UNIVERSITY OF CALIFORNIA,
IRVINE

A Genomic Analysis Pipeline and Its Application to Pediatric Cancers

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Michael Dylan Zeller

Dissertation Committee:
Professor Pierre Baldi, Chair
Associate Professor Xiaohui Xie
Professor Bogi Andersen

2014

DEDICATION

In memory of my father, Donald G. Zeller.

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vii**

ACKNOWLEDGMENTS **viii**

CURRICULUM VITAE **ix**

ABSTRACT OF THE DISSERTATION **xii**

1 Introduction **1**
1.1 Heterogeneity of cancer 2
1.2 Gene expression in cancer 3
1.3 The rise of personalized medicine 4
1.4 Pediatric cancer 5
1.5 Genomic analysis pipeline 6

2 Sequence processing methods **8**
2.1 DNA-seq 9
2.1.1 Genome assembly 11
2.1.2 Identifying variations 12
2.1.3 Variant location and effect 15
2.2 RNA-seq 18
2.2.1 Small variations in transcripts 20
2.2.2 *De novo* assembly 22
2.3 ChIP-seq 23
2.3.1 TCF1E ChIP-seq in a colon cancer cell line 25
2.3.2 Sequence features of GRHL3 ChIP-seq peaks 27

3 Expression analysis **30**
3.1 Transcription factor binding sites 31
3.1.1 Searching for motifs 32
3.1.2 Incorporating conservation 32
3.1.3 Phosphorylation site prediction from motifs 33
3.1.4 Alternative target recognition in the Wnt signaling pathway 37

3.1.5 Enrichment analysis 42
3.1.6 Defining the transcriptional regulatory network of skin wounding . . . 43
3.2 Differential expression analysis 45
3.2.1 Microarray differential analysis 46
3.2.2 Identifying differential antigens 49
3.2.3 RNA-seq differential analysis 50
3.2.4 RNA-seq differential analysis in pediatric patients 53
3.2.5 Neuron-specific chromatin regulatory subunit BAF53b 55
3.3 Hybrid approach 59

4 Integrative analysis 62
4.1 Network analysis 62
4.1.1 Circadian clock regulates the host response to Salmonella 64
4.2 Cancer-specific integrative analysis 67
4.2.1 Curating literature 68
4.2.2 GEO datasets 70
4.2.3 Gene list overlaps 74
4.2.4 Normal tissue expression 76
4.2.5 Mitelman fusions 76
4.2.6 Prognosis 78
4.3 Cancer pathways 79
4.4 Identifying drug candidates 81

5 Reporting 83
5.1 Framework 84
5.2 Ranking pathways and genes in pediatric cancer 85
5.3 Interesting findings in pediatric cancer 87
5.3.1 Patient CHOC23 (AML) 87
5.3.2 Patient CHOC33 (Neuroblastoma) 90
5.3.3 Patients CHOC36 and CHOC03 (AML) 91

6 Conclusion 95
6.1 Applications of pipeline 96
6.2 Comparison to other work 96
6.3 Importance of automation 97
6.4 Value of integrative approach 98
6.5 Future work 99

Bibliography 100

A Supplementary Information 117
A.1 Hardware requirements 117
A.2 Software requirements 118
A.3 Pediatric cancer 119

B Collaborative interface 123

LIST OF FIGURES

Page

1.1 Heterogeneity of cancer 2

2.1 Pipeline overview 8
2.2 RNA-seq coverage for Grhl3 transcript in knock-out 20
2.3 k-medoid clustering of time-course RNA-seq reads 21
2.4 dnTCF1E ChIP-seq reads in Colon cancer cells 26
2.5 Identifying enriched features in ChIP-seq peaks 28

3.1 Transcriptional regulation overview 31
3.2 Visual representation of position-weight matrix 33
3.3 Predicted phosphorylation sites using PWM search 34
3.4 Motifs identified in TCF1 ChIP-seq peaks 39
3.5 Histogram of frequency of C-clamp motif in ChIP-seq peaks 40
3.6 Identifying functional ChIP-seq peaks based on PWMs 41
3.7 Enriched transcription factors in human skin wounding response 45
3.8 Example microarray image for two samples 49
3.9 Identifying differential antigens 51
3.10 Performance of predicting antigens using protein microarrays 52
3.11 Expression differences between patients 54
3.12 RNA-seq differential genes in learning and memory 57
3.13 Regression of rat RNA-seq with microarray for the same samples 60
3.14 Yeast hybrid differential analysis results 61

4.1 Transcription factor network integrated with microarray data 66
4.2 Quantile normalized microarray datasets for each cancer type 72
4.3 Diagram of the processes used to prioritize expression variations 74
4.4 Distribution of genetic variants among RPKM values 75
4.5 Detecting false positive differential genes 77
4.6 Prognostic markers and survival curves 79
4.7 Example of a network with drug, interaction, and transcription databases . . 80
4.8 Portion of KEGG AML pathway for patient CHOC36 81
4.9 CHOC36 drug-target edges in AML pathway 82

5.1 Example org-mode markup for reports 85
5.2 UCSC Genome Browser displaying genetic variations in PARK2 transcripts . 91

5.3 PARK2 gene expression in patient tumors 92
5.4 Enriched transcription factor LMO2 in AML 93

6.1 Technology biases in different sequence technologies 99

LIST OF TABLES

Page

2.1 Typical distribution of the small variations 13
2.2 Distribution of extracted gene lists per patient 16

3.1 Contingency table 43
3.2 Enriched transcription factors in skin wounding 44

4.1 Distribution of curated gene lists 68
4.2 Number of gene symbols in curated gene lists 69
4.3 Number of microarray samples per type of cancer 71
4.4 Significance of gene list overlaps with missense mutations 75
4.5 Hits in Mitelman fusion database for sarcoma patient 78

5.1 Top 10 ranked pathways for CHOC23 (AML) 88
5.2 Top 10 curated gene ranking for CHOC23 (AML) 89

ACKNOWLEDGMENTS

I would like to acknowledge all of our past and present collaborators. Without their data, none of this work would have been possible: Olivier Cinquin Lab, Marcelo Wood Lab, Bogi Andersen Lab, Philip Felgner Lab, Lee Bardwell Lab, Paolo Corsi Lab, Alan Barbour Lab, Susan Sandmeyer Lab, Peter Kaiser Lab, and Dr. Sender and his colleagues at CHOC.

In particular I would like to acknowledge Amy Hopkin, Liz and Will Gordon, Ramzi Nasr, Marina Bellet, Arlo Randall, Christophe Magnan, Matthew Kayala, Paul Rigor, Kenny Daily, Pietro di Lena, Vishal Patel, Thomas Kristensen, Annie Vogel, James Yu, Nate Hoverter, Yu Liu, Nick Ceglia, Peter Sadowski, and last but certainly not least, Stacey Borrego.

I would also like to thank those entities whose funding made this work possible:

1. Bioinformatics Training Grant (BIT). NIH Grant 5T15LM007743. 2009-2013
2. UC Irvine Center for Complex Biological System's Opportunity Award. Defining the Transcriptional Regulatory Network of Acute Human Epidermal Wound Response in Time and Space. Collaboration with William Gordon, Bogi Andersen Lab. 2011-2012

This work has been made possible due to support by a grant from the Hyundai Foundation to Dr. Leonard Sender and grants NIH LM01, NIH NLM T15 LM07, and NSF IIS-0513376 to Pierre Baldi.

We acknowledge also the support of the CHOC, the UCI Institute for Genomics and Bioinformatics, and the UCI Genomics High-Throughput Facility. Additional support of our computational infrastructure has been provided by Jordan Hayes and Yuzo Kanomata.

Professors Rick H. Lathrop and Charless Fowlkes are acknowledged for providing guidance while serving on my advancement committee.

Professor Xiaohui Xie for providing exceptional training for the BIT program coursework.

Professor Bogi Andersen for providing insightful discussion during the course of my PhD through regular research meetings.

I thank Professor Pierre Baldi for his mentorship and guidance during my graduate career and for his valuable advice on all aspects of my thesis.

In addition, the inclusion of our previous articles discussed in this document was made possible by the generous policies of IEEE, Springer-Verlag, the National Academy of Sciences, Elsevier, Inc., and the Nature Publishing Group.

CURRICULUM VITAE

Michael Dylan Zeller

EDUCATION

Doctor of Philosophy in Computer Science	**2014**
University of California, Irvine	*Irvine, California*
Bachelor of Science in Computer Science	**2008**
Washington University	*St. Louis, Missouri*
Bachelor of Science in Biomedical Engineering	**2008**
Washington University	*St. Louis, Missouri*

RESEARCH EXPERIENCE

Graduate Research Assistant	**2008–2014**
University of California, Irvine	*Irvine, California*

TEACHING EXPERIENCE

Microarray Analysis Short Course	**2012**
University of California, Irvine	*Irvine, California*
Reader	**2011–2011**
University of California, Irvine	*Irvine, California*
Teaching Assistant	**2009–2009**
University of California, Irvine	*Irvine, California*

WORK EXPERIENCE

Part-time Bioinformatician	**2013–2014**
Zymo Research	*Irvine, California*
Software Engineer Co-op	**2007–2007**
St. Jude Medical CRMD	*Sylmar, California*
Undergraduate Researcher (NSF)	**2006–2006**
Washington University	*St. Louis, Missouri*
Web developer	**2005–2012**
Red Rhinoceros	*New York, New York*

AWARDS

sbv IMPROVER Challenge 2 Subchallenge 2 **2013**
2nd place with Peter Sadowski. Results presented in Athens, Greece

sbv IMPROVER Challenge 1 **2012**
5th place overall with Peter Sadowski and William Gordon, Andersen Lab

UCI CCBS's Opportunity Award **2011-2012**
In collaboration with William Gordon, Bogi Andersen Lab

Bioinformatics Training Grant **2009-2013**
NIH Grant 5T15LM007743

LEADERSHIP

External Liaison **2012–2013**
Palo Verde Resident's Council *Irvine, California*

Organizer **2010–2011**
UCI Wine Night *Irvine, California*

President **2009–2010**
Bioinformatics Training Grant (BIT) Program *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

A Genomic Analysis Pipeline and Its Application to Pediatric Cancers **2014**
M. Zeller, C. N. Magnan, V. R. Patel, P. Rigor, L. Sender, and P. Baldi. *IEEE TCBB* Accepted

The TCF C-clamp DNA Binding Domain Expands the Wnt Transcriptome via Alternative Target Recognition **2014**
N Hoverter, **M Zeller**, MM McQuade, A Garibaldi, A Busch, KJ Hertel, P Baldi, and M Waterman. Submitted

A GRHL3-regulated repair pathway suppresses immune-mediated epidermal hyperplasia **2014**
Gordon W, **Zeller M**, Herndon Klein RH, Swindell W, Ho H, Espetia F, Gudjonsson JE, Baldi P, and Andersen B. Submitted

Inter-species prediction of protein phosphorylation in the sbv IMPROVER species translation challenge **2014**
M, Biehl, P. Sadowski, G. Bhanot, E. Bilal, A. Dayarian, P. Meyer, R. Norel, K. Rhrissorrakrai, **M. Zeller**, and S. Hormoz. Submitted

Combining docking site and phosphosite predictions to find new substrates: Identification of Smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate **2013**

Gordon EA, Whisenant T, **Zeller M**, Kaake R, Gordon W, Krotee P, Patel V, Huang L, Baldi P, Bardwell L. *Cellular signalling* August 24 2013

Inferring Epitopes of a Polymorphic Antigen Amidst Broadly Cross-Reactive Antibodies Using Protein Microarrays: A Study of OspC Proteins of Borrelia burgdorferi **2013**

Baum E, Randall AZ, **Zeller M**, Barbour AG. *PLoS One* 8(6), June 24 2013

The neuron-specific chromatin regulatory subunit BAF53b is necessary for synaptic plasticity and memory **2013**

Vogel-Ciernia, A., Matheos, D. P., Barrett, R. M., Kramr, E. A., Azzawi, S., Chen, Y., Magnan, C. N., **Zeller, M.**, Sylvain, A., Haettig, J., Jia, Y., Tran, A., Dang, R., Post, R. J., Chabrier, M., Babayan, A. H., Wu, J. I., Crabtree, G. R., Baldi, P., Baram, T. Z., Lynch, G., and Wood, M. A. *Nature neuroscience* 16(5), 552561 May 2013

Circadian clock regulates the host response to Salmonella **2013**

Bellet MM, Deriu E, Liu JZ, Grimaldi B, Blaschitz C, **Zeller M**, Edwards RA, Sahar S, Dandekar S, Baldi P, George MD, Raffatellu M, Sassone-Corsi P. *PNAS* May 28 2013

GRHL3/GET1 and Trithorax Group Members Collaborate to Activate the Epidermal Progenitor Differentiation Program **2012**

A. S. Hopkin, W. Gordon, R. H. Klein, F. Espitia, K. Daily, **M. Zeller**, P. Baldi, and B. Andersen. *PLoS Genetics* 2012

High-throughput prediction of protein antigenicity using protein microarray data **2012**

C. Magnan, **M. Zeller**, M. Kayala, A. Vigil, A. Randall, P. Felgner, and P. Baldi. *Bioinformatics* Sept 2010

REFEREED CONFERENCE PUBLICATIONS

A Signature Compiler for the Edinburgh LF **2007**

M. Zeller, A. Stump, and M. Deters. *Logical Frameworks and Meta-Languages: Theory and Practice (LFMTP)*, 2007. Affiliated with the Conference on Automated Deduction (CADE). Bremen, Germany, July 15, 2007

ABSTRACT OF THE DISSERTATION

A Genomic Analysis Pipeline and Its Application to Pediatric Cancers

By

Michael Dylan Zeller

Doctor of Philosophy in Computer Science

University of California, Irvine, 2014

Professor Pierre Baldi, Chair

We present a cancer genomic analysis pipeline which takes sequencing reads for both germline and tumor genomes as input and outputs prioritized lists of the most affected genes in the tumor genome. Using publicly available datasets and literature specific to each patient, we extract out clinically relevant information to be used in a novel reporting and ranking system in order to identify the most affected genes and pathways within a patient. Network-based approaches that integrate protein-protein, protein-TF, and protein-drug interaction data are used to identify potentially therapeutic drugs and their targets. Effects of genetic variations on gene expression, as profiled by RNA-seq in tumor samples, are used to provide further evidence of "driver" mutations – those mutations responsible for tumor progression. Additionally, previously implicated small and large variations (including gene fusions) are reported.

Results are presented in a collaborative interface that combines all evidence for the top ranking genes and pathways. Affected genes in and around protein coding sequences are investigated further using sequence-level features such as predicted secondary structure, solvent accessibility, phosphorylation status, and protein domains. By using an integrative approach, effects of genetic variations on gene expression are used to provide further evidence of "driver" mutations.

This pipeline has been developed with the aim to be used in assisting in the analysis of pediatric tumors, as an unbiased and automated method. We present results that agree with previous literature and highlight specific findings in a few patients. Portions of this pipeline have been successfully reused in the analysis of other high-throughput sequencing data in non-cancer related projects. This work provides a basis for which future personalized medicine pipelines can be systematically performed in order to assist in the treatment of newly diagnosed cancer patients.

Chapter 1

Introduction

Cancers are ranked among the leading causes of death worldwide, with 8.2 million deaths in 2012 as reported by the World Health Organization (WHO). The American Society of Clinical Oncology predicts that within the next 16 years, cancer will overtake heart disease as the number one killer. The rise of cancer as a leading cause of death can be attributed to an increase in the average lifespan, which was 46.3 in 1900 and rose to 73.8 in 1998 for males in the US. As an individual ages, and risks such as accidental death and heart attacks decrease, the occurrence of cancer within an individual is inevitable. In fact, about 80 percent of men who reach age 80 have prostate cancer cells in their prostate, although in many cases this will not lead to death [39]. It is imperative that we address the burden that the rise of cancer will have on the clinician and in the research lab, through the development of standard methods for analyzing cancer genomes that lead to easily interpretable results.

1.1 Heterogeneity of cancer

At the most fundamental level, cancer is a disease of the DNA. Mutations are acquired in your cells, either through environment or genetic susceptibility over time. Normally, these mutations are sequestered through normal cell machinery, but in some cases they sneak through. Over time, these changes to the DNA sequence and the molecules that interact with it ultimately lead to uncontrolled cell proliferation. The multiply affected pathways through which these mutations are acquired have been well characterized, and no single pathway or mutation gives rise to the cancer phenotype [70].

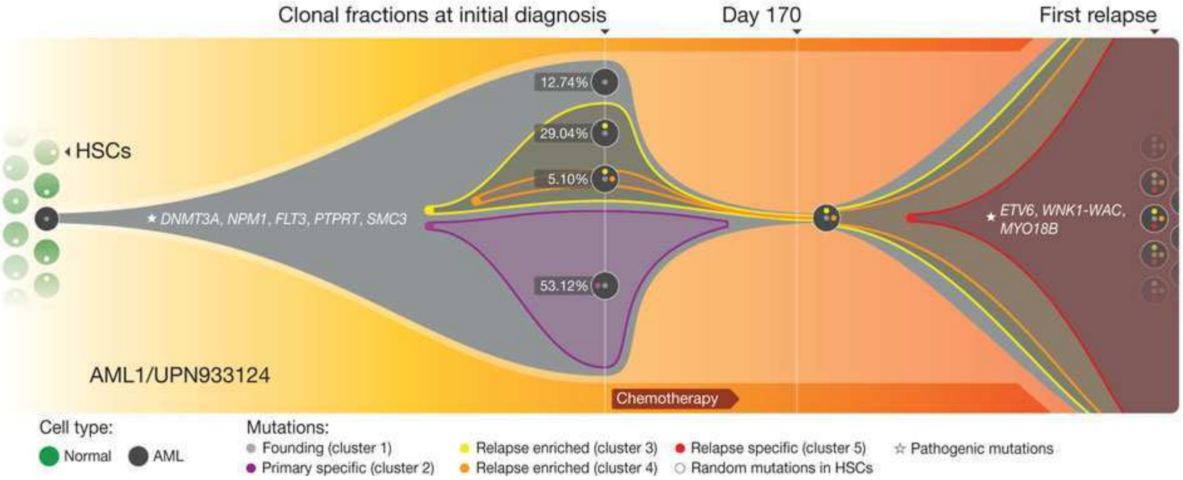

Clonal fractions at initial diagnosis
Day 170
First relapse
HSCs
12.74%
29.04%
5.10%
53.12%
★ DNMT3A, NPM1, FLT3, PTPRT, SMC3
★ ETV6, WNK1-WAC, MYO18B
AML1/UPN933124
Chemotherapy
Cell type:
● Normal
● AML
Mutations:
● Founding (cluster 1)
● Primary specific (cluster 2)
● Relapse enriched (cluster 3)
● Relapse enriched (cluster 4)
● Relapse specific (cluster 5)
○ Random mutations in HSCs
☆ Pathogenic mutations

Figure 1.1: **Cancer is naturally heterogeneous. As a tumor progresses, multiple mutations are acquired. A subset of mutations allow a tumor to survive chemotheraphy and result in relapse after treatment.**
Figure and legend adapted from [46].

It has been shown that cancer evolves through a reiterative process of clonal expansion [66]. As shown in Figure 1.1, this process in AML involves multiple tumor cell subpopulations, that during the normal course of treatment, become resistant to chemotherapy through selective pressure. The resulting population of cells contain multiple acquired mutations, some of which can be identified through analytic approaches [187]. Typically, 2 to 8 mutations are attributed to "driving" the tumor progression, non-essential mutations in the tumor are referred to a “passenger” [169]. Thus, current standard therapies could benefit from a more targeted, personal approach that takes into account the specific subclonal populations present in an individual’s tumor. Sorting out the “driver” from “passenger” mutations in an individual is essential in meeting this goal.

1.2 Gene expression in cancer

The role of gene expression in cancer is less understood than that of the acquired DNA mutations, especially those that have an obvious affect on the resulting protein sequence. Mutations in DNA regions that have unknown function may still play a role in the development of uncontrolled cell growth, through the increase in transcription of certain key genes and pathways. The ENCODE project annotated 80% of the genome as functional, much of which is out-side of protein coding sequences ([2, 3]). As far back as 2000, researchers have been able to identify specific and common aberrant gene expression profiles among cancers [139].

More recently, others have been able to successfully extract out prognostic genes from gene expression data in pancreatic cancer [180], supporting the role of aberrant gene expression playing an important role in tumor progression. While researchers have been successful in analyzing the human genome for DNA alterations using high-throughput sequencing technologies, the challenge of integrating multiple sources of data, such as gene expression, still remains [47].

1.3 The rise of personalized medicine

With respect to our current understanding of cancer biology, our approach to cancer has become more personalized. It has become standard procedure to screen for specific BRCA1 mutations in breast cancer cases, in addition to other low-throughput methods such as immunohistochemistry approaches for HER2 positive or negative staining in breast cancer [114]. These approaches are limited to a few handful of genes and cannot, for instance, identify novel mutations in BRCA1. Thus high-throughput sequencing technologies capable of identifying not only the DNA sequence (DNA-seq) but, for instance, also epigenetic states (e.g. Methyl-seq) and gene expression levels (RNA-seq), hold the promise to help better understand cancer in all its various forms in addition to holding the key information needed for a more personalized approach to medicine.

Indeed large-scale cancer sequencing projects, such as the Cancer Genome Atlas [6], have already started and produced volumes of data that are already well beyond what can be transferred over the Internet. However, these projects are still at a relatively early stage of development and are fraught with numerous challenges associated with the complexity of the sequencing technology, the lack of standardization, the sheer volume of data, the heterogeneity of cancers, the complexity of cancer biology, and the problem of obtaining proper control samples, to name only a few. Although incomplete by necessity, problems, solutions, and results from these projects ought to be shared periodically in order to move the field towards more standardized solutions and accelerate the pace of discovery.

As recent as 2009, personalized medicine treatments for AML based on markers identified using microarray profiling [151] had been suggested. In more recent years we have seen cases where high-throughput sequencing was used during patient treatment ([175, 103]). In an even more recent paper, researchers outline the challenges and necessary components of a personalized medicine pipeline [164]. Despite a decline in sequencing costs as we approach reaching the goal of a $1000 genome, significant barriers to entry still exist in getting personalized medicine into the clinic, including the high-cost of analyzing these genomes [108].

1.4 Pediatric cancer

Worldwide, it is estimated that childhood cancer has an incidence of more than 175,000 per year, and a mortality rate of approximately 96,000 per year. In the United States, cancer is the second most common cause of death among children between the ages of 1 and 14 years, exceeded only by accidents, with an incidence of about 12,000 of newly diagnosed cases per year and 1,300 deaths. The most common cancers in children are (childhood) leukemia (34%), brain tumors (23%), and lymphomas (12%). Other, less common childhood cancer types are: Neuroblastoma (7%), Wilms tumor (5%), NonHodgkin lymphoma (4%) , Rhabdomyosarcoma (3%), Retinoblastoma (3%), Osteosarcoma (3%), Ewing sarcoma (1%), Germ cell tumors, Pleuropulmonary blastoma, Hepatoblastoma, and hepatocellular carcinoma. White and Hispanic children are more likely than children from any other racial or ethnic group to develop cancer. The causes of most childhood cancers are unknown.

In this thesis we describe the ongoing development of a computational pipeline for the analysis of high-throughput sequencing cancer data that is currently being applied to pediatric cancer data that is regularly being sequenced, and further re-sequenced on recurrence, as a result of a collaboration between the University of California, Irvine (UCI) and the Children Hospital of Orange County (CHOC). The CHOC receives on the order of 100 new cases per year, and a project was started in 2012 to sequence the genome from healthy and cancer tissues of a subset of newly diagnosed cases – and therefore with no emphasis on particular tumors or tissue types – together with high-throughput gene expression measurements from cancer cells using RNA-seq.

1.5 Genomic analysis pipeline

The impact of high-throughput sequencing (HTS) has not been confined to cancer biology research. While sequencing analysis tools have been around for a long time [143], the impact of high-throughput sequencing on genetics research in recent years cannot be denied. For any research question you can now probe the genome or transcriptome of your samples and identify significant differences between a control and an experimental condition.

We have successfully used high-throughput sequencing in a number of collaborations on campus at the University of California, Irvine. In doing so, we have acquired standard analysis pipelines for dealing with each of the major types of sequencing data being obtained currently:

- DNA-seq
- RNA-seq
- ChIP-seq

Our goal has been to develop an analysis pipeline comprising a combination of in-house and third party software to manage and analyze the raw data produced by these experiments. In the case of pediatric cancer, we intend to so in a timely manner after sequencing data becomes available. This includes the identification and ranking of affected genes containing both small and large variants, and their integrative systems biology analyses against the large background of omic, literature, and other data available to us in order to derive inferences of clinical relevance specific to the cancer types of the patients sequenced. In the case of other projects, we aim to explore novel biological questions with an emphasis on assessing the differences in DNA composition or gene expression profiles between two biological conditions.

In the subsequent chapters we outline a genomic analysis pipeline that was developed to explore cancer, and in particular pediatric cancers. Portions of this pipeline are reused to study other related diseases such as canonical Wnt signaling in colon cancer [76] (submitted), predicting phosphorylation sites in the human proteome [63], and the role of transcription factor Grhl3 in skin wounding and related diseases [64] (submitted), to name a few applications. This pipeline is used to analyze sequencing reads directly from Illumina sequencers, which is commonly delivered by the Genomics High-throughput Sequencing Facility (GHTF) here on campus, in addition to variant calling services performed by two popular vendors – Illumina, Inc. and Complete Genomics.

Chapter 2

Sequence processing methods

High-throughput sequence processing is by no-means standardized as of yet. Solutions exist, but vary in details from lab to lab, and even from person to person within a single lab. Many tools exist for processing this data, and each have their own pros and cons, as well as varying levels of technical support. Within our lab, we have multiple ways to process different types of high-throughput sequencing data, with multiple steps depending on the project.

An overview of the workflow for the pediatrics cancers project is presented in Figure 2.1. It begins by collecting two different samples for each patient participating in the CHOC pediatric cancers project. The first sample is collected in the tissue affected by the cancer

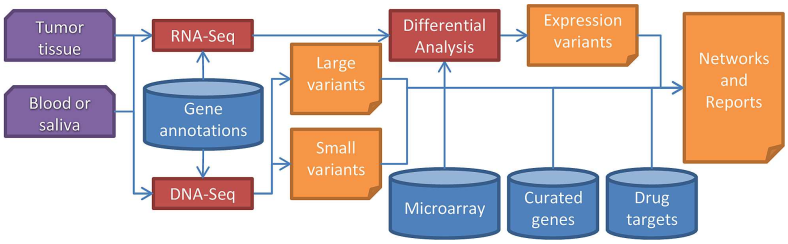


Figure 2.1: **Overview of the genomics analysis pipeline which starts from raw sequencing reads derived from two biological samples per patient and results in a HTML report with ranked genes and pathways.**

and the second sample is collected either from blood or saliva, depending on the patient, to be used as a control sample during the analysis. The samples are processed each by the DNA-seq step, and the tumor sample is additionally processed by the RNA-seq analysis step.

The DNA-seq step in this case provides list of variants and the RNA-seq provides expression values which feeds into the Differential Analysis step (see Section 3.2.4). An additional pipeline step exists for processing ChIP-seq datasets, which were not obtained in the pediatrics project. We describe the pipeline steps with respect to the pediatrics project, but also highlight a few examples throughout the thesis where these steps are used in other non-cancer related projects.

2.1 DNA-seq

DNA-seq is the most common method used to probe the genome in order identify the DNA alterations responsible for a disease, such as p53 mutations in cancer [122]. When the mutations have been previously well characterized, DNA microarrays can be used to detect a wide range of DNA alterations, but this fails to characterize novel mutations. More recently, approaches such as exome sequencing, which sequences only the 1% of the genome responsible for coding sequences, have been successful in identifying the mutations responsible for disease [123]. Additionally, DNA-seq is the method by which we build reference genomes of model organisms, including human [1]. Variations of this technology exist, such as the Pacific Bioscience's RS II sequencer, which can perform single DNA molecule sequencing producing read lengths of up to 20kb [23].

For our pediatrics project, we take advantage of commercial solutions that exist for rapidly processing DNA-seq data in order to identify DNA variations compared to the human reference genome. Complete Genomics's (Mountain View, CA) *Cancer Sequencing Service* and

Illumina, Inc.'s (San Diego, CA) *RapidTrack WGS Service* are used to sequence the tumor and control samples. Sequencing platforms, data vendors, patient description, and data obtained for each patient are reported in Table A.1.

Sequencing reads provided by Complete Genomics are paired 35 bp reads where each 35 bp read is made of four shorter reads close but not necessarily contiguous on the sequenced genome. Sequencing reads for all the datasets generated on an Illumina instrument are paired 100 bp reads. Each vendor perform their own variant calling. Illumina, Inc. provides DNA-seq variants called using their software CASAVA. Open-source solutions exist in addition to these commercial solutions, such as VarScan2 [93] or Gatk [45], for identifying DNA mutations in a wide-range of DNA-seq data, and can be used in place of these commercial vendors for identifying variants within DNA-seq reads.

Quality controls and data filtering The sequencing data quality is assessed based on the standard PHRED quality scores predicted during the base calling step on each sequencing platform and on the base call distribution for each sequencing cycle. Typically, reads mapping to the mitochondrial or nuclear ribosomal RNA genes and PhiX control reads are removed from the original datasets when analysis is performed in-house.

Alignment to reference genome hg19/GRCh37 Both Complete Genomics and Illumina deliver the short-read alignment results as part of their sequencing service. Alignments delivered by Complete Genomics are performed using the CGA tools developed by the same company to handle the specific structure of the short-reads. Alignments delivered by Illumina are performed using their short-read aligner Eland v2e. Alignments to the reference genome can also be performed using open-source software Bowtie [99] or BWA [100] which are based on the Burrows-Wheeler transformation [26], although for Complete Genomics these do not perform well with short-reads of length 35bp.

2.1.1 Genome assembly

Standardized multi-genome assemblies Rapidly evolving technologies and softwares, lack of standardization, and large volumes of heterogeneous data are common issues in genomic analyses and are paired here with the necessity to compare several assemblies for the same individuals to extract relevant differences potentially correlated with the corresponding disease. Our strategy to limit the effects of these problems and provide uniform downstream analyses for all the patients is to adopt a fixed representation of genome assemblies consisting in three major components: 1) the features needed for the downstream analysis; 2) a fixed level of description and annotation; and 3) a standardized scoring system and data format allowing multiple genome assemblies and RNA-seq experiments to be rapidly combined with each other and compared.

Features selected to describe each assembly include a unique call for each allele, the called allele sequence, zygosity, ploidy, call confidence level, read counts (coverage), and genomic annotations corresponding to each position.

Control and cancer genomes comparison The two assembled genomes for each patient can directly be compared from the assemblies described in the previous section. First, positions not fully called on both genomes and both alleles have been excluded from the rest of the analysis as they do not allow a reliable comparison between the two genomes. Around 95% of the known positions in the reference sequences are fully called on both genomes regardless of the sequencing platform.

2.1.2 Identifying variations

Small variations are usually defined as the DNA differences with the reference genome sequences that can be directly observed in and predicted from short sequencing reads, i.e. of very limited size. These variations can be accurately predicted in many cases, are widely studied, reported in numerous databases, and many of them are already documented for their possible implication in diseases together with their frequency in the population. They are thus of great interest for genomic analyses and the focus of many studies worldwide.

Comparative analysis The small variations called during the genome assembly (see Section 2.1.1) are classified into four categories: SNPs, insertions, deletions, and substitutions. Between 4 and 4.5 million such variations with the reference sequences are called for each assembled genome in the pediatrics cancers project. The large number of small variations results in many false calls and systematic biases in many cases, and some methods to address this have been developed [29].

However, in our case, the two samples for each patient in the CHOC pediatric cancers project are sequenced on the same platform and the genomes are assembled using the same methods and software, hence a significant part of the biases and false calls is thus likely to be repeated on each assembly. By comparing both genomes and extracting only the differences between the cancer genome and the control genome, we can reasonably assume these issues to affect the resulting set of variations significantly less.

Variations observed between the cancer genome and the control genome only represent a very small fraction of the variations called on both genomes, less than 0.01% in most cases (example provided for one patient in Table 2.1), reducing drastically the number of variations to further analyze for each patient. Variations observed on both genomes are not studied further regardless of the effect these variations may have on proteins when they occur in

Table 2.1: **Mixed variations include cases where different small variants are called between the two alleles of a genome or between the two genomes.**
Table adapted from [188] (submitted).

Variation	Control DNA vs Reference		Cancer DNA vs Reference		Cancer DNA vs Control DNA	
SNPs	3,359,243	76.08%	3,356,920	76.11%	36,931	40.15%
Substitutions	207,760	4.71%	209,452	4.74%	20,545	22.33%
Deletions	637,146	14.43%	633,823	14.37%	31,024	33.73%
Insertions	201,883	4.57%	200,976	4.56%	3,472	3.77%
Mixed	9,381	0.21%	9,193	0.21%	14	0.02%
Total	4,415,413		4,410,364		91,986	

gene regions.

Absence of control samples in other diseases In some cases there is an absence of a control sample for DNA-seq, such as is the case for non-cancer diseases that arise due to environmental changes or a genetic disorder from birth. When dealing with high-throughput sequencing data for these cases, we still usually have access to a cohort of normal patient samples, typically a random subset of the population that does not have the disease in question (e.g. 1000 genomes project [4]). This cohort of normal patients can be processed in exactly the same way as the disease samples, so that it can be used to infer a set of commonly occurring small and large variations.

These variations are defined to be those variations identified by our pipeline that are also present in a significant subset of the normal population. These significant common variants are used to filter the disease patient variants, in much the same way as if we had a single normal sample for that patientas in the case of our pediatric patients. Additionally, we can still make use of our other data sources to filter out other commonly observed SNPs, particularly the SNPs found in dbSNP, to further filter the disease patient variants.

Large variations Large variations are the large-scale chromosomal variations or rearrangements leading to a significant change in the classical organization of the DNA in the genome.

- **Novel junctions** are observed junctions between distant parts of the genome (intra- or inter-chromosomal).

- **Gene fusions** are a special case of novel junctions leading to the fusion of two distantly located genes resulting in a new, functionally different protein product. This analysis is performed in-house for the data delivered by Illumina.

- **Copy number variations (CNVs)** are relatively large deletion or duplication events leading to a different number of copies observed for specific regions of the genome.

- **Chromosome duplications** or deletions are a particular case of CNVs where an entire chromosome is either missing or observed with more than two copies. These variations are detected using in-house software and further validated based on the expression results obtained following the protocol described in Section 3.

We thus implemented a case-by-case set of rules based on the overlap length between the large variations predicted for the baseline genome and the ones predicted for the cancer genome (not detailed here) to decide if a large variation is likely to be unique to the cancer genome or not. Similarly to the small variations, we list all of the genes affected by large variations in the cancer genome considering the eight following categories:

- Deletion
- Inversion
- Tandem-duplication
- Distal-duplication
- Inter-chromosomal rearrangement
- Gene fusion
- Higher CNV
- Lower CNV

The sizes of each of these gene lists are summarized in Table 2.2.

2.1.3 Variant location and effect

Small variations observed only in the cancer genome and in genic regions can be further analyzed as their effect on the resulting proteins can be directly deduced from their location in many cases. Eight distinct types of disruptions or changes in the proteins are considered in our pipeline. Seven of them are inspired by the classification of the variant effects performed by the CGA tools (Complete Genomics) and we added the loss of heterozygosity between the germline and cancer genomes, frequently reported in cancer cases [91], resulting in the following classification:

- Missense (change of amino-acid)
- Nonsense (premature stop codon)
- Nonstop (stop codon altered)
- Misstart (start codon altered)
- Splicing (variation in a donor or acceptor site)
- Frameshift (indels changing the reading frame)
- Inframe (indels not changing the reading frame)
- LOH (loss of heterozygosity)

For each of the eight variant effects listed above, we use the hg19 gene coordinates to extract the list of genes overlapping the called small variations. The confidence for a gene to be actually affected by such variation is directly given by the confidence of the small variation call. The sizes of the gene lists for each patient are summarized in Table 2.2 and range from a few genes for the most deleterious variations to a few hundred genes for missense mutations, which are more common and less likely to be deleterious than other variations.

Table 2.2: **Mean size and standard deviation of the gene lists extracted during the various stages of the analysis for each patient. These lists are used in computing a ranking score for each gene in the final reports.**
Table adapted from [188] (submitted).

Small Variations			**Large Variations**			**Gene Expression**		
Gene List	**Mean**	**StdDev**	**Gene List**	**Mean**	**StdDev**	**Gene List**	**Mean**	**StdDev**
Missense	589.9	437.1	Deletion	406.2	555.1	Under expr. HIGH	52.0	67.1
Nonsense	26.2	27.8	Inversion	234.3	325.1	Under expr. MED	433.2	172.3
Nonstop	2.0	3.0	Tandem-Dup	122.8	188.1	Under expr. LOW	14.2	21.7
Misstart	1.2	2.0	Distal-Dup	8.3	27.6	Over expr. HIGH	24.8	42.7
Splicing	31.4	28.0	Inter-Chr	51.0	51.2	Over expr. MED	438.2	99.6
Frameshift	113.3	161.2	Gene Fusion	24.4	21.7	Over expr. LOW	4.9	11.2
Inframe	61.4	51.0	Higher CNV	782.2	1221.8	All tumors HIGH	54.9	71.1
LOH	179.6	321.5	Lower CNV	488.4	1423.3	All tumors MED	457.0	147.7
						All tumors LOW	74.0	82.1
						Tumor sig.	1430.4	489.8
						Control sig.	1087.2	1208.5
						Contrast sig.	2455.7	3011.0
UNION	781.1	563.6	UNION	2030.3	1774.7	UNION	4897.9	2570.6

Flagged small variations Putative small variations, specifically single nucleotide polymorphisms (SNPs), have been categorized into three subsets – unique, common, or flagged – with respect to the latest dbSNPs (version 137) tracks from the UCSC Genome Browser [113]. Specifically, when creating these subsets we used the curated subsets of dbSNPs referred to as Common SNPs and Flagged SNPs.

SNPs that have a minor allele frequency of at least 1% and that are mapped to a single location in the reference genome assembly are included in the Common SNPs subset. Taken as a set, these commonly occurring SNPs should be less likely to be associated with severe genetic diseases.

Further, for the Flagged SNPs, only SNPs flagged as clinically associated by dbSNP, that map to a single location in the reference genome assembly, and not known to have a minor allele frequency of at least 1%, are included. SNPs that do not fit into either the common or flagged SNPs subsets are categorized as unique SNPs, specific to the patient. Additionally, COSMIC annotations [56] are added to the small variation calls whenever available.

Protein domains Besides the commonly characterized affect of small variations on protein coding sequence detailed in Section 2.1.3, we characterized the location of small variants based on predicted secondary structure and solvent accessibility using the SCRATCH software suite [33]. In addition, protein domain families predicted by Pfam [129] are used to identify if the small variation affects a protein family domain, which in many cases can identify important functional portions of a protein such as protein binding domains. This information is incorporated into our final report in order to manually investigate the consequences of small variations.

Phosphorylation sites Phosphorylation sites identify locations of serine, threonine, and tyrosine residues which are targeted by kinases, which account for about 500 human proteins. PhosphoSite Plus [75] is used as a database of validated phosphorylation sites across all human proteins for 155,588 non-redundant sites. Sequence variations overlapping the portions of the coding sequences responsible for coding for these affected residues are identified using simple overlaps of coordinates, as they may play an important role in many diseases, particularly cancer [133] (see Section 3.1.3).

Variant transcription factor binding sites Putative transcription factor binding sites (TFBS) for the human reference genome build hg19 are predicted using MotifMap ([184, 37]). A conservation score of at least 2 (the bayesian branch length score (BBLS)), along with a FDR score of at most 0.20 (computed using randomly permuted motifs) are used to filter potential binding sites down to a total of 3,523,896 sites across the 717 transcription factors annotated by TRANSFAC (version 9) and JASPAR. TFBS are overlapped for variants falling within 5bp of the consensus to identify possible deleterious regulatory connections in our network analysis.

Mitelman fusions The Mitelman database [116] contains 3752 entries corresponding to gene fusions implicated in different types of cancer. To identify and prioritize these gene fusions in our patients, we cross this database with all of the gene fusions found for each patient to identify high-priority fusions and to present the relevant literature in our final reports that are of clinical relevance. Three of the patients in our study contained fusions previously described. These fusions were originally identified in the same tumor type as each of the patients. All identified Mitelman fusions are listed in Table A.3.

2.2 RNA-seq

RNA-seq measures the abundance of transcripts within a sample. RNA-seq correlates with cDNA microarrays, another method of measuring transcript abundance, in addition to providing a better estimate of protein levels as measured using mass spectrometry [57]. It is more sensitive, particularly for low-level transcripts, and linearly related to the abundance of the target transcript [121]. It also has the ability to identify novel transcripts and differentially expressed isoforms. On the other hand, the analysis of RNA-seq is much more computationally intensive.

RNA-seq in pediatric patients When the sample extracted at the tumor tissue is not exhausted by the DNA extraction, RNA sequencing is also performed using an Illumina HiSeq 2500 instrument either by the Scripps Research Next Generation Sequencing Core Facility (San Diego, CA) or by the Genomics High-Throughput Facility of the University of California, Irvine (Irvine, CA). The RNA sequencing data is subject to the same quality controls and is pre-filtered to remove common contaminants in RNA-seq libraries. The RNA sequencing data is aligned to the reference genome hg19 together with the known splice junction sequences extracted from the RefSeq database using Eland v2e. Sequencing data obtained and the vendors used for each patient is shown in A.1.

Quantifying expression Gene expression levels are computed directly from the read alignment results. Standard RPKM values (reads per kilobase of exon model per million mapped reads) [121] are computed directly for each exon, splice junction, and isoform covered by the sequencing data. In some cases, as is the case for our pediatric patients, no RNA sequencing data is available for the baseline genome samples, or any other control tissue samples, and therefore standard differential analysis of the gene expression levels cannot be performed for each patient. This problem is discussed further in Section 3.2.4.

Other tools used Additional methods for quantifying expression exist, and another common set of tools is used to process RNA-seq datasets outside of the pediatric cancers project, in some cases. Typical measurements of expressions are made by counting the number of reads that align to different mRNA sequences, i.e UTRs and exons, and tools exist to construct the reference transcriptome from annotation files, such as TopHat2 [88].

In order to use TopHat2, mean fragment lengths are computed over FASTQ files by first aligning all reads using Bowtie [99] and then using SAMtools [101] to compute the mean distance between mate pairs in a paired-end RNA-seq. This distance is then used to prevent discordant reads and to pick the best read alignments based on the expected read-length distribution. In some cases, especially with longer reads such as the 100bp reads from Illumina, fragment lengths below 200bp can cause dove tailing (i.e. paired-end reads overlap), and in most cases should be allowed. After aligning to the reference transcriptome, TopHat2 uses cufflinks [162] to probabilistically assign reads to different isoforms, based on the abundance of reads aligning to unique exon, in addition to identifying novel transcripts based on reads aligning to the exons of known transcripts.

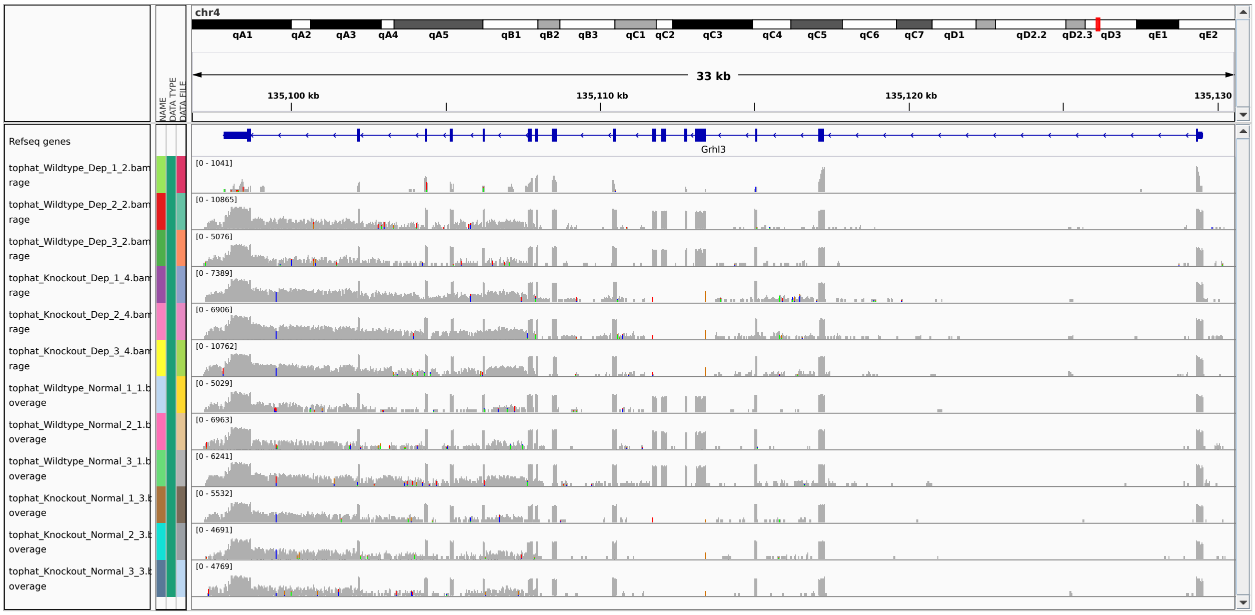

chr4
qA1 qA2 qA3 qA4 qA5 qB1 qB2 qB3 qC1 qC2 qC3 qC4 qC5 qC6 qC7 qD1 qD2.2 qD2.3 qD3 qE1 qE2
33 kb
135,100 kb
135,110 kb
135,120 kb
135,130
NAME
DATA TYPE
DATA FILE
Refseq genes
Grhl3
tophat_Wildtype_Dep_1_2.bam rage
tophat_Wildtype_Dep_2_2.bam rage
tophat_Wildtype_Dep_3_2.bam rage
tophat_Knockout_Dep_1_4.bam rage
tophat_Knockout_Dep_2_4.bam rage
tophat_Knockout_Dep_3_4.bam rage
tophat_Wildtype_Normal_1_1.b overage
tophat_Wildtype_Normal_2_1.b overage
tophat_Wildtype_Normal_3_1.b overage
tophat_Knockout_Normal_1_3.b overage
tophat_Knockout_Normal_2_3.b overage
tophat_Knockout_Normal_3_3.b overage
[0 - 1041]
[0 - 10865]
[0 - 5076]
[0 - 7389]
[0 - 6906]
[0 - 10762]
[0 - 5029]
[0 - 6963]
[0 - 6241]
[0 - 5532]
[0 - 4691]
[0 - 4769]

Figure 2.2: **Grhl3 transcript visualized in IGV [160]. RNA-seq reads align to exons for Grhl3 mutant and normal samples. Deleted exon denoted by lack of reads on exons 4 through 7 for mutant samples. Abnormal sample is identified by a lack of reads across all Grhl3 examples (first sample).**

Visualizing read counts across transcripts The accuracy of the read counts in this context can be used to easily differentiate between mutant and wildtype mice harboring a deletion in the coding sequence of the gene, Grhl3 (Figure 2.2). Additionally, tools exist such as cummerRbund [61], which can assist in providing visualizations of RNA-seq data. Clustering of transcripts can be performed for time-course RNA-seq experiments, as was performed for a Carbon limited time-course in *Y. lipolytica* (Figure 2.3).

2.2.1 Small variations in transcripts

RNA sequencing data is available for a large portion of the patients in the pediatric cancers project. Small variations can also be called for the transcripts based on the alignment results (Section 2.1). We use the software developed by Illumina, Casava Variant Detection and Counting (VDC), to extract SNPs and indels following the same protocol as the one

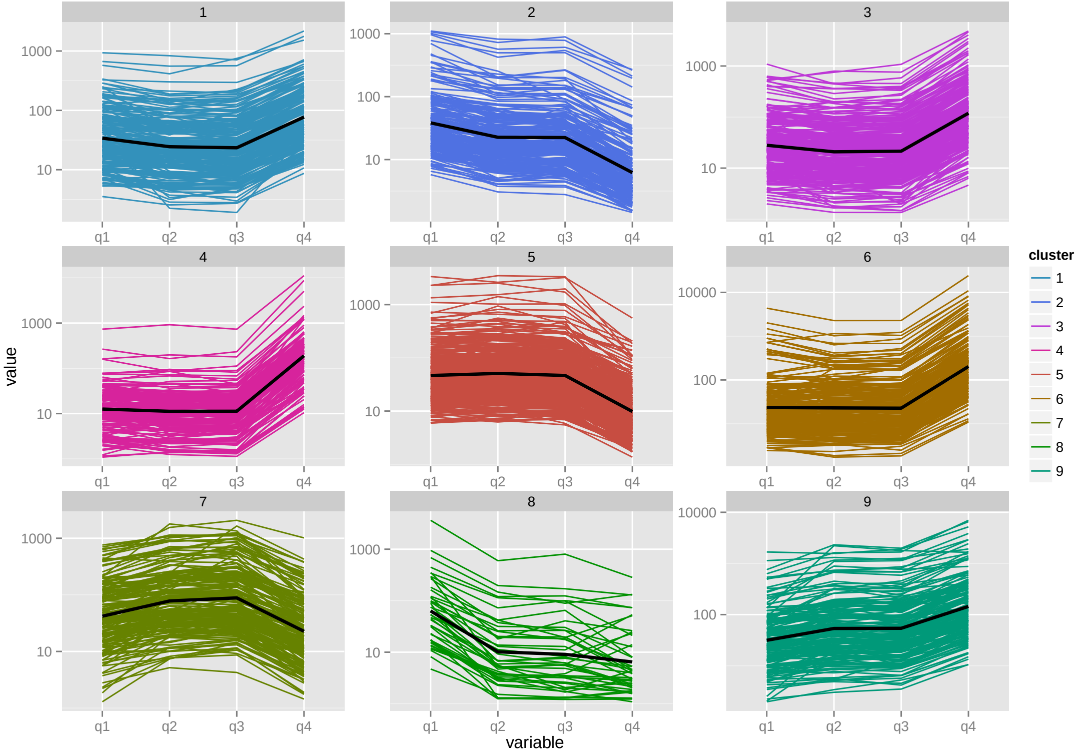


Figure 2.3: **The cummeRbund R package [61] can be used to extract out significantly changing transcripts from cuffdiff results and performs** ***k*****-medoid clustering on expression profiles. Nine clusters of Yarrowia RNA-seq time-course are identified, showing that the majority of transcripts change expression at the last time point.**

used by Illumina for their DNA sequencing service.

Additionally, we make use of TopHat-Fusion [89] for Illumina, Inc. reads. TopHat-Fusion wraps the TopHat tool to align each end of a 100bp read to the genome to discordant regions of the transcriptome. After identifying such reads, TopHat-Fusion can identify the junction involved in a fusion between two transcripts. These fusion transcript can be used to validate fusions identified in the DNA-seq in such cases.

2.2.2 *De novo* assembly

De novo assembly of transcripts is performed in some cases in order to be reference independent. Many tools exist to solve this problem, such as Oases [147], SOAPdenovo-Trans [185], and Trinity [65]. *De novo* assembly of RNA-seq reads provides the ability to characterize previously uncharacterized transcripts, using tools such as ORF Finder [137] to identify coding sequences within assembled reads, and then subsequently to BLAST these ORFs to identify likely functions for these transcripts. RNA-seq *de novo* assembly can be used to identify 3' and 5' UTRs of transcripts, which would otherwise be unknown for a reference genome that has been previously unannotated.

Expression profiling in Yarrowia lipolytica time-course *De novo* assembly of RNA-seq reads is necessary when a good reference genome does not currently exist, as is the case for *Y. lipolytica*. While assemblies do exist ([86, 49]), they are poorly annotated and lack gene expression information aside from predicted transcripts based on sequence alone, using tools such as GLIMMER [43], in addition to tools specifically designed for yeast such as YGAP [176], which are only able to explain approximately 20-30% of reads derived from the RNA-seq (data not shown).

We performed *de novo* assembly of RNA-seq reads in *Y. lipolytica* RNA-seq time-course shown in Figure 2.3 using Trinity [65]. After obtaining *de novo* transcripts, we identify ORFs and BLAST to the 32 yeast genomes to identify the transcript based on protein sequence alignment results. In this case, we then aligned these transcripts back to two versions of the reference genome using Exonerate [150], in order to compare the quality of the genome assemblies (based on DNA-seq reads).

We evaluated the quality of our own in-house assembly and found that a larger number of transcripts align back to this reference compared to the previous genome assembly performed in 2004 [49]. On average, 134 more *de novo* transcripts align back with 95% or better sequence identity to the new assembly and on average 2600 *de novo* transcripts align better than the previous assembly.

2.3 ChIP-seq

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is used to pull-down fragments of DNA bound to specific factors, usually transcription factors, and then sequence them. ChIP-seq, like the other high-throughput sequencing methods, can be performed using an microarray instead of sequencing (ChIP-chip) but is rarely performed anymore. ChIP-seq almost always uses a reference genome in order to align reads, and *de novo* methods are almost unheard of, nor are they needed, as the quality of the genome assembly does not have a large impact on ChIP-seq results [25].

Peak calling ChIP-seq relies on identifying the DNA fragments that were pulled down by antibodies attached to beads. Identification of these fragments is done through a statistical technique known as peak calling, which looks for higher than normal read count distributions along a region (i.e. peaks).

A number of solutions exist for doing so, including MACS [189] and QuEST [165]. p-value cutoffs (typically p<10-5, or a peak score of 50) are used to assess the confidence in binding based on the parameters of a Poisson modeled estimated on equivalent regions of the Input sample, which is the cross-linked DNA before the antibody pull-down. Typically, a false discovery rate is estimated by calling peaks within this Input sample for each p-value cutoff.

ChIP-seq peaks are analyzed for enriched sequences (i.e. motifs). This process is described in further detail in Section 3.1.4. The tool HOMER [72] is used to assign peaks to their nearest gene or to identify overlapping genomic features. Enrichment of these features can be assessed by sampling random fragments of the genome and comparing the distributions of features such as conservation, methylation, CpG%, among others (see Section 2.3.2) within the called peaks.

Quality control Besides the use of an input DNA sample, a number of quality control metrics exist, such as those laid out by the ENCODE project [97]. The ENCODE project recommends comparing the peak sets of the replicates such that Rep A and B peak lists are truncated to the same length, followed by comparing top 40% of the replicate A peak list with the entire replicate B peak list (and vice versa). If 80% of the top 40% set is contained in the larger set, then the data is considered to be very reproducible. Another metric, the irreproducibility discovery rate [102], has been shown to work well and can be used to determine a cutoff for the number of peaks to call based on the concordance among ChIP-seq biological replicates and can be implemented using MACS. Lastly, peaks between samples or between experiments can be overlapped to determine a significance using the hypergeometric distribution. HOMER [72] is able to compute this statistic based on the overlaps of peak BED files.

2.3.1 TCF1E ChIP-seq in a colon cancer cell line

ChIP-seq was performed between a mutant and wildtype TCF1E in a colon cancer cell line DLD-1 to profile the DNA binding profile in the presence or absense of a second binding domain within this protein (See Section 3.1.4) under doxycycline-induced expression at 0hr, 2hrs, and 9hrs.

ChIP-seq sequencing reads were analyzed by first aligning 50bp single-end reads to the hg19 reference genome (GRCh37/hg19) with Bowtie v2.0.0-beta6 [98]; 70-80% of the reads uniquely mapped to the genome), allowing at most three mismatches per read. MACS v1.4 [189] was used to call ChIP-seq peaks where only one unique read per position was retained to avoid PCR artifacts and a default cut-off for a peak score greater than 50 was used to discard weak binding events. Replicate mapped reads were either pooled together before peak calling (pooled analysis) or replicates were kept separate and peak calling was performed independently (replicate analysis).

A comparison of the peaks called at each time point exhibited significant overlap for each of the dnTCF1EWT and dnTCF1Emut biological replicates. For example, the reciprocal overlap of the top 20% of wildtype 2 hour peaks was 61% and 54% and for mutant 2hrs it was 80% and 86%. We further assessed reproducibility using IDR analysis, a stringent rank order approach that compares the rank order p-values of peaks in biological replicates. IDR analysis showed that a cut-off of the top 1,000 peaks had a False Discovery Rate of 0.1 for dnTCF1EWT and .04 for dnTCF1Emut. Following the ENCODE protocol for ChIP-seq, we therefore pooled the biological replicates for each time point.

The top 1000 peaks identified in our wildtype are shown in Figure 2.4A. The fainter heatmap signals for the mutant samples at these sites indicate weaker binding, a general pattern also reflected in the fewer number of total peaks called for the mutant. Read distributions within peaks are visualized using IGV [160] (Figure 2.4C), by computing the read coverage at every

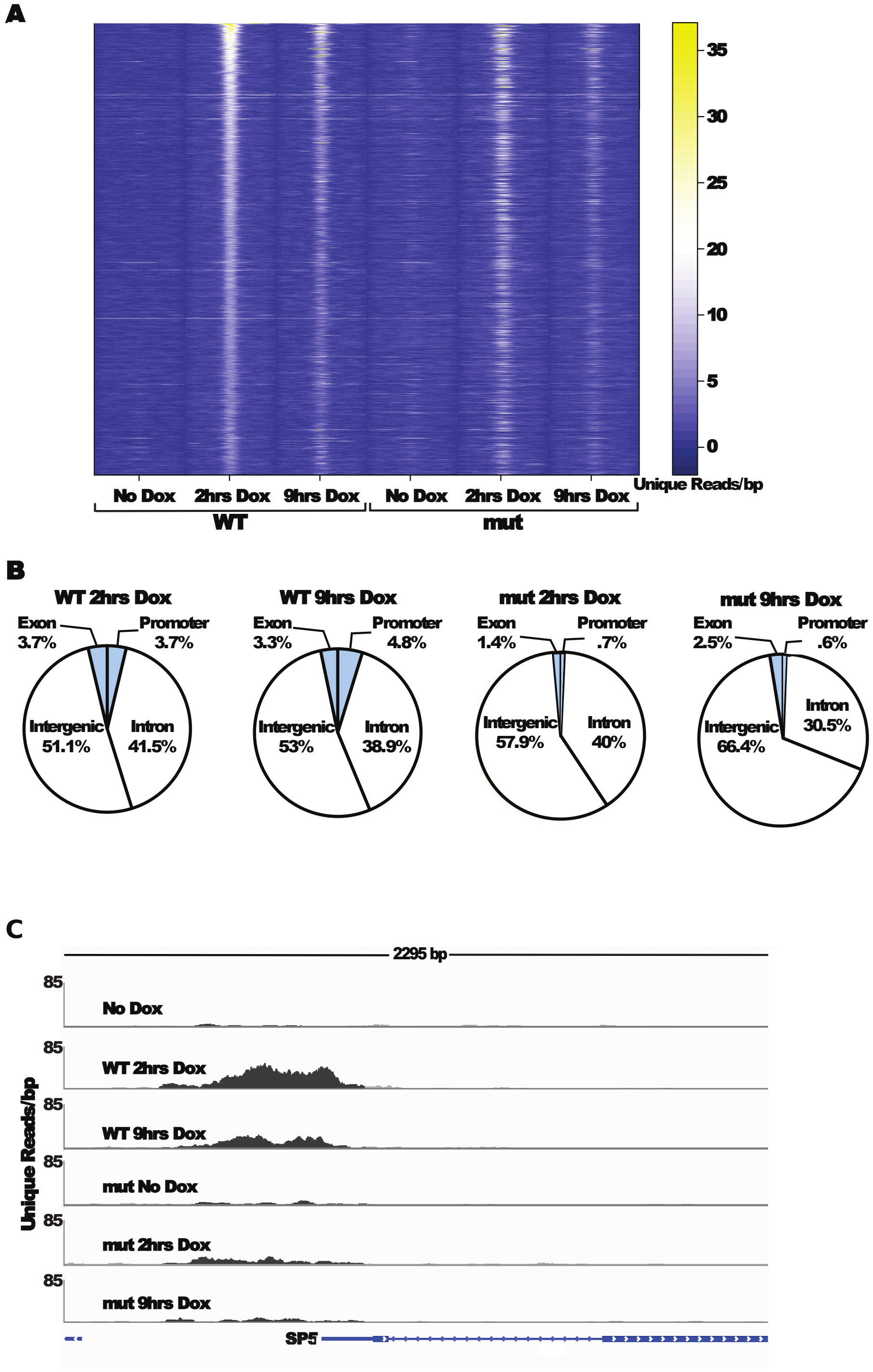


Figure 2.4: **(a) Heatmap of reads aligned to genome for the top 1000 peaks in wildtype at 2 hours. (b) Pie charts of location of peaks for wildtype and mutant samples. (b) Coverage of reads for a peak upstream of known Wnt target SP5. This peak appears to be C-clamp dependent.**
Figure adapted from [76] (submitted).

base pair of the genome using bedtools [130].

We also asses the significance of overlap of our ChIP-seq, with another recently published β-catenin ChIP-seq experiment in a different colon cancer cell line [172]. As we would expect, we found significant enrichment of the 10614 β-catenin ChIP-seq peaks within our mutant and wildtype peaks for 2 and 9 hours (p=10E-587 for WT 2hrs, 10E-125 for MUT 2hrs), but not our 0hr time points (p=0.74 and p=0.25, respectively). Additionally, these peaks can be analyzed later using the methods outlined in Chapter 3 and are presented in Section 3.1.4.

2.3.2 Sequence features of GRHL3 ChIP-seq peaks in wounding experiments [64]

We explore the sequence level features of Grhl3 ChIP-seq peaks in various conditions related to skin development. We identified 4,035 (common to two independent experiments), 4,820 and 9,294 significant GRHL3 peaks (FDR<0.05), respectively, in embryonic day 16.5 mouse skin, recovery from wax stripping, and on day 4 of 5% IMQ treatment. Unexpectedly, the perfect consensus GRHL3 binding site ([96, 179]) was found in a relatively small fraction of the peaks (for further discussion on searching for motifs and PWMs within our pipeline see Section 3.1.1), and many computationally identified [37] GRHL3 sites across the genome were not bound by GRHL3. To test for significance of motif occurrence, PWMs for GRHL3 were searched across all peaks using HOMER and MotifMap in the ChIP-seq peaks as well as for 100 random shuffles of peaks (controlling for the peak size and chromosomal distribution), and assessed significance using the Z-score at α <0.05. We found that GRHL3 PWMs were significant within the embryonic and wax stripping ChIP-seq but not the IMQ, while the GRHL3 perfect consensus was not significant in any of the ChIP-seq.

We assessed a number of peak level features for significance in peaks containing a GRHL3 PWM site versus genome-wide sites containing the GRHL3 PWM: distance to nearest TSS,

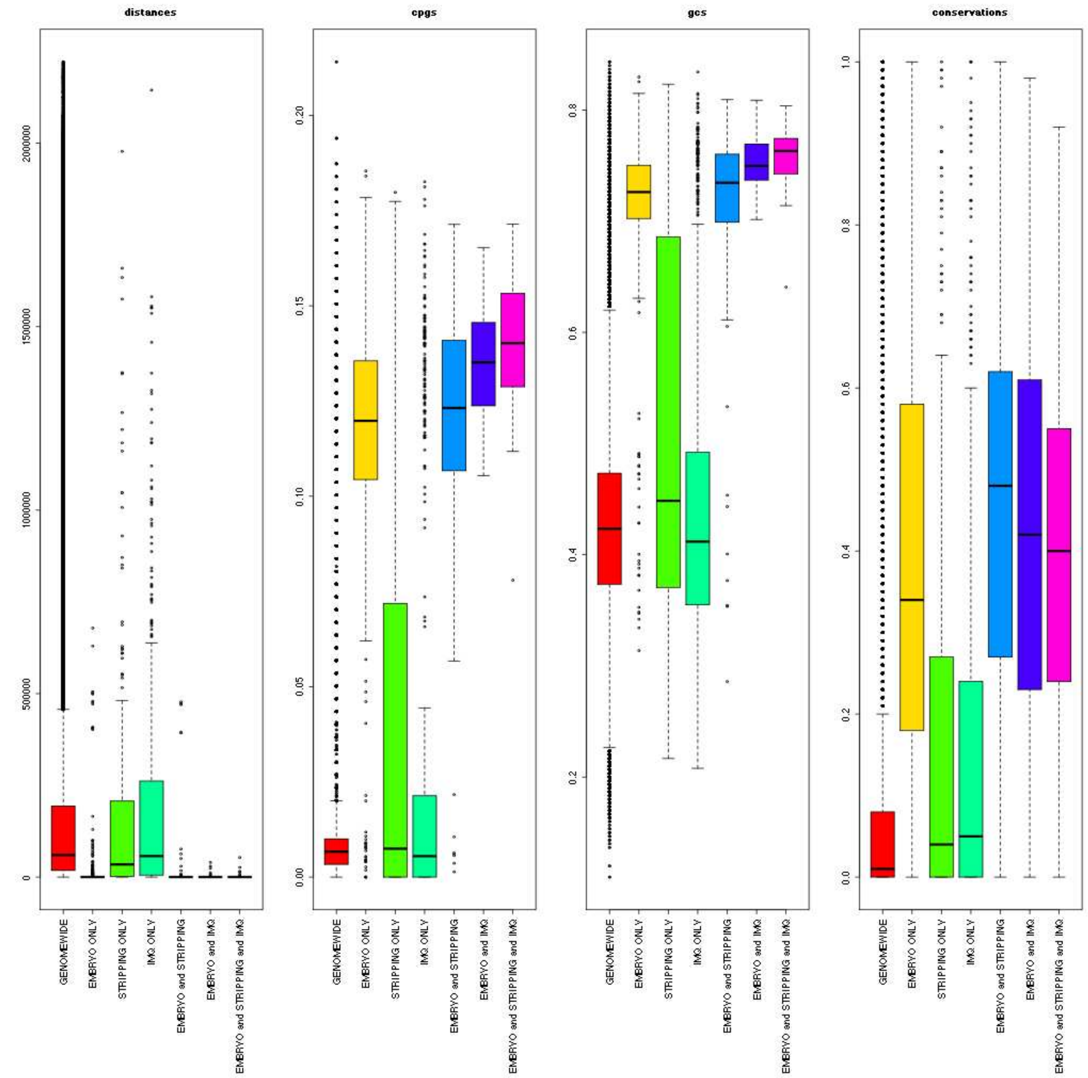


Figure 2.5: **Distance to nearest TSS, CpG%, GC%, and mean conservation for ChIP-seq peaks containing the GRHL3 PWM as compared to the 300bp around all GRHL3 binding sites within the genome.**
Figure adapted from [64] (submitted).

CpG%, GC%, Conservation, CpG methylation overlap, RepeatMask overlap. Using a log-odd score cutoff tuned to a known GRHL3 site, we used MotifMap to first identify all locations of the GRHL3 PWM in the genome (mm9) above the cutoff, expanded to 300bp (roughly the average peak length for the ChIP-seq peaks). Similarly, ChIP-seq peaks were filtered for those containing at least one GRHL3 site above the cutoff. For each peak level feature, we tested for a significance difference in the ChIP-seq peaks containing GRHL3 as compared to the genome background using a one-sided Wilcoxon rank-sum test for the numeric features (distance, CpG%, GC%, and conservation) and a Fishers exact test on the CpG methylation and RepeatMasker overlap counts. We found that both evolutionary sequence conservation and DNA methylation showed highly significant enrichment in the bound versus non-bound GRHL3 PWM sites and that the highest confidence peaks common between the different experiments showed even greater enrichment of these features (Figure 2.5). This data suggests that GRHL3 preferentially associates with those sites found in regions marked by sequence conservation and DNA methylation, indicative of gene-regulatory regions.

Chapter 3

Expression analysis

As described previously (see Section 1.2), the role of gene expression in cancer is important, but not as well understood as the role of DNA variations falling within protein coding sequences. Figure 3.1 shows an overview of the various transcriptional regulators in the cell. The role of chromatin, depicted as beads on a string in this figure, is explored in Section 3.2.5, as an application of our standard expression analysis pipeline for RNA-seq. The other components of this figure, transcription factors (either proximal or distal), are explored via the sequence composition of their observed binding sites, known as transcription factor binding sites (TFBS), and how we search for and quantify them. We explore how the distribution of these sites within the sequences upstream of a gene can be used to identify causal relationships between transcription factors and a list of genes.

We then move onto describe the portion of the pipeline that identifies the most significant expression differences between two groups of samples. With respect to the pediatric cancer patients, we show how this is used to identify transcripts with aberrant gene expression in the tumor. One of the more novel aspects of our genomics analysis pipeline is our incorporation of expression data into a DNA-seq pipeline, as described later in Section 4.2.2. In order to obtain these results, we have a standard pipeline for the differential analysis of the two primary sources of expression data: (1) microarray and (2) RNA-seq. We also include a few examples of where this analysis step was performed to investigate non-cancer related research questions.

3.1 Transcription factor binding sites

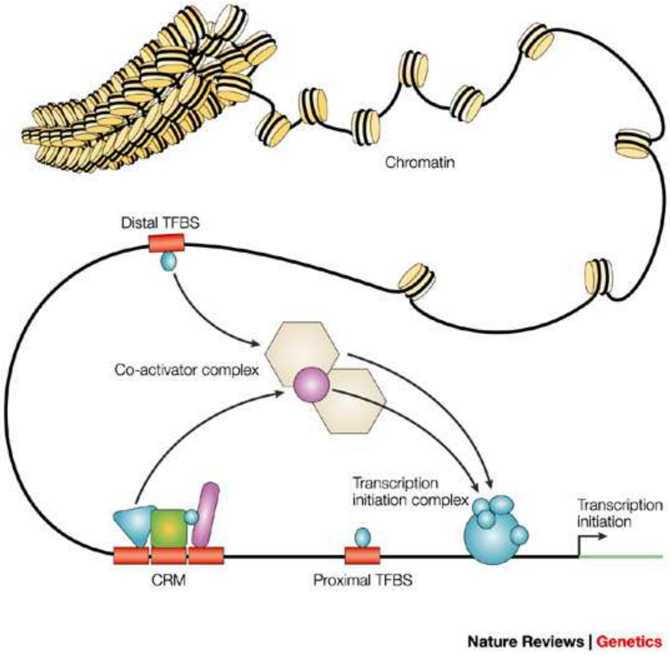


Figure 3.1: **Transcription initiation is regulated by multiple factors, such as the organization of chromatin as well as the presence or absence of co-factor complexes. Transcription factor binding sites (TFBS) directly upstream or distal to the transcription initiation sites are used to identify sets of sequences to which a particular transcription factor binds.**
Figure and legend text adapted from [171].

3.1.1 Searching for motifs

The sequences for which a transcription factor binds to, usually identified using either protein binding affinity assays or ChIP-seq for that factor, are referred to as transcription factor binding sites (TFBS). TFBS can be summarized for a factor in the form of a position-weight matrix (PWM). These PWMs denote the relative base-pair frequencies of each base position of a TFBS, usually with respect to a background base pair frequency. Such a model is limited in that it cannot model gaps or conditional probabilities between bases in a TFBS, but in practice it works quite well and has become ubiquitous in the field. A visual representation of an example PWM is shown in Figure 3.2, as rendered by the software WebLogo [36].

These log-odds matrices for a particular transcription factor are referred to as motifs. A number of tools exist for working with transcription factor motifs in the form of PWMs, including tools such as PoSSuMsearch [16] and MOODS [94] which are regularly used to identify sequences that match with a certain probability within a large set of sequences, in addition to the general ChIP-seq analysis software suite HOMER [72], which identifies these sequences within a smaller subset of regions of interest in the genome. MEME [11] and HOMER can even extract a PWM from a set of sequences by identifying sets of enriched k-mers (sequences of length k) for various k.

3.1.2 Incorporating conservation

Our pipeline relies heavily on MotifMap ([184, 37]), which making use of the MOODS pipeline [94] in order to search for all annotated motifs obtained from the databases TRANSFAC [111] and JASPAR [24]. These sites are annotated for popular genome builds, for which we currently support mouse (mm9), human (hg19), and yeast (sacCer2). In addition to a probability of binding, MotifMap provides a measure of conservation, the bayesian branch length score (BBLS), that has been shown to increase the accuracy of identifying true binding

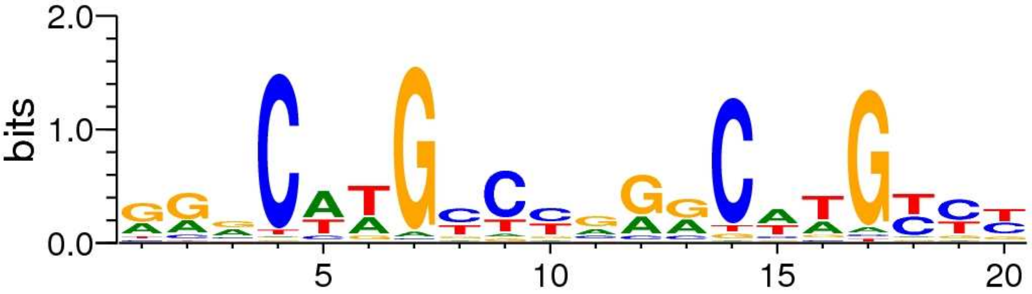


Figure 3.2: **The position-weight matrix (PWM) for p53 is visualized using a sequence logo, which denotes the frequency of binding at each base position of a TFBS. The height of each vertical stack represents the relative frequency at which that base appears at that position. PWMs are summarized as a consensus sequence, i.e. RRNCWWGBYYVRRCHWGYYB for P53.**

sites, as measured by ChIP-seq, for a number of factors. We make use of this conservation score to help remove false positive TFBS, using a conservation cut-off (see Section 2.1.3).

3.1.3 Phosphorylation site prediction from motifs

Phosphorylation sites are a second important regulatory mechanism in the cell. These sites are present on proteins and when activated can inhibit the normal function of that protein, or cause a protein to be recognized by the ATP-dependant ubiquitin-proteasome degradation pathway. Recent studies have shown that 90% of tumors contain phosphorylation related mutations [133], and these are observed at a higher frequency in cancer datasets compared to the background population [132]. Identifying these sites within protein sequences, and then mapping them to the genome is an important aspect in annotating the DNA-seq variants (see Section 2.1.3). We describe a recent paper published in Cellular Signalling in which we identify genuine phosphylation sites by a number of kinases using PWMs.

Ranking of putative phosphoproteins with MAPK docking-sites We started with a list of 394 human proteins with putative docking sites of the D-site class, identified by

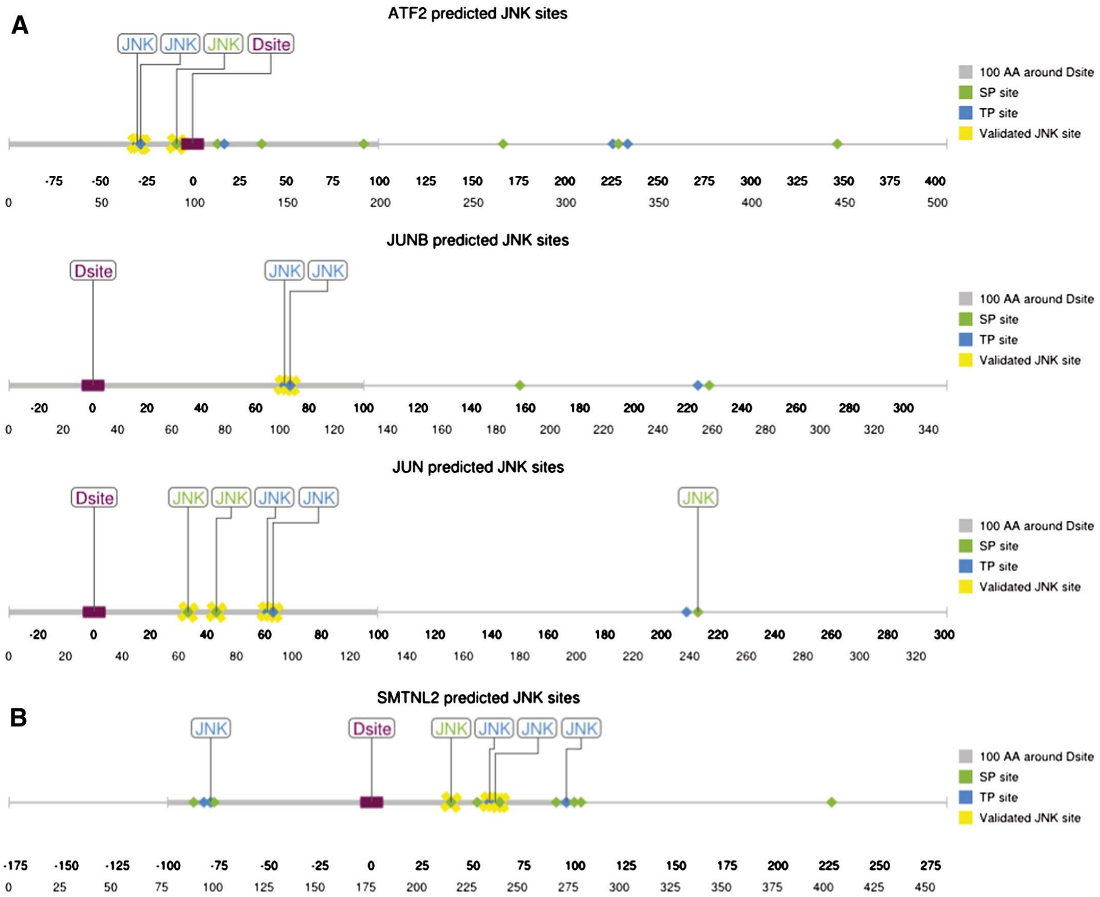


Figure 3.3: **A) Linear representation of known JNK targets with all potential ST/TP phosphosites shown. Those sites predicted by our PWM method labeled as JNK. Literature-validated phosphosites are indicated with yellow crosses. The D-finder-predicted D-site is shown as a purple rectangle; B) Linear representation of full-length SMTNL2 protein and all potential ST/TP phosphosites. Phosphosites identified in the mass spectrometry/mutagenesis experiments are indicated with yellow crosses.** Figure adapted from [63].

D-finder and published previously [178]. D-finder was developed using a training set of literature-verified D-sites found in JNK substrates and binding partners, and thus D-finder's predictions should be enriched for JNK substrates relative to ERK/p38 substrates.

Each protein on the list of D-finder predictions was scanned for SP and TP sites, as these constitute potential MAPK phosphosites. Proteins were scored based on the number of S/TP sites they contained within a 100 amino acid window on either side of their putative docking site. In this manner, 308 proteins were identified as potential substrates, meaning that they contained at least one potential phosphosite within 100 residues of the predicted docking site. Of these 308, 232 contained a cluster of 2 or more S/TP sites near the D-site.

To further analyze these 308 putative substrates for potential JNK phosphosites, we used a probability-weight matrix (PWM)-based approach to scan for S/TP sites that were in a local sequence context that indicated that they might be efficient JNK phosphosites.

Construction of position weight matrices for the identification of potential JNK phosphorylation sites We compiled known in vitro and in vivo JNK1/2/3 phosphosites, found in the current database of PhosphoSitePlus [75], and used these data to create position weight matrices (PWMs). PWMs were computed for JNK1, JNK2 and JNK3, for phosphoserine and phosphothreonine sites independently, thus resulting in a total of 6 PWMs. These PWMs take as input 15-residue peptide substrings centered on the SP or TP residue being evaluated. We calculated background frequencies for each amino acid from the set of every phosphoserine- and phosphothreonine-containing peptide within PhosphoSitePlus. This approach for determining background frequencies (as opposed to using global coding sequence frequencies, for example) should in principle reduce any bias towards over-scoring a peptide based on global properties (charge, surface accessibility, intrinsic disorder) that the set of all phosphoacceptor peptides may be enriched for.

These PWMs were then scanned over known JNK substrates, using LAsearch as implemented in PoSSuM 1.3 [16], which calculates a p-value for each S/T site based on the expected number of matches for all permutations of the PWM for a background sequence of the same length as the input (E-value). The p-value threshold used when predicting JNK phosphosites was tuned so that 80% of the input JNK1 and JNK2 peptides would be predicted correctly as JNK phosphosites. This threshold was chosen as it minimized the sum of false negatives plus false positives in known JNK substrates.

The output resulting from applying these PWMs to three known JNK targets (ATF2, JUNB, and JUN [21]) is shown in Figure 3.3A. For ATF2, the JNK PWMs correctly identified the known JNK phosphosites T69, T71 and S112, while also correctly rejecting the many other ST/TP sites found throughout the protein. For JUNB, the JNK PWMs correctly identified the known JNK phosphosites T102 and T104, while also correctly rejecting the 3 other ST/TP sites in the polypeptide. Finally, for the canonical JNK substrate JUN, the JNK PWMs displayed a similar degree of accuracy, although it did falsely predict a single SP site (this lone false positive, however, was not within 100 amino acids of the D-site). From these examples it can be concluded that the JNK PWMs we constructed are both sensitive and specific.

SMTNL2 is a predicted MAPK substrate The human SMTNL2 protein was found to contain 11 minimal putative MAPK phosphoacceptor sites (S/TP) within 100 amino acids of the D-site; furthermore, 5 of these were predicted to be likely JNK phosphosites by the JNK PWM approach (Figure 3.3B), 3 of which were subsequently validated through mass-spectrometry.

3.1.4 Alternative target recognition in the Wnt signaling pathway [76]

The Wnt signaling pathway is one of several vital developmental pathways conserved in all phyla of the animal kingdom ([74, 153]). In abnormal settings, such as when mutations in Wnt pathway components cause overactive signaling, gene expression patterns of proliferation are unbalanced. For example, early development of the majority (80%) of colon cancer cases are driven by overactive Wnt/β-catenin signaling ([120, 5]). The targets and gene programs that are misregulated in these cells are specified by the DNA binding specificities of the LEF/TCFs. The factor studied here, TCF1E, contains two DNA binding domains, a WRE binding domain and a C-clamp binding domain [77].

Although the C-clamp has been shown to interact with select GC-rich Helper sites in mammals [77] and Drosophila [20], its role in the genome-wide binding of C-clamp isoforms of TCFs is unknown. We have previously shown that the human C-clamp interacts with a short Helper site (5-RCCG-3) with an unusual degree of flexibility in that the site can be recognized on the 5 or 3 side of a Wnt Response Element with a tolerance for varied spacing between the elements [77]. We also determined that the C-clamp recognizes a 7 nucleotide extended Helper site (5-GCCGCCR-3), a motif first identified in Drosophila as occurring adjacent to WREs for dTCF/pangolin recognition [30]. We therefore searched for enrichment of a slightly shorter version of the extended Helper site (5-RCCGCC-3) in our ChIP-seq peaks because the C-clamp is highly conserved between humans and Drosophila and because the short 4bp Helper site occurs too frequently for meaningful searches.

ChIP-seq to profile binding of wildtype and mutant TCF1 We performed ChIP-seq experiments to determine the binding profile of a wildtype and C-clamp mutant version of TCF1 in DLD-1 colon cancer cells. Our results indicate that the C-clamp-Helper site interaction widely contributes to binding site selection and strength.

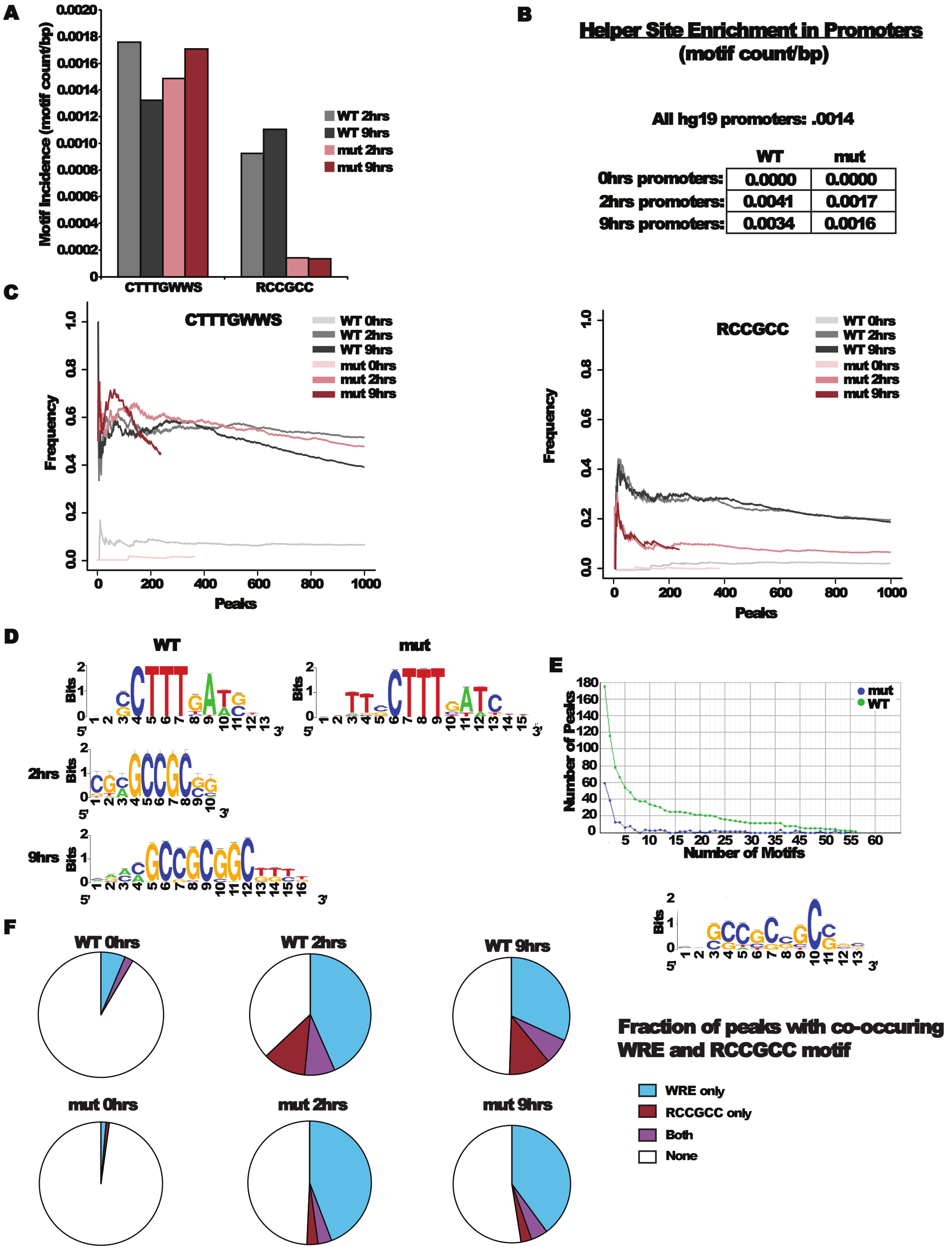
In agreement with previous cap analysis gene expression (CAGE) results (not shown), a significant difference in binding of the helper site consensus sequence was observed (RCCGCC; Figure 3.4A). The frequency of RCCG within wildtype peaks is much higher than observed in human genome promoters (Figure 3.4B).

We find that wildtype TCF1 bound to a greater number of gene bodies and promoters compared to mutant TCF1 (Figure 2.4B), suggesting that the C-clamp helps position TCF1 for transcriptional regulation.

To account for the issue that promoters and gene bodies tend to be more GC-rich than intergenic regions, we evaluated the occurrence of the Helper site in peaks within promoter regions (defined as -1kb to +100bp from the TSS) for both wildtype and mutant as compared with the frequency of the Helper site in all human RNA polymerase II promoters. We used the incidence of (5-RCCGCC-3) per base pair to determine if the Helper site occurrence is significantly higher than expected. Despite a higher background GC content within promoters, the incidence of RCCGCC per bp is still significantly greater within promoters bound by dnTCF1EWT (Figure 3.4B) (Mann-Whitney-Wilcoxon test; p=2.3E-12). For comparison, the incidence of the Helper site in promoters occupied by dnTCF1Emut is not significantly greater than the genome background promoters (p¿0.05; Figure 3.4B).

Using frequency plots for each factor, we readily observe a different once in RCCGCC frequency among ChIP-seq peaks in wildtype and mutant (Figure 3.4C), in addition to localization of these sites near the peak centers as shown in Figure 3.5. We also observed a greater percentage of dnTCF1EWT ChIP-seq peaks that have at least one occurrence of the Helper site compared to dnTCF1Emut, whereas the percentage of peaks containing a WRE was similar between the two. (Figure 3.4C).

Our bioinformatics analysis also indicates that the C-clamp can interact with a Helper site



Helper Site Enrichment in Promoters
(motif count/bp)

All hg19 promoters: .0014

	WT	mut
0hrs promoters:	0.0000	0.0000
2hrs promoters:	0.0041	0.0017
9hrs promoters:	0.0034	0.0016

Fraction of peaks with co-occuring WRE and RCCGCC motif

Figure 3.4: **Differences in C-clamp helper site motif occurrences. A. Occurrence of helper site, but not WRE, is found to be lower in the mutant. B. Helper site frequencies are normal in Mutant TCF1 peaks compared to all human promoters, but is 2-3 times more frequent in wildtype peaks. C. Frequency plots showing a lack of RCCGCC in mutant TCF1 peaks. D. Motifs identified in peaks. E. Histogram of distribution of multiple helper site motifs. F. Fraction of WRE and helper site motifs in peaks.**
Figure adapted from [76] (submitted).

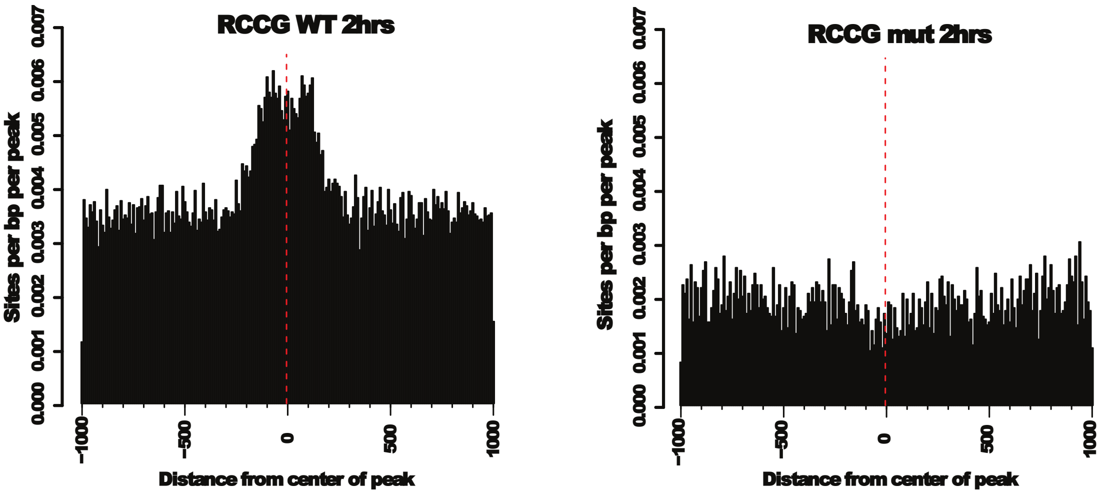


Figure 3.5: **Histograms of the number of sites observed per base position away from the center of all peaks in wildtype vs. mutant peaks. RCCG sequence is located at a higher frequency near center of peaks only in wildtype.**
Figure adapted from [76] (submitted).

independently of a concurrent HMG-WRE interaction. Surprisingly, we find that these sites can contribute to TCF1EWT-mediated transcriptional regulation, suggesting that the C-clamp-Helper site interaction can contribute to transcription without a neighboring HMG-WRE interaction. These sites have been validated using mutagenesis analysis for a few examples.

Changes in gene expression We also assessed early changes in gene expression in tandem with ChIP-seq experiments using a metabolic labeling and high throughput RNA-seq technique called 4Thiouridine-seq. This technique selectively labels actively transcribed nascent RNAs, and by comparing 4Thiouridine-seq transcript changes with ChIP-seq data, we identified new C-clamp dependent Wnt target genes that showed direct regulation by TCF1, including histone genes, and other genes with high relevance to cell proliferation. We conclude that the C-clamp can work independent of WRE interactions in modulating the DNA activities and transcriptional output of TCF1E. However, our results also indicate that the DNA binding specificities of the HMG box and C-clamp synergize to control gene

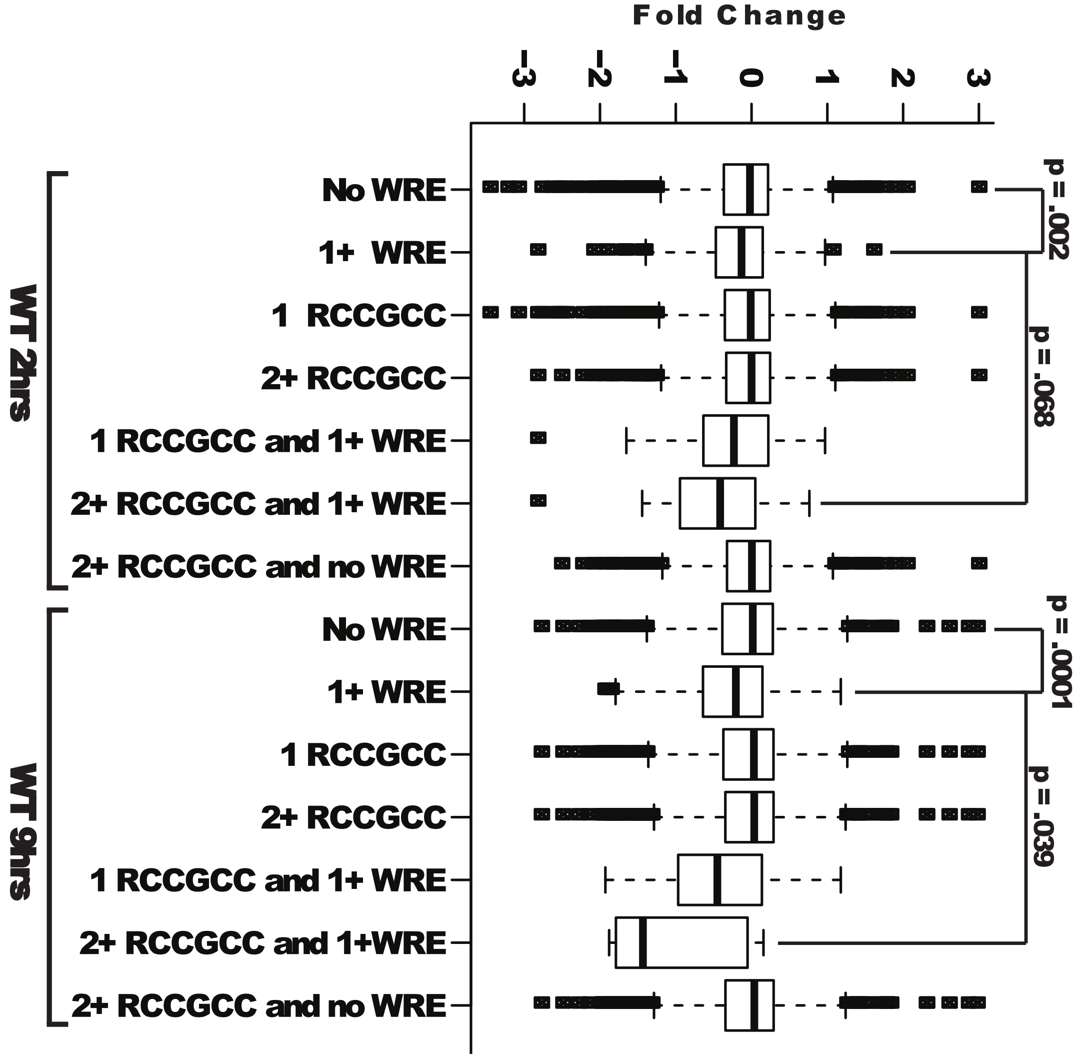


Figure 3.6: **ChIP-seq peaks for WT 2 and 9 hours are combined with gene expression as measured through 4'Thiouridine-seq. Subsets of peaks containing both a WRE and a C-clamp helper site have more significant changes in gene expression of transcripts within 30kb of the peak than those with neither or only one site. Multiple RCCGCC sites appear to magnify this change in expression.**
Figure adapted from [76] (submitted).

expression of a subset of Wnt target genes (See Figure 3.6).

3.1.5 Enrichment analysis

Given a set of genes, possibly obtained by the differential analysis outlined in the next section, a common question is what transcription factors are responsible for regulating a subset or all of these genes? An unfortunate fact is that the gene responsible for regulating a set of differentially expressed genes is not always identified as changing itself, as is the case for IL-12 in the transcriptional response to UV light [118].

To identify which transcription factors may be regulating a set of genes, we use a Fisher's Exact test to determine significance of the number of binding sites within the list, as compared to the 36742 transcripts annotated in the human genome, and subsequently rank transcription factors by this p-value. Fisher's exact test uses the formula below, along with a contingency table as shown in Table 3.1. Using the TFBS identified by MotifMap (see Section 2.1.3), and for a specific set of distances (usually 3kb or 10kb from the transcription start site of transcripts), we identify all bases containing a TFBS for each transcription factor both within the transcriptome and within the transcripts of our gene list. This analysis can be extended to co-localization of >1 TF using 2x4 contingency tables, whose probabilities can be calculated using Monte Carlo Methods [112].

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

This analysis is incorporated into our pipeline as part of a standard gene list enrichment analysis step using the same framework as in Section 5.1 to present enrichment results.

Table 3.1: **Contingency table used for Fisher's exact test calculation for identifying enriched TFBS. We assume each base pair sequence in the genome has a background probability of being a TFBS and are independent**

Bases in promoters	Gene set	Genomewide
Is a TFBS	a	b
Is not a TFBS	c	d

Additionally, all analysis is performed using a server-client interface which keeps in memory a database of all TFBS within a fixed distance and cutoffs from all transcripts in the human genome, for returning enrichment results in less than a minute for any sized list. This step additionally computes DAVID GO term enrichment [44], GO terms listed per gene, GSEA [158] for gene lists of interest, and the TFBS enrichment described here. Additionally, this step reconstructs the original network with PPI and MotifMap edges along with the enriched network, listing all identified transcription factors in the initial list using a network approach shown in the section below, as well as described further in Section 4.1.

3.1.6 Defining the transcriptional regulatory network of skin wounding

As an example of the application of the TFBS enrichment outside of the context of pediatric cancers, we identify the transcriptional regulators of the human response to skin wounding. We analyzed RNA-seq data obtained by collaborators and their colleges at the UCI Medical Center for skin samples extracted pre- and post-surgery, in order to investigate the human skin wounding response in human tissue. Differential transcripts were identified (496 up- and 412 down-regulated) using the methods described in Section 3.2.4. These genes were found to be related to wounding response through enriched GO terms found with DAVID [78] as part of the standard analysis pipeline. The list of enriched TFBS within this list are shown in Table 3.2, among which NFkB, STAT4, SRF, CREB, PURA had enrichment for binding sites, but were not found to be differentially expressed. ETS2 was enriched in addition to

Table 3.2: Enriched transcription factors in skin wounding

Factor	**TFBS in**		**Bases in**		**Motif**			
	Targets	**Genome**	**Targets**	**Genome**	**Uniprot**	**Accession**	**Log-odds**	**p-Value**
SRF	107	612	3376844	147004742	P11831	M00152	7.6125E+00	3.8771E-53
NF-kappaB	121	1593	3376844	147004742	P19838	M00774	3.3066E+00	2.9822E-27
HNF4	1070	35844	3376844	147004742	P41235	M00134	1.2995E+00	5.1235E-16
RELA	42	416	3376844	147004742	NA	MA0107	4.3952E+00	3.3643E-14
c-Rel	49	669	3376844	147004742	Q04864	M00053	3.1885E+00	1.6288E-11
AP-1	174	4480	3376844	147004742	P01100	M00517	1.6908E+00	2.6488E-10
REL	44	633	3376844	147004742	NA	MA0101	3.0260E+00	7.6351E-10
CREB	216	6006	3376844	147004742	P16220	M00916	1.5656E+00	1.3625E-09
NF-kappaB(p65)	28	339	3376844	147004742	NA	NA	3.5955E+00	3.1232E-08
MTF-1	7	19	3376844	147004742	Q14872	M00650	1.6042E+01	1.3003E-06
RelB:p52(NF-kappaB)	27	389	3376844	147004742	NA	NA	3.0215E+00	1.3076E-06
AR	25	347	3376844	147004742	P10275	M01201	3.1363E+00	1.7020E-06
TEF-1	1113	41789	3376844	147004742	P28347	M00704	1.1595E+00	1.9339E-06
ATF	90	2279	3376844	147004742	NA	M00017	1.7192E+00	2.8737E-06
NeuroD	1826	71172	3376844	147004742	Q13562	M01288	1.1169E+00	4.5150E-06
STAT1	37	706	3376844	147004742	P42224	M00224	2.2815E+00	1.0032E-05
STAT5A(homodimer)	22	333	3376844	147004742	NA	NA	2.8760E+00	2.4072E-05
MAFA	394	13789	3376844	147004742	Q8NHW3	M01709	1.2439E+00	3.4531E-05
SOX10	412	14495	3376844	147004742	P56693	M01131	1.2374E+00	3.7450E-05
ETS2	726	27080	3376844	147004742	P15036	M01207	1.1671E+00	6.0842E-05
OCT1	48	1115	3376844	147004742	P14859	M00342	1.8741E+00	9.3469E-05
ATF4	9	80	3376844	147004742	P18848	M00514	4.8969E+00	1.8285E-04
STAT5B(homodimer)	13	187	3376844	147004742	NA	NA	3.0263E+00	6.6460E-04
Bach1	14	230	3376844	147004742	O14867	M00495	2.6498E+00	1.4096E-03
TCF-4	62	1765	3376844	147004742	Q9NQB0	M00671	1.5292E+00	1.9624E-03
NF-kappaB(p50)	26	571	3376844	147004742	NA	NA	1.9822E+00	1.9781E-03
Bach2	20	400	3376844	147004742	Q9BYV9	M00490	2.1767E+00	2.3502E-03
SOX9	18	357	3376844	147004742	P48436	M00410	2.1950E+00	2.6158E-03
LEF1	137	4578	3376844	147004742	Q9UJU2	M01022	1.3028E+00	3.1424E-03
TATA	3	12	3376844	147004742	P20226	M00252	1.0884E+01	4.2074E-03
ATF3	26	619	3376844	147004742	P18847	M00513	1.8284E+00	4.7724E-03

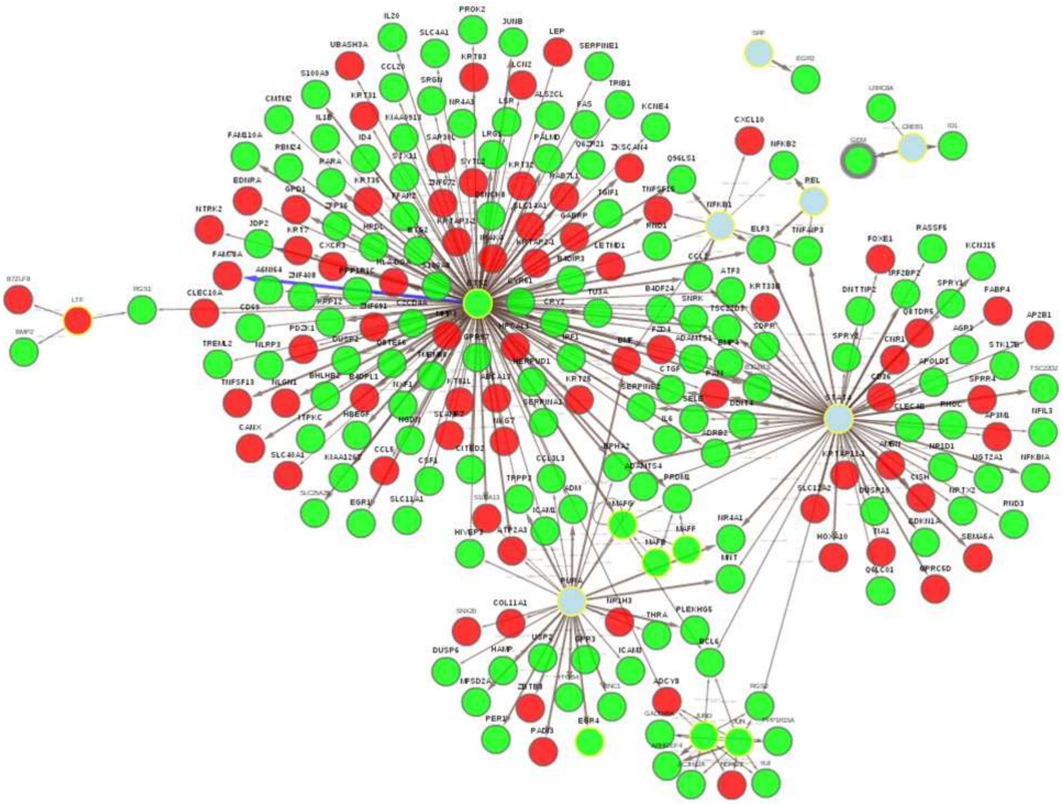


Figure 3.7: **Network reconstruction using only differentially expressed transcripts identified in the RNA-seq experiment along with the enriched transcription factors identified using the TFBS enrichment analysis pipeline step. Significant genes are shown in red or green. Only a handful of enriched factors are differentially expressed themselves.**

being differentially expressed pre- and post-wounding and having the most edges between nodes in the network representation show in Figure 3.7. It is currently being investigated for its role in skin differentiation and the wounding repair pathway.

3.2 Differential expression analysis

Differential expression analysis steps aim to identify significant differences in gene expression between two, or more, conditions. In order to do so, a reliable measure of gene expression in a sample is required. Currently, two methods are employed to analyze gene expression:

microarray and RNA-seq. The analysis steps overlap between the two, although significant differences exist between both methods, leaning in favor of RNA-seq in almost all cases [170].

3.2.1 Microarray differential analysis

We need to process a number of microarray datasets in the pediatric cancers project (see Section 4.2.2, which makes use of this step). Despite the rapid adoption of RNA-seq over microarray [7], microarrays are still regularly performed and in some cases are the only form of expression data available for older experiments. Microarrays measure the strength of binding between two complementary strands of DNA, or even between protein and DNA, using florescent probes (for an example, see Figure 3.8).

We process microarray data using the following steps:

- Background normalization
- Mean-variance bias normalization
- Quantile normalization
- Determining significant transcripts
- Multiple-test correction

Background normalization Different probes on an array have different binding affinities to their complementary DNA sequences. To measure this non-specific binding, mismatch probes are typically used to assess this error. One popular algorithm, MAS5 [127], is used to normalize each probe in an Affymetrix array using these mismatch probes. A second popular algorithm, RMA [81], does not make use of these spots but is not used in this pipeline. The RMA method additionally performs quantile normalization, which we do perform subsequently.

Additionally, every spot has a background intensity around the individual spots due to the fluorescence markers used in their quantification. This background intensity is usually subtracted from the spot intensity average, and an additional step that is performed by MAS5.

Mean-variance bias normalization Microarray data is known to have a significant mean-variance bias, meaning that as the mean intensity of a probe increases, so does the variance. This is intuitive, because as intensity values approach 0, so does the variance. A similar problem is faced when attempting to apply a t-test to any correlation values (bounded by 0 and 1) which lead to Fisher's z' transformation for correlations [55]. Typically, this variance in microarrays is adjusted for using a log transformation on the expression values after background normalization. Alternatively, the R package (vsn) for variance stabilizing normalization in microarray data are used [79, 50] (http://Bioconductor.org/). This method performs a variance correction that is log for large expression values but becomes linear as expression values approach 0, where the log transformation breaks down. In addition to the variance correction, using known non-differentially expressed spots this method calculates maximum likelihood shifting and scaling calibration parameters for different arrays. This type calibration has been shown to minimize experimental effects [95].

Quantile normalization Generally, it is desirable that the distributions of expression profiles across multiple samples is similar. In most experiments, the majority of genes are not expected to be changing, so this is reasonable. In order to fix what are likely sample-specific biases, for example, overall lower expression in one sample compared to others, quantile normalization is employed [22]. This is used in a number of other methods, such as RMA [81]. When microarrays are of wildly different microarray platforms, such as those between Affymetrix HG-U133plus vs. HG-U133A, and possibly even between Affymetrix microarray intensities and Illumina BeadArray, one can use quantile normalization as a way to compensate for different technology biases as is done in the case of the many pediatric datasets obtained for the pediatric cancer project (see Section 4.2.2).

Determine significance of differential expression To identify significant differences between two measurements, a t-test is usually employed. In our pipeline, differential analysis is done using the Cyber-T software described in [13] and available online at http://cybert.ics.uci.edu. Cyber-T calculates a Bayesian regularized estimate of the variance of the signal intensity levels. Then theses variance estimates are used to compare the groups with a t-test. Cyber-T has been shown to perform well under low-replicate conditions.

Multiple test correction When test many hypotheses at the same time (i.e. testing for differential expression in each gene of the human transcriptome), and using a standard p-value cutoff of 0.05 for each test, one will be left with approximately 1500 incorrect tests due to the 5% error rate per test. In order to obtain an overall error rate of 5%, i.e. a false discovery rate (FDR) of 5%, a stricter p-value threshold needs to be used for each test. At the limit, if a 5% error rate is desired, then a p-value cutoff of 0.05/(number of tests) needs to be used, or similarly, we can multiply our p-values by the number of tests to correct our p-values and allow use of the original 0.05 p-value cutoff. This approach is known as Bonferroni correction [174], and is considered quite conservative. An alternative, and the one that we employ regularly is Benjamini-Hochberg correction [19], which uses the Benjamini-Hochberg step-up method to adjust p-values such that choosing an adjusted p-value cutoff of 0.05 controls the false discovery rate by rejecting the null hypothesis of all tests below the first test that has an uncorrected p-value less than it's percentile-rank times 0.05.

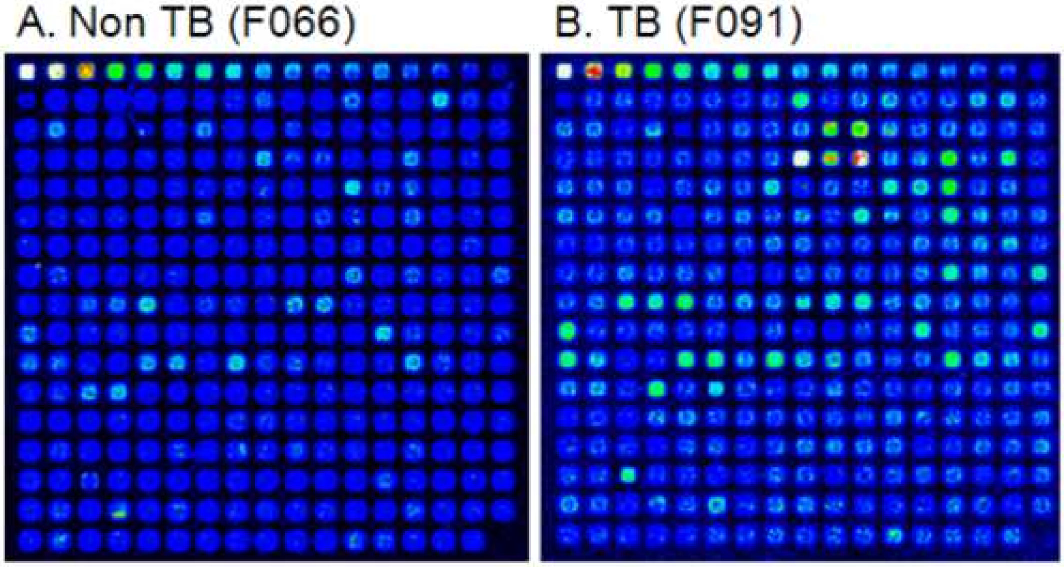


Figure 3.8: **Example microarray image for two samples: one obtained from a non-tuberculosis infected patient and another from an infected patient from Colombia. Intensity of florescence quantifies the reactivity of that patient's blood to the protein within each spot of the array, and differential analysis identifies the most reactive proteins.**

3.2.2 Identifying differential antigens

Besides identifying differential transcripts in cancer microarray datasets, as outlined in Section 4.2.2, we have processed numerous other microarray datasets, including protein microarray datasets, using a very similar pipeline. These protein microarrays were developed by the Felgner Lab [40] to profile immune response by exposing blood serum to translated proteins in each spot of the array. Spots are printed onto an ELISA [53], which is used to measure the reactivity between antibodies in the sera to proteins on the array.

The *M. tuberculosis* dataset consists of sera samples collected from Colombia where the exposed group tests positive using both sputum smear and bacterial culture tests for *M. tuberculosis* and the naive control group is negative for both tests. Additionally, all sera are HIV negative in order to minimize the confounding relationship between HIV positive sera and the antibody response to the *M. tuberculosis* antigens.

We applied our standard analysis pipeline to these datasets to obtain a ranking of antigens based on their significance in infected vs. uninfected patients (Figure 3.9). We use the same methods used to previously analyze the immune response to *B. henselae* infection [167]. Multiple antigen classifiers were built using Support Vector Machines (SVMs) [12, 35]. The e1071 and ROCR packages in R were utilized to train the SVMs and to produce receiver operating characteristic curves, respectively, and we observed AUC values of 82% after only including the top 10 differential antigens. Afterwards, performance values leveled off and there was no advantage to including any more than 10 antigens in our predictor.

This analysis was further extended to not only the prediction of patient infection status, but to the prediction of protein antigenicity using sequence-level features and training an ensemble method on the significant antigens identified for *M. tuberculosis* and four other infections [106] (Figure 3.10).

3.2.3 RNA-seq differential analysis

The underlying distribution of the expression values for RNA-seq differ significantly from that of microarrays, but much of the same analysis still applies. Background normalization is no longer needed, but the distribution of read counts differs from that of microarray intensity values. While microarray intensity values are drawn from a normal distribution, arising from measurement error, counts of RNA-seq reads aligning within a transcript have a poisson distribution [154], which has variance equal to mean. To allow for variance, a negative binomial model is commonly used, which assumes RNA-seq measurement error arises from a beta distribution.

Packages designed to handle perform differential analysis on read count data assume this

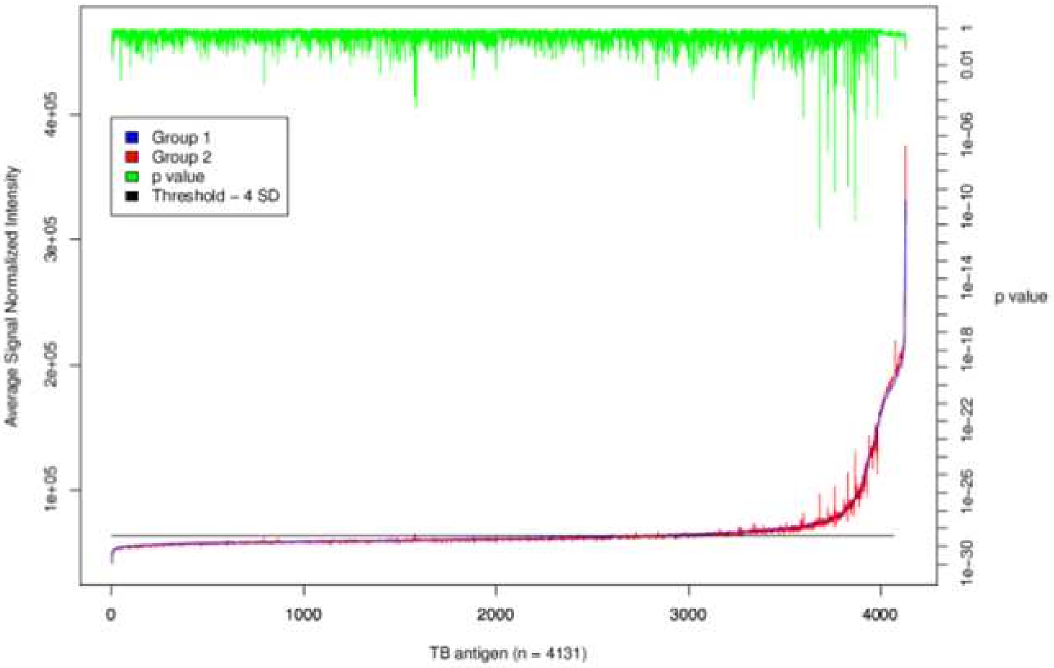


Figure 3.9: **Identifying differential antigens through the use of a standard microarray analysis pipeline. Significance of differences is plotted along the upper x-axis as -log_{10}(p-values). Comparison with average signal intensity from the array identifies a handful of antigens with significant differences between infected and uninfected patients.**

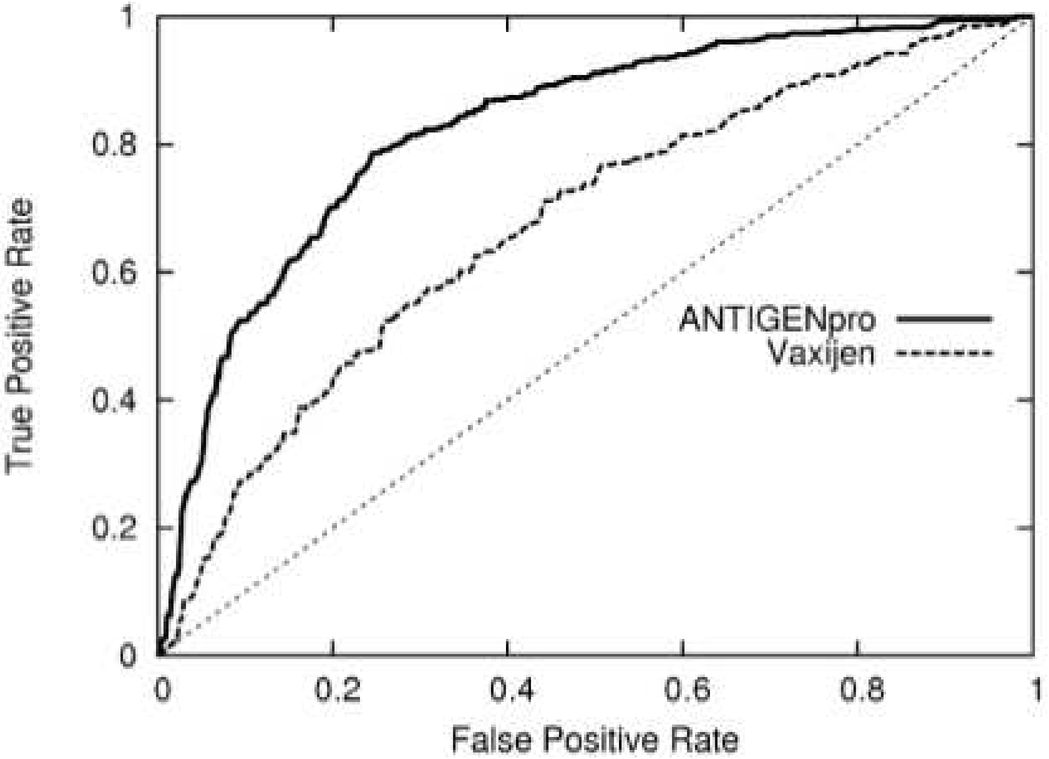


Figure 3.10: **ROC curves calculated from the 10-fold cross-validation of ANTIGENpro and Vaxijen [48].**
Figure adapted from [106].

negative binomial model (DEseq [9], EdgeR [136], cuffdiff [161]). On the other hand, RPKM-normalized [121] values do not fit the poisson model, coming from a real valued distribution. By using RPKM values, we can still make use of a bayesian t-test [85] or non-parametric tests, as was performed in the microarray differential expression analysis portion of the pipeline (Section 3.2.1). As CyberT has been shown to work well across a number of different datasets, and particularly well for low replicate datasets, we make use of it extensively in our RNA-seq differential analysis portion of the pipeline.

3.2.4 RNA-seq differential analysis in pediatric patients

In the absence of tissue-matched control RNA-seq samples for each patient – which in many cases is not feasible to obtain – each patient's RPKM values are compared to a pooled sample created by combining the other patients' RPKM values. We perform a pooled analysis, similar to what is done in [175, 103]. Differential analysis of RPKM-normalized read counts is performed using CyberT [13] which was recently upgraded to handle both DNA microarray and RNA-seq data [85]. A confidence in the Bayesian prior of 3 is used instead of the default of 10 within CyberT to estimate the variance in gene expression. Rather than use strict p-value cutoffs, the top 5% most significantly over- or under-expressed genes, as well as the top 5% least significantly changing genes, are retained for down-stream analysis. The sizes of each of these gene lists are summarized under the gene expression column in Table 2.2. In Figure 4.5, we show the RPKM values obtained for each patient for the DCC transcript, using the analysis steps in Section 2.2. Patient CHOC08 can be seen to have a much higher expression of this transcript than other patients, despite a relatively flat expression profile across various tissue types for which tumors were obtained, and is highlighted in our differential analysis as having a very significant p-value (p=2.49E-11; top 1%).

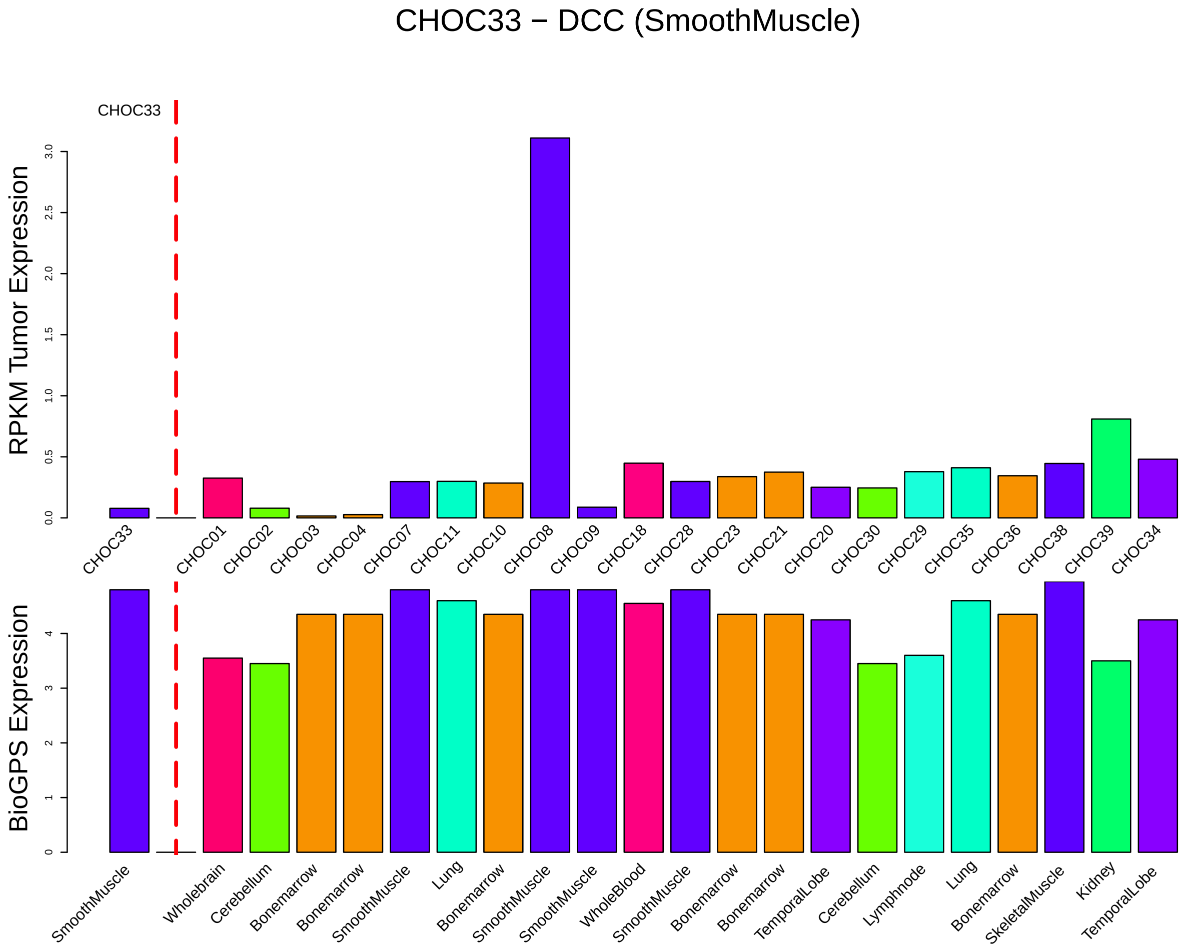


Figure 3.11: **Expression for DCC as measured by RPKM is significantly higher than other patients (p=2.49E-11; upper), yet with no detected DNA variants present within DCC coding sequence. BioGPS GCRMA expression values are relatively constant across tissue types for which tumor samples were obtained (lower) [188].**

Variant transcription factors Transcription factors have been shown to have a large role in tumor progression, as evidenced by a large number of transcription factors that are known tumor suppressors. We identify potentially important affected transcription factors by making use of the predicted TFBS described in Section 2.1.3. For each transcription factor, we determine the number of binding sites predicted within 3kb upstream and 1kb downstream of the transcription start site (TSS) of all transcripts in the human genome. We compare these counts to those within the same distance to genes in each of the following three lists:

1. The top 5% under-expressed genes in the patient vs. other patient RNA-seq differential analysis
2. The top 5% over-expressed genes in the patient vs. other patient RNA-seq differential analysis
3. The top 5% differential genes in the control vs. tumor microarray data obtained for this patient's cancer type, as described in Section 4.2.2.

We use a Fisher's Exact test to determine significance of the number of binding sites within the above lists, as compared to the 36742 transcripts annotated in the human genome, and subsequently rank transcription factors by p-value (see Section 3.1.5). For each enriched transcription factor with p-value less than 0.05, we determine if the protein for that transcription factor is affected by any small or large variations or has abnormal gene expression for that patient. This results in lists of approximate 0-20 variant transcription factors per patient. In conjunction with the expression of the putative targets of these factors, we can identify what are likely causal relationships between over- or under-expression of certain factors and subsequent over- or under-expression of their targets.

3.2.5 The neuron-specific chromatin regulatory subunit BAF53b is necessary for synaptic plasticity and memory [168]

Recent exome sequencing studies have implicated polymorphic Brg1-Associated Factor (BAF) complexes (mammalian SWI/SNF chromatin remodeling complexes) in several human intellectual disabilities and cognitive disorders. However, it is currently unknown how mutations in BAF complexes result in impaired cognitive function. Mice harboring selective genetic manipulations of BAF53b have severe defects in long-term memory and long-lasting forms of hippocampal synaptic plasticity. These findings suggest that the BAF nucleosome remodeling complex regulates gene expression required for proper neuronal function, synaptic plasticity and memory processes.

Changes in gene expression during memory consolidation Thus, we set out to identify which genes BAF53b regulates during memory consolidation. We performed an RNA sequencing experiment using dorsal hippocampal tissue from four groups of animals: Baf53b+/- mice taken directly from their home cage without training, Baf53b+/- mice taken 30 min after OLM training, wild-type mice taken directly from their home cage without training and wild-type mice taken 30 min after OLM training. We have previously observed substantial gene expression changes in the dorsal hippocampus 30 min after OLM training [14]. Mean PHRED quality scores indicate high-quality sequencing data for each replicate (not shown).

We first compared the expression profiles of the wild-type and Baf53b+/- mice home cage groups and found that the majority of genes (19,524) were equivalently expressed at baseline in the two groups (Figure 3.12A). There were also groups of genes that showed increased expression (80) in wild-type compared with Baf53b+/- mice and vice versa (57) at home cage. We next examined differences in gene expression following training in the wild-type mice. Consistent with numerous studies [8], wild-type mice showed robust changes in gene expression, including many immediate early genes (IEGs), following OLM training (compared with home cage; Figure 3.12B), indicating that the training period was sufficient to induce activity-dependent gene expression during memory consolidation. In addition, many of the activity-regulated genes (124) increased in the wild type were also significantly induced in Baf53b+/- mice following training ($P<0.05$). These genes were enriched for Gene Ontology [10] terms for regulation of transcription, RNA processing and intracellular signaling, and

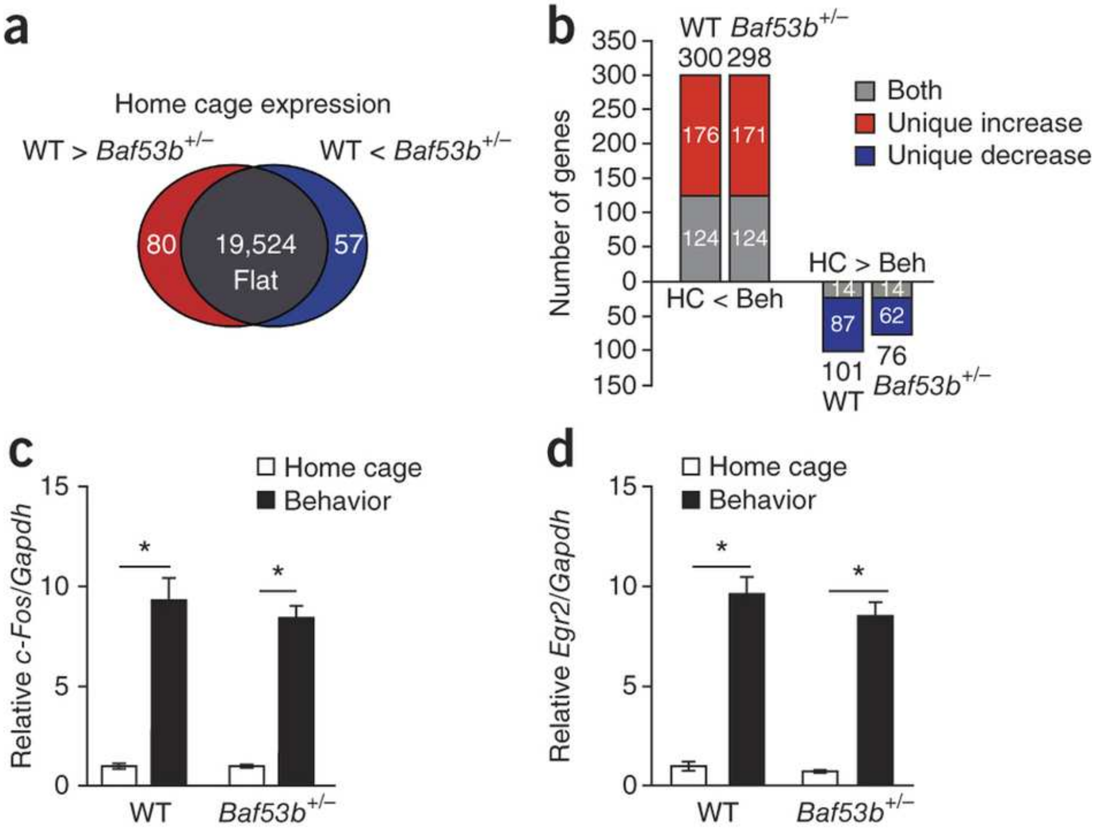


Figure 3.12: **(a) Gene expression diagram for wild-type compared with Baf53b+/- mice taken directly from the home cage. (b) Gene expression for genes that increased or decreased expression following behavior (30 min post training) compared with home cage. Genes with differential expression at home cage were removed before analysis. 'Both' indicates genes regulated similarly in wild-type and Baf53b+/ mice. 'Unique Increase' comprises genes that increased in only the indicated genotype. 'Unique Decrease' comprises genes that decreased in only the indicated genotype. Groups: Baf53b+/ mice taken from home cage (HC, n = 6), Baf53b+/ mice taken 30 min post training (Beh, n = 6), wild-type taken from home cage (n = 6) and wild-type taken 30 min post training (n = 6). Total gene counts for each genotype given above or below each column. (c) qRT-PCR validation of the IEG c-Fos (ANOVA; main effect of behavior, F1,20 = 157.6, P <0.0001; no effect of genotype, F1,20 = 0.49, P = 0.49; no interaction, F1,20 = 0.45, P = 0.51). Expression is shown relative to Gapdh and normalized to wild-type mice taken from home cage. (d) qRT-PCR validation of the IEG Egr2 (ANOVA; main effect of behavior, F1,20 = 224.2, P ¡ 0.0001; no effect of genotype, F1,20 = 1.53, P = 0.23; no interaction, F1,20 = 0.55, P = 0.47). Expression is shown relative to Gapdh and normalized to wild-type mice taken from home cage. *P <0.05, n values refer to number of mice.**

Figure adapted from [168].

included the majority of IEGs (Figure 3.12C,D). This suggests that BAF53b and nucleosome remodeling do not affect IEG expression during memory consolidation and that the long-term memory impairments observed in Baf53b+/- mice are caused by different mechanisms. Of the 300 genes with increased expression in the wild type following OLM training, the expression of 176 failed to significantly increase in the Baf53b+/- mice ($P>0.05$; Figure 3.12B). These genes were enriched for Gene Ontology terms involving transcription regulation and neurogenesis, as well as chromosome organization and chromatin modification, indicating a potential role for BAF53b in organizing higher order chromatin structure. In the Baf53b+/- mice, there were also a group of genes (171) that were induced following behavior that were not normally increased in the wild type (Figure 3.12B). These genes were enriched for Gene Ontology terms involving regulation of cell death, glutamate release, behavioral response to drugs of abuse, synaptic transmission and regulation of neurogenesis.

In addition to increases in gene expression, there were 101 genes whose expression decreased in the wild type following OLM training compared with home cage (Figure 3.12B) and 76 genes that decreased in the Baf53b+/- mice. Of the 101 genes that decreased in the wild type, 14 also decreased in the Baf53b+/- mice and 87 that did not. The 87 genes that failed to show an activity-dependent decrease in expression in the Baf53b+/- mice were enriched for Gene Ontology terms involving cell homeostasis, postsynaptic cell membrane and cytoskeleton. In addition to the impaired decrease in gene expression, the Baf53b+/- mice also had 62 genes whose decrease in expression following behavior was not matched in the wild type. These aberrantly decreased genes were enriched for Gene Ontology terms involving mitochondria function.

To further explore the link between the impairments in maintenance of long-term potentiation and cofilin phosphorylation, we examined gene expression for Gene Ontology terms involved in actin cytoskeleton and the postsynaptic density. Most genes examined showed similar expression between the Baf53b+/- mice and their wild-type littermates. However, there were several genes that showed misregulation either at baseline (home cage) or following OLM training that are involved in regulating the Rac-PAK and RhoA-LIMK pathways that both culminate in phosphorylation of cofilin and actin cytoskeleton reorganization [134]. For example, mir132 has been shown to regulate spine plasticity ([173, 80]) and long-term OLM41 by regulating Rac1 activity through translational repression of p250GAP [173]. Additional regulators of this pathway were also disrupted in the Baf53b+/- mice, including Citron (Rho interacting kinase) and Fhl2 (a member of the four-and-a-half LIM only protein family implicated in linking signaling pathways to transcriptional regulation). Components upstream of the Rac-PAK and RhoA-LIMK were also altered in the Baf53b+/- mice, including the NMDA receptor subunits Grin2b and Grin2a, as well as Efna4.

3.3 Hybrid approach

Additionally, we explored the possibility of using microarray data as a control for RNA-seq datasets, in the absence of RNA-seq controls, as was the case for our pediatrics patients. To do so, we obtained the only two datasets available to us with both microarray and RNA-seq for the same samples. One is in yeast [124] and another in rat [157], each containing 3 and 4 control/experiment samples each.

We obtained the microarray samples and RNA-seq samples for both papers from GEO and the short read archive (SRA), respectively. We aligned FASTQ files for the RNA-Seq data to the respective reference genomes using Tophat [88] and cufflinks [162] with annotation GTFs from UCSC, to derive log_2(RPKM) for annotated transcripts. We processed the microarray CEL files using MAS5 [127], quantile normalization, and log_2 transformation, to derive equivalent log_2(expression) values.

Next, we used Affymetrix probe annotations to map to SGD and RGD gene IDs, and paired

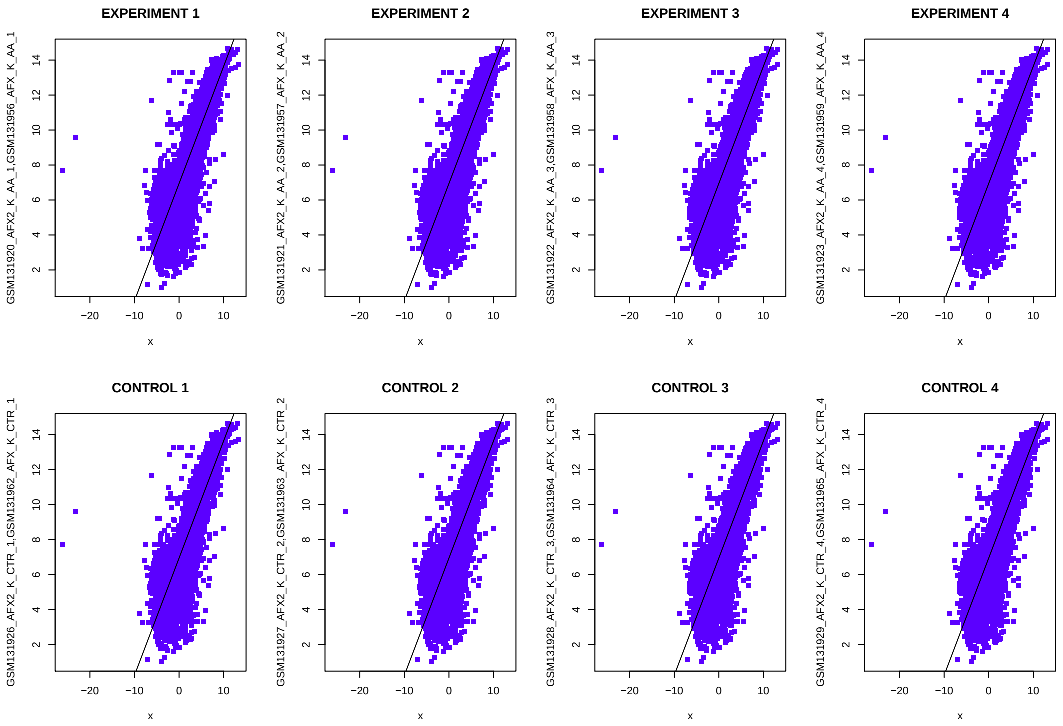


Figure 3.13: **Regression of rat RNA-seq data with microarray data for the same samples. log_2 normalized microarray expression is plotted vs. log_2 normalized RPKM values and a robust linear model is fit to the data. RNA-seq and microarray agree, except for lowly expressed transcripts.**

up probes with transcripts, taking the max probe log_2(expression) value across multiple probes per transcript. Mean expression was computed for each set of experiment/control biological replicates, to be used in the next step (see Figure 3.13).

We then linearly transformed the log_2(RPKM) data to log_2(expression) using robust linear models, rlm in R [82] (with init="lts" and method="MM"), on the mean across transcripts for our two groups being compared. Lastly, we quantile normalized the two groups being compared, after transformation. To identify differential transcript, CyberT [13] was ran with a confidence of 3, window size of 101, and a p-value cutoff of 0.05, and found that while more differential genes were called (Figure 3.14), it appears possible to combine RNA-seq data with GEO data reusing the original fit. It is unclear if this trend persists across species, and in the absence of human samples with both microarray and RNA-seq data available,

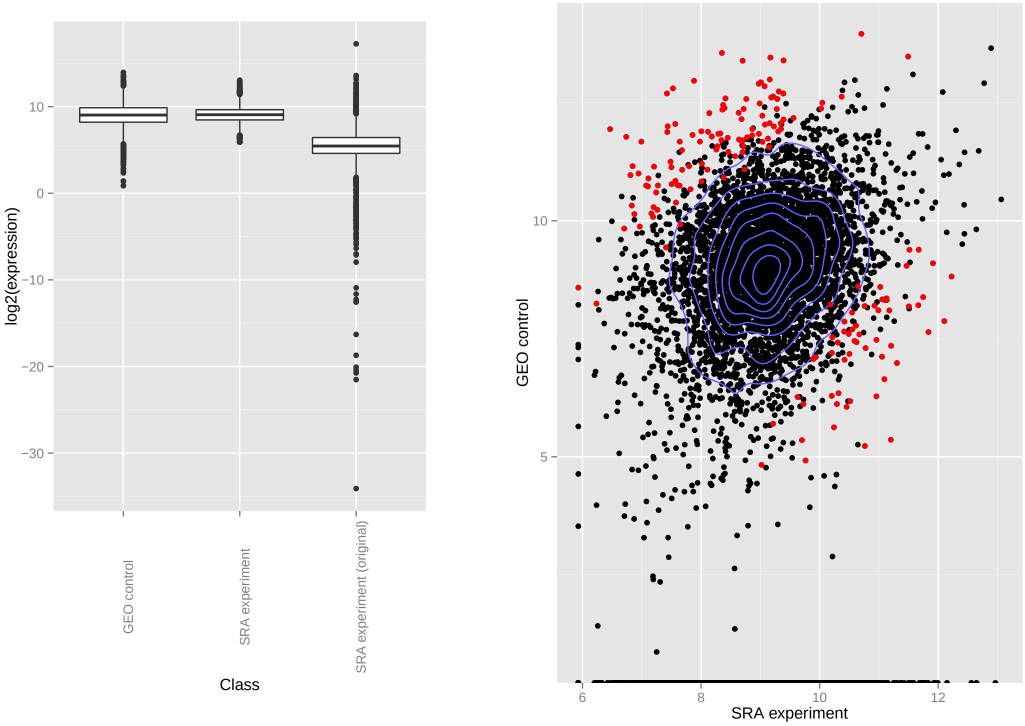


Figure 3.14: **Yeast RNA-seq experimental condition data obtained from SRA is transformed in order to be compared against GEO control data in order to perform differential analysis (left). Differential genes are shown in red (right), where more genes are significantly differential than in the GEO control vs. experiment, representing a higher false positive rate.**

this hybrid approach was not incorporated into the final genomic analysis pipeline for the pediatric cancers project.

Chapter 4

Integrative analysis

4.1 Network analysis

One of the fundamental problems facing modern analysis of high-throughput sequencing data is the integration of multiple -omic datasets [47]. More and more frequently we encounter research questions that involve multiple -omic datasets such as those that combine DNA-seq and RNA-seq (our CHOC pediatrics project), RNA-seq and ChIP-seq (role of TCF1E in Wnt signaling; Section 3.1.4), and a time-course RNA-seq across multiple conditions (Learning and memory in BAF53b knockout mice; Section 3.2.5), among other common combinations.

On top of this, we have access to publicly available datasets from GEO ([51, 15]), haplotyping projects such as dbSNP [149] and COSMIC [56], and high-throughput drug-interaction databases ([181, 90]), to name a few. Integrating these databases with our individual results from different portions of our genomic analysis pipeline requires pulling data from many of these publicly available databases. We maintain copies of many such publicly available datasets for immediate use in the various projects that need them. We keep this information within a single data repository and have wrapped access to each of them into a single unifying framework. We created this framework for easily performing overlaps of these databases and the results of our -omic datasets analyses.

To further integrate our various analyses, we make use of in-house software that we developed to visualize the interactions among our -omic datasets and our databases of interactions. The central component in this portion of the pipeline is a network-based view of our data. To achieve this, we needed to obtain information on the various ways the pieces of our data interact. Such information is able to summarize our results by tying together the most basic elements of our networks, our nodes: transcripts, proteins, and genes. Edges between nodes can denote that protein X interacts with protein Y (a protein-protein interaction), or that protein A has a predicted binding site within a certain distance to transcript A of gene B. We use this framework to build such networks, and visualize them using Cytoscape Web [105]. Besides determining the input set of nodes in our networks, our -omics datasets can be visually overlaid within the network by making use of publicly available plotting software (Google Visualization API) to plot the gene expression values for each gene, for example, as the node itself.

In total, we incorporate data from the following sources in this portion of our pipeline, none of which are specific to cancer:

Nodes:

- Proteins: UniProtKB [107]
- Genes/mRNA: UCSC [113]
- microRNA: mirbase [68]
- Protein complexes: CORUM [141]

Edges:

- microRNA-RNA: miRanda [67]

- Protein-protein: BioGRID [155], String [83], MIPS [125]
- Drug-protein: PharmGKB [177], BindingDB [104], DrugBank ([182, 181, 90])
- TFBS: MotifMap ([184, 37]), CENTIPEDE [128]

Annotations:

- Epigenetic marks: ENCODE [3]
- Pathways: KEGG [84], NCI PID [144], UniPathway [119]
- Abundance: ProteinAtlas [163], BioGPS [183]
- Protein features: Pfam [129], PhosphositePlus [75], SCRATCH [33]

4.1.1 Circadian clock regulates the host response to Salmonella [18]

In a paper published in PNAS [18], we help identify the connections between circadian transcription as regulated by the transcription factor, CLOCK, and the response to Salmonella infection within mice affected either at night or during the day. We integrate our -omic dataset processed using the steps outlined in Section 3.2 into a network view, incorporating transcripts with transcription factor binding sites using the in-house software described in Section 4.1.

Regulatory Network analysis Our network was initialized with the proteins identified as belonging to cluster 1 from the analysis of microarray data containing expression profiles for mice affected and unaffected during the night or during the day with salmonella (see Figure 4.1A), together with the clock protein aryl hydrocarbon receptor nuclear traslocator-like (ARNTL) and the NF-κB proteins NF-κB1, -κB2, v-rel reticuloendotheliosis viral oncogene homolog A (RelA), and reticuloendotheliosis oncogene (cRel). MotifMap ([184, 37]) was used to search for putative transcription factor binding sites within an 8-kb region centered on the translation start site of each gene in our network, excluding exons. Using known positional-weight matrices for clock and NF-kB proteins, as well as hypoxia inducible factor 1, alpha subunit (HIF1-α) (which was already present in cluster 1), we assigned to every site the following two scores: (i) a motif matching score (Z-score); and (ii) a conservation score [Bayesian Branch Length Score; BBLS [184]] calculated by using a multiple alignment of 30 genomes from mouse to zebrafish. We filtered these sites by using a Z-score threshold of 4.27 (P=0.00001), along with a modest amount of conservation by using a BBLS cutoff of 1.0. Directed edges were drawn between the transcription factors and the proteins whose gene had at least one binding site satisfying the above criteria. The resulting network was also further pruned (Figure 4.1D) by progressively removing those that were not annotated as transcription factors in either JASPAR ([24, 110]) or TRANSFAC 9.4 [111] databases, which jointly comprise >800 binding matrices corresponding to >400 distinct transcription factors in mouse.

Computational Analysis Reveals Connections Between Circadian Transcription and Inflammatory Response To extend our comprehension of the transcriptional pathways participating in the circadian activation of the host defense against infection, we used the above computational modeling approach to predict the transcriptional regulatory networks involved in the control of genes in each of the four clusters identified in our genomic profiling analysis. In cluster 1, we identified main synergistic nodes connecting transcription driven by BMAL1:CLOCK (ARNTL:CLOCK) to critical inflammatory pathways (Figure 4.1D). The gene node graphics show transcription factors with significant changes in expression between different conditions (blue, WT infected vs. uninfected day; green, WT infected vs. uninfected night; brown, Clock mutant infected vs. uninfected day; orange, Clock mutant infected vs. uninfected night; P<0.05; Figure 4.1). In agreement with previous

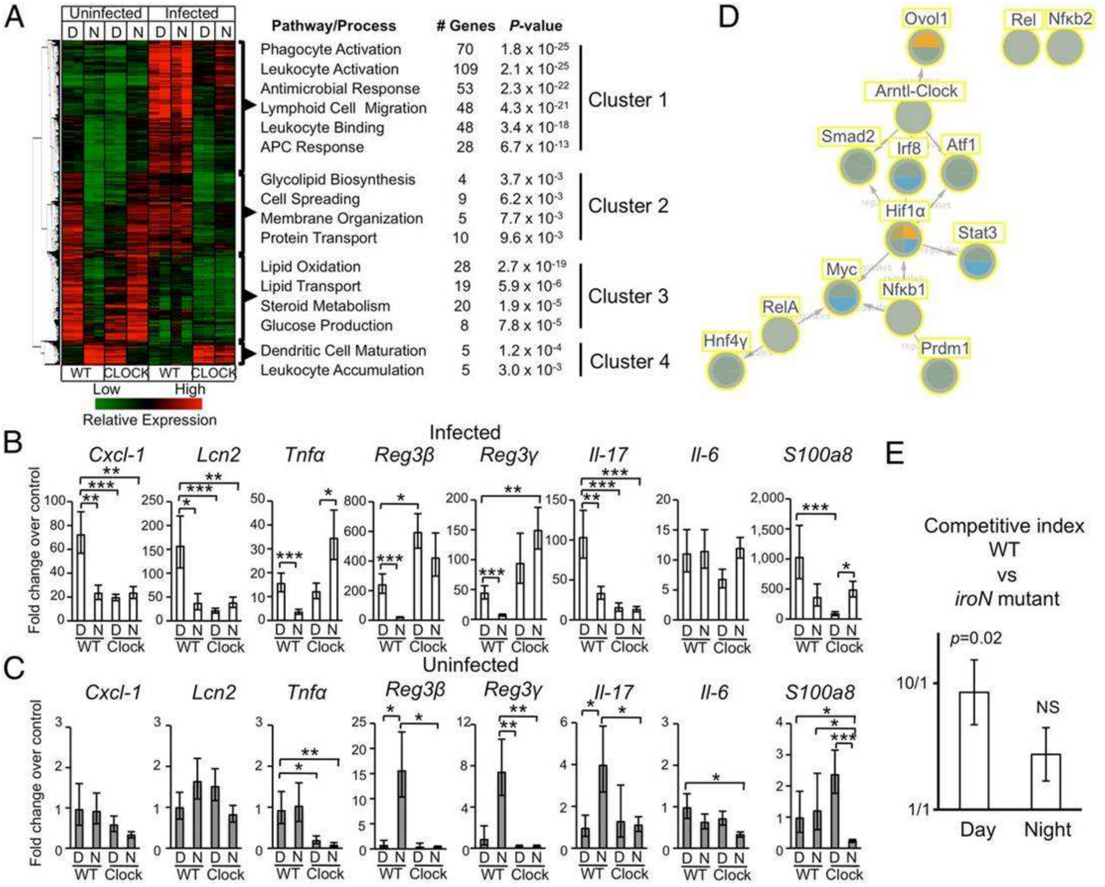


Figure 4.1: **Microarray analysis from cecum of mice infected with S. Typhimurium reveals a circadian mechanism modulating the response to acute bacterial infection. (A) Heat diagram showing changes in gene expression detected in mice infected with S. Typhimurium at day or night in WT and Clock mutant mice, compared with uninfected controls. A list of the most represented subcategories of genes from each cluster, the number of genes included in each subcategory, and the relative P value are shown. (B and C) Transcriptional profiles of selected proinflammatory/antimicrobial genes identified in cluster 1. (D) Network of transcription factors involved in regulation of subsets of genes included in cluster 1. Significant changes ($P < 0.05$) are shown as colored circles (blue, WT infected vs. uninfected day; green, WT infected vs. uninfected night; brown, Clock mutant infected vs. uninfected day; orange, Clock mutant infected vs. uninfected night). (E) Competitive infection with a mixture of S. Typhimurium WT and iroN mutant at two different times of the day.**
Figure and legend text adapted from [18].

reports, NF-B and hypoxia inducible factor 1, alpha subunit (HIF-1) are the transcription factors with the largest numbers of connections ([73, 148]). Both NF-B and HIF-1 share many target genes with the transcription factor BMAL1 (ARNTL), which cooperates with CLOCK in regulating circadian transcription. Notably, HIF-1 appeared to mediate significant changes in all four conditions that we analyzed. Nodes were much larger in WT mice compared with Clock mutant mice, thus indicating that many pathways regulated by HIF-1 also require a functional clock system. This observation is of particular interest because cross-talk between hypoxic and circadian pathways has been proposed, and Hif-1 is thought to be a clock-controlled gene [52].

4.2 Cancer-specific integrative analysis

Without context, calling variants within an individual's genome is not enough to identify the most relevant mutations in a patient. Variants provided by commercial solutions such as Complete Genomics or Illumina, Inc., or by open-source pipelines such as VarScan2 [93], do not solve the problem of ranking the most important genic mutations. Rather, they rank the most confident of such mutations, and in most cases an overwhelming number of potential driver mutations are identified. Common solutions exist, such as using predefined lists to screen variants, usually obtained for domain-specific knowledge such as the gene lists present within QIAGEN's Ingenuity Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). In addition to gene lists, other domain-specific knowledge can be incorporated and can even provide a ranking of the most important genes (see Chapter 5).

Solving this problem requires us to perform a few steps in our genomic analysis pipeline that are specific to each type of cancer, in order to provide a context in which to identify the most affected genes for each pediatrics patient. We integrate this information with the variants identified from DNA-seq, along with the transcripts identified as differentially expressed in the RNA-seq, to build a network view of the most important pathways within each individual.

Table 4.1: **Mean size and standard deviation of the gene lists curated using the three different methods for multiple types of cancer.**
Table from [188] (submitted).

Curated Gene Lists		
Gene List	**Mean**	**StdDev**
Entrez	154.6	134.5
MEDLINE	70.8	69.5
GeneRIF	63.7	74.1
UNION	243.3	196.2

4.2.1 Curating literature

To narrow down our variants to those contained within genes known to be involved in a certain type of cancer, gene lists are automatically curated from three primary sources. These three sources are (1) NCBI MEDLINE abstract and titles, (2) NCBI GeneRIF [115], and (3) NCBI Entrez queries.

For the first source, we perform text pattern matching on the corpus of abstracts and titles from NCBI Medline, using the UCSC hg19 genome annotation tables for a list of all known gene symbols in our search. Using the PubMed API, we retrieve a list of articles matching a specific type of cancer and extract all gene symbols in the titles and abstracts of these articles. Secondly, we cross reference the same articles with NCBI GeneRIF in order to find all genes that have been manually annotated for these articles. NCBI GeneRIF contains 800,000 gene symbols annotated to MEDLINE articles, 477,417 of which are for Homo sapiens. Thirdly, the NCBI Entrez web API is used to return a list of genes for any query related to each type of cancer. The final sizes of each of our curated gene list for each type of cancer are shown in Table 4.1.

Genes affected in multiple cancers Additionally, lists of genes which are known to be affected in or related to cancer are pulled from three public sources, in order to create a list of genes commonly implicated in a wide range of cancers. These sources, along with the number of symbols in each, are: (1) The Bushman Lab Cancer Gene List [27] (2032), (2)

Table 4.2: **Number of unique gene symbols for each type of cancer extracted using the three curated gene methods**
Table from [188] (submitted).

PubMed Query	Entrez Query	Total	Entrez	GeneRIF	PubMed
(all AND (pediatric OR leukemia)) OR (acute AND lymphoblastic AND leukemia)	acute lymphoblastic leukemia	349	224	96	125
(aml AND (pediatric OR leukemia)) OR (acute AND myeloid AND leukemia)	acute myeloid leukemia	622	412	177	203
atrt OR (atypical AND teratoid AND rhabdoid)	atypical teratoid rhabdoid	10	3	2	7
(clearcell OR (clear AND cell)) AND sarcoma	clearcell sarcoma	300	294	2	12
ependymoma	ependymoma	17	11	1	6
hlh OR hemophagocytic lymphohistiocytosis	hemophagocytic lymphohistiocytosis	51	27	9	20
(hodgkins OR hodgkin OR hodgkin's) AND lymphoma	hodgkins lymphoma	207	12	120	94
(jpa OR pediatric OR pilocytic OR juvenile) AND astrocytoma	juvenile astrocytoma	35	19	5	13
mds OR Myelodysplastic syndrome	myelodysplastic syndrome	244	171	52	71
medulloblastoma	medulloblastoma	189	171	28	16
melanoma	melanoma	580	286	252	213
neuroblastoma	neuroblastoma	505	337	133	144
osteosarcoma	osteosarcoma	361	253	79	104
rhabdomyosarcoma	rhabdomyosarcoma	178	143	20	40
(testicular OR 'germ cell') AND (cancer OR tumor)	germ cell tumor	134	26	26	41
wilms AND tumor	wilms tumor	111	84	17	23

The Cancer Gene Census [58] (489), and (3) Network of Cancer Genes 3.0 [38] (1495). In total, 450 genes symbols were found in all three sources, across the 2916 genes present in at least one source.

Gene list benchmark using expert knowledge We compared our results with those from a manually curated list of gene symbols suspected to be involved in germline tumors provided by our colleagues at CHOC:

DCC, FHIT, ALPPL2, HRAS, CGB, MGMT, FHL2, KITLG, KIT, ALPP, MXI1, MYBL2, MADH4, MAGEA4, LLGL2, EPCAM, MYCL1, CCNE1, CDKN2A, CDKN2C, POU5F1, CDKN2D, PIWIL1, KLK10, FAS, KLK13, TP53, RB1, DDX4, MYCN, JUP, AFP, NRAS, CCND2, PDGFRA, MDM2, GRB7, TNFRSF6, PLAP

We proceeded to overlap this gene list with the list of 134 total gene symbols extracted by our automated methods. For the extraction method from PubMed we found that 5 of the 41 symbols overlap: AFP, FAS, KIT, TP53, PLAP. For the Entrez queries method we found that 1 of the 26 symbols overlap: KITLG. For the GeneRIF method we found that none of the 26 genes were within the manually curated list.

Using the gene lists of genes affected in multiple cancers, of the 2916 genes present in any single source, 27 of the 30 manually curated genes were found. Using just the 450 genes confirmed in all three sources, 17 of 30 of the manually curated gene lists were found.

These overlaps suggest that our extraction method from PubMed queries performs relatively well, compared to the other methods. Although, in the case of the Entrez queries for germline tumors, we find an addition gene that was not picked up by the literature based method.

4.2.2 GEO datasets

Microarray datasets are automatically obtained from the Gene Expression Omnibus (GEO) [15] for the cancer types of our pediatric patient samples using the cancer type as a query. Control datasets and samples for each cancer type are also obtained if available. These control datasets are used in lieu of proper control sample RNA-seq datasets for each patient, which are not realistically obtainable for the pediatric patients being sequenced. We use these control datasets in the methods below to complement the RNA-seq differential analysis in such cases. The datasets obtained and the number of samples per type of cancer are shown in Table A.2 and Table 4.3.

Table 4.3: **Number of microarray samples per type of cancer**
Table from [188] (submitted).

ALL (64)	**AML** (436)	**AML_CONTROL** (39)
ATRT (18)	**CLEARCELLSARCOMA** (14)	**CLEARCELLSARCOMA_CONTROL** (3)
EPENDYMOMA (75)	**EPENDYMOMA_CONTROL** (75)	**EWINGSSARCOMA** (16)
HLH (17)	**HLH_CONTROL** (33)	**HODGKINS** (14)
HODGKINS_CONTROL (10)	**HODGKINS_OTHER** (45)	**JPA** (23)
JPA_CONTROL (8)	**MDS** (13)	**MDS_CONTROL** (15)
MEDULLOBLASTOMA (53)	**MELANOMA** (7)	**MELANOMA_OTHER** (11)
NEUROBLASTOMA (93)	**OSTEOSARCOMA** (25)	**OSTEOSARCOMA_OTHER** (12)
RHABDOMYOSARCOMA (12)	**SARCOMA** (43)	**SARCOMA_CONTROL** (15)
TESTICULAR (67)	**TESTICULAR_CONTROL** (3)	**WT** (44)
WT_CONTROL (26)		

Differential genes

Following standard microarray practices, each microarray dataset obtained is background normalized using the MAS5 algorithm [127]. Using platform annotations provided by GEO, probes are matched to gene symbols, and for cases where there are multiple probes per gene symbol, the probe with the maximum expression is retained. In the cases where raw expression is available, expression values are log-normalized to correct for the variance-mean bias commonly observed in microarray data. For any preprocessed datasets where raw microarray data is not available, data is log normalized if it was not already, to keep scales consistent.

After all datasets for all types of cancer present within our patients are preprocessed, gene symbols are then matched across all microarray samples. After removing samples and symbols that are missing more than 75% of data, 17011 unique gene symbols remain, for which any missing data is imputed using k-nearest neighbors. Lastly, quantile normalization is used to normalize between all arrays and the distribution of expression values across tumor types is shown in Figure 4.2. This pre-processing step is performed again if any additional patients with distinct types of cancer are obtained.

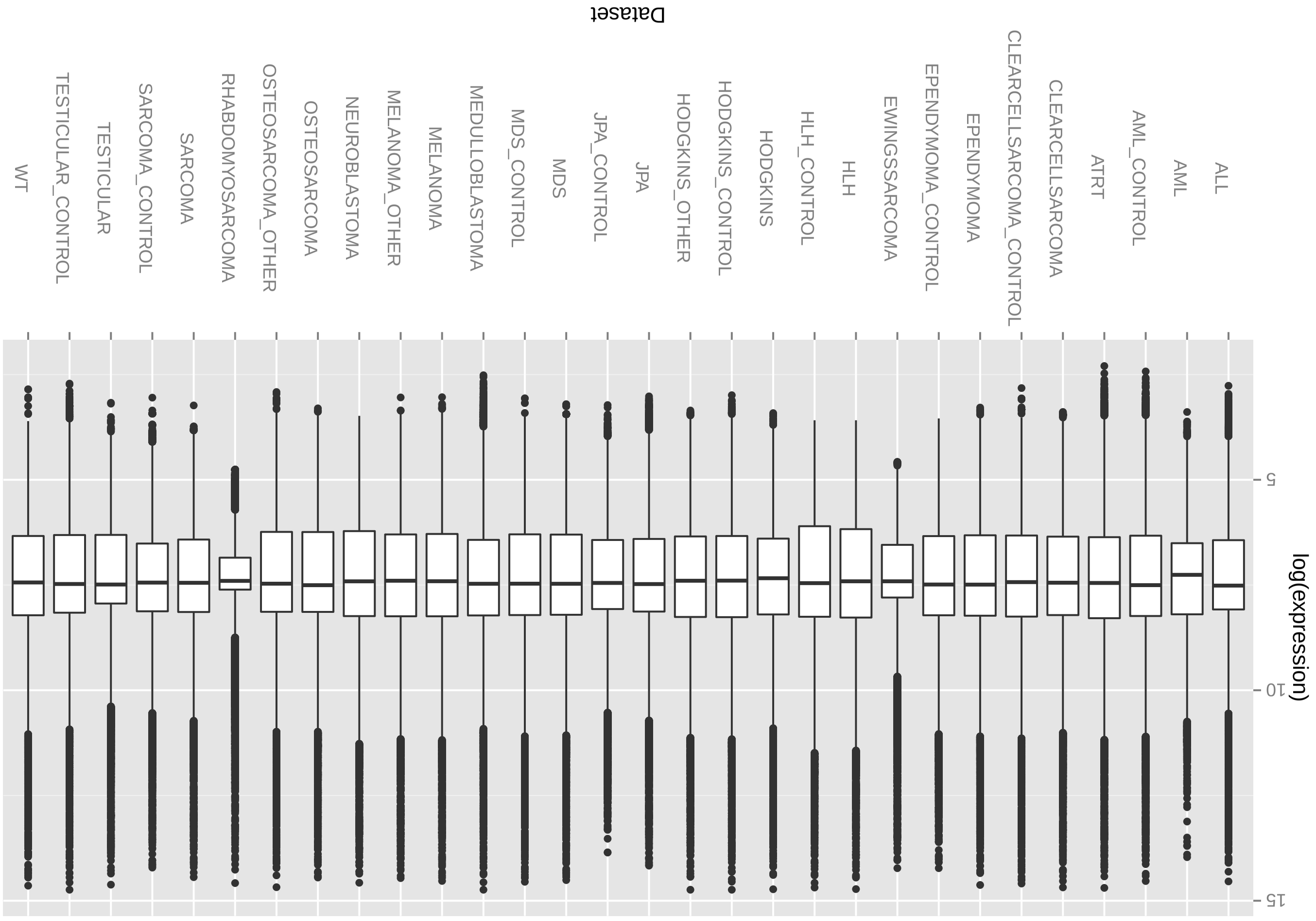
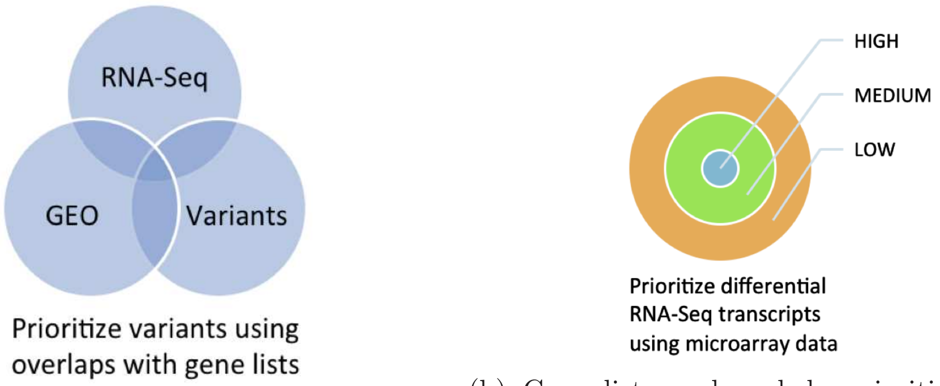


Figure 4.2: **Quantile normalized microarray datasets for each cancer type**
Figure from [188] (submitted).

To test for differential transcripts, Cyber-T ([13, 85]) with a Benjamini-Hochberg multiple test corrected p-value cutoff of 0.05 is performed on a number of different contrasts utilizing the microarray data. In particular, when control samples exist for a type of cancer, Cyber-T is used to identify differentially expressed transcripts specific to that type of cancer which can be used to prioritize (1) variants within those transcripts or (2) those transcripts that were also identified by the RNA-seq analysis differential analysis in patients afflicted with that cancer.

Cyber-T is additionally used to perform the exact same analysis as is done using the pooled cancer patient RNA-seq samples described in Section 3.2.4. In lieu of the patient samples, the median expression values for each gene symbol are used across all microarray samples for that type of cancer. The median microarray sample for each patient is tested for differential expression against the set of median microarray samples derived for each other patient as was done for the RNA-seq data. Additionally, in the types of cancer where control data is available, we perform the same differential analysis for all patients with that type of cancer using the median control microarray data instead of the median tumor microarray data. Lastly, using all of the tumor microarray data for all types of cancer, we use Cyber-T to identify transcripts that are commonly expressed, or unexpressed, in cancer. In summary, we define the following gene lists using microarray data:

1. GEO Control vs GEO Cancer (if applicable) for each tumor type
2. GEO Control vs GEO Matched Cancer (if applicable) for each tumor type
3. GEO Cancer vs GEO Matched Cancers for each tumor type
4. Common in GEO Cancers Expressed
5. Common in GEO Cancers Unexpressed


RNA-Seq
GEO
Variants
Prioritize variants using overlaps with gene lists
HIGH
MEDIUM
LOW
Prioritize differential RNA-Seq transcripts using microarray data

(a) Venn diagram of the overlap performed

(b) Gene list overlaps help prioritize RNA-seq transcripts

Figure 4.3: **Diagram of the processes used to prioritize expression variations**

4.2.3 Gene list overlaps

When we attempt to prioritize RNA-seq differential genes based on the expression or lack of expression of transcripts, we observe no clear separation between transcripts with variants and transcripts without variants. Instead, we must make use of the list of genes identified from our differential analysis of the microarray data for each type of cancer, in addition to gene lists curated from literature for each type of cancer, to prioritize transcripts identified by the RNA-seq differential analysis in patients with each type of cancer.

We first investigated the significance of various overlaps using a Fisher's Exact test to identify the overlaps with the most significant enrichment for small variants within patients. We observe significant ($P<0.05$) overlap of three of the cancer specific gene lists with affected genes within patients. The most informative, and significant, are the list of genes curated from literature, the differential transcripts identified using microarray data, and the transcripts with high expression compared to other patients that also fall within the microarray differential transcripts.

The significance of the last list above prompted us to prioritize our RNA-seq differential transcripts using a similar gene list overlap approach, since this overlap was found to enrich

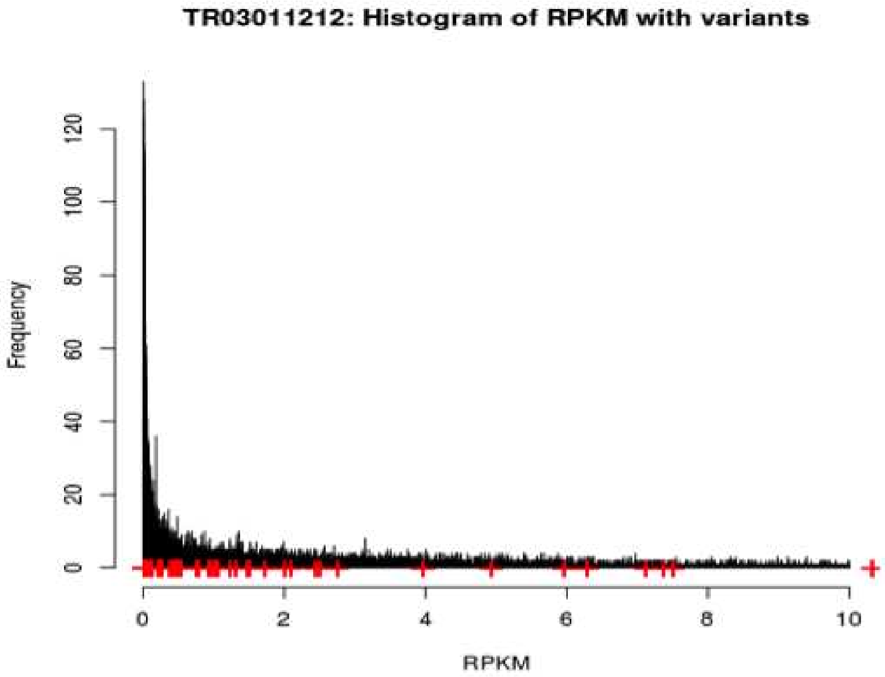


Figure 4.4: **Distribution of RPKM values for patient TR. Red ticks denote variants' RPKM values in this same patient.**

Table 4.4: **Significance of overlap with missense mutations for the various gene-expression differential analysis approaches for patient CHOC03**
Table from [188] (submitted).

	List Size	**# Variants**	**Overlap**	$-log_{10}(P)$	**P <0.05**
Curated from Literature	117	1319	10	1.36	TRUE
Bottom 5% RNA-Seq p-value HIGH	206	1319	15	1.12	FALSE
Bottom 5% RNA-Seq p-value LOW	3	1319	0	0.82	FALSE
Bottom 5% RNA-Seq p-value MEDIUM	354	1319	15	0.11	FALSE
Control vs Tumor Top 100	100	1319	7	0.81	FALSE
Control vs Tumor p-value <0.05	8595	1319	593	15.95	TRUE
Negative fold Top 5% RNA-Seq p-value HIGH	185	1319	13	0.97	FALSE
Negative fold Top 5% RNA-Seq p-value LOW	10	1319	1	1.03	FALSE
Negative fold Top 5% RNA-Seq p-value MEDIUM	242	1319	13	0.41	FALSE
Positive fold Top 5% RNA-Seq p-value HIGH	154	1319	15	2.15	TRUE
Positive fold Top 5% RNA-Seq p-value LOW	1	1319	0	1.28	FALSE
Positive fold Top 5% RNA-Seq p-value MEDIUM	165	1319	10	0.60	FALSE
Bottom 100 p-values for cancer GEO ANOVA	100	1319	6	0.57	FALSE

for transcripts with small variants – which likely influences the expression of those affected genes. For gene lists (2) and (3) from our microarray analysis in Section 4.2.2, we can identify tissue-specific genes and genes we would expect to change, respectively, in the RNA-seq analysis against the pooled patient samples for patients with that type of cancer. These help further prioritize the differential RNA-seq transcripts into `HIGH`, `MEDIUM`, and `LOW` gene lists based on overlaps with microarray gene lists (1), (3), and (2), respectively, where the `HIGH` category corresponds to the same overlap we found a significant enrichment of small variants in. The average size of these lists are summarized in Table 2.2.

4.2.4 Normal tissue expression

The Human U133A Gene Atlas dataset [156] is obtained from BioGPS [183] to be used as a measure of normal tissue expression for the tissues most similar to the tumor sample obtained in each patient. This determines a baseline gene expression profile in healthy tissue to be used as a control. This dataset contains GCRMA values as a result of normalizing the microarray samples obtained from 79 human tissues. Combining these with the RPKM values from the RNA-seq analysis, we generate profiles of gene expression in (1) all patient tumor tissue samples and (2) all of the matched normal tissue samples, in order to identify abnormal patterns of expression in patients, i.e. those that would not be expected due to normal differences between tissues from which tumors were obtained. An example of this profile is shown in Figure 4.5.

4.2.5 Mitelman fusions

The Mitelman database [116] contains 3752 entries corresponding to gene fusions implicated in different types of cancer. To identify and prioritize these gene fusions in our patients, we cross this database with all of the gene fusions found for each patient to identify high-priority

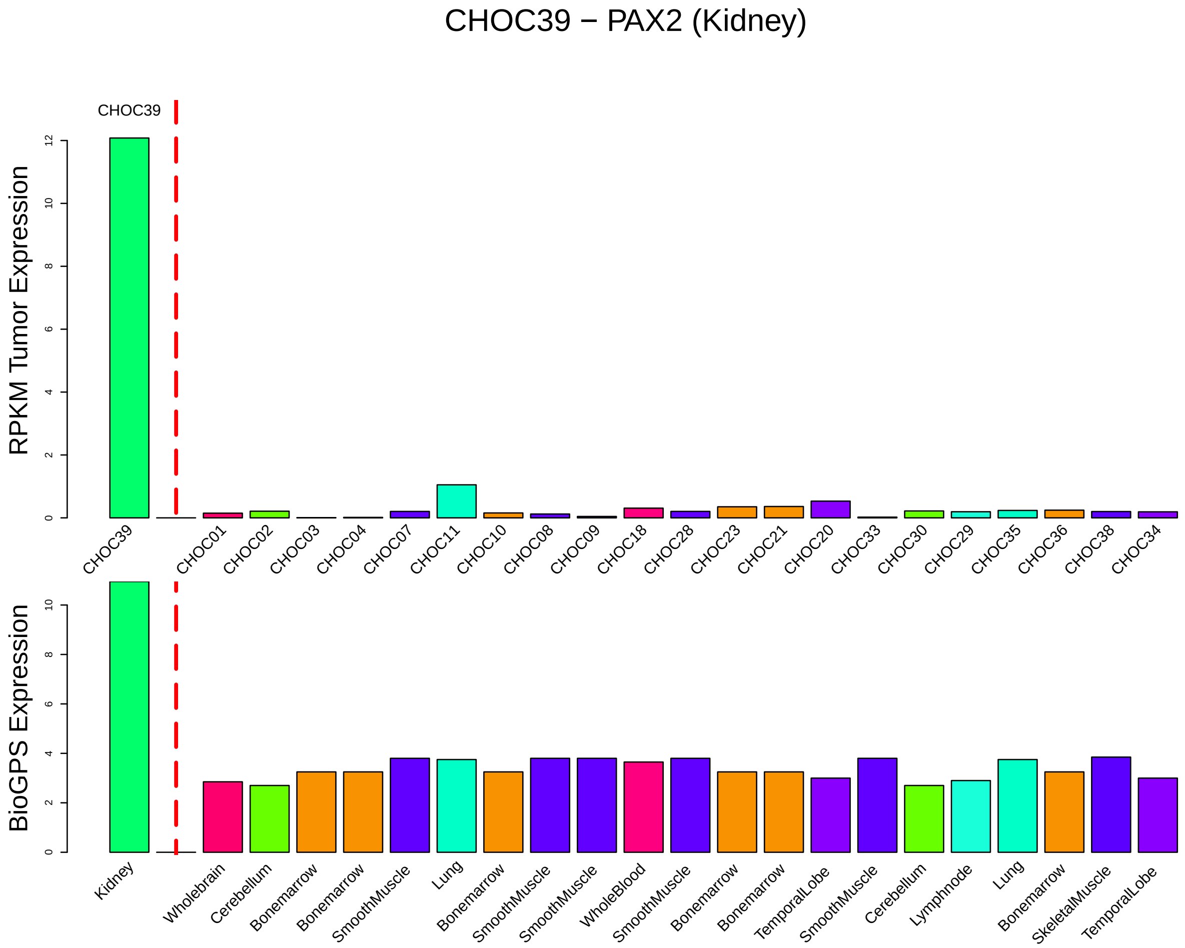


Figure 4.5: **BioGPS expression (lower) suggests that high expression of PAX2 in Wilms tumor patient (upper) is normal considering the normal gene expression within the kidney.**

fusions and to present the relevant literature in our final reports that are of clinical relevance. Three of the patients in our study contained fusions previously described. These fusions were originally identified in the same tumor type as each of the patients. All identified Mitelman fusions are listed in A.3. The hits identified for our clear cell sarcoma patient are listed in Table 4.5.

Table 4.5: **Entries in the Mitelman fusion database [116] for our Sarcoma patient. Most identified EWSR1/ATF1 entries have been previously found in clear cell sarcomas.** Table from [188] (submitted).

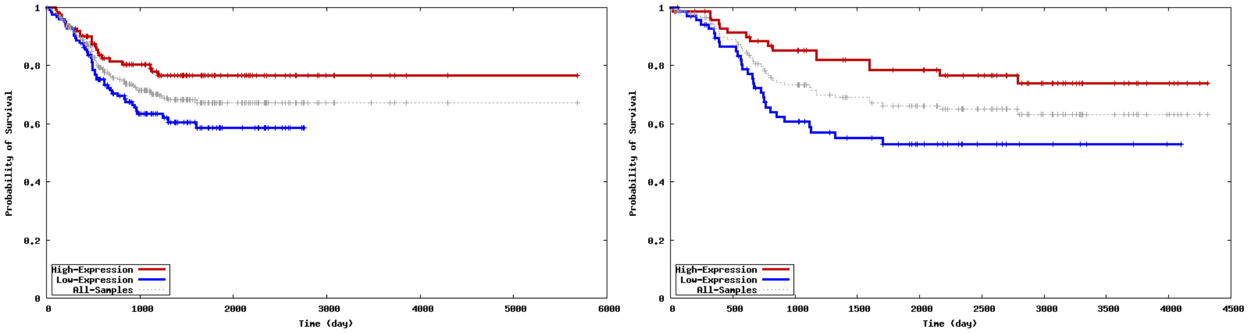
Author, Year	Journal	Morphology	Gene
Antonescu et al, 2011	Genes Chromosomes Cancer	Malignant epithelial tumor, special type	EWSR1/ATF1
Dunham et al, 2008	Am J Surg Pathol	Angiomatoid malignant fibrous histiocytoma	EWSR1/ATF1
Friedrichs et al, 2005	Int J Surg Pathol	Clear cell sarcoma	EWSR1/ATF1
Fukuda et al, 2000	Pathol Int	Clear cell sarcoma	EWSR1/ATF1
Hallor et al, 2007	Cancer Lett	Angiomatoid malignant fibrous histiocytoma	EWSR1/ATF1
Hansen Hallor et al, 2005	Genes Chromosomes Cancer	Angiomatoid malignant fibrous histiocytoma	EWSR1/ATF1
Hiraga et al, 1997	Virchows Arch	Clear cell sarcoma	EWSR1/ATF1
Panagopoulos et al, 2002	Int J Cancer	Clear cell sarcoma	EWSR1/ATF1
Rossi et al, 2007	Clin Cancer Res	Angiomatoid malignant fibrous histiocytoma	EWSR1/ATF1
Somers et al, 2005	Am J Surg Pathol	Osteogenic/bone tumor, NOS	EWSR1/ATF1
Speleman et al, 1997	Mod Pathol	Clear cell sarcoma	EWSR1/ATF1
Taminelli et al, 2005	Virchows Arch	Clear cell sarcoma	EWSR1/ATF1
Zucman et al, 1993	Nat Genet	Clear cell sarcoma	EWSR1/ATF1

4.2.6 Prognosis

Much work has been done to identify the most prognostic gene markers, specific to a type of cancer [180]. Such genes identify the difference in survival odds, usually captured using the Kaplan-Meier estimate [60], between subsets of the patients with low- or high-expression for that gene. Prognostic data is acquired by researchers at hospitals who profile incoming patients with high-throughput gene expression methods (either microarray or RNA-seq based) and have also measured the survival statistics for these same patients.

For rhabdomyosarcoma and neuroblastoma, the web site Oncogenomics has processed neuroblastoma data from the Oberthuer Lab. Our neuroblastoma patient, has the highest ranked gene DCC, with numerous mutations and low RPKM compared to other patients (see Figure 4.5). We observe a significant (p=0.012) difference in survival rates between patients in the low expression group (125) vs the high expression group (126) for DCC as indicated in Figure 4.6(a). Another high ranking gene, PAX3, is significant in Rhabdomyosarcoma (fused in our patient and high expression compared to others) (p=2.960e-03; Figure 4.6(b)).

PrognoSan [117] is a publicly available source of manually curated and publicly available survival and gene expression profile data obtained from GEO. At it's current date, it provides



(a) DCC

(b) PAX3

Figure 4.6: **Prognostic curves using Oncogenomics datasets for Neuroblastoma and Rhabdomyosarcoma highlight significant survival differences for patients with low or high expression of key affected genes for two patients. DCC (left) has low expression in a neuroblastoma patient, in addition to numerous genetic mutations within its coding sequence and PAX3 (right) is involved in a gene fusion in a Rhabdomyosarcoma patient in addition to higher overall expression.**

raw data for extracting out Kaplan-Meier estimates for AML and a few other cancer types that are not present without our pediatric cohort. Currently, there is a lack of prognostic data for other common pediatric cancers. Availability of future prognostic datasets for all types of cancers is essential for a proper personalized medicine pipeline, such as is being done in the area of pharmacogenetics [138], as in the example of certain detectable DNA alterations in HIV patients affected drug efficacy [126].

4.3 Cancer pathways

It has been shown that cancer cells share in common multiple acquired capabilities that enable the cell to proliferate uncontrollably. These hallmarks of cancer have been highlighted previously ([70, 71]) and show a wide range of known pathways to be affected across different types of cancer. To visualize the connections between affected genes for each patient within known pathways – as well their connections to unaffected proteins – networks are created using in-house software which are then rendered in a web browser using CytoscapeWeb [105].

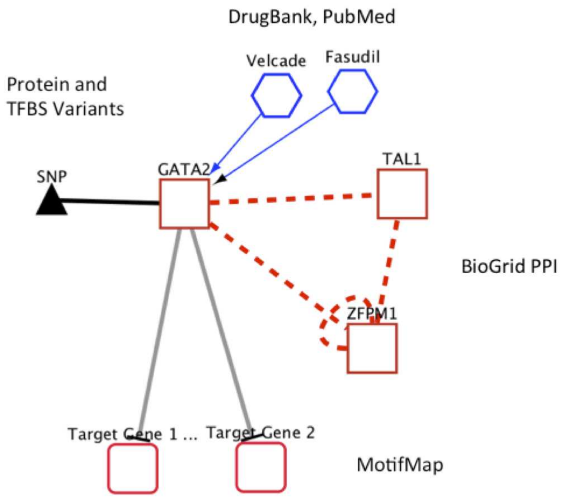


Figure 4.7: **Network is used to relate transcripts to each other and to potential drugs.** Figure from [188] (submitted).

In order to initialize networks with proteins related to specific pathways, 478 known pathways are downloaded from KEGG Pathways [84] and the NCI Pathway Interaction Database [144]. Subsequently, transcription factor (TF)-DNA, TF-TF, protein-protein edges are added to the network based on the publicly available datasets from MotifMap ([184, 37]) and BioGRID [155], respectively. Variants on proteins, as well as the proteins identified as differential in the microarray and RNA-seq analyses, are used to highlight portions of the network and help visually interpret the biological role of the mutations. Variant TFBS are visualized by highlighting the edges between transcription factors and the genes that contain a site for that factor within its promoter. Further, drug-protein interactions are added to the network, as described in the next section. Taken together, this network approach assists in investigating potential driver mutations with a focus on identifying potential drug candidates and their targets. A simplified example of such a network is shown in Figure 4.7.

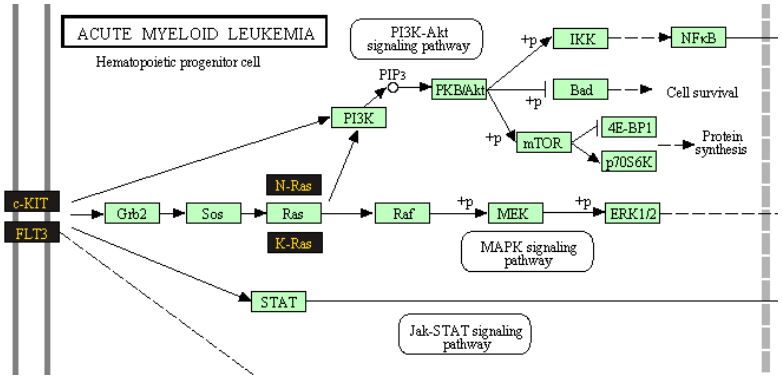


Figure 4.8: **Dark boxes indicate proteins with clinically associated variations.** Figure from [188] (submitted).

4.4 Identifying drug candidates

In order to elucidate potentially druggable therapeutic targets, we have integrated several publicly accessible databases of drugs into our network analysis. We have included well-characterized and predicted drug-effects, binding affinities, and drug-efficacy. These databases include the following resources:

- DrugBank [90] [181] [182]
- BindingDB [104]
- PharmGKB [177]

Each database provides an orthogonal set of annotations from which one can infer potential attenuation of known drug-effect, or perhaps novel drug interaction. Additional drug and drug-target information were also incorporated using semantic web resources for open drug data. These include Bio2RDF [17] [28], Chem2Bio2RDF [31], and Linked-Open Drug Data (LODD) [142]. Figure 4.8 shows the original AML pathway from the KEGG database. Figure 4.9 shows the corresponding auto-generated drug-target network used in exploring potential therapeutic targets.

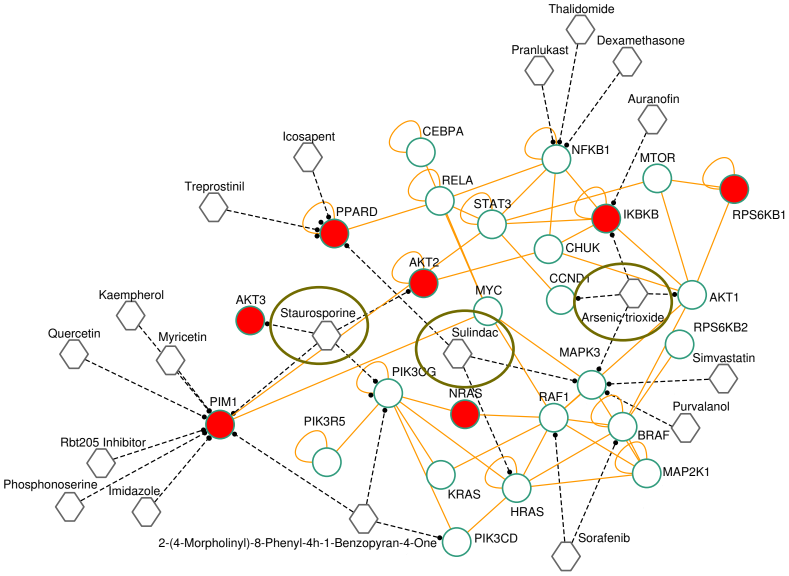


Figure 4.9: **Network limited to variants. Circles denote proteins and hexagons denote drugs. Filled circles denote affected proteins, with identified potentially therapeutic drugs circled.**
Figure from [188] (submitted).

To assist in identifying potential therapeutic drugs, we use a network-based approach which leverages the auto-generated networks for all pathways. For any gene target, we identify which KEGG or NCI pathways it is present in, and perform a breadth-first search starting at the gene target until we find a drug with an affected gene target. Additionally, if multiple such drug-targets exist at the same distance from our initial target, we choose the drug that targets the most genes, with preference given to drugs with a greater number of affected targets. Figure 4.9 shows the set of drugs reached by this method via a search originating from each of the affected genes in the AML pathway for one patient. Using the pathway ranking method described in Section 5, we additionally prefer drugs obtained from the top ranked pathways that contain our gene target.

Chapter 5

Reporting

Collaborative interfaces are one of the five essential elements of a personalized genome analysis pipeline as outlined by Valencia and Hidalgo [164]. Such an interface displays the results of the various analysis steps distilled down to just the most important pieces of information that a clinician or non-computational biologist can easily interpret, usually in some kind of interactive interface such as the Integrative Genome Viewer (IGV) ([135, 160]) but sometimes in a static PDF report.

In the cases of identifying genetic variations in an individual, identifying a list of differentially expressed genes between two conditions, or in calling ChIP-seq peaks, the task of identifying the most important results to present for further study is non-obvious, particular such list sizes are on the order of 100s and 1000s. For differential genes, most tools such as CyberT [13, 85] provide p-values which can rank genes to identify the most significant, but in many cases domain knowledge is incorporated such that the a subset of the most significantly differential genes is retained for further study, such as a list of genes involved in a certain pathway, or genes having an enriched GO terms identified using tools such as DAVID [44]. In the case of pediatric cancers, we have a ranking of DNA variations based on the number of reads supporting the altered DNA sequence in that patient, but the problem of identifying the most important of such mutations still remains.

We developed a novel approach to reporting for our pediatric pipeline in order to sift through the still many affected genes found in each patient, despite having removed variations present in the germline control samples. When looking at the overlap with the 487 KEGG and NCI pathways in our network analysis, on average 342 pathways had at least one small variation per patient. Additionally, we find that over 100 curated genes per patient are present within one of these pathways. To reduce this to a more clinically relevant and manageable number of affected pathways, it is necessary to limit pathways based on their importance in a each specific type of cancer. Similarly, we must limit the list of affected genes that contain genetic variants or have aberrant gene expression. After doing so, we then build networks for the pathways most affected pathways in each patient in order to visualize the interactions between affected genes within the same pathway, and to identify potential drug candidates.

5.1 Framework

Our distilled reports are presented to collaborators and clinicians through in-house software that renders reports into HTML, allowing for easily presenting tables of results and downloadable figures to collaborators. We adhere to the approach outlined by Knuth in his book Literate Programming [92], and implement all of our reports using a markup language embedded in so-called “runnable reports”. Our in-house software is an extension of previously developed software, org-mode [146], combined with the popular statistical computation language R [131]. Figure 5.1 shows an example of such mark-up, and how it is commonly used in our reports.

Custom PHP code is used to wrap calls to emacs which is used to render the org-mode

```
* Curated Genes (=Curated=)

For src_R[:session *R* :results raw]{disease},
src_R[:session *R* :results raw]{length(curated)}
genes were curated from PubMed and Entrez searches.

Out of these,
src_R[:session *R* :results raw]{length(curated.variants)}
were found to have some type of variation present
and are listed below.

#+BEGIN_SRC R :session :results verbatim raw :exports results
capture.output(write.org(load.table(’curated.txt’)))
#+END_SRC
```

Figure 5.1: **Example org-mode markup for reports**

reports to HTML. This webserver implements a read-only group account, in addition to user accounts with permission to edit the reports and rerun. This allows for a one-way sharing of reports using a uniform, and secure, interface. HTML reports are styled using modern software packages, including jQuery, Bootstrap. A pop-up dialog using fancybox and an in-house node-js server wraps the R package shiny [140] as shown in Figure B.6. See Appendix B for screenshots of our collaborative interface generated for each patient using this framework. All software requirements are listed in Appendix A.2.

5.2 Ranking pathways and genes in pediatric cancer

In order to filter genes with variations down to the ones most probable to contain driver mutations, we develop a ranking method for both pathways and genes based on enrichment scores. Using the list of curated genes described previously – those specific to a type of cancer – we use a Fisher's exact test to determine the statistical significance of the overlap between the curated gene list and the list of genes in each pathway. Pathways are then ranked based on this significance value. This ranking is specific to each type of cancer but not specific to any individual patient. Additionally, for each patient, the ranked pathways for their type of cancer are filtered for only those pathways containing at least one genetic variation (small or large) within the curated list of genes for that cancer. In most patients, this reduces the average affected pathways from 342 down to less than 50 affected pathways. Specifically, we compute the pathway enrichment p-value as the probability of observing the overlap between the pathway gene list and our curated gene lists, assuming 39131 genes in the human genome.

$$\text{Score(Pathway)} = -\log_{10}(\text{pathway enrichment p-value})$$

Similarly, for each patient, we compute an enrichment score for any single gene based on the list of variants which are affecting that gene. The enrichment score of each individual type of variant listed in Table 2.2 (under the small variations, large variations, and gene expression columns) is determined using the overlap of variants of a particular type within the table with the list of curated genes for that patient's cancer, calculated using a Fisher's Exact test.

$$\text{Score(Gene)} = \sum -\log_{10}(\text{variant enrichment p-value})$$

To justify this approach, assuming the curated lists reflect the genes we expect to be mutated in patients with this type of cancer, we should observe more variations within this list than in a random gene list of the same size. As remarked on in Section 4.2.3, we find this to be the case across patients. Types of genetic variations that score higher will be ones that contain a larger number of affected genes within the curated list, and therefore we might expect our driver mutations to be carried by the same categories of mutations in those genes, and others, within the same patient.

To assess the robustness of our gene ranking method to variations in the previously curated gene lists, we used a leave-one-out approach. After the initial ranking of genes based on the initial curated gene list for each patient, we removed each of the top 50 curated genes from the curated gene list and re-ranked that gene in order to measure its change in rank. We found that across all patients, 81% of the top 50 genes for each patient moved less than 25 ranks, with a median change in rank of 3 for the top 25 curated genes. Further, within the top 10 genes for each patient we observed a median change in rank of only 1. This suggests that the top ranked curated genes are influenced less by their own contribution to the ranking score than those further down the list and that the ranking of genes is relatively stable with respect to the composition of the curated gene list.

Lastly, using the list of 450 symbols common to most cancers as was defined in Section 4.2.1, we look for affected genes within this more general list that rank highly but are not contained within the curated list of genes. These are genes that have been implicated in any type of cancer. Therefore, any affected genes within this list warrant further consideration aside from those in our curated gene lists for each type of cancer. Our final reports are in the form of network views of the top ranked pathways, along with tables of the top ranked genes along with their associated pathways, drug candidates, and expression profiles.

5.3 Interesting findings in pediatric cancer

5.3.1 Patient CHOC23 (AML)

To demonstrate the effectiveness of our pipeline in identifying genes affected by likely driver mutations, we explore the ranking observed for one of our patients with acute myeloid leukemia (AML), CHOC23. Our objective is to identify the affected genes within the tumor genome of CHOC23 that most directly relate to AML. We employed the ranking method

Table 5.1: **Top 10 ranked pathways for CHOC23 (AML)**
Table from [188] (submitted).

Pathway Description	**Score**	**Size**	**Curated Overlap**	**Curated Affected**
PI3K-Akt signaling pathway	7.07	881	35	3
Chronic myeloid leukemia	5.66	179	13	1
Acute myeloid leukemia	4.19	180	11	3
Signaling events mediated by HGFR (c-Met)	3.81	80	7	1
Pathways in cancer	3.09	890	26	3
Hepatitis B	2.94	374	14	1
Small cell lung cancer	2.93	215	10	2
Pancreatic cancer	2.73	191	9	1
ATR signaling pathway	2.63	39	4	1
Toll-like receptor signaling	2.64	236	10	1

described previously to rank the pathways that would be of most interest in AML, the results of which are presented in Table 5.1. The top 3 pathways for this patient were PI3K-Akt signaling pathway, Chronic myeloid leukemia, and Acute myeloid leukemia. This initial ranking of pathways is not specific to this patient and is shared with all AML patients. As we would expect, the leukemia pathways for CML and AML rank near the top.

The gene ranking method also performs well for this patient, and in contrast to the pathway ranking, is specific to this patient. As shown in Table 5.2, this method ranks MLL3 the highest. The score for MLL3 is calculated as follows:

$$\begin{aligned}
\text{Score(MLL3)} &= \text{Score(Fusion)} + \text{Score(Deletion)} \\
&+ \text{Score(LowerCN)} + \text{Score(Missense)} \\
&= 0.5767 + 1.5911 + 0.6887 + 0.6484 \\
&= 3.51
\end{aligned}$$

The higher value for Score(Deletion) reflects the fact that, in this patient, deletions are more enriched within the curated list of genes for AML than any of the other variations. For the majority of patients, the genetic variations that score highest are: (1) microarray

Table 5.2: **Top 10 curated gene ranking for CHOC23 (AML)**
Table from [188] (submitted).

Gene	Score	RPKM	Variants (counts)
MLL3	3.51	0.71703	fusion (6); deletion (1); lowerCN (13); missense (1)
ERCC1	2.92	0.8563	under expr. LOW; inversion (3)
NCOR1	2.78	0.78821	inversion (3); deletion (1); higherCN (3)
TCF7L1	2.55	0.42467	deletion (1); unique inframe (1)
NCOR2	2.51	0.76177	under expr. MEDIUM; deletion (6)
HPR	2.24	0.41816	deletion (1); missense (1)
LEPR	2.24	0.30763	deletion (12); missense (1)
NRAS	2.16	0.46848	flagged missense (1)
HSPA1A	1.70	0.55796	under expr. MEDIUM; tandemdup (1)
MUT	1.65	0.37452	tandemdup (1); missense (1); loh (1)

differential genes, (2) fusions, (3) deletions, and (4) genes with lower RPKM compared to other patients. The fact that this gene ranks at the top is of no small consequence, it is one of the few identified Mitelman fusions in all of the patients. The other mutations identified provide further evidence that MLL3 is significantly altered in this patient, and contribute to it being the highest ranked.

We make use of our automated method for drug recommendations to address the problem of a lack of directly druggable targets. For this patient, none of the top ten ranking curated genes have any drugs that directly target them, making therapeutic recommendations for these genes problematic. To address this, we search over each of the top ranked genes and identify which of the top ranked pathways, if any, that gene is contained within. If one or more pathways exists, we perform a graph-based search for the nearest best drug candidates, as described earlier in Section 4.4.

For this patient, the first such targets with drug candidates are TCF7L1 and NCOR2. In the absense of directly druggable targets, we find that TCF7L1 has two potential drug candidates. The first of which, Staurosporin – a potent protein kinase inhibitor – is identified in the Prostate Cancer pathway through TCF7L1′s interaction with CTNNB1, which interacts with Staurosporin′s direct target GSK3B. Interestingly, we find a number of additional targets for Staurosporin in the AML pathway (shown in Figure 4.9 for a different AML patient)

– AKT2, AKT3, KIK3CG, and PIM1, all containing genetic variants within this patient. The second drug candidate, Vorinostat, a HDAC1 inhibitor, is identified in the Regulation of β-catenin pathway, again through TCF7L1's interaction with CTNNB1, which interacts with HDAC1.

We also identify Vorinostat as the best drug candidate for our next highest ranked gene, NCOR2, through an entirely different pathway, the Notch signalling pathway where HDAC1 and NCOR2 have a protein-protein interaction. This drug has been in Phase II clinical trials for AML patients, and while it was not shown to be effective alone, it shows promise as a potential drug for high-risk patients in conjunction with other drugs [145]. It may be the case that only a subset of patients, such as this one, would respond to this drug. In the absense of any direct drug candidates in the top ranked genes, we are able to identify reasonable drug candidates through this pathway-based approach.

5.3.2 Patient CHOC33 (Neuroblastoma)

Another interesting case is patient CHOC33, a neuroblastoma patient. Neuroblastoma is a tumor derived from neural crest cells from the sympathetic nervous system. Using this patient as an example, we explore how we relate the gene expression data to the top ranked genetic variations found. Our pipeline focuses on the curated list of genes associated with neuroblastoma (505 genes), for which the top ranking in this patient are as follows: PTPRD (2.68), PARK2 (2.44), DCC (2.34), and ALK (2.22). All of these genes contain genetic variants within their coding sequences.

Our second ranked gene, PARK2, contains a deletion in the first intron as well as an exonic region of higher copy number, as shown in Figure 5.2, which contributes to its high rank. PARK2 is also identified as having a relatively higher expression in this patient compared to others (Figure 5.3). The gene expression profile provides strong evidence that PARK2

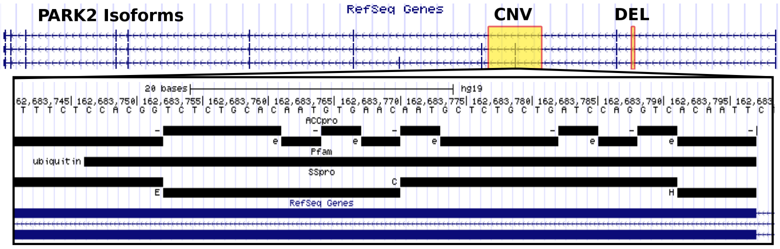


Figure 5.2: **PARK2 variations in patient CHOC33. Zoomed in region highlights features of exon 2 within the copy number variant, which includes a functional ubiquitin domain identified by Pfam.**
Figure from [188] (submitted).

gene expression is being altered as a result of its genetic variants. This lends credibility to the called genetic variants, as well as informing on the direction of change of PARK2 expression in this patient's tumor. Additionally, by overlapping Pfam predicted domains for PARK2, we identified a portion of the ubiquitin domain that confers PARK2 a role in the ubiquitin-ligase pathway (Figure 5.2). This further suggests that PARK2 is functioning in tumor progression in this patient. Recently, PARK2 has been shown to have an emerging role in cancer [186].

5.3.3 Patients CHOC36 and CHOC03 (AML)

We perform a meta analysis on our two primary AML patients, CHOC36 and CHOC03, in which we attempt to find genes that had common variations: either genetic or in their expression. One such gene is EPOR, which stood out as having significantly higher expression in both patients compared to other patients' tumors (Figure 5.4). Despite a known higher expression in healthy bone marrow as compared to other tissues (data not shown), the level of EPOR expression observed for these two patients is not observed in other patients for which the RNA-seq data was also obtained from the patient's bone marrow.

EPOR, known as erythropoietin receptor, is involved in the Jak-STAT signaling pathway,

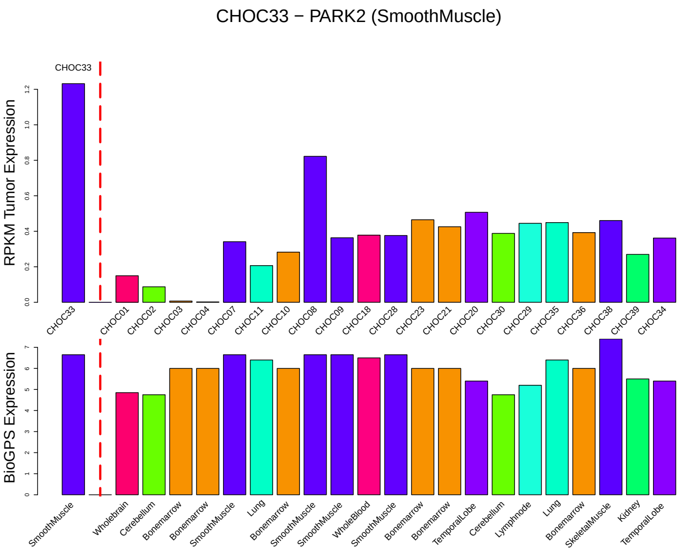


Figure 5.3: **PARK2 gene expression in patient tumors (top) and BioGPS normal tissue gene expression in tumor-matched tissues (bottom), where patient CHOC33 is the first bar on far left. We observe higher expression for PARK2 in CHOC33 as compared to other patients, in contrast to a relatively constant gene expression across healthy tissues.**
Figure from [188] (submitted).

which ranks within the top five pathways for both patients. Additionally, for CHOC03, the top 5% most highly differentially expressed genes were enriched within the list of curated genes for AML, indicating a strong increase in expression in a subset of curated genes for this patient as compared to other patients. This not only has the effect of ranking EPOR highly (#19 ranked curated gene, #1 ranked curated gene within a top 25 ranked pathway), but also of highlighting specific pathways that are over-expressed, mainly the PI3K-Akt and Jak-STAT signaling pathways. The Jak-STAT signaling pathway for CHOC03 contains high expression variants in a number of highly connected genes, namely: STAT5A, EPOR, PTPN6, IL6ST, CSF2RF, JAK3, TPOR, and PIM1 all show much higher expression than other patients. These variants are readily visible using the network approach (network not shown).

While not differentially expressed between our curated AML control vs. tumor microarray samples (P=0.6354), it has been shown previously that in approximately 60% of AML pa-

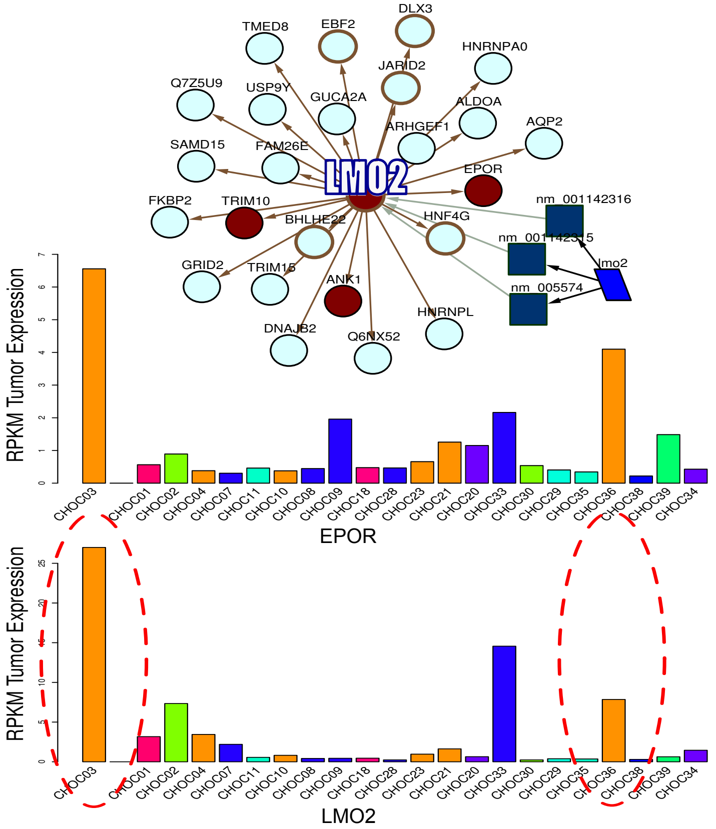


Figure 5.4: **Enriched transcription factor LMO2, along with 3 of 25 of its predicted targets (top; dark circles for genes with high expression), has high expression compared to other patients for both of our primary AML patients (bottom; primary AML patients circled). One of these targets, EPOR, was identified as being to be significantly higher in both primary AML patients and not others (middle).**
Figure from [188] (submitted).

tients, EPOR is unexpressed [32]. Additionally, remission times for patients with higher EPOR expression is significantly lower compared to those without EPOR expression [159] which is likely the case for these patients since all of the patients sequenced are after recurrence of the primary tumor. Additionally, in some cases patients with AML are being treated with erythropoiesis-stimulating agents, but it is believed that this could cause proliferation in a subset of AML patients with EPOR expression ([54, 32]), suggesting that these AML patients fall into a specific subtype of AML, differentiating them from our secondary AML patients CHOC23 and CHOC26, for which we do not observe an increase in EPOR expression.

Additionally, we investigated what transcription factors were enriched in each of our AML patients based on the location of predicted binding sites upstream of our differential gene lists (see Section 3.2.4). Further validating the importance of EPOR in these patients, we identify a significant enrichment for transcription factor LMO2 in our list of over-expressed transcripts (rank #3 for CHOC03; p-value = 4.5E-5). LMO2 and three out of its 25 targets predicted by MotifMap (EPOR, ANK1, and TRIM10) all have high expression in this patient (Figure 5.4). LMO2 has been previously found to be involved in AML [34], and its high expression, particularly in CHOC03, compared with other patients is further evidence of a subtype of AML within our primary AML patients.

Chapter 6

Conclusion

What we have developed for our pediatric cancers project is a complete genomic analysis pipeline starting from raw sequencing reads leading to clinically interpretable results in the form of short (1-100) ranked lists of the most important affected genes. In practice, the turn around time is a day for processing of the raw sequencing reads and generating the final reports – for patients with cancer types for which curated gene lists have already been obtained.

Our pipeline adheres to the five steps of a cancer analysis pipeline outlined by Valencia and Hidalgo [164]: 1. Genome analysis: We analyze DNA-seq and RNA-seq data from commercial vendors using a uniformed format for calling variants. 2. Consequences of mutations and genomics alterations: For small variations we identify the affect on protein sequence in addition to protein secondary structure, solvent accessibility, and known protein domains. 3. Network level analysis: We make use of NCI and KEGG pathways to identify the most relevant pathways for each type of cancer. 4. Drug: We make use of the ranked pathways and genes to identify potential drug candidates for each patient. And lastly, 5. Collaborative interfaces: We integrate multiple sources of information into a network view that includes regulatory information across all patients and tissue types for exploring the interactions among affected genes and potential drugs.

6.1 Applications of pipeline

Our pipeline has an emphasis on pediatric cancers, but we present results where we can reuse portions of our pipeline in the analysis of other high-throughput sequencing projects. In Section 3.1.3 we show how the search for transcription factor binding sites is adapted to predicting phosphorylation sites in the human proteome. In Section 3.1.4 we developed a ChIP-seq analysis pipeline in parallel with our DNA-seq and RNA-seq portions of the pipeline, which attempts to answer a specific research question in colon cancer cells using novel analysis approaches. In Section 3.1.6 we show how we can define a transcriptional regulatory network of skin wounding using the TFBS enrichment portion of our pipeline, combined with the network analysis described in Section 4.1. In Section 3.2.2 we use our differential analysis pipeline for microarray datasets to identify differential antigens in development of a vaccine for Tuberculosis. In Section 3.2.5 we use our differential analysis pipeline developed for RNA-seq data to investigate BAF53b knockout mice. Overall, we outlined a standard set of analyses performed for various types of high-throughput sequencing data.

6.2 Comparison to other work

In contrast to other published pipelines [42], our pipeline successfully integrates expression data into our ranking, in addition to giving priority to mutated variant transcription factors. A recent opinion paper [109], highlights the importance of the integration of multiple -omic datasets, which we have demonstrated.

Our method differs from other methods such as IntOGen [62] or MuteProc [87], which only attempt to use a cohort of patients to identify a list of driver mutations across patients. These are not specifically aimed at identifying the driver mutations within a single patient. Similarly, the commercial packages offered by Tgen or Cypher Genomics only screen for mutations within a predefined and pre-ranked set of mutations, and are typically biased towards a handful of well studied genes (i.e. TP53). When a rough comparison is made to reports obtained for a few of these patients, it was found that these commercial offerings list genes almost always falling within our curated list of genes, that predominantly had missense mutations and aberrant copy number variations, all of which are reported by our pipeline as well.

6.3 Importance of automation

Most importantly, after initially obtaining the datasets used in our integrative approach, our pipeline is automated up to and including the identification of potential drug candidates, and handles newly diagnosed patient with cancer types we have already seen without any intervention. This is an important aspect when working with pediatric cancer patients where the time from diagnosis to treatment is critical. In fact, the main reason multiple sequencing technologies were required for this project was to balance the turn around time of the sequencing technology and the cost of the sequencing technology on a per patient basis. Being able to return clinically relevant results immediately after sequencing results are obtained is an important aspect of a complete genomics analysis pipeline such as this, and will be critical of any personalized genomics pipeline that is to have widespread adoption.

6.4 Value of integrative approach

By combining RNA-seq, DNA-seq, and microarray data, in addition to numerous sources of annotations on the reference genome, we were able to identify likely driver mutations in pediatric cancers. We found that such an integrative approach is essential, and information from gene expression data in particular, can complement a search for genetic variants, making results more robust. Typically, we observe many mutations within the top ranked pathways, indicating that multiple genes are likely affected in a tumor cell in order to effectively knockout critical pathways, as shown to be required in cancer ([70, 71]).

In some cases, gene expression data alone can stratify patients with different subtypes of cancer, such as was the case for our primary AML samples and EPOR expression. In other cases, gene expression data was found to agree with the DNA-seq variants, giving stronger evidence that this particular variant could be considered a driver mutation. The gene lists derived from microarray control vs. tumor data (when available) are found to overlap well with the set of genes affected by variants (P=1E-15 for AML patient CHOC03). These curated lists allow for screening of variants within a set of the most important genes and pathways, by making use of multiple sources of patient data.

We found that using integrative approaches in the form of gene and drug networks along with gene expression profiles helped improve the interpretation of genetic variants. Our novel ranking methods quickly identify the most important mutations for the cancer specific to each patient and we showed in a few example patients that the most highly ranked genes and pathways had interesting results that agreed with literature. What we have developed thus far is a general genomic pipeline, which we demonstrated a use for in identifying likely driver mutations in pediatric cancer. This same pipeline can be readily adapted to the study of any genetic variants associated with any trait or disease of interest (e.g. the “driver” mutations of schizophrenia).

6.5 Future work

During the course of the development of our pipeline we observed specific biases in some of the results depending on the sequencing platform used, necessitating in some cases correcting for these biases, as is the case for a few fusions that appeared in multiple Illumina patients that at first appeared clinically relevant. Such technology biases are an aspect of our future work in this pipeline, and with more patients we will be able to identify the full scope of such biases and correct them in a systematic way. Given the advantage of the network representation for interpreting results and identifying relationships between variations, we also see the advantage in implementing some network-based inference to complement our enrichment-based approach, in order to increase the quality of the rankings of pathways and likely driver mutations.

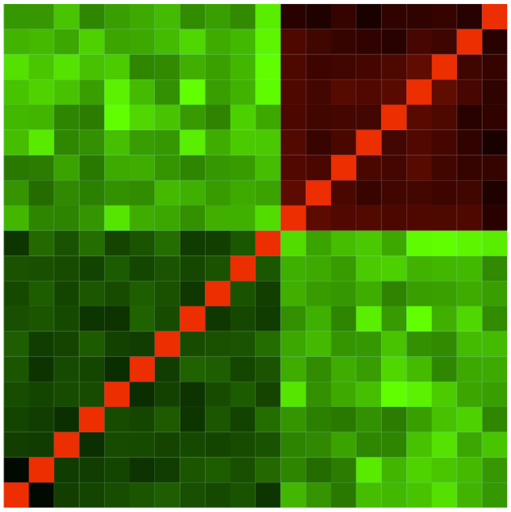


Figure 6.1: **Biases in sequencing technology are observed when patients are clustered using hierarchical clustering on the variants within each patient, two correlated clusters are observed which separate the sequencing technologies exactly**

Bibliography

[1] Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001.

[2] Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

[3] An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sept. 2012.

[4] An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Oct. 2012.

[5] Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, July 2012.

[6] The cancer genome atlas homepage, Sept. 2013.

[7] When can we expect the last damn microarray paper?, May 2014.

[8] C. M. Alberini. Transcription factors in long-term memory and synaptic plasticity. *Physiological reviews*, 89(1):121–145, Jan. 2009.

[9] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106+, Oct. 2010.

[10] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000.

[11] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208, July 2009.

[12] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach, Second Edition (Adaptive Computation and Machine Learning)*. A Bradford Book, second edition edition, Aug. 2001.

[13] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, June 2001.

[14] R. M. Barrett, M. Malvaez, E. Kramar, D. P. Matheos, A. Arrizon, S. M. Cabrera, G. Lynch, R. W. Greene, and M. A. Wood. Hippocampal focal knockout of CBP affects specific histone modifications, long-term potentiation, and long-term memory. *Neuropsychopharmacology*, 36(8):1545–1556, July 2011.

[15] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data setsupdate. *Nucleic Acids Research*, 41(D1):D991–D995, Jan. 2013.

[16] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7(1):389+, Aug. 2006.

[17] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706 – 716, 2008.

[18] M. M. Bellet, E. Deriu, J. Z. Liu, B. Grimaldi, C. Blaschitz, M. Zeller, R. A. Edwards, S. Sahar, S. Dandekar, P. Baldi, M. D. George, M. Raffatellu, and P. Sassone-Corsi. Circadian clock regulates the host response to Salmonella. *Proceedings of the National Academy of Sciences*, 110(24):9897–9902, June 2013.

[19] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[20] P. Bhanot, M. Brink, C. H. Samos, J. C. Hsieh, Y. Wang, J. P. Macke, D. Andrew, J. Nathans, and R. Nusse. A new member of the frizzled family from Drosophila functions as a Wingless receptor. *Nature*, 382(6588):225–230, July 1996.

[21] M. A. Bogoyevitch and B. Kobe. Uses for JNK: the Many and Varied Substrates of the c-Jun N-Terminal Kinases. *Microbiology and Molecular Biology Reviews*, 70(4):1061–1095, Dec. 2006.

[22] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan. 2003.

[23] S. Brakmann. Single-molecule analysis: A ribosome in action. *Nature*, 464(7291):987–988, Apr. 2010.

[24] J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, 36(suppl 1):D102–D106, Jan. 2008.

[25] N. Buisine and L. Sachs. Impact of genome assembly status on ChIP-Seq and ChIP-PET data mapping. *BMC Research Notes*, 2(1):257+, Dec. 2009.

[26] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical Report 124, 1994.

[27] F. Bushman. Bushman lab: Genelists, Sept. 2013.

[28] A. Callahan, J. Cruz-Toledo, and M. Dumontier. Ontology-based querying with bio2rdf's linked open data. *Journal of Biomedical Semantics*, 4(Suppl 1):S1, 2013.

[29] B. L. Cantarel, D. Weaver, N. McNeill, J. Zhang, A. J. Mackey, and J. Reese. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC bioinformatics*, 15(1):104+, Apr. 2014.

[30] M. V. Chang, J. L. Chang, A. Gangopadhyay, A. Shearer, and K. M. Cadigan. Activation of wingless targets requires bipartite recognition of DNA by TCF. *Current biology : CB*, 18(23):1877–1881, Dec. 2008.

[31] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. Wild. Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 11(1):255, 2010.

[32] G.-L. L. Cheng, W. Wang, H.-Y. Y. Wang, and Z.-G. G. Cui. Expression of EPOR on acute leukemia cells and its clinical significance. *Journal of experimental hematology / Chinese Association of Pathophysiology*, 19(1):15–18, Feb. 2011.

[33] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research*, 33(suppl 2):W72–W76, July 2005.

[34] U. Cobanoglu, M. Sonmez, H. M. M. Ozbas, N. Erkut, and G. Can. The expression of LMO2 protein in acute B-cell and myeloid leukemia. *Hematology (Amsterdam, Netherlands)*, 15(3):132–134, June 2010.

[35] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sept. 1995.

[36] G. E. Crooks, G. Hon, J.-M. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, June 2004.

[37] K. Daily, V. Patel, P. Rigor, X. Xie, and P. Baldi. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics*, 12(1):495+, 2011.

[38] M. D'Antonio, V. Pendino, S. Sinha, and F. D. Ciccarelli. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic acids research*, 40(Database issue), Jan. 2012.

[39] T. J. Daskivich, K.-H. Fan, T. Koyama, P. C. Albertsen, M. Goodman, A. S. Hamilton, R. M. Hoffman, J. L. Stanford, A. M. Stroup, M. S. Litwin, and D. F. Penson. Effect of Age, Tumor Risk, and Comorbidity on Competing Risks for Survival in a U.S. PopulationBased Cohort of Men With Prostate Cancer. *Annals of Internal Medicine*, 158(10):709+, May 2013.

[40] D. H. Davies, X. Liang, J. E. Hernandez, A. Randall, S. Hirst, Y. Mu, K. M. Romero, T. T. Nguyen, M. Kalantari-Dehaghi, S. Crotty, P. Baldi, L. P. Villarreal, and P. L. Felgner. Profiling the humoral immune response to infection by using proteome microarrays: High-throughput vaccine and diagnostic antigen discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):547–552, Jan. 2005.

[41] S. Davis and P. Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 14:1846–1847, 2007.

[42] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding. MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, 22(8):1589–1598, July 2012.

[43] A. L. Delcher, K. A. Bratke, E. C. Powers, and S. L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)*, 23(6):673–679, Mar. 2007.

[44] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. David: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[45] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, May 2011.

[46] L. Ding, T. J. Ley, D. E. Larson, C. A. Miller, D. C. Koboldt, J. S. Welch, J. K. Ritchey, M. A. Young, T. Lamprecht, M. D. McLellan, J. F. McMichael, J. W. Wallis, C. Lu, D. Shen, C. C. Harris, D. J. Dooling, R. S. Fulton, L. L. Fulton, K. Chen, H. Schmidt, J. Kalicki-Veizer, V. J. Magrini, L. Cook, S. D. McGrath, T. L. Vickery, M. C. Wendl, S. Heath, M. A. Watson, D. C. Link, M. H. Tomasson, W. D. Shannon, J. E. Payton, S. Kulkarni, P. Westervelt, M. J. Walter, T. A. Graubert, E. R. Mardis, R. K. Wilson, and J. F. DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, Jan. 2012.

[47] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Human molecular genetics*, 19(R2):R188–R196, Oct. 2010.

[48] I. A. Doytchinova and D. R. Flower. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC bioinformatics*, 8(1):4+, Jan. 2007.

[49] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.-M. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.-M. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.-F. F. Richard, M.-L. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J.-L. L. Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, July 2004.

[50] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18(suppl 1):S105–S110, July 2002.

[51] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, Jan. 2002.

[52] M. Egg, L. Köblitz, J. Hirayama, T. Schwerte, C. Folterbauer, A. Kurz, B. Fiechtner, M. Möst, W. Salvenmoser, P. Sassone-Corsi, and B. Pelster. Linking oxygen to time: the bidirectional interaction between the hypoxic signaling pathway and the circadian clock. *Chronobiology international*, 30(4):510–529, May 2013.

[53] E. Engvall and P. Perlmann. Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G. *Immunochemistry*, 8(9):871–874, Sept. 1971.

[54] M. Feng and Y.-C. C. Li. Expression of erythropoietin receptor in leukemia cells and relation of erythropoietin level with leukemic anemia. *Journal of experimental hematology / Chinese Association of Pathophysiology*, 16(6):1265–1270, Dec. 2008.

[55] R. A. Fisher. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507–521, 1915.

[56] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39(suppl 1):D945–D950, 2011.

[57] X. Fu, N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng, and P. Khaitovich. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, 10(1):161+, Apr. 2009.

[58] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, Mar. 2004.

[59] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[60] M. K. K. Goel, P. Khanna, and J. Kishore. Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, 1(4):274–278, Oct. 2010.

[61] L. Goff, C. Trapnell, and D. Kelley. *cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data.*, 2012. R package version 2.0.0.

[62] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas. IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082, Sept. 2013.

[63] E. A. Gordon, T. C. Whisenant, M. Zeller, R. M. Kaake, W. M. Gordon, P. Krotee, V. Patel, L. Huang, P. Baldi, and L. Bardwell. Combining docking site and phosphosite predictions to find new substrates: Identification of smoothelin-like-2 (SMTNL2) as a c-Jun N-terminal kinase (JNK) substrate. *Cellular Signalling*, 25(12):2518–2529, Dec. 2013.

[64] W. Gordon, Z. M, R. Herndon Klein, W. Swindell, H. Ho, F. Espetia, J. Gudjonsson, P. Baldi, and B. Andersen. A GRHL3-regulated repair pathway suppresses immune-mediated epidermal hyperplasia. *In submission*, 2014.

[65] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, July 2011.

[66] M. Greaves and C. C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, Jan. 2012.

[67] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140–D144, Jan. 2006.

[68] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl 1):D154–D158, Jan. 2008.

[69] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[70] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan. 2000.

[71] D. Hanahan and R. A. Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, Mar. 2011.

[72] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589, May 2010.

[73] A. Hoffmann and D. Baltimore. Circuitry of nuclear factor kappaB signaling. *Immunological reviews*, 210(1):171–186, Apr. 2006.

[74] T. W. Holstein. The evolution of the Wnt pathway. *Cold Spring Harbor perspectives in biology*, 4(7), July 2012.

[75] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, and B. Zhang. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, 4(6):1551–1561, June 2004.

[76] N. Hoverter, M. Zeller, M. M. McQuade, A. Garibaldi, A. Busch, K. J. Hertel1, P. Baldi, and M. Waterman. The TCF C-clamp DNA Binding Domain Expands the Wnt Transcriptome via Alternative Target Recognition. *In submission*, 2014.

[77] N. P. Hoverter, J.-H. H. Ting, S. Sundaresh, P. Baldi, and M. L. Waterman. A WNT/p21 circuit directed by the C-clamp, a sequence-specific DNA binding domain in TCFs. *Molecular and cellular biology*, 32(18):3648–3662, Sept. 2012.

[78] d. W. Huang, B. T. Sherman, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. David gene id conversion tool. *Bioinformation*, 2(10):428, 2008.

[79] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)*, 18 Suppl 1(suppl 1):S96–S104, July 2002.

[80] S. Impey, M. Davare, A. Lesiak, A. Lasiek, D. Fortin, H. Ando, O. Varlamova, K. Obrietan, T. R. Soderling, R. H. Goodman, and G. A. Wayman. An activity-induced microRNA controls dendritic spine formation by regulating Rac1-PAK signaling. *Molecular and cellular neurosciences*, 43(1):146–156, Jan. 2010.

[81] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, Apr. 2003.

[82] D. A. James, W. N. Venables, and B. D. Ripley. Modern Applied Statistics with S-PLUS. *Technometrics*, 38(1):77+, Feb. 1996.

[83] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*, 37(Database issue):D412–D416, Jan. 2009.

[84] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280, Jan. 2004.

[85] M. A. Kayala and P. Baldi. Cyber-T web server: differential analysis of high-throughput data. *Nucleic acids research*, 40(Web Server issue):W553–W559, July 2012.

[86] S. Kerscher, G. Durstewitz, S. Casaregola, C. Gaillardin, and U. Brandt. The complete mitochondrial genome of yarrowia lipolytica. *Comparative and functional genomics*, 2(2):80–90, 2001.

[87] A. H. Khodabakhshi, A. Fejes, I. Birol, and S. Jones. Identifying cancer mutation targets across thousands of samples: MuteProc, a high throughput mutation analysis pipeline. *BMC Bioinformatics*, 14(1):167+, May 2013.

[88] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36+, 2013.

[89] D. Kim and S. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12(8):R72+, Aug. 2011.

[90] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, 39(Database issue):D1035–1041, Jan 2011.

[91] A. G. Knudson. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, Apr. 1971.

[92] D. E. Knuth. Literate Programming. *The Computer Journal*, 27(2):97–111, Jan. 1984.

[93] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576, Mar. 2012.

[94] J. Korhonen, P. Martinmäki, C. Pizzi, P. Rastas, and E. Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, 25(23):3181–3182, Dec. 2009.

[95] D. P. Kreil, N. A. Karp, and K. S. Lilley. DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. *Bioinformatics (Oxford, England)*, 20(13):2026–2034, Sept. 2004.

[96] E. I. Kudryavtseva, T. M. Sugihara, N. Wang, R. J. Lasso, J. F. Gudnason, S. M. Lipkin, and B. Andersen. Identification and characterization of Grainyhead-like epithelial transactivator (GET-1), a novel mammalian Grainyhead-like factor. *Developmental dynamics : an official publication of the American Association of Anatomists*, 226(4):604–617, Apr. 2003.

[97] S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831, Sept. 2012.

[98] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Meth*, 9(4):357–359, Apr. 2012.

[99] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25–10, Mar. 2009.

[100] H. Li and R. Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[101] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug. 2009.

[102] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, Sept. 2011.

[103] D. C. Link, L. G. Schuettpelz, D. Shen, J. Wang, M. J. Walter, S. Kulkarni, J. E. Payton, J. Ivanovich, P. J. Goodfellow, M. Le Beau, D. C. Koboldt, D. J. Dooling, R. S. Fulton, R. H. Bender, L. L. Fulton, K. D. Delehaunty, C. C. Fronick, E. L. Appelbaum, H. Schmidt, R. Abbott, M. O'Laughlin, K. Chen, M. D. McLellan, N. Varghese, R. Nagarajan, S. Heath, T. A. Graubert, L. Ding, T. J. Ley, G. P. Zambetti, R. K.

Wilson, and E. R. Mardis. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *JAMA : the journal of the American Medical Association*, 305(15):1568–1576, Apr. 2011.

[104] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35(Database issue):198–201, Jan 2007.

[105] C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader. Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18):2347–2348, Sept. 2010.

[106] C. N. Magnan, M. Zeller, M. A. Kayala, A. Vigil, A. Randall, P. L. Felgner, and P. Baldi. High-throughput prediction of protein antigenicity using protein microarray data. *Bioinformatics*, 26(23):2936–2943, Dec. 2010.

[107] M. Magrane and U. Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009+, Jan. 2011.

[108] E. R. Mardis. The $1,000 genome, the $100,000 analysis? *Genome medicine*, 2(11):84+, 2010.

[109] E. R. Mardis. Genome sequencing and cancer. *Current Opinion in Genetics & Development*, 22(3):245–250, June 2012.

[110] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(Database issue):gkt997–D147, Nov. 2013.

[111] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–D110, Jan. 2006.

[112] C. R. Mehta and N. R. Patel. ALGORITHM 643: FEXACT: A FORTRAN Subroutine for Fisher's Exact Test on Unordered R&Times;C Contingency Tables. *ACM Trans. Math. Softw.*, 12(2):154–161, June 1986.

[113] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead, B. J. Raney, A. Pohl, V. S. Malladi, C. H. Li, B. T. Lee, K. Learned, V. Kirkup, F. Hsu, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, T. R. Dreszer, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, 41(Database issue):D64–69, Jan 2013.

[114] D. M. Minot, J. Voss, S. Rademacher, T. Lwin, J. Orsulak, B. Caron, R. Ketterling, A. Nassar, B. Chen, and A. Clayton. Image analysis of HER2 immunohistochemical staining. Reproducibility and concordance with fluorescence in situ hybridization of a laboratory-validated scoring technique. *American journal of clinical pathology*, 137(2):270–276, Feb. 2012.

[115] J. A. Mitchell, A. R. Aronson, J. G. Mork, L. C. Folk, S. M. Humphrey, and J. M. Ward. Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annual Symposium proceedings*, pages 460–464, 2003.

[116] J. B. Mitelman F and M. F. (Eds.). Mitelman database of chromosome aberrations and gene fusions in cancer, 2013.

[117] H. Mizuno, K. Kitada, K. Nakai, and A. Sarai. PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Medical Genomics*, 2(1):18+, Apr. 2009.

[118] M. Molenda, L. Mukkamala, and M. Blumenberg. Interleukin IL-12 blocks a specific subset of the transcriptional profile responsive to UVB in epidermal keratinocytes. *Molecular immunology*, 43(12):1933–1940, May 2006.

[119] A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic acids research*, 40(Database issue):D761–D769, Jan. 2012.

[120] P. J. Morin, A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein, and K. W. Kinzler. Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science (New York, N.Y.)*, 275(5307):1787–1790, Mar. 1997.

[121] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, July 2008.

[122] P. A. J. Muller and K. H. Vousden. p53 mutations in cancer. *Nat Cell Biol*, 15(1):2–8, Jan. 2013.

[123] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1):30–35, Nov. 2009.

[124] I. Nookaew, M. Papini, N. Pornputtpong, G. Scalcinati, L. Fagerberg, M. Uhlén, and J. Nielsen. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, 40(20):gks804–10097, Sept. 2012.

[125] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, A. Ruepp, and D. Frishman. The MIPS mammalian proteinprotein interaction database. *Bioinformatics*, 21(6):832–834, Mar. 2005.

[126] L. A. Panther, R. W. Coombs, S. A. Aung, C. dela Rosa, D. Gretch, and L. Corey. Unintegrated HIV-1 circular 2-LTR proviral DNA as a marker of recently infected cells: relative effect of recombinant CD4, zidovudine, and saquinavir in vitro. *Journal of medical virology*, 58(2):165–173, June 1999.

[127] S. Pepper, E. Saunders, L. Edwards, C. Wilson, and C. Miller. The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics*, 8(1):273+, 2007.

[128] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, Mar. 2011.

[129] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, Jan. 2012.

[130] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar. 2010.

[131] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[132] P. Radivojac, P. H. Baenziger, M. G. Kann, M. E. Mort, M. W. Hahn, and S. D. Mooney. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, 24(16):i241–i247, Aug. 2008.

[133] J. Reimand, O. Wagih, and G. D. Bader. The mutational landscape of phosphorylation signaling in cancer. *Scientific reports*, 3, Oct. 2013.

[134] C. S. Rex, L. Y. Chen, A. Sharma, J. Liu, A. H. Babayan, C. M. Gall, and G. Lynch. Different Rho GTPase-dependent signaling pathways initiate sequential steps in the consolidation of long-term potentiation. *The Journal of cell biology*, 186(1):85–97, July 2009.

[135] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, Jan. 2011.

[136] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140, Jan. 2010.

[137] I. T. Rombel, K. F. Sykes, S. Rayner, and S. A. Johnston. ORF-FINDER: a vector for high-throughput gene identification. *Gene*, 282(1-2):33–41, Jan. 2002.

[138] A. D. Roses. Pharmacogenetics and the practice of medicine. *Nature*, 405(6788):857–865, June 2000.

[139] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*, 24(3):227–235, Mar. 2000.

[140] RStudio, Inc. *Easy web applications in R.*, 2013. URL: `http://www.rstudio.com/shiny/`.

[141] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. N. Doudieu, V. Stümpflen, and H. W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(Database issue):D646–D650, Jan. 2008.

[142] M. Samwald, A. Jentzsch, C. Bouton, C. Kallesoe, E. Willighagen, J. Hajagos, M. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens. Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*, 3(1):19, 2011.

[143] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448, May 1975.

[144] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay, and K. H. Buetow. PID: the Pathway Interaction Database. *Nucleic acids research*, 37(Database issue):D674–D679, Jan. 2009.

[145] E. W. Schaefer, A. Loaiza-Bonilla, M. Juckett, J. F. DiPersio, V. Roy, J. Slack, W. Wu, K. Laumann, I. Espinoza-Delgado, S. D. Gore, and Mayo P2C Phase II Consortium. A phase 2 study of vorinostat in acute myeloid leukemia. *Haematologica*, 94(10):1375–1382, Oct. 2009.

[146] E. Schulte, D. Davison, T. Dye, and C. Dominik. A Multi-Language Computing Environment for Literate Programming and Reproducible Research. *Journal of Statistical Software*, 46(3):1–24, Jan. 2012.

[147] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8):1086–1092, Apr. 2012.

[148] G. L. Semenza. Hypoxia-inducible factors in physiology and medicine. *Cell*, 148(3):399–408, Feb. 2012.

[149] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, Jan. 2001.

[150] G. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31+, 2005.

[151] F. O. Smith. Personalized medicine for AML? *Blood*, 116(15):2622–2623, Oct. 2010.

[152] G. K. Smyth. Limma: linear models for microarray data. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420, 2005.

[153] M. Srivastava, E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, A. Salamov, M. L. Carpenter, A. Y. Signorovitch, M. A. Moreno, K. Kamm, J. Grimwood, J. Schmutz, H. Shapiro, I. V. Grigoriev, L. W. Buss, B. Schierwater, S. L. Dellaporta, and D. S. Rokhsar. The Trichoplax genome and the nature of placozoans. *Nature*, 454(7207):955–960, Aug. 2008.

[154] S. Srivastava and L. Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38(17):e170, Sept. 2010.

[155] C. Stark, B.-J. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database issue):D535–D539, Jan. 2006.

[156] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, Apr. 2004.

[157] Z. Su, Z. Li, T. Chen, Q.-Z. Li, H. Fang, D. Ding, W. Ge, B. Ning, H. Hong, R. G. Perkins, W. Tong, and L. Shi. Comparing Next-Generation Sequencing and Microarray Technologies in a Toxicological Study of the Effects of Aristolochic Acid on Rat Kidneys. *Chem. Res. Toxicol.*, 24(9):1486–1493, Aug. 2011.

[158] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, Oct. 2005.

[159] A. Takeshita, K. Shinjo, K. Naito, K. Ohnishi, M. Higuchi, and R. Ohno. Erythropoietin receptor in myelodysplastic syndrome and leukemia. *Leukemia & lymphoma*, 43(2):261–264, Feb. 2002.

[160] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, Mar. 2013.

[161] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotech*, 31(1):46–53, Jan. 2013.

[162] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, May 2010.

[163] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12):1248–1250, Dec. 2010.

[164] A. Valencia and M. Hidalgo. Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome Medicine*, 4(7):61+, July 2012.

[165] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834, Aug. 2008.

[166] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

[167] A. Vigil, R. Ortega, A. Jain, R. Nakajima-Sasaki, X. Tan, B. B. Chomel, R. W. Kasten, J. E. Koehler, and P. L. Felgner. Identification of the feline humoral immune response to Bartonella henselae infection by protein microarray. *PloS one*, 5(7), 2010.

[168] A. Vogel-Ciernia, D. P. Matheos, R. M. Barrett, E. A. Kramar, S. Azzawi, Y. Chen, C. N. Magnan, M. Zeller, A. Sylvain, J. Haettig, Y. Jia, A. Tran, R. Dang, R. J. Post, M. Chabrier, A. H. Babayan, J. I. Wu, G. R. Crabtree, P. Baldi, T. Z. Baram, G. Lynch, and M. A. Wood. The neuron-specific chromatin regulatory subunit BAF53b is necessary for synaptic plasticity and memory. *Nat Neurosci*, 16(5):552–561, May 2013.

[169] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, Mar. 2013.

[170] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, Jan. 2009.

[171] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr. 2004.

[172] K. Watanabe, J. Biesinger, M. L. Salmans, B. S. Roberts, W. T. Arthur, M. Cleary, B. Andersen, X. Xie, and X. Dai. Integrative ChIP-seq/Microarray Analysis Identifies a CTNNB1 Target Signature Enriched in Intestinal Stem Cells and Colon Cancer. *PLoS ONE*, 9(3):e92317+, Mar. 2014.

[173] G. A. Wayman, M. Davare, H. Ando, D. Fortin, O. Varlamova, H.-Y. Y. Cheng, D. Marks, K. Obrietan, T. R. Soderling, R. H. Goodman, and S. Impey. An activity-regulated microRNA controls dendritic plasticity by down-regulating p250GAP. *Proceedings of the National Academy of Sciences of the United States of America*, 105(26):9093–9098, July 2008.

[174] D. L. Weeks and R. G. Miller. Simultaneous Statistical Inference. *Biometrics*, 23(4):857+, Dec. 1967.

[175] J. S. Welch, P. Westervelt, L. Ding, D. E. Larson, J. M. Klco, S. Kulkarni, J. Wallis, K. Chen, J. E. Payton, R. S. Fulton, J. Veizer, H. Schmidt, T. L. Vickery, S. Heath, M. A. Watson, M. H. Tomasson, D. C. Link, T. A. Graubert, J. F. DiPersio, E. R. Mardis, T. J. Ley, and R. K. Wilson. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA : the journal of the American Medical Association*, 305(15):1577–1584, Apr. 2011.

[176] E. P. Wera, D. Armisen, K. Byrne, and K. Wolfe. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics*, 13(1):237+, 2012.

[177] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.*, 92(4):414–417, Oct 2012.

[178] T. C. Whisenant, D. T. Ho, R. W. Benz, J. S. Rogers, R. M. Kaake, E. A. Gordon, L. Huang, P. Baldi, and L. Bardwell. Computational Prediction and Experimental Verification of New MAP Kinase Docking Sites and Substrates Including Gli Transcription Factors. *PLoS Comput Biol*, 6(8):e1000908+, Aug. 2010.

[179] T. Wilanowski, A. Tuckfield, L. Cerruti, S. O'Connell, R. Saint, V. Parekh, J. Tao, J. M. Cunningham, and S. M. Jane. A highly conserved novel family of mammalian developmental transcription factors related to Drosophila grainyhead. *Mechanisms of development*, 114(1-2):37–50, June 2002.

[180] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, M. Niedergethmann, W. Weichert, M. Bahra, H. J. Schlitt, U. Settmacher, H. Friess, M. Büchler, H.-D. Saeger, M. Schroeder, C. Pilarsky, and R. Grützmann. Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. *PLoS Comput Biol*, 8(5):e1002511+, May 2012.

[181] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, 36(Database issue):D901–906, Jan 2008.

[182] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, 34(Database issue):D668–672, Jan 2006.

[183] C. Wu, I. Macleod, and A. I. Su. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic acids research*, 41(Database issue):D561–D565, Jan. 2013.

[184] X. Xie, P. Rigor, and P. Baldi. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, 25(2):167–174, Jan. 2009.

[185] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G. K. Wong, and J. Wang. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, pages btu077+, Feb. 2014.

[186] L. Xu, D.-c. Lin, D. Yin, and Koeffler. An emerging role of PARK2 in cancer. *Journal of Molecular Medicine*, 92(1):31–42, 2014.

[187] A. Youn and R. Simon. Estimating the order of mutations during tumorigenesis from tumor genome sequencing data. *Bioinformatics*, 28(12):1555–1561, June 2012.

[188] M. Zeller, C. N. Magnan, V. R. Patel, P. Rigor, L. Sender, and P. Baldi. A Genomic Analysis Pipeline and Its Application to Pediatric Cancers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2014. Accepted.

[189] Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, C. Nusbaum, R. Myers, M. Brown, W. Li, and X. S. Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137+, 2008.

Appendix A

Supplementary Information

A.1 Hardware requirements

Our pipeline is run on a 26 node 16-core 2.2 GHz AMD Opteron(TM) Processor 6274 cluster consisting of 48GB RAM per machine, running CentOS release 5.9 (Final).

Storage requirements per patient include approximately 400-500GB for storing the sequencing alignments for both the DNA-seq and the RNA-seq, and 5-10GB in output files including gene lists, variant details per transcript, final networks and reports. Additionally, approximately 100GB is needed for the necessary database files, etc., that are shared across all patients as well as those specific to each type of cancer.

The entire pipeline runs in approximately 2 days compute time per patient, distributed using Sun Grid Engine (SGE 6.2u5).

A.2 Software requirements

We present a list of all software packages we make use of in the pipeline, including versions and references where applicable.

Stand-alone packages:

- picard-tools (1.81)
- samtools (0.1.18) [101]
- casava (1.8.2)
- SCRATCH (1.0) [33]
- blast (2.2.26)
- MACS (1.4) [189]
- Bowtie (2.0.0-beta6) [98]
- TopHat2/TopHat-Fusion (2.0.3) [88, 89]
- cufflinks/cuffdiff (2.1.1) [162, 161]
- Trinity (2014-04-13p1) [65]
- GSEA (2.2.08) [158]
- MOODS (1.0) [94]
- PoSSuMsearch (1.3) [16]
- VarScan2 (2.3.6) [93]
- IGV (2.2.13) [160]
- Exonerate (2.2.0) [150]
- Weblogo (2.8.2) [36]

Collaborative interfaces:

- Cytoscape Web (1.0.4) [105]

- Apache (2.2.15)
- Django (1.4.1)
- Flask (0.10.1)
- Bootstrap (2.2.1)
- jQuery (1.8.3)
- fancybox (2.1.5)

Python (2.6.5) modules:

- networkx (1.7) [69]
- pybedtools (0.6.2)
- cybtpy [85]

R (2.15.0) packages:

- DAVIDQuery [78]
- GEOquery [41]
- affy [59]
- limma [152]
- shiny [140]
- MASS [166]
- ggbio
- cummeRbund
- biovizBase
- impute

A.3 Pediatric cancer

Table A.1: **Corresponding disease and tumor sample source, DNA and RNA sequencing data vendors for each patient. GHTF = UC Irvine GHTF.**

Participant	Gender	Diagnosed Disease	Tumor Sample Source	WGS Vendor	RNA
CHOC33	Female	Neuroblastoma	Third Rib Left Side	Illumina	Scripps
CHOC07	Female	Rhabdomyosarcoma	Soft Tissue, Posterior Tibia	C. Genomics	Scripps
CHOC20	Male	Atypical Teratoid Rhabdoid Tumor	Brain, Left Temporal	Illumina	Scripps
CHOC35	Male	Metastatic Osteosarcoma	Lung, Right Upper Lobe	C. Genomics	Scripps
CHOC18	Male	Hemophagocytic Lymphohistiocytosis	Blood	C. Genomics	Scripps
CHOC01	Male	Ependymoma	Brain Tissue	C. Genomics	Scripps
CHOC36	Female	Acute Myeloid Leukemia (AML)	Bone Marrow	C. Genomics	Scripps
CHOC11	Male	Rhabdomyosarcoma	Pleural, Left Pleural Base	C. Genomics	Scripps
CHOC43	Female	Acute Lymphoblastic Leukemia (ALL)	Cerebralspinal Fluid	Illumina	N/A
CHOC38	Male	Hodgkin's Lymphoma	Right Neck Mass	C. Genomics	Scripps
CHOC02	Female	Large Cell/Anaplastic Medulloblastoma	Brain Tissue, Posterior Fossa	C. Genomics	Scripps
CHOC10	Male	Myelodysplastic Syndrome	Bone Marrow	C. Genomics	Scripps
CHOC41	Male	Germ Cell Tumor	Medialstinal Mass	C. Genomics	Scripps
CHOC21	Male	Acute Lymphoblastic Leukemia (ALL)	Bone Marrow	Illumina	Scripps
CHOC23	Female	Ewing's Sarcoma, AML	Bone Marrow	Illumina	Scripps
CHOC08	Female	Clear Cell Sarcoma	Soft Tissue, Left Chest Wall	C. Genomics	Scripps
CHOC04	Female	Acute Lymphoblastic Leukemia (ALL)	Bone Marrow	C. Genomics	Scripps
CHOC30	Female	Medulloblastoma	Posterior Fossa	Illumina	Scripps
CHOC34	Male	Pilocytic Astrocytoma	Temporal Lobe/Intrav. Nodule	Illumina	Scripps
CHOC39	Female	Wilm's tumor	Left Kidney	C. Genomics	Scripps
CHOC13	Female	Ewing's Sarcoma	Left Popliteal Mass	C. Genomics	GHTF
CHOC09	Female	Melanoma	Soft Tissue, Right Buttock	C. Genomics	Scripps
CHOC28	Female	Osteosarcoma	Right Chest Wall Mass	Illumina	Scripps
CHOC29	Male	Hodgkin's Lymphoma	Periportal Lymph Node	C. Genomics	Scripps
CHOC26	Female	Acute Lymphoblastic Leukemia, AML	Bone Marrow	Illumina	N/A
CHOC25	Female	Embryonal Tumor with ANTR	Brain, R. Perieto-Occipital	Illumina	Scripps
CHOC24	Female	Atypical Teratoid Rhabdoid Tumor	Brain, Left High Parietal	C. Genomics	N/A
CHOC03	Male	Acute Myeloid Leukemia (AML)	Bone Marrow	C. Genomics	Scripps

Table A.2: GEO microarray datasets used

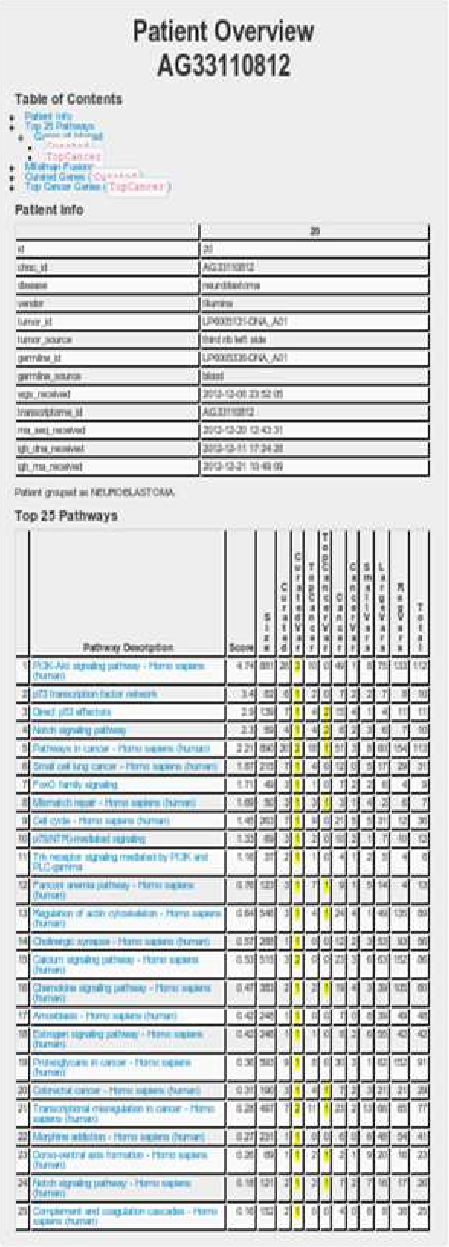
GEO ID	Samples
GSE10615	TESTICULAR (27)
GSE10899	ALL (4) AML (6)
GSE12417	AML (405)
GSE12453	HODGKINS (12) HODGKINS_CONTROL (10) HODGKINS_OTHER (45)
GSE12512	OSTEOSARCOMA (25) OSTEOSARCOMA_OTHER (12)
GSE12907	JPA (21) JPA_CONTROL (4)
GSE16254	NEUROBLASTOMA (88)
GSE1825	EWINGSSARCOMA (5) NEUROBLASTOMA (5)
GSE2223	JPA (2) JPA_CONTROL (4)
GSE22696	WT (26) WT_CONTROL (26)
GSE26050	HLH (11) HLH_CONTROL (33)
GSE2657	HODGKINS (2) SARCOMA (4)
GSE2712	CLEARCELLSARCOMA (14) CLEARCELLSARCOMA_CONTROL (3) WT (18)
GSE2719	SARCOMA (39) SARCOMA_CONTROL (15)
GSE27283	EPENDYMOMA (75) EPENDYMOMA_CONTROL (75)
GSE2779	MDS (13) MDS_CONTROL (15)
GSE28026	ATRT (18)
GSE30074	MEDULLOBLASTOMA (30)
GSE4587	MELANOMA (7) MELANOMA_OTHER (11)
GSE468	MEDULLOBLASTOMA (23)
GSE4698	ALL (60)
GSE8607	TESTICULAR (40) TESTICULAR_CONTROL (3)
GSE9476	AML (25) AML_CONTROL (39)
GSE967	EWINGSSARCOMA (11) RHABDOMYOSARCOMA (12)

Table A.3: Mitelman fusions identified in all patients

Patient ID	Author..Year	Morphology	Karyotype	Gene
CHOC11	Galili et al 1993	Alveolar rhabdomyosarcoma	t(2;13)(q36;q14)	PAX3/FOXO1
CHOC11	Gordon et al 2003	Pleomorphic rhabdomyosarcoma	t(2;13)(q36;q14)	PAX3/FOXO1
CHOC11	Shapiro et al 1993	Alveolar rhabdomyosarcoma	t(2;13)(q36;q14)	PAX3/FOXO1
CHOC23	Bain et al 1998	Myelodysplastic syndrome	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Goto et al 1994	Acute myelomonocytic leukemia	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Iida et al 1993	AML without differentiation	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Iida et al 1993	AML	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Iida et al 1993	Acute myelomonocytic leukemia	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Joh et al 1996	Acute myeloid leukemia	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Jun et al 2011	Acute myeloid leukemia	t(1;9;11)(p34;p21;q23)	MLL/MLLT3
CHOC23	Matsuo et al 1997	AML without differentiation	ins(11;9)(q23;p21p23)	MLL/MLLT3
CHOC23	Nakamura et al 1993	ALL/lymphoblastic lymphoma	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Nakamura et al 1993	Acute myeloid leukemia	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Negrini et al 1993	AML	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Poirel et al 1996	AML without maturation	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Shago et al 2004	AML	ins(9;11)(p21;q23q23)	MLL/MLLT3
CHOC23	Soler et al 2008	AML without maturation	ins(9;11)(p21;q13q23)	MLL/MLLT3
CHOC23	Super et al 1997	AML	t(9;11;13)(p21;q23;q34)	MLL/MLLT3
CHOC23	Voskova et al 2004	AML with minimal differentiation	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Yamamoto et al 1994	AML without maturation	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Yamamoto et al 1994	Acute myelomonocytic leukemia	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Yamamoto et al 1994	AML without differentiation	t(9;11)(p21;q23)	MLL/MLLT3
CHOC23	Yamamoto et al 1994	ALL/lymphoblastic lymphoma	t(9;11)(p21;q23)	MLL/MLLT3
CHOC08	Antonescu et al 2011	Malignant epithelial tumor, special type	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Dunham et al 2008	Angiomatoid malignant fibrous histiocytoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Friedrichs et al 2005	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Fukuda et al 2000	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Hallor et al 2007	Angiomatoid malignant fibrous histiocytoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Hansen Hallor et al 2005	Angiomatoid malignant fibrous histiocytoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Hiraga et al 1997	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Panagopoulos et al 2002	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Rossi et al 2007	Angiomatoid malignant fibrous histiocytoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Somers et al 2005	Osteogenic/bone tumor	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Speleman et al 1997	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Taminelli et al 2005	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1
CHOC08	Zucman et al 1993	Clear cell sarcoma	t(12;22)(q13;q12)	EWSR1/ATF1

Appendix B

Collaborative interface



Patient Overview
AG33110812

Table of Contents

- Patient Info
- Top 25 Pathways
  - Genes of Interest
    - Curated
    - TopCancer
- Mitelman Fusion
- Curated Genes (Curated)
- Top Cancer Genes (TopCancer)

Patient Info

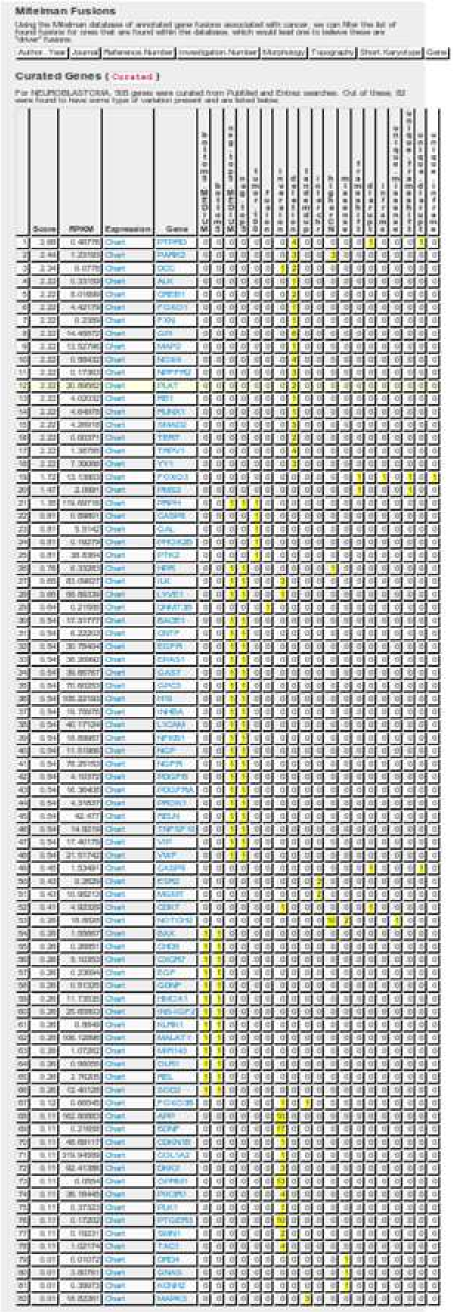
	20
id	20
chrc_id	AG33110812
disease	neuroblastoma
vendor	Illumina
tumor_id	LP6005131-DNA_A01
tumor_source	third rib left side
germline_id	LP6005336-DNA_A01
germline_source	blood
wgs_received	2012-12-06 23:52:05
transcriptome_id	AG33110812
rna_seq_received	2012-12-20 12:43:31
igh_dna_received	2012-12-11 17:24:28
igh_rna_received	2012-12-21 10:49:09

Patient grouped as NEUROBLASTOMA

Top 25 Pathways

	Pathway Description	Score	Size	Curated	CuratedVar	TopCancer	TopCancerVar	Cancer	CancerVar	SmallVars	LargeVars	RegVars	Total
1	PI3K-Akt signaling pathway - Homo sapiens (human)	4.74	351	26	3	10	0	49	1	8	75	133	112
2	p73 transcription factor network	3.4	82	6	1	2	0	7	2	2	7	8	10
3	Direct p53 effectors	2.9	139	7	1	4	2	15	4	1	4	11	11
4	Notch signaling pathway	2.3	59	4	1	4	2	6	2	3	6	7	10
5	Pathways in cancer - Homo sapiens (human)	2.21	390	20	2	18	1	51	3	8	60	154	112
6	Small cell lung cancer - Homo sapiens (human)	1.87	215	7	1	4	0	12	0	5	17	29	31
7	FoxO family signaling	1.71	49	3	1	1	0	7	2	2	6	4	9
8	Mismatch repair - Homo sapiens (human)	1.69	30	3	1	3	1	3	1	4	2	6	7
9	Cell cycle - Homo sapiens (human)	1.45	263	7	1	9	0	21	5	5	31	12	36
10	p75(NTR)-mediated signaling	1.33	69	3	1	2	0	10	2	1	7	10	12
11	Trk receptor signaling mediated by PI3K and PLC-gamma	1.16	37	2	1	1	0	4	1	2	5	4	6
12	Fanconi anemia pathway - Homo sapiens (human)	0.76	123	3	1	7	1	9	1	5	14	4	13
13	Regulation of actin cytoskeleton - Homo sapiens (human)	0.64	546	3	1	4	1	24	4	1	46	135	69
14	Cholinergic synapse - Homo sapiens (human)	0.57	285	1	1	0	0	12	2	3	53	93	56
15	Calcium signaling pathway - Homo sapiens (human)	0.53	515	3	2	0	0	23	3	6	63	152	86
16	Chemokine signaling pathway - Homo sapiens (human)	0.47	383	2	1	2	1	19	4	3	39	105	60
17	Amoebiasis - Homo sapiens (human)	0.42	248	1	1	0	0	7	0	6	39	49	45
18	Estrogen signaling pathway - Homo sapiens (human)	0.42	246	1	1	1	0	9	2	6	55	42	42
19	Proteoglycans in cancer - Homo sapiens (human)	0.38	503	9	1	8	0	30	3	1	62	152	91
20	Colorectal cancer - Homo sapiens (human)	0.31	190	3	1	4	1	7	2	3	21	21	29
21	Transcriptional misregulation in cancer - Homo sapiens (human)	0.28	487	7	2	11	1	23	2	13	66	65	77
22	Morphine addiction - Homo sapiens (human)	0.27	231	1	1	0	0	6	5	6	46	54	41
23	Dorso-ventral axis formation - Homo sapiens (human)	0.26	69	1	1	2	1	2	1	9	20	16	23
24	Notch signaling pathway - Homo sapiens (human)	0.18	121	2	1	2	1	7	2	7	16	17	26
25	Complement and coagulation cascades - Homo sapiens (human)	0.16	152	2	1	0	0	4	0	6	8	38	25

Figure B.1: **The top 25 affected pathways are displayed with summary information for a patient.**



Mitelman Fusions

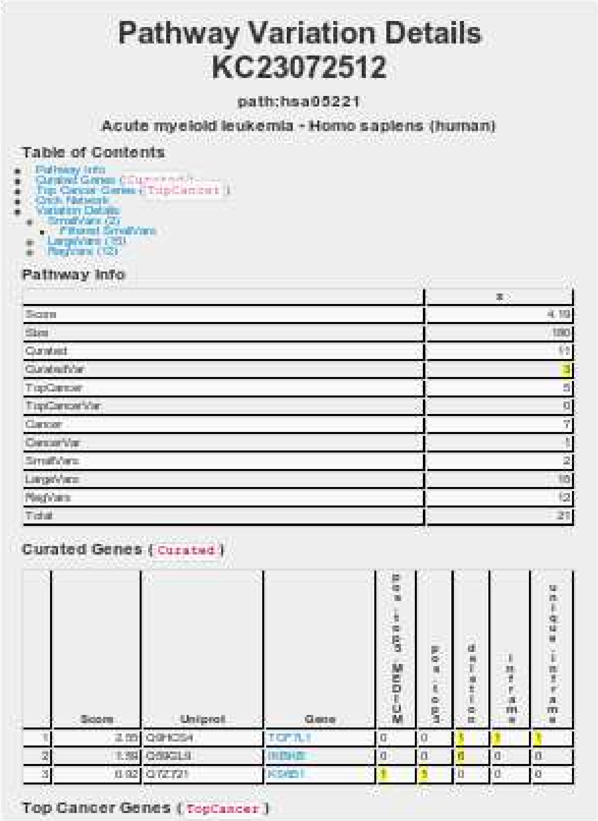
Using the Mitelman database of annotated gene fusions associated with cancer, we can filter the list of found fusions for ones that are found within the database, which would lead one to believe these are "driver" fusions.

Curated Genes (Curated)

For NEUROBLASTOMA, 305 genes were curated from PubMed and Entrez searches. Out of these, 82 were found to have some type of variation present and are listed below.

	Score	RPKM	Expression	Gene
1	2.88	0.48776	Chart	PTPRD
2	2.46	1.23193	Chart	PARK2
3	2.34	8.0778	Chart	DCC
4	2.22	0.33159	Chart	ALK
5	2.22	8.01686	Chart	CREB1
6	2.22	4.42179	Chart	FOXO1
7	2.22	0.2389	Chart	FXN
8	2.22	14.46572	Chart	GIR
9	2.22	13.52796	Chart	MAPT
10	2.22	0.59432	Chart	NEDD9
11	2.22	0.17363	Chart	NPFFR2
12	2.22	20.89562	Chart	PLAT
13	2.22	4.02032	Chart	RB1
14	2.22	4.64678	Chart	RUNX1
15	2.22	4.28918	Chart	SMAD2
16	2.22	0.68371	Chart	TERT
17	2.22	1.38798	Chart	TRPV1
18	2.22	7.30088	Chart	YY1
19	1.72	13.13803	Chart	FOXO3
20	1.47	2.1881	Chart	PMS2
21	1.35	119.69718	Chart	PRPH
22	0.81	0.89801	Chart	CASP8
23	0.81	5.5142	Chart	GAL
24	0.81	0.19279	Chart	PHOX2B
25	0.81	38.8394	Chart	PTK2
26	0.78	6.33283	Chart	HPR
27	0.65	83.05827	Chart	ILK
28	0.60	56.88339	Chart	LYVE1
29	0.54	0.21585	Chart	DNMT3B
30	0.54	17.31777	Chart	BACE1
31	0.54	6.22203	Chart	CNTF
32	0.54	30.79404	Chart	EGFR
33	0.54	36.26960	Chart	EPAS1
34	0.54	38.66767	Chart	GAS7
35	0.54	70.60250	Chart	GPC3
36	0.54	935.22193	Chart	H19
37	0.54	18.75575	Chart	INHBA
38	0.54	40.17124	Chart	L1CAM
39	0.54	18.89683	Chart	NFKB1
40	0.54	11.51866	Chart	NGF
41	0.54	78.25153	Chart	NGFR
42	0.54	4.10372	Chart	PDGFB
43	0.54	16.36485	Chart	PDGFRA
44	0.54	4.31637	Chart	PROK1
45	0.54	42.477	Chart	RELN
46	0.54	14.8219	Chart	TNFSF10
47	0.54	17.46179	Chart	VIP
48	0.54	21.51742	Chart	VWF
49	0.46	1.53491	Chart	CASP9
50	0.43	0.2629	Chart	ESR2
51	0.43	10.95213	Chart	MGMT
52	0.41	4.92325	Chart	CEBT
53	0.28	18.8528	Chart	NOTCH2
54	0.28	1.55867	Chart	BAX
55	0.28	0.28851	Chart	CHD5
56	0.28	5.10352	Chart	CXCR7
57	0.28	0.23694	Chart	EGF
58	0.28	0.51325	Chart	GDNF
59	0.28	11.13535	Chart	HMGA1
60	0.28	25.65852	Chart	INS-IGF2
61	0.28	0.5548	Chart	NUB1
62	0.28	506.12586	Chart	MALAT1
63	0.28	1.07262	Chart	MIR140
64	0.28	0.98055	Chart	OLR1
65	0.28	2.76205	Chart	REL
66	0.28	52.46128	Chart	SOD2
67	0.12	0.68545	Chart	FOXO3B
68	0.11	562.80983	Chart	APP
69	0.11	0.21692	Chart	BDNF
70	0.11	46.69117	Chart	CDKN1B
71	0.11	319.94589	Chart	COL1A2
72	0.11	92.41386	Chart	DKK3
73	0.11	0.0554	Chart	[illegible]
74	0.11	26.18446	Chart	[illegible]
75	0.11	0.37323	Chart	PLK1
76	0.11	0.17202	Chart	PTGER3
77	0.11	0.19231	Chart	SMN1
78	0.11	1.02174	Chart	TAC1
79	0.01	0.01072	Chart	[illegible]
80	0.01	3.80781	Chart	GNAS
81	0.01	0.35073	Chart	KCNH2
82	0.01	18.82381	Chart	MAPK3

Figure B.2: **The list of Mitelman fusions along with the ranked genes in the curated list. Yellow highlights denote variants present for that gene.**



Pathway Variation Details KC23072512

path:hsa05221

Acute myeloid leukemia - Homo sapiens (human)

Table of Contents

- Pathway Info
- Curated Genes (Curated)
- Top Cancer Genes (TopCancer)
- Crick Network
- Variation Details
  - SmallVars (2)
    - Filtered SmallVars
  - LargeVars (10)
  - RegVars (12)

Pathway Info

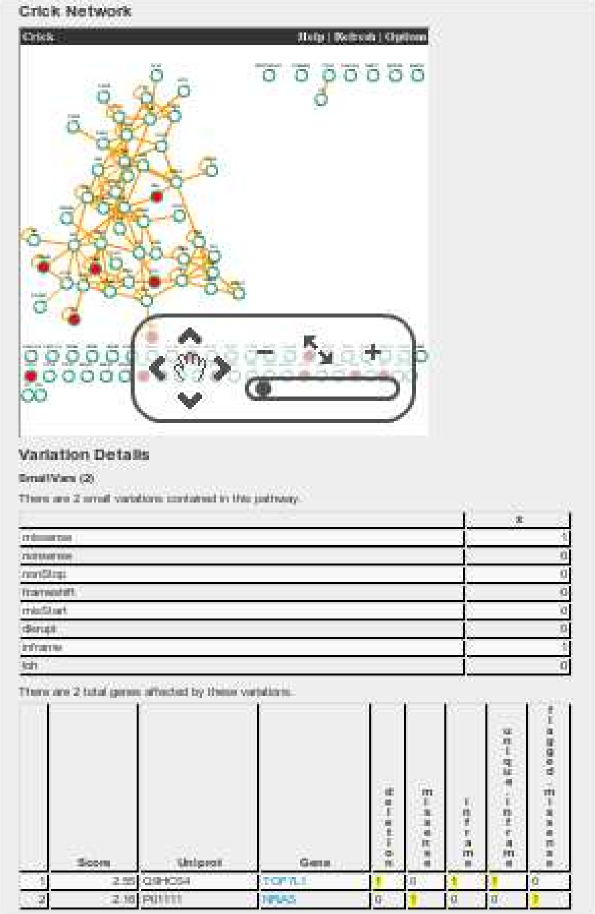
	x
Score	4.19
Size	180
Curated	11
CuratedVar	3
TopCancer	5
TopCancerVar	0
Cancer	7
CancerVar	1
SmallVars	2
LargeVars	10
RegVars	12
Total	21

Curated Genes (Curated)

	Score	Uniprot	Gene	pos_top3_MEDIUM	pos_top3	deletion	inframe	unique_inframe
1	2.55	Q9HCS4	TCF7L1	0	0	1	1	1
2	1.59	Q59GL9	IKBKB	0	0	6	0	0
3	0.92	Q7Z721	KSR1	1	1	0	0	0

Top Cancer Genes (TopCancer)

Figure B.3: **The list of variants within a single pathway can be obtained via a hyperlink in the top-level report. Each of the affected curated genes within the pathway are listed along with their variants.**



Crick Network

Variation Details

SmallVars (2)

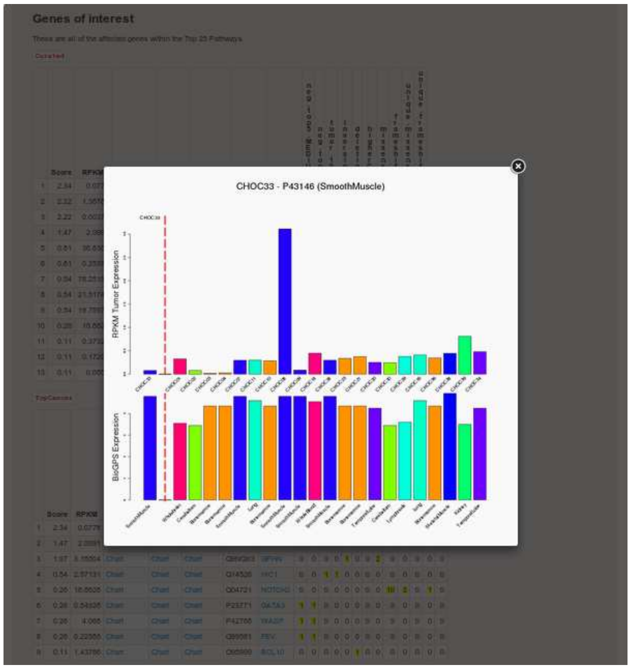
There are 2 small variations contained in this pathway.

	x
missense	1
nonsense	0
nonStop	0
frameshift	0
misStart	0
disrupt	0
inframe	1
loh	0

There are 2 total genes affected by these variations.

	Score	Uniprot	Gene	deletion	missense	inframe	unique_inframe	flagged_missense
1	2.55	Q9HCS4	TCF7L1	1	0	1	1	0
2	2.16	P01111	NRAS	0	1	0	0	1

Figure B.4: **An interactive network representation combining protein-protein, TF-protein, and drug-protein interactions for all top 25 pathways is generated and included in the report. Red nodes denote affected genes and details are obtained by clicking on each node.**



Genes of interest

These are all of the affected genes within the Top 23 Pathways

	Score	RPKM
1	2.34	0.077
2	2.22	1.3676
3	2.22	0.0037
4	1.47	2.088
5	0.81	36.630
6	0.81	0.2338
7	0.54	78.2510
8	0.54	21.5178
9	0.54	19.7597
10	0.28	10.660
11	0.11	0.3732
12	0.11	0.1720
13	0.11	0.000

	Score	RPKM																		
1	2.34	0.0776																		
2	1.47	2.0881																		
3	1.97	3.15004	Chart	Chart	Chart	[illegible]	[illegible]	0	0	0	0	1	0	0	2	0	0	0	0	0
4	0.54	2.57131	Chart	Chart	Chart	Q14526	HIC1	0	0	1	1	0	0	0	0	0	0	0	0	0
5	0.26	16.6626	Chart	Chart	Chart	Q04721	NOTCH2	0	0	0	0	0	0	0	0	10	2	0	1	0
6	0.26	0.54326	Chart	Chart	Chart	P23771	GATA3	1	1	0	0	0	0	0	0	0	0	0	0	0
7	0.26	4.066	Chart	Chart	Chart	P42768	WASP	1	1	0	0	0	0	0	0	0	0	0	0	0
8	0.26	0.22366	Chart	Chart	Chart	Q99581	FEV	1	1	0	0	0	0	0	0	0	0	0	0	0
9	0.11	1.43786	Chart	Chart	Chart	O95999	BCL10	0	0	0	0	0	1	0	0	0	0	0	0	0

Figure B.5: **Expression profiles for tumor (top) and normal (bottom) tissues is displayed with the patient on the far left as compared to all other patients on the right. Figures are generated on demand using the shiny package in R along with fancybox to implement the pop-up dialog.**

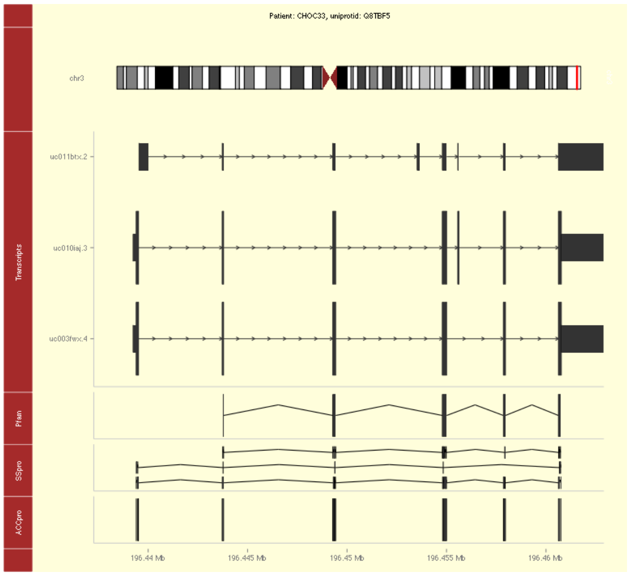


Figure B.6: **Details for each transcript are obtained on demand using the ggbio and biovizBase packages for R. Protein sequence features (solvent accessibility and secondary structure from SCRATCH and Pfam domains) are highlighted.**