INTEGRATIVE NEXT GENERATION AND SINGLE CELL SEQUENCING APPROACHES TO UNDERSTAND THE PROGRESSION AND TREATMENT RESPONSE OF HEMATOLOGICAL MALIGNANCIES

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Priyanka Vijay

January 2018

INTEGRATIVE NEXT GENERATION AND SINGLE CELL SEQUENCING APPROACHES TO UNDERSTAND THE PROGRESSION AND TREATMENT RESPONSE OF HEMATOLOGICAL MALIGNANCIES

Priyanka Vijay, Ph.D.

Cornell University 2018

Despite recent advances in cancer therapy, most hematological malignancies remain incurable. Mantle cell lymphoma (MCL), a B-cell non-Hodkin's lymphoma, and the myelodysplastic syndromes (MDS), a cancer of the bone marrow resulting in peripheral blood cytopenias and increased leukemia risk, have a median survival rate of 5 years. Standard of care for both elicits variable responses in patients, and even patients who initially respond often acquire resistance. The development of improved therapeutic strategies is hindered by the limited understanding of molecular mechanisms underlying innate and acquired resistance. Here, we investigate these mechanisms using next generation sequencing approaches in MCL and MDS in the context of specific treatment regimens.

We show that in MCL, resistance to combination of palbociclib, a CDK4/6 inhibitor, and bortezomib, a proteasome inhibitor, in a phase I clinical trial was correlated with copy number variations, and altered gene expression, which were identified through a longitudinal approach integrating exome and transcriptome sequencing of purified MCL cells from lymph node biopsies.

Notably, we identified a loss of type I interferon genes in non-responders that may cause innate resistance in these patients, and a relapse-specific hemizygous loss in chromosome 13q in an initially responding patient that may reflect an alternate mechanism of resistance.

In MDS, we used single cell RNA-sequencing to evaluate transcriptional heterogeneity in hematopoietic stem cells pre-/post-therapy with decitabine, a hypomethylating agent. To investigate dysregulated splicing at the single cell resolution, we developed DISCO, a single cell splicing analysis platform, and identified heterogeneous altered splicing of immune genes and ribosomal proteins in spliceosome-mutated cells. Subsets of ribosomal proteins were differentially spliced and down-regulated in MDS cells, with even greater dysregulation detected in non-responders. Up-regulated pathways include p53 signaling and myeloid differentiation. Heterogeneous transcriptional patterns defined distinct stem cell states, and state dynamics pre- to post-therapy highlighted differentiating features (ex. CTBP1 expression) of populations likely driving resistance.

Overall, these results leverage the latest sequencing technologies and new computational methods to study purified MCL and MDS patient cells during treatment, identify important mechanisms regulating pathogenesis and therapy response, and offer opportunities for developing a precision medicine framework in treating these diseases.

BIOGRAPHICAL SKETCH

Priyanka Vijay spent her childhood between India, Saudi Arabia, and the United States. She graduated from Rutgers University in 2012 with a major in Genetics and a minor in Mathematics where she worked on *C. elegans* reproductive biology in Dr. Andy Singson's lab. She joined the Tri-Institutional Training Program in Computational Biology and Medicine in 2012, and began her graduate research in Dr. Chris Mason's lab in August 2013, where she sought to understand more about cancer biology and impact treatment of cancer using a combination of experimental and computational next generation sequencing approaches.

ACKNOWLEDGMENTS

I thank my advisor, Dr. Chris Mason, for his continued support and unflagging enthusiasm, and my committee members Drs. Selina Chen-Kiang, Chris Park, and Olivier Elemento for guiding me and teaching me throughout my graduate career. I also thank past and present members of the Mason lab including Dr. Pedro Blecua for helping me in my infant days in the lab, Dr. Sheng Li for her advice on projects and careers, and Matt Mackay and Delia Tomoiaga for their help with the MDS project and beyond. I am thankful to Drs. Francine Garret-Bakelman, Maurizio Di Liberto, Xiangao Huang, Steve Chung, and Virginia Klimek for their collaborations. I am grateful to Dr. David Christini, Margie Hinonangan-Mendoza, Kathleen Pickering, and the Tri-I CBM program. Support for work described here was provided by Tri-I CBM through NIH training grant T32GM083937, the World Quant Foundation, the Starr Cancer Consortium, the Evans Foundation, and the National Institutes of Health.

Thank you to all my friends for making me laugh always, especially Dr. Neel Madhukar for making even the most stressful times fun. Special thanks to my pup May for reminding me to face each day with new excitement for adventure. And finally, my family - I would not be here today without the encouragement, advice, support, and love of my parents and sister.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgments	iv
List of Figures	viii
List of Tables	x
List of Abbreviations	xi
Chapter 1: Introduction	1
Chapter 2: Clinical Genomics – Challenges and Opportunities	8
Challenges	9
Opportunities	22
Chapter 3: Longitudinal Genomic and Transcriptomic Analysis of Mantle Cell Lymphoma in a Targeted Combination Trial of a Selective CDK4/6 Inhibitor	32
Preamble	32
Introduction	32
Results	34
Integrative genomic and transcriptomic examination of responders and non-responders	34
CDK4 amplifications do not affect response to PALBOR	37
Multiple large-scale copy number variations differentiate responders from non- responders	37
Two-hit loss of ATM and TP53 in non-responders	50
BCL6 amplifications do not affect response to PALBOR	52
Relapse-specific chromosome 13 deletion	54
Clonal architecture and evolution of single nucleotide variants	57
Discussion	63

Chapter 4: Single Cell Isoform Dynamics and Transcriptomics in Myelodysplastic Syndromes Stem Cells during Therapy 70

Preamble 70

Introduction 70

Results 73

Single cell RNA-seq of MDS patient bone marrow biopsies 73

MDS and Normal HSCs exhibit unique transcriptional landscapes 73

DISCO: a novel method for analyzing alternative splicing in single cell and other large scale RNA-seq data 80

DISCO recapitulates known effects of SRSF2 mutations 82

SRSF2 mutant MDS HSCs exhibit unique splicing changes 83

Aberrant expression of immune signaling and hematopoiesis genes in SRSF2 mutated cells 87

MDS HSCs exhibit differential isoform usage 87

Dysregulated ribosomal protein expression and p53 signaling in MDS stem cells 89

Ribosomal proteins down-regulated and differentially spliced in decitabine non-responders 93

Analysis of single cell heterogeneity identifies distinct cell state distributions differing by response group and time point 96

MDS HSCs exhibit decreased expression of stem cell regulators and enrichment of myeloid genes 99

Distinct stem cell states identified using semi-supervised pseudotemporal ordering of MDS and Normal HSCs 101

Decitabine induces shifts in stem cell transcriptional states 105

Transcriptional cell states identify signatures of decitabine resistance 105

Candidate resistance genes include transcriptional repressor CTBP1 107

Discussion 114

Appendix	115
Materials and Methods – MCL Exome And Transcriptome Sequencing (Chapter 3)	115
Materials and Methods – MDS Single Cell RNA Sequencing (Chapter 4)	117
References	123

LIST OF FIGURES

CHAPTER 1

Figure 1.1	Major events in a decade of cancer genomics	3

CHAPTER 2

Figure 2.1	Omics workflows	10
Figure 2.2	Integrative omics analysis pipelines	21
Figure 2.3	Single cell sequencing for clinical applications.	29

CHAPTER 3

Figure 3.1	PALBOR combination therapy protocol	36
Figure 3.2	Copy number variations	39
Figure 3.3	Expression of NR-specific deletions genes	42
Figure 3.4	Expression of NR-specific amplification genes	43
Figure 3.5	Expression of R-specific amplification genes	44
Figure 3.6	Expression of R-specific deletion genes	46
Figure 3.7	Subset of R-specific deletions	47
Figure 3.8	MYC amplifications and over-expression	49
Figure 3.9	NR-specific ATM and TP53 alterations	51
Figure 3.10	BCL6 amplifications in NR and R	53
Figure 3.11	Relapse-specific chr13 deletions	55
Figure 3.12	Expression of chr13 deletion genes	56
Figure 3.13	SNV cluster dynamics 1	59
Figure 3.14	SNV cluster dynamics 2	60
Figure 3.15	Clonal evolution across 3 time points	62
Figure 3.16	Protein interactions between interferon and TRAIL	69

LIST OF FIGURES (CONTINUED)

CHAPTER 4

Figure 4.1	Experimental design and samples	74
Figure 4.2	FACS sorting of MDS stem cells	75
Figure 4.3	Mapping rates of scRNA-seq data	76
Figure 4.4	Visualizing MDS transcriptomes	78
Figure 4.5	Cell cycle phases of single cells	79
Figure 4.6	Single cell isoform analysis with DISCO	81
Figure 4.7	DISCO recapitulates known effects of SRSF2 mutations	84
Figure 4.8	Differentially expressed isoforms in SRSF2 mutated HSCs	85
Figure 4.9	DISCO results for SRSF2 mutated	86
Figure 4.10	Differential gene expression in SRSF2 mutated	88
Figure 4.11	Altered splicing in MDS HSCs	90
Figure 4.12	Differential splicing of ADGRE5 and TAGAP	91
Figure 4.13	Differential gene expression in MDS HSCs	92
Figure 4.14	Differential gene expression between R and NR	94
Figure 4.15	Altered splicing patters between R and NR	95
Figure 4.16	Ribosomal protein gene and isoform dysregulation	97
Figure 4.17	Lineage ordering and cell state identification	98
Figure 4.18	HSC and myeloid regulators	100
Figure 4.19	Semi-supervised lineage ordering	102
Figure 4.20	Lineage ordering of each patient sample	103
Figure 4.21	Pseudotime DEGs and pathway enrichment	104
Figure 4.22	Identifying candidate resistance genes	108

LIST OF TABLES

CHAPTER 3

Table 3.1 GO terms enriched in R- or NR-specific CNVs 40

LIST OF ABBREVIATIONS

AML	Acute myeloid leukemia
CNV	Copy Number Variant
DEG	Differentially expressed gene
DEI	Differentially expressed isoform
DSG	Differentially spliced gene
GO	Gene Ontology
HSC	Hematopoietic Stem Cell
IFN	Interferon
MCL	Mantle Cell Lymphoma
MDS	Myelodysplastic Syndromes
NGS	Next Generation Sequencing
NR	Non-responder
NR-Br3	Non-responder branch 3 cells
PALBOR	Combination treatment of palbociclib and bortezomib
PCA	Principal components analysis
PD	Progression disease
pG1	Prolonged G1 arrest
pG1-S	Synchronous progression through S phase
R	Responder
R-Br3	Responder branch 3 cells
RNA-seq	RNA sequencing
RPG	Ribosomal protein gene
scRNA-seq	Single-cell RNA sequencing
SD	Stable disease
SNV	Single Nucleotide Variant
t-SNE	t-distributed stochastic neighbor embedding
TCGA	The Cancer Genome Atlas
VAF	Variant Allele Frequency
WES	Whole exome sequencing
WTS	Whole transcriptome sequencing

CHAPTER 1

INTRODUCTION

Some of the first insights into the etiology of cancer came from the discovery of oncoviruses, such as rous sarcoma virus (Rous, 1911), and carcinogens, such as coal tar (Itchikawa & Baum, 1925). With the structure of DNA resolved in 1953 and increased understanding of how genes function in cells, focus shifted towards the oncogenic transformation of genes, and the identification of such genes, sarc for example (Stehelin, Varmus, Bishop, & Vogt, 1976), provided an understanding of cancer as a genetic process. The existence of these specific genes, termed oncogenes and tumor suppressor genes, implicated targeting these mechanisms as a promising therapeutic strategy. Among the earliest targeted therapies was imatinib for chronic myeloid leukemia (CML). A key oncogenic mechanism in CML was identified in 1960 to be a shortened chromosome 22, the result of a t(9;22) reciprocal translocation producing a constitutively active tyrosine kinase, BCR-ABL (Deininger, Goldman, & Melo, 2000; Hungerford & Nowell, 1960). Imatinib was developed to target this kinase, gained FDA approval in 2001, and was estimated to improve overall survival to 95% in a 5 year study (Druker et al., 2006).

The identification of genomic alterations central to the oncogenic process motivated in part the vast concerted effort of sequencing a complete human genome (Dulbecco, 1986). The idea gained traction in the 1980s, was formally launched as the human genome project in 1990, and was completed in 2003. This milestone coupled with the invention of faster and cheaper sequencing technologies (next generation sequencing, NGS) heralded in an explosion of cancer sequencing efforts (Figure 1.1) (Wheeler & Wang, 2013).

Through the immense amount of data offered by NGS technologies, we now understand that cancers arise from the accumulation of inherited and somatic genetic alterations that encompass several “hallmarks” necessary for neoplastic transformation (Hanahan & Weinberg, 2011; Vogelstein et al., 2013). Many of these studies assayed DNA, RNA, and methylation, revealing far greater complexity of the cancer genome than previously anticipated (McLendon et al., 2008). Further complexity is added by the fact that most tumors comprise of heterogeneous subpopulations, which has been experimentally observed through cytogenetic, Sanger sequencing, and NGS experiments (Landau, Carter, Getz, & Wu, 2014). As originally proposed by Nowell in 1976, these subpopulations then compete with each other for space and resources in Darwinian fashion, and the clones better equipped to survive and proliferate in the tumor’s microenvironment progress (Nowell, 1976). Thus, heterogeneity lends tumors an inherent plasticity, allowing them to adapt to changing conditions and combat host defenses and external therapy

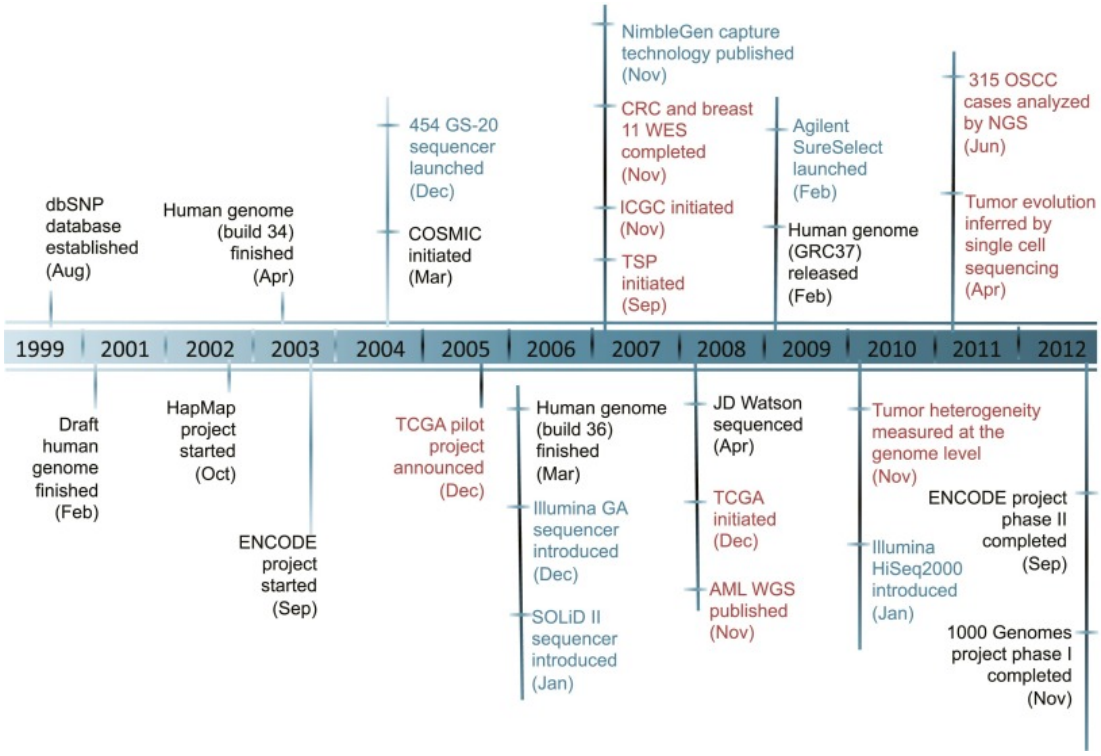


Figure 1.1. Major events in a decade of cancer genomics. Taken from Figure 1 of Wheeler and Wang's review on cancer genomics titled "From human genome to cancer genome: the first decade." (Dark blue) Major advances in massively parallel sequencing platforms and targeted enrichment technologies; (black) major large-scale projects designed to catalog genomic variations of normal human individuals; (red) cancer genomics. (dbSNP) Database of single nucleotide polymorphism; (HapMap) haplotype map of the human genome; (ENCODE) Encyclopedia of DNA Elements; (COSMIC) Catalog of Somatic Mutations in Cancer; (TCGA) The Cancer Genome Atlas; (GA) genome analyzer; (CRC) colorectal carcinoma; (WES) whole-exome sequencing; (ICGC) International Cancer Genome Consortium; (TSP) tumor sequencing project; (AML) acute myeloid leukemia; (WGS) whole-genome sequencing; (OSCC) ovarian small cell carcinoma (Wheeler & Wang, 2013).

(Landau et al., 2014). This principle is evident in the example of imatinib. Despite being an incredibly effective drug, imatinib is unable to eradicate all cancer cells and discontinuation of the drug often results in recurrence of the leukemia that has been attributed to residual cells that are not dependent on BCR-ABL for survival and can proliferate in the absence of the inhibitor (Corbin et al., 2011). Overall, understanding the complexity of cancer genomes at the level of molecular mechanisms underlying a disease and the heterogeneous presentations of these mechanisms within a tumor is paramount to defining diagnostic subtypes, predicting prognosis, developing new targeted therapies and combinations of therapies, and implementing therapeutic regimens leveraging knowledge of these mechanisms within a patient, i.e. precision medicine.

Investigating this complexity is a problem aptly addressed by NGS approaches. NGS has revolutionized biomedical research, enabling study of entire genomes, exomes, transcriptomes, epigenomes, epitranscriptomes, proteomes, metabolomes, microbiomes, and other “-omes” across large cohorts orders more quickly and cheaply than what was previously possible. The opportunities presented by the various types of NGS technologies for impacting human health as well as the challenges posed in implementing them are detailed in Chapter 2.

With respect to cancer, whole genome and exome sequencing are powerful for characterizing somatic single nucleotide variants, insertions and

deletions, and copy number variants. Although exome sequencing restricts data to coding exons of the genome, its lower cost and often increased interpretability is appealing. Mutations outside of protein coding parts of the genome are difficult to interpret since not enough is known about the function of these regions, but they likely have important consequences for regulation of gene and protein expression (Dunham et al., 2012). As such, the combination of exome and transcriptome sequencing is particularly powerful since it enables identification of protein coding variants and direct measurement of global gene expression changes that are downstream of more difficult to assay features such as non-coding variants and tumor microenvironment. Our work leveraging this strategy in mantle cell lymphoma (MCL) to study response and resistance to the combination targeted therapy of palbociclib and bortezomib (PALBOR) is discussed in Chapter 3.

The mutation profiles produced by genome and exome sequencing can be coupled with computational methods to infer cellular prevalence of variants, i.e. percentage of cells that contain a variant in a tumor sample (Roth et al., 2014). However single cell resolution is required to accurately infer intratumoral heterogeneity and subclonal compositions (Hughes et al., 2014; Navin et al., 2011). Further, heterogeneity at the transcriptional level is impossible to resolve with traditional transcriptome sequencing (RNA-seq), prompting the development of single cell RNA-seq technologies (scRNA-seq). These technologies have revealed new levels of heterogeneity previously

undetectable by bulk sequencing in a variety of healthy and disease tissues, suggesting that single cell resolution is necessary to accurately characterize complex tissue samples (Buganim et al., 2012; Dalerba et al., 2011; A. P. Patel et al., 2014; Shalek et al., 2013; Tirosh et al., 2016; Treutlein et al., 2014; Wagner, Regev, & Yosef, 2016). Comparative analyses of the different scRNA-seq technologies were recently published (Svensson et al., 2017; Wu et al., 2014). Parallel to the technological advances, several computational methods have been developed to characterize transcriptional heterogeneity and identify cell types (Brennecke et al., 2013; Grün, Kester, & van Oudenaarden, 2014; Haghverdi, Buettner, & Theis, 2014; Ji & Ji, 2016; J. K. Kim et al., 2015; Qiu et al., 2017; Trapnell et al., 2014; Welch, Hartemink, & Prins, 2016). Motivated by the successes and demonstrated utility of scRNA-seq, we used it to study purified MDS stem cells before and after hypomethylating treatment. Since mutations in the cell's splicing machinery are detected in half of all MDS cases (Bejar & Steensma, 2014), we were also interested in exploring single cell heterogeneity in alternative splicing and isoform expression. To date, few studies have been done on the topic, partly due to the fact that many of the single cell sequencing platforms assay only 3' ends of mRNAs as opposed to full-length transcripts and thus, do not allow isoform characterization. Early work confirmed the presence of splicing heterogeneity by identifying bimodal distributions of isoform ratios (Shalek et al., 2013), and more recent efforts either ignore these bimodalities in statistical

testing (Welch, Hu, & Prins, 2016) or are confounded by variable gene expression (Qiu et al., 2017). Chapter 4 describes our work in developing DISCO (Distributions of Isoforms in Single Cell Omics), a platform for analyzing alternative splicing in scRNA-seq data, and leveraging DISCO and other methods in analyzing the MDS transcriptome at the single cell level.

By leveraging integrative exome and transcriptome sequencing in MCL and scRNA-seq in MDS, we identified important mechanisms regulating pathogenesis and therapy response which can be tested in larger cohorts and implemented in the clinic in guiding treatment of these diseases. This effort was aided in both cases by longitudinal sampling of purified cancer cells.

CHAPTER 2

CLINICAL GENOMICS: CHALLENGES AND OPPORTUNITIES

PREAMBLE

This chapter was modified from a published review[1]. PV, AM, and SL planned content and wrote the manuscript. PV and SL made figure 2.1, PV made figures 2.2 and 2.3. All authors reviewed and edited the content.

INTRODUCTION

Next generation sequencing and other high-throughput approaches have become ubiquitous over the past few years, producing a deluge of new data at an unprecedented rate. However, incorporating novel insights from these data into clinical practice is not always obvious. Here, we review the current challenges of the field, specifically with respect to different genomic, transcriptomic, and epigenomic sequencing approaches and platforms, and important concepts for sequencing study designs that maximize statistical power and clinical utility. We also describe the current applications of these technologies across a range of topics including integrative study designs,

[1] Vijay P, McIntyre A, Mason CE, Greenfield J, Li S. (2016). Clinical Genomics: Challenges and Opportunities. Critical Reviews in Eukaryotic Gene Expression, 26(2):97-113. doi: 10.1615/CritRevEukaryotGeneExpr.2016015724

cancer genomics, precision medicine, and computational and machine learning analyses. Collectively, this provides a resource on experimental and computational methods to further enable researchers to leverage next generation sequencing (NGS) in their studies. These technologies have already improved medical interventions and will continue to transform medicine in the clinic and at a personal level by offering individuals increased agency in managing their health throughout a lifetime.

CHALLENGES

Sequencing is promising for clinical utility. The development of next-generation sequencing has lowered the cost of sequencing from 100 million dollars in 2001, to the 1000-dollar genome in 2014. The lowered cost has made sequencing more accessible to the medical community for diagnosis support (**Figure 2.1**). Sequencing can generate a wide variety of data types, which favors its use over other existing techniques, including PCR and microarrays. Emerging sequencing techniques from the past few years provide alternative approaches to traditional NGS platforms that produce shorter reads, such as the Illumina HiSeq sequencer, which are widely used for DNA-seq, RNA-seq, and bisulfite sequencing. Overall, genome, transcriptome, and DNA methylome sequencing can provide extensive clinical information for diagnosis and prognostic classifications.

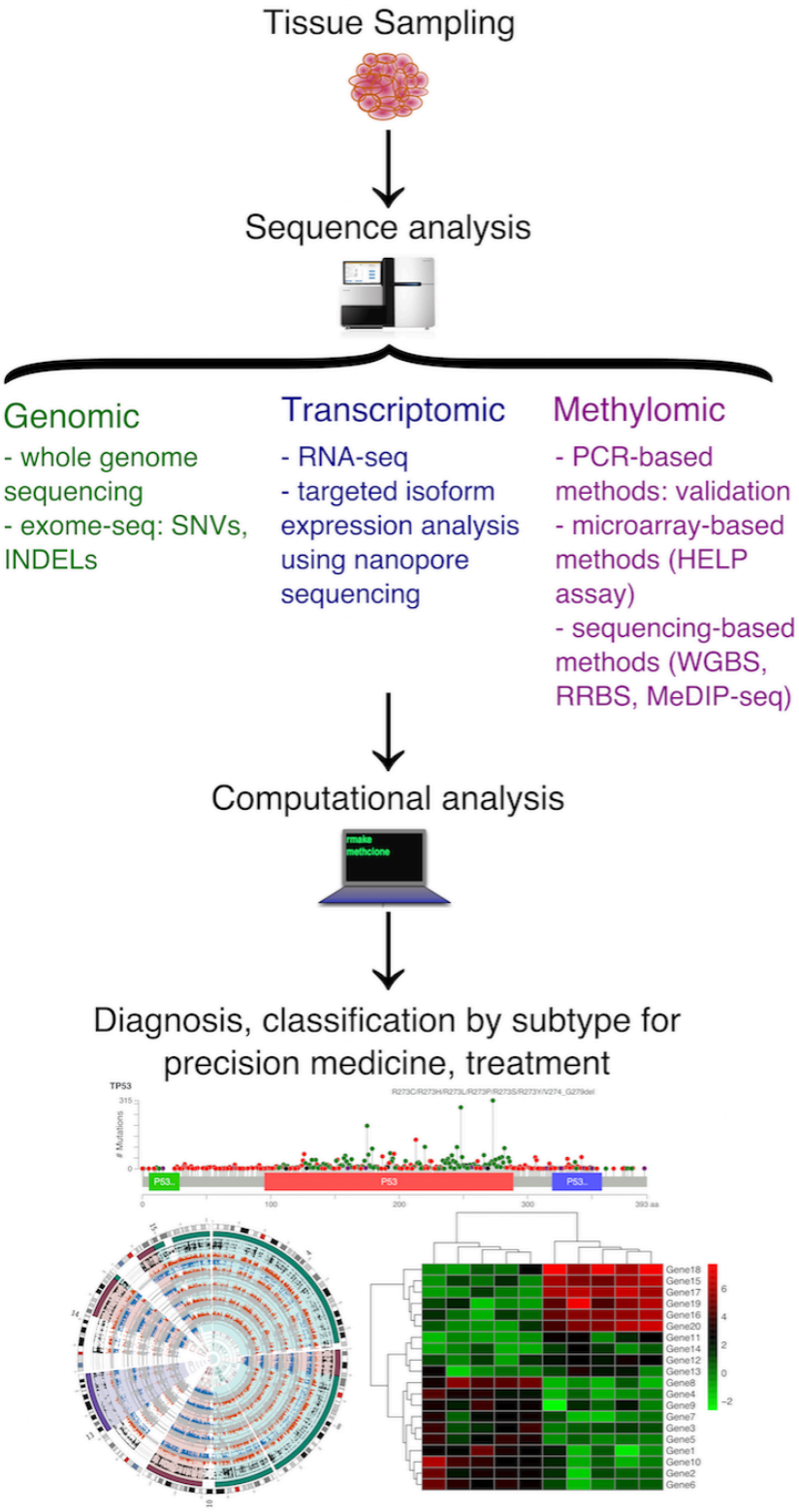


Figure 2.1. Omics workflows. Overview of genomic, transcriptomic, and methylomic sequence analysis workflows for disease characterization and precision medicine.

The advent of single molecule sequencing further decreases cost but significant challenges in data production and analysis remain. Both the PacBio RS and the MinION™ nanopore sequencer offer longer read lengths than other sequencing technologies, on the order of kilobases or tens of kilobases (Mason, Porter, & Smith, 2014). Pacific Biosciences released the PacBio RS sequencer in 2010, and although accuracy was initially poor at 86%, repeated sequencing of each read can increase accuracy to 99%. While specific bioinformatics tools have been developed over the past few years to cope with the error rate (Chaisson & Tesler, 2012), the high cost of the instrument has limited its adoption.

Oxford Nanopore Technologies first introduced the MinION™ sequencer in 2012, but only began distributing the sequencers to researchers through an early-access program in 2014. Unlike the PacBio RS, the MinION™ is highly portable at the size of a large USB stick, and requires a relatively small investment of around $1000. These features could help avoid the time and cost of sending samples to reference laboratories by bringing sequencing to clinics themselves, particularly in remote locations. However, nanopore technology is still in the nascent stages of development. Estimates have placed per-base error rates at 10-15% (Jain et al., 2015), which would need to improve drastically before nanopore sequencers could be considered a viable tool for many diagnostic applications. Efforts to demonstrate the clinical potential of the MinION™ have focused on pathogen identification and

characterization, including sequencing of the influenza virus (J. Wang, Moore, Deng, Eccles, & Hall, 2015), antibiotic resistance genes in *Salmonella enterica* serovar Typhi (Ashton et al., 2014), and the Ebola virus to identify transmission patterns during the recent outbreak (Gardy, Loman, & Rambaut, 2015). Researchers have also taken advantage of long read lengths to analyze isoform expression of alternatively spliced RNA (Bolisetty, Rajadinakaran, & Graveley, 2015). Current coverage depths allow for targeted studies of RNA isoforms but not whole transcriptome analysis. Yet, as the chemistry and analysis continue to evolve, nanopore sequencing shows increasing promise as an accessible and powerful means for evaluating patients and the pathogens that affect them.

DNA sequencing for clinical applications. There are many consortiums devoted to standardizing sequencing performance. Accuracy and reproducibility are the two key factors for sequencing technology to be widely used in clinical practice. DNA sequencing enables detection of germline and somatic mutations, which calls for comprehensive standardization. A cross-platform performance comparison of whole-genome sequencing revealed that 88.1% of SNVs detected were shared by Illumina and Complete Genomics. The concordance of insertion and deletion (indel) calling is much lower, with just 26% shared (Lam et al., 2012). Another current study comparing Illumina MiSeq, 454 GS Junior and Ion Torrent PGM from Life Technology for bacteria

genome shows that Illumina has the lowest error rate and has no homopolymer-associated indel errors (Loman et al., 2012). Whole exome sequencing (WES), which captures only genic regions, provides a more cost-efficient alternative to whole genome sequencing. WES has shown high accuracy for detecting Single Nucleotide Variants (SNVs) and short indels, although, when compared to high coverage WGS, WES shows limited power for CNV detection (Tan et al., 2014). A recent assessment of WES and exome array comparative genomic hybridization (CGH) using clinical samples showed that WES has the potential for clinical CNV detection, but currently, the combination of an array based approach with WES improves the accuracy of CNV calling, especially for intergenic regions and single-exon changes (Retterer et al., 2015). If using WES, the choice of exome-seq protocols affects the results. A comprehensive comparison between Agilent, Roche, and Illumina platforms showed varying strengths in the detection of variants across genic and untranslated regions (M. J. Clark et al., 2011a). Specifically, Nimblegen, the only platform that uses high-density overlapping baits, has higher sensitivity in variant detection. A concurrent study also confirmed that the Nimblegen platform has higher coverage of exonic regions, with at least 20x coverage (Sulonen et al., 2011). The Agilent and Illumina platforms, however, target a wider range of genomic regions, and with deeper sequencing, these two platforms detect more variants (M. J. Clark et al., 2011b). Another advantage of Illumina's capture method is that it provides

coverage for untranslated areas, which might be of interest to researchers who would like to include noncoding variants in their analyses. For an even more targeted and affordable method than WES, specific cancer panels are commonly used. These require prior domain knowledge, which include recurrent genetic or epigenetic lesions. Recurrent somatic mutations have been identified in many cancer types and used to predict risk levels of the disease (Ding et al., 2012). In acute myeloid leukemia (AML), 15 biomarkers have been used to further stratify patients who were previously all placed in the intermediate risk group by cytogenetic classification. This helps to develop treatment plans for AML patients tailored to the risk for each group (J. P. Patel et al., 2012). Indeed, targeted sequencing provides a much deeper view of the known genes and hotspots for mutations. However, with ever decreasing sequencing cost and as more possible drug targets are identified for clinical treatments, exome-seq covering larger areas of the genome has the potential to be more widely applied in clinical diagnosis and prognostic decisions.

RNA sequencing is a promising candidate for clinical applications (Van Keuren-Jensen, Keats, & Craig, 2014). This technique enables whole transcriptome examination, including detection of gene expression, alternative isoforms, fusion genes, and expressed variants (S. Li, Tighe, et al., 2014). However, RNA sequencing is also very sensitive to systematic bias (S. Li,

Łabaj, et al., 2014; Risso, Ngai, Speed, & Dudoit, 2014). Previously, we and others have defined multiple quality metrics that flag samples with potential gene expression quantification issues, including gene body coverage evenness, GC content, insert size, and base error rate (Risso, Schwartz, Sherlock, & Dudoit, 2011). In the FDA-led Sequencing Quality Control (SEQC) study for RNA-seq performance evaluation, gene body coverage evenness, GC content, and insert size have been shown to be related to library preparation, and base error rate dependent on the sequencer used (S. Li, Łabaj, et al., 2014). Multiple software packages for gene expression normalization were compared. EDAseq, which corrects for both the intra-group variations and quantification bias caused by GC-content and gene length, was found to perform the best for accurate differential gene expression analysis (Risso et al., 2011). PEER and sva have shown higher potency in detection of latent variables for the quantification of gene expression among different sites of sequencing data (Stegle, Parts, Piipari, Winn, & Durbin, 2012). For a statistically powerful RNA-seq study design, it is always recommended to use consistent experimental strategies, including sequencer, read length, sequencing depth, and protocol (Z. Su et al., 2014). Studies have shown that sequencing depth is critical for discovery of new genes and accurate gene expression profiling (Toung, Morley, Li, & Cheung, 2011). Later follow-up studies focused on differential gene expression analysis have shown that increasing the biological replicates more efficiently improves the accuracy of

gene quantifications (Rapaport et al., 2013). Therefore, experimental design for RNA-seq analysis is critical for accurate differential gene expression analysis.

DNA methylation provides a complementary approach to clinical measures for patient classification. DNA methylation is the addition of a methyl group to the 5th position of cytosine, which has the specific effect of suppressing gene expression. DNA methylation has been defined as one of the hallmarks of cancers and aging (Rodríguez-Paredes & Esteller, 2011; Teschendorff et al., 2010). Many different types of cancers have consistently shown the dysregulation of DNA methylation (Figueroa et al., 2010; Mack et al., 2014; Noushmehr et al., 2010; Sandoval et al., 2013). The Cancer Genome Atlas (TCGA) consortium and many other research studies have shown that cancers can be classified based on their degree of DNA methylation (Noushmehr et al., 2010; Weisenberger, 2014). Subgroups of many cancers exhibit CpG island methylator phenotype (CIMP), including breast cancer (Fang et al., 2011), brain cancer (Mack et al., 2014; Noushmehr et al., 2010; Pajtler et al., 2015), blood cancer (Figueroa et al., 2010), gastric cancer (Bass et al., 2014), liver cancer (Shen, 2002), and lung cancer (Sandoval et al., 2013). Groups of patients classified based on DNA methylation patterns show distinct clinical outcomes, including overall survival and disease free progression (Figueroa et al., 2010; Noushmehr et al., 2010).

The CIMP positive group has been shown to differentiate and stratify patients into groups with distinct clinical outcomes. For example, in glioblastoma patients, a CIMP positive phenotype is usually associated with distinct copy number changes, appears exclusively in the proneural subtypes, and is associated with IDH1 mutations and improved clinical outcomes (Noushmehr et al., 2010). In a recent study of ependymoma, which is the third most common pediatric brain tumor, researchers showed that CIMP positive patients with posterior fossa ependymoma have worse clinical outcome than CIMP negative patients (Mack et al., 2014; Pajtler et al., 2015). The genetic background of CIMP positive patients presents a blended picture and indicates the importance of DNA methylation as an alternative approach for patients risk stratifications (Mack et al., 2014).

There are many advantages to using DNA methylation analysis for clinical profiling. First, it does not rely heavily on the genetic alterations of the diseases, and so, it can be applied to diseases where there are sparse somatic mutations. Second, the material under analysis is DNA, which is advantageous because DNA is less sensitive to heat or enzymatic degradation than RNA, resulting in more accurate profiling.

Several methods have emerged to quantitatively measure DNA methylation, grouped here into three categories: (1) PCR-based methods, (2) microarray-based methods, and (3) sequencing-based methods. The PCR-based methods are usually used as a validation approach for high-throughput

quantification. The second are microarray-based methods. The HpaII tiny fragment Enrichment by Ligation-mediated PCR Assay (HELP Assay) was a commonly used regional DNA methylation quantification approach for research and clinical sample profiling (Pan et al., 2012). It is based on the restriction enzyme Hpall's ability to exclusively recognize and cleave methylated CpG DNA sites. Another commonly used single base resolution microarray-based DNA methylation quantification approach is Illumina Infinium BeadChip Kit. The BeadChip array platform utilizes two different bead types to measure DNA methylation levels at single cytosine. Infinium HumanMethylation450 BeadChip Kit (450K array) is one of the Infinium Kits that cover the most number of methylation sites for human samples (485,000 sites). It covers 99% of the RefSeq genes, which, on average, have 17 CpG sites per gene. The 450K array has been widely used in DNA methylation quantification over the past few years, with more than 10,000 entries in the GEO database, providing a valuable international resource for comparison between different cohorts of patient samples (Lowe & Rakyan, 2013).

The third class of methods are sequencing-based, and can either provide single base resolution or regional quantifications of DNA methylation levels (Down et al., 2008; Frommer et al., 1992; A. Meissner, 2005). The single base resolution methods mainly use bisulfite conversion sequencing, where the cytosine of a CpG site without methylation will be converted into uracil while the cytosine with methylation will remain as a cytosine. So, in the

final sequencing readout, a CpG site without methylation will be thymidine instead of cytosine (Frommer et al., 1992; A. Meissner, 2005). Eventually, the single site CpG methylation level will be calculated based on the percentage of reads with cytosine among the total number of reads mapped. These methods include whole genome bisulfite sequencing (WGBS) (Frommer et al., 1992), reduced representation bisulfite sequencing (RRBS) (A. Meissner, 2005), and targeted methylation sequencing (TMS) (Deng et al., 2009). WGBS requires a large amount of sequencing depth, as each base in the whole genome needs to be covered by at least 4 reads in order to achieve accurate quantification. This enables the inclusion of regions with both high and low CG density. However, RRBS and TMS only cover a subset of regions in the genome, providing a cheaper alternative, and enable accurate quantification of about 15% of CpG sites from higher CG density regions, including CpG islands, and promoter regions (Deng et al., 2009; A. Meissner, 2005). This makes it possible to profile more patients with regions that are of particular interest in transcriptome regulation. The regional level quantification approaches mainly use affinity-based DNA methylation sequencing, such as methylation DNA immunoprecipitation sequencing (MeDIP-seq) (Down et al., 2008). This approach is similar to CHIP-seq, using antibodies that recognize genomic locations with methylated CpGs. Comparison of the 450K array and whole genome sequencing approaches showed decent correlation (C. Clark et al., 2012). The 450K array was also shown to generate highly reproducible data

between seven technical replicates of clinical samples (Pan et al., 2012). But for clinical utility, a large effort comparing different platforms remains to be done.

Computational Analysis of Multi-Omics Data. Perhaps one of the biggest challenges in going from bench to bedside in sequencing studies is the accurate and reproducible analysis of the resulting terabytes of data. Sequence data analysis is a multistep process that often needs to be adapted for a specific experiment or scientific question. For any sequencing data, computational analysis generally begins with aligning reads to a reference genome (**Figure 2.2**). Commonly used programs include BWA for DNA reads, STAR for RNA-seq data, and Bismark, BSMAP, or BSmapper for methylation data. The choice between these programs depends on such factors as sequencing platform, read length, and desired SNP tolerance, with various programs optimized for different read characteristics (Mielczarek & Szyda, 2015). After mapping to the genome, analysis depends on the scientific question, with specific programs designed broadly to call variants, identify differential expression, or quantify the extent of methylation. When multiple data types are available, their integration can identify a network of interacting and interdependent processes contributing to disease states. Indeed, clustering patient samples using models that computationally combine different data types has revealed novel subtypes unseen when evaluating a

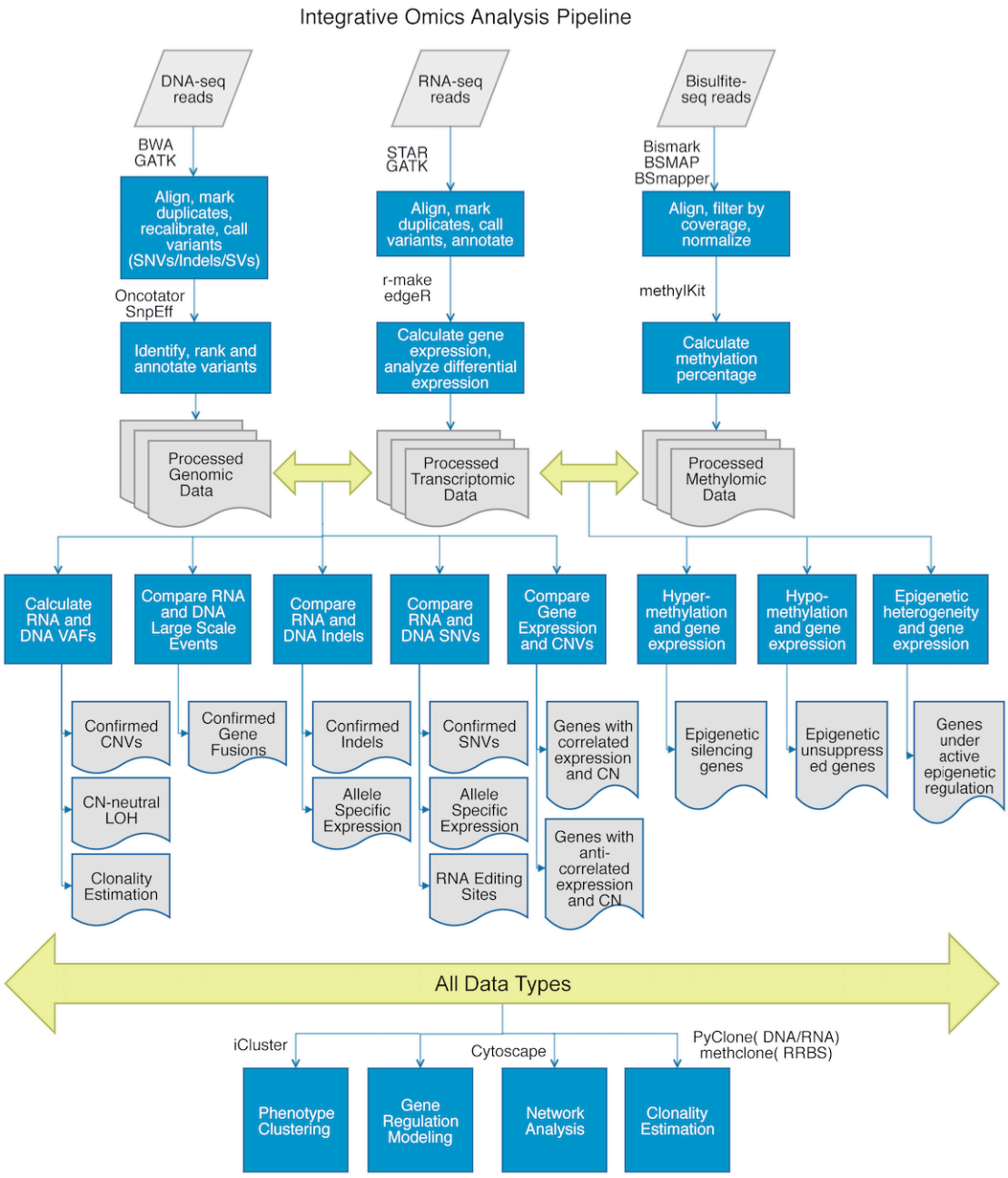


Figure 2.2. Integrative omics analysis pipeline. Computational methods for genomic, transcriptomic, and methylomic data analysis for identifying variations at all levels and integrating different data types for increased confidence.

single data type (Mo et al., 2013; M. Su, Dou, Cheng, & Han, 2015). Despite challenges in cost, cross-platform comparisons, technical standards, analysis methods, etc., advances in massively parallel sequencing techniques present new opportunities to improve clinical research, which we explore in the next section.

OPPORTUNITIES

Leveraging Electronic Health Records Data. Genomics and informatics are being incorporated into many aspects of patient care, especially with the transition to electronic health records (EHR). Despite the relatively new shift to EHR, the data is already being leveraged for large-scale studies using machine learning and data mining methods, as it offers unprecedented access to large sample sizes and diverse patient cohorts. These studies include mining for adverse drug effects (G. Wang, Jung, Winnenburg, & Shah, 2015), and developing a classifier for disease phenotype severity (Boland, Tatonetti, & Hripcsak, 2015). The implications of this transition to EHR for clinical genomics, including genetic testing, have been previously reviewed (Marsolo & Spooner, 2013).

Genomics and Chronic Illnesses. Genomic approaches are becoming important for both preventing and managing chronic illnesses, such as diabetes and inflammatory bowel disease. The human microbiome project and

other metagenomic studies have revealed the importance of healthy gut microbiota, which has now been translated to clinical practice through fecal microbiota transplants for treating *Clostridium Difficile* infections, ulcerative colitis, Crohn's disease, and other digestive illnesses (Drekonja et al., 2015; Mandal, Saha, & Das, 2015).

Personalized Healthcare and Direct-to-Consumer Genomics. Statistical models can incorporate genomic features and family history, coupled with factors such as age, weight, and ethnicity, for disease risk prediction in healthy individuals. This has been especially useful for early intervention in individuals at high risk for diabetes and cardiovascular disease. Clinical genomics platforms such as Foundation Medicine, Ingenuity, and Personalis facilitate implementation of genetic testing in clinical platforms (C. J. Patel et al., 2013). As of August 2015, the NIH's genetic testing registry catalogued 28,542 tests spanning 4,726 genes for the purpose of diagnosing any of 9,927 conditions. These not only include classical Mendelian diseases, such as Huntington's chorea, but also predict predisposition to complex diseases, such as Alzheimer's, and drug response, for example sensitivity to the anticoagulant Warfarin. With the prevalence of direct-to-consumer tools like 23andme and ancestry.com that make this type of information accessible to interested individuals, people are more empowered than ever to advocate for their own health. Computational methods utilizing patients' genetic information, coupled

with EHR in some cases, for disease risk prediction are actively being researched (L. Li et al., 2014). Federal policies have been changing to reflect the shift to clinical genomics, as evidenced by the 2013 shutdown of 23andme's genetic testing arm and the 2015 repeal of the ban, and by the 2013 landmark supreme court case that barred the previously common practice of patenting genes (Rosenfeld & Mason, 2013). Other challenges to consider are the legal and ethical issues surrounding genetic testing in children and adolescents, previously reviewed by Botkin *et al.* (Botkin et al., 2015).

Genomics and Cancer. Despite challenges, genomics has introduced a paradigm shift in medicine, especially in the treatment of cancer. Where historically cancer has been categorized by the tissue type it affects, it is now increasingly being defined by genetic alterations. The vast breadth of knowledge gained from large national and international cancer sequencing efforts, mainly The Cancer Genome Atlas and the International Cancer Genome Consortium, has immeasurably increased our understanding of the genetic mechanisms, molecular subtypes, and heterogeneity of cancers (Ciriello et al., 2013; Hudson et al., 2010). These data have been made easily accessible to the scientific community. Tools like the cBio Portal, for example, allow anyone to query the mutation load of any given gene in all assayed cancer types. Thus, cancer genomics is continuously being translated to

clinical settings (Cerami et al., 2012). One such case is in recurrent Mantle Cell Lymphoma, where the authors utilized an integrative genomics and transcriptomics approach coupled with extensive functional studies to attribute the cause of relapse after Ibrutinib treatment to a relapse-specific single nucleotide variation in the drug target, BTK (Chiron et al., 2014). This can now be incorporated into the therapeutic decision-making pipeline by testing for this BTK mutation. Similar efforts in a wide variety of cancers have categorized subtypes of cancers based on genetic information, and these classifications are actively used in diagnosis, prognosis, and therapeutics.

A classical success story of the use of genomics in cancer therapy is of the BRAF inhibitor vemurafenib in metastatic melanoma. Genomic screening of metastatic melanoma patients identified BRAF V600 mutations in half of all patients that increased the sensitivity of cancer cells to BRAF inhibitors (Chapman et al., 2011). One of the common challenges of targeted therapies, however, has been the development of resistance, which was seen in cases of melanoma treated with BRAF inhibitors. Combinations of drugs as opposed to monotherapies lower the risk of resistance and relapse. For example, dabrafenib in combination with trametinib was found to prolong progression-free survival and increase response rates in BRAF V600 melanoma compared to monotherapy (Robert et al., 2014).

Combination therapies are often found to perform more successfully, as developing resistance is less likely. Computational methods for predicting

effective drug combinations alleviate the enormous cost of exhaustive experimental testing in every cancer model. Instead, these machine learning methods can use data from cell line assays as training sets and predict successful combinations for genetically defined subtypes that can then be tested in patient-derived xenograft models (Costello et al., 2014). Some of the experimental data sets currently available for use in computational models are the NCI 60 cancer cell line and drug screening data (Shoemaker, 2006), NIH's Library of Integrated Cellular Signatures (LINCS), and the Broad Institute's connectivity map (Lamb et al., 2006). By modeling drug-gene interactions coupled with the genomic alterations of a patient's tumor, doctors are now able to predict the efficacy of different chemotherapies or targeted therapies in a personalized manner. These not only include rule based decision tree methods, but also more complex computational models. In addition to predicting the efficacy of combination therapies, computational methods for drug repositioning are also continuously gaining popularity and producing effective therapies (Shameer, Readhead, & T. Dudley, 2015).

Since many of these drug development and prediction approaches rely on accurate and detailed patient stratification based on genomic data, whole genome, exome, and transcriptome sequencing are more and more routinely performed for clinical samples, either at time of collection for rapid turnaround or banked for future analysis. This vast amount of sequencing data has also enabled better prognosis assessment in many cases, although this is not new

to the sequencing era. By 2000, microarrays were being used for molecular stratification of cancer samples that identified gene signatures defining differential survival (Perez-Diez, Morgun, & Shulzhenko, 2000). Unsurprisingly, the advent of next generation sequencing methods increased studies in this vein.

Even with applications to all aspects of human health and disease, cancer remains the one disease (really an innumerable collection of diseases) where genomics has had the biggest impact. This is owed to the genetic nature of cancer, since cancers arise from the accumulation of inherited and somatic genetic alterations (Vogelstein et al., 2013). Heterogeneous subpopulations comprising tumors have been experimentally observed through cytogenetic, Sanger sequencing, and next generation sequencing experiments (Landau et al., 2014). As originally proposed by Nowell in 1976, these subpopulations compete with each other for space and resources, and the clones better equipped to survive and proliferate in the tumor's microenvironment will progress (Nowell, 1976). Genomics enabled researchers to assess the compositions of tumors and infer the molecular characteristics of distinct subpopulations.

The main challenge in accurately inferring heterogeneity and clonal evolution is that most tumor profiling methods involve a bulk sample of cells, effectively masking intratumoral variability. With novel technological developments in single cell sequencing, we can now measure these

subpopulations directly and at a previously unprecedented resolution. Single cell sequencing will add a new level to clinical applications of tumor sequencing by increasing the resolution with which we can model complex tumor dynamics and incorporate that into prognosis assessment and drug efficacy prediction (**Figure 2.3**). The development of single cell sequencing methods, especially single cell RNAseq, which has been previously used in immune cells (Shalek et al., 2013), breast cancer (Navin et al., 2011), melanoma circulating tumor cells (Hou et al., 2013), and glioblastoma (A. P. Patel et al., 2014), has addressed this issue. Each of these cases revealed new levels of heterogeneity that are undetectable in bulk samples, suggesting that single cell resolution is necessary to accurately characterize complex tissue samples.

An added benefit is that all of these sequencing data are submitted to curated repositories with publication, such as the database of Genotypes and Phenotypes, the Sequencing Reads Archive, and the Gene Expression Omnibus. This helps alleviate the problem of small sample sizes common in clinical settings and/or rare diseases. Researchers interested in any of this publicly available data can download it and apply their own analyses. For those unfamiliar with computational and bioinformatics methods, there are also pipelines with guided user interfaces that facilitate these steps, such as STORMseq (Karczewski et al., 2014), Genesifter, Ingenuity variant analysis software, and more. There is also current research in software design for use

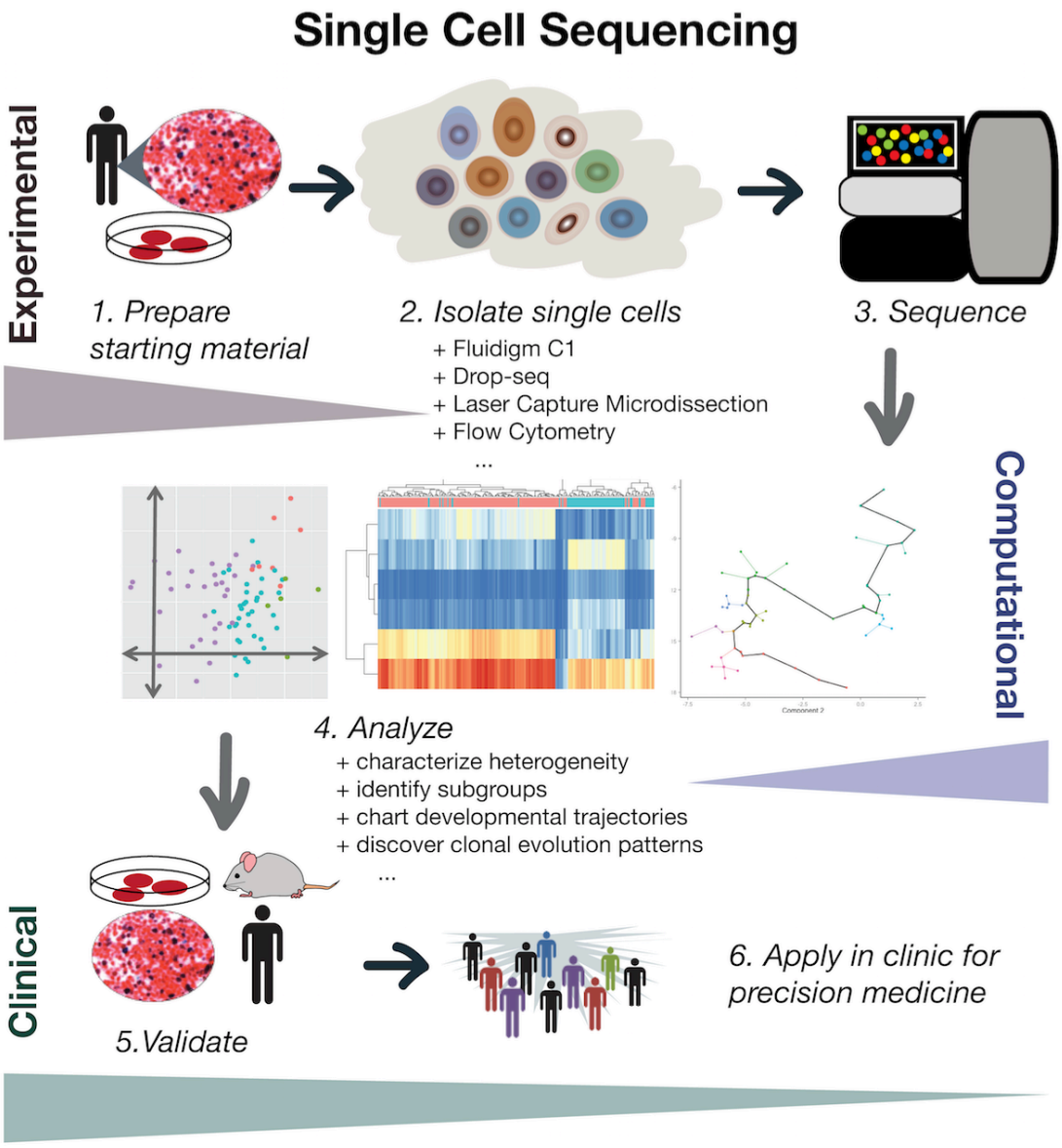


Figure 2.3. Conceptual overview of single cell sequencing for clinical applications.

by non-computational clinical scientists (Shyr, Kushniruk, van Karnebeek, & Wasserman, 2015). Although the data repositories in place are serving a much-needed purpose, there are opportunities here for better infrastructure, support for IRB approvals, ease of submission, and ease of access.

Advances in Genomics Approaches for Neurobiology

Molecular stratification with genomic sequencing advising patient therapy is not limited to cancer drugs. Although less understood, genomic approaches are also used in neurobiology, especially in the study of Alzheimer's and autism spectrum disorders (Han et al., 2014). With large-scale efforts in mapping the human brain using cutting edge brain imaging techniques, big data approaches are becoming increasingly useful in understanding neurodegenerative diseases. Understanding mutations and predispositions to these diseases would allow for early intervention, which is often the only hope for therapy.

National and International Personalized Medicine Initiatives. Overall, clinical genomics has pervasively affected human health and disease, especially in the field of oncology. The paradigm shift in the understanding and treatment of cancer is mirrored by federal policy changes, most notably though President Obama's Precision Medicine Initiative, announced in his 2015 State

of the Union Address. This initiative includes increased funding to the National Cancer Institute for researching genomic drivers in cancer and for streamlining the design and testing of targeted therapies based in genetics. Relatedly, the prototype of clinical trials is transforming to better reflect the shift to personalized medicine as seen by the success of the IMPACT and following IMPACT2 studies. It is important to note that these changes in clinical genomics are happening on a global scale, inspiring international cooperation to advance medicine (Manolio et al., 2015).

CHAPTER 3

LONGITUDINAL GENOMIC AND TRANSCRIPTOMIC ANALYSIS OF MANTLE CELL LYMPHOMA IN A TARGETED COMBINATION TRIAL OF A SELECTIVE CDK4/6 INHIBITOR

PREAMBLE

This chapter is modified from a manuscript in preparation[2]. PV performed computational analysis, generated figures, and wrote text with input from SCK, MDL, XH, and CEM. SE generated immunohistochemistry images. All authors reviewed data and content. SCK and CEM conceived the project.

INTRODUCTION

Mantle cell lymphoma (MCL) is a rare subtype of B-cell non-Hodgkin's lymphoma with a median survival time of 5 years (Pérez-Galán, Dreyling, & Wiestner, 2011). A defining characteristic of MCL is a t(11;14)(q13;q32) translocation resulting in aberrant expression of cyclin D1, a cell cycle protein normally undetected in B-cells (Pérez-Galán et al., 2011). Cyclin D1 complexes with cyclin dependent kinases CDK4 and CDK6 to drive G1/S

[2]Vijay P, Di Liberto M, Huang X, Ely S, Blecua P, Chiron D, Elemento O, Martin P, Leonard JP, Mason CE, Chen-Kiang S. Longitudinal genomic and transcriptomic analysis of mantle cell lymphoma in a targeted combination trial of a selective CDK4/6 inhibitor. (In preparation).

transition. Although CDK6 is silenced in MCL, CDK4 is expressed, and previous work has implicated functional inhibition of CDK4 as a therapeutic strategy for MCL (Di Liberto et al., n.d.; Leonard et al., 2012; Marzec et al., 2006).

Palbociclib is a selective CDK4/6 inhibitor that received accelerated FDA approval for doubling progression-free survival in metastatic breast cancer in combination with letrozole in a phase II trial (Baughn et al., 2006; Finn et al., 2014). It has also been applied as a single agent in MCL achieving remarkable efficacy in cell lines and patients (Leonard et al., 2012). However, as resistance is common for single agent cancer therapies, a combination approach of palbociclib in conjunction with the proteasome inhibitor bortezomib was tested and found effective in myeloma xenografts. Palbociclib was shown to induce reversible, prolonged G1 arrest (pG1), which consequently sensitized cells to cytotoxic killing by bortezomib (Huang et al., 2012). This combination approach (PALBOR) was clinically tested in a phase I trial in recurrent MCL with favorable results (Di Liberto et al., n.d.). Here, we leveraged integrative DNA and RNA sequencing of 6 patients from the trial, 3 responders (R) and 3 non-responders (NR) in order to identify mutation profiles of different patients, genes involved in determining sensitivity and resistance, and longitudinal shifts of subclonal composition of mutations during one cycle of treatment.

Previous work in MCL genomics has identified several recurrent mutations (Beà et al., 2013; Colomer & Campo, 2014; B. Meissner et al., 2013; Rahal et al., 2014). But, it is unclear how these mutations affect response to palbociclib or how palbociclib affects cellular processes at the transcriptional level. A longitudinal examination of both DNA and RNA in patients undergoing identical treatment protocols, as shown here, is rare in MCL and other cancer sequencing efforts. For normal controls, we sequenced resting and activated peripheral B cells from healthy donors. With this unique data set, we have discovered multiple mechanisms potentially explaining resistance in non-responders and acquired resistance in a relapse patient, pointing to future avenues of investigation for improving precision medicine approaches in combination therapy of palbociclib and bortezomib.

RESULTS

Integrative genomic and transcriptomic examination of responders and non-responders. Normal cell cycle progression through G1/S transition necessitates the dimerization of D type cyclins with CDK4/6, which then inactivate retinoblastoma protein (Rb) by phosphorylation (**Figure 3.1A**). Palbociclib selectively and potently inhibits CDK4 at precisely early G1 phase, and in the presence of functional Rb, causes cells to arrest and synchronize in G1 (prolonged G1 arrest, pG1). Cell cycle is resumed with the discontinuation of the inhibitor as cells progress through G1/S (**Figure 3.1A**) (Baughn et al.,

2006; Huang et al., 2012; Leonard et al., 2012). The deregulation of gene expression profiles as a consequence of pG1 is known in cell lines to improve the efficacy of the proteasome inhibitor bortezomib (Huang et al., 2012). A phase I clinical trial testing this combination (PALBOR) in patients with previously treated recurrent MCL was undertaken. One cycle of therapy lasts 21 days with palbociclib being administered for the first 12 days, inducing pG1 by day 8 when bortezomib treatment is started and repeated on days 11, 15, and 18 (**Figure 3.1B**). Results of the clinical trial are detailed in Di Liberto and Martin *et al*. (Di Liberto et al., n.d.).

Using lymph node biopsies from patients, we performed a longitudinal examination of DNA and RNA changes during the first treatment cycle. For DNA, we used whole exome sequencing (WES) to assay copy number variations and single nucleotide variations before treatment (D1) and the end of the first cycle (D21). For RNA, we used whole transcriptome sequencing (WTS) prior to treatment (D1), after palbociclib when the cells are in pG1 phase (D8), and the end of the first cycle (D21) (**Figure 3.1C**). We did not perform WES at D8 because cellular populations are expected to be similar between D1 and D8 as cells are not yet killed. The D21 time point not only reveals the effects of one cycle of the combination therapy, but also represents the population of cells entering the next cycle and may be used to predict response in the next cycle. This high quality sequencing data following 6 patients with different responses to identical treatment regimens enabled

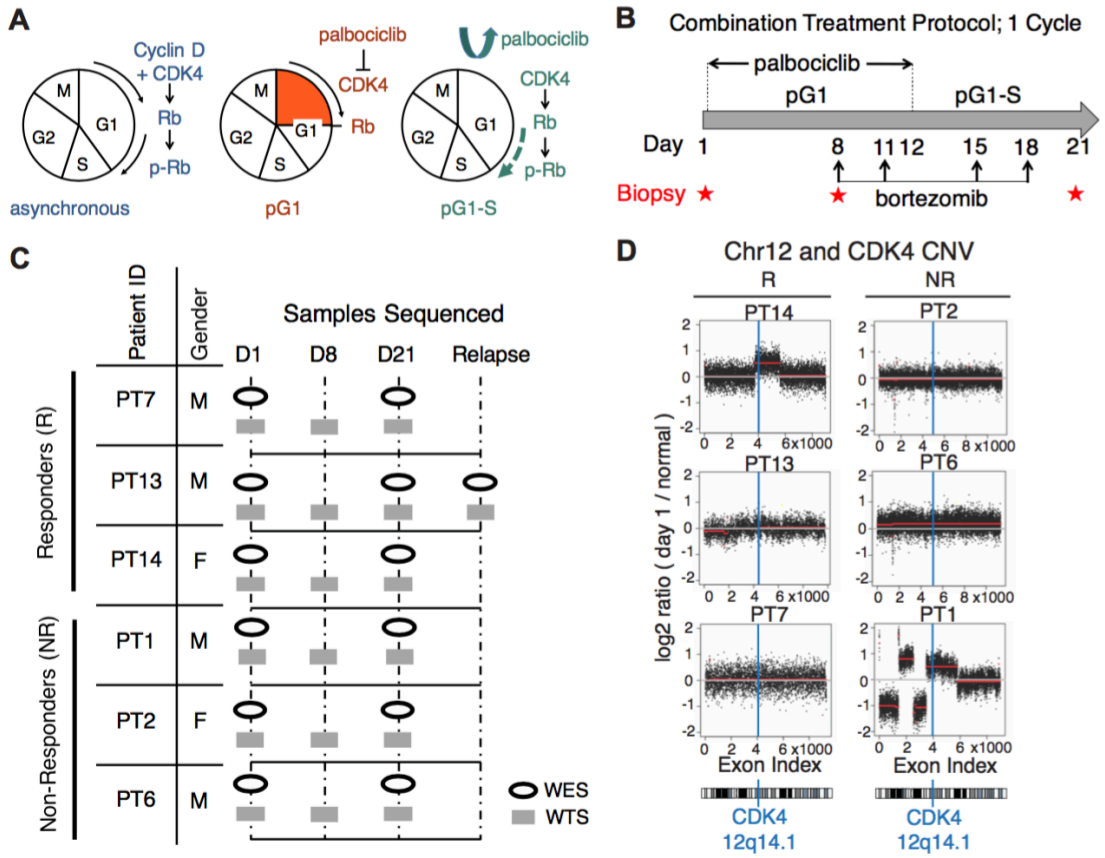


Figure 3.1. PALBOR combination therapy protocol and patient information. A. Molecular mechanism of palbociclib's reversible effect on cell cycle progression. **B.** Clinical protocol for one cycle of palbociclib and bortezomib combination therapy. **C.** Patient lymph node samples with whole exome and whole transcriptome sequencing comprising 3 responders and 3 non-responders. **D.** Chromosome 12 amplifications affecting the target of palbociclib, CDK4, found in complete responder PT14 and non-responder PT1. pG1, prolonged G1 arrest; pG1-S, synchronous progression through S phase; WES, whole exome sequencing; WTS, whole transcriptome sequencing; R, responder, i.e., had greater than 50% reduction of tumor volume; NR, non-responder.

identification of pathways likely involved in conferring innate resistance in non-responders, and acquired resistance in relapse by using an integrative analysis pipeline to study clonal architecture and evolution during therapy of copy number and single nucleotide variants. Matched normal tissue samples were available for a subset of patients (responders PT7, PT13, and PT14), allowing us to confidently isolate MCL-specific mutations in these patients. For the remaining patients (PT1, PT2, and PT6), matched normal tissue was unavailable, and therefore we were unable to establish somatic status of mutations in these patients, an issue we approached differently for CNVs and SNVs, as described in the respective sections below.

CDK4 amplifications do not affect response to PALBOR. Chromosome 12 amplifications spanning CDK4, the target of palbociclib, were present at baseline in complete responder PT14 and non-responder PT1, with similarly high expression of CDK4 mRNA across all MCL patients (**Figure 3.1D**). This suggests that palbociclib remains a therapeutically effective CDK4 inhibitor even in the presence of genomic amplifications of the target.

Multiple large-scale copy number variations differentiate responders from non- responders. CNVs were detected using 3 parallel methods: xhmm, VarScan 2 copycaller with circular binary segmentation using DNAcopy, and genome-wide SNP allele frequencies (Fromer & Purcell, 2014; Koboldt et al.,

2012; Venkatraman & Olshen, 2007). Both xhmm and SNP allele frequencies allow for CNV characterization in the absence of normal controls. For DNAcopy, which uses the log ratio of normalized read depth between tumor and normal, we used the average of the read depth in normal tissues as the control for the patients lacking matched normal samples. Consensus CNVs across the different methods were deemed 'high confidence' and calculated for all samples. Genome-wide visualization of these CNVs at D1 revealed 2 main points: (1) all genomes in the cohort have several large scale (>1 Mb) copy number aberrations, and (2) a subset of these CNVs are specific to either the responder group, defined as patients with partial (PT7, PT13) or complete response (PT14), or the non-responder group, defined as patients with stable (PT1) or progression disease (PT2, PT6), and may indicate mechanisms of sensitivity or resistance (**Figure 3.2**).

Non-responder (NR) specific deletions, i.e. hemizygous deletions found in at least 2/3 non-responders and 0/3 responders, span 99 megabases and a total of 760 genes. These genes are significantly enriched for gene ontology processes involving type I interferon signaling, which includes Gene Ontology (GO) terms for STAT phosphorylation, B cell proliferation, and a host of other immune related functions (**Table 3.1**) (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009). The enrichment of these pathways can largely be attributed to a chromosome 9 deletion spanning 16 interferon genes. Other cancer genes in NR-specific deletions, selected for their overlap with the COSMIC Cancer

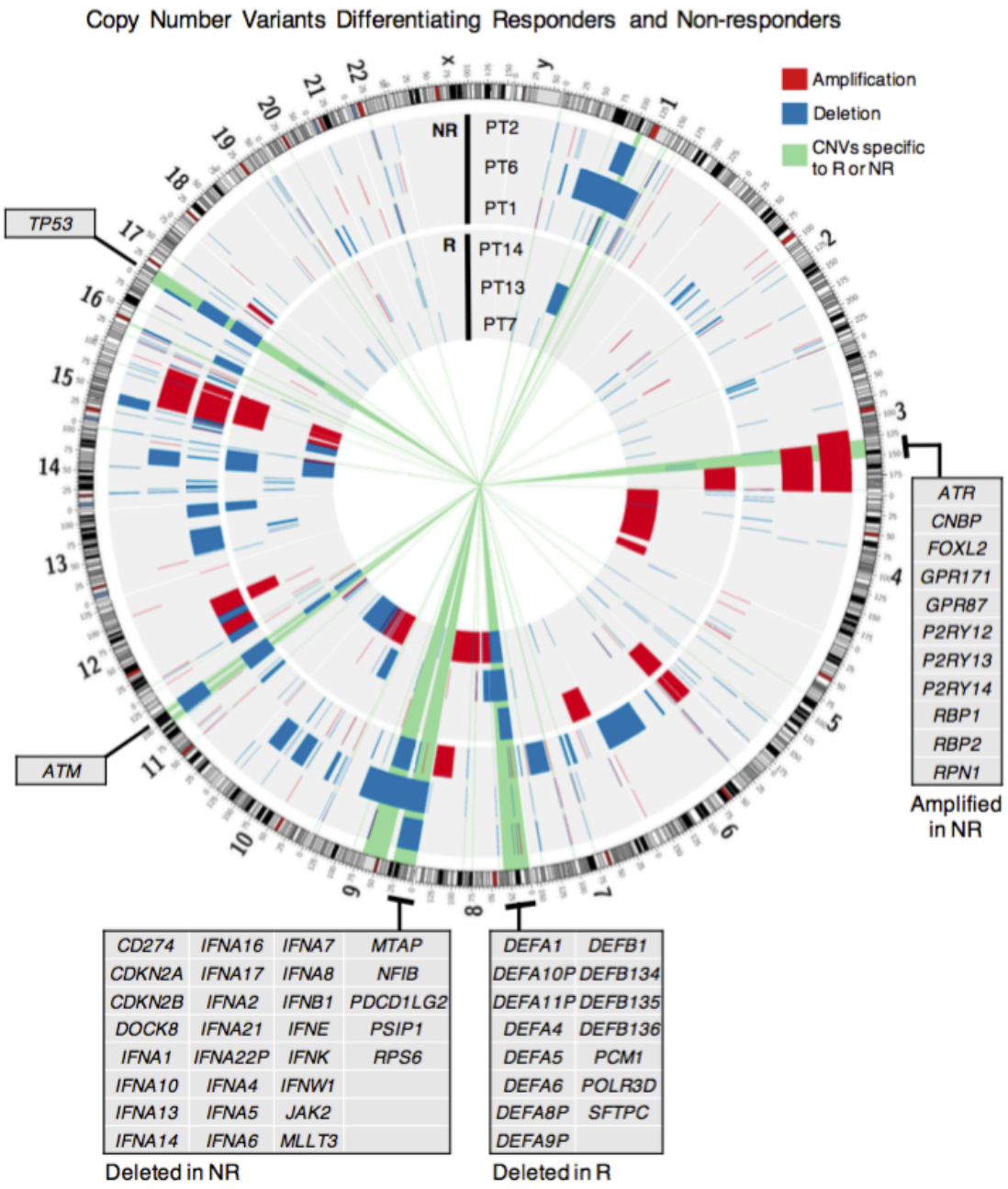


Figure 3.2. Copy number variations. Genome-wide visualization of CNVs before treatment (D1) where blue is deletion and red is amplification. The 3 outside tracks show non-responders (NR) and the 3 inner tracks show responders (R). Green highlights refer to regions of CNVs specific to R or NR (i.e. shared by at least 2/3 of one group and 0/3 of the other) with genes contributing to enriched Gene Ontology (GO) terms or known to be causally mutated in cancer from the COSMIC database shown in tables.

Table 3.1. Gene ontology terms enriched in large-scale CNVs that differentiate responders and non-responders.

Gene group	GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
Deleted in 2/3 or 3/3 responders and 0/3 non-responders	GO:0042742	defense response to bacterium	1.65E-10	2.19E-06	7.38 (18207,195,215,17)
	GO:0009617	response to bacterium	1.23E-09	8.21E-06	6.48 (18207,222,215,17)
	GO:0098542	defense response to other organism	3.10E-08	1.38E-04	4.65 (18207,346,215,19)
	GO:0051707	response to other organism	7.41E-07	2.47E-03	3.63 (18207,467,215,20)
Deleted in 2/3 or 3/3 non-responders and 0/3 responders	GO:0033141	positive regulation of peptidyl-serine phosphorylation of STAT protein	3.82E-23	5.08E-19	30.14 (18207,18,537,16)
	GO:0033139	regulation of peptidyl-serine phosphorylation of STAT protein	3.82E-23	2.54E-19	30.14 (18207,18,537,16)
	GO:0002323	natural killer cell activation involved in immune response	1.67E-20	7.40E-17	24.66 (18207,22,537,16)
	GO:0045343	regulation of MHC class I biosynthetic process	1.67E-20	5.55E-17	24.66 (18207,22,537,16)
	GO:0002286	T cell activation involved in immune response	1.07E-15	2.84E-12	12.45 (18207,49,537,18)
	GO:0042100	B cell proliferation	1.91E-15	4.24E-12	14.66 (18207,37,537,16)
	GO:0043330	response to exogenous dsRNA	1.37E-14	2.61E-11	13.23 (18207,41,537,16)
	GO:0043331	response to dsRNA	3.34E-14	5.56E-11	12.62 (18207,43,537,16)
	GO:0030101	natural killer cell activation	7.27E-13	1.07E-09	10.64 (18207,51,537,16)
	GO:0070661	leukocyte proliferation	1.63E-12	2.17E-09	7.45 (18207,91,537,20)
	GO:0046651	lymphocyte proliferation	2.72E-12	3.29E-09	7.76 (18207,83,537,19)
	GO:0060338	regulation of type I interferon-mediated signaling pathway	3.66E-12	4.06E-09	11.87 (18207,40,537,14)
	GO:0032943	mononuclear cell proliferation	4.29E-12	4.39E-09	7.58 (18207,85,537,19)
	GO:0002285	lymphocyte activation involved in immune response	1.76E-11	1.67E-08	7.53 (18207,81,537,18)
	GO:0033138	positive regulation of peptidyl-serine phosphorylation	7.68E-11	6.81E-08	6.94 (18207,88,537,18)
	GO:0002366	leukocyte activation involved in immune response	9.17E-11	7.63E-08	6.05 (18207,112,537,20)
	GO:0002263	cell activation involved in immune response	1.08E-10	8.49E-08	6.00 (18207,113,537,20)
	GO:0002250	adaptive immune response	1.14E-10	8.42E-08	6.78 (18207,90,537,18)
	GO:0033135	regulation of peptidyl-serine phosphorylation	5.76E-10	4.04E-07	5.80 (18207,111,537,19)
	GO:0060337	type I interferon signaling pathway	1.20E-09	7.96E-07	7.37 (18207,69,537,15)
	GO:0030183	B cell differentiation	1.24E-09	7.85E-07	6.78 (18207,80,537,16)
	GO:0051122	hepoxilin biosynthetic process	2.19E-08	1.33E-05	33.91 (18207,5,537,5)
	GO:0051121	hepoxilin metabolic process	2.19E-08	1.27E-05	33.91 (18207,5,537,5)
	GO:0006959	humoral immune response	3.06E-08	1.70E-05	4.60 (18207,140,537,19)
	GO:0001959	regulation of cytokine-mediated signaling pathway	4.15E-08	2.21E-05	5.37 (18207,101,537,16)
	GO:0060759	regulation of response to cytokine stimulus	9.56E-08	4.89E-05	5.07 (18207,107,537,16)
	GO:0018916	nitrobenzene metabolic process	7.49E-07	3.69E-04	33.91 (18207,4,537,4)
	GO:0030098	lymphocyte differentiation	9.76E-07	4.64E-04	3.70 (18207,174,537,19)
	GO:0042113	B cell activation	1.05E-06	4.80E-04	4.27 (18207,127,537,16)
	GO:0051607	defense response to virus	1.97E-06	8.74E-04	3.87 (18207,149,537,17)
	GO:0070489	T cell aggregation	2.04E-06	8.77E-04	3.39 (18207,200,537,20)
	GO:0042110	T cell activation	2.04E-06	8.50E-04	3.39 (18207,200,537,20)
	GO:0071593	lymphocyte aggregation	2.38E-06	9.61E-04	3.36 (18207,202,537,20)
	GO:0050817	coagulation	3.34E-06	1.31E-03	2.37 (18207,487,537,34)
	GO:0007596	blood coagulation	3.34E-06	1.27E-03	2.37 (18207,487,537,34)
	GO:0070486	leukocyte aggregation	3.74E-06	1.38E-03	3.26 (18207,208,537,20)
	GO:0007599	hemostasis	4.17E-06	1.50E-03	2.34 (18207,492,537,34)
	GO:0002252	immune effector process	4.35E-06	1.52E-03	2.50 (18207,407,537,30)
	GO:0042759	long-chain fatty acid biosynthetic process	4.88E-06	1.67E-03	16.95 (18207,10,537,5)
	GO:0050878	regulation of body fluid levels	7.22E-06	2.40E-03	2.14 (18207,618,537,39)
	GO:0019372	lipoxygenase pathway	2.32E-05	7.52E-03	13.04 (18207,13,537,5)
	GO:0070458	cellular detoxification of nitrogen compound	2.55E-05	8.09E-03	33.91 (18207,3,537,3)
	GO:0051410	detoxification of nitrogen compound	2.55E-05	7.90E-03	33.91 (18207,3,537,3)
	GO:0034109	homotypic cell-cell adhesion	2.82E-05	8.53E-03	2.76 (18207,258,537,21)
	GO:0007159	leukocyte cell-cell adhesion	3.30E-05	9.77E-03	2.81 (18207,241,537,20)
Amplified in 2/3 or 3/3 non-responders and 0/3 responders	GO:0035589	G-protein coupled purinergic nucleotide receptor signaling pathway	7.15E-08	9.52E-04	42.78 (18207,14,152,5)
	GO:0035588	G-protein coupled purinergic receptor signaling pathway	5.32E-07	3.54E-03	29.95 (18207,20,152,5)
	GO:0035590	purinergic nucleotide receptor signaling pathway	6.94E-07	3.08E-03	28.52 (18207,21,152,5)
	GO:0035587	purinergic receptor signaling pathway	2.64E-06	8.79E-03	22.18 (18207,27,152,5)

Enrichment calculated by GOrilla as (b/n) / (B/N), where N is the total number of genes, B is the total number of genes associated with a specific GO term, n is the number of genes in the input group, and b in the number of genes in the intersection.

Gene Census database, include include TP53, CDKN2A, and JAK2 (Futreal et al., 2004). Deletions of CDKN2A and TP53 have previously been identified as predictive of poorer prognosis in MCL (Delfau-Larue et al., 2015). CDKN2A, also known as $p16^{INK4A}$, inhibits CDK4 and CDK6 and is required for palbociclib's function. Loss of function of JAK2 underlies leukemic transformation in subtypes of Acute Myeloid Leukemia, and JAK2 is known to play an important role in interferon signaling (Beer et al., 2010; Darnell, Kerr, & Stark, 1994). These NR-specific deletions affect mRNA expression, further implicating their functional potential (**Figure 3.3**).

Non-responder specific amplifications (found in at least 2/3 non-responders and copy-neutral in all responders) span 255 genes, and are enriched for GO terms involving G-protein coupled receptor signaling pathways (**Table 3.1**). Genes responsible for this GO enrichment are found in a shared chr3 amplification unique to PT2 and PT6 (**Figure 3.2**). Responder specific amplifications span 12 genes and are not statistically significantly enriched for any GO terms. Similar to large-scale deletions, NR-specific and R-specific amplifications correspondingly affect mRNA expression (**Figure 3.4, 3.5**).

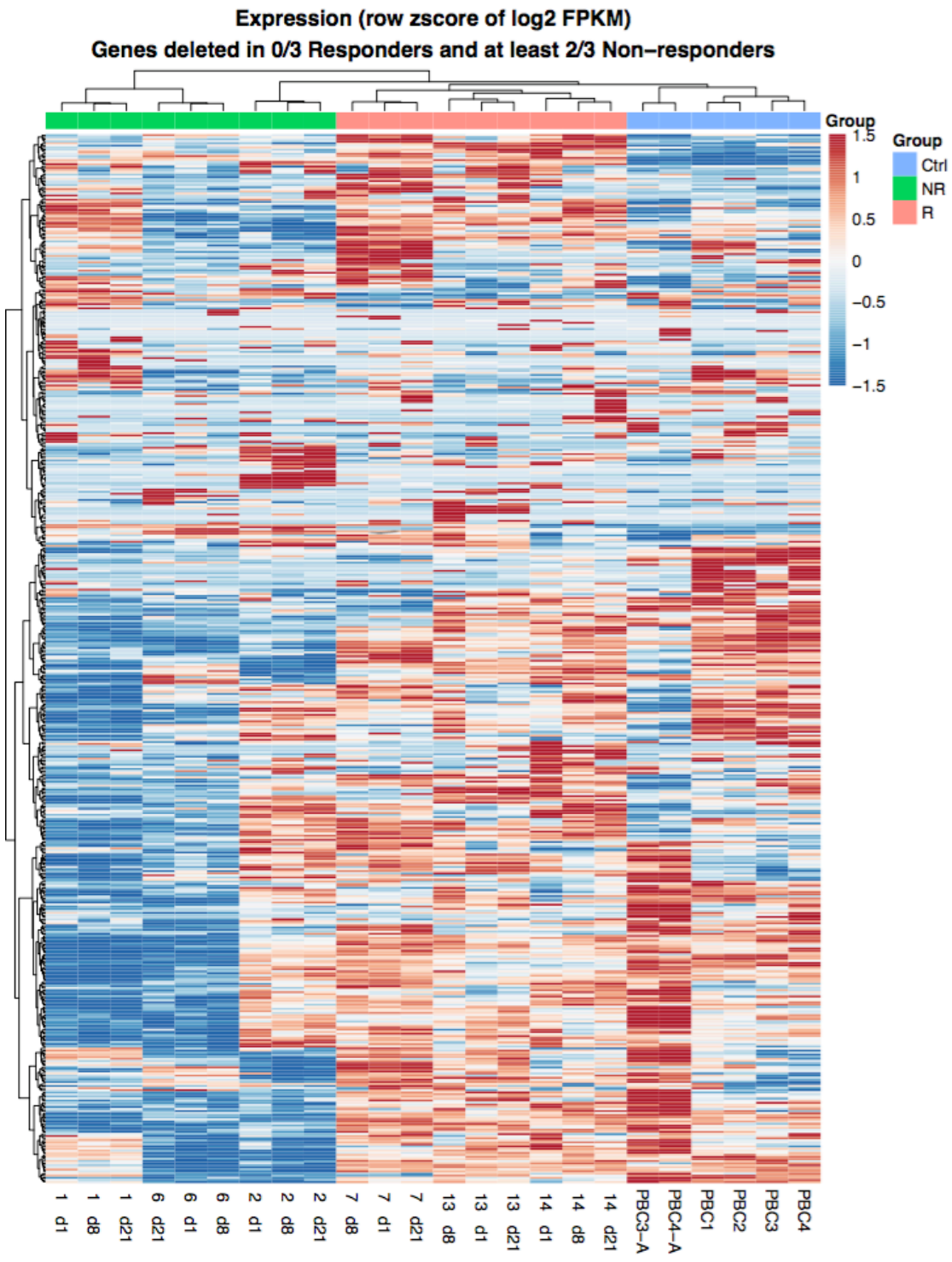


Figure 3.3. NR-specific deletions. Heatmap of mRNA expression of genes hemizygously deleted in at least 2/3 non-responders and 0/3 responders. Expression value plotted is the z-score of log2-transformed FPKM of each gene.

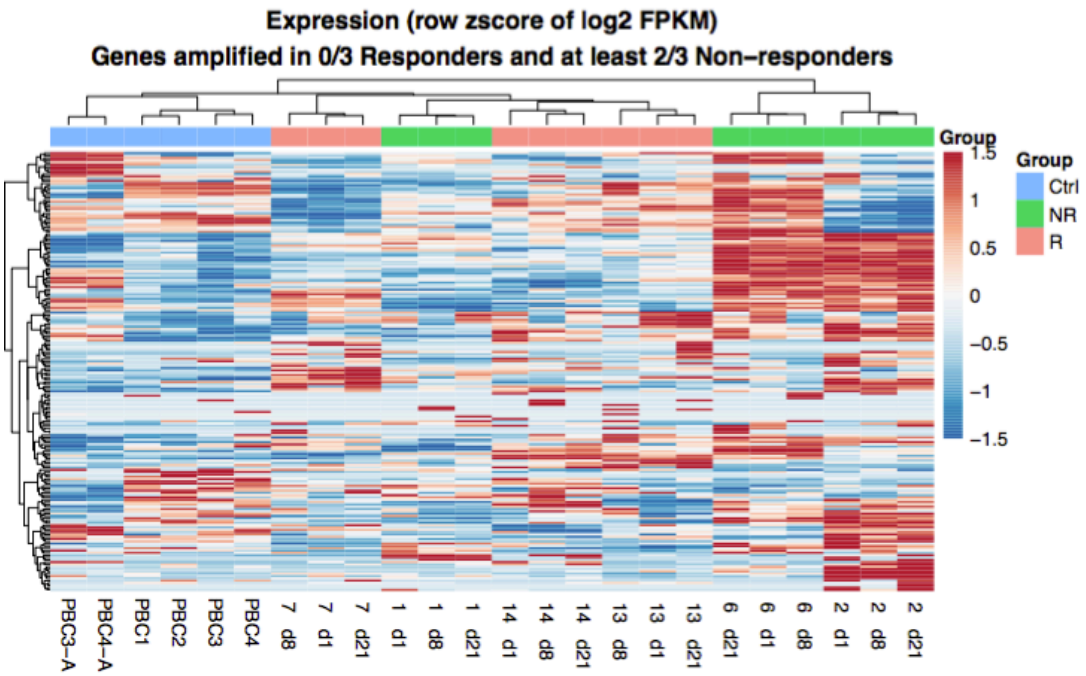


Figure 3.4. NR-specific amplifications. Heatmap of mRNA expression of genes in regions of chromosomal amplifications found in at least 2/3 non-responders and 0/3 responders. Expression value plotted is the z-score of log2-transformed FPKM of each gene.

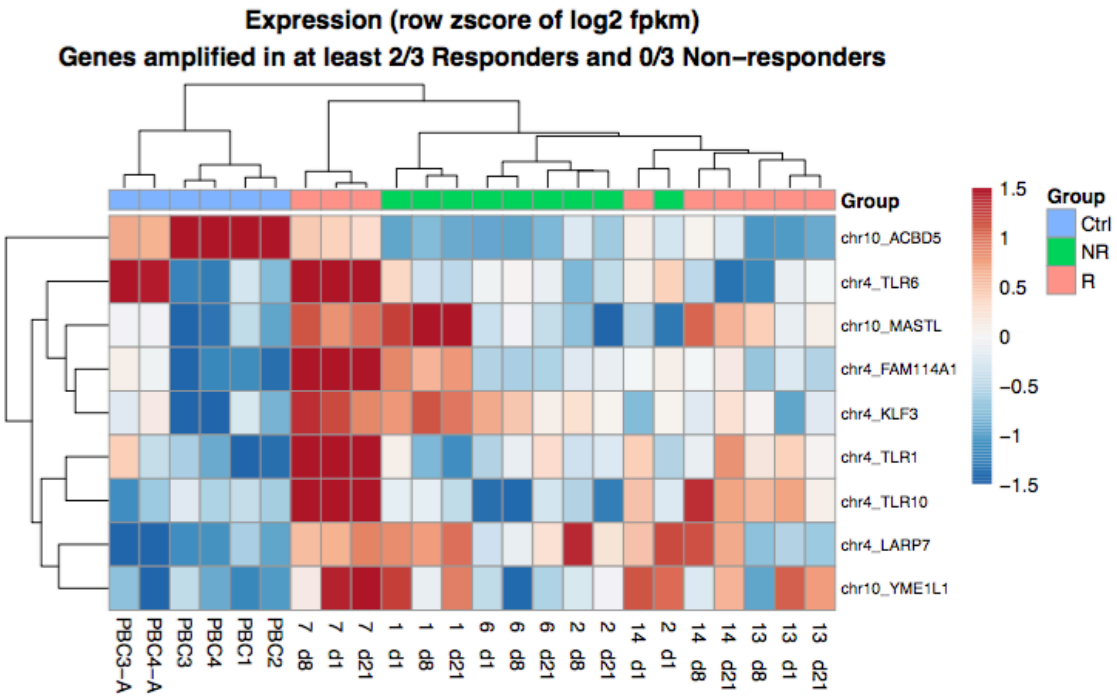


Figure 3.5. R-specific amplifications. Heatmap of mRNA expression of genes in regions of chromosomal amplifications found in at least 2/3 responders and 0/3 non-responders. Expression value plotted is the z-score of log2-transformed FPKM of each gene.

In parallel with NR-specific hemizygous deletions being enriched for interferon genes, responder (R) specific hemizygous deletions are enriched for GO terms involving defensive response to bacteria owing to the deletion of DEF genes on chr8 (**Table 3.1, Figure 3.2**). Both of these implicate an important role for immune response in the success of PALBOR treatment. Cancer genes in this list include PCM1, a gene recurrently translocated in a number hematological malignancies along with translocation partners RET and JAK2 (Futreal et al., 2004). RET is amplified in partial responder PT7 and JAK2 deleted in non-responders PT2 and PT6. R-specific deletions also span hundreds of other genes and lower the expression levels of these genes (**Figure 3.6**). Interestingly, lower expression of a cluster of R-specific deleted genes brings their expression levels closer to normal PBC controls than those of the other MCL patients, causing responders and PBCs to cluster interchangeably with each other and distinctly from non-responders (**Figure 3.7**). This suggests that deletions of these genes results in a less severe disease in the context of PALBOR.

Comparing all identified copy-number altered genes in our patient cohort to the COSMIC Cancer Gene Census list highlighted the presence of several recurrently amplified and deleted genes. Commonly deleted in cancer genes found in our data include ATM, BIRC3, BRCA2, CDKN2A, CDKN2C, PTEN, RB1, and, TP53. TP53 and ATM are not only deleted but also non-synonymously mutated in our sample cohort, discussed further below.

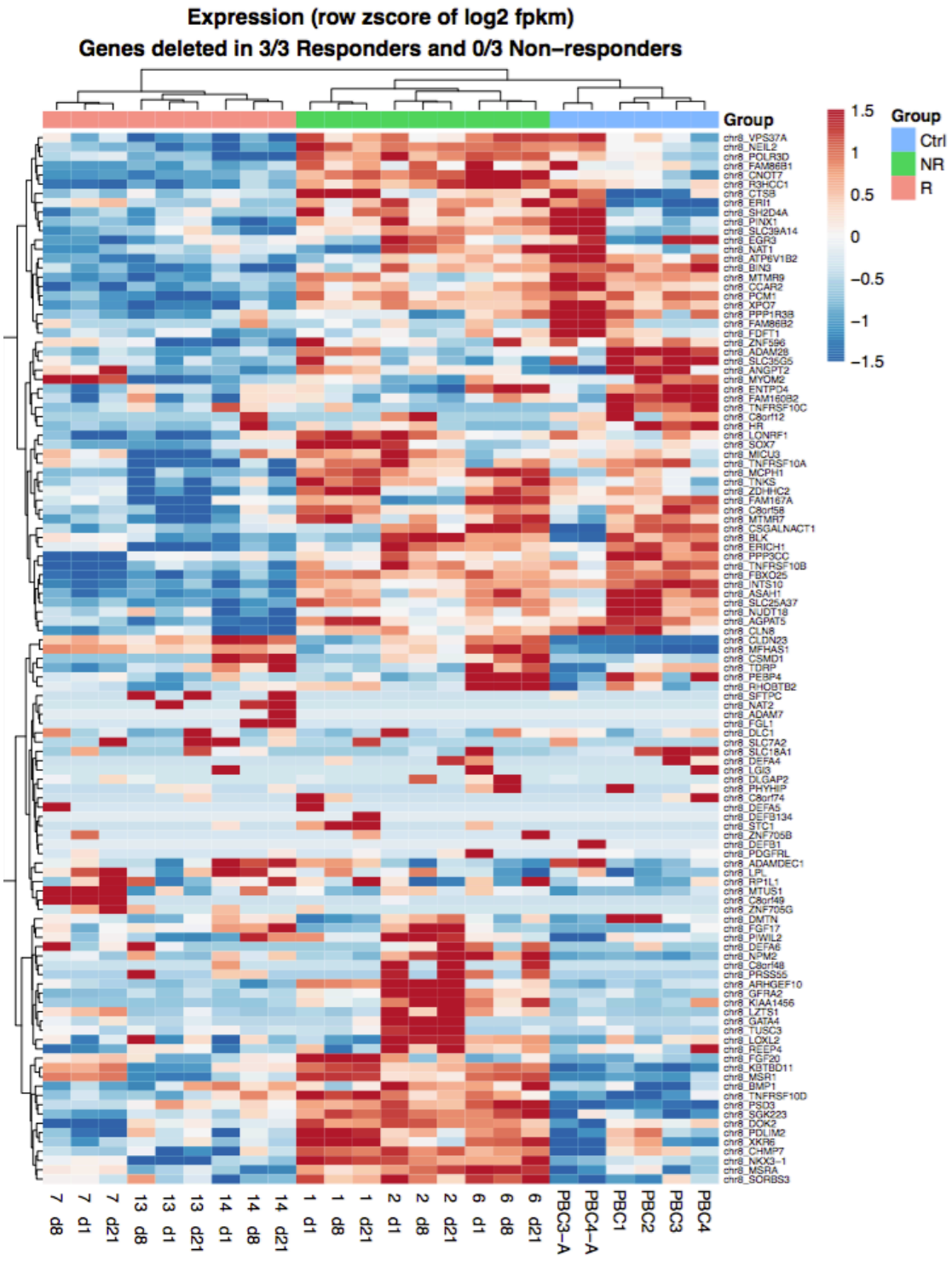


Figure 3.6. R-specific deletions. Heatmap of mRNA expression of of genes in regions of hemizygous deletions found in all 3 responders and 0/3 non-responders. Expression value plotted is the z-score of log2-transformed FPKM of each gene.

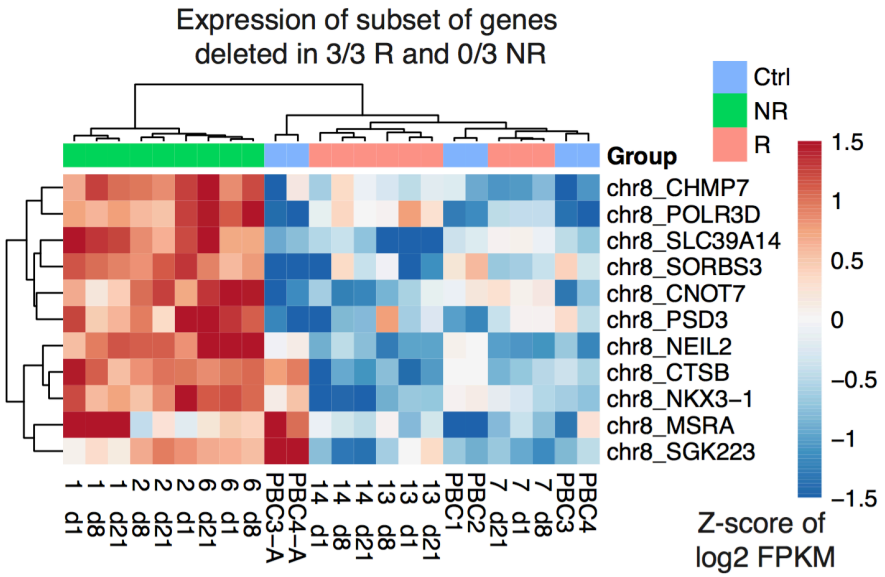


Figure 3.7. Subset of R-specific deletions. Subset of genes shown in Supplementary Figure 5 where deletion in responders results in expression levels similar to normal controls. Samples from patients lacking any deletion (i.e. non-responders) have elevated levels compared to resting normal B cells and comparable to a subset of activated normal B cells.

Commonly amplified in cancer genes also found in our list include JUN, MDM2, MYC, and SOX2. MYC overexpression is known to provide a growth advantage resulting in a more aggressive disease in the double-hit subtype of MCL (Setoodeh et al., 2013). We detected a 33 Mb chromosome 8 amplification in non-responder PT1, present at D1 and D21, spanning the MYC gene (**Figure 3.8A**). PT1, along with non-responder PT2, also had higher MYC RNA levels compared to the other patients and normal peripheral blood controls (PBCs) (**Figure 3.8B**). Responders exhibited lower MYC RNA levels than PBCs. Higher MYC RNA in PT1 and PT2 also correlated with higher protein levels as seen by immunohistochemistry (**Figure 3.8C**). WES revealed amplification of chr8 in partial responder PT7, but this was not reflected in the RNA levels as PT7 had similarly low MYC expression as the other responders and lower expression than the PBCs.

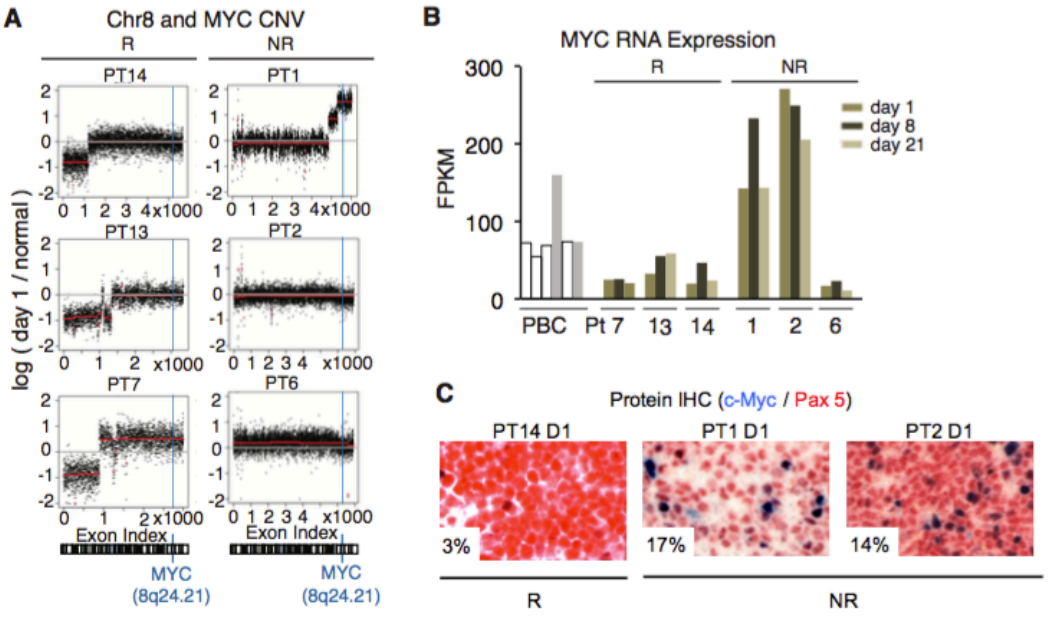


Figure 3.8. MYC amplifications and over-expression. A. Amplification of chr8 spanning MYC in 2/6 patients. **B.** Increased RNA expression of MYC in PT1 (has DNA amplification) and PT2 (no CNV). **C.** Immunohistochemistry showing increased protein expression of MYC in PT1 and PT2 correlating with RNA expression.

Two-hit loss of ATM and TP53 in non-responders. Integrative WES and WTS analysis revealed that all 3 NRs have deletions, mutations, or both of ATM and TP53. PT2 has a mutation and hemizygous deletion of ATM resulting in 100% expression of the mutated allele and reduced overall expression, PT6 has a similar combined deletion and mutation in TP53 again resulting in 100% expression of the mutated allele although mRNA expression is increased instead of decreased, and PT1 has hemizygous deletions of both ATM and TP53 resulting in lower expression of ATM but higher expression of TP53, similar to PT6 (**Figure 3.9A, B**). ATM alterations were also found in responders of the cohort: complete responder PT14 has an ATM SNV and partial responders PT13 and PT7 have hemizygous losses of ATM. However, 50% expression of the wild type allele remains in both of these cases and similar overall gene expression as normal controls and the other responders, suggesting that heterozygous mutations of ATM, the most commonly mutated gene in MCL, is tolerated by the treatment, but loss of either both copies in one of the genes or hemizygous loss of both genes, ATM and TP53, may affect response to PALBOR.

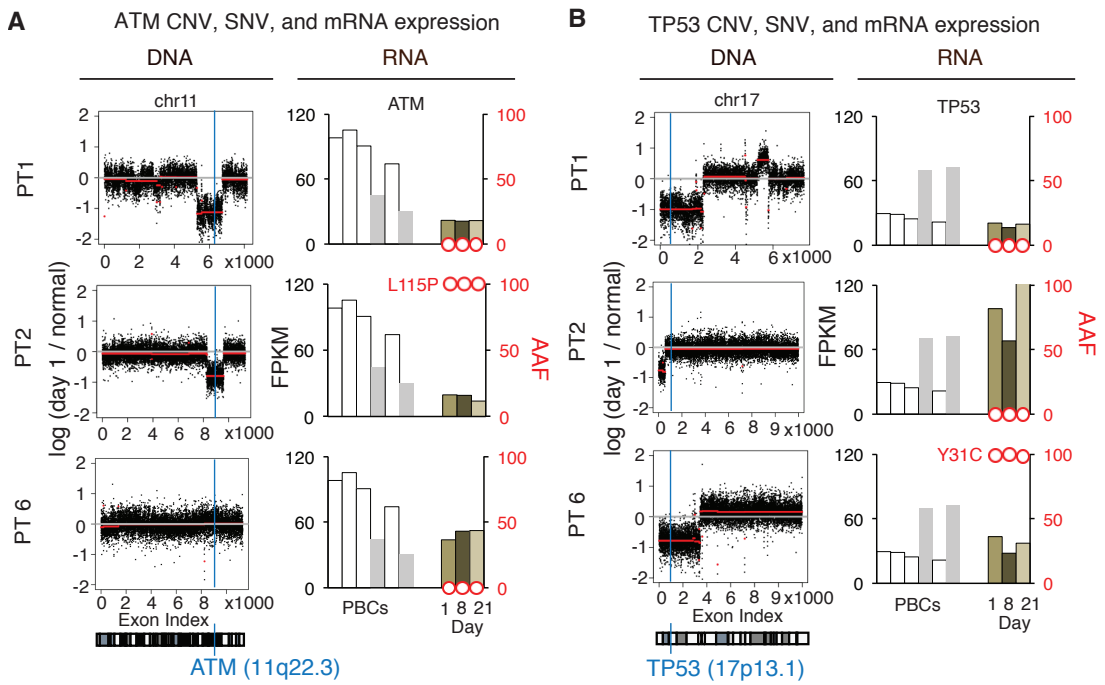


Figure 3.9. Non-responder specific SNVs and hemizygous deletions of ATM and TP53. A, B. Integrative copy number, mutation, and RNA expression data show that 3/3 NRs have ATM and/or TP53 alterations.

BCL6 amplifications do not affect response to PALBOR. Chromosome 3 amplifications spanning BCL6 were found in non-responders PT2 and PT6 and the complete responder PT14 (**Figure 3.10A**) suggesting that gain of BCL6, a transcription factor important in germinal centers and B-cell lymphomagenesis and expressed in 12% of MCL cases, does not generate resistance to PALBOR (Basso & Dalla-Favera, 2010; Gualco, Weiss, Harrington, & Bacchi, 2010). The effect of BCL6 amplifications in PT14, PT2, and PT6 was confirmed at the RNA (**Figure 3.10B)** and protein levels (**Figure 3.10C**); PT14 had higher RNA expression than all other patients and both PT14 and PT2 had higher protein expression than PT1, a patient who lacked BCL6 CNV. BCL6 protein was compared to hematoxylin instead of PAX5, which identifies MCL tumor cells, since PAX5 IHC was found to inhibit BCL6 expression resulting in unreliable images. PAX5 staining for comparison can be found in figure 3.8.

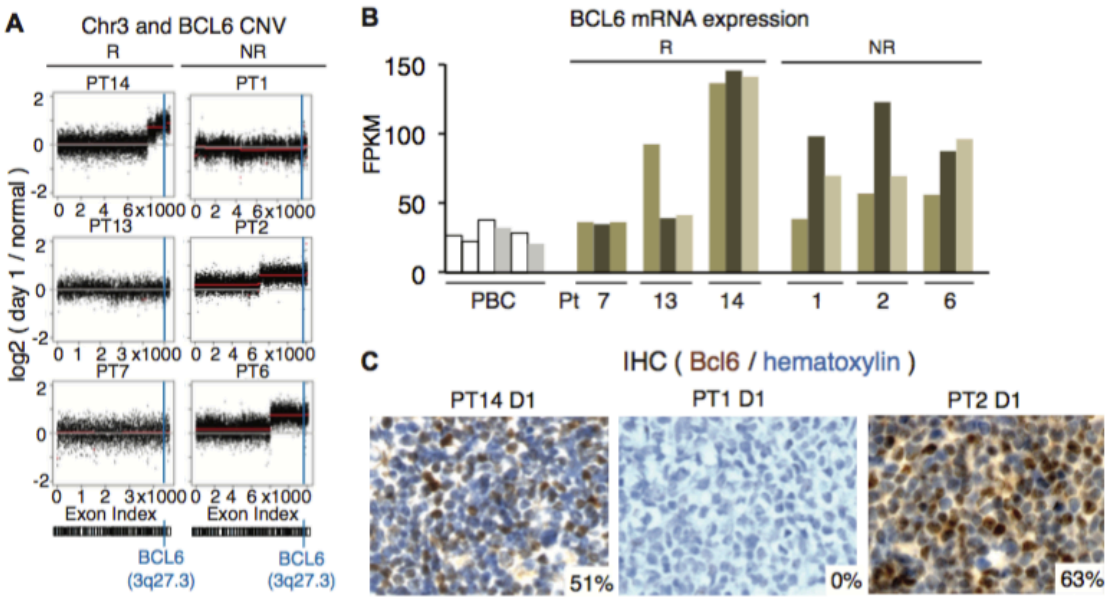


Figure 3.10. BCL6 amplifications in non-responders and responders. A. CNV of chr3 spanning BCL6 in R PT14 and NRs PT2 and PT6. **B.** RNA expression of BCL6 in all patients showing higher than PBC control levels in 5/6 patients and especially high levels in PT14. **C.** Immunohistochemistry of BCL6 at D1 shows high protein levels in PT14 and PT2, both of whom have a DNA amplification, and lower levels in PT1, who does not have a CNV.

Relapse-specific chromosome 13 deletion. As one of the responding patients, PT13, later relapsed, we sequenced the tumor exome at 3 time points: prior to treatment D1, end of first cycle D21, and a relapse sample biopsied prior to any additional treatment. We found the relapse time point to lack the hemizygous deletions of interferon genes thought to contribute to innate resistance in non-responders PT2 and PT6. Non-responder PT1 also lacked this chromosome 9 deletion, pointing to alternate potential mechanisms of resistance in the relapse sample and in PT1. We discovered the relapse sample to contain the same CNVs at D1 and D21 with the exception of a relapse-specific chromosome 13 hemizygous deletion spanning 62.7 Mb (**Figure 3.11A**). A chromosome 13 deletion largely overlapping this region was also found in PT1 (**Figure 3.11A**). This hemizygous deletion spans RB1, a gene essential to palbociclib's mechanism of action. The effects of the deletion are compensated by regulatory mechanisms however, and RB1 RNA levels remain similar across patient samples with and without the deletion and normal controls (**Figure 3.11B**). Other genes affected by the chr13 deletion are lowered in expression and may contribute to; these genes include COSMIC cancer related genes BRCA2, FOXO1, LCP1, and LHFP (**Figure 3.12**).

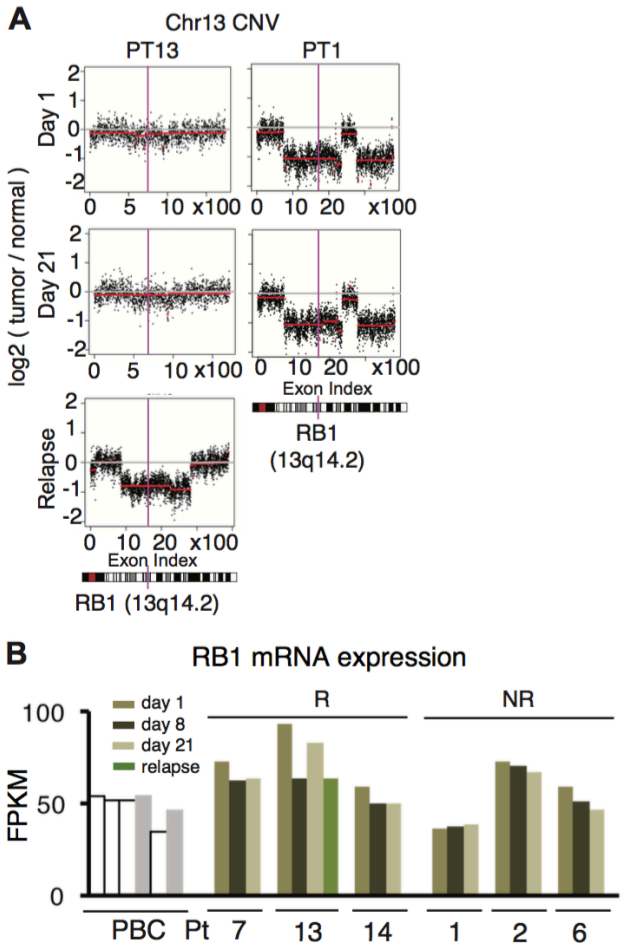


Figure 3.11. Chromosome 13 deletion unique to NR PT1 and relapse in PT13 and relapse-specific shifts in somatic SNV. A. Plots of read depth showing large-scale hemizygous loss in chromosome 13 in PT1 (all timepoints) and PT13 (relapse only). **B.** RB1 mRNA expression in samples with chr13 deletion (blue) and those without (pink) showing similar expression in patients with and without CNV.

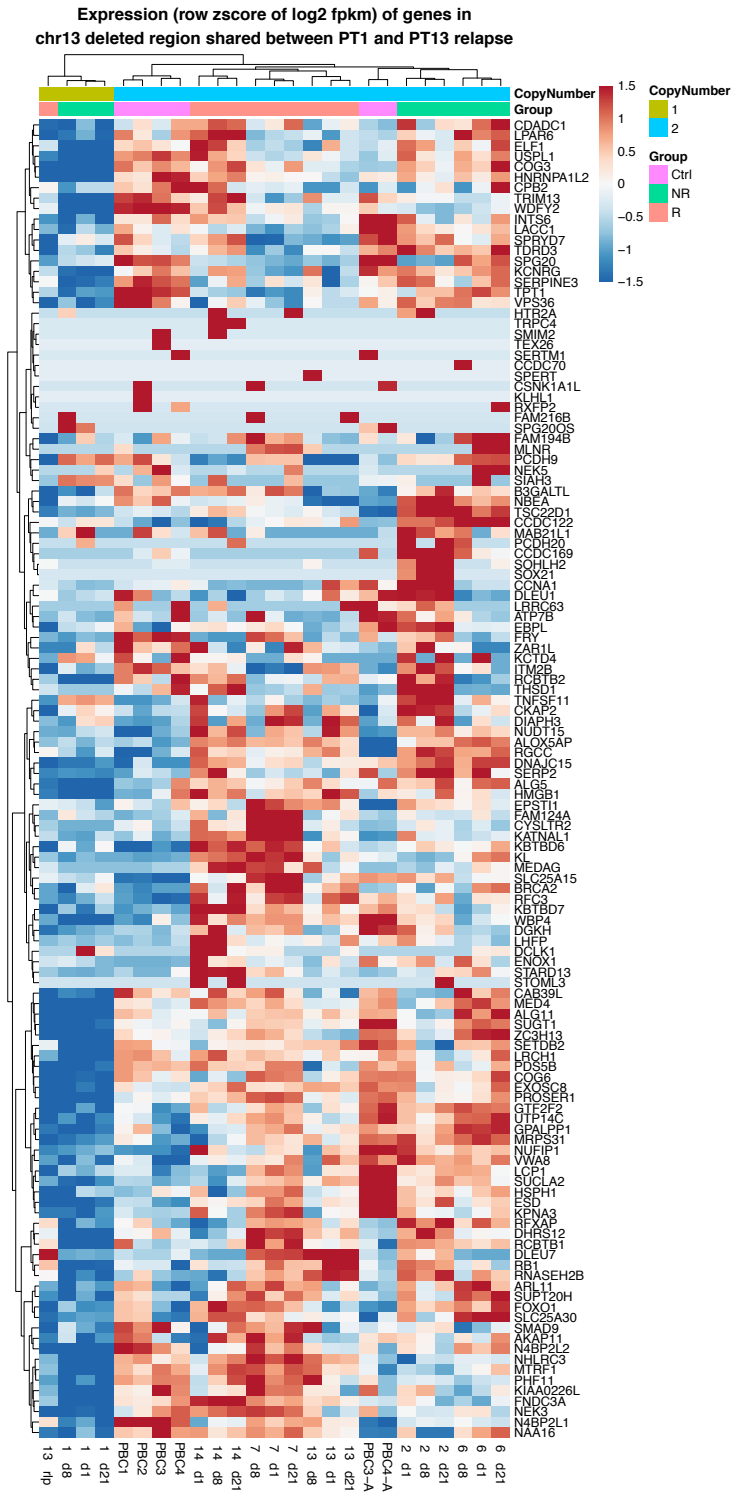


Figure 3.12. Chromosome 13 hemizygous deletion. Heatmap of z-score of log2 FPKM values of all expressed genes in the region of deletion shared by both PT1 and PT13 relapse.

Clonal architecture and evolution of single nucleotide variants. Although the chr13 deletion was the only large CNV undetectable before relapse, comparison of the cellular frequencies of SNVs between the two primary time points and the relapse time point revealed complex subclonal architecture including clusters of Single Nucleotide Variants (SNVs) that expanded in relapse. SNVs were identified using GATK UnifiedGenotyper for all tumor and germline samples simultaneously. SNVs with low read depth (<10) and low quality scores (<20) as well as those present in 1000 genomes phase I data derived from Oncotator annotations were filtered out (Ramos et al., 2015). For patients with matched normal samples (PT7, PT13, PT14), SNVs were considered somatic if greater than 10% of reads mapped to the variant allele in the tumor while 0 reads mapped the variant allele in the normal.

To infer subclonal composition of SNVs, we used pyclone, a probabilistic graphical model that infers cellular prevalence of mutations using allele frequency and copy number (Roth et al., 2014). In order to identify the presence of subclones that shift during therapy with high confidence, we limited this analysis to variants identified in WES at bases covered by at least 30 reads at all time points within a patient, have an alternate allele frequency of at least 10% in one or more time points, and have a minimum change in allele frequency of 25%. This includes coding and noncoding mutations since either provide evidence for the presence of a subclone. These filters were applied to patients with and without matched normal controls alike, and for

patients with matched normal, we further filtered any germline SNPs (single nucleotide polymorphisms).

The inferred clonal dynamics during the first cycle of treatment for all patients (R and NR) revealed clusters that decrease in prevalence, increase, and stay constant, demonstrating therapy-related clonal evolution (**Figure 3.13**). Although this identified interesting SNVs, especially in CR PT14 where SNVs in CCT5, CPT1C, and MUC4 were found to decrease after D1, the lack of overlap in clonal and subclonal SNVs across patients likely due to the overall low number of SNVs used in this analysis motivated us to expand our analysis to all somatic SNVs with minimum coverage of 10 reads at both time points and minimum variant allele frequency of 10% in at least one time point. This analysis (limited to patient samples with germline data) revealed a larger list of SNVs that expand or decrease during the treatment timeline (**Figure 3.14A-F**). However, overlap between the genes with SNVs in decreasing clusters and increasing clusters remained low, with only one shared gene, ADAMTS18, identified (**Figure 3.14G, H**). These genes also lack shared GO enrichments confirming they are not different genes functioning in the same molecular pathways. Given these results, we conclude that SNVs are not useful in this case for identifying signatures of resistance or sensitivity that could be translated to other patients, and the shared CNVs discussed above offer more insight.

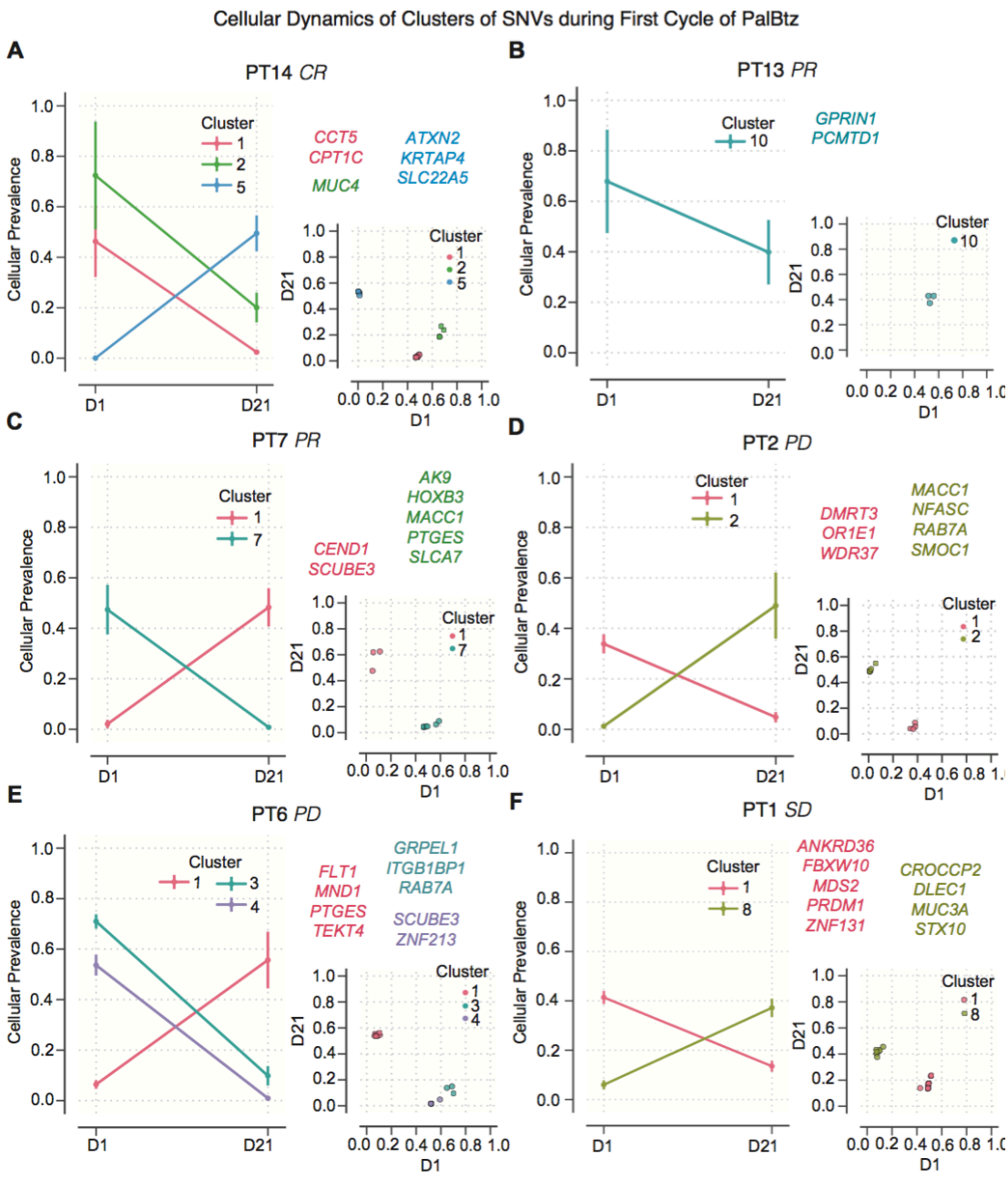


Figure 3.13. SNV cluster dynamics of highly covered SNVs. Cellular prevalence of inferred clusters (left) and the corresponding SNVs in each cluster (right) prior to treatment (D1) and after first cycle (D21) in each patient. Input SNVs have a minimum coverage of 30 reads, minimum variant allele frequency of 10%, and a minimum frequency shift between time points of 25%.

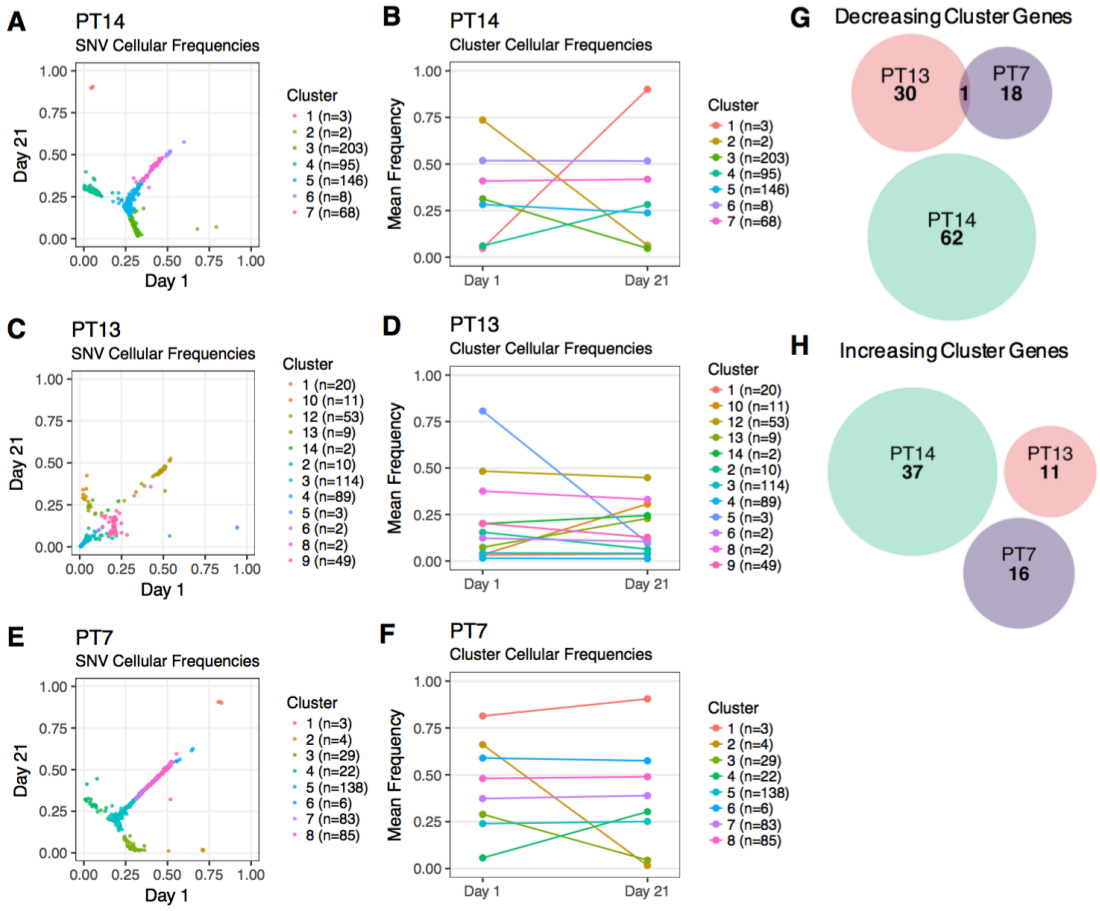


Figure 3.14. SNV cluster dynamics. A-F. Cellular prevalence of SNVs passing minimum coverage of 10 and minimum variant allele frequency of 10% filters and inferred clusters prior to treatment (D1) and after first cycle (D21) in responders PT14 (**A, B**), PT13 (**C, D**), and PT7 (**E, F**). **G.** Overlap between genes in clusters decreasing in frequency in each responder. **H.** Overlap between genes in clusters increasing in frequency in each responder.

Longitudinal analysis of SNVs in PT13 with these more lenient filters and larger list of SNVs and comparison of the cellular frequencies of SNVs between the two primary time points and the relapse time point revealed complex subclonal architecture and dynamics. Several SNVs remain constant in frequency between the primary and relapse time points implying a clonal origin for the relapse sample, and there was a cluster of SNVs that decrease in frequency during the first cycle of treatment, reflecting the effect of PALBOR on overall clonal composition (**Figure 15**). In addition, we detected 3 clusters of SNVs that were present at less than 10% cellular prevalence at D1 and D21 and expand to greater than 25% in relapse (**Figure 15**). These clusters encompass 223 total SNVs and were inferred to be present in an average of 4% of cells at the D1 time point suggesting that the dominant relapse clone was present pre-treatment.

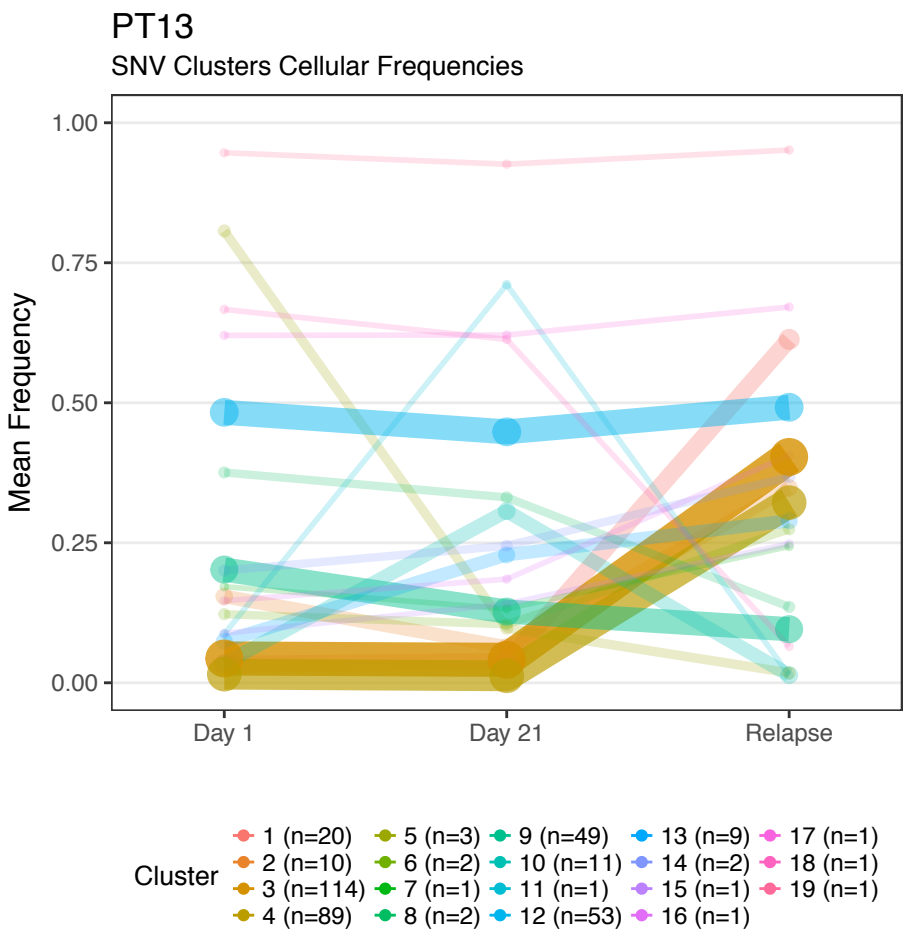


Figure 3.15. Clonal evolution across three time points. Cellular frequencies of highly covered SNVs (minimum coverage 10 at each time point, minimum variant allele frequency 10% at one or more time points) at D1, D21, and relapse in PT13 showing both clusters of shared SNVs and clusters of SNVs expanding in relapse.

DISCUSSION

This study describes novel information gained from longitudinal sequencing of exomes and transcriptomes of patients in a phase I clinical trial in recurrent Mantle Cell Lymphoma. The trial consisted of the combination of the CDK4/CDK6 inhibitor palbociclib and the proteasome inhibitor bortezomib (PALBOR). Our unique data profiling 3 responders and 3 non-responders at multiple time points of before treatment, mid cycle while cells are in prolonged G1 arrest (pG1), and after one cycle of treatment where the sampled cells have escaped both CDK4 and proteasome inhibition, have enabled us to identify candidate genomic and transcriptomic predictors of response to PALBOR. For the 3 responders, we were also able to sequence matched normal tissue acquired from buccal swabs, which were unavailable for the 3 non-responders.

We investigated subclonal architecture and clonal evolution in the specific context of PALBOR treatment, identifying clusters of SNVs in responders and non-responders that decrease in prevalence, indicative of sensitive cells, or increase in prevalence, indicative of cells escaping treatment. However, overlap between patients on genes and pathways comprising these clusters was absent, suggesting features other than SNVs should be used to identify signatures of sensitivity and resistance. Instead, we found large-scale copy number aberrations and corresponding gene expression changes to likely affect response to PALBOR.

Palbocilib was first applied as a therapeutic agent in MCL because of the aberrant expression of CDK4 and cyclin D1 that is a hallmark of the disease. CDK6, the other target of palbociclib, is silenced in MCL (Di Liberto et al., n.d.). Interestingly, 2/6 patients in the sequenced cohort had a CDK4 amplification, one of whom was the complete responder PT14, suggesting that palbociclib remains effective even with an amplification of its substrate, and the lack of somatic SNVs and indels in coding regions of CDK4 suggests that CDK4 is a favorable drug target. Similarly, BCL6 amplifications were found in the complete responder and 2/3 non-responders, implicating that response to treatment is not governed by this aberration, as is the case in many other lymphomas such as DLBCL (Gualco et al., 2010; Karube et al., 2008). BCL6 is known to be regulated at the RNA and protein levels, mainly plays a role in germinal center lymphoma cells, and amplifications are uncommon in MCL, a pre-germinal center cancer (Camacho et al., 2004; Gualco et al., 2010). BCL6 is also known to lower TP53 expression, which, along with ATM, is hemizygously deleted and mutated in multiple patients in our cohort (Phan & Dalla-Favera, 2004).

Unlike CNVs spanning CDK4 and BCL6, there are several large CNVs that do correlate with patient response group; most notably, chromosome 9 deletions affect 3/3 non-responders and 0/3 responders. In 2 of the non-responders, PT2 and PT6, chr9 hemizygous deletions span a large set of genes coding for type I interferons (IFNs), a class of cytokines produced and

released by immune cells in response to viruses. IFN loss has been implicated in other cancers for causing resistance to chemotherapy and may explain resistance to PALBOR in PT2 and PT6 (Sistigu et al., 2014). Chromosome 9p loss was also recurrently detected in microarray copy number profiling of 77 MCL primary tumors (Hartmann et al., 2010). A synergistic link between interferon and bortezomib was previously reported in other cancers; interferon-α, when used as a therapeutic agent, synergizes with bortezomib in melanoma (Lesinski et al., 2008; Markowitz et al., 2014) and bladder cancer (Papageorgiou, Kamat, Benedict, Dinney, & McConkey, 2006), and interferon-γ overcomes bortezomib resistance in hematological cells (Niewerth et al., 2014). Not only are these immune pathways related to the function of bortezomib, but may also play an integral role in CDK4/6 inhibition therapy. Recent work in breast cancer mouse models uncovered that CDK4/6 inhibition activates expression of endogeneous retroviral elements by the tumor cell, in turn triggering an anti-tumor immune response facilitated by type III interferons and T-cell-mediated cytotoxicity (goel et al., 2017). Our data suggests that hemizygous loss of interferon expression through a recurrent CNV may antagonize PALBOR activity. A potential mechanism for this action is through the Interferon Regulatory Factor (IRF) family of transcription factors and TRAIL-mediated cell death as these pathways share multiple points of functional interactions (**Figure 3.16**). Cell-cycle coupled loss of IRF4 as a result of palbociclib-induced pG1 is known to sensitize cells to bortezomib

killing (Huang et al., 2012). Functional studies have shown that TRAIL gene expression is regulated by IRF genes (Yoshida et al., 2005), and microarray studies have revealed TRAIL-mediated induction of interferon genes as well as interferon-induced up regulation of of TRAIL (Kumar-Sinha, Varambally, Sreekumar, & Chinnaiyan, 2002). Thus, loss of interferon genes may reduce sensitivity to PALBOR by lowering TRAIL-mediated apoptosis (Di Liberto et al., n.d.).

The lack of any deletion of interferon genes in PT1 or the relapse sample in PT13 indicates alternate mechanisms of resistance. Resistance in PT1, the patient with stable disease, may stem from other CNVs as this patient has the highest number of deletions and duplications of any in the sequenced cohort. These include amplification of MYC and hemizygous deletions of both ATM and TP53. Genomic amplification and corresponding elevated transcript and protein levels of the MYC oncogene may contribute to resistance since MYC plays an important role in B cell proliferation and is commonly mutated or overexpressed in lymphomas (Hao et al., 2002; Oberley et al., 2013; Setoodeh et al., 2013).

For PT13, who relapsed after initial response, we not only sequenced the D1 and D21 time points as we did for the other patients but also the relapse sample, providing insight into both initial response and acquired resistance. The relapse sample lacked the chromosome 9 deletion and subsequent IFN loss seen in non-responders PT2 and PT6. Exploring clonal

evolution in this patient revealed a large list of mutations and CNVs that remain constant in frequency at D1, D21, and relapse implying a clonal origin for the relapse tumor. SNV clusters that expand in the relapse time point were detected at low levels (4% of cells) in the earlier time points, suggesting that the dominant relapse clone was present from the start of treatment and therapy selected for the resistant clone. This is consistent with the fact that this patient was a partial responder. The only CNV that changed between D21 and the relapse sample was a chromosome 13 deletion which largely overlaps with the chromosome 13 deletion observed in non-responder PT1. This deletion spans RB1 although RNA levels of RB1 are compensated in both patients, emphasizing the importance of integrative DNA and RNA approaches. However, the hundreds of other genes in this region where loss of one copy affected RNA expression may play a role in driving resistance in the relapse tumor as well as in PT1. Further indication of the functional importance of this deletion is that we previously identified a similar relapse-specific chr13 deletion in the context of ibrutinib treatment in MCL (Chiron et al., 2014). Overall, we leveraged a unique longitudinal and integrative genomic and transcriptomic study design to understand molecular changes underlying response to the combination therapy of palbociclib and bortezomib. Exploring genomic alterations, validating their effect on RNA expression, and further testing effects on protein level through immunohistochemistry enabled us to generate hypotheses about mechanisms important for response to PALBOR

therapy. We demonstrated a potential role for loss of interferon genes in innate resistance coupled with combined mutations and hemizygous deletions of ATM and TP53 and amplification and concordantly increased mRNA expression of MYC. After more extensive functional validation and analysis of data from larger cohorts, these can then be applied towards precision medicine efforts for targeting patients likely to respond while furthering understanding of MCL biology.

These data have implications for not just treating MCL but cancer therapy in general. Research in recent years has lent promise to a class of drugs called immunotherapies where the body's immune system is leveraged for targeting tumor cells. As this study and other recent work on the mechanism of CDK4/6 inhibition therapy have highlighted an important role for immune signaling in the efficacy of this therapy, combining PALBOR with immunotherapy such as interferon therapy, may further improve response to PALBOR. Moving forward, testing these combinations preclinically and implementing in clinic may offer a new set of therapeutic possibilities in MCL and other cancers.

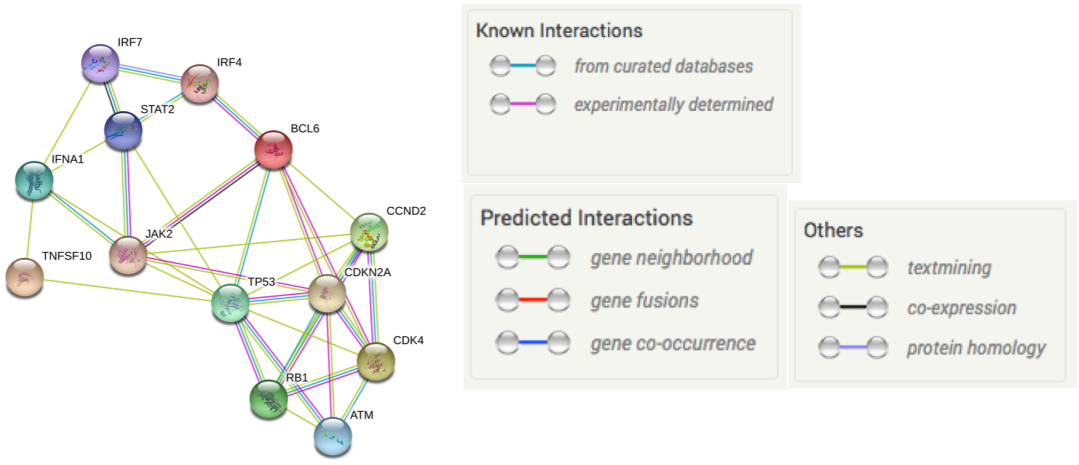


Figure 3.16. Protein-protein interaction network from the STRING database (Szklarczyk et al., 2015) showing shared functions between interferon signaling, TNFSF10 (TRAIL) cell death, and other genes involved in CNVs or response to bortezomib.

CHAPTER 4

SINGLE CELL ISOFORM DYNAMICS AND TRANSCRIPTOMICS IN MYELODYSPLASTIC SYNDROMES STEM CELLS DURING THERAPY

PREAMBLE

This chapter is an expanded version of a manuscript in preparation[3]. SC performed FACS sorting. PV performed all single cell experiments and sequencing library preparations, computational analysis, generated figures, and wrote text with input from all authors. MM contributed to analysis in figure 4.x and text describing it. All authors reviewed data and content. VK, CEM, and CYP conceived the project.

INTRODUCTION

MDS are a class of neoplastic bone marrow failure disorders that frequently progress to acute myeloid leukemia (AML) and affect an estimated 30,000 to 50,000 patients annually in the U.S. (Bejar & Steensma, 2014). Current FDA-approved therapies include the DNA methyltransferase inhibitors decitabine and azacitidine, and the immunomodulatory agent lenalidomide,

[3] Vijay P, Chung SS, MacKay M, Tomoiaga D, Gonzalez MDR, Stern D, O'Sullivan D, Klimek V, Mason CE, Park CY. Single Cell Isoform Dynamics and Transcriptomics in Myelodysplastic Syndromes Stem Cells During Therapy. (in preparation)

which produce short-term remissions in a subset of patients. However, disease always remerges, and the only potentially curative treatment remains allogeneic stem cell transplantation (SCT), which results in long-term disease free survival in 30-40% of the patients who qualify for and receive SCT (Bejar & Steensma, 2014). Further understanding of the complex mechanisms underlying MDS pathogenesis and how MDS disease-initiating cells evade current treatments through innate or acquired mechanisms would improve treatment of MDS. Transcriptional profiling using next generation sequencing approaches is particularly promising given the epigenetic dysregulation and aberrant splicing frequently observed in MDS through recurrent mutations in both epigenetic regulators and splicing factors (Bejar et al., 2012; Haferlach et al., 2014; Papaemmanuil et al., 2013). Hematopoietic stem cells (HSCs, Lin-CD34+CD38-CD90+CD45RA-) were previously identified as the MDS-initiating population (Pang et al., 2013; Woll et al., 2014). Here, we report the first transcriptome-wide study of MDS HSCs, the first application of single cell RNA-sequencing (scRNA-seq) in MDS, and (to our knowledge) the first single-cell delineation of isoform changes of a cancer during therapy.

Recent advances in single cell sequencing have demonstrated its utility in identifying heterogeneous cell types in a number of normal and disease contexts (Gawad, Koh, & Quake, 2016; Macosko et al., 2015; Pollen et al., 2014; Shalek et al., 2013; Svensson et al., 2017; Trapnell et al., 2014). However, studies exploring splicing heterogeneity at the single cell level are

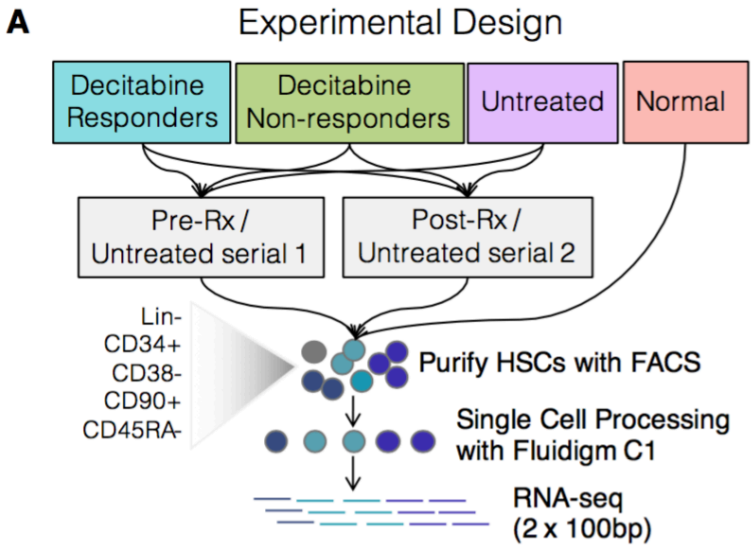
limited (Shalek et al., 2013). Here, we sequenced full-length mRNA transcripts as opposed to the 3'-end counting used by the majority of high-throughput scRNA-seq platforms currently available, which provided a unique insight into the heterogeneous expression of splice variants in MDS HSCs. Moreover, we developed an open-source tool, DISCO (distributions of isoforms in single cell omics; https://pbtech-vc.med.cornell.edu/git/mason-lab/disco/tree/master), a novel method described here for the analysis of alternative splicing in scRNA-seq or other large sets of RNA-seq data. DISCO provides an easy to use method to compare relative isoform abundances between groups of samples, perform non-parametric statistical testing, corrects for multiple testing, and visualizes significant shifts in splice variant distributions. Using DISCO, we verified previously reported findings of SRSF2 mutations altering preference of exonic splice elements and demonstrate novel, therapy-specific responses.

Integrative analysis of gene and isoform expression heterogeneity of MDS HSCs pre- and post-decitabine therapy in responders, non-responders, untreated patients with stable disease, and age-matched normal donors recapitulated previously known pathogenic mechanisms, such as dysregulated ribosome biogenesis (Narla & Ebert, 2010), and identified novel pathways differentiating disease states and response. Single cell resolution provided a unique view into the heterogeneous cell states occupied by MDS HSCs and their dynamics during treatment, highlighting differences between populations likely sensitive or resistant to therapy.

RESULTS

Single cell RNA-seq of MDS patient bone marrow biopsies. We FACS-purified hematopoietic stem cells (HSCs) from MDS patients and normal age-matched controls, and used the Fluidigm C1 platform to isolate 684 single cells from bone marrow biopsies of 3 groups of MDS patients (n=8): decitabine responders, decitabine non-responders, and MDS patients untreated (**Figure 4.1**). HSCs were defined using FACS markers Lin-CD34+CD38-CD90+CD45RA- (**Figure 4.2**) (Pang et al., 2013). Longitudinal pre- and post-decitabine treatment samples (and serial samples from the untreated patient) were sequenced from 4 patients, providing a cell-to-cell view of the effects of decitabine on MDS stem cells and MDS progression. Similarly processed single HSCs were sequenced from age-matched normal donors as controls. All cells were sequenced at an average depth of 4.6 million mapped reads per cell using paired-end 100bp sequencing, which we and others have shown improves accurate mapping across splice junctions and detection of novel junctions compared to shorter and single reads (Chhangawala, Rudy, Mason, & Rosenfeld, 2015) (**Figure 4.3**).

MDS and Normal HSCs exhibit unique transcriptional landscapes. A heatmap of highly expressed genes (mean log2(FPKM) of 2 or higher across all samples) reveals clustering of patient and response groups but also



B

Samples

Patients	Pre	Post	Response	SRSF2 mutation
	Number of HSCs			
1	79	-	Untreated	none
2	63	85	Untreated	none
3	19	71	Non-Responder	none
4	56	31	Responder	R94_S101del
5	68	-	Responder	P95H
6	33	27	Non-Responder	P95_P96>RPP
7	-	17	Responder	none
9	61	-	Responder	none
Control1	55	-	Normal	none
Control2	82	-	Normal	none

Figure 4.1. Experimental Design and Samples. A. single cell RNA-seq was used to assay transcriptome signatures of FACS purified hematopoietic stem cells (HSCs) from MDS patients before and after decitabine treatment. **B.** The number of single cells captured from each patient sample at different time points (pre- and post-decitabine for treated patients, and serial time points for untreated) as well as response status and SRSF2 mutations independently assessed with a targeted sequencing panel.

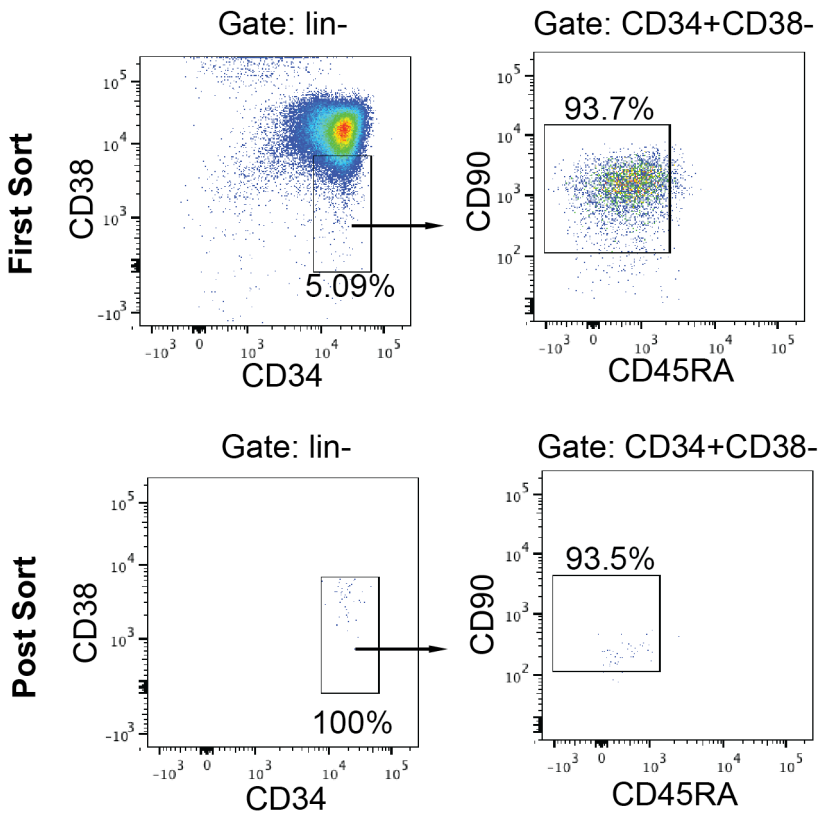


Figure 4.2. FACS purification of MDS stem cells. An example sort (Lin-CD34+CD38-CD90+CD45RA-)

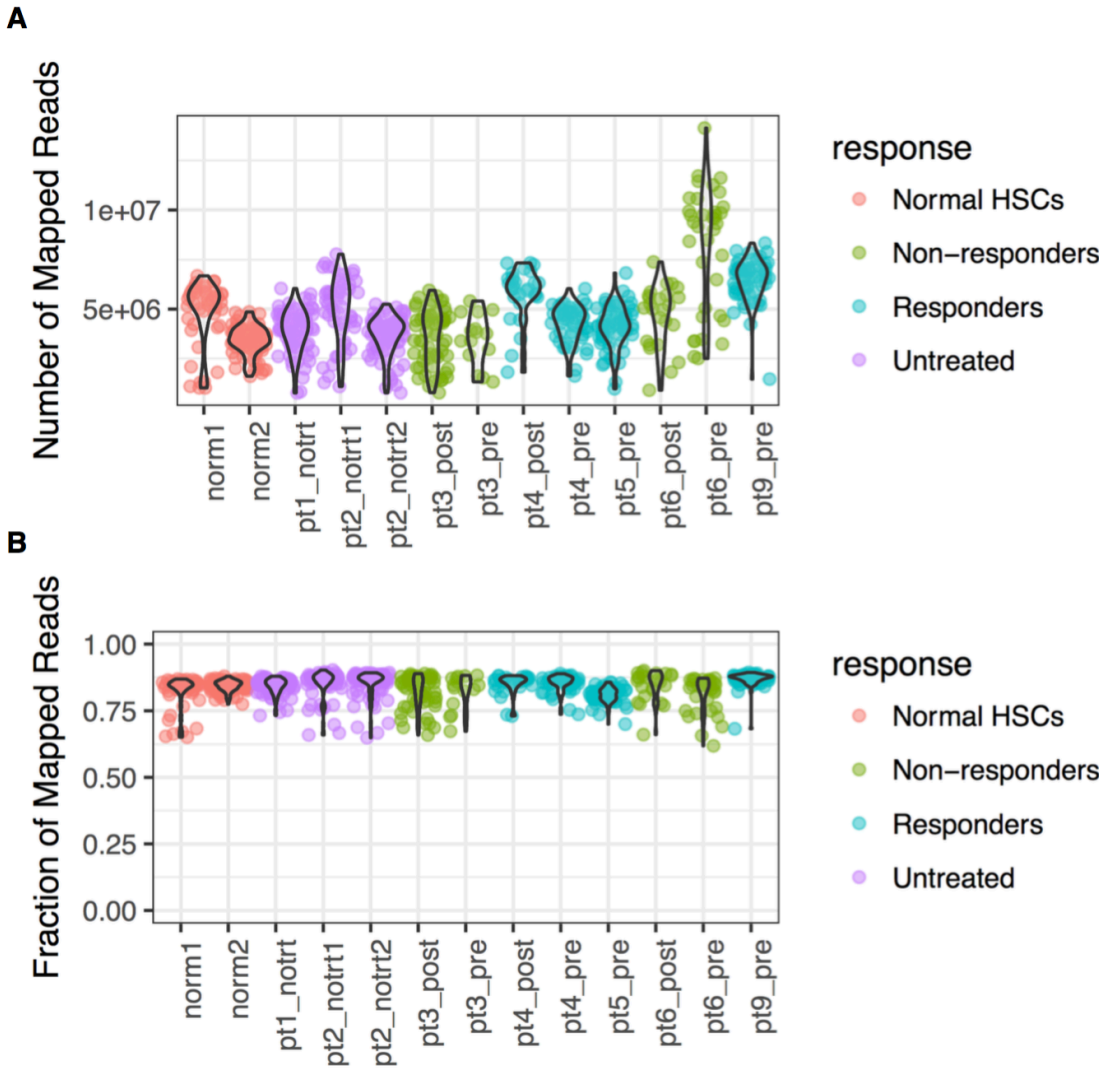


Figure 4.3. Mapping rates of scRNA-seq data. A. Number of mapped reads in each cell in each sample (mean: 4.6 million). **B.** Fraction of reads mapped in each cell (mean: 0.84)

extensive heterogeneity (**Figure 4.4A**). To tease out these signals, explore the transcriptional landscape of MDS and normal HSCs, and visualize relationships between different groups of cells, we used the dimensionality reduction method t-distributed stochastic neighbor embedding (t-SNE), and found that the majority of MDS and normal stem cells are well differentiated by transcriptional signals even in reduced 2-dimensional space (**Figure 4.4B**). Moreover, HSCs from patients who did not respond to decitabine therapy cluster more distantly from normal HSCs than those from responders even prior to therapy, suggestive of response biomarkers. To test whether the primary source of variation observed on the t-SNE projection is derived from differences in cell cycle phase, we calculated phase scores based on expression of known cell cycle markers, and found that signals other than cell cycle phase must explain the observed separation (**Figure 4.5**). These signals are explored further below using differential expression analysis, pseuodotemporal lineage ordering, and cell state identification. As several spliceosome genes are recurrently mutated in MDS (Alderton, 2015) and in our patient cohort (**Figure 4.1B**), aberrant alternative splicing, a likely key driver of MDS pathogenesis, may contribute to differentiating these groups in addition to gene expression.

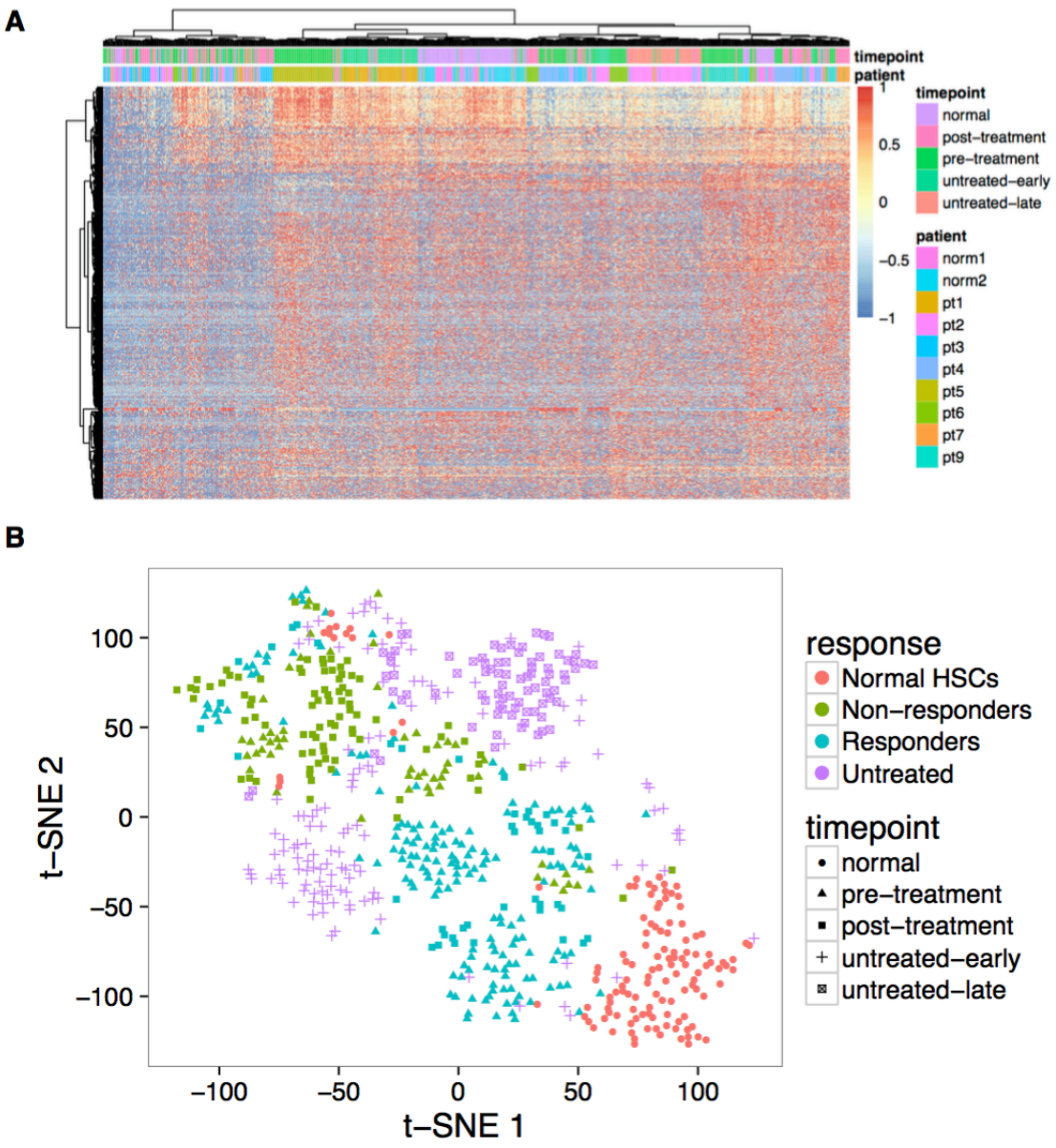


Figure 4.4. Visualizing MDS transcriptomes. A. Heatmap of the Z-score of expression (log2(FPKM+1)) of highly expressed genes (mean log2(FPKM) of 2 or higher) across all samples. **B.** T-distributed stochastic neighbor embedding (t-SNE) projection of single cell transcriptomes showing clusters separating by response group, with non-responders furthest from the cluster of normal HSCs.

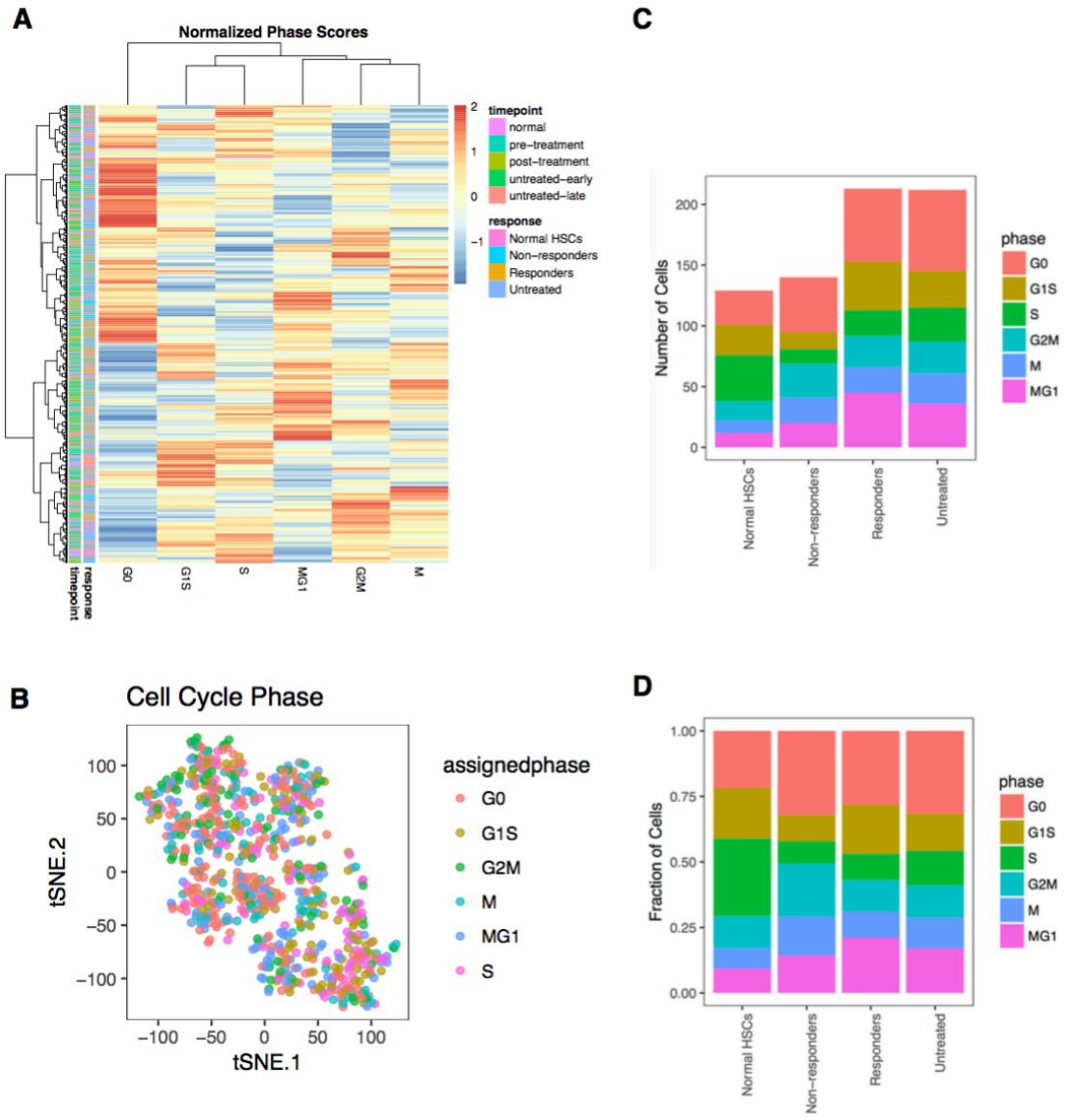


Figure 4.5. Cell cycle phases of single cells. A. Heatmap of normalized scores calculated in each cell for each cell cycle phase. **B.** Assigned phases of cells on t-SNE projection **C.** Number of cells and **D.** Fraction of cells of each response group assigned to each phase.

DISCO: a novel method for analyzing alternative splicing in single cell and other large scale RNA-seq data. scRNA-seq yields a unique view into the distributions of isoforms in single cells otherwise unobservable with bulk RNA-sequencing. For example, for a gene with two isoforms, a 50% measured expression of each isoform from bulk data may refer to all cells expressing both isoforms in equal proportions or half the cells in the population expressing only isoform 1 and the other half expressing only isoform 2 (**Figure 4.6A**). Similar to reports on bulk vs. single-cell gene expression (Shalek et al., 2013), isoform abundance quantifications (relative to other isoforms of the same gene) averaged across single cells recapitulates the signal observed in bulk sequencing measurements (r^2=0.778), while any single cell reveals heterogeneous isoform expression otherwise missed by bulk sequencing (**Figure 4.6B**). This heterogeneous expression of isoforms may be integral to defining functional cell types, cancer progression, and response to therapies. With the advent of single cell technologies enabling sequencing of full-length mRNA transcripts, heterogeneous isoform expression can now be studied in a transcriptome-wide manner.

Current computational methods for analyzing alternative splicing from bulk RNAseq data do not translate well to the single cell setting since they either do not model comparisons between multiple sets of samples (Katz, Wang, Airoldi, & Burge, 2010), are too slow to scale to hundreds of samples (Trapnell et al., 2012), or do not address the complex multi-modal distributions

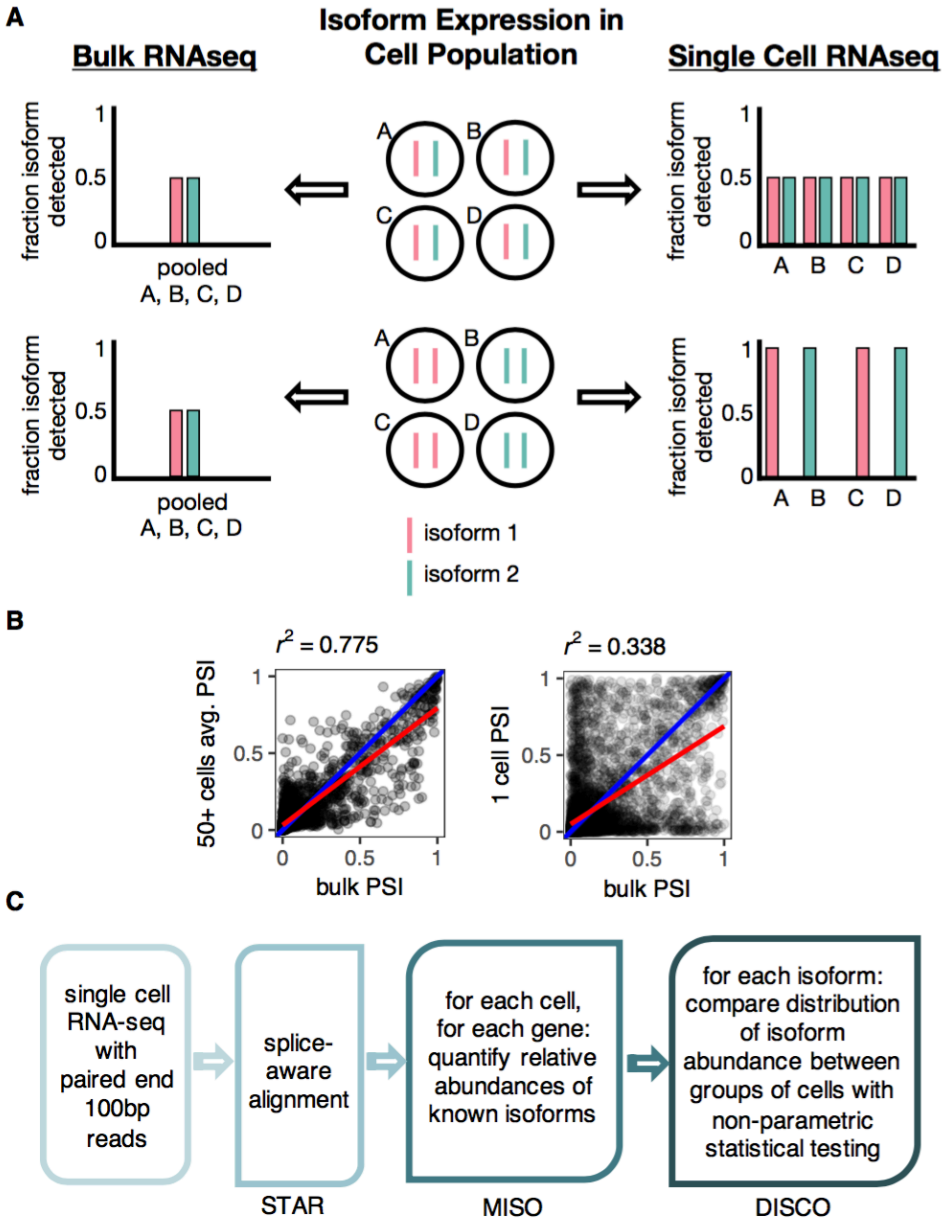


Figure 4.6. Single Cell Isoform Analysis with DISCO. A. Theoretical example of new information gained by isoform analysis at single cell resolution compared to bulk sequencing. **B.** Comparison of isoform abundance relative to other isoforms of the same gene (PSI; percent spliced isoform) between bulk (~1000 pooled cells) RNA-seq and the average across at least 50 cells shows high correlation, similar to gene expression (left); comparing bulk with a single HSC shows low correlation, suggesting a high degree of cell-cell heterogeneity (right). **C.** Schematic of the analysis pipeline for DISCO (Distributions of Isoforms in Single Cell Omics), a novel method for comparing relative isoform abundances in sample groups of single cell transcriptomes.

of isoform expression observed at the single cell level (Shalek et al., 2013). To address this issue, we developed DISCO (Distributions of Isoforms in Single Cell Omics), a software package to facilitate the identification and visualization of differential isoform usage from the hundreds or more samples of scRNA-seq experiments. Our analysis pipeline consists of splice-aware alignment with STAR, inferring relative abundances of each isoform in each cell with software such as MISO, and using Kolmogorov-Smirnov tests with correction for multiple testing to identify significant shifts in the distributions of isoform expression between two sample groups of interest (**Figure 4.6C**).

DISCO recapitulates known effects of SRSF2 mutations. SRSF2 is mutated in 20-30% of MDS cases (E. Kim, Ilagan, Bradley, & Abdel-Wahab, 2015). Previous studies in mouse models have shown that SRSF2 P95H mutations alter the protein's preference for exonic splice enhancers (ESEs), and identified an increased occurrence of CCNG motifs over GGNG in the mutated samples (E. Kim et al., 2015). To investigate whether these changes are preserved in MDS HSCs, we used DISCO to identify differentially spliced cassette exons between patients with and without SRSF2 P95 mutations and tested for the enrichment of 4-, 5-, and 6-mers in these exons above a background distribution measured from 1000 randomly selected, expressed exons. We observed an increase in the CCNG motif in exons included more in SRSF2 mutated samples as well, suggesting not only that the functional

effects of SRSF2 P95 mutations seen in mouse cells and human cell lines are conserved in MDS patient HSCs but also offering a proof of principle of the DISCO method (**Figure 4.7**).

SRSF2 mutant MDS HSCs exhibit unique splicing changes. In addition to shifts in cassette exon inclusion, we detected heterogeneous expression of isoforms within single cells and significant shifts in isoform expression between SRSF2 mutated (mut) and not mutated (wt) in 10 genes at FDR< 0.05 after filtering for coverage and mean shifts (**Figure 4.8**). These differentially spliced genes (DSGs) are enriched for gene ontology (GO) terms involving ribosomal functions (GO:0006614, GO:0000184, GO:0006364, GO:0006412, GO:0006413, GO:0019083) and interferon signaling (GO:0035455, GO:0034341, GO:0046597, GO:0060337, GO:0045071, GO:0035456, GO:0009615) at FDR< 0.05 since they include a set of RPGs (RPS28, RPL15, RPL29, and RPL17) and interferon-induced transmembrane proteins (IFITM2, IFITM3). IFITM3 and SPNS3 are shown in more detail (**Figure 4.9**) with violin plots revealing bimodal distributions of cellular prevalence of each isoform (**Figure 4.9A, D**), coverage plots across exons of DSGs confirming mean shifts in isoform expression (**Figure 4.9B, E**), and FPKM quantifications showing complementary information on absolute shifts (**Figure 4.9C, F**).

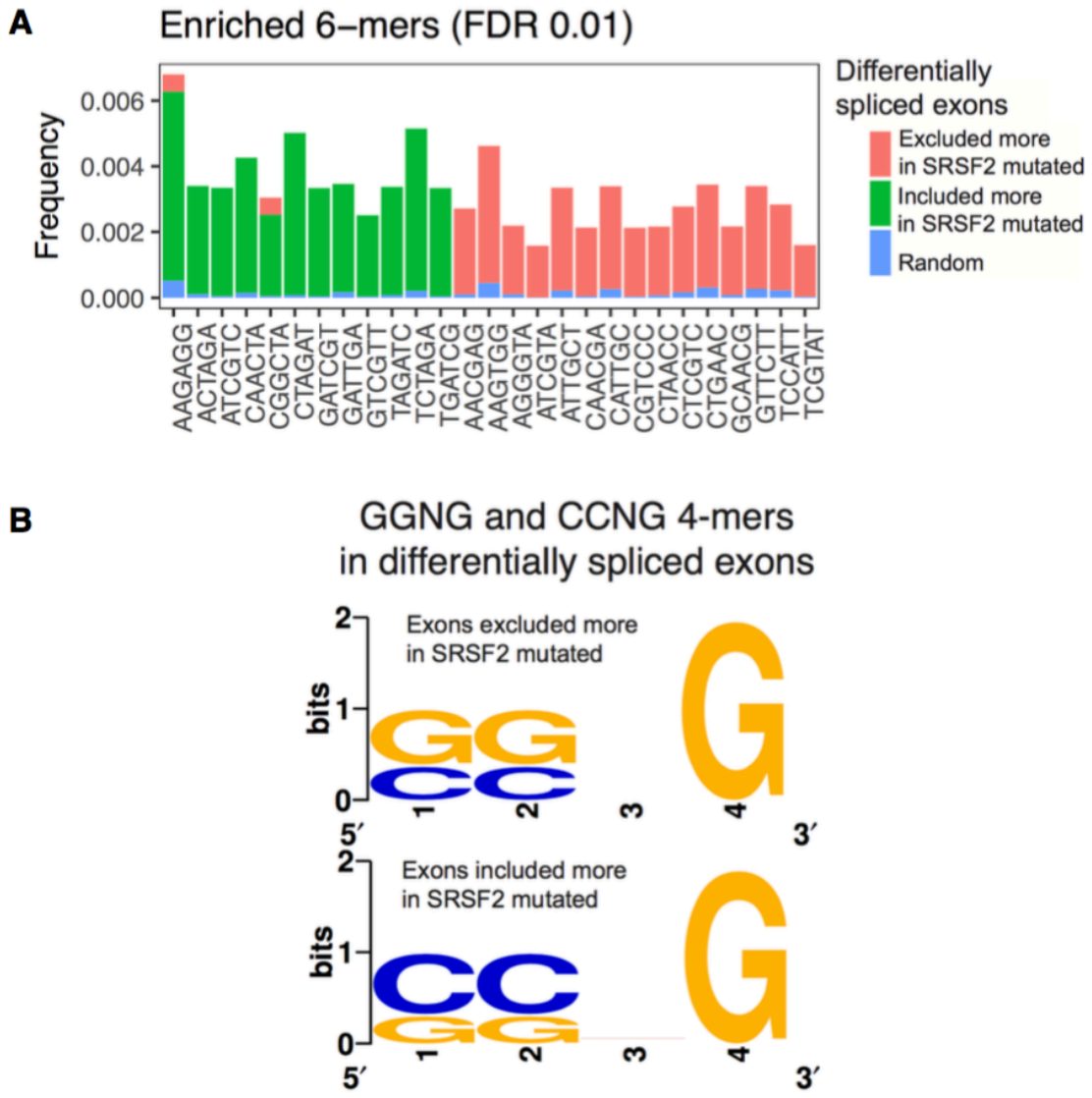


Figure 4.7. DISCO recapitulates known effects of SRSF2 mutations. A. 6-mers significantly enriched (Fisher's exact test, FDR 0.01) above background (k-mer distribution across 1000 random non-differentially spliced exons) in exons differentially spliced between cells from patients with SRSF2 mutations and patients without SRSF2 mutations. **B**. Enrichment of GGNG and CCNG 4-mers in differentially spliced exons showing that exons spliced in more in SRSF2 mutated are more enriched for CCNG than GGNG.

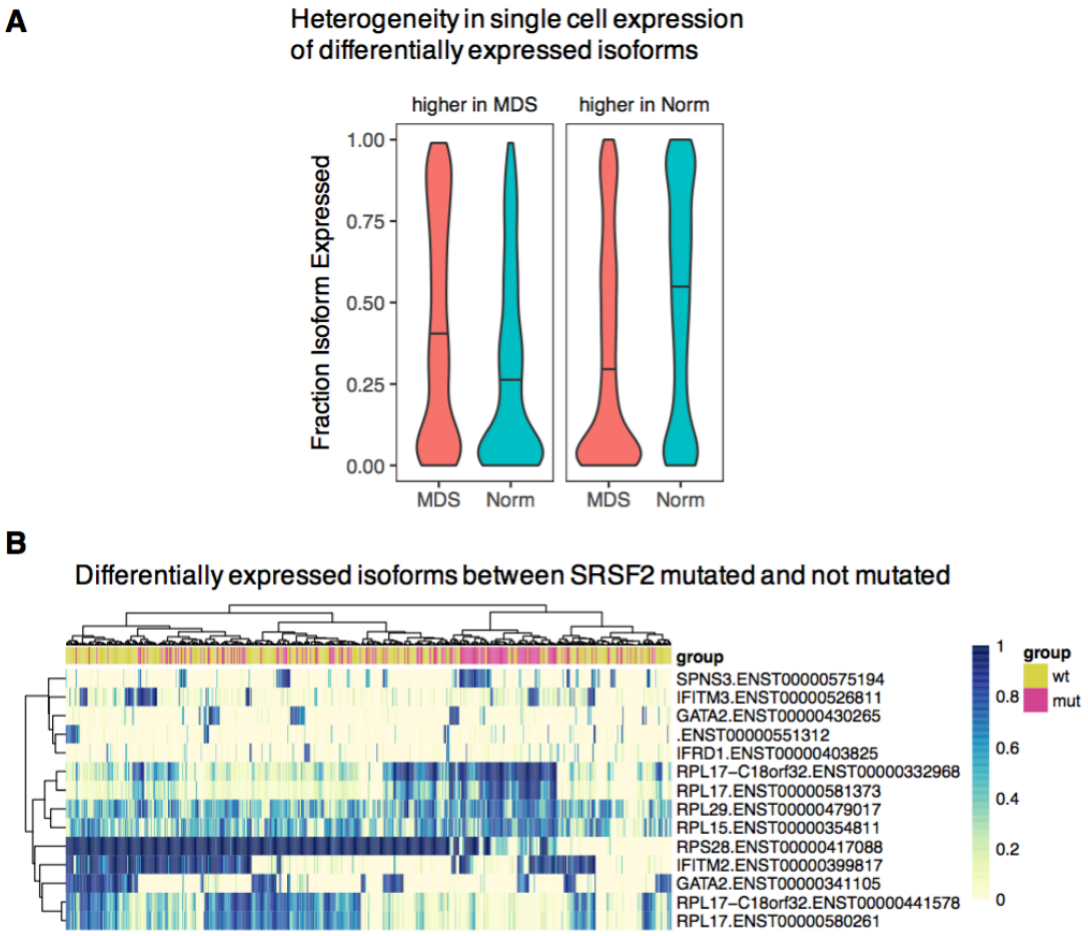


Figure 4.8. Differentially expressed isoforms in SRSF2 mutated HSCs. A. Distribution of fraction isoform expressed values aggregated across isoforms significantly higher in MDS (left) and lower in MDS (right). **B.** Heatmap of fraction isoform expressed of each DEI in SRSF2 mutated (pink) compared to WT (yellow).

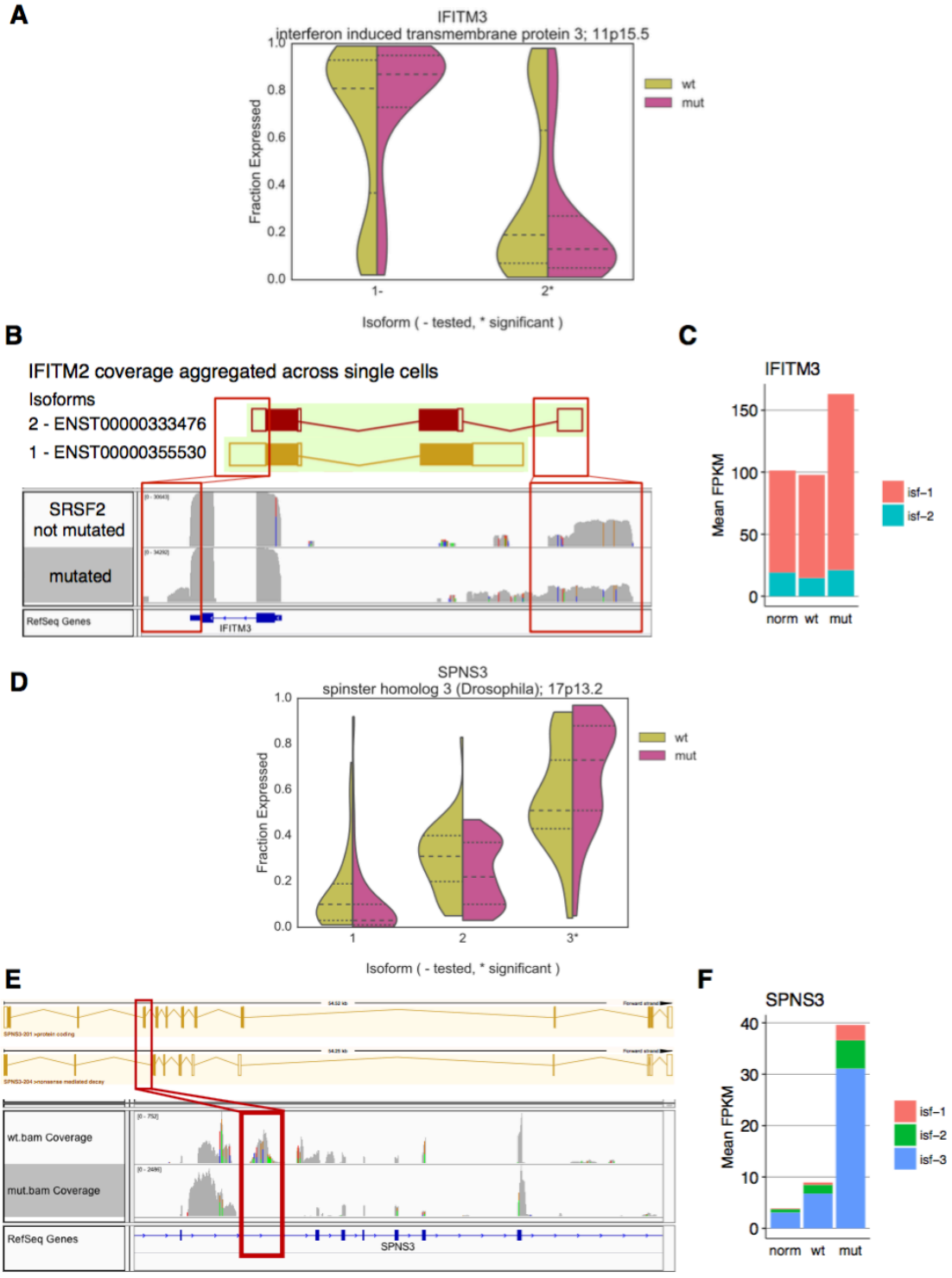


Figure 4.9. DISCO results for 2 DSGs in SRSF2 mutated. A. Distribution of fraction isoform expressed for the 2 isoforms of IFITM3. **B.** Read coverage across gene body of IFITM3. **C.** Mean expression (FPKM) of each IFITM3 isoform in normal HSCs, SRSF2 mutated, and WT. **D-F.** Like A-C for SPNS3.

Aberrant expression of immune signaling and hematopoiesis genes in SRSF2 mutated cells. Since SPNS3 gene expression drastically increases in SRSF2 mutated (**Figure 4.9F**) cells, we explicitly tested differential gene expression to identify other similarly affected genes and found 53 differentially expressed genes (DEGs) with increased expression in SRSF2 mutated and 100 DEGs with decreased expression at FDR< 0.05 (**Figure 4.10A**). Of the DSG list, these include SPNS3 and GATA2. Genes more highly expressed in SRSF2 mutated are enriched for GO terms involving interleukin signaling (IL2RA, IL2RG) and myeloid differentiation (HOXA7, CCL3, HIST1H4H) (**Figure 4.10B**). Dysregulated innate immune and inflammatory signaling has been implicated in MDS pathogenesis (Ga??n-G?mez et al., 2015; Keerthivasan et al., 2014); our findings of aberrant splicing and expression of genes in these pathways suggest these as possible disease-causing mechanisms of mutated SRSF2.

MDS HSCs exhibit differential isoform usage. To identify genes and molecular pathways that distinguish MDS stem cells from normal HSCs irrespective of SRSF2 mutations, we analyzed differentially expressed isoforms and genes between pre-treatment MDS cells and age-matched normal controls. Of the 8,517 isoforms passing minimum coverage and average PSI filters in at least 50 cells in each group, DISCO identified 45 isoforms to be significantly differentially expressed spanning 38 DSGs at an

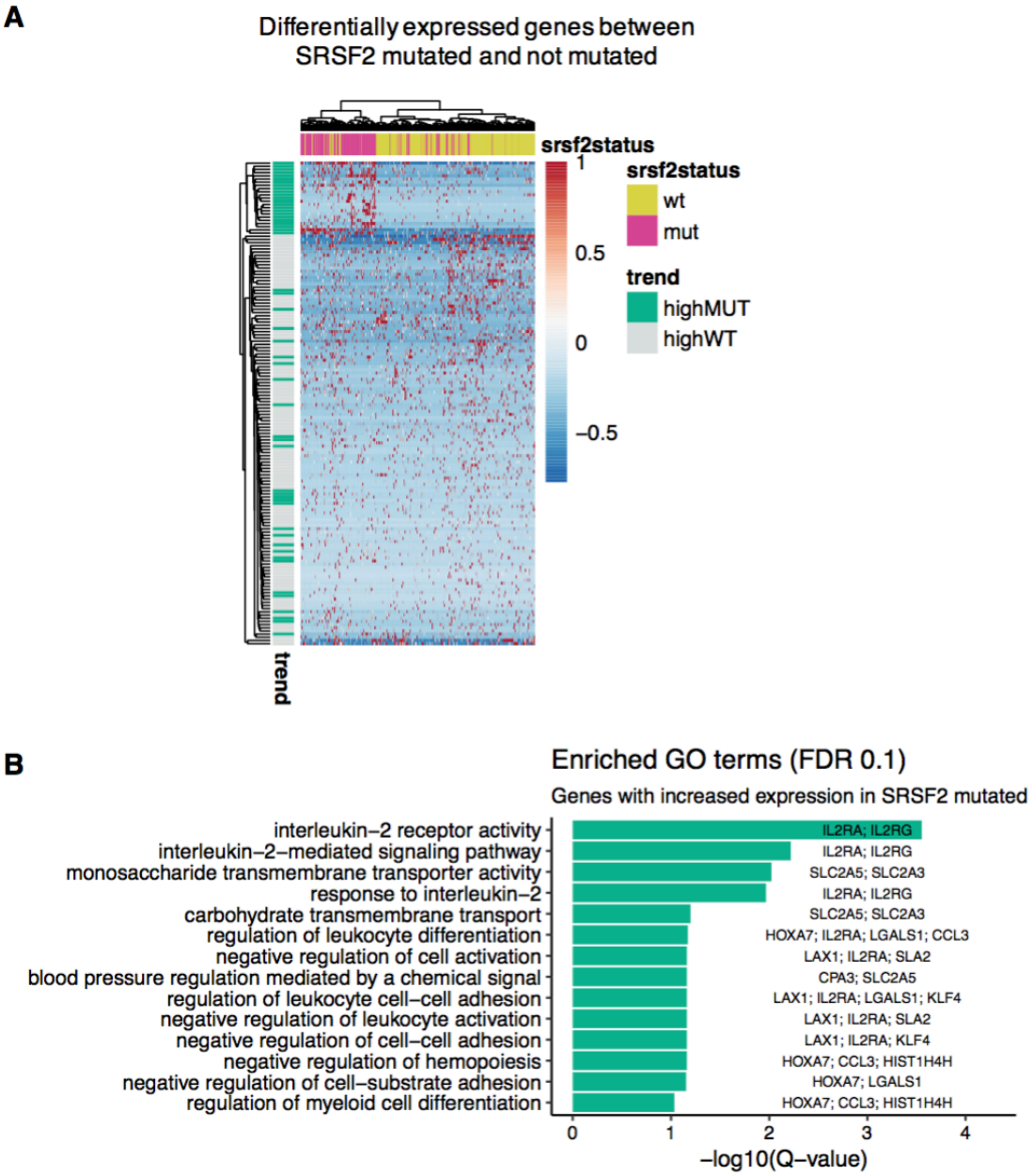


Figure 4.10. Differential gene expression in SRSF2 mutated HSCs. A. Heatmap of mRNA expression (z-score of log2(FPKM +1)) SRSF2 mutated DEGs. **B.** GO terms enriched in SRSF2 mutated DEGs.

FDR <0.05 and a >10% shift in mean isoform level (**Figure 4.11**). Two of the DSGs are TAGAP and ADGRE5, which demonstrate the observed cellular heterogeneity in isoform expression patterns (**Figure 4.12A,C**). Visualizing alignment of reads aggregated across cells of each group (**Figure 4.12B,D**) and mean FPKM quantifications of each isoform (**Figure 4.12E,F**) confirm the shift in expression identified by DISCO. The 38 DSGs are enriched for GO terms involving ribosomal protein functions, transcription regulation / elongation through RNA polymerase II promoter (GO:0045944, GO:0006357, GO:0006368), golgi organization (GO:0007030), and DNA methylation (GO:0006306). The genes in the DNA methylation GO term are transcription factors ATF7IP and FOS, of which FOS is a proto-oncogene that functions in cell proliferation and differentiation through interactions with the JUN family. FOS also belongs in the RNA polymerase II GO terms along with PPP3CA, CCNH, TBL1XR1, NCOA7, SUB1, RAD21, SOD2, and MLLT3.

Dysregulated ribosomal protein expression and p53 signaling in MDS stem cells. As suggested by the t-SNE projection of transcriptional features, MDS stem cells occupy distinct transcriptional landscapes and at baseline (pre-treatment) express 491 genes at significantly differentially expressed levels from normal controls (FDR< 0.01) (**Figure 4.13A**). The KEGG pathways significantly enriched (FDR 0.1) in genes overexpressed in MDS HSCs are p53 signaling, cancer, viral carcinogenesis, hematopoietic cell lineage, and

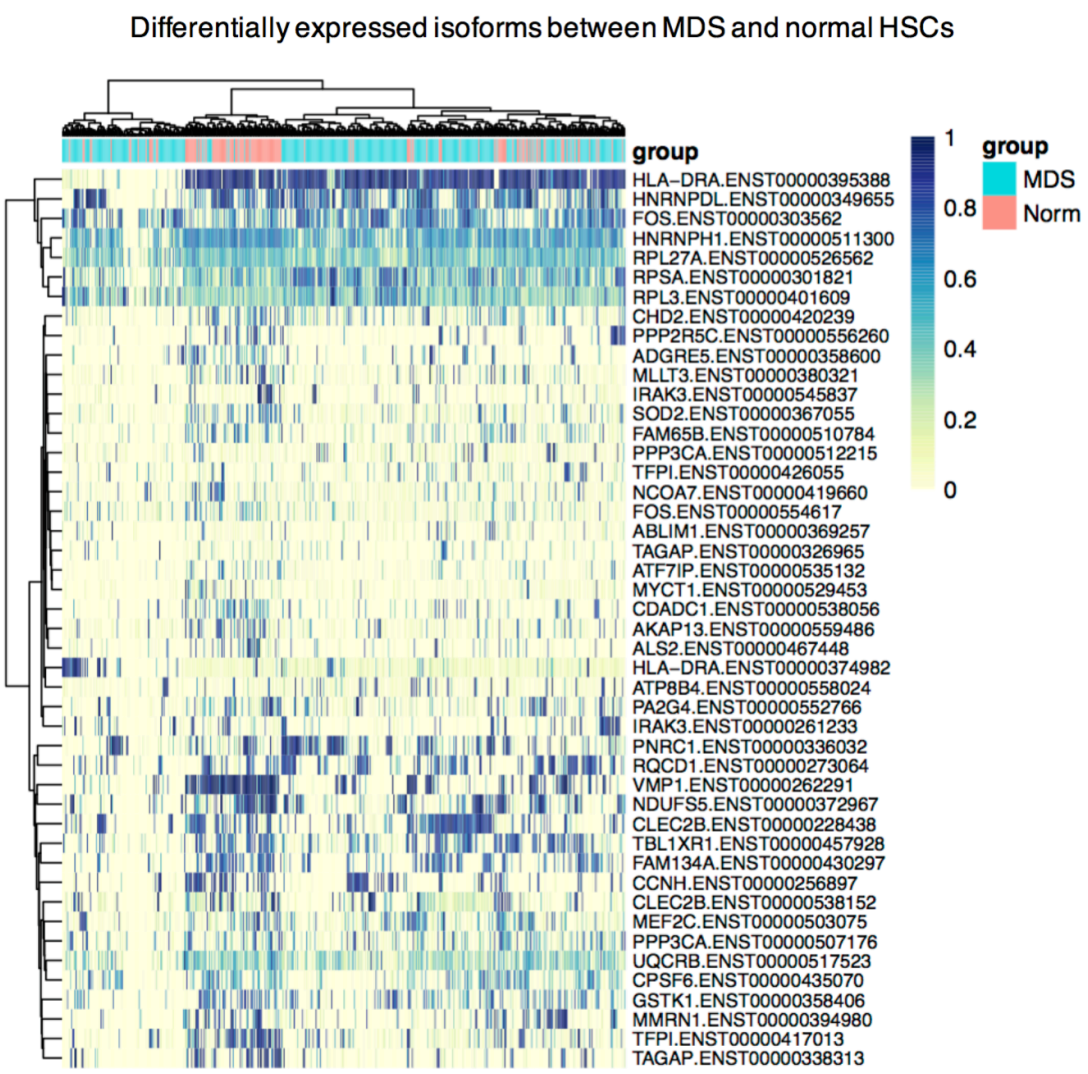


Figure 4.11. Altered splicing in MDS HSCs. Heatmap of isoform ratios of isoforms differentially expressed between normal HSCs and pre-treatment MDS HSCs.

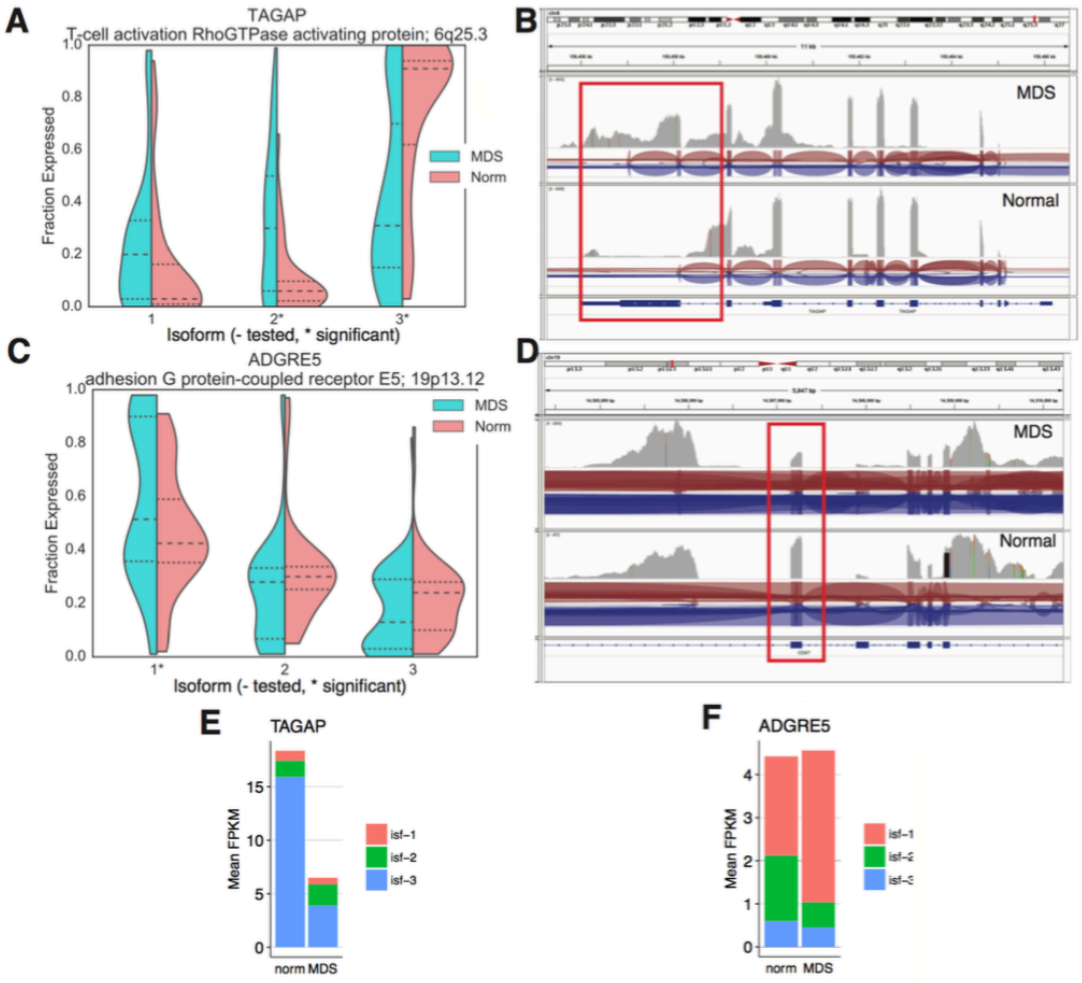


Figure 4.12. Differential splicing of ADGRE5 and TAGAP in MDS HSCs. A. Distribution of fraction isoform expressed for TAGAP. **B.** Coverage across TAGAP gene body aggregated across cells of each group (MDS, top; Normal, bottom). **C.** Distribution of fraction isoform expressed for ADGRE5. **D.** Coverage across ADGRE5 gene body aggregated across cells of each group (MDS, top; Normal, bottom). **E.** Average expression (FPKM) of each isoform for TAGAP in normal and MDS HSCs. **F.** Average expression (FPKM) of each isoform for ADGRE5 in normal and MDS HSCs.

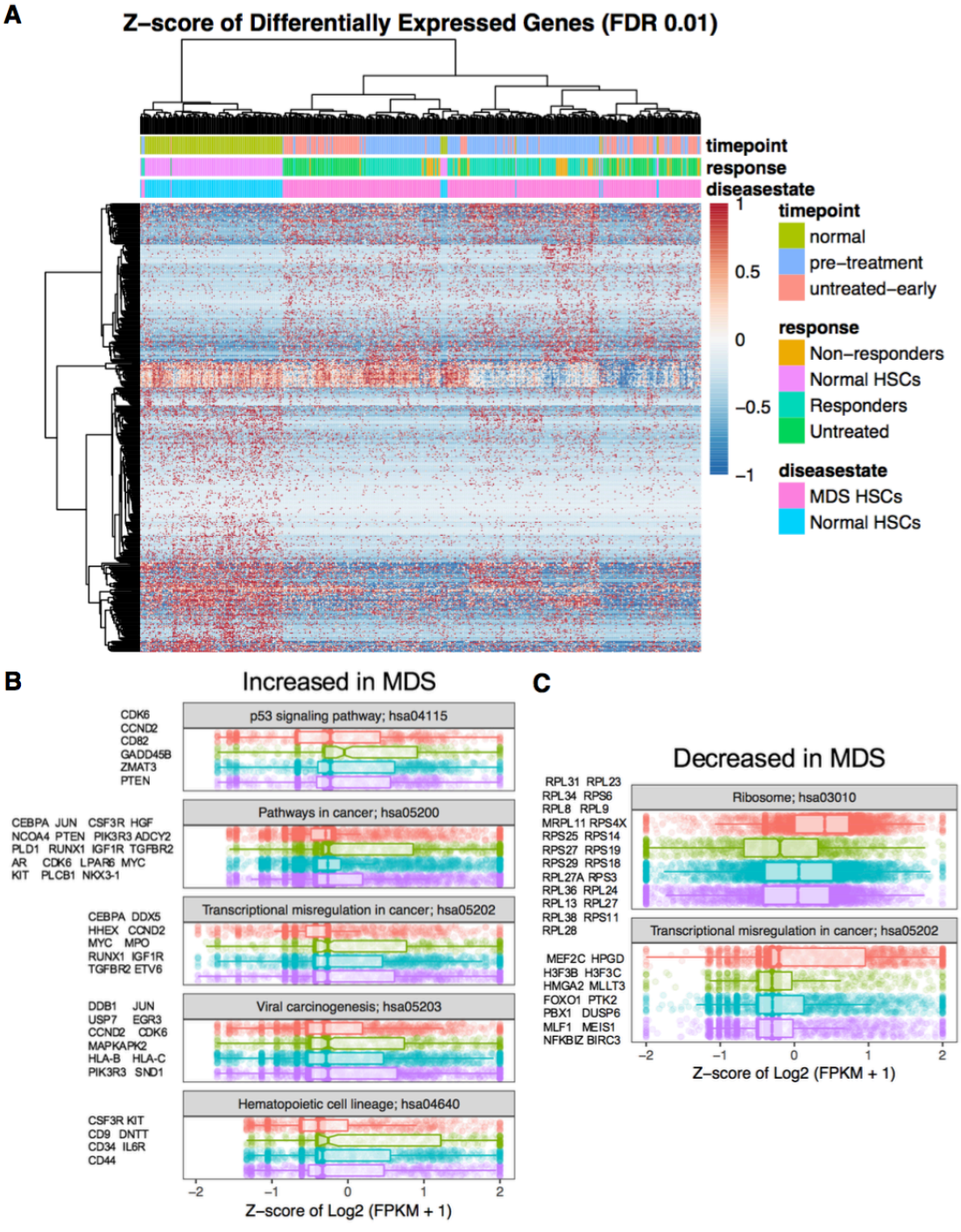


Figure 4.13. Differential gene expression in MDS HSCs. A. Heatmap of mRNA expression (z-score of log2(FPKM+1)). **B., C.** KEGG pathways significantly enriched (FDR 0.1) in DEGs.

transcriptional misregulation in cancer (**Figure 4.13B**). Genes under expressed in MDS HSCs are also enriched for the KEGG pathways transcriptional misregulation in cancer as well as ribosome function (**Figure 4.13C**). Decreased expression of RPGs in MDS HSCs compared to age-matched normal controls is noteworthy, as defects in ribosome function have been previously implicated in MDS pathogenesis (McGowan et al., 2008; Raza & Galili, 2012; Rinker et al., 2016) and recent work has shown that ribosome heterogeneity affects the transcriptional efficiencies of specific mRNAs [23,24].

Ribosomal proteins down-regulated and differentially spliced in decitabine non-responders. Differences in ribosomal protein gene and isoform expression were also observed between MDS patients, with non-responders expressing even lower levels of RPGs than responders (**Figure 4.14**). Of the 41 isoforms identified by DISCO as differentially expressed, 25 are RPGs (**Figure 4.15A**). Isoform shifts are again observed to follow heterogeneous patterns with clusters of cells shifting from low to high isoform proportions (**Figure 4.15B**), and these shifts likely have functional effects since the vast majority of these isoforms have been studied and are known to be either protein coding or undergo nonsense mediated decay (NMD). Overall, 40 out of the 80 RPGs are differentially expressed or spliced between MDS stem cells and normal controls, and 17 RPGs are differentially expressed or spliced within MDS stem cells between pre-treatment responders and non-

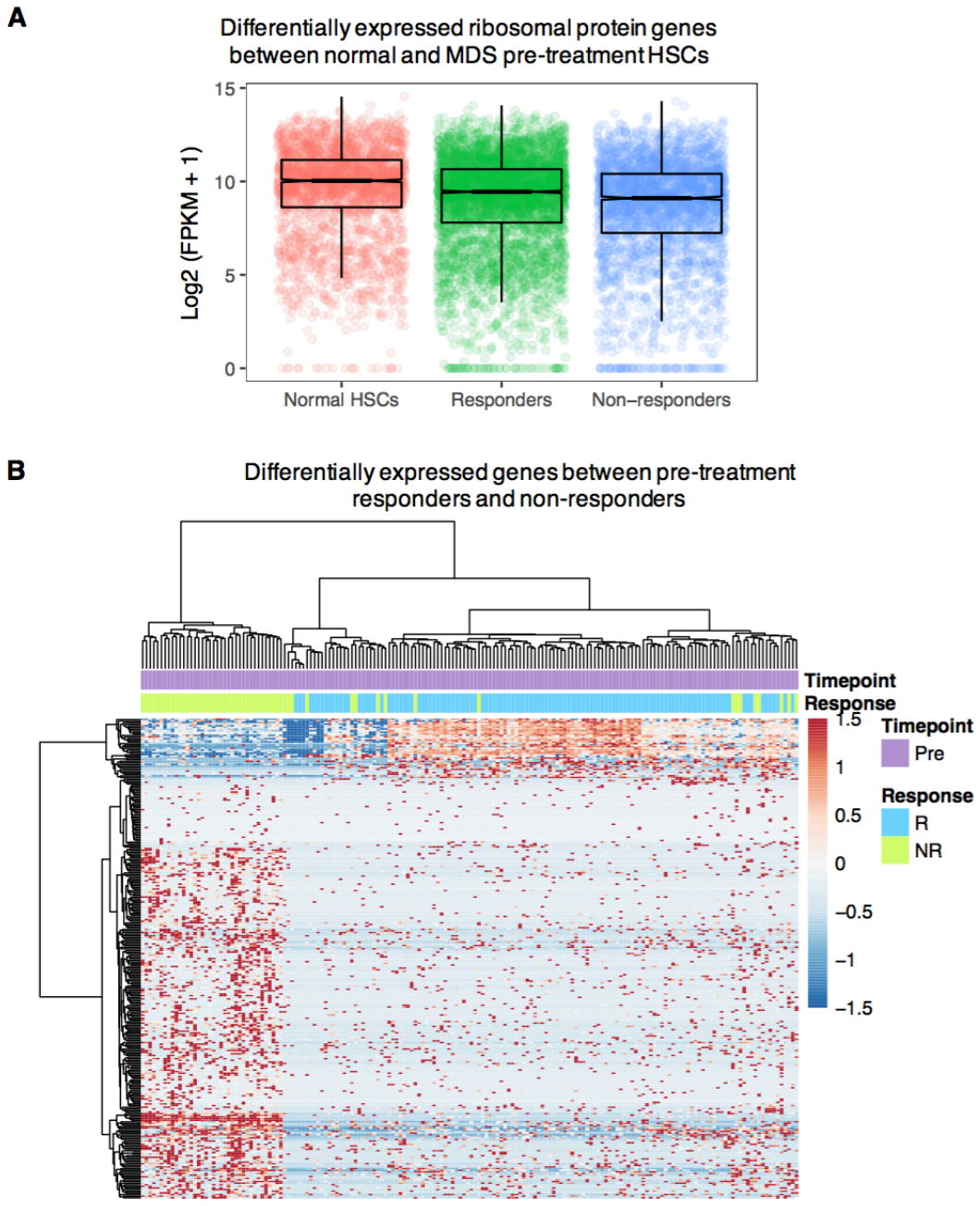


Figure 4.14. Differential gene expression between responders and non-responders. A. Ribosomal protein genes differentially expressed between normal and MDS HSCs are even lower in non-responders than responders. **B.** Heatmap of gene expression (z-score of log2[FPKM + 1]) of DEGs between pre-treatment responders and non-responder HSCs.

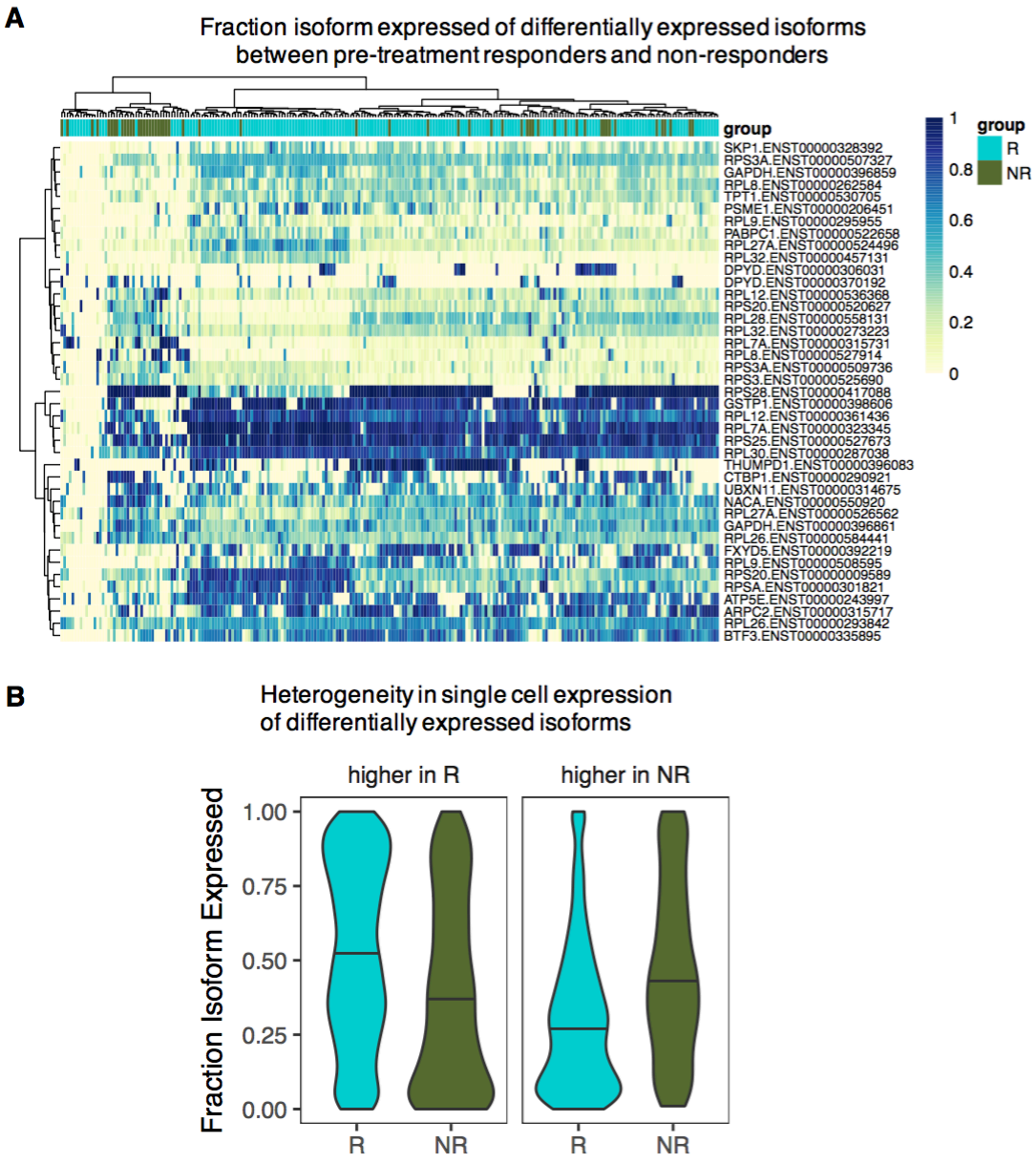


Figure 4.15. Altered splicing patters between responders and non-responders. A. Heatmap of fraction isoform expressed of DEIs between pre-treatment R (blue) and NR (green). **B.** Distribution of fraction isoform expressed aggregated across all DEIs higher in R (left) and higher in NR (right).

responders (**Figure 4.16**). These results suggest that ribosomal dysregulation not only plays a role in MDS disease mechanisms but also in response to decitabine. This is consistent with the fact that the clinical subtype of MDS patients with a 5q deletion (del5q), where the deletion of RPS14 is thought to be a key driver of pathogenesis, do not respond well to cytidine analogs (ex. decitabine) and are usually treated with lenalidomide (Raza & Galili, 2012). It stands to reason based on our data, since none of the patients studied here have del5q, that this effect is not limited to del5q-related loss of RPS14 and can potentially be generalized to decreased expression of a wide variety RPGs through mechanisms besides chromosomal abnormalities.

Analysis of single cell heterogeneity identifies distinct cell state distributions differing by response group and time point. Investigating gene and isoform expression in MDS stem cells identified dysregulated immune signaling, p53 activation, and ribosome biogenesis. Single cell resolution of our data enables us to extend these results a step further by treating each cell as a unique time point to understand dynamics of these changes and how they fluctuate within heterogeneous cell populations across patient samples. We ordered cells based on transcriptional profiles on a singular axis, termed “pseudotime”, and constructed a lineage hierarchy of all cells using a graph-based tree building algorithm (Qiu et al., 2017; Trapnell et al., 2014) (**Figure 4.17**). This recapitulated the t-SNE result of non-responders

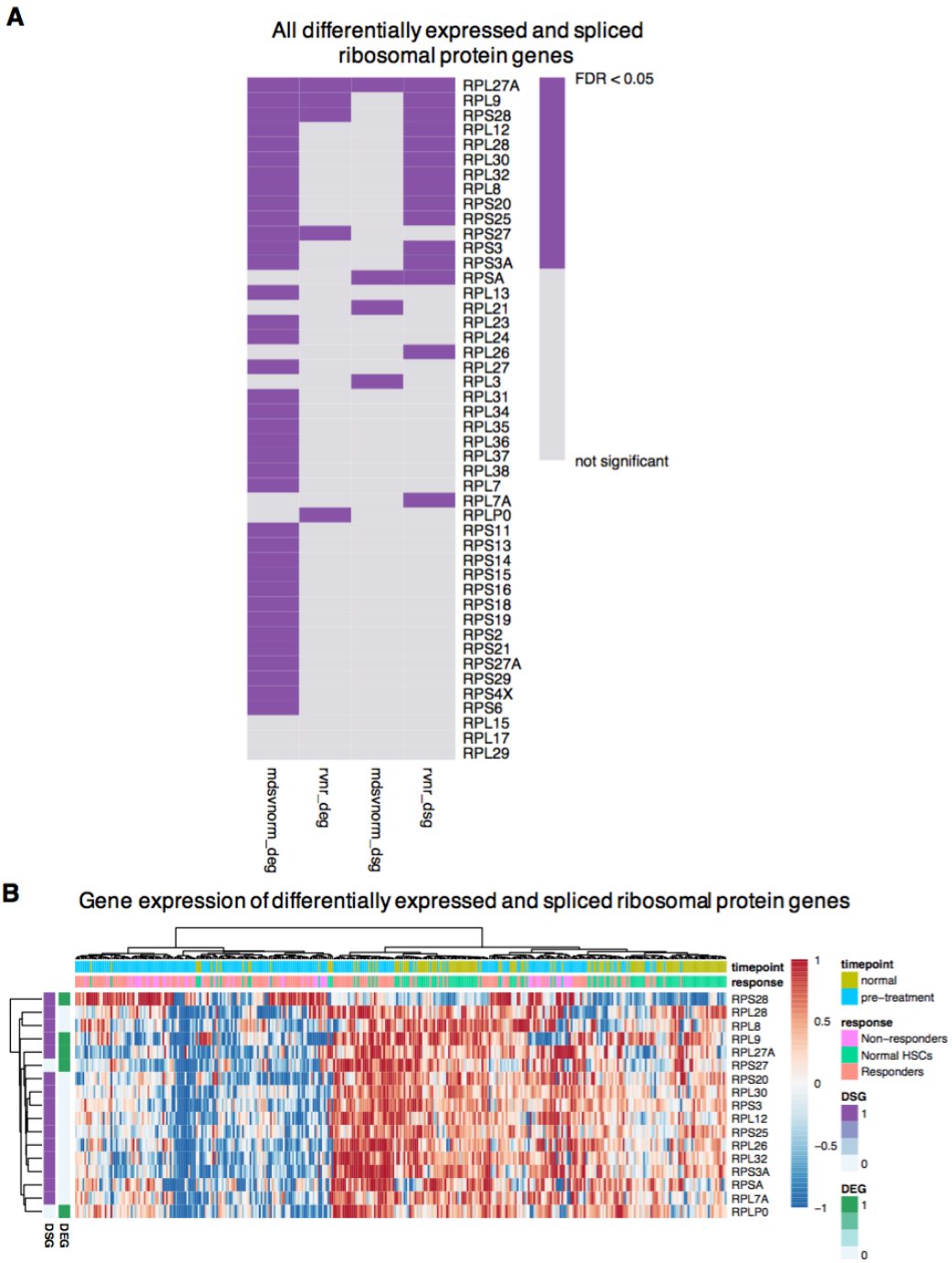


Figure 4.16. Ribosomal protein gene and isoform dysregulation. A. RPGs (rows) that are differentially expressed between MDS and normal, differentially expressed between R and NR pre-treatment, differentially spliced between MDS and normal, and differentially spliced between R and NR (columns from left to right). **B.** Heatmap of the subset of RPGs that are either differentially expressed or differentially spliced between R and NR showing the z-score of gene expression (log[FPKM+1]).

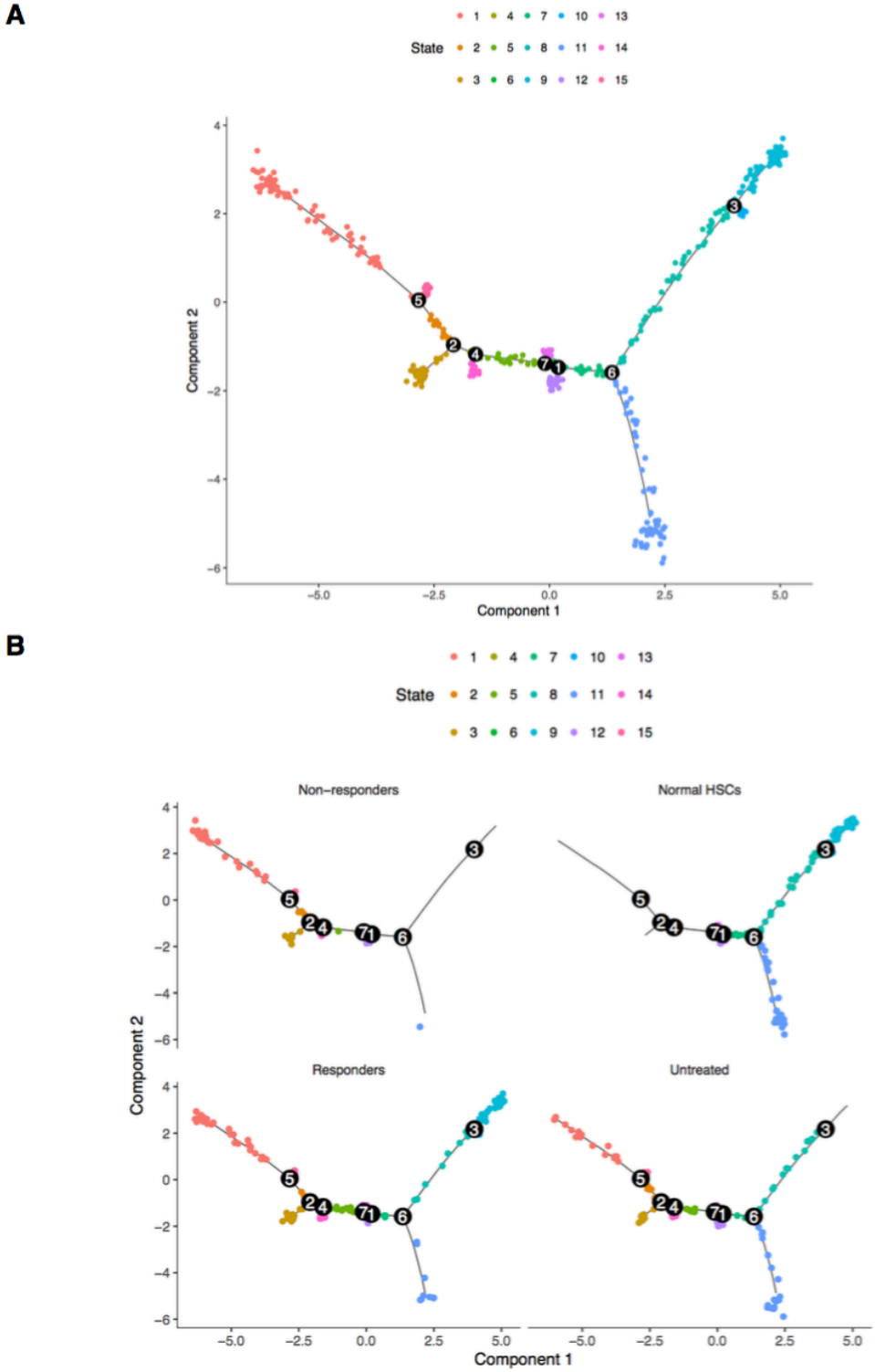


Figure 4.17. Lineage ordering and cell state identification. A. Pseudotemporal lineage projection showing cell states (colors) and branch points (black numbered circles). **B.** As in A, but with cells separated by response group.

having transcription profiles most distant from normal HSCs, and assigned cells to 21 distinct cell states. To tease out the most biologically relevant signals from the resulting complex branching of cells, we seeded the algorithm with the disease relevant hematopoietic lineage markers (stem cell regulators and myeloid differentiation genes) due to their differential expression between MDS and normal stem cells, and reassessed pseudotemporal ordering.

MDS HSCs exhibit decreased expression of stem cell regulators and enrichment of myeloid genes. We compared DEGs to a list of hematopoietic lineage markers manually curated from literature, and identified 5 HSC genes (CXCR4, HLF, MECOM, MEIS1, and NR4A2), and 2 myeloid lineage associated genes (CEBPA and MPO) to be significantly differentially expressed between MDS and normal HSCs (**Figure 4.18A**). Expression of lymphoid marker genes were detected in both MDS and normal stem cells but were not differentially expressed between MDS and normal. Notably, MDS samples contained a higher proportion of HSCs exhibiting a significant increase in myeloid gene expression, consistent with a more myeloid-biased HSC pool (**Figure 4.18A**). To further explore this signature and its implications for the MDS transcriptional landscape, we used the collective expression of the HSC and myeloid genes to assign cells a stem cell score and a myeloid score, and defined criteria for separating cells into stem cell and myeloid categories (**Figure 4.18B,C**). This classification was then used to identify a

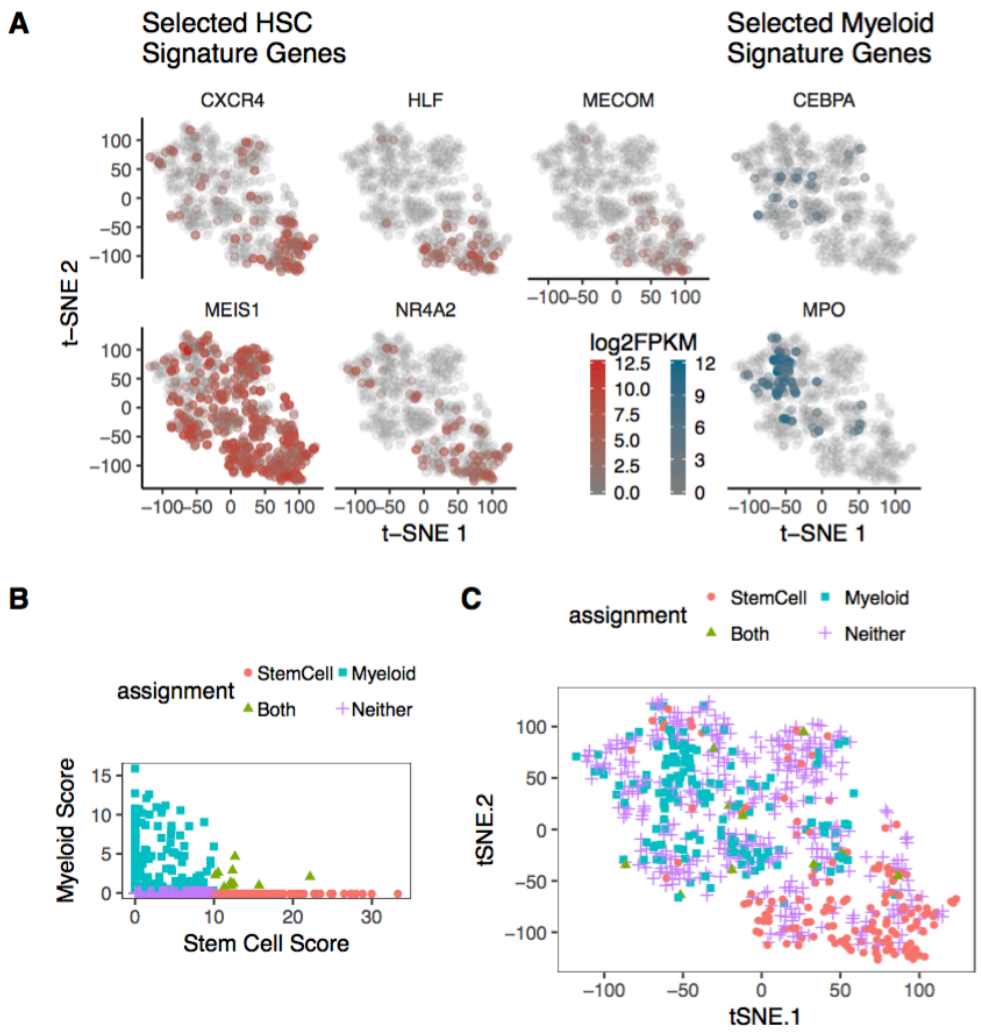


Figure 4.18. Expression of stem cell and myeloid lineage regulators differentiating MDS HSCs. A. t-SNE projections of cells colored by expression level of stem cell and myeloid regulators. **B.** Expression (log2(FPKM+1) of DEGs known to be stem cell (D) or myeloid (E) signature genes on t-SNE projection. **C.** Stem cell and myeloid scores calculated as a function of the expression of genes shown in E; cells with stem cell scores of more than 10 and myeloid score less than 1 were marked as stem cell, and cells with myeloid score more than 1 and stem cell score less than 10 were marked as likely belonging to a more myeloid state. G. Distribution of cells marked as belonging to a stem cell or myeloid state on the t-SNE projection showing that MDS HSCs are enriched for myeloid state cells and normal HSCs for stem cell state cells.

larger set of genes co-expressed with these marker genes, i.e. differentially expressed between the two categories, to use as the basis for lineage ordering of cells (80 genes at FDR <0.01).

Distinct stem cell states identified using semi-supervised pseudotemporal ordering of MDS and Normal HSCs. The semi-supervised pseudotemporal ordering of all MDS and normal HSCs based on the 80 genes identified through the above method resulted in a lineage tree with 3 main branches (**Figure 4.19A**). Normal HSCs exclusively occupied the low pseudotime branch (branch 1), HSCs from non-responders occupied the higher pseudotime branches (branches 2 and 3), and HSCs from responders and patients with stable disease occupied all branches (**Figure 4.19B, C**). Assignment of cells to states based on branch positions reveals shifts in cell states between the different groups. Proportion of state 1 cells was highest in normal and lowest in non-responders, and proportion of state 7 cells increased from responders to stable disease to non-responders (**Figure 4.19D**). These cell state assignments are robust across patient samples (**Figure 4.20**). 166 genes were found to be significantly differentially expressed along the pseudotime axis (FDR 0.01) and were enriched for pathways in systemic lupus erythematosus, cytokine–cytokine receptor interaction, p53 signaling, and more (**Figure 4.21**).

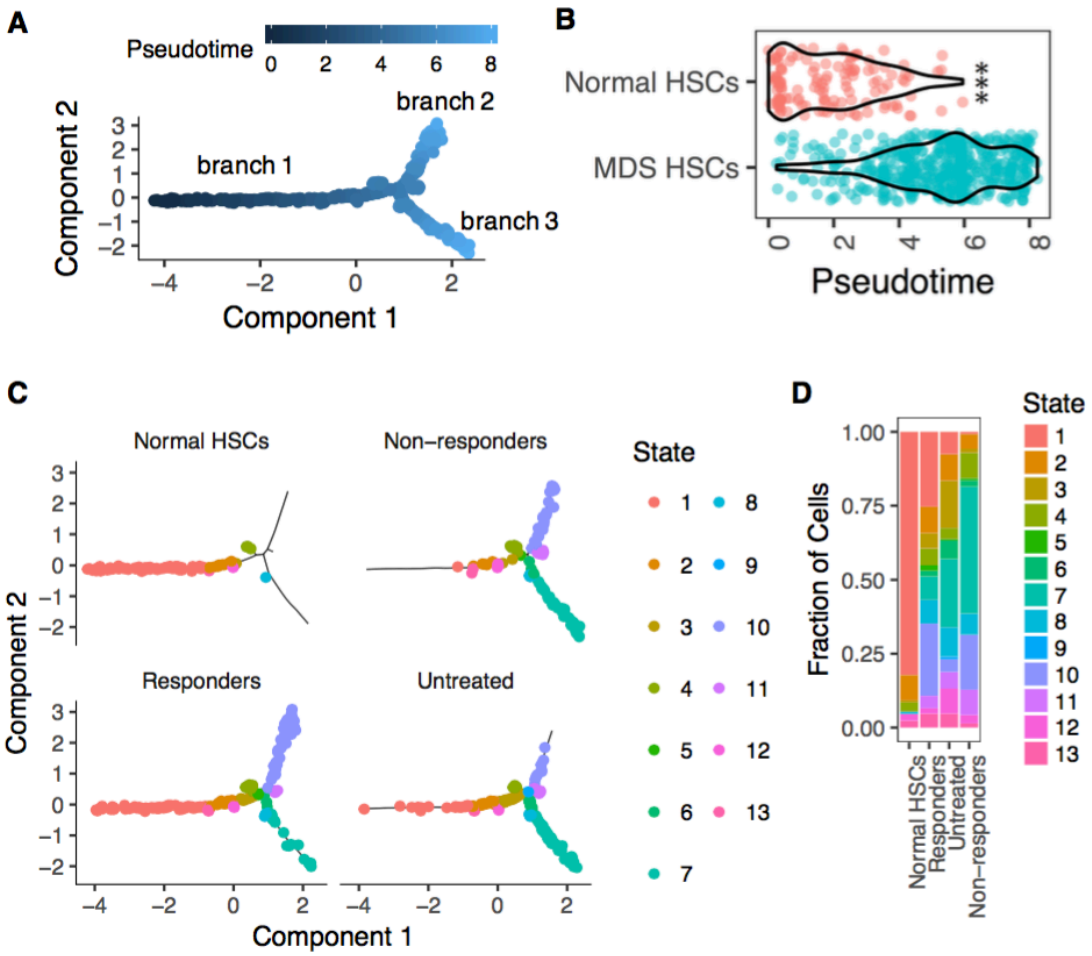


Figure 4.19. Semi-supervised lineage ordering. A. Graph-based inference of a pseudotemporal ordering of single cells based on the expression of the stem cell and myeloid signature genes shows in figure 4.18. **B.** Pseudotime distribution between MDS and normal HSCs showing a significant increase in pseudotime in MDS HSCs (t-test, $p < 0.001$). **C.** Pseudotemporal ordering of cells separated by response group. **D.** Distribution of cell state assignments based on position on lineage ordering tree.

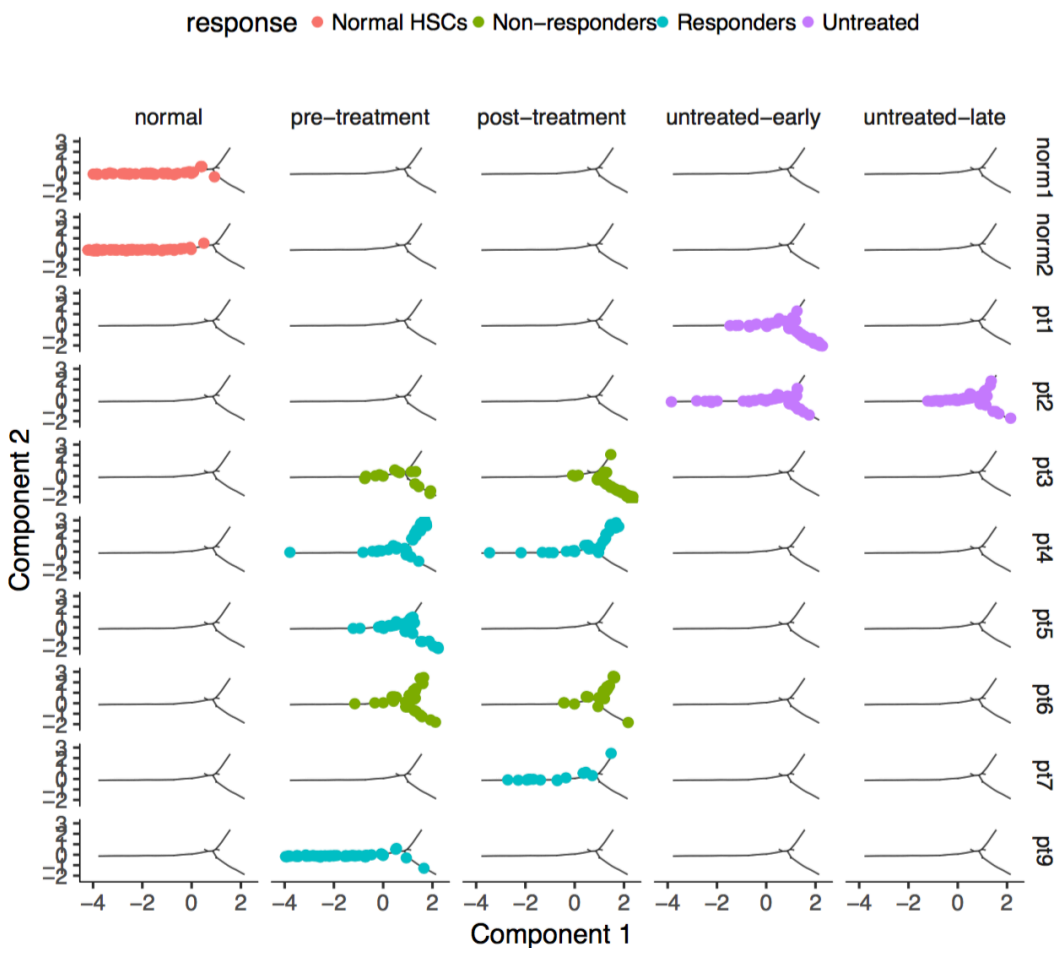


Figure 4.20. Lineage ordering of each patient sample. Pseudotemporal lineage hierarchy of each patient sample with patients listed as rows and time points in columns. Each tree represents an individual C1 run and so an individual experimental batch.

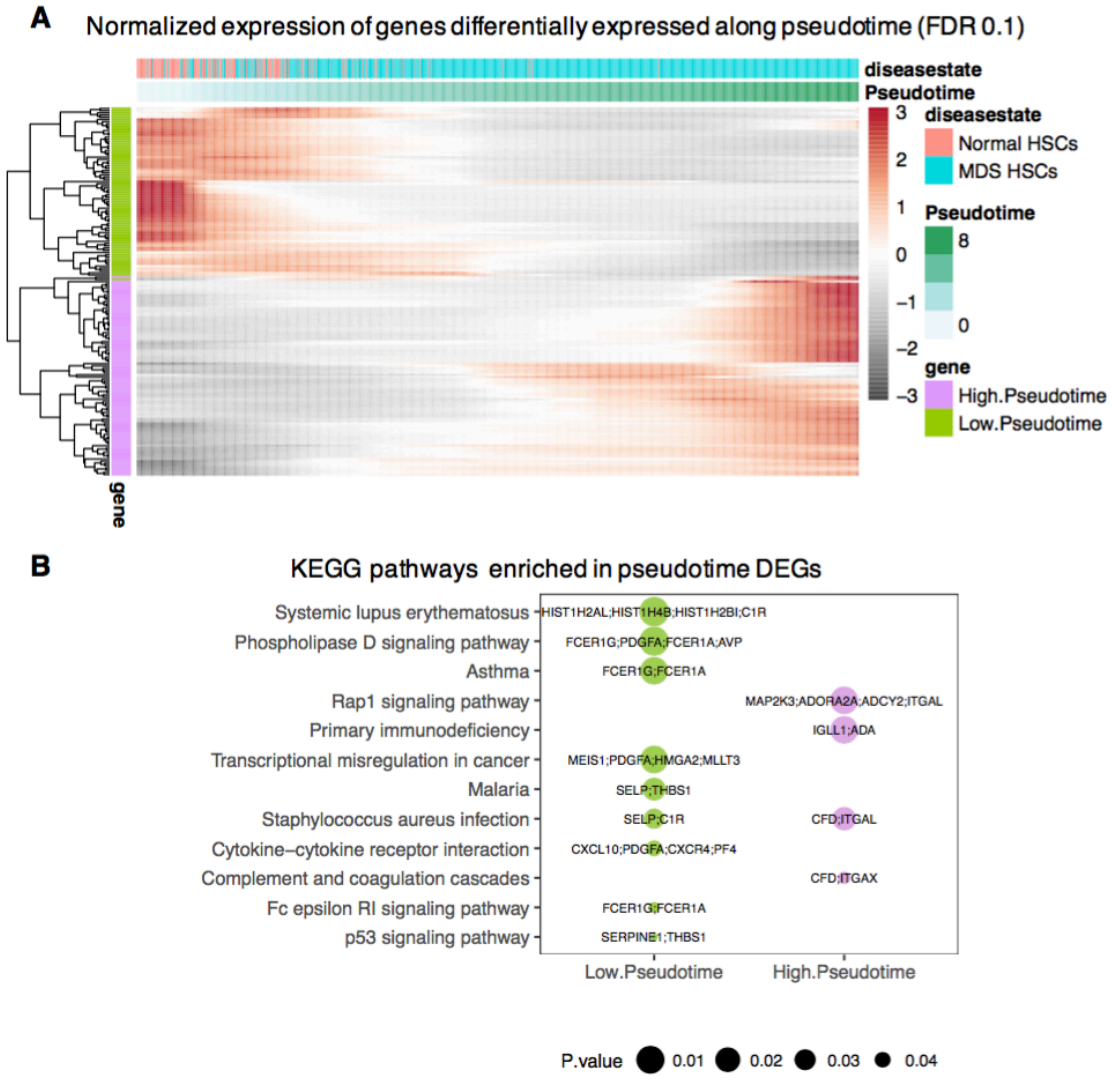


Figure 4.21. Pseudotime DEGs and pathway enrichment. E. Heatmap of normalized expression (using variance-stabilizing transformation and fitting to spline curves) of genes significantly differentially expressed at FDR 0.01 across the pseudotime axis. **F.** Enriched KEGG pathways in genes in panel E, separated by genes highly expressed in low pseudotime or high pseudotime cells.

Decitabine induces shifts in stem cell transcriptional states. Stem cell states not only differ in distributions between decitabine responders and non-responders but are also dynamic on a treatment timescale. Following therapy, the non-responder populations increased in proportion of state 7 cells, a state unique to MDS cells, and lost all cells from state 1, the state occupied by the vast majority of normal HSCs (**Figure 4.22A**). Conversely, in responders, there is a slight increase in the number of cells in state 1, although the proportion of cells in MDS-specific states (states 5-9) remains high, suggesting that decitabine treatment may not be affecting all disease-causing populations of cells. Differential expression testing comparing cells in the two high pseudotime branches revealed a significant decrease of genes in B cell receptor signaling, PI3K-Akt signaling, inflammatory bowel disease, primary immunodeficiency, non-homologous end-joining, and hematopoietic cell lineage pathways in state 7 and the other states on its branch.

Transcriptional cell states identify signatures of decitabine resistance. To determine active resistance mechanisms and potential response biomarkers we compared non-responder branch 3 cells (NR-Br3) to responder branch 3 cells (R-Br3). As mentioned previously, branch 3 cells have the most dramatic shifts from treatment with NR-Br3 and R-Br3 holding opposite trends–responders decreased from 38 to 1 cell and non-responders increased from 18 cells to 60 cells (**Figure 4.22B, C**), suggestive of a molecular

difference between NR-Br3 and R-Br3 cells capable of conferring decitabine resistance and disease progression. 553 genes were found to be more highly expressed within a subset of pre-treatment NR-Br3s, and a larger subset of post-treatment NR-Br3s, compared to all pre-treatment R-Br3s, as outlined in figure 6D.

Principal component analysis (PCA) of these 553 genes shows responder and normal cells clustering together and away from non-responder cells. Further, post-treatment NR-Br3s are almost completely separated from pre-treatment R-Br3s, with pre-treatment NR-Br3s existing between both states, but with more similarity to post-treatment NR-Br3s (**Figure 4.22D, E**). GO analysis of these 553 non-responder genes show statistical enrichment of regulation of type I interferon production, nucleocytoplasmic transport, RNA splicing, mRNA processing, ribonucleoprotein complex biogenesis, negative regulation of transcription, and others. The $\log_2$ fold change of these genes relative to the maximum expression within the opposing group across all NR-Br3s and RN3s are shown in figure 4.22G.

We repeated the analysis outlined in figure 6D using isoform expression across all branch 3 cells. We identified 2,284 NR-specific isoforms, 218 expressed in multiple NR-Br3 pre-treatment cells above the maximum expression within R-Br3 pre-treatment cells. 42 GO terms were statistically enriched (FDR <0.01) using ConsensusPathDB outlined in figure 4.22H. 10 genes had more than 2 isoforms enriched within NR-Br3 cells out of the

identified 218 isoforms (**Figure 4.22I**). Each of the previously described analysis resulted in recurrent genes identified as response-specific, including C-terminus binding protein 1(CTBP1). CTBP1 was the highest ranked gene by differential gene expression (**Figure 4.22G**), occurs within 11 GO categories identified by the differentially expressed isoform analysis (**Figure 4.22H**), and has 5 isoforms which are higher within NR-Br3 cells than the R-Br3 cohort (**Figure 4.22J**).

Candidate resistance genes include transcriptional repressor CTBP1. CTBP1 has been implicated as an oncogene in many other cancer types and is necessary for increased growth and abnormal differentiation of murine hematopoietic cells. Further, CTBP1 has been shown to be a master transcriptional repressor of tumor suppressors through direct binding of target genes, histone deacetylation, and direct inhibition of RNA polymerase II. Beyond its nuclear roles in transcription, CTBP1 has cytoplasmic roles dealing with membrane fission, pinocytosis, transportation, stabilization of pH levels through amomonia production and increasing glutamine supply. CTBP1 is enriched within non-responders compared to responders, with its expression increasing post-treatment non-responders, peaking within NR-Br3s. Further, CTBP1 is elevated within both non-responder patients compared to all other patients. Taken together, CTBP1 is a prime resistance candidate based off of our analysis and current literature.

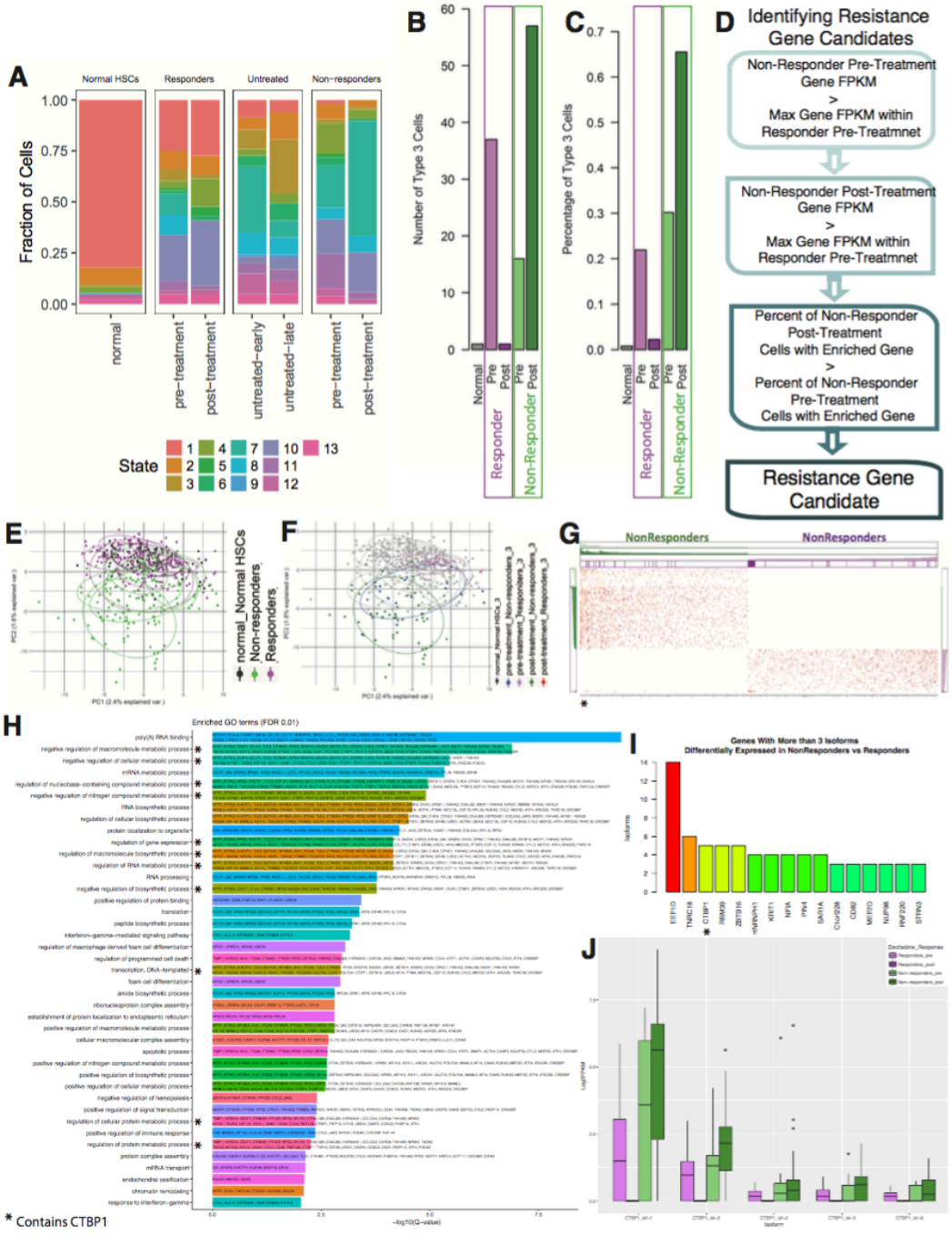


Figure 4.22. Identifying candidate resistance genes. A. Cell state distribution, separated by pre-treatment and post-treatment or early and late untreated. **B., C.** Branch 3 cell numbers. **D.** Methodology for identifying genes likely driving resistance. **E., F.** PCA based on identified genes separates response groups. **G.** Heatmap of identified genes. **H.** GO enrichment. **I.** Top candidates from isoform-centric analysis. **J.** Isoforms of CTBP1.

DISCUSSION

Our work provides a comprehensive view of the transcriptional landscape of human MDS stem cells at single cell resolution by identifying dysregulated expression and splicing patterns implicating specific pathways in MDS pathogenesis, disease causing mechanisms of mutations in splice factor genes, biomarkers of response to hypomethylating therapy, and therapy resistance mechanisms. Single cell resolution placed these findings in the context of distinct cellular transcriptional states and their dynamics during treatment. In order to investigate patterns of dysregulated splicing at the single cell level, we developed DISCO (Distributions of Isoforms in Single Cell Omics), a novel method enabling statistical comparisons of proportions of isoforms expressed that scales well to groups of hundreds of samples (i.e. cells). This approach enables characterizing shifts in splicing patterns, regardless of the overall expression of a gene (given a minimum number of reads for accurate quantification), and so is not confounded by up- or down-regulation of the gene (and all its isoforms), thus providing unique results from a differential gene expression analysis.

Integrated analysis of gene and isoform expression in stem cells from patients with known SRSF2 mutations, all of which were at the commonly mutated P95 site, compared to patients lacking any SRSF2 mutations, identified over-expression and altered splicing of genes involved in immune signaling and inflammation (IL2RA, IL2RG, CCL3, IFITM2, IFITM3, IFRD1).

Both the differential splicing and differential gene expression analyses reveal extensive cell-to-cell heterogeneity, suggesting that SRSF2 mutations do not uniformly affect all cells. A recent body of work has provided evidence for activated innate immunity playing a key role in MDS pathogenesis (Ga??n-G?mez et al., 2015; Keerthivasan et al., 2014; Starczynowski, 2014). Our data supports this and suggests that these pathways may be related to the molecular reasons underlying the high frequency of splice factor mutations observed in MDS and other hematological malignancies.

Transcriptional signals differentiating MDS stem cells from healthy controls, irrespective of SRSF2 status, were largely dominated by decreased expression of ribosomal protein genes (RPGs). Defective ribosome biogenesis through inactivating mutations and deletions of RPGs are known to cause MDS (specifically the 5q deletion subtype), among a variety of other hematological disorders known as ribosomopathies (Galili, Qasim, & Raza, 2009). One of the theories for the malignant mechanism of ribosomal haploinsufficiency is the activation of p53 signaling through binding of MDM2, the main p53 suppressor, by the excess uncomplexed ribosomal proteins that build up in the cell when ribosome biogenesis is inhibited (McGowan et al., 2008; Zhang et al., 2003). Other effects of RPG down-regulation are also likely since disrupted stoichiometries of the 80 core RPGs affect the heterogeneous compositions of ribosomes, which in turn modifies their preferences for translating specific mRNA transcripts, likely resulting in widespread

transcriptional dysregulation (Shi et al., 2017). Our data show that a large subset of RPGs are down-regulated and differentially spliced in MDS stem cells from patients with normal 5q karyotype. Two of these RPG isoforms are known to undergo nonsense-mediated decay suggesting that alternative splicing may lower RPG protein abundance even more than what can be measured with RNA-seq.

Comparison of decitabine non-responders to responders revealed even greater down-regulation of RPGs as well as a larger proportion of differentially spliced RPGs, suggesting that ribosome biogenesis plays a role in not only MDS pathogenesis but also response to decitabine, a mainstay in MDS and AML therapeutics. Concordant to this finding, p53 signaling genes are significantly over-expressed in MDS cells and even more in non-responders. Patients with 5q deletion syndrome, where the disease gene has been traced to RPS14, are generally treated with lenalidomide, as opposed to hypomethylating agents decitabine and azacitidine (List et al., 2005). Also, stoichiometric imbalances of ribosomal proteins have been linked to activating p53 signaling, potentially through binding of MDM2 by uncomplexed ribosomal proteins (McGowan et al., 2008; Zhang et al., 2003). Thus, defective ribosome biogenesis is not only implicated in MDS pathogenesis but also a larger set of hematological disorders broadly termed ribosomopathies (Narla & Ebert, 2010). Our findings suggest that a subset of patients with normal 5q karyotype

but significantly decreased RPG expression may also benefit from therapeutic alternatives to hypomethylating agents.

Single cell granularity further resolved these findings through the identification of distinct transcriptional stem cell states, and using matched pre- and post-treatment samples, we investigated the dynamics of these cell states during treatment. This revealed a striking expansion of a specific stem cell state post-treatment in non-responders. We examined the distinct features of this state using a custom pipeline (methods) designed to detect gene and isoform expression changes contributed by either the majority of cells or a small subset of cells that may drive resistance. RPGs were again identified in this analysis, providing confirmation of their involvement in the cells most likely to be responsible for driving resistance. A large set of other genes spanning a number of pathways involving metabolic processing, RNA processing, and transcriptional regulation were also identified by this analysis, illustrating the complex resistance mechanisms implored by cancer cells to escape decitabine mediated cell death.

Decitabine is a cytidine analog with a nitrogen in the 5-carbon position functions primarily as a hypomethylating agent by being incorporated into DNA during replication and inhibiting DNA methyltransferases. Though it also has a secondary genotoxic affects, cellular death from this primary mechanism may be evaded by either 1) not integrating decitabine into the DNA through either catabolizing it, packaging and removing it from the cell, or increasing de novo

pathway of nucleotide synthesis; or 2) if integrated into the DNA, affects may be evaded by any mechanism which does not allow the expression of the functional form of the gene – albeit at transcription, post-transcriptional modification, translational, or post-translational steps. The previously shown depression of RPGs may allow for decreased translation whereas differential splicing may lead to altered functions. CTBP1, identified as the top resistance candidate gene within non-responders, may confer resistance to azanucleotide treatments in many ways, including epigenomic regulation of histones. A recent study identified the small molecule NSC95397 as a CTBP1 inhibitor, with evidence of inducing cellular apoptosis (Blevins et al., 2015). Interestingly, this small molecule was also shown to inhibit spliceosomal activity – though it has not been shown if this is a result of CTBP1 inhibition or a different mechanism. Given the overwhelming evidence of differential isoform usage and splicing within non-responder vs responder cells, these pathways may be more connected than previously thought – with a potential role of CTBP1 as an important regulator.

In addition to insights on transcriptional heterogeneity in MDS, this work highlights important future avenues of research. These include: 1) the effects of RPG down-regulation in cell lines and mouse models on response to decitabine, and 2) preclinical studies to test the efficacy of combined therapy between azanucleotides and CTBP1 inhibitors. More evidence on the impact of RPG down-regulation on desensitizing patients to azanucleotide therapy

could readily be implemented in clinic by choosing alternate therapeutic strategies for patients with RPG deletions. After more testing on its mechanism of efficacy, combination therapy with CTBP1 inhibitors may be an important addition to the therapeutic options available in MDS.

APPENDIX

MATERIALS AND METHODS – MCL EXOME AND TRANSCRIPTOME SEQUENCING (CHAPTER 3)

Sample collection. Lymph node biopsies were collected from patients in a phase I trial of palbociclib and bortezomib for recurrent MCL (Di Liberto et al., n.d.). Primary MCL cells were purified using MACS CD19 MicroBeads (Miltenyi Biotec) with > 90% yield of tumor cells (CD19+, CD5+) as assessed by flow cytometry. Matched normal controls for exome seq was available for some patients with buccal swabs. Peripheral blood B cells (PBCs) from 3 healthy volunteers isolated using the same protocol served as the normal controls for RNA seq.

Sequencing. DNA was isolated from purified MCL cells, and a sequencing library was created with the Illumina TruSeq (v3) DNA Preparation kits (FC-121-1031). Following isothermal cluster generation (PE-401-3001) and 75x75 paired-end (PE) sequencing on the HiSeq2500 (FC-401-3001), the samples underwent primary analysis with the Illumina base calling and primary analysis software (HCS 1.4, CASAVA 1.8.2, and RTA 1.2).

Immunohistochemistry. Immunoperoxidase staining was performed on a Leica Bond III automated immunostainer, using antibodies for PAX5 and BCL6 supplied by the manufacturer and according to the manufacturers instructions (Leica Microsystems, Bannockburn, IL) as previously described.[1] Staining was performed with PAX5 and a red chromogen (no counterstain), with BCL6 and a brown chromogen (hematoxylin counterstain), as well as multiplex staining for PAX5 (red chromogen)/BCL6 (blue chromogen), where dual staining resulted in a purple signal. Image analysis with signal quantitation of the above detailed IHC slides was performed on the Ariol 50 (Leica Microsystems, Bannockburn, IL) image analysis system, according to the manufacturers specifications, as previously described (Chiron et al., 2014).

DNA and RNA Analysis. WES reads were aligned to hg19 reference and the quality assessed using our in-house pipeline *g-make*, which uses BWA for alignment (H. Li & Durbin, 2009), picard for duplicate removal (http://picard.sourceforge.net), and GATK for single nucleotide and indel variant calling (McKenna et al., 2010). Copy number variants were identified using XHMM, VarScan somatic copycaller with circular binary segmentation with R DNAcopy, and SNP allele frequencies from GATK results. As VarScan somatic copycaller relies on a matched normal control to identify regions of significant gain or loss of read dept, for the patient samples lacking these, we used average read depth across the normal controls of the other patients, in

effect assuming that the normal cells lack large-scale (>1 Mb) CNVs. PyClone v1.0 was used to infer subclonal clusters based on variant allele frequencies of SNVs and copy number (Roth et al., 2014). WTS reads were processed through Genesifter software (http://www.geospiza.com/Products/WTA.shtml) with limma-voom in R for differential gene expression analysis (Law, Chen, Shi, & Smyth, 2014). Integrative DNA and RNA analysis was performed using custom R and python scripts.

MATERIALS AND METHODS – MDS SINGLE CELL RNA SEQUENCING (CHAPTER 4)

Sample Collection. HSCs were purified from viably frozen patient bone marrow biopsies for scRNA-seq using FACS markers Lin-CD34+CD38-CD90+CD45RA-, following the protocol detailed by Pang *et al.* (Pang et al., 2013).

Single Cell RNA-seq. Single cells were captured and mRNA isolated using Clontech's SMARTer chemistry for ultra low input on the Fluidigm C1 Single Cell Auto Prep system as per manufacturer's protocol. Illumina's Nextera XT kit was used for library preparation prior to 100bp paired-end sequencing on the HiSeq2500 platform. Q-values for all bases were >30 and image analysis was performed with Illumina's CASAVA pipeline.

Alignment, quality control, and gene quantification. Raw sequence reads were analyzed through the r-make pipeline for quality control and alignment with STAR to the hg19 human reference genome (Dobin et al., 2013; S. Li, Tighe, et al., 2014). Quality metrics assessed include mapping rates, distribution of reads of across different regions (exons, introns, mitochondrial, intergenic, and ribosomal), coverage across gene body to detect extent of 3' bias in poly-A RNA preps, number of genes detected as a function of read depth, GC content, and strand distribution. Gene counts were calculated using HTseq and the RefSeq gene annotation, and FPKM (fragments per kilobase per million) measurements were used for normalized transcript abundance quantification (Anders, Pyl, & Huber, 2014). Prior to analyzing and comparing transcriptomes, the following preprocessing steps were taken within each sample: (1) filter out samples from wells known to have captured more than 1 cell based on microscopic examination at the time of cell capture, (2) filter out samples with more than 20% reads mapping to mitochondrial regions, which can indicate dying cells, (3) filter out samples with less than 50,000 reads mapped to genes, (4) filter out genes expressed in less than 3 cells at FPKM greater than 1, and (5) filter out samples with extremely low expression of GAPDH and ACTB (less than 2 standard deviations below the mean). These steps enabled downstream analysis to be performed on high quality data.

Alternative splicing analysis. We have developed a method enabling single cell isoform analysis called DISCO (Distributions of Isoforms in Single Cell Omics). The DISCO pipeline consists of splice-aware alignment with STAR (Dobin et al., 2013), isoform quantification relative to all isoforms of a gene using MISO (Katz et al., 2010), and non-parametric statistical testing, multiple testing correction, and visualization of significant results with DISCO. Depending on the reference database used with MISO, DISCO can also be applied to analyzing skipped exons, mutually exclusive exons, retained introns, and any other set of alternative RNA processing events that can be defined in an annotation file. DISCO is not limited to being run with MISO output; it can be used downstream of any method that provides relative quantifications, i.e. PSI (percent spliced in) for exons and PI (percent isoform) for isoforms. DISCO is publicly available along with installation and usage instructions and example runs here: https://pbtech-vc.med.cornell.edu/git/mason-lab/disco/tree/master.

We used DISCO with MISO quantifications for full-length isoforms defined by the ENSEMBL annotation for comparing MDS and normal, responders and non-responders, and patients with and without SRSF2 mutations. For the last case, we also analyzed skipped exons and tested for enrichment of exonic splice enhancers by calculating 4-, 5-, and 6-mer distributions in exons identified by DISCO and significantly alternatively spliced and using fisher's exact tests to compare the proportion of each k-mer in

exons more highly included in SRSF2 mutated against a background proportion calculated across a thousand random non-significantly alternatively spliced exons of the genome. The same was also calculated for exons more highly excluded in SRSF2 mutated cells, thus enabling a measure of which k-mers are enriched and may function as enhancer units. This is similar to the approach used by Kim *et al.* (E. Kim et al., 2015).

Gene expression heterogeneity and lineage ordering. We characterized gene expression heterogeneity by integrating several complementary approaches: (1) t-SNE (t-distributed stochastic neighbor embedding) for visualizing relationships between cells in a reduced dimension space, (2) graph-based ordering and clustering of cells (3) differential expression analysis to identify genes driving cell state differences (Trapnell et al., 2014; Van Der Maaten & Hinton, 2008).

The t-SNE dimensionality reduction algorithm was applied to $\log_2$ FPKM values across all single cells after filtering outlier cells and lowly expressed genes (as described above). We used the Rtsne pacakage's Barnes-Hut implementation with a perplexity of 10, maximum number of iterations of 10,000, theta of 0.1, and random seed of 42 (https://github.com/jkrijthe/Rtsne).

Differential expression analysis between MDS stem cells and normal HSC controls was performed using the monocle R package (Trapnell et al., 2014). Input cells were restricted to the pre-treatment and first serial untreated time

points for the MDS group and all cells for the normal group. Genes expressed at a minimum FPKM of 1 in less than 10 cells were excluded, the FPKM matrix was transformed to absolute RNA counts using the “relative2abs” monocle function, and the resulting counts were modeled using a negative binomial distribution for differential expression testing. Benjamini-hochberg multiple testing correction was used with a significance threshold of 0.01.

As we observed hematopoietic stem cell regulators and myeloid lineage markers among the genes differentially expressed between MDS and normal, we further explored this by using genes co-expressed with these stem cell and myeloid markers to seed a semi-supervised pseudotemporal lineage ordering analysis. These genes were identified by assigning cells to a stem cell or myeloid state and performing a differential expression test between the two groups. Cells were categorized as “stem cell” if the sum expression across stem cell marker genes was more than an FPKM of 10, “myeloid” if the sum expression across myeloid marker genes was more than an FPKM of 0.1, “both” if both are true, and “neither” if neither are true. Genes differentially expressed between the “stem cell” and “myeloid” categories were then identified similar to the MDS vs. normal test described above. Cells are then ordered based on expression of these genes using monocle’s orderCells function, which uses the DDRtree package’s reversed graph embedding framework (Qiu et al., 2017). The resulting tree structure separates cells into

three main branches and 13 distinct clusters, and genes responsible for the clustering were identified through differential expression tests.

Identifying candidate resistance markers. To discover resistance markers, we examined the genes distinct to each psuedotime branch. Branch 3 genes are largely depleted in responders post-treatment but show increases within non-responders post-treatment. We used the following criteria for identifying resistant candidate cells and genes: A non-responder pre-treatment branch 3 cell must have a higher expression of a given gene than the maximum expression of that gene within responder pre-treatment branch 3 cells. Further, the percentage of non-responder post-treatment branch 3 cells which have this gene enriched must be higher than the percentage of non-responder pre-treatment branch 3 cells.

Pathway Enrichment Analysis. Pathway and gene ontology enrichment analyses of differentially expressed and differentially spliced genes were performed using a combination of EnrichR (Kuleshov et al., 2016), ConsensusPathDB, GOrilla, and GSEA. Correction for multiple testing and q-values are listed with each analysis in the main text and supplemental tables.

REFERENCES

Alderton, G. K. (2015). Haematological malignancies: Splicing the MDS genome. *Nature Reviews. Cancer*, *15*(7), 393. http://doi.org/10.1038/nrc3975

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166–169. http://doi.org/10.1093/bioinformatics/btu638

Ashton, P. M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., … O'Grady, J. (2014). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology*, *33*(3), 296–300. http://doi.org/10.1038/nbt.3103

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., … Liu, J. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202–209. Retrieved from http://dx.doi.org/10.1038/nature13480

Basso, K., & Dalla-Favera, R. (2010). BCL6: master regulator of the germinal center reaction and key oncogene in B cell lymphomagenesis. *Advances in Immunology*, *105*, 193–210. http://doi.org/10.1016/S0065-2776(10)05007-8

Baughn, L. B., Di Liberto, M., Wu, K., Toogood, P. L., Louie, T., Gottschalk, R., … Chen-Kiang, S. (2006). A novel orally active small molecule potently induces G1 arrest in primary myeloma cells and prevents tumor growth by specific inhibition of cyclin-dependent kinase 4/6. *Cancer Res*, *66*(15), 7661–7667. http://doi.org/10.1158/0008-5472.CAN-06-1098

Beà, S., Valdés-Mas, R., Navarro, A., Salaverria, I., Martín-Garcia, D., Jares, P., … Campo, E. (2013). Landscape of somatic mutations and clonal evolution in mantle cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18250–5. http://doi.org/10.1073/pnas.1314608110

Beer, P. A., Delhommeau, F., LeCouédic, J.-P., Dawson, M. A., Chen, E., Bareford, D., … Green, A. R. (2010). Two routes to leukemic transformation after a JAK2 mutation–positive myeloproliferative

neoplasm. *Blood*, *115*(14). Retrieved from http://www.bloodjournal.org/content/115/14/2891?sso-checked=true

Bejar, R., & Steensma, D. P. (2014). Recent developments in myelodysplastic syndromes. *Blood*, *124*(18).

Bejar, R., Stevenson, K. E., Caughey, B. A., Abdel-Wahab, O., Steensma, D. P., Galili, N., … Ebert, B. L. (2012). Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *30*(27), 3376–82. http://doi.org/10.1200/JCO.2011.40.7379

Blevins, M. A., Kouznetsova, J., Krueger, A. B., King, R., Griner, L. M., Hu, X., … Zhao, R. (2015). Small Molecule, NSC95397, Inhibits the CtBP1-Protein Partner Interaction and CtBP1-Mediated Transcriptional Repression. *Journal of Biomolecular Screening*, *20*(5), 663–672. http://doi.org/10.1177/1087057114561400

Boland, M. R., Tatonetti, N. P., & Hripcsak, G. (2015). Development and validation of a classification approach for extracting severity automatically from electronic health records. *Journal of Biomedical Semantics*, *6*, 14. http://doi.org/10.1186/s13326-015-0010-8

Bolisetty, M. T., Rajadinakaran, G., & Graveley, B. R. (2015). Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, *16*(1), 204. http://doi.org/10.1186/s13059-015-0777-z

Botkin, J. R., Belmont, J. W., Berg, J. S., Berkman, B. E., Bombard, Y., Holm, I. A., … McInerney, J. D. (2015). Points to Consider: Ethical, Legal, and Psychosocial Implications of Genetic Testing in Children and Adolescents. *The American Journal of Human Genetics*, *97*(1), 6–21. http://doi.org/10.1016/j.ajhg.2015.05.022

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. a, Zhang, X., Proserpio, V., … Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–5. http://doi.org/10.1038/nmeth.2645

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., … Jaenisch, R. (2012). Single-cell expression analyses during cellular

reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, *150*(6), 1209–22. http://doi.org/10.1016/j.cell.2012.08.023

Camacho, F. I., García, J. F., Cigudosa, J. C., Mollejo, M., Algara, P., Ruíz-Ballesteros, E., … Piris, M. A. (2004). Aberrant Bcl6 protein expression in mantle cell lymphoma. *The American Journal of Surgical Pathology*, *28*(8), 1051–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15252312

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., … Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–4. http://doi.org/10.1158/2159-8290.CD-12-0095

Chaisson, M. J., & Tesler, G. (2012). Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics*, *13*, 238. http://doi.org/10.1186/1471-2105-13-238

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., … McArthur, G. A. (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England Journal of Medicine*, *364*(26), 2507–16. http://doi.org/10.1056/NEJMoa1103782

Chhangawala, S., Rudy, G., Mason, C. E., & Rosenfeld, J. A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. *Genome Biology*, *16*, 131. http://doi.org/10.1186/s13059-015-0697-y

Chiron, D., Di Liberto, M., Martin, P., Huang, X., Sharman, J., Blecua, P., … Chen-Kiang, S. (2014). Cell-cycle reprogramming for PI3K inhibition overrides a relapse-specific C481S BTK mutation revealed by longitudinal functional genomics in mantle cell lymphoma. *Cancer Discovery*, *4*(9), 1022–35. http://doi.org/10.1158/2159-8290.CD-14-0098

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., & Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, *45*(10), 1127–1133. http://doi.org/10.1038/ng.2762

Clark, C., Palta, P., Joyce, C. J., Scott, C., Grundberg, E., Deloukas, P., …

Coffey, A. J. (2012). A comparison of the whole genome approach of MeDIP-seq to the targeted approach of the Infinium HumanMethylation450 BeadChip(®) for methylome profiling. *PloS One*, *7*(11), e50233. http://doi.org/10.1371/journal.pone.0050233

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., … Snyder, M. (2011a). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–14. http://doi.org/10.1038/nbt.1975

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., … Snyder, M. (2011b). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–14. Retrieved from http://dx.doi.org/10.1038/nbt.1975

Colomer, D., & Campo, E. (2014). Unlocking new therapeutic targets and resistance mechanisms in mantle cell lymphoma. *Cancer Cell*, *25*(1), 7–9. http://doi.org/10.1016/j.ccr.2013.12.011

Corbin, A. S., Agarwal, A., Loriaux, M., Cortes, J., Deininger, M. W., & Druker, B. J. (2011). Human chronic myeloid leukemia stem cells are insensitive to imatinib despite inhibition of BCR-ABL activity. *The Journal of Clinical Investigation*, *121*, 396–409. http://doi.org/10.1172/JCI35721

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*. http://doi.org/10.1038/nbt.2877

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. a, … Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, *29*(12), 1120–7. http://doi.org/10.1038/nbt.2038

Darnell, J. E., Kerr, I. M., & Stark, G. R. (1994). Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science (New York, N.Y.)*, *264*(5164), 1415–21. http://doi.org/10.1126/science.8197455

Deininger, M. W. N., Goldman, J. M., & Melo, J. V. (2000). The molecular biology of chronic myeloid leukemia. *Blood*, *96*, 3343–3356.

Delfau-Larue, M.-H., Klapper, W., Berger, F., Jardin, F., Briere, J., Salles, G., … Hoster, E. (2015). CDKN2A and TP53 deletions predict adverse outcome in younger mantle cell lymphoma patients, independent of treatment and MIPI. *Blood*.

Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E. M., Antosiewicz-Bourget, J., … Zhang, K. (2009). Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology*, *27*(4), 353–60. Retrieved from http://dx.doi.org/10.1038/nbt.1530

Di Liberto, M., Martin, P., Chiron, D., Vijay, P., Huang, X., Blecua, P., … Chen-Kiang, S. (n.d.). Longitudinal integrative whole transcriptome and exome sequencing identifies genes that reprogram lymphoma cells for clinical response to CDK4/6 inhibition in combination therapy.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., … DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, *481*(7382), 506–10. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3267864&tool=pmcentrez&rendertype=abstract

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635

Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., … Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, *26*(7), 779–85. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644410&tool=pmcentrez&rendertype=abstract

Drekonja, D., Reich, J., Gezahegn, S., Greer, N., Shaukat, A., MacDonald, R., … Wilt, T. J. (2015). Fecal Microbiota Transplantation for Clostridium difficile Infection: A Systematic Review. *Annals of Internal Medicine*, *162*(9), 630–8. http://doi.org/10.7326/M14-2693

Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H.,

Gattermann, N., … Larson, R. A. (2006). Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *The New England Journal of Medicine*, *355*, 2408–2417. http://doi.org/10.1056/NEJMoa062867

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. *Science (New York, N.Y.)*, *231*(4742), 1055–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3945817

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*, 48. http://doi.org/10.1186/1471-2105-10-48

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L. G. T., … Chan, T. A. (2011). Breast cancer methylomes establish an epigenomic foundation for metastasis. *Science Translational Medicine*, *3*(75), 75ra25. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146366&tool=pmcentrez&rendertype=abstract

Figueroa, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., … Melnick, A. (2010). DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*, *17*(1), 13–27. Retrieved from http://www.sciencedirect.com/science/article/pii/S1535610809004206

Finn, R. S., Crown, J. P., Lang, I., Boer, K., Bondarenko, I. M., Kulyk, S. O., … Slamon, D. J. (2014). The cyclin-dependent kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer (PALOMA-1/TRIO-18): a randomised phase 2 study. *The Lancet Oncology*, *16*(1), 25–35. http://doi.org/10.1016/S1470-2045(14)71159-3

Fromer, M., & Purcell, S. M. (2014). Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Current Protocols in*

Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.], *81*, 7.23.1-7.23.21. http://doi.org/10.1002/0471142905.hg0723s81

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., … Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(5), 1827–31. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=48546&tool=pmcentrez&rendertype=abstract

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., … Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews. Cancer*, *4*(3), 177–83. http://doi.org/10.1038/nrc1299

Ga??n-G?mez, I., Wei, Y., Starczynowski, D. T., Colla, S., Yang, H., Cabrero-Calvo, M., … Garcia-Manero, G. (2015). Deregulation of innate immune and inflammatory signaling in myelodysplastic syndromes. *Leukemia*, *29*(7), 1458–1469. http://doi.org/10.1038/leu.2015.69

Galili, N., Qasim, S. A., & Raza, A. (2009). Defective ribosome biogenesis in myelodysplastic syndromes. *Haematologica*, *94*(10), 1336–8. http://doi.org/10.3324/haematol.2009.012021

Gardy, J., Loman, N. J., & Rambaut, A. (2015). Real-time digital pathogen surveillance — the time is now. *Genome Biology*, *16*(1), 155. http://doi.org/10.1186/s13059-015-0726-x

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. http://doi.org/10.1038/nrg.2015.16

goel, shom, DeCristo, M. J., Watt, april C., BrinJones, H., sceneay, J., Li, B. B., … Zhao, J. J. (2017). CDK4/6 inhibition triggers anti-tumour immunity. *Nature Publishing Group*, *548*. http://doi.org/10.1038/nature23465

Grün, D., Kester, L., & van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods*, *11*(6), 637–40. http://doi.org/10.1038/nmeth.2930

Gualco, G., Weiss, L. M., Harrington, W. J., & Bacchi, C. E. (2010). BCL6, MUM1, and CD10 expression in mantle cell lymphoma. *Applied*

Immunohistochemistry & Molecular Morphology : AIMM / Official Publication of the Society for Applied Immunohistochemistry, *18*(2), 103–8. http://doi.org/10.1097/PAI.0b013e3181bb9edf

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., … Ogawa, S. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia*, *28*(2), 241–247. http://doi.org/10.1038/leu.2013.336

Haghverdi, L., Buettner, F., & Theis, F. J. (2014). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, *31*(18). http://doi.org/10.1093/bioinformatics/btv325

Han, G., Sun, J., Wang, J., Bai, Z., Song, F., & Lei, H. (2014). Genomics in neurological disorders. *Genomics, Proteomics & Bioinformatics*, *12*(4), 156–63. http://doi.org/10.1016/j.gpb.2014.07.002

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*. http://doi.org/10.1016/j.cell.2011.02.013

Hao, S., Sanger, W., Onciu, M., Lai, R., Schlette, E. J., & Medeiros, L. J. (2002). Mantle cell lymphoma with 8q24 chromosomal abnormalities: a report of 5 cases with blastoid features. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *15*(12), 1266–72. http://doi.org/10.1097/01.MP.0000037310.82136.99

Hartmann, E. M., Campo, E., Wright, G., Lenz, G., Salaverria, I., Jares, P., … Rosenwald, A. (2010). Pathway discovery in mantle cell lymphoma by integrated analysis of high-resolution gene expression and copy number profiling. *Blood*, *116*(6), 953–61. http://doi.org/10.1182/blood-2010-01-263806

Hou, Y., Fan, W., Yan, L., Li, R., Lian, Y., Huang, J., … Qiao, J. (2013). Genome analyses of single human oocytes. *Cell*, *155*(7), 1492–506. http://doi.org/10.1016/j.cell.2013.11.040

Huang, X., Di Liberto, M., Jayabalan, D., Liang, J., Ely, S., Bretz, J., … Chen-Kiang, S. (2012). Prolonged early G(1) arrest by selective CDK4/CDK6 inhibition sensitizes myeloma cells to cytotoxic killing through cell cycle-coupled loss of IRF4. *Blood*, *120*(5), 1095–106. http://doi.org/10.1182/blood-2012-03-415984

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., … Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–8. http://doi.org/10.1038/nature08987

Hughes, A. E. O., Magrini, V., Demeter, R., Miller, C. a., Fulton, R., Fulton, L. L., … Graubert, T. a. (2014). Clonal Architecture of Secondary Acute Myeloid Leukemia Defined by Single-Cell Sequencing. *PLoS Genetics*, *10*(7). http://doi.org/10.1371/journal.pgen.1004462

Hungerford, D. A., & Nowell, P. C. (1960). A minute chromosome in human chronic granulocytic leukemia. *Science*, *132*, 1457–1501. http://doi.org/10.1126/science.132.3438.1488

Itchikawa, K., & Baum, S. M. (1925). The Rapid Production of Cancer in Rabbits by Coal-Tar. *The Journal of Cancer Research*, *9*(1). Retrieved from http://cancerres.aacrjournals.org/content/9/1/85

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, *12*(4), 351–356. http://doi.org/10.1038/nmeth.3290

Ji, Z., & Ji, H. (2016). TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, *44*(13). http://doi.org/10.1093/nar/gkw430

Karczewski, K. J., Fernald, G. H., Martin, A. R., Snyder, M., Tatonetti, N. P., & Dudley, J. T. (2014). STORMSeq: An Open-Source, User-Friendly Pipeline for Processing Personal Genomics Data in the Cloud. *PLoS ONE*, *9*(1), e84860. http://doi.org/10.1371/journal.pone.0084860

Karube, K., Ying, G., Tagawa, H., Niino, D., Aoki, R., Kimura, Y., … Ohshima, K. (2008). BCL6 gene amplification/3q27 gain is associated with unique clinicopathological characteristics among follicular lymphoma without BCL2 gene translocation. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *21*(8), 973–8. http://doi.org/10.1038/modpathol.2008.75

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–15. http://doi.org/10.1038/nmeth.1528

Keerthivasan, G., Mei, Y., Zhao, B., Zhang, L., Harris, C. E., Gao, J., … Ji, P.

(2014). Aberrant overexpression of CD14 on granulocytes sensitizes the innate immune response in mDia1 heterozygous del(5q) MDS. *Blood*, *124*(5). Retrieved from http://www.bloodjournal.org/content/124/5/780?ijkey=ed16e8e50216a643fb538a7e8d11fc589b092321&keytype2=tf_ipsecsha

Kim, E., Ilagan, J. O., Bradley, R. K., & Abdel-Wahab, O. (2015). SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell*, *27*, 617–630. http://doi.org/10.1016/j.ccell.2015.04.006

Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Illicic, T., Teichmann, S. A., & Marioni, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nature Communications*, *6*, 8687. http://doi.org/10.1038/ncomms9687

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., … Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. http://doi.org/10.1101/gr.129684.111

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., … Ma'ayan, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, *44*(W1), W90–W97. http://doi.org/10.1093/nar/gkw377

Kumar-Sinha, C., Varambally, S., Sreekumar, A., & Chinnaiyan, A. M. (2002). Molecular cross-talk between the TRAIL and interferon signaling pathways. *The Journal of Biological Chemistry*, *277*(1), 575–85. http://doi.org/10.1074/jbc.M107795200

Lam, H. Y. K., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., … Snyder, M. (2012). Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*. http://doi.org/10.1038/nbt.2065

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science (New York, N.Y.)*, *313*, 1929–1935.

http://doi.org/10.1126/science.1132939

Landau, D. a, Carter, S. L., Getz, G., & Wu, C. J. (2014). Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia*, *28*(1), 34–43. http://doi.org/10.1038/leu.2013.248

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. http://doi.org/10.1186/gb-2014-15-2-r29

Leonard, J. P., LaCasce, A. S., Smith, M. R., Noy, A., Chirieac, L. R., Rodig, S. J., … Shapiro, G. I. (2012). Selective CDK4/6 inhibition with tumor responses by PD0332991 in patients with mantle cell lymphoma. *Blood*, *119*(20), 4597–607. http://doi.org/10.1182/blood-2011-10-388298

Lesinski, G. B., Raig, E. T., Guenterberg, K., Brown, L., Go, M. R., Shah, N. N., … Carson, W. E. (2008). IFN-alpha and bortezomib overcome Bcl-2 and Mcl-1 overexpression in melanoma cells by stimulating the extrinsic pathway of apoptosis. *Cancer Research*, *68*(20), 8351–60. http://doi.org/10.1158/0008-5472.CAN-08-0426

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

Li, L., Ruau, D. J., Patel, C. J., Weber, S. C., Chen, R., Tatonetti, N. P., … Butte, A. J. (2014). Disease risk factors identified through shared genetic architecture and electronic medical records. *Science Translational Medicine*, *6*(234), 234ra57. http://doi.org/10.1126/scitranslmed.3007191

Li, S., Łabaj, P. P., Zumbo, P., Sykacek, P., Shi, W., Shi, L., … Mason, C. E. (2014). Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology*, *32*(9), 888–95. http://doi.org/10.1038/nbt.3000

Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., … Mason, C. E. (2014). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nature Biotechnology*, *32*(9), 915–25. http://doi.org/10.1038/nbt.2972

List, A., Kurtin, S., Roe, D. J., Buresh, A., Mahadevan, D., Fuchs, D., … Zeldis, J. B. (2005). Efficacy of lenalidomide in myelodysplastic

syndromes. *The New England Journal of Medicine*, *352*(6), 549–57. http://doi.org/10.1056/NEJMoa041668

Loman, N. J., Misra, R. V, Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*. http://doi.org/10.1038/nbt.2198

Lowe, R., & Rakyan, V. K. (2013). Marmal-aid--a database for Infinium HumanMethylation450. *BMC Bioinformatics*, *14*, 359. http://doi.org/10.1186/1471-2105-14-359

Mack, S. C., Witt, H., Piro, R. M., Gu, L., Zuyderduyn, S., Stütz, A. M., … Taylor, M. D. (2014). Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. *Nature*, *506*(7489), 445–50. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4174313&tool=pmcentrez&rendertype=abstract

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–14. http://doi.org/10.1016/j.cell.2015.05.002

Mandal, R. S., Saha, S., & Das, S. (2015). Metagenomic Surveys of Gut Microbiota. *Genomics, Proteomics & Bioinformatics*. http://doi.org/10.1016/j.gpb.2015.02.005

Manolio, T. A., Abramowicz, M., Al-Mulla, F., Anderson, W., Balling, R., Berger, A. C., … Ginsburg, G. S. (2015). Global implementation of genomic medicine: We are not alone. *Science Translational Medicine*, *7*(290), 290ps13-290ps13. http://doi.org/10.1126/scitranslmed.aab0194

Markowitz, J., Luedke, E. A., Grignol, V. P., Hade, E. M., Paul, B. K., Mundy-Bosse, B. L., … Carson, W. E. (2014). A phase I trial of bortezomib and interferon-α-2b in metastatic melanoma. *Journal of Immunotherapy (Hagerstown, Md.: 1997)*, *37*(1), 55–62. http://doi.org/10.1097/CJI.0000000000000009

Marsolo, K., & Spooner, S. A. (2013). Clinical genomics in the world of the electronic health record. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *15*(10), 786–91.

http://doi.org/10.1038/gim.2013.88

Marzec, M., Kasprzycka, M., Lai, R., Gladden, A. B., Wlodarski, P., Tomczak, E., … Wasik, M. A. (2006). Mantle cell lymphoma cells express predominantly cyclin D1a isoform and are highly sensitive to selective inhibition of CDK4 kinase activity. *Blood*, *108*(5), 1744–50. http://doi.org/10.1182/blood-2006-04-016634

Mason, C. E., Porter, S. G., & Smith, T. M. (2014). Characterizing Multi-omic Data in Systems Biology. *Adv Exp Med Biol*, *799*, 15–38. http://doi.org/10.1007/978-1-4614-8778-4_2

McGowan, K. A., Li, J. Z., Park, C. Y., Beaudry, V., Tabor, H. K., Sabnis, A. J., … Barsh, G. S. (2008). Ribosomal mutations cause p53-mediated dark skin and pleiotropic effects. *Nature Genetics*, *40*(8), 963–70. http://doi.org/10.1038/ng.188

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. http://doi.org/10.1101/gr.107524.110

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., M. Mastrogianakis, G., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. http://doi.org/10.1038/nature07385

Meissner, A. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, *33*(18), 5868–5877. Retrieved from http://nar.oxfordjournals.org/content/33/18/5868.full

Meissner, B., Kridel, R., Lim, R. S., Rogic, S., Tse, K., Scott, D. W., … Gascoyne, R. D. (2013). The E3 ubiquitin ligase UBR5 is recurrently mutated in mantle cell lymphoma. *Blood*, *121*, 3161–3164. http://doi.org/10.1182/blood-2013-01-478834

Mielczarek, M., & Szyda, J. (2015). Review of alignment and SNP calling algorithms for next-generation sequencing data. *Journal of Applied Genetics*. http://doi.org/10.1007/s13353-015-0292-7

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., … Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(11), 4245–50. http://doi.org/10.1073/pnas.1208949110

Narla, A., & Ebert, B. L. (2010). Ribosomopathies: human disorders of ribosome dysfunction. *Blood*, *115*(16), 3196–205. http://doi.org/10.1182/blood-2009-10-178129

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., … Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*, 90–94. http://doi.org/10.1038/nature09807

Niewerth, D., Kaspers, G. J. L., Assaraf, Y. G., van Meerloo, J., Kirk, C. J., Anderl, J., … Cloos, J. (2014). Interferon-γ-induced upregulation of immunoproteasome subunit assembly overcomes bortezomib resistance in human hematological cell lines. *Journal of Hematology & Oncology*, *7*(1), 7. http://doi.org/10.1186/1756-8722-7-7

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., … Aldape, K. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, *17*(5), 510–22. http://doi.org/10.1016/j.ccr.2010.03.017

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science (New York, N.Y.)*, *194*, 23–28. http://doi.org/10.1126/science.959840

Oberley, M. J., Rajguru, S. A., Zhang, C., Kim, K., Shaw, G. R., Grindle, K. M., … Yang, D. T. (2013). Immunohistochemical evaluation of MYC expression in mantle cell lymphoma. *Histopathology*, *63*(4), 499–508. http://doi.org/10.1111/his.12207

Pajtler, K. W., Witt, H., Sill, M., Jones, D. T. W., Hovestadt, V., Kratochwil, F., … Pfister, S. M. (2015). Molecular Classification of Ependymal Tumors across All CNS Compartments, Histopathological Grades, and Age Groups. *Cancer Cell*, *27*(5), 728–743. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25965575

Pan, H., Chen, L., Dogra, S., Teh, A. L., Tan, J. H., Lim, Y. I., … Holbrook, J. D. (2012). Measuring the methylome in clinical samples: improved

processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics : Official Journal of the DNA Methylation Society*, *7*(10), 1173–87. http://doi.org/10.4161/epi.22102

Pang, W. W., Pluvinage, J. V, Price, E. a, Sridhar, K., Arber, D. a, Greenberg, P. L., … Weissman, I. L. (2013). Hematopoietic stem cell and progenitor cell mechanisms in myelodysplastic syndromes. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 3011–6. http://doi.org/10.1073/pnas.1222861110

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., … Campbell, P. J. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood*, *122*(22), 3616–27; quiz 3699. http://doi.org/10.1182/blood-2013-08-518886

Papageorgiou, A., Kamat, A., Benedict, W. F., Dinney, C., & McConkey, D. J. (2006). Combination therapy with IFN-alpha plus bortezomib induces apoptosis and inhibits angiogenesis in human bladder cancer cells. *Molecular Cancer Therapeutics*, *5*(12), 3032–41. http://doi.org/10.1158/1535-7163.MCT-05-0474

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., … Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, (June), 1–9. http://doi.org/10.1126/science.1254257

Patel, C. J., Sivadas, A., Tabassum, R., Preeprem, T., Zhao, J., Arafat, D., … Gibson, G. (2013). Whole genome sequencing in support of wellness and health maintenance. *Genome Medicine*, *5*(6), 58. http://doi.org/10.1186/gm462

Patel, J. P., Gönen, M., Figueroa, M. E., Fernandez, H., Sun, Z., Racevskis, J., … Levine, R. L. (2012). Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *The New England Journal of Medicine*, *366*(12), 1079–89. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3545649&tool=pmcentrez&rendertype=abstract

Perez-Diez, A., Morgun, A., & Shulzhenko, N. (2000). Microarrays for Cancer Diagnosis and Classification. Landes Bioscience. Retrieved from

http://www.ncbi.nlm.nih.gov/books/NBK6624/

Pérez-Galán, P., Dreyling, M., & Wiestner, A. (2011). Mantle cell lymphoma: biology, pathogenesis, and the molecular basis of treatment in the genomic era. *Blood*, *117*(1), 26–38. http://doi.org/10.1182/blood-2010-04-189977

Phan, R. T., & Dalla-Favera, R. (2004). The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature*, *432*(7017), 635–639. http://doi.org/10.1038/nature03147

Pollen, A. a, Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. a, Lui, J. H., … West, J. a a. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, *32*(10). http://doi.org/10.1038/nbt.2967

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., & Trapnell, C. (2017). Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, *14*(3), 309–315. http://doi.org/10.1038/nmeth.4150

Rahal, R., Frick, M., Romero, R., Korn, J. M., Kridel, R., Chan, F. C., … Stegmeier, F. (2014). Pharmacological and genomic profiling identifies NF-κB-targeted treatment strategies for mantle cell lymphoma. *Nature Medicine*, *20*, 87–92. http://doi.org/10.1038/nm.3435

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., … Getz, G. (2015). Oncotator: cancer variant annotation tool. *Human Mutation*, *36*(4), E2423-9. http://doi.org/10.1002/humu.22771

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., … Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, *14*(9), R95. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4054597&tool=pmcentrez&rendertype=abstract

Raza, A., & Galili, N. (2012). The genetic basis of phenotypic heterogeneity in myelodysplastic syndromes. *Nature Reviews. Cancer*, *12*(12), 849–59. http://doi.org/10.1038/nrc3321

Retterer, K., Scuffins, J., Schmidt, D., Lewis, R., Pineda-Alvarez, D., Stafford, A., … Haverfield, E. (2015). Assessing copy number from exome

sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(8), 623–9. Retrieved from http://dx.doi.org/10.1038/gim.2014.160

Rinker, E. B., Dueber, J. C., Qualtieri, J., Tedesco, J., Erdogan, B., Bosompem, A., & Kim, A. S. (2016). Differential expression of ribosomal proteins in myelodysplastic syndromes. *Journal of Clinical Pathology*, *69*(2), 176–180. http://doi.org/10.1136/jclinpath-2015-203093

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896–902. Retrieved from http://dx.doi.org/10.1038/nbt.2931

Risso, D., Schwartz, K., Sherlock, G., & Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, *12*(1), 480. Retrieved from http://www.biomedcentral.com/1471-2105/12/480

Robert, C., Karaszewska, B., Schachter, J., Rutkowski, P., Mackiewicz, A., Stroiakovski, D., … Schadendorf, D. (2014). Improved Overall Survival in Melanoma with Combined Dabrafenib and Trametinib. *New England Journal of Medicine*, *372*(1), 141116004513004. http://doi.org/10.1056/NEJMoa1412690

Rodríguez-Paredes, M., & Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, *17*(3), 330–9. Retrieved from http://dx.doi.org/10.1038/nm.2305

Rosenfeld, J. A., & Mason, C. E. (2013). Pervasive sequence patents cover the entire human genome. *Genome Medicine*, *5*(3), 27. http://doi.org/10.1186/gm431

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., … Shah, S. P. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, *11*(4), 396–8. http://doi.org/10.1038/nmeth.2883

Rous, P. (1911). A Sarcoma of the Fowl Transmissable by an Agent Separable from the Tumor Cells. *The Journal of Experimental Medicine*, *13*(4), 397–411. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19867421

Sandoval, J., Mendez-Gonzalez, J., Nadal, E., Chen, G., Carmona, F. J., Sayols, S., … Esteller, M. (2013). A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *31*(32), 4140–7. Retrieved from http://jco.ascopubs.org/content/early/2013/09/27/JCO.2012.48.5516

Setoodeh, R., Schwartz, S., Papenhausen, P., Zhang, L., Sagatys, E. M., Moscinski, L. C., & Shao, H. (2013). Double-hit mantle cell lymphoma with MYC gene rearrangement or amplification: a report of four cases and review of the literature. *International Journal of Clinical and Experimental Pathology*, *6*(2), 155–67. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3544229&tool=pmcentrez&rendertype=abstract

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., … Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, *498*(7453), 236–40. http://doi.org/10.1038/nature12172

Shameer, K., Readhead, B., & T. Dudley, J. (2015). Computational and experimental advances in drug repositioning for accelerated therapeutic stratification. *Current Topics in Medicinal Chemistry*, *15*(1), 5–20. http://doi.org/10.2174/1568026615666150112103510

Shen, L. (2002). DNA Methylation and Environmental Exposures in Human Hepatocellular Carcinoma. *CancerSpectrum Knowledge Environment*, *94*(10), 755–761. Retrieved from http://jnci.oxfordjournals.org/content/94/10/755

Shi, Z., Fujii, K., Kovary, K. M., Genuth, N. R., Röst, H. L., Teruel, M. N., … al., et. (2017). Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. *Molecular Cell*, *32*(0), 710–714. http://doi.org/10.1016/j.molcel.2017.05.021

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews. Cancer*, *6*(10), 813–23. http://doi.org/10.1038/nrc1951

Shyr, C., Kushniruk, A., van Karnebeek, C. D. M., & Wasserman, W. W.

(2015). Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors. *Journal of the American Medical Informatics Association : JAMIA*. http://doi.org/10.1093/jamia/ocv053

Sistigu, A., Yamazaki, T., Vacchelli, E., Chaba, K., Enot, D. P., Adam, J., … Zitvogel, L. (2014). Cancer cell-autonomous contribution of type I interferon signaling to the efficacy of chemotherapy. *Nature Medicine*, *20*(11), 1301–1309. http://doi.org/10.1038/nm.3708

Starczynowski, D. T. (2014). Errant innate immune signaling in del(5q) MDS. *Blood*, *124*(5). Retrieved from http://www.bloodjournal.org/content/124/5/669?sso-checked=true

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3398141&tool=pmcentrez&rendertype=abstract

Stehelin, D., Varmus, H. E., Bishop, J. M., & Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, *260*(5547), 170–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/176594

Su, M., Dou, X., Cheng, H., & Han, J.-D. J. (2015). Integrative Epigenomics. In A. E. Teschendorff (Ed.), *Computational and Statistical Epigenomics* (Vol. 7, pp. 127–139). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-017-9927-0

Su, Z., Łabaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., … Shi, L. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, *32*(9), 903–914. Retrieved from http://dx.doi.org/10.1038/nbt.2957

Sulonen, A.-M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., … Saarela, J. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology*, *12*(9), R94.

Retrieved from http://genomebiology.com/2011/12/9/R94

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., … Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, *14*(4), 381–387. http://doi.org/10.1038/nmeth.4220

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., … von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(D1), D447–D452. http://doi.org/10.1093/nar/gku1003

Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., … Zhu, M. (2014). An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human Mutation*, *35*(7), 899–907. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24599517

Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., … Widschwendter, M. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Research*, *20*(4), 440–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847747&tool=pmcentrez&rendertype=abstract

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, *352*(6282), 189–196. http://doi.org/10.1126/science.aad0501

Toung, J. M., Morley, M., Li, M., & Cheung, V. G. (2011). RNA-sequence analysis of human B-cells. *Genome Research*, *21*(6), 991–8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106332&tool=pmcentrez&rendertype=abstract

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., … Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*, 381–6. http://doi.org/10.1038/nbt.2859

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., &

Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. http://doi.org/10.1038/nbt.2450

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., … Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, *509*(7500), 371–5. http://doi.org/10.1038/nature13173

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. Retrieved from http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Van Keuren-Jensen, K., Keats, J. J., & Craig, D. W. (2014). Bringing RNA-seq closer to the clinic. *Nature Biotechnology*, *32*(9), 884–5. Retrieved from http://dx.doi.org/10.1038/nbt.3017

Venkatraman, E. S., & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, *23*, 657–663. http://doi.org/10.1093/bioinformatics/btl646

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*, *339*, 1546–58. http://doi.org/10.1126/science.1235122

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, *34*(11). http://doi.org/10.1038/nbt.3711

Wang, G., Jung, K., Winnenburg, R., & Shah, N. H. (2015). A method for systematic discovery of adverse drug events from clinical notes. *Journal of the American Medical Informatics Association : JAMIA*. http://doi.org/10.1093/jamia/ocv102

Wang, J., Moore, N. E., Deng, Y.-M., Eccles, D. A., & Hall, R. J. (2015). MinION nanopore sequencing of an influenza genome. *Frontiers in Microbiology*, *6*, 766. http://doi.org/10.3389/fmicb.2015.00766

Weisenberger, D. J. (2014). Characterizing DNA methylation alterations from The Cancer Genome Atlas. *The Journal of Clinical Investigation*, *124*(1), 17–23. Retrieved from http://www.jci.org/articles/view/69740

Welch, J. D., Hartemink, A. J., & Prins, J. F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biology*, *17*(1). http://doi.org/10.1186/s13059-016-0975-3

Welch, J. D., Hu, Y., & Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, *44*(8), e73. http://doi.org/10.1093/nar/gkv1525

Wheeler, D. A., & Wang, L. (2013). From human genome to cancer genome: The first decade. *Genome Research*, *23*(7), 1054–1062. http://doi.org/10.1101/gr.157602.113

Woll, P. S., Kjällquist, U., Chowdhury, O., Doolittle, H., Wedge, D. C., Thongjuea, S., … Jacobsen, S. E. W. (2014). Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells In Vivo. *Cancer Cell*, 794–808. http://doi.org/10.1016/j.ccr.2014.03.036

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., … Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, *11*(1), 41–6. http://doi.org/10.1038/nmeth.2694

Yoshida, K., Yamamoto, K., Kohno, T., Hironaka, N., Yasui, K., Kojima, C., … Matsuyama, T. (2005). Active repression of IFN regulatory factor-1-mediated transactivation by IFN regulatory factor-4. *International Immunology*, *17*(11), 1463–71. http://doi.org/10.1093/intimm/dxh324

Zhang, Y., Wolf, G. W., Bhat, K., Jin, A., Allio, T., Burkhart, W. A., & Xiong, Y. (2003). Ribosomal protein L11 negatively regulates oncoprotein MDM2 and mediates a p53-dependent ribosomal-stress checkpoint pathway. *Molecular and Cellular Biology*, *23*(23), 8902–12. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC262682