3	6	9	9	18	15	6	7	TD	1
A52	9	12	PTPRD	LRMP	umich	43	5	9	9	18	12	2	8	InterC	1
C004	3	3	UBA5	MRAS	umich	10	2	9	9	18	12	2	13	IntraC	1
H1792	12	12	SRGAP1	MSRB3	umich	3	2	9	9	18	79	14	119	IntraC	1
H23	2	2	THADA	MTA3	umich	38	9	9	9	18	33	12	16	IntraC	1
H441	1	1	MEAF6	SCMH1	umich	5	9	4	7	11	7	27	4	IntraC	1
C051	15	15	MYO5C	TNFAIP8L3	umich	12	1	9	9	18	15	52	15	IntraC	1
A25	3	3	IP6K1	TRAIP	umich	4	5	9	9	18	63	46	58	TD	1
A85	8	8	RBM12B	MMP16	umich	3	6	2	2	4	2	4	3	IntraC	0

3.2.5 Mutation calling

UMICH cohort: Single nucleotide variants (SNVs) were called using Varscan2 (Varscan2/2.2.8)[48] on the ssRNAseq libraries of the UMICH cohort. Because, we did not have matched normal for each tumor sample, we consider only SNVs that were previously reported in the Catalogue of Somatic Mutations database (COSMIC version 56). Single nucleotide mutations in other positions were not considered for reporting or downstream analysis. SNVs present in dbSNP (v135) were filter out, as well as SNVs with variant fraction smaller than 10%, or with less than six reads covering the position. Insertions and deletions were not called from the RNAseq data, because currently there are not available algorithms to efficiently assess these genetic aberrations on RNASeq libraries. SNVs for all tumor samples were aggregated and annotated using variant-tools[141].

TCGA cohort: All somatic mutations both SNVs and indels called on Exome sequencing data for the TCGA consortium were extracted from aggregated Mutation Annotation Format (MAF) files available at the Broad institute firehose Genome Data Analysis Center MAF dashboard on May 11 of 2013.

Seoul cohort: All SNV and insertion/deletion somatic mutations reported by Seo et al (2012) were used[23].

3.2.6 Sample annotation

We annotated the mutation status of oncogenes and tumor suppressor well known to be involved in lung adenocarcinoma and squamous carcinomas. On one hand, known activating mutations were considered for *KRAS, NRAS,*

HRAS, EGFR, BRAF, PIK3CA, and *MET*, while missense or non-sense mutations were considered for *TP53, STK11, NF1, PTEN, SMARCA4, CDKN2A,* and *APC*. Mutations reported in COSMIC were considered for *AKT, MEK, ATM, AKT1, KEAP1, U2AF1, RBM10, ARID10,* and *MYC* which have been recently implicated on these indications[3,142]. Finally, we used the somatic mutation information to divide the combined cohort in two groups: samples with known drivers and samples of unknown drivers. The first group corresponds to samples with somatic mutations in *KRAS, NRAS, HRAS, EGFR, BRAF* and/or *PIK3CA*, while the second group to samples that do not harbor alterations in those well-known driver genes.

3.3 Experimental methods

3.3.1 Sample acquisition

We collected tumor samples from 67 patients with lung adenocarcinomas and 36 patients with lung squamous carcinoma. Matched normal lung tissues samples were collected at the edge of cut lung lobe, as far as possible and at least 3 cm far away from tumor, following surgery at the University of Michigan. The recruitment of subjects and informed consent were reviewed and approved by our IRB. These tissues were preserved by flash freezing immediately following surgical resection, and clinical and follow-up data have been collected. None of the patients used in this study received preoperative chemotherapy or radiation therapy. The 24 cell lines included in this study were all acquired from The American Type Culture Collection (ATCC) and grown according to the ATCC suggested media conditions.

3.3.2 Total RNA isolation

Regions of tumor tissue containing a minimum of 70% tumor cellularity defined by cryostat sectioning were utilized for RNA isolation. Tissues or cell lines RNAs were isolated using miRNeasy mini kit (Qiagen). RNA quality was analyzed using the 2100 Bioanalyzer (Agilent Santa Clara, CA). Only samples with RNA integrity number (RIN) >8.0 were subjected to RNA sequencing.

3.3.3 Preparation of RNASeq libraries

Transcriptome libraries were prepared following a modified protocol previously described for generating strand specific RNASeq libraries[24]. Briefly 2.5 µg of total RNA was subjected to polyA selection using oligodT beads (Invitrogen, Carlsbad, CA). Purified polyA RNA was fragmented and reverse transcribed using SuperscriptII (Invitrogen, Carlsbad CA). Second strand synthesis was performed with DNA Polymerase I (New England Biolabs, Ipswich, MA) in the presence of dNTP mix containing dUTP instead of dTTP. The product was then subjected to end repair, A base addition and adaptor ligation steps. Libraries were next size selected in the range of 350 bps after resolving in a 3% Nusieve 3:1 (Lonza, Basel, Switzerland) agarose gel and DNA recovered using QIAEX II gel extraction reagent (Qiagen, Valencia, CA). Libraries were barcoded during the 14-cycle PCR amplification with Phusion DNA polymerase (New England Biolabs, Ipswich, MA) and purified using AMPure XP beads (Beckman Coulter, Brea, CA). Library quality was estimated with Agilent 2100 Bioanalyzer for size and concentration. The paired end libraries were sequenced with Illumina HiSeq 2000 (2x100 bases, read legth). Reads that passed the filters on Illumina BaseCall software were used for further analysis.

3.3.4 PCR fusion validation

We validated a subset of nominated fusion genes by THF from UMICH cohort using real-time (RT-PCR). Of the 27 attempted fusions, 24 were validated,

2 had inconclusive results and 1 was not validated, representing a validation rate of 89%.

3.3.5 RNA isolation, cDNA synthesis and quantitative RT-PCR

Total RNA was isolated using either QIAzol reagent or RNAeasy micro kit (QIAGEN, Valencia, CA). cDNA was synthesized from total RNA using Superscript III in presence of random primers (Invitrogen, Carlsbad, CA). Quantitative Real-time PCR (qPCR) was performed using SYBR Green Master mix on the StepOne Real-Time PCR System (Applied Biosystems). All oligonucleotide primers for the qPCR assays were obtained from Integrated DNA Technologies (Coralville, IA); NRG1 forward 5'GATTCCTACCGAGACTCTCCTC3' and reverse 5'TGGAAGGCATGGACACCGTCAT3' and GAPDH forward 5'GTCTCCTCTGACTTCAACAGCG3' and reverse primer 5'ACCACCCTGTTGCTGTAGCCAA3'. Fold changes were calculated relative to GAPDH and normalized to the non-targeting control.

3.4 siRNA knockdown studies

Lung cancer cell line NCI-H1793 were plated in 6-well plates at a desired numbers and transfected with 2 nmol of NRG1 siRNAs (J-004608-11; and J-004608-12) or non-target control siRNA (Thermo Scientific). Transfection with oligofectamine reagent (Invitrogen, Carlsbad, CA) was performed twice over a period of 48 hours. Knockdown efficiency was determined by qPCR. Cell proliferation assessed by Incucyte, 24 hours after transfection, cells were trypsinized and plated in triplicate at 8,000 cells per well in 24-well plates. The plates were incubated in the IncuCyte live-cell imaging system (Essen Biosciences) at 37°C with 5% CO_2 atmosphere. Cell proliferation rate was assessed by kinetic imaging confluence measurements at 3-hour time intervals.

3.4.1 Cloning and expression of CD74-NGR1 fusions and cell proliferation and migration assays

CD74-NRG1 fusion transcript was amplified from the Index lung cancer sample tissue cDNA with forward 5'CACCATGCACAGGAGGAGAAGCAGGAGCTGT3' and reverse primers 5'TTCAGGCAGAGACAGAAAGGGAGTGGA3' using Hi-fidelity polymerase (Qiagen, Valencia, CA). The PCR product was gel purified and cloned into plenti-TOPO cloning vector (Invitrogen, Carlsbad, CA) and the DNA sequence was independently verified by Sanger sequencing. The control LacZ or C-terminal V5 tagged CD74-NRG1 constructs were transfected into the normal lung epithelial cell line BEAS-2B cells. The stable cells generated after selection in BEBM media (Lonza, Basel, Switzerland) containing 3 micrograms of blasticidin (Invitrogen, Carlsbad, CA). For proliferation assays, 50,000 cells were plated in 12-well plates and grown in regular media. Cells were harvested by trypsinization and counted manually at indicated time points. All assays were performed in quadruplicates. For migration assays, stable cells were re-suspended in medium without growth factors, then seeded at 50,000 cells per well into Boyden chambers (8 μm pore size, BD Biosciences) and were incubated for 24 hours in a humidified incubator at 37°C, 5% CO_2 atmosphere. The bottom chamber contained medium with growth factors as chemo-attractant. The top non-migrating cells were removed with a cotton swab moistened with medium and the lower surface of the membrane was stained with Diff-Quick Stain Set (Siemens). The number of cells migrating to the basal side of the membrane was visualized with an Olympus microscope at 20x magnification. Pictures of five random fields from 4 wells were obtained and the number of cells stained manually quantified.

3.4.2 Protein isolation and western blot analysis

Cells were plated in 100 mm plates and incubated at 37ºC in 5% CO_2 overnight to allow cells to adhere. Cells were washed with ice cold PBS twice. Whole-cell extracts from treated or untreated cell lines were harvested using cell lysis buffer (Cell Signaling), according to the manufacturer's protocol. Protein concentrations in cell lysates were measured using the protein assay quantification (bicinchoninic) (Pierce, Rockford, IL). Equal amounts of protein were loaded in each lane. Cell lysates were resolved under reducing conditions by 10% SDS-PAGE and then transferred to PDVF membranes. After being blocked with 5% milk in tris-buffered saline (TBS) with 0.1% Tween20, the membranes were incubated with antibodies against activated or total forms of protein overnight at 4°C, washed three times with 0.1% Tween 20 - TBS and then incubated for 60 minutes with 2000:1 peroxidase-conjugated anti-rabbit IgG. Antibodies against E-Cadherin, Vimentin, phospho-Erbb3, phospho-Erbb3, phosho-ERK and total-ERK were purchased from Cell Signaling Technology Inc. (Beverly, MA). Total Erbb3 and Erbb4 were purchased from Santa Cruz Biotechnology Inc. (Dallas, TX). The membrane-bound peroxidase activity was detected using ECL Prime Western Blotting Detection kits (Amersham, Arlington Heights, IL) and chemiluminescent images were captured by exposing film.

3.4.3 Chemicals and cell proliferation assays

Two MEK inhibitors (AZD6244 and GSK1120212), an EGFR/ERBB2 inhibitor (Afatinib) and a MET/ALK inhibitor (Crizotinib) were obtained from SelleckChem. The effect of these drugs on proliferation of NCI-H1793, NCI-H1299 and NCI-H1792 was measured. Cells were plated at 1.5 – 2.5 × 10^3 cells/well in 100µl of appropriate culture medium using 96-well plates and incubated at 37ºC in 5% CO_2 overnight and then treated 24 hours later with the respective drugs. Each inhibitor was prepared at 7 serial dilutions ranging from 0.03 to 30µM at the final concentration. On day 3, cell viability was assessed using 10µl/well of WST-1 reagent (Roche), according to manufacturer's

instructions. The absorbance at 450 nm and reference at 630 nm were measured using an automated plate reader (ELx808 Bio-Tek) at different time-points. Cell proliferation was estimated by dividing the mean absorbance of the treatment group divided by the mean absorbance of the vehicle-treated control X 100%. Inhibitory concentration 50% (IC_{50}) was calculated using GraphPad Prism 6 software.

3.5 Results

3.5.1 Patient cohort description

We have assembled a cohort of 732 patient samples, which includes lung adenocarcinoma and squamous carcinoma patients, by combining our UMICH cohort with public available data from TCGA and the recently published cohort by Seo *et al.,* 2012[23] (from here on the Seoul cohort).

We sequenced mRNA from 133 samples by using strand specific RNA paired-end sequencing (ssRNASeq) technology. The UMICH cohort includes 67 LUAD, 36 LUSC, 24 lung cancer cell lines and 6 matched nonmalignant lung samples. Moreover this cohort included 64 stage I, 17 stage II and 22 stage III patients. Eighty-nine patients were smokers, whereas 8 were never-smokers and in 4 cases the smoking status was unknown. The median smoking pack years was 45 (range, 2 – 300) and practically all patients were heavy smokers (more than 10 pack years). The average follow up was 5.05 years. Sample acquisition details were described provided in the methods section. The TCGA cohort used in this study encompasses 305 LUAD and 216 LUSC samples. This includes 250 stage I, 112 stage II, 101 stage III, and 19 stage IV cases as well as 39 with unknown stage. This cohort includes 4 never-smokers, 20 light smokers (defined by less than 10 pack years of tobacco use) and 365 heavy smokers (more than ten pack years of tobacco smoking), and the average follow up was 1.72 years. Finally, the Seoul cohort includes 79 LUAD, which did not have public available

clinical information. The Seoul cohort includes 79 matched normal samples; fusions called in these normal samples were used for filtering as described in the methods.

In summary, the combined cohort used in this study includes 451 lung adenocarcinomas, 251 lung squamous carcinomas and 24 NSCLC cell lines, making this the most comprehensive RNA-Sequencing cohort of lung cancers assembled so far. A summary of the Clinic-pathological characteristics is provided in Table 3.3-3.

Table 3.3-3. Clinic-pathological characteristics of the combined lung cohort used in this study.

SAMPLES				
	LUAD	**LUSC**	**LUCL**	**TOTAL**
UMICH	**67**	**36**	**24**	127
SEOUL	79	0	0	79
TCGA	305	216	0	521
TOTAL	451	251	24	**727**
SEX				
	MALE	**FEMALE**		
UMICH	56	55		
SEOUL	48	31		
TCGA	298	223		
TOTAL	402	309		
FOLLOW UP TIME				
	MIN	**MEDIAN**	**MAX**	**AVAILABLE**
UMICH	0.26	4.6411	17.3726	111
SEOUL	NA	NA	NA	0
TCGA	0	0.9233	18.6630	436
TUMOR STAGE				
	STAGE I	**STAGE II**	**STAGE III**	**STAGE IV**
UMICH	64	17	22	0
SEOUL	NA	NA	NA	NA
TCGA	250	112	101	19
SMOKING				
	NEVER	**LIGHT**	**HEAVY**	
UMICH	8	NA	89	
SEOUL	NA	NA	NA	
TCGA	4	20	365	

3.5.2 Global overview of the fusions' landscape

Fusion calling has lagged behind single nucleotide variant calling, and currently there are not best practices for fusion identification, removal of false positives neither benchmarking comparison of different algorithms on public available dataset with golden truth positives. In order to have comparable results among samples and cohorts it is important to develop unified and data driven fusion prediction pipelines. We used the workflow described in Figure 3.1 (See Methods) to identify fusions, quantify the total number of observed fusions in each of patient, and integrate mutation and clinical data for each of the 732 patients in our combined cohort

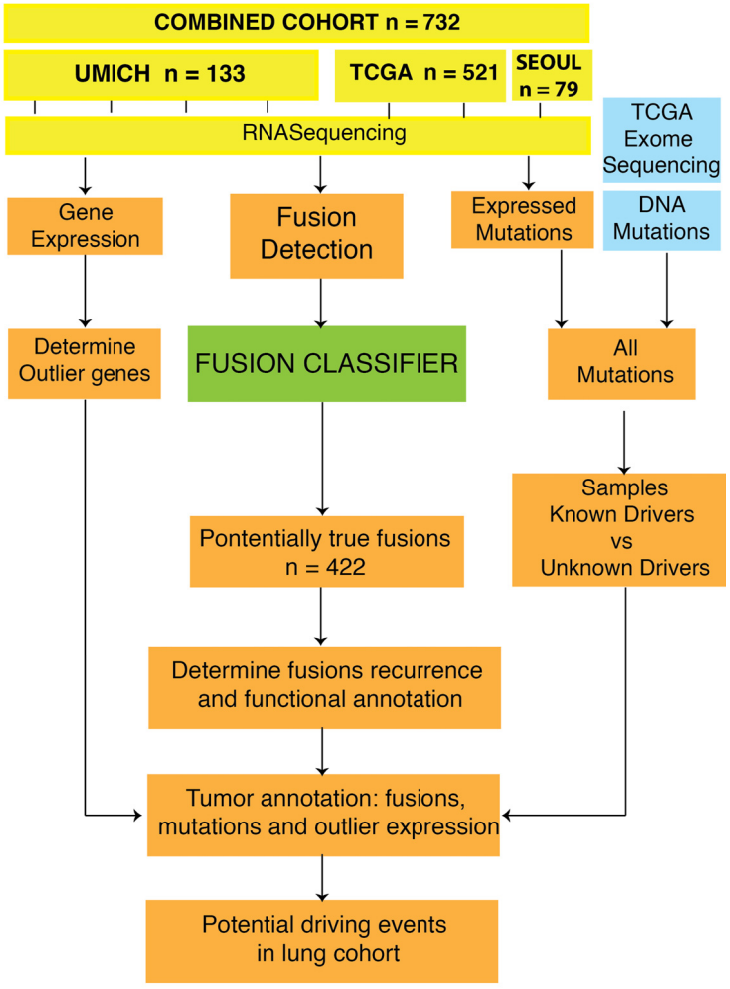


Figure 3.1 Schematic diagram of the data generation and analysis workflow of lung cancer RNASeq data.

A total of 732 lung cancer samples that include 708 clinical specimens and 24 cell lines, representing 451 LUAD and 251 LUSC, were interrogated for gene fusions and somatic mutations. The cohort was assembled combining 133 University of Michigan samples (UMICH), 79 Seoul National University samples (SEOUL), and 521 Cancer Genome Atlas samples (TCGA). The RNASeq data was mapped to human RefSeq Hg19 using TopHat2. Fusion calls were made with TopHat-Fusion (THF). In all cases fusions present in normal samples were considered false positives and filtered out. We developed and applied a fusion classifier that retained 422 gene fusions for further downstream analysis. The 422 fusions were classified into recurrent (>2 samples) and private fusions and further divided into inter chromosomal, intra chromosomal or fusions resulting from potential tandem duplication events. In addition somatic mutations in well-known lung oncogene drivers and tumor suppressors were determined and annotated for each sample. Finally, both LUAD and LUSC cohorts were divided into either samples harboring oncogene driving mutations or samples without known driver genes.

We detected an average of 13 fusions per tumor sample (range, 0 - 67). Although, both lung adenocarcinoma and squamous carcinoma have similarly high single nucleotide mutation rate of about 8.1 mutations/Mb[3,130], they showed different average number of fusions per sample. We observed an average of 11 fusions in lung adenocarcinoma tumors, while 17 in squamous carcinoma (student t-test p-value < 2.2×10^{-16}). Moreover, we did not observe statistically significant differences on the average number of fusions between heavy and light smokers (LUAD student t-test p-value= 0.75; LUSC student t-test p-value=0.42); nor among different clinical stages regardless of the tissue type (Table B-2, Table B-3). However, we did find that tumors harboring missense or nonsense mutations in *TP53* showed greater average number of fusions as compared to samples with *TP53* wild type (Supplementary Figures 1a, 1b, p-value = 0.0012). Because > 80% of lung squamous carcinomas have somatic mutations in *TP53*[3]; that difference is consistent with the one observed on the average number of fusions between LUAD and LUSC carcinomas. In LUAD, we also observed a significant correlation among the presence of oncogenic mutations (e.g. *KRAS* activating mutations) and *TP53* deleterious mutations (stop codon or splice site mutations), and the number of fusions (Fisher's exact test p-value=0.0089). This correlation could not be tested in LUSC because in this indication there were a very few number of samples with mutations in *KRAS*, *EGFR* or other oncogenes.

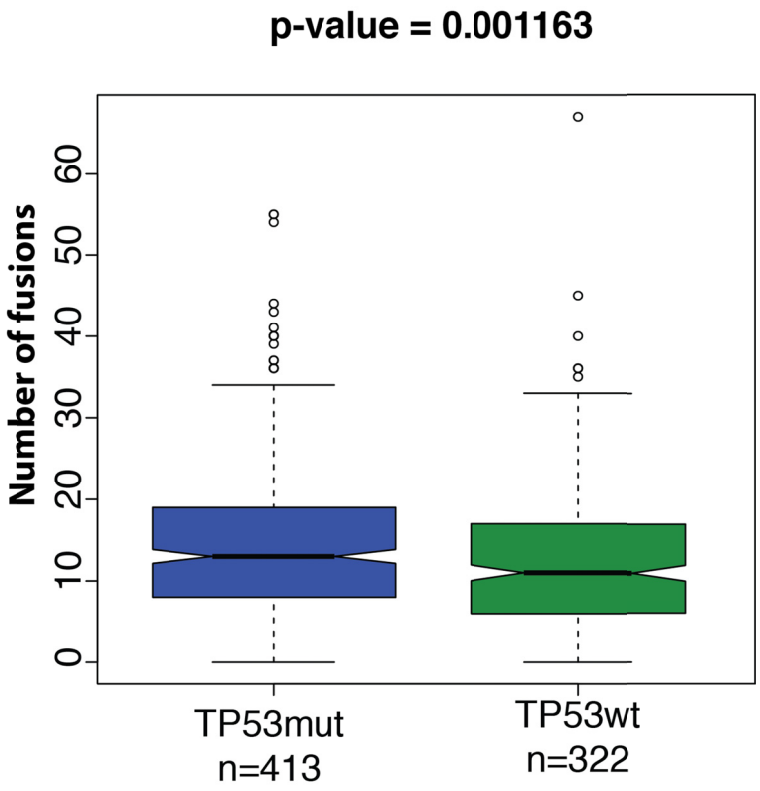


Figure 3.2. Comparison between the number of fusions in samples with TP53 wild type vs TP53 mutated samples.

3.5.3 Number of fusions is associated with prognosis

Because fusions usually result as a consequence of genomic rearrangements, the number of fusions might indicate the level of rearrangement undergone by the cancer patient's genome. Therefore, we investigate the relation between number of fusions by patient and their prognosis. Patients in our combined cohort were classified into three categories: patients with low (0-7), intermediate (8-17), or high (≥18) number of fusions and performed a 10-year Kaplan-Meyer survival analysis. Patients with high number of fusions had significantly shorter median overall survival (35.6 months, 95% confidence interval (CI) 27.2 – 43.9) as compared to tumors with intermediate (49.5 months, 95% CI 23.9 – 75.1) or low number of fusions (62.3 months, 95% CI 44.6 – 80.1; log-rank p-value 0.017, Figure 3.3).

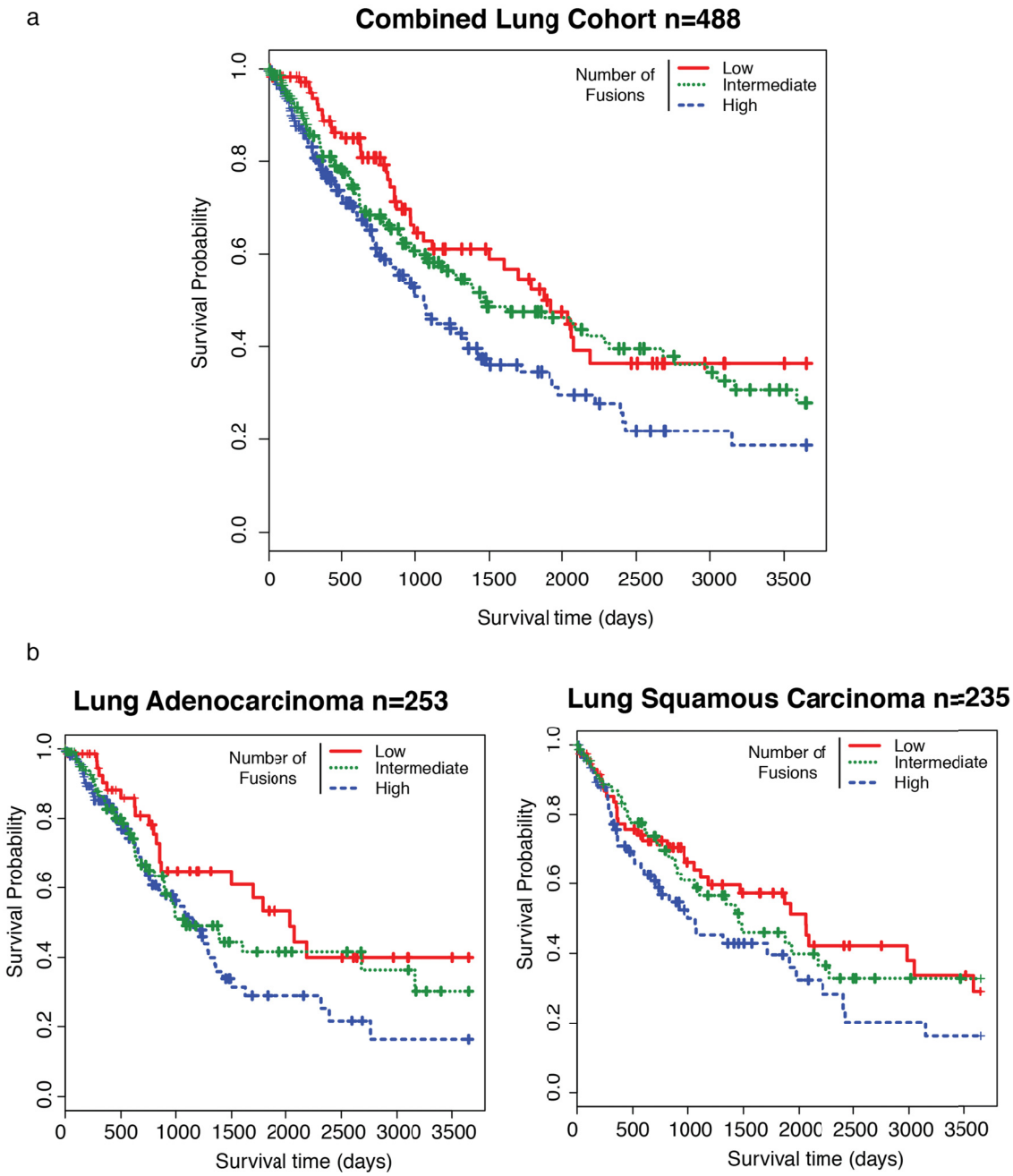


Figure 3.3. Gene fusion frequency is a prognostic indicator in both LUAD and LUSC.

A) Kaplan-Meyer survival curve for the combined cohort samples with low (0-7), intermediate (8-16), or high (≥17) number of fusions (p-value=p=0.0089). Samples with high number of fusions have worst prognosis (Cox survival analysis p-value=0.0053) **B)** Kaplan-Meyer survival curve for LUAD samples with low (0-6), intermediate (7-12), or high (≥13) number of fusions (p-value=0.076). Samples with high number of fusions have worst prognosis (Cox survival analysis p-value=0.029) b) Kaplan-Meyer survival plot for LUSC samples with low (0-11), intermediate (12-18) and high (≥19) number of fusions (p=0.169). Samples with high number of fusions have worst prognosis (Cox survival analysis p-value= 0.0717).

Statistically significant clinical covariates in the univariate Cox model (Table 3.3-4) were used in a multivariate analysis examining the prognostic value of number of fusions.

Table 3.3-4. Univariate Cox regression for overall survival according to clinical variables (n = 621).

	Overall survival		
	HR	**95% CI**	**p-value**
Age, continuous	1.03	1.01 – 1.04	< 0.001
Sex			
Female	1.00	--	
Male	1.33	1.02 – 1.74	0.037
Stage, continuous	1.55	1.35 – 1.76	< 0.001
Smoking status			
Non-smoker	1.00	--	
Smoker (<35 pack-year)	1.31	0.52 – 3.30	0.565
Smoker (≥35 pack-year)	1.49	0.61 – 3.67	0.378
Histology			
Adenocarcinoma	1.00	--	
Squamous cell carcinoma	0.99	0.76 – 1.29	0.989
***TP53* status**			
Wild-type	1.00	--	
Mutant	0.94	0.66 – 1.33	0.717
***KRAS* status**			
Wild-type	1.00	--	
Mutant	0.94	0.66 – 1.33	0.717
***EGFR* status**			
Wild-type	1.00	--	
Mutant	1.01	0.77 – 1.33	0.924

Strikingly, high number of fusions was independently associated with worse overall survival (HR = 1.56, 95% CI 1.13 – 2.15, p-value = 0.007, Table 3.3-5), after adjusting for gender and disease stage. When mutation statuses of

TP53, KRAS and *EGFR* or smoking status were included in the analysis, number of fusions remained independently associated with worse outcome as well (p-value =0.005).

Table 3.3-5. Multivariate Cox regression for overall survival according to number of fusions in 621 NSCLC patients adjusted by age, gender and stage.

Covariates in the model		Hazard Ratio	95% confidence intervals	p-value
Age, continuous		1.04	1.02 – 1.05	<0.001
Gender	Female	1.00	–	
	Male	1.17	0.89 – 1.54	0.270
Stage, continuous		1.64	1.43 – 1.88	<0.001
Number of fusions	Low	1.00		
	Intermediate	1.11	1.78 – 1.59	
	High	1.56	1.13 – 2.15	0.007

3.5.4 Lung fusions landscape is dominated by low recurrence and private fusions

In order to prioritize fusion candidates and discriminate potentially true fusions from spurious ones, we developed a classifier to distinguish potentially genuine fusions from false positive ones (See Methods). This classifier uses structural and functional annotation features of each fusion in order to predict whether a fusion is potentially genuine or not.

Remarkably, our classifier has a recovery rate greater than 90% and it automatically recapitulates our intuitive knowledge about the important structural properties defining *bona fide* fusions (Methods). In our fusions' dataset, the top five features contributing the most to the fusion classifier are, in decreasing order

of importance, fusion type (Inter-chromosomal, Intra-chromosomal, Tandem-duplication), sum of the median alignment quality of reads supporting the fusions, number of spanning and encompassing reads across the fusion junction and the cohort normalized expression value for the 3'-partner gene (Figure 3.4).

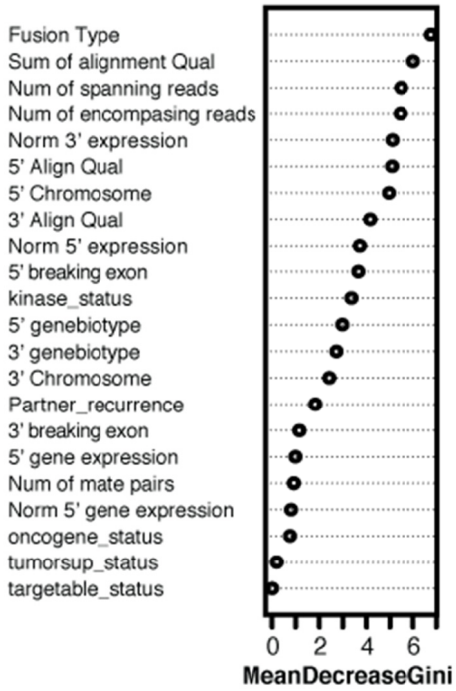


Figure 3.4. Features used for the fusion's classifier.

Features are in decreasing order of importance according to the mean decrease in the Gini score.

Using this classifier, 422 potentially genuine fusions were nominated across the whole cohort (Fusions Table[3]). Sixty-four out those 422 fusions (15%) involved kinases (either as 3' or 5'-partner) including known ROS1, RET and ALK, 52 fusions involved oncogenes and 63 involved tumor suppressors (Fusions Table). Moreover, of those fusions involving cancer-related genes we found 61 productive fusions (fusion protein is produced), 63 disruptive (ORFs of original genes are destroyed) and 6 promoter fusions (3' partner full wild type

[3] The complete fusion table is available upon request as an excel file. It is not included in this document because of the large number rows (> 400), making inefficient to paste it into a text document.

sequence is preserved). Given the size of our cohort, we could better estimate the recurrence of different gene fusions and so we distinguish between three types of recurrence: molecular, functional and family recurrence. Molecular recurrence refers to fusions in which the same 5' and 3' partners are observed in different samples such as SLC34A2-ROS1; functional recurrence refers to cases in which either the 5' or 3' partner is the same (CCDC6-RET and KIF5B-RET); and gene family recurrence correspond to gene fusions in which 5' or 3' partner belongs to the same gene family (FGFR3-TACC3, FGFR2-CCDC6, BAG4-FGFR1). Functionally recurrent kinase fusions ROS1, RET and ALK were found on 0.86%, 0.29%, and 0.14% across this combined cohort (Table B-4,Table B-5). Other functionally recurrent gene fusions include BCAS3-MAP3K3, MRC2-MAP3K3; and GOSR1-NF1 and NLK-NF1 and NF1-PSMD11. The recurrent gene fusions involving the tumor suppressor Neurofibromin 1 (NF1) do not generate productive fusion proteins (GOSR1-NF1, NLK-NF1, NF1-PSMD11) and instead destroy the functional activity of NF1, suggesting that this could be an additional mechanism for NF1 inactivation in lung cancers. NF1 inactivation leads to activation of the *PIK3CA* pathway.

Our results confirm the high heterogeneity and low recurrence of lung cancer fusions, as most fusions found were private fusions or appeared at very low frequency (Fusions Table and Table B-4,Table B-5)).

Although present in a small percentage of patients, known targetable fusions are preferentially observed in samples lacking any other known oncogenic drivers. We therefore determine for each sample in our combined cohort the mutation status of well-known oncogenes and tumor suppressor playing a role in lung cancers (Methods), reproducing previous results about the mutational landscape of LUAD and LUSC (Figure 3.5) and confirming that known fusions involving ROS1, RET and ALK are exclusively found in samples without other oncogenic drivers. In this set of samples the frequency of those fusions was 1.3%, 0.52 and 0.26 respectively. Moreover, our integrative analysis

combining mutational and fusion status with gene expression also showed that for fusions such as ROS1 in some index samples the expression of the fusion kinase was an outlier across the combined cohort (e.g., 3 out 6 in ROS1). Interestingly, we also identified the presence of samples with outlier expression of ROS1, and FGFR3 almost exclusively in samples without other oncogenic drivers (Fisher exact test p-values= 0.0048 and 0.0864 respectively, Figure 3.5). While the mechanism of overexpression remains to be delineated, the outlier kinase expression may have a potential driving role and this patient subset may also benefit from the available tailored drug therapies.

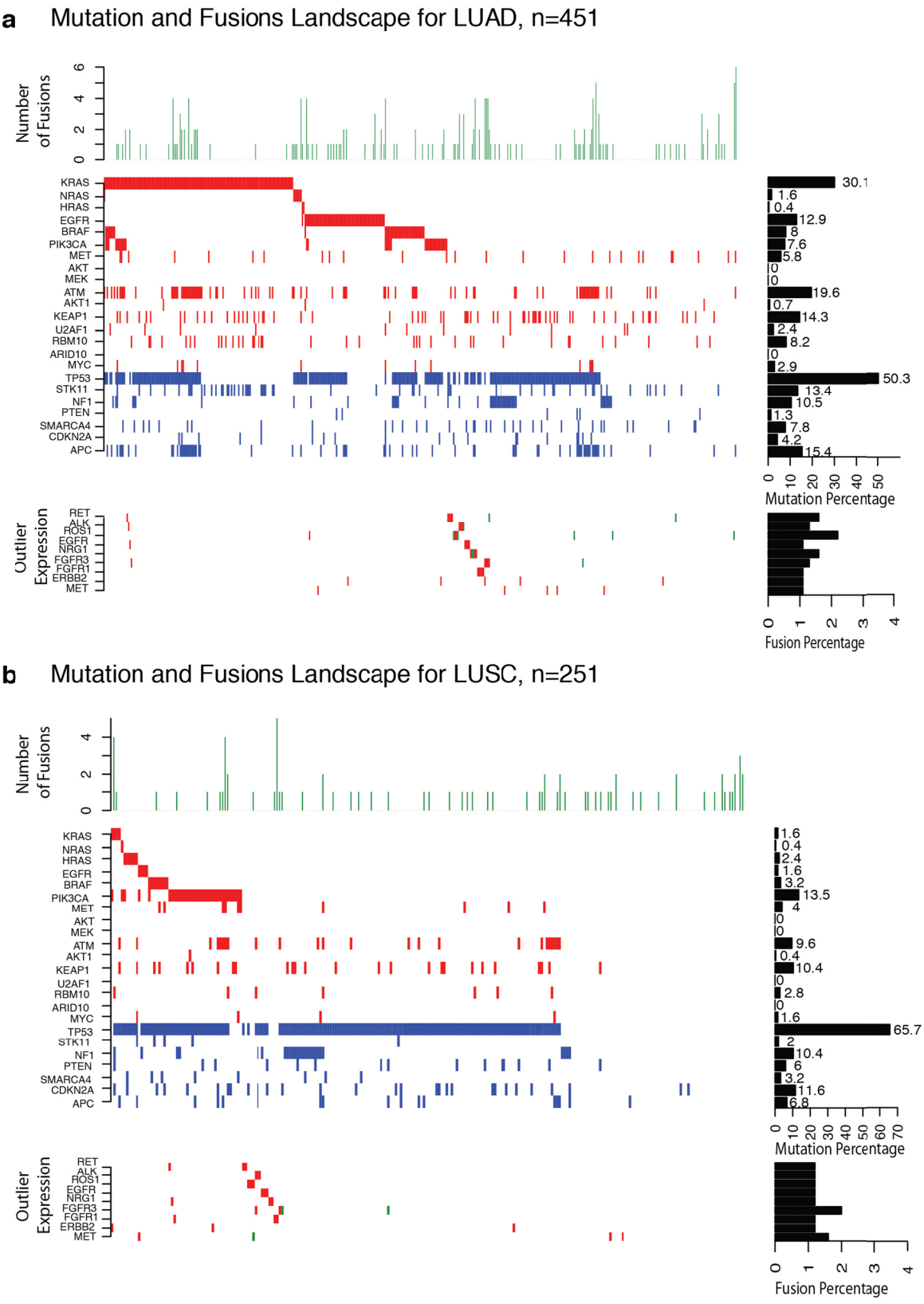


Figure 3.5. The gene fusion and mutational landscape of lung cancers.

A) Lung adenocarcinoma (LUAD, n=451). **B)** Lung squamous carcinoma (LUSC, n=251). *Top panel:* Histograms represent the number of high quality fusions identified in each sample. *Central Panel:* Heatmap denotes the presence or absence of activating mutations in known oncogenes (red) and deleterious mutations in tumor suppressors (blue). Samples are presented in columns and genes are presented in rows. *Right middle panel*: Bar plot summarizes the number of samples harboring activating or deleterious mutations for each gene. *Bottom Panel:* Heatmap displays samples harboring both known and novel gene fusions (green) involving either receptor kinase genes or NRG1. Samples in red indicate outlier expression pattern observed in the respective genes. The ordering of samples in center panels was dictated by mutation status in *KRAS, NRAS, HRAS, EGFR, BRAF, PIK3CA*, and *TP53* genes in that order. Remarkably, *NRG1* gene fusions were observed in samples that lack other driver events similar to the RTK fusions involving *ROS1, RET, ALK* and *FGFR3* genes. The NRG1 fusion index samples exhibited outlier *NRG1* expression and the outlier samples harbored no other known driver events in both LUAD and LUSC.

3.5.5 Recurrent NRG1 fusions in lung cancers

Remarkably, we found a novel functionally recurrent gene fusion where the common 3'-gene Neuregulin 1 (NRG1) was fused to different 5' partners (Figure 3.6a). The gene fusions, CD74-NRG1, RBPMS-NRG1 and WRN-NRG1, occurred in both LUAD and LUSC samples. While both CD74-NRG1 and RBPMS-NRG1 fusion events resulted in the production of chimeric fusion proteins, the WRN-NRG1 fusion results in the overexpression of full length NRG1 controlled by the WRN gene promoter. As a member of EGF family of ligands, the growth factor NRG1 transduces its signal through the HER/ErbB receptor tyrosine kinases[143,144]. NRG1 protein contains various domains such as kringle like, immunoglobulin like domains and the EGF domain that is located in the C-terminal region[143]. Notably the EGF domain that is essential for receptor interaction[145] is preserved in all the NRG1 fusions identified (Figure 3.6a). All NRG1 fusion index samples were found in the driver negative group (0.78% frequency) and displayed outlier expression of NRG1 specifically in the tumor sample and not in the matched normal tissue (Figure 3.6b and Figure 3.6c). Strikingly similar to the pattern described above for known kinases fusions such as ROS1. Therefore, we reason that NRG1 overexpression could be implicated in its dysregulation. Among all samples in our combined cohort, the driver negative lung cancer cell line H1793 exhibited the highest expression of NRG1 (more than 250 FPKM) (Figure 3.6d), but no NRG1 fusion was detected either by RNASeq or FISH. To understand NRG1 functionality in this cell line we resorted

to siRNA mediated gene knock down. A 70% knock down achieved with two independent NRG1 siRNAs (Figure 3.6e) affected cell proliferation rate as indicated by cell growth assay (Figure 3.6f). Outlier NRG1 expression was also observed in 10 other driver negative samples (Table 3.3-6), elevating the frequency of samples with NRG1 dysregulation to 13/314 (4.14% recurrence in the driver negative group) implicating a potential causal role for NRG1 in this patient subpopulation.

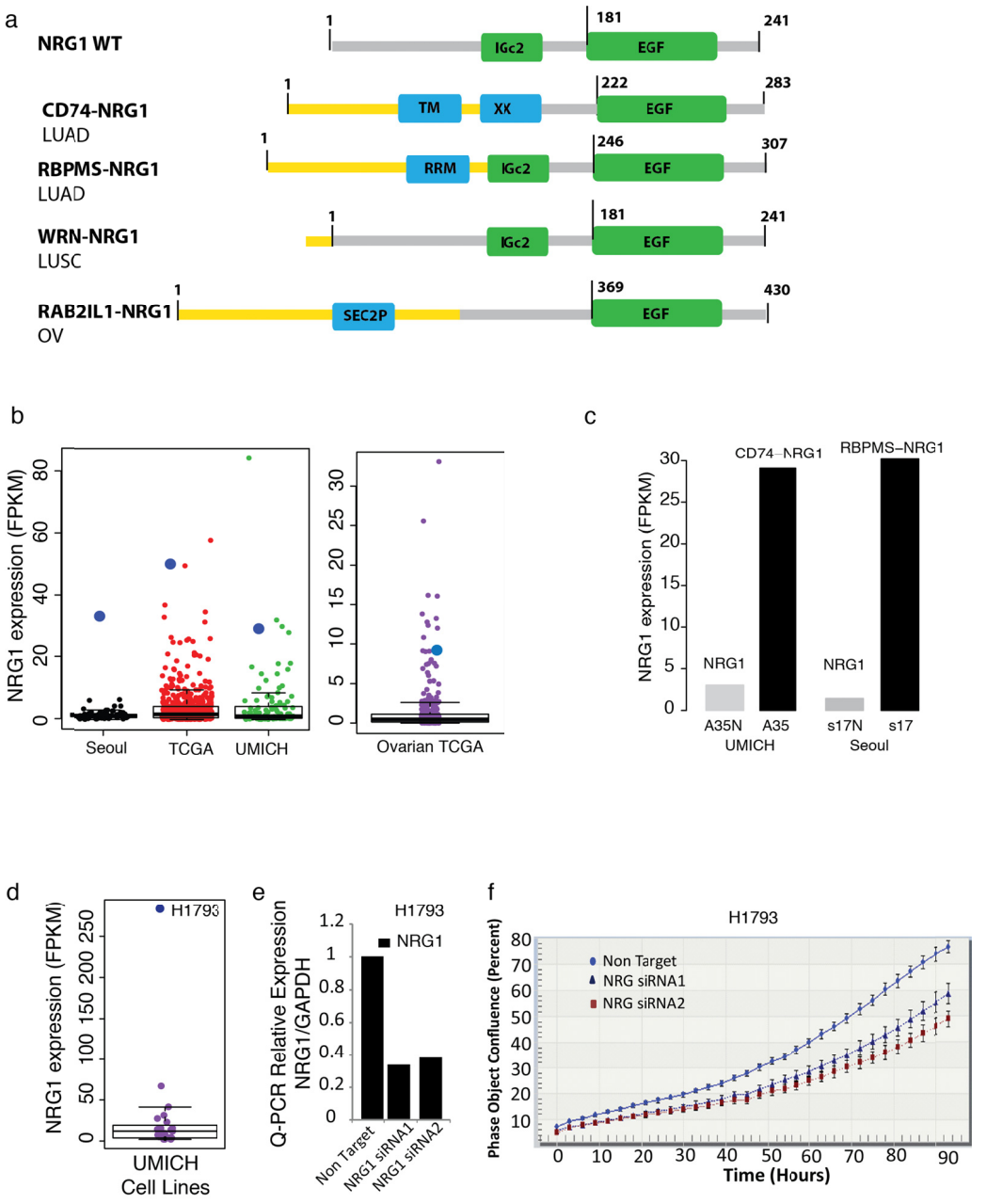


Figure 3.6. Recurrent cancer specific NRG1 fusions in lung cancer.

A) Recurrent fusions involving NRG1 as a 3' partner were detected in lung adenocarcinoma and lung squamous carcinoma in the three cohorts included in this study. Schematic representation of functional domains present in the NRG1 fusion proteins namely CD74-NRG1; RBPMS-NRG1 (LUAD); WRN-NRG1 (LUSC) and RAB2IL1-NRG1 (Ovarian) compared to the wild type NRG1 (Top). The receptor binding EGF domain is preserved in all fusions. TM- Transmembrane domain; RRM- domain; IGc2- domain; SEC2P-domain; **B)** Analysis of RNASeq expression values, revealed outlier NRG1 mRNA expression in all index cases within each cohort. **C)** High NRG1 mRNA expression driven by the fusion event in the index tumor tissue compared to matched normal, in both UMICH and Seoul LUAD represented as bar plot. **D)** Boxplot showing outlier expression of NRG1 in H1793 in UMICH lung cell line cohort. **E)** Two independent siRNAs mediated knockdown of NRG1 in H1793 cell reduced NRG1 transcript expression (Q-PCR) and decreased cell proliferation as monitored by Incucyte confluence analysis.

Table 3.3-6. Lung cancer samples harboring fusions and/or outlier expression of NRG1.

Patient	Cohort	Disease	Fusion	Oncogene	Mutation	Outlier Percentile	FPKM	NRG1 Cohort
A35	umich	LUAD	YES	NO	TP53 p.P33R p.P72R; RBM10 p.A630P p.A696P; SMARCA4 p.R513W; APC p.V1822D p.V1804D; ATM p.N1983S	99%	29.08	0.9273
lc_s17	seoul	LUAD	YES	NO	TP53 p.PXXXR	99%	33.08	0.9273
C028	umich	LUAD	NO/ TBD	NO	SMARCA4 p.E1056X; TP53 p.R248L p.R209L; APC p.V1822D p.V1804D; ATM p.N1983S;	99%	83.92	0.9273
0232d299-4cdf-4fd7-9a5e-8d13c208b40c	tcga	LUAD	NO/ TBD	NO	TP53 p.R156P; KEAP1 p.D236N; RBM10 p.S781L;	99%	21.32	0.9273
7b0622ab-63ea-483f-ae40-d3ea587bdbba	tcga	LUAD	NO/ TBD	NO	-	99%	25.86	0.9273
H1793	umich	LUAD_cl	NO/ TBD	NO	SMARCA4 p.E514X; TP53 p.P33R p.R141H; APC p.V1822D p.E1991D; EGFR p.C311F; ATM p.N1983S;	99%	281.86	10.1265
a3e1ac67-a1f2-44fb-8343-a7e8239fc24a	tcga	LUSC	YES	NO	TP53 p.G244C; PIK3CA p.D1045V	99%	49.5573	4.2247
ce8612ab-3149-4a6a-b424-29c0c21c9b8b	tcga	LUSC	NO/ TBD	NO	TP53 p.S314fs; CDKN2A p.P3fs; APC p.S966G; NF1 p.E1734V.	99%	34.5314	4.2247
7e691df8-8ea6-472c-86bf-504c7ba6983d	tcga	LUSC	NO/ TBD	NO	APC p.S966G; CDKN2A p.P3fs; TP53 p.S314fs; NF1 p.E1734V	99%	49.3324	4.2247
791f1b21-695e-4db1-b41d-80590c09d257	tcga	LUSC	NO/ TBD	NO	KEAP1 p.R320Q p.R470C; PIK3CA p.E453K	99%	31.2416	4.2247
14a4a93a-e24d-46f2-bee3-18bd792ef95a	tcga	LUSC	NO/ TBD	NO	TP53 p.E271*	99%	36.7394	4.2247
6394fe4a-6034-4c79-b28f-aa43e3753730	tcga	LUSC	NO/ TBD	NO	-	99%	57.5317	4.2247

In addition to characterizing NRG1 fusions, we used normal lung BEAS-2B cells to generate stable overexpression of CD74-NRG1 fusion protein. Fusion overexpression significantly increased both cell proliferation (Figure 3.7a) and migration (Figure 3.7b, Figure 3.7c) and induced a notable phenotypic alteration in cell shape (Figure 3.7d). Western blot analysis revealed evidence for epithelial mesenchymal transition (EMT) upon CD74-NRG1 overexpression as supported by increased vimentin protein expression (Figure 3.7e). In order to identify potential pathways activated by the CD74-NRG1 fusion, we performed gene expression microarray profiling of CD74-NRG1 and LacZ clones. Significant analysis of microarrays (SAM) shows vimentin as one of the top overexpressed genes in CD74-NRG1 confirming the western blots (Figure B.1), as well as, down regulation of cadherins, supporting the hypothesis of EMT in CD74-NRG1 positive cells. Gene set enrichment analysis identified down regulation of cell adhesion pathways (Figure B.1) and, interestingly, up-regulation of SRC (Figure 3.7f) and ERBB (Figure 3.7g) pathways in CD74-NRG1 cells. In light of these results we assessed the activation of those pathways by western blot and confirmed that, compared to LacZ control, the CD74-NRG1 cells showed substantially increased levels of phosphorylated ERBB3 and phosphorylated JNK, while a modest increase in phospho-ERK (Figure 3.7h). Having functionally characterized CD74-NRG1 fusion in lung cancers, we looked for productive and outlying NRG1 fusions in other cancer types and found the presence of RAB2IL1-NRG1 in ovarian cancer. As noted in the lung cancer fusions, functional EGF domain is retained in RAB2IL1-NRG1 and the fusion index case exhibited outlier NRG1 expression (Figure 3.7a and Figure 3.7b).

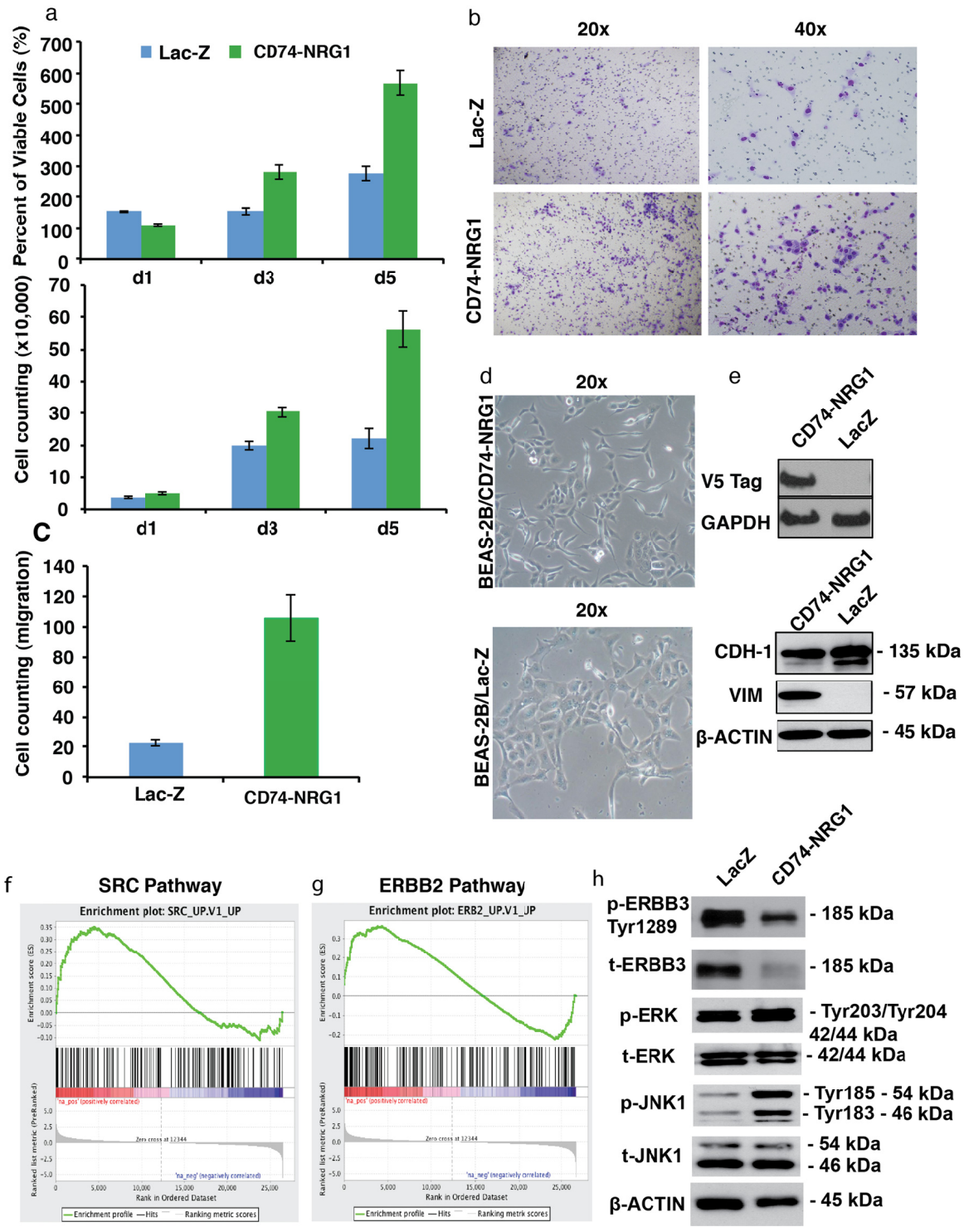


Figure 3.7. Functional Characterization of NRG1 fusion.

A) Cell proliferation by WST-1 assay (upper panel) and cell counting (lower panel) on BEAS-2B cells stably transfected with Lac-Z or CD74-NRG1 fusion. Both assays demonstrated that cells expressing the CD74-NRG1 fusion had significantly higher proliferation rate at day 3 and 5 ($p<0.001$ for both time-points) as compared to Lac-Z. **B)** Representative pictures of cells migrating to the basal side of the Boyden chamber

membrane after Diff-Quick staining with an Olympus microscope at 20x magnification. **C)** Cell counting for cells migrating through the Boyden chamber membrane after 24 hours. BEAS-2B cells expressing CD74-NRG1 fusion showed a higher migration rate as compared to Lac-Z (=0.0014). **D)** Representative pictures of BEAS-2B cells expressing the CD74-NRG1 fusion or Lac-Z. Cells expressing the CD74-NRG1 fusion appeared smaller and more fusiform as compared to Lac-Z, suggesting that they acquired a more mesenchymal phenotype. **E)** Western blot analysis of V5 Tag, E-cadherin (CDH-1) and Vimentin in transfected BEAS-2B cells. V5 Tag was expressed only in CD71-NRG1 transfected cells, which showed a slightly lower expression of CDH-1 and a significant increase of Vimentin expression. **F,G)** Gene set enrichment analysis based on differentially-expressed genes among BEAS-2B cells transfected with the CD74-NRG1 fusion or Lac-Z. Significant up-regulation of SRC and ERBB2 pathways was observed in CD74-NRG1 cells. **H)** Western blot analysis of ERBB3, ERK and JNK1 activation. CD74-NRG1 cells showed a noticeably activation of ERBB3 and JNK1 as compared to Lac-Z cells, whereas ERK activation was discreetly higher in the cells harboring the fusion.

3.6 Discussion

Treatment and diagnosis of NSCLC, especially LUAD, has been transformed by the use of targeted therapies and companion diagnostics tests. For example, EGFR activating mutations in exons 18, 19 and 21 are now routinely assessed before recommending treatments with Gefitinib and Erlotinib; as the response rate is close to 70%[146] in the mutation positive subpopulation of advanced NSCLC. More recently, fusions involving tyrosine kinases such as ROS1, ALK and RET[8,132,147] have been identified primarily in young lung adenocarcinomas patients with no other driver mutations and no history of tobacco smoking. Despite the low frequency of those fusions in the population, phase I clinical trials have shown that patients with EML4-ALK fusions respond well to Crizotinib[148,149] a drug targeting ALK, demonstrating the efficacy of targeting these kinases in the rearrangement positive subpopulation of patients. In this study, we use RNA sequencing to characterize, in an unbiased manner, the fusions' landscape of lung adenocarcinoma and lung squamous carcinoma indications in order to identify potentially new oncogenic fusions.

We showed that the fusions landscape is highly heterogeneous and dominated by low recurrence and private fusions (Figure 3.5); with on average higher number of fusions per sample being observed in LUSC than LUAD (t-test p-value $< 2.2 \times 10^{-16}$), but not statistically significant differences with respect to

any other clinical characteristics such as smoking history or disease stage (Table 3.3-4). We also found that tumors harboring missense or nonsense mutations in *TP53* have greater average number of fusions than *TP53* wild type, although this result is potentially confounded by the high prevalence of TP53 mutations in squamous carcinomas. Remarkably, high number of fusions was independently associated with worse overall survival (Table 3.3-5, Figure 3.3), after adjusting for gender, disease stage and mutation status of *TP53*, *KRAS* and *EGFR*. As RNA sequencing becomes widely adopted for profiling transcript expression and gene fusions, our results suggest that the number of fusions could be used as an independent prognostic marker in lung cancers.

The recurrent tyrosine kinase fusions, previously reported, are found almost exclusively in driver negative lung adenocarcinomas, and have not been reported in squamous carcinoma. Here, we found recurrent fusions involving 3'gene Neuregulin 1 (NRG1) (CD74-NRG1, RBPMS-NRG1 and WRN-NRG1) and NRG1 outlier expression in both LUAD and LUSC indications (Figure 3.6). NRG1, a growth factor receptor that transduces its signal through the HER/ErbB receptor tyrosine kinases pathway, is expressed in a subset cancers, namely breast, lung and others[150]. CD74 is a known 5'-fusion partner in lung cancer involved in ROS1 kinase fusions. While CD74-NRG1 and WRN-NRG1 fusions contains the signal peptide and type II transmembrane domain to locate NRG1 in the plasma membrane, cellular location of RAB2IL1-NRG1 and RBPMS-NRG1 needs to be further characterized. Nevertheless, it has been previously reported that of the 20 NRG1 transcript variants (several transcripts lack the N-terminal signal sequence required for transport to extracellular space and for membrane localization. In these instances an internal hydrophobic amino acid stretch is speculated to substitute the N-terminal signal sequence[143,145].

Remarkably, NRG1 fusions are present in samples with no other driver events (Table 2, Figure 3) and the index samples display outlying expression of the NRG1 gene (Figure 3.6), resembling oncogenic fusions such as ROS1.

Moreover, we also found 10 additional cases of outlying NRG1 expression in driver negative samples, suggesting NRG1 potential role as a driver on those samples. We demonstrated that abrogating NRG1 outlying expression affects cell proliferation (Figure 3.6) and more importantly we showed that the fusion construct (CD74-NRG1) increased proliferation and migration in cell line models (Figure 3.7). Taken together, NRG1 fusions and outlying expression of NRG1 account for 4.14% of the driver negative lung cancer patients.

The therapeutic potential of NRG1-ERBB autocrine loop has been previously suggested[151] and more recently blocking NRG1 and other ligand-mediated Her4 signaling were shown to be useful in enhancing the magnitude and duration of the chemotherapeutic response of NSCLC[152]. Further studies of the therapeutic opportunities for LUAD and LUSC patients with NRG1 rearrangements are warranted.

In conclusion, the previously documented success of targeted therapies against low recurrence oncogenic fusions in lung cancer and the high heterogeneity of the fusions' landscape, shown in this study, reinforce the demand for more personalized and tailored drug therapies for this disease.

3.7 Contributions

Science is a collective enterprise and it is much more fun when done with friends and good collaborators. The results presented in this chapter were made possible for the great collaboration and support of a team of people in the Chinnayian and Nesvizhskii labs.

O. Alejandro Balbin: Omics data integration, RNASeq data processing, development of fusion classifier, fusions database, mutation analysis, additional bioinformatics and clinical analysis, NRG1 functional experiments design. Manuscript writing. Saravana M. Dhanasekaran: Beautiful RNASeq strand

specific libraries for the UMICH lung cohort, CD74-NRG1 construct and transfection, ORFs construction of the NRG1 fusion proteins, NRG1 functional experiments design. Manuscript writing. Ernest Nadal: Functional characterization of the CD74-NRG1 fusion protein, proliferation and invasion assays for BEAS-2B transformed cells with CD74-NRG1, clinical analysis, Western blots for JNK and ERBB3 activation. Manuscript writing. Guoan Chen: Tissue collection and PCR gene fusion validation. Matthew Iyer: RNASeq data processing. Dan Robinson: RNA Sequencing. Xuhong Cao: RNA Sequencing. David Beer: Experiment design and overall scientific project oversee. Arul M. Chinnayian: Experiment design and overall scientific project oversight.

Chapter 4
Antisense gene expression in human cancers: understanding cis-acting mechanisms of transcript regulation

4.1 Background

High throughput RNA sequencing has revealed widespread transcription in the human genome[153]. However, the extent to which both DNA strands (forward and reverse) are transcribed in regions of the genome with overlapping genes is less well characterized. This lack of understating is in part due to the fact that initial RNASeq protocols did not preserve the strand of the original RNA. Overlapping transcripts originating from the same locus of DNA but on opposite strands are known as sense-antisense transcript pairs (S-AS) and they have been described in eukaryotes and bacteria. Natural antisense transcripts (NATs) are transcribed from the opposite strand to that of the sense transcript of either protein-coding or non-protein-coding genes[154,155]. In this study, the originally annotated transcript will be considered as the sense transcript and the more recently identified one on the opposite strand as the antisense transcript, following Pelachano and Steinmetz (2013)[155].

4.1.1 Natural antisense transcripts classification

NATs may arise from independent transcriptional units including cryptic promoters situated within genes, typically in intronic regions, or near

transcriptional start sites of neighboring genes. According to the orientation of the transcripts involved, overlapping pairs are classified in three groups: head-to-head (HTH), tail-to-tail (TTT) and embedded (EMB) pairs. In the HTH group the 5'-region of both transcripts are overlapping, while the 3'-regions in the TTT group. In EMB pairs one of the transcripts is fully contained within the other (Figure 4.1). Intronic antisense transcripts are a special category as they overlap partially or totally the introns of another gene in the opposite strand. In general terms all pairs are referred as **cis-NAT pairs**.

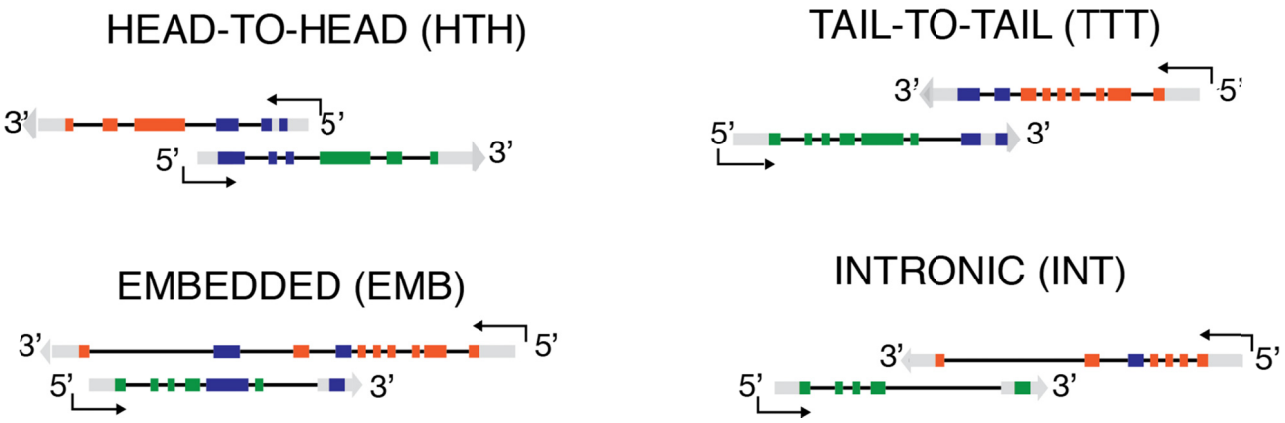


Figure 4.1. Schematic representing different types of cis-NAT pairs, according to the orientation of the overlapping genes.

4.1.2 Natural antisense transcripts function

The mechanisms of NAT functioning are just beginning to be elucidated. NATs can function locally or distally and in cis or in trans to regulate the expression of other genes. Cis-acting regulation happens when the regulatory element and the target gene are transcribed from the same locus. This regulation can occur locally (in the nucleus) or distally (in the cytoplasm). Local cis-regulation tends to involve epigenetic changes in the target gene, while distal cis-regulation involves RNA-RNA interactions between transcripts transcribed from the same locus. Trans-acting regulation occurs in the nucleus when antisense

transcripts affect the expression of genes transcribed from different loci[155] (Figure 4.2).

NATs cis-regulation can activate[156,157] or silence[158] the corresponding sense mRNAs. The mechanisms of this regulation are complex, diverse and it can be exerted at different levels including transcriptional activation[156,157] or silencing[159,160], mRNA stabilization[161,162], alternative splicing[163], and post-translational regulation among others.

Moreover, several small-scale studies suggest that dysregulation of antisense transcript expression play a role in diseases such as Alzheimer[164], Schizophrenia[165], Parkinson[166] and cancers[167-175], highlighting the relevance of antisense transcription in disease. Recently, Modarresi *et al.,* (2012)[158] used antago-NATs to demonstrate that inhibition of natural antisense transcripts *in vivo* caused gene specific up-regulation of the sense gene; providing a proof of principle for manipulating NATs in order to specifically regulate the gene expression of the cognate genes and opening the possibility for new therapeutic opportunities.

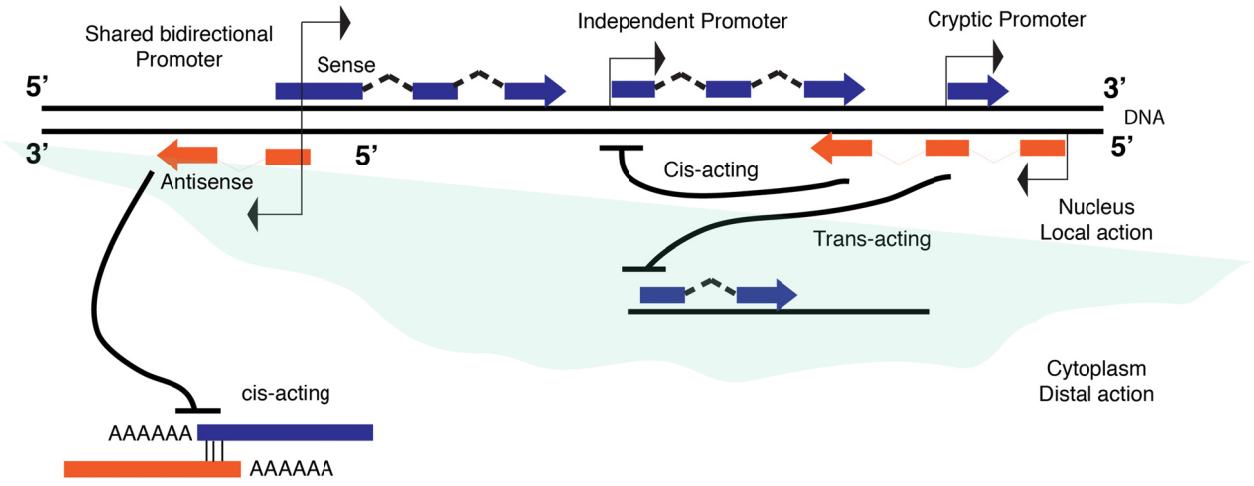


Figure 4.2. Schematic representation of cis-NAT mechanisms of action.

4.1.3 Antisense expression

Several attempts to describe the patterns and magnitude of antisense expression suggested that antisense expression might be far more extensive than previously anticipated; with 10-20% of loci exhibiting antisense expression in humans[176], though up to 72% in mice[177]. However, previous studies attempting to characterize the magnitude of antisense expression used methods such as ASSAGE[4] [176] or SAGESeq[5] for identifying NATs and quantifying antisense expression. These methods are intrinsically limited in their accuracy and coverage of the entire transcriptome, allowing only the assessment of a small fraction of the total number of genes and missing transcripts expressed at low levels. Due to these limitations and their laborious experimental protocols, those methods have been applied only to small datasets. This has limited our understanding of the landscape of antisense expression, the patterns of expression between overlapping transcripts and more importantly the role of antisense expression in cancer.

4.1.4 Strand specific RNA sequencing

RNA sequencing (RNASeq) has opened the way for a comprehensive analysis of the entire transcriptome. However, standard libraries for RNASeq do not preserve the information about which DNA strand was originally transcribed. This information is lost during the synthesis of randomly primed double stranded cDNA followed by addition of adaptors for next-generation sequencing[24]. In some cases, the strand is inferred by computational methods relying in known open reading frames and splice-site orientation in eukaryotic genomes, but these

[4] ASSAGE: asymmetric strand-specific analysis of gene expression
[5] SAGESeq: Serial analysis of gene expression coupled with sequencing

methods are not able to accurate resolve the expression of locus with overlapping genes.

Strand specific RNA sequencing (ssRNASeq) solves those problems by providing direct information on the DNA strand that was originality transcribed. Several methods for strand specific RNA library preparation has been proposed, but dUTP second strand labeling and the illumina RNA ligation are the leading protocols for ssRNASeq (see Chapter 1)[24]. ssRNASeq enhances the value of RNASeq experiments by allowing an accurate characterization of antisense transcripts, demarcation of the exact boundaries of adjacent genes in opposite strands and accurate resolving the expression of overlapping transcripts.

4.1.5 Aims of this study

In this study we use strand specific RNA paired-end sequencing (ssRNASeq) to comprehensively characterize the landscape of antisense expression. We applied ssRNASeq on a cohort of 376 patients including 9 different cancer tissue types, making this the biggest cohort of ssRNASeq data assembled so far. Our results reveal that greater than 60% of human annotated transcripts have measureable expression coming from the opposite strand of the DNA. We also demonstrate that cis-NAT gene pairs have in general a positive correlation between their levels of expression, and that this pattern is stronger for head-to-head overlapping pairs. Moreover, by analyzing CpG islands localization with respect to the regions of overlap between transcripts, we suggest that the high gene expression correlation of HTH pairs would reflect shared bidirectional promoters between the sense and antisense transcripts.

Furthermore, according to the expression across tissues, four groups of antisense loci were identified: tissue-specific, tissue-enriched/non-specific, ubiquitously and cancer specific (in tumor but not in normal samples). Finally, this

study creates and makes available a catalogue of tumor suppressors and oncogenes with significant antisense expression (oncoNATdb), which will allow cancer researchers to investigate the mechanisms of sense-antisense regulation in cancer.

4.2 Methods

4.2.1 Bio-repository description

The Michigan Center for Translational Pathology (MCTP) strand specific RNASeq repository included in this study has 376 samples. Most of the samples correspond to cancer tissues, being the largest tissue cohorts: breast, lung adenocarcinoma, lung squamous carcinoma, prostate cancer, ovarian cancer, pancreatic cancer, meningioma, rare cancers, and lung cell lines. Table 4.4-1 presents a break down of major and minor cohorts included in this study.

Table 4.4-1.Tissues and number of samples the MCTP ssRNASeq cohort.

Major Cohorts		
Tissue	**Abbreviation**	**Number of samples**
Breast cancer	BRCA	66
Lung adenocarcinoma	LUAD	66
Lung squamous carcinoma	LUSC	37
Lung cell lines	LUCL	31
Prostate cancer	PRCA	27
Ovarian cancer	OVARIAN	23
Pancreas cancer	PANC	17
Meningioma	MENINGIOMA	13
Rare cancers	RARE	39
Minor Cohorts		
Tissue	**Abbreviation**	**Number of samples**
Cholangiocarcinoma	CHOLANGIO	8
Lung large cell carcinoma	LULC	8
Merkel cell carcinoma	MERKEL	8
Lung match normals	LUNO	7
Sarcomas	SARCOMA	7
Osteosarcoma	OSTEOSARCOMA	5
Adrenocortical carinoma	ADRENOCORTICAL	4
Hodgkin's lymphoma	HODGKINS	4
Rhabdomyosarcoma	RHABDOMYOSARCOMA	3
Combined Cohort		**376**

4.2.2 Preparation of RNASeq libraries

Transcriptome libraries were prepared following a modified protocol previously described for generating strand specific RNASeq libraries[24]. Briefly 2.5 micrograms of total RNA was subjected to polyA selection using oligodT beads (Invitrogen, Carlsbad, CA). Purified polyA RNA was fragmented and reverse transcribed using SuperscriptII (Invitrogen, Carlsbad CA). Second strand synthesis was performed with DNA Polymerase I (New England Biolabs, Ipswich, MA) in presence of dNTP mix containing dUTP instead of dTTP. The product was then subjected to end repair, A base addition and adaptor ligation steps. Libraries were next size selected in the range of 350 bps after resolving in a 3%

Nusieve 3:1 (Lonza, Basel, Switzerland) agarose gel and DNA recovered using QIAEX II gel extraction reagent (Qiagen, Valencia, CA). Libraries were barcoded during the 14-cycle PCR amplification with Phusion DNA polymerase (New England Biolabs, Ipswich, MA) and purified using AMPure XP beads (Beckman Coulter, Brea, CA). Library quality was estimated with Agilent 2100 Bioanalyzer for size and concentrations. The paired end libraries were sequenced with Illumina HiSeq 2000 (2x100 bases, read length). Reads that passed the filters on Illumina BaseCall software were used for further analysis. Importantly, because of the nature of this protocol the second read in each pair is complementary to the original mRNA and therefore indicates what DNA strand was transcribed.

4.2.3 Sequence Alignment

Sequence alignment was performed using the Tuxedo pipeline: Bowtie2 (Bowtie2/2.0.2) and Tophat2 (TopHat/2.0.4)[47]. We supplied TopHat with the set of transcript models annotated in the Homo sapiens Ensembl database version 69. The option fr-firststrand was used for the strand specific RNASeq libraries while all other parameters were used with default values. When provided with ssRNASeq data TopHat2 annotates aligned reads with the tag XS indicating the strand of origin in the DNA.

4.2.4 Transcript summarization

We used Ensembl v69 as the reference transcriptome to reconstruct the longest annotation for each gene based on the transcript and exon information provided by this assembly. We only included transcript isoforms that satisfied the following criteria: gene and transcript biotypes were annotated with the same type; transcript isoform annotation level was manual or automatic followed by manual revision (annotation levels 1 or 2) and transcript isoform was not reported as a problematic in the Encode-Gencode attributes table (e.g: transcript biotype

is retained_intron, to be experimentally confirmed, or disruptive_domain). Moreover, each transcript isoform used for our gene models was annotated with their "isoform expression rank" across the tissue cohorts, and their support level (**tsl**) provided by the Genecode project. Tsl equal to 1 indicates that all splice junctions of that transcript are supported by at least one non-suspect mRNA; any other number suggests that the transcript is supported by suspicious ESTs. These final gene models were used as the reference loci, or features, for downstream analysis.

4.2.5 Strand specific expression

The final gene models in the summarized transcriptome obtained in 4.2.4 were used to compute strand specific expression. Paired-end reads mapping to the forward or opposite strand of a feature were counted in order to quantify the raw amount of forward and reverse transcription on a particular locus. In order to determine what DNA strand a read pair was originated from, we first used the reads' XS tag, provided by TopHat2, to identify the strand for each read in the pair. Then, we use the fact that in our ssRNASeq protocol the second read is complementary to original mRNA and therefore the second read has to be on the same strand than the feature, while the first read on the opposite strand. These criteria unambiguously define a read pair DNA strand of origin.

We discard all pairs in which one or both reads map to multiple locations in the genome, and all read pairs in which any of the reads was improperly mapped or did not have the XS flag provided by TopHat2 to indicate the strand of origin.

4.2.6 Read counts normalization

Read counts normalization was performed using DESeq[52], which models the read counts data using a negative binominal distribution and estimates the variance by modeling the sum of the shot or Poisson noise and the sample-to-sample variation. DESeq first estimates the effective library size, and then divide the counts by the effective library size in order to bring counts into a common scale. Given the size of our cohort, we used the following parameters to estimate the variance (dispersion): method="per-condition", sharingMode="gene-est-only", fitType="local". Normalized counts were used for all other downstream analysis.

4.2.7 ssRNASeq strand specificity estimation

ssRNASeq protocol's strand specificity is defined as the number or reads mapping to known transcribed regions at the expected strand. Assuming that most genes are transcribed in the sense direction, Levin *et al.,* (2010) measured the strand specificity or protocol error rate of a library, as the fraction of reads mapped to the opposite strand generated by the forward gene. This fraction constitutes a measure of the protocol error rate, ranging from 0.5 for the best method to 12% for the less specific one[24].

In order to determine the protocol error rate, we select loci that do not overlap any other transcripts in our reference transcriptome, and do not have any other neighboring gene within 20Kb in either side (3' or 5' ends). We reason that the fraction of reads mapped to the opposite strand of those loci would constitute an estimate of our ssRNASeq protocol error rate (*pe*). The average *pe* in our cohort of 376 samples is 0.64**%** (min=0.17%, max=0.69%, sd=0.0055), demonstrating the high strand specificity of our libraries.

4.2.8 Detection of transcripts with expression in both strands

In order to determine how many loci consistently express both forward and reverse strands, we first determined for each sample the protocol error rate as describe above as an estimate of the background noise in the opposite strand of a particular loci. For each sample, loci with read counts in both the forward and reverse strands, and opposite strand ratio (OPSratio= Reverse read counts/(Forward read counts + R read counts)) greater than the *pe* for the sample were considered as expressing both strands. Next, we leveraged the size of our full cohort (n=376) aiming at identifying those loci that are expressing both strands consistently across multiple samples. We reason that recurrent expression is an indicator of genuine transcripts; therefore antisense transcript expressed above the protocol error rate in different samples across the cohort would have a higher chance of being genuine. We defined a locus as having measurable antisense expression if that locus has OPSratio > pe in at least 5% (n=20) of the cohort samples.

4.2.9 Antisense loci identification

We used a probabilistic method for natural antisense transcripts identification using RNASeq (NASTI-seq[178]). That method incorporates the protocol error rate (*pe*) of the ssRNASeq procedure and employs a model comparison framework to identify loci with significant antisense expression. Briefly, for each locus in a reference transcriptome the method first calculates the probability of the observed read count data under a sense only model, in which it is assumed that the sense gene is the only one being expressed and the reads mapped to the opposite strand are due to the *pe*. Then, the method calculates the probability of the observed data under a second model, antisense model, in which the reads mapping to the opposite strand of a particular locus in the genome come from two different sources: *pe* and the *bona fide* expression of an antisense transcript overlapping the locus. The NASTIscore, a type of Bayesian information criteria (BIC) score is calculated to determine what model fits best the

data. Finally, a training dataset of true positives and true negative pairs was used to determine the minimum NASTIscore score that distinguishes between potentially true antisense loci from the background noise.

We build the training datasets as follows. First, we determine all overlapping pairs in our reference transcriptome. Then, we annotated each pair according to the Ensembl gene biotype of the genes involved and the length of the overlapping regions. The true positive pairs dataset was conformed by all pairs of genes, with an exonic overlapping greater than one base pair and involving a protein coding gene and a known antisense transcript. The true negative dataset was conformed by all protein coding genes that do not overlap exons or introns of other genes, the closest neighboring gene in either direction is more than 20 Kb away and the mean number of reads mapping to the opposite strand of that gene is less than 50.

4.2.10 Identification of lineage- and cancer-specific antisense loci

We computed the NASTIscore for all loci with expression in both strands in at least one sample of a tissue cohort. The NASTIscore calculation was performed for each tissue type independently. Then loci with significant antisense expression, as determined by a NASTIscore greater than the minimum NASTIscore for each cohort (see 4.2.9), were aggregated. These loci will be denoted as antisense loci (ASloci) from here on. Furthermore, we defined three groups of antisense loci according to their presence in different cohorts. Antisense loci identified in all cohorts were termed ubiquitous ASloci, while ASloci identified in only one tissue type were name tissue or lineage specific. We also observed ASloci that were expressed in more than two tissue types but not in all and we treated this group as tissue enriched or non-specific ASloci.

Because our compendium is substantially enriched for cancer samples and all benign samples correspond to match normal of lung adenocarcinoma

(LUAD) and squamous carcinoma (LUSC) patients, we identified lung cancer specific ASloci, that is, ASloci identified as such in the LUAD and LUSC tumor cohorts but not in the benign cohort. We further determined what ASloci were detected in our cohort of 31 lung cell lines and which ones were also identified in other cancers. The first of these groups correspond to lung cancer specific ASloci while the second one to cancer non-specific ASloci.

4.2.11 Correlation between sense and antisense transcripts

In order to determine the patterns of expression between genes of a cis-NAT pair, we calculated different measures of correlation between the expression of the sense and antisense genes. The measures of correlation calculated were the Spearman correlation coefficient, the Pearson correlation coefficient, the coefficient of robust correlation and the mutual information. All these metrics produced very similar correlation results; therefore we chose the Spearman correlation for further analyses. The correlation was computed independently for each cis-NAT pair in each tissue cohort, as well as, across the combined cohort of 376 samples. The statistical significance of each correlation was corrected for multiple hypotheses testing using the Hochberg's procedure.

We compute the null or random distribution, as the distribution of correlations between any two random genes. Similar to the cis-NAT calculation, we compute the null distribution for each tissue type and across our combined cohort.

4.2.12 CpG islands analysis

CpG islands have been found in 30 to 60% of unidirectional and 80% to 95% of bidirectional promoters[179]. Bidirectional promoters refer to intergenic sequence between the transcription start sites of bidirectional genes pairs. Bidirectional gene pairs, in turn, are defined as two non-overlapping genes

arranged in head-to-head configuration and separated by less than 1000 bp[180]. Therefore, we used the presence or absence of CpG islands in a genomic region as proxy for the presence or absence of a gene promoter in that region.

In order to determine whether overlapping regions of head-to-head cis-NAT pairs harbor potential bidirectional promoters, we downloaded, from UCSC genome browser, tracks providing CpG island strength predictions, and mapping of *bona fide* CpG islands for the human genome hg19. These tracks are based on large-scale epigenome predictions described by [181]. Next, for each pair of cis-NAT genes, we defined the DNA regions of overlap between those genes and then tabulated how many CpG islands are found within that overlapping region. We reason that if a gene promoter exists within the overlapping regions of cis-NAT pairs, we should observe an enrichment of CpG islands in those regions. As a positive control we identified a set of bidirectional protein coding gene pairs using the definition presented above and including only gene pairs with gene expression correlation greater than 0.2 across our cohort of 376 samples. A mean correlation of 0.2 between the expressions of bidirectional genes was previously described[180] and it is confirmed in by our own analyses.

4.2.13 Differential expression analysis of sense/antisense pairs

DESeq normalized read counts as described in 4.2.6 were used for differential expression analysis between lung adenocarcinoma and lung squamous tumor samples and their match normal samples. We reasoned that the forward and reverse expression of a particular locus could change in a consistent or inconsistent fashion between tumor and normal samples. In a consistent change, the expression of forward and opposite strands will be over or under expressed between tumor and normal samples. On the other hand in an inconsistent change, the expression of forward and reverse strands will change in opposite directions between tumor and normal samples. Therefore, when the

forward strand is over expressed the opposite strand will be under-expressed and vice versa, suggesting potential mechanisms of interference between sense and antisense genes.

In order to identify loci with consistent or inconsistent changes in the expression of forward and reverse strands, we used a negative binomial test as described by Anders *et al*., (2010)[52] to determine differential expressed cis-NAT pairs between tumor and normal samples. We first identified cis-NAT pairs for which both sense and antisense genes were differentially expressed with adjusted p-value <= 0.1. Then we defined a log fold change threshold (lfcth) of 1 and select as consistent pairs differentially expressed cis-NAT pairs for which the absolute log fold change expression of sense and antisense genes were >= lfcth. Inconsistent cis-NAT pairs were defined as differentially expressed pairs for which the log fold change expression of the sense gene was >= lfcth, while the expression of antisense gene <= -1* lfcth, or vice versa.

4.3 Results

4.3.1 Development of a bioinformatics analysis workflow for antisense transcript analysis

Strand specific RNA paired sequencing (ssRNASeq) data from a compendium of 376 samples (303 tissue and 69 cell lines samples), representing both cancer and benign from 9 different tissue types recently generated for our laboratory, was used to develop a bioinformatics workflow for the analysis and characterization of antisense expression in human cancers (Figure 4.1, Methods).

First, sequencing reads were mapped to the human genome (hg19, GRh37) using TopHat2 (TopHat/2.0.4)[47]. Then, a summarized transcriptome was build by reconstructing the longest annotation for each gene, using transcript and

exon information provided in the Ensembl.v69 assembly. Only high quality transcript isoforms were included, while problematic and miss-annotated transcripts were filtered out (see Methods). This procedure generated 42,129 gene models. Second, these gene models were used as reference loci to compute the number of strand specific pair-end reads mapping to the forward or reverse strand of each locus; and then to calculate the expression level of each strand in that locus (see Methods). Loci expression was then normalized using DESeq[52]. Third, strand specificity was calculated for each library in order to determine the protocol error or background noise affecting our estimation of the expression coming from the opposite strand (see Methods). Fourth, loci consistently expressing both, forward and reverse, strands across our cohort were identified. Moreover, a locus that has OPSratio > pe in at least 5% (n=20) of the cohort samples (Methods) was considered as a locus with measurable antisense expression. Fifth, a probabilistic method was used for natural antisense transcripts identification using RNASeq (NASTI-seq[178]). This method accounts for the variable protocol error in order to identify loci with significant antisense expression (Methods).

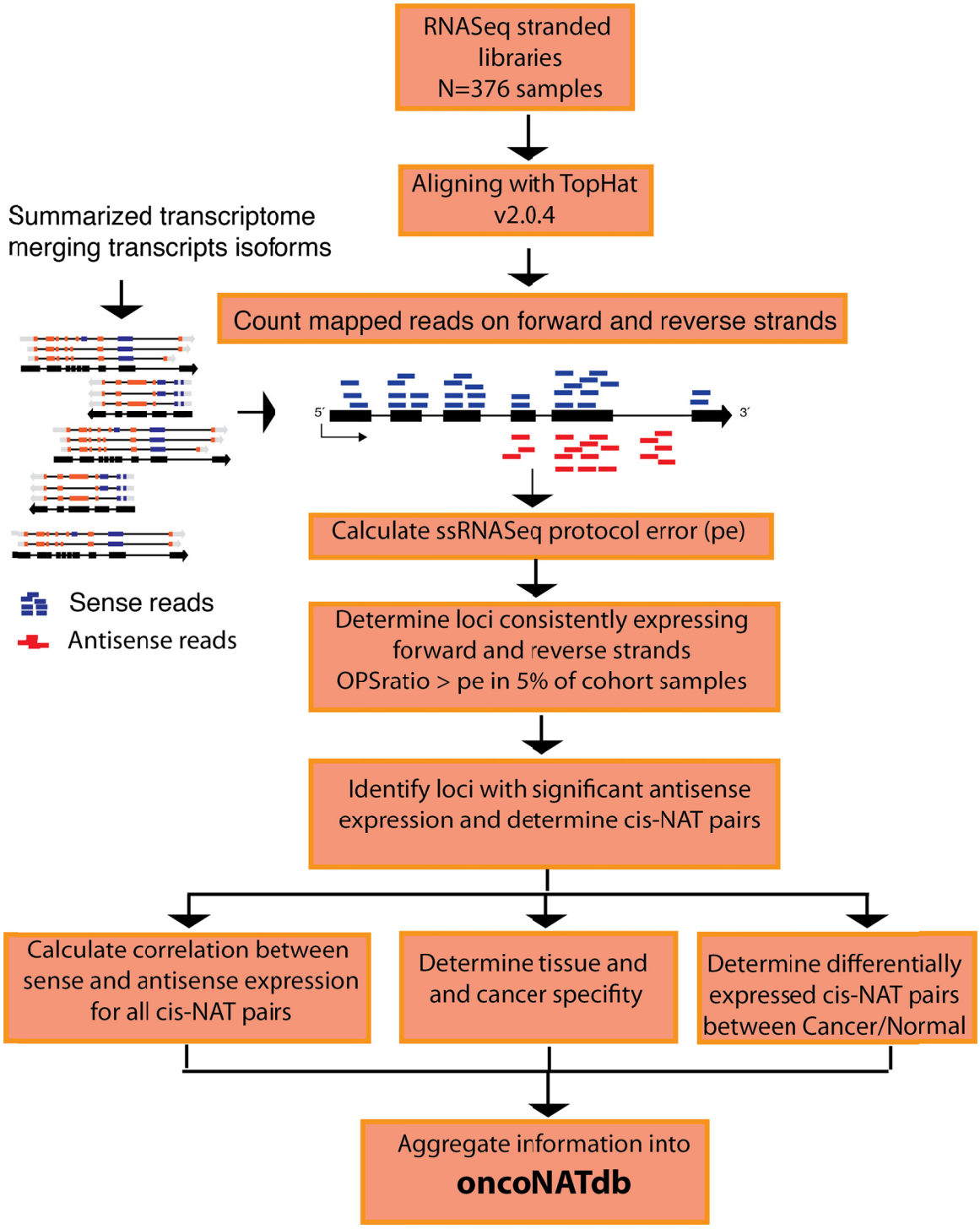


Figure 4.3. Bioinformatics workflow for characterization of Antisense loci.

Finally, we calculated the correlation between sense and antisense transcripts forming cis-NAT pairs and determined tissue specific, tissue-enriched/non-specific, ubiquitous and cancer specific antisense loci. Taken together this bioinformatics pipeline nominates expressed antisense loci across 9 different tissue types and establishes their pattern of expression. The pipeline further aggregates tumor suppressor and oncogenes with significant antisense expression in a single catalogue, oncoNATdb.

4.3.2 Antisense expression is pervasive across the human transcriptome.

Figure 4.4a and Figure 4.4b show that in any given locus most of the observed expression originates on the annotated or forward strand; the expression originating from the opposite or reverse strand is overall two to three orders of magnitude lower (median of reverse/forward = 0.001). Accurate quantification of strand-specific expression is further complicated by the ssRNASeq protocol error (*pe*)[24]. To address this, we calculated the protocol error for each of our samples and then determined the fraction of the transcriptome with measurable expression in the opposite strand. *pe* ranges from 0.5 for good ssRNASeq libraries to 12% for the less specific ones[24]. The average *pe* in our cohort of 376 samples is 0.64**%** (min=0.17%, max=0.69%, sd=0.0055), which indicates a high strand specificity of our libraries (Methods) and supports the use of these data for identifying loci harboring expression of both strands. We defined a locus as having measurable antisense expression if that locus has opposite strand ratio (OPSratio) greater than *pe* in at least 5% (n=20) of the cohort samples.

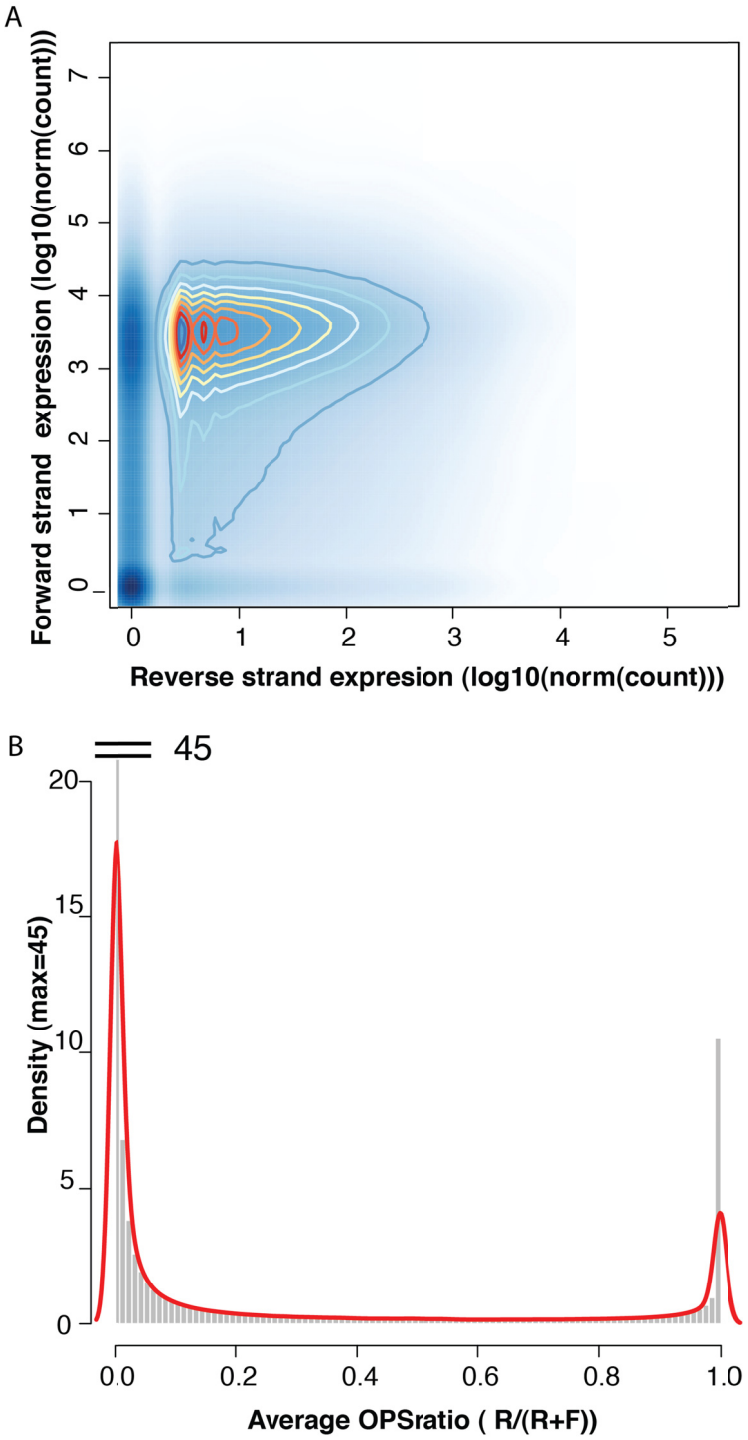


Figure 4.4. Forward and reverse expression.

A) Density plot of the expression over the forward and reverse strand across all samples in our cohort. On any given locus most of the expression is coming from the annotated or forward strand, and the expression coming from the opposite strand is in general two orders of magnitude lower (median of reverse/forward = 0.001). **B**) Average OPSratio density. R=reverse strand, F=forward strand. Loci expressing predominantly the forward strand have OPS ratio close to 0, while loci expressing mainly the reverse strand have OPSratio close to 1.

Our results indicate that greater than 60% (mean=62.41%, n=26,289) of annotated gene loci consistently express the reverse strand (Figure 4.5). Here, we observed a similar number regardless of the tissue of origin (mean=61.81%, sd=0.07, Figure 4.6). These results reveal that the human genome is pervasively transcribed in both strands.

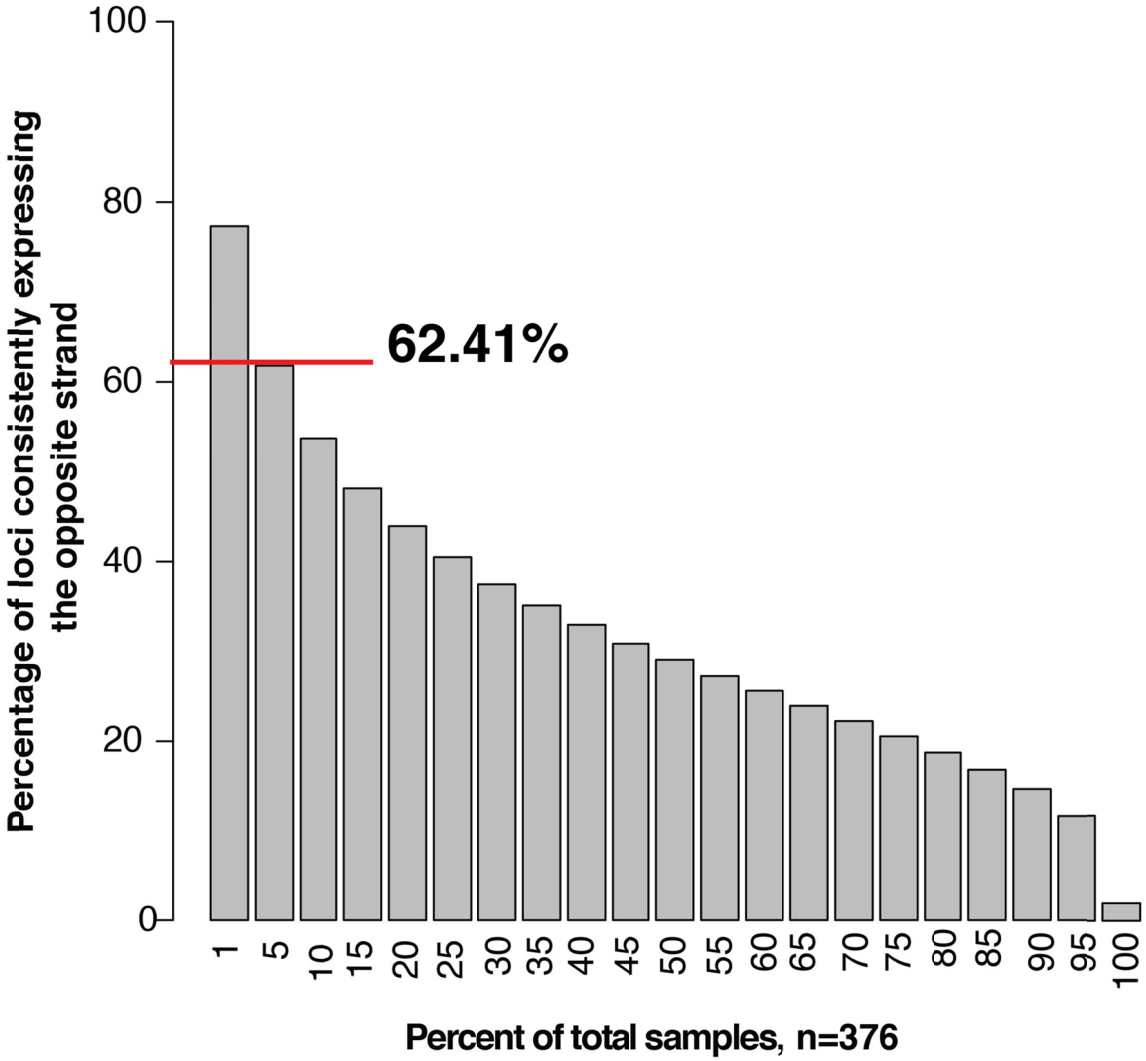


Figure 4.5. Percentage of loci with measurable expression in the opposite strand.

A loci is considered to be measurable if it express the opposite strand above the protocol error rate in at least 5% of the samples in the cohort (n=20).

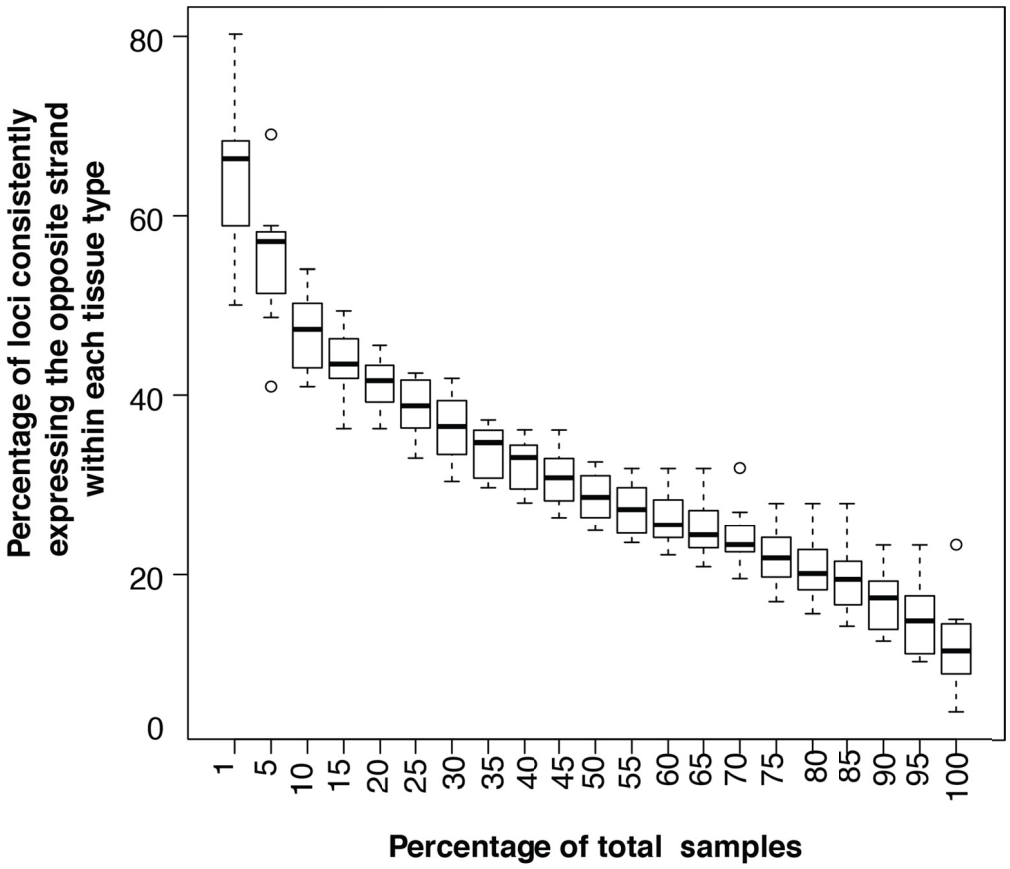


Figure 4.6. The percentage of loci with measurable expression in the opposite strand tissue distribution.

The percentage of loci with measurable expression in the opposite strand is similar regardless of the tissue of origin (mean=61.81%, sd=0.0693, Supplementary Figure 2). Dots inside the boxplot represent the percentage of loci consistently expressing the opposite strand in the indicated fraction of each tissue cohort independently.

In prior work by He *et al*., 2008, Katayama *et al*., 2005 and Maruyama *et al*., 2012[177,182,183], the OPSratio was a useful tool in order to classify a DNA locus according to the relative contribution of each strand to the total expression observed in the locus. They defined three main categories, forward loci with OPSratio <= 20%, reverse loci with OPSratio >=80% and forward-reverse loci with OPSratio values between 20% and 80% (Figure 4.4b supports this classification). Figure C.1a shows the OPSratio distribution across a wide range of expression values in our study, supporting the groups defined by those

authors. Figure C.1b indicates that forward loci would represent 87% of transcripts, reverse loci 4%, and forward-reverse loci 9% respectively.

Although useful as an initial assessment, thresholding approaches, which use the OPSratio or the minimum number of reads in the opposite strand[176,184-186], as the only criteria to determine antisense loci, can be affected by biological variation, library size differences and the efficiency of the strand specific libraries. Thus, these approaches introduce error into the identification of antisense loci. More importantly, these methods do not account for the generally two-orders-of-magnitude lower expression of reverse strand compared to the forward strand (Figure 4.4) Therefore, using only the OPSratio likely underestimates the number of antisense loci; missing loci with significant antisense expression, especially in those cases in which the sense strand is expressed at intermediate or high levels.

In order to overcome these limitations, we used a probabilistic method for natural antisense transcript's identification using RNASeq (NASTI-seq[178]) that incorporates the variable *pe* of ssRNASeq protocols and employs a model comparison framework to identify loci with significant antisense expression (Methods). Briefly, for each locus in a reference transcriptome the method calculates both the probability of the observed read count data under a sense only model and an antisense model. In the first model, reads mapped to the opposite strand are due to the *pe* only, while in the second one reads mapping to the opposite strand of a particular locus come from two different sources: the *pe* and the *bona fide* expression of an antisense transcript overlapping the locus. Therefore, an antisense locus is defined as a region of DNA in which the antisense model explains better than the sense only model the read count data observed over that region (Methods).

Out of all transcribed loci consistently expressing the reverse strand across our entire cohort, an average of 6398 (sd=1019.30) genes were identified

as *bona fide* antisense loci for each tissue type (Figure 4.7a). This corresponds to 36.82% (26.77%-44.33%) out of all loci with measurable expression in the opposite strand. Out of the total number of antisense loci predicted by NASTI-seq across all tissue types (n=11773), 71.38% (n=8403) loci correspond to predicted *cis*-NAT overlapping pairs of genes based on the reference transcriptome (Methods), while 28.62% (n=3370) are potentially new antisense loci . Out of these 3370, 526 have a gene neighbor within 500bp of either side (Figure 4.7B), representing potentially cis-NAT pairs in which the gene annotation is shorter than what is observed from sequencing the data. The remaining 2844 loci could represent un-annotated novel overlapping antisense transcripts.

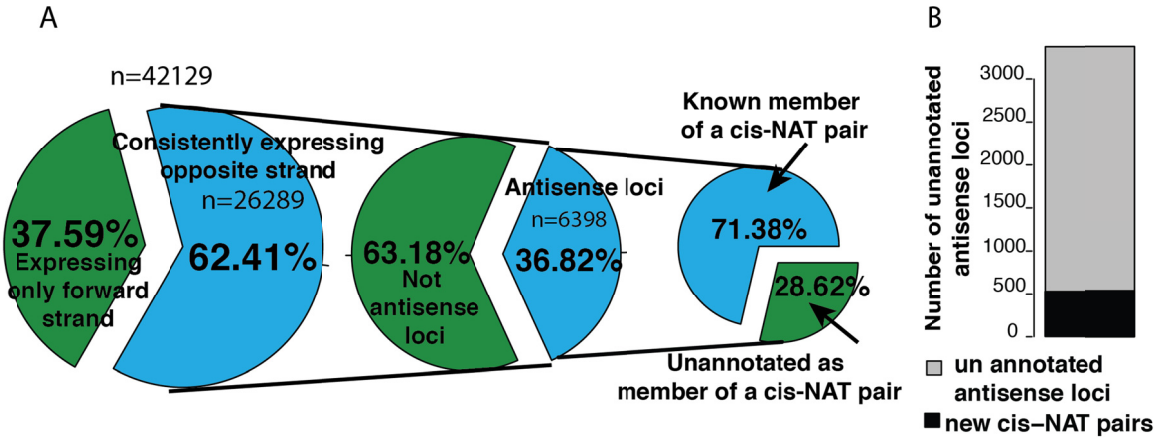


Figure 4.7. Landscape of antisense expression.

A) The percentage of loci with measurable expression in the opposite strand tissue distribution. The percentage of *bona fide* antisense loci was 36.82% (26.77%-44.33%) on average across all cohorts. *Bona fide* antisense loci are loci with measurable and significantly high (as determined by the NASTISeq statistical modeling) expression in the opposite strand. **B**) Previously un-annotated antisense loci. We report ~3500 new potential antisense loci.

4.3.3 Widespread positive correlation between sense and antisense transcripts

Although limited and anecdotal evidence suggests the co-expression of sense and antisense gene pairs, previous studies have been limited by the size

of the cohorts and the inaccuracy and low coverage of previous technologies. We therefore systematically characterize the relationship between sense and antisense expression across our cohort of 376 samples. Figure 4.8a shows that the expression of overlapping genes is in general positively correlated and greater than what would be expected by chance, with a median Spearman coefficient correlation of 0.27. Interestingly, cis-NAT gene pairs with a Head-to-Head (HTH) orientation showed the strongest positive correlation of their gene expression (Figure 4.8b, median R=0.40); while gene pairs in tail-to-tail (TTT), embedded (EMD) and intronic (INTRONIC) configurations had very similar and weaker levels of expression correlation (median R=0,23; 0,22 and 0,26 respectively)(Table C-1).

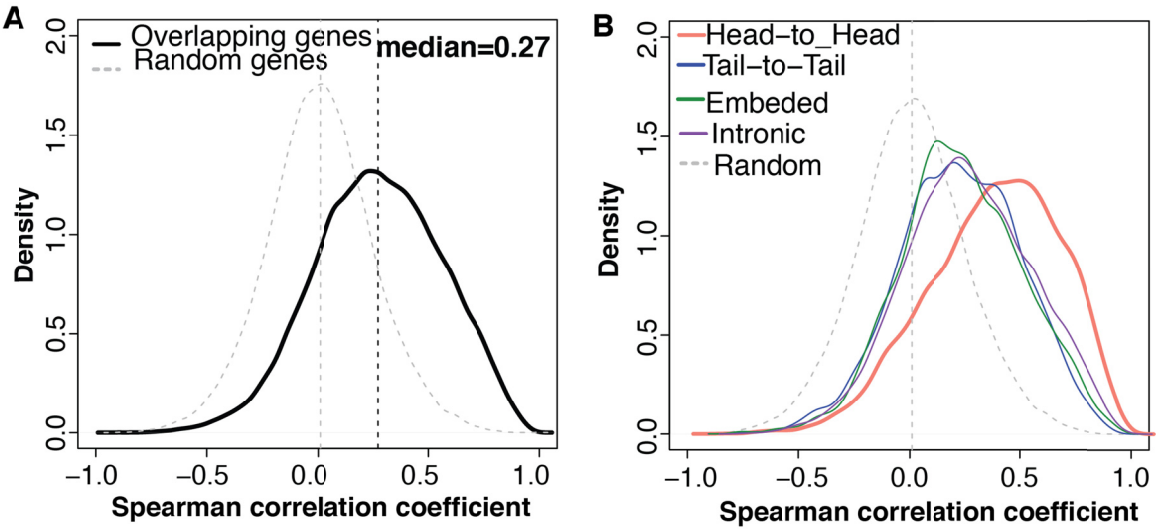


Figure 4.8. cis-NAT pairs genes expression is highly correlated.

A) Sense and antisense expression of overlapping genes is mainly positive correlated for all overlapping types (median R=0.27). Correlation between random pairs of genes is shown in the gray dashed line. **B)** Head-to-Head cis-NAT gene pairs show the highest positive correlation among all overlapping types (median R=0.4). Correlation between random pairs of genes is shown in the gray dashed line.

These observations hold across different tissue types (Figure 4.9), suggesting that a common mechanism may be responsible for the degree of co-expression observed between sense and antisense genes.

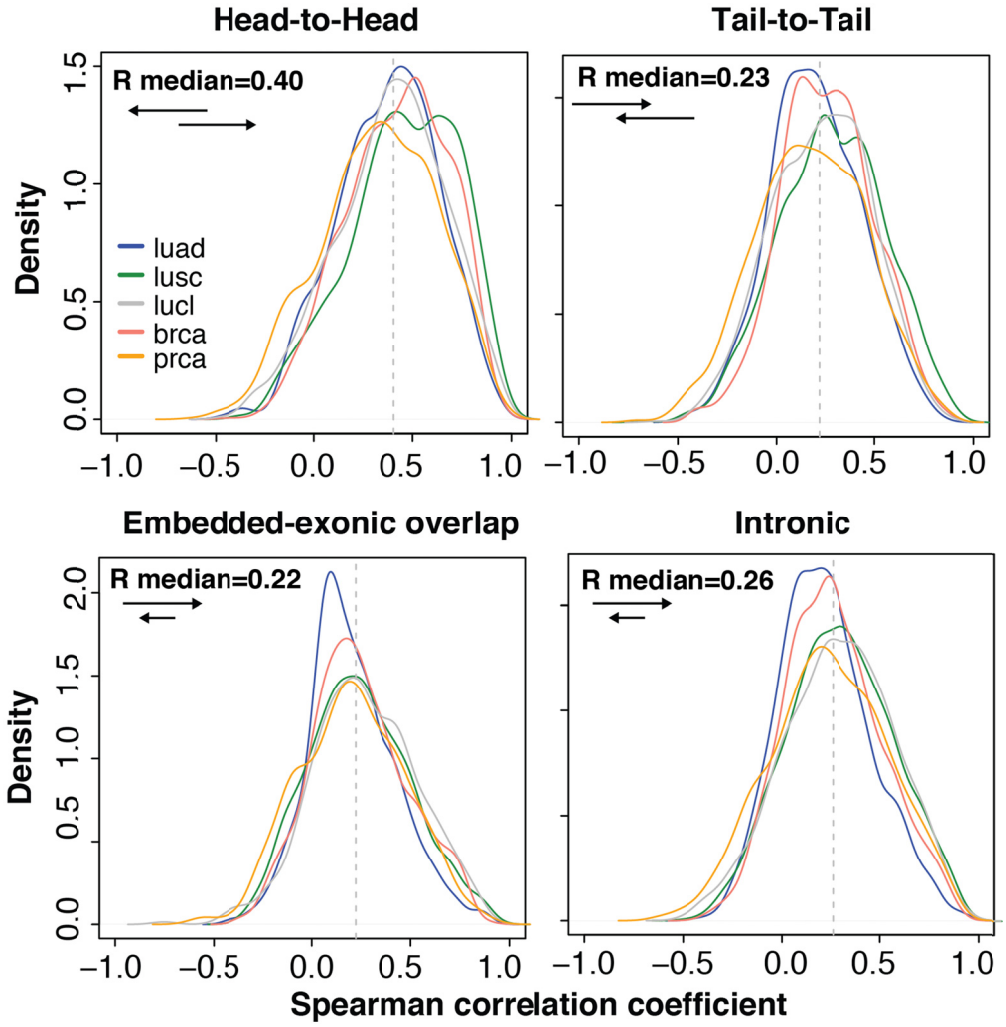


Figure 4.9. cis-NAT pairs genes expression correlation across different tissue types.

Co-expression of sense and antisense gene pairs have been experimentally demonstrated for WT1/WT-AS1[187], TP53/WRAP53[161], WDR83/DHPS[162], CDKN2A/CDKN2A-AS[188] among others; while an anticorrelation was recently suggested for BDNF/BDNF-AS[158]. Our results recapitulated those well-studied cases (Figure 4.10) as well as generate novel examples of highly co-expressed cis-NAT gene pairs.

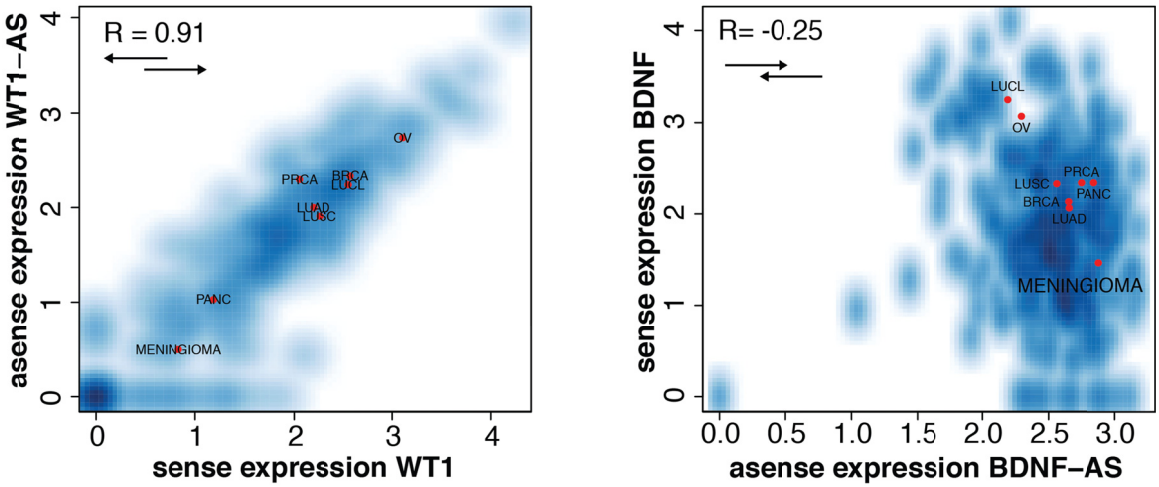


Figure 4.10. cis-NAT gene pairs previously reported in which the antisense regulates the cognate sense gene.

Density scatter plots for as WT1/WT1-AS (left) and BDNF/BDNF-AS (right). Two previously reported examples of cis-NAT pairs with positive (left) and negative (right) correlation. Red points indicate the average value for the each cohort.

We focused our attention on the cis-NAT pair formed by NKX2-1 homeobox 1 and NKX2-1-AS, because this pair has one the highest gene expression correlation in our dataset and the potential interaction between NKX2-1 and NKX2-1-AS has not been described yet. These genes show a positive correlation between their expression of R=0.94 across all tissues (Figure 4.11a), with lung adenocarcinomas and lung squamous carcinomas as the linages with highest expression of both genes across our combined cohort. Interestingly, this pattern is clearly observed in the ssRNASeq lung cell lines data (Figure 4.11a inset), allowing us to further validate this observation and credential our bioinformatics analysis. We confirmed this positive correlation in expression in a panel of 29 lung cell lines using qRT-PCR(Figure 4.11b).

Remarkably, NKX2-1, the thyroid transcription factor 1 (also known as TTF-1) is master regulator essential for lung development and peripheral lung cells known as terminal respiratory units (TRU) and it plays a role as a “linage-survival” oncogene in lung adenocarcinomas [189]. Moreover, NKX2-1 expression

is inversely associated with p53 and KRAS mutations, and NKX2-1 positive lung adenocarcinomas are dependent on sustained expression of this gene for survival[190]. The strong co-expression between NKX2-1 and NKX2-1-AS suggests NKX2-1-AS potential for regulating NKX2-1, and demonstrates the utility of our antisense compendia for uncovering potentially new aspects of tumor biology.

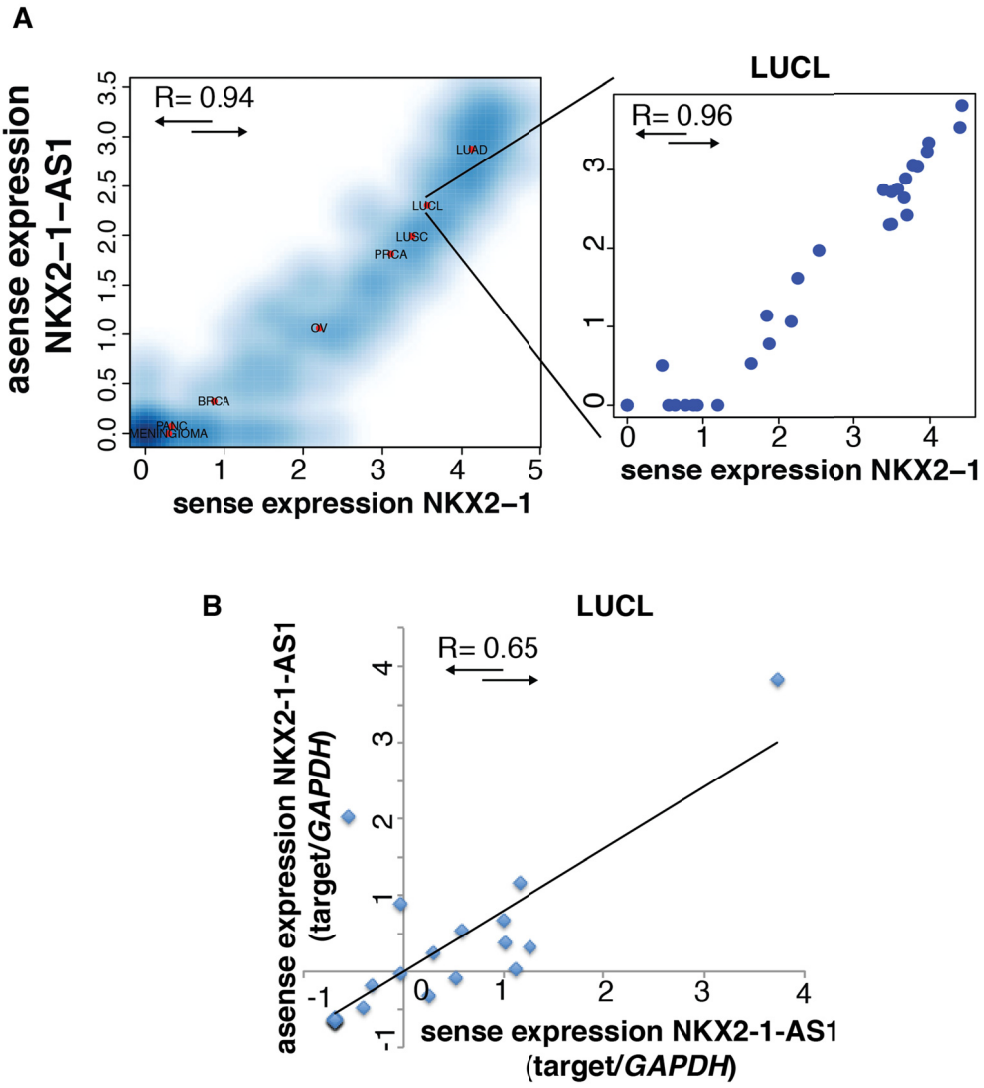


Figure 4.11. Novel cis-NAT gene pairs with high gene expression correlation.

A) Density scatter plot for NKX2-1 and NKX2-1-AS. This is a novel example of highly and positively correlated cis-NAT pair. The thyroid transcription factor 1 (also known as TTF-1) plays a role as a "linage-survival" oncogene in lung adenocarcinomas[189]. In the inset, a scatter plot showing only the sense and antisense expression for the samples in the lung cell line cohort. **B**) Quantitative PCR validating the expression of NKX2-1 and NKX2-1-AS across a lung cell line cohort of 29 samples.

4.3.4 Bidirectional promoters would direct the expression of head-to-head cis-NAT pairs

As shown in Figure 4.8b, HTH cis-NAT pairs have the highest positive gene expression correlation and this expression pattern is observed across different tissues types, suggesting that a common structural mechanism coordinates the expression of both genes in the pair. Similarly, co-expression patterns have also been observed for divergent but not overlapping genes driven by bidirectional promoters[180]. Inspecting the structural properties of the HTH cis-NAT pairs, we realized that close to 60% of those pairs involve overlapping regions between the 5'UTR (5UTR-5UTR) regions of each gene or the 5'UTR and the first exon (5UTR-exon, specially between protein coding and ncRNAs where UTRs are not defined). Taken all these together, we hypothesize that HTH cis-NAT pairs may share bidirectional like promoters that direct the concerted expression of both genes in the pair.

Bidirectional promoters are genomic regions that initiate transcription in both directions[191]. In metazoans, bidirectional promoters have typically been associated with the intergenic sequence between the transcription start sites of two non-overlapping genes arranged in divergent orientation and separated by less than a 1000 bp[180]. Recent studies have estimated that about 10% of protein-coding genes would share a bidirectional promoter[179]. Bidirectional promoters are CG rich and CpG islands are present in 80% to >95% of bidirectional promoters, while only present in 30 to 60% of unidirectional promoters[192]. Other marks of active transcription such as RNA polymerase II occupancy and modified histones H3K4me2, H3K4me3, and H3K9ac have also been observed in bidirectional promoters[179]. Because of their high association with bidirectional promoters, we used the presence or absence of CpG islands in a genomic region as proxy for the presence or absence of a gene promoter in that region.

We reasoned that if a bidirectional promoter exists between the genes of a HTH cis-NAT pair, then a CpG island should be present in the overlapping region between those two genes (Figure 4.12a). We found that 78% of the HTH cis-NAT pairs have CpG islands in the region of overlap between the two the genes. This percentage increases to 85% when only pairs involving HTH genes overlapping in the 5UTRs (5UTR-exon) regions are considered (Figure 4.12b). Similarly, we found that 83% of bidirectional but not overlapping protein-coding genes had CpG islands in their promoters. In contrast to these, we only observed up to 25% of CpG islands in the overlapping regions of cis-NAT pairs with tail-to-tail (TTT) or embedded (EMB) configurations (Figure 4.12b). Taken together, those results support our hypothesis that shared bidirectional promoters between HTH overlapping genes direct the tight co-expression of these cis-NAT pairs.

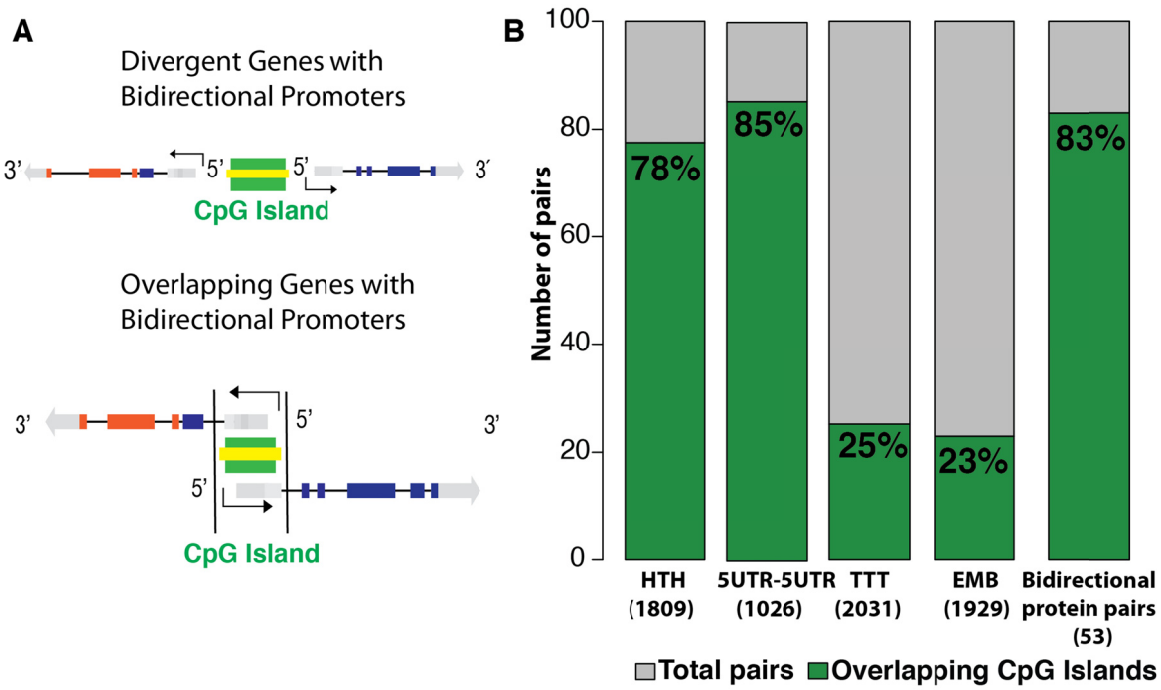


Figure 4.12. Bidirectional promoters direct transcription of Head-to-Head cis-NAT gene pairs.

A) Schematics representing a bidirectional promoter. CpG islands (green) are present in 80% - 95% of bidirectional promoters (yellow) that initiate transcription of not overlapping protein-coding genes in opposite directions (Top). Schematics representing a HTH cis-NAT gene pair sharing a bidirectional promoter in the overlapping region (Bottom). **B**) Number of gene pairs with CpG islands found in their overlapping regions. >78% of HTH cis-NAT present CpG islands in their overlapping region.

A representative example of the effect a bidirectional promoter has over the co-expression of cis-NAT gene pairs is shown in Figure 4.13, which illustrates 3 neighboring genes, HOXD3, HOXD-AS1 and HOXD1. On one hand, HOXD1 and HOXD-AS1 form a HTH cis-NAT pair, which is supported by CpG islands as well as H3K27Ac marks in the overlapping region between them. On the other, HOXD3 and HOXD-AS1 are arranged in a TTT orientation and do not have CpG islands or any other marks of active transcription in their overlapping region. Remarkably, although these three transcripts are very close to each other, within a 30Kb region, the genes expression correlation was R=0.90 (p-value<=2.2e-16) for HOXD1 and HOXD-AS1, while only R=0.36 (p-value=1.88e-10) for HOXD3 and HOXD-AS1, suggesting stronger transcriptional co-regulation between HOXD1 and HOXD-AS1 than between HOXD3 and HOXD-AS1.

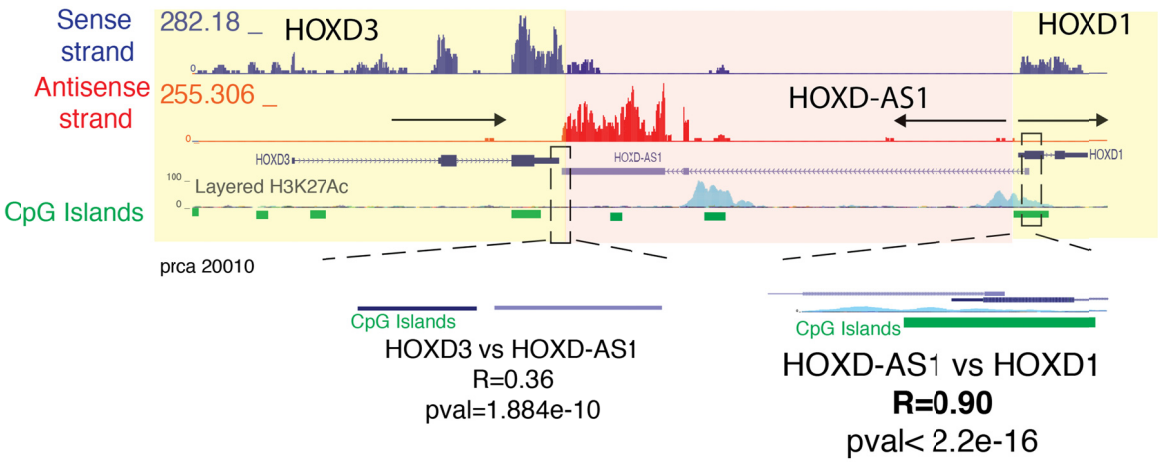


Figure 4.13. Example of cis-NAT with bidirectional promoter.

This example shows how bidirectional promoters would determine high correlation between HTH cis-NAT pairs, even when other close neighbors exist in different orientation. HOXD1/HOXD-AS1 HTH gene pairs would share a bidirectional promoter, as there is an enrichment in CpG islands in their overlapping region. The correlation in their gene expression is greater than the correlation observed for the HOXD3-HOXD-AS1 TTT gene pair.

Of note, recent studies have shown that promoters of protein coding genes can initiate transcription in both directions, generating non-coding RNAs transcripts, such as paRNAs (promotor associated RNAs) and TSSaRNAs

(transcription start site associate RNAs) from the opposite strand[154,193]. Increasing evidence also suggests that these non-coding transcripts influence the expression of their cognate protein coding genes through multiple mechanisms that are still incompletely understood[191]. For NATs in particular, several examples of highly co-expressed HTH cic-NAT pairs have been reported in which the antisense gene regulates the expression of their sense counterpart. Sessa *et al.,* 2007[156] demonstrated that the expression of antisense transcripts to human HOXA genes promotes a collinear activation of the corresponding cognate HOXA sense genes. Extending these observations, Zhang *et al.,* 2009[157] showed that HOTAIRM1, a non-coding RNA, was co-expressed with the HOXA gene locus and HOTAIRM1 positively regulated the expression levels of the HOXA gene.

Our results give additional support to these observations regarding the pattern of expression of HOXA genes and their respective antisense genes (Figure 4.14a, Figure C.2). Moreover based on the similarities with those examples, both in the genomic structure and the high genes expression correlation, we illustrate several representative examples of other gene pairs in the HOXD (chr2), HOXC (chr12) and HOXB (chr17) clusters that exhibit similar co-expression patterns to the one described for the HOXA (chr7) cluster (Figure 4.14b, Figure C.2a,b). These data suggests that a similar regulation mechanism between sense and antisense transcripts could exist in those other clusters.

Importantly, these regulation patterns are not restricted to homeotic genes, as our results also nominate other known examples such as WT1/WT1-AS (Figure 4.10a), as well as novel cis-NAT pairs with the characteristics described above and functions as diverse as cell adhesion and migration (BVES) (Figure 4.14b), Ras guanine nucleotide-releasing factors (RASGRF2) (Figure C.2c), transmembrane proteins (TMEM220, TMEM176B, TMEM176A) and transcription factors (NKX2-1, WT1, TBX5, HAND2, FOXD3) among others (Figure 4.14b). Similarly to NKX2-1, we validated the positive correlation between

sense and antisense transcripts for representative cis-NAT pairs HOXD1/HOXD1-AS1, HOXC10/HOX10-AS3 and BVES/BVES-AS (Figure 4.15).

Taken together, this study thus characterizes the landscape of antisense expression in the human transcriptome and the genes expression patterns of different cis-NAT pairs.

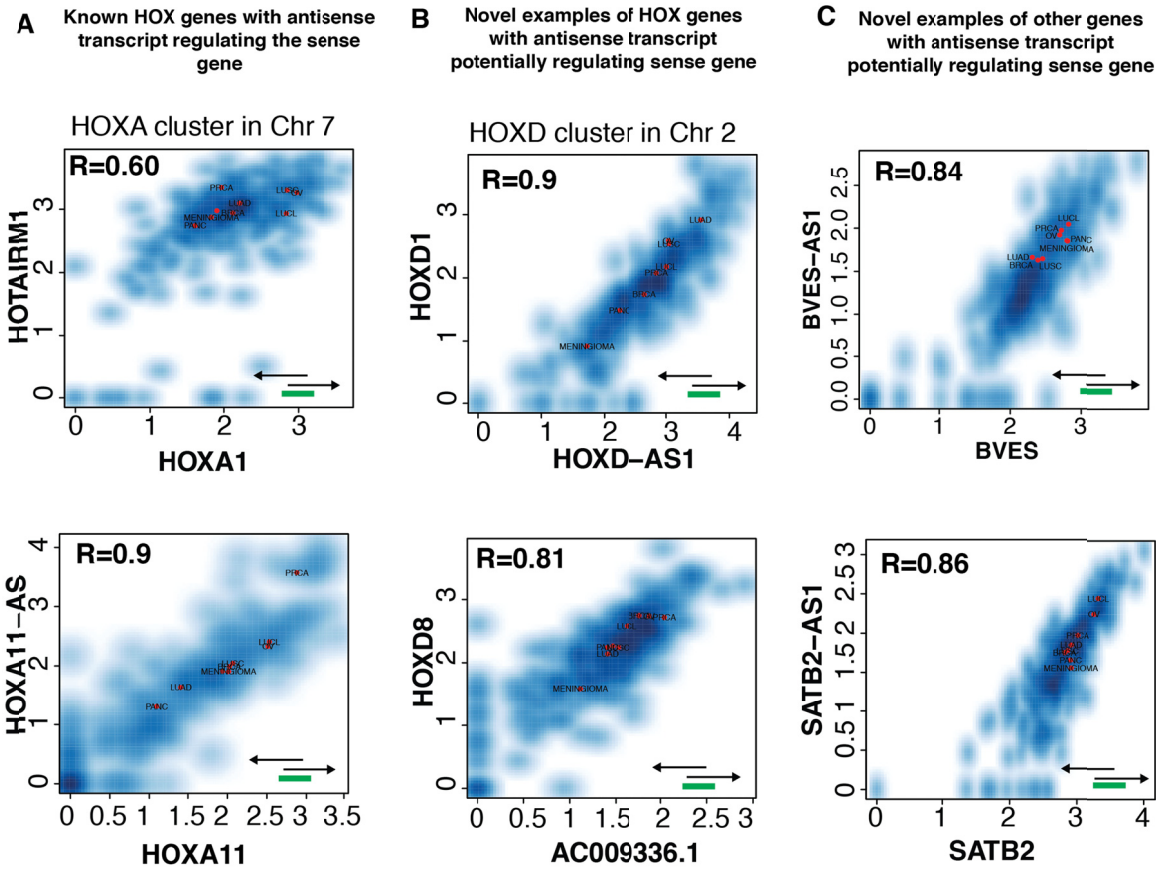


Figure 4.14. Known and novel examples of head-to-head cis-NAT with bidirectional.

A) Sense and antisense transcripts in the HOXA cluster (chr7), such as HOTAIRM1-HOXA and HOXA11-HOXA11-AS, have a strong co-expression pattern, and antisense genes are known to positively regulate the expression of their cognate sense gene. HOTAIRM1-HOXA **B**) The HOXD cluster (chr 2) is a novel example of HTH cis-NAT HOX genes with antisense transcripts potentially acting as positive regulators of the sense gene. **C**) Representative pairs of other HTH cis-NAT pairs, outside the HOX family, for which the antisense transcript could positively regulate the cognate gene.

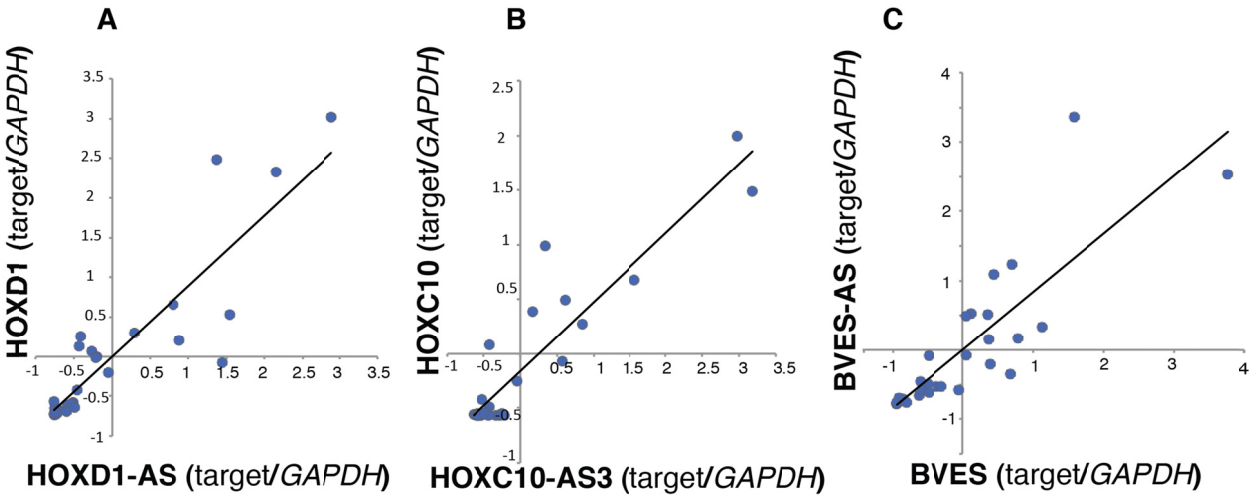


Figure 4.15. Experimental validation of positive gene expression correlation for representative HTH cis-NAT pairs.

Quantitative PCR validating the expression of **A) HOXD1-HOXD1-AS1; B) HOXC10-HOX10-AS3 and C) BVES-BVES-AS** across a lung cell line cohort of 29 samples.

4.3.5 Patterns of antisense expression in human cancer tissues

Analyzing the expression of antisense loci for tissue types with more than 10 samples in our compendia (Table 4.4-1, Major cohorts), we observed three broad groups of loci according to the expression of their antisense strand across different tissue types.

We identified a group of 3025 antisense loci that are expressed in all tissue types in our cohort and were therefore named as ubiquitous antisense loci (Figure 4.16a). These loci represent 39.27% - 65.04% out of all antisense loci identified in each tissue type. Protein coding genes involved in ubiquitous antisense loci are enriched for functions such as DNA repair and response to DNA damage, protein transport and localization, regulation of ncRNA metabolic process and phosphorylation (Figure 4.16c (left)). Out the 2052 protein coding genes in this category, 116 were known cancer related genes (tumor suppressors or oncogenes) such as FLI1, ATM, BCL2L11, NF1, TP53BP1, KRAS, PI3KCA and RAF1 among others. We noted that FLI1, ATM1 and

BCL2L11 represent particular cases of the HTH cis-NAT pattern described before, while KRAS, PIK3CA and RAF1 correspond to tail-to-tail configurations. In these last cases, the transcription of a neighboring protein coding gene overlaps or runs into the 3UTR and the body of those oncogenes (Figure C.3).

A second group of, an average, 2881 (sd=967.96) antisense loci were expressed at high levels in several tissue types and display low or absent expression in the others. These tissue enriched/non-specific antisense loci account for 26.21% - 53.23% out of all antisense loci identified in each cohort (Figure 4.16a). Protein coding genes in this group were enriched for cell adhesion, activation of protein kinases and embryonic morphogenesis (Figure 4.16c, middle). Notably, 133 known cancer genes were also found in this category including AXL, MTUS1, E2F2, TET2, JAK2, STK11 MAP4K1, BCAS1 and CCND1.

Finally, we identified a third group of antisense loci that are mainly expressed at high levels in only one tissue type. We consider these to be lineage-specific (Figure 4.16b), with the possibility that such transcripts contribute to tissue specific processes. This category of transcripts represented the smallest group, with an average of 244 (sd=166) loci by tissue, representing only 1.8 - 7% out of all antisense loci identified in each cohort. In contrast with the ubiquitous group, tissue specific antisense loci indeed were enriched for functions related with tissue development, morphogenesis and differentiation (Figure 4.16c, right. Out of 1563 linage specific loci, 113 involved tumor suppressors or oncogenes (Figure 4.16a) such as GTSE1, ERCC6 and GSK3B in LUAD; ABL2 in LUSC; ROS1, LCK and BCL2 in BRCA; TP53 and KLK10 in PRCA; CREBL2 and CDK2 in PANC; and RET, ABL1, TBX1 and VAV1 in the lung cell lines (Figure 4.16a). By inspecting the coverage maps of these examples, we found that ROS1, RET, VAV1, ABL2 and BLC2 do not have annotated overlapping transcripts; however we observed clear evidence of embedded antisense transcription in all of them (Figure C.5).

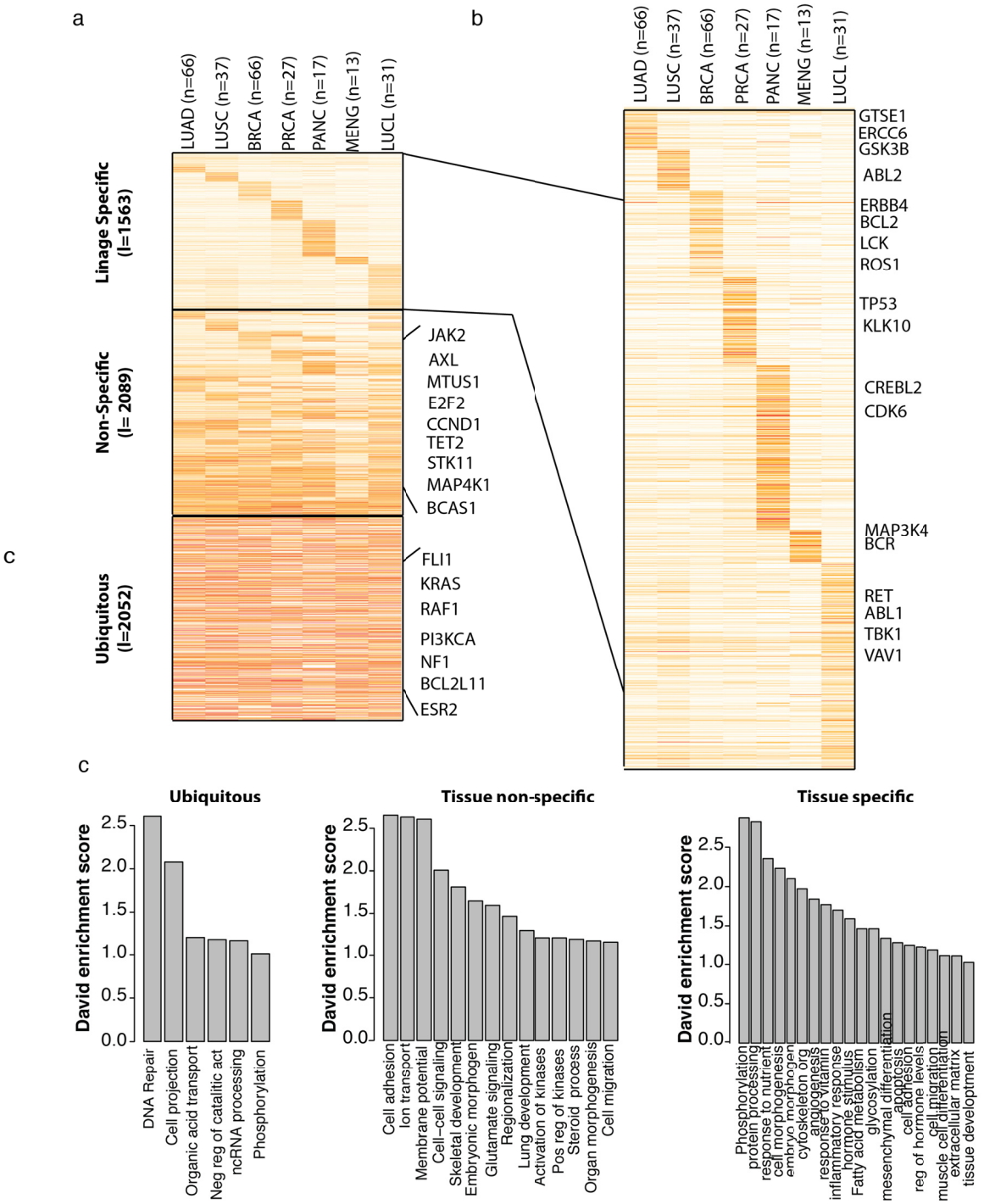


Figure 4.16. Antisense loci according to their expression across tissues.

A) Expression in the opposite strand of ubiquitous (n=2052), non-specific (n=2089) and linage specific (n=1563) antisense loci. **B**) Zoom in of linage specific antisense loci. Several tumor suppressor and oncogenes are identified as antisense loci in different cancer tissue types. **C**) Functional analysis of all protein coding genes identified as antisense sense loci, ubiquitous (left), tissue non-specific (middle) and tissue-specific (right).

4.3.6 Antisense loci in lung cancers

Because our compendium is substantially enriched for cancer samples, and benign samples correspond only to match normal samples of lung cancer patients (lung adenocarcinoma (LUAD) and squamous carcinoma (LUSC)), we investigate lung cancer specific antisense loci. 2101 cancer-specific antisense loci were found expressed in LUAD or LUSC but not in the benign samples (Figure 4.17a); 1456 were expressed in both tissues and 1212 out of those were also found in our cohort of lung cell lines. Out of those 1456 loci, 1260 were found in lung and at least another tissue type (Figure 4.17a) whereas 196 were lung cancer specific (Figure 4.17b).

Interrogating antisense loci involving cancer related genes, 88 cancer-related genes were found in which the expression of the opposite strand was statistically significant according to the NASTIseq score. Interestingly several of those genes do not have a previously annotated antisense transcript; however our ssRNASeq data suggest the presence of promoter associated, intronic and 3UTR antisense expression. E2F2 antisense transcript that locates to the 3UTR region of this gene has not been previously and is preferentially observed in lung cancers. ABL2, MTAP and GTSE1 display unannotated antisense expression originating from an embedded intronic transcript and they were mainly observed in LUAD and LUSC respectively.

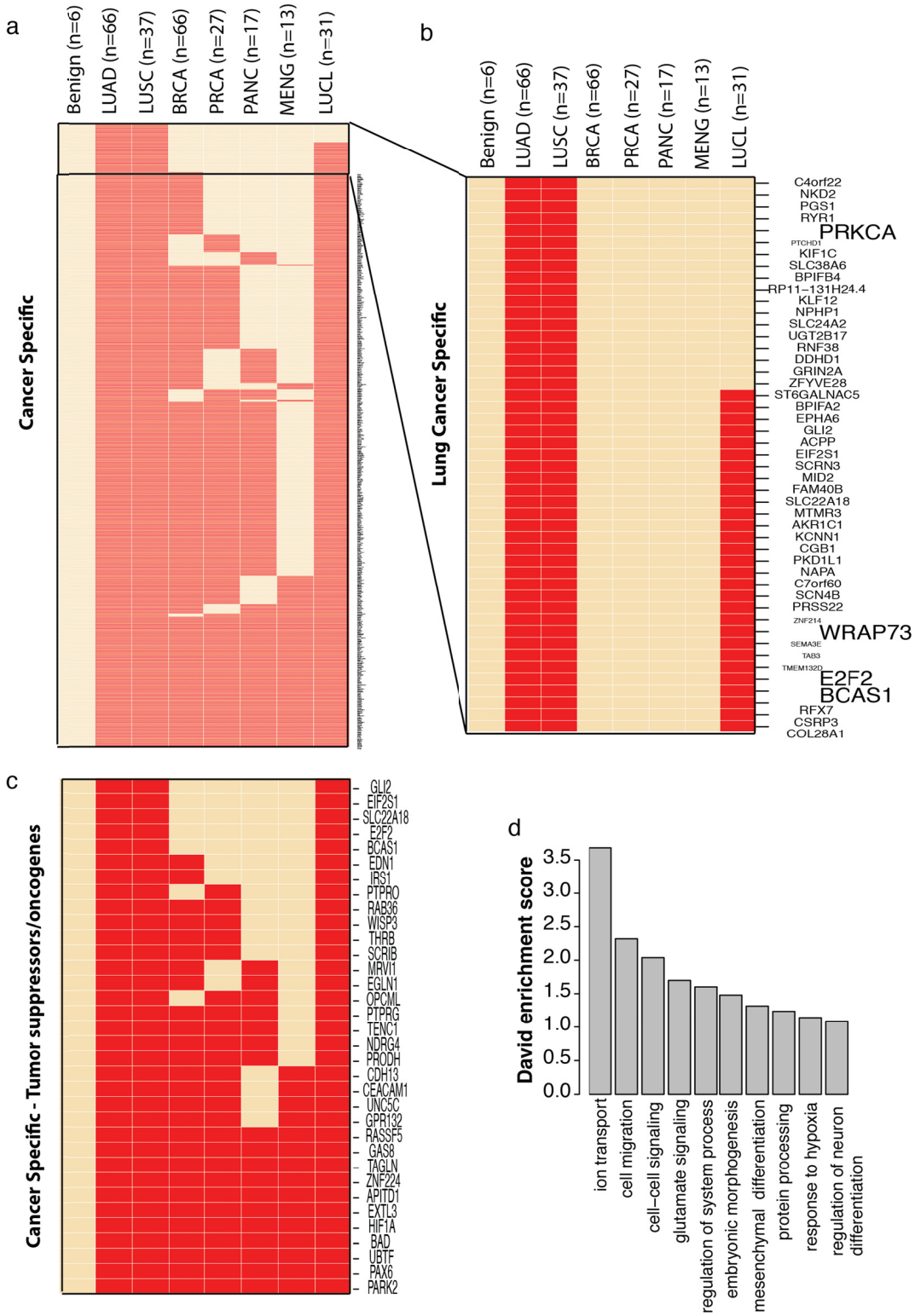


Figure 4.17. Cancer specific antisense loci.

A) Cancer specific antisense loci. Loci determined as antisense loci in lung and other cancer tissues but not in lung normal samples. **B**) Lung cancer specific antisense loci. Loci determined as antisense loci in lung

adenocarcinoma or lung squamous carcinoma but not in lung normal samples. **C**) Tumor suppressors and oncogenes specific antisense loci. **D**) Functional analysis of all protein coding genes identified as cancer specific antisense sense loci. The heatmaps represent presence/absence of antisense the Antisense loci in a particular tissue type.

Next, we focused on analyzing how the sense and antisense expression on a given locus changed between tumor and normal samples. We reasoned that the forward and reverse expression of a particular locus could change in a consistent or inconsistent fashion between tumor and normal samples. In a consistent change, the expression of forward and opposite strands will be over or under expressed between tumor and normal samples. In an inconsistent change, the expression of forward and reverse strands will change in opposite directions between tumor and normal samples. Therefore, when the forward strand is over expressed the opposite strand will be under-expressed and vice versa, suggesting different potential mechanisms of regulation between sense and antisense genes.

In order to identify loci with consistent or inconsistent expression changes, we used DESeq normalized read counts over forward and reverse strand of a locus to perform differential expression analysis between tumor samples and normal samples (Methods). A negative binomial test[52] was used to determine loci whose forward and reverse strands were differentially expressed. Both strands were required to have an absolute log fold change (lfc) greater than 1 with identical signs for consistent loci (lfc>=1 or lfc<=-1), while opposite signs for inconsistent loci (forward lfc>=1 and reverse lfc<=-1; or forward lfc<=-1 and reverse lfc>=1) (Methods).

First an analysis of 3 pairs of matched LUAD tumor and normal samples was performed, revealing the four groups of loci that we hypothesized (Figure 4.18a) and then this proof of concept analysis was extended to the full lung adenocarcinoma (n=66) and lung squamous carcinoma (n=36) cohorts. Figure 4.18b identified those groups by showing the average log fold change for each

locus across the LUAD cohort, while Figure 4.19 demonstrates that these consistent or inconsistent changes are uniformly observed in all samples.

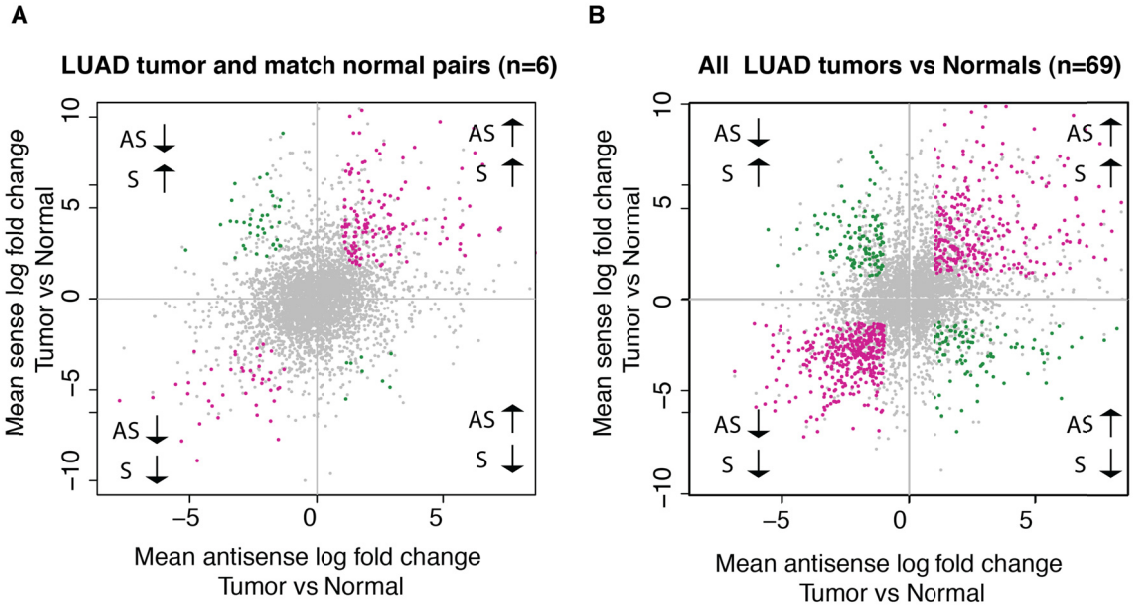


Figure 4.18. Antisense loci dysregulation in cancer.

A) Mean sense log fold change vs mean antisense log fold change in expression between tumor and normal matched pairs of lung samples (n=6 samples, 3 pairs). Gray dots represent unchanged pairs. Green dots represent antisense loci that show an opposite expression of the sense and antisense strand expression between tumor and normal. Purple dots represent loci showing a consistent change of the sense and antisense expression between tumor and normal. A dot is colored salmon if the relationship is observed in only one of the tumor-normal matched pairs. **B**) Mean sense log fold change vs mean antisense log fold change between Tumor and Normal Samples (N=69). Color code as in A.

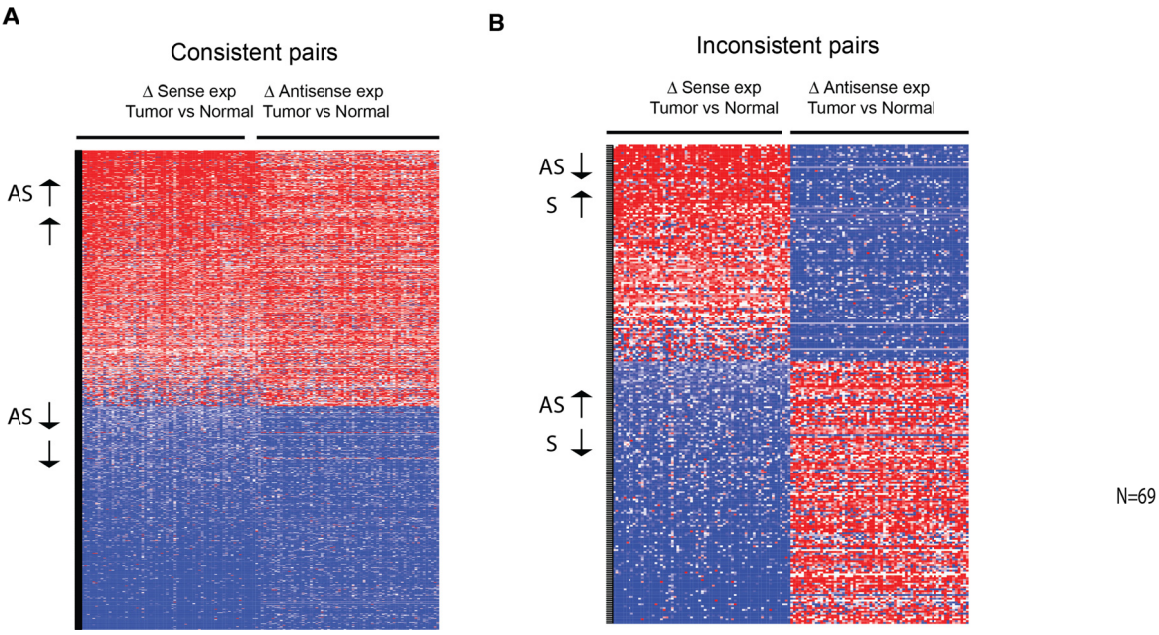


Figure 4.19. Heat maps of cancer's specific consistent and inconsistent sense-antisense pairs.

A) Heat maps of cancer's specific consistent, **B)** inconsistent sense/antisense pairs.

Overall, three were about three times more consistent than inconsistent loci in both LUAD and LUSC (Table 4.4-2). Intriguingly, by focusing on those differentially expressed loci that are annotated as cis-NAT gene pairs, we observed that consistent loci are enriched for head-to-head pairs, while embedded and tail-to-tail configurations are over-represented in the inconsistent group (Table 4.4-3). This is particularly interesting in light of the results discussed in previous sections, and suggests that indeed HTH cis-NAT gene pairs change in a coordinate fashion during cancer progression.

Table 4.4-2. Number of consistent and inconsistent differential expressed antisense loci.

	LUAD		LUSC	
	All	Cancer-related	All	Cancer-related
Consistent loci	831	55	1258	71
Inconsistent loci	258	10	417	18

Table 4.4-3. Break down by configuration of consistent and inconsistent differential expressed antisense loci.

	LUAD			LUSC		
	EMB	HTH	TTT	EMB	HTH	TTT
All cis-NAT	6807	1944	2152	6807	1944	2152
Consistent loci	125	133	65	171	131	84
Inconsistent loci	60	14	37	74	11	64

Analyzing each tissue independently shows that HTH cis-NAT pairs are over-represented in consistent loci fisher test p-value=1e-5 and for both LUAD and LUSC 3x3 contingency table. Fisher test p-value<2.2e-16 and 1.41e-15 for the 2x2 contingency table including the consistent loci.

The Hypoxia-inducible factor 1-alpha (HIF1A) and the Tubulin Polymerization Promoting Protein (TPPP) are representative examples of consistent and inconsistent loci respectively. In the HIF1A locus the expression of forward and reverse strands increase in tumors samples with respect to normals (Figure 4.20a); while in the TPPP locus the expression of the antisense transcript increases in tumors while TPPP sense expression decreases Figure 4.20b). Our data shows HIF1A and TPPP sense/antisense expression changes are rather general phenomena that is observed in both match tumor-normal pairs (Figure C.4) and the rest of tumor samples (Figure 4.21a, Figure 4.21b).

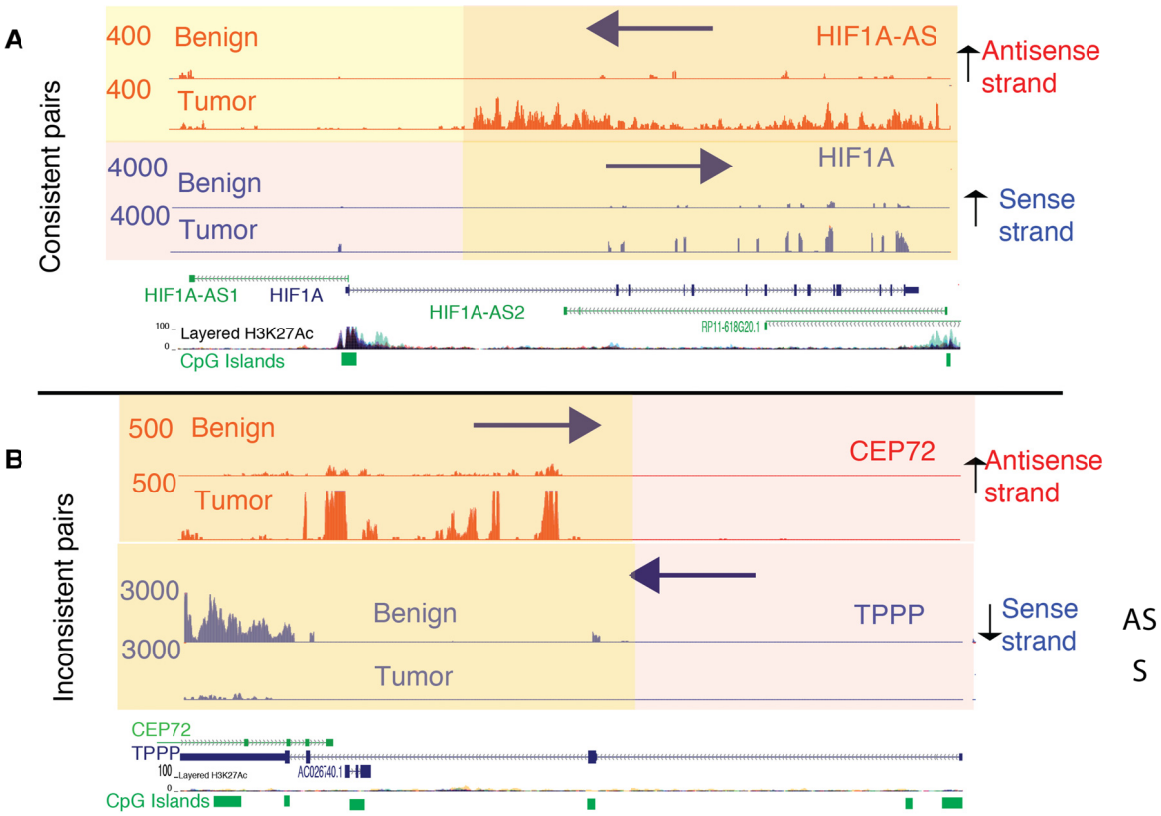


Figure 4.20. Examples of consistent and Inconsistent genes.

A) Coverage map for HIF1A, an example of a consistent antisense locus. **B**) Coverage map for TPPP, an example of an inconsistent antisense locus.

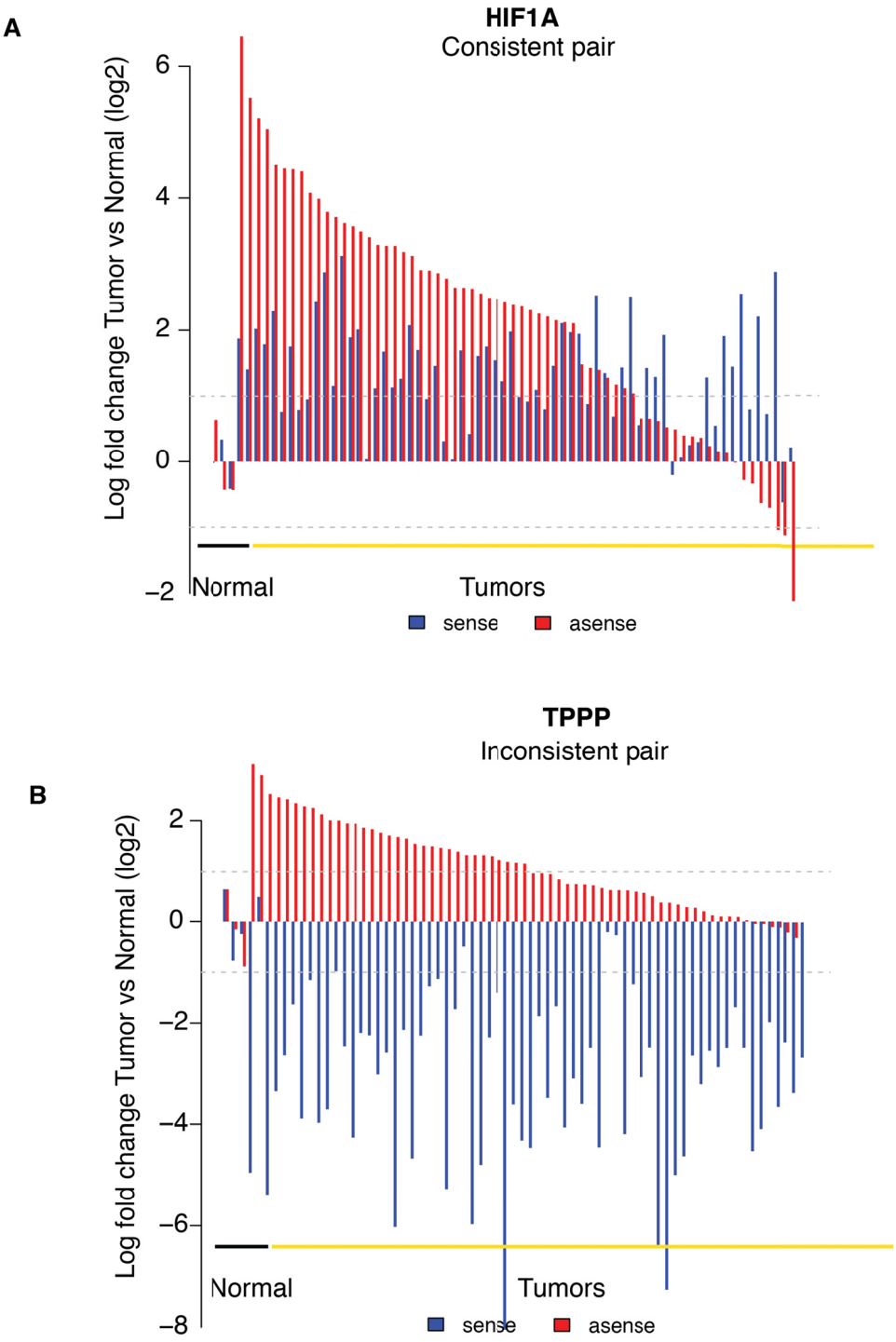


Figure 4.21. Log fold change between tumors and normals for consistent and Inconsistent genes.

A) HIF1A, consistent pair, sense and antisense RNASeq log fold change between tumor and the mean of the normal samples. The barplot shows sense and antisense expression changing in the same direction for this gene. **B**) TPPP, consistent pair, sense and antisense RNASeq log fold change between tumor and the mean of the normal samples. The barplot shows sense and antisense expression changing in opposite directions for this gene. An RTQ-PCR validation in one of our match tumor-normal pairs is presented in the **Figure C.4**.

4.3.7 oncoNATdb: a catalogue of antisense loci involving tumor suppressor and oncogenes

Given the increasing evidence for the role of antisense dysregulation in cancer[167-175,194], recent studies have suggested that targeting antisense transcripts in the clinical setting may represent a promising technology for modulating the expression of specific genes[158,195]. The first step in bringing these emerging therapies into the cancer arena is to catalogue and characterize all cancer related genes involved in cis-antisense regulation.

We therefore created, oncoNATdb, the first catalogue of cis-NAT pairs involving cancer related genes. To do so, we first performed an unbiased search of cis-NAT pairs in which at least one of overlapping genes was a known tumor suppressor or oncogene and calculated the gene expression correlation for the cis-NAT pair across our combined cohort of 376 cancer samples. 51% of tumor suppressors and 46% of oncogenes were found overlapping with another gene in the opposite direction (Table 4.4-4). Given that 46% of other protein coding genes harbor overlapping transcripts, these data suggest that tumor suppressors are slightly enriched for overlapping antisense transcripts (Fisher exact test p-value=0.0027), raising the possibility that antisense transcription could play a key role in modulating the expression of those genes.

Table 4.4-4. Number of tumor suppressors, oncogenes and other protein coding forming overlapping pairs.

	Protein coding genes	Tumor suppresors	Oncogenes
Overlapping other transcript	8650	379**	168
Not overlapping other transcript	10072	357	200
Total	18722	736	368

Next, we focused in HTH cis-NAT pairs involving tumor suppressors or oncogenes that had evidence of bidirectional promoters, high gene expression correlation and statistically significant expression of the antisense strand (high NASTIscore). A representative list of these candidates is presented in Table 4.4-5. Remarkably, our bioinformatics analyses capture the majority of cancer related genes known to be regulated by antisense transcripts. Furthermore our approach nominates new cis-NAT pairs involving tumor suppressors and oncogenes such as CCND2, MYCN, TP73, ATM and ETV7. An assessment of the mechanisms of regulation in these cis-NAT pairs will be informative for deciphering the role of antisense regulation in cancer.

Then, we look for TTT cis-NAT pairs involving cancer related genes within the linage specific, tissue enriched, ubiquitous and cancer specific antisense loci groups described earlier. We observed known oncogenes such as KRAS, PIK3CA and RAF1, in which the transcription of a neighboring protein coding gene overlaps or runs into the 3'-UTR and body of those oncogenes. Moreover, we applied the same analysis for annotated EMB cis-NAT pairs involving cancer related genes and found cases such as HIF1A, a cancer specific antisense locus that changes consistently between tumor and normal samples, and NF1. A list of representative examples for these categories is presented in Table 4.4-6.

Finally, we use our ssRNASeq data to directly examine the antisense expression on cancer related genes that did not have annotated overlapping transcripts. We found additional examples of oncogenes and tumor suppressors with significant expression of the antisense strand, suggesting potential novel transcripts that are overlapping and might regulate those genes. Such genes included RET, VAV1, E2F2, and BLC2. A representative list of those cases is presented in Table 4.4-7.

4.4 Discussion

In this study, we used strand specific RNA sequencing on a cohort of 376 samples to describe the magnitude and patterns of antisense expression in the human cancer transcriptome. Based on our analyses, we have further created oncoNATdb, a catalogue of antisense loci involving tumor suppressor and oncogenes. Our results indicated that more than 60% of all human loci have measurable expression of the antisense strand, suggesting that antisense transcription is widespread phenomenon across the genome. In addition, we also show that, on average, 37% of those loci would correspond to *bona fide* expressed cis-NAT pairs (Figure 4.7). Our estimates expand upon earlier limited assessments of the extent of the antisense transcriptome[86,87].

Moreover, by analyzing the expression patterns of overlapping genes, we confirmed that gene expression of overlapping genes is positively correlated (median Spearman correlation coefficient R=0.27), and in particular, that HTH cis-NAT pairs have the highest correlation (median Spearman correlation coefficient R=0.4) among all other configurations types (Figure 4.8). This high correlation of HTH pairs, we hypothesize, is due to bidirectional promoters that direct the expression of both genes in the pair. Supporting this, greater than 78% of HTH cis-NAT pairs have CpG islands in their overlapping regions, suggesting bidirectional promoters; similarly 83% of bidirectional but not overlapping genes had CpG islands in their intergenic regions (Figure 4.12). This hypothesis is further supported by detailed analyses of known examples in the HOXA cluster and experimental validation of the co-expression pattern of novel candidates in a panel of lung cell lines (Figure 4.14, Figure 4.15). Remarkably, differentially expressed cis-NAT pairs between tumor and normal samples that exhibit a consistent behavior are enriched for HTH pairs (Table 4.4-3), implying a common mechanism of regulation.

Furthermore, by analyzing the expression of antisense loci in the major tissue types in our compendia, we observed three broad groups of loci according to the expression of their antisense strand across different tissue types. A first group of 3025 ubiquitously expressed antisense loci were found present in all tissue types in our cohort (Figure 4.16a). Those loci were enriched for functions such as DNA repair, phosphorylation and ncRNA processing. Notably we found 116 cancer related genes such as Fli-1 Proto-Oncogene ETS-Transcription Factor (FLI1), which forms a HTH cis-NAT pair with FLI1-AS1 transcript, and KRAS, PIK3CA and RAF1 oncogenes that form TTT cis-NAT pairs with neighboring protein-coding genes. Although the potential functional consequences of such tail-to-tail configurations remain largely unknown, a recent study in gastric cancer showed that two TTT overlapping protein-coding genes could concordantly regulate each other by forming a RNA duplex at the overlapping 3'-UTRs which increased their mutually stability[162]. Our resource provides potential new candidates for this phenomenon, which merit further investigation and raise the exciting possibility of new avenues for regulating the expression of well-known oncogenes.

A second group of, on average, 2881 (sd=967.96) tissue enriched antisense loci expressed at high levels in several tissue types and absent in the others (Figure 4.16b) was also found. Within this group 133 cancer-related genes displayed significant antisense expression. The last group corresponds to lineage specific antisense loci that are mainly expressed in only one tissue type (Figure 4.16b). Despite representing only 1.8 - 7% out of all antisense loci identified in each cohort, lineage specific antisense loci were enriched by tissue specific morphogenesis functions and thus have the potential of regulating biological processes unique to distinct tissue types. We found 113 cancer-related genes in this group, such as ROS1, ABL2, and BLC2. Interestingly, several of those genes do not have annotated overlapping transcripts, however our ssRNASeq shows clear evidence of embedded or promoter associated antisense transcription in all of them (Figure C.5). These observations demonstrate the advantages of using

ssRNASeq to resolve the expression of complicated regions with overlapping transcripts and to discovered potential new loci with significant but unannotated antisense expression.

In addition, comparing tumor and benign lung adenocarcinoma and squamous carcinoma samples we found cancer specific loci (Figure 4.17) and showed that the expression of the two genes in a cis-NAT pair can change in the same (consistent) or opposite (inconsistent) direction when comparing tumor and normal samples (Figure 4.18, Figure 4.19). Noteworthy examples of consistently regulated cancer related genes loci in LUAD were Zinc Finger E-Box Binding Homeobox 2 (ZEB2) and Polo-Like Kinase 4 (PLK4). ZEB2 and ZEB2-AS1 form a bidirectional HTH cis-NAT pair that is essential for down regulation of E-cadherin during epithelial-mesenchymal transition. Beltran *et al* 2008[196] elegantly showed that ZEB2 and ZEB2-AS1 transcription is directed by a bidirectional promoter and more importantly that ZEB2-AS1 up-regulates Zeb2 protein expression, which in turn down regulates E-cadherin expression. On the other hand PLK4 is essential for centriole duplication and when overexpressed is important in tumorigenesis by inducing centrosome aberrations. Notably, the antisense transcript overlapping PLK4 has not been annotated yet, but according to our ssRNASeq data it is oriented in a HTH configuration overlapping the 5 prime region of PLK4. Elucidating the biological implications of those very different expression patterns would deepen our understanding of antisense regulation and their role in cancer.

Finally, our study comprehensively examined, for the first time, the extent of antisense expression in cancer related genes and aggregated these findings in oncoNATdb, a catalogue of cancer-related genes with significant antisense expression. We show that 608 (50.08%, out of 1214) of cancer-related genes have annotated overlapping transcripts and 296 out of those 608 have significant antisense expression. 48.64% of the overlapping pairs formed by those 296 genes are HTH cis-NAT examples, with high correlation (R>=0.3) and evidence

of bidirectional promoters, whereas 25.02% are TTT and 24.34% are EMB gene pairs. In addition, 155 cancer-related genes have significant antisense expression, but do not have annotated overlapping transcripts. 27 out those 155 have a very close gene nearby (<=500bp) whose UTRs transcription could extend into the neighboring gene. The reaming 128 cancer-related genes could have putative novel antisense transcripts.

Antisense transcripts regulate several well-studied tumor suppressors and oncogenes and there is increasing evidence of antisense dysregulation in cancer[167-175]. The molecular mechanisms of this regulation are multiple and poorly understood. Nevertheless controlled modulation of natural antisense transcripts, in order to modify the expression of sense genes[158], is an emerging technology that promises to deliver gene specific targeted therapies.

This study characterizes the landscape of antisense expression in human cancers and provides a resource, oncoNATdb, which will enable cancer researchers to investigate sense-antisense regulation and its role in cancer.

4.5 Contributions

Science is a collective enterprise and it is much more fun when done with friends and good collaborators. The results presented in this chapter were made possible for the great collaboration and support of a team of people in the Chinnayian and Nesvizhskii labs.

O. Alejandro Balbin: Sense and antisense bioinformatics analysis pipeline, Omics data integration, ssRNASeq data processing, statistical analysis, Manuscript writing. John Prenser: PCR validation of sense and antisense gene pairs. Rohit Malik: Experimental validation of sense/antisense gene pairs. Saravana M. Dhanasekaran, Dan Robinson and Yi-Mi Wu: Beautiful RNASeq strand specific

libraries for the MCTP cohort. Xuhong Cao: RNA Sequencing. Alexey Nesvizhskii: Data analysis oversee and manuscript writing. Arul M. Chinnayian: Overall scientific project oversight and manuscript writing.

Table 4.4-5. Representative tumor suppressors and oncogenes head-to-head cis-NAT pair with bidirectional promoters.

The column Status indicates that, according to previous reports, the antisense transcript regulates the actionable gene expression.

Status	Gene Pair	Overlap region	Transcript biotypes	Overlap Type*	Spearman correlation	Adj-pvalue	NASTI score	Gene Type*
Reported	CDKN2B-AS1_CDKN2A	chr9:21994787-21995300	Antisense-protein_coding	HTH	0.8042	1.51E-83	799633.48	TS/ONC
Reported	HIF1A_RP11-618G20.4	chr14:62162255-62162557	protein_coding-lincRNA	HTH	0.5203	1.28E-24	165798.13	TS/ONC
Reported	WRAP53_TP53	chr17:7589616-7592397	protein_coding-protein_coding	H2H	0.3658	1.09E-10	23555.416	TS/ONC
Reported	HOTAIRM1_HOXA1	chr7:27135175-27135615	Antisense-protein_coding	H2H	0.6145	1.66E-37	12623.264	TS/ONC
Reported	ZEB2-AS1_ZEB2	chr2:145277418-145277677	Antisense-protein_coding	H2H	0.6727	6.26E-48	10518.263	TS/ONC
Reported	WT1-AS_WT1	chr11:32456243-32457392	Antisense-protein_coding	H2H	0.8939	1.92E-129	-	TS/ONC
-	CCND2_RP11-264F23.4	chr12:4361898-4414516	protein_coding-antisense	H2H	0.7983	6.56E-81	144673.84	TS/ONC
-	MYCN_MYCNOS	chr2:16082067-16082976	protein_coding-antisense	H2H	0.6500	5.42E-44	122128.44	TS/ONC
-	TP73_WRAP73	chr1:3547328-3652765	protein_coding-protein_coding	H2H	0.4131	1.59E-13	13637.53	TS/ONC
-	PROX1_PROX1-AS1	chr1:213992975-214214853	protein_coding-antisense	H2H	0.8537	1.41E-104	4072.8672	TS
-	CAV2_AC002066.1	chr7:116139405-116139985	protein_coding-antisense	H2H	0.8354	2.73E-96	2231070.7	TS
-	PDX1_PDX1-AS1	chr13:28403903-28500368	protein_coding-antisense	H2H	0.7379	2.18E-62	46064.246	TS
-	ATM_NPAT	chr11:108093208-108093913	protein_coding-protein_coding	H2H	0.6919	6.49E-52	7131.8909	TS
-	RP1-50J22.4_ETV7	chr6:36322416-36359771	Antisense-protein_coding	H2H	0.7268	1.64E-59	13477.113	ONC
-	ARHGEF5_RP4-798C17.6	chr7:144052378-144052613	protein_coding-antisense	H2H	0.7634	5.09E-70	25937.039	ONC

-	TYMS_C18orf56	chr18:657601-658340	protein_coding-protein_coding	H2H	0.7484	1.03E-65	458624.8249	ONC

***HTH=Head-to-Head, TS=Tumor suppressor, ONC=Oncogene. All gene pairs in this table have had CpG islands in the overlapping regions between the genes and the loci was called as antisense loci by NASTISeq.**

Table 4.4-6. Representative tumor suppressors and oncogenes tail-to-tail and embedded cis-NAT pair with bidirectional promoters.

The column Status indicates that, according to previous reports, the antisense transcript regulates the actionable gene expression.

Status	Gene Pair	Overlap region	Transcript biotypes	Overlap Type*	Spearman correlation	Adj-pvalue	NASTI score	Gene Type*
-	PIK3CA_KCNMB3	chr3:178951879-178957881	protein_coding&protein_coding	T2T	0.3179	9.74E-08	48034.56	ONC
-	LYRM5_KRAS	chr12:25357022-25362845	protein_coding&protein_coding	T2T	0.3097	3.53E-07	977660.66	TS/ONC
-	MKRN2_RAF1	chr3:12623612-12626156	protein_coding&protein_coding	T2T	0.4221	7.80E-15	62380.91	ONC
-	ESR1_SYNE1	chr6:152011628-152958936	protein_coding&protein_coding	T2T	0.3356	5.52E-08	430288.52	TS
-	CREB1_METTL21A	chr2:208394458-208490652	protein_coding&protein_coding	T2T	0.2866	3.26E-05	934027.76	TS
-	FLI1_FLI1-AS1	chr11:128562386-128563286	protein_coding&antisense	EMB	0.8385	4.54E-98	12775.55	TS
-	TPPP2_NDRG2	chr14:21484919-21539031	protein_coding&protein_coding	EMB	0.6476	1.32E-42	26275.06	TS
-	WNT5A-AS1_WNT5A	chr3:55499740-55523973	antisense&protein_coding	EMB	0.5328	1.65E-25	36668.81	ONC
-	NF1_EVI2B	chr17:29421942-29708905	protein_coding&protein_coding	EMB	-0.2357	0.007099035	881428.7	TS
-	DLG3_DLG3-AS1	chrX:69664708-69725337	protein_coding&antisense	EMB	0.5592	7.27E-29	54784.17	
-	NF1_EVI2A	chr17:29421942-29708905	protein_coding&protein_coding	EMB	-0.2789	7.99E-05	881428.7	TS

-	NTRK1_INSRR	chr1:156811870-156812063	protein_coding&protein_coding	EMB	0.5510	1.70E-28	47464.80	ONC

Table 4.4-7. Representative tumor suppressors and oncogenes with significant antisense expression but no annotated overlapping transcripts.

Status	Gene Pair	Overlap region	Transcript biotypes	Overlap Type*	Spearman correlation	Adj-pvalue	NASTI score	Gene Type*
-	RET	chr10:43572473-43625799	protein_coding	EMB	-	-	179.81	ONC
-	VAV1	Chr19:6772720-6857371	protein_coding	EMB	-	-	186.39	ONC
-	E2F2	chr1:23832920-23857712	protein_coding	TTT	-	-	5405.11	ONC
-	BCL2	chr18:60790577-60987361	protein_coding	EMB	-	-	1270.35	TS
-	PTEN	chr10:89622868-89731687	protein_coding	HTH	-	-	1118.89	TS
-	VAV2	chr9:136627014-136857726	protein_coding	EMB	-	-	1611.13	ONC
-	CDKN2C	chr1:51426415-51440305	protein_coding	HTH, paRNA?	-	-	4239.83	ONC

Chapter 5
Conclusions and future directions

Omics technologies for high-throughput profiling of human genome, transcriptome and proteome are revolutionizing cancer research and nourishing a nascent paradigm in clinical care. The success of this new precision medicine paradigm will depend on our ability to combine diverse omics-based measurements to distill clinically relevant information that can be acted upon. This thesis developed bioinformatics approaches to integrate multi-omics datasets and applied these approaches in three distinct studies that identified novel actionable genes and pathways in cancer.

In Chapter 2, alternative targetable proteins were found in non-small cell lung cancers (NSCLC) with activating mutations in KRAS (a well-know but undruggable oncogene) by profiling their transcriptome, proteome and phosphoproteome. By reconstructing targetable networks associated with KRAS dependency, we nominated lymphocyte-specific protein tyrosine kinase (*LCK*) as a critical gene for cell proliferation in these samples, suggesting LCK as a novel druggable protein in KRAS-dependent NSCLC.

In Chapter 3, novel oncogenic gene fusions were identified in NSCLC patients with previous to this work unknown driver genes. By characterizing the landscape of fusions in NSCLC, this study revealed that gene fusions incidence is an independent prognostic factor for poor outcome. It was also discovered that Neuregulin 1 (*NRG1*) is a novel low recurrence 3' fusion partner present

exclusively in patients with an unknown driver; resembling previously reported and targetable kinase fusions in lung cancers.

Chapter 4 focused on the characterization of cancer-related genes that are involved in sense-antisense gene pairs and could be regulated by natural antisense transcripts. By determining the extent of antisense gene expression across human cancers and comparing with well-documented sense-antisense pairs, our results raise the possibility that antisense transcripts could modulate the expression of well-known tumor suppressors and oncogenes. This study provided a resource, oncoNATdb, a catalogue of cancer related genes with significant antisense transcription. The oncoNATdb catalogue will enable researchers to investigate the mechanisms of sense-antisense regulation and further advance our understanding of their role in cancer, which may lead to the discovery of novel therapies.

Collaborative projects such as the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) are generating vast amounts of omics-based datasets. These projects are profiling the genome, transcriptome and proteome for thousands of patients, providing an unprecedented molecular characterization of multiple cancer types. These datasets also provide an exceptional opportunity to discover novel targetable genes and pathways on patient populations with currently unmet needs. Integrative analyses would be essential to translate this molecular information into informative findings that points towards new therapies and novel targets. The bioinformatics methods presented in this thesis illustrated different approaches to integrate these multi-omics datasets.

In the future, I anticipate building scalable bioinformatics approaches based upon the computational methods presented in this thesis to integrate the multi-omics datasets produced by projects such as the TCGA and the ICGC. Such a system should be able to generate a detailed molecular profile for each

patient in the cohort (using all omics-based measurements), integrate relevant clinical information and determine patient communities based on those molecular and clinical profiles. As fundamental feature of this system should include a patient's molecular profile oriented search, allowing researchers (and patients) to use a patient's molecular profile in order to retrieve those other patients (community of patients) with molecular profiles that closely resemble the query. The community of patients generated by this "patients like me-molecular" search, may facilitate the discovery of novel targets and unappreciated therapeutic opportunities.

The discovery of LCK kinase as druggable target in KRAS-dependent NSCLC merits additional experimental and bioinformatics studies to explore its specific role in these cancers and potential avenues to inhibit its activity. Preliminary results, not shown in this dissertation, indicate that LCK localizes to the nucleus of KRAS dependent cells. A previous report in T-acute lymphoblastic leukemia (T-ALL) also showed nuclear localization of LCK. In T-ALL LCK binds to the promoter of LIM domain only protein (Lmo2)[197], which is a critical transcription factor in the development of this disease[197]. Therefore, LCK could be exerting unanticipated roles in KRAS dependent NSCLC by directly regulating the activity of transcription factors. In order to study this hypothesis, Chip-Seq experiments for LCK could be performed in order to determine what DNA regions LCK binds and identify the genes that are regulated, if any exists. Coupling those experiments with RNASeq or microarray profiling after LCK knockdown could also reveal the precise links between LCK and the apoptosis pathways that were suggested in chapter 2. In addition, it is essential to extend the clinical significance of LCK in disease free survival in order to determine the prognostic value of LCK in lung cancer. For this, we could use the currently available TCGA NSCLC dataset to evaluate the significance of LCK as a prognostic marker. This analysis is, however, complicated by the overall poor prognosis of lung cancers, but we anticipate that detailed clinical follow-up of the TCGA and our internal cohort of patients would allow us to disentangle the effect of LCK in prognosis.

The identification of low recurrence NRG1 fusions, as well as NRG1 overexpression, in NSCLC driver negative patients suggest that close to 4% of NSCLC driver negative patients could benefit from further studies on the role of NRG1 in NSCLC and the development of directed therapies for targeting NRG1. Chemotherapy is the first line treatment for more than 50% of patients with NSCLC, regardless of the stage; however, in some cases chemotherapy cannot remove the tumor or prevent disease recurrence. A study recently published demonstrated that residual tumor cells after chemotherapy express high levels of NRG1; moreover, inhibition of NRG1 signaling significantly enhanced the magnitude and response to chemotherapy[152]. A deep characterization of all NRG1 fusions presented in this study (localization, and interaction partners), as well as the common signaling pathways activated in both fusion index samples and outlier expression samples would help to determine the mechanism of action of NRG1. Chapter 3 also presented a novel approach for identifying and filtering out the vast amount of false positive fusions produced by any of the fusion algorithms. The fusion classifier developed in chapter 3 could be further improved by including information about the presence or absence of an open reading frame (ORF) in the fusions formed. In order to include this, we would need to extend the algorithm to determine the sequence of all potential fusion transcripts formed between the 5' and 3' fusion genes and then determine the longest ORF that extends beyond the fusion breakpoint. A categorical value, 1/0, would then be included as an additional feature in the classification step. Including ORF information would focus the results on rearrangements producing fusion proteins, as the previously reported kinase fusions. Finally, the fusions database generated in this study could be extended to include additional lung datasets and additional fusion events called with improved fusion detection algorithms. This database could constitute a reference point for other researchers looking for low recurrence fusions in NSCLC.

Chapter 4 focused on the characterization of cancer-related genes that are involved in sense-antisense gene pairs and could be regulated by natural antisense transcripts. This is a very new field and the future directions are unlimited. To begin with, this study suggests a relationship between HTH-cis-NAT pairs and bidirectional promoters, which could be further enhanced by integrating omics-based measurements of additional chromatin marks of histone modification and nucleosome free regions. Generating these datasets for the cell types used in this study is the only limiting factor to making progress in this direction. The ssRNASeq utilized in this study also confirms a widespread expression of antisense transcripts from the promoter of many genes as it was previously observed in yeast[193]. We have not address the extension of those specific type of ncRNAs, neither its relation with cancer genes. An immediate follow up study would characterize this phenomenon as preliminary observations of the coverage maps of gene expression shows that promoter ncRNA (paRNA) are highly transcribed from several cancer genes across tissue types. oncoNATdb could be further extended to include those examples of paRNA that have not been annotated but found in cancer-related genes. More importantly, the study presented in chapter 4 suggested that many cancer-related genes could be regulated for antisense transcripts. Therefore, designing clever experiments to disentangle the mechanism of regulation should be at the forefront of future follow up studies. In particular, it would be essential to demonstrate what antisense transcripts activate or silence their respective cognate gene targets. Stabilization of oncogenes mRNA may lead to increased activity in cancer cells, while interference in tumor suppressors expression may abolish their activity promoting cancer development.

In conclusion, the computational methods for integrating omics-based datasets developed in this thesis will assist others with similar tasks and challenges. More importantly, these approaches nominated novel targetable genes and pathways for patient populations with “undruggable” cancers,

warranting further studies of the therapeutic opportunities provided by these discoveries.

Appendix A
Additional analyses for chapter 2

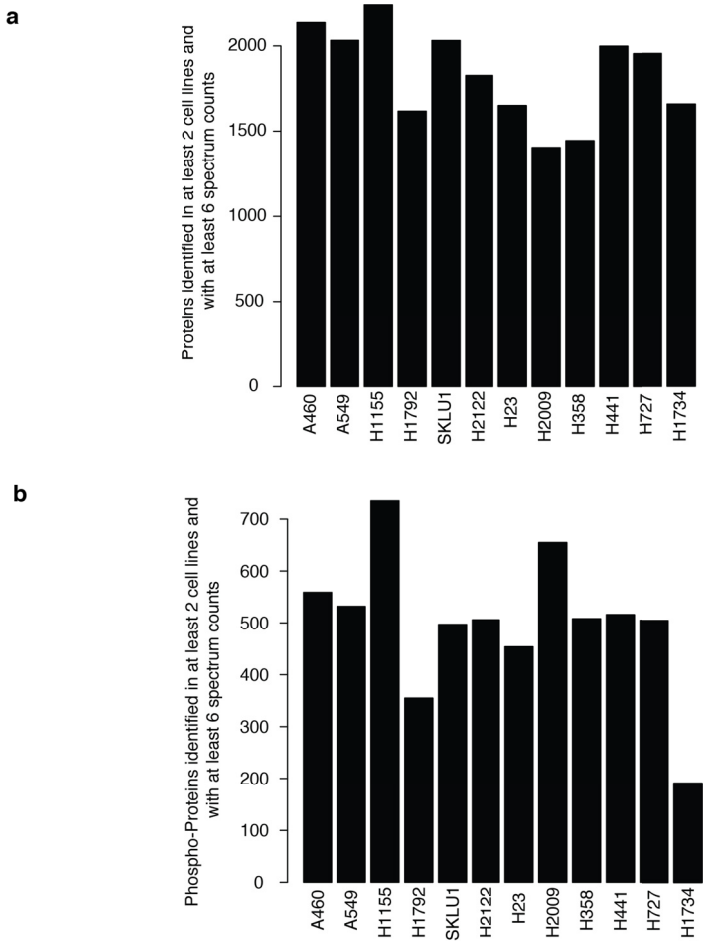


Figure A.1. Number of proteins and phosphoproteins identified by LC-MS/MS.

A) Proteins identified in the flow through dataset in at least two different cell lines, with at least one unique peptide and a minimum of 6 spectrum counts across all cell lines. **B)** Phospho-proteins identified in the enrich fraction in at least two different cell lines, with at least one unique phospho-peptide and a minimum of 6 spectrum counts across all cell lines.

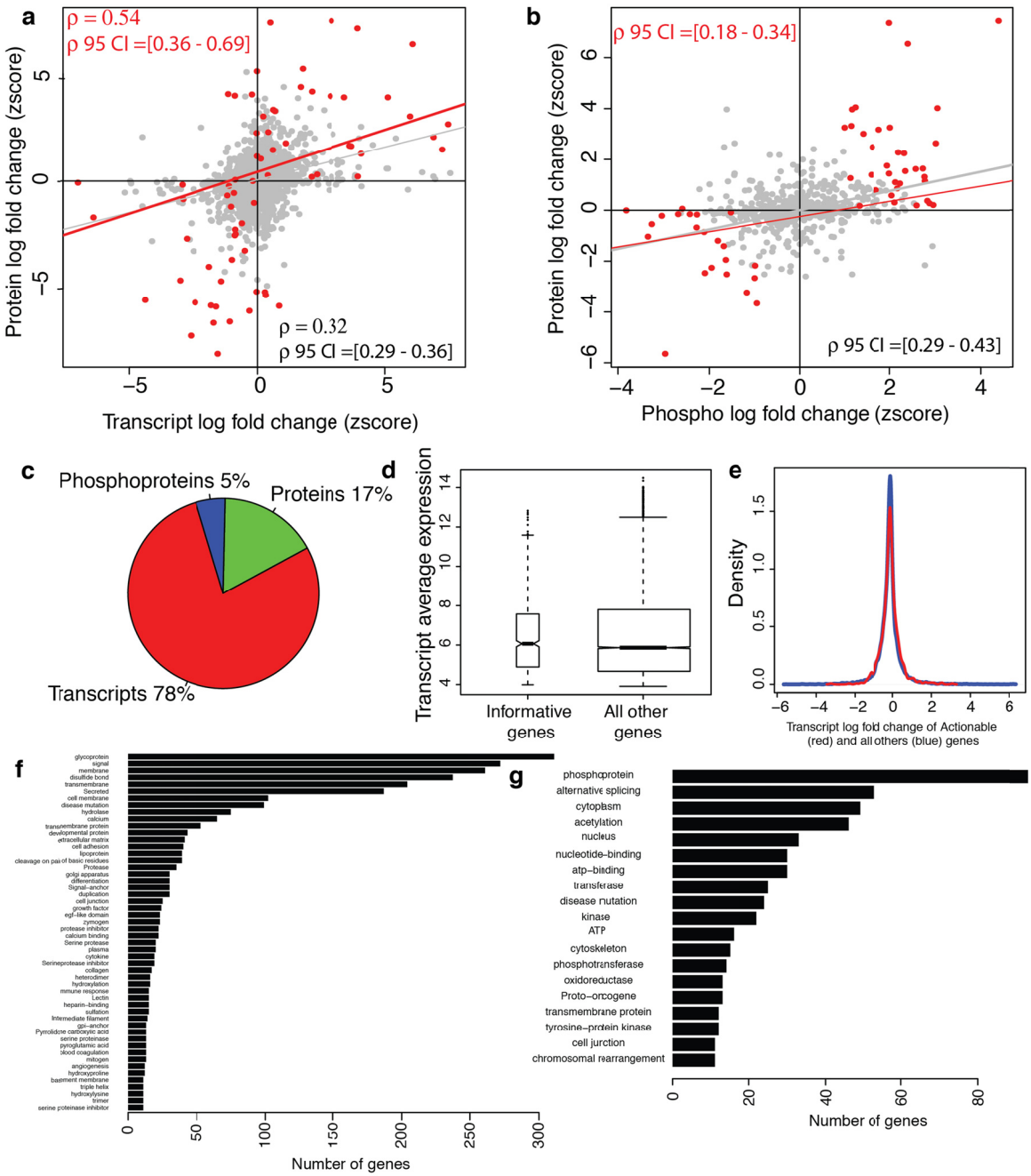


Figure A.2. Analysis of omics signatures.

A) Overall correlation between LFC of transcript and LCF of unmodified protein. Differential abundance was calculated as the log fold change between the transcript or protein expression in the KRAS-Dep vs KRAS-Ind cell lines. In red proteins identified as differentially expressed by the combined S score and gray all other proteins. The 95% confidence interval (CI) for the Pearson correlation (ρ) is shown in gray for all proteins and in red for proteins detected as differentially expressed by the S score. Only proteins with both transcript and protein abundance are plotted. Only proteins with both transcript and protein abundance are plotted. **B**) Overall correlation between LFC for unmodified protein and the LFC phosphorylated protein. Differential abundance was calculated as in a). Colors and 95% CI as in a). Only proteins with both protein and phosphoprotein abundance are plotted. **C)** A naïve method for integrating transcript, protein and phosphoprotein signatures produces a set of differentially expressed proteins overrepresented by genes

found only in the transcriptome dataset (~78%). Only 17% and 5% of the proteins would be found as differentially expressed in the proteomics or phospho-proteomics datasets respectively (A t-test adjusted pvalue <=0.05 for the LFC was used to select differentially expressed proteins). **D)** Average transcript expression of informative and non-informative genes across the panel of cell lines. Informative genes have a smaller dynamic range of expression than the non-informative genes. Whiskers correspond to the data point +/- 1.5 of the interquartile range of each box. The widths of the boxes are drawn proportional to the square-roots of the number of genes in each group. **E)** Distribution of the differential expression values for informative (red) and non-informative (blue) genes when comparing KRAS-Dep vs KRAS-Ind cell lines. The longer tails in the distribution of non-informative genes determines the set of genes that are selected as differential expressed genes by a naïve approach leaving out most of the informative genes. **F)** Proteins found as differentially expressed only in the transcriptome dataset have very general and unspecific functions. Proteins in this dataset are mainly glicoproteins, transmembrane or secreted proteins, which are characterized by a wide dynamic range of expression but are not necessary related with KRAS dependency phenotype. **G)** Proteins found as differentially expressed using the S score are enriched on very specific molecular functions such as phosphorylation, alternative splicing, and acetylation.

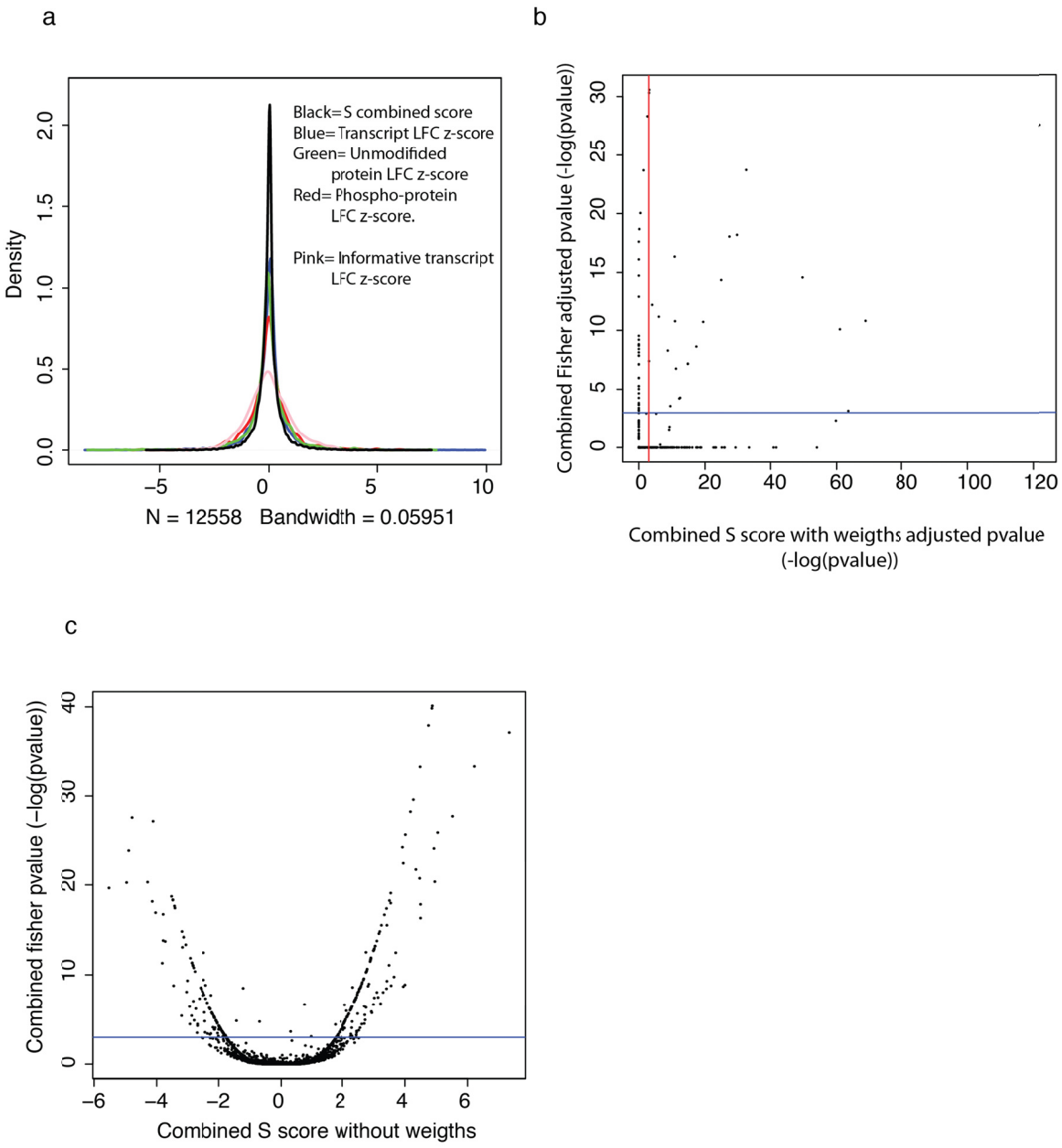
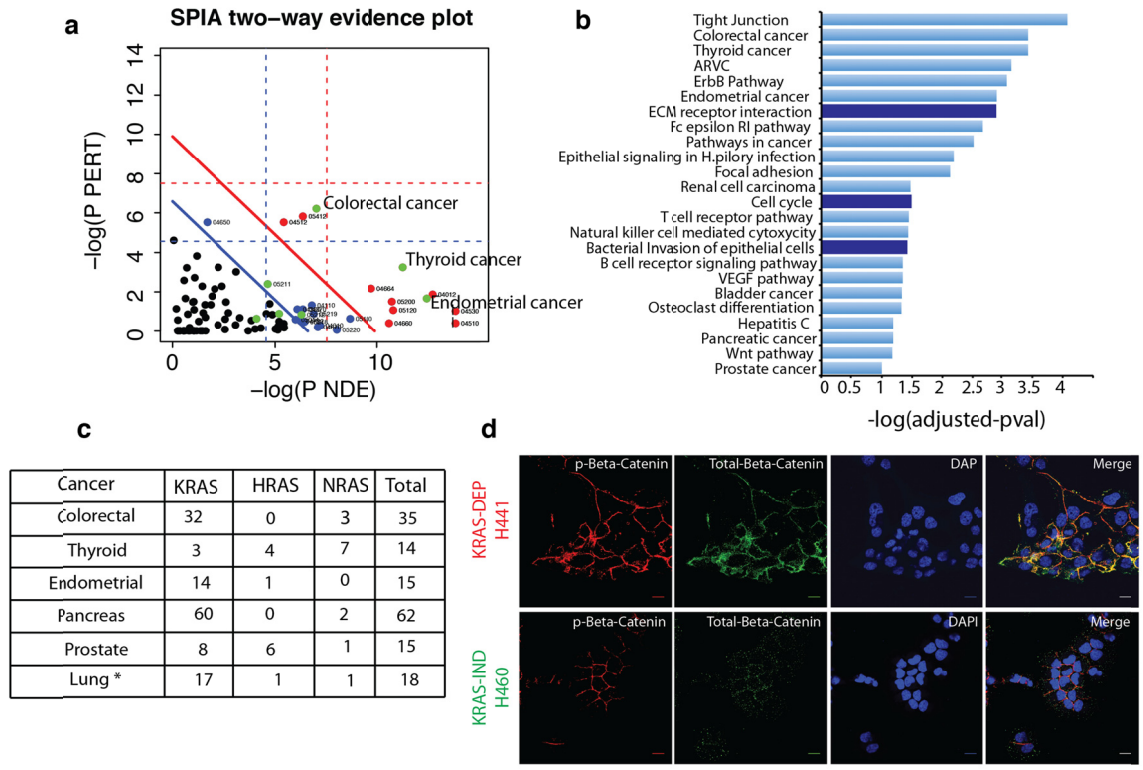


Figure A.3. Comparison between integrative scores.

A) Density plot for all z-score normalized log fold change (LFC). LFC was calculated with respect to the comparison KRAS-Dep vs KRAS-Ind. **B)** Combined fisher pvalue vs Combined S score pvalue. Both pvalues were adjusted using the Hochberg method. Blue and red lines indicates pvalues=-log(0.05). **C)** Volcano plot for the combined fisher pvalue and the combined S score. Blue line marks pvalue = -log(0.05).



Cancer	KRAS	HRAS	NRAS	Total
Colorectal	32	0	3	35
Thyroid	3	4	7	14
Endometrial	14	1	0	15
Pancreas	60	0	2	62
Prostate	8	6	1	15
Lung *	17	1	1	18

Figure A.4. Network enrichment analysis of integrated datasets.

A) SPIA network enrichment analysis. Horizontal axis corresponds to the enrichment probability of obtaining at least the observed number of genes (NDE) on the given pathway just by chance (-log(p-value)). The vertical axis corresponds to the perturbation enrichment, and represents the probability of obtaining the observed total perturbation or more extreme on the given pathway just by chance. Pathways at the right of the red oblique line are significant after Bonferroni correction of the global p-values obtained by combining the pPERT and pNDE evidence, while pathways at the right of the blue oblique line are significant after a FDR correction of the global p-values. The KEGG pathway-id is shown close to each bullet point. **B)** SPIA enrichment analysis reveals activation of main signaling programs in KRAS-Dep vs KRAS-Ind cells, such as ERBB signaling and cancer specific associated pathways, tight junctions and cell adhesion. Interestingly, KRAS-Dep cells also activates immune related signaling modules such as T cell receptor signaling, natural killer cell mediated cytotoxicity and Fc epsilon RI pathway. Pathways displayed are significant after FDR corrections with global adjusted p-values <= 0.05. Light and dark blue represent activated and inhibited pathways, respectively, in KRAS-Dep vs KRAS-Ind cells. **C)** Cancer specific pathways activated in KRAS-Dep vs KRAS-Ind cell lines correspond to cancers in which activating mutations in RAS oncogenes are present in a significant percentage of cases. Pathways displayed were significant in the SPIA analysis after a FDR correction with global adjusted p- <= 0.05. * The lung pathway had adjusted-pvalue <= 0.17. Data obtained from reference 1. **D)** Confocal microscopy shows high levels of phosphorylated as well as total B-catenin localized on the cell membrane of H441 KRAS-Dep cell line. This localization was not observed in H460 KRAS-Ind cell line. High levels of B-catenin or phospho B-catenin were not observed in the nucleus of those cell lines. The scale bar corresponds to 10um.

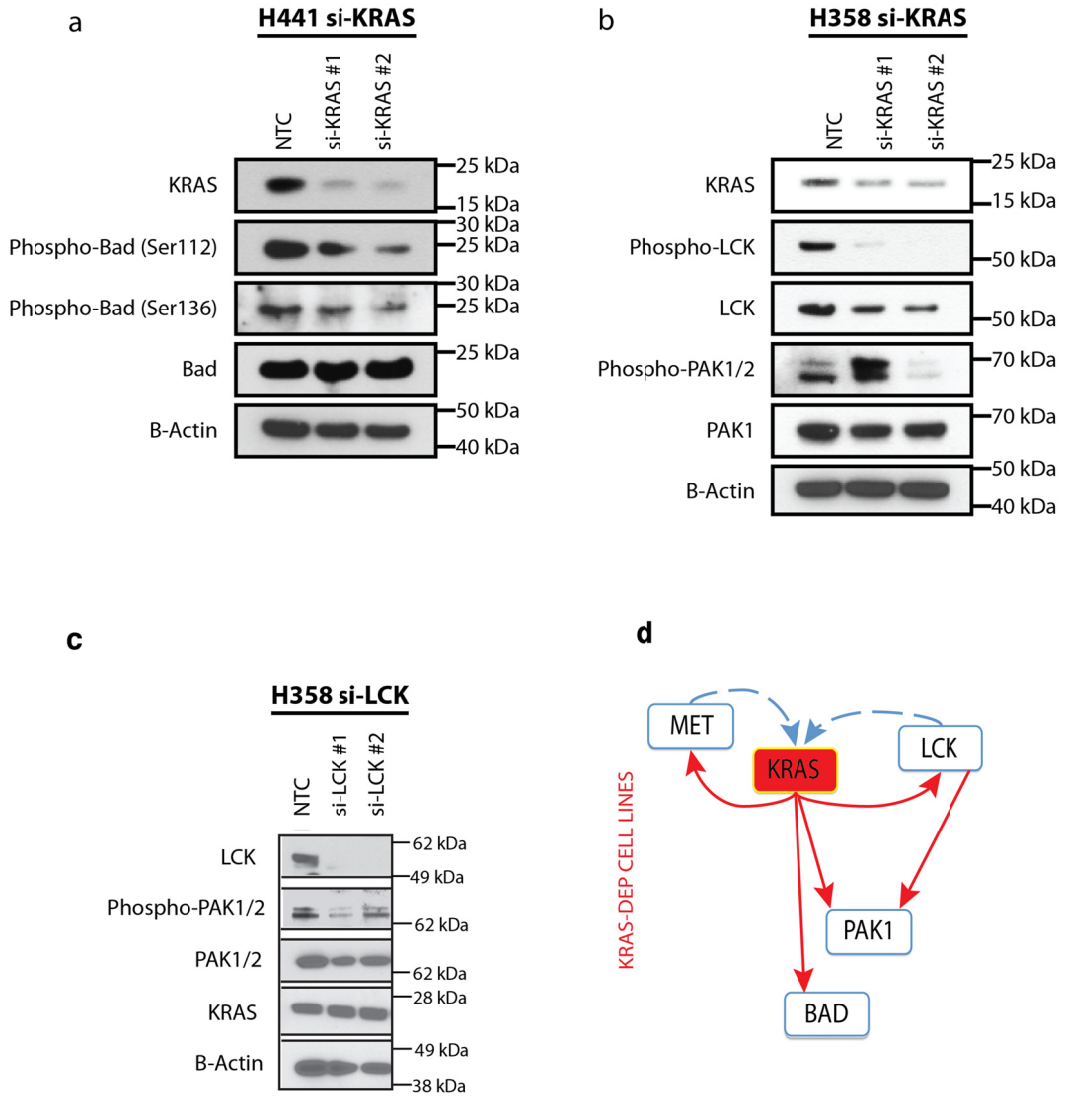


Figure A.5. Feed forward loop between KRAS and LCK.

A) KRAS influences phosphorylation of BAD on sites 112 and 136/ Phosphorylation of BAD-112, BAD136 causes inactivation of this pro-apoptotic protein. KRAS knock down with two independent siRNAs in H441 dependent cell line decreases phosphorylation of BAD (residues 112 and 136), but did not affect BAD protein levels. **B)** KRAS influences total and phosphorylated protein level of potential druggable kinases LCK, PAK1/2 in KRAS-Dep cell lines. Knock down of KRAS with two independent siRNAs reduces phosphorylation levels of LCK, and PAK1/2 in H358 cell line. KRAS-KD also reduced total protein levels of LCK, but not PAK1/2. **C)** LCK influences PAK1/PAK2 activation in KRAS dependent cell lines. Knock down of LCK using two independent siRNAs reduces phosphorylation levels of PAK1/2 but not their protein level in H358 KRAS-Dep cell line. **D)** Simplified model of the signaling pathway characterized in KRAS dependent cell lines. Red arrows represent phosphorylation (activation) events that are influenced by KRAS, and revealed by this study as present in NSCLC KRAS dependent cell lines. Blue arrows represent potential

existent interactions between MET, LCK, PAK1 and BAD that were previously known but not assessed in this study.

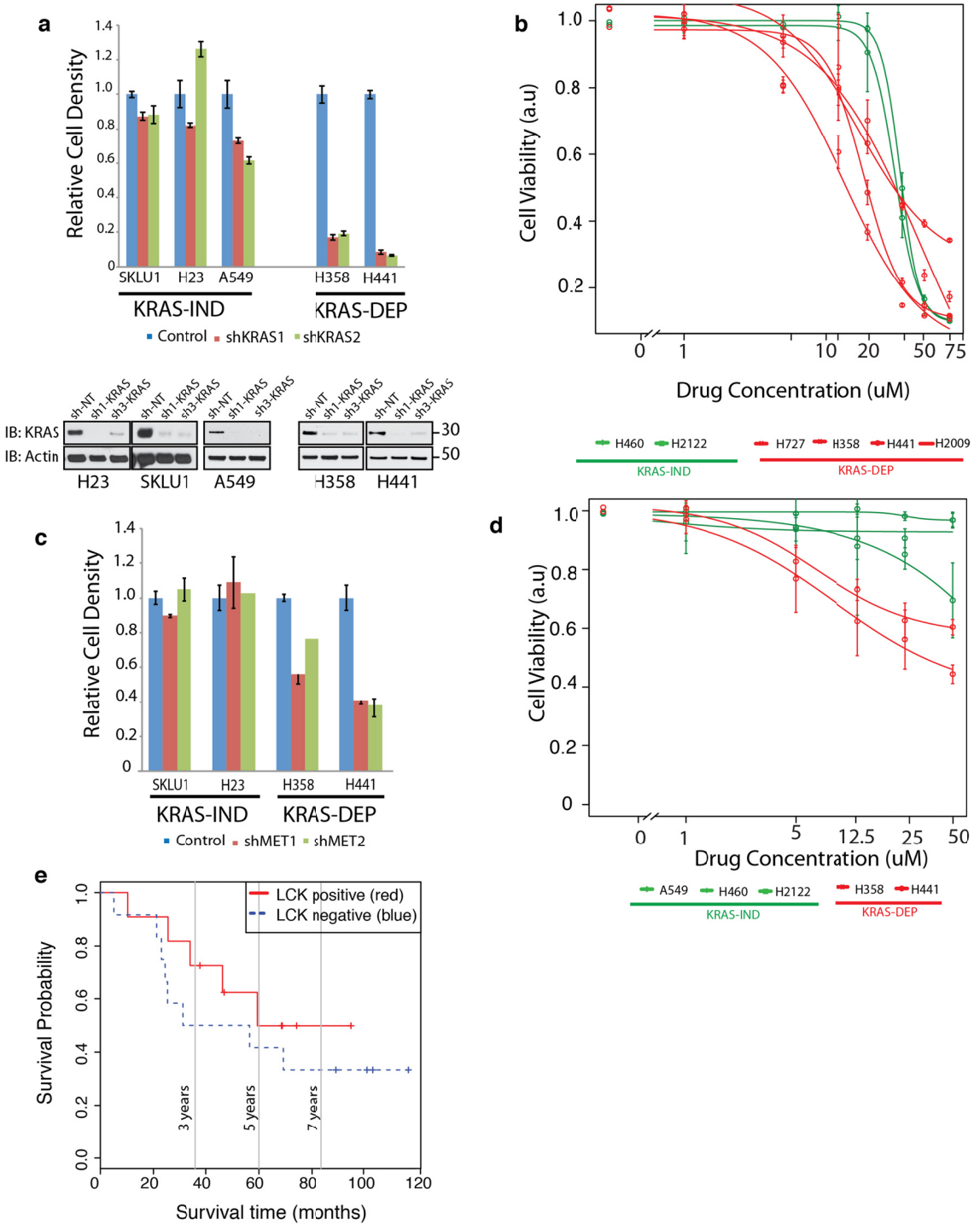


Figure A.6. KRAS dependent cells would be dependent on LCK for survival.

A) KRAS knock down impairs proliferation in NSCLC KRA-DEP but not in KRAS-Ind cell lines. Two independent lentivirus shRNAs significantly decrease cell proliferation in KRAS-Dep cell lines but not in KRAS-Ind ones. The error bars correspond to the standard error calculated over three independent replicates. **B)** Inhibition of LCK using small molecule inhibitor (LCK Inhibitor II, Millipore, CAS 918870-43-6), preferentially impaired cell proliferation in KRAD-Dep (red) but not in KRAS-Ind cell lines (green). **C)** MET knock down impairs proliferation in NSCLC KRAS-Dep but not in KRAS-Ind cell lines mimicking the effect observed by KRAS knock down. The error bars correspond to the standard error calculated over three independent replicates. **D)** Inhibition of MET using small molecule inhibitor selectively, but mildly, decreases cell proliferation in KRAS-Dep (H441, H358) but not in KRAS-Ind (H460, H2122, A549) cell lines. **E)** Analysis of LCK staining with respect to patient overall survival. LCK positive samples = 11, LCK negative samples = 12. We observed the largest difference in survival probability between LCK-positive and LCK-negative patients at 3 years after diagnosis. At this point in time, the survival probability of LKC positive samples (KRAS-Dep) is above 75% while only 50% for the LCK negative (KRAS-Ind) samples as shown in the figure below. The Chi-square test p-value for the difference in the survival probability at 3 years is p=0.23. Moreover, the overall survival curves for LCK positive and LCK negative are not statistically significant over the full course of time (Chi-square test p= 0.379). This is not surprising given the small number of samples available for the analysis, which translates in low power for detecting differences in survival time, and the overall low survival rates of all lung cancer patients.

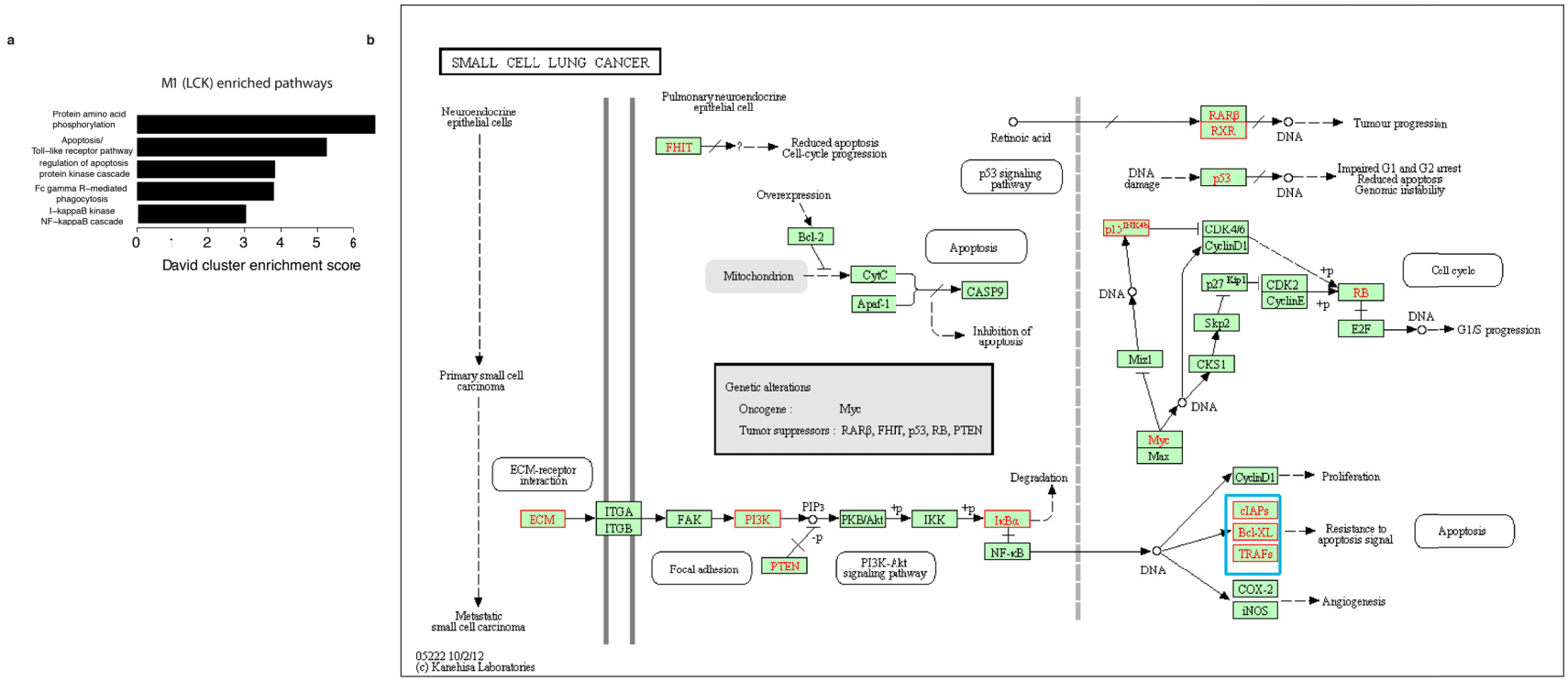


Figure A.7. LCK module would be involved in modulating apoptosis.

A) Module M1 containing LCK and PAK1 is enriched on proteins belonging to the apoptosis or regulation of apoptosis pathways. David analysis of the proteins belonging to M1 reconstructed by the PCST algorithm. The David cluster enrichment score is plotted for all pathways with a Bonferroni corrected pvalue <=0.05. **B)** Small Cell lung cancer pathway enriched using the gene expression signature obtained after knock down of LCK in H441 and H358 cell lines. In blue box, a small module of proteins (TRAF1, BIRC3 and BCL2L1) controlling apoptosis is highlighted.

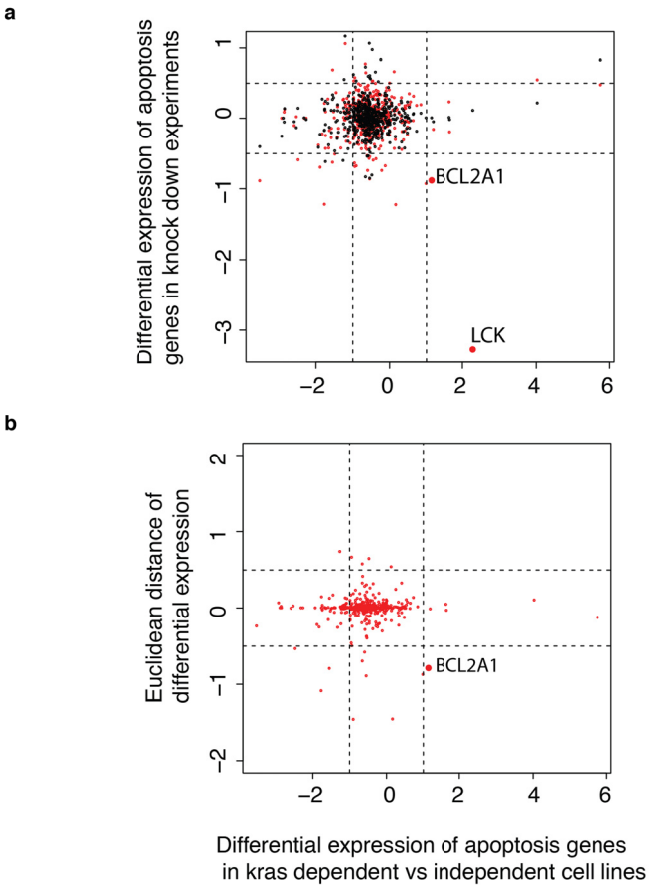


Figure A.8. BLC2A1, an apoptosis gene that is specifically regulated by LCK in NSCLC KRAS-Dep cell lines.

A) BCL2A1 nomination. The x-axis shows the differential expression of apoptosis genes when comparing KRAS-Dep vs KRAS-Ind cell lines. The y axis shows the average differential expression of the same genes when comparing a siRNA knock down of LCK in H441 and H358 cell lines with resect to the targeting control (red dots), or the average differential expression when comparing the knock down of a "random" gene compared to its respective control (black dots) in three unrelated prostate cell lines. Genes affected by the overall siRNA treatment, but not specifically by LCK itself would by overlapping or very close in this plot. **B)** BLC2A1 is an apoptosis gene specifically affected by LCK in NSCLC KRAS-Dep cells. The y axis represents the Euclidean distance between a red and black dot representing the same gene in the Supplementary Figure S7C. Genes that are specifically affected by LCK have positive or negative Euclidean distances according to the magnitude of their perturbation, while genes nonspecifically affected by the siRNA treatment would have Euclidean distances close to 0. LCK was left out of the plot for convenience of the scale.

Table A-1 SPIA analysis on the differentially abundant proteins identified by the S score. Status A=Activated, I=Inhibited

	KEGG ID	Size	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Tight junction	4530	121	8	7.4E-06	1.4E+01	2.3E-01	1.6E-04	7.5E-03	1.3E-02	A
Epithelial cell signaling in bacterial infection	5120	63	6	1.3E-05	1.8E+01	2.8E-01	3.5E-04	7.5E-03	2.7E-02	A
Focal adhesion	4510	193	10	4.6E-06	4.6E+00	9.1E-01	1.5E-02	7.7E-02	1.0E+00	A
Thyroid cancer	5216	29	4	9.2E-05	2.3E+01	1.7E-01	4.3E-04	7.5E-03	3.3E-02	A
Pathways in cancer	5200	312	9	1.2E-03	8.8E+01	5.2E-02	4.8E-04	7.5E-03	3.8E-02	A
ARVC	5412	69	3	2.0E-02	1.2E+01	4.0E-03	4.3E-04	7.5E-03	3.4E-02	A
Bacterial invasion of epithelial cells	5100	64	5	1.9E-04	1.6E+01	5.9E-01	9.4E-03	5.7E-02	7.4E-01	I
Colorectal cancer	5210	61	3	1.4E-02	2.9E+01	1.3E-02	8.9E-04	1.1E-02	7.0E-02	A
ECM-receptor interaction	4512	83	4	5.0E-03	2.0E+01	3.7E-02	1.0E-03	1.1E-02	8.0E-02	I
Endometrial cancer	5213	51	4	8.4E-04	2.1E+01	4.1E-01	8.7E-03	5.6E-02	6.8E-01	A
Osteoclast differentiation	4380	125	5	3.9E-03	2.8E+01	1.0E-01	2.8E-03	2.7E-02	2.2E-01	A
Fc epsilon RI signaling pathway	4664	73	4	3.2E-03	3.6E+01	1.7E-01	4.5E-03	3.9E-02	3.5E-01	A
Vascular smooth muscle contraction	4270	104	2	2.1E-01	5.8E+01	4.0E-03	7.4E-03	5.3E-02	5.8E-01	A
Natural killer cell mediated cytotoxicity	4650	120	3	7.8E-02	1.1E+02	1.3E-02	5.0E-03	3.9E-02	3.9E-01	A
Fc gamma R-mediated phagocytosis	4666	87	4	5.9E-03	2.6E+01	2.3E-01	1.0E-02	5.8E-02	8.2E-01	A
ErbB signaling pathway	4012	83	4	5.0E-03	2.5E+01	3.9E-01	2.2E-02	9.9E-02	1.0E+00	A
Wnt signaling pathway	4310	135	3	1.0E-01	2.5E+01	4.0E-02	1.6E-02	8.0E-02	1.0E+00	A
Renal cell carcinoma	5211	70	2	1.2E-01	2.4E+01	5.2E-02	2.3E-02	9.9E-02	1.0E+00	A
Cell cycle	4110	112	4	1.4E-02	1.5E+01	4.4E-01	4.9E-02	1.8E-01	1.0E+00	I

Table A-2 Data provenance for chapter 2.

DataSet	Provenance	Used in this study in section:	Availability	Notes
Proteome data for all NSCLC cancer cell lines	Generated by this study	Integration of omics dataset to nominate actionable proteins. Figure 1	Supplementary Tables ST5, ST8	
Phospho-Proteome data for all NSCLC cancer cell lines	Generated by this study	Integration of omics dataset to nominate actionable proteins. Figure 1	Supplementary Tables ST6, ST9	
Gene expression data for all NSCLC cencer cell lines	Sanger Cell Line Project. BROAD Institute.	Integration of omics dataset to nominate actionable proteins. Figure 1	http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-783/	
Gene expression data for Outlier analysis	Sanger Cell Line Project. BROAD Institute.	KRAS-Dep NSCLC cell lines are also LCK-Dep for proliferation. Outlier expression analysis Figure 5a	http://www.ebi.ac.uk/arrayexpress/files/E-MTAB-783/	
Microarray data for LCK-KD H441, H358	Generated by this study	LCK could be associated with apoptosis pathways. Supplementary Figure 7, 8. Supp Table ST3	Available upon request	
Microarray data for MET-KD H441, H358	Generated by this study	LCK could be associated with apoptosis pathways. Supplementary Figure 7, 8. Supp Table ST3	Available upon request	
TMA with genotype information for KRAS	Generated by this study	LCK activation is observed in clinical samples.	Supplementary Table ST2. Available upon request	
Databases	**Provenance**	**Used in this study in section:**	**Availability**	**Notes**

KEGG database	Kyoto University Bioinformatics Center	Used for Network analysis with the SPIA and PCST algorithms	www.genome.jp/kegg-bin/download	* Download was free at the time of this study, but it is not anymore
STRING Database		Used for Network analysis with PCST algorithm in order to obtain the weigth of each protein-protein interaction	http://string-db.org	
Informative Genes	Generated by this Study. Sanger cancer concensus genes, List of Kinases, common genes involve in genomic rearreagments	Used for classifying each gene as informative or not in the integration of omics datasets.	Supplementary Tables ST4	
Source Code	**Provenance**	**Used in this study in section:**	**Availability**	**Notes**
X!Tandem	The global proteome machine	mzXML search	www.thegpm.org/tandem/	
PeptideProphet and ProteinProphet	Transproteomic Pipeline	Post-processing of X!Tandem Searches	tools.proteomecenter.org/TPP.php	
Abacus	Nesvizhskii Lab Universit of Michigan	Aggregation and summarization of spectral counts for each protein and phosphoprotein across all cell lines	nesvilab.org/software	
SPIA	Tarca, A.L. et al. A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75-82 (2009)	Network Analysis	http://bioconductor.org/biocLite.R, biocLite("SPIA")	
MSGSTEINER	Bailly-Bechet, M. et al. Finding undetected protein associations in cell signaling by belief propagation. Proceedings of the National Academy of Sciences of the United States of America 108, 882-887 (2011)	Prize Collecting Steiner Tree Algorithm solution	http://areeweb.polito.it/ricerca/cmp/code/bpsteiner	

KEGGGraph	Zhang and Wiemann, KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor. Bioinformatics 2009, 1.	Merging KEEG pathways in order to create the meta-pathway (G)	http://www.bioconductor.org/packages/2.11/bioc/html/KEGGgraph.html	

Table A-3 Mutation status of the cell lines used in this study

Cell line	KRAS_STATUS	KRAS_MUT	KRAS	NRAS	HRAS	PIK3CA	EGFR	BRAF	TP53	TP53_p.t
SW900	KRAS_IND	p.G12V,c.35G>T	1	0	0	0	0	0	0	1
A549	KRAS_IND	p.G12S,c.34G>A	1	0	0	0	0	0	0	0
H460	KRAS_IND	p.Q61H c.183A>T	1	0	0	1	0	0	0	0
H2122	KRAS_IND	p.G12C,c.34G>T	1	0	0	0	0	0	1	0
SK-LU-1	KRAS_IND	p.G12D,c.35G>A	1	0	0	0	0	0	1	0
H1792	KRAS_IND	p.G12C,c.34G>T	1	0	0	0	0	0	0	1
H23	KRAS_IND	p.G12C c.34G>T	1	0	0	0	0	0	1	0
H1155	KRAS_IND	p.Q61H,c.183A>T	1	0	0	0	0	0	1	0
H1734	KRAS_DEP	p.G13C,c.37G>T	1	0	0	0	0	0	1	0
H2009	KRAS_DEP	p.G12A,c.35G>C	1	0	0	0	0	0	1	0
H358	KRAS_DEP	p.G12C c.34G>T	1	0	0	0	0	0	0	0
H441	KRAS_DEP	p.G12V,c.35G>T	1	0	0	0	0	0	1	0
H727	KRAS_DEP	p.G12V,c.35G>T	1	0	0	0	0	0	1	0

Table A-4 TMA KRAS genotype and IHC pLCK staining.

ID	Sample	Exon 1 Genotype	Zygosity	Exon 2 Genotype	Zygosity	pLCK_Antibody_1500
1	C004	12: GGT -> TGT	Heterozygous	WT	Homozygous	1
2	C006	12: GGT -> GAT	Heterozygous	WT	Homozygous	1
3	C008	12: GGT -> GAT	Heterozygous	WT	Homozygous	1
4	C016	12: GGT -> GTT	Heterozygous	WT	Homozygous	1
5	C026	12: GGT -> TGT	Heterozygous	WT	Homozygous	1
6	C035	12: GGT -> TGT	Heterozygous	WT	Homozygous	1
7	C038	12: GGT -> TGT	Heterozygous	WT	Homozygous	1
8	C045	12: GGT -> GTT	Heterozygous	WT	Homozygous	1
9	C046*	WT	Homozygous	WT	Homozygous	1
10	C053	12: GGT -> GTT	Heterozygous	WT	Homozygous	1
11	C067	12: GGT -> GTT	Homozygous	WT	Homozygous	1
12	C087	12: GGT -> TGT	Heterozygous	WT	Homozygous	1
13	C112	12: GGT -> GAT	Heterozygous	WT	Homozygous	1
14	C113	13: GGC -> GAC	Heterozygous	WT	Homozygous	1
15	C116	12: GGT -> GAT	Heterozygous	WT	Homozygous	1
16	C117	WT	Homozygous	34: G -> T	Heterozygous	1
17	C081	WT	Homozygous	34: G -> T	Heterozygous	1
18	C083	12: GGT -> AGT	Heterozygous	WT	Homozygous	-1
19	C082	12: GGT -> GTT	Heterozygous	WT	Homozygous	-1
20	C037	12: GGT -> TGT	Heterozygous	WT	Homozygous	-1
21	C071	12: GGT -> TGT	Heterozygous	WT	Homozygous	-1
22	C096	12: GGT -> TGT	Heterozygous	WT	Homozygous	-1

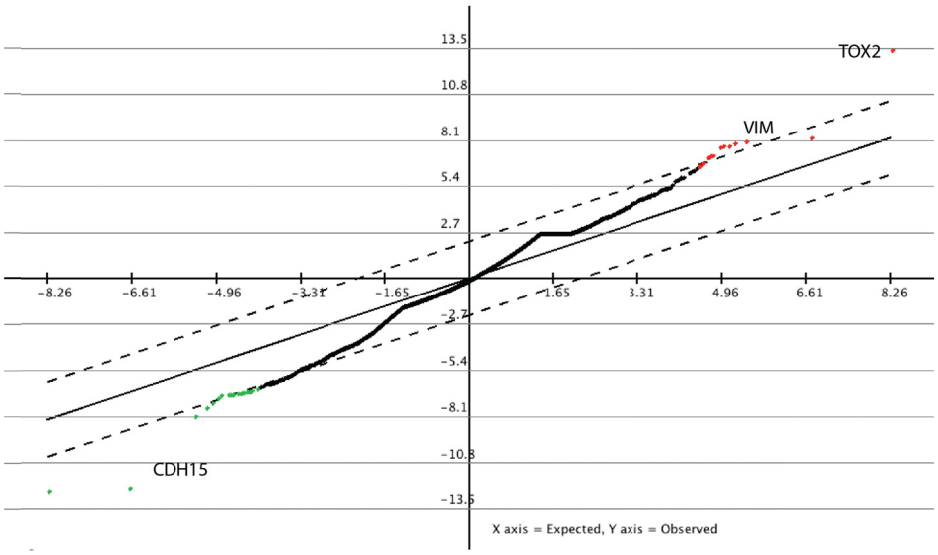
23	C021	12: GGT -> GCT	Heterozygous	WT	Homozygous	-1
24	C047	12: GGT -> TGT	Heterozygous	WT	Homozygous	-1
25	C062	12: GGT -> TGT	Heterozygous	WT	Homozygous	-1
26	C084	WT	Homozygous	34: G -> T	Heterozygous	-1
27	C033	WT	Homozygous	35: G -> A	Heterozygous	-1
28	C040	WT	Homozygous	35: G -> C	Heterozygous	-1
29	C058	WT	Homozygous	34: G -> A	Heterozygous	-1
11	C067	12: GGT -> GTT	Homozygous	WT	Homozygous	1

Table A-5 Differentially activated pathways determined by the SPIA algorithm after knock down of LCK or MET. Status A=Activated, I= Inhibited.

	KEGGID	Size	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Rheumatoid arthritis	5323	90	16	1.4E-08	0.0E+00	1.0E+00	2.8E-07	1.4E-05	2.9E-05	I
Amoebiasis	5146	104	13	1.9E-05	-3.2E+00	6.1E-01	1.4E-04	3.0E-03	1.5E-02	I
Malaria	5144	51	8	1.5E-04	2.2E+00	3.3E-01	5.5E-04	8.3E-03	5.8E-02	A
Lysosome	4142	119	13	7.9E-05	0.0E+00	1.0E+00	8.2E-04	1.1E-02	8.6E-02	I
Natural killer cell mediated cytotoxicity	4650	122	8	3.5E-02	-8.9E+01	3.0E-03	1.1E-03	1.2E-02	1.1E-01	I
NOD-like receptor signaling pathway	4621	53	7	1.1E-03	-8.1E+00	3.5E-01	3.6E-03	3.1E-02	3.8E-01	I
Axon guidance	4360	127	10	6.0E-03	1.8E+01	7.5E-02	3.9E-03	3.1E-02	4.1E-01	A
Vibrio cholerae infection	5110	53	6	5.6E-03	-4.6E+00	1.0E-01	4.8E-03	3.4E-02	5.0E-01	I
Small cell lung cancer	5222	83	8	4.0E-03	-1.8E+01	1.8E-01	6.0E-03	3.7E-02	6.3E-01	I
African trypanosomiasis	5143	32	5	2.8E-03	2.2E+00	4.8E-01	1.0E-02	5.8E-02	1.0E+00	A
Antigen processing and presentation	4612	71	7	6.2E-03	-3.0E+00	3.7E-01	1.6E-02	8.4E-02	1.0E+00	I
Bile secretion	4976	69	7	5.3E-03	-2.6E+00	3.0E-01	1.2E-02	6.5E-02	1.0E+00	I
Toll-like receptor signaling pathway	4620	95	9	2.6E-03	1.6E+01	2.8E-01	6.0E-03	3.7E-02	6.3E-01	A
Specific pathways associated with MET Knock down										
B cell receptor signaling pathway	4662	72	8	8.0E-04	-8.7E+00	3.5E-01	2.5E-03	4.3E-02	2.6E-01	I
Renal cell carcinoma	5211	69	5	4.1E-02	-1.8E+01	1.3E-02	4.6E-03	5.2E-02	4.7E-01	I
ErbB signaling pathway	4012	83	8	2.0E-03	-1.6E+01	3.9E-01	6.4E-03	6.5E-02	6.5E-01	I
Focal adhesion	4510	192	12	7.1E-03	-2.6E+01	1.7E-01	9.3E-03	8.6E-02	9.5E-01	I
Bile secretion	4976	69	9	1.1E-04	2.2E+00	5.6E-01	6.6E-04	1.3E-02	6.7E-02	A

Appendix B
Additional analyses for chapter 3

A



B

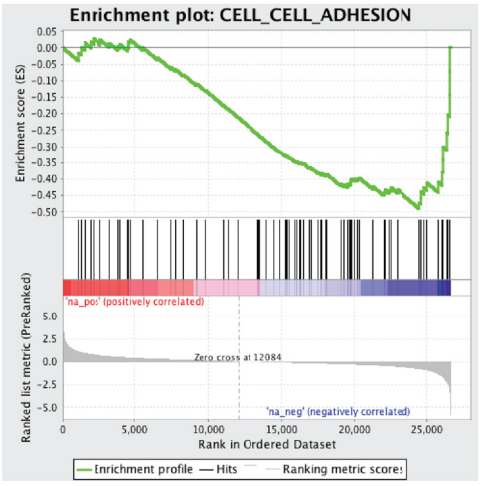


Figure B.1. Significant Analysis of microarrays (A) and Gene Set Enrichment Analysis for the expression of the fusion construct CD74-NRG1 in BEAS-2B cells (B).

Table B-1.Fusions used as true positives for the random forest classifier.

Other attributes of each fusion such as functional annotation and gene expression of both partners are not shown.

Patient	5'Chr	3'Chr	3'Gene	5'Gene	Cohort	5'exon	3'exon	5'AlignQ	3'AlignQ	AlignQS	SpanReads	MatePairs	Encompassing	FusionType	PCR Validated
H441	2	2	EIF2AK2	SULT6B1	umich	14	2	9	9	18	355	408	132	TD	TRUE
H1793	21	12	RUNX1	PTPRR	umich	5	13	9	8	17	10	5	9	InterC	TRUE
H1734	17	17	MRC2	MAP3K3	umich	1	2	9	9	18	63	34	68	IntraC	TRUE
H1734	12	14	FAM60A	DPF3	umich	6	8	9	9	18	268	18	1	InterC	TRUE
C057	9	6	DAPK1	GMDS	umich	2	7	5	5	10	5	6	3	InterC	TRUE
C011	5	5	TTC1	DOCK2	umich	2	28	9	9	18	20	11	21	IntraC	TRUE
A63	9	19	TSC1	SMARCA4	umich	16	12	3	4	7	4	1	5	InterC	TRUE
A35	5	8	CD74	NRG1	umich	3	6	9	9	18	275	139	150	InterC	TRUE
A35	5	8	CD74	NRG1	umich	5	6	8	5	13	9	139	2	InterC	TRUE
A34	3	3	RAF1	TMEM40	umich	6	11	9	9	18	22	4	7	TD	TRUE
A25	5	5	SLC12A7	TERT	umich	3	12	9	9	18	56	50	41	TD	TRUE
A25	9	20	CDK9	AHCY	umich	5	9	2	3	5	3	7	7	InterC	TRUE
A25	9	20	CDK9	AHCY	umich	3	9	1	1	2	1	7	1	InterC	TRUE
lc_s51	X	X	EDA	MID1	seoul	1	7	8	8	16	9	1	3	IntraC	NA
lc_s48	4	6	SLC34A2	ROS1	seoul	13	12	9	9	18	390	35	203	InterC	NA
lc_s42	10	10	KIF5B	RET	seoul	12	12	9	9	18	56	3	23	IntraC	NA

lc_s39	5	6	CD74	ROS1	seoul	3	10	9	9	18	440	19	372	InterC	NA
lc_s39	5	6	CD74	ROS1	seoul	3	9	2	2	4	3	19	2	InterC	NA
lc_s26	2	2	MAP4K3	PRKCE	seoul	34	2	7	8	15	8	1	6	IntraC	NA
lc_s26	2	2	EML4	ALK	seoul	13	10	9	9	18	27	4	30	IntraC	NA
lc_s20	17	17	BCAS3	MAP3K3	seoul	22	2	5	5	10	5	1	7	IntraC	NA
lc_s13	10	12	FGFR2	CIT	seoul	2	25	9	9	18	25	5	23	InterC	NA
a8d6694c-a213-	10	10	CCDC6	RET	tcga	9	12	5	2	7	8	17	2	IntraC	NA
36bf02f8-c1c8-	4	6	SLC34A2	ROS1	tcga	13	12	9	0	9	86	2	44	InterC	NA
028e99e9-5b9a-	6	6	EZR	ROS1	tcga	5	10	8	9	17	31	223	16	IntraC	NA
H441	2	2	EIF2AK2	SULT6B1	umich	14	2	9	9	18	355	408	132	TD	TRUE

Table B-2 Comparison of the number of fusions among different tumor stages in LUAD.

Student t-test p-values.

LUAD	Stage I	Stage II	Stage III	Stage IV
Stage I		0.2937	0.7833	0.1472
Stage II			0.2194	0.04902
Stage III				0.2008

Table B-3 Comparison of the number of fusions among different tumor stages in LUSC.

Student t-test p-values.

LUSC	Stage I	Stage II	Stage III	Stage IV
Stage I		0.01409	0.7258	0.09339
Stage II			0.06015	0.01956
Stage III				0.07492

Table B-4 3' Representative fusions recurrence across the combined cohort and in the driver positive and driver negative samples.

The full list includes 400 unique 3'genes.

3'gene	# Rec Samples	In Driver +	In Driver -	Overall recurrence	Driver + recurrence	Driver - recurrence	Total number samples	Driver +	Driver -
ROS1	6	0	5	0.857	0.000	1.295	700	314	386
NRG1	3	0	3	0.429	0.000	0.777	700	314	386
MAP3K3	2	0	1	0.286	0.000	0.259	700	314	386
RET	2	0	2	0.286	0.000	0.518	700	314	386
ALK	1	0	1	0.143	0.000	0.259	700	314	386
TERT	1	1	0	0.143	0.318	0.000	700	314	386
DZIP1	2	0	0	0.286	0.000	0.000	700	314	386
WWOX	2	0	0	0.286	0.000	0.000	700	314	386
ABCC5	2	0	2	0.286	0.000	0.518	700	314	386
C1orf22	2	1	0	0.286	0.318	0.000	700	314	386
LILRB2	2	1	1	0.286	0.318	0.259	700	314	386
RABGAP	2	0	0	0.286	0.000	0.000	700	314	386
ZNF585	2	0	0	0.286	0.000	0.000	700	314	386
AFF3	2	1	0	0.286	0.318	0.000	700	314	386
PEMT	2	2	0	0.286	0.637	0.000	700	314	386
FGFR3	2	0	2	0.286	0.000	0.518	700	314	386
PSMD11	2	1	1	0.286	0.318	0.259	700	314	386

Table B-5 3' Representative fusions recurrence across the combined cohort and in the driver positive and driver negative samples.

The full list includes 391 unique 5'genes.

5'gene	# Rec Samples	Overall recurrence	Driver + recurrence	Driver - recurrence	Total number samples	Driver +	Driver -
SLC34A2	3	0.429	0.955	0.777	700	314	386
MYH9	3	0.429	0.955	0.777	700	314	386
TXNRD1	3	0.429	0.955	0.777	700	314	386
GPR98	3	0.429	0.955	0.777	700	314	386
DAPK1	2	0.286	0.637	0.518	700	314	386
CD74	2	0.286	0.637	0.518	700	314	386
RAF1	2	0.286	0.637	0.518	700	314	386
SLC12A7	2	0.286	0.637	0.518	700	314	386
CCDC6	2	0.286	0.637	0.518	700	314	386
UCHL5	2	0.286	0.637	0.518	700	314	386
PPP1CC	2	0.286	0.637	0.518	700	314	386
FOXK2	2	0.286	0.637	0.518	700	314	386
POLD3	2	0.286	0.637	0.518	700	314	386
SAMD12	2	0.286	0.637	0.518	700	314	386
PTPN14	2	0.286	0.637	0.518	700	314	386

Appendix C
Additional analyses for chapter 4

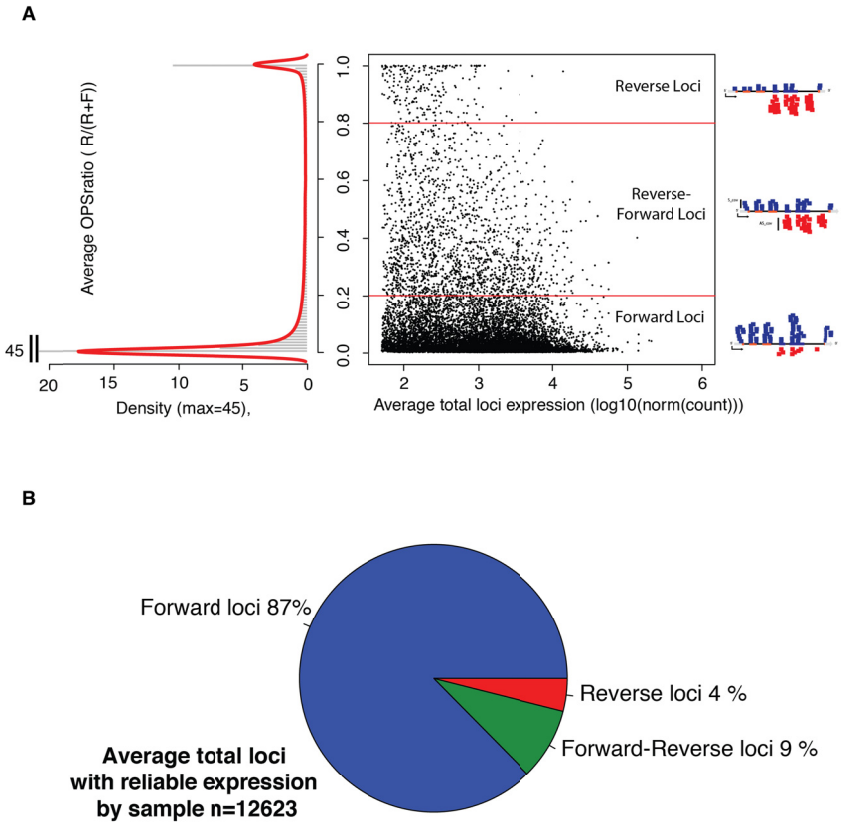


Figure C.1. Average OPSratio for all loci across the cohort.

A) Average loci OPSratio vs loci total expression. **B)** Distribution of loci according to thei OPSratio.

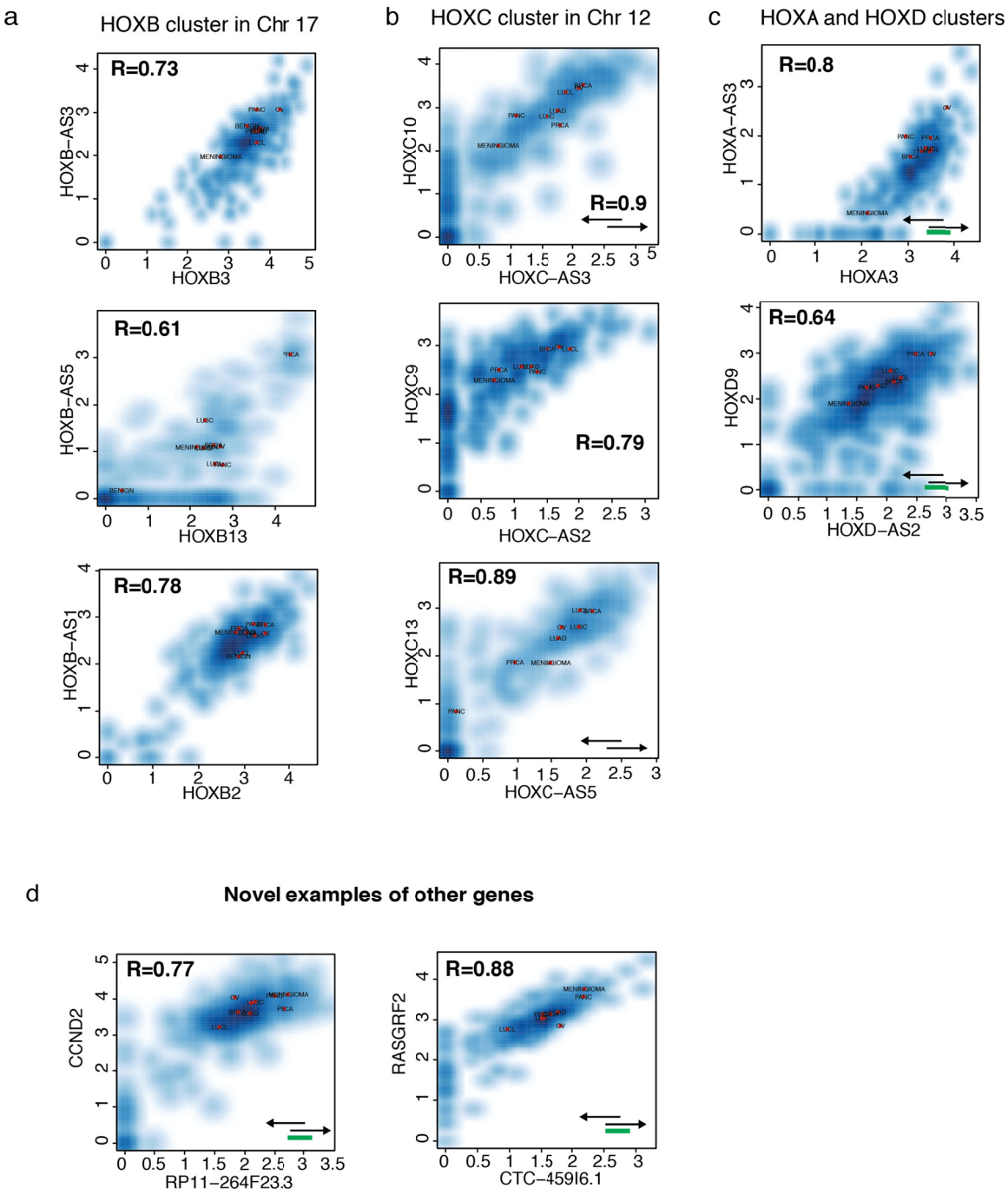


Figure C.2. Gene expression correlation plots HTH-cisNAT pais.

A) HOXB cluster examples. **B)** HOXC cluster examples. **C)** Additional HOXA and HOXD examples. **D)** Additional genes that are not homebox genes.

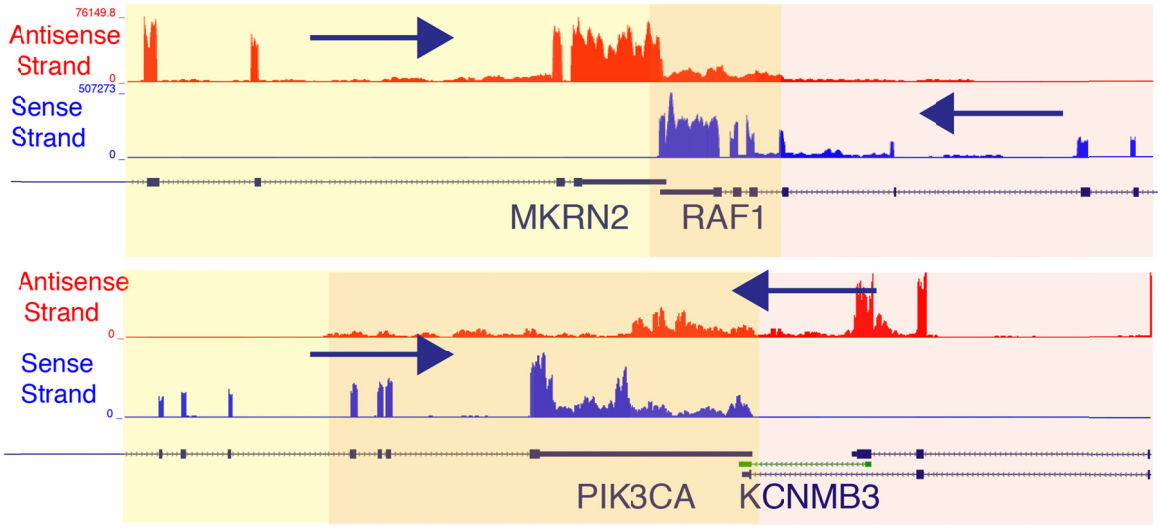


Figure C.3. Example of TTT ubiquitous genes.

Top: MRKN2 3UTR runs into RAF1 3UTR and gene body. Sum coverage map over the lung adenocarcinoma cohort (LUAD)**. Bottom:** KCNMB3 3UTR runs into PIK3CA 3UTR and gene body. Sum coverage map over the lung adenocarcinoma cohort (LUAD).

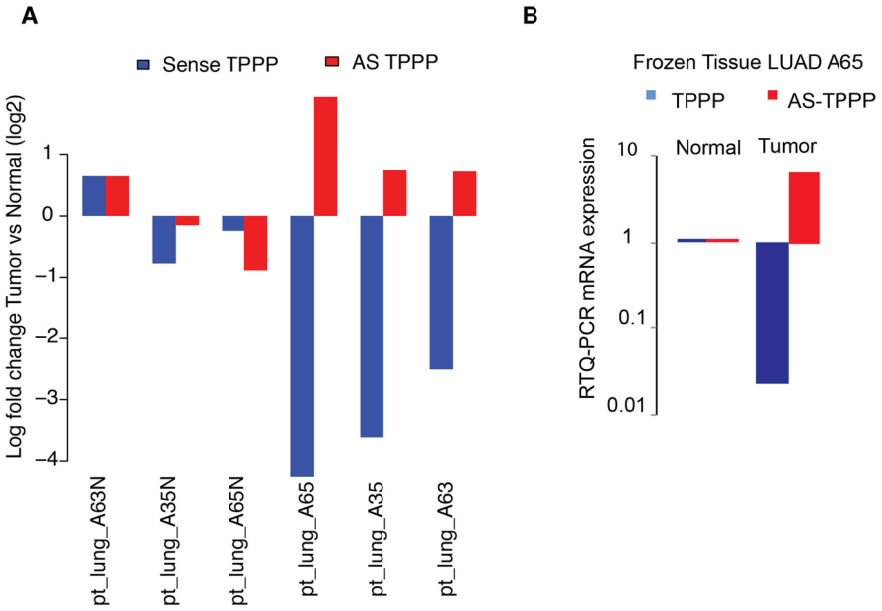


Figure C.4. Log fold change between tumors and normals Inconsistent genes.

A) TPPP inconsistent pair, sense and antisense RNASeq log fold change between tumor and the mean of the normal samples. The barplot shows sense and antisense expression changing in opposite directions for this gene. **B)** An RTQ-PCR validation in one of our match tumor-normal pairs.

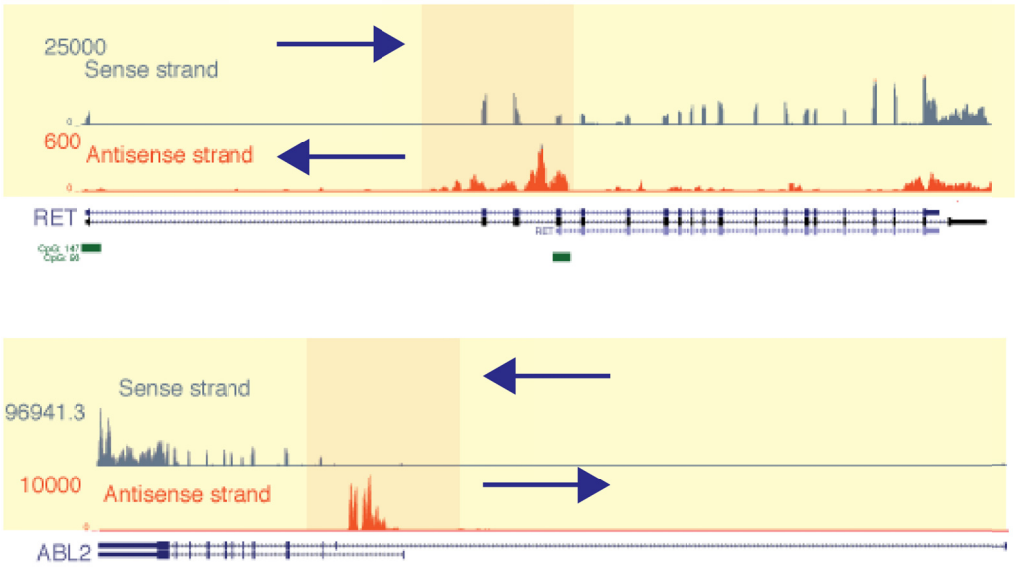


Figure C.5. Coverage maps for embedded unannotated antisense transcripts.

Table C-1 Comparison of gene expression correlation distribution for different configurations.

The head-to-head configuration is the one with highest correlation among all other ones. Student t-test p-values. The comparison are directional and therefore not symmetric.

	Tail-to-Tail	Head-to-Head	Embedded	Random pairs
Tail-to-Tail		1	0.9999	2.73E-272
Head-to-Head	1.23E-144		4.16E-42	0
Embedded	1.38E-05	1		2.03E-149

Bibliography

1 Society, C. Cancer Statistics in the United States 2009., (2009).

2 Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075, doi:10.1038/nature07423 (2008).

3 Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

4 Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561-566, doi:10.1038/nature05945 (2007).

5 Inamura, K. *et al.* EML4-ALK lung cancers are characterized by rare other mutations, a TTF-1 cell lineage, an acinar histology, and young onset. *Mod Pathol* **22**, 508-515, doi:10.1038/modpathol.2009.2 (2009).

6 Rikova, K. *et al.* Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131**, 1190-1203, doi:10.1016/j.cell.2007.11.025 (2007).

7 Ju, Y. S. *et al.* A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Research* **22**, 436-445, doi:10.1101/gr.133645.111 (2012).

8 Takeuchi, K. *et al.* RET, ROS1 and ALK fusions in lung cancer. *Nature Medicine* **18**, 378-381, doi:10.1038/nm.2658 (2012).

9 Karnoub, A. & Weinberg, R. Ras oncogenes: split personalities. *Nature Reviews Molecular Cell Biology* **9**, 517-531 (2008).

10 Singh, A. *et al.* A Gene Expression Signature Associated with K-Ras Addiction Reveals Regulators of EMT and Tumor Cell Survival. *Cell* **15**, 489-500 (2009).

11 Weinberg, R. A. *The biology of cancer*. (Garland Science, 2007).

12 Loboda, A. *et al.* A gene expression signature of RAS pathway dependence predicts response to PI3K and RAS pathway inhibitors and expands the population of RAS pathway activated tumors. *BMC Medical Genomics* **3**, 26 (2010).

13 Luo, J. *et al.* A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. *Cell* **137**, 835-848, doi:10.1016/j.cell.2009.05.006 (2009).

14 Barbie, D. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112 (2009).

15 Guo, A. *et al.* Signaling networks assembled by oncogenic EGFR and c-Met. *Proceedings of the National Academy of Sciences* **105**, 692-697 (2008).

16 Guha, U. *et al.* Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS. *Proceedings of the National Academy of Sciences* **105**, 14112-14117 (2008).

17 Bertotti, A. *et al.* Only a subset of Met-activated pathways are required to sustain oncogene addiction. *Science signaling* **2**, ra80 (2009).

18 Carretero, J. *et al.* Integrative Genomic and Proteomic Analyses Identify Targets for Lkb1-Deficient Metastatic Lung Tumors. *Cancer Cell* **17**, 547-559.

19 Wu, C.-J. *et al.* A Predictive Phosphorylation Signature of Lung Cancer. *PLoS ONE* **4**, e7994 (2009).

20 Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497-1500, doi:10.1126/science.1099314 (2004).
21 Luo, J. *et al.* A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. *Cell* **137**, 835-848 (2009).
22 Engelman, J. A. *et al.* Effective use of PI3K and MEK inhibitors to treat mutant Kras G12D and PIK3CA H1047R murine lung cancers. *Nat Med* **14**, 1351-1356, doi:nm.1890 [pii] 10.1038/nm.1890 (2008).
23 Seo, J. S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Research* **22**, 2109-2119, doi:10.1101/gr.145144.112 (2012).
24 Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**, 709-715, doi:10.1038/nmeth.1491 (2010).
25 Moghaddas Gholami, A. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep* **4**, 609-620, doi:10.1016/j.celrep.2013.07.018 (2013).
26 Beausoleil, S., Villen, J., Gerber, S., Rush, J. & Gygi, S. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **24**, 1285-1292 (2006).
27 Bodenmiller, B. & Aebersold, R. Vol. 470 317-334 (Elsevier).
28 Choi, H., Fermin, D. & Nesvizhskii, A. Significance Analysis of Spectral Count Data in Label-free Shotgun Proteomics. *Molecular & Cellular Proteomics* **7**, 2373-2385 (2008).
29 Choudhary, C. & Mann, M. Decoding signalling networks by mass spectrometry-based proteomics. *Nature Reviews Molecular Cell Biology* **11**, 427-439, doi:10.1038/nrm2900 (2010).
30 Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nature Biotechnology* **28**, 710-721 (2010).
31 Griffin, N. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nature Biotechnology* **28**, 83-89.
32 Keshamouni, V. *et al.* Temporal Quantitative Proteomics by iTRAQ 2D-LC-MS/MS and Corresponding mRNA Expression Analysis Identify Post-Transcriptional Modulation of Actin-Cytoskeleton Regulators During TGF-Î²-Induced Epithelial-Mesenchymal Transition. *Journal of Proteome Research* **8**, 35-47 (2009).
33 Mueller, L., Brusniak, M.-Y., Mani, D. R. & Aebersold, R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *Journal of proteome research* **7**, 51-61 (2008).
34 Mueller, L. *et al.* SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470-3480 (2007).
35 Rush, J. *et al.* Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature biotechnology* **23**, 94-101 (2005).
36 Schreiber, T., MÃ¤usbacher, N., Breitkopf, S., Grundner-Culemann, K. & Daub, H. Quantitative phosphoproteomics--an emerging key technology in signal-transduction research. *Proteomics* **8**, 4416-4432 (2008).
37 Wong, J., Sullivan, M. & Cagney, G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. *Briefings in bioinformatics* **9**, 156-165 (2008).
38 Zhang, B. *et al.* Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of proteome research* **5**, 2909-2918 (2006).
39 Zhu, W., Smith, J. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics. *Journal of biomedicine & biotechnology* **2010**, 1-7.

40 Rikova, K. *et al.* Global Survey of Phosphotyrosine Signaling Identifies Oncogenic Kinases in Lung Cancer. *Cell* **131**, 1190-1203 (2007).
41 Xie, X. *et al.* A comparative phosphoproteomic analysis of a human tumor metastasis model using a label-free quantitative approach. *Electrophoresis* **31**, 1842-1852 (2010).
42 Mayya, V. *et al.* Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Signal* **2**, ra46, doi:2/84/ra46 [pii]
10.1126/scisignal.2000007 (2009).
43 Geiger, T., Wehner, A., Schaab, C., Cox, J. & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & cellular proteomics : MCP* **11**, M111 014050, doi:10.1074/mcp.M111.014050 (2012).
44 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
45 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
46 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
47 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).
48 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
49 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).
50 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
51 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
52 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).
53 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31**, 46-53, doi:10.1038/nbt.2450 (2013).
54 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
55 Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology* **12**, R72, doi:10.1186/gb-2011-12-8-r72 (2011).
56 Iyer, M. K., Chinnaiyan, A. M. & Maher, C. A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903-2904, doi:10.1093/bioinformatics/btr467 (2011).

57 McPherson, A. *et al.* deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS computational biology* **7**, e1001138, doi:10.1371/journal.pcbi.1001138 (2011).
58 Nesvizhskii, A., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Analytical chemistry* **75**, 4646-4658 (2003).
59 Keller, A., Nesvizhskii, A., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry* **74**, 5383-5392 (2002).
60 Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352, doi:10.1038/nature10983 (2012).
61 Zhang, S., Li, Q., Liu, J. & Zhou, X. J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics* **27**, i401-409, doi:10.1093/bioinformatics/btr206 (2011).
62 Brink-Jensen, K., Bak, S., Jorgensen, K. & Ekstrom, C. T. Integrative analysis of metabolomics and transcriptomics data: a unified model framework to identify underlying system pathways. *PLoS ONE* **8**, e72116, doi:10.1371/journal.pone.0072116 (2013).
63 Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906-2912, doi:10.1093/bioinformatics/btp543 (2009).
64 Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005-1017, doi:10.1016/j.cell.2010.11.013 (2010).
65 Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS ONE* **7**, e35236, doi:10.1371/journal.pone.0035236 (2012).
66 IBM, C. U. a. *DREAM5 Network inference challenge*, 2010).
67 Chari, R., Coe, B., Vucic, E., Lockwood, W. & Lam, W. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology* **4**, 67 (2010).
68 Gatza, M. *et al.* A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences* **107**, 6994-6999.
69 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-i245, doi:10.1093/bioinformatics/btq182 (2010).
70 Balbin, O. A. *et al.* Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat Commun* **4**, 2617, doi:10.1038/ncomms3617 (2013).
71 Lopes, T. J. S. *et al.* Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases. *Bioinformatics* **27**, 2414-2421, doi:10.1093/bioinformatics/btr414 (2011).
72 Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Molecular systems biology* **3**, 140, doi:10.1038/msb4100180 (2007).
73 Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* **22**, 398-406, doi:10.1101/gr.125567.111 (2012).

74 Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* **21**, 1109-1121, doi:10.1101/gr.118992.110 (2011).

75 Huang, S. s. C. & Fraenkel, E. Integrating Proteomic, Transcriptional, and Interactome Data Reveals Hidden Components of Signaling and Regulatory Networks. *Science signaling* **2**, ra40-ra40, doi:10.1126/scisignal.2000350 (2009).

76 Goldenberg, A., Mostafavi, S., Quon, G., Boutros, P. C. & Morris, Q. D. Unsupervised detection of genes of influence in lung cancer using biological networks. *Bioinformatics* **27**, 3166-3172, doi:10.1093/bioinformatics/btr533 (2011).

77 Huang, S.-s. & Fraenkel, E. Integrating Proteomic, Transcriptional, and Interactome Data Reveals Hidden Components of Signaling and Regulatory Networks. *Sci. Signal.* **2**, ra40 (2009).

78 Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Molecular systems biology* **3** (2007).

79 Chang, J. *et al.* A Genomic Strategy to Elucidate Modules of Oncogenic Pathway Signaling Networks. *Molecular Cell* **34**, 104-114 (2009).

80 Rhodes, D. *et al.* Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology* **23**, 951-959 (2005).

81 Bebek, G. & Yang, J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC bioinformatics* **8**, 335 (2007).

82 Bild, A. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353-357 (2005).

83 Chu, L.-H. & Chen, B.-S. Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC systems biology* **2**, 56 (2008).

84 Dittrich, M., Klau, G., Rosenwald, A., Dandekar, T. & Muller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24**, i223-231 (2008).

85 Keller, A. *et al.* A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics (Oxford, England)* **25**, 2787-2794 (2009).

86 Klammer, M., Godl, K., Tebbe, A. & Schaab, C. Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC bioinformatics* **11**, 351.

87 Missiuro, P. V. *et al.* Information flow analysis of interactome networks. *PLoS computational biology* **5**, e1000350 (2009).

88 Taylor, I. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology* **27**, 199-204 (2009).

89 Vellaichamy, A. *et al.* "Topological Significance" Analysis of Gene Expression and Proteomic Profiles from Prostate Cancer Cells Reveals Key Mechanisms of Androgen Response. *PLoS ONE* **5**, e10936.

90 Wu, Z., Zhao, X. & Chen, L. Identifying responsive functional modules from protein-protein interaction network. *Molecules and cells* **27**, 271-277 (2009).

91 Zhao, X.-M., Wang, R.-S., Chen, L. & Aihara, K. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic acids research* **36** (2008).

92 Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology* **11**, R53.

93 Soong, T.-t., Wrzeszczynski, K. & Rost, B. Physical protein-protein interactions predicted from microarrays. *Bioinformatics* **24**, 2608-2614 (2008).

94 Yeger-Lotem, E. *et al.* Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nature Genetics* **41**, 316-323 (2009).
95 Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences* **107**, 6286-6291.
96 Ljubic, I. *et al.* An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Mathematical Programming* **105**, 427-449, doi:DOI 10.1007/s10107-005-0660-x (2006).
97 Dogan, S. *et al.* Molecular Epidemiology of EGFR and KRAS Mutations in 3026 Lung Adenocarcinomas: Higher Susceptibility of Women to Smoking-related KRAS-mutant Cancers. *Clin Cancer Res* **18**, 6169–6177, doi:10.1158/1078-0432.CCR-11-3265 (2012).
98 Riely, G. J. *et al.* Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma. *Clin Cancer Res* **14**, 5731-5734, doi:10.1158/1078-0432.CCR-08-0646 (2008).
99 Society, A. C. *American Cancer Society Figures and Facts*. (2012).
100 Cox, A. D. & Der, C. J. Ras history: The saga continues. *Small GTPases* **1**, 2-27, doi:10.4161/sgtp.1.1.12178 (2010).
101 Scholl, C. *et al.* Synthetic Lethal Interaction between Oncogenic KRAS Dependency and STK33 Suppression in Human Cancer Cells. *Cell* **137**, 821-834, doi:10.1016/j.cell.2009.03.017 (2009).
102 Singh, A. *et al.* TAK1 inhibition promotes apoptosis in KRAS-dependent colon cancers. *Cell* **148**, 639-650, doi:10.1016/j.cell.2011.12.033 (2012).
103 Cheriyath, V. *et al.* Phosphoproteomics Identifies Oncogenic Ras Signaling Targets and Their Involvement in Lung Adenocarcinomas. *PLoS ONE* **6**, e20199, doi:10.1371/journal.pone.0020199 (2011).
104 Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467, doi:10.1093/bioinformatics/bth092 (2004).
105 Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207-214, doi:10.1038/nmeth1019 (2007).
106 Organization, T. G. P. M. <http://www.thegpm.org/crap/index.html> (
107 Fermin, D., Basrur, V., Yocum, A. K. & Nesvizhskii, A. I. Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. *Proteomics* **11**, 1340-1345, doi:10.1002/pmic.201000650 (2011).
108 Petrak, J. *et al.* Déjà vu in proteomics. A hit parade of repeatedly identified differentially expressed proteins. *Proteomics* **8**, 1744-1749, doi:10.1002/pmic.200700919 (2008).
109 Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75-82, doi:10.1093/bioinformatics/btn577 (2009).
110 Bailly-Bechet, M. *et al.* Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 882-887, doi:10.1073/pnas.1004751108 (2011).
111 Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 365, doi:1471-2164-10-365 [pii]
10.1186/1471-2164-10-365 (2009).
112 Shankavaram, U. T. *et al.* Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol Cancer Ther* **6**, 820-832, doi:1535-7163.MCT-06-0650 [pii]
10.1158/1535-7163.MCT-06-0650 (2007).

113 Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117, doi:10.1186/gb-2003-4-9-117 (2003).
114 Fleiss, J. Review papers : The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2**, 121-145, doi:10.1177/096228029300200202 (1993).
115 Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine* **5**, e184, doi:10.1371/journal.pmed.0050184 (2008).
116 He, H. *et al.* P-21 activated kinase 1 knockdown inhibits beta-catenin signalling and blocks colorectal cancer growth. *Cancer Lett* **317**, 65-71, doi:10.1016/j.canlet.2011.11.014 (2012).
117 Jensen, L. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research* **37**, D412-416 (2009).
118 Ong, C. C. *et al.* Targeting p21-activated kinase 1 (PAK1) to induce apoptosis of tumor cells. *Proceedings of the National Academy of Sciences* **108**, 7177-7182, doi:10.1073/pnas.1103350108 (2011).
119 Harr, M. W. *et al.* Inhibition of Lck enhances glucocorticoid sensitivity and apoptosis in lymphoid cell lines and in chronic lymphocytic leukemia. *Cell Death and Differentiation* **17**, 1381-1391, doi:10.1038/cdd.2010.25 (2010).
120 Shi, M. A Constitutively Active Lck Kinase Promotes Cell Proliferation and Resistance to Apoptosis through Signal Transducer and Activator of Transcription 5b Activation. *Molecular Cancer Research* **4**, 39-45, doi:10.1158/1541-7786.mcr-05-0202 (2006).
121 Giglione, C., Gonfloni, S. & Parmeggiani, A. Differential actions of p60c-Src and Lck kinases on the Ras regulators p120-GAP and GDP/GTP exchange factor CDC25Mm. *European journal of biochemistry / FEBS* **268**, 3275-3283 (2001).
122 Gherardi, E., Birchmeier, W., Birchmeier, C. & Vande Woude, G. Targeting MET in cancer: rationale and progress. *Nat Rev Cancer* **12**, 89-103, doi:10.1038/nrc3205 (2012).
123 Chen, H. Y. *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* **356**, 11-20, doi:10.1056/NEJMoa060096 (2007).
124 Vogler, M. BCL2A1: the underdog in the BCL2 family. *Cell Death and Differentiation* **19**, 67-74, doi:10.1038/cdd.2011.158 (2011).
125 Datta, S. R. *et al.* Survival factor-mediated BAD phosphorylation raises the mitochondrial threshold for apoptosis. *Dev Cell* **3**, 631-643 (2002).
126 Fang, X. *et al.* Regulation of BAD phosphorylation at serine 112 by the Ras-mitogen-activated protein kinase pathway. *Oncogene* **18**, 6635-6640, doi:10.1038/sj.onc.1203076 (1999).
127 Ferlay, J. *et al.* Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International journal of cancer. Journal international du cancer* **127**, 2893-2917, doi:10.1002/ijc.25516 (2010).
128 Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J Clin* **63**, 11-30, doi:10.3322/caac.21166 (2013).
129 Couraud, S., Zalcman, G., Milleron, B., Morin, F. & Souquet, P. J. Lung cancer in never smokers--a review. *Eur J Cancer* **48**, 1299-1311, doi:10.1016/j.ejca.2012.03.007 (2012).
130 Weir, B. A. *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898, doi:10.1038/nature06358 (2007).

131 Takeuchi, K. *et al.* KIF5B-ALK, a novel fusion oncokinase identified by an immunohistochemistry-based diagnostic system for ALK-positive lung cancer. *Clin Cancer Res* **15**, 3143-3149, doi:10.1158/1078-0432.CCR-08-3248 (2009).
132 Drilon, A. *et al.* Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas. *Cancer Discov* **3**, 630-635, doi:10.1158/2159-8290.CD-13-0035 (2013).
133 Wang, X. S. *et al.* An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nature Biotechnology* **27**, 1005-1011, doi:10.1038/nbt.1584 (2009).
134 Wu, Y. M. *et al.* Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discov* **3**, 636-647, doi:10.1158/2159-8290.CD-13-0050 (2013).
135 Li, Y., Chien, J., Smith, D. I. & Ma, J. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics* **27**, 1708-1710, doi:10.1093/bioinformatics/btr265 (2011).
136 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research* **38**, e178, doi:10.1093/nar/gkq622 (2010).
137 Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98, doi:10.1038/nrg2934 (2011).
138 Carrara, M. *et al.* State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? *BMC bioinformatics* **14 Suppl 7**, S2, doi:10.1186/1471-2105-14-S7-S2 (2013).
139 Carrara, M. *et al.* State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int* **2013**, 340620, doi:10.1155/2013/340620 (2013).
140 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).
141 San Lucas, F. A., Wang, G., Scheet, P. & Peng, B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28**, 421-422, doi:10.1093/bioinformatics/btr667 (2012).
142 Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120, doi:10.1016/j.cell.2012.08.029 (2012).
143 Falls, D. L. Neuregulins: functions, forms, and signaling strategies. *Exp Cell Res* **284**, 14-30 (2003).
144 Holmes, W. E. *et al.* Identification of heregulin, a specific activator of p185erbB2. *Science* **256**, 1205-1210 (1992).
145 Wen, D. *et al.* Structural and functional aspects of the multiplicity of Neu differentiation factors. *Mol Cell Biol* **14**, 1909-1919 (1994).
146 Sholl, L. M. *et al.* EGFR mutation is a better predictor of response to tyrosine kinase inhibitors in non-small cell lung carcinoma than FISH, CISH, and immunohistochemistry. *Am J Clin Pathol* **133**, 922-934, doi:10.1309/AJCPST1CTHZS3PSZ (2010).
147 Lipson, D. *et al.* Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. *Nature Medicine* **18**, 382-384, doi:10.1038/nm.2673 (2012).
148 Koivunen, J. P. *et al.* EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin Cancer Res* **14**, 4275-4283, doi:10.1158/1078-0432.CCR-08-0168 (2008).
149 Shaw, A. T. *et al.* Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. *Lancet Oncol* **12**, 1004-1012, doi:10.1016/S1470-2045(11)70232-7 (2011).

150 Montero, J. C. *et al.* Neuregulins and cancer. *Clin Cancer Res* **14**, 3237-3241, doi:10.1158/1078-0432.CCR-07-5133 (2008).
151 Gollamudi, M., Nethery, D., Liu, J. & Kern, J. A. Autocrine activation of ErbB2/ErbB3 receptor complex by NRG-1 in non-small cell lung cancer cell lines. *Lung Cancer* **43**, 135-143 (2004).
152 Hegde, G. V. *et al.* Blocking NRG1 and other ligand-mediated Her4 signaling enhances the magnitude and duration of the chemotherapeutic response of non-small cell lung cancer. *Science Translational Medicine* **5**, 171ra118, doi:10.1126/scitranslmed.3004438 (2013).
153 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108, doi:10.1038/nature11233 (2012).
154 Guil, S. & Esteller, M. Cis-acting noncoding RNAs: friends and foes. *Nat Struct Mol Biol* **19**, 1068-1075, doi:10.1038/nsmb.2428 (2012).
155 Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat Rev Genet* **14**, 880-893, doi:10.1038/nrg3594 (2013).
156 Sessa, L. *et al.* Noncoding RNA synthesis and loss of Polycomb group repression accompanies the colinear activation of the human HOXA cluster. *Rna* **13**, 223-239, doi:10.1261/rna.266707 (2007).
157 Zhang, X. *et al.* A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **113**, 2526-2534, doi:10.1182/blood-2008-06-162164 (2009).
158 Modarresi, F. *et al.* Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nature Biotechnology* **30**, 453-459, doi:10.1038/nbt.2158 (2012).
159 Yu, W. *et al.* Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**, 202-206, doi:10.1038/nature06468 (2008).
160 Congrains, A., Kamide, K., Ohishi, M. & Rakugi, H. ANRIL: Molecular Mechanisms and Implications in Human Health. *Int J Mol Sci* **14**, 1278-1292, doi:10.3390/ijms14011278 (2013).
161 Mahmoudi, S. *et al.* Wrap53, a natural p53 antisense transcript required for p53 induction upon DNA damage. *Molecular Cell* **33**, 462-471, doi:10.1016/j.molcel.2009.01.028 (2009).
162 Su, W. Y. *et al.* Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. *Cell Res* **22**, 1374-1389, doi:10.1038/cr.2012.57 (2012).
163 Morrissy, A. S., Griffith, M. & Marra, M. A. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Research* **21**, 1203-1212, doi:10.1101/gr.113431.110 (2011).
164 Faghihi, M. A. *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nature Medicine* **14**, 723-730, doi:10.1038/nm1784 (2008).
165 Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human Molecular Genetics* **9**, 1415-1423 (2000).
166 Scheele, C. *et al.* The human PINK1 locus is regulated in vivo by a non-coding natural antisense RNA during modulation of mitochondrial function. *BMC Genomics* **8**, 74, doi:10.1186/1471-2164-8-74 (2007).

167 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-419, doi:10.1016/j.cell.2010.06.040 (2010).
168 Kogo, R. *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer research* **71**, 6320-6326, doi:10.1158/0008-5472.CAN-11-1021 (2011).
169 Niinuma, T. *et al.* Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. *Cancer research* **72**, 1126-1136, doi:10.1158/0008-5472.CAN-11-1803 (2012).
170 Luo, J. H. *et al.* Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology* **44**, 1012-1024, doi:10.1002/hep.21328 (2006).
171 Schmidt, L. H. *et al.* The long noncoding MALAT-1 RNA indicates a poor prognosis in non-small cell lung cancer and induces migration and tumor growth. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **6**, 1984-1992, doi:10.1097/JTO.0b013e3182307eac (2011).
172 Geng, Y. J., Xie, S. L., Li, Q., Ma, J. & Wang, G. Y. Large intervening non-coding RNA HOTAIR is associated with hepatocellular carcinoma progression. *J Int Med Res* **39**, 2119-2128 (2011).
173 Kim, K. *et al.* HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* **32**, 1616-1625, doi:10.1038/onc.2012.193 (2013).
174 Silva, J. M., Boczek, N. J., Berres, M. W., Ma, X. & Smith, D. I. LSINCT5 is over expressed in breast and ovarian cancer and affects cellular proliferation. *RNA Biol* **8**, 496-505 (2011).
175 Han, Y., Liu, Y., Gui, Y. & Cai, Z. Long intergenic non-coding RNA TUG1 is overexpressed in urothelial carcinoma of the bladder. *J Surg Oncol* **107**, 555-559, doi:10.1002/jso.23264 (2013).
176 He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855-1857, doi:10.1126/science.1163853 (2008).
177 Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564-1566, doi:10.1126/science.1112009 (2005).
178 Li, S., Liberman, L. M., Mukherjee, N., Benfey, P. N. & Ohler, U. Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data. *Genome Research* **23**, 1730-1739, doi:10.1101/gr.149310.112 (2013).
179 Wang, G. *et al.* Identification of regulatory regions of bidirectional genes in cervical cancer. *BMC Medical Genomics* **6 Suppl 1**, S5, doi:10.1186/1755-8794-6-S1-S5 (2013).
180 Trinklein, N. D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Research* **14**, 62-66, doi:10.1101/gr.1982804 (2004).
181 Bock, C., Walter, J., Paulsen, M. & Lengauer, T. CpG island mapping by epigenome prediction. *PLoS computational biology* **3**, e110, doi:10.1371/journal.pcbi.0030110 (2007).
182 He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The Antisense Transcriptomes of Human Cells. *Science* **322**, 1855-1857, doi:10.1126/science.1163853 (2008).

183 Maruyama, R. *et al.* Altered antisense-to-sense transcript ratios in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 2820-2824, doi:10.1073/pnas.1010559107 (2012).
184 Yassour, M. *et al.* Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biology* **11**, R87, doi:10.1186/gb-2010-11-8-r87 (2010).
185 Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic acids research* **37**, e123, doi:10.1093/nar/gkp596 (2009).
186 Klostermeier, U. C. *et al.* A tissue-specific landscape of sense/antisense transcription in the mouse intestine. *BMC Genomics* **12**, 305, doi:10.1186/1471-2164-12-305 (2011).
187 Dallosso, A. R. *et al.* Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer. *Rna* **13**, 2287-2299, doi:10.1261/rna.562907 (2007).
188 Pasmant, E. *et al.* Characterization of a germ-line deletion, including the entire INK4/ARF locus, in a melanoma-neural system tumor family: identification of ANRIL, an antisense noncoding RNA whose expression coclusters with ARF. *Cancer research* **67**, 3963-3969, doi:10.1158/0008-5472.CAN-06-2004 (2007).
189 Yamaguchi, T., Hosono, Y., Yanagisawa, K. & Takahashi, T. NKX2-1/TTF-1: an enigmatic oncogene that functions as a double-edged sword for cancer cell survival and progression. *Cancer Cell* **23**, 718-723, doi:10.1016/j.ccr.2013.04.002 (2013).
190 Tanaka, H. *et al.* Lineage-specific dependency of lung adenocarcinomas on the lung development regulator TTF-1. *Cancer research* **67**, 6007-6011, doi:10.1158/0008-5472.CAN-06-4774 (2007).
191 Wei, W., Pelechano, V., Jarvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet* **27**, 267-276, doi:10.1016/j.tig.2011.04.002 (2011).
192 Liu, B., Chen, J. & Shen, B. Genome-wide analysis of the transcription factor binding preference of human bi-directional promoters and functional annotation of related gene pairs. *BMC Systems Biology* **5 Suppl 1**, S2, doi:10.1186/1752-0509-5-S1-S2 (2011).
193 Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037, doi:10.1038/nature07728 (2009).
194 Takayama, K. *et al.* Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. *The EMBO Journal* **32**, 1665-1680, doi:10.1038/emboj.2013.99 (2013).
195 Wahlestedt, C. Targeting long non-coding RNA to therapeutically upregulate gene expression. *Nat Rev Drug Discov* **12**, 433-446, doi:10.1038/nrd4018 (2013).
196 Beltran, M. *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes Dev* **22**, 756-769, doi:10.1101/gad.455708 (2008).
197 Venkitachalam, S., Chueh, F. Y. & Yu, C. L. Nuclear localization of lymphocyte-specific protein tyrosine kinase (Lck) and its role in regulating LIM domain only 2 (Lmo2) gene. *Biochem Biophys Res Commun* **417**, 1058-1062, doi:10.1016/j.bbrc.2011.12.095 (2012).