THE UNIVERSITY OF CHICAGO

DISCOVERY AND CHARACTERIZATION OF NOVEL SOMATIC AND WIDESPREAD INHERITED POLYMORPHIC FUSION GENES IN HUMANS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS

BY
CHAITANYA BANDLAMUDI

CHICAGO, ILLINOIS
MARCH 2016

Copyright © 2016 by Chaitanya Bandlamudi

All Rights Reserved

ACKNOWLEDGMENTS

I am incredibly grateful to everyone who has made this work possible. I especially want to thank my advisor, Kevin White, for welcoming me into his lab nearly six years ago, and, for his support and mentorship throughout my graduate school experience. Kevin is a brilliant and creative scientist. But his approach to a scientific question, his relentless desire to pursue the next natural question despite the scope and complexity, and, his ability to effectively communicate, both in presentations and in writing, are what I admire the most and what I hope to emulate in the later phases of my career. I also would like to thank the members of my committee Michelle Le Beau, Andrey Rzhetsky and Dan Nicolae for providing valuable insights throughout the progression of my research.

I have benefited greatly from many members of the White lab as well as the broader UChicago community. Early in my graduate career, I had the invaluable opportunity to work with the then-postdocs in the lab Megan McNerney, Casey Brown, Thomas Stricker and David Vanderweele. I learned so much from them about how to think productively about a scientific question and how to best pursue it. I also would like to especially thank Barbara Stranger as well as the members of Institute for Genomics and Systems Biology (IGSB)'s weekly genomics meeting for helpful advice in the latter part of my thesis. I also benefited a great deal from general discussions with my lab-mates Jason Pitt and Aashish Jha. Due to shared interest in cancer genomics, Jason and I shared many intellectually stimulating and often productive discussions on our research projects. A substantial portion of this work relied heavily on access to computational resources. For that, I thank the Computation Institute and the Laboratory of Advanced Computing for providing the necessary infrastructure. I especially would like to thank Joe Urbanski for his support in running analyses on the Beagle supercomputer.

Of course, my graduate pursuits would have never been successful without the help of my family. My parents Radha and Govardhan worked tremendously hard to provide a secure foundation for me and my sister, Harita, to pursue our career aspirations. Without their

constant love and support, this work would not have been possible. I especially would like to thank my wife Gowthami for her relentless love, support and encouragement throughout the better part of my graduate school. During both extreme ups and downs of my research, her emotional and thoughtful support helped me keep my sanity between the two tail-ends of the "sanity" distribution. I could not have asked for a better partner in life and I cannot imagine what my life would be without her. I am also thankful to my cousins, Praveen and Lahari Kanneganti. I am also extremely thankful to have the friendship of my dear friend David Kondru. Finally, I would like to dedicate my work to my late grandfather, Ramaiah Patchala, who raised me and instilled the values of hardwork, determination and ambition.

CONTENTS

ACKNOWLEDGMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xi

LIST OF SUPPLEMENTARY TABLES xii

ABSTRACT xiv

1 INTRODUCTION 1
- 1.1 Overview 1
- 1.2 Somatic fusion genes in cancer 2
  - 1.2.1 Fusion genes in hematological malignancies 4
  - 1.2.2 Fusion genes in sarcomas 9
  - 1.2.3 Fusions in epithelial cancers 10
  - 1.2.4 Motivation for further research 15
- 1.3 Inherited polymorphic fusion genes 16
- 1.4 Methods to detect fusion genes 19

2 A HIGHLY SENSITIVE AND SPECIFIC APPROACH FOR DETECTING GENE FUSIONS FROM TRANSCRIPTOME SEQUENCING DATA 23
- 2.1 Abstract 23
- 2.2 Introduction 23
- 2.3 Results 26
  - 2.3.1 MOJO overview 26
  - 2.3.2 Anchor read criteria for high confidence fusion calls 26
  - 2.3.3 Evaluation of sensitivity and specificity of MOJO 28
  - 2.3.4 Runtime performance comparisons 31
- 2.4 Discussion 33
- 2.5 Methods 36
  - 2.5.1 Minimum Overlap Junction Optimizer (MOJO) Algorithm 36
  - 2.5.2 Simulation to estimate the effect of anchor read length on specificity . 39
  - 2.5.3 Framework for evaluating MOJO 40
- 2.6 Appendix: Figures 41
- 2.7 Appendix: Supplementary Tables 48
- 2.8 Contributions 52

3 DISCOVERY AND CHARACTERIZATION OF NOVEL RECURRENT FUSION GENES IN 33 HUMAN CANCERS 53
- 3.1 Abstract 53
- 3.2 Introduction 53
- 3.3 Results 55

3.3.1 Somatic fusion transcript discovery . . . 55
3.3.2 Characteristics of somatic fusion transcripts . . . 57
3.3.3 Detection of somatic fusion transcripts in normal samples . . . 58
3.3.4 Characteristics of fusions across 33 cancers . . . 59
3.3.5 Enrichment of cancer associated genes in fusion transcripts . . . 60
3.3.6 Kinase fusions . . . 62
3.3.7 Landscape of recurrent fusion genes across human cancers . . . 63
3.3.8 Novel recurrent chimeric proteins . . . 64
3.3.9 Novel recurrent out-of-frame fusions . . . 66
3.3.10 Genes recurrently mis-regulated by fusion events . . . 68
3.3.11 Functional validation of novel recurrent fusion genes . . . 70
3.4 Discussion . . . 71
3.5 Methods . . . 75
3.5.1 Transcriptome sequencing data . . . 75
3.5.2 Gene expression quantification . . . 76
3.5.3 Gene fusion discovery . . . 76
3.5.4 Evaluation of sensitivity and specificity . . . 78
3.5.5 Evaluation of specificity using RT-PCR validations . . . 81
3.5.6 Enrichment analysis of cancer associated genes among fusion genes . . . 82
3.5.7 Protein domain analysis of kinase fusion genes . . . 83
3.5.8 Somatic copy number alterations supporting fusion transcripts . . . 83
3.5.9 Focal and gene level copy number alterations . . . 83
3.5.10 Known recurrent fusion genes . . . 84
3.5.11 Functional validations . . . 84
3.5.12 Cell migration and invasion assays . . . 85
3.6 Appendix: Figures . . . 87
3.7 Appendix: Supplementary Tables . . . 114
3.8 Contributions . . . 125

4 DISCOVERY OF WIDESPREAD INHERITED POLYMORPHIC FUSION GENES 126
4.1 Abstract . . . 126
4.2 Introduction . . . 126
4.3 Results . . . 129
4.3.1 Fusion gene discovery from healthy tissues . . . 129
4.3.2 Characteristics of polymorphic fusion genes . . . 131
4.3.3 Population genetic properties of polymorphic fusion genes . . . 133
4.3.4 Functional characteristics of fusion genes . . . 137
4.3.5 Ectopic expression of 3' genes of polymorphic fusion transcripts . . . 138
4.4 Discussion . . . 140
4.5 Methods . . . 142
4.5.1 RNA and DNA sequencing data . . . 142
4.5.2 Gene fusion detection from RNA-seq . . . 143
4.5.3 SV detection from WGS . . . 144
4.5.4 Targeted fusion junction search in TCGA/HapMap . . . 144
4.5.5 Gene expression quantification . . . 145

4.6 Appendix: Figures 145
4.7 Appendix: Supplementary Tables 163
4.8 Contributions 166

5 DISCUSSION AND FUTURE DIRECTIONS 167
5.1 Summary 167
5.2 Gene fusion discovery 167
5.2.1 Discussion 167
5.2.2 Future Directions 169
5.3 Fusion genes in cancer 170
5.3.1 Discussion 170
5.3.2 Future Directions 173
5.4 Inherited polymorphic fusion genes 176
5.4.1 Discussion 176
5.4.2 Future Directions 177

BIBLIOGRAPHY 179

LIST OF FIGURES

1.1 Structural rearrangements resulting in fusion genes 3

1.2 Potential functional consequences of fusion genes 17

2.1 Effect of anchor length on specificity of the anchor read 42

2.2 Evaluation of sensitivity and specificity using cell lines 43

2.3 Evaluation of sensitivity and specificity two primary tumors 44

2.4 Runtime comparison of various fusion callers including MOJO 45

2.5 Illustration of terms to describe fusion discovery 45

2.6 MOJO algorithm overview 46

2.7 Evaluation of sensitivity and specificity using cell lines – non-canonical 47

3.1 Fusion discovery workflow and performance evaluation 88

3.2 Characteristics of somatic fusions across 33 cancers 90

3.3 Enrichment of cancer associated genes among somatic fusions 92

3.4 Recurrent fusions in the TCGA 94

3.5 Functional evaluation of novel recurrent fusion events 96

3.6 Distribution of TCGA and CCLE transcriptomes analyzed in this study 98

3.7 TCGA PanCancer gene fusion calling workflow 99

3.8 Saturation curve showing the effect of filtering out GTEx fusion calls 100

3.9 Extended comparisons of MOJO, MapSplice and FusionCatcher using 120 tumor transcriptomes 101

3.10 Distribution of fusion calls supported by segment breaks in "Both" partner genes 102

3.11 Distribution of fusion calls supported by segment breaks in "either" partner gene 103
3.12 Proportion of fusion-positive tumors per cancer type 104
3.13 Degree of genomic instability is correlated with the number of fusions per sample 104
3.14 Schematic of *ITPK2-LTBP4* fusion detected in primary, metastatic and normal tissue of one individual 105
3.15 Proportion of individuals with at least one fusion involving a known cancer gene 105
3.16 Proportion of individuals with at least one fusion involving a known kinase gene 106
3.17 Frequency distribution of recurrent fusions 106
3.18 Known fusion genes 107
3.19 Expression characteristics of recurrent fusion genes 108
3.20 Expression characteristics of genes dysregulated by fusions 109
3.21 Westernblot of fusion proteins 111
3.22 Cellular localization of proteins 112
3.23 Wound healing assay to measure migration of fusion-positive cells 113
4.1 Polymorphic fusion genes identified in the GTEx donors 147
4.2 Population-specific enrichment of polymorphic fusion genes 148
4.3 Domain and ectopic expression characteristics of polymorphic gene fusions . . . 151
4.4 Distribution of number of fusion calls per donor and per tissue 153
4.5 Characteristics of multi-partner gene fusions that are filtered out 154
4.6 QC plot showing distribution of various categories of fusion gene calls across tissues types in GTEx 155

4.7 Detection of a fusion in all tissues depends on its expression level 157
4.8 Expression patterns of individual genes in the polymorphic fusion gene set . . . 158
4.9 Expression profile of wild-type *NAIP* across tissues 159
4.10 Expression profile of *CHEK2* across tissues in fusion positive donors 159
4.11 Expression profiles of individual genes of *TFG-GPR128* fusion across tissues in genes 160
4.12 Expression profiles of individual genes of *MAEA-FAM9B* fusion across tissues in genes 161
4.13 Expression profiles of individual genes of *CASP4-CARD18* fusion across tissues in genes 162

LIST OF TABLES

2.1 List of RT-PCR validated fusions in the Primary and Relapse tumors 32

4.1 List of 63 polymorphic fusion genes discovered in this study 134

LIST OF SUPPLEMENTARY TABLES

2.1 Description of cell lines used in this study for performance evaluation of 9 fusion callers 48

2.2 Configurations of published algorithms used for comparisons 49

2.3 List of previously reported true positives in the 18 cancer cell lines 50

2.4 Fusion calls nominated by eight different methods within 18 cell lines 51

2.5 Fusion calls nominated by seven different methods within two primary tumors . 51

3.1 TCGA and CCLE samples analyzed in this study 114

3.2 TCGA samples analyzed in this study 115

3.3 CCLE cell line transcriptomes analyzed in this study 115

3.4 GTEx normal tissues analyzed in this study 115

3.5 HapMap cell line transcriptomes analyzed in this study 115

3.6 Primary tumor transcriptomes used in this study for comparisons 116

3.7 Workflow comparisons - 55 transcriptomes - canonical fusions 116

3.8 Workflow comparisons - 55 transcriptomes - non-canonical fusions 117

3.9 Workflow comparisons - 126 transcriptomes - canonical fusions 117

3.10 Workflow comparisons - 126 transcriptomes - non-canonical fusions 118

3.11 RT-PCR validation results from 12 cell lines 118

3.12 True positives detected by MOJO-PC and Yoshihara et al. 119

3.13 True positives detected by MOJO-PC and Stransky et al. 119

3.14 List of previously reported fusions (COSMIC/Mitelman) 120

3.15 Fusion calls nominated by MOJO-PC from 9,360 primary tumors 120

3.16 Fusion calls nominated by MOJO-PC from relapse/metastatic tumors 121

3.17 Fusion calls nominated by MOJO-PC from adjacent normal tissues 121

3.18 Correlations between fusions and genomic instability across cancer types 122

3.19 Gene categories enriched for somatic fusion genes 122

3.20 Recurrently fused cancer associated genes 122
3.21 Fusions with intact kinase domains 123
3.22 Recurrent chimeric proteins 123
3.23 Recurrent out-of-frame fusions 124
3.24 Recurrently dysregulated genes 124

4.1 Manifest of 9,126 GTEx transcriptomes analyzed in this study 163
4.2 Multi-partner fusion genes that are filtered out 164
4.3 Low expressed fusion genes that are filtered out 164
4.4 Low recurrence fusion genes that are filtered out 164
4.5 High confidence polymorphic fusion genes identified in this study 165
4.6 WGS-based structural variants supporting polymorphic fusion genes 165

ABSTRACT

Fusion genes have played a seminal role in cancer biology both clinically as well as substantially promoting our understanding of the mechanisms in cancer. Although fusions genes have been studied primarily in the context of cancer, the shared DNA repair mechanisms between mitotic and meiotic cells suggests that such events may exist as polymorphic events in healthy populations. In chapter 1, I present a comprehensive overview of our understanding of somatic as well as germline gene fusions. In chapter 2, I present a new fusion discovery method that is central to the discoveries in my research. In chapter 3, I survey nearly ten thousand cancer patients across various tumor types and study the pan-cancer characteristics of fusions. I identify some of the most recurrently identified novel fusions and followup with functional validations to investigate their tumorigenic properties. In chapter 4, I surveyed more than five hundred healthy individuals and discovered many novel polymorphic fusion genes. I show that a substantial proportion of the recurrent polymorphic fusion genes display population specific differentiation. Using gene expression profiles, I demonstrate that many of these fusion genes may have functional consequences. This previously unreported class of genetic variation may have far reaching implications in explaining a proportion of missing heritability. In the final chapter, I will discuss the key findings and highlight several future directions that will further our understanding of the role of gene fusions in both cancer as well as healthy individuals.

CHAPTER 1

INTRODUCTION

1.1 Overview

Variation in the somatic and germline genomes is driven by mutations introduced during DNA replication or spontaneously that evade the cell's DNA repair mechanisms. Types of mutations include point mutations (single nucleotide variants, SNVs), short insertions/deletions (indels) and structural rearrangements such as deletions, insertions, duplications and translocations. Much progress has been made in our understanding of the SNVs in both the context of variation in human populations (Roach et al. 2010; Kloosterman et al. 2015; Genomes Project et al. 2015) as well as in cancer (Lawrence et al. 2013; Kandoth et al. 2013). Type of substitutions and the genome context in which they occur are used to evaluate the functional consequences. Recent DNA sequencing and copy number analyses have also characterized the landscape of structural rearrangements in both germline (Sudmant et al. 2015a; Sudmant et al. 2015b) and somatic (Zack et al. 2013) genomes. In addition to gene amplifications and deletions, other consequences of structural rearrangements include juxtaposition of partial regions of two genes resulting in generation of a novel fusion gene (Figure 1.1). The past thirty years of cancer biology research has firmly established the significance of fusion genes in tumor initiation and progression (Mitelman et al. 2007). However, recent technological advances present new opportunities to find novel biologically significant fusions as well as, for the first time, study their global characteristics across cancers. In contrast to somatic fusion genes, our understanding of germline fusions in human populations is extremely limited and exploring this novel type of genetic variation could open a new field of characterizing the impact of inherited gene fusions on human biology and evolution, and even might explain a portion of the much discussed missing heritability of complex traits(Manolio et al. 2009).

In the following sections of this chapter, I will provide an overview of our current understanding of the prevalence and the significance of somatic and germline gene fusions as well

as provide my motivation for this research.

1.2 Somatic fusion genes in cancer

Current understanding for the genetic basis of cancer emerged nearly a century ago from the remarkable insights of Theodore Boveri while studying abnormally segregating chromosomes in sea urchins and malignant cancer cells (Boveri 2008). Before the concept of 'gene' was developed, he hypothesized that individual chromosomes carry distinct information and that a certain assortment and numbers of chromosomes underlie malignant cells. Even more prescient are his postulations of 'growth stimulatory' and 'growth inhibitory' chromosomes that are analogous to oncogenes and tumor suppressors, respectively, that were discovered more than six decades later (Knudson 1971; Tabin et al. 1982; Varmus 1984). Studies in the following decades reaffirmed Boveri's concept of chromosomal instability but it wasn't until the discovery that DNA is the genetic material of inheritance and the determination of its structure that the different types of mutations underlying cancer genomes started to emerge. In 1960, advances in cytogenetics allowed the discovery of an unusually small chromosome, termed Philadelphia (Ph) chromosome, present in only the malignant cells of chronic myelogenous leukemia (CML) (Nowell et al. 1960). Through novel chromosomal staining techniques and meticulous screening, thirteen years later, Janet Rowley discovered that this small chromosome is the product of reciprocal translocation between chromosomes 9 and 22, referred to as t(9;22) (Nowell et al. 1960). Nearly a decade later, gene mapping studies showed that this rearrangement generated breaks within breakpoint cluster region (*BCR*) and abelson tyrosine kinase receptor 1 (*ABL1*) genes (Heisterkamp et al. 1983; Groffen et al. 1984). Around the same time, the mapping of the first proto-oncogene *HRAS* and tumor suppressor *RB1* genes further solidified that cancer is a genetic disease and thus initiated the quest to discover and characterize genes that are mis-regulated in cancer.

The predominant model for tumorigenesis is that cancer is an evolutionary process guided by Darwinian principles in a clone of deregulated cells that have somatically acquired "ad-

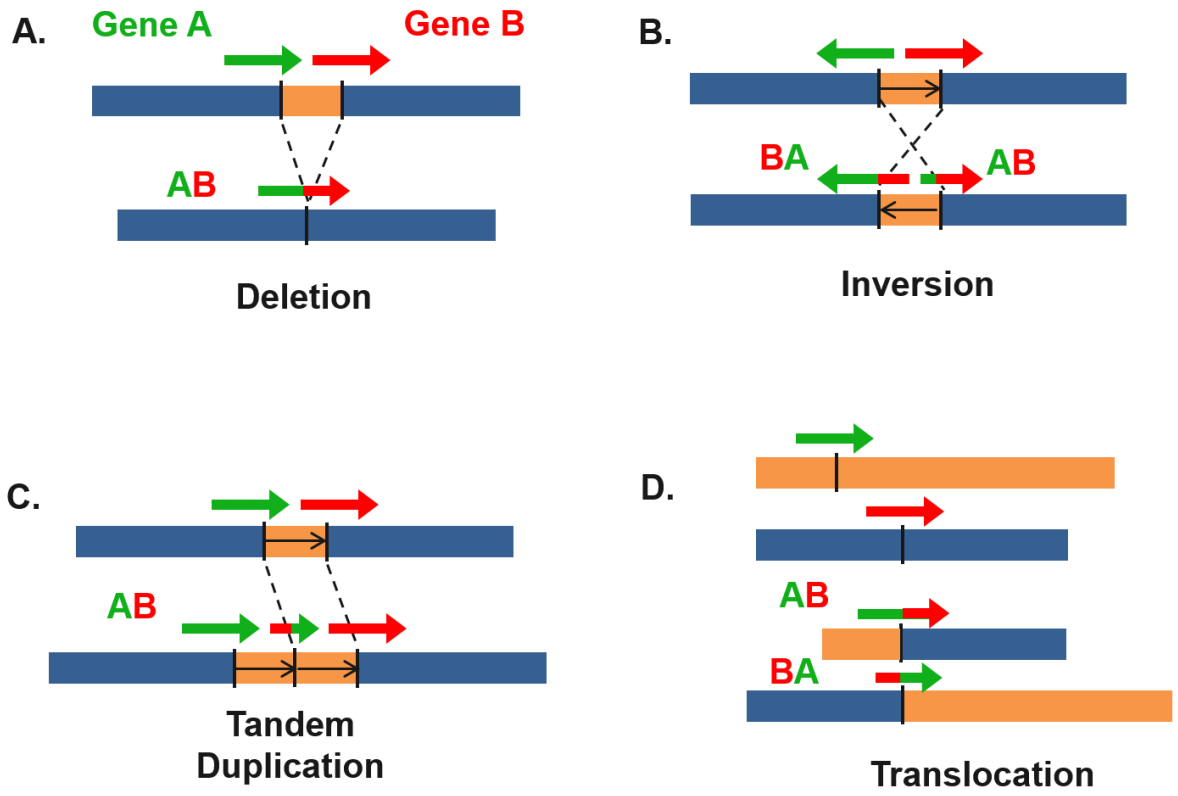


Figure 1.1: Structural rearrangements resulting in generation of novel fusion genes. (a) A deletion spanning (but not encompassing) two genes in same orientation can generate a fusion gene comprising 5' of gene A and 3' of gene B. (b) An inversion event spanning (but not encompassing) two genes on opposite strands can generate two fusion genes with swapped upstream and downstream regions. (c) A tandem duplication event spanning (but not encompassing) two genes on the same strand can generate a fusion gene with the 5' of gene B fused to the 3' of gene A. (d) Translocation events with breakpoints within the genic regions can generate fusion genes. A balanced translocation has the potential to generate a reciprocal fusion.

vantageous" mutations (Stratton et al. 2009). Selectively advantageous mutations ("driver" alterations) acquired by clones can result in fixation of the mutation by selective sweep. Alternately, mutations that have neutral or detrimental effect on fitness ("passenger" mutations) that are already present before the selective sweep are also fixed within the clone (Greaves et al. 2012). Recent studies show that tumors progress through a series of clonal expansions and in the process acquire a multitude of "driver" as well as "passenger" mutations (Maley et al. 2004; Bozic et al. 2010; Beerenwinkel et al. 2007). Although many types of "driver" mutations have been identified in various cancers (Vogelstein et al. 2013), gene fusions have an elevated significance in cancer biology due to their strong potential to "drive" tumorigenesis and their ability to serve as effective diagnostic, prognostic and therapeutic markers. More than four decades since the discovery of the translocation that generated *BCR-ABL1* gene fusion in CML, many oncogenic fusions have been discovered across leukemias, sarcomas and carcinomas. In the following sections, I will present a brief overview of our understanding of fusions across various types of cancers.

1.2.1 Fusion genes in hematological malignancies

Many of the first reports of translocations in cancer came from studies of hematological malignancies primarily enabled by the relatively simple karyotype of leukemic cells that allowed for easier detection. In 1972, a year before Janet Rowley discovered that the Ph chromosome is a consequence of t(9;22) translocation, she also discovered another translocation t(8;21) in leukemic samples which constituted the first reported recurrent translocation in cancer (Rowley 1973b; Rowley 1973a). Following t(8;21) and t(9;22), intense search for recurrent translocations in the following years led to the discovery of t(8;14) and t(8;22) in Burkitt lymphoma (Zech et al. 1976; Berger et al. 1979; Miyoshi et al. 1979; Fraisse et al. 1981), t(4;11) in acute lymphoblastic leukemia (Oshimura et al. 1977), t(15;17) in acute promyelocytic leukemia (Rowley et al. 1977), and t(14;18) in follicular lymphoma (Ohno et al. 1979).

Emerging and ultimately, overwhelming evidence of recurrent chromosomal abnormalities in subtypes of leukemia strongly implied a biological role for these events. The next significant development in understanding their role came in 1982 from the cloning of the first translocation that identified immunoglobulin heavy chain (*IGH*) and Myc-oncogene (*MYC*) genes involved in t(8;14) in Burkitt lymphoma (Dalla-Favera et al. 1982; Taub et al. 1982). *MYC*, at that time, had recently been mapped to chromosome 8 (Neel et al. 1982) and its role in acutely transforming retroviruses had been clearly established (Sheiness et al. 1979; Payne et al. 1981; Neel et al. 1981; Hayward et al. 1981). Mouse plasmacytomas were also discovered to have *Myc* translocations involving the *IGH* locus as the partner, further implying the importance of *MYC* fusion events (Shen-Ong et al. 1982; Crews et al. 1982). Soon after, transgenic mice overexpressing *MYC* were reported to develop B-cell lymphomas, strongly suggesting an oncogenic role for this translocations in Burkitt lymphoma (Adams et al. 1985).

The discovery of *IGH-MYC* translocations provided renewed emphasis to the importance of translocations and the potential for discovering novel cancer associated genes from uncloned translocations. A year after the cloning of *MYC* translocations, *BCR* and *ABL1* genes in the t(9;22) translocation in CML were mapped (Heisterkamp et al. 1983; Groffen et al. 1984)]. *ABL1*, the human homolog of the transforming sequence (v-ABL) of Abelson murine leukemia virus (A-MuL V) had already been mapped to chromosome 9. However, the discovery of a novel fusion transcript generated by truncated regions of *BCR* and the *ABL1* oncogene was significant (Shtivelman et al. 1985). Although the oncogenic role of the activated murine Abl gene had been reported prior to this discovery (Witte et al. 1980), the retention of this phosphorylation activity in the truncated *ABL1* gene strongly suggested a potential mode of action in CML (Konopka et al. 1984). Mice infected with retrovirus encoding the *BCR-ABL1* fusion were found to develop myeloproliferative syndrome that closely resembled the chronic phase of CML (Daley et al. 1990). However, interestingly, this fusion required supporting mutations to transform cells (Cuenco et al. 2001; Huntly et al.

2004). Nevertheless, mouse models suggested that suppressing the kinase activity might be an effective therapeutic option. A decade of searching for kinase inhibitors to target this fusion gene culminated in 2001, with the approval for clinical use of a highly specific inhibitor called imatinib (Druker et al. 2001). Five-year follow-up study showed a 95% progression-free survival for CML patients (Druker et al. 2006). This seminal milestone is not only marked by the first successful targeted treatment of a cancer but also by presenting a proof of principle for the effective targeting of kinase-encoding genes as a potent therapeutic option.

In the same year that the genes in t(9;22) were mapped, cloning the t(14;18) translocation characteristic of follicular lymphoma led to the identification of *IGH* gene on chromosome 14 joined to a novel gene called *BCL2* (B-Cell CLL/Lymphoma 2) (Tsujimoto et al. 1984). *BCL2* was later shown to be from a family of genes that regulate and contribute to apoptosis (Vaux et al. 1988; Yin et al. 1994). More than a decade earlier, the concept of programmed cell death - apoptosis - had been introduced from the observations of tissue sections of certain cancers (Kerr et al. 1972). However, the characterization of *BCL2* and its role in cancer-associated translocations highlighted, for the first time, the importance of programmed cell death and its deregulation in cancer.

Cloning of the t(15;17), characteristic aberrations of acute promyelocytic leukemias (APL), identified promyelocytic leukemia gene (*PML*) joined to the retinoic acid receptor alpha (*RARA*) gene (The et al. 1990; Lemons et al. 1990). Expression of *PML-RARA* fusion in the myeloid-promyelocytic lineage of transgenic mice induced abnormal hematopoiesis, eventually leading to acute myeloid leukemia (He et al. 1997; Grisolano et al. 1997; Brown et al. 1997). The fusion blocks proper differentiation of hematopoietic progenitors, resulting in accumulation of blast cells resulting in tumorigenesis. A serendipitous discovery lead to successful development of a therapeutic plan that involves all-trans retinoic acid to target and degrade the fusion protein in patients with APL, with a 12-year progression-free survival rate of 70% (Huang et al. 1988; Ades et al. 2010). Together with *BCR-ABL1*, these two fusions have become a paradigm for targeted treatments in cancer.

Nearly 20 years after the discovery of the first recurrent translocation, t(8;21), in leukemia, the genes involved were identified as *RUNX1* and *RUNX1T1* (Miyoshi et al. 1991; Gao et al. 1991; Miyoshi et al. 1993; Nucifora et al. 1993). *RUNX1* (also referred to as *AML1*, *CBFA2*) is a hematopoietic transcription factor that is essential for normal hematopoiesis. RUNX1T1 has been reported to interact with DNA binding domains. In vitro and in vivo studies show that *RUNX1-RUNX1T1* fusion inhibits the wild type function of *AML1* resulting in blocking of normal myeloid and erythroid differentiation leading to leukemogenesis (Ahn et al. 1998; Kitabayashi et al. 1998; Yergeau et al. 1997; Okuda et al. 1998). Interestingly, *RUNX1* has been found to be a frequent translocation partner in leukemias with more than 55 partners discovered to date (De Braekeleer et al. 2011).

Mapping the t(4;11) in acute lymphoblastic leukemias identified a novel gene, *MLL* (mixed lineage leukemia, synonym: *KMT2A*). Similar to *RUNX1* but in a more striking fashion, this gene is frequently reported in different types of leukemias with multiple partners. In addition to t(4;11), other common translocations involving the 11q23.3 locus are t(9;11), t(11;19) and t(6;11) (Hagemeijer et al. 1982; Prasad et al. 1993). To date, over 121 distinct translocation partners have been identified for *MLL* (Meyer et al. 2013), of which, only 79 partner genes have been characterized. Despite the number of translocation partners observed, all MLL-positive leukemias share a similar clinical feature - very poor prognosis compared to other rearrangement types (Munoz et al. 2003). *MLL* is a DNA-binding protein that methylates histone H3 lysine 4 (H3K4) and upregulates the expression of multiple genes including the Hox cluster genes (Milne et al. 2002). *MLL* fusion-positive cases typically encode a MLL protein that lose their H3K4 methyltransferase activity. This can result in reprogramming of differentiated myeloid cells that can activate stem-like properties (Cozzio et al. 2003; Okada et al. 2005; Krivtsov et al. 2007).

Clinically, translocations have played a crucial role in establishing the diagnosis of new leukemia cases. The strong correlation of recurrent cytogenetic abnormalities with the morphological features of leukemic cells have enabled the refinement of classification systems

for leukemias, referred to as the WHO classification for the tumors of the hematopoietic system (Vardiman et al. 2009). For example, recurrent translocations resulting in *BCR-ABL1*, *PML-RARA* and *IGH-BCL2* fusions are highly specific to CML, APL and follicular lymphoma, respectively. In addition, these classifications have also been shown to be correlated with the overall prognosis. For example, some rearrangements such as t(8;21), t(15;17) and inv(16) (*CBFB-MYH11*) show good response to therapy and survival but events such as *MLL* translocations, deletion of parts of chromosomes 5 or 7 are associated with poor outcome (Grimwade et al. 2010). The WHO classification is an internationally recognized classification system that relies on a combination of morphology, immunophenotype, genetics and clinical features.

It is estimated that approximately 45% of all leukemia cases have normal karyotypes (no detectable cytogenetic abnormality) (Mrozek et al. 2004). However, the introduction of high-resolution approaches to detect chromosomal aberrations such as single-nucleotide polymorphism (SNP) arrays and next-generation sequencing have enabled the detection of many smaller scale rearrangements that evaded cytogenetic detection (Bullinger et al. 2010; Walter et al. 2009; Ley et al. 2008; Suela et al. 2007). Fusions generated from these events are referred to as cryptic fusions. In some cases, these rearrangements were found to manifest in biologically relevant outcomes such as the inter-chromosomal translocation of PML gene on chromosome 15 to the RARA gene on chromosome 17, or vice-versa, resulting in the generation of the classic PML-RARA fusion (Grimwade et al. 1997; Welch et al. 2011). Similarly, cryptic rearrangements resulting in novel fusions have also been reported in both adult as well as pediatric AMLs (Reader et al. 2007). However, the overall incidence of cryptic fusions appears to be substantially lower.

To date, more than 952 fusions have been reported in various types of hematological disorders (Mertens et al. 2015). However, only 161 of these have been detected recurrently, demonstrating the heterogeneity of the disease in presenting a diverse set of chromosomal alterations. A recent survey of fusion transcripts from 179 AML samples using transcriptome

sequencing identified 29 fusion proteins of which 15 are novel (Cancer Genome Atlas Research 2013). All 15 fusions are detected in only one tumor sample suggesting that we may be approaching saturation to discover novel recurrent fusions in AML. However, larger sample sizes in future studies may provide better evidence for this hypothesis.

1.2.2 Fusion genes in sarcomas

Sarcomas are rare (0.7% of all cancer incidences) mesenchymal malignancies arising from bone, cartilage or connective tissues such as adipose, muscle, etc. Nearly a quarter of the sarcomas are now estimated to have a recurrent translocation event (Mertens et al. 2009). Following the discoveries of translocations in leukemias, cytogenetic analysis of some common sarcomas quickly identified recurrent translocations such as t(2;13) and t(1;13) in aveolar rhabodomyosarcoma (ARMS) (Biegel et al. 1991; Douglass et al. 1991; Douglass et al. 1987; Turc-Carel et al. 1986), t(11;22) in Ewing sarcoma (EWSR) (Whang-Peng et al. 1986) and t(X;18) in synovial sarcoma (Smith et al. 1987; Turc-Carel et al. 1987). Gene mapping studies identified *PAX3-FOXO1* and *PAX7-FOXO1* as the genes fused in ARMS (Barr et al. 1993; Shapiro et al. 1993; Galili et al. 1993). Recent chromatin immunoprecipitation assays have shown that *PAX3-FOXO1* fusion proteins bind proximal to transcriptional start sites and co-regulate many cancer associated target genes such as myogenic differentiation 1 (*MYOD1*), myogenic factor 5 (*MYF5*), fibroblast growth factor receptor 4 (*FGFR4*), anaplastic lymphoma tyrosine kinase (*ALK*), etc (Cao et al. 2010). Breakpoint cloning identified Ewing sarcoma breakpoint region 1 (*EWSR1*) and friend leukemia virus integration 1 (*FLI1*) as the genes involved in the t(11;22) (Delattre et al. 1992). *EWSR1-FLI1* was found to block the differentiation of mesenchymal stem cells by inducing the expression of embryonic stem cell genes such as *POUF1*, *SOX2* and *NANOG*, and as well as upregulating the polycomb group transcriptional repressor *EZH2* (Taylor et al. 2011). Interestingly, the *SYT-SSX* fusion generated by the t(X;18) fusion (Clark et al. 1994) in synovial sarcomas was also shown to induce the expression of pluripotent genes *POU5F1*, *SOX2* and *NANOG*,

suggesting that nuclear reprogramming by fusion proteins may be a common mechanism in sarcomas to disrupt the mesenchymal lineages (Naka et al. 2010). Overall, transcription factors involved in lineage differentiation are frequent targets of chromosomal rearrangements in leukemias and sarcomas (Rosenbauer et al. 2007).

Therapeutically, unlike in some leukemias, targeted therapies for fusion-positive sarcomas remain elusive. However, the significant understanding of molecular mechanisms of some recurrent fusions in sarcomas hold promise for development of targeted drugs or better clinical management (Bennani-Baiti et al. 2012).

1.2.3 Fusions in epithelial cancers

Although solid tumors account for 90% (American Cancer Society, 2015) of all cancer incidences, the vast majority of recurrent fusions reported to date have been in hematological malignancies and sarcomas. This discrepancy is primarily driven by the technical limitations in analyzing solid tumor cells that tend to be highly aneuploid and highly rearranged with complex karyotypes. Before the introduction of high throughput and unbiased methods such as next generation sequencing (NGS) to detect translocations, conventional methods such as karyotyping had been moderately successful in identifying translocations in solid cancers. The first reported transformative rearrangements in solid tumors involved the *RET* proto-oncogene and neurotrophic tyrosine kinase receptor type 1 (*NTRK1*) in thyroid cancer (Pierotti et al. 1992; Takahashi et al. 1985; Grieco et al. 1990). Many different translocation partners were identified for *RET* and *NTRK1* in the following years (Pierotti 2001). Rearrangements involving other members of the NTRK family have also been reported (Greco et al. 1997). Interestingly, a fusion involving another member of the NTKR family, *NTKR3* was also found fused to ETS family transcription factor, *ETV6*, in pediatric mesenchymal and secretory breast cancers (Knezevich et al. 1998; Tognon et al. 2002). In follicular thyroid carcinoma, the mapping of recurrent t(2;3) identified a novel fusion gene involving *PAX8* and *PPARG* genes (Kroll et al. 2000). In salivary gland tumors, fusions

involving *PLAG1* and *HMGA2* have been reported (Geurts et al. 1997; Kas et al. 1997). For example, *CTNNB1-PLAG1* results in ubiquitous expression of a normally tissue specific gene, pleomorphic adenoma gene 1 (*PLAG1*) (Willert et al. 1998; Peifer 1997). The role of *HMGA2-NFIB* fusion remains to be elucidated. Apart from *RET/NTRK1* fusions, many of the early fusions in solid tumors are extremely rare <0.5%, demonstrating the challenge of interpreting these events. Despite the slow progress of molecular and functional characterization of these events, the discovery of recurrent fusions in solid cancers strongly suggests biological significance.

Application of unbiased gene fusion detection approaches such as gene expression microarrays and NGS significantly increased the number of fusions reported in solid tumors. The discovery of a highly recurrent fusion *TMPRSS2-ERG* that is incident in $\tilde{4}5\%$ of all prostate cancer patients rejuvenated the efforts to search for highly frequent fusions in solid tumors (Tomlins et al. 2005). However, apart from the *TMPRSS2-ERG*, an emerging theme from the hundreds of fusions reported in the past decade across various cancer types is that, unlike in hematological malignancies, fusions that define morphologically distinct subtypes of cancers (such as *BCR-ABL1* in CML and *PML-RARA* in APL) have been very rare. Such fusions have only been reported in extremely rare types of cancers. For example, in solitary fibrous tumor (SFT), a rare soft tissue sarcoma with an incidence rate of 0.013% (200 cases/year), all SFT cases (51/51) are predicted to have a *NAB2-STAT6* fusion generated by an inversion event in 12q13 (Robinson et al. 2013; Chmielecki et al. 2013). In vitro assays showed that this fusion induced proliferation and activated early growth response genes (Robinson et al. 2013). Fibrolamellar hepatocellular carcinoma (FL-HCC) is also a rare (<0.01% incidence rate) but malignant liver tumor affecting adolescents and young adults without any prior liver disease. A deletion event was identified to generate *DNAJB1-PRKACA* in all 15/15 patients with this disease (Honeyman et al. 2014). *PRKACA* is predicted to be overexpressed and activated. A *C11orf95-RELA* fusion was identified in more than two-thirds of patients with supratentorial ependymoma – a rare tumor of brain

and spinal cord (0.04% incidence rate) (Parker et al. 2014). The *C11orf95-RELA* fusion was shown to transform neuronal stem cells and form tumors in mice. Although the incidence rates are low, the high frequency of these fusions within a particular tumor type strongly indicates specificity. An exception to this 'high frequency fusion – rare cancer type' paradigm are the *TMPRSS2/ERG* fusions in cancer. Interestingly, despite their high frequency, the biological significance of *TMPRSS2/ERG* fusions in prostate cancer remains challenging to elucidate (Tomlins et al. 2008).

Except for a select few, the majority of fusions reported in solid tumors are reported at frequencies $\tilde{5}\%$ or lower and involve genes that are proximal to each other on the same chromosome. This latter characteristic is attributable to the greater degree of genomic instability within the solid tumors that manifests as amplificaitons and deletions (median size 0.7Mb) (Zack et al. 2013). Biologically and clinically, one of the more significant examples of a recurrent fusion recently reported in solid tumors is the fusion of echinoderm microtubule-associated protein-like 4 (*EML4*) to anaplastic lymphoma receptor tyrosine kinase (*ALK*) in 5-7% of non-small cell lung cancers (Soda et al. 2007; Rikova et al. 2007; Seo et al. 2012). The fusion retains a coiled-coil domain of *EML4* and the kinase domain of *ALK*, yielding a protein that results in constitutive dimerization and activation of the kinase domain. Transgenic mice expressing the fusion were shown to develop adenocarcinoma nodules within weeks (Soda et al. 2007; Soda et al. 2008). *EML4-ALK* fusions were later also identified in colorectal (Lipson et al. 2012), breast (Lin et al. 2009) and renal cell carcinomas (Sugawara et al. 2012) at frequencies <1%. The striking aspect about this fusion is the successful application of an ALK inhibitor (crizotinib) that results in reduction of phosphorylated *ALK* and suppresses key downstream proliferation pathways. Follow-up studies showed that crizotinib more than doubled the progression-free survival and tripled the response rate compared to chemotherapy (Shaw et al. 2013a). Interestingly, crizotinib was originally developed to target receptor tyrosine kinase *MET* (Zou et al. 2007; Christensen et al. 2007), further reinforcing that successful kinase targeting may be an effective approach to targeting

cancers with kinase-addiction.

Following the discovery of *EML4-ALK*, other kinase fusions have been reported in various solid tumors. ROS proto-oncogene 1 (*ROS1*), a receptor tyrosine kinase with key role in epithelial cell differentiation, was found fused in carcinomas of non-small cell lung (Rikova et al. 2007; Takeuchi et al. 2012; Rimkunas et al. 2012; Suehara et al. 2012), gastric (Lee et al. 2013) and bile duct (Gu et al. 2011). Targeting *ROS1* with crizotinib resulted in inhibition of proliferation and cell death evasion pathways in cell lines (Davies et al. 2012). Other kinase fusions such as *RET* and *NTRK1* that were detected first in thyroid carcinomas are now identified in various cancers. In addition to the first reported translocation in solid tumors, *CCDC6-RET* in thyroid cancers, *RET* fusions were also discovered in lung cancers fused to various partners such as *NCOA4* and *KIF5B* (Takeuchi et al. 2012; Wang et al. 2012; Lipson et al. 2012; Kohno et al. 2012). RET inhibitor Cabozantinib was found to be an effective target in a preliminary clinical trial (Drilon et al. 2013). Apart from the first reported *TPR/TPM3-NTRK1* fusion in papillary thyroid carcinomas (Greco et al. 1997), *NTRK1* fusions were also reported in colon and lung cancers. In vitro testing of *NTRK1* inhibitors showed reasonable success in abrogating the auto-phosphorylation activity of TRKA and inhibition of cell growth (Vaishnavi et al. 2013).

More recently, another recurrent fusion between *FGFR3*, another receptor tyrosine kinase, and coiled-coil domain containing *TACC3* was first reported in glioblastomas ($< 5\%$), and then subsequently shown at lower frequencies in lung, bladder and oral cancers (Singh et al. 2012; Wu et al. 2013b). The fusion protein is shown to acquire constitutive kinase activity, localize to mitotic spindle poles and induce chromosomal segregation defects leading to aneuploidy (Singh et al. 2012). Overexpression of the fusion protein alone was sufficient to induce anchorage independent growth in rat fibroblast cells. Fusion expressing cells were shown to respond to kinase inhibitors in vitro as well as in vivo, suggesting potential for developing targeted drugs.

Apart from kinases, other types of fusion-genes have also been recently reported. 3% of

colon cancer patients are estimated to have a fusion between *VTI1A* and T-cell transcription factor 4 encoding *TCF7L2* (Bass et al. 2011). Wild-type TCF7L2 plays a key role in Wnt signaling in association with Catenin Beta-1 (*CTNNB1*) to regulate intestinal epithelial cell differentiation and proliferation genes. siRNA knock-down of this fusion in a fusion-positive cell line resulted in reduction of anchorage independent growth suggesting a potential role for this fusion in oncogenic activity (Bass et al. 2011). In another example, R-spondin family member genes, agonists of the canonical Wnt signaling pathway, were found to be involved in fusions in 10% of colon cancers (Shinmura et al. 2014; Seshagiri et al. 2012). Interestingly, R-spondin fusions were also recently reported in prostate cancer (Robinson et al. 2015). Despite the clear evidence for *TCF7L2* and R-spondin fusions, the mechanistic principles behind their potential oncogenic effects remain to be elucidated.

Two over-arching patterns of recurrent fusions are shared between solid tumors and hematological malignancies. First, similar to *MLL* (Meyer et al. 2013) and *RUNX1* (De Braekeleer et al. 2009) in hematological malignancies, key genes in solid tumors such as *ALK/RET/ROS1/NTRK1* (Takeuchi et al. 2012) are found fused to multiple distinct partners. Second, also as in leukemias, but in a more striking fashion, families of genes are found recurrently fused in solid tumors. Examples include FGFR-family genes, *FGFR1/2/3*, in various cancers (Wu et al. 2013b); *BRAF/RAF1/CRAF* fusions in tumors of brain, gastric, melanoma and prostate (Jones et al. 2008; Palanisamy et al. 2010; Cin et al. 2011)]; *NTRK1/NTRK3* fusions in papillary, lung, oral and breast (Bongarzone et al. 1998; Vaishnavi et al. 2013; Skalova et al. 2010; Tognon et al. 2002); ETS family transcription factors *ERG/ETV1-5*, primarily in prostate (Tu et al. 2007; Tomlins et al. 2007). Another observation is that the higher degree of genomic instability in solid tumors introduces a substantially greater degree of intertumor heterogeneity that manifests as a diverse mutational spectrum, including a more complex landscape of fusions in epithelial malignancies (Burrell et al. 2013).

1.2.4 Motivation for further research

Despite significant progress made in understanding the role of gene fusions in cancer in the past four decades, several key aspects remain to be explored further. I will highlight some key areas that are the central focus of my research in somatic fusions (Chapter 3). This pursuit is enabled by large cancer sequencing projects, such as The Cancer Genome Atlas, that allow systematic identification and evaluation of gene fusions across cancers.

Motivation 1. Discovery and functional characterization of novel recurrent fusions

Given the clinical and biological significance of fusion genes, a comprehensive catalog of fusions would substantially improve our understanding of the disease as well as ultimately affect patient care. The low frequency of recently reported fusions suggests that many rare but biologically significant fusions remain to be discovered. We hypothesized that with large sample sizes spanning multiple cancer types, we can identify recurrent fusions that are rare within tumor type but when aggregated across cancers will reach appreciable recurrence levels to justify further research, including functional validations. Identifying "driver" fusions will promote our understanding of tumorigenesis and tumor maintenance as well as enable the potential development of targeted treatments. In chapter 3, I analyze more than nine thousand tumors from over 30 cancer types for fusion genes. I discover a number of novel recurrent fusion genes and demonstrate their tumor inducing characteristics by functionally validating the most recurrent ones among them.

Motivation 2. Comprehensive characterization of the diversity of fusion events

Chimeric proteins such as *BCR-ABL1*, *EML4-ALK* and *PML-RARA* that contain partial domains from two genes have been the central focus of investigation of fusions in cancer. Other consequences of fusion events that have been not well-studied include: gene dysregulation, truncated proteins and gene disruptions (loss-of-function) (Figure 1.2). Fusions resulting in dysregulation of gene by swapping of regulatory sequences between the fusion partners have been reported in Burkitt lymphoma (Zech et al. 1976) and medulloblastoma (Northcott

et al. 2012). Similarly, in glioblastoma, *FGFR3-TACC3* is shown to evade miRNA regulation by swapping the 3' untranslated region (Parker et al. 2013). However, apart from a few specific examples, we do not have a thorough understanding of the degree to which these events occur. Truncated proteins can be generated by out-of-frame fusions that result in C-terminal protein of the upstream gene or the N-terminal truncated protein of the downstream gene (Figure 1.2B). For example, *ESR1-CCDC170* fusions in breast cancer have been reported to use a novel translational start-site within *CCDC170* to generate a N-terminal truncated protein that was shown to induce metastatic phenotypes in breast cancer cells (Veeraraghavan et al. 2014). Finally, fusion events resulting in loss-of-function of one or both partner genes have not been thoroughly evaluated. The existence and pervasiveness of these different types of events in various cancers remain poorly understood. Although challenges remain in determining truncations or classifying a fusion as a loss-of-function event, with large sample sizes, recurrent events implicating biological functions may emerge.

Motivation 3. Determining the landscape of fusions across cancers

Large collection of tumors allow for investigation of various characteristics of fusions across cancer types such as: (i) the varying rate at which fusions occur in various types of cancer and the factors associated with this variability, (ii) the spatial characteristics of genes involved in fusions, (iii) enrichment of gene categories among fusions identified, and (iv) possible insights into the mechanisms that generate recurrent fusions.

1.3 Inherited polymorphic fusion genes

Birth of new genes with novel functions is a major contributor to the adaptive evolution of the species (Conant et al. 2008; Kaessmann 2010; Chen et al. 2013). New genes can be introduced as a second copy of of a gene resulting from duplications or retrotranspositions, or by a fusion event. Population genetics studies to date have focused exclusively on the origin and consequences of gene duplications, primarily due to relative ease in detecting them compared to fusion events. In a famous perspective in 1970, Susumo Ohno, building upon

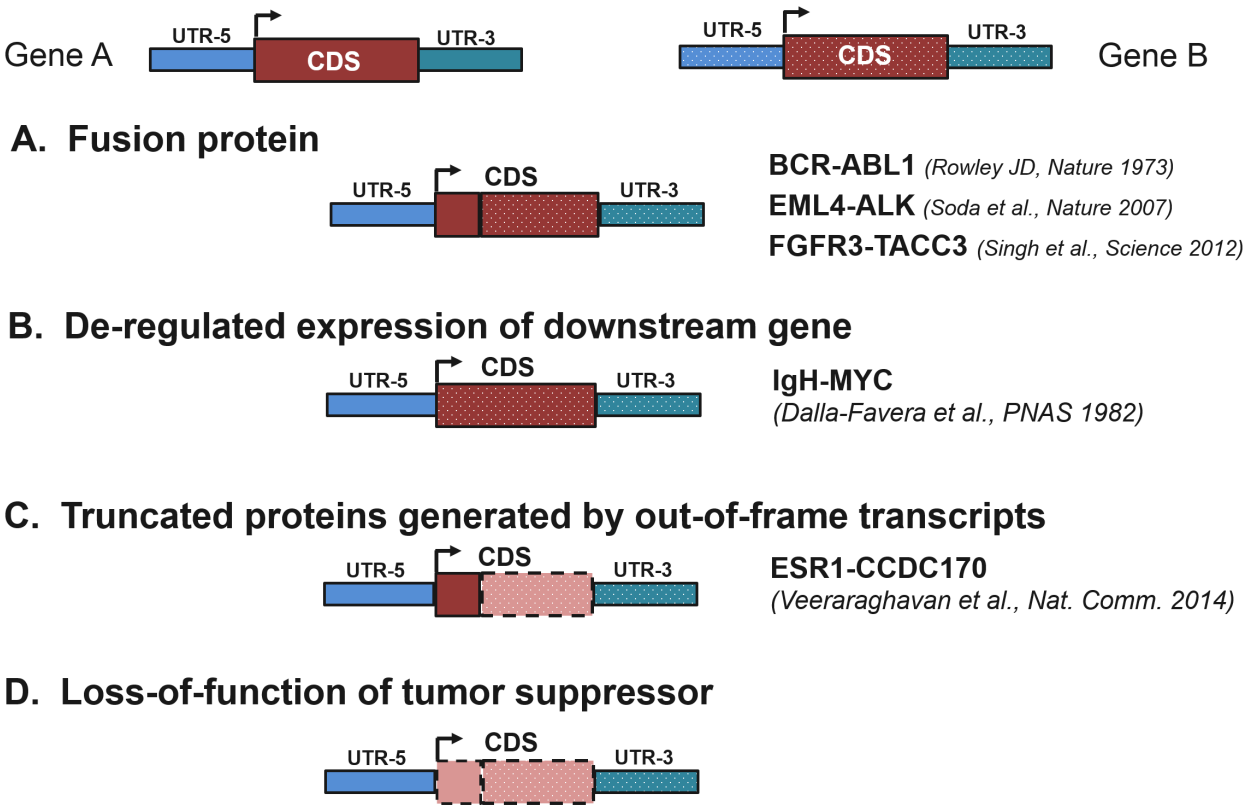


Figure 1.2: Potential functional consequences of four common types of fusion genes (a) A canonical fusion protein containing partial coding sequences from both the partner genes. (b) Swapping of UTR-5 of genes A and B (c) Fusion gene is out of frame if the downstream partial transcript sequence is not in-frame with the annotated coding sequence. A C-terminal truncated peptide comprising only the partial protein from the 5' gene or a C-terminal truncated peptide comprising only the partial protein from the 3' gene can be generated (d) Out-of-frame fusion possibly disrupts one or both genes that have an anti-apoptotic role

years of earlier work (Taylor et al. 2004), postulated that gene duplication is a single most important factor in evolution. A duplicated gene can acquire novel function that is distinct from that of the parental gene in a process call neofunctionalization. In a definitive example, Ross et al showed that the duplicated copy of parental *HP1B* gene, Umbrea, acquires a novel role in chromosomal segregation by losing the heterochromatin localizing domain, rewiring protein interaction networks and species-specific changes in the C-terminal domain (Ross et al. 2013). This example demonstrates how a functionally redundant gene can rapidly become essential through subsequent changes.

Similar to gene duplications, a number of fusion genes with important functions have been discovered that originated before the speciation or lineage diversification events (Long et al. 2003; Zhou et al. 2008; Kaessmann 2010). In a classic example, a chimeric gene, *jingwei*, was identified to be generated by the insertion of retrotransposed sequence of alcohol dehydrogenase gene within the gene body of yellow-emperor (*yande*) gene in the common ancestor of *Drosophila yakuba* and *Drosophila teissieri* more than 3 million years ago. The evolutionary significance of the resulting testis-specific fusion protein is demonstrated by the rapid evolution of *jingwei* before and after the divergence of *D. yakuba* and *D. teissieri* (Long et al. 1993). Similarly, in the hominidae lineage, two chimeric genes *PMCHL1* and *PMCHL2* were derived melanin-concentrating hormone (*MCH*) gene. *PMCHL1* is predicted to be generated by a complex mechanism of retrotransposition and exon shuffling of *MCH* gene coupled with de novo creation of splice sites nearly 20 million years ago. *PMCHL2* is generated by a duplication event involving *PMCHL1* (Courseaux et al. 2001; Viale et al. 2000). Both genes show tissue-specific expression patterns. Many chimeric genes have been found to evolve under positive selection, which strongly implies that the adaptation model of neofunctionalization is responsible for their maintenance.

In contrast to fusions that are fixed within the respective species, we have limited understanding of fusions that have been recently introduced within species and that are segregating as polymorphic variants. A recent study investigating the structural variation within dif-

ferent strains of *Drosophila melanogaster* identified 16 chimeric genes that are segregating within the species. In humans, the select few examples of fusion genes discovered are as a consequence of search for fusion transcripts in the genomes of cancer and autistm spectrum disorders (Chase et al. 2010; Holt et al. 2012). A comprehensive survey of these events in humans remains constrained by technical limitations. Recent analyses of 2,500 whole genomes identified extensive novel structural variation in the human genome (Sudmant et al. 2015a), suggesting that large proportion of structural variation and its transcriptional consequences remains to be elucidated. In chapter 4 of this thesis, I perform a survey of fusion genes across more than five hundred individuals of different ethnicities and report 59 novel fusion genes in the human genome with a substantial proportion of them segregating differentially across populations.

1.4 Methods to detect fusion genes

Chromosome banding methods developed in late 1960s and early 1970s have made it possible to identify each chromosome separately. Banding techniques that label guanine rich (quinacrine mustard, Q-banding) (Caspersson et al. 1968; Caspersson et al. 1969; Caspersson et al. 1970) and thymine-rich (Giemsa, G-banding) (Drets et al. 1971; Hsu et al. 1971; Arrighi et al. 1971) regions on the chromosome have enabled, for the first time, the identification of smaller-scale chromosomal abnormalities including inversions that were not previously detectable. These approaches also paved way for a formalized nomenclature for describing the chromosomal aberrations ("Paris Conference (1971): Standardization in human cytogenetics" 1972). These developments subsequently led to the discovery and subsequent confirmation of recurrence by independent groups, of many translocations in cancers. The next significant development came in 1988 with the development of fluorescence *in situ* hybridization (FISH) technique that allows for labeling of whole or parts of chromosomes with DNA specific probes (Lichter et al. 1988). This technique allowed for discovery of many novel fusion genes including previously reported genes involved in fusions that were subsequently

determined to be fused to multiple partners (Rowley et al. 1990). Further improvements of this method allows labeling each chromosome with different colored probes and subsequently, simultaneous visualization of all chromosomes on the metaphase spreads of the tumor cells (Schrock et al. 1996).

These cytogenetic karyotyping techniques have been invaluable in characterizing the genetic basis of many cancer types. However, they suffer from key limitations that confound comprehensive fusion discovery (Mitelman et al. 2007). Foremost, these methods have very low resolution and preclude the discovery of smaller scale rearrangements which are characteristic of many solid tumors. Furthermore, solid tumors tend to have complex genomes with high degree of aneuploidy, along with large number of smaller scale rearrangements, and, thereby limiting the applicability of these techniques. However, in the early 2000s, the introduction of array-based platforms for gene expression and, subsequently, next-generation sequencing have provided higher resolution approaches to detect potential fusion genes in an unbiased fashion. In 2005, Tomlins et al. 2007 used individual probes on the gene expression arrays to identify genes that showed substantially different expression patterns between the 5' and 3' ends of the same gene. Using this approach they discovered a novel fusion, *TMPRSS2-ERG*, in prostate cancer. Later on, the same group demonstrated the applicability of next-generation sequencing based approaches to detect fusions in an unbiased fashion from the transcriptomes of cancer patients (Maher et al. 2009).

In theory, paired-end whole genome sequencing (WGS) allows for accurate identification of structural variants (SVs) that can be used to identify fusion genes that may or may not be expressed. Multiple strategies are used to identify different types of SVs. For example, copy number variants can be accurately nominated by read-depth approaches that assume a random (Poisson) distribution of reads mapping to the genome, and this approach can identify regions that show significant enrichment or depletion of read coverage to nominate amplifications and duplications, respectively. In contrast, detection of a fusion gene requires identification of a junction sequence comprising two distinct regions of the genome. A combi-

nation of read-pair and split-read approaches are used to identify these junctions. Read-pair methods are based on identifying paired-end reads that are inconsistent with the expected insert size and orientiation. These reads are called discordant reads. Split-read methods then attempt to identify paired-end reads with one end mapping to the breakpoint junction and the other end supporting the predicted proximal breakpoint regions. Many methods have been developed to identify SVs from WGS (Ding et al. 2014).

However, WGS-based fusion discovery is not trivial. In addition to the numerous technical issues, two primary challenges confound the fusion discovery from WGS. First, the breakpoints tend to involve highly repetitive regions in the genome. Most sequencing approaches use small insert size of <500bp which are not large enough to span most repetitive regions. Second, biochemical studies characterizing the DNA double strand repair process and the sequencing studies that use long reads have highlighted an increasing complexity of the breakpoint region comprising sequences from distinct loci introduced during the repair process. For example, Simsek et al. 2010 identified that NHEJ repair enzymes can introduce sequences in lengths ranging from a few bp to several Kb originating from distant loci during the repair process. Consistent with this finding, the recent WGS analysis of 2,500 whole genomes supplemented with long-read sequencing (up to 10kb) highlighted the pervasiveness of this complexity (Sudmant et al. 2015a). For example, a distal sequence ranging from 3bp to 10's of kb was inserted in nearly 45% of all breakpoints that were characterized. In addition, they also observed that the regions with SVs tend to have may breakpoints clustered in close proximity. These challenges limit the applicability of short-read WGS to identify fusion genes.

Transcriptome sequencing (RNA-seq) mitigates these two primary limitations by virtue of the endogenous cellular process of splicing out the intronic regions in which majority of the breakpoints occur. Using RNA-seq, fusions are identified by first nominating the discordant reads where each end maps to two different genes, and, second, by identification of anchor reads that support the exon-exon junctions involved in the fusion. RNA-seq based fusion

discovery presents its own challenges with high numbers of false positives contributed by both technical and biological artifacts. Technical artifacts, such as random chimeras during the library preparation or alignment artifacts due to homology in the exome, can confound fusion discovery. In addition, several sources of biological artifacts can be introduced by biological processes such as back-splicing and intragenic rearrangements. Accurately accounting for these artifacts without compromising sensitivity is a primary challenge in the RNA-seq based fusion discovery field. In chapter 2, I will present a more thorough overview of current methods in RNA-seq based fusion discovery and also present a novel algorithm to detect fusions from transcriptomes with high sensitivity and specificity.

CHAPTER 2

A HIGHLY SENSITIVE AND SPECIFIC APPROACH FOR DETECTING GENE FUSIONS FROM TRANSCRIPTOME SEQUENCING DATA

2.1 Abstract

Although transcriptome sequencing (RNA-seq) allows for unbiased detection of gene fusions, the varying sensitivity and specificity of the existing methods to detect gene fusions has been a limiting factor in discovering novel fusions in cancers. I introduce an algorithm, MOJO (Minimum Overlap Junction Optimizer), to identify fusion genes from paired-end RNA-seq. MOJO relies on mapping of discordant read pairs and on extensive filtering for exon junction sequence reads that meet a minimum overlap criteria anchored within the exons from two fused genes. I compare MOJO with eight other methods using two primary tumor and 18 cell line transcriptomes, and I demonstrate its superior sensitivity and specificity. Unlike existing methods, I show that MOJO performs consistently better across diverse set of transcriptomes and also demonstrates faster runtimes. MOJO is freely available at: https://github.com/cband/MOJO.

2.2 Introduction

Gene fusions are generated by chromosomal rearrangements that result in fusion of genomic fragments of two distinct transcriptional loci. These events are enriched in tumors due to the inherent genomic instability of cancer genomes. Many fusion genes have been discovered in cancer that serve as specific diagnostic markers, prognostic indicators and therapeutic targets (Mitelman et al. 2007; Rowley 2008). NGS approaches such as whole genome sequencing (DNA-seq) and transcriptome sequencing (RNA-seq) have allowed for discovery of numerous fusions in cancers (Mertens et al. 2015). Unlike, DNA-seq, RNA-seq allows

for detection of expressed and, therefore, potentially gain-of-function fusions with functional consequences. Numerous methods have been developed to detect fusions from RNA-seq (Sboner et al. 2010; McPherson et al. 2011; Kim et al. 2011; Iyer et al. 2011; Asmann et al. 2011; Wu et al. 2013a; Wang et al. 2010b; Jia et al. 2013; Kinsella et al. 2011) but their sensitivities and specificities vary substantially (Carrara et al. 2013). Given that the majority of the oncogenic fusions discovered over the past decade are found to occur at frequencies less than 5% within a given tumor type, it is essential for a fusion discovery process to be sensitive enough to detect fusions consistently across tumors while controlling for spurious calls.

Two strategies for fusion discovery from RNA-seq exist. First, an assembly based approach aligns all RNA-seq reads into short contigs and then aligns them onto reference genomes (Robertson et al. 2010; Grabherr et al. 2011). Theoretically, this breakpoint agnostic strategy is the most precise approach to identify fusions. However, the requirements for longer run-times and deeper sequencing limit, as well as the overall lack of proper evaluation of these methods, remains a limiting factor for broad applicability. The second and the most widely implemented strategy is alignment based. Maher et al. 2009, in a proof-of-concept study, first described the fundamental algorithmic approach to identify fusions from paired-end RNA-seq. In that study, they first identify candidate gene pairs from ends of paired-end reads that map to two different genes, and they then nominate fusions based on the reads that map to exon-exon junctions between the two genes. Later this approach was extended to detect fusions in a gene model annotation independent manner that allows detection of fusions with breakpoint junctions within introns or intra-exonic regions (McPherson et al. 2011; Kim et al. 2011). Sboner et al. 2010 first identified and implemented the filters for the various sources of false positives in RNA-seq based fusion discovery resulting from PCR duplicates, repetitive regions, homology in the genome and random chimeras between highly expressed genes.

Many alignment-based methods that have been published implement variations of these

fundamental discovery and filtering approaches developed by Maher et al. 2009 and Sboner et al. 2010. However, studies comparing a subset of the methods show that substantial variation in sensitivity and specificity remains (Carrara et al. 2013). Non-uniformity of the approaches used for comparisons could also result in differences among methods. Of note, we identify three criteria, in addition to consistent parameterization, that when unaccounted for can lead to imprecise evaluations. First, different gene annotation models used by different methods should be integrated to consider only the genes that can be nominated by all methods. Second, synthetic data and cell lines do not sufficiently capture the complexity of a primary tumor transcriptomes that tend to be heterogeneous populations of cells with varying degrees of genomic instability. And, finally, evaluating methods with different library preparations (read length, library size, protocol differences) is essential to demonstrate the resiliency of a given method to various artifacts. In most methods developed so far, the confidence level in a fusion call is typically established by the number and distribution of reads supporting the fusion junction. The stringency with which the junction mapping reads is defined can affect the sensitivity to detect low expressed fusions, as well as the specificity due to spurious alignments.

In this study, I introduce MOJO, an algorithm to identify fusions from paired-end RNA-seq data. The basic approach begins by identifying clusters of reads with each end mapping to two different genes. Next, candidate fusion genes are identified from these clusters and exon junctions are constructed between the candidate exons that are predicted to be fused between the two genes. Reads are aligned to these junctions and candidate fusion supporting reads are identified with one end aligning to exon-exon fusion junction and the other end aligning to one of the two genes. MOJO maximizes sensitivity by iteratively searching all potential alignments to construct discordant clusters. MOJO then maximizes specificity by applying rigorous filters designed to control for both technical and biological sources of artifacts. A key determinant of specificity is the uniqueness with which the anchor read aligns to the fusion junction. I evaluated the extent to which an anchor read that maps to

any of the fusion junctions between all possible pairs of genes also maps else where in the genome and transcriptome. Findings from this analysis are used to construct a set of specific criteria for users to optimize for specificity.

We evaluated MOJO along site eight published methods, using 18 cell line transcriptomes with previously validated fusions. To determine the performance of MOJO on primary tumors, we use a primary and the corresponding relapse transcriptomes of a patient with high grade serous ovarian carcinoma. These evaluations indicate superior performance of MOJO, and furthermore, demonstrate that MOJO's performance scales with sequencing depth and library complexity. Finally, we show that MOJO is computationally faster than the six next best performing methods, and the computational run-times scale linearly with sequencing depth.

2.3 Results

2.3.1 MOJO overview

MOJO is designed to discover gene fusions from paired-end transcriptome sequencing (Figures 2.5-2.6, see Methods). Briefly, discordant reads with each end supporting a distinct gene are identified. An iterative alignment approach is used to maximize the number of discordant reads nominated for further analysis. Reads mapping to repetitive or homologous regions are excluded. Next, candidate exons in each of the two genes supported by discordant reads that are predicted to be involved in the fusion junction are identified. Anchor reads supporting the candidate junctions are then identified. Following stringent filtering for various biological and technical artifacts, candidate fusion genes are nominated.

2.3.2 Anchor read criteria for high confidence fusion calls

A primary source of false positives in fusion discovery from transcriptome sequencing is the incorrect assignment of anchor reads to a predicted fusion junction. Such spurious

anchor reads can be generated by an alternatively spliced transcript of one of the two genes in the fusion pair (alternatively mapping anchor read, ALT-AR), or from a distinct transcriptional locus (ambiguous anchor read, A-AR) that shares extended homology with the fusion junction (Figure 2.1a). If a fusion junction is entirely supported by ALT-ARs or A-ARs, then such ambiguous fusion junctions could lead to false positives when not accounted for, or they could lead to false negatives if filtered out. Understanding the degree to which homology in the transcriptome could confound fusion discovery can be used to construct criteria to nominate high confidence fusion calls. We therefore sought to estimate what fraction of all possible fusions between all genes in the transcriptome have fusion junctions supported by ambiguous anchor reads (ALT-ARs and A-ARs). Briefly, we first constructed fusion junctions between all possible gene-gene fusions in the human genome. From each fusion junction, we then generated 50p split-end reads of varying anchor lengths (10, 15, 20 and 25bp) covering the junction. For example, for a 10bp anchor length we generated two split-end reads, one with a 10bp left overhang and one with a 10bp right overhang. Similarly, we generated 9 different split-end reads (two each for 10, 15 and 20bps, and, one for a 25bp overhang) for each fusion junction. All split-end reads were then aligned to both the genome and the transcriptome. For each split-end read alignment position, we search for homology between the 500bp proximal to the aligned region and the region from which the split-end read was generated. Identification of shared homology would suggest that the anchor read originating from the fusion junction could also be generated from the canonical genome/transcriptome, and therefore, would introduce ambiguity.

We find that the number of gene pairs with ambiguous junctions is strictly dependent on the anchor length (Figure 2.1b). For 10bp anchor length, we find that more than 12% (22 million gene pairs) of all possible gene fusions, have at least one fusion junction that can incorrectly be nominated by a spurious anchor read. In contrast, for a 25bp anchor length, we find that only 0.02% (38,000 gene pairs) produce an ambiguous junction. Our analysis, for the first time, demonstrates the affect of anchor length on specificity and the importance

of properly classifying ambiguous anchor reads.

2.3.3 Evaluation of sensitivity and specificity of MOJO

2.3.3.1 Comparisons using 18 cell line transcriptomes

Here we evaluated MOJO along with eight existing methods (Supplementary Table 2.2), using 18 cell line transcriptomes in which experimentally validated fusions have been reported previously (Supplementary Table 2.1). We constructed a true positive set comprising 58 fusions reported by the respective studies and an additional 115 fusions that have been reported by other studies using these cell lines (Supplementary Table 2.3). A subset of these 173 fusions may not be detected in these transcriptomes due to sequencing depth or cell line heterogeneity. All 9 methods including MOJO were executed on this data using the criteria specified in Supplementary Table 2.2. To ensure uniform comparisons, we applied several filtering steps (see Section 2.5.3) to control for high false positive rate of some methods that may not be reflective of their true performance. We note that no true positive (Supplementary Table 2.3) is missed during this post-processing. Finally, we stratified these calls into two categories based on whether the fusion junction involves annotated exon-exon boundaries (canonical fusions) or intronic/intra-exonic regions (non-canonical fusions). Within the 18 transcriptomes, an aggregate of 3,058 fusions were nominated by all methods, of which only 360 (11%) are canonical (Supplementary Tables 2.4).

Among the canonical fusions, we identified 102 true positives nominated by at least one method. MOJO demonstrated the highest sensitivity identifying 97% (99 out of 102) of the true positives, followed by deFuse and SOAPfuse at 85% and 80%, respectively (Figure 2.2a, Supplementary Table 2.4). We find substantial differences in the fusions that are detected by various methods. The six best performing previously published methods (deFuse, SOAPfuse, FusionCatcher, ChimeraScan, Tophat and MapSplice) collectively identified 99 (98%) true positives but shared a median overlap of only 61 (60%) true fusions between any two methods.

All three false negatives in MOJO are fusions supported by three or fewer reads (Figure 2.2b). Although MOJO identified highest number of low-expressed fusions, the three false negatives are due to poor base qualities or mismatches in the few anchor reads supporting these fusions. Lower sensitivity of the other methods could be attributable to the low expression level of the fusion gene (low number of anchor reads supporting the fusion). However, we find that all methods except MOJO missed at least three fusion calls that were supported by five or more reads (Figure 2.2b). Despite an overall false negative rate of 14.8% for deFuse and 33.6% with MapSplice, both methods correctly identified 100% and 94%, respectively, of all true positives (n=53) supported by 10 or more anchor reads. Overall, we found that MOJO exhibits higher sensitivity across the expression spectrum.

Of the 258 previously unreported canonical fusions (referred to here as "other" fusions) that have been nominated by the eight methods, only 36 (13.8%) are called by at least two methods and 23 (8.8%) by four or more methods (Figure 2.2c). While the fusion calls nominated by multiple methods are high confidence novel fusions in the respective cell lines, the majority of the singletons are strong candidates for false positives. In our comparisons singletons ranged from 0 in FusionHunter to 78 in deFuse. Anchor reads supporting a fusion junction can be a proxy for confidence level in a fusion call with higher number of anchor reads strongly indicating a fusion event. However, higher number of anchor reads can also be generated when the various sources of alignment artifacts are incorrectly accounted for. We find that all eight singleton fusion calls by MOJO are supported by only one anchor and, therefore, can be easily filtered out. In contrast, for all other methods, singletons are supported by reads across the spectrum highlighting the challenges associated with controlling specificity within the existing methods.

Non-canonical fusions are chimeric transcripts that retain intronic sequences or with junctions involving intra-exonic regions. A recently discovered fusion *CLDN18-ARHGAP26* uses a cryptic splice site internal to exon 5 of *CLDN18* to generate a fusion transcript that was discovered recurrently in multiple cohorts suggesting biological significance (Can-

cer Genome Atlas Research 2014). Six methods compared here (deFuse, SOAPfuse, FusionCatcher, Tophat, MapSplice and FusionMap) are designed to detect non-canonical fusions (Figure 2.7). In all, 2,807 non-canonical fusions have been nominated by the six methods with 33 (1.3%) nominated by two or more methods and 6 (0.2%) by three or more (Figure 2.7c). We find 5 true positives among these with only one fusion reported by multiple methods, indicating a high false negative rate to detect these types of fusions. Although, deFuse successfully identified 3 out of 5 non-canonical true positives, it also nominated 1,297 singleton fusion calls (72/sample). In contrast to the "other" category of fusions identified for canonical fusions, here find 98.8% of all "other" fusions are nominated by one method alone suggesting that these are likely false positives.

2.3.3.2 Comparisons using primary and relapse tumors of one individual

We next evaluated MOJO using a primary (125 million, 2x100bp reads) and a relapse (87 million, 2x100bp reads) transcriptome from a patient with high grade serous ovarian carcinoma using seven methods (Figure 2.3). FusionCatcher and FusionHunter were excluded due to execution failures. All seven methods were run under parameters requiring at least two anchor reads to support each fusion call (Supplementary Table 2.2). An aggregate of 2,310 fusion calls were nominated with 161 (6.9%) canonical fusion calls. We identified and successfully validated all 32 canonical fusions that were nominated by three or more methods using RT-PCR (Figure 2.3a, Supplementary Table 2.5).

We identified 32 canonical fusion calls (12 in primary and 20 in relapse) nominated by three or more methods. All 32 fusions were successfully validated using RT-PCR. These 32 events comprise 22 unique fusions with 10 fusions detected in both primary and relapse (Table 2.1). Interestingly, 10/12 fusions in the primary are generated by intra-chromosomal rearrangements of chromosome 19 suggesting a possible chromothripsis event (Stephens et al. 2011). In addition, of the 10 fusions introduced after sampling the primary tissue, only one event is from chromosome 19, further suggesting a potential catastrophic event introduced

on chromosome 19 early on.

Among the seven methods compared here, only MOJO successfully identified all 32 fusion calls (Table 2.1). One of the challenges presented by the existing methods is the inconsistent performance across different, but closely related, RNA-seq datasets. Demonstrating this, we find that for 9/10 fusions shared between primary and relapse, at least one of the methods that detected the fusion in one of the two tumors failed to detect the same fusion in the other (Table 2.1). For example, *ELL-SMARCA4* is supported by 1,125 and 2,507 anchor reads in the primary and the relapse tumors, respectively. ChimeraScan and SOAPfuse identified this fusion in the primary but failed to detect it in the relapse tumor. And vice versa with *NUSAP1-CASC4*, with both these methods identifying the fusion in the relapse but missing it in the primary tumor. Similarly, *RAB11B-USHBP1* was detected by deFuse in the relapse but not in the primary. These observations suggest that the inherent design principles adopted by existing methods may not account for the heterogeneity observed in the primary tumors and library preparations. Overall, apart from MapSplice where the five false negatives are due to low expressed fusions, all other methods failed to identify fusions with a broad range of expression levels (Figure 2.3b).

Using the "other" category of fusion calls as a proxy for false positives, we find that deFuse and FusionMap nominated the highest number of these events (Figure 2.3c). Although we cannot exclude the possibility this category of "other" fusions may contain true fusions, given we infer that only 3/103 of these "other" events are shared by multiple callers, this set of fusions is enriched for false positives. We show that the two "other" fusion calls in MOJO are supported by two or fewer reads suggesting that higher threshold of anchor reads is an accurate parameter for controlling false positives (Figure 2.3d).

2.3.4 Runtime performance comparisons

Fusion discovery is a computationally intensive task that involves aligning to the normal transcriptome, followed by a series of alignments to the rearranged transcriptomes. Fur-

5' Gene				3' Gene				Primary sample			Relapse sample		
Name	Chrom	Strand	TSS (Mb)	Name	Chrom	Strand	TSS (Mb)	Is fusion detected	# Methods	# Anchor reads	Is fusion detected	# Methods	# Anchor reads
SKI	1	+	2.16	PRKCZ	1	+	1.98	N	-	-	Y	5	8
RNF181	2	+	85.82	KCMF1	2	+	85.2	N	-	-	Y	3	3
FAM20C	7	+	0.19	EIF3B	7	+	2.39	N	-	-	Y	4	4
TAX1BP1	7	+	27.78	RMI2	16	+	11.34	Y	3	8	Y	3	5
SURF1	9	-	136.22	REXO4	9	-	136.27	N	-	-	Y	3	5
REXO4	9	-	136.27	SERPINB10	18	+	61.56	N	-	-	Y	4	7
ATP11A	13	+	113.34	CUL4A	13	+	113.86	N	-	-	Y	3	2
NUSAP1	15	+	41.62	CASC4	15	+	44.58	Y	5	45	Y	7	37
IQGAP1	15	+	90.93	ZNF774	15	+	90.89	N	-	-	Y	5	12
SKAP1	17	-	46.21	SPOP	17	-	47.68	N	-	-	Y	6	5
KANK3	19	-	8.39	DOCK6	19	-	11.31	Y	6	13	N	-	-
RAB11B	19	+	8.45	USHBP1	19	-	17.36	Y	3	21	Y	4	6
LDLR	19	+	11.2	HAS1	19	-	52.22	Y	6	109	Y	7	186
AKAP8L	19	-	15.49	GATAD2A	19	+	19.5	Y	3	42	Y	4	104
ELL	19	-	18.55	SMARCA4	19	+	11.07	Y	7	1125	Y	5	2607
URI1	19	+	30.41	CYP4F11	19	-	16.02	Y	4	13	N	-	-
FBXO17	19	-	39.43	C19orf18	19	-	58.47	Y	6	34	Y	7	99
SEPW1	19	+	48.28	CTC-459F4.3	19	+	28.28	Y	4	162	Y	4	237
MYBPC2	19	+	50.94	SIRT2	19	-	39.37	N	-	-	Y	5	16
HAS1	19	-	52.22	ANGPTL4	19	+	8.43	Y	3	6	Y	4	5
ZNF787	19	-	56.6	ANO8	19	-	17.43	Y	5	39	Y	7	12
CCDC117	22	+	29.17	HSCB	22	+	29.14	N	-	-	Y	4	13

Table 2.1: List of RT-PCR validated fusions in the Primary and Relapse tumors. 5' and 3' gene information is shown. TSS - transcription start site. 'Is fusion detected' indicates if the fusion was nominated by at least one of the methods from RNA-seq. '# Methods' indicates the number of methods identifying the fusion. '# Anchor reads' indicates the highest number of anchor reads identified by the methods nominating the fusion.

thermore, the downstream filtering steps incorporated by varying methods could contribute to substantial computational burden, resulting in higher costs in both time and equipment. MOJO attempts to alleviate this by implementing a design that is optimized for parallelization as well as reduced I/O burden. Three aspects of the algorithm substantially improve the runtimes. First, to generate discordant reads, we perform a series of iterative alignments using pre-built indexes for bowtie2 and bwa against a specified gene model (Figure 2.6a). During each iteration, the sensitivity of the alignments in increased. Second, to filter out homology based candidate gene fusion nominations, we use a pre-built index consisting of the homology map between all possible genes in the human genome. Third, to identify spurious alignments, we iteratively search for alternate alignments using a combination of bowtie2 and blat that is designed to incrementally increase sensitivity after each iteration (Figure 2.6c-d).

We compare the runtime performance of seven best performing methods using 18 cell line transcriptomes used for comparisons above, as well as, for 36 additional primary tumor transcriptomes from The Cancer Genome Atlas (TCGA) project with a median sequencing depth of 86 million reads (Figure 2.4). We found that MOJO consistently runs 5x faster than deFuse, 7x faster than FusonCatcher and 6x faster than MapSplice. All methods were run on identical hardware and with comparable parameters. We note that a subset of the methods including deFuse, FusionCatcher and MapSplice are also designed to detect non-canonical fusions and as a result may confer additional computational burden.

2.4 Discussion

We have implemented a new algorithm called MOJO for identifying gene fusions from transcriptome sequencing data. The superior performance of MOJO in comparison with existing methods is derived from the efficient implementation of a central tenet of the algorithm that an anchor read supporting a fusion junction cannot be generated by biological or technical artifacts. We therefore developed rigorous filters that attempt to evaluate whether

each anchor read can be aligned canonically to annotated/unannotated transcriptome or whether it can be generated by post-transcriptional events such as back-splicing or intragenic splicing. This downstream stringent filtering allowed us to maximize sensitivity at the initial step of the algorithm by considering all possible candidate gene fusions nominated by discordant clusters.

We evaluated MOJO using a panel of 20 cell line and primary tumor transcriptomes with true positives experimentally determined either in this study or previously. In comparison with eight existing methods, MOJO consistently demonstrated better performance (Figures 2.2, 2.3). MOJO successfully identified 97.7% of all true positives within the transcriptomes analyzed here. We note that the limited number of false negatives are primarily due to the small number of low quality reads supporting the events (Figure 2.3b). Controlling for false positives is a significant challenge within existing methods. Using singleton fusion calls (nominated by only one method) as proxy for false positives, we show that MOJO nominates the lowest number of singletons compared to the next best performing methods, deFuse, SOAPfuse and FusionCatcher. A striking demonstration of our accurate accounting of the various technical artifacts is that, in contrast to other methods, all 10 singleton fusion nominations by MOJO are supported by one (n=8) or two (n=2) anchor reads. This suggests that anchor reads can be effectively used as a threshold to modulate specificity. In contrast, other methods nominated singletons supported by anchor reads ranging across the expression spectrum.

Apart from the number of anchor reads, another key criteria that affects specificity is the anchor length (Figure 2.5). Larger anchor lengths can increase specificity but at the cost of sensitivity when detecting low expressed fusions or transcriptomes with lower sequencing depth. Here, we showed that increasing the anchor length criteria from 10bps to 15bps reduced the number of junctions with ambiguously aligned anchor reads from 12% to <0.2%. MOJO's default parameters require only one anchor read with an anchor length of 10bps to nominate a fusion. However, for transcriptomes with moderate to high sequencing depth, we

recommend increasing this threshold to two anchor reads with an anchor length of 15bps.

One of the limitation of MOJO is its dependency on gene model annotations to identify fusions involving junctions at annotated exon-exon boundaries. Fusions involving novel exons or those that retain intronic sequences are rare, and even if generated, majority of them are likely to be out-of-frame, resulting in loss of function. However, a recent report of a novel recurrent fusion in gastric cancer that uses a novel intra-exonic splice site suggests that such events cannot be ignored. Another limitation of MOJO and transcriptome based fusion discovery, in general, is the inability to correctly classify fusion transcripts generated by technical artifacts such as incomplete genome annotation or biological artifacts such as trans-splicing. Such fusions can manifest at high frequencies in a large analyses and can limit the power of novel fusion discovery. We suggest that such artifacts can accounted for using a sufficiently large panel of normal transcriptomes and by excluding fusions discovered in normal samples from those identified in tumors.

In summary, MOJO provides a significant improvement in balancing sensitivity and specificity compared to existing methods, in detecting gene fusions from transcriptomes. The computationally efficient and scalable implementation of MOJO will enable discovery of fusions from large panel of cancer as well as healthy tissue transcriptomes towards identification of both somatic as well as germline gene fusions.

MOJO is freely available at http://github.com/cband/MOJO along with reference indexes for various gene models including UCSC knownGene, GAF3.0 and Ensembl along with different human genome builds (hg19 and hg38).

2.5 Methods

2.5.1 Minimum Overlap Junction Optimizer (MOJO) Algorithm

2.5.1.1 Definitions

The terms essential to describe the fusion discovery process are first defined in Figure 2.5. If A-B represents the fusion gene with 5' of gene A fused to 3' of gene B, then a discordant read is a paired-end read with one end mapping to gene A and the other to gene B. A discordant cluster comprises all discordant reads that span a predicted fusion junction for A-B. An anchor read is a paired-end read mapping to the fusion junction with the split-read end mapping to the fusion junction while the other-read end maps to either gene A or gene B. The minimum length of overhang of the split-read across the junction is defined as anchor length. An anchor read is defined as an ambiguous anchor read if it can be mapped concordantly to a canonical transcriptome or genome in addition to mapping to the candidate fusion junction.

2.5.1.2 Algorithm Overview

The MOJO algorithm comprises four main steps (Figure 2.6): (i) Identification of candidate gene fusions from clusters of discordant reads, (ii) constructing exon-exon junction library for each gene pair and identifying anchor reads, (iii) filters to remove spurious anchor reads, and (iv) filters to remove spurious junctions.

1. Identification of candidate gene fusions

This step begins with identification of discordant pairs of reads (Figure 2.6a). Raw FastQ paired-end reads are trimmed to remove poor quality bases using the trimming function implemented in BWA (Li et al. 2009). To optimize for speed, all concordantly aligning reads are identified and excluded by aligning all reads to the spliced transcriptome comprising all

possible isoforoms. The remaining unmapped reads are retained for downstream analysis. Next, discordant reads are identified in two iterative steps to improve sensitivity. First, unmapped reads are trimmed to 36bps and aligned to the unspliced transcriptome. In the second iteration, to account for trimmed reads containing splice junctions, the residual full-length unmapped reads are trimmed on their 5' ends and the 36bp reads are realigned to the unspliced transcriptome. Discordant reads are identified as reads with each end aligning to two different genes. Additional filters are applied to remove homology driven artifacts, PCR duplicates and reads mapping to repetitive regions. To improve speed, homology driven artifacts are filtered out using a pre-built index of homology between all possible combinations of genes in the human genome.

An initial set of candidate gene fusions is constructed by clustering discordant reads by breakpoint junction they are predicted to span. To increase sensitivity, multi-mapping reads are assigned to multiple candidate fusion genes. We next control for random chimeras generated by homology driven template switching during RT-PCR or by an unknown process that randomly ligates fragments of transcripts between two genes that are highly expressed. We hypothesize that a real breakpoint should have at least two discordant reads with the mapping positions of reads on either gene not farther than the mean fragment size. We note that this filter does not eliminate all random chimeras but this simple geometric rule reduces the total number of discordant clusters by more than 80% without affecting overall sensitivity. We find that alternate approaches for controlling randomly ligated fragments that are based on abnormal insert size tend to reduce sensitivity. A candidate fusion gene is nominated if the discordant cluster is supported by at a minimum number of discordant reads.

2. Identification of junction mapping anchor reads

For each candidate fusion junction, a maximal set of exons of both partner genes that are likely involved in the fusion junction is predicted from the exons involved in discordant reads.

All exons proximal (3 exons up/downstream) to the exons supported by discordant reads are nominated as junction-exons. We then construct a library of sequences for all possible exon-exon junctions between the predicted junction-exons of both the genes. Fusion junctions with low sequence complexity (di-nucleotide entropy <1.2) are filtered out. All reads not mapped to the spliced transcriptome are realigned to this library to identify potential anchor reads. Reads with anchor length shorter than a user defined length (default=10bp) are filtered out. The affect of the length of anchor region on sensitivity and specificity is discussed below. The other-read end (non junction mapping end) of the anchor read is required to map to one of the two genes of the fusion candidate. Multi-mapping other-end reads are allowed at this stage. In addition, a number of mismatches (default=2%) are allowed in the anchor region to accommodate sequencing errors and single nucleotide polymorphisms. PCR duplicates are filtered out based on alignment positions of the reads.

3. Filters to remove spurious anchor reads

Spurious anchor reads can manifest due to technical and biological artifacts. Technical artifacts can mis-classify a concordantly mapping read as an anchor read in two ways. First, the split-read can map concordantly to a known or unannotated splice junction in the same gene that the other-read maps to. We check for this alternative splicing by aligning the split-read to the fusion partner genes using Smith-Waterman, allowing at most one gap without a gap extension penalty. Second, due to the homology in the genome, the other-read and split-read pair can map concordantly to another region of the genome or transcriptome. We evaluate each anchor read using a series of iterative blat alignments to the genome and transcriptome. In each iteration, we tune the blat parameters to increase sensitivity to properly split a read and search for alternative alignments for the anchor reads. Biological artifacts can be introduced by intra-genic rearrangements or non-canonical splicing processes. Tandemly duplicated exons or inverted exons can result in incorrectly nominating a fusion between two genes that share sequence identity at the rearranged exon but limited overall homology. Recently discovered splicing processes such as back splicing and circular RNAs

can generate transcripts with scrambled exons which can also generate spurious fusion calls similar to the consequences of intra-genic rearrangements. We control for these artifacts by creating a library of back spliced exons for each of the fusion partner genes and filter out anchor reads if the split-read ends align properly to this scrambled exon-junction library.

4. Filters to remove spurious fusion junctions

We identify and correct for three types of spurious junctions. First, a skewed distribution of split-reads at the fusion junction suggests an alignment artifact or inadequate annotation in the reference genome. For a true fusion junction, the distribution of midpoints of split-reads is expected to be centered at the breakpoint fusion junction. We filter out these junctions if the peak of the midpoints of split-reads is more than two standard deviations (default=2) away from the breakpoint junction. We incorporate additional heuristic criteria (exon position in the transcript) to accommodate for depleted coverage at the ends of the transcripts and for low expressed genes. Second, we attempt to control for random chimeras between highly expressed genes that are generated during library preparation. Third, low complexity junction regions could increase false positives. Di-nucleotide entropy of 20bps of sequence from 5' and 3' genes at the fusion junction is calculated individually. If the entropy of either sequence is less than 1.8 (99th-percentile value of entropy of all 20), then at least one high confidence read is required to retain this fusion.

2.5.2 Simulation to estimate the effect of anchor read length on specificity

We next sought to understand what fraction of the gene-pairs are confounded by an ambiguous fusion junction, that is, a fusion junction supported by a spurious anchor read originating from the canonical transcriptome. We identify these ambiguous fusion junctions in two steps. In the first step, for each fusion junction of a gene-pair, 50bp reads are constructed with varying anchor lengths (10, 15, 20 and 25bp). For each of the 31 billion possible junctions between 27,000 non-paralogous pairs of genes, we constructed 9 split-reads

comprising 37TB of single-end reads. All reads were then aligned to the spliced transcriptome and the genome to find all regions (excluding the two genes that generated the read) where split-end reads map. We find 112 million fusion junctions that have at least one split-end read that can be mapped in its entirety to a distinct genomic or transcriptomic locus. In the second step, we compare the ±500bp sequences surrounding the fusion junction and the split-end read aligned target region. We then classify a fusion junction as ambiguous for a given anchor length if a shared identity of at least 50bp contiguous sequence, with at most three mismatches, is detected.

2.5.3 Framework for evaluating MOJO

We evaluated MOJO along with eight other methods (Supplementary Table 2.1). To allow for proper comparisons, we identified and accounted for four criteria that contribute to differences in calls between algorithms that may not be indicative of their performance. First, we make three parameters consistent across the methods: number of discordant reads, number of anchor reads and anchor length (see Supplementary Table 2.2). Second, the methods compared use different annotations: RefSeq, UCSC knownGene and Ensembl. To mitigate the differences due to annotation (mainly singleton calls), we use hg19 coordinates provided by the individual algorithms to collapse all calls onto the Ensembl_v78 reference. We accept a method's fusion call if the predicted breakpoint is ±50kb of the transcription start and end sites of the Ensembl annotated genes. Third, we filter out all fusion calls that are predicted to be read-through events defined by the 5' gene located upstream to the 3' gene in the same orientation and <250kb apart. Finally, a subset of the methods, including MOJO, are not designed to identify fusion transcripts that retain intronic regions or partial exons. Therefore, we exclude all fusion calls with breakpoints not involving annotated exon-exon boundaries that are nominated by various callers in this study. Comparisons involving non-canonical fusions are discussed in supplementary (Figure S3).

2.6 Appendix: Figures

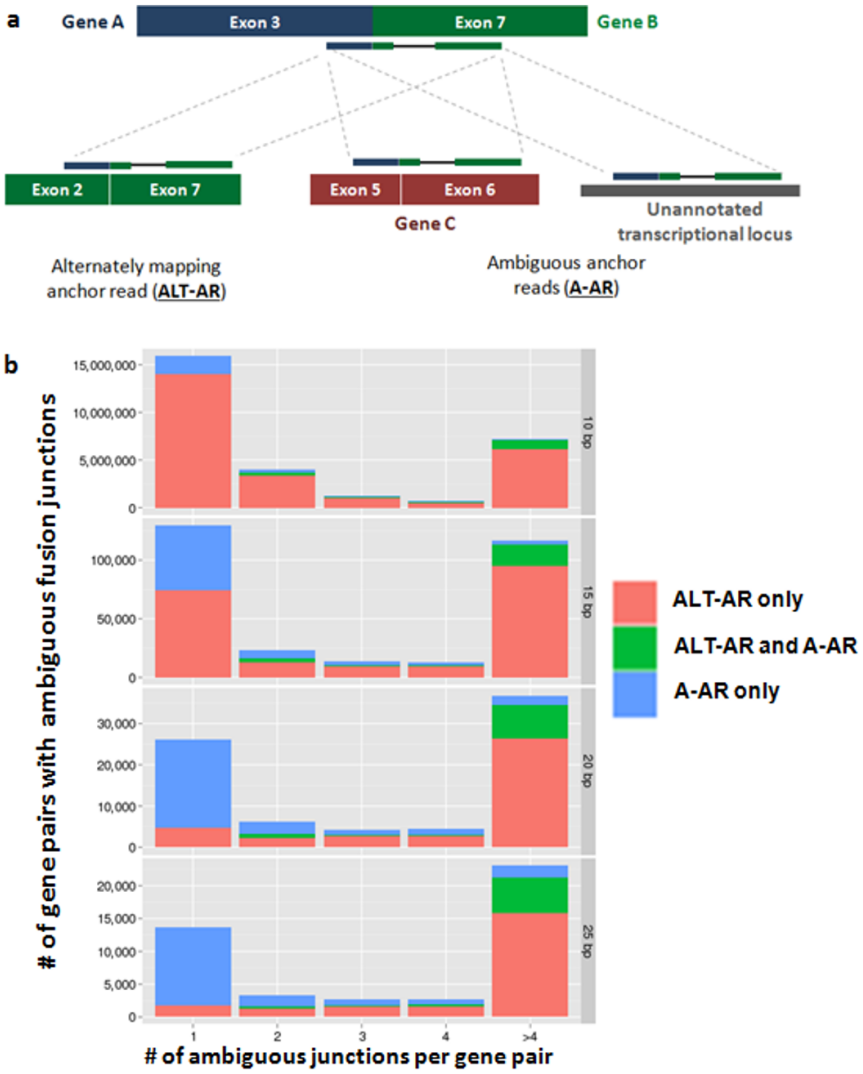


Figure 2.1: Effect of anchor length on specificity of the anchor read. (a) an anchor read that supports a potential fusion junction can also be generated by the canonical transcript from one of the partner genes (alternately mapping anchor reads, ALT-AR), from a distinct annotated/unannotated transcriptional locus (ambiguous anchor reads, A-AR). (b) Estimation of the number of ALT-AR and A-AR reads from the analysis of all possible fusion junctions between all genes in the transcriptome. Ambiguous anchor reads (ALT-ARs and A-ARs) are paired-reads originating from a function junction but that can also be mapped to one of the partner genes (ALT-AR) or to a distinct transcriptional locus (A-AR). # of gene pairs with ALT-AR and A-ARs are shown on the y-axis. # of ambiguous fusion junctions per gene pair are shown on the x-axis. Gene-pairs with more than four ambiguous junctions are grouped into >4 category. Each panel corresponds to anchor reads of varying anchor lengths. Different scales on y-axis is used for each of the anchor length panel to allow for visualization.

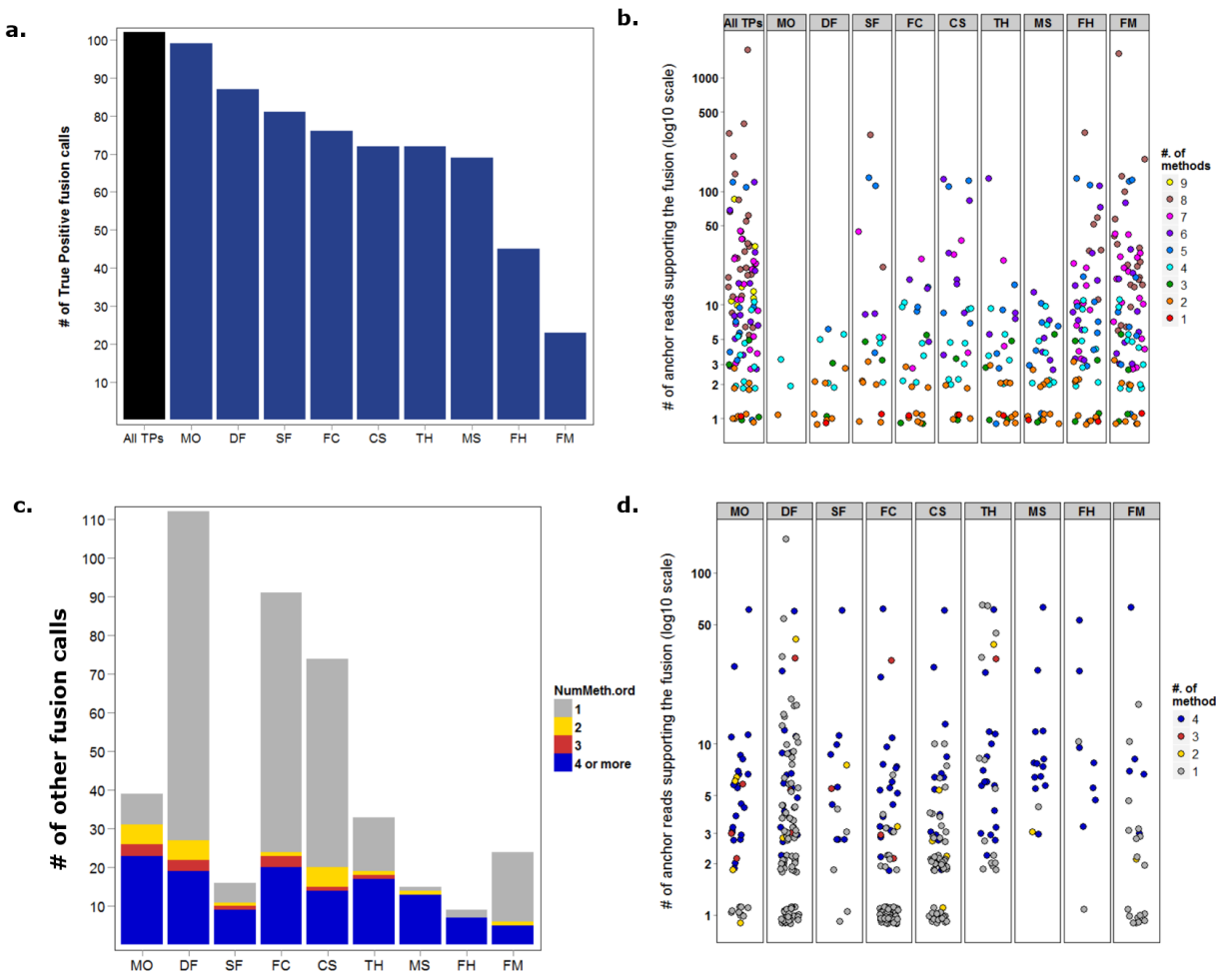


Figure 2.2: Evaluation of sensitivity and specificity of methods to detect fusions involving "annotated" exon-exon boundaries using 18 cancer cell lines. a. Total number of true positives called by each of the eight methods. "All TPs" shows the aggregate number of true positives detected by at least one method. b. False negatives by each of the methods is shown here. X-axis shows the number of reads supporting each of the false negative fusion. The anchor read count used for each fusion is the highest reported by the methods supporting it. Each dot is colored by the number of different methods supporting. Each data point is plotted with 5% noise to allow for proper visualization of overlapping points (geom_jitter in R). c. Total number of "other" fusion calls (a proxy for false positives) reported by each of the methods. Coloring indicates the total number of methods supporting the fusion. Singleton calls are represented in grey. d. Number of anchor reads supporting the "other" fusion calls shown in c. "All TPs" - all true positives (see Supplementary Table 2.3), "MO" - MOJO, "DF" - deFuse, "SF" - SOAPfuse, "FC" - FusionCatcher, "CS" - ChimeraScan, "TH" - tophat-fusion, "MS" - MapSplice, "FH" - FusionHunter, "FM" - FusionMap

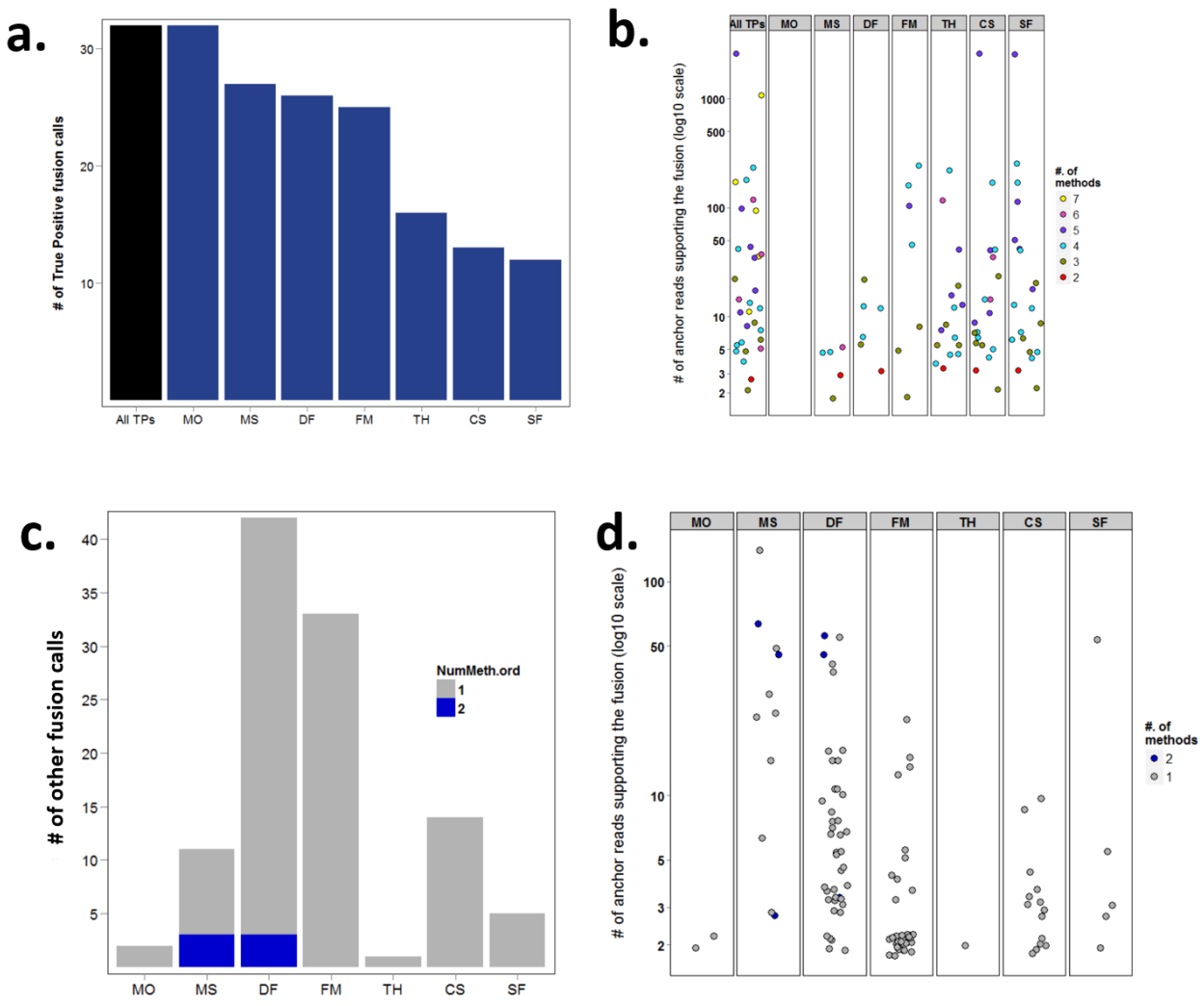


Figure 2.3: Evaluation of sensitivity and specificity of methods to detect fusions involving "annotated" exon-exon boundaries using 2 primary tumor transcriptomes. a. Total number of true positives called by each of the eight methods. "All TPs" shows the aggregate number of true positives detected by at least one method. b. False negatives by each of the methods is shown here. X-axis shows the number of reads supporting each of the false negative fusion. The anchor read count used for each fusion is the highest reported by the methods supporting it. Each dot is colored by the number of different methods supporting. Each data point is plotted with 5% noise to allow for proper visualization of overlapping points (geom_jitter in R). c. Total number of "other" fusion calls (a proxy for false positives) reported by each of the methods. Coloring indicates the total number of methods supporting the fusion. Singleton calls are represented in grey. d. Number of anchor reads supporting the "other" fusion calls shown in c. "All TPs" - all true positives (see Supplementary Table 2.3), "MO" - MOJO, , "MS" - MapSplice, "DF" - deFuse, "FM" - FusionMap, "TH" - tophat-fusion, "CS" - ChimeraScan, "SF" - SOAPfuse

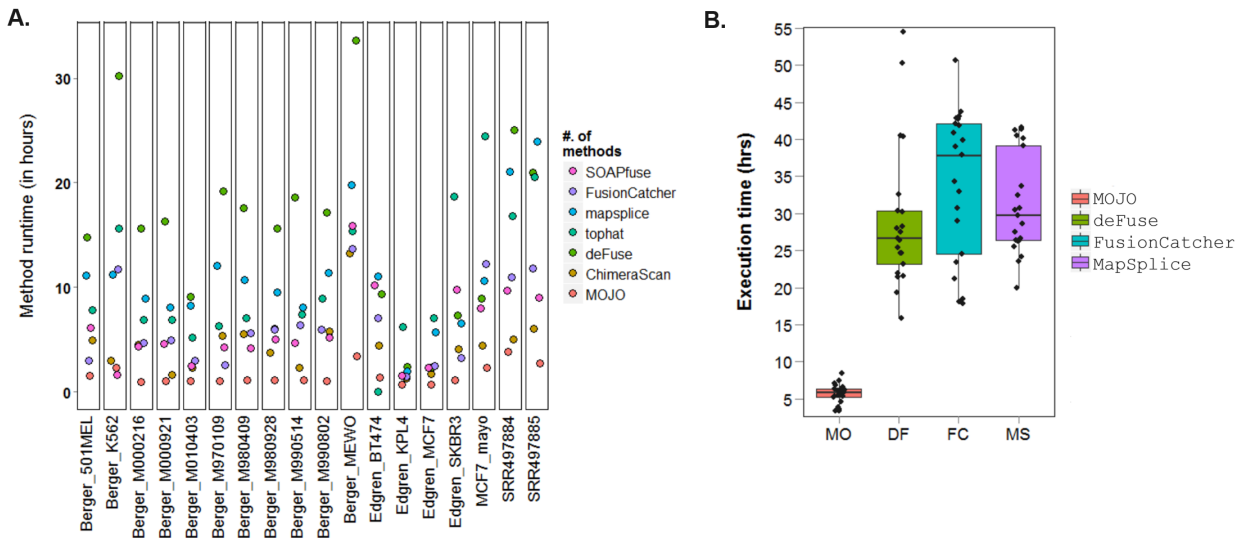


Figure 2.4: Runtime comparison of various fusion callers including MOJO . a. 18 cell line transcriptomes with sequencing depths ranging from 14 to 43 million reads were used for comparisons. b. 36 TCGA transcriptomes with a median sequencing depth of 86 million reads were used to evaluate runtime performance. Only the top four best performing methods were compared here. All comparisons in (a) and (b) were made on identical hardware with 16 cores and 64GB memory.

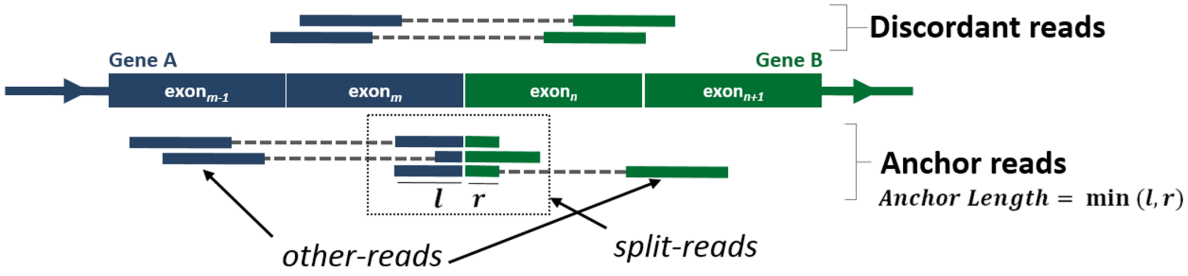


Figure 2.5: Illustration of terms to describe fusion discovery. Discordant reads are paired-end reads with each end mapping to distinct genes. Paired-end reads with one end mapping the junction (split-end) and the other end mapping to one of the two partner genes (other-end) are defined as Anchor reads. If the left and the right overhang of the split-end mapping to the junction are defined as l and r, respectively, then the anchor length is the minimum of the two overhangs.

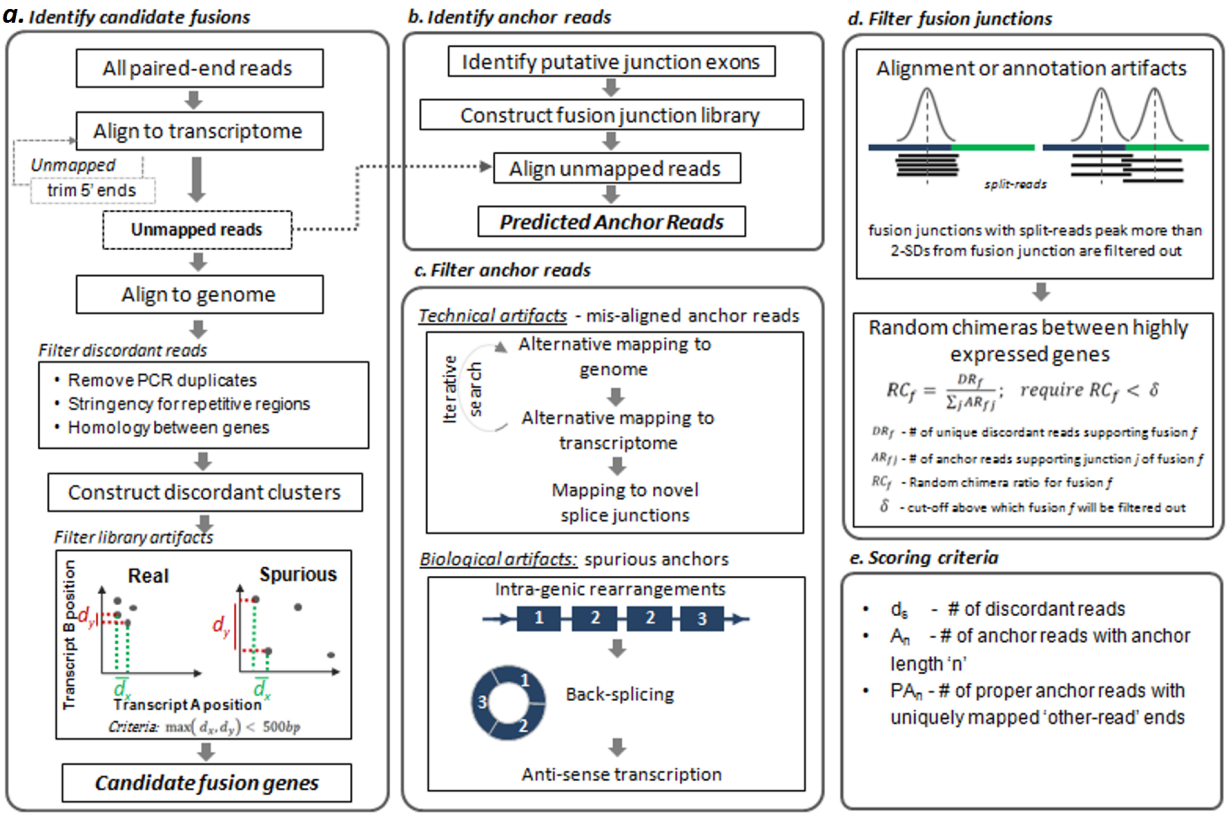


Figure 2.6: MOJO algorithm overview. See text.

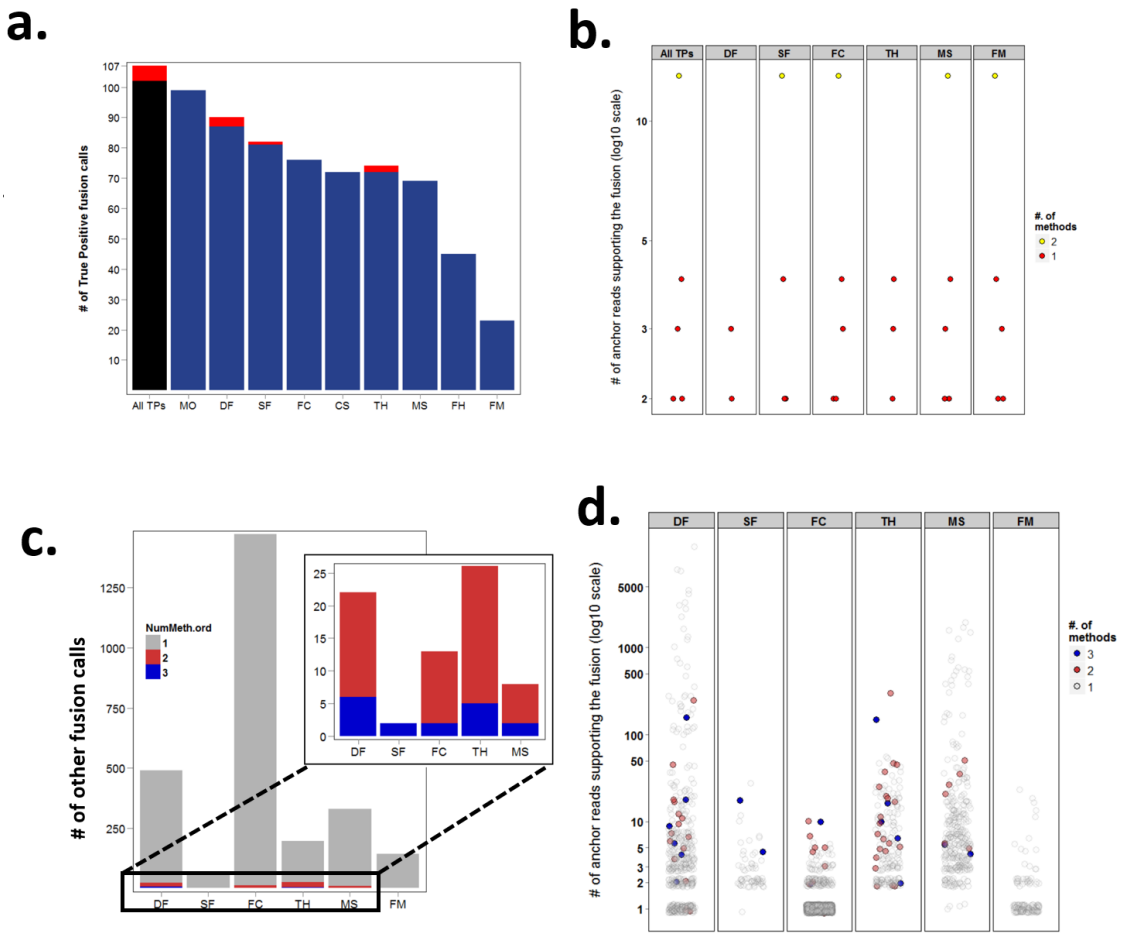


Figure 2.7: Evaluation of sensitivity and specificity of methods to detect fusions involving "unannotated" exon-exon boundaries using 18 cancer cell lines. a. Total number of true positives called by each of the eight methods. Fusions involving unannotated exon-exon boundaries are shown in red. "All TPs" shows the aggregate number of true positives detected by at least one method. b. False negatives by each of the methods is shown here. X-axis shows the number of reads supporting each of the false negative fusion. The anchor read count used for each fusion is the highest reported by the methods supporting it. Each dot is colored by the number of different methods supporting. Each data point is plotted with 5% noise to allow for proper visualization of overlapping points (geom_jitter in R). Only the six methods that can detect non-canonical fusions are shown here. c. Total number of "other" fusion calls (a proxy for false positives) reported by each of the methods. Coloring indicates the total number of methods supporting the fusion. Singleton calls are represented in grey. d. Number of anchor reads supporting the "other" fusion calls shown in c. Singleton calls are shown with transparency to allow for visualization of "other" calls shared between methods. "All TPs" - all true positives (see Supplementary Table 2.3), "MO" - MOJO, "DF" - deFuse, "SF" - SOAPfuse, "FC" - FusionCatcher, "CS" - ChimeraScan, "TH" - tophat-fusion, "MS" - MapSplice, "FH" - FusionHunter, "FM" - FusionMap

2.7 Appendix: Supplementary Tables

Study	Cell line	Tissue type	Paired-end reads		Total # of fusions reported by:	
			Type	Depth (mill)	primary paper[a]	other studies[b]
Edgren et al.	MCF7	Breast	2x50bp	12.8	3	58
	BT474	Breast	2x50bp	29.7	11	12
	KPL4	Breast	2x50bp	10.2	3	0
	SKBR3	Breast	2x50bp	42.3	9	13
Berger et al.	MEWO	Melanoma	2x50bp	42	0	0
	501MEL	Melanoma	2x50bp	14.9	4	0
	K562	CML	2x50bp	31.3	2	1
	Melanoma patient derived short-term cultures from 8 individuals		2x50bp	Mean: 14.2 Median: 15.1	7	0
Asmann et al.	MCF7	Breast	2x50bp	29.9	8	53
Jia et al.	T24	Bladder	2x90bp	32.2	8	0
	5637	Bladder	2x90bp	36.8	9	0

Supplementary Table 2.1: Description of cell lines used in this study for performance evaluation of 9 fusion callers. [a] number of fusions reported by the primary study that generated the data. [b] fusions identified by other studies in the respective cell lines.

Method	Version	Configuration	
		High sensitivity	High confidence
Tophat	2.0.9	Tophat2 --fusion-search --fusion-anchor-length 10 --fusion-min-dist 250000 Tophat-fusion-post --num-fusion-reads 1 --num-fusion-pairs 2 --num-fusion-both 3	Tophat2 --fusion-search --fusion-anchor-length 20 --fusion-min-dist 250000 Tophat-fusion-post --num-fusion-reads 2 --num-fusion-pairs 2 --num-fusion-both 4
deFuse	0.6.1/0.5.0**	config.txt -span_count_threshold 2 -split_min_anchor 10	config.txt -span_count_threshold 2 -split_min_anchor 20
FusionHunter	1.4	MINOVLP 10	MINOVLP 20
FusionMap	6.0.0.14	MinimalFusionAlignmentLength=15	MinimalFusionAlignmentLength=20
SOAPfuse	1.25	-PA_s08_min_sum_reads=3 -PA_s05_the_minimum_span_reads _for_candidate=2 -PA_s07_junc_read_map_both_ sides_at_least=10 -PA_s08_min_bases_covered_both_ sides_around_fuse_point=10	-PA_s08_min_sum_reads=3 -PA_s05_the_minimum_span_reads _for_candidate=2 -PA_s07_junc_read_map_both_ sides_at_least=20 -PA_s08_min_bases_covered_both_ sides_around_fuse_point=20
MapSplice	2.1.2	<all default>	<all default>
Chimerscan	0.4.5a	-anchor-min 10 –anchor-length 10	-anchor-min 20 –anchor-length 20
FusionCatcher	0.99.2	-s 2 -r 1 -a 10	-s 2 -r 2 -a 20

Supplementary Table 2.2: Configurations of published algorithms used for comparisons. Parameters specified for each of the methods is indicated for the 'High sensitivity' (for 18 cell lines) and 'High specificity' (for two primary tumors) modes. ** - A subset (4/20) of tumors failed to run with version 0.6.1. Older version of deFuse was used in those cases.

Supplementary Table 2.3: (See workbook "Table S2.3" in supplementary file Supplementary.Tables.Chapter2.xlsx associated with this dissertation). List of previously reported true positives in the 18 cancer cell lines. Specific columns are:

Column 1: Cell line name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: PubmedID of study reporting the fusion
Column 5: Name of the reference reporting the fusion
Column 6: Chromosome of 5' gene
Column 7: Strand of 5' gene
Column 8: Chromosome of 3' gene
Column 9: Strand of 3' gene
Column 10: Flag to indicate if the fusion event is a read-through event. (1: if distance between genes on the same strand is <200kb and the 5' gene is upstream of the 3' gene, 0: otherwise)
Column 11: Distance between the two genes
Column 12: Is the fusion a non-canonical event. (1: if the fusion junction involves an unannotated exonic boundary or an intronic region, 0: otherwise)

Supplementary Table 2.4: (See workbook "Table S2.4" in supplementary file Supplementary.Tables.Chapter2.xlsx associated with this dissertation). Fusion calls nominated by eight different methods within 18 cell lines. Specific columns are:

Column 1: Sample name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: MOJO (1: fusion nominated; 0: not nominated)
Column 5: ChimeraScan (1: fusion nominated; 0: not nominated)
Column 6: deFuse (1: fusion nominated; 0: not nominated)
Column 7: tophat (1: fusion nominated; 0: not nominated)
Column 8: SOAPfuse (1: fusion nominated; 0: not nominated)
Column 9: MapSplice (1: fusion nominated; 0: not nominated)
Column 10: FusionMap (1: fusion nominated; 0: not nominated)
Column 11: FusionHunter (1: fusion nominated; 0: not nominated)
Column 12: FusionCatcher (1: fusion nominated; 0: not nominated)
Column 13: Total # of methods nominating the fusion
Column 14: Is fusion previously reported (1: yes; 0: no)
Column 15: # of anchor reads supporting the fusion (if nominated by multiple methods, the maximum value is used)
Column 16: Distance between the two genes
Column 17: Is canonical fusion? (1: yes, fusion junction involves annotated exon-exon boundaries; 0: no)

Supplementary Table 2.5: (See workbook "Table S2.5" in supplementary file Supplementary.Tables.Chapter2.xlsx associated with this dissertation). Fusion calls nominated by seven different methods within two primary tumors. Specific columns are:

Columns 1-17: identical to 2.4

2.8 Contributions

I designed and developed the algorithm with input from my advisor, Kevin White. I wrote all the code, debugged the algorithm and wrote documentation. I would like to thank Christopher Brown, Megan McNerney and Thomas Stricker for constructive discussions that improved specific aspects of the algorithm. I performed all the comparisons with other published methods in both cell lines and primary tumors. Reanne Bowlby, Andy Mungall and Karen Mungall at British Columbia Cancer Agency's Genome Sciences Center performed and analyzed the de novo assembly of the genomes and transcriptomes.

I performed all the RT-PCR validations for fusions nominated in the primary tumors. I would like to thank Thomas Stricker for offering expertise on optimizing assays to detect fusion transcripts from picograms of RNA. I also would like to thank April Peterson for performing RT-PCR validations that helped identify certain sources of false positives that eventually led to improvement of the algorithm. Transcriptome and whole genome sequencing of the primary and relapse tumors were generated by the High-throughput Genome Analysis Core (HGAC) at the Institute for Genomics and Systems Biology (IGSB).

Compute resources during development and performance comparisons with other methods were provided by Bionimbus (Robert Grossman), Center for Research Informatics (CRI) and Argonne National labs (Beagle). I would like to especially thank the CRI for making available the 1TB memory machine that was extensively used for memory profiling, optimization and scalability testing of the algorithm.

I would also like to thank the late Dr. Janet Rowley for inspiration and encouragement for the development of these methods.

CHAPTER 3

DISCOVERY AND CHARACTERIZATION OF NOVEL RECURRENT FUSION GENES IN 33 HUMAN CANCERS

3.1 Abstract

The diagnostic, prognostic and therapeutic potential of oncogenic fusion events in cancer has been well established. Here, using a highly sensitive and specific method, we analyzed 9,360 primary tumors in The Cancer Genome Atlas (TCGA) and identified 19,818 somatic fusion transcripts. We find substantial variability in the rate of fusions across cancers that is strongly correlated with the overall genomic instability within the tumors. In addition to finding enrichment for known cancer genes, we discover a number of novel recurrently fused genes that may be candidates for novel gain- or loss-of-function events. Notably, we find fusions involving known tumor suppressor genes in 1.3% of all patients for which we did not find a corresponding copy number alteration on SNP arrays. Across all cancers, we identified 1,144 (94% novel) recurrent fusion genes that can generate chimeric proteins, result in potentially truncated proteins or cause dysregulation. Functional evaluation of three of the most recurrent novel fusion events identified cancer-like phenotypes in various genetic backgrounds. Taken together, our findings lay the groundwork for further understanding of the role of fusions in cancer and for developing new diagnostics and, ultimately, personalized cancer treatment.

3.2 Introduction

Fusion genes constitute an important class of somatic mutations in cancer. Since the discovery of the translocation that generated the *BCR-ABL1* gene fusion in chronic myelogenous leukemia (CML) (Rowley 1973b), many oncogenic fusions have been discovered that: confer strong transformative potential (Singh et al. 2012; Soda et al. 2007), serve as mark-

ers for highly specific and successful drug targeting (Druker et al. 2006; Kwak et al. 2010), and, strongly correlate with the morphological and pathological characteristics of subtypes of leukemias (Vardiman et al. 2009), sarcomas (Taylor et al. 2011) and carcinomas (French et al. 2004; Behboudi et al. 2006; Honeyman et al. 2014). Although sequencing studies have narrowed the discrepancy in recurrent fusions reported in leukemias and sarcomas in comparison with carcinomas, reports of highly recurrent oncogenic fusions in the latter remain rare. A more complex understanding of oncogenic fusions for epithelial tumors is emerging with the discovery of many rare but potentially clinically actionable fusions in morphologically distinct cancer types. For example, the oncogenic and kinase inhibitor responsive fusion gene, *FGFR3-TACC3*, discovered in glioblastomas ($\tilde{3}$%), is also reported at frequencies of less than 1% in squamous cell carcinomas of lung, esophagus, and, head and neck (Singh et al. 2012; Wu et al. 2013b). Similarly, druggable *EML4-ALK* is detected in non-small cell carcinomas (4-7%) as well as at <1% frequencies in colorectal, breast and renal cell cancers (Shaw et al. 2013b; Soda et al. 2007). Furthermore, key genes such as *ALK*, *ROS1*, *RET* or those within gene families, such as *FGFR* and *NTRK*, have been identified to be fused to multiple partners across multiple cancers (Shaw et al. 2013b). The established clinical relevance of the fusions highlighted above (Shaw et al. 2013b) demonstrates the significance of finding novel fusions that may be less frequent within cancer types but when observed across multiple cancers, collectively constitute an appreciable number of cases to compel further study and functional validations.

The primary focus of fusions in cancer has been predominantly on events that generate chimeric proteins. However, other less studied consequences of fusion events include protein truncations (Veeraraghavan et al. 2014), transcriptional de-regulation by juxtaposition of regulatory events (Northcott et al. 2012) and those resulting in loss-of-function of one or both genes (Perner et al. 2007). To our knowledge, no study has evaluated these various types of fusions across a wide range of human cancers. We hypothesized that with large sample sizes and efficient computational approaches to accurately identify fusions, we could

comprehensively characterize the landscape of various types of fusion events as well as identify novel recurrent fusions that may be candidates for events affecting tumor initiation and progression.

In this study, we used a highly sensitive and specific approach to identify and study the characteristics of fusions across 33 human cancers. We demonstrate a >92% validation rate and at least 28% higher sensitivity for our pipeline as compared to previous approaches (Stransky et al. 2014; Yoshihara et al. 2014). We show that the frequency of fusions within each patient varies by cancer type and is strongly correlated with overall genomic instability. We also find that somatic fusions are enriched for known cancer associated genes. Specifically, fusions disrupting tumor suppressors or those involved in transcriptional de-regulation of the respective downstream partner gene are common events across all cancers. Furthermore, we identify a number of novel recurrent fusions and show that they induce proliferation, invasion and migration in specific genetic contexts. Collectively, our findings present a thorough overview of fusions across cancers and demonstrate that functional characterization of novel events will be an important step towards further expanding the catalogue of actionable fusions.

3.3 Results

3.3.1 Somatic fusion transcript discovery

We performed fusion discovery on 9,360 tumor transcriptomes across 33 cancer types in TCGA using Minimum Overlap Junction Optimizer (MOJO) in the highest sensitivity mode (Figure 3.6, Supplementary Tables 3.1, 3.2). To account for spurious recurrent fusions that could manifest in such large pan-cancer analysis, we applied three post processing filters (Figure 3.1A, Figure 3.7, See Section Section 3.5.3.1 for details). First, for each fusion junction, we required uniform and stringent coverage of anchor reads (reads with one end mapping to the junction and the other to one of the two genes of the fusion pair). If a

fusion junction is recurrently nominated, the anchor read evidence is pooled across samples to enhance sensitivity. Second, to further account for alignment artifacts, we performed sensitive gapped alignment of the junction sequence and filtered out the fusion junction if the junction sequence aligns concordantly in three contiguous segments or less. A fusion gene is retained only if at least one of the fusion isoforms detected in the tumors satisfies the previous two criteria. Third, using 10,340 transcriptomes comprising tissues from healthy donors (Supplementary Table 3.4, 3.5), we filtered out all TCGA fusion calls that were also detected in two or more normal samples (Figure 3.8, Methods).

We evaluated the performance of our pan-cancer fusion discovery pipeline (MOJO-PC, Figure 3.1A, Figure 3.7) using a panel of 55 primary tumors from six cancers in TCGA in which recurrent fusions have been previously reported (Supplementary Table 3.6). We compared MOJO-PC, as well as the standalone MOJO (MOJO-S) alongside other commonly used fusion algorithms, deFuse (McPherson et al. 2011), MapSplice (Wang et al. 2010b) and FusionCatcher. Both MOJO-PC and MOJO-S recovered 56/57 (98.2%) of true positives (Figure 3.1B, Supplementary Tables 3.7-3.10, see Methods for extended comparisons). Among the 1,311 fusions that are collectively nominated by all methods, and that are not supported by prior experimental evidence ('other' fusion calls, Figure 3.1C), a subset could be real fusion events. However, singleton fusions among them are strong candidates for false positives (see Chapter 2). Although, deFuse and FusionCatcher called 52/57 (91.2%) of the true positives, both methods nominated a large number of 'other' singleton fusions (Figure 3.1C). 90% of MOJO-PC's 'other' fusion calls are supported by at least one other method and 59% by all four methods, demonstrating its high specificity.

To further evaluate specificity, we selected twelve cancer cell line transcriptomes in the Cancer Cell Line Encyclopedia (CCLE) and performed RT-PCR validations on all fusions nominated using the MOJO-PC pipeline. We achieved an experimental validation rate of 91.8% (123/134) (Figure 3.1D, Supplementary Table 3.11). We also compared MOJO analysis with two previous studies, Yoshihara et al. 2014 and Stransky et al. 2014, that reported

fusions in subsets of samples analyzed in this study. For these comparisons, we used fusions previously reported primarily in COSMIC and Mitelman databases as true positives ('known fusions', Supplementary Table 3.14). We demonstrate 27% (Figure 3.1E, Supplementary Table 3.12) and 39% (Figure 3.1F, Supplementary Table 3.13) higher sensitivity when compared to these studies and using the same criteria, with an overall false negative rate of <3%, suggesting that the rigorous series of filters we applied did not compromise sensitivity.

3.3.2 Characteristics of somatic fusion transcripts

We nominated 19,818 fusion transcripts across 9,360 primary tumors (2.12/tumor), as well as an additional 243 fusions in 64 recurrent and metastatic tumors (3.8/tumor) (Figure 3.7, Supplementary Table 3.15-3.17). 3.4% of fusion transcripts in primary tumors (669/19,818) have been previously reported in non-TCGA cohorts (Figure 3.2A) (known fusions in COSMIC and Mitelman databases, see Supplementary Table 3.14). An additional 19.7% (3,898) of fusion events involve a gene that is among the 832 genes previously reported to be involved in fusions. Protein coding potential and expression levels of the fusions are important determinants of potential function. Of the 67% of fusions that are generated by the fusion of partial coding sequence regions of two genes (CDS-CDS), two-thirds are predicted to be in-frame (Figure 3.2B). 16.3% of all events involve the fusion of the upstream partner gene to the 5'-UTR of the downstream gene, resulting in potential dysregulation of the latter. Overall, at least 60% of novel and 82.3% of the previously reported fusions are predicted to generate an in-frame transcript.

We next investigated the expression characteristics of the fusions. Using normalized anchor read count as a proxy for expression of the fusion, we find that 43.8% of all fusions are supported by 5 or fewer anchor reads. However, we also find that 32.5% of known fusions, including known oncogenic events such as *FGFR3-TACC3* (n=9), *EML4-ALK* (n=3), *CCDC6-RET* (n=3) and *ETV6-NTKR3* (n=2) are expressed below this threshold, indicat-

ing that these fusions can still be biologically significant (Figure 3.2C). Furthermore, we observed multiple splice isoforms for a fusion within the same tumor for 33.3% of novel and 54% of known fusions, indicating that this phenomenon is not a consequence of aberrant splicing (Figure 3.2D). We next assessed the proportion of fusion transcripts with corresponding segment breaks on the copy number profile of the partner genes. We observed a median of 38% of fusion transcripts across cancers have segment breaks within 100kb of both genes involved in the fusion (Figures 3.10-3.11). A similar overall concordance of 26% is observed for known fusion genes. Although it is expected that all of these fusion events correspond to genomic rearrangements, these results highlight that, sub-clonality, copy number neutral events and sensitivity issues associated with SNP array-based copy number calling can contribute significantly to the false negatives in fusion discovery.

3.3.3 Detection of somatic fusion transcripts in normal samples

Somatic fusion transcripts that are detected in both the primary as well as the corresponding adjacent normal tissues can be indicative of an early event or simply a consequence of infiltrating tumor cells. Our approach to use a large panel of independent control samples allowed us to identify 137 fusion events across 74 out of 678 tumor adjacent normal tissues. Among the 75 fusion transcripts that were detected in both the tumor-normal pairs, we find four *TMPRSS2-ERG* and one *NDRG1-ERG* fusion events. *TMPRSS2-ERG* has been previously reported as an early event in prostate cancer tumor progression (Perner et al. 2007). We also find one fusion transcript, *ITPKC-LTBP4*, in primary, metastatic and adjacent normal tissues of one breast cancer patient (TCGA-BH-A18V). 11 fusions are detected in each of the primary and metastatic tissues with 8 shared between them. However, we find only one fusion, *ITPKC-LTBP4*, in the adjacent normal despite five other fusion transcripts in the primary tumor that are expressed at levels higher than this fusion. This suggests that our finding in the adjacent normal is more likely to be an early event rather than a consequence of contaminating tumor tissue. Across the cohort, we find this fusion in three tumors, one

each of BRCA, PAAD and UCS, with the nuclear export signal containing N-terminus of *ITPKC* fused to C-terminus of *LTBP4* containing all the EGF-like domains (Figure 3.14). *LTBP4* is previously shown to regulate the activity of *TGFB1* but its role in tumorigenesis remains to be evaluated.

3.3.4 Characteristics of fusions across 33 cancers

We find a substantial variability in frequency of fusions per tumor, ranging from a mean of 8.6 (median 7) in ovarian cancer to 0.18 (median 0) per tumor in uveal melanoma (UVM) (Figure 3.2E). Although only 55% of tumors in TCGA are predicted to have at least one fusion, we find significant variation in the proportion of tumors with fusions across cancers (Figure 3.12). 96% of ovarian tumors had at least one fusion while only 11.3% of the UVM tumors were fusion-positive. Consistent with the cancer-type specific genomic instability as a driving factor for this variation, we find a significant correlation between numbers of fusions and the number of somatic copy number alterations (sCNAs) for 22 out of 33 tumor types with highest number of fusions (overall correlation, rho=0.57, p <2.2e-16, Figures 3.2E, 3.13, Supplementary Table 3.18).

Fusions generated by duplications are the most frequent (44.3%) events followed by those resulting from translocations, inversions and deletions at 23.3%, 20.1% and 11.3%, respectively (Figure 3.2F). Consistent with previous estimates for median size of sCNAs at 0.7Mb (Zack et al. 2013), we find that 48.3% of all fusions are between genes that are less than one megabase apart, ranging from 68.8% in kidney chromophobe to 14.7% in leukemia. Four tumor types with the highest proportion of long distance fusions (>1Mb between fused genes or translocations), LAML (14.7%), SARC (15.3%), PRAD (22.1%) and THCA (22.4%) also comprise the four out of five tumor types with the highest number of known fusions.

3.3.5 Enrichment of cancer associated genes in fusion transcripts

16.8% (3,349) of all fusion transcripts in primary tumors involve a gene that is previously associated with cancer (Futreal et al. 2004; Vogelstein et al. 2013) (empirical p-value <1e-6, Figure 3.3A, Supplementary Table 3.19, see Methods). At least 22.8% of all patients across 33 cancer types (median 16.7%) in TCGA are predicted to have a novel fusion transcript involving a known cancer gene (Figure 3.15). Of the 514 events involving known oncogenes, 78% are predicted to be in-frame (Figure 3.3B, Supplementary Table 3.20). *ERBB2* (n=101) and *EGFR* (n=67) are well known oncogenes that are frequently amplified across multiple cancers, which indicates that focal genomic instability may be the driver of some recurrently fused genes. Fifteen distinct fusion partners were identified for the MET proto-oncogene; *CAPZA2* (n=9), *CAV1* (n=6) and *ST7* (n=3) are the most common partners among which only five *CAPZA2* and one *ST7* fusions retain an intact kinase domain, suggesting that fusions in this locus could also be a consequence of recurrent somatic alterations. However, it remains to be determined if MET fusions with intact kinase domains confer a phenotype.

We find 472 fusion transcripts that involve well known tumor suppressor genes such as *NF1* (n=27), *PTEN* (n=24), *CREBP* (n=24), *RB1* (n=13) and *TP53* (n=12) (Figure 3.3C, Supplementary Table 3.18). We also find 78 fusion events comprising four SWI/SNF pathway tumor suppressor genes: *SMARCA4* (n=37), *ARID1B* (n=21), *ARID1A* (n=10) and *PBRM1* (n=10). For 29% of putative tumor suppressor fusion events, we did not find a corresponding copy number alteration from SNP arrays (Supplementary Table 3.20). Three out of top four in this category include *RUNX1* (n=13), *CREBBP* (n=8) and *MLL3* (n=6), classic translocation partners in hematological malignancies that are primarily detected in solid tumors in this study (23/27). Our findings imply that the current understanding of loss of known tumor suppressor genes in cancers may be underestimated due to the limitations of copy number arrays (Supplementary Table 3.20).

Excluding known oncogenes and tumor suppressors, among other cancer genes, *TM-*

PRSS2 and *ERG* were the most recurrently fused genes, with 191 out of 259 fusions involving the classic *TMPRSS2-ERG* fusion in prostate cancer (Tomlins et al. 2005). Only 7 out of remaining 68 fusions involving *TMPRSS2/ERG* are detected in tissues other than prostate cancer demonstrating a high degree of tissue-specificity of fusions involving these genes. Three out of next top five most recurrently fused genes, *RARA*, *MYH9* and *CDK12* are candidates for loss of function events (Figure 3.3D). *PML-RARA* fusions are characteristic of acute promyelocytic leukemia. Here, we find 44 additional *RARA* fusions with 31 different partner genes in 35 tumors across 8 cancer types. 35 out of 44 of these *RARA* fusions are predicted to retain only the first exon of *RARA* indicating a possible loss of function mechanism. Retinoic acid (RA) bound retinoic acid receptor alpha (*RARA*) has been shown to upregulate *RARB*, inducing growth arrest and apoptosis (Liu et al. 1996; Lin et al. 2000). Prior work from our lab has also characterized an antagonistic interplay between anti-proliferative effects of retinoic acid and proliferative effects of estrogen receptor signaling pathways in maintaining a balanced gene expression in breast cancer cells (Hua et al. 2009). This indicates that loss of *RARA* may result in dsyregulation of estrogen receptor signaling. Myosin heavy chain 9 (*MYH9*) has been recently reported as a candidate tumor suppressor in squamous cell carcinoma (Schramek et al. 2014). Here, *MYH9* is found to be fused to 14 different partner genes in 43 tumors across 14 different cancer types. In 19 of these cases, the *MYH9* fusion retains only its 5'-UTR, further suggesting a loss-of-function mechanism. Lastly, cyclin-dependent kinase 12 (*CDK12*), with role in maintaining genomic stability by regulating DNA damage response genes (Blazek et al. 2011), is found recurrently fused in 39 tumors with 26 different partners indicating a loss of function event.

One of the challenges in determining the loss-of or gain-of function potential for the fusions is the context dependent role in which various fusions may confer phenotype in different cancers, as both tumor suppressors and oncogenes. For example, PTPRK has been previously reported as a candidate tumor suppressor in various cancers (Xu et al. 2015; Sun et al. 2013; Flavell et al. 2008). However, a recent study also reported potential Wnt

signaling activation by PTPRK-RSPO3 fusions in colon cancer cells (Seshagiri et al. 2012). This highlights the significance of evaluating each fusion independently to determine its gain-of or loss-of function activity.

3.3.6 Kinase fusions

24.2% (499/2067) of all kinase fusions we identified are predicted to generate a transcript with an ORF that retains an intact kinase domain. As such, 4.9% of all tumors in TCGA are predicted to express a fusion with an intact kinase domain (Supplementary Table 3.21). Constitutive activation of the kinase domain through dimerization motifs such as coiled-coil domains are characteristics of hallmark fusions such as *BCR-ABL1* (Shaw et al. 2013b). Here, we find a coiled-coil domain in 206 out of 499 fusions with in-tact kinase domains (Figure 3.3E). Among these, we find seven *FGFR2* events fused to six different partners, including three novel partners *KIAA1598* (n=2), *SMN2* (n=1) and *C10orf118* (n=1). We also find one pancreatic adenocarcinoma sample with *EML4* fused to *NTRK3*, both genes that are partners of other well characterized oncogenic fusions; *EML4-ALK* in non-small cell lung carcinoma (Soda et al. 2007) and *ETV6-NTKR3* in secretory breast cancer (Tognon et al. 2002). *EML4-NTRK3* has also been previously reported in a glioma cell line that when targeted with crizotinib reduced cell viability (Klijn et al. 2015).

Among the kinase domain fusions, we find a novel *KLK2-FGFR2* fusion in three primary tumors, as well as in one corresponding metastatic tumor. We also find higher recurrence for previously reported singleton fusions, *WHSC1L1-FGFR1* (n=3) and *GRB7-ERBB2* (n=7), indicating that they may be biologically significant. Although both *WHSC1L1* and *ERBB2* are regions of focal amplifications, all 10 fusions involving these genes fuse the 5'-UTR of the upstream gene partner to the full length downstream kinase-involved gene, suggesting canonical overexpression of the kinases. Inhibition of over-expressed *MERTK*, a MER family receptor tyrosine kinase consisting of a transmembrane, helical and a kinase domain in non-small cell lung carcinoma, has been previously shown to increase apoptosis in cell lines

(Linger et al. 2013). Here we find signal peptide containing region of transmembrane protein 87B (*TMEM87B*) fused to *MERTK* in 9 tumors. Of the 7 tumors with in-frame *TMEM87B-MERTK* fusion transcripts, only four are predicted to have the full length kinase domain with the rest retaining a partial kinase domain. The fusion of the signal peptide to the kinase domain within these fusions indicates a potential biological mechanism involving mis-localization of this kinase.

3.3.7 Landscape of recurrent fusion genes across human cancers

We identified 2,070 fusions recurrently identified in at least two samples across the cohort (Figure 3.17). 805 of these have been previously reported within a subset of overlapping samples analyzed by Yoshihara et al (Yoshihara et al. 2014). We find a higher recurrence level for at least 41.7% of these fusions within the same samples, demonstrating higher sensitivity of our approach (Figure 3.19). Overall 31% of all recurrent fusion genes and 59.6% of those that are highly recurrent (>=5 tumors) are also detected in at least one cancer cell line in CCLE.

Across 24 out of 33 cancer types analyzed here, we identified 59 recurrent fusion genes that were previously reported. Interestingly, 17 out of 23 most highly recurrent fusion genes (>=5 tumors) across all cancers, excluding prostate and leukemia, are detected in multiple cancer types (Figure 3.18). For example, we find targetable *FGFR3-TACC3* fusion in twelve cancer types, seven more than previously reported (Wu et al. 2013b; Yoshihara et al. 2014). In addition, we observe *ESR1-CCDC170* in uterine corpus endometrial carcinoma, uterine carcinosarcoma and ovarian cancer, in contrast to where it was initially reported in breast cancer (Veeraraghavan et al. 2014). *PVT1-MYC* fusion, first reported in medulloblastoma as a *MYC* overexpression event (Northcott et al. 2012), is detected in four tumors in breast cancer and in 15 additional tumors across seven other cancers. Both *PVT1* and *MYC* are located in chromosome band 8q24.21, which is the most frequently amplified region across human cancers (Beroukhim et al. 2010). Wnt-signaling activating *PTPRK-RSPO3* fusions

were first reported in colon and rectal adenocarcinomas (Seshagiri et al. 2012). Here, we also find this fusion in gastric (n=3) and esophageal cancers (n=1). Including both novel and known fusion genes, we find 92.5% (322/348) of all highly recurrent (>=5 tumors) fusion genes in multiple cancer types.

A subset of the 2,070 recurrent fusions are likely a consequence of recurrent genomic instability within a given locus. We hypothesized that such fusions will display a random distribution of breaks along the lengths of both the genes involved in the fusion. Using this criteria, we enriched for fusion genes that demonstrated selective retention/loss of domains in at least one of the partner genes across the tumors in which the fusion is detected. We further classified 1,144 such recurrent fusions into three categories based on: (i) whether the fusion event generates a canonical chimeric protein with domains from both genes (chimeric proteins, n=369, Supplementary Table 3.22), (ii) if the fusion event generates out-of-frame transcripts in at least 20% of the tumors (out-of-frame fusions, n=685, Supplementary Table 3.23), or, (iii) if the upstream gene fuses to the 5'-UTR of the downstream gene resulting in potential mis-regulation (dysregulated genes, n=90) (Supplementary Table 3.24).

3.3.8 Novel recurrent chimeric proteins

44 out of 70 highly recurrent (>=5 tumors) fusions proteins have been previously reported within the TCGA or other cohorts (Figure 3.4A, Supplementary Table 3.23). The most highly recurrent novel event fuses the cell surface glycoprotein *CD44* to pyruvate dehydrogenase complex, component X (*PDHX*) in 21 tumors (Figure 3.4A). *PDHX* is an essential E3-binding subunit of the pyruvate dehydrogenase (PDH) complex that plays a key role in the citric acid cycle by catalyzing the decarboxylation of pyruvate to acetyl-CoA and the breakdown of glucose. All 15/17 *CD44-PDHX* fusion transcripts that are predicted to be in-frame retain the E3 binding and catalytic domains of *PDHX* (Figure 3.5A). The most predominant isoform, observed in 11 tumors, comprises the signal peptide and X link domain of *CD44* fused to the 2-oxoacid dehydrogenase acetyltransferase domain of PDHX. This in-

dicates a potential mis-localization of mitochondrian specific *PDHX* gene to extracellular matrix or the golgi apparatus (Binder et al. 2014). Interestingly, another fusion involving *CD44* and a proximal gene, solute carrier family 1 member 2 (*SLC1A2*), has been recently reported to gastric cancer and was shown to augment metabolism in cancer cells (Tao et al. 2011). The overall high recurrence of *CD44-PDHX* fusion along with the potential subcellular mis-localization of the catalytic domain of an essential glyolysis enzyme indicates a potential functional consequence that remains to be elucidated.

Septin family genes have been shown to play a multitude of roles such as cytokinesis, vesicle transport as well as microtubule and actin dynamics (Russell et al. 2005). Among these genes, septin 9 (*SEPT9*) has been previously reported as a fusion partner of *KMT2A* in therapy-related acute myeloid leukemia (Osaka et al. 1999). *SEPT9* overexpression has also been reported in a wide range of cancers including breast and ovarian (Burrows et al. 2003; Montagna et al. 2003; Scott et al. 2005). *SEPT9* overexpression has been shown to promote mesenchymal-like migration of renal cells and correspondingly, its knockdown decreased migration (Dolat et al. 2014; Estey et al. 2010). Interestingly, the N-terminus truncated isoform of *SEPT9* (referred to as *SEPT9_v4*), has been previously shown to enhance cell motility (Chacko et al. 2005). In our study, we find a translocation event in seven breast tumors that fused the two thymidylate synthase domains of thymidylate synthetase gene (*TYMS*) on 18p11.2 to full length septin GTPase domain of *SEPT9* on 17q25.3. Four more tumors in different cancers also are predicted to have this fusion transcript (Figure 3.4A). All 11 fusions generate a *TYMS-SEPT9* transcript that contains the full length *SEPT9_v4* isoform that has been shown to increase migration. Interestingly, we find that all seven breast samples with this fusion have a *TP53* mutation ($p<2.45e-4$, FDR 0.05). However, the biological relevance of this co-occurrence remains to be determined. Overall, the high recurrence level of this fusion and prior association of *SEPT9* with cancer strongly imply a functional role for this fusion.

We also report here two other recurrent fusion genes that involve genes previously im-

plicated in cancer. In nine tumors across six cancer types, we found fusions involving ras homolog family member D (*RHOD*) as downstream partner to *KDM2A* (n=7) and two other genes. In 8/9 events, the upstream gene partner is fused to the second exon of RHOD, resulting in retention of the majority of the GTPase catalytic domain. Overexpression of *RHOD* has been previously shown to be sufficient to induce increased cell proliferation in endothelial cells (Kyrkou et al. 2013). In seven tumors across four cancer types, we find the first two exons of E2F-related transcription factor (*TFDP1*) fused to third (n=6) and fifth (n=1) exons of Cullin 4A (*CUL4A*). Higher expression of *CUL4A* has been shown to be associated with poor prognosis in breast and ovarian cancers (Schindl et al. 2007; Birner et al. 2012). Overexpressed *CUL4A* has been shown to induce proliferation in non-small cell lung cancer cell lines and promote migration and invasion in breast cancer cells (Wang et al. 2014a; Wang et al. 2014b).

3.3.9 Novel recurrent out-of-frame fusions

Out-of-frame transcripts may be enriched for loss-of-function events such as *SLC45A3-ERG* (n=13 tumors) (Perner et al. 2013), or, passenger events such as *RPS6KB1-VMP1* (n=65) (Kalyana-Sundaram et al. 2012) that result from recurrent amplifications across cancers (Figure 3.4B, Supplementary Table 3.23). In addition, many of the fusions may play a context dependent role with respect to tissue type and mutational background. Frequent consequence of out-of-frame fusions is the generation of truncated proteins. C-terminal truncated peptides can be generated by introduction of an early stop codon, or N-terminal truncated peptides can be generated by the activation of a novel translational start-site that is internal to the downstream gene. Truncated proteins generated by nonsense mutations or aberrant splicing have been previously reported to have loss-of (Ostler et al. 2007; Spain et al. 1999) or gain-of-function (Dehm et al. 2011) consequences in different cancers. While the majority of 685 recurrent out-of-frame fusions identified in this study have the potential to generate a C-terminal truncated protein of the upstream gene, it remains to be evaluated on

a case-by-case basis if the N-terminal truncated proteins of the downstream partner genes in the fusions have translational potential. An example of the latter is a fusion between *ESR1* and *CCDC170* that has been shown to generate an N-terminal truncated CCDC170 using a novel translational start-site downstream of the annotated start site. Overexpression of this truncated peptide was shown to induce migration and invasion in breast cancer cells (Veeraraghavan et al. 2014). Hypothesizing the gain- or loss-of-function of a truncated protein or the translational potential of an novel ORF requires experimental evaluations.

In this study, we identify a novel fusion between the bone morphogenetic receptor 1b (*BMPR1B*) and PDZ and LIM domain containing 5 (*PDLIM5*) in 19 breast, 5 prostate and 4 ovarian tumors. In 24/28 cases, the 5'-UTR of *BMPR1B* is fused to the 3'-exons of *PDLIM5* that retain an ORF with all (n=22) or a subset (n=2) of the three LIM domains (Figure 3.5B). The fusion is predicted to result in a truncated PDLIM5 that loses its PDZ domain at the N-terminal resulting in a LIM-domain-only protein. Interestingly, LIM-domain-only proteins have been implicated in the onset or progression of various cancers. For example, LIM-domain-only containing proteins such as LIM-domain-only 4 and LIM-domain-only 2 have been previously associated with tumor progression in breast and prostate cancers, respectively (Lu et al. 2006; Ma et al. 2007). Strikingly, truncated PDLIM5 with only the three LIM domains (ΔPDLIM5) was previously shown to interact with and activate multiple isoforms of protein kinase C, including *PRKCB* (Kuroda et al. 1996; Maturana et al. 2011). Overexpression of *PRKCB* has been previously implicated in proliferation and invasion phenotypes in multiple cancers including breast and prostate (Garg et al. 2014). This presents an intriguing hypothesis that the fusion generates truncated PDLIM5 proteins that may activate the protein kinase C signaling pathway which results in activation of various downstream tumor initiation or progression pathways.

3.3.10 Genes recurrently mis-regulated by fusion events

We identified 90 genes in 490 tumors that are likely recurrently dysregulated by the juxtaposition of the 5' UTR with their upstream partners (Figure 3.4C, Supplementary Table 3.24). The most recurrent among these is the fusion of 19 different 5' partner genes (8 on distinct chromosomes) to the breast cancer anti-estrogen resistance 4 (*BCAR4*) gene located on 16p13.13. BCAR4 fusions are detected in 57 tumors across 14 cancer types including cervical (3.9%) and gastric (3%) cancers (Figure 3.4C, Supplementary Figure 3.20A). First identified in a tamoxifen resistance screen, over-expression of BCAR4 protein in the estrogen dependent ZR-75-1 breast cancer cell line has been shown to induce anchorage independent growth in vitro and tumor formation in xenograft models (Godinho et al. 2011; Meijer et al. 2006). BCAR4 was shown to induce AKT and ERK1/2 phosphorylation mediated by ERBB2/3 (Agthoven et al. 2015; Agthoven et al. 2012). More recently, higher BCAR4 gene expression has been shown to be associated with late stage metastasis and poor survival in breast cancer (Xing et al. 2014). The same study characterized functional role of *BCAR4* as a long non-coding RNA (lncRNA) that interacts with SNIP1 and PNUTS, members of Hedgehog signaling pathway. This interaction has been shown to be essential to promote migration. Both the SNIP1 and PNUTS interaction sequences in the first and last exons, respectively, of *BCAR4* were shown to be essential for this migratory activity (Xing et al. 2014). However, in all *BCAR4* fusion-positive tumors, we find the 5' gene fused to full length open-reading frame (ORF) containing the last exon (42/57) or the 5' UTR in the second to last exon (15/57) of *BCAR4* (Figure 3.4B, Figure 3.5C), resulting in loss of SNIP1 interacting regions but retaining the PNUTS interacting regions in all the isoforms of this fusion transcript.

In 17 tumors, itchy E3 ubiquitin protein ligase (*ITCH*) is found fused to the 5'-UTR of agouti signaling protein (*ASIP*) (Figure 3.20B). In two other tumors, *ASIP* is up-regulated by two distinct genes. Fusions involving *TERT* have been previously reported in two SARC

tumors in TCGA and one lung adenocarcinoma from a different cohort (Stransky et al. 2014). Here, we find seven different 5' partner genes in 15 tumors across eight different cancer types fused to the 5'-UTR of *TERT*. In all cases, we found *TERT* to be significantly overexpressed; indicating that in addition to DNA copy number amplifications, somatic mutations and epigenetic modifications, *TERT* expression driven by fusions is a common event across cancers (Figure 3.20C). In 13 tumors across seven cancers types, we find three 5' partner genes fused to *IFI6* (Interferon, Alpha-Inducible Protein 6) (Figure 3.20D). Anti-apoptotic characteristics of overexpressed *IFI6* have been previously demonstrated in gastric and breast tumors (Cheriyath et al. 2012; Tahara et al. 2005). Interestingly, we find two other *IFI6* proximal genes, *AHDC1* (AT-Hook binding DNA motif Containing 1) and *FGR* also potentially dysregulated by multiple 5' partners in 17 additional tumors (Figure 3.20E-F). We do not find focal rearrangements in any of the cancer types with these fusions suggesting that these are not driven by overall genomic instability in that locus. In 10 tumors, enolase 1 (*ENO1*) was found fused to the 5'-UTR of Atrophin-1-Related Protein (*RERE*). *ENO1* overexpression was shown to increase invasion, while its inhibition abrogated the transforming activity in oral cancer cell lines (Tsai et al. 2010) (Figure 3.20G). In four breast and one lung squamous cell carcinomas, we find follicular dendritic cell secreted protein (*FDCSP*) potentially dysregulated by 5'-UTR swapping with a proximal gene (n=4) and a translocation (n=1) (Figure 3.20H). A role for *FDCSP* in invasion and motility in ovarian cancer cell lines was previously suggested (Wang et al. 2010a). Dysregulation of ras homolog family member-A (*RHOA*) was observed in 9 tumors across four cancer types (Figure 3.20I). Overexpression of *RHOA* has been associated with progression of various solid tumors, including all four cancer types in which *RHOA* 5'-UTR fusions are identified (Parri et al. 2010). In 8 tumors, the 5'-UTR of ubiquitin conjugating enzyme E2D 2 (*UBE2D2*) is juxtaposed by the upstream regions of two proximal genes (Figure 3.20J). *UBE2D2* has been previously shown to interact with the E3-ubiquitin ligase *MDM2* and to maintain low levels of TP53 and MDM2 in normal cells (Saville et al. 2004). This suggests that overexpression of *UBE2D2* may augment TP53

degradation, a hypothesis that remains to be evaluated. We find two liver and one gastric tumors with two different upstream partners *SFPQ* (n=2) and UNC5CL (n=1) fused to the first coding exon of micropthalmia transcription factor (MiT) family gene, *TFEB*. MiT genes are frequent translocation partners in renal cell carcinomas (RCC), with particularly poor prognosis associated with *SFPQ-TFE3* fusion (Kauffman et al. 2014). *MALAT1* has been previously reported as the only recurrent upstream partner of *TFEB*. However, no MiT family fusions have been previously reported in non-kidney tumors. Taken together, these observations indicate that gene mis-regulation by fusion events is a common theme in cancer.

3.3.11 Functional validation of novel recurrent fusion genes

Recurrence level of the fusion gene implies potential biological significance but experimental evaluations are essential to determine the functional consequences of the fusions. We selected the most highly recurrent novel fusion event from each of the three categories of fusions (Figure 3.4): *CD44-PDHX*, *BMPR1B-PDLIM5* and *BCAR4*, and sought to experimentally evaluate the effects of their overexpression on proliferation, invasion and migration of benign and cancer cell lines. For each fusion event, we selected the most predominant protein isoform observed in our study (Figure 3.5A-C). For *CD44-PDHX*, the most predominant isoform comprises the signal peptide and X link domain of CD44 fused to the 2-oxoacid dehydrogenase acetyltransferase of PDHX in 11 out of 21 tumors (Figure 3.5A). For *BMPR1B-PDLIM5*, we selected the most common peptide isoform containing the three LIM domains (ΔPDLIM5) (Figure 3.5B). Finally, for *BCAR4*, we selected the 122aa ORF in the last exon of *BCAR4*, that is expressed in all tumors with this fusion. An HA-tag is attached to the C-terminal end of each of the three sequences and the constructs are synthesized on a microfluidics gene synthesis platform (Quan et al. 2011). Each of the constructs are cloned into the plasmid that expresses GFP under the elongation factor-1 (EF1) promoter and the fusion protein under the cytomegalovirus (CMV) promoter. We sought to generate cell lines that stably transfected with the fusion-positive plasmids as well as the

GFP-only control plasmid.

Cellular context may play a role in promoting the phenotype of these fusions. Given that all three fusions were detected in at least one breast tissue (Figure 3.4) and the robust characterization of breast cancer subtypes to-date, we evaluated each of the fusions in MCF10A benign breast epithelial cells, MCF7 estrogen receptor positive breast cancer cells and MDA-MB-231 triple negative breast cancer cells. Immunofluorescence showed that CD44-PDHX protein localized to the cytoplasm and endoplasmic reticulum but BCAR4 and ΔPDLIM5 are seen in both nucleus and cytoplasm (Figures 3.21-3.22). Proliferation assays showed an increase in proliferation for both BCAR4 and ΔPDLIM5 in MCF7 and MDA-MB-231 but CD44-PDHX cells were proliferating at a faster rate only in MCF7 cells (Figure 3.5D-F). Interestingly, BCAR4 over-expression did not have an effect on invasion and migration (Figure 3.5G-L, Figure 3.23). This suggests a potential dual-role for BCAR4 as both a protein inducing proliferation (Meijer et al. 2006; Agthoven et al. 2015; Agthoven et al. 2012) as well as lncRNA in promoting invasion and migration (Xing et al. 2014). We also find the majority of BCAR4 fusions (49/57) in 13 cancer types other than breast, suggesting a potential mode of activation of proliferation pathways that is shared across cancers. Only ΔPDLIM5 cells showed an increase in invasion and migration in the already metastatic MDA-MB-231 cells (Figure 3.5G-L). A wound healing assay showed that benign MCF10A cells exhibited augmented healing when compared to GFP-only control suggesting that overexpressed ΔPDLIM5 may alone be sufficient to induce metastatic phenotypes in certain genetic backgrounds (Figure 3.23).

3.4 Discussion

In this study, we developed a highly sensitive and specific approach to identify fusions from tumor transcriptomes and performed a comprehensive survey of gene fusions across 33 human cancers from TCGA. We identified a striking variability in the number of fusions per tumor, ranging from 0 in 44% of all tumors to 20 or more in 0.2% of tumors, and we found this

variation strongly correlated with the overall genomic instability (Figure 3.2E) suggesting that a substantial proportion of the fusions, as with somatic point mutations (Greaves et al. 2012; Pon et al. 2015), may be passenger events segregating by drift or neutral evolution. However, we also find an enrichment for cancer associated genes among the fusions identified suggesting that the genomic instability creates a pool of mutational diversity for the already neoplastic cells to selectively acquire phenotypically advantageous fusions.

A primary challenge in the analysis of recurrent fusions is distinguishing between gain-of-function events and those that are loss-of-function or passenger events. Here, we hypothesized that recurrent fusions in the latter category will show a random distribution of breaks along the lengths of both the partner genes and, using this criteria, we excluded them from further analysis. We note that due to our less stringent classification, an additional proportion of the remaining 1,144 recurrent fusions (Figure 3.4, Supplementary Table 3.22-3.24) may still be non-driver events generated by focal sCNAs. However, such events cannot be excluded from this criteria alone given that some of the oncogenic fusion partners such as *FGFR3* and *CCDC6* that are within regions of focal amplifications and deletions, respectively. Alternately, our classification criteria may also exclude non-random fusions that are low recurrent and the distribution of breaks along the lengths of the genes cannot be distinguished from random distribution of breaks.

Chimeric proteins are an important class of fusions due to the potential for highly specific drug targeting. 89% (329/369) of the recurrent fusions discovered here have not been previously reported in non-TCGA cohorts. However, 98% (322/329) of them are found in 9 or fewer tumors across the cohort demonstrating the highly heterogeneous combinations in which recurrent, and potentially functional fusions are generated. Instead, we find a striking number of low frequency fusions involving a biologically interesting gene fused to novel partners. For example, we find 24.6% (91/369) of the recurrent chimeric proteins and 1,126 in-frame singletons, collectively across 14.8% of all tumors, that involve a gene that is reported in the COSMIC database of fusions. These findings demonstrate the combinatorial

nature of biologically relevant fusions and highlight the significance of low frequency but functionally relevant fusions.

Recurrent out-of-frame fusions are likely loss-of-function events but can also generate C- and N-terminal truncated proteins from 5' and 3' partner genes, respectively. Some of these proteins may have functions as well. Similar to the N-terminal truncated CCDC170 protein generated by the recently reported ESR1-CCDC170 fusion in breast cancer, we find a novel *BMPR1B-PDLIM5* fusion that has the potential to generate a C-terminal LIM-domain-only PDLIM5 peptides. Interestingly, we find this fusion only in breast (1.7%), ovarian (1%) and prostate (1%), all hormone driven cancers. We show that ΔPDLIM5 augmented the proliferation of MCF7 and MDA-MB-231 cells as well as metastatic phenotypes of the latter. LIM domains of PDLIM5 have been previously shown to interact with and activate PRKCB (Kuroda et al. 1996; Maturana et al. 2011). Activated PRKCB has been shown to increase proliferation of MCF7 cells (Li et al. 2006), phosphorylate RB1 in endothelial cells (Suzuma et al. 2002) and promote invasion in Ras/Mek/PKC1/Rac1-dependent pathway (Zhang et al. 2004). Our results clearly demonstrate the migratory effect of over-expressed ΔPDLIM5 in MCF7 and MDA-MB-231 cells. However, further analysis is required to determine if this is mediated through PRKCB activity.

Deregulation of gene expression by UTR-5 swapping is another consequence of fusions that has been primarily observed in hematological malignancies but also has been described in some epithelial cancers (Mertens et al. 2015). Here, we report 90 genes with their UTR-5s fused to the 5' partner gene in at least two tumors across the cohort, collectively in 5.1% of all tumors analyzed. The most frequent gene in this category is the *BCAR4* (n=57 tumors), first reported as a protein coding gene with its ORF sufficient to increase proliferation in breast cancer cells (Meijer et al. 2006; Agthoven et al. 2012). However, a recent study characterized BCAR4 as lncRNA with potential role in promoting breast cancer metastasis through the activation of the non-canonical hedgehog/GLI2 signaling pathway mediated by BCAR4's interaction with both SNIP1 and PNUTS. All 57 *BCAR4* fusions identified in this

study result in loss of the SNIP1 binding region, shown to be essential for inducing invasion, suggesting that the ORF containing region of BCAR4 may by the essential portion of this gene within these tumors (3.5C). Overexpression of BCAR4 protein increased proliferation of both MCF7 and MDA-MB-231 cells but did not have an effect on invasion and migration, suggesting a potential bifunctional role for *BCAR4* as lncRNA as well as a protein in conferring distinct phenotypes (3.5). Reports of bifunctional mRNAs have been rare (Rinn et al. 2012), likely limited by the ability to identify dual-role mRNAs. Recent ribosome profiling studies provide robust evidence for small peptides translated from significant numbers of non-coding RNA, suggesting that this phenomenon is not uncommon (Andrews et al. 2014; Calviello et al. 2015). Although further experiments simultaneously evaluating *BCAR4*'s transcriptional and translational consequences within the same experimental system are required to establish it as first bifunctional mRNA in cancer, our in vitro analysis of BCAR4 protein, in conjunction with prior work (Meijer et al. 2006; Agthoven et al. 2015; Agthoven et al. 2012), demonstrates its effect on promoting tumor proliferation. In contrast to prior studies evaluating BCAR4 in breast cancer, we find its fusions at high frequency in 13 other cancers including cervical (4%), gastric (4%), esophageal (2.8%) and uterine (2.3%), suggesting a potential shared mechanism of function.

The genetic context and cellular environment of the cell lines used for experimental validations play a crucial role in determining the functional effects of a fusion gene. More than 85% of the tumors in which *CD44-PDHX* and *BCAR4* fusions are identified are from tissues other than breast. In addition, the stage of tumor progression and acquired supporting mutations also influence the phenotypes of the fusion. These factors could explain the lack of phenotype observed in the MCF10A benign breast epithelial cells. Another limitation of our study is our dependence on current gene models to identify fusions at annotated exon boundaries and potentially missing very rare but biologically significant types of fusions such as those generated using novel splice sites (Cancer Genome Atlas Research 2014).

Multi-partner gene fusions such as those involving key genes such as *MLL* in leukemias

(Meyer et al. 2013) and *RET/ROS/ALK* in solid tumors (Takeuchi et al. 2012) have been well described. In our study, we uncovered substantial diversity with which genes are fused to multiple partners. We find the most recurrently fused individual genes (>=20 fusions involving the gene) are found have a median of 14 distinct fusion partners. We do not find a statistically significant difference in number of distinct partners for the most recurrently fused oncogenes and tumor suppressors (median: 10 vs. 12, p¡0.89) suggesting that this criteria may not necessarily indicate a loss of function characteristic. These findings demonstrate the heterogeneity of the fusions and highlights the importance of evaluating recurrence by gene, as well.

Collectively, our study substantially improves our understanding of fusions across cancers and underscores the significance of functionally evaluating the highly recurrent fusions towards developing potential targeted therapies.

3.5 Methods

3.5.1 Transcriptome sequencing data

RNA-seq data was downloaded from TCGA, GTEx and NCBI SRA. Raw unaligned sequence files for 11,506 tumor and normal tissue transcriptomes in the TCGA were downloaded from cgHub (cghub.ucsc.edu) (Supplementary Table 3.1). Aligned bams for 934 tumor cell lines from the Cancer Cell Line Encyclopedia (CCLE) were also downloaded from cgHub (Barretina et al. 2012). .bam files were converted to fastq for downstream analysis using custom scripts. Genotype-Tissue Expression (GTEx) transcriptomes (.sra files) were downloaded from the Database of Genotypes and Phenotypes (dbGaP, project id: phs000424.v3.p1). 9,675 .sra files representing 53 different tissues from 550 individuals were downloaded and converted to fastq using SRA Toolkit v2.3.4 (https://github.com/ncbi/sra-tools/wiki/Downloads). 665 additional normal lymphoblastoid cell line transcriptomes from 464 healthy HapMap individuals were downloaded from EMBL-EBI's ArrayExpress portal

(project id: ERP001942).

3.5.2 Gene expression quantification

Transcript quantification for 10,189 transcriptomes in the TCGA was performed using Kallisto (http://arxiv.org/abs/1505.02710). Transcripts per million (TPM) values were extracted from each sample and a matrix of gene x sample was constructed. Across sample normalization was performed using Trimmed Mean of M-values (TMM) method implemented in edgeR (Dillies et al. 2013).

3.5.3 Gene fusion discovery

Fusion discovery on GTEx, TCGA and CCLE transcriptomes was performed using MOJO-v.0.5.4 in the highest sensitivity mode requiring only two discordant reads and one anchor read to nominate a fusion. Read-through transcripts between genes <250kb are filtered out.

3.5.3.1 TCGA pan-cancer fusion calling workflow

A detailed schematic of the fusion discovery workflow is shown in Figure 3.7. All TCGA tumors (n=9,704) were first analyzed using MOJO (Minimum Overlap Junction Optimizer) in the highest sensitivity mode requiring only two discordant reads and one anchor read (minimum anchor length of 10bps). 213,767 fusion transcripts comprising 79,756 distinct fusion genes were nominated. We next applied four filters to account for technical and biological artifacts.

We next applied two filters to account for alignment and annotation based artifacts. For both filters, to increase sensitivity and specificity, we pooled read evidence for fusion junctions that are recurrently detected across multiple samples. To ensure specificity, only the anchor reads for which the non-junction mapping end maps uniquely to one of the two

genes in the fusion pair are pooled. A fusion gene is required to have at least one fusion junction that passes both the filters. The first filter accounts for non-uniform coverage of anchor reads at the fusion junction – a characteristic of alignment artifact. Each fusion junction is required to be supported by at least 3 anchor reads (minimum anchor length of 15bps) with each of those anchor reads' starting and ending alignment positions spaced by at least 3bp (Figure 3.7). For example, for a read length of 50bp, each junction-read (junction mapping end of the anchor read) is required to have at least 15bp mapping to 5' exon (left_overhang) and 35bp to the 3' exon (right_overhang) of the fusion. And, each of the, at least 3, junction-reads are required to have a minimum overhang of at least 15, 18 and 21 bps on both sides of the fusion junction.

The second filter accounts for fusion junctions comprising exonic sequences that cannot be uniquely mapped. That is, if sequences from one of the exons involved in the fusion junction (last 40bps of 5' exon and fist 40bp of 3' exon) maps to multiple regions in the genome, then we require at least two anchor reads with the non-junction mapping ends mapping uniquely to both the genes of the fusion pair to retain this junction. Alternately, if the full 80bp junction sequence aligns to the genome as three or fewer contiguous blocks, then the junction is filtered out. For both steps of this filter (individual 40bp exonic sequences and full 80bp junction sequence), alignments to the genome and transcriptome are performed using Blat.

We next constructed a filter to account for chimeric artifacts from highly expressed genes that are likely to be generated during library preparation. The ratio of anchor reads to discordant reads is used to filter out high expression artifacts. For a given fusion gene f, we determine the median of anchor read to discordant read ratios (mADratiof) of all the fusion calls for f. A fusion gene is filtered out if mADratiof <0.01. If either of the two genes are highly expressed (max(mRpkmA, mRpkmB) >100) across the samples containing the fusion, then a fusion gene is filtered out if mADratiof <0.05. mRpkmA and mRpkmB are median of expression values of genes A and B within the samples containing the fusion.

The fourth filter attempts to filter out non-somatic events such as germline fusions and artifacts of splicing that can manifest as chimeric transcripts in the analysis of tumor transcriptomes. A panel of 10,340 normal transcriptomes from Genotype Tissue Expression project and NCBI's SRA are used as controls to remove these artifacts. MOJO is run in the highest sensitivity mode on all normal transcriptomes. A fusion gene in the TCGA is filtered out if two or more fusion events are detected between the two genes within the control samples. 11,653 fusion calls comprising 319 distinct fusion genes were filtered out by this filter (Figure 3.8).

Finally, an additional 2,148 fusion calls (833 fusion genes) with either of the genes classified as immunoglobulin, HLA-locus, pseudogenes or genes not annotated in RefSeq were filtered out.

3.5.4 Evaluation of sensitivity and specificity

We evaluated the performance of the fusion discovery pipeline in this study (MOJO-PC) and standalone MOJO (MOJO-S) against three other previously published methods deFuse (v0.6.2), MapSplice (v2.1.9) and FusionCatcher (v0.99.3e). We selected 123 primary tumor transcriptomes from LAML (54 tumors), LUAD (3), KIRC (13), GBM (14) and THCA (39) in which 128 fusion transcripts comprising 27 recurrent fusion genes have been previously reported (Supplementary Table 3.6). In addition to the 128 fusion transcripts, a fusion gene that is a previously reported in COSMIC or Mitelman is also classified as a true positive in this comparison (Supplementary Table 3.14).

We attempted to run all methods with parameters configured to nominate fusions supported by at least two discordant reads and two anchor reads with the anchor length (minimum overhang of the junction mapping read) of at least 15bp. MOJO-S and MapSplice are run with default parameters using GAF3.1 annotation. deFuse is run with the parameters: span_count_threshold=2 and split_min_anchor=15. FusionCatcher with -s 2,2,2,2 -r 1,1,1,1 and -a 15,15,15,23

All methods were run on identical hardware comprising 32 cores and 64GB RAM. However, each method was only allowed to use 16 cores primarily due memory constraints associated with further scaling up. Only 55 out of 123 tumors were successfully analyzed by all four methods. deFuse, FusionCatcher and MapSplice were unsuccessful in fully processing 65, 2 and 1 of the 127 transcriptomes samples, respectively. 62 of the 65 deFuse failures are due to one previously reported but unresolved version specific issue. All other failures, including the two failed runs for FusionCatcher, are primarily due to memory limitations. Comparisons using 55 tumors and 120 tumors between the four methods is shown in Figures 3.1 and 3.14, respectively.

Differences in algorithm implementations and gene annotations can account for significant differences in the number of fusion calls nominated. For example, MOJO nominates fusions only if the fusion junction is at annotated exon-exon boundaries (defined here as canonical fusions). In contrast, the other three methods nominate fusions with junctions involving intronic or intra-exonic breaks (non-canonical fusions), and as a result, due to the increased search space, these methods may nominate large numbers of fusions. We therefore, applied four post-processing steps to account for factors that conflate the false positives and as a consequence may negatively affect the performance of the methods (true positive nominations are automatically retained). First, to account for differences in gene models used by the various methods, we collapsed all fusion calls onto GAF3.0 annotation. Any fusion call supporting transcripts that are outside the boundaries of GAF3.0 annotations (except for true positives) are filtered out. Second, we filtered out fusion calls if at least one of the genes involved does not have an Entrez gene id, or if it belongs to pseudogene, ribosomal or immune system (Ig and HLA loci) classes of genes. Third, we enforced a read-through filter to remove all fusion calls between genes that are <250kb on the same strand and are candidates for a read-through event. Fourth, we filtered out all fusion transcript calls with fusion junctions that do not involve annotated exon boundaries. This filters out all fusion transcript nominations with junctions involving intra-exonic/intronic regions. However, we

retain all true positives irrespective of the type of fusion junction. Finally, we ensured that no true positive call has been inadvertently filtered out by using the breakpoint coordinates for the raw fusion calls for each method and rescued a fusion if the two breakpoints are 50kb up/downstream of the transcription start/end sites of the true positive fusion gene pair.

3.5.4.1 Comparison with MOJO-PC, MOJO-S, deFuse, FusionCatcher and MapSplice (55 tumors)

After applying the above filters, 1,368 canonical fusion gene calls have been collectively nominated by at least one of the four methods with MOJO-PC, MOJO-S, deFuse, FusionCatcher and MapSplice nominating 1.8, 2.3, 19.6, 6.2 and 1.7 fusions per sample, respectively (Figure 3.1, Supplementary Table 3.7). A total of 57 true positives have been detected by at least one of the four methods. The only false negative by MOJO, *NSD1-STIM1* (detected by deFuse), is reported in the COSMIC/Mitelman databases. The resulting *NSD1-STIM1* fusion transcript is predicted to retain a partial intron and therefore, is not predicted to be coding. In contrast to canonical fusions, 7,313 non-canonical fusion calls have been nominated by deFuse (81.4 calls/sample), MapSplice (36/sample) and FusionCatcher (16.1/sample) (Supplementary Table 3.8). Only 33 non-canonical fusion calls (0.05%) were nominated by more than one method and only one fusion (0.015%) is nominated by all three, suggesting that this category of fusion calls could be enriched for false positives.

3.5.4.2 Comparison with MOJO-PC, MOJO-S, FusionCatcher and MapSplice (120 tumors)

745 fusions have been nominated among the 120 samples by MOJO-PC (1.8/sample), MOJO-S (2.4/sample), FusionCatcher (5.3/sample) and MapSplice (2.1/sample). 130 true positives were nominated by at least one of the three methods within these samples (Figure 3.9A). Both MOJO-PC and MOJO-S successfully re-called all (100%) of the true positives

while FusionCatcher and MapSplice detected 124 (95.4%) and 115 (88.5%), respectively (Figure 3.9). MOJO-PC also nominated the fewest 'other' fusions with 87 (0.725/sample) while MOJO-S, FusionCatcher and MapSplice nominated 1.35, 4.3 and 1.1 'other' fusions per sample (Figure 3.9B). MOJO-PC also nominated the fewest singleton 'other' fusion calls with 9 nominations.

3.5.5 Evaluation of specificity using RT-PCR validations

The twelve cell lines selected for RT-PCR validations of the fusion comprise four breast (HCC1187, MDA-MB-157, UACC-893 and HCC-202), 3 ovarian (CAOV4, CAOV3, SK-OV3) and one each of endometrium (KLE), skin (SK-MEL3), pancreas (Hs-766T), low grade glioma (T98-G) and multiple myeloma (OPM-2) cell lines. Fusion analysis was performed using the MOJO-PC pipeline. 176 fusion transcripts comprising 130 fusion genes were nominated within these cell lines. We chose candidates for validations based on strict criteria for primer design. The primers are required to have GC content between 45-55%, melting temperature between 55 °C-61 °C, a GC-clamp in the last three base-pairs of each primer and a product size of 100-500bp. Based on this specific criteria, we successfully generated primers using Primer3 for 134 candidates (Supplementary Table 3.11). The selected fusions are supported by varying anchor reads ranging from 1 to 687 reads.

Four cell lines SK-OV3, CAOV3, OPM-2 and T98-G were acquired from the Cellular Screening Center core facility (University of Chicago, IL). The remaining eight lines were ordered from ATCC (Manassas, VA). All cell lines except CAOV4 were cultured in growth media specified by ATCC. CAOV4 cells were cultured in RPMI and normal CO2. RNA from each cell line was extracted using RNeasy Mini Kit from QIAGEN. Reverse transcription was performed to generate the cDNA library for each line using SuperScript III First-Strand Synthesis System from Invitrogen (Carlsbad, CA) and OligodT primers. PCR reaction for each primer pair was performed using Q5 Hot Start High-Fidelity 2X Master Mix from New England BioLabs, Inc (Ipswich, MA). PCR reaction conditions: (1) initial denaturation at

98 °C for 30 seconds, (2) 35 cycles of: denaturation at 98 °C for 10 seconds, annealing at 65 °C for 30 seconds, extension at 72 °C for 25 seconds, and (3) final extension at 72 °C for 2 minutes. PCR products were run on agarose gel and bands were cut and purified using ZymocleanTM Gel DNA Recovery kit from Zymo Research (Irvine, CA). Purified product was Sanger sequenced using the same primers in the PCR reaction. DNA sequences from Sanger sequencing were aligned to the genome using BLAT and fusion candidates for which the sequence supports the expected fusion junction were marked as successful.

3.5.6 Enrichment analysis of cancer associated genes among fusion genes

Enrichment analysis of genes involved in fusions was performed using the gene ontology categories for biological process (n=825), molecular function (n=396) and cellular component (n=233) that are annotated in Molecular Signatures Database (MSigDB) (Subramanian et al. 2005). In addition, we also included four additional gene sets comprising 558 genes in COSMIC cancer gene census (Futreal et al. 2004), 509 kinases (http://www.uniprot.org/docs/pkinfam) (Manning et al. 2002), 71 tumor suppressors and 54 oncogenes (Vogelstein et al. 2013).

We empirically determined the enrichment of each of the 1,459 gene sets by generating millions of simulated sets of 19,961 fusion transcripts from the human transcriptome (GAF 3.0 annotation with 26,627 genes and 73,000 transcripts). For the simulated fusion transcript sets we retained five properties of the observed distribution of the fusion transcripts. First, the distribution of the lengths of the transcripts involved in the fusions are kept identical between observed and simulated datasets. Second, the observed spatial characteristics such as distance between the partner genes and the number of fusions per chromosome are also maintained. Third, recurrent fusions are also constructed based on the observed distribution of recurrence levels. Fourth, a fusion gene is excluded if the partner genes shared 50% or greater transcript homology. Finally, we filtered out pseudogenes, HLA/Ig and non-RefSeq transcripts (see Figure 3.7). Using these criteria, we constructed 100 million simulated sets

of fusion transcript and determined the likelihood of observing genes from a given gene set by chance (Supplementary Table 3.19).

3.5.7 Protein domain analysis of kinase fusion genes

For each of the fusion transcripts involving a kinase gene, we identified open-reading frames using TransDecoder-2.0.1 (https://transdecoder.github.io/). We then used HMMSCAN program in the HMMER package to annotate PFAM-A domains in the identified peptide sequence (Eddy 2009; Finn et al. 2014). All default options were used. We identified peptides containing any of the three annotated kinase domains: PF00069 (Pkinase), PF00433 (Pkinase_C) or PF07714 (Pkinase_Tyr). We next identified coiled-coil domains using the ncoils program using all default options (Lupas et al. 1991).

3.5.8 Somatic copy number alterations supporting fusion transcripts

Level 3 processed segmented copy number data generated using Affymetrix SNP6.0 was downloaded for 8,241 samples from the TCGA data portal. Segment calls supported by fewer than five SNPs or with the segment mean between -0.3 and +0.3 are excluded for consideration. For a gene to be supported by a segment break, the start or end position of the break should be within 100kb upstream of the transcriptional start site and 100kb downstream of the end site. The two partner genes can be supported by independent segments.

3.5.9 Focal and gene level copy number alterations

GISTIC (Genomic Identification of Significant Targets in Cancer) focal amplification and deletion were downloaded from Broad Institute's FireHose (http://gdac.broadinstitute.org). Calls generated in the April 2015 run were downloaded for all 33 cancer types analyzed in this study. GISTIC calls representing gene level copy number alterations were downloaded from the "TCGA Pan-Cancer" dataset in http://genome-cancer.ucsc.edu. A segment mean

value of <-0.3 and >0.3 is designated as deleted or amplified, respectively.

3.5.10 Known recurrent fusion genes

Previously reported recurrent fusions were compiled from COSMIC (http://cancer.sanger.ac.uk/cosmic, rel. v70) and Mitelman (http://cgap.nci.nih.gov/Info/CGAPDownload) databases. A list of 390 fusion genes that wre reported recurrently in 40,838 samples in COSMIC and 64,479 samples in Mitelman database are compiled in Supplementary Table 3.19.

3.5.11 Functional validations

3.5.11.1 Fusion gene synthesis and lentiviral production

Synthetic open reading frame (ORF) constructs corresponding to *CD44-PDHX* fusion gene, truncated *PDLIM5* and *BCAR4* (below) were synthesized by General Biosystems, Inc (Morrisville, NC) using an approach described by Quan et al (Quan et al. 2011). A HA-tag is added to the C-terminal end of each of the open reading frame. Lentiviral packaging was done by DNA/RNA Delivery Core facility at the Skin Disease Research Center (Northwestern University, Chicago IL).
Full sequences below:

CD44-PDHX (671aa + 9aa HAtag):

```
ATGGACAAGTTTTGGTGGCACGCAGCCTGGGGACTCTGCCTCGTGCCGCTGAGCCTGGCGCAGATCGATTTGAATATAACCTGCCGCTTT
GCAGGTGTATTCCACGTGGAGAAAAATGGTCGCTACAGCATCTCTCGGACGGAGGCCGCTGACCTCTGCAAGGCTTTCAATAGCACCTTG
CCCACAATGGCCCAGATGGAGAAAGCTCTGAGCATCGGATTTGAGACCTGCAGGTATGGGTTCATAGAAGGGCACGTGGTGATTCCCCG
GATCCACCCCAACTCCATCTGTGCAGCAAACAACACAGGGGTGTACATCCTCACATCCAACACCTCCCAGTATGACACATATTGCTTCAAT
GCTTCAGCTCCACCTGAAGAAGATTGTACATCAGTCACAGACCTGCCCAATGCCTTTGATGGACCAATTACCATAACTATTGTTAACCGTGA
TGGCACCCGCTATGTCCAGAAAGGAGAATACAGAACGAATCCTGAAGACATCTACCCCAGCAACCCTACTGATGATGACGTGAGCAGCGGC
TCCTCCAGTGAAAGGAGCAGCACTTCAGGAGGTTACATCTTTTACACCTTTTCTACTGTACACCCCATCCCAGACGAAGACAGTCCCTGGAT
CACCGACAGCACAGACAGAATCCCTGCTACCAGTGATCCCATTAAGATACTAATGCCATCACTGTCTCCTACAATGGAAGAAGGAAACATTG
TGAAATGGCTGAAAAAGGAAGGTGAAGCGGTGAGTGCTGGAGATGCATTATGTGAAATTGAGACTGACAAAGCTGTGGTTACCTTAGATG
CAAGTGATGATGGAATCTTGGCCAAAATCGTGGTTGAAGAAGGAAGTAAAAATATACGGCTAGGTTCACTAATTGGTTTGATAGTAGAAGAA
GGAGAAGATTGGAAACATGTTGAAATTCCCAAAGACGTAGGTCCTCCACCACCAGTTTCAAAACCTTCAGAGCCTCGCCCCTCACCAGAAC
CACAGATTTCCATCCCTGTCAAGAAGGAACACATACCCGGGACACTACGGTTCCGTTTAAGTCCAGCTGCCCGCAATATTCTGGAAAAACAC
TCACTGGATGCTAGCCAGGGCACAGCCACTGGCCCTCGGGGGATATTCACTAAAGAGGATGCTCTCAAACTTGTCCAGTTGAAACAAACGG
GCAAGATTACCGAGTCCAGACCAACTCCAGCCCCCACAGCCACTCCCACAGCACCTTCGCCCCTACAGGCCACAGCTGGACCATCTTATCC
CCGGCCTGTGATCCCACCAGTATCAACTCCTGGACAACCCAATGCAGTGGGCACATTCACTGAAATCCCCGCCAGCAATATTCGAAGAGTTA
TTGCCAAGAGATTAACTGAATCTAAAAGTACTGTACCTCATGCATATGCTACTGCTGACTGTGACCTTGGAGCTGTTTTAAAAGTTAGGCAAGA
TCTGGTCAAAGATGACATTAAAGTATCAGTAAATGATTTTATCATCAAGGCAGCAGCTGTTACCCTTAAACAAATGCCAGATGTTAATGTAAGCT
GGGATGGAGAGGGCCCAAAGCAACTGCCATTTATTGACATTTCAGTGGCTGTGGCAACAGATAAAGGCTTACTTACTCCAATCATAAAAGATG
CTGCTGCTAAAGGTATCCAGGAAATTGCTGACTCTGTAAAGGCTCTATCAAAGAAAGCAAGAGATGGAAAATTGTTGCCTGAAGAATACCAAG
GAGGATCTTTTAGTATTTCCAACTTGGGGATGTTTGGCATCGACGAATTTACTGCAGTGATTAACCCTCCTCAGGCCTGCATTTTGGCGGTTG
GGAGGTTCCGACCTGTGCTGAAGCTCACTGAGGATGAAGAGGGAAATGCCAAACTGCAGCAGCGCCAGCTCATAACAGTCACAATGTCAA
GTGACAGTCGAGTGGTTGATGACGAACTGGCAACCAGGTTTCTTAAAAGTTTTAAAGCAAACCTAGAGAATCCTATCCGACTTGCCTACCCAT
ACGATGTTCCAGATTACGCTTAA
```

BCAR4 (121aa + 9aa HAtag):

```
ATGTACCAACCTATCCAAACTTATCCATGGATGAATCTATCCAGAAGACGGGAGTTCCGATGCTTGTCTTGCTCTGAATGTCTGCTTGTCAC
CTGCTTAGGGTTATCGACTGTGATTCTGGGACTCATTGTTGTTCTACAGGACCCCTCTGACTCTGTGGTTTTCTCTACTGGATTAACAATGAT
AGCCATAGGTGCTTTTTTTGTTGTCCTCACTGGAGTGACAGCCCTGTGTACGGTTACAGTCGACGAGAACTTGCAGAAAACCACGAGGCTA
AGACTAGGAGTGATACGAAAAAGCGGAAGTCTCCAAGGAACTACAGAGCCTTCCATGACTCACTCAATAATCGCTAGCACCTCGCTGTACC
CATACGATGTTCCAGATTACGCTTAA
```

Truncated PDLIM5 (272aa + 9aa HAtag):

```
ATGCCCGAGAGCCTGGACAGCCCAACCTCTGGCAGACCAGGGGTTACCAGCCTCACAGCTGCAGCTGCCTTCAAGCCTGTAGGATCCAC
TGGCGTCATCAAGTCACCAAGCTGGCAACGGCCAAACCAAGGAGTACCTTCCACTGGAAGAATCTCAAACAGCGCTACTTACTCAGGATC
AGTGGCACCAGCCAACTCAGCTTTGGGACAAACCCAGCCAAGTGACCAGGACACTTTAGTGCAAAGAGCTGAGCACATTCCAGCAGGGA
AACGAACTCCGATGTGCGCCCATTGTAACCAGGTCATCAGAGGACCATTCTTAGTGGCACTGGGGAAATCTTGGCACCCAGAAGAATTCAA
CTGCGCTCACTGCAAAAATACAATGGCCTACATTGGATTTGTAGAGGAGAAAGGAGCCCTGTATTGTGAGCTGTGCTATGAGAAATTCTTTG
```

CCCCTGAATGTGGTCGATGCCAAAGGAAGATCCTTGGAGAAGTCATCAGTGCGTTGAAACAAACTTGGCATGTTTCCTGTTTTGTGTGTGT
AGCCTGTGGAAAGCCCATTCGGAACAATGTTTTTCACTTGGAGGATGGTGAACCCTACTGTGAGACTGATTATTATGCCCTCTTTGGTACTAT
ATGCCATGGATGTGAATTTCCCATAGAAGCTGGTGACATGTTCCTGGAAGCTCTGGGCTACACCTGGCATGACACTTGCTTTGTATGCTCAG
TGTGTTGTGAAAGTTTGGAAGGTCAGACCTTTTTCTCCAAGAAGGACAAGCCCCTGTGTAAGAAACATGCTCATTCTGTGAATTTTTACCCAT
ACGATGTTCCAGATTACGCTTAA

3.5.11.2 Immunofluorescence microscopy

Stable MCF10A cell lines expressing each of the three synthetic constructs were seeded on cover glass were fixed in Buffer 1 (3% paraformaldehyde in phosphate-buffered saline (PBS)) at room temperature for 10 min, followed by a 2-min permeablization process in Buffer 2 (0.5% Triton X-100 in 20 mm HEPES, pH 7.5, and 50 mm NaCI, 3 mm MgCl2). For italic gamma-Tubulin staining, cells were fixed in methanol for 5 min following incubation with 3% paraformaldehyde. The cells were then placed in blocking Buffer 3 (Buffer 2 with 0.1% Triton X-100 and 2% bovine serum albumin) for 15 min, incubated with the primary antibody in Buffer 3 for 30 min and washed four times with Buffer 2. After incubation with the secondary antibody and subsequent washes, the cells were mounted in VectorShield medium containing 4',6-diamidino-2-phenylindole (Vector Laboratories, Burlingame, CA, USA). The slides were analysed by fluorescence microscopy (Leica DMIRBE inverted microscope and Openlab 3.1.4 software).

3.5.12 Cell migration and invasion assays

Cell migration and invasion assays were performed using the Corning BioCoat Tumor Migration (Cat# 351164) and Invasion (Cat# 354167) systems and according to manufacturer's instructions. Briefly, cells were pre-treated with 10ug/ml of DilC12(3) Fluorescent dye (Corning, Cat# 354218) for 1 hr. Cells were then collected and 1.25 x 10e4 cells were seeded into each well of both BioCoat systems in serum deprived conditions, containing DMEM media supplemented with 0.1% BSA. 200ul of DMEM media supplemented with 10% FBS was used as a chemoattractant and added to each basal chamber. Both the migration and invasion BioCoat systems were incubated for 22 hrs at 37C, 5% CO2 atmosphere

before readings were acquired using a bottom-reading fluorescent plate reader at 549/565 (Ex/Em). All data is normalized to GFP controls for each respective cell line.

3.6 Appendix: Figures

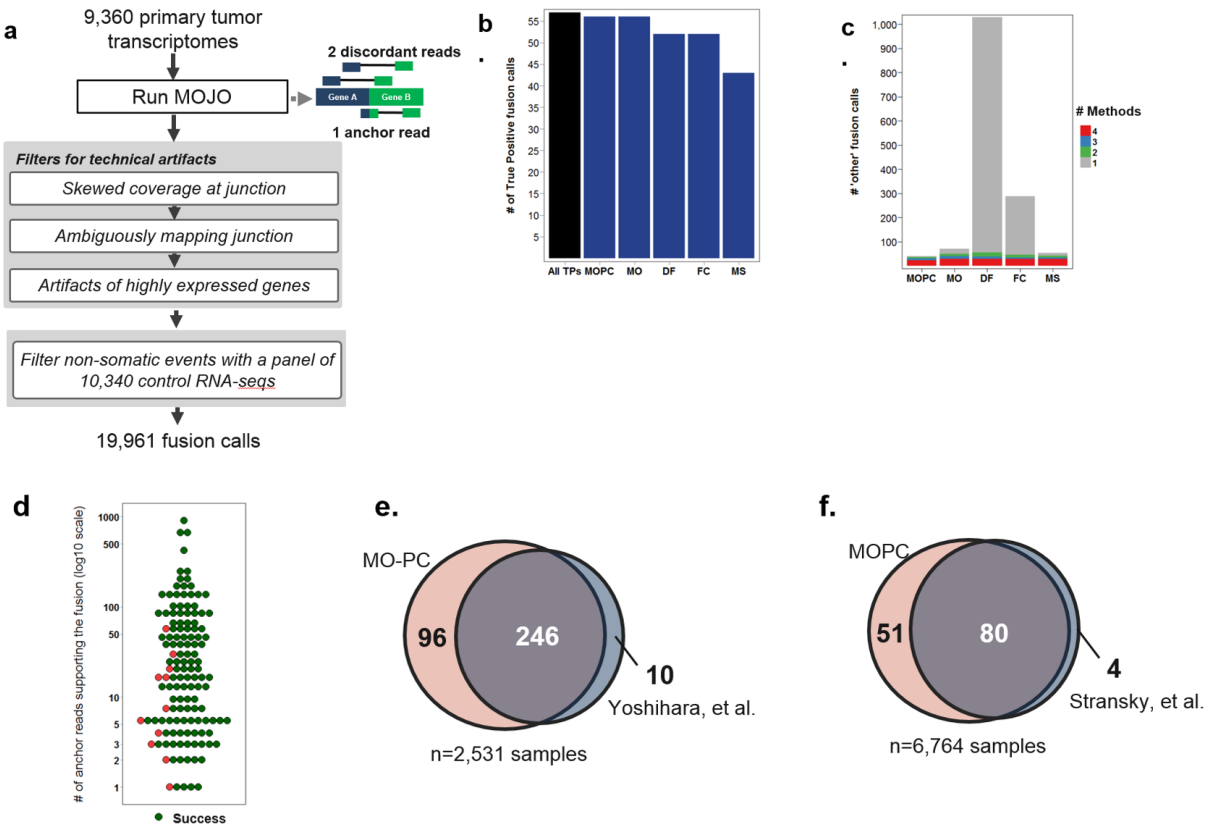


Figure 3.1: Fusion discovery workflow and performance evaluation.

Figure 3.1: (Continued from previous page). Fusion discovery workflow and performance evaluation. (a) Brief schematic of the fusion discovery workflow. MOJO is run on all samples in highest sensitivity mode requiring only two discordant reads and one anchor read supporting each fusion. Evidence for each recurrent fusion is pooled across samples and filters are applied to account for technical artifacts. See Figure 3.7 and Section 3.5.3.1 for detailed steps. (b-c) Sensitivity b and specificity c are evaluated using 55 primary tumor transcriptomes in the TCGA in which recurrent fusions have been previously reported. True positives (TPs) in (b) are fusions that have been previously reported in Mitelman or COSMIC gene fusion databases. MOPC: MOJO PanCancer filtering pipeline, MO: MOJO standalone (without the filtering steps in Figure 3.7), DF – deFuse, FC – FusionCatcher, MS – MapSplice. In (c), "other" fusion calls represents fusions that have not been previously validated. Post-processing filters are applied for each method to exclude fusion call nominations that are not reflective of true differences between the callers (see Methods). No true positive is excluded. Fusions in red are nominated by all four methods (MO, DF, FC and MS) representing a high confidence set of true positive fusions not previously reported. Singleton fusions (in gray) represent a high confidence set of false positives. (d) Distribution of number of anchor reads supporting each of the 134 cell line fusions on which RT-PCR validations were attempted. Validations are designated as successful (green) with evidence from sanger sequencing and as failed (red) if otherwise. (e-f) comparison of true positive fusion calls with two previous studies analyzing fusions within subsets of samples studied here. True positive fusions are primarily compiled from COSMIC and Mitelman databases (Supplementary Table 3.14). (e) comparison of sensitivity between MOPC fusion calls and those reported in Yoshihara et al. Only 2,531 out of 4,366 samples that have at least one fusion reported in Yoshihara et al. are considered for comparisons. (f) comparison of sensitivity to detect true positive kinase fusions between MOPC fusion calls and Stransky et al.

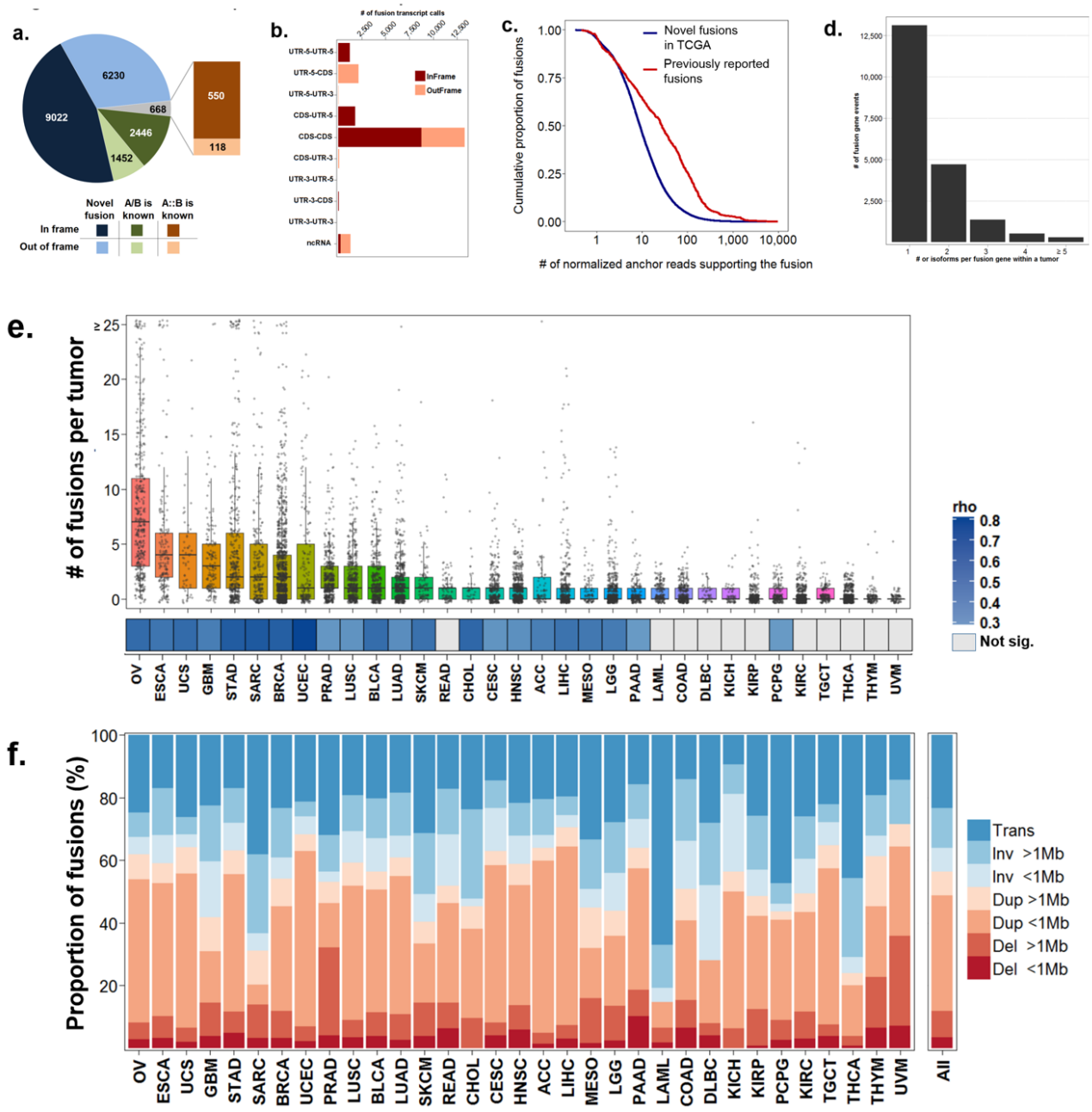


Figure 3.2: Characteristics of somatic fusions across 33 cancers

Figure 3.2: (Continued from previous page). Characteristics of somatic fusions across 33 cancers (a) Fusion transcripts are classified according to their 'known' status. 'A::B is known' indicates the fusion gene is previously reported. 'A/B is known' indicates that one of the two genes is involved in a previously reported fusion. And, 'Novel fusion' if the fusion gene is not previously reported. A fusion transcript is determined as in-frame if it contains an open-reading frame (ORF) with an annotated translational start site and the annotated translational stop site of either the upstream or downstream partner gene. For a given fusion transcript call with multiple isoforms in a given sample, it is classified as in-frame if any one of the isoforms are in-frame. (b) The distribution of the fusion transcripts according to the types of regions in which the breaks occurred in the 5' and 3' genes is shown. UTR-5: 5' untranslated region, UTR-3: 3' untranslated region, CDS – coding sequencing, ncRNA: gene annotated as non-coding RNA. (c) Cumulative distribution of fusions supported by number of normalized anchor reads for fusions reported previously and novel fusions in the TCGA. 24% of fusions in both novel and previously reported categories are supported by 8 reads or fewer. (d) Distribution of number of fusions with multiple isoforms detected in a given tumor sample is shown. On the x-axis is the number of distinct isoforms detected per fusion gene within a tumor sample. (e) Frequency spectrum of fusions per sample across 33 cancer types. On the x-axis are cancer types sorted by the mean number of fusions per tumor detected. On y-axis are the number of fusions within the respective tumor. Note that tumors with more than 25 fusions are binned into one group for visualization purpose. Correlation with number of fusions and the degree of genomic instability (measured as number of genomic segment breaks in copy number array) is shown in the panel below the boxplots (Spearman rank correlation). All significant correlations are shown in shades of blue. P-values are corrected for multiple testing. (f) Spatial characteristics of genes involved in fusions across the multiple cancer types is shown with respect to the type of rearrangement that is predicted to generate them. Trans – translocations, Inv – inversions, Del – deletions, Dup – tandem duplications. Intra-chromosomal rearrangements are further stratified by >1Mb or <1Mb to distinguish between large and small rearrangements.

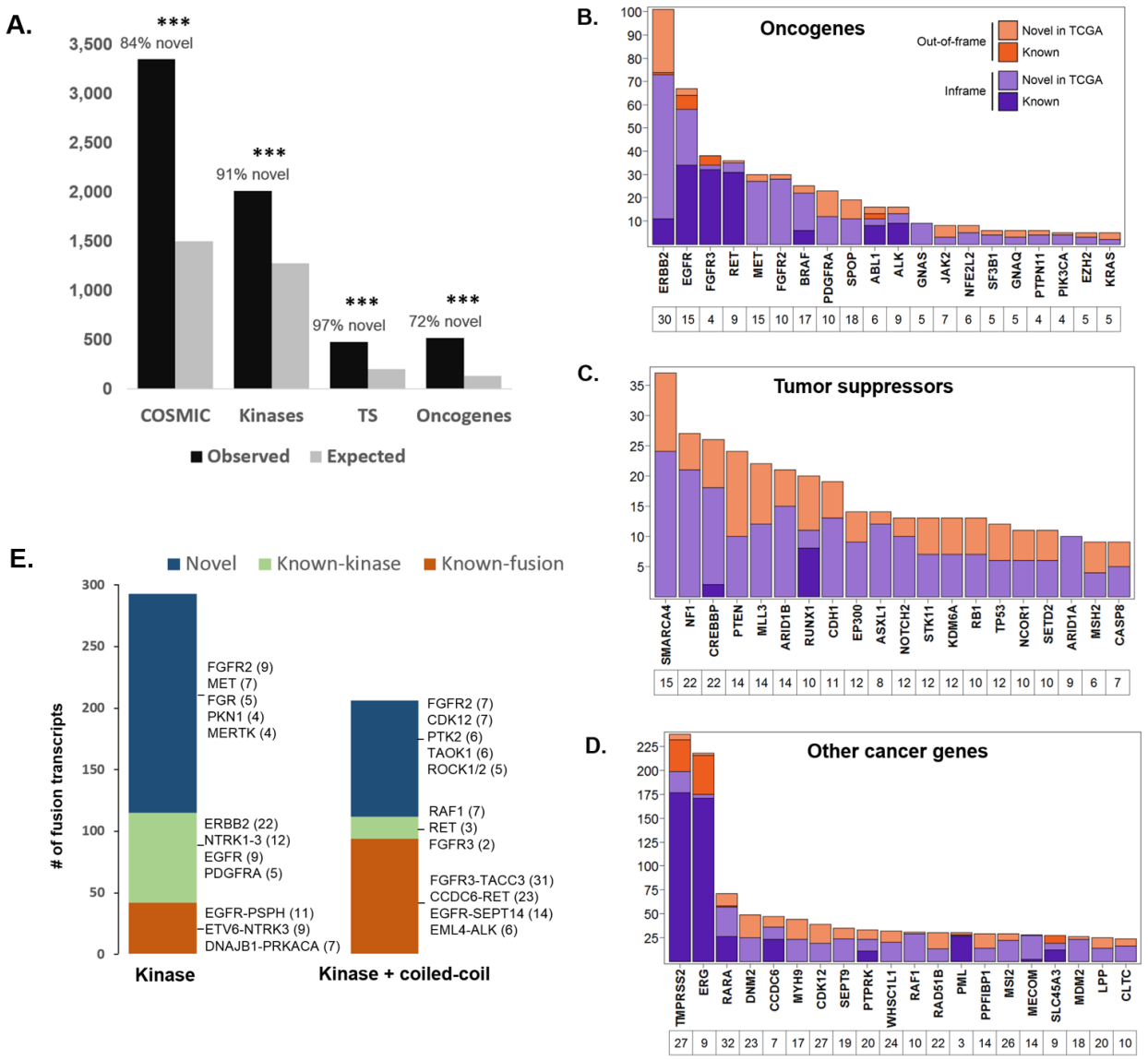


Figure 3.3: Enrichment of cancer associated genes among somatic fusions.

Figure 3.3: (Continued from previous page). Enrichment of cancer associated genes among somatic fusions. (a) Enrichment of four overlapping sets of cancer associated genes is shown. COSMIC: 558 genes, Kinases: 509, Tumor Suppressors, TS: 71, Oncogenes: 54 (Vogelstein et al., Science 2003). 'Expected' number of fusions in each category is determined by sampling the observed number of fusion events after retaining key properties of the distribution of the genes involved in fusions. These include: lengths of the genes, spatial characteristics with respect to the two genes involved and recurrence level of the fusion gene. *** - p-value <0.001, FDR 0.01 (see Methods). (b) Top 20 recurrently fused oncogenes. 90% of the fusions involving oncogenes involve one of the 20 genes shown here. (c) Top 20 tumor suppressors genes involving in fusions. (d) Top 20 genes that are in COSMIC that are not oncogenes or tumor suppressors. For (b-d), fusions are indicated as 'Known' if the fusion gene has been presorted previously. Table below each panel represents the number of distinct partners observed for the given recurrently fused gene. (e) Classification of 499 fusion transcripts with in-tact kinase domains. 206 have both kinase and a coiled-coil domain. Kinase and coiled-coil domains are identified by searching for Pfam domains within the individual fusion transcripts identified in tumors. 'Known-fusion' – fusion gene is previously reported, 'Known-kinase' kinase is involved in a previously reported fusion gene. Number of fusion transcripts corresponding to each event is indicated in parenthesis.

A.

			OV (412)	ESCA (180)	STAD (410)	UCS (57)	SARC (258)	GBM (155)	BRCA (1093)	UCEC (177)	LUSC (501)	BLCA (408)	ACC (79)	LUAD (512)	SKCM (103)	PRAD (497)	LIHC (371)	CESC (304)	READ (94)	CHOL (36)	HNSC (501)	MESO (87)	LGG (514)	PAAD (178)	DLBC (48)	LAML (174)	COAD (289)	PCPG (182)	TGCT (150)	KIRP (292)	KIRC (533)	THCA (505)	THYM (119)	KICH (66)	UVM (80)	Total	% InFrame	Mean NARs	# CellLines	IsKnown
TMPRSS2	ERG	DEL													190											1								191	85	42	2	1
FGFR3	TACC3	DUP		2	2			3	1		5	10		1			1	5			2		2						2						36	89	608	4	1
CCDC6	RET	INV												1												1					21				23	100	24		1
CD44	PDHX	DUP	1	1	2		1		3		3	1						1			7										1					21	81	6	43	
PML	RARA	TRA																							16										16	94	7.1	1	1
EGFR	SEPT14	INV						12													1		3												16	94	260		1
FLNB	SLMAP	DUP	1	1	2		1		3			2					1	2						1					1						15	80	4.6	5	
LYPD6	LYPD6B	DUP	9		2							1								1															13	100	3.8	2	
EGFR	PSPH	INV						12				1																							13	85	382		1
ERBB2	PPP1R1B	DUP	2	1					3		1	1					1	2						1											12	92	250		1
CBFB	MYH11	INV																							11										11	100	9.4	1	1
RAF1	AGGF1	TRA														1															10				11	100	6.3		
RARA	PML	TRA																							11										11	100	10	1	1
TYMS	SEPT9	TRA		1					7		1										1					1									11	100	1.8	4	
PTPRK	RSPO3	INV		1	3														3							4									11	100	44	1	1
ERBB2	STARD3	DUP		3	2				1		1							2			1														10	80	3.5	3	
RGNEF	BTF3	DUP	2	2	2						1							1			2														10	90	5.4	2	
ETV6	NTRK3	TRA							1						1											2					6				10	100	8.7		1

B.

RPS6KB1	VMP1	DUP	5	3	6				7	1	9	4		17				3			5					1			1	3					65	65	15	41	1
BMPR1B	PDLIM5	DUP	4						19							5																			28	58	2.4	5	
ESR1	CCDC170	DUP	1			1			18	4																									24	13	4.4	4	1
GTF2I	GTF2IRD1	DUP	11	1	2				3	1		1		2			1							1												23	53	16		
COL14A1	DEPTOR	DUP	14							4				2																					20	45	32		
ITGB6	RBMS1	DUP	1	1					3		1	7									5			1												19	69	7.7	3	
CCDC6	ANK3	DUP	11						3	3				1																					18	62	30	4	
MLL5	LHFPL3	DUP	3	1	5	1		1	1	2				2	1		1																			18	6	15	2	
HOXB3	HOXB-AS3	INV	2		1				4		1															3			1	2			4		18	0	1.1	5	
TTC6	MIPOL1	DUP							15							3																			18	0	3.5	2	
SH3PXD2A	OBFC1	DUP		3	2				1		1	3		1				1			4															16	57	9.9	1	
PDE4D	DEPDC1B	DUP	7	1					1		3			1		3																				16	44	6.2	2	
ANKHD1	CYSTM1	DUP	2		4	1	1		2	1				1		1										1	1				1				16	44	6.2	4	
NCOR2	SCARB1	DUP	3	3	3		1		1	1			1	2											1											16	50	36	2	
FRMD6	LOC283553	DUP		1							6							2			6															15	0	2.8	2	
LRBA	SH3D19	DUP	2		1				1		2	3				1					1				1	1			1						14	72	4.4	3	
THSD4	LRRC49	DUP	2	1					10									1																		14	79	2.2	3	
PPP1CB	PLB1	DUP	7		1				2	1	1		1																							13	62	11	1	
ASCC1	MICU1	DUP			4	1			5					1				1			1															13	47	3.1	3	

C.

	# 5' Partners	OV (412)	ESCA (180)	STAD (410)	UCS (57)	SARC (258)	GBM (155)	BRCA (1093)	UCEC (177)	LUSC (501)	BLCA (408)	ACC (79)	LUAD (512)	SKCM (103)	PRAD (497)	LIHC (371)	CESC (304)	READ (94)	CHOL (36)	HNSC (501)	MESO (87)	LGG (514)	PAAD (178)	DLBC (48)	LAML (174)	COAD (289)	PCPG (182)	TGCT (150)	KIRP (292)	KIRC (533)	THCA (505)	THYM (119)	KICH (66)	UVM (80)	Total	Mean NARs	# CellLines	IsKnown
BCAR4	19	5	5	13	1			8	4		2		2	1	1		12				1				1		1							57	38	5	
MYC	4		2	2				3	1	3			1		2		5							1											20	41	24	1
ASIP	3	4		1				2			4	3	1			2	1	1																	19	20	3		
TERT	7		1			4	2		1							3												2	1					14	7.5			
IFI6	3	7	1	1				1	1	1			1																					13	6.9	1		
ERBB2	6		1	3				3			2						3																	12	18	5		
WIPF2	8	1	1	6				4																										12	20			
AHDC1	1	2		1				1		5										2														11	9.9			
ENO1	2	3	2				1	2	1				1																					10	6.2	2		
WDR74	1	1			1	1		2		1			1				1												1					9	6.4	1		
PRKCH	1			2				2		1			1							1									2					9	2.2	10		
RHOA	5	3		2				2			2																							9	65	1		
ZNF774	1	1	2		1			1		1																			1		1			8	2.9	1		
ELOVL7	2	1	1	1				1							4																			8	4	3		
UBE2D2	2	4		1					2											1														8	19	3		
ETV4	1														7																			7	130		1	
PTRH2	2	1		1	1			3					1																					7	2.8	6		
CDC42EP4	3	2						1	2	2																								7	7.2	1		

Figure 3.4: Recurrent fusions in the TCGA

Figure 3.4: (Continued from previous page). Recurrent fusions in the TCGA. Top most recurrent fusions that are predicted to generate chimeric proteins (A), out-of-frame events (B) and result in gene dysregulation (C) are shown. Number in each cell represents the number of fusions found in the given tumor type. For (A-B), first two columns represent 5' and 3' gene names, followed by the type of rearrangement predicted to generate the fusion event. DUP: duplication, TRA: translocation, INV: inversion, DEL: deletion. Columns 3-37 correspond to the tumor type (see Supplementary Table X for definitions of acronyms). '% InFrame' shows the proportion of fusion transcripts that are predicted to be protein coding using annotated translational start site. 'Mean NARs' – mean normalized anchor reads is the mean value of anchor reads supporting each of the fusion transcript calls that are normalized to the library size in the respective tumor sample in which it is found. '# CellLines' - # of CCLE cell lines predicted to have the corresponding fusion. 'IsKnown' - indicates if the fusion has been previously reported. (C) Most recurrent fusions that involve an upstream gene fused to the 5'-UTR of the downstream gene are shown. '# 5' partners' - number of distinct upstream partner genes that fuse to the indicated gene.

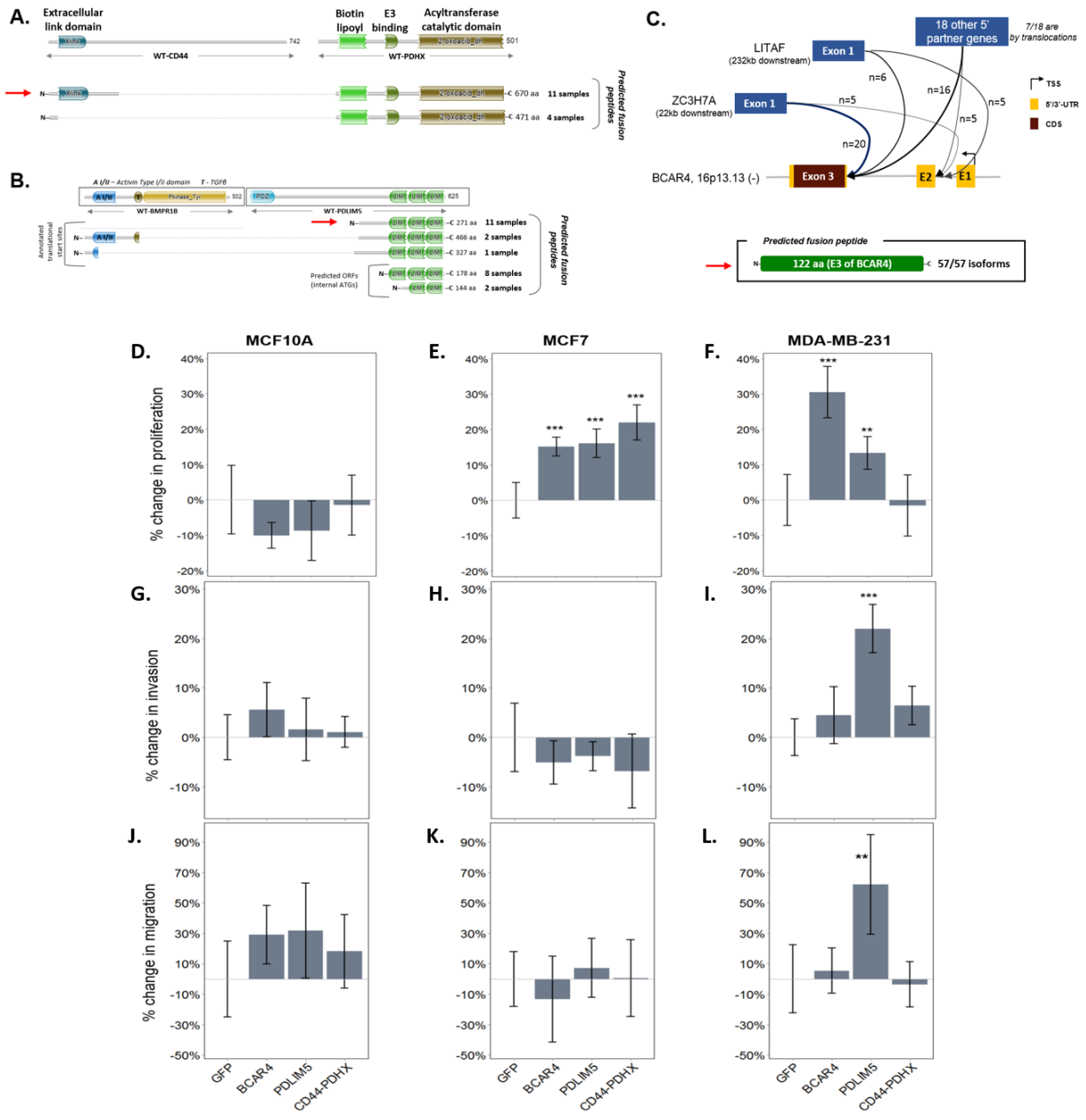


Figure 3.5: Functional evaluation of novel recurrent fusion events

Figure 3.5: (Continued from previous page). Functional evaluation of novel recurrent fusion events (A-C) schematics showing the translational consequences of CD44-PDHX (A), PDLIM5 (B) and BCAR4 (C) fusion events. (A-B) domains in wildtype CD44, PDHX, BMPR1B and PDLIM5 are shown at the top of the schematic. Below are the fusion proteins predicted to be generated. Dotted indicates the portion of the wildtype domains that are lost in the final fusion proteins. Predicted open-reading frames (ORFs) in (B) indicate the potential for generating a protein using an unannotated translational start site. Red arrow highlights the isoform selected for functional evaluations (A-B). Assays for proliferation (D-F), invasion (G-I) and migration (J-L) of the three fusion constructs along with the control-GFP expressed in different genetic backgrounds (MCF10A, MCF7 and MDA-MB-231). (D-F) Cell proliferation assays were performed by seeding six replicates of 5 million cells of each of the cell lines and measuring percent change in cell number w.r.t GFP-control after 96 hours. Transwell assays were used to measure the invasion and migration of the 12 cell lines as described in Methods (G-L). Six replicates were used for each cell line. Results for both invasion and migration are shown relative to the GFP-control. Only PDLIM5 in the MDA-MB-231 cell line showed an increase in invasion and migration. Bar: s.d. (* P <0.05, ** P <0.01, *** P <0.001)

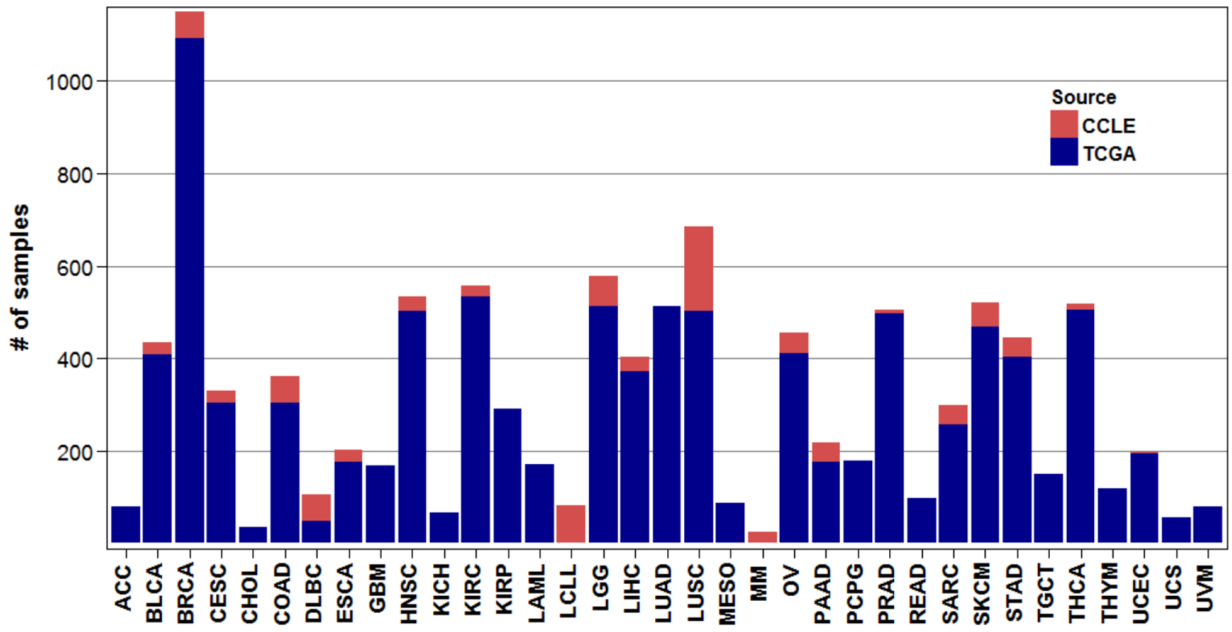


Figure 3.6: Distribution of TCGA and CCLE transcriptomes analyzed in this study. See Table 3.1 for acronym definitions.

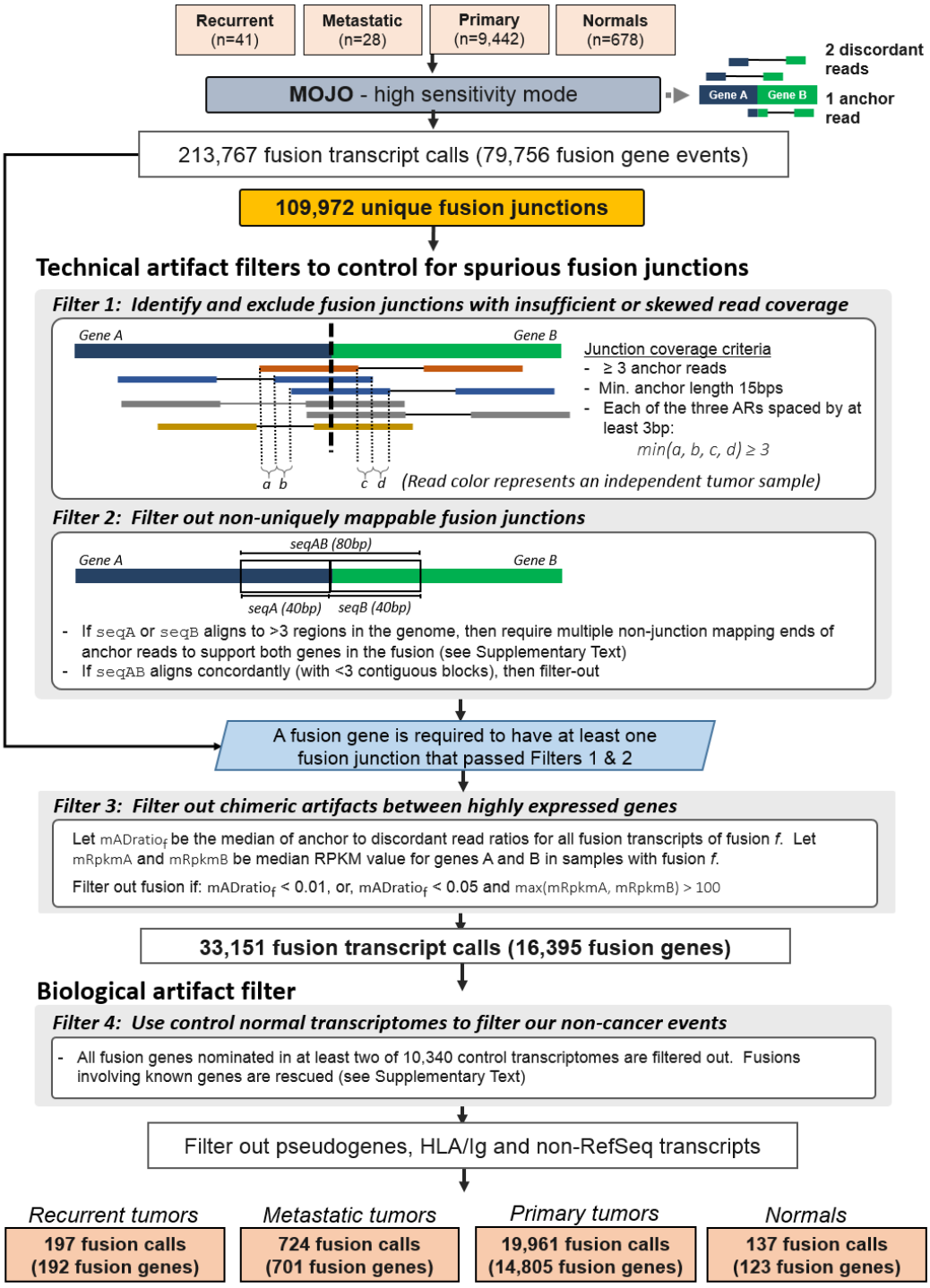


Figure 3.7: TCGA PanCancer gene fusion calling workflow (MOJO-PC pipeline). See Section 3.5.3.1 for further details.

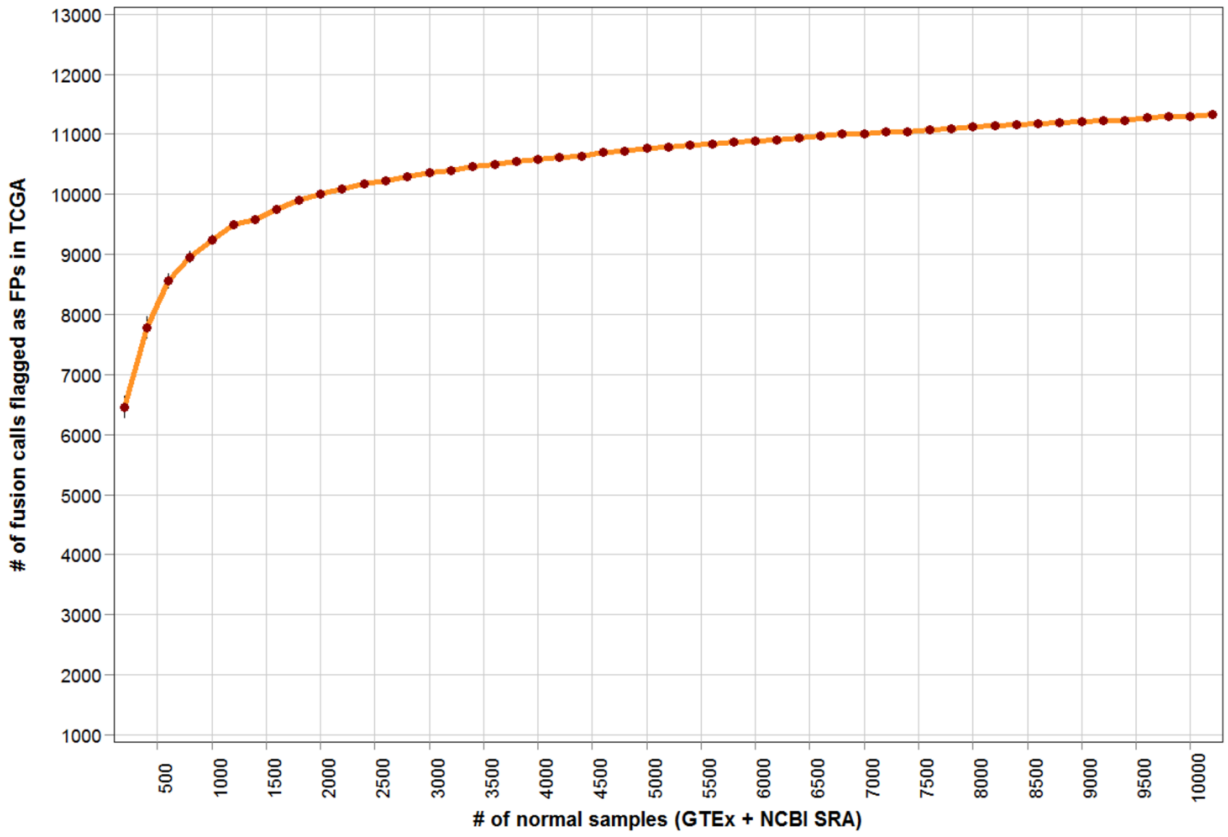


Figure 3.8: Saturation curve showing the effect of filtering out GTEx fusion calls from TCGA calls (Filter # 4 in Figure 3.7). For varying number of control normal samples (x-axis), the number of fusion calls in the TCGA that are filtered out is shown on y-axis. At each step, the control set of transcriptomes was re-sampled 100 times and the mean number of fusion calls that were filtered out is shown.

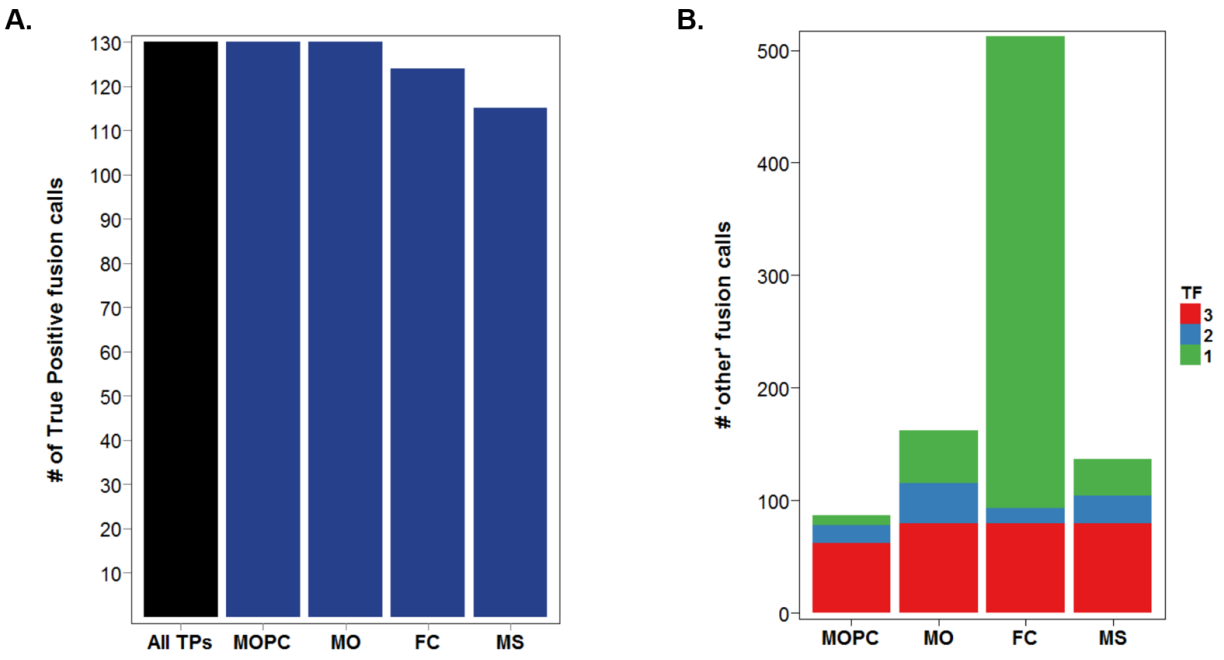


Figure 3.9: Extended comparisons of MOJO, MapSplice and FusionCatcher using 120 tumor transcriptomes. This figure corresponds to Figure 3.1B-C. As described in Methods, deFuse failed to run on 65/120 tumor samples considered for comparisons. Here comparisons of sensitivity (a) and specificity (b) using the entire set of 120 transcriptomes is shown for MOJO, MapSplice and FusionCatcher. True positives (TPs) in (a) are fusions that have been previously reported in Mitelman or COSMIC gene fusion databases (Supplementary Table 3.11). MOPC: MOJO PanCancer filtering pipeline, MO: MOJO standalone, DF – deFuse, FC – FusionCatcher, MS – MapSplice. In (c), "other" fusion calls represents fusions that have not been previously validated. Post-processing filters are applied for each method to exclude fusion call nominations that are not reflective of true differences between the callers (see Methods). No true positive is excluded. Fusions in red are nominated by all four methods (MO, DF, FC and MS) representing a high confidence set of fusions not previously reported. Singleton fusions (in gray) represent a high confidence set of false positives.

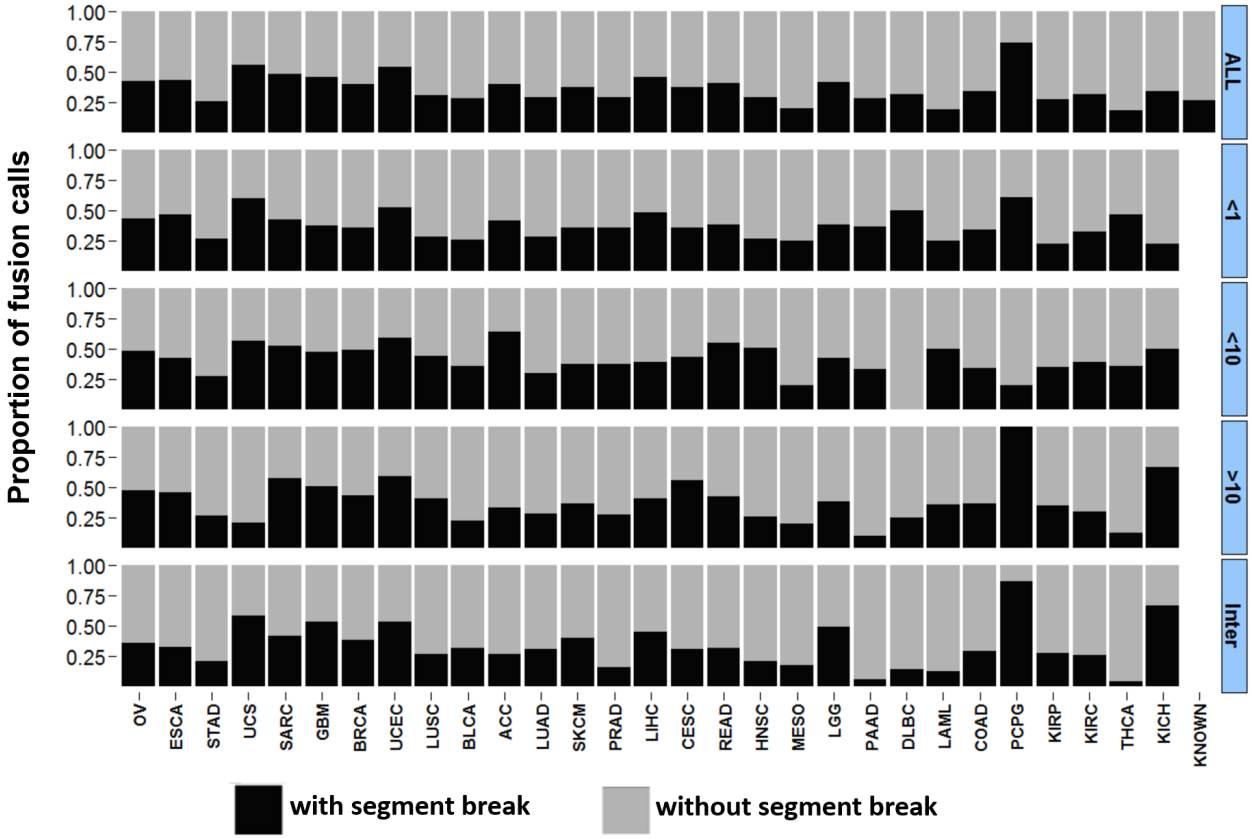


Figure 3.10: Proportion of fusion calls supported by at least one segment break in the somatic copy number profile of "both" the genes involved in the fusion. Breaks 100kb upstream and downstream of the gene transcriptional start and stop sites are considered. Proportions for various classes of fusions "<1 Mb", "<10Mb", etc are shown along with aggregate 'ALL'. "KNOWN" comprises fusions previously reported (Supplementary Table 3.14). Overall, a median of 34% of all fusions are supported by one or more segment breaks within both the partner genes. Tumor acronyms are defined in Supplementary Table 3.1.

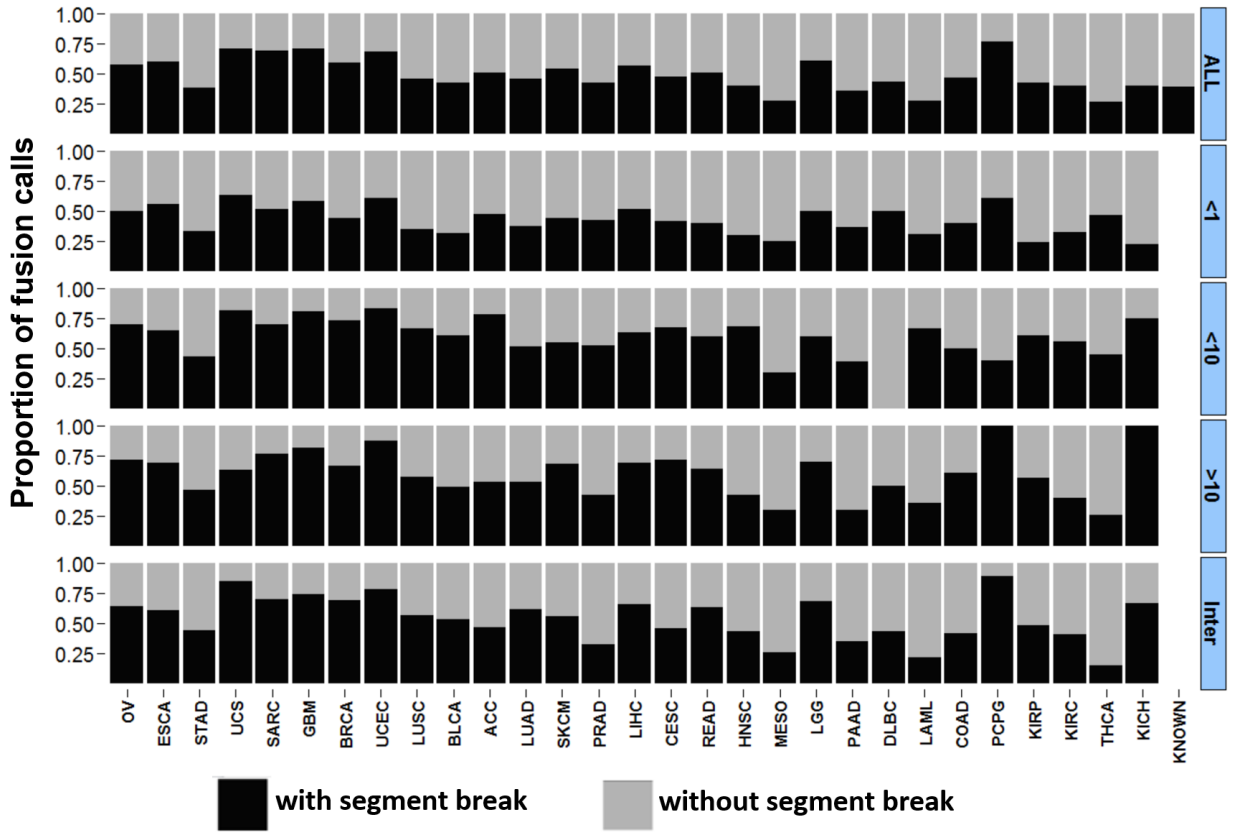


Figure 3.11: Proportion of fusion calls supported by at least one segment break in the somatic copy number profile of "either" the genes involved in the fusion. Breaks 100kb upstream and downstream of the gene transcriptional start and stop sites are considered. Proportions for various classes of fusions "<1 Mb", "<10Mb", etc are shown along with aggregate 'ALL'. "KNOWN" comprises fusions previously reported (Supplementary Table 3.14). Overall, a median of 46% of all fusions are supported by one or more segment breaks within "either" of the partner genes. Tumor acronyms are defined in Supplementary Table 3.1.

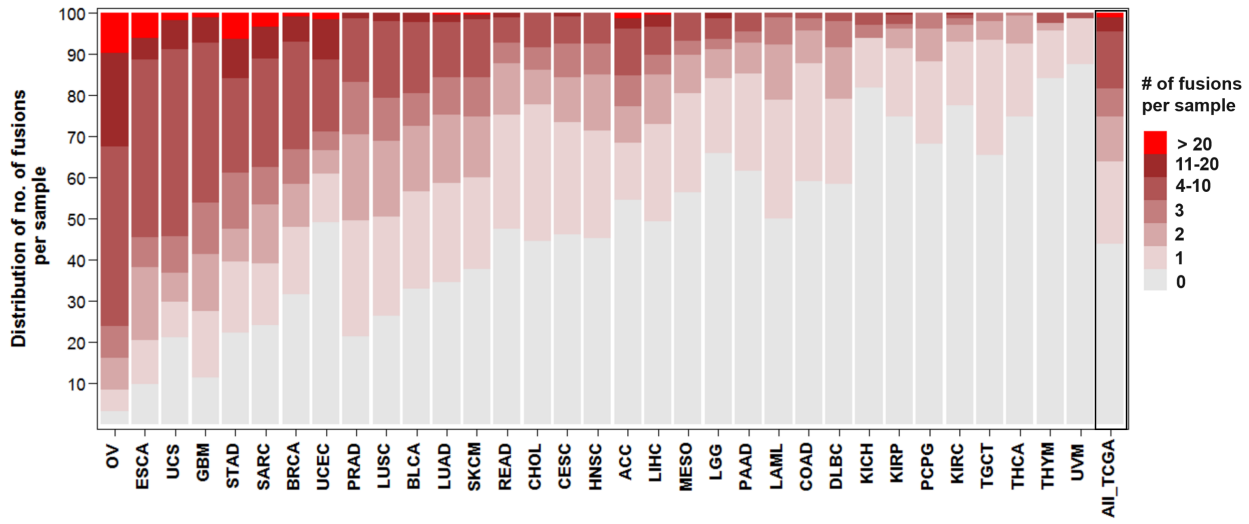


Figure 3.12: Proportion of tumor samples with varying number fusions are shown. On the y-axis, the proportion of samples within the corresponding tumor type that have varying number of fusions (0 fusions, 1, 2, 3, 4-10, 11-20 >20) is shown. 'All_TCGA' shows this distribution across the entire cohort of 9,360 primary tumors. Tumor acronyms are defined in Supplementary Table 3.1

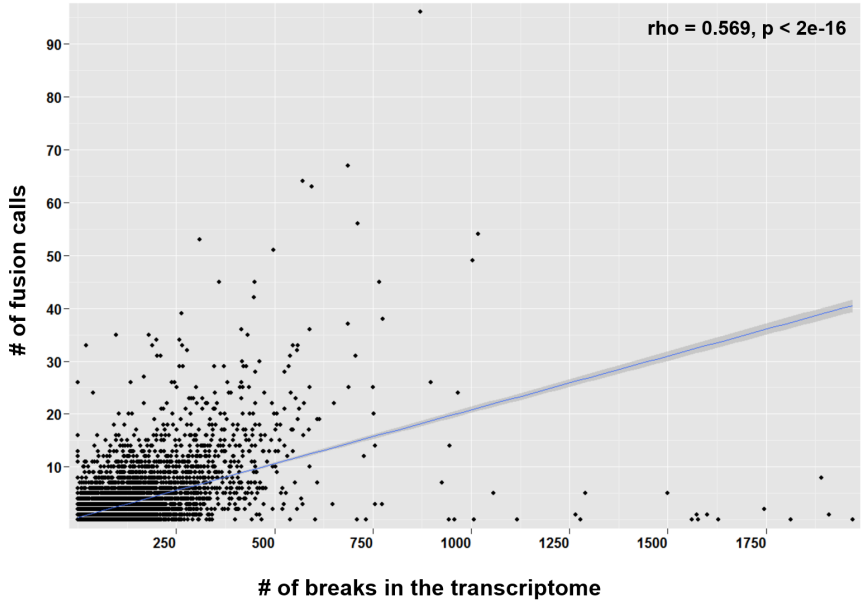


Figure 3.13: Correlation between the number of fusions per sample and the number of segment breaks in transcriptional regions determined from copy number arrays (Level-3 TCGA data). Each dot corresponds to one of the 9,350 tumors (10 tumors with >2000 segment breaks are excluded). Spearman rank correlation is shown at the top of the panel.

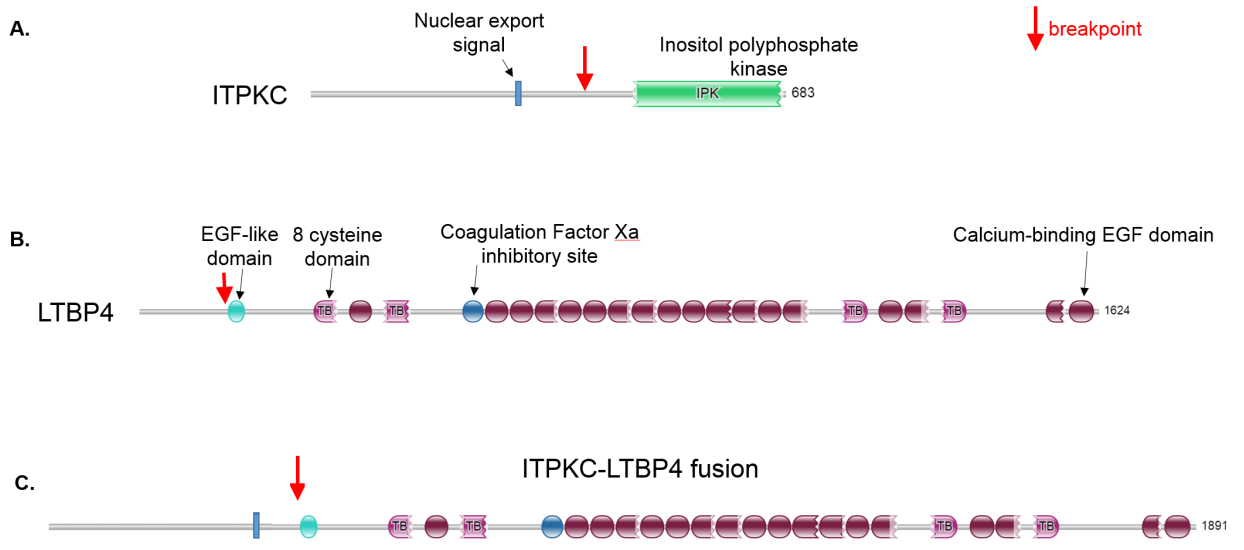


Figure 3.14: Schematic of *ITPKC-LTBP4* fusion detected in primary, metastatic and adjacent normal tissue of one breast cancer patient in TCGA. Protein domain composition of (a) *ITPKC*, (b) *LTBP4* and (c) *ITPKC-LTBP4* fusion is shown. Breakpoint in the two proteins in (a-b) is highlighted by red vertical arrow. Fusion retains the nuclear export signal of *ITPKC* and all the functionally relevant domains of *LTBP4*. Red arrow in (c) indicates fusion junction.

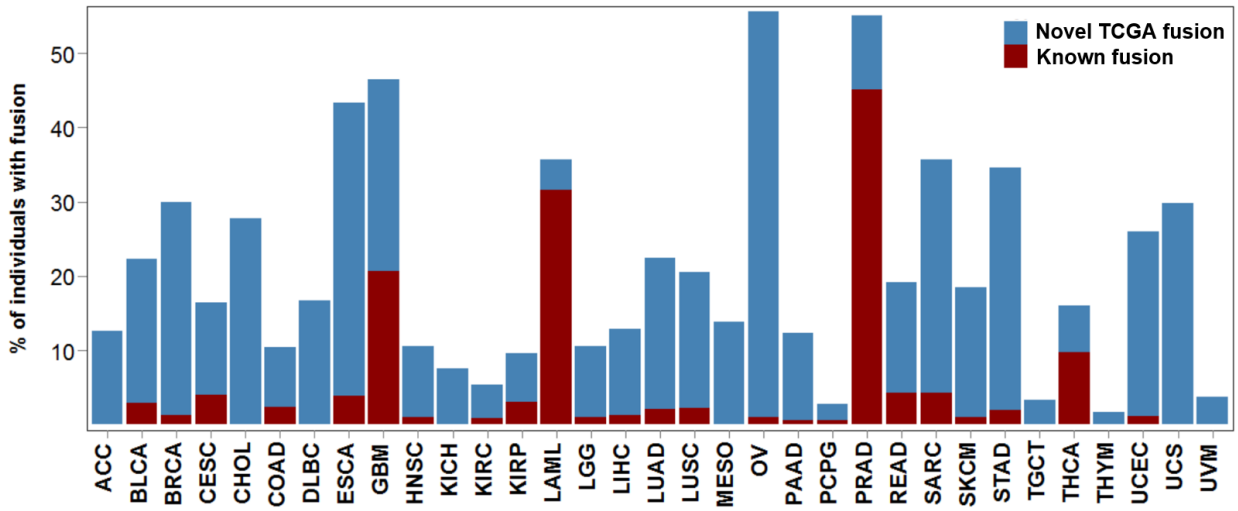


Figure 3.15: Proportion of individuals that have at least one fusion involving a gene that is previously implicated in cancer (COSMIC cancer gene census). 'Known fusions' are those that were previously reported in COSMIC and Mitelman databases (Supplementary Table 3.14). See Supplementary Table 3.20 for complete list of fusions. Tumor acronyms are defined in Supplementary Table 3.1

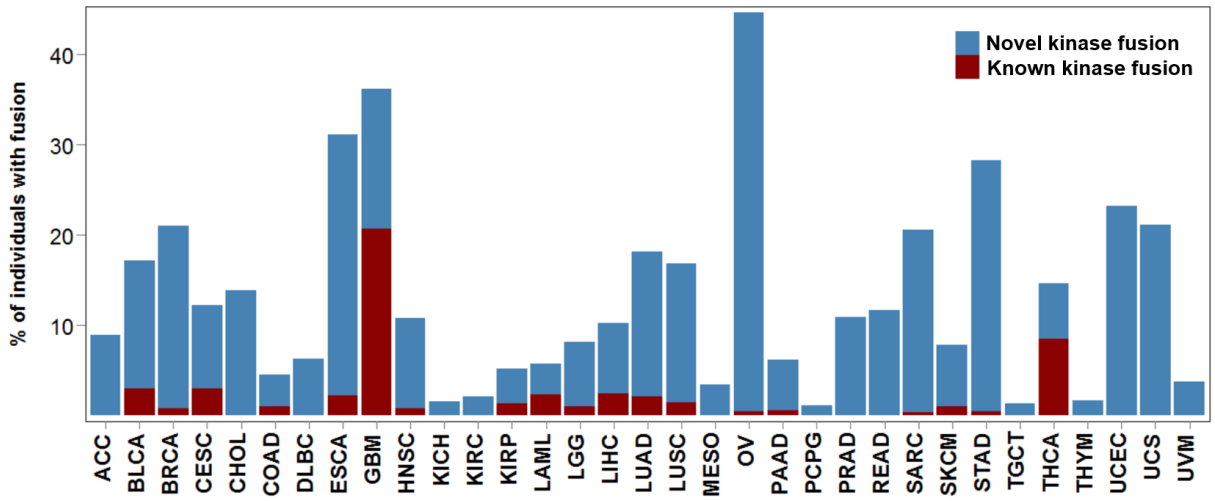


Figure 3.16: Proportion of individuals that have at least one fusion involving a known kinase gene. Only fusions containing any one of the 571 known kinases are shown here. 'Known kinase fusions' are those that involve a kinase gene and the fusion pair is previously reported in COSMIC and Mitelman databases (Supplementary Table 3.14). 'Novel kinase fusion' indicates that the fusion has not been previously reported. See Supplementary Table 3.20 for complete list of fusions. Tumor acronyms are defined in Supplementary Table 3.1

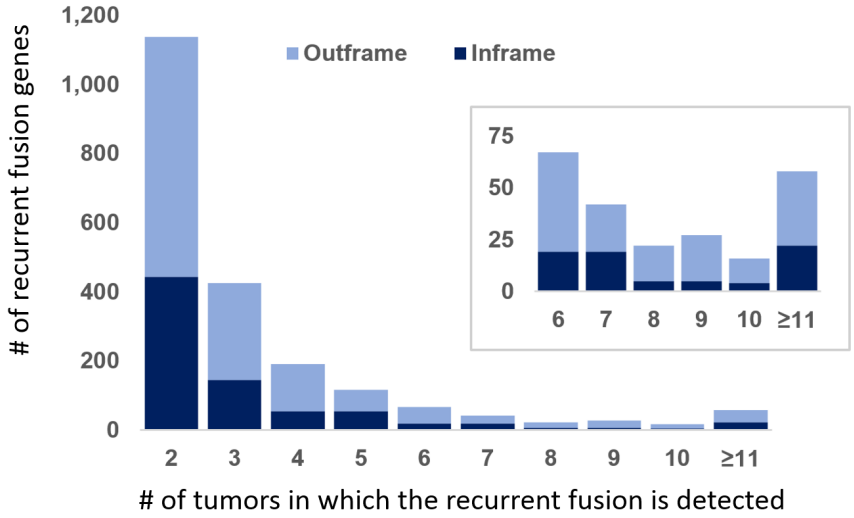
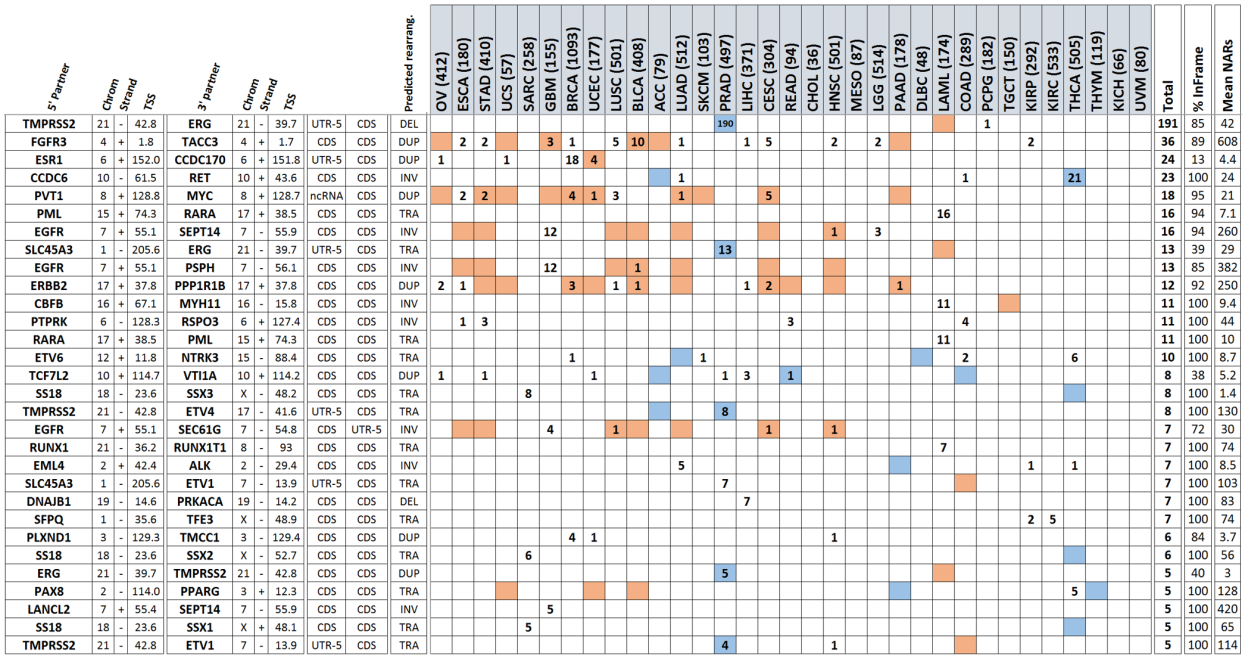


Figure 3.17: Frequency distribution of recurrent fusions. y-axis shows the number of distinct recurrent fusion genes observed at varying recurrence level (x-axis) in the TCGA. A fusion gene is classified as 'InFrame' if >80% of the fusion transcript calls are predicted to be in-frame.



5' Partner	Chrom	Strand	TSS	3' partner	Chrom	Strand	TSS			Predicted rearrang.	OV (412)	ESCA (180)	STAD (410)	UCS (57)	SARC (258)	GBM (155)	BRCA (1093)	UCEC (177)	LUSC (501)	BLCA (408)	ACC (79)	LUAD (512)	SKCM (103)	PRAD (497)	LIHC (371)	CESC (304)	READ (94)	CHOL (36)	HNSC (501)	MESO (87)	LGG (514)	PAAD (178)	DLBC (48)	LAML (174)	COAD (289)	PCPG (182)	TGCT (150)	KIRP (292)	KIRC (533)	THCA (505)	THYM (119)	KICH (66)	UVM (80)	Total	% InFrame	Mean NARs
TMPRSS2	21	-	42.8	ERG	21	-	39.7	UTR-5	CDS	DEL														190																		191	85	42
FGFR3	4	+	1.8	TACC3	4	+	1.7	CDS	CDS	DUP		2	2			3	1		5	10		1			1	5			2		2					2						36	89	608
ESR1	6	+	152.0	CCDC170	6	+	151.8	UTR-5	CDS	DUP	1			1			18	4																								24	13	4.4
CCDC6	10	-	61.5	RET	10	+	43.6	CDS	CDS	INV												1													1					21			23	100	24
PVT1	8	+	128.8	MYC	8	+	128.7	ncRNA	CDS	DUP		2	2				4	1	3			1				5																18	95	21
PML	15	+	74.3	RARA	17	+	38.5	CDS	CDS	TRA																								16								16	94	7.1
EGFR	7	+	55.1	SEPT14	7	-	55.9	CDS	CDS	INV						12													1		3											16	94	260
SLC45A3	1	-	205.6	ERG	21	-	39.7	UTR-5	CDS	TRA														13										16									13	39	29
EGFR	7	+	55.1	PSPH	7	-	56.1	CDS	CDS	INV						12				1																						13	85	382
ERBB2	17	+	37.8	PPP1R1B	17	+	37.8	CDS	CDS	DUP	2	1					3		1	1					1	2					1											12	92	250
CBFB	16	+	67.1	MYH11	16	-	15.8	CDS	CDS	INV																								11								11	100	9.4
PTPRK	6	-	128.3	RSPO3	6	+	127.4	CDS	CDS	INV		1	3														3								4								11	100	44
RARA	17	+	38.5	PML	15	+	74.3	CDS	CDS	TRA																								11								11	100	10
ETV6	12	+	11.8	NTRK3	15	-	88.4	CDS	CDS	TRA							1						1												2					6			10	100	8.7
TCF7L2	10	+	114.7	VTI1A	10	+	114.2	CDS	CDS	DUP	1		1					1						1	3		1															8	38	5.2
SS18	18	-	23.6	SSX3	X	-	48.2	CDS	CDS	TRA					8																											8	100	1.4
TMPRSS2	21	-	42.8	ETV4	17	-	41.6	UTR-5	CDS	TRA														8																			8	100	130
EGFR	7	+	55.1	SEC61G	7	-	54.8	CDS	UTR-5	INV						4				1						1			1														7	72	30
RUNX1	21	-	36.2	RUNX1T1	8	-	93	CDS	CDS	TRA																								7								7	100	74
EML4	2	+	42.4	ALK	2	-	29.4	CDS	CDS	INV												5																	1		1		7	100	8.5
SLC45A3	1	-	205.6	ETV1	7	-	13.9	UTR-5	CDS	TRA														7																			7	100	103
DNAJB1	19	-	14.6	PRKACA	19	-	14.2	CDS	CDS	DEL															7																			7	100	83
SFPQ	1	-	35.6	TFE3	X	-	48.9	CDS	CDS	TRA																														2	5			7	100	74
PLXND1	3	-	129.3	TMCC1	3	-	129.4	CDS	CDS	DUP							4	1											1														6	84	3.7
SS18	18	-	23.6	SSX2	X	-	52.7	CDS	CDS	TRA					6																												6	100	56
ERG	21	-	39.7	TMPRSS2	21	-	42.8	CDS	CDS	DUP														5																			5	40	3
PAX8	2	-	114.0	PPARG	3	+	12.3	CDS	CDS	TRA																																5			5	100	128
LANCL2	7	+	55.4	SEPT14	7	-	55.9	CDS	CDS	INV						5																												5	100	420
SS18	18	-	23.6	SSX1	X	+	48.1	CDS	CDS	TRA					5																												5	100	65
TMPRSS2	21	-	42.8	ETV1	7	-	13.9	UTR-5	CDS	TRA														4					1														5	100	114

Figure 3.18: Highly recurrent (5 or more) fusion genes that are previously reported (known fusions in Supplementary Table 3.14). Red and Blue backgrounds indicate that at least one of the two genes involved in fusions is in a region of focal amplification or deletion, respectively, within the cancer type. Focal amplifications and deletions are derived from GISTIC (see Section) calls generated by Zack et al. 2013 (see Methods). First 8 columns corresponding to gene information (name, chromosome, strand, transcription start site) for the 5' and 3' partner genes. Column 9 indicates the region in which the breaks occur in the two genes: UTR-5 and UTR-3 correspond to 5' and 3' untranslated regions, CDS - coding sequence, ncRNA - a non coding transcript. Column 10 shows the type of rearrangement predicted to generate the fusion event. DUP: duplication, TRA: translocation, INV: inversion, DEL: deletion. Columns 11-44 correspond to the tumor type (see Supplementary Table 3.1 for definitions of acronyms). Column 45 indicates the total number of tumors in which the fusion is found. Column 46, '% InFrame', shows the proportion of fusion transcripts that are predicted to be protein coding using annotated translational start site. Column 47 indicates the expression level of the fusion that is represented as 'Mean NARs' – mean normalized anchor reads is the mean value of anchor reads supporting each of the fusion transcript calls that are normalized to the library size in the respective tumor sample in which it is found.

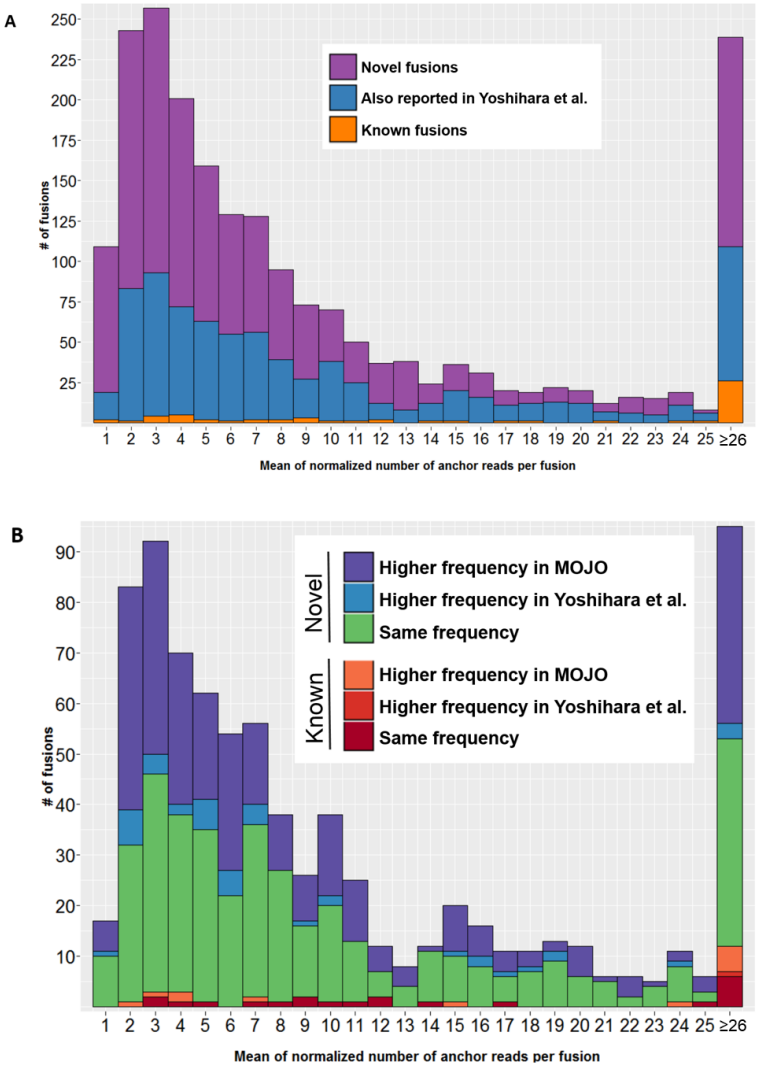


Figure 3.19: Expression characteristics of recurrent fusion genes. Expression level is determined as the mean of normalized anchor reads (NARs) for all the fusion transcripts corresponding to the fusion gene. NAR is the anchor read count normalized to the sequencing depth. For sample with multiple isoforms, the isoform with the highest number of isoform is considered. On x-axis is the discretized expression level from 1 - 25 reads. All recurrent fusions with mean NAR >25 are binned into >=26. (a) expression characteristics for all 2,070 recurrent fusions classified by whether the fusion gene is previously reported (Supplementary Table 3.19) or if the fusion is reported previously within the TCGA by Yoshihara et al., or if it is a novel fusion in TCGA discovered here. (c) Expression characteristics for recurrent fusions identified within the samples analyzed by both our study and Yoshihara et al. is shown. Each fusion is classified into two broad categories of Known (Supplementary Table 3.19) and Novel, and further into three additional categories each representing whether the fusion is reported at higher frequency in our study, in Yoshihara et al., or at identical frequencies. For both (a-b), recurrent fusions are identified across the expression spectrum suggesting that increased sensitivity of our approach is not exclusively due to our power to detect low expressed fusions.

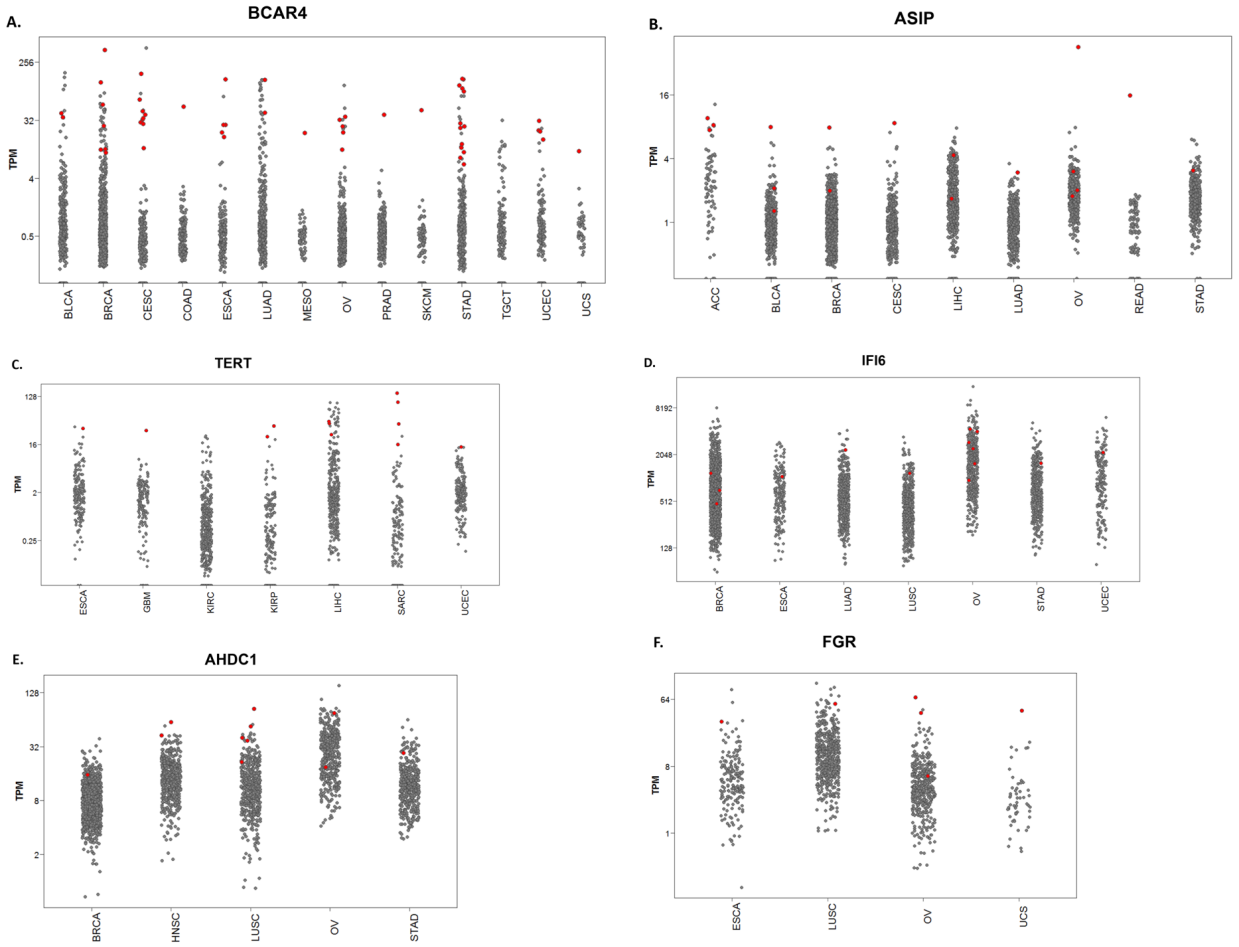


Figure 3.20: Expression characteristics of genes most frequently dysregulated by fusions

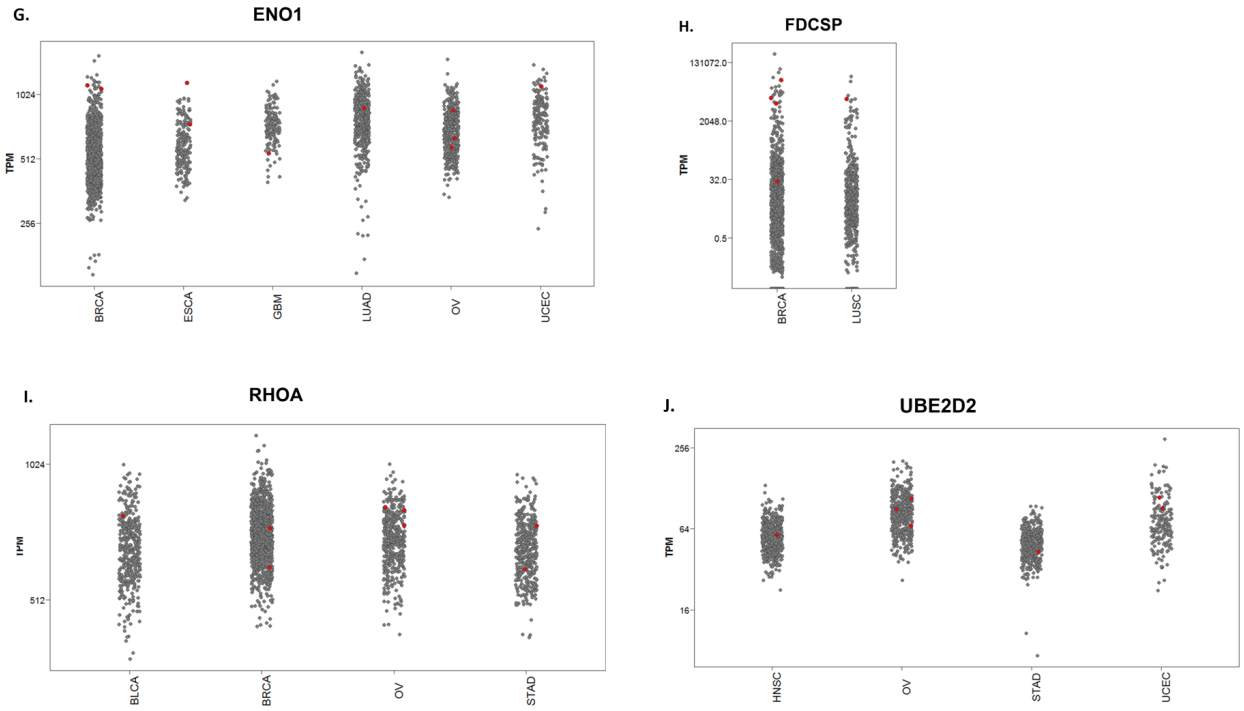


Figure 3.20: (Continued from previous page). Expression characteristics of genes most frequently dysregulated by fusions is shown here (corresponds to Figure 3.4C). For each of the panels, the expression of the gene is represented on y-axis as transcripts per million (TPM). See Section 3.5.2 for details on transcript quantification and normalization. On x-axis are select cancer types in which at least one tumor is predicted to have a fusion that results in dysregulation of the corresponding gene. Fusion-positive tumors are indicated in 'red' dots. (A) *BCAR4*, (B) *ASIP*, (C) *TERT*, (D) *IFI6*, (E) *AHDC1*, (F) *FGR*, (G) *ENO1*, (H) *FDSCP*, (I) *RHOA*, (J) *UBE2D2*. (see Supplementary Table 3.1 for definitions of tumor acronyms)

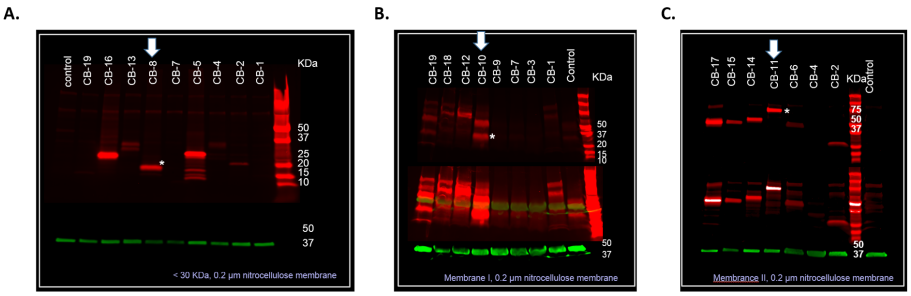


Figure 3.21: Western blots analysis of HA-tagged fusion proteins extracted from cell lysate of stable MCF10A cell lines expression BCAR4 (A), truncated PDLIM5 (B), and CD44-PDHX (C) fusion proteins. In each of the panels, the lane highlighted with arrow indicates the lane corresponding to the fusion protein. Protein lysates were separated on SDS-PAGE and transferred onto nitrocellulose membrate for immunoprecipitation with HA and anti-beta-actin antibodies. Red: HA, Green: beta-actin. Asterisk indicates the band corresponding to expected size for each protein. BCAR4: 14.2kDa, truncated PDLIM5: 30.7kDa, CD44-PDHX: 73.7 kDa. Expression of the fusion protein was also confirmed using immuno-fluorescence (Figure 3.22.

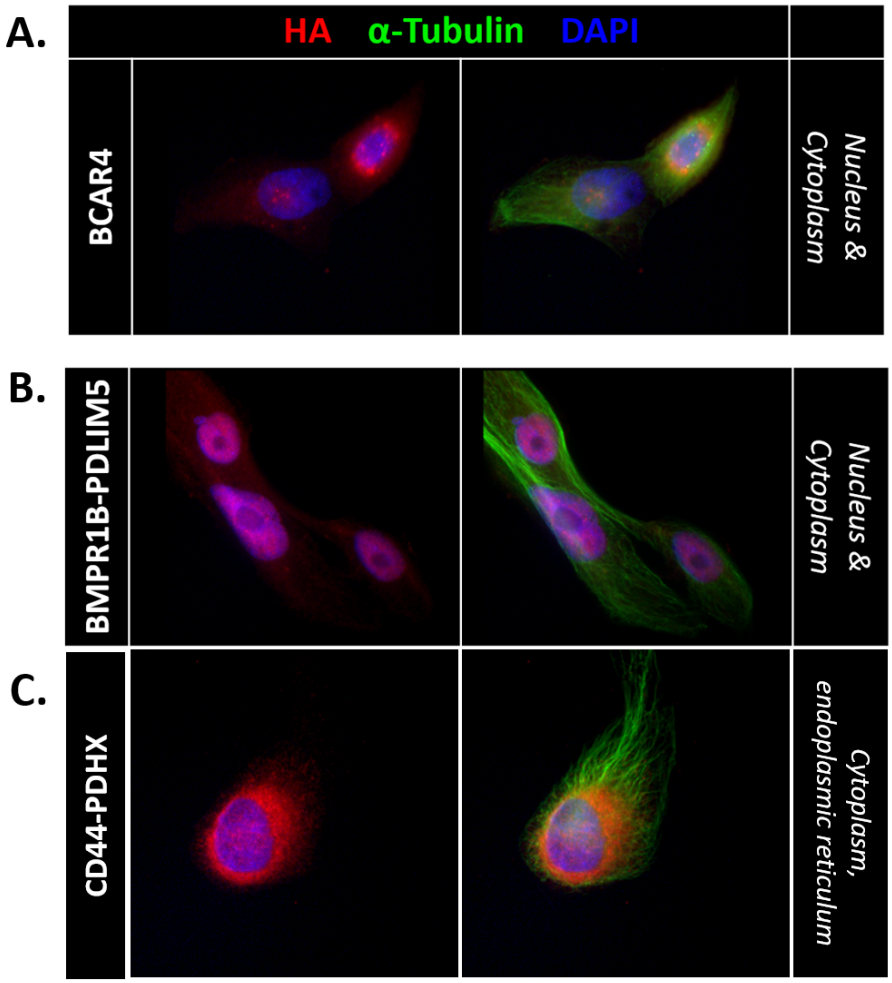


Figure 3.22: Cellular localization of HA-tagged fusion proteins: BCAR4 (A), truncated PDLIM5 (B) and CD44-PDHX (C). MCF10A stable cell lines expressing each of the three HA tagged proteins were immunostained with anti-alpha-tubulin antibody, anti-HA antibody and 6-diamidino-2-phenylindole (DAPI). Red: HA, Green: alpha-tubulin, DNA: blue.

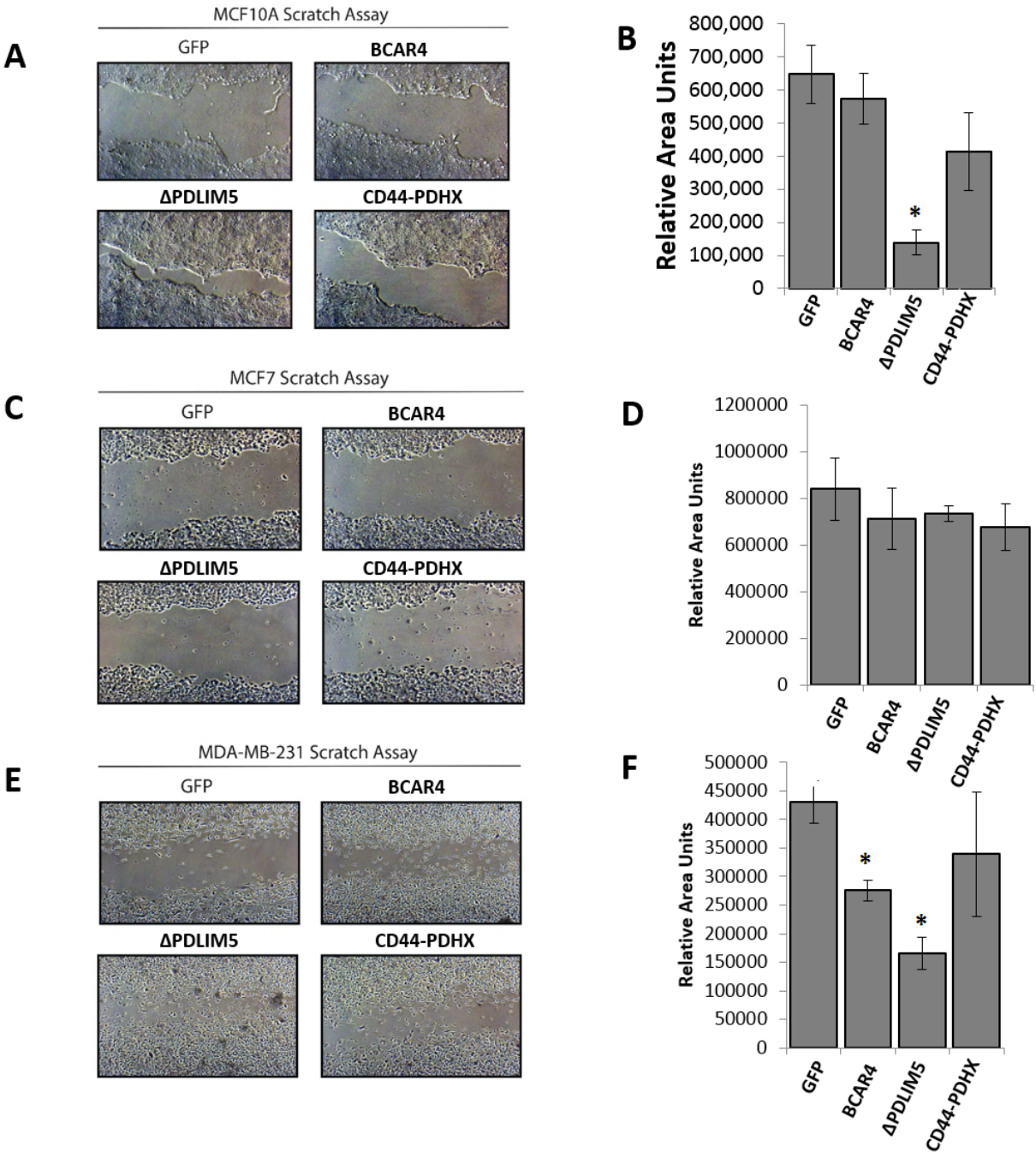


Figure 3.23: Wound healing assay to measure migration of MCF10A (A-B), MCF7 (C-D) and MDA-MB-231 (D-E) cells transfected with the three fusion constructs. In each of the three cell lines, the migratory phenotype of fusion-positive cells is quantified with respect to the corresponding GFP control. A scratch was applied to the cell monolayers of triplicates of each of the 12 cell lines (including controls) and migration of cells towards the would is measured at Day 0 and Day 2. Representative photographs after 48 hours are shown in (A,C,E). The extent of wound recovery was determined as described under Methods. Barcharts (B, D, F) show the mean and standard error of the amount of gap closed by each of the cell lines. Column: Mean of three experiments. Bar: s.e. (*P¡0.01).

3.7 Appendix: Supplementary Tables

		TCGA				
Tumor abbreviation	Tumor description	Primary	Recurrent	Paired Metastatic	Paired Normal	CCLE
ACC	Adrenocortical carcinoma	79				
BLCA	Bladder Urothelial Carcinoma	408			19	26
BRCA	Breast invasive carcinoma	1093		7	112	56
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	304		2	3	25
CHOL	Cholangiocarcinoma	36			9	
COAD	Colon adenocarcinoma	289	1	1	26	58
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48				57
ESCA	Esophageal carcinoma	180		1	13	26
GBM	Glioblastoma multiforme	155	6			
HNSC	Head and Neck squamous cell carcinoma	501		2	42	33
KICH	Kidney Chromophobe	66			25	
KIRC	Kidney renal clear cell carcinoma	533			72	25
KIRP	Kidney renal papillary cell carcinoma	290			32	
LAML	Acute Myeloid Leukemia	174				
LCLL	Chronic Lymphocytic Leukemia					81
LGG	Brain Lower Grade Glioma	514	14			65
LIHC	Liver hepatocellular carcinoma	371	2		50	32
LUAD	Lung adenocarcinoma	512	2		57	
LUSC	Lung squamous cell carcinoma	501			51	184
MESO	Mesothelioma	87				1
MM	Multiple Myeloma Plasma cell leukemia					25
OV	Ovarian serous cystadenocarcinoma	412	6			45
PAAD	Pancreatic adenocarcinoma	178		1	4	41
PCPG	Pheochromocytoma and Paraganglioma	179		2	3	
PRAD	Prostate adenocarcinoma	497		1	52	7
READ	Rectum adenocarcinoma	94	1		6	
SARC	Sarcoma	258	3	1	2	40
SKCM	Skin Cutaneous Melanoma	103		2		52
STAD	Stomach adenocarcinoma	410			32	41
TGCT	Testicular Germ Cell Tumors	150				
THCA	Thyroid carcinoma	505		8	59	12
THYM	Thymoma	119			2	
UCEC	Uterine Corpus Endometrioid Carcinoma	177	1		7	3
UCS	Uterine Carcinosarcoma	57				
UVM	Uveal Melanoma	80				
	Total Sample Counts:	**9360**	**36**	**28**	**678**	**935**

Supplementary Table 3.1: Statistics of primary tumor and cell line transcriptomes analyzed in this study.

Supplementary Table 3.2: (See workbook 3.2 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Sample manifest of TCGA samples analyzed in this study. Specific columns are:

Column 1: Tumor type
Column 2: Patient ID
Column 3: Primary tumor sample ID (empty if not exists)
Column 4: Corresponding relapse tumor sample ID (empty if not exists)
Column 5: Corresponding metstatic tumor sample ID (empty if not exists)
Column 6 : Tumor adjacent normal sample ID (empty if not exists)

Supplementary Table 3.3: (See workbook 3.3 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Sample manifest CCLE cell line transcriptomes analyzed in this study. Specific columns are:

Column 1: Tumor type
Column 2: Cellline ID
Column 3: Bam file name

Supplementary Table 3.4: (See workbook 3.4 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Sample manifest of GTEx normal tissues analyzed in this study. Specific columns are:

Column 1: Subject ID (also referred to as Donor ID)
Column 2: Run ID (transcriptome)
Column 3: TissueSampleID
Column 4: BodySite (specific tissue type)
Column 5: HistologicalType (organ)

Supplementary Table 3.5: (See workbook 3.5 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Sample manifest of HapMap cell line transcriptomes analyzed in this study. Specific columns are:

Column 1: ProjectID (NCBI generated ID)
Column 2: RunID (NCBI generated ID)
Column 3: Sex
Column 4: Population
Column 5: URL to raw sequence (first end)
Column 6: URL to raw sequence (second end)

Supplementary Table 3.6: (See workbook 3.6 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Primary tumor transcriptomes used in this study for comparisons. Fusions previously reported within these samples are also listed here. Specific columns are:

Column 1: Tumor type
Column 2: Sample name
Column 3: 5' partner of the fusion
Column 4: 3' partner of the fusion
Column 5: Primary reference reporting the fusion
Column 6: Notes

Supplementary Table 3.7: (See workbook 3.7 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). MOJO-PC workflow comparison with deFuse, FusionCatcher and MapSplice. Dataset: 55 transcriptomes. Only canoncial fusions are compared here. Specific columns are:

Column 1: Sample name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: MOJOPancancer - MOJO-PC workflow (1: fusion nominated; 0: not nominated)
Column 4: MOJO - MOJO standalone(1: fusion nominated; 0: not nominated)
Column 5: deFuse (1: fusion nominated; 0: not nominated)
Column 6: MapSplice (1: fusion nominated; 0: not nominated)
Column 7: FusionCatcher (1: fusion nominated; 0: not nominated)
Column 13: Total # of methods nominating the fusion
Column 14: Is fusion previously reported (1: yes; 0: no)
Column 15: # of anchor reads supporting the fusion (if nominated by multiple methods, the maximum value is used)
Column 16: Distance between the two genes

Supplementary Table 3.8: (See workbook 3.8 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). MOJO-PC workflow comparison with deFuse, FusionCatcher and MapSplice. Dataset: 55 transcriptomes. Only non-canoncial fusions are compared here. Specific columns are:

Column 1: Sample name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: MOJOPancancer - MOJO-PC workflow (1: fusion nominated; 0: not nominated)
Column 4: MOJO - MOJO standalone(1: fusion nominated; 0: not nominated)
Column 5: deFuse (1: fusion nominated; 0: not nominated)
Column 6: MapSplice (1: fusion nominated; 0: not nominated)
Column 7: FusionCatcher (1: fusion nominated; 0: not nominated)
Column 8: Total # of methods nominating the fusion
Column 9: Is fusion previously reported (1: yes; 0: no)
Column 10: # of anchor reads supporting the fusion (if nominated by multiple methods, the maximum value is used)
Column 11: Distance between the two genes

Supplementary Table 3.9: (See workbook 3.9 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). MOJO-PC workflow comparison with deFuse, FusionCatcher and MapSplice. Dataset: 126 transcriptomes. Only canoncial fusions are compared here. Specific columns are:

Column 1: Sample name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: MOJOPancancer - MOJO-PC workflow (1: fusion nominated; 0: not nominated)
Column 4: MOJO - MOJO standalone(1: fusion nominated; 0: not nominated)
Column 5: deFuse (1: fusion nominated; 0: not nominated)
Column 6: MapSplice (1: fusion nominated; 0: not nominated)
Column 7: FusionCatcher (1: fusion nominated; 0: not nominated)
Column 8: Total # of methods nominating the fusion
Column 9: Is fusion previously reported (1: yes; 0: no)
Column 10: # of anchor reads supporting the fusion (if nominated by multiple methods, the maximum value is used)
Column 11: Distance between the two genes

Supplementary Table 3.10: (See workbook 3.10 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). MOJO-PC workflow comparison with deFuse, FusionCatcher and MapSplice. Dataset: 126 transcriptomes. Only non-canoncial fusions are compared here. Specific columns are:

Column 1: Sample name
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: MOJOPancancer - MOJO-PC workflow (1: fusion nominated; 0: not nominated)
Column 4: MOJO - MOJO standalone(1: fusion nominated; 0: not nominated)
Column 5: deFuse (1: fusion nominated; 0: not nominated)
Column 6: MapSplice (1: fusion nominated; 0: not nominated)
Column 7: FusionCatcher (1: fusion nominated; 0: not nominated)
Column 8: Total # of methods nominating the fusion
Column 9: Is fusion previously reported (1: yes; 0: no)
Column 10: # of anchor reads supporting the fusion (if nominated by multiple methods, the maximum value is used)
Column 11: Distance between the two genes

Supplementary Table 3.11: (See workbook 3.11 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). RT-PCR validation results from 12 cell lines. Specific columns are:

Column 1: Cell line name
Column 2: 5' gene name
Column 3: 5' gene chrom
Column 4: 5' gene strand
Column 5: 5' gene breakpoint
Column 6: 3' gene name
Column 7: 3' gene chrom
Column 8: 3' gene strand
Column 9: 3' gene breakpoint
Column 10: # of anchor reads
Column 11: 5' primer
Column 12: 3' primer
Column 13: Validation successful (1: yes, 0: no)

Supplementary Table 3.12: (See workbook 3.12 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). True positives detected by MOJO-PC and Yoshihara et al. Specific columns are:

Column 1: Tumor sample ID
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: Fusion detected by MO-PC (1: yes, 0: no)
Column 5: Fusion detected by Yoshihara et al.(1: yes, 0: no)

Supplementary Table 3.13: (See workbook 3.13 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). True positives detected by MOJO-PC and Stransky et al. Specific columns are:

Column 1: Tumor sample ID
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: Fusion detected by MO-PC (1: yes, 0: no)
Column 5: Fusion detected by Stransky et al.(1: yes, 0: no)

Supplementary Table 3.14: (See workbook 3.14 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). List of previously reported fusions compiled from COSMIC, Mitelman and other sources). Specific columns are:

Column 1: 5' gene name
Column 2: 3' gene name
Column 3: Source

Supplementary Table 3.15: (See workbook 3.15 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Fusion calls nominated by MOJO-PC from 9,360 primary tumors. Specific columns are:

Column 1: Tumor type
Column 2: Patient ID
Column 3: Tumor sample type (Primary)
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6: # of normalized anchor reads supporting the fusion (Column 5 normalized to sequencing depth)
Column 7-11: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 12-16: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 17: # of isoforms observed for the fusion
Column 18: Is fusion in-frame (1: yes, 0: no)
Column 19: Known fusion (1: yes, 0: no)
Column 20: Fusion detected by Yoshihara et al? (1: yes, 0: no)
Column 21: Fusion detected by Stransky et al? (1: yes, 0: no)
Column 22: Predicted rearrangement type (TRA - translocation, DEL - deletion, INS - insertion, INV - inversion)

Supplementary Table 3.16: (See workbook 3.16 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Fusion calls nominated by MOJO-PC from relapse/metastatic tumors. Specific columns are:

Column 1: Tumor type
Column 2: Patient ID
Column 3: Tumor sample type (Relapse/Metastatic)
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6: # of normalized anchor reads supporting the fusion (Column 5 normalized to sequencing depth)
Column 7-11: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 12-16: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 17: # of isoforms observed for the fusion
Column 18: Is fusion in-frame (1: yes, 0: no)
Column 19: Known fusion (1: yes, 0: no)
Column 20: Fusion detected by Yoshihara et al? (1: yes, 0: no)
Column 21: Fusion detected by Stransky et al? (1: yes, 0: no)
Column 22: Predicted rearrangement type (TRA - translocation, DEL - deletion, INS - insertion, INV - inversion)

Supplementary Table 3.17: (See workbook 3.17 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Fusion calls nominated by MOJO-PC from adjacent normal tissues. Specific columns are:

Column 1: Tumor type
Column 2: Patient ID
Column 3: Tumor sample type (Normal)
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6: # of normalized anchor reads supporting the fusion (Column 5 normalized to sequencing depth)
Column 7-11: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 12-16: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 17: # of isoforms observed for the fusion
Column 18: Is fusion in-frame (1: yes, 0: no)
Column 19: Known fusion (1: yes, 0: no)
Column 20: Fusion detected by Yoshihara et al? (1: yes, 0: no)
Column 21: Fusion detected by Stransky et al? (1: yes, 0: no)
Column 22: Predicted rearrangement type (TRA - translocation, DEL - deletion, INS - insertion, INV - inversion)

Supplementary Table 3.18: (See workbook 3.18 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Correlations between fusions and genomic instability across cancer types Specific columns are:

Column 1: Tumor type
Column 2: Spearman rank correlation between number of fusions and the genomic instability within tumors
Column 3: bonferroni corrected p-value

Supplementary Table 3.19: (See workbook 3.19 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Gene categories enriched for somatic fusion genes. 1453 gene categories in Molecular Signature Database were used to compute enrichment. Specific columns are:

Column 1: Gene Ontology category (pathway)
Column 2: Observed # of fusions involving genes within the category
Column 3: Expected # of fusions involving genes within the category
Column 4: Fold change (Observed/Expected)
Column 5: Empirical p-value (100 million permutations)
Column 6: Q-value

Supplementary Table 3.20: (See workbook 3.20 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Recurrently fused cancer associated genes. Specific columns are:

Column 1: Gene name
Column 2: Total # of fusions
Column 3: # of fusions with gene as upstream partner
Column 4: # of fusions with gene as downstream partner
Column 5: Is COSMIC cancer gene (1: yes)
Column 6: Is Tumor Suppressor (1: yes, 0: no)
Column 7: Is Oncogene (1: yes, 0: no)
Column 8: Is Kinase (1: yes, 0: no)
Column 9: Is InFrame (1: yes, 0: no)
Column 10: # fusions with copy number data available from SNP arrays
Column 11: # fusions supported by copy number alterations on SNP arrays

Supplementary Table 3.21: (See workbook 3.21 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Fusions with intact kinase domains. Specific columns are:

Column 1: Tumor Sample ID
Column 2: 5' gene name
Column 3: 3' gene name
Column 4: 5' Exon Id
Column 5: 3' Exon Id
Column 6: Kinase-only fusion - Known fusion
Column 7: Kinase-only fusion - involving known kinase
Column 8: Kinase-only fusion - novel fusion
Column 9: Kinase + coiled-coil domain containing fusion - Known fusion
Column 10: Kinase + coiled-coil domain containing fusion - involving known kinase
Column 11: Kinase + coiled-coil domain containing fusion - novel fusion

Supplementary Table 3.22: (See workbook 3.22 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Recurrent chimeric proteins. Specific columns are:

Columns 1-4: 5' gene info (name, chromosome, strand, transcriptional start site)
Columns 5-8: 3' gene info (name, chromosome, strand, transcriptional start site)
Column 9: 5' break position
Column 10: 3' break position
Column 11: predicted rearrangement type
Column 12-45: 33 cancers
Column 46: Total # of fusions
Column 47: % of fusions that are predicted to be inframe
Column 48: Mean # of normalized anchor reads (NARs).
Column 49: Is Known fusion
Column 50: Reported by Yoshihara et al.
Column 51: Reported by Stransky et al.

Supplementary Table 3.23: (See workbook 3.23 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Recurrent out-of-frame fusions. Specific columns are:

Columns 1-4: 5' gene info (name, chromosome, strand, transcriptional start site)
Columns 5-8: 3' gene info (name, chromosome, strand, transcriptional start site)
Column 9: 5' break position
Column 10: 3' break position
Column 11: predicted rearrangement type
Column 12-45: 33 cancers
Column 46: Total # of fusions
Column 47: % of fusions that are predicted to be inframe
Column 48: Mean # of normalized anchor reads (NARs).
Column 49: Is Known fusion
Column 50: Reported by Yoshihara et al.
Column 51: Reported by Stransky et al.

Supplementary Table 3.24: (See workbook 3.24 in Supplementary.Tables.Chapter3.xlsx associated with this dissertation). Recurrently dysregulated genes. Specific columns are:

Columns 1-4: 5' gene info (name, chromosome, strand, transcriptional start site)
Column 1: Gene name
Column 2: # of distinct 5' partners
Column 3-36: 33 cancers
Column 37: Total # of fusions
Column 38: Mean # of normalized anchor reads (NARs).
Column 39: Is Known fusion
Column 40: Reported by Yoshihara et al.
Column 41: Reported by Stransky et al.

3.8 Contributions

I designed the project with my advisor, Kevin White. I performed the data acquisition, quality control, analysis, compiled the figures and wrote the manuscript with inputs from Kevin. I identified the cell lines for RT-PCR validations for evaluating specificity, designed the primers and perform validations along with Vineet Dhiman, who validated majority of them.

I identified select recurrent fusion genes for functional validations. Jennifer Moran (IGSB's Recombineering Core) and Jingdong Tian (General Biosystems, Inc) synthesized the fusion gene constructs. Alex Yemelyanov and Pankaj Bhalla at the Northwestern Skin and Disease Research core packaged the fusion gene plasmids into lenti-viral particles. Pei-Chun optimized the protocol for generating stable fusion gene expressing cell lines. Functional validations for proliferation, invasion and migration were performed by Pei-Chun Lin, Shenglang Gao and Vineet Dhiman. I would like to acknowledge Siquan Chen and Sam Bettis of Cellular Screening Center at the University of Chicago for maintaining cell lines in culture and providing the robotics and support to do high throughput screening of fusion cell lines.

ASCAT copy number calls for the TCGA tumors was provided by Peter Van Loo at the Sanger Institute. I would like to acknowledge Lorenzo Pesce and Joe Urbanski at the Computation Institute for patiently responding to queries, setting up block reservations for compute nodes and facilitating large scale analyses on the Beagle super computer. I also would like to thank Jason Pitt, Michael Bolt, Megan McNerney and the members of IGSB's TCGA working group for helpful discussions during the course of this project.

CHAPTER 4

DISCOVERY OF WIDESPREAD INHERITED POLYMORPHIC FUSION GENES

4.1 Abstract

Recent studies have extensively characterized structural variations (SV) such as deletions, duplications, insertions and inversions across the human genome. However, the landscape of fusion genes, generated by SVs that span multiple genes, is poorly understood, primarily due to technical limitations in identifying them. Here, using a highly specific pipeline, we performed a survey of fusions from a median of 15 tissue transcriptomes derived from each of 524 individuals. We identified 63 high confidence inherited fusion genes. We successfully identified a SV supporting the fusion gene for 21 out of 24 fusions within the donors for whom genome sequencing available. 44/63 fusion genes are also detected in a transcriptome of at least one individual in the TCGA (n=9,270 individuals) and HapMap (n=464) projects. Six recurrent fusion genes showed statistically significant difference is fusion frequency between different populations. We identified 26 fusion genes that result in potential dsyregulation of the 3' partner gene resulting in its ectopic expression. For example, swapping of regulatory regions between ubiquitously expressed *CASP4* and tissue-specific capsase-1 inhibitor, *CARD18*, resulted in constitutive overexpression of the latter. Our findings, for the first time, demonstrate the pervasiveness of fusion genes in human populations and highlight their potential contribution to population variation.

4.2 Introduction

Copy number variants (CNVs) such as amplifications and deletions contribute to more than 10 fold higher inter-individual genomic variation compared to single nucleotide variants (SNVs) (Weischenfeldt et al. 2013). While CNVs can be detected using comparative

genomic hybridization or SNP arrays, other types of structural variants (SVs) such as inversions and inter-/intra- chromosomal translocations of sequences can only be identified with whole genome sequencing (WGS). Recent SV analysis of over 2,500 whole genomes identified 55%, 76% and 94% novel deletions, duplications and inversions that have not been previously reported. This indicates that current estimates of the structural variation in the human genome are conservative (Sudmant et al. 2015a). Many CNVs have been identified as associated with Mendelian disorders or de-novo and inherited diseases (Zhang et al. 2009). Recent evidence of strong positive and negative selective constraints on deletions and duplications of dosage sensitive genes, as well as non-coding regions, demonstrate their contribution to phenotypic variation and human adaptation (Iskow et al. 2012a).

Fusion genes are generated by SVs that juxtapose regulatory or coding domains of two distinct genes. Such events are a hallmark of cancer genomes, with numerous somatic fusion genes discovered with strong oncogenic potential (Mertens et al. 2015). Despite the higher degree of genomic instability that generates fusion genes in cancer, the shared mechanisms for DNA repair between cancer and normal cells led us to hypothesize that many fusion genes may have been be generated during human evolution and therefore may be segregating in human populations. To date, only four fusion genes have been reported in healthy individuals. *TFG-GPR128*, generated by a tandem duplication event in 3q12.2 was discovered in a screen for pathogenic somatic variants in myeloproliferative neoplasms and is detected at 2% frequency in healthy individuals (Chase et al. 2010). Similarly, in an assay to identify CNVs associated with autism, *MAPKAPK5-ACAD10* was found at 0.5% frequency in both cases and controls (Holt et al. 2012). Two other fusion transcripts, *KANSL1-ARL17A* and *NAIP-OCLN*, are predicted to be generated from regions of complex structural variation in 17q21.3 and 5q13.2, respectively (Courseaux et al. 2003; Boettger et al. 2012). Despite their discovery, the functional significance as well as the population genetic properties of these events have not been evaluated.

The ability to comprehensively characterize the fusion genes from WGS studies is con-

strained by technical limitations. The complexity of the breakpoint region due to the propensity for breaks in highly repetitive regions (Argueso et al. 2008), or clustering of multiple breakpoints in a short region (Sudmant et al. 2015b) or inclusion of additional sequence from distant loci (Simsek et al. 2010) can be challenging to overcome when linking two different genes, especially with short paired-end sequencing data. These factors are mitigated by a transcriptome based approach that detects exon-exon fusion junctions of properly spliced transcripts. We hypothesize that a transcriptome based approach can be used to detect expressed polymorphic fusion genes from healthy tissues. Limitations of this approach include accurately classifying true polymorphic events from those generated by trans-splicing (Horiuchi et al. 2006) or random chimeras between highly expressed genes (Frenkel-Morgenstern et al. 2012). In addition, low expressed fusions or those that are subject to non-sense mediated decay may evade detection in a transcriptome based approach. Despite these limitations, an RNA-seq based approach remains a powerful approach to identify fusions that are expressed at moderate to high levels.

In this study, we sought to identify fusion genes from over five hundred donors in the Genotype Tissue Expression (GTEx) project. Leveraging 15 tissues per donor allowed us to account for false positives with high specificity. We identified population-specific enrichment for some of the fusion genes, suggesting population bottlenecks or recent selection at these loci. Gene expression analysis showed that the fusion genes have the potential to induce ectopic expression in a wide range of tissues. Overall, our results demonstrate the potential functional significance of the fusion genes and provide emphasis for experimental and population genetics analyses.

4.3 Results

4.3.1 Fusion gene discovery from healthy tissues

We sought to identify expressed fusion genes within 524 donors in the GTEx project. A unique aspect of the GTEx project is the availability of transcriptome data from multiple tissues for each donor (median 15 tissues/donor, Figure 4.1A), allowing us to treat each tissue as a biological replicate for identifying high confidence fusion genes. By leveraging the frequency of the fusion within the multiple tissues of an individual and its overall frequency across the donors, many of the technical and biological artifacts that confound fusion discovery can be mitigated. To this effect, we hypothesized that spurious fusions can be pervasively detectable across donors but with low frequency within the tissues of each donor, whereas, germline fusion genes can be detected in many tissues of a subset of donors.

We performed fusion transcript discovery using MOJO (see chapter 2) on 8,800 tissue transcriptomes in GTEx (Figure 4.1A, 4.4, Supplementary Table 4.1). To further control for specificity, we pooled anchor read evidence for recurrent fusion transcripts across the transcriptomes and controlled for technical artifacts resulting from spurious alignments, ambiguous junctions due to repetitive regions and random chimeras resulting from highly expressed genes or chimeric artifacts (Figure 4.1B, Methods). After excluding fusion genes that are predicted to be read-through events (fusions between genes that are <200kb in the same orientation) we identified 422 fusion transcripts recurrently detected in at least three tissues of one donor (Supplementary Table 4.2). 99.1% of all tissues nominated five or fewer fusion transcripts, suggesting that our findings are not enriched by outlier transcriptomes (Figure 4.5). We next applied two filters to control for potential characteristics of post-transcriptional events: fusion transcripts involving genes with multiple partners and those that are low expressed. First, we filtered out 119 fusion transcripts involving genes fused to multiple partners that are >5Mb apart (multi-partner fusions, Supplementary Table 4.2). 17 distinct genes comprised 87% (104/119) of the multi-partner fusions (Figure 4.5A,

Supplementary Table 4.3). Among the tissues with multi-partner genes, 93% of them have only one fusion in this category, indicating that this phenomenon is not a consequence of aberrant splicing (Figure 4.5B). However, we find a subset of tissue types showing marginal enrichment for these events (Figure 4.6A). The source of these events (that we excluded from further analysis) remains to be explored. Second, we hypothesized that the low expressed fusion transcripts may be challenging to disambiguate from those that are due to potential background level of aberrant transcription. We therefore filtered out 188 fusion transcripts with low overall expression (mean normalized anchor read count <2) across all tissues (Figure 4.6B, Supplementary Table 4.3).

We next leveraged the multiple tissues per donor to further classify the remaining 115 fusions according to their recurrence within each donor. We identified 63 fusion genes that were detected in a median of three or more tissues per donor and represent a high-confidence set of candidate polymorphic fusion genes (Supplementary Tables 4.4-4.5, Figure 4.6C-D). 82.5% (52/63) of the fusion genes are detected in a single donor and supported by a median of 10 tissues demonstrating that these are unlikely to be artifacts (Figure 4.1C). Among the donors with fusion genes, we found the corresponding fusion transcript in a median of 72% of tissues/donor (ranging from 10 to 100%) and found a correlation between expression level of the fusion and the fraction of fusion-positive tissues within the donor (rho=0.73, p<1e-11, Figure 4.7). This, in addition to the correlation between sequencing depth and the number of fusion calls (rho: 0.16, p<2e-16), suggests that the false negatives in the fraction of tissues without the fusion could be due to a sensitivity issue that is a function of sequencing depth and the expression level of the fusion. Higher frequency for random chimeras between highly expressed genes has been previously reported (Frenkel-Morgenstern et al. 2012). We examined the expression characteristics of genes involved in fusion genes and find that, for 84% of the fusion genes, the median expression of the 5' and 3' partner genes in the fusion-positive samples are within two standard deviations (s.d.) from their pan-cohort expression mean and within 1 s.d. for 63% of the fusion genes, indicating that our findings are not

driven by highly expressed genes (Figure 4.8).

For a subset of donors in GTEx with WGS data available, we sought to find evidence for structural variants supporting the fusion genes. We identified a supporting SV for 22 out of 24 fusion genes within 21 donors for which WGS is available (Table 4.1, Supplementary Table 4.6). The two fusion genes without a supporting SV could be generated by complex rearrangements and, thus, false negatives of the SV analysis. For example, one of these two fusions, *NAIP-OCLN*, has been previously reported as chimeric transcript originating from a complex segmentally duplicated region on chromosome 5. An alternate source of SVs is the database of genomic variants (DGV) that is compiled from large scale copy number analyses (MacDonald et al. 2014). We note that the variants in DGV are enriched for those that are predicted to be copy number altering such as amplifications and deletions that can be detected on SNP arrays. We find evidence for a SV in DGV spanning the two fused genes for 39.6% (25/63) of the fusion genes including 8 out of 11 recurrent ones. The three fusion genes with no evidence in DGV are predicted to be generated by copy number neutral events such as inversions (n=2) and inter-chromosomal events (n=1) (Table 4.1). Taken together, our findings demonstrate the high specificity of our fusion discovery approach.

4.3.2 Characteristics of polymorphic fusion genes

94% (59/63) of fusions reported in this study are novel and 17% (11/63) are detected in at least two donors in GTEx (Figure 4.1C). We identified all four previously reported fusion genes, *KANSL-ARL17A* (n=177 donors), *NAIP-OCLN* (n=58), *TFG-GPR128* (n=13) and *MAPKAPK5-ACAD10* (n=2). Spatial characteristics show that 87% (55/63) of the fusions are between proximal genes (<1Mb) on the same chromosome (Figure 4.1D). 68% and 17% of these are predicted to be generated by duplication and inversion events, respectively. We also find evidence for complex structural rearrangements resulting in a fusion of one gene to two distinct proximal genes. For example, we find one donor with *C9orf86* fused to *KIAA1984* as well as *TMEM141* (all of which are on 9q34.3). Similarly, fusions involving

DLG1 (3q29) and *INTS4* (11q14.1) fused to proximal genes are identified in one donor each (Table 4.1). We find evidence for reciprocal fusion transcripts for three fusion genes, *PDE1C-DNAJC6* (n=2 donors), *AUH-SYK* (n=1) and *S100PBP-YARS* (n=1). Since we detect a fusion based on the junction involved, it remains to be determined if the entire downstream and upstream regions of the genes are maintained in the balanced reciprocal fusion. For example, inspection of the orientation of the genes and the exons involved in the inter-chromosomal reciprocal *DNAJC6* (1p31.3)-*PDE1C* (7p14.3) event suggests that the genomc region in *DNAJC6* between introns 8 and 15 may be copied or translocated downstream of exon 20 of *PDE1C*. We also find two additional translocation derived fusion events, *MAEA* (4p16.3)-*FAM9B* (Xp22.32) and *MARK3* (14q32.32)-*MKNK2* (19p13.3) in one donor each. Loci for these four genes do not have any known mobile element insertions or sequences that show active retrotransposon activity (Stewart et al. 2011; Sudmant et al. 2015a).

Rare *de novo* translocation events have been previously reported at a frequency of 1/2000 (Warburton 1991). Given that 82.5% (52/63) fusion genes identified here are detected in only one individual, we investigated if these events are low frequency polymorphic fusions (VAF <0.2%) or of *de novo* origin that are found in only related donors. We performed targeted fusion analysis on the 9,499 transcriptomes in the Cancer Genome Atlas (TCGA) project. Although TCGA tissues are derived from tumors, we expect to see transcripts from polymorphic fusion genes expressed in the tumor transcriptomes as well. Supporting fusion transcripts in TCGA are required to share exon-exon junctions identical to TCGA. 68% (40/59) of all novel fusions are detected in at least one individual in the TCGA and 54% in multiple individuals. 7/10 highly recurrent fusions in GTEx are also identified in at least one of the 464 donors in HapMap project. Although the frequencies of the fusions detected in both TCGA and GTEx are concordant, the frequencies in TCGA are slightly but consistently lower. This can be attributed to the parameters limiting the sensitivity to detect these events in TCGA such as single-tissue per individual and varying sequencing

depth. Overall, our findings indicate that a substantial proportion of the fusions discovered here are polymorphic fusion genes segregating at low frequencies in the human population.

4.3.3 Population genetic properties of polymorphic fusion genes

44% of all donors in GTEx have at least one fusion gene and 18% have a low frequency fusion gene (variant allele frequency, VAF <2.5%, Figure 4.2A). We investigated the population genetic characteristics of the 11 recurrently detected fusion genes (>=2 donors, Table 4.1) across three different populations: European Americans (EA), African Americans (AA), Asians (Asian), in three different cohorts: GTEx (EA, AA only), TCGA (all three populations) and HapMap (EA, AA only). We sought to find fusion genes that show significant difference in the frequency of the fusion between pairs of populations within each of the cohorts (binomial test). We find 6 out of 11 recurrent fusion genes with significant differences in frequencies between at least two populations (Figure 4.2). The most striking differences in allele frequencies are for two previously known fusions *KANSL1-ARL17A* (17q21.31) and *NAIP-OCLN* (5q13.2) that are in the regions of complex genomic rearrangements in each loci. Nine different haplotypes involving two independent inversions and multiple duplications have been previously reported in 17q21.31 (Boettger et al. 2012; Steinberg et al. 2012). Interestingly, this locus is also shown previously to be under strong positive selection in EA (Stefansson et al. 2005). Multiple phenotypes such as 17q21.31 microdeletion syndrome and various neurodegenerative diseases such as Parkinson's disease have been previously shown to be associated with haplotypes in this locus (Sharp et al. 2006; Koolen et al. 2006; Zody et al. 2008; Koolen et al. 2008; Simon-Sanchez et al. 2009; Skipper et al. 2004; Tobin et al. 2008). Here we detect *KANSL1-ARL17A* fusion transcript at 7 fold higher frequency in EA vs. AA and 6 fold higher frequency in EA vs. Asian (Figure 4.2). Although out-of-frame, *KANSL1-ARL17A* can generate a partial 430aa peptide of *KANSL1* (KAT8 Regulatory NSL Complex Subunit 1) that retains the coiled-coil domain but the KAT8 interaction and catalytic domains are lost. It remains to be investigated if this truncated protein has a

							GTEx				TCGA				HapMap LCLs					
Fusion	Type	BreakType	InFrame?	GTEx CNV	DGV CNV	WGS SV	GTEx Total	White (n=443)	AA (n=71)	Asian (n=6)	TCGA-total	White (n=7164)	AA (n=810)	Asian (n=622)	LCLs-total	CEU (n=92)	FIN (n=95)	GBR (n=95)	TSI (n=93)	YRI (n=89)
KANSL1_ARL17A	Dup_<1	CDS-CDS			1	1	**177**	172	4		**2898**	2445	54	39	**73**	26	16	20	11	
NAIP_OCLN	Inv_<1	CDS-UTR-5	1			0	**58**	25	33		**266**	103	125	6	**9**					9
* TFG_GPR128	Dup_<1	CDS-CDS	1	1	1	1	**13**	13			**192**	172	4		**4**			2	2	
SPSB1_H6PD	Dup_<1	UTR-5-CDS		2	1		**4**	4			**27**	25								
GATC_COX6A1	Dup_<1	CDS-CDS			1	1	**3**	2	1		**40**	32	2	1	**1**			1		
TFDP2_XRN1	Dup_<1	CDS-CDS	1		1	1	**3**	3			**15**	12								
TRPM4_PPFIA3	Dup_<1	CDS-CDS	1		1		**3**		3		**8**	1	5							
PARG_BMS1	Inv_>1	CDS-CDS	1				**2**		2		**53**	13	8	31	**3**					3
MAPKAPK5_ACAD10	Dup_<1	CDS-CDS			1		**2**	2			**32**	29			**1**			1		
* VPS33A_CLIP1	Dup_<1	CDS-CDS		1	1	1	**2**	2			**18**	14	1		**1**	1				
PDE1C_DNAJC6	Trans	CDS-CDS					**2**		2											
LPHN2_ODF2L	Inv_>1	CDS-UTR-5	1				**1**	1			**79**	59	4	13						
* SEMA4B_IDH2	Inv_<1	UTR-5-UTR-5	1				**1**	1			**12**	6	1	1						
SNX30_KIAA1958	Dup_<1	CDS-CDS			1		**1**				**11**	1	9							
ARRB1_GDPD5	Dup_<1	CDS-UTR-5	1			1	**1**	1			**9**	4	3							
* TTC28_CHEK2	Dup_<1	CDS-CDS	1		1		**1**	1			**7**	5	1	1						
COMMD10_AP3S1	Dup_<1	CDS-CDS	1		1	1	**1**	1			**7**	7								
RSF1_INTS4	Dup_<1	CDS-CDS	1			1	**1**	1			**7**	6		1						
LOC100129917_CPLX1	Dup_<1	CDS-CDS				1	**1**		1		**7**	7								
ZNF555_ZNF554	Dup_<1	CDS-CDS			1		**1**	1			**5**	5								
INTS4_NDUFC2	Dup_<1	CDS-CDS			2	1	**1**	1			**5**	4								
DLG1_BDH1	Dup_<1	CDS-UTR-5	1	1		1	**1**	1			**5**	2								
DNAJC6_PDE1C	Trans	CDS-CDS	1				**1**		1		**4**		3							
RNF138_RNF125	Dup_<1	CDS-CDS			1		**1**			1	**4**	1		3						
RPH3AL_NXN	Dup_<1	CDS-CDS	1				**1**	1			**4**	3		1						
C5orf42_NUP155	Dup_<1	CDS-CDS			1		**1**	1			**4**	3	1							
YARS_S100PBP	Inv_<1	CDS-CDS			2		**1**	1			**3**	3								
*BCL7A_PSMD9	Dup_<1	CDS-CDS					**1**	1			**3**	3								
C9orf86_KIAA1984	Dup_<1	CDS-CDS	1		1	1	**1**	1			**3**	2	1							
* MAP2K1_DIS3L	Dup_<1	CDS-UTR-5	1		1		**1**	1			**2**	1								
CUL5_RAB39A	Dup_<1	CDS-CDS	1		1		**1**	1			**2**	1	1							
* SYK_AUH	Inv_<1	CDS-CDS					**1**	1			**2**	1		1						
DLG1_LOC220729	Dup_<1	CDS-ncRNA			1	1	**1**	1			**2**	1								

(a)

Table 4.1: List of 63 polymorphic fusion genes discovered in this study and their incidence in TCGA and HapMap cohorts.
"Type" - type of structural rearrangement predicted to generate this fusion . Inv - inversion, Dup - duplication, Del - deletion, Tra - translocation. SVs are represented as <1 and >1 corresponding to whether the genes are within or further than a megabase, respectively.
"BreakType" - the breakpoint location within the upstream and downstream genes. CDS - coding sequence, UTR-5 - 5' untranslated region, UTR-3 - 3' untranslated region, ncRNA - if the gene is an annotated non-coding RNA.
"InFrame" - if both the CDS sequences from the 5' and 3' genes are in the annotated open-reading frame. Continued...

							GTEx				TCGA				HapMap LCLs					
Fusion	Type	BreakType	InFrame?	GTEx CNV	DGV CNV	WGS SV	GTEx Total	White (n=443)	AA (n=71)	Asian (n=6)	TCGA-total	White (n=7164)	AA (n=810)	Asian (n=622)	LCLs-total	CEU (n=92)	FIN (n=95)	GBR (n=95)	TSI (n=93)	YRI (n=89)
DLG1_LOC220729	Dup_<1	CDS-ncRNA			1	1	1	1			2	1								
PABPC1L_YWHAB	Dup_<1	CDS-UTR-5	1		1		1	1			2	2								
ADNP2_RBFA	Dup_<1	CDS-CDS			1	1	1		1		2		2							
*HIP1_CCL26	Dup_<1	CDS-UTR-5	1		1	1	1	1			2	1		1						
* AUH_SYK	Inv_<1	CDS-UTR-5	1				1	1			1	1								
P4HTM_ARIH2	Dup_<1	CDS-CDS	1	1		1	1	1			1		1							
TTBK2_UBR1	Dup_<1	UTR-5-CDS			1		1	1			1		1							
ARMCX6_ARMCX4	Inv_<1	UTR-5-UTR-3				1	1	1			1	1								
PER2_HDAC4	Dup_<1	CDS-CDS					1		1		1	1								
SLC29A3_UNC5B	Dup_<1	CDS-CDS		1	1		1	1			1	1								
C9orf86_TMEM141	Dup_<1	CDS-CDS	1			1	1	1			1	1								
S100PBP_YARS	Inv_<1	CDS-CDS			2		1	1			1		1							
DOCK8_AY343892	Dup_<1	CDS-ncRNA					1	1			1	1								
FSCN1_ZNF815P	Del_<1	CDS-CDS	1			1	1	1												
GPR35_RNPEPL1	Dup_<1	CDS-CDS			2		1	1												
RBM10_SSX5	Inv_<1	CDS-UTR-3					1	1												
CHD5_KCNAB2	Inv_<1	CDS-UTR-5	1		1		1	1												
ACSS1_NINL	Dup_<1	CDS-CDS	1	1		1	1	1												
CASP4_CARD18	Dup_<1	CDS-CDS	1				1	1												
NMNAT1_CTNNBIP1	Inv_<1	CDS-UTR-5	1				1	1												
AMMECR1_CHRDL1	Dup_<1	CDS-CDS					1	1												
MARK3_MKNK2	Trans	CDS-UTR-5	1				1		1											
* AY070437_MLF1	Dup_<1	UTR-5-UTR-5	1				1	1												
UNC5A_EIF4E1B	Dup_<1	CDS-UTR-5	1				1	1												
* P2RY8_PLXNA3	Inv_>1	UTR-5-UTR-5	1				1	1												
WIZ_ARMC6	Inv_>1	CDS-CDS	1			1	1	1												
TMEM57_SYF2	Inv_<1	CDS-CDS				1	1	1												
SARDH_VAV2	Dup_<1	CDS-CDS	1	1	1		1	1												
CD109_MTO1	Dup_<1	CDS-CDS	1				1	1												
MAEA_FAM9B	Trans	CDS-UTR-5	1				1	1												
ZNFX1-AS1_CSE1L	Dup_<1	ncRNA-UTR-5	1	1		0	1	1												

(b)

Table 4.1: Continued....

"GTEx CNV" - For a subset of donors with copy number variant data available, we determined if there is CNV that supports the fusion gene. If 'GTEx CNV' is 1, then a single CNV is found that supports both the genes in the fusion.

"DGV CNV" - Similar to GTEx CNV, this indicates whether a previously reported SV (in the database of genomic variants) supports the fusion gene.

"WGS SV" - Indicates whether the fusion is supported by a structural variant identified within the donors identified with the fusion. See Supplementary Table 4.6 for information on SVs supporting the fusion. 1: Yes, 0: no, ' ': no WGS data available for analysis.

"GTEx", "TCGA" and "HapMap LCLs" - number of individuals detected with the corresponding fusion in each of the three cohorts. Individuals are stratified by race. Numbers in parenthesis in the header indicate total number of individuals.

* - fusions involving genes previously associated with cancer (COSMIC)

functional role of if this fusion transcript is a consequence of mulitple rearrangements in this locus.

Similar to the *KANSL1* locus, 5q13.2 contains mutliple segmental duplications involving genes such as *NAIP* (NLR family apoptosis inhibitory protein) and *OCLN* (occludin) (Courseaux et al. 2003). Mulitple isoforms of *NAIP-OCLN* fusion transcripts also have been reported. However, here we find one of the isoforms that fuses the CDS of *NAIP* to 5' untranslated region (UTR-5) of *OCLN* at 8 fold higher frequency in AA compared to both EA or Asian. The fusion can generate a truncated NAIP with just the nucleoside-triphosphatase (NACHT) domain or a full length OCLN protein.

We find another previously reported fusion *TFG-GPR128* (3q12.2) at 6-fold higher frequency in EA compared both AA and Asian (Figure 4.2). Predicted to be generated by a tandem duplication event, this fusion fuses the Phox and Bem1 (PB1) domain of TFG to G-protein coupled receptor domain of GPR128 (Figure 4.11A, see below). Two out of three remaining population enriched fusion genes are found exclusively in AA donors (*PDE1C-DNAJC6*) or at higher frequency in AA (*TRPM4-PPFIA3*) (Table 4.1). *TRPM4-PPFIA3* is seen in six AA and one EA individual across both GTEx and TCGA. The resulting in-frame fusion transcript is predicted to retain first 4 exons of transient receptor potential cation channel, subfamily M (*TRPM4*) and last 7 exons of Protein tyrosine phosphatase, receptor type, F polypeptide, alpha 3 (*PPFIA3*). The fusion protein loses the catalytic domain of *TRPM4* but retains the three protein-protein interaction sterile alpha motifs (SAM). Interestingly, we find *PARG-BMS1* at significantly higher frequency of 4.9% in Asians compared to 0.1% in EA or 0.9% in AA. In 31 Asians, and, 23 EA and AA donors across both GTEx and TCGA, we find the A-domain of poly ADP-Ribose glycohydrolase (*PARG*) fused to the SAM motifs of *BMS1*. Although not statistically significant, we and others observed the low frequency fusion gene *MAPKAPK5-ACAD10* exclusively in the EA population (n=31) suggesting another candidate that is enriched in EA (Holt et al. 2012). We note that the differences in allele frequencies for fusions between the cohorts is primarily driven by dif-

ferences in sensitivty of ascertaining the fusion event and under-powers estimation in an unbiased fashion across the populations. Our finding that 9 out of 11 top recurrent fusions show either a statistically significant population specific enrichment (n=6) or are exclusively detected in one population but at rare enough frequencies to evade statistical significance (n=3) opens the possibility that there may be selection pressures on these loci.

4.3.4 Functional characteristics of fusion genes

Functional consequences of a fusion gene depends upon the domains retained/lost, its protein coding potential and its expression level. 57% (36/63) of fusions are predicted to generate an in-frame transcript (Figure 4.3A). Out-of-frame fusions may result in haploinsufficiency of one or both of the genes, or alternately, can translate a truncated protein. Determining functional significance remains challenging due to the multitude of ways in which a novel fusion can affect cells. Interestingly, we found an enrichment for COSMIC cancer associated genes among the genes involved in fusion genes (10/60, p <6e-7, three reciprocal events are not included). However, we did not find any of these fusions at higher frequency in TCGA than expected by chance (possibly under-powered by potential false negatives in TCGA), as might be expected if they were inherited risk factors/alleles. But, it is not necessary that these 10 fusions are involved in initiation or progression of cancer, but rather can potentially contribute to an array of phenotypes that manifest in alteration of cell cycle pathways, a hallmark of cancer. Among the 10 cancer gene fusions, we find those involving *HIP1*, *CHEK2*, *BCL7A*, *IDH2*, *TFG*, etc involved in the fusions here (highlighted in Table 4.1). In one donor in GTEx and two individuals in TCGA we find Huntington Interacting Protein 1 (*HIP1*) fused to chemokine (C-C motif) ligand 26 (*CCL26*). Transforming activity of HIP1 has been implicated in brain, prostate and lymphoid malignancies (Carbone et al. 2008; Bradley et al. 2005; Bradley et al. 2007a; Bradley et al. 2007b). Potential tumor suppressor Checkpoint kinase 2 (*CHEK2*) (Hirao et al. 2002) is found fused in 8 individuals to the upstream Tetratricopeptide Repeat Domain 28 (*TTC28*). Interestingly,

the fusion involves second exon of *CHEK2* and the retention of its entire kinase domain in the final transcript. We also found the fusion-positive donor to express this gene at the 99th percentile in most tissues (Figure 4.10). Another fusion involving B-Cell CLL/Lymphoma 7A (*BCL7A*), a candidate tumor suppressor gene in lymphoma (Carbone et al. 2008), results in a loss of 85% of BCL7A protein. Although each fusion genes requires careful evaluation to hypothesize its potential biological significance, these observations highlighted here suggest that a substantial proportion of the fusions identified here may have a phenotypic effect.

Fusion events that result in the fusion of an upstream gene to the 5' untranslated region (UTR-5) of the downstream gene comprise the most prevalent class of protein coding fusions (n=17). Fusions of this category can significantly alter the expression of the downstream gene. For example, in one donor, we find that the fusion of *MAEA* (macrophage erythroblast attacher) to the UTR-5 of *FAM9B* (family with sequence similarity 9, member B) resulted in >14 s.d. increase in expression level of the downstream gene (Figures 4.3A, 4.8). In another example, we find *UNC5A*, a member of netrin-1 family of receptors fused to the UTR-5 of *EIF4E1B* (eukaryotic translation initiation factor 4E family member 1B) resulting in >12 s.d. increase in expression of the *EIF4E1B*. In one donor, we find a truncated MARK3 (MAP/Microtubule Affinity-Regulating Kinase 3) without its KA-1 domain fused to the UTR-5 of *MKNK2* (MAP Kinase Interacting Serine/Threonine Kinase 2). This fusion transcript is predicted to retain translational start sites from both the genes resulting in expression of either truncated MARK3 or full length MKNK2. While KA-1 domain of MARK3 is required to maintain the specificity of the kinase to bind to regulators only at the plasma membrane, increased expression of *MKNK2* can induce phosphorylation of eIF4E (eukaryotic translational initiation factor 4E).

4.3.5 Ectopic expression of 3' genes of polymorphic fusion transcripts

In addition to the 17 UTR-5 fusions, we identify 9 additional CDS-CDS in-frame fusions that result in retention of >90% of the downstream gene (Figure 4.3A). Hypothesizing a

potential for ectopic expression of the downstream gene due to the fusion event that resulted in its regulatory region swapped with its upstream partner, we constructed tissue-specific expression profiles for each of the genes involved in these 26 fusions. A gene is designated as expressed in the given tissue if its median RPKM >=1 across all donors within the cohort. Using this criteria, we identified 20 out of 26 fusion genes for which the number of tissues in which the wild-type expression of the upstream and downstream genes differs by five or more (Figure 4.3B). For each of the 20 fusion genes, we find a median of 3.5 tissues where the individual with the fusion showed the highest expression for the downstream gene compared to the entire cohort (Figure 4.3B). Given that only a median of 20 tissues were sequenced for the fusion-positive donors, the number of tissues showing mis-regulation could be an underestimate.

In the most striking example, the known fusion gene *TFG-GPR128* results in ubiquitous expression of the G-protein coupled receptor domain (GPS) of *GPR128* (Figure 4.11). The biological role of *GPR128* is poorly understood. Apart from the GPS domain, no other known domains are found in this 797 amino acid peptide (Fredriksson et al. 2003). Deletion of *GPR128* in mice showed reduced body weight and increased intestinal contraction frequency (Ni et al. 2014). Although the potential biological function remains unclear, the ubiquitous expression pattern suggests a biological function for this fusion. In another example, the *MAEA-FAM9B* fusion of a ubiquitously expressed gene *MAEA* to the full length *FAM9B*, a tissue-specific gene expressed in testis, uterus and liver (Figure 4.12). FAM9 family of genes have been reported to be localized to the nucleus and play a role in meiotic recombination in testis (Martinez-Garay et al. 2002). The exact function of all three FAM9 family of genes (A/B/C) remain unknown. In another interesting case, we find the entire *CARD18*, a caspase-1 inhibitor, under the regulation of *CASP4* (caspase-4) gene. Caspase-4 is ubiquitously expressed across tissues with predominant expression in blood and plays a role in inflammation response (Figure 4.13). In tissues with this fusion, the expression of *CASP4* is not abrogated but we find *CARD18* expressed ectopically in 21 tissues including

blood. *CARD18* binds to caspase-1 and subsequently blocks it from proteolyically processing IL-1 to its active form (Humke et al. 2000). Interestingly, *CASP4* is also shown to play a critical role in interacting with *CASP1* to convert the inactive IL-1 precursor to its active state (Sollberger et al. 2012). Overall, our findings suggest that ectopic expression may be a frequent mechanism with which polymorphic fusion genes induce functional effects.

4.4 Discussion

Fusion genes have long been considered a hallmark of cancer genomes in somatic cells. Despite substantial evidence for structural variation that is emerging from large scale population genetics studies (Sudmant et al. 2015a; Sudmant et al. 2015b), the evidence for polymorphic fusion genes segregating in human lineage remained sparse. This is primarily attributable to the technical limitations in identifying them from genome sequencing based approaches. Here, we used a transcriptome guided approach to identify expressed fusion genes in 524 individuals in the GTEx project. By considering the multiple tissue transcriptomes of each donor as biological replicates, we developed heuristic filters to identify fusion genes with high specificity. Among the donors with WGS data available, we successfully found a supporting SV for 22 out of 24 fusion genes detected within those donors demonstrating the high specificity of our approach (Supplementary Table 4.6). In all, we identified 63 inherited fusion genes (59 novel) with at least one fusion detected in 44% (230/524) of the donors (Table 4.1). We find that 70% of the fusion genes are also detected in at least one individual in TCGA and HapMap cohorts suggesting that a significant proportion are polymorphic events.

Although a majority (87%) of the fusion genes identified here are predicted to be generated by intra-chromosomal rearrangements involving genes <1 megabase apart, we also find events implicating complex rearrangements such as one gene fused to multiple proximal partners within the same donor. Interestingly, such complex events have been previously reported in cancer and associated with regions of multiple copy number amplifications during

the life cycle of the tumor (Kangaspeska et al. 2012). Haplotype reconstruction of the loci is required to accurately determine the structure of genes following these complex events including translocations. In addition, identifying the breakpoint sequence will allow for investigation of the types of DNA repair mechanisms that generated them.

Multiple loci under strong positive selection have been recently identified from genome-wide scans of human populations (Pickrell et al. 2009; Perry et al. 2007; Bersaglieri et al. 2004). Here, we sought to investigate if any of the fusions show population specific enrichment. We note that this analysis may be under-powered due to the overall low frequency of the fusions combined with small sample sizes in GTEx, and, false negatives due to single tissue sequencing in TCGA/HapMap cohorts. Despite this, we identify six fusion genes segregating differentially between Europeans (EA), African Americans (AA) and Asians. For example, we find *KANSL1-ARL17A* and *TFG-GPR128* are observed at more than six-fold higher frequency each in EA than AA. *KANSL1/ARL17* genes are within a loci (17q21.31) that is previously reported to be under strong positive selection in EA (Stefansson et al. 2005). We also find fusions such as *NAIP-OCLN* and *PARG-BMS1* enriched in AA and Asian populations, respectively. In addition to the six fusions that reached statistical significance for population specific enrichment (Figure 4.2), we find three other recurrent fusions, *SPSB1-H6PD*, *TFDP2-XRN1* and *MAPKAP5-ACAD10* that are detected exclusively in EA suggesting that larger samples are required to determine their enrichment. However, our observation that six out of eleven most recurrent fusions are segregating differentially across populations suggests that these loci may be under recent positive selection – a hypothesis that remains to be evaluated.

Phenotypic consequence of a fusion depends on on a multitude of factors including whether the resulting transcript is in-frame or out-of-frame. In-frame events can generate novel chimeric proteins or simply result in dsyregulation of one or both of the partner genes. Out-of-frame events can be loss of function events resulting in haplo-insufficiency of one or both partner genes. 36/63 fusions reported here are predicted to be in-frame. Among

these, we find 26 fusions with potential to induce ectopic expression of the downstream partner gene. Interesting among these is the EA-enriched *TFG-GPR128* fusion that results in ubiquitous expression of a highly tissue specific G-protein signaling factor (Figure 4.11). The functional consequences of this ubiquitous expression of the G-protein signaling protein remain to be evaluated. Another striking example is the ubiquitous expression of testis-specific *FAM9B* gene in one *MAEA-FAM9B* fusion-positive donor (Figure 4.12). Ectopic expression has been previously reported as a mechanism to induce dominant phenotypes in various species (Cerwenka et al. 2001; Bender et al. 2000; Brand et al. 1993; Hinds et al. 1992). Overall, our observation that more than one-third of fusion genes result in ectopic expression suggests that these events may have the potential to introduce novel phenotypic changes.

In summary, we identified a widespread prevalence of a novel class of genetic variation that can introduce novel proteins as well as induce striking alterations in tissue specific expression profiles of the genes. The findings reported here open new research possibilities in population genetics as well as in phenotype association, both of which can only begin with the identification of haplotypes that contain these fusions. Cloning of the breakpoint regions of the fusions and subsequent identification of the genotypes on the fusion containing haplotype will enable investigation into evolutionary selection pressures as well as potential associations with GWAS phenotypes.

4.5 Methods

4.5.1 RNA and DNA sequencing data

Short read archive (.sra) files corresponding to 9,126 tissues and whole genome sequencing .sra files for 21 donors were downloaded from NCBI dbGaP (phs000424.v4). .sra to .fastq conversion was done using NCBI's SraToolkit. GTEx RNA-seq and DNA-seq reads are 2x76 and 2x150, respectively. RNA-seqs for 464 HapMap LCL cell lines was down-

loaded from NCBI's short read archive. See Table S3.5 in the attachment to this thesis: SupplementaryTables.Chapter3.xlsx.

4.5.2 Gene fusion detection from RNA-seq

Fusion discovery was performed using Minimum Overlap Junction Optimizer (MOJO) in the highest sensitivity mode requiring two discordant reads and one supporting read. Prior to applying filters described in Figure 4.1B, we applied five filters to exclude potential technical artifacts that can manifest from pan-cohort fusion calling. First, to increase confidence level in a fusion call, we pooled read evidence for fusion junctions (exon-exon junctions between two genes) that are recurrently detected across multiple tissues (within and across donors). For a fusion gene to be nominated, at least one fusion junction between the two genes is required to be supported by at least 3 unique anchor reads (each with a minimum anchor length of 15bps). Each of the three supporting anchor reads's starting and ending alignment positions are required to be spaced by at least 3bps. For example, for a read length of 50bp, each junction-read (junction mapping end of the anchor read) is required to have at least 15bp mapping to 5' exon (left_overhang) and 35bp to the 3' exon (right_overhang) of the fusion. And, each of the, at least 3, junction-reads are required to have a minimum overhang of at least 15, 18 and 21 bps on both sides of the fusion junction.

Second, we applied a filter that accounts for fusion junctions comprising exonic sequences that cannot be uniquely mapped. That is, if sequences from one of the exons involved in the fusion junction (last 40bps of 5' exon and fist 40bp of 3' exon) maps to multiple regions in the genome, then we require at least two anchor reads with the non-junction mapping ends mapping uniquely to both the genes of the fusion pair to retain this junction. Alternately, if the full 80bp junction sequence aligns to the genome as three or fewer contiguous blocks, then the junction is filtered out. For both steps of this filter (individual 40bp exonic sequences and full 80bp junction sequence), alignments to the genome and transcriptome are performed using Blat.

Third, we next constructed a filter to account for chimeric artifacts from highly expressed genes that are likely to be generated during library preparation. The ratio of anchor reads to discordant reads is used to filter out high expression artifacts. For a given fusion gene f, we determine the median of anchor read to discordant read ratios ($mADratiof$) of all the fusion calls for f. A fusion gene is filtered out if $mADratiof$ <0.01. If either of the two genes are highly expressed (max($mRpkmA$, $mRpkmB$) >100) across the samples containing the fusion, then a fusion gene is filtered out if $mADratiof$ <0.05. $mRpkmA$ and $mRpkmB$ are median of expression values of genes A and B within the samples containing the fusion.

Fourth, we excluded fusions between genes that share >50% homology and are further than 1 megabase apart. Finally, we excluded all fusion genes that are nominated between proximal gene on the same strand and are potentially generated by a read through event. After applying the five filters, we identified 422 fusion genes that are detected in at least three tissues in at least one donor in the entire cohort.

4.5.3 SV detection from WGS

WGS fastq files are aligned to the human genome (hg19) using bwa-mem. Sorted bams are input into LUMPY (v0.2.11) to identify structural rearrangements (Layer et al. 2014). A structural variant is determined to be supporting a fusion event if the rearrangement is predicted to generate a fusion transcript that is consistent with the event detected from the RNA-seq. SVs overlapping the two genes but not supporting the event were not considered (Eg: complex rearrangements in the NAIP-OCLN locus).

4.5.4 Targeted fusion junction search in TCGA/HapMap

For each the 63 fusion genes, we constructed the fusion transcripts comprising the precise fusion junction detected in GTEx. Fusion transcripts comprising all possible isoforms of both genes are considered (if exon involved in the fusion junction is shared by multiple isoforms). We next constructed a bowtie2 index of this fusion transcript library. All samples in TCGA

and HapMap were aligned to the fusion transcript library. Anchor reads with one end mapping to the fusion junction (anchor length >=15bp) and the other end mapping to one of the two genes, are identified. At least two anchor reads are required to nominate a fusion within the sample.

4.5.5 Gene expression quantification

Gene expression for 9,126 tissues was performed using Kallisto (http://arxiv.org/abs/1505.02710). Estimated read count values were extracted from each sample and a matrix of gene x sample was constructed. Across sample normalization was performed using Trimmed Mean of M-values (TMM) method implemented in edgeR (Dillies et al. 2013).

4.6 Appendix: Figures

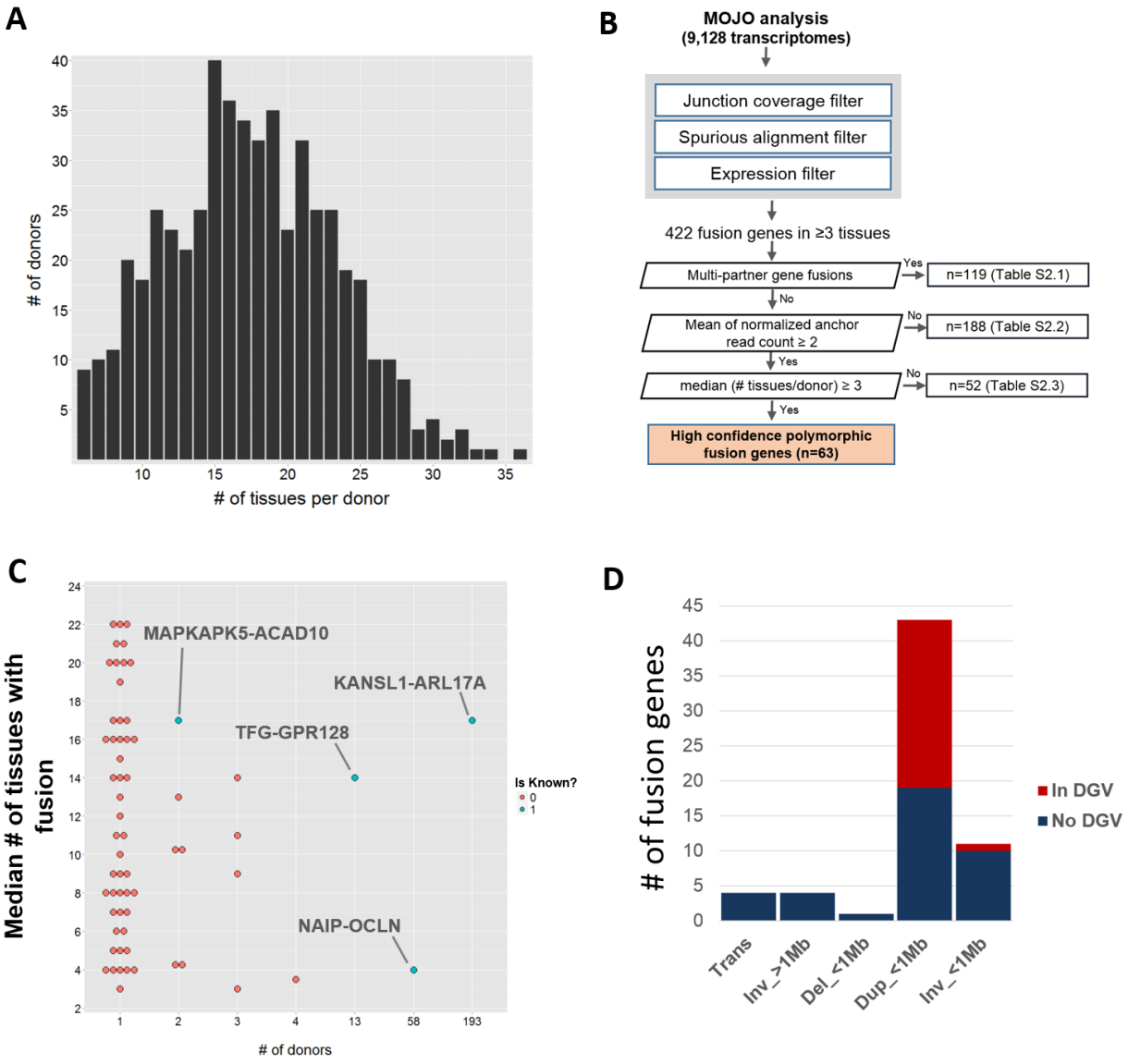
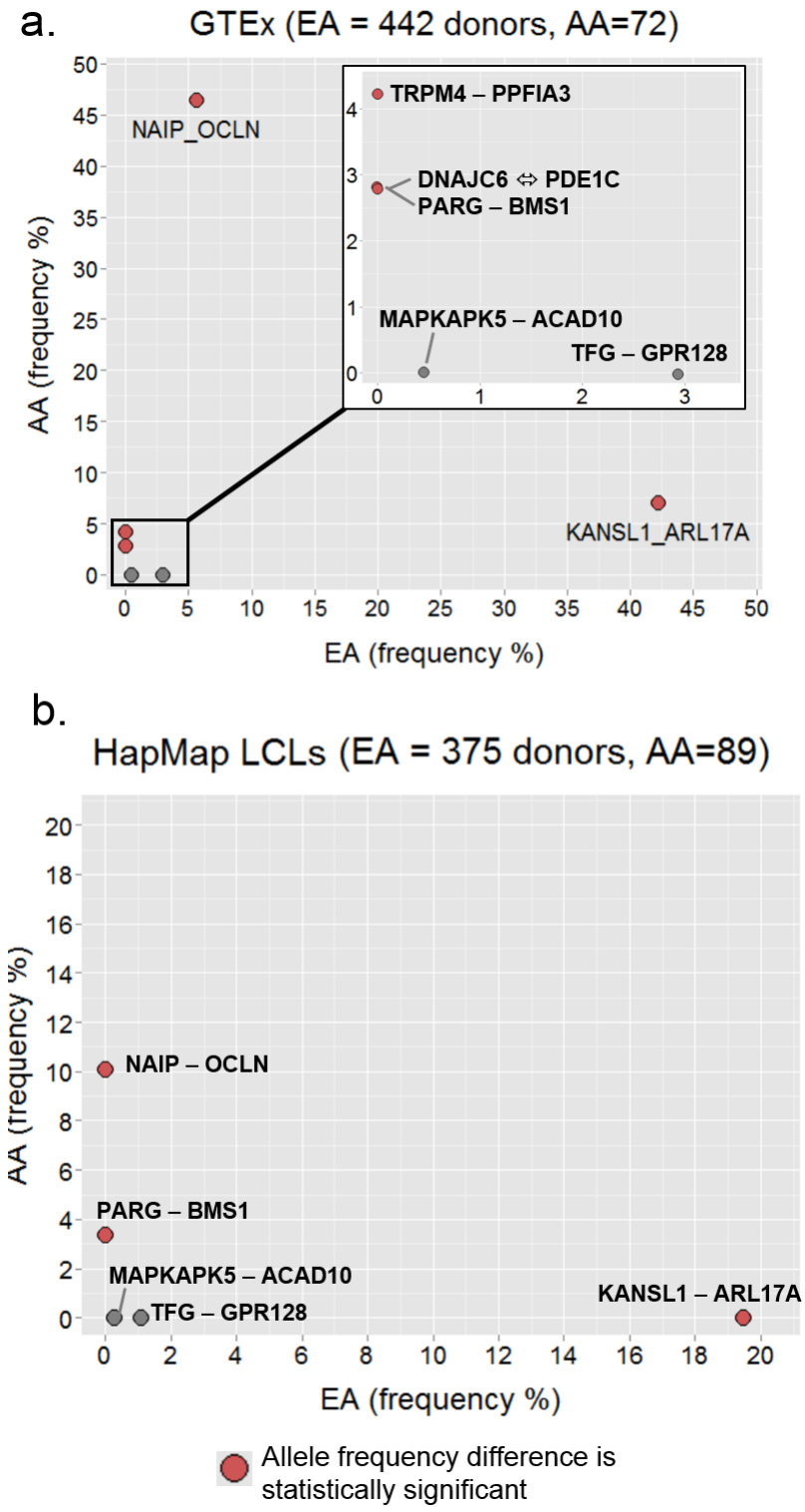


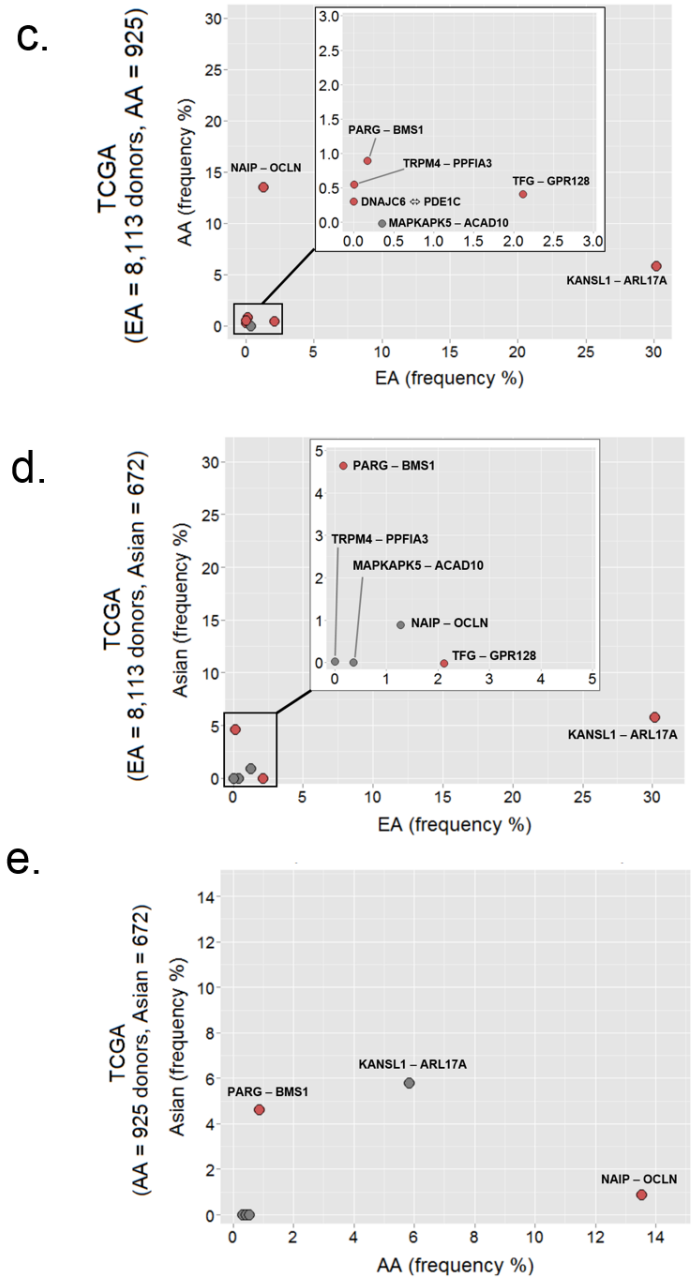
Figure 4.1: Polymorphic fusion genes identified in the GTEx donors

Figure 4.1: (Continued from previous page). Polymorphic fusion genes identified in GTEx donors. (a) Distribution of number of distinct tissues per donor for which transcriptome sequencing data is available is shown. (b) Workflow to identify fusion genes has four primary components. After the fusion calls are nominated using MOJO, a series of filters are applied to remove technical artifacts introduced by ambiguity in alignment or incomplete annotation (see Methods). Next, fusions that are recurrently fused to multiple partners (Supplementary Table 4.2) are excluded along with low expressed fusions (Supplementary Table 4.3) that may not be detected consistently across all tissues. Finally, fusion genes that are detected in a median number of at least three tissues are classified as high confidence polymorphic fusion genes. (c) Frequency distribution of the 63 polymorphic fusion genes shown with the number of tissues in which they are detected in. All four known fusion genes (highlighted in green) are detected in multiple donors and multiple tissues. *NAIP-OCLN* is detected in only a subset of tissues due to the tissue-specific expression pattern of *NAIP* as well as its overall low expression (Figure 4.9). (d) Fusion genes are overlapped with the structural variants annotated in the database of genomic variants (DGV). An SV in DGV is deemed as supporting a fusion gene if each end of the SV falls within the gene bodies of the individual genes involved in the fusion. An SV is required to span both genes to be deemed as corroborating. The fusion genes are are stratified according to the rearrangement that is predicted to generate the event. Trans: translocation, Inv: inversion, Del: deletion, Dup: duplication. Inv/Del/Dup are further stratified into whether the genes are less than 1 megabase apart (<1Mb) or further than 1 megabase apart (>1Mb)



(a)

Figure 4.2: Population-specific enrichment of polymorphic fusion genes



(b)

Figure 4.2: Population-specific enrichment of polymorphic fusion genes within GTEx.

Figure 4.2: (Continued from previous page). Population-specific enrichment of polymorphic fusion genes within GTEx (a), HapMap (b) and TCGA (c) cohorts is shown for seven fusion genes. x- and y-axes show the frequency of the fusion within the two populations compared. A binomial test is used to determine if the frequency differences between the two populations are significant. Fusions showing significant differences are highlighted in red. EA - European Americans, AA - African Americans. *DNAJC6-PDE1C* is a reciprocal event. Although MAPKAPK5-ACAD10 did not reach significance in any of the cohorts, it is shown here to demonstrate that it is a low frequency event detected exclusively in AA population.

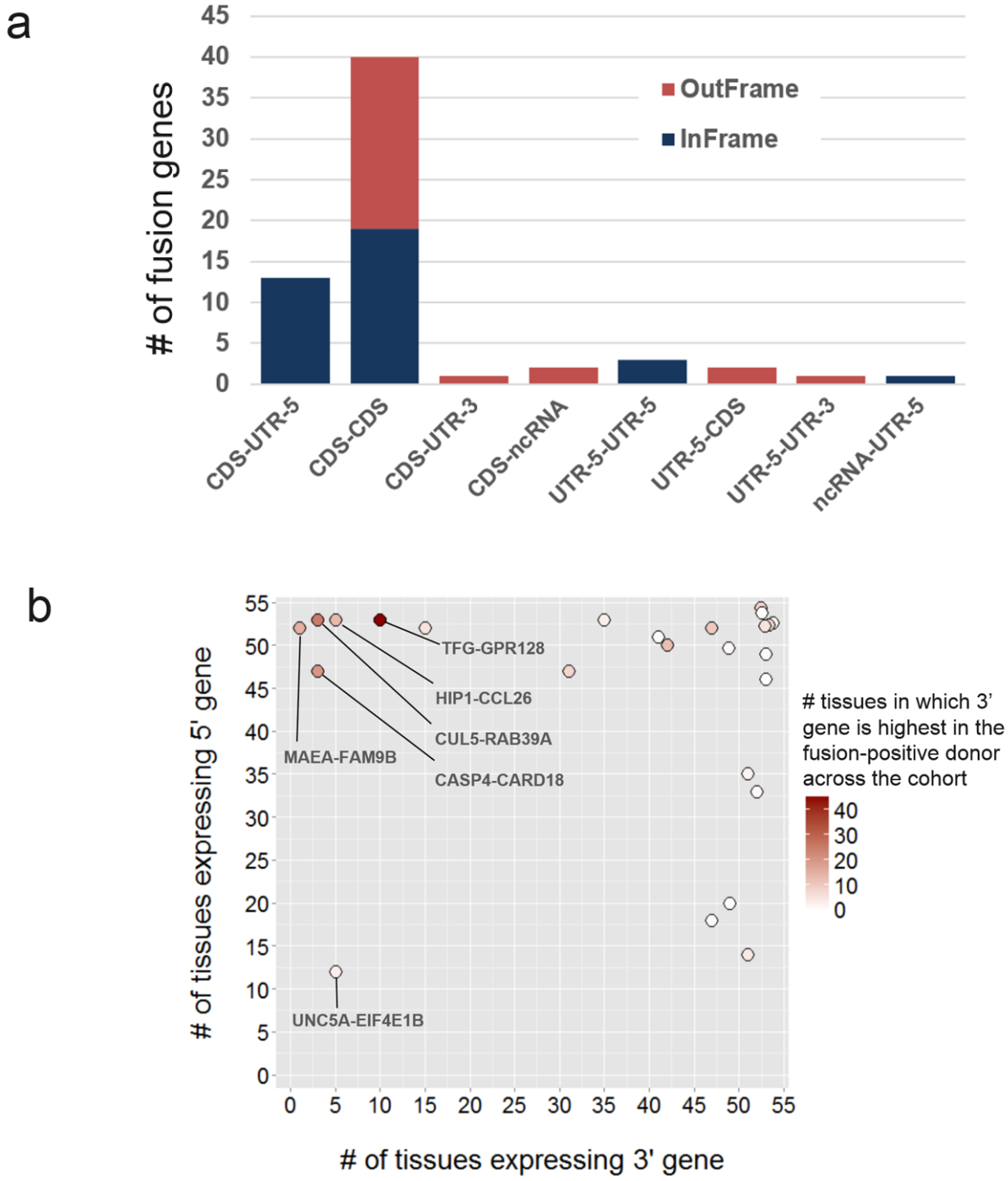


Figure 4.3: Domain and ectopic expression characteristics of polymorphic gene fusions.

Figure 4.3: (Continued from previous page). Domain and ectopic expression characteristics of polymorphic gene fusions. (a) Fusion genes are classified by the regions in which the breakpoints occur in the two genes. (CDS - coding sequence, UTR-5/UTR-3 - 5' and 3' untranslated regions, ncRNA - gene annotated as a non-coding RNA). Fusion gene is designated as InFrame if the entire transcript with regions from both genes is translatable. OutFrame fusions are those in which the sequence contributed by the downstream gene is out of frame. OutFrame fusion can still generated C-terminal truncated peptides of the upstream partner gene. (b) The degree of ectopic expression of the 3' partner gene induced by the fusion event is shown here. x- and y-axes show the number of tissues in which the wildtype 5' and 3' genes are expressed, respectively (a gene is classified as expressed if its median expression across all donors within the tissue is >1 RPKM). Gradient color coding of each data point indicates the number of tissues in which the the 3' gene is expressed at the highest level among the entire cohort within the respective tissue type. For example, in case of *TFG-GPR128*, the normal copy of *TFG* is expressed ubiquitously in 53 tissues where as the expression of normal copy of *GPR128* is detected in 10 tissues. However, the fusion results in the highest expression of GPR128 in 42 additional tissues within the donors that tested positive for the fusion.

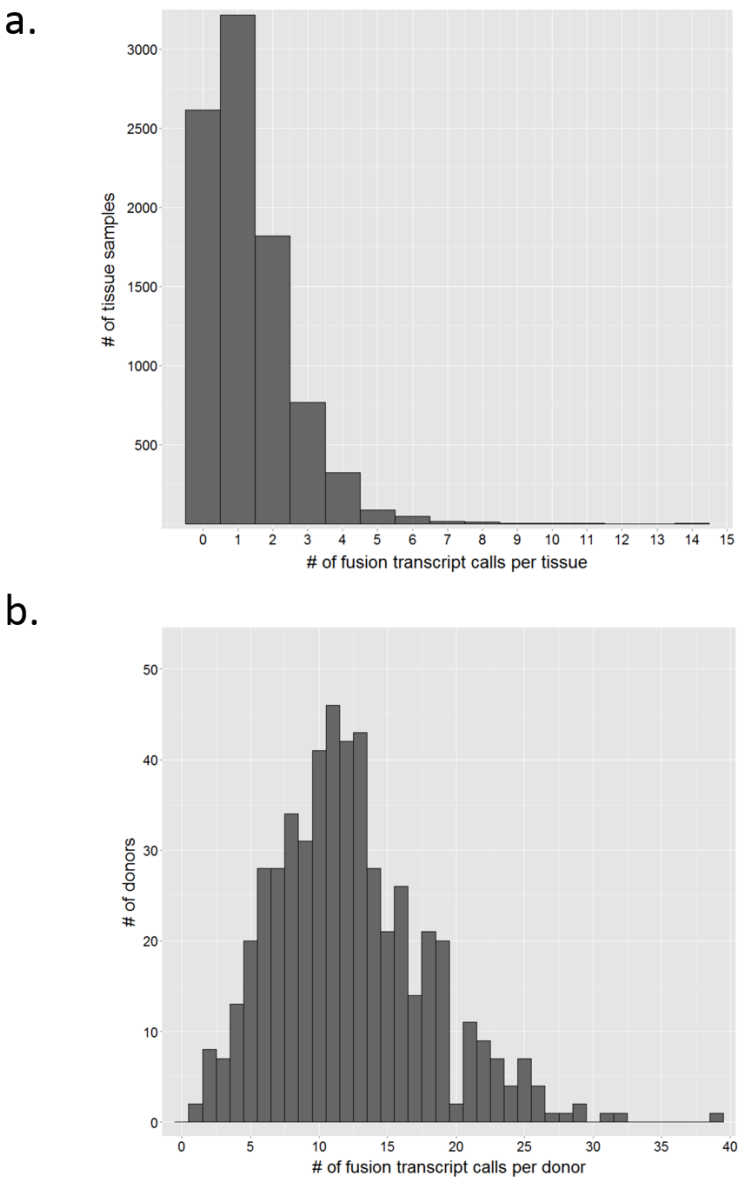


Figure 4.4: Distribution of number of fusion calls per tissue (a) and per donor (b). This demonstrates that our fusion nominations are not driven by outlier tissue samples or donors.

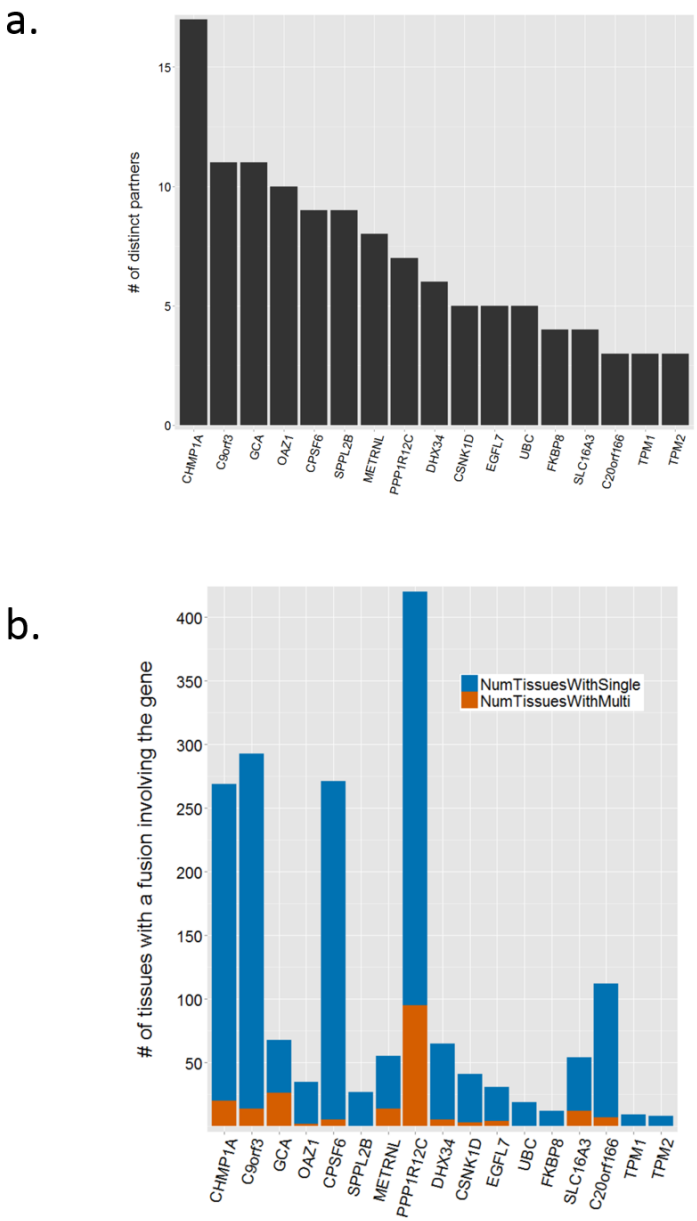
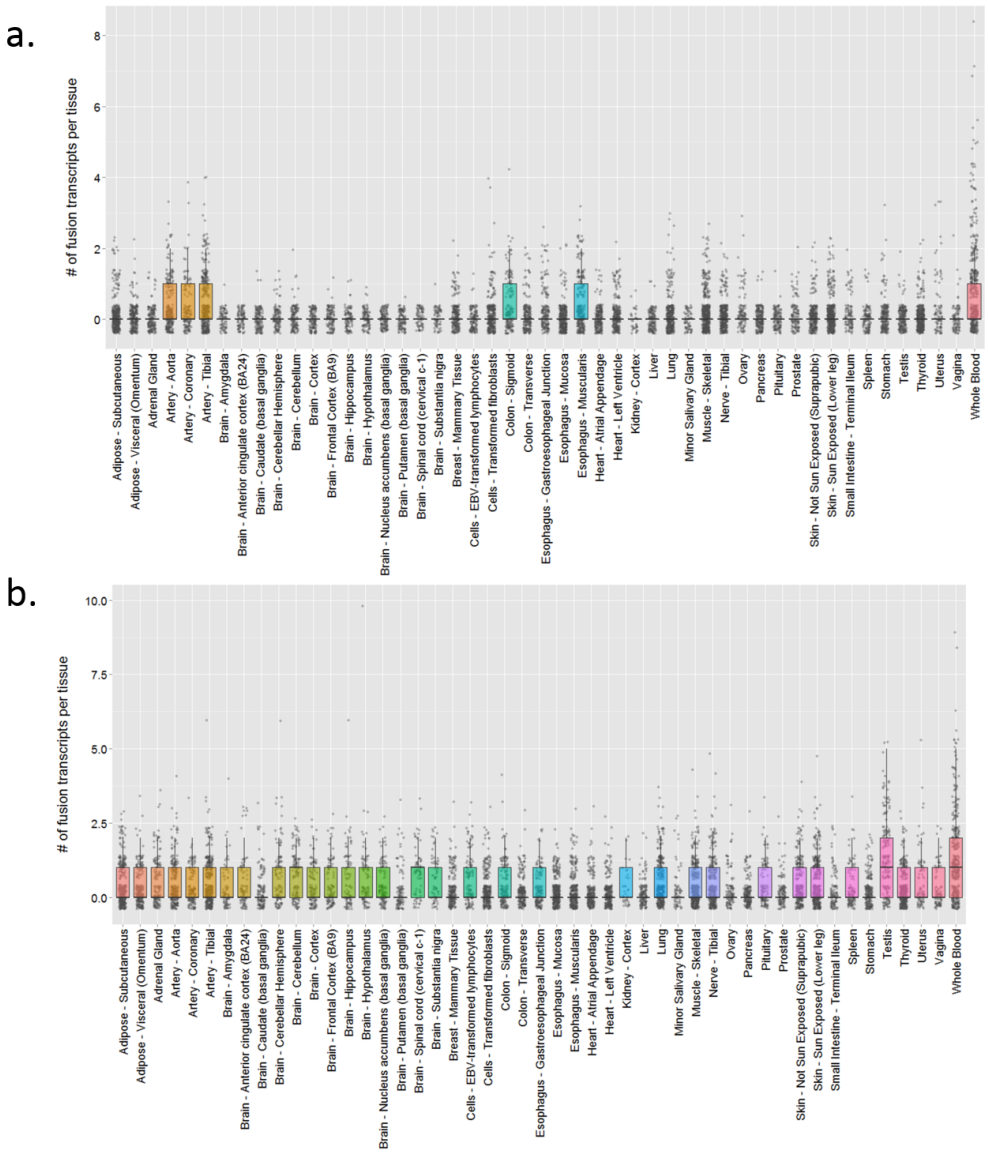
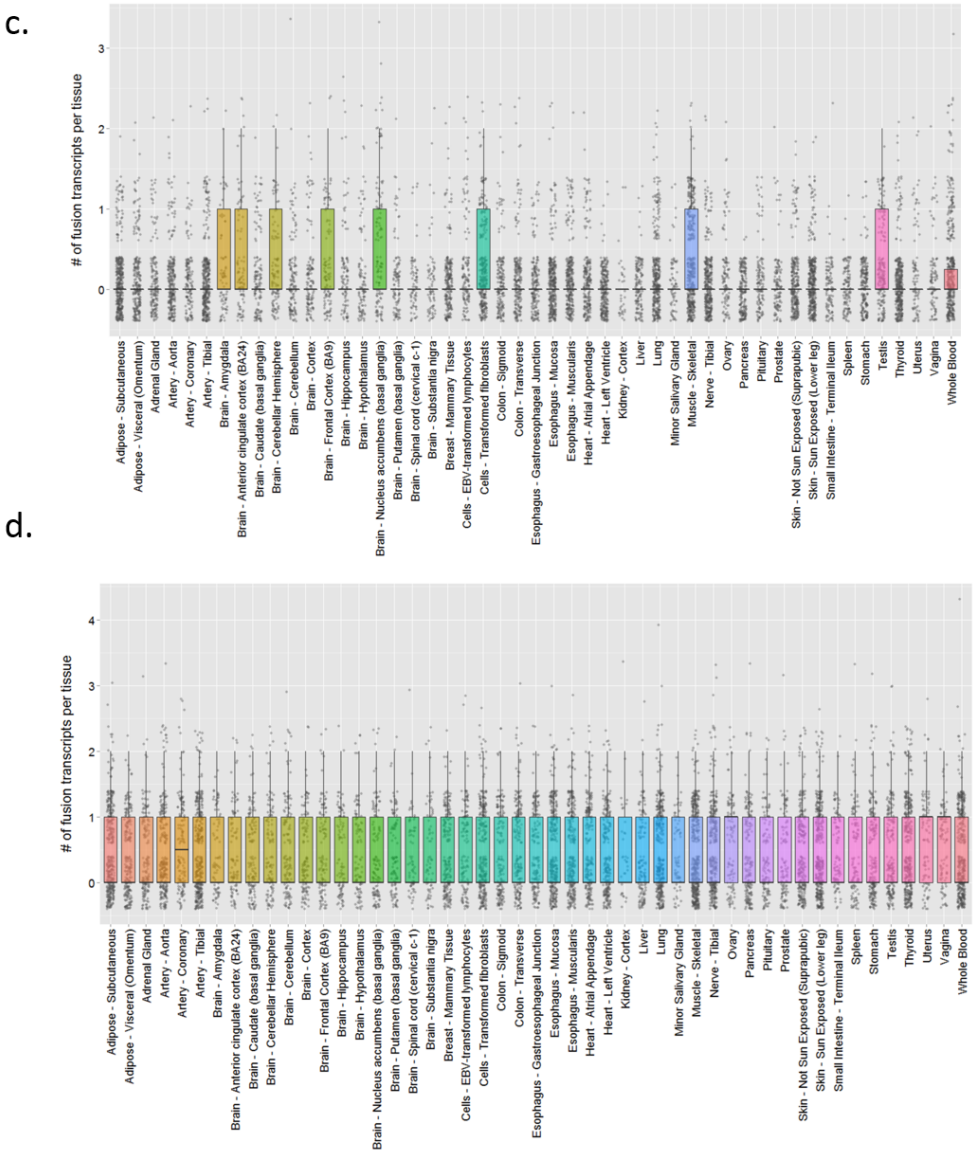


Figure 4.5: Characteristics of multi-partner gene fusions that are filtered. (a) On x-axis are the 17 most recurrently genes identified among the fusion calls. # of donors in which a fusion involving a corresponding multi-partner gene is on y-axis. These 17 comprise 84% of the multi-partner fusions identified. (b) # of distinct tissue transcriptomes (y-axis) in which each of the multi-partner genes (x-axis) is identified is shown. The y-axis is stratified by whether the tissue contains multiple distinct fusions involving the same multi-partner gene (NumTissuesWithMulti) or if only one fusion involving the corresponding gene is found in the tissue (NumTissuesWithSingle). We find that only 14% of tissues containing one of the 117 multi-partner fusion genes contain more than one such fusion gene suggesting that these are not a consequence of outlier tissues. This plot demonstrates that multi-partner tissues are not enriched in a small sub-set of tissues and therefore, are not necessarily driven by technical artifacts.



(a)

Figure 4.6: QC plot showing the distribution of # of fusion calls of different categories across various tissue type



(b)

Figure 4.6: (Continued from previous page). QC plot showing the distribution of # of fusion calls of different categories across various tissue type. Data shown corresponds to Figure 4.1B. On x-axis are the different types and y-axis represents the number of fusion transcripts of a given category that are detected in the corresponding tissue. The different categories of fusions are: (a) multi-partner fusion gene calls that are fusions involving genes found fused to 3 or more distinct partners across the cohort, (b) low expressed fusion gene calls (see main text), (c) low within-donor frequency fusions that are detected in less than a median of three tissues per donor, and, (d) the 63 high confidence polymorphic fusion genes identified in this study

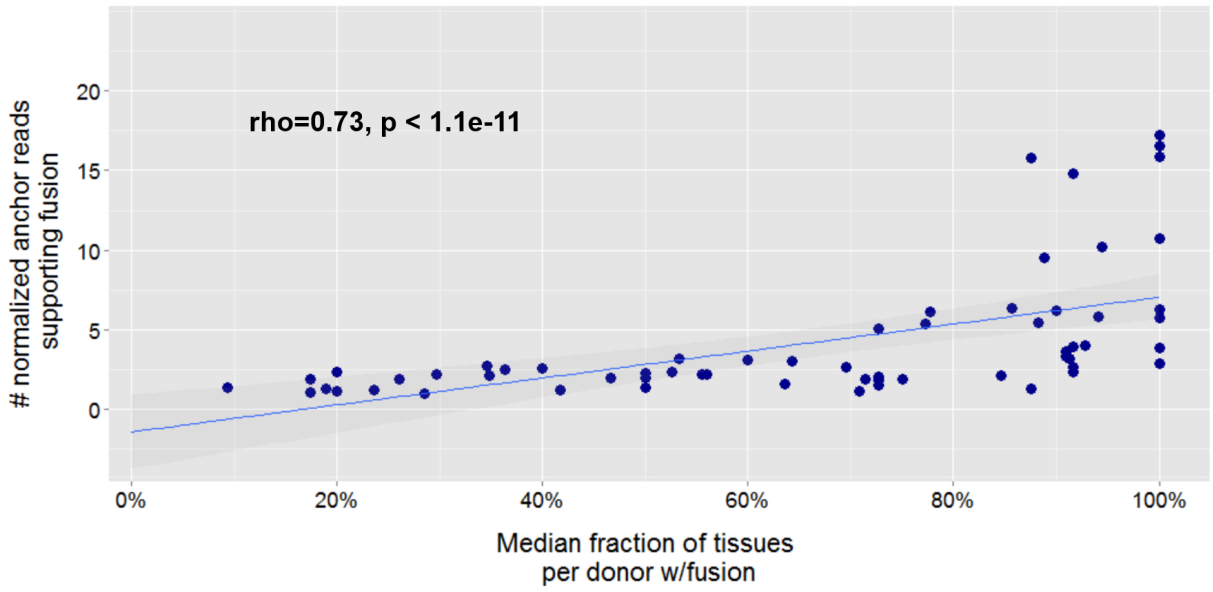


Figure 4.7: Detection of a fusion in all tissues depends on its expression level. On x-axis is the proportion of tissues within a donor in which the fusion is detected. On y-axis is the median number (median across all tissues with the fusion) of normalized anchor reads supporting the fusion. Anchor reads are normalized to sequencing depth to compute the normalized anchor read counts. 54% of the fusion genes are detected in less than 75% of all tissues sequenced for that given donor. Spearman rank correlation of expression level of the fusion gene with the proportion of tissues in which the fusion is identified: rho 0.73, p<1.1e-11

Figure 4.8: Expression of individual partner genes within the donors with the polymorphic fusion genes. Normalized expression values are transformed into Z-score to quantify the deviation of a given gene's expression within a donor from its global mean. For each of the 63 fusion genes, the median z-score of the individual partner genes within the donors in which they are found, is shown. x- and y-axes show the median z-score for 3' and 5' partner genes, respectively. Known fusions are highlighted in red. For 81% of all fusion genes, the genes are expressed within two standard deviations demonstrating that majority of these are not likely to be driven by expression artifacts.

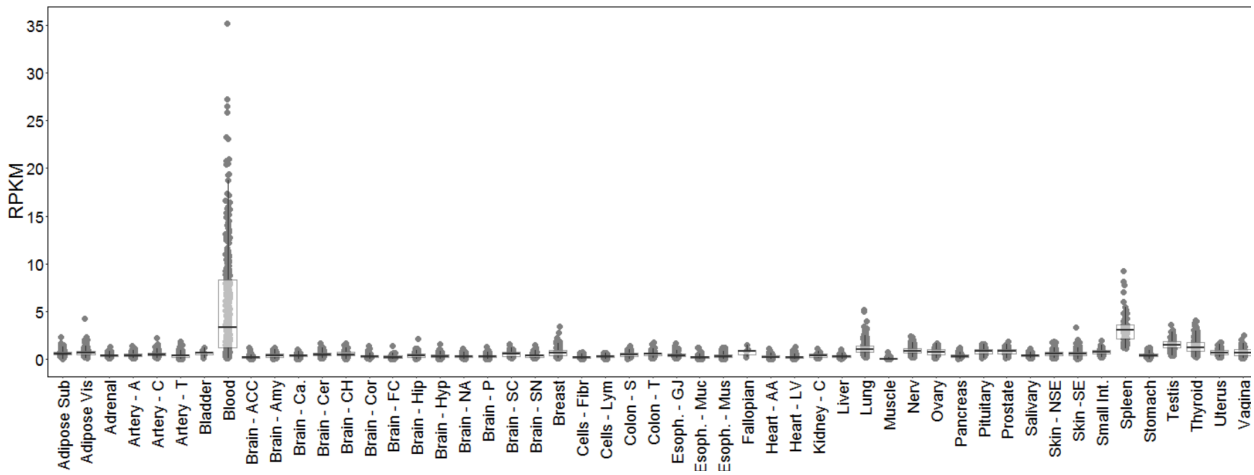


Figure 4.9: Expression profile of wild-type *NAIP* across tissues. Each point corresponds to a donor. Figure illustrates the highly tissue-specific nature of NAIP gene.

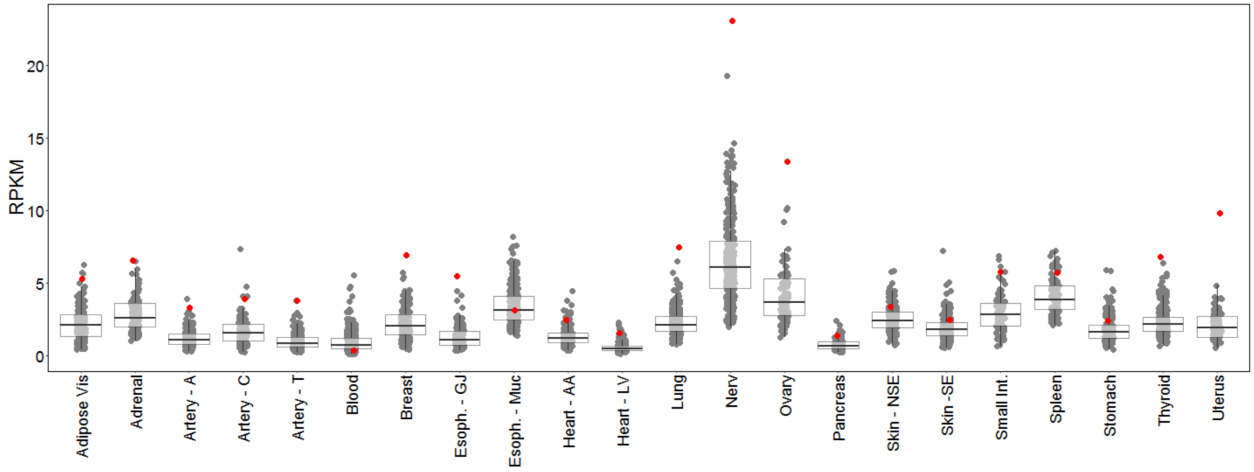


Figure 4.10: Expression profile of *CHEK2* across tissues in fusion positive donors. Expression level of *CHEK2* (checkpoint kinase 2) across all tissues and donors in the GTEx. Fusion-positive samples are shown in red. Only the tissue types in which the fusion is detected is shown.

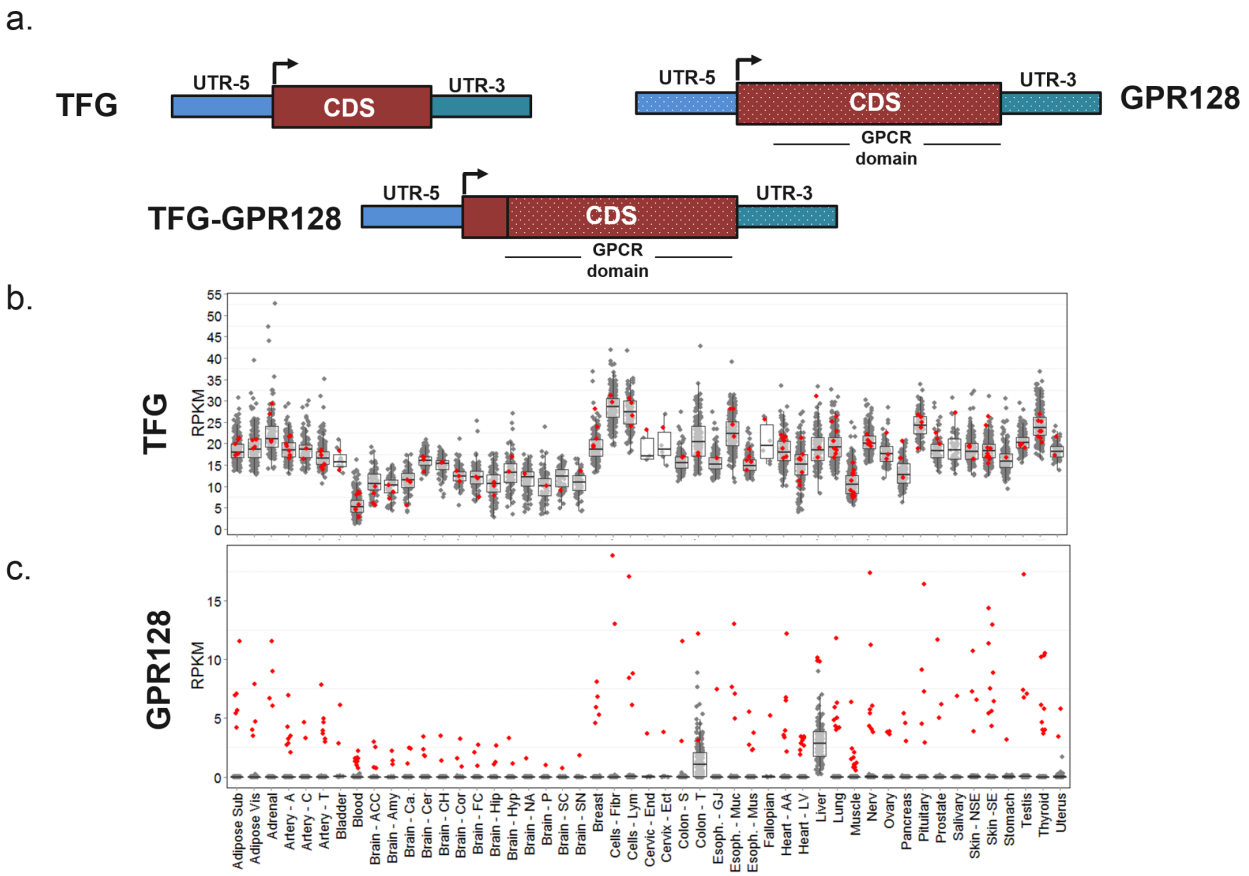


Figure 4.11: Expression profiles of individual genes of *TFG-GPR128* fusion across tissues in genes (a) Schematic of the fusion event. The resulting fusion protein contains the PB1 domain of *TFG* and the full length G-protein coupled receptor domain of *GPR128* (b-c) Expression levels of *TFG* (b) and *GPR128* (c) across all tissues and donors in the GTEx. Fusion-positive samples are shown in red. Only the tissue types in which the fusion is detected is shown.

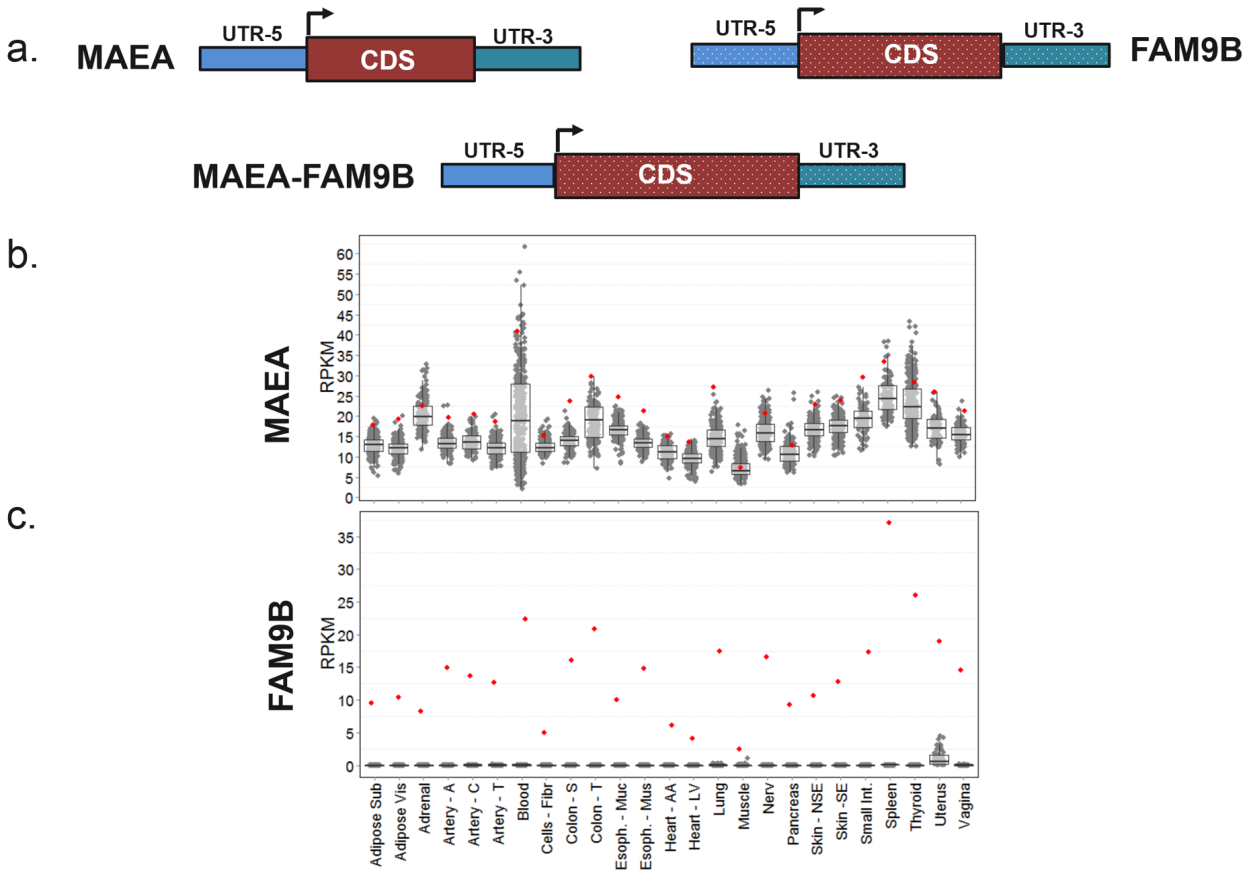


Figure 4.12: Expression profiles of individual genes of *MAEA-FAM9B* fusion across tissues in genes. (a) Schematic of the fusion event showing swapping of the UTR-5 of *MAEA* and *FAM9B*. (b-c) Expression levels of *MAEA* (b) and *FAM9B* (c) across all tissues and donors in the GTEx. Fusion-positive samples are shown in red. Only the tissue types in which the fusion is detected is shown.

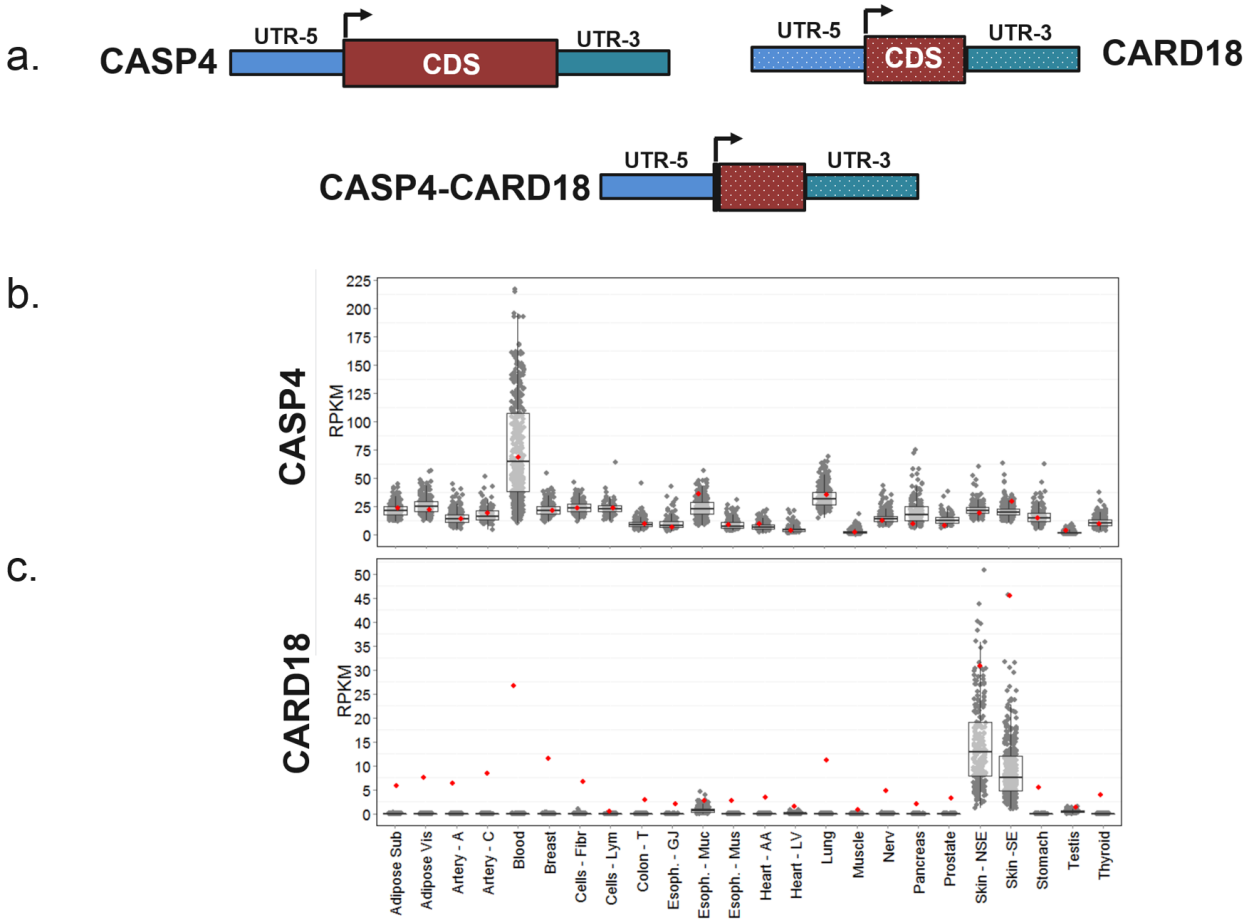


Figure 4.13: Expression profiles of individual genes of *CASP4-CARD18* fusion across tissues (a) Schematic of the fusion event. The fusion swaps the UTR-5 and the first two amino acids of *CARD18* and *CASP4*. (b-c) Expression levels of *CASP4* (b) and *CARD18* (c) (c) across all tissues and donors in the GTEx. Fusion-positive samples are shown in red. Only the tissue types in which the fusion is detected is shown.

4.7 Appendix: Supplementary Tables

Supplementary Table 4.1: See workbook 4.1 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). Manifest of 9,126 GTEx transcriptomes analyzed in this study. Specific columns are:

Column 1: SRA Run ID

Column 2: SRA Sample ID

Column 3: GTEx Donor ID

Column 4: GTEX Tissue ID

Column 5: Body Site ID (specific tissue type)

Column 6: Histological type (organ)

Column 7: Sex

Supplementary Table 4.2: See workbook 4.2 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). Multi-partner fusion genes that are filtered out. Refer to Figure 4.1B and main text. Specific columns are:

Column 1: Tissue type
Column 2: Donor ID
Column 3: Tissue Sample ID
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6-10: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 11-15: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 16: Is fusion in-frame (1: yes, 0: no)

Supplementary Table 4.3: See workbook 4.3 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). Low expressed fusion genes that are filtered out. Refer to Figure 4.1B and main text. Specific columns are:

Column 1: Tissue type
Column 2: Donor ID
Column 3: Tissue Sample ID
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6-10: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 11-15: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 16: Is fusion in-frame (1: yes, 0: no)

Supplementary Table 4.4: See workbook 4.4 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). Low recurrence fusion genes that are filtered out. Refer to Figure 4.1B and main text. Specific columns are:

Column 1: Tissue type
Column 2: Donor ID
Column 3: Tissue Sample ID
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6-10: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 11-15: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 16: Is fusion in-frame (1: yes, 0: no)

Supplementary Table 4.5: See workbook 4.5 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). High confidence polymorphic fusion genes identified in this study. Refer to Figure 4.1B-D and main text. Specific columns are:

Column 1: Tissue type
Column 2: Donor ID
Column 3: Tissue Sample ID
Column 4: # of discordant reads supporting the fusion
Column 5: # of anchor reads supporting the fusion
Column 6-10: 5' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 11-15: 3' gene info (name, chrom, strand, breakpoint, breakpoint region)
Column 16: Is fusion in-frame (1: yes, 0: no)

Supplementary Table 4.6: See workbook 4.6 in Supplementary.Tables.Chapter4.xlsx associated with this dissertation). WGS-based structural variants supporting polymorphic fusion genes. Specific columns are:

Column 1: GTEx Run ID
Column 2: Donor ID
Column 3: Tissue Sample ID
Column 4: 5' gene name
Column 5: 3' gene name
Column 7: SV chromosome
Column 8: SV start position
Column 9: SV end position
Column 10: SV predicted rearrangement type (DUP: duplication, INS: insertion, TRA: translocation, DEL: deletion
Column 11: SV size (in bp)
Column 12: # of discordant reads supporting the SV
Column 13: # of unique discordant reads supporting the SV
Column 14: # of anchor reads supporting the SV

4.8 Contributions

I designed the project with my advisor, Kevin White. I performed the data acquisition, quality control, analysis, compiled the figures and wrote the manuscript with inputs from Kevin. I identified select HapMap and cancer cell lines expressing candidate fusion genes. Vineet Dhiman performed RT-PCR and genomic PCR assays to validate selected fusions. Cancer cell lines were acquired from the Cellular Screening Center. HapMap cell line validations are performed by Vineet in collaboration with the Pharmacogenomics of Anticancer Agents Research (PAAR) group. I would like to thank Barbara Stranger for helpful discussions.

CHAPTER 5

DISCUSSION AND FUTURE DIRECTIONS

5.1 Summary

In this thesis, I have presented a comprehensive survey of fusions in cancer genomes and human populations. Both objectives are enabled by the development of a new algorithm to detect fusions from transcriptome data. Through the analysis of more than nine thousand tumors, I presented a deeper understanding of the complexity and heterogeneity of the fusions across cancers. I discovered several novel recurrent fusion genes that showed features of tumors when introduced into various immortalized tumor cells, highlighting the significance of recurrent fusion genes. In chapter 4, through the analysis of over five hundred healthy individuals, I present the first reported catalog of polymorphic fusion genes in humans. Population specific enrichment of these fusion genes suggested a possible role for recent selection at these loci. Using gene expression analysis, I show that many of the polymorphic fusion genes result in ectopic expression of the genes involved suggesting a strong potential to affect phenotype. In the sections below, I will present key findings of my work in the context of prior work and also discuss the limitations as well as highlight the future directions in pursuing some of the significant questions pertaining to fusion genes in both cancer and germline genomes.

5.2 Gene fusion discovery

5.2.1 Discussion

The central aim of my thesis is to discover novel fusion genes in cancer that play a role in tumor initiation or progression. This pursuit is motivated by technological advances in transcriptome sequencing that enabled discovery of fusions that previously evaded conventional cytogenetic techniques. However, the poor performance of existing methods to accurately

identify fusions from such large cohorts prompted me to develop a new algorithm.

In chapter 2, I present a new algorithm called MOJO (Minimum Overlap Junction Optimizer) that is designed to detect fusions from paired-end transcriptome sequencing. Through extensive comparisons with existing methods using a compendium of 20 cell line and primary tumor transcriptomes, I show that MOJO demonstrates superior sensitivity and specificity. In principle, MOJO is designed to detect fusion junctions (exon-exon junctions) that are supported by only two discordant and one anchor read with a sensitivity of >99%. Specificity is achieved by efficiently implementing a central tenet to MOJO that an anchor read that aligns to the fusion junction cannot align anywhere else in the genome or annotated/unannotated transcriptome. Rigorous filters are implemented to also determine if the anchor read can originate from some of the novel splicing events and other intra-genic rearrangements that can manifest as false positives (Carrara et al. 2013).

Despite substantial improvements in sensitivity and specificity, due to the limitations of RNA-seq based fusion discovery, two main classes of false positives that cannot be accurately modeled still manifest in MOJO's fusion calls. First, fusion transcripts are detected between genes that are abnormally highly expressed. Random chimeras have been previously suggested to be generated during library preparation. However, instead of detecting multiple exon-exon junctions between the highly expressed genes, a phenomenon consistent with random chimeras, only a select few exon-exon junctions are observed, suggesting a possible unknown source of post-transcriptional events generating these transcripts. This type of artifact can be simply mitigated by interpreting the MOJO's fusion call in context of the expression level of the two genes (also provided in MOJO's output). The second source of false positives is the unknown splicing mechanisms that remain to be fully characterized. In addition to the recently reported back-splicing/circularRNAs (Guo et al. 2014), other types of splicing events such as trans-splicing (Li et al. 2008) have been attributed to generate chimeric transcripts but the evidence for this type of event remains to be demonstrated. Such splicing artifacts can manifest as recurrent fusions during the somatic fusion discovery

in a large panel of cancer samples. These can be accounted for by incorporating a control panel of "healthy" tissue (those confirmed by pathology to not have neoplastic transformation) transcriptomes and simply filtering out any somatic fusion call that is also detected in the "healthy" transcriptome. However, "healthy" tissues may also contain a sub-clonal population of transformed cells that are not detected by pathology but identified through transcriptome sequencing. Therefore, caution must be used when applying filters to remove "healthy" tissue fusion calls from a "tumor" tissue. The development of MOJO and the insights I gained into the nuances of transcriptome based fusion discovery allowed me to investigate the prevalence and characteristics of fusions across large number of human cancers in chapter 3.

5.2.2 Future Directions

One of the key challenges in transcriptome based fusion detection is determining the sequencing depth needed to achieve saturation for fusion discovery within the tumor sample. This is a function of the degree of genomic instability within the tumor, its ploidy, purity of the sample (fraction of tumor cells within the tissue sample), the sub-clonality of the fusion gene and lastly, its expression level. Higher degree of genomic instability and ploidy increases the complexity of transcriptome. Similarly, deeper sequencing is required to capture sub-clonal fusions, those from low purity samples or low expressed events. An algorithmic framework that determines the saturation level in lieu of these various parameters can be extremely valuable in controlling for false negatives, especially, in the context of diagnostics.

Quantification of the expression level of the fusion transcript can be extremely useful in determining its overall significance. Currently, we and others have used the anchor reads (normalized to library size) supporting the fusion junction as a proxy for its expression. However, the sequence biases associated with RNA-seq assay may introduce substantial variance is inferring its true expression. For example, GC-biases at the exon-exon junction can affect the sequencing coverage (Benjamini et al. 2012). This can effect read counts when comparing

a fusion gene detected in two distinct tumors but each involving a different exon-exon junction. Recent improvements in alignment-free isoform quantification methods suggest that such an approach may hold promise. These algorithms use an expectation-maximization method to determine the maximum-likelihood estimate for read counts supporting each of the isoforms for a given gene (Patro et al. 2014). Theoretically, such a concept can be applied to fusions by incorporating the fusion gene into the transcriptome along with the various isoforms of the wild-type copies of both the individual genes. However, the applicability and the accuracy of this approach remains to be evaluated.

5.3 Fusion genes in cancer

5.3.1 Discussion

Fusion genes have long held a prominent role in cancer biology. The overwhelming majority of the first reported fusion genes have been found to induce phenotypes resulting in the perception that such events are clear indicators of drivers. As a result, the discovery of a novel fusion gene in routine genomics analyses provokes strong interest in finding its association to cancer. In chapter 3, through the analysis of nearly ten thousand tumors, we find that fusions are as pervasive as point mutations albeit at substantially lower frequencies, suggesting that some of driver/passenger characteristics associated with point mutations are also attributable to fusions. In contrast to ∼300-5,000 point mutations per tumor across various cancer types (Lawrence et al. 2013), we find that fusions occur at mean frequencies between 0.2 and 8 per tumor across cancers. Overall, we find that 55% of all tumors have at least one fusion and ∼1.3% have more than 20 per tumor. We also found that the rate of fusions is strongly correlated with the overall genomic instability within the tumor.

Recurrence levels of somatic genetic alterations play a key role in interpreting their biological relevance. We identified 1,144 recurrent (>=2 tumors) fusion genes comprising various categories of fusion such as chimeric proteins, out-of-frame events and gene dysregulation

events. 95% of these are novel demonstrating the enormous potential for understanding the landscape of recurrent fusions in cancer biology. Overall, at least one of these recurrent events are detected in 45% of all patients in TCGA. Although we partially accounted for genomic instability by filtering out recurrent fusions demonstrating random distribution of breaks, we note that subset of these fusions may still be a consequence of regions in the genome that may be susceptible to breaks in certain tissue and environmental contexts. However, we find an enrichment of fusions involving cancer associated genes among these (224 out of 1,144) suggesting that a substantial proportion may be biologically interesting (Futreal et al. 2004). Overall, ∼18% of all patients in the TCGA are predicted to have a recurrent fusion involving a known cancer gene.

Interestingly, we observed a striking difference between the frequency spectrum of recurrent fusions and other types of alterations such as point mutations. Apart from the well-known *TMPRSS2-ERG* fusion that is characteristic of $\tilde{4}5\%$ of prostate cancers, the top five most recurrent fusions discovered here are found at substantially lower frequencies across cancers: *RPS6KB1-VMP1* (0.7%), BCAR4 fusions (0.6%), *FGFR3-TACC3* (0.4%), *BMPR1B-PDLIM5* (0.3%) and *ESR1-CCDC170* (0.25%). In contrast, nearly 43 genes that were previously associated with cancer have been found to be mutated in at least 10% of all human cancers. These include tumor suppressors such as *TP53* (30%), *NRG1* (19%), *PTPRK* (13%), *RAD51B* (12%) and *PTEN* (10%), and oncogenes such as *ALK* (15%), *PIK3CA* (13%), *BRAF* (13%) and *KRAS* (2%) (Chang et al. 2015; Kandoth et al. 2013).

Our results reaffirmed and substantially expanded the scope of two emerging trends in the recent studies investigating gene fusions in solid tumors. First, we find that recurrent fusion genes are generated within tumors across different morphological and histological types. Excluding prostate, leukemias and sarcomas, we find that 207 out of 224 highly recurrent fusions (>=5 tumors) are detected across multiple cancer types. These include 14/16 known fusions. For example, we discover fusions with biological and clinical significance such as *FGFR3-TACC3*, *EML4-ALK*, *ETV6-NTRK3* and *PTPRK-RSPO3* in three to twelve dif-

ferent cancer types, respectively, suggesting a shared activation of oncogenic pathways. Such phenomenon of shared similarities in etiology of cancer across organs has been previously reported. *TP53*-mutated high grade serous ovarian, serous endometrial and triple negative breast cancers have been shown to share a global transcriptional profile involving activation of similar oncogenic pathways (Hoadley et al. 2014; Cancer Genome Atlas Research et al. 2013; Cancer Genome Atlas 2012). Second, we find a striking number of recurrently fused genes that pair with a large number of partner genes. For example, the top 20 most fused oncogenes partner with a median of 8 distinct genes. These include *BRAF* (n=17 partners), *EGFR* (n=15), *MET* (n=15) and *FGFR2* (n=10). Similar diversity is observed for other classes of genes including tumor suppressors and kinases. The ubiquitous and combinatorial nature of fusions across cancers strongly emphasize a need for unbiased diagnostic approaches to identify all biologically significant fusions irrespective of the tumor type.

We sought to functionally evaluate many novel recurrent fusions discovered in this study. In chapter 3, we present our findings on the cancer inducing potential of three of the most recurrent novel fusions discovered. A challenge in functional evaluations is determining the suitable genetic context and environmental conditions that are reflective of context in which the fusions are introduced into the primary tumors during cancer development. For example, the three fusion events, *CD44-PDHX*, *BCAR4* dysregulation and *PDLIM5* truncation are detected across 17 different tissue types. In addition, the intertumor heterogeneity that manifests as different co-occurring mutations that may aid these fusions presents further challenges in experimentally determining the significance of these events. Here, we chose to evaluate these fusions in three different breast genetic backgrounds primarily because each of the three fusions are detected in at least 3 breast tumors and also due to our substantial understanding of different subtypes of breast cancer. Each fusion was evaluated in one benign and two different neoplastic subtypes of breast cancer. Our results show that the fusions can induce proliferation, invasion and migration phenotypes in a context dependent manner. We are currently performing in vivo assays by injecting these fusion-positive cells into mouse

mammary fat pad and evaluating their tumorigenic potential. Our findings demonstrates the biological and clinical significance of evaluating recurrent fusions towards understanding cancer progression as well as presenting options that affect patient care.

5.3.2 Future Directions

5.3.2.1 Novel fusion genes

The substantial inter-tumor heterogeneity, even within same tissues, leads to every cancer having its own neoplastic environment comprising various combinations of somatic alterations. Identification of the recurrent mutations and commonly altered genes has been a primary endeavor for genomics in cancer. A recent study estimated that a sample size of 600-5,000 within cancer type is required to reach saturation to identify all cancer associated genes that are mutated at >1% (Lawrence et al. 2014). Due to the overall low frequency of fusions in cancer and the diversity of fusion partners observed, we hypothesize that the required sample size to fully understand the landscape of recurrent fusions in cancer, is substantially larger. Such analyses would start yielding interesting insights into novel classes of genes other than oncogenes/kinases/transcription factors that are not previously associated with cancer. For example, the three recurrent fusions that we demonstrated to show cancer phenotypes in various cell backgrounds involve such novel genes (see Chapter 3).

5.3.2.2 New classification scheme for gene fusions

In contrast to other types of mutations, analyzing and inferring the transcriptional and function consequences of fusion genes remains a daunting challenge. Fusion analyses to date, including ours, have focused primarily on studying recurrent fusions and classified them into several categories based on their protein coding potential (in-frame vs. out-of-frame fusions) for future follow-ups. Here, we also identify and characterize a previously under-reported class of fusions resulting in gene dysregulation. However, limiting the analysis to recurrent

fusions and these three criteria severely constrains our ability to gain a broader understanding of the consequences of fusions. In addition, our observation of substantial diversity in the fusion partners and the long tail of low frequency fusions (for example, 86% of all fusion genes identified here are singletons) prompts for a need for a new approach to classify fusions. Perhaps, a classification based on the protein domains and families of domains, rather than individual genes. In chapter 3, our discovery of novel kinase fusions comprising the canonical kinase domain fused to the coiled-coil domain, a combination that is characteristic of oncogenic fusions such as *BCR-ABL1* and *FGFR3-TACC3*, as well as observations of fusions where a catalytic domain of a signaling protein is fused to protein-protein interaction domains of a different gene suggesting a convenient mechanism for disruption of signaling (data not shown) suggests that such a domain-based approach may yield novel insights. We suggest that this new language for studying fusions will allow for identification of novel classes of recurrent fusion peptides or recurrent fusion transcripts that will enable deeper understanding into the broad consequences of these events in cancer.

5.3.2.3 Genomic instability and sub-clonal fusions

We identified a strong correlation between the number of fusions within a tumor and the overall number of segment breaks in the transcriptome (a proxy for genomic instability). The interplay between the two is likely driven by the increasing genomic instability during tumor progression, a hallmark of cancer, that in itself is acquired due to a series of mutations in the DNA repair pathways. This instability manifests in the form of both increasing number of point mutations as well as disproportionately large number of genomic rearrangements. A consequence of this is that large proportion of genetic variation introduced after the onset of genomic instability may provide a repertoire for the neoplastic clone to selective acquire advantageous mutations. Such mutations may include those that aid tumor progression rather than tumor initiation. Evaluating this hypothesis requires the determination of clonality (variant allele fraction, VAF) of the fusion observed. Numerous studies have evaluated and

identified sub-clonal point mutations of clinical significance aided by straightforward inference of proportion of variant sequence reads supporting the fusion (Kirkizlar et al. 2015; Yates et al. 2015). Some previous studies have also investigated the subclonal copy number alterations across multiple cancers (McGranahan et al. 2015; Ha et al. 2014; Landau et al. 2013). In contrast, determining the VAF or proportion of tumor cells with the fusion is extremely challenging, primarily due to the difficulty in identifying fusion breakpoints from whole genome sequencing. Advances in sequencing technologies and algorithms that allow for accurate estimation of VAF of the fusion genes will enable testing the compelling hypothesis presented here that genomic instability induced recurrent fusions may be enriched for those that induce tumor progression phenotypes such as migration and invasion.

5.3.2.4 Novel functional assays

Determining the functional significance of the fusion gene is essential to establish its diagnostic, prognostic and therapeutic relevance. Enrichment of known cancer associated genes among the fusions strongly suggests that many of the fusions discovered here, including low frequency or singleton events, can have a causal role in cancer. However, apart from a select few that are highly recurrent, comprehensive functional evaluation of the vast majority of them is prohibitively expensive as well as time consuming. Advances in genome engineering techniques such as the CRISPR/Cas9 system that can introduce highly specific mutations in the genome have been incorporated into forward genetics studies to identify mutations that induce cancer phenotypes (Chen et al. 2015). Also, the recent demonstration that this system can introduce specific translocations in the genome suggests that such a system might be suitable to evaluate larger number of fusions simultaneously (Maddalo et al. 2014; Torres et al. 2014; Choi et al. 2014). Briefly, such a high throughput experimental system would involve multiple pairs of guide RNA (gRNA), each designed to introduce a double strand break in the fusion partner genes, introduced into a population of cells. Cells are then cultured and outgrowing colonies are identified and sequenced to determine their rearrangement status.

In addition to simultaneously assaying large numbers of fusions, an advantage of this assay is that the generated fusion gene will be under the endogenous promoter, and thereby, allows for a more representative evaluation of the fusion consequences on tumor phenotypes.

5.3.2.5 Novel therapeutic options

Given the low frequency of the vast majority of the fusions identified, it is clear that the drug development paradigm that brought imatinib to the clinic is not sustainable. Alternately, other therapeutic options such as cancer immunotherapies may provide opportunities to tune the patients' own T-cells to target the unique sequence at the breakpoint junction of the fusion gene with high specificity and degrade it. A recent study demonstrated the potential for these therapies in targeting tumor-specific mutations (Carreno et al. 2015). Such therapeutics, in addition to being targeted, also have the potential to be less toxic than conventional chemotherapy.

5.4 Inherited polymorphic fusion genes

5.4.1 Discussion

Copy number variants (CNVs) such as gene amplifications and deletions have been the primary focus of studies evaluating the functional and population genetic consequences of structural variation (SV) in the human genome. One of the first examples is the cytokine gene *CCL3L1* that was reported to show increased copy number in Africans compared to non-Africans (Redon et al. 2006). Evidence for recent selection and association of this copy number with diseases such as AIDS and systemic lupus demonstrated the significance of this CNV [MamtaniAnnRheum2008, DolanNatImm2007, GonzalezScience2005]. Similarly, amylase gene, *AMY1*, was later shown to demonstrate population specific differentiation between populations with high- and low-starch diets (Perry et al. 2007). These two findings demonstrated the impact of CNVs on human adaptation and evolution. Numerous studies

have further discovered and characterized many CNVs in the human genome (Iskow et al. 2012b).

Fusion genes comprise another class of genetic variation that is introduced by SVs. However, in contrast to CNVs, reports of novel fusion genes remain limited to two examples that were discovered as a consequence of studying different disease phenotypes. Here, we comprehensively surveyed more than five hundred individuals and discovered 63 polymorphic fusion genes in the genome. Interestingly, found that 6 out of 11 recurrent fusion genes showed population specific enrichment. For example, the fusion *TFG-GPR128* is seen at six-fold higher frequency in Europeans compared to African Americans. Analysis of gene expression showed that this fusion results in ectopic expression of a highly tissue-specific *GPR128* in all tissues within the donors with the fusion. Expanding on this, we find that 26 out of 36 in-frame fusions result in ectopic expression of the downstream partner gene. We find a number of interesting fusions involving genes such as *CARD18*, *BCL7A*, *CHEK2*, etc that present interesting hypotheses for functional followups. Our findings here, with respect to both population and expression analyses, draw parallels to early reports of interesting CNVs such as *CCL3L1* and *AMY1* that have substantially expanded our understanding of genetic variation contributing to phenotypic diversity. We suggest that the majority of the fusions reported here are excellent candidates for further functional and evolutionary analyses.

5.4.2 Future Directions

Uncovering this new class of genetic variation yields exciting opportunities for future research pursuits. All of which begin with first constructing the haplotype contain the fusion genes. The genotypes identified on these haplotypes can be used to perform detailed population and human genetics analyses. To do this, we are currently attempting to sequence the breakpoints for all fusion genes that were identified from WGS analysis, using long range PCR.

Genotype information on the fusion containing haplotypes will enable detailed investi-

gations into various types of selection pressures that may be acting upon the fusion gene. In addition, the genotypes can also be used to interrogate the catalog of GWAS associated SNPs for any phenotype or disease associations including cancer. Although we found an enrichment for cancer associated genes within the 63 fusions, we did not find an enrichment for any of these polymorphic fusion genes in the TCGA cancer cohort. However, due to the sensitivity issues associated with finding the germline fusions in the TCGA, we cannot exclude this without further study. Another future direction is the functional evaluation of some of the fusion genes discovered here. Gene expression analyses identified some striking examples such as in one case the fusion event resulted in the UTR-5 of a caspase 4 gene (*CASP4*) fused to the full coding domain of caspase 1 inhibitor (*CARD18*) with potential to effect immune response. Functional assays can be designed with specific hypothesis and specific quantitative or qualitative phenotype for evaluating the read out.

BIBLIOGRAPHY

Adams, J. M. et al. (1985). "The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice". In: *Nature* 318.6046, pp. 533–8.

Ades, L. et al. (2010). "Very long-term outcome of acute promyelocytic leukemia after treatment with all-trans retinoic acid and chemotherapy: the European APL Group experience". In: *Blood* 115.9, pp. 1690–6.

Agthoven, T. van et al. (2012). "Protein pathway activation mapping reveals molecular networks associated with antiestrogen resistance in breast cancer cell lines". In: *IntJCancer* 131.9, pp. 1998–2007.

Agthoven, T. van et al. (2015). "Breast Cancer Anti-Estrogen Resistance 4 (BCAR4) Drives Proliferation of IPH-926 lobular Carcinoma Cells". In: *PLoSOne* 10.8, e0136845.

Ahn, M. Y. et al. (1998). "Negative regulation of granulocytic differentiation in the myeloid precursor cell line 32Dcl3 by ear-2, a mammalian homolog of Drosophila seven-up, and a chimeric leukemogenic gene, AML1/ETO". In: *ProcNatlAcadSciUSA* 95.4, pp. 1812–7.

Andrews, S. J. et al. (2014). "Emerging evidence for functional peptides encoded by short open reading frames". In: *NatRevGenet* 15.3, pp. 193–204.

Argueso, J. L. et al. (2008). "Double-strand breaks associated with repetitive DNA can reshape the genome". In: *ProcNatlAcadSciUSA* 105.33, pp. 11845–50.

Arrighi, F. E. et al. (1971). "Localization of heterochromatin in human chromosomes". In: *Cytogenetics* 10.2, pp. 81–6.

Asmann, Y. W. et al. (2011). "A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines". In: *NucleicAcidsRes* 39.15, e100.

Barr, F. G. et al. (1993). "Rearrangement of the PAX3 paired box gene in the paediatric solid tumour alveolar rhabdomyosarcoma". In: *NatGenet* 3.2, pp. 113–7.

Barretina, J. et al. (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". In: *Nature* 483.7391, pp. 603–7.

Bass, A. J. et al. (2011). "Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion". In: *NatGenet* 43.10, pp. 964–8.

Beerenwinkel, N. et al. (2007). "Genetic progression and the waiting time to cancer". In: *PLoSComputBiol* 3.11, e225.

Behboudi, A. et al. (2006). "Molecular classification of mucoepidermoid carcinomas-prognostic significance of the MECT1-MAML2 fusion oncogene". In: *GenesChromosomesCancer* 45.5, pp. 470–81.

Bender, F. C. et al. (2000). "Caveolin-1 levels are down-regulated in human colon tumors, and ectopic expression of caveolin-1 in colon carcinoma cell lines reduces cell tumorigenicity". In: *Cancer Res* 60.20, pp. 5870–8.

Benjamini, Y. et al. (2012). "Summarizing and correcting the GC content bias in high-throughput sequencing". In: *Nucleic Acids Res* 40.10, e72.

Bennani-Baiti, I. M. et al. (2012). "Lysine-specific demethylase 1 (LSD1/KDM1A/AOF2/BHC110) is expressed and is an epigenetic drug target in chondrosarcoma, Ewing's sarcoma, osteosarcoma, and rhabdomyosarcoma". In: *HumPathol* 43.8, pp. 1300–7.

Berger, R. et al. (1979). "A new translocation in Burkitt's tumor cells". In: *HumGenet* 53.1, pp. 111–2.

Beroukhim, R. et al. (2010). "The landscape of somatic copy-number alteration across human cancers". In: *Nature* 463.7283, pp. 899–905.

Bersaglieri, T. et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene". In: *Am J Hum Genet* 74.6, pp. 1111–20.

Biegel, J. A. et al. (1991). "Chromosomal translocation t(1;13)(p36;q14) in a case of rhabdomyosarcoma". In: *GenesChromosomesCancer* 3.6, pp. 483–4.

Binder, J. X. et al. (2014). "COMPARTMENTS: unification and visualization of protein subcellular localization evidence". In: *Database (Oxford)* 2014, bau012.

Birner, P. et al. (2012). "Human homologue for Caenorhabditis elegans CUL-4 protein overexpression is associated with malignant potential of epithelial ovarian tumours and poor outcome in carcinoma". In: *JClinPathol* 65.6, pp. 507–11.

Blazek, D. et al. (2011). "The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes". In: *Genes Dev* 25.20, pp. 2158–72.

Boettger, L. M. et al. (2012). "Structural haplotypes and recent evolution of the human 17q21.31 region". In: *NatGenet* 44.8, pp. 881–5.

Bongarzone, I. et al. (1998). "RET/NTRK1 rearrangements in thyroid gland tumors of the papillary carcinoma family: correlation with clinicopathological features". In: *ClinCancerRes* 4.1, pp. 223–8.

Boveri, T. (2008). "Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris". In: *JCellSci* 121 Suppl 1, pp. 1–84.

Bozic, I. et al. (2010). "Accumulation of driver and passenger mutations during tumor progression". In: *ProcNatlAcadSciUSA* 107.43, pp. 18545–50.

Bradley, S. V. et al. (2005). "Serum antibodies to huntingtin interacting protein-1: a new blood test for prostate cancer". In: *Cancer Res* 65.10, pp. 4126–33.

Bradley, S. V. et al. (2007a). "Aberrant Huntingtin interacting protein 1 in lymphoid malignancies". In: *Cancer Res* 67.18, pp. 8923–31.

Bradley, S. V. et al. (2007b). "Huntingtin interacting protein 1 is a novel brain tumor marker that associates with epidermal growth factor receptor". In: *Cancer Res* 67.8, pp. 3609–15.

Brand, A. H. et al. (1993). "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes". In: *Development* 118.2, pp. 401–15.

Brown, D. et al. (1997). "A PMLRARalpha transgene initiates murine acute promyelocytic leukemia". In: *ProcNatlAcadSciUSA* 94.6, pp. 2551–6.

Bullinger, L. et al. (2010). "Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis". In: *Leukemia* 24.2, pp. 438–49.

Burrell, R. A. et al. (2013). "The causes and consequences of genetic heterogeneity in cancer evolution". In: *Nature* 501.7467, pp. 338–45.

Burrows, J. F. et al. (2003). "Altered expression of the septin gene, SEPT9, in ovarian neoplasia". In: *J Pathol* 201.4, pp. 581–8.

Calviello, L. et al. (2015). "Detecting actively translated open reading frames in ribosome profiling data". In: *NatMethods*.

Cancer Genome Atlas, Network (2012). "Comprehensive molecular portraits of human breast tumours". In: *Nature* 490.7418, pp. 61–70.

Cancer Genome Atlas Research, Network (2013). "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia". In: *N Engl J Med* 368.22, pp. 2059–74.

— (2014). "Comprehensive molecular characterization of gastric adenocarcinoma". In: *Nature* 513.7517, pp. 202–9.

Cancer Genome Atlas Research, Network et al. (2013). "Integrated genomic characterization of endometrial carcinoma". In: *Nature* 497.7447, pp. 67–73.

Cao, L. et al. (2010). "Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer". In: *CancerRes* 70.16, pp. 6497–508.

Carbone, A. et al. (2008). "Array-based comparative genomic hybridization in early-stage mycosis fungoides: recurrent deletion of tumor suppressor genes BCL7A, SMAC/DIABLO, and RHOF". In: *Genes Chromosomes Cancer* 47.12, pp. 1067–75.

Carrara, M. et al. (2013). "State-of-the-art fusion-finder algorithms sensitivity and specificity". In: *Biomed Res Int* 2013, p. 340620.

Carreno, B. M. et al. (2015). "Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells". In: *Science* 348.6236, pp. 803–8.

Caspersson, T. et al. (1968). "Chemical differentiation along metaphase chromosomes". In: *Exp Cell Res* 49.1, pp. 219–22.

Caspersson, T. et al. (1969). "DNA-binding fluorochromes for the study of the organization of the metaphase nucleus". In: *Exp Cell Res* 58.1, pp. 141–52.

Caspersson, T. et al. (1970). "Identification of human chromosomes by DNA-binding fluorescent agents". In: *Chromosoma* 30.2, pp. 215–27.

Cerwenka, A. et al. (2001). "Ectopic expression of retinoic acid early inducible-1 gene (RAE-1) permits natural killer cell-mediated rejection of a MHC class I-bearing tumor in vivo". In: *Proc Natl Acad Sci U S A* 98.20, pp. 11521–6.

Chacko, A. D. et al. (2005). "SEPT9_v4 expression induces morphological change, increased motility and disturbed polarity". In: *JPathol* 206.4, pp. 458–65.

Chang, M. T. et al. (2015). "Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity". In: *Nat Biotechnol*.

Chase, A. et al. (2010). "TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals". In: *Haematologica* 95.1, pp. 20–6.

Chen, S. et al. (2013). "New genes as drivers of phenotypic evolution". In: *NatRevGenet* 14.9, pp. 645–60.

Chen, S. et al. (2015). "Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis". In: *Cell* 160.6, pp. 1246–60.

Cheriyath, V. et al. (2012). "G1P3, an interferon- and estrogen-induced survival protein contributes to hyperplasia, tamoxifen resistance and poor outcomes in breast cancer". In: *Oncogene* 31.17, pp. 2222–36.

Chmielecki, J. et al. (2013). "Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors". In: *NatGenet* 45.2, pp. 131–2.

Choi, P. S. et al. (2014). "Targeted genomic rearrangements using CRISPR/Cas technology". In: *Nat Commun* 5, p. 3728.

Christensen, J. G. et al. (2007). "Cytoreductive antitumor activity of PF-2341066, a novel inhibitor of anaplastic lymphoma kinase and c-Met, in experimental models of anaplastic large-cell lymphoma". In: *MolCancerTher* 6.12 Pt 1, pp. 3314–22.

Cin, H. et al. (2011). "Oncogenic FAM131B-BRAF fusion resulting from 7q34 deletion comprises an alternative mechanism of MAPK pathway activation in pilocytic astrocytoma". In: *ActaNeuropathol* 121.6, pp. 763–74.

Clark, J. et al. (1994). "Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma". In: *Nat Genet* 7.4, pp. 502–8.

Conant, G. C. et al. (2008). "Turning a hobby into a job: how duplicated genes find new functions". In: *NatRevGenet* 9.12, pp. 938–50.

Courseaux, A. et al. (2001). "Birth of two chimeric genes in the Hominidae lineage". In: *Science* 291.5507, pp. 1293–7.

Courseaux, A. et al. (2003). "Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation". In: *GenomeRes* 13.3, pp. 369–81.

Cozzio, A. et al. (2003). "Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors". In: *GenesDev* 17.24, pp. 3029–35.

Crews, S. et al. (1982). "Mouse c-myc oncogene is located on chromosome 15 and translocated to chromosome 12 in plasmacytomas". In: *Science* 218.4579, pp. 1319–21.

Cuenco, G. M. et al. (2001). "Cooperation of BCR-ABL and AML1/MDS1/EVI1 in blocking myeloid differentiation and rapid induction of an acute myelogenous leukemia". In: *Oncogene* 20.57, pp. 8236–48.

Daley, G. Q. et al. (1990). "Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome". In: *Science* 247.4944, pp. 824–30.

Dalla-Favera, R. et al. (1982). "Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells". In: *ProcNatlAcadSciUSA* 79.24, pp. 7824–7.

Davies, K. D. et al. (2012). "Identifying and targeting ROS1 gene fusions in non-small cell lung cancer". In: *ClinCancerRes* 18.17, pp. 4570–9.

De Braekeleer, E. et al. (2009). "RUNX1 translocations in malignant hemopathies". In: *AnticancerRes* 29.4, pp. 1031–7.

De Braekeleer, E. et al. (2011). "RUNX1 translocations and fusion genes in malignant hemopathies". In: *FutureOncol* 7.1, pp. 77–91.

Dehm, S. M. et al. (2011). "Alternatively spliced androgen receptor variants". In: *Endocr Relat Cancer* 18.5, R183–96.

Delattre, O. et al. (1992). "Gene fusion with an ETS DNA-binding domain caused by chromosome translocation in human tumours". In: *Nature* 359.6391, pp. 162–5.

Dillies, M. A. et al. (2013). "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis". In: *Brief Bioinform* 14.6, pp. 671–83.

Ding, L. et al. (2014). "Expanding the computational toolbox for mining cancer genomes". In: *Nat Rev Genet* 15.8, pp. 556–70.

Dolat, L. et al. (2014). "Septins promote stress fiber-mediated maturation of focal adhesions and renal epithelial motility". In: *JCellBiol* 207.2, pp. 225–35.

Douglass, E. C. et al. (1987). "A specific chromosomal abnormality in rhabdomyosarcoma". In: *CytogenetCellGenet* 45.3-4, pp. 148–55.

Douglass, E. C. et al. (1991). "Variant translocations of chromosome 13 in alveolar rhabdomyosarcoma". In: *GenesChromosomesCancer* 3.6, pp. 480–2.

Drets, M. E. et al. (1971). "Specific banding patterns of human chromosomes". In: *Proc Natl Acad Sci U S A* 68.9, pp. 2073–7.

Drilon, A. et al. (2013). “Response to Cabozantinib in patients with RET fusion-positive lung adenocarcinomas”. In: *CancerDiscov* 3.6, pp. 630–5.

Druker, B. J. et al. (2001). “Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome”. In: *NEnglJMed* 344.14, pp. 1038–42.

Druker, B. J. et al. (2006). “Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia”. In: *NEnglJMed* 355.23, pp. 2408–17.

Eddy, S. R. (2009). “A new generation of homology search tools based on probabilistic inference”. In: *GenomeInform* 23.1, pp. 205–11.

Estey, M. P. et al. (2010). “Distinct roles of septins in cytokinesis: SEPT9 mediates midbody abscission”. In: *JCellBiol* 191.4, pp. 741–9.

Finn, R. D. et al. (2014). “Pfam: the protein families database”. In: *NucleicAcidsRes* 42.Database issue, pp. D222–30.

Flavell, J. R. et al. (2008). “Down-regulation of the TGF-beta target gene, PTPRK, by the Epstein-Barr virus encoded EBNA1 contributes to the growth and survival of Hodgkin lymphoma cells”. In: *Blood* 111.1, pp. 292–301.

Fraisse, J. et al. (1981). “Variant translocation in Burkitt’s lymphoma: 8;22 translocation in a French patient with an Epstein-Barr virus-associated tumor”. In: *CancerGenetCytogenet* 3.2, pp. 149–53.

Fredriksson, R. et al. (2003). “There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini”. In: *BiochemBiophysResCommun* 301.3, pp. 725–34.

French, C. A. et al. (2004). “Midline carcinoma of children and young adults with NUT rearrangement”. In: *JClinOncol* 22.20, pp. 4135–9.

Frenkel-Morgenstern, M. et al. (2012). “Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts”. In: *GenomeRes* 22.7, pp. 1231–42.

Futreal, P. A. et al. (2004). “A census of human cancer genes”. In: *NatRevCancer* 4.3, pp. 177–83.

Galili, N. et al. (1993). "Fusion of a fork head domain gene to PAX3 in the solid tumour alveolar rhabdomyosarcoma". In: *NatGenet* 5.3, pp. 230–5.

Gao, J. et al. (1991). "Isolation of a yeast artificial chromosome spanning the 8;21 translocation breakpoint t(8;21)(q22;q22.3) in acute myelogenous leukemia". In: *ProcNatlAcadSciUSA* 88.11, pp. 4882–6.

Garg, R. et al. (2014). "Protein kinase C and cancer: what we know and what we do not". In: *Oncogene* 33.45, pp. 5225–37.

Genomes Project, Consortium et al. (2015). "A global reference for human genetic variation". In: *Nature* 526.7571, pp. 68–74.

Geurts, J. M. et al. (1997). "Expression of reciprocal hybrid transcripts of HMGIC and FHIT in a pleomorphic adenoma of the parotid gland". In: *CancerRes* 57.1, pp. 13–7.

Godinho, M. et al. (2011). "Characterization of BCAR4, a novel oncogene causing endocrine resistance in human breast cancer cells". In: *JCellPhysiol* 226.7, pp. 1741–9.

Grabherr, M. G. et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome". In: *NatBiotechnol* 29.7, pp. 644–52.

Greaves, M. et al. (2012). "Clonal evolution in cancer". In: *Nature* 481.7381, pp. 306–13.

Greco, A. et al. (1997). "Chromosome 1 rearrangements involving the genes TPR and NTRK1 produce structurally different thyroid-specific TRK oncogenes". In: *GenesChromosomesCancer* 19.2, pp. 112–23.

Grieco, M. et al. (1990). "PTC is a novel rearranged form of the ret proto-oncogene and is frequently detected in vivo in human thyroid papillary carcinomas". In: *Cell* 60.4, pp. 557–63.

Grimwade, D. et al. (1997). "Characterization of cryptic rearrangements and variant translocations in acute promyelocytic leukemia". In: *Blood* 90.12, pp. 4876–85.

Grimwade, D. et al. (2010). "Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities

among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials". In: *Blood* 116.3, pp. 354–65.

Grisolano, J. L. et al. (1997). "Altered myeloid development and acute leukemia in transgenic mice expressing PML-RAR alpha under control of cathepsin G regulatory sequences". In: *Blood* 89.2, pp. 376–87.

Groffen, J. et al. (1984). "Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22". In: *Cell* 36.1, pp. 93–9.

Gu, T. L. et al. (2011). "Survey of tyrosine kinase signaling reveals ROS kinase fusions in human cholangiocarcinoma". In: *PLoSOne* 6.1, e15640.

Guo, J. U. et al. (2014). "Expanded identification and characterization of mammalian circular RNAs". In: *Genome Biol* 15.7, p. 409.

Ha, G. et al. (2014). "TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data". In: *Genome Res* 24.11, pp. 1881–93.

Hagemeijer, A. et al. (1982). "Translocation (9;11)(p21;q23) in three cases of acute monoblastic leukemia". In: *CancerGenetCytogenet* 5.2, pp. 95–105.

Hayward, W. S. et al. (1981). "Activation of a cellular onc gene by promoter insertion in ALV-induced lymphoid leukosis". In: *Nature* 290.5806, pp. 475–80.

He, L. Z. et al. (1997). "Acute leukemia with promyelocytic features in PML/RARalpha transgenic mice". In: *ProcNatlAcadSciUSA* 94.10, pp. 5302–7.

Heisterkamp, N. et al. (1983). "Localization of the c-ab1 oncogene adjacent to a translocation break point in chronic myelocytic leukaemia". In: *Nature* 306.5940, pp. 239–42.

Hinds, P. W. et al. (1992). "Regulation of retinoblastoma protein functions by ectopic expression of human cyclins". In: *Cell* 70.6, pp. 993–1006.

Hirao, A. et al. (2002). "Chk2 is a tumor suppressor that regulates apoptosis in both an ataxia telangiectasia mutated (ATM)-dependent and an ATM-independent manner". In: *Mol Cell Biol* 22.18, pp. 6521–32.

Hoadley, K. A. et al. (2014). "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin". In: *Cell* 158.4, pp. 929–44.

Holt, R. et al. (2012). "CNVs leading to fusion transcripts in individuals with autism spectrum disorder". In: *EurJHumGenet* 20.11, pp. 1141–7.

Honeyman, J. N. et al. (2014). "Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma". In: *Science* 343.6174, pp. 1010–4.

Horiuchi, T. et al. (2006). "Alternative trans-splicing: a novel mode of pre-mRNA processing". In: *BiolCell* 98.2, pp. 135–40.

Hsu, T. C. et al. (1971). "Distribution of constitutive heterochromatin in mamallian chromosomes". In: *Chromosoma* 34.3, pp. 243–53.

Hua, S. et al. (2009). "Genomic antagonism between retinoic acid and estrogen signaling in breast cancer". In: *Cell* 137.7, pp. 1259–71.

Huang, M. E. et al. (1988). "Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia". In: *Blood* 72.2, pp. 567–72.

Humke, E. W. et al. (2000). "ICEBERG: a novel inhibitor of interleukin-1beta generation". In: *Cell* 103.1, pp. 99–111.

Huntly, B. J. et al. (2004). "MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors". In: *CancerCell* 6.6, pp. 587–96.

Iskow, R. C. et al. (2012a). "Exploring the role of copy number variants in human adaptation". In: *TrendsGenet* 28.6, pp. 245–57.

— (2012b). "Exploring the role of copy number variants in human adaptation". In: *Trends Genet* 28.6, pp. 245–57.

Iyer, M. K. et al. (2011). "ChimeraScan: a tool for identifying chimeric transcription in sequencing data". In: *Bioinformatics* 27.20, pp. 2903–4.

Jia, W. et al. (2013). "SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data". In: *GenomeBiol* 14.2, R12.

Jones, D. T. et al. (2008). "Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas". In: *CancerRes* 68.21, pp. 8673–7.

Kaessmann, H. (2010). "Origins, evolution, and phenotypic impact of new genes". In: *GenomeRes* 20.10, pp. 1313–26.

Kalyana-Sundaram, S. et al. (2012). "Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer". In: *Neoplasia* 14.8, pp. 702–8.

Kandoth, C. et al. (2013). "Mutational landscape and significance across 12 major cancer types". In: *Nature* 502.7471, pp. 333–9.

Kangaspeska, S. et al. (2012). "Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms". In: *PLoS One* 7.10, e48745.

Kas, K. et al. (1997). "Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations". In: *NatGenet* 15.2, pp. 170–4.

Kauffman, E. C. et al. (2014). "Molecular genetics and cellular features of TFE3 and TFEB fusion kidney cancers". In: *NatRevUrol* 11.8, pp. 465–75.

Kerr, J. F. et al. (1972). "Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics". In: *Br J Cancer* 26.4, pp. 239–57.

Kim, D. et al. (2011). "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts". In: *GenomeBiol* 12.8, R72.

Kinsella, M. et al. (2011). "Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs". In: *Bioinformatics* 27.8, pp. 1068–75.

Kirkizlar, E. et al. (2015). "Detection of Clonal and Subclonal Copy-Number Variants in Cell-Free DNA from Patients with Breast Cancer Using a Massively Multiplexed PCR Methodology". In: *Transl Oncol* 8.5, pp. 407–16.

Kitabayashi, I. et al. (1998). "The AML1-MTG8 leukemic fusion protein forms a complex with a novel member of the MTG8(ETO/CDR) family, MTGR1". In: *MolCellBiol* 18.2, pp. 846–58.

Klijn, C. et al. (2015). "A comprehensive transcriptional portrait of human cancer cell lines". In: *NatBiotechnol* 33.3, pp. 306–12.

Kloosterman, W. P. et al. (2015). "Characteristics of de novo structural changes in the human genome". In: *GenomeRes* 25.6, pp. 792–801.

Knezevich, S. R. et al. (1998). "A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma". In: *NatGenet* 18.2, pp. 184–7.

Knudson A. G., Jr. (1971). "Mutation and cancer: statistical study of retinoblastoma". In: *Proc Natl Acad Sci U S A* 68.4, pp. 820–3.

Kohno, T. et al. (2012). "KIF5B-RET fusions in lung adenocarcinoma". In: *NatMed* 18.3, pp. 375–7.

Konopka, J. B. et al. (1984). "An alteration of the human c-abl protein in K562 leukemia cells unmasks associated tyrosine kinase activity". In: *Cell* 37.3, pp. 1035–42.

Koolen, D. A. et al. (2006). "A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism". In: *NatGenet* 38.9, pp. 999–1001.

Koolen, D. A. et al. (2008). "Clinical and molecular delineation of the 17q21.31 microdeletion syndrome". In: *JMedGenet* 45.11, pp. 710–20.

Krivtsov, A. V. et al. (2007). "MLL translocations, histone modifications and leukaemia stem-cell development". In: *NatRevCancer* 7.11, pp. 823–33.

Kroll, T. G. et al. (2000). "PAX8-PPARgamma1 fusion oncogene in human thyroid carcinoma [corrected]". In: *Science* 289.5483, pp. 1357–60.

Kuroda, S. et al. (1996). "Protein-protein interaction of zinc finger LIM domains with protein kinase C". In: *JBiolChem* 271.49, pp. 31029–32.

Kwak, E. L. et al. (2010). "Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer". In: *NEnglJMed* 363.18, pp. 1693–703.

Kyrkou, A. et al. (2013). "RhoD participates in the regulation of cell-cycle progression and centrosome duplication". In: *Oncogene* 32.14, pp. 1831–42.

Landau, D. A. et al. (2013). "Evolution and impact of subclonal mutations in chronic lymphocytic leukemia". In: *Cell* 152.4, pp. 714–26.

Lawrence, M. S. et al. (2013). "Mutational heterogeneity in cancer and the search for new cancer-associated genes". In: *Nature* 499.7457, pp. 214–8.

Lawrence, M. S. et al. (2014). "Discovery and saturation analysis of cancer genes across 21 tumour types". In: *Nature* 505.7484, pp. 495–501.

Layer, R. M. et al. (2014). "LUMPY: a probabilistic framework for structural variant discovery". In: *Genome Biol* 15.6, R84.

Lee, J. et al. (2013). "Identification of ROS1 rearrangement in gastric adenocarcinoma". In: *Cancer* 119.9, pp. 1627–35.

Lemons, R. S. et al. (1990). "Cloning and characterization of the t(15;17) translocation breakpoint region in acute promyelocytic leukemia". In: *GenesChromosomesCancer* 2.2, pp. 79–87.

Ley, T. J. et al. (2008). "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome". In: *Nature* 456.7218, pp. 66–72.

Li, H. et al. (2006). "Protein kinase C beta enhances growth and expression of cyclin D1 in human breast cancer cells". In: *CancerRes* 66.23, pp. 11399–408.

Li, H. et al. (2008). "A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells". In: *Science* 321.5894, pp. 1357–61.

Li, H. et al. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". In: *Bioinformatics* 25.14, pp. 1754–60.

Lichter, P. et al. (1988). "Rapid detection of human chromosome 21 aberrations by in situ hybridization". In: *Proc Natl Acad Sci U S A* 85.24, pp. 9664–8.

Lin, B. et al. (2000). "Orphan receptor COUP-TF is required for induction of retinoic acid receptor beta, growth inhibition, and apoptosis by retinoic acid in cancer cells". In: *MolCellBiol* 20.3, pp. 957–70.

Lin, E. et al. (2009). "Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers". In: *MolCancerRes* 7.9, pp. 1466–76.

Linger, R. M. et al. (2013). "Mer or Axl receptor tyrosine kinase inhibition promotes apoptosis, blocks growth and enhances chemosensitivity of human non-small cell lung cancer". In: *Oncogene* 32.29, pp. 3420–31.

Lipson, D. et al. (2012). "Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies". In: *NatMed* 18.3, pp. 382–4.

Liu, Y. et al. (1996). "Retinoic acid receptor beta mediates the growth-inhibitory effect of retinoic acid by promoting apoptosis in human breast cancer cells". In: *MolCellBiol* 16.3, pp. 1138–49.

Long, M. et al. (1993). "Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila". In: *Science* 260.5104, pp. 91–5.

Long, M. et al. (2003). "The origin of new genes: glimpses from the young and old". In: *NatRevGenet* 4.11, pp. 865–75.

Lu, Z. et al. (2006). "LMO4 can interact with Smad proteins and modulate transforming growth factor-beta signaling in epithelial cells". In: *Oncogene* 25.20, pp. 2920–30.

Lupas, A. et al. (1991). "Predicting coiled coils from protein sequences". In: *Science* 252.5009, pp. 1162–4.

Ma, S. et al. (2007). "The significance of LMO2 expression in the progression of prostate cancer". In: *JPathol* 211.3, pp. 278–85.

MacDonald, J. R. et al. (2014). "The Database of Genomic Variants: a curated collection of structural variation in the human genome". In: *Nucleic Acids Res* 42.Database issue, pp. D986–92.

Maddalo, D. et al. (2014). "In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system". In: *Nature* 516.7531, pp. 423–7.

Maher, C. A. et al. (2009). "Transcriptome sequencing to detect gene fusions in cancer". In: *Nature* 458.7234, pp. 97–101.

Maley, C. C. et al. (2004). "Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus". In: *CancerRes* 64.10, pp. 3414–27.

Manning, G. et al. (2002). "The protein kinase complement of the human genome". In: *Science* 298.5600, pp. 1912–34.

Manolio, T. A. et al. (2009). "Finding the missing heritability of complex diseases". In: *Nature* 461.7265, pp. 747–53.

Martinez-Garay, I. et al. (2002). "A new gene family (FAM9) of low-copy repeats in Xp22.3 expressed exclusively in testis: implications for recombinations in this region". In: *Genomics* 80.3, pp. 259–67.

Maturana, A. D. et al. (2011). "LIM domains regulate protein kinase C activity: a novel molecular function". In: *CellSignal* 23.5, pp. 928–34.

McGranahan, N. et al. (2015). "Clonal status of actionable driver events and the timing of mutational processes in cancer evolution". In: *Sci Transl Med* 7.283, 283ra54.

McPherson, A. et al. (2011). "deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data". In: *PLoSComputBiol* 7.5, e1001138.

Meijer, D. et al. (2006). "Functional screen for genes responsible for tamoxifen resistance in human breast cancer cells". In: *MolCancerRes* 4.6, pp. 379–86.

Mertens, F. et al. (2009). "Translocation-related sarcomas". In: *SeminOncol* 36.4, pp. 312–23.

Mertens, F. et al. (2015). "The emerging complexity of gene fusions in cancer". In: *NatRevCancer* 15.6, pp. 371–81.

Meyer, C. et al. (2013). "The MLL recombinome of acute leukemias in 2013". In: *Leukemia* 27.11, pp. 2165–76.

Milne, T. A. et al. (2002). "MLL targets SET domain methyltransferase activity to Hox gene promoters". In: *MolCell* 10.5, pp. 1107–17.

Mitelman, F. et al. (2007). "The impact of translocations and gene fusions on cancer causation". In: *NatRevCancer* 7.4, pp. 233–45.

Miyoshi, H. et al. (1991). "t(8;21) breakpoints on chromosome 21 in acute myeloid leukemia are clustered within a limited region of a single gene, AML1". In: *ProcNatlAcadSciUSA* 88.23, pp. 10431–4.

Miyoshi, H. et al. (1993). "The t(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript". In: *EMBOJ* 12.7, pp. 2715–21.

Miyoshi, I. et al. (1979). "2/8 translocation in a Japanese Burkitt's lymphoma". In: *Experientia* 35.6, pp. 742–3.

Montagna, C. et al. (2003). "The Septin 9 (MSF) gene is amplified and overexpressed in mouse mammary gland adenocarcinomas and human breast cancer cell lines". In: *Cancer Res* 63.9, pp. 2179–87.

Mrozek, K. et al. (2004). "Cytogenetics in acute leukemia". In: *Blood Rev* 18.2, pp. 115–36.

Munoz, L. et al. (2003). "Acute myeloid leukemia with MLL rearrangements: clinicobiological features, prognostic impact and value of flow cytometry in the detection of residual leukemic cells". In: *Leukemia* 17.1, pp. 76–82.

Naka, N. et al. (2010). "Synovial sarcoma is a stem cell malignancy". In: *StemCells* 28.7, pp. 1119–31.

Neel, B. G. et al. (1981). "Avian leukosis virus-induced tumors have common proviral integration sites and synthesize discrete new RNAs: oncogenesis by promoter insertion". In: *Cell* 23.2, pp. 323–34.

Neel, B. G. et al. (1982). "Two human c-onc genes are located on the long arm of chromosome 8". In: *ProcNatlAcadSciUSA* 79.24, pp. 7842–6.

Ni, Y. Y. et al. (2014). "Deletion of Gpr128 results in weight loss and increased intestinal contraction frequency". In: *WorldJGastroenterol* 20.2, pp. 498–508.

Northcott, P. A. et al. (2012). "Subgroup-specific structural variation across 1,000 medulloblastoma genomes". In: *Nature* 488.7409, pp. 49–56.

Nowell, P. C. et al. (1960). "Chromosome studies on normal and leukemic human leukocytes". In: *JNatlCancerInst* 25, pp. 85–109.

Nucifora, G. et al. (1993). "Detection of DNA rearrangements in the AML1 and ETO loci and of an AML1/ETO fusion mRNA in patients with t(8;21) acute myeloid leukemia". In: *Blood* 81.4, pp. 883–8.

Ohno, S. et al. (1979). "Nonrandom chromosome changes involving the Ig gene-carrying chromosomes 12 and 6 in pristane-induced mouse plasmacytomas". In: *Cell* 18.4, pp. 1001–7.

Okada, Y. et al. (2005). "hDOT1L links histone methylation to leukemogenesis". In: *Cell* 121.2, pp. 167–78.

Okuda, T. et al. (1998). "Expression of a knocked-in AML1-ETO leukemia gene inhibits the establishment of normal definitive hematopoiesis and directly generates dysplastic hematopoietic progenitors". In: *Blood* 91.9, pp. 3134–43.

Osaka, M. et al. (1999). "MSF (MLL septin-like fusion), a fusion partner gene of MLL, in a therapy-related acute myeloid leukemia with a t(11;17)(q23;q25)". In: *ProcNatlAcadSciUSA* 96.11, pp. 6428–33.

Oshimura, M. et al. (1977). "Chromosomes and causation of human cancer and leukemia. XXVI. Binding studies in acute lymphoblastic leukemia (ALL)". In: *Cancer* 40.3, pp. 1161–72.

Ostler, K. R. et al. (2007). "Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins". In: *Oncogene* 26.38, pp. 5553–63.

Palanisamy, N. et al. (2010). "Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma". In: *NatMed* 16.7, pp. 793–8.

"Paris Conference (1971): Standardization in human cytogenetics" (1972). In: *Cytogenetics* 11.5, pp. 317–62.

Parker, B. C. et al. (2013). "The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma". In: *JClinInvest* 123.2, pp. 855–65.

Parker, M. et al. (2014). "C11orf95-RELA fusions drive oncogenic NF-kappaB signalling in ependymoma". In: *Nature* 506.7489, pp. 451–5.

Parri, M. et al. (2010). "Rac and Rho GTPases in cancer cell motility control". In: *CellCommunSignal* 8, p. 23.

Patro, R. et al. (2014). "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms". In: *Nat Biotechnol* 32.5, pp. 462–4.

Payne, G. S. et al. (1981). "Analysis of avian leukosis virus DNA and RNA in bursal tumours: viral gene expression is not required for maintenance of the tumor state". In: *Cell* 23.2, pp. 311–22.

Peifer, M. (1997). "Beta-catenin as oncogene: the smoking gun". In: *Science* 275.5307, pp. 1752–3.

Perner, S. et al. (2007). "TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion". In: *AmJSurgPathol* 31.6, pp. 882–8.

Perner, S. et al. (2013). "Loss of SLC45A3 protein (prostein) expression in prostate cancer is associated with SLC45A3-ERG gene rearrangement and an unfavorable clinical course". In: *IntJCancer* 132.4, pp. 807–12.

Perry, G. H. et al. (2007). "Diet and the evolution of human amylase gene copy number variation". In: *Nat Genet* 39.10, pp. 1256–60.

Pickrell, J. K. et al. (2009). "Signals of recent positive selection in a worldwide sample of human populations". In: *Genome Res* 19.5, pp. 826–37.

Pierotti, M. A. (2001). "Chromosomal rearrangements in thyroid carcinomas: a recombination or death dilemma". In: *CancerLett* 166.1, pp. 1–7.

Pierotti, M. A. et al. (1992). "Characterization of an inversion on the long arm of chromosome 10 juxtaposing D10S170 and RET and creating the oncogenic sequence RET/PTC". In: *ProcNatlAcadSciUSA* 89.5, pp. 1616–20.

Pon, J. R. et al. (2015). "Driver and passenger mutations in cancer". In: *Annu Rev Pathol* 10, pp. 25–50.

Prasad, R. et al. (1993). "Cloning of the ALL-1 fusion partner, the AF-6 gene, involved in acute myeloid leukemias with the t(6;11) chromosome translocation". In: *CancerRes* 53.23, pp. 5624–8.

Quan, J. et al. (2011). "Parallel on-chip gene synthesis and application to optimization of protein expression". In: *Nat Biotechnol* 29.5, pp. 449–52.

Reader, J. C. et al. (2007). "A novel NUP98-PHF23 fusion resulting from a cryptic translocation t(11;17)(p15;p13) in acute myeloid leukemia". In: *Leukemia* 21.4, pp. 842–4.

Redon, R. et al. (2006). "Global variation in copy number in the human genome". In: *Nature* 444.7118, pp. 444–54.

Rikova, K. et al. (2007). "Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer". In: *Cell* 131.6, pp. 1190–203.

Rimkunas, V. M. et al. (2012). "Analysis of receptor tyrosine kinase ROS1-positive tumors in non-small cell lung cancer: identification of a FIG-ROS1 fusion". In: *ClinCancerRes* 18.16, pp. 4449–57.

Rinn, J. L. et al. (2012). "Genome regulation by long noncoding RNAs". In: *AnnuRevBiochem* 81, pp. 145–66.

Roach, J. C. et al. (2010). "Analysis of genetic inheritance in a family quartet by whole-genome sequencing". In: *Science* 328.5978, pp. 636–9.

Robertson, G. et al. (2010). "De novo assembly and analysis of RNA-seq data". In: *NatMethods* 7.11, pp. 909–12.

Robinson, D. et al. (2015). "Integrative clinical genomics of advanced prostate cancer". In: *Cell* 161.5, pp. 1215–28.

Robinson, D. R. et al. (2013). "Identification of recurrent NAB2-STAT6 gene fusions in solitary fibrous tumor by integrative sequencing". In: *NatGenet* 45.2, pp. 180–5.

Rosenbauer, F. et al. (2007). "Transcription factors in myeloid development: balancing differentiation with transformation". In: *NatRevImmunol* 7.2, pp. 105–17.

Ross, B. D. et al. (2013). "Stepwise evolution of essential centromere function in a Drosophila neogene". In: *Science* 340.6137, pp. 1211–4.

Rowley, J. D. (1973a). "Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia". In: *AnnGenet* 16.2, pp. 109–12.

— (1973b). "Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining". In: *Nature* 243.5405, pp. 290–3.

— (2008). "Chromosomal translocations: revisited yet again". In: *Blood* 112.6, pp. 2183–9.

Rowley, J. D. et al. (1977). "15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia". In: *Lancet* 1.8010, pp. 549–50.

Rowley, J. D. et al. (1990). "Mapping chromosome band 11q23 in human acute leukemia with biotinylated probes: identification of 11q23 translocation breakpoints with a yeast artificial chromosome". In: *Proc Natl Acad Sci U S A* 87.23, pp. 9358–62.

Russell, S. E. et al. (2005). "Do septins have a role in cancer?" In: *Br J Cancer* 93.5, pp. 499–503.

Saville, M. K. et al. (2004). "Regulation of p53 by the ubiquitin-conjugating enzymes UbcH5B/C in vivo". In: *JBiolChem* 279.40, pp. 42169–81.

Sboner, A. et al. (2010). "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data". In: *GenomeBiol* 11.10, R104.

Schindl, M. et al. (2007). "Overexpression of the human homologue for Caenorhabditis elegans cul-4 gene is associated with poor outcome in node-negative breast cancer". In: *AnticancerRes* 27.2, pp. 949–52.

Schramek, D. et al. (2014). "Direct in vivo RNAi screen unveils myosin IIa as a tumor suppressor of squamous cell carcinomas". In: *Science* 343.6168, pp. 309–13.

Schrock, E. et al. (1996). "Multicolor spectral karyotyping of human chromosomes". In: *Science* 273.5274, pp. 494–7.

Scott, M. et al. (2005). "Multimodality expression profiling shows SEPT9 to be overexpressed in a wide range of human tumours". In: *Oncogene* 24.29, pp. 4688–700.

Seo, J. S. et al. (2012). "The transcriptional landscape and mutational profile of lung adenocarcinoma". In: *GenomeRes* 22.11, pp. 2109–19.

Seshagiri, S. et al. (2012). "Recurrent R-spondin fusions in colon cancer". In: *Nature* 488.7413, pp. 660–4.

Shapiro, D. N. et al. (1993). "Fusion of PAX3 to a member of the forkhead family of transcription factors in human alveolar rhabdomyosarcoma". In: *CancerRes* 53.21, pp. 5108–12.

Sharp, A. J. et al. (2006). "Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome". In: *NatGenet* 38.9, pp. 1038–42.

Shaw, A. T. et al. (2013a). "Crizotinib versus chemotherapy in advanced ALK-positive lung cancer". In: *NEnglJMed* 368.25, pp. 2385–94.

Shaw, A. T. et al. (2013b). "Tyrosine kinase gene rearrangements in epithelial malignancies". In: *NatRevCancer* 13.11, pp. 772–87.

Sheiness, D. et al. (1979). "DNA and RNA from uninfected vertebrate cells contain nucleotide sequences related to the putative transforming gene of avian myelocytomatosis virus". In: *JVirol* 31.2, pp. 514–21.

Shen-Ong, G. L. et al. (1982). "Novel myc oncogene RNA from abortive immunoglobulin-gene recombination in mouse plasmacytomas". In: *Cell* 31.2 Pt 1, pp. 443–52.

Shinmura, K. et al. (2014). "RSPO fusion transcripts in colorectal cancer in Japanese population". In: *MolBiolRep* 41.8, pp. 5375–84.

Shtivelman, E. et al. (1985). "Fused transcript of abl and bcr genes in chronic myelogenous leukaemia". In: *Nature* 315.6020, pp. 550–4.

Simon-Sanchez, J. et al. (2009). "Genome-wide association study reveals genetic risk underlying Parkinson's disease". In: *NatGenet* 41.12, pp. 1308–12.

Simsek, D. et al. (2010). "Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4-ligase IV during chromosomal translocation formation". In: *NatStructMolBiol* 17.4, pp. 410–6.

Singh, D. et al. (2012). "Transforming fusions of FGFR and TACC genes in human glioblastoma". In: *Science* 337.6099, pp. 1231–5.

Skalova, A. et al. (2010). "Mammary analogue secretory carcinoma of salivary glands, containing the ETV6-NTRK3 fusion gene: a hitherto undescribed salivary gland tumor entity". In: *AmJSurgPathol* 34.5, pp. 599–608.

Skipper, L. et al. (2004). "Linkage disequilibrium and association of MAPT H1 in Parkinson disease". In: *AmJHumGenet* 75.4, pp. 669–77.

Smith, S. et al. (1987). "A consistent chromosome translocation in synovial sarcoma". In: *CancerGenetCytogenet* 26.1, pp. 179–80.

Soda, M. et al. (2007). "Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer". In: *Nature* 448.7153, pp. 561–6.

Soda, M. et al. (2008). "A mouse model for EML4-ALK-positive lung cancer". In: *ProcNatlAcadSciUSA* 105.50, pp. 19893–7.

Sollberger, G. et al. (2012). "Caspase-4 is required for activation of inflammasomes". In: *JImmunol* 188.4, pp. 1992–2000.

Spain, B. H. et al. (1999). "Truncated BRCA2 is cytoplasmic: implications for cancer-linked mutations". In: *Proc Natl Acad Sci U S A* 96.24, pp. 13920–5.

Stefansson, H. et al. (2005). "A common inversion under selection in Europeans". In: *Nat Genet* 37.2, pp. 129–37.

Steinberg, K. M. et al. (2012). "Structural diversity and African origin of the 17q21.31 inversion polymorphism". In: *Nat Genet* 44.8, pp. 872–80.

Stephens, P. J. et al. (2011). "Massive genomic rearrangement acquired in a single catastrophic event during cancer development". In: *Cell* 144.1, pp. 27–40.

Stewart, C. et al. (2011). "A comprehensive map of mobile element insertion polymorphisms in humans". In: *PLoS Genet* 7.8, e1002236.

Stransky, N. et al. (2014). "The landscape of kinase fusions in cancer". In: *NatCommun* 5, p. 4846.

Stratton, M. R. et al. (2009). "The cancer genome". In: *Nature* 458.7239, pp. 719–24.

Subramanian, A. et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *ProcNatlAcadSciUSA* 102.43, pp. 15545–50.

Sudmant, P. H. et al. (2015a). "An integrated map of structural variation in 2,504 human genomes". In: *Nature* 526.7571, pp. 75–81.

Sudmant, P. H. et al. (2015b). "Global diversity, population stratification, and selection of human copy-number variation". In: *Science* 349.6253, aab3761.

Suehara, Y. et al. (2012). "Identification of KIF5B-RET and GOPC-ROS1 fusions in lung adenocarcinomas through a comprehensive mRNA-based screen for tyrosine kinase fusions". In: *ClinCancerRes* 18.24, pp. 6599–608.

Suela, J. et al. (2007). "DNA profiling by arrayCGH in acute myeloid leukemia and myelodysplastic syndromes". In: *Cytogenet Genome Res* 118.2-4, pp. 304–9.

Sugawara, E. et al. (2012). "Identification of anaplastic lymphoma kinase fusions in renal cancer: large-scale immunohistochemical screening by the intercalated antibody-enhanced polymer method". In: *Cancer* 118.18, pp. 4427–36.

Sun, P. H. et al. (2013). "Protein tyrosine phosphatase kappa (PTPRK) is a negative regulator of adhesion and invasion of breast cancer cells, and associates with poor prognosis of breast cancer". In: *J Cancer Res Clin Oncol* 139.7, pp. 1129–39.

Suzuma, K. et al. (2002). "Characterization of protein kinase C beta isoform's action on retinoblastoma protein phosphorylation, vascular endothelial growth factor-induced en-

dothelial cell proliferation, and retinal neovascularization". In: *ProcNatlAcadSciUSA* 99.2, pp. 721–6.

Tabin, C. J. et al. (1982). "Mechanism of activation of a human oncogene". In: *Nature* 300.5888, pp. 143–9.

Tahara E., Jr. et al. (2005). "G1P3, an interferon inducible gene 6-16, is expressed in gastric cancers and inhibits mitochondrial-mediated apoptosis in gastric cancer cell line TMK-1 cell". In: *CancerImmunolImmunother* 54.8, pp. 729–40.

Takahashi, M. et al. (1985). "Activation of a novel human transforming gene, ret, by DNA rearrangement". In: *Cell* 42.2, pp. 581–8.

Takeuchi, K. et al. (2012). "RET, ROS1 and ALK fusions in lung cancer". In: *NatMed* 18.3, pp. 378–81.

Tao, J. et al. (2011). "CD44-SLC1A2 gene fusions in gastric cancer". In: *SciTranslMed* 3.77, 77ra30.

Taub, R. et al. (1982). "Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells". In: *ProcNatlAcadSciUSA* 79.24, pp. 7837–41.

Taylor, B. S. et al. (2011). "Advances in sarcoma genomics and new therapeutic targets". In: *NatRevCancer* 11.8, pp. 541–57.

Taylor, J. S. et al. (2004). "Duplication and divergence: the evolution of new genes and old ideas". In: *AnnuRevGenet* 38, pp. 615–43.

The, H. de et al. (1990). "The t(15;17) translocation of acute promyelocytic leukaemia fuses the retinoic acid receptor alpha gene to a novel transcribed locus". In: *Nature* 347.6293, pp. 558–61.

Tobin, J. E. et al. (2008). "Haplotypes and gene expression implicate the MAPT region for Parkinson disease: the GenePD Study". In: *Neurology* 71.1, pp. 28–34.

Tognon, C. et al. (2002). "Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma". In: *CancerCell* 2.5, pp. 367–76.

Tomlins, S. A. et al. (2005). "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer". In: *Science* 310.5748, pp. 644–8.

Tomlins, S. A. et al. (2007). "Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer". In: *Nature* 448.7153, pp. 595–9.

Tomlins, S. A. et al. (2008). "Role of the TMPRSS2-ERG gene fusion in prostate cancer". In: *Neoplasia* 10.2, pp. 177–88.

Torres, R. et al. (2014). "Engineering human tumour-associated chromosomal translocations with the RNA-guided CRISPR-Cas9 system". In: *Nat Commun* 5, p. 3964.

Tsai, S. T. et al. (2010). "ENO1, a potential prognostic head and neck cancer marker, promotes transformation partly via chemokine CCL20 induction". In: *EurJCancer* 46.9, pp. 1712–23.

Tsujimoto, Y. et al. (1984). "Cloning of the chromosome breakpoint of neoplastic B cells with the t(14;18) chromosome translocation". In: *Science* 226.4678, pp. 1097–9.

Tu, J. J. et al. (2007). "Gene fusions between TMPRSS2 and ETS family genes in prostate cancer: frequency and transcript variant analysis by RT-PCR and FISH on paraffin-embedded tissues". In: *ModPathol* 20.9, pp. 921–8.

Turc-Carel, C. et al. (1986). "Consistent chromosomal translocation in alveolar rhabdomyosarcoma". In: *CancerGenetCytogenet* 19.3-4, pp. 361–2.

Turc-Carel, C. et al. (1987). "Involvement of chromosome X in primary cytogenetic change in human neoplasia: nonrandom translocation in synovial sarcoma". In: *ProcNatlAcadSciUSA* 84.7, pp. 1981–5.

Vaishnavi, A. et al. (2013). "Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer". In: *NatMed* 19.11, pp. 1469–72.

Vardiman, J. W. et al. (2009). "The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes". In: *Blood* 114.5, pp. 937–51.

Varmus, H. E. (1984). "The molecular genetics of cellular oncogenes". In: *Annu Rev Genet* 18, pp. 553–612.

Vaux, D. L. et al. (1988). "Bcl-2 gene promotes haemopoietic cell survival and cooperates with c-myc to immortalize pre-B cells". In: *Nature* 335.6189, pp. 440–2.

Veeraraghavan, J. et al. (2014). "Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers". In: *NatCommun* 5, p. 4577.

Viale, A. et al. (2000). "Structure and expression of the variant melanin-concentrating hormone genes: only PMCHL1 is transcribed in the developing human brain and encodes a putative protein". In: *MolBiolEvol* 17.11, pp. 1626–40.

Vogelstein, B. et al. (2013). "Cancer genome landscapes". In: *Science* 339.6127, pp. 1546–58.

Walter, M. J. et al. (2009). "Acquired copy number alterations in adult acute myeloid leukemia genomes". In: *Proc Natl Acad Sci U S A* 106.31, pp. 12950–5.

Wang, C. et al. (2010a). "C4orf7 contributes to ovarian cancer metastasis by promoting cancer cell migration and invasion". In: *OncolRep* 24.4, pp. 933–9.

Wang, K. et al. (2010b). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". In: *NucleicAcidsRes* 38.18, e178.

Wang, R. et al. (2012). "RET fusions define a unique molecular and clinicopathologic subtype of non-small-cell lung cancer". In: *JClinOncol* 30.35, pp. 4352–9.

Wang, Y. et al. (2014a). "CUL4A induces epithelial-mesenchymal transition and promotes cancer metastasis by regulating ZEB1 expression". In: *CancerRes* 74.2, pp. 520–31.

Wang, Y. et al. (2014b). "CUL4A overexpression enhances lung tumor growth and sensitizes lung cancer cells to erlotinib via transcriptional regulation of EGFR". In: *MolCancer* 13, p. 252.

Warburton, D. (1991). "De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints". In: *Am J Hum Genet* 49.5, pp. 995–1013.

Weischenfeldt, J. et al. (2013). "Phenotypic impact of genomic structural variation: insights from and for human disease". In: *NatRevGenet* 14.2, pp. 125–38.

Welch, J. S. et al. (2011). "Use of whole-genome sequencing to diagnose a cryptic fusion oncogene". In: *JAMA* 305.15, pp. 1577–84.

Whang-Peng, J. et al. (1986). "Cytogenetic characterization of selected small round cell tumors of childhood". In: *CancerGenetCytogenet* 21.3, pp. 185–208.

Willert, K. et al. (1998). "Beta-catenin: a key mediator of Wnt signaling". In: *CurrOpinGenetDev* 8.1, pp. 95–102.

Witte, O. N. et al. (1980). "Abelson murine leukaemia virus protein is phosphorylated in vitro to form phosphotyrosine". In: *Nature* 283.5750, pp. 826–31.

Wu, J. et al. (2013a). "SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads". In: *Bioinformatics* 29.23, pp. 2971–8.

Wu, Y. M. et al. (2013b). "Identification of targetable FGFR gene fusions in diverse cancers". In: *CancerDiscov* 3.6, pp. 636–47.

Xing, Z. et al. (2014). "lncRNA directs cooperative epigenetic regulation downstream of chemokine signals". In: *Cell* 159.5, pp. 1110–25.

Xu, Y. et al. (2015). "Notch and TGF-beta pathways cooperatively regulate receptor protein tyrosine phosphatase-kappa (PTPRK) gene expression in human primary keratinocytes". In: *Mol Biol Cell* 26.6, pp. 1199–206.

Yates, L. R. et al. (2015). "Subclonal diversification of primary breast cancer revealed by multiregion sequencing". In: *Nat Med* 21.7, pp. 751–9.

Yergeau, D. A. et al. (1997). "Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene". In: *NatGenet* 15.3, pp. 303–6.

Yin, X. M. et al. (1994). "BH1 and BH2 domains of Bcl-2 are required for inhibition of apoptosis and heterodimerization with Bax". In: *Nature* 369.6478, pp. 321–3.

Yoshihara, K. et al. (2014). "The landscape and therapeutic relevance of cancer-associated transcript fusions". In: *Oncogene*.

Zack, T. I. et al. (2013). "Pan-cancer patterns of somatic copy number alteration". In: *NatGenet* 45.10, pp. 1134–1140.

Zech, L. et al. (1976). "Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas". In: *IntJCancer* 17.1, pp. 47–56.

Zhang, F. et al. (2009). "Copy number variation in human health, disease, and evolution". In: *AnnuRevGenomicsHumGenet* 10, pp. 451–81.

Zhang, J. et al. (2004). "Protein kinase C (PKC) betaII induces cell invasion through a Ras/Mek-, PKC iota/Rac 1-dependent signaling pathway". In: *JBiolChem* 279.21, pp. 22118–23.

Zhou, Q. et al. (2008). "On the origin and evolution of new genes–a genomic and experimental perspective". In: *JGenetGenomics* 35.11, pp. 639–48.

Zody, M. C. et al. (2008). "Evolutionary toggling of the MAPT 17q21.31 inversion region". In: *NatGenet* 40.9, pp. 1076–83.

Zou, H. Y. et al. (2007). "An orally available small-molecule inhibitor of c-Met, PF-2341066, exhibits cytoreductive antitumor efficacy through antiproliferative and antiangiogenic mechanisms". In: *CancerRes* 67.9, pp. 4408–17.